Characterization and Mitigation of High-Confidence Errors Through the Use of Human-In-The-Loop Methods

Pavel Hoogland

Master Computer Science
Data Science and Technology Track
2021-2022

TU Delft

Characterization and Mitigation of High-Confidence Errors Through the Use of Human-In-The-Loop Methods

Domain Expert Driven Approach to Model Development

by

Pavel Hoogland

to obtain the degree of Master of Science MSc Computer Science
Track: Data Science and Technology
at the Delft University of Technology,
to be defended publicly on February 8, 2022.

Supervisors: Jie Yang, Oana Inel

External Supervisors: Margje Schuur, Jasper van Vliet

Student number:	4450892	
Thesis committee:	Prof. dr. ir. Geert-Jan Houben,	TU Delft, chair
	Dr. Jie Yang,	TU Delft, supervisor
	Dr. Lydia Chen,	TUDelft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Abstract

In the use of Machine Learning systems, attaining the trust of those that are the end-users can often be difficult. Many of the current state-of-the-art systems operate as Black-Boxes. Errors produced by these Black-Box systems, without further explanation as to why these decisions were made, will deteriorate trust. This effect is especially strong when these erroneous decisions are generated with high confidence. This thesis presents both a data-driven as well as a human-in-the-Loop based methodology to characterize and mitigate high-confidence errors. We propose an Iterative Expert Session based methodology. By engaging domain experts through a series of interaction sessions, we aim to reduce the disconnect and knowledge gap between data scientists and domain experts, and to ultimately increase trust in the model. A practical approach was taken working in close connection with the practice of the data scientists of the ILT, helping them in improving their model and providing a direct contribution. We study the problem in the context of Road and Transportation law violations, by engaging inspectors (i.e., domain experts) in day-in-the-life and in-house interview sessions. A thorough analysis is performed of the most important features for data instances that were in error with a high degree of confidence. A method is presented that helps in characterizing these errors by predicting errors. We show that by careful removal of biased data features, proper data selection and by bridging the knowledge gap between domain experts and data scientists, we can improve the performance of the machine learning model. We show an increase of model Precision from 0.56800 with a baseline of 0.32968 to a Precision of 0.52077 with a baseline of 0.23473. Considering the baseline, this is an increase of 28.9% in Precision. We reduce biases existent in the data by reducing variables that predict on inspector practice. The magnitude of High Confidence Errors in the top 20% errors went from 0.70435 to 0.70465 showing an improvement taking into account the reduced baseline and removal of overfitted variables.

Preface

This Thesis was written as part of the research during my internship at the Inspectie Leefomgeving en Transport (ILT), which is part of the Dutch Ministry of Infrastructure and Water Management.

During the course of my Master's programme in Computer Science I have gained an affinity for the interplay between humans and machines within the context of Human Driven Machine Learning. This peaked my interest for the topic of this thesis. My Research Questions were formulated together with and under the supervision of Dr. Jie Yang and Dr. Oana Inel at the Delft University of Technology as well as Margje Schuur at the ILT.

I want to thank Jie Yang and Oana Inel for supervising my thesis. They have always been supportive in guiding my thesis in the right direction. Without their help this thesis would not be the same. During the course of my internship I have had great help and support from my fellow colleagues at the IDLab of the ILT. I would like to thank Margje Schuur, Jasper van Vliet and Victor Ciulei for their continuous help in supervising my progress during the course of my thesis. Their knowledge and expertise helped me tremendously in my work. Additionally, for the practical side of things, working on the model and organising the expert sessions, I would like to thank Piet van Hoffen and Paul Ozkohen for their knowledge and insights.

I would like to thank and acknowledge the Road Transportation inspectors that participated together with us during the case study sessions: Gert Bosman, Willem van Dijk, Jan van der Laarse[†], Govert Pelkmans, Erik Rozenbrand, Bert van Voorthuizen, Björn Weekhout. I also want to thank Mark van der Ham, the team leader of the ILT Road and Transportation team, for helping schedule the sessions with the inspectors.

Last but not least, I would like to express my thanks to all my family and friends that supported me through the course of my thesis and through all the ups and downs of my work.

Pavel Hoogland
Delft, February 2022

Contents

1 Introduction

The use of Machine Learning(ML) systems in real world scenarios is ever increasing. These systems help us in making decisions based on large sets of data that are increasingly becoming available. Traditionally, the effectiveness of these Machine Learning systems has been measured in terms of accuracy. However, in recent years it has been shown to be too simple of a view when creating models for real-life scenarios. A good example of needing to go beyond mere accuracy is in the field of Content Recommender Systems [26]. These systems can be evaluated by how the end-users themselves experience the recommendations they produce. To properly suit the user's needs, the system should take into account the novelty or serendipity of certain items in order to increase their effectiveness.

Similar considerations should be directed to the area of predictive Machine Learning systems. As the sophistication and accuracy of these systems increase, so does their complexity. This increase in complexity makes these systems suffer from issues of trust. Many of these trust issues can mainly be attributed to their Black-Box nature. A Black-Box model is essentially a model of which the creator is only aware of the input and the output, while only having vague or no knowledge about the inner workings. A popular example of such a system is the Deep Neural Network, but there are many others like this, such as Random Forest models or XGBoost models. This means that when you work with such a model for classification it will be unknown as to what the prediction was based on and how it was made. This is especially detrimental for trust when the predictions are erroneous.

This thesis covers these problems in the context of a wide area of contexts and topics. Among others, the thesis touches on ML model development, Interpretable Machine Learning, High Confidence Errors, Domain Expert interactions with Data scientists as well as End-User Trust.

1.1. Interpretable Machine Learning

The challenges of understanding and explaining the choices that a Machine Learning model makes have long been an issue [43], and this has spawned a whole new field of Interpretable and Explainable Machine Learning. With new privacy and data regulations laws such as the GDPR [2] in 2018, the issue of interpretable systems has become an even more pressing one. Creators of high-stakes Machine Learning systems can be held accountable for the choices which the system makes [2, art. 21, 22]. However, the degree to which there is an actual 'right to explanation' is still a controversial topic, as you could make the case that the creators are only required to provide meaningful information such that the user can choose to opt-out from an automated decision making system. New proposed regulation [57] on Artificial Intelligence systems does cover the aspect of a user's right to interpretation of model output, but this is still in an early stage. Either way, for proper model use it is essential to understand the choices thoroughly. The issue of complex Black-Box models has become so troublesome in interpretation that some advocate for avoiding them entirely and using more interpretable methods instead [59]. Using simpler, more interpretable models is a common approach, however as Vaughan & Wallach stated: "... many of these techniques are based on commonly accepted assumptions about intelligibility, which can be wrong" [65, p.13]. Instead they call for researchers to leverage other human-centered fields such as social sciences to come to a better view of what is needed for interpretable systems. They state that instead we need a solid understanding of who the relevant stakeholders are, what their goals are, and how interpretability techniques can be designed to facilitate these.

1.2. User Trust

There is an ever-increasing need for trustworthiness of Machine Learning models, Chandler, Foltz, and Elvevåg [18] discuss the issues of trustworthiness of Machine Learning models in the field of psychiatry and call for discussion on new policies. They mention the need for greater transparency, generalizabilty and *explainability*. Thus we see that for these systems there exists a real lack of *transparency*, which is rife ground for issues of trust. Secondly, due to this low transparency, when errors or biases are detected, it is usually a difficult task to find the causes and have them fixed, i.e. there is no *accountability*. We can see that transparency, accountability, and explainability are interconnected issues.

Xiong et al. [69] point out that not each system requires the same level of trust. As an example, they make the distinction of the trust required for something as a Netflix ML algorithm as opposed to ML implemented in autonomous vehicles. They lay out 4 best design principles for system adoption and fostering user trust and propose further study into these. One of their most important principles is that of User Visibility. This factor means that the end-user's trust depends on how much insight they have in the system and how much control they have on the data that goes into it. Although they also make it clear that there is as of yet still very little study done into the issue of trust in the sphere of Machine Learning and security. Hengstler, Enkel, and Duelli [30], who studied trust in the field of autonomous vehicles, found that early and proactive communication, concrete and tangible information, and transparency in the development process are all of great importance in the pursuit of achieving the end user's trust.

1.2.1. High Confidence Errors

To give a better indication of how much u can trust a decision made by a Machine Learning system, you can provide a measure of confidence of the prediction. This way, the user can be more certain of how seriously it should take the prediction. With errors that happen when the system has low confidence in its prediction, the user will judge more leniently. However, only a few wrong predictions that are generated with a high level of confidence may already have a severe impact on the trustworthiness of the model. The psychological concept of Negativity Bias is a well studied one [15, 4]. This shows that people have a tendency to remember and linger much more on negative events rather than positive ones. This means that it is vital to reduce the impact of these events. Such errors produced with a high confidence are commonly rooted in the incompleteness of the model, and arise due to the biases in the training data. The issue of bias is well illustrated in Lakkaraju et al. [38] where an example is shown of an image classification task where the training data is comprised only of images of black dogs and of white and brown cats. The model trained on this data will label a white dog as a cat with high confidence. These errors are challenging to detect because the system itself can not give us any useful information.

1.2.2. Disconnect

Trust and proper interaction with the *domain experts* in a field and the *data scientists* is often a difficult and interconnected issue. Thus we can see problems arising as a consequence of a disconnect between *data scientists* and *domain experts* which is often present when using machine learning systems [66]. This disconnect can be present on different levels, i.e., the concept or the process (how-to) level [51, 19, 21] and can lead to decreased *domain expert* trust in the system.

Providing model outcomes to experts without insights into why the (potentially erroneous) predictions were made could harm user trust even further and inhibit user-developer interactions. Understanding why certain predictions are made is key to user engagement, system adoption, and sustainability [62].

The Data Scientists that design the model can usually only have their assumptions and hypotheses on the underlying reality of a domain, which entirely depends on their own knowledge of the domain. This makes this gap in knowledge a bottleneck in the process of developing a good model. Therefore, the data scientists lack a lot of the vital domain knowledge that the domain experts do have. In order to make the model fit the practice of the domain experts, there is a need for ways to transfer the domain expert knowledge to the data scientist and consequentially to the model. In Section 3.2 of the Background we will go into further detail into ways that this has previously been studied.

In the real world these models are mostly meant to aid the decision making of real-life experts, whether it be medical professionals, legal professionals, or, as in our case, inspectors. It is thus essential to not only tunnel vision on the data but to also focus on the needs of the user base. Combining all data into one risk model may make sense from a purely data centered approach, as more vehicles inspected can serve for a higher amount of data and thus it can improve the accuracy of the model, however this way you are merely

generalizing over the data rather than facilitating the inspectors in their work.

1.3. Human-In-The-Loop Machine Learning

A concept increasing in popularity is that of Human-in-the-loop Machine Learning, which is fundamentally the process of leveraging the knowledge of both machine and human intelligence in the process of creating machine learning models. Holzinger [34] shows that in the health-informatics domain, where data is not always abundant, an interactive machine learning approach with a 'doctor-in-the-loop' development cycle can be better than automated machine learning models. Classic development cycles of machine learning systems can be tedious and with little results. Xin et al. [68] discuss some of the challenges and give opportunities for making the cycle more efficient. Challenges of development workflows include the reuse of intermediate results, as this can often be a difficult process especially for novice data scientists. This leads to the cumbersome rerunning of the entire workflow and model, which can take a lot of time. Also the impact of certain changes on the performance of the model is not always entirely clear. Using an optimized Human-In-The-Loop development cycle with intermediate result reuse and impact assessment of changes, they try to tackle some of these challenges. Although they show this solution to be promising in speeding up the machine learning development workflow, it still only leverages the knowledge of the data scientists themselves.

1.4. Research Context

Now that we have outlined the problem of our research, we will discuss the context in which this research was done. We work on this problem in the context of an internship at the Netherlands Human Environment and Transport Inspectorate (ILT)[1]. Here we work on a predictive classification model that predicts for the risk of law violations for the Road and Transportation Inspectors.

1.4.1. Vehicle Violation Inspections

The Road and Transportation inspectors have as a task to make sure that all trucks driving on the Dutch roadways adhere to the laws that are in place for these vehicles. Those vehicles that do not adhere to a certain law are referred to as Non-Compliant, or Violator. With the thousands of company vehicles that drive on the Dutch roadways each day, the inspectors have the difficult job to select those that are at high risk of being in violation of any laws. Their criteria for selection are based on, for the most part, visual characteristics of the vehicle and/or the driver. Using their knowledge and experience, they in real-time and on-site make decisions on whether a vehicle is to be selected for inspection or not.

In order to aid the inspectors with this task and augment their experience, the ILT has set out to develop a Road and Transportation risk classification model. This model serves to assist the inspectors in those areas where data is available but not directly available to the inspectors. Based on this data, we can provide the inspectors with a score indicating the risk of non-compliance of the vehicle.

These inspectors are experienced professionals, sometimes working at the ILT already for decades, this means that they prefer to rely entirely on their experience. This can be an issue when we want to introduce new technologies to them especially if they are based on difficult to grasp concepts such as is the case with Machine Learning. When we are dealing with technologies that might be wrongly interpreted as trying to replace the inspectors or telling them how to do their work, the mistrust that this creates can only increase. Rather, the technology should be there to aid the inspectors by increasing the precision of their inspections. We need to gradually make the inspectors accustomed to these technologies such that they can clearly understand the use cases and the advantages that the model provides.

1.4.2. The Problem

We now come to the problem of our study. Essentially, the problem we have is two-sided. On the one hand we are working with a model that we want to be as accurate as possible. This comes with plentiful of intricacies such as choice of model, choice of data, and even issues of data bias. This choice of data and data biases can then be a cause for high confidence errors to occur.

On the other hand, we have the human factor. As many Machine Learning models are being used as decision making aids for end-users, these end-users should be willing to use the model and trust it. Since we believe the model to be beneficial for the work of the ILT we want the inspectors to trust the model that we are providing and encourage the inspectors to use it. However, as is often the case with use cases for machine

[1] Inspectie Leefomgeving en Transport (ILT), Ministerie van Infrastructuur en Waterstaat: `https://www.ilent.nl/`

learning systems, there exists a disconnect between what we as the data scientists know and what the domain experts (i.e. inspectors) know and do in their practice.

Working on this problem brings with it various challenges, it can be a difficult task to exactly know how complex models will react to changes to the data, which makes characterizing those areas where improvement is necessary particularly arduous. On top of this we have the human component. Introducing humans into the development of a system is never an easy task, as all might have differing opinions on topics both in their expertise as well as outside their expertise. Therefore, using their knowledge should be done in a careful and concise way.

To work on these challenges, we will both provide a careful analysis of the model and the data, as well as outline a methodology to interact with the experts.

1.4.3. Practical Approach

During the course of this thesis as part of the internship of the ILT a more practical approach was taken over a purely theoretical one. During the internship a strong focus was put on helping the ILT with their Risk Classification model as well as helping them in their interactions with the Domain Experts and how to improve these interactions. One of the core aims of this thesis was to build a bridge between the data scientists practice and the practice of the Inspectors. The work that was done was continually put into practice for the actual model and for the experts. Data biases were made clear from the findings of this thesis, which helped in improving the risk classification model. This also paves a new starting point for the ILT and how they are looking to deal with data bias issues into the future.

HCOMP Paper During the course of the thesis, to complement and present the work that has been done, a 2-page Work-In-Progress paper was written focusing on the domain expert based exploration of model errors[35]. This paper focuses only on the domain expert driven part of this thesis and can be found in appendix D.

1.5. Research questions:

We see that the issues of knowledge gaps, lack of trust in the model and high confidence errors are interconnected problems in Machine Learning that are difficult to overcome. This research explores these problems using both a data- and human-driven approach. Thus, we state the main research question for this thesis as the following:

> *How can we best characterize and mitigate predictive errors that are produced with a high model confidence?*

In this research the important distinction is that of "What the model has learned vs. What the model should know" this leads us to the following sub research questions:

- RQ 1: *What Instances Best Characterize System Knowledge?*

To answer this question, we need to perform a data driven analysis of the model itself. Through this analysis, we will get a better idea of what features are important for the model, what features can cause issues and what instances resemble what the model has learned best.

- RQ 2: *How to best interpret what the model has learned?*

Answering this research question means we have to perform a literature study on what ways of interpreting the model would be best. Then applying what we have learned from the study to our model, combined with the feedback from the experts, helps us to better answer this question.

- RQ 3: *What knowledge do we want the model to have?*

This research question we answer by involving the domain experts. By doing the Expert Sessions with the domain experts we want to better our knowledge of what the model should know.

- RQ 4: *What does the model not know?*

Finally, this question will essentially be answered by combining what we have learned in the process of answering RQs 1 and 3. By combining these two we can get an idea of the areas where the model is still lacking and where it needs improvement.

1.6. Contributions

This thesis explores the domains of High Confidence errors and human-driven development by studying the effect of iterative experts' engagement in a series of interaction sessions. It serves as an exploration for aspects, among others, of data bias, machine learning errors, human-driven design, user trust and model interpretability. Among others, some of the most noteworthy contributions that this thesis presents are the following:

- A novel generalized methodology through the use of Iterative Expert Sessions that has as an aim to close the knowledge gap between data scientists and domain experts.
- A data-driven methodology for characterizing High Confidence Errors. This allows us to better characterize High Confidence Errors that are predicted both as False Positives and as False Negatives.
- The sessions performed during the thesis aid us in understanding the requirements and challenges, and help us in characterizing:
  - Model Errors
  - False Positive HCEs
  - False Negative HCEs
  - Correctly predicted instances

  We show these sessions to be a promising method to facilitate model development and build trust in the model.
- Through the direct involvement of domain experts in the process of building the model, we provide an interactive modeling method to reduce the presence of data biases.
- Scientific contribution through the submission of a paper for HCOMP 2021: The Ninth AAAI Conference on Human Computation and Crowdsourcing[35]
- Direct Practical contribution to the ILT by improvement of the model and lessons learned from the sessions. This provides a direct contribution to the ILT and their model with the following:
  - Improvement of the model performance with regards to the baseline performance.
  - Creation of a general expert interaction model for the ILT's interactions with their domain experts
  - Reduction in features biased on inspection practice such as Location and Time.
  - Making the experts of the ILT aware of feature bias issues, providing a new starting point for the organization.

1.7. Thesis Overview

In the next section 2 we elaborate on the context in which the research has been performed under the guidance of the ILT, we also discuss the workings of the model that was developed during the study and the data that was used for training and validating it. Following this in section 3 we give an overview of the most important concepts to grasp and the previous related work that has been performed in this area of research. We then give an overview of the data driven methodology in chapter 4, followed by the methodology of the iterative expert sessions in chapter 5. Next in section 6 there will follow an extensive analysis of the domain and system knowledge by a data driven analysis of the model and the important data variables. It will include an overall model performance analysis, feature explanations, High Confidence Error analyses and model bias studies. In chapter 7 we give a validation of our expert session methodology by providing the results from the sessions performed with the inspectors. Using the findings from these sessions as well as the insights from chapter 6, in chapter 8 we present the changes that were made to the model and how these changes improved the model. We finish the thesis with a conclusion covering the main conclusions, recommendations, limitations and future work.

2

Research Context and Model

This section covers the context in which this research was performed. We first outline the context and motivation of the use of the model and the research. Then, we motivate the choice of model, explain the confidence measure and the evaluation metrics used, and lastly we give a description of the dataset used for the training of the model.

2.1. Context

The Research is performed in the context of the work of the Inspectie Leefomgeving en Transport (ILT). This is the Human Environment and Transport Inspectorate of the Dutch Government. The ILT has as its job, ensuring road and transportation safety while also providing a level playing field on the market and protecting companies and drivers from unfair circumstances or bad working conditions. In doing this, the ILT aims to have an effective approach where only those owners that are law non-compliant (violators) are punished while law compliant owners are minimally impacted by the ILT.

The core laws and violations that the inspectors of the ILT check for are the following:

- **Vehicle Licenses:** All vehicles that are transporting goods on the Dutch Roadways have certain licenses they legally have to carry, depending on what goods they transport and what route they make.
- **Rest and Driving times:** All drivers of transport vehicles have a legally determined amount of hours that they can drive until they must have rest time. This is to ensure safe road conditions and fair competition between companies.
- **Manipulation of Tachograph:** A Tachograph is a small device fitted inside of a truck or vehicle that records all of the vehicle's speed and distance travelled. Sometimes these can be manipulated to be able to drive for longer amounts of distances.
- **Overweight of the vehicle:** To ensure safe road conditions, each vehicle has a maximum amount of load it is allowed to carry.
- **Transport of dangerous goods and substances(ADR):** Dangerous goods and substances can pose a threat to health and environment, which means that there is a large amount of rules and regulations to adhere to.

From the regulatory bodies of the European Union there has been a call for the wider adoption of digital technologies to promote road safety. An example of this is the introduction of the use of Smart Tachographs [1]. As part of the EU Commission Road safety policy framework [53] the ILT is also required to adhere to road safety standards, which includes making use of technologies such as risk assessment ratings.

In 2019, as part of the Smart and Safe programme the Innovation and Data Lab (IDLab), which is a branch of the ILT working on innovative and data driven technologies, started on the development of a risk classification system for road transport safety and fairness. There is a clear need for more advanced data-driven models for the inspectors of the ILT. These inspectors traditionally work mostly based on inspection experience and decision rules. The risk model is developed by the data scientists at the IDLab.

As this is a novel technology introduced to the inspectors, the research of this thesis is particularly relevant for the IDLab as it aims to help in generating trust and encouraging the smooth adoption of the model.

2.2. Risk Model

The risk model that is used during the course of this thesis is the predictive Road and Transport risk classification model developed by the data scientists at the ILT. The choice of model is a Random Forest based model. Random Forest is a popular choice for models used by the ILT due to its good characteristics, some of these include great versatility with various types of data, good performance with high dimensionality datasets, easily parallelizable and being relatively easy to understand. Here we explain how the Random Forest algorithm works. As the Random Forest belongs to the category of Tree-based models, we first explain Tree-based models.

2.2.1. Tree-Based Models

Tree-based models work by the use of a series of if-then decision rules that are generated from training data. These rules can then be used to generate predictions. These decision rules are created in the form of a branching tree structure. Tree-based models are particularly strong when used in an ensemble learning setting where many (weak learner) trees are combined to get better results, such as Random Forests and Gradient Boosting Machines(e.g. XGBoost).

Both Random Forest and Gradient Boosting methods work by combining decision trees, where the former does this only at the end as an average(Bagging) while the latter combines the trees sequentially(Boosting). Gradient Boosted trees in general show better performance than Random Forests [17]. However, Random Forests can be more flexible as well as better resistant to noisy data.

The classic implementation of Decision Trees that can be used for classification purposes were first proposed by Breiman et al. [14]. The tree structure has a root node and leaf nodes. The leaf nodes indicate the final model decision, so in the case of binary classification either a 1 or 0. The non-leaf nodes indicate decisions of the model based on feature criteria. Each non-leaf node makes a decision how to split based on the feature value. For the classical [14] implementation, the splitting criterium is calculated based on Gini Impurity. The formula for the Gini impurity is given in equation 2.1:

$$Gini = \sum_{i=1}^{C} \_i(1 - p_i) \tag{2.1}$$

Here C is the total amount of classes and p_i is the probability of class i. Thus for a binary two class scenario this is simply:

$$p_0(1 - p_0) + p_1(1 - p_1)$$

The Gini impurity is essentially a metric for the probability of incorrectly classifying a randomly chosen datapoint when randomly classifying the instance according to the class distributions. For each class (i.e. leaf-node), the Gini impurity is calculated. For the splitting node the final Gini impurity is then the weighted average of the two Gini impurities calculated. The Gini impurity obtained will then be compared to the impurity of the parent node. The aim is to have a decrease in Gini impurity that is as substantial as possible. Meaning that the splitting of the features must be chosen such that you obtain the largest decrease in Gini Impurity.

The use of tree-based models is popular in both classification as well as regression settings. The relative simplicity of these models yet good performance makes them a desirable choice for models. These models are flexible in their use as they can be applied to numerous types of data. They can deal with sparse data, dense data, discrete data, continuous data, while also being able to deal with missing data points.

Though simple to explain, simple decision trees are not very robust, the model is quite naive in its use and small changes in the dataset used can already cause the entire tree to have to be changed. Thus, most applications prefer to use them as part of a more complex model, a popular implementation being the Random Forest algorithm.

2.2.2. Random Forest

The Random Forest [13] is what is known as an Ensemble method. Here essentially, to make the model more stable and powerful, you take many decision trees and combine their result into one model. This is done using a technique called Bagging, which stands for Bootstrapping and aggregating. For bootstrapping you take your original sample and repeatedly draw random samples from this set. You can do this until you have the same number of random samples as items in your original sample set. The aggregating part is that you take the results from many different decision trees, each trained on their own bootstrapped set, and then

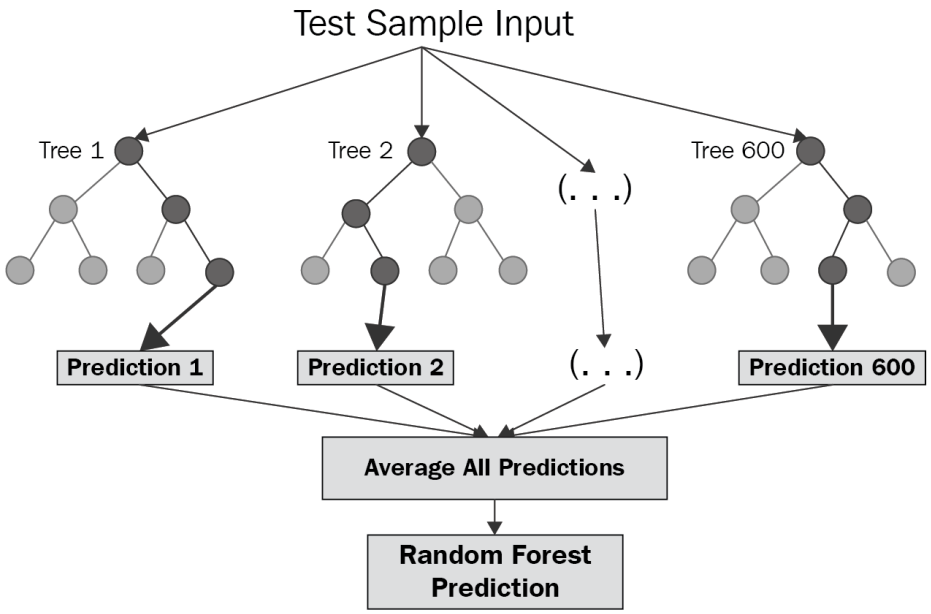


Figure 2.1: Simplified representation of a Random Forest model (Source: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454 by Afroz Chakure)

.

combine these results into a weighted final result. A simplified representation of how the random forest works is shown in figure 2.1

The choice of how many trees to aggregate over and how many bootstrap samples to take can have a big effect on each model. The number of trees used can be set and tuned as a hyperparameter. Even though with more trees accuracy could increase, so can storage and computation time requirements. Additionally, using too many trees has shown to sometimes cause unnecessary increases in variance [29, p. 596]. This use of bagging also consequently increases the complexity of the model, making it less explainable.

Another issue with this approach is that if you use all the same features on each tree, they will be heavily correlated with each other. To solve this issue, we can take for each tree a random subset of only m features to use for the training and splitting, this is a hyperparameter and is often also referred to as *mtry*. It has been shown [29] that a common good choice for *mtry* is close to the square root of the total amount of features. Note that this optimal choice of mtry differs when the random forest is used for regression instead of classification.

The model for the ILT is created using the R package of randomForest [42] which is used in combination with the caret [37] package which is a useful package that streamlines the pipeline of training for classification and regression purposes.

2.3. Model Confidence

In order to give a better sense of the reliability of predictions, it is useful to accompany these with a confidence score. Most commonly this confidence score is represented as a percentage score from 0 to 1. For the Random Forest model we use a probability based confidence score. As we can not know the real class probabilities of the classes, we need a way to estimate them. The simplest method using a Random Forest is to estimate based on the proportion of the tree ensemble that predicts a certain class. This proportion is the total number of votes of all trees that predict the specific class. This proportion then forms the probability score for the prediction. There are other confidence metrics besides the simple probability score. However, Bakker [7] has shown that for more complex models with high overall accuracies, such as a Random Forest, the probability score can not be improved upon by using more complex confidence metrics such as the conformal prediction framework presented in Bakker [7].

Alternatively, other methods exist for estimating the probability. A different approach to estimation is using the random forest in a regression setting, returning a value between 0 and 1 as output. Malley et al. [50] show that for any non-parametric statistical method, this regression method, also referred to as a Probability

Machine, returns consistent probability estimates. Another method is using the Out-of-Bag samples to estimate the probability [12]. Due to the fact that when using bootstrapping you can choose the same samples more than once, the Out-of-Bag samples are those samples in the original dataset which were not included in the bootstrapped set. Using these samples, you can calculate the Out-of-Bag Error (see section 2.4.2). All these have been shown to perform fairly similarly, with, on some datasets, the Out-of-Bag method performing better [41, 29]. As the focus of this thesis is not to compare the effectiveness of probability estimations and confidence scores but rather to use these probabilities as a tool for analyzing issues with the data we choose to stick with the simplest method of the tree majority vote.

2.4. Evaluation Metrics

Once you have a model that works satisfactory you want to have a way to measure its performance. In order to evaluate the performance of a model you need to have metrics to evaluate them on.

2.4.1. Precision and Recall

To measure the performance of a predictive model popular metrics are those of Precision and Recall. Loosely speaking Precision is a metric showing the quality of predictions and Recall the quantity. Precision is essentially a measure of what number of positively classified items are actually positive. In our case this means what fraction of vehicles selected for being high risk were actually in violation. The equation for Precision can thus be written as in 2.2:

$$\frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2.2}$$

True Positives and False Positives are calculated by comparing the true label of the instances with the predicted label given by the model. If the model correctly predicts a positive class, this will be a True Positive. In our case, the positive class is the violator classification. This means that if the model predicts the instance to be a violator but the true label is a compliant vehicle, this will be a False Positive. If the opposite is the case, meaning it predicts as a compliant vehicle but the vehicle was in violation, this would be a False Negative.

Another Metric that is often used in combination with Precision is the Recall. The Recall measures the amount of true positives out of the set of all relevant items. In our case this means the total proportion of violating vehicles that are selected out of the set of all vehicles that are in violation. The equation for Recall can be written as in 2.3:

$$\frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2.3}$$

A summary of these metrics and their relation are also shown and explained in figure 2.2.

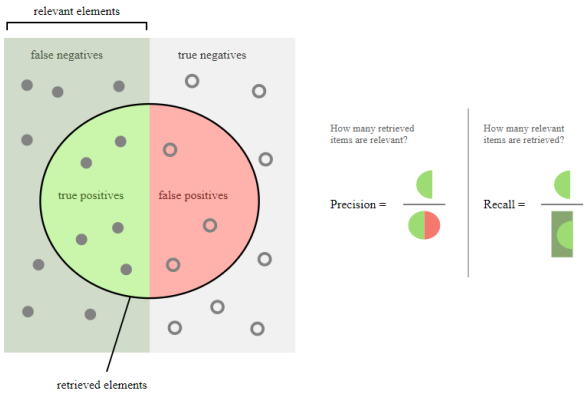


Figure 2.2: Precision and Recall (Edited, https://upload.wikimedia.org/wikipedia/commons/2/26/Precisionrecall.svg by Walber, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

.

For our model we optimize based on Precision. Since the aim of the model is to have as large as possible a proportion of non-compliant vehicles out of all vehicle that are selected, optimizing for the precision is the best way to achieve this. The time of the road inspectors is valuable, they can only inspect a small amount of vehicles per day, meaning that each vehicle that has been selected has to matter. If the vehicle was compliant

after all, the inspectors have spent valuable time and resources and achieved no improvement in road safety. You could make the argument that you want to catch all non-compliant vehicles and thus want higher Recall. However, this is simply intractable due to the large amount of vehicles on the road. As we are optimizing for Precision, the recall metric is usually on the lower end as these metrics are often in competition with each other.

2.4.2. OOB Error

To measure how accurately the model performs on new data not previously seen, you can use a test set on which the trained classifier is to be tested. The error calculated on this set is the generalization error of the model. Another way to estimate this error is to use the Out-of-Bag(OOB) error using the Out-Of-Bag samples. To calculate this, you take for each Out-of-Bag sample all trees that were not trained on this sample. Where the Out-of-Bag samples are those samples in the original dataset which were not included in the bootstrapped set. Taking the majority of votes of all these models (i.e. trees) and comparing this to the true value of the Out-of-Bag sample will give the OOB error. You can repeat this for the whole data set, i.e. all Out-of-Bag samples, and then take the average. Using the relative frequencies of the class, we can estimate the probability by averaging this over all trees. Breiman [12] has shown the OOB error method to be a good estimate for the generalization error and can thus be seen as an estimate for the error measured on a test set of the same size as the training set. Thus, to assess the overall quality of the model, the OOB error is a good metric. We can use it to compare models and see how they have improved after changes have been applied.

2.4.3. AUC Score

Additionally, we also look at the AUC score to evaluate our model performance. AUC stands for Area under the curve. The often used metric for this is Area under the ROC curve(AUROC). The ROC curve essentially looks at the rate of True Positives vs False Positives. This metric is effective at looking at how many true positives are drawn before a false positive. In situations where false positives are equally as troublesome as false negatives, this is a good metric to optimize. Although, as we have pointed out earlier, as our cost of false positives is quite high, we choose to optimize for Precision. Nevertheless, AUC score can reflect the overall performance of the model.

2.4.4. High Confidence Evaluation Metrics

Since we are trying to make changes to impact the presence of High Confidence errors, we need a way to measure the actual magnitude of these errors. To measure the magnitude we look at the top 20% of the errors and the top 5%. The distributions of the prediction probabilities have approximately a normal distribution. For this normal distribution the top 20% of errors are approximately all those errors that are more than 1 standard deviation away from the mean, i.e. 80% of the predictions are within 1σ of the mean. For the the top 5% we have all those errors that are more than 2 standard deviations away from the mean, i.e. 95% of the predictions are within 2σ of the mean. This means that of all errors that the model has made, we take the (two-sided) bottom and top highest confidence errors. As the model predicts a probability score between 0 and 1, close to 0 means high confidence of compliance and close to 1 means high confidence of non-compliance/violation. This means that for 20% we take the bottom 10% of highest confidence error probabilities of compliance and the top 10% most confident errors of non-compliance. The top 20% looks at most of the errors that can be seen as highly confident. Usually we find that the distribution is skewed to predicting vehicles as compliant, meaning that the distributions are skewed to the left, this means that values of 60% confidence are usually already seen as high confidence on the top end.

Alternatively, the measurement of the top 5% of errors should really give us an indication of the magnitude of the highest confidence errors which have the highest chance of causing doubts in the performance of the model. Equation 2.4 shows the formula for calculating the magnitude of the top 20% and equation 2.5 the formula for the top 5%

$$\frac{\sum 1 - Bottom\ Confidence + \sum Top\ Confidence}{Total\ Errors\ in\ 20\%} \tag{2.4}$$

$$\frac{\sum 1 - Bottom\ Confidence + \sum Top\ Confidence}{Total\ Errors\ in\ 5\%} \tag{2.5}$$

2.5. Dataset

The Data used to train the risk model consists of 5 main parts. Here we will list and explain each part.

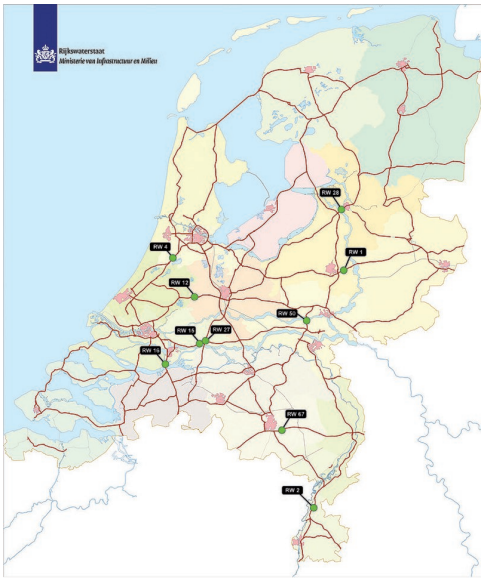


Figure 2.3: Map of The Netherlands showing locations of WIM passages

2.5.1. Vehicle Inspections

The core part of the dataset is the Vehicle Inspection data. The data set is based on the data in the Holmes system that the inspectors themselves use during their inspections. They annotate for each vehicle they inspect important characteristics of the driver and the vehicle. For vehicles that are found to be violating, it is more likely that more data about them will be annotated. The data contains the information of all vehicle and company inspections from May 2017 until October 2020. This dataset includes a total of 24.505 inspected license plates/vehicles, of which a total of 8989 are Dutch vehicles. Additionally it includes 296 company inspections which are linked based on the license plate and owner data.

2.5.2. Weigh-In-Motion Passages

The Weigh-in-Motion(WIM) Passages are a collection of 10 locations[36] on the Dutch roadways where automated weighing systems are used. Figure 2.3 shows a map of where these locations are located. These systems are a collection of cameras, sensors and weighing platforms built into the road surface. These work together to provide inspectors with an extensive insight into the weight distribution of the vehicle as well as keeping track of different information such as the location,time and speed of the vehicle.

The WIM passage data includes data from January 2018 up to August 2019. In the data the license plate of the vehicle is linked with the raw data received from the WIM passages.

2.5.3. Vehicle Registration

The vehicle registration data comes from the Dutch oversight of the Roadtransport registration, the Rijksdienst voor het Wegverkeer(RDW)[1]https://opendata.rdw.nl/ The open data of the RDW includes all general vehicle aspects and characteristics. On top of that, it also includes all data on vehicle general maintenance and previously found maintenance flaws. The open data is up to date up to the end of 2020.

Additionally, also a manually obtained dataset from the RDW from 2019 is used, which includes all vehicle owner assignments. Note that this includes only the front truck part of the vehicles and not the trailer part, which is registered separately.

2.5.4. Vehicle Licenses

From the national and international road transportation organisation (i.e. NIWO) we obtained the data of the European and Dutch vehicle licenses required for road transportation vehicles and owners. This data consist of data until the end of 2019.

[1]https://opendata.rdw.nl/

2.5.5. Company Registration
The company registration data consists of all data that is available through the Dutch Chamber of Commerce register, where owners of the vehicle can be linked through their company association, often based on the address of the vehicle and the address of the company. This includes all data up to the end of 2020.

2.5.6. Missing Data
Much of the valuable data that is available for Dutch vehicles is missing for foreign vehicles, license and company information is not available from foreign vehicles. This means that it is decided to leave the foreign vehicles out of the model. This is an issue as the inspectors have the responsibility of checking all vehicles on the Dutch Roadways, so including foreign vehicles. This means that the model is not as useful for these.

In this chapter we have outlined the entirety of the context in which the work of this thesis is performed. We explained the role of the ILT in this and how the model that they are developing works. We explain the predictive classification model that is used by giving an explanation of tree-based models and Random Forests. We show and explain which metrics are important in this research as well as the optimization of the model. Lastly we outlined the datasets and its constituents that are used for the training of the model.

3

Background and Related Work

As covered in the introduction, this thesis incorporates many areas of Machine Learning and Human-Centered Computing. The combination of Human Sciences and Computer Sciences makes it such that there is much background to understand on how these two touch. To get a grasp of the previous work performed in the literature and to understand the concepts required during the reading of this thesis, this chapter outlines some of the concepts. We cover the work on Unknown-Unknowns and High confidence errors, this is a novel area of research in the Machine Learning space and is yet to be thoroughly researched. Since we work on human interactions with Machine Learning models, we cover previous work on Domain Expert Interaction. Further, we lay out some concepts about the different kinds of distinctions in data bias that exist in the field of Machine Learning. Lastly, we do an extensive literature survey on model interpretability and explainability, which helps us in our goal of understanding our model outputs.

3.1. High-Confidence Errors

Errors produced by machine learning systems can be seen to belong to two broad categories: **(1)** Errors near the decision boundaries of a model. These errors are mostly caused by the inherent variances within the data. These errors are understandable from a practical point of view. Since the model is uncertain about how to categorize an instance, it can cause an error in the prediction. **(2)** Errors made far from the decision boundary. These errors usually hint at inherent issues with the data used to train the model. This can be a lack of data or also an under-representation of certain types of data points. These errors are produced with a high model confidence and we will refer to these as High-Confidence Errors (HCEs). These errors are detrimental to the trust of a system and should as such be avoided. The first type of errors can also be referred to as *Known-Unknowns*. The second type can be referred to as *Unknown-Unknowns*. These Unknown-Unknowns are the High Confidence Errors.

At the center of the question of High Confidence Errors and how to characterize and mitigate these is the fact that they happen at the outer edges of the domain space rather than near the decision boundary. The fact that the decision is made with a high confidence but is still an error can hint at either the fact that there are some vital *Unknown-Unknowns* in the domain space that we are missing or that the model over relies on certain indicators and thus starts overfitting on these.

3.1.1. Unknown-Unknowns

Previous work has gone into the field of detection of errors, particularly in the detection of *Unknown-Unknowns* by Attenberg, Ipeirotis, and Provost [6]. Here, participants were challenged to 'beat' the model by providing it with examples that would result in a high predictive error. The participants were able to identify many errors where the model assumed it was correct, but was wrong due to missing data and/or biases. A system similar to Beat The Machine has recently been developed as Facebook's Dynabench.[1]

Vandenhof and Law [64] tried to improve on the method of Attenberg, Ipeirotis, and Provost [6] by using a hybrid approach where the crowdsource method of [6] is combined with a set of decision rules on how high confidence predictions are made –obtained from interpretability methods– as well as an algorithm that intelligently serves the instances to the participants.

[1] https://dynabench.org/about

Liu et al. [45] show that it is possible to extrapolate certain examples of *Unknown-Unknowns* by looking for patterns of errors and using these as data for a classifier that can be used to find more examples of *Unknown-Unknowns*.

3.2. Domain Expert Interaction

From the literature it looks to be the case that a vital aspect in combatting high confidence errors made by the model is some intervention by a human participant in the loop. The human factor can help in determining those areas where data bias or lack of proper data could be an issue. For each domain where Machine Learning is applied, there are always also experts in these domains. These are the so called Domain Experts. The interaction between these experts and those Data Scientists creating the model is a critical aspect to proper model development.

Data selection and feature construction can be seen as the main crux of classical machine learning. Previous work [24] has shown that for certain applications, a continuously interactive way of machine learning can lead to improvements. A study into development tools for statistical ML by Patel et al. [56] has shown that there is a need for an exploratory and iterative process in the process of data, and feature selection. In the natural language processing domain, Park et al. [55] proposed interactive tools that enable sharing domain knowledge through domain concept extraction and label justifications. Nonetheless, the role of the Domain Experts themselves is often overlooked in the development of trustworthy machine learning systems.

Stumpf et al. [63] investigated the interactions between end-users and machine learning algorithms and performed three experiments to study the potential for these interactions in increasing the users' knowledge of, and trust in the system while also aiding in improving the performance. They studied how to incorporate feedback from the end-users into the model without negatively affecting the model's accuracy. The participants of the experiments made suggestions and critiques on what to change: "... participants made a wide variety of reasoning suggestions, including reweighting features, feature combinations, relational features, and even wholesale changes to the algorithms."[63, p.23]. The researchers showed that the algorithm that took the user feedback into account outperformed a simple online learning algorithm. Seymoens et al. [60] propose a methodology of co-creation workshops for the development of decision support systems in the health domain of rheumatology. We will come back to this in our chapter on the Expert Session Methodology in chapter 5.

3.3. Data Biases

A Machine Learning model is only as powerful as the data you provide it. This means that if the data has biased indicators in it this will also reflect itself in the output of the model. Throughout this thesis we deal with different types of biases at different points of the development cycle. One of the main biases that is often present in machine learning systems is called sample bias or selection bias. This bias is created by the environment and training instances that the model is trained on being different from the real world scenario. A model that is trained only on inspected vehicles will surely be at the mercy of the judgement that an inspector puts on these vehicles. There is no way to know the true 'risk' of a vehicle, only risk in the eyes of the inspector, as the inspector is also the annotator for much of the data that is used. This leads to a generalization error, where the model may perform well on data similar as to that which it is trained on, i.e. the inspection data. But when this model is then applied in a real life environment the performance can differ.

This also points at the obvious other source of bias, the inspector or expert themselves. Each Inspector works with their own set of presuppositions and assumptions of what a non-compliant vehicle is like. Moreover, not each Inspector is as strict as the other. This is essentially a form of Measurement Bias, since now each data instance of the same type may be judged differently, causing inconsistent and conflicting information for the model.

Furthermore, the data itself can have issues. The data can be overrelying on certain indicators which do not give an accurate representation of the domain space. This is typically a consequence of the sample bias, since the model overfits on these biased indicators.

3.4. Model Interpretability

The rise in the complexity of machine learning models has highlighted a new problem, namely that of understanding the decisions that these models make. It has become increasingly unclear what features in the data are at the core of certain decisions, which has become especially troublesome for applications using pattern

recognition or clustering models on unstructured data[8]. For there to be trust in and accountability of these systems, we require both experts and users to understand how these systems work and what they base their decisions on. This has given rise to the fields of interpretable and explainable AI(XAI)[27].

In order to answer our second research question, namely *"How to best interpret what the model has learned?"*, It is necessary to get a sense of what it means to interpret a model and what best approaches and techniques currently exist for this purpose. Because of this, we preceded our research with a literature study that included a study on the topic of interpretability.

In this section we outline the struggle of clearly defining interpretability, we give an overview of what it takes to make an interpretable design and we outline some of the common interpretability techniques, such as SHAP, LIME, feature dependencies and counterfactuals. We weigh their benefits while also giving the justifications for our choice of techniques and how these relate to answering our research questions.

Right to Explanation As Machine Learning systems are being applied in real life applications, regulations have been created in order to ensure proper use. Data regulations laws such as the GDPR [2], address the issue of systems' need to be interpretable. Systems that are in use are having real life consequences. This is especially an issue when significant amounts of money or even human lives are at stake [9, 40]. Additionally, in recent years, much controversy has been created by the introduction of ML systems in fields such as hiring or risk assessment, where biases in training data have a big influence on citizens lives [22, 39]. This is also relevant for the case of the ILT, where the decision making models can have an impact on the way inspections are performed.

As explained also in the introduction chapter, the designers of the high-stakes Machine Learning systems can be held accountable for the choices which the system makes[2, art. 21, 22]. This poses new challenges for the creators, as they now have to take this interpretability factor into account when creating a model, this is especially troublesome with more complex models such as Random Forests and Neural Networks [28]. However the degree to which there is an actual 'right to explanation' is still a controversial topic, some argue that articles 21 and 22 of the GDPR merely requires creators of automated decision making systems to provide the user with meaningful information such that the user can make an educated choice on their right to opt-out from an automated decision making system. There have been new proposals in the European Commission[57] on Artificial Intelligence systems that do cover the aspect of a user's right to interpretation of model output, although this is still only a proposal. Either way we can envision a future where creators are required to explain their model's decisions, meaning that it is even more important now with the future in mind. Since we are using a Random Forest model for our case study, we need ways to explain what the model is 'thinking'. Since we perform sessions where we present domain experts with model outputs, we require some way to convey these outputs in a valuable and comprehensible way to the domain experts.

3.4.1. Interpretability vs. Explainability

Before we discuss methods for interpretability, it is important to clearly define what it means for a system to be interpretable and to also point out the distinction between interpretability and explainability, as these are often wrongly conflated. Commonly, interpretability can be defined as "the ability to explain or to present in understandable terms to a human" [23]. However in reality a clear definition is difficult to set as there are many distinct aspects to interpretability. Interpretability touches on many topics such as trust in the system, causal relationships, informativeness of the model, robustness of the model to changes in data, privacy of user data and even fairness [44]. These topics, often termed *desiderata* in the interpretable ML literature, have been clearly summarised by Marcinkevičs and Vogt [52].

Differentiating the term *Explainability* from *Interpretability* is continuously under discussion in the Machine Learning community [52, 44, 59]. Some authors such as Lipton [44] state that in the case of interpretability we want to know "How does the model work?", whereas explainability methods try to answer "What else can the model tell me?". Alternatively Rudin [59] states that Interpretable ML is concerned with inherently interpretable Machine Learning models, while Explainable ML is mostly concerned with providing post-hoc explanations of existing models. Throughout this thesis we mostly stick to the more loose definition of Lipton [44] where more focus is put on the transparency of the model. Since the focus of our research is on High Confidence Errors rather than specific interpretability study, we need a definition which aids us in this pursuit. Lipton [44] definition of "How does the model work" sticks close to our RQ2 of *How to best interpret what the model has learned*.

Local vs. Global Explanations There exist many different approaches to providing model explanations. Before we discuss several explanation methods, it is important to note the distinction between local and global explanations. When we talk about local explanations, these are simply single data instance explanations, concerned with only the feature values of that specific instance and potentially as it relates to the rest of the feature set. Global explanations are explanations of the model as a whole, i.e. what features and concepts give rise to the totality of the model classifications.[23, p. 7] In the case of our research we want to characterize specific instances, namely those instances predicted with a high confidence but which were in error. This means that we are not necessarily interested in interpreting the model as a whole globally, but rather we want to interpret those instances that are of interest to our goal. This means that we put more focus on explaining locally.

3.4.2. Interpretable Design

As shown in the previous section, interpretability is a broad concept that can be concerned with many different aspects of the model. For the purpose of understanding models more clearly, we need to identify those explanation methods that are most effective towards our aim. Visual analytics tools have been developed for fields such as Deep Learning [33]. There have as well been surveys for more general predictive visual analytics tools [46, 47]. Here we discuss an influential paper in the field of interpretability visualisations, namely that of GAMUT [32]. For the creation of their Visual Analytics, GAMUT Hohman et al. [32] explored how interactive interfaces could better support model interpretation. Their work helps in answering questions about human-centered approaches to model interpretability as well as some of its requirements.

This paper touches on those topics which are essential for our research in regard to interpretability. For our research we are working on Human-centered approaches for helping us characterize and mitigate errors created by the model. Interpretability is a tool in our toolkit that can help us in this effort. The paper of [32] gives a clear outline of what interpretability techniques are needed for a proper interaction with human participants. It clearly explains the benefits and drawbacks of these techniques and how to best use them in a human-centered setting.

Based on the literature as well as interviews with both ML researchers and ML practitioners, Hohman et al. [32] generate the following capabilities for an interpretation interface to have:

- **Local instance explanation**: Given a single data instance, what features contribute to what degree to the ultimate prediction.
- **Instance explanation comparison**: Given a set of instances, what distinguishes them from each other, what features are higher and what features are lower.
- **Counterfactuals**: What is the effect of modifying certain features on the ultimate prediction.
- **Nearest Neighbours**: Given an instance, what are other instances sharing similar features, predictions, etc.
- **Regions of (un)certainty**: What regions of feature values cause higher uncertainty of the model.
- **Feature importance**: In the entirety of the model, what are those features that contribute the most to the classification.

To evaluate their tool Hohman et al. [32] performed a user study involving data scientists and non data scientists, 12 participants in total. During the study, the focus was on how these users understand the models best.

The user study covered three important main topics:

1. Reasons for interpretability:

   During the study participants used the model to confirm prior beliefs they had about the data. This way they increased their trust in the model, by confirming these beliefs. Participants also mentioned that there are trade-offs to be made between simplicity and completeness of the explanations, making some participants promoting the use of simpler models which are easier to explain, even though they might impact accuracy.

2. The use of global vs. local Explanations:

During the study novices to machine learning gravitated more to local explanations. Intuitively these explanations are easier to understand, but ultimately both global and local explanations were complementary.

3. The effect of interactivity:

   From the user study it became quite clear that the interactivity of the tool is a vital aspect in the exploring, comparing and understanding of different areas of the data and the model. Making the participants claim that they could barely imagine the tool without it

3.4.3. Interpretation Methods

In this section we explain some of the most commonly used methods for the interpretation of ML models, as well as pointing out their disadvantages and advantages and how they relate to our research.

SHAP SHapley Additive exPlanations or SHAP, as it is more commonly known, first proposed by Lundberg and Lee [49], is an interpretability method for showing the feature contributions for a single data instance, a set of instances, or the whole model. It calculates the feature contributions for a data instance by using the concept of Shapley values [61]. Shapley values is a game theoretical concept of coalitional games of how to fairly distribute a payout over a set of players. Shapley Values work by taking a coalition of all players ϕ and removing the player ϕ_i for who you want to calculate the Shapley value. The difference between the value of the coalition that includes ϕ_i and the one that does not include ϕ_i is the marginal contribution of ϕ_i. Then perform this step for all permutations of possible coalitions that differ only by containing or not containing ϕ_i. We now take the mean of all the marginal contributions to obtain the Shapley Value of ϕ_i. This Shapley Value is the average contribution of player ϕ_i to the total value of the coalition.

Now to apply this method to the domain of model interpretability we can see the feature values acting as players and the prediction output as the payout. Using this approach Lundberg and Lee [49] propose their implementation named SHAP. They show that SHAP satisfies all properties of local accuracy, missingness and consistency [49].

To calculate the individual contributions of the features to the final model output we calculate the Shapley Value for each feature. The Shapley Value ϕ for feature value i can be calculated using equation 3.1:

$$\phi_i(f,x) = \sum_{s \subset x} \frac{s|!(C-|s|-1)!}{C!}(f_x(s) - f_x(s \backslash i) \tag{3.1}$$

Here f is the model, x is the datapoint, s is a subset of features from that datapoint and C is the total number of features. The first term in the equation is a weighting term for the total amount of features that are included in the subset. This is performed since a large change in the model output is more significant when the subset is large rather than when the subset is small. The second term calculates the difference in model output with and without the feature included. Here we note that, to obtain a model output, you can not simply exclude feature values and calculate the function value, for this reason, instead of excluding the features, SHAP takes random feature values from the dataset. On average this will remove the relevance of these features and thus be the same as removing the feature values. Since performing these calculations for all permutations would be computationally impossible when the feature set is large, we need a way to approximate the values. Lundberg and Lee [49] do this with their KernelShap method. KernelShap fits a linear regression model, where the output of the model is the prediction value, the variables are given a value of 0 or 1 if the feature is present and the coefficients of this model are the approximations of the Shapley Values. Other model-specific kernels are also proposed, such as TreeSHAP [48] for tree-based models.

Overall SHAP is an excellent interpretation method with a good theoretical backbone, it satisfies the three properties of local accuracy, missingness and consistency and produces fairly distributed explanations over the feature values. With the TreeShap method, SHAP has a fast implementation that can quickly and efficiently calculate the feature importances for individual instance level explanations. Since our model is a Random Forest model, SHAP is an excellent method to use. It allows us to get a clear view of each features contribution while also providing a fast implementation. SHAP explanations are based on a relatively simple to understand concept of including the feature or not including it and observing the impact. Since we want to characterize certain specific instances that are in error we have a need to explain local instances and analyze what features most contribute to these. Weighing the benefits, we find SHAP to be a good option for this. SHAP is also a good option for our Expert sessions, as the SHAP feature contributions can help us in making the model clearer to the experts which helps build trust in the system.

Local Surrogate Models A different approach, but also widely used is the use of Local Surrogate Models. This is essentially the use of simpler, more interpretable models which are trained on local neighbourhoods of the data, to approximate the interpretation of the whole model. Ribeiro, Singh, and Guestrin [58] propose their implementation of such a surrogate model, namely Local interpretable model-agnostic explanations (LIME). LIME works by taking variations of the data and training the surrogate model on these, while weighing those instances by how close they are to the instance of interest. The advantage of this method is that you can train a more interpretable model, such as, for example, a decision tree on this local neighbourhood. Figure 3.1 shows an example[2] of the process of selecting the instances, reweighing them and training the local model.

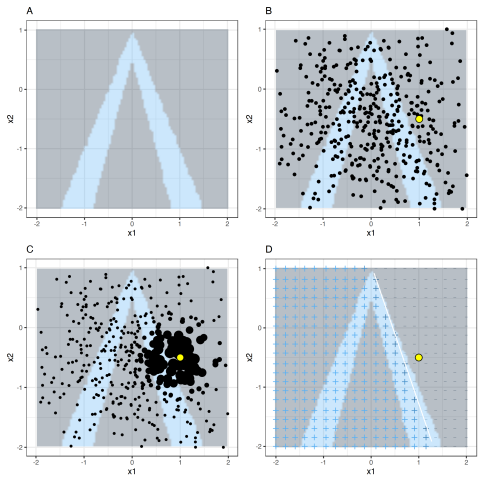


Figure 3.1: LIME: Process of selecting and reweighing the neighbourhood. The yellow dot indicates the instance of interest. The bottom right square shows the reweighing of the instances in the local neighbourhood, the bottom left square shows the new classifications based on training the model in the local neighbourhood.[2]

An advantage of this method is that it is quite flexible. You can change your underlying model while keeping your surrogate model the same. The explanations produced by the surrogate can even use different features from the ones used by the original model, making it a more flexible choice. Depending on what type of surrogate model you choose, the explanations can be more intuitive for people not so proficient in machine learning.

Despite these advantages, the main disadvantage is that doing local surrogate models right can be quite challenging. The choice of neighbourhood can be quite tricky, the explanations are not always as robust, the choice of neighbourhood can have a large impact on the quality of explanations, even two neighbourhoods that are very close to each other can still have significantly different explanations [5], making it quite unstable. Moreover, due to the fact that you are free to choose the surrogate model, this requires a lot of knowledge of both the problem space and what models would perform well, making it a method that is difficult to apply correctly.

Feature Interactions Sometimes we are not only interested in individual contributions of features, but also their interactions with each other. One simple option to do this is through Partial Dependency Plots (PDP). Here you would plot the values of a variable against the values of the target variable and see the effect of changes on the output. A problem with this approach is that you can only perform it with 2 features at a time. Doing this on a model where there is a large number of features will get unruly, for that reason this method is only useful for small single feature impact analyses. However, it is still particularly useful for showing some clear data biases for certain features, such as certain variable values that greatly impact the performance of the model. For this reason we employ this method for analysing some of the model's variables in chapter 6.

Another method that can be used to look at the interaction between two features is the Friedman H-statistic [25]. This can also be extended to the interaction of a feature with all other features. Using the partial dependence between variables, the H-statistic can give an interaction strength for each variable, plotting these interaction strengths for all features can give a good sense of the importance of some features.

[2]Source: https://christophm.github.io/interpretable-ml-book/lime.html

Counterfactual Explanation To get a sense of the causality of some features, we can perform a counterfactual explanation analysis. Counterfactuals look at the effect of changing individual feature values on the prediction output. More specifically; an explanation using counterfactuals describes the smallest change to the feature value that is needed in order to change the prediction to a certain output. An example of an implementation of counterfactuals is that of Wachter, Mittelstadt, and Russell [67], where they minimize a loss function of how far the predicted outcome of a counterfactual is from the the desired outcome

Counterfactuals can be a suitable choice, especially for when trying to give explanations for decisions of automated decision-making systems, such as when causality explanations are needed for the GDPR [67].

SIRUS: Decision Rules Lastly, we also looked at what methods are especially good for interpreting Random Forests, since we are working with a random forest. For this we found an approach by Bénard et al. [11] named SIRUS (Stable and Interpretable Rulee Set). This is a method of generating a set of decision rules from the Random Forest. The decision rules are generated by aggregating over the entire Random Forest itself, rather than aggregating over the predictions generated by the trees. The most frequently occurring nodes in the trees are used to create the set of decision rules. An example[3] of such a set of rules, created using the Titanic dataset is shown in figure 3.2.

This approach can give insights in the Random Forest as it can provide more insight on the specific feature values that the model uses to make its decisions.

Average survival rate $p_s = 39\%$.

if	`sex` is male	**then**	$p_s = 19\%$	**else**	$p_s = 74\%$
if	1^{st} or 2^{nd} `class`	**then**	$p_s = 56\%$	**else**	$p_s = 24\%$
if	1^{st} or 2^{nd} `class` & `sex` is female	**then**	$p_s = 95\%$	**else**	$p_s = 25\%$
if	`fare` $< 10.5£$	**then**	$p_s = 20\%$	**else**	$p_s = 50\%$
if	no `parents or children aboard`	**then**	$p_s = 35\%$	**else**	$p_s = 51\%$
if	2^{st} or 3^{nd} `class` & `sex` is male	**then**	$p_s = 14\%$	**else**	$p_s = 64\%$
if	`sex` is male & `age` ≥ 15	**then**	$p_s = 16\%$	**else**	$p_s = 72\%$

Figure 3.2: Set of SIRUS Decision Rules produced using the Titanic Dataset

In this chapter we have shown the previous work that has gone into characterizing and mitigating high confidence errors through research into Unknown-Unknowns. We highlighted how domain expert involvement has previously been used to great effect. Furthermore, we also provided a distinction of the different types of bias that can exist for a model. Lastly, we ended with a literature study on model interpretatbility, where we outlined some of the most important concepts and techniques that are used in the ever-emerging field of Interpretable Machine Learning. We gave a description of the definitions of interpretability and explaibility that are out there in the literature while also noting the difficulty in defining these. As we are interested in studying the workings of our model with the study of High Confidence Errors in mind, we chose to go with the simpler defined definition of interpretability as being "How does the model work", which is better suited to our goal of characterizing High-Confidence errors. We outlined the design of interpretability interfaces as a tool for researching what techniques are most essential for us to use in our case study. Finally, we outlined the interpretation methods, where we chose to go with SHAP as well as feature interaction as our main techniques that we use. Feature interactions can be a great tool for showing certain data biases such as outlier variable values that have a large impact on model performance.

[3] Source: https://github.com/cran/sirus

4

Data Methodology

To improve the model and reduce the magnitude of high confidence errors, we need to first understand what the model itself has learned. This section presents a methodology for a data driven approach for understanding the model's knowledge. In such a way we answer our first research question RQ 1: *What Instances Best Characterize System Knowledge?* For the Data methodology part we focus on understanding the model instances both through an analysis of highly contributing variables as well as characterizing the High Confidence errors. Furthermore, through the use of an error prediction model we characterize what instances are more prone to be in error. A full graphical summary of the Data-driven methodology is shown in figure 4.1.

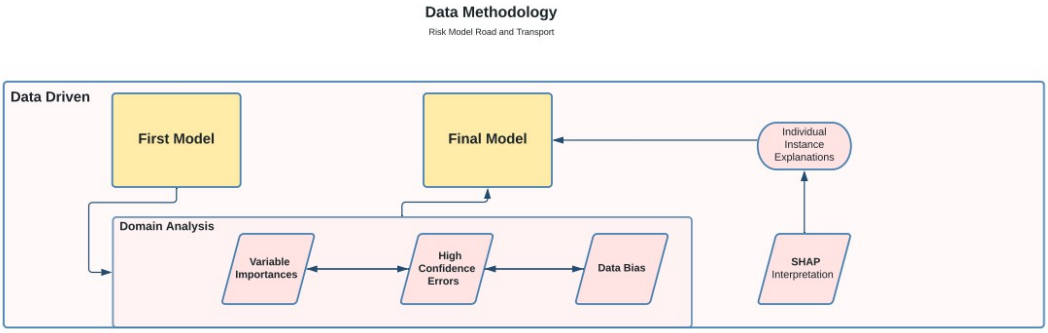


Figure 4.1: Graphical summary of Data driven model methodology

4.1. Dataset Split

For our efforts of characterizing the High Confidence Errors created by the model, we need a sufficiently large set of test instances on which we can make predictions. For this a train/test split was made on the dataset. We chose a split of 80/20 percent for train and test sets respectively. This choice was made weighing the size (8584 data instances) of the set as well as the need for a representative test set. We need a sufficiently large testset since we also need a sufficiently large amount of errors. The distribution of the confidence scores of the errors generally follows a normal distribution or positively skewed normal distribution, meaning that there will be, relative to the entire set, only a small amount of errors on the tail ends of the distribution. For the purpose of the Analysis sessions, which are further explained in chapter 5, a 90/10 train test split was used to keep the performance of the model as close as possible to that of the model trained on the full set.

The Train/Test splits are performed using a stratified split, this means that the proportions of the prediction categories are preserved, meaning that each split will have the same proportion of non-compliant vehicles vs. compliant vehicles as in the original dataset.

Alternatively, an option would be to use a Leave-One-Out Cross Valdation (LOOCV) approach, where you train the model on the full dataset leaving one instance (or a set of instances) out on which you will test.

4.2. Variable Importances

The risk model, as explained in section 2.2, learns from the training data what variables are most likely to contribute highly to the risk of violation for a vehicle. In the case of the Random Forest model, the importance of a variable is for the most part determined by its mean decrease in the Gini Impurity over all trees. Those decision nodes of the features with the highest decrease will have the largest contribution to correctly classifying the instances. This gives us a good indication of what the model has learned. However, to further deepen our understanding of these specific variables we perform a more thorough analysis for the top contributing variables. For this, we take the top 20 highest contributing variables and we analyse what feature values have the greatest occurrences. To see how these feature values contributed to the predictions, we compare these feature values with those values for only the instances that were erroneously predicted. This way, by comparing between the different sets of instances, we get a better understanding of the choices the model makes and how these choices lead to errors.

4.3. High Confidence Error Analysis

In our research we are mostly interested in understanding not only what the model has learned but also how this relates to the characterization of high confidence errors. How do the instances differ from the errors that occur with a lower confidence score. For this we compare both the errors and the full set instances with a subset of the highest confidence errors. Here we are faced with the difficulty of characterizing what is "High Confidence". For this we want to take a subset of the errors which are produced with a sufficiently large confidence score such that it would be troublesome for the trust in the model if these occur in a real world scenario. We define our High Confidence errors as being on both sides of the distribution, meaning both the False Positive (High probabilities) instances as well as the False Negative instances (Low probabilities). Generally the confidence scores of the model predictions follow a (skewed) normal distribution, meaning that there are a relatively low amount of High Confidence errors, as most predictions, and thus also most errors aggregate around the mean. For this reason, to have a statistically sufficient amount of errors, we choose to define High Confidence as relatively broad, where we take all errors produced with a confidence score 1 standard deviation away from the mean on both tails of the distribution.

Those features that correlate with high occurrences of High Confidence errors are subjects for further analysis and improvement. If the model is overfitting on these biased variables, this can cause the model to be overly confident (wrongly) in real world scenarios, causing High Confidence errors.

4.4. Error Prediction

To answer our research questions of in particular *RQ1: What instances best characterize system knowledge* and *RQ4: What does the model not know?* we need a data-driven approach to more reliably see what instances and variables contribute to High Confidence errors. This helps us to better characterize these and as such to find ways to mitigate these.

In the previous sections we have shown how we aim to find and analyse those important variables that contribute to errors. Here we take a more proactive approach, where our aim is not to only post-hoc find those contributing variables, but to predict those instances that are most likely to be predicted as error in the original model. This way we want to get to the core of what characterizes a typical 'erroneous instance'. To do this, we will train a new Random Forest model on a dataset where instead of the original violator/compliant labels, we use the labels of Correct (correctly predicted) and Error (wrongly predicted). This means that we build a model that has the job of predicting whether an instance is likely to be wrongly predicted. Looking at the choices that this new model makes can give us insights into what pitfalls there are for our Risk model and what variables are troublesome. We relabel all instances that were in error to have the label "Error" while we label all other instances as "Correct" instances. This way we have a new binary classification problem. We train the new Random Forest model using the relabeled Test Set data as the training data. We do note that because of the size of the Testset, the number of instances that the model has been trained on is relatively small.

4.5. SHAP Interpretation

As covered in section 3.4, model interpretation can serve as a strong tool for understanding the decisions a model makes. Thus, model interpretability helps us in answering RQ1 and RQ4. Furthermore, for our expert sessions, model interpretability helps us in conveying the model's knowledge to the experts. This eases the

interactions between data scientists and domain experts.

For our model interpretation we use a SHAP variable contribution analysis (as explained in section 3.4.3) for the top 30 highest contributing variables. Since our focus is on characterizing High Confidence Errors, we perform this SHAP analysis on those instances predicted with a high confidence, both being a False Positive as well as False Negative. This allows us to better see what variables and variable values can cause these errors to occur. We then compare these with SHAP contributions from correctly predicted instances.

This chapter covered the Data driven methodology for characterizing errors and high confidence errors. This methodology is applied on the ILT risk model to make improvements in the next iterations of the model. The results of applying this methodology and also how it leads to the desired improvements are covered in chapters 6 and 8.

5

Session Methodology

In many of the applications of Machine Learning data scientists have the job of creating a model for a use case of which they will not necessarily be the final end-users. At the same time they also may or may not be domain experts in the relevant field. Because of this disparity, it means that there often exists a disconnect between, on the one side, what the data scientists know and are aware of, and on the other side, the knowledge and expertise of experts in the domain. This has also been referred to in the literature as the *knowledge acquisition bottleneck*[31].

To bridge the gap between Data Scientists and Domain Experts we propose a framework consisting of expert interaction sessions to guide the Human-Centered interaction between the Machine Learning model, Data Scientists and Domain Experts (i.e. inspectors).

This framework consists of a set of interaction sessions with the experts. Each session is to serve as a stepping stone for the next session such as to create a 'pipeline' of sessions where you build on knowledge from the previous session. This process is meant to be an iterative process where at the end of the series of sessions you can perform the process again but with new found knowledge and new found areas for improvement. This method of interaction is meant to create an optimal environment to gather knowledge from the experts while also familiarizing the experts with the workings of the model.

In this section, we start by giving an elaboration on the motivation for the methodology. Then, for each type of session, we give a generalized outline of the structure of the sessions. Namely, the sessions are the Exploratory sessions, In Depth sessions and Analysis sessions. We will explain the purpose of each session, how they work, and give their justifications. Finally a case study was performed in order to validate the methodology. We will give a detailed setup of each of these sessions that were performed with the Road and Transport Inspectors. The results gathered from the case study are outlined in Section 7.

5.1. Methodology Motivation

An interdisciplinary team consisting of members with diverse sets of backgrounds and knowledge can struggle with the issue of how to achieve common ground among the participants. Achieving common ground is vital in the process of closing the knowledge gap between multidisciplinary experts [54, 16]. Research into common ground building and negotiation by Beers et al. [10] is particularly relevant for this. They explain how participants have their own internal representations of ideas which have to be transferred to external representations that the other participants can understand. In the research of Beers et al. [10] they propose the use of formalisms, which are essentially a shared language and ways of communication that help in understanding what is being discussed. Formalisms are a framework for the participants to take a set of objects and rules and transform this to an external representation of the knowledge. This helps in moving from 'unshared' knowledge to 'external' knowledge through the externalisation of the participants internal ideas. The receivers of the knowledge should then, through the formalism and their internalisation, come to a shared knowledge.

In order to maintain common ground, an active effort needs to be made to fix misunderstandings between the participants. Formalisms help in facilitating this. From the study it was shown that formalisms were beneficial in the process of obtaining common ground. Nevertheless, the researchers admit that in a professional real world setting, other factors such as interpersonal factors or competition factors may be at

play.

These formalisms of how to transfer knowledge are part of the process common ground. Convertino et al. [19] give a distinction between two types of common ground: Content Common Ground and Process Common Ground. Content common ground involves the actual knowledge of domain subjects and the working practice. Process common ground, on the other hand, is concerned with all the shared rules, strategies and manners of interaction.

Convertino et al. [19] looked at the development of process common ground over time. They found that process common ground increases over time as explanations of rules and strategies of interaction become clearer. They also studied the development of content common ground over time. They showed that content common ground is formed through the continuous repetition, revision and clarification of relevant concepts in the domain. Mao et al. [51] studied, through a set of interviews, the scientific collaborations and interactions between data scientists and bio-medical scientists. They showed that the increase of process common ground allows for an increased ability of updating the content common ground as well. Similar conclusions came from the study of Convertino et al. [20].

These findings lead us to the conclusion that for the purpose of closing the knowledge gap between data scientists and experts, there is a need for an iterative collaboration cycle with multiple consecutive sessions. This is since, with each session and session iteration, the process common ground between the data scientists and the experts increases, and as a consequence so does the ability to update the shared content common ground.

5.1.1. Closing the Knowledge Gap

The methodology proposed by Seymoens et al. [60] shows a previous attempt at integrating domain expert interactions into a generalized methodology by proposing a series of subsequent co-creation workshops performed in the health domain. In their conclusion they call for the development of different methods for efficiently involving the domain experts in knowledge capture processes. This research, even though it has similarities to our approach, it differs in the sense that the researchers put more focus on increasing the interpretability of the model with as their objective "... to enable the discovery of underlying skills and know-how as they are provided through dialogue by professionals" [60, p. 212] in the field". On the other hand, our aim is to build forward on this approach and to iteratively perform continuous interactions with the experts, eventually leading to modifications to the model. This is used to then reduce the magnitude of high confidence errors, allowing us to preserve and increase trust in the system.

Seymoens et al. [60] propose their methodology through a series of workshops, starting with acquisition workshops where experts propose and discuss certain examples of scenarios where they apply their tacit domain knowledge. These sessions result in a series of flowcharts presenting the explicit and tacit knowledge. Following this there is a series of Validation workshops to clarify any ambiguity that still exists in the flowcharts regarding the knowledge needed for effective decision support systems. This methodology differs from our proposed methodology in its open nature. The methodology has no pre-set goal besides the aim of tackling the *knowledge acquisition bottleneck* for decision support systems. Our methodology has the aim of using these sessions to not only learn implicit and explicit knowledge from the domain experts but to also relate this to the trust in the model itself as well as explicitly reducing the (high confidence) errors that the model makes.

Besides closing the knowledge gap, we are also interested in achieving user trust in the system. [30] call for proactive communication, concrete and tangible information and transparency in the development process in order to better achieve this trust. It is important to already early on in the development process involve the end-users. Xiong et al. [69] point out in their design principle of User Visibility, that the more insight the user has, the more they are willing to trust the system. This also means that there is a need to introduce the experts to the actual inner workings of the model, calling for a session where we present the model results to the experts and interpret the results together with them, making clarifications where needed.

5.1.2. Methodology Requirements

Based on the findings from the literature, we lay out a set of requirements that our methodology must adhere to. These are the following:

- A generalized methodology providing a framework for the reduction of the knowledge gap between Domain Experts and Data Scientists

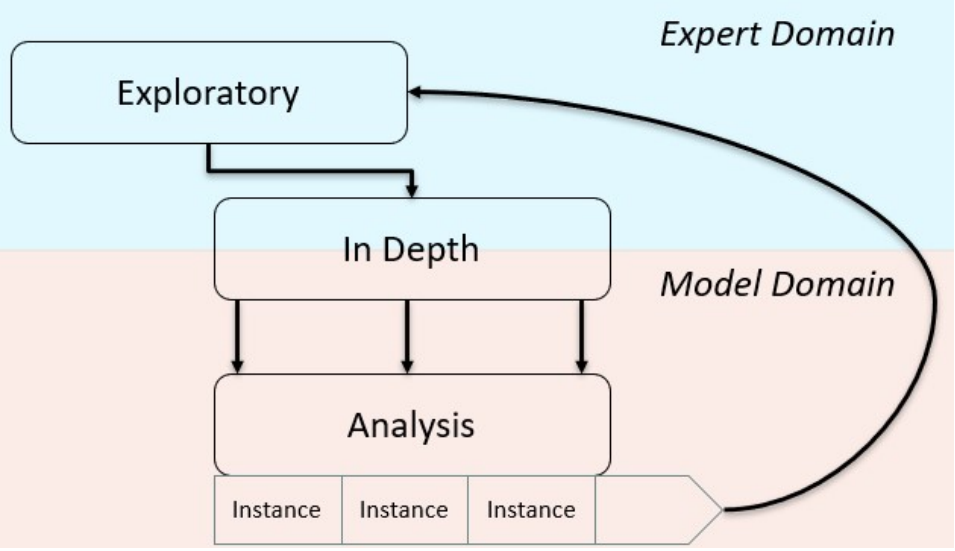


Figure 5.1: Graphical representation of Expert Sessions showing the relation between each. In the Expert domain the Exploratory sessions, moving through the In-Depth sessions to the Model domain and finally presenting specific model instances in the Analysis sessions. In the end the sessions can be repeated in an iterative fashion

.

- An iterable process that aids in the continuous increase of process common ground between participants, with as a consequence also an increase in content common ground.
- An improvement of the model through the reduction of High Confidence errors
- An increase of user trust in the model through early domain expert involvement by means of interactions with the model, as well as the fostering of a better understanding of the model.

5.1.3. Session Methodology

Now we introduce our own proposed Methodology. It consists of three types of sessions, namely *Exploratory*, *In-Depth*, and *Analysis* sessions, which have as their aim to address the requirements outlined in section 5.1.2. Each session type serves a different purpose which we outline below. These sessions help us in answering research questions 3 and 4, namely:

RQ 3: *What knowledge do we want the model to have?*

RQ 4: *What does the model not know?*

Through the experts knowledge we learn what indicators are essential to the model's good functioning as well as allowing us to pinpoint areas in the data that are still lacking. The sessions are structured such that we start from a general point, exploring in the domain of the experts, gradually moving to a more model-based domain with each session. Each type of session can be done in a single session or over the course of several separate sessions. Multiple separate sessions give the opportunity to go deeper into the topics and achieve more common ground, but also come at the cost of more time investment from the Domain Experts and Data Scientists. This is a trade-off to be made which can be different for each domain. The sessions are structured in a way that after each type of session changes are made to the model. The changes to the model carry on to the next session such that there essentially is a 'pipeline' of model improvements. The sessions are meant to work iteratively, meaning that once the final session has finished and all model changes have been performed it is up to the data scientists to reevaluate the state of the model and see if a new round of sessions is justified. Those areas of the domain that are yet unexplored can serve as a starting point for the next round of sessions. A graphical summary of all the sessions is displayed in Figure 5.1. A distinction is made between those sessions that happen in the domain of the experts and those that are based on knowledge from the domain of the model and data scientists.

5.2. Exploratory Sessions

The first session is the Exploratory session. As this is the starting point of the sessions, we assume only a basic understanding of the area of interest. However, even for those more advanced in the field, these sessions can always give the opportunity for new insights. The aim of the Exploratory session is to get a deeper understanding of the domain knowledge and working practice of the *Domain Experts*. We want to learn which are the indicators or features that experts use to ground their decisions on, and how they compare to how the machine learning model is built.

During the course of the Exploratory sessions the driving question is *"Where do the domain experts put their focus when making their decision?"*. Due to the open nature of the Exploratory session the way of answering this question can come in different forms. A good option to really get a feeling of what the working practice of the experts is like, is a day-in-the-life style session where the *Data Scientists* join the experts in their practice. During such a session the data scientists have ample time and opportunity to question the experts on the knowledge they want to acquire from the session. Besides this, it also gives the data scientists a great opportunity to evaluate and get a better sense of in what areas of the working practice the machine learning system can help the experts the most. The downside of this type of session is that, due to its more personal nature, it is less formal and thus it can be harder to stick to a predetermined format of what you want to learn.

The other choice for the exploratory session is to perform it in the form of a semi-structured interview [3]. This is a common form of interview in qualitative research where the interviewer asks mostly open questions where there is only a general framework of the topics that need to be discussed. This type of interview is a good method for situations where you are mostly exploring uncharted territory [3], as it leaves room for a broad set of ideas to be brought up by the participants. This setting leaves much room open for discussion between the experts that are being interviewed. This approach, however, requires a good deal of sophistication from the interviewers end, as these interviews can be time and labor intensive and are prone to drift off topic. The interviewer has to be able to adjust to the discussion smartly such that the topics are all covered while also allowing the experts to put forth their ideas. Furthermore, it can be strenuous to extract from the interviews all relevant information to your use case, as this can include going through many notes and/or recordings of the session.

During the interview session it is the job of the data scientists to bring up those topics of expertise on which they know they need more insight. This means that there has to be a preparation beforehand to analyse what areas the data scientists think they are still missing valuable knowledge. During the sessions, domain experts have the task of providing the data scientists with what they pay attention to with regards to the topic presented by the data scientists. They need to accompany these insights with their reasoning as to why they focus on these. It is the data scientists/interviewers job to probe the expert further if they feel that not enough reasoning is provided. The insights from these sessions will provide the data scientists with a better understanding of what features and data are essential for predictions.

Not all indicators that the experts give are always implementable within the data, it is up to the data scientists to decide what indicators are necessary to support the work of the expert and what indicators should be left for the expert's judgement only. The experts can help in making this decision in the following type of session, namely the In-Depth Session.

5.3. In-Depth Sessions

The In-Depth sessions are the next set of sessions, they follow the corresponding exploratory sessions. They have the aim of focusing on those points from the exploratory session(s) that require further study. The focus lies on understanding what data is useful for the model and how much it contributes to the prediction: extracting and ranking features that are vital to the models' goal. The driving question of this session is *"How does the available data contribute to the classification goal?"*. Answering this question sometimes requires further elaboration of information gathered from domain experts during the previous session. We are essentially in a search to extract and pinpoint from all the data that we have access to, those features that are vital to our goal, while also leaving room for new data sources to be incorporated in the model. Perhaps the domain experts themselves have suggestions on what kind of data they would like to see represented in the model, making suggestions like this is part of the In-Depth sessions. In this way, we incorporate the knowledge of the domain experts in the process of feature selection and engineering. As opposed to the Exploratory sessions where we were really open to any indicators that the experts propose, here we are only concerned with those which are representable in actual data and features. In such a way these sessions serve as the bridge between

the Expert Domain and the Model Domain.

The In-Depth Sessions are optimally performed in the form of Semi-Structured interviews, to offer space for discussion. The data scientists choose those areas in the data that they still are uncertain about whether they are beneficial for the model or not. They then ask from the domain experts to assess the relevance of these to the classification. Domain experts have the job to evaluate whether this data is actually relevant or not. When it is relevant, it is up to the data scientists to probe for further reasoning. This includes asking questions on why it is relevant, what parts of it exactly are relevant, and how this should translate to the classification goal. To give an example in the field of road inspectors, you could look at the relevance of company structures and nationality associations. Then it would be up to the experts to determine if this is relevant for classifying non-compliance risk. In case it is relevant, they should elaborate on what specific aspects of the company or nationality makes it a higher risk. Furthermore, they should clarify how it relates to the classification, meaning if different nationalities have a higher risk or rather lower risk of non-compliance. This method should make it clear to data scientists how the model should behave, which helps them in assessing the performance of the model when actually evaluating it. However, as we are proposing an expert-driven development cycle, it should not only be up to the data scientists to evaluate. We also want to involve the experts in this model evaluation. This leads us to the next type of session, the Analysis session.

5.4. Analysis sessions

The final session in the iterative interaction cycle is the Analysis session. It is called Analysis session since the main goal here is to analyze specific predictions of the model and evaluate whether they perform as expected. Essentially, these sessions serve two main purposes:

1. *Present experts with data of real-world cases and model classifications to better compare the model decisions with how experts make decisions in the same scenario*

2. *Provide the experts with a look into the decision making process of the model as well as hands-on experience to facilitate building of trust in the model*

By studying experts decisions and model classifications on actual data instances, we can better compare the model performance with how experts make decisions in real-world scenarios. Furthermore, we want to see how much the experts trust these model classifications. The driving question of this session is *"Do the domain experts trust the model predictions?"*. The predictions made by the model need to be understandable to the experts. Simply providing the experts with prediction results or confidence scores will not enlighten them much. The model predictions need to be accompanied with appropriate interpretation methods. In section 3.4 we discussed model interpretability and gave some common interpretability methods. The data scientist should choose those methods that are appropriate for their domain and their model. In general the use of a SHAP value feature contribution analysis [49] is a good approach as these give a clear indication of which are the highest contributing variables to present to the experts on top of more benefits mentioned in section 3.4.

The Analysis sessions are performed in the form of structured interviews. This means that beforehand the data scientists preselect data instances for which experts' input is informative, this can include data instances that the model has trouble with such as high confidence errors. However this can also include correct prediction by the model, to evaluate whether the experts would make a similar prediction. The choice of what instances to present is an important part of these sessions and can decide entirely the value that these sessions bring. Choosing these instances also means choosing what impression you want to give the experts of the model. Picking only erroneous predictions may give the experts the impression that the model performs badly in general. While only providing correct predictions may be not as valuable to get new knowledge or can even make the experts suspicious why the model is always correct. The key here is good communication. The data scientists need to communicate to the experts the reasons why they chose these instances in particular and what information they hope to get from the experts.

During the session the domain experts are presented with the data instances and their representation. The correct choice and presentation of the instances is vital for these sessions, as this choice highly influences the decision making of the experts. Choosing those variables and aspects of the data to show case also inherently introduces some bias from the data scientists' side, making the judgement of the experts reliant on this. Maybe the data scientists have an idea of what variables the experts would pay attention to and only choose to showcase these, this would essentially only confirm the data scientists presupposed hypotheses rather than leaving it up to the experts entirely. The other choice is to leave the variable selection entirely up to the interpretation method you are using, meaning in the case of SHAP values, as an example, you only

show the experts the top 10 contributing variables. This also entails, however, the issue of the session design. How to present the data instance can differ. You can simply give a list of features, or you can give the features in the form of a story which is easier to understand for the experts and perhaps fits their normal working setting better. Whatever decision you make, it is necessary to be aware of its consequences and the impact it has on the final result of the session.

Having presented the experts with the data instance it is up to them to first give their assessment and classification of the instance. This means that they give their judgement as to what class the instance belongs to, which can also be accompanied with how certain they are about this judgement. Following this, the model prediction is revealed accompanied with other relevant information about the prediction such as model confidence scores. Now a discussion should follow on any similarities or differences in the judgement of the experts vs. the model. From these discussions new indicators can arise or indicators that the model used to make its judgement can be reevaluated if the experts do not agree with them.

By finishing the Analysis sessions and answering this session's question we can finish this iteration of the sessions and apply what we have learned to the model. We can reevaluate and iterate over the session model again.

In the previous chapter 4 we explained our data driven approach. We can now combine this approach with the approach described in this chapter to have a full model improvement methodology. In figure 5.2 we show a graphical summary of the full model methodology. Instead of going straight from the first model to the final model, we have now added an additional intermediate model phase. This model will use the findings from the data-driven analysis as well as the findings from the previous expert sessions to generate an improved model. In such a way, we have a pipeline for iterative model improvement.

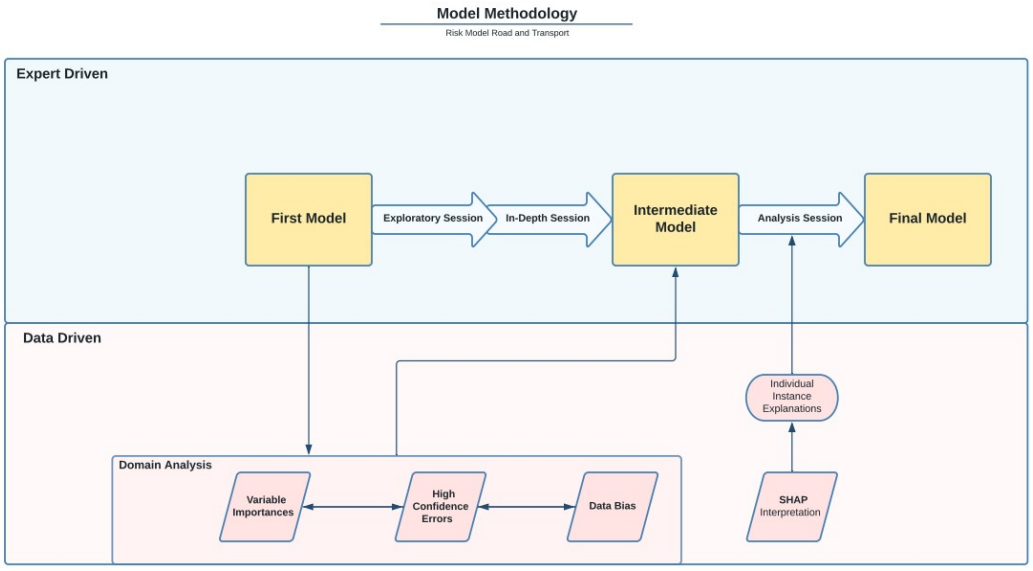


Figure 5.2: Graphical summary of Expert driven methodology from the sessions combined with the Data driven methodology. The Methodology is separated between the Expert Driven methodology(blue) and the Data Driven methodology(red).

5.5. Case Study

The case study sessions with the Road and Transportation inspectors of the ILT. We performed 2 Exploratory Sessions, 1 In-Depth session and 1 Analysis session. We will outline each session and some of the decisions that were made for each.

5.5.1. Exploratory

For the exploratory sessions a total of two sessions were held, one preliminary day-in-the-life style session with the inspectors of the dangerous goods inspectors (ADR) and one semi-structured interview session with the other road inspectors.

For the day-in-the life session we had the aim to gauge the need for a risk model for the ADR inspectors. We as the data scientists spent the day with the group of 4 ADR inspectors at their location in Geldrop and observed their working practice. The initial model was built with all vehicle violations included, this also means violations on Dangerous substance (ADR) laws. However, as this area is a bit niche, whether a general risk model is needed is up to debate. We observed only the inspection of the vehicles and not the selection process. The selection process of the vehicles was performed by a separate inspector who goes on the road to choose a vehicle to inspect. Due to COVID-19 regulations, we were not able to join to observe this process.

For the other exploratory session, a semi-structured interview session was held with 7 Road and Transportation inspectors, 4 data scientists and 2 interview coordinators. The interview coordinators were there to facilitate a proper interview flow and to guide and keep the discussions on topic. During this session the data scientists proposed 3 areas of exploration to the inspectors: overload, rest and driving times, and cabotage. For each area the inspectors are asked the question:

> "What aspects and features do you pay attention for when looking for non-compliance in this area?"

The inspectors brainstorm in order to identify the most important indicators. After this they have 10 minutes to write down their most important indicators using a collaborative writing tool (e.g. Mural[1]) and rank them according to their judgement. At this stage, data scientists can ask clarifying questions.

5.5.2. In-Depth

An In-Depth session was held in the form of a Semi-Structured interview. This session was performed with 4 inspectors and 4 data scientists. As this session was held with a smaller set of inspectors there were no interview coordinators present. Inspectors were queried on the following areas of interest regarding model data: moment and location of inspection, cargo, vehicle maintenance, vehicle ownership, and company structures. For each separate topic they were given the opportunity to discuss the topic with each other while the data scientists could ask clarifying questions.

For the moment of inspection, inspectors were asked:

> "Which temporal aspects are important for the compliance of a vehicle?"

Examples were given from previous sessions being: Time, day/night time, day of the week, weekend, month, specific periods, and holidays.

For the location of inspection the inspectors were asked the question:

> "Which elements of the location may say something about compliance of a vehicle?"

Examples were provided of elements that were found from the previous Exploratory session. These were: Region of the country and specific common routes.

On the topic of vehicle cargo, the inspectors were asked:

> "What types of transport have an impact on the cargo that is being transported and how do these relate to the compliance of a vehicle?"

Again, examples found from previous sessions were provided such as: Open transport vehicles, animal transport, construction transport.

For vehicle ownership we make the distinction between the tractor (pulling) part of the vehicle and the truck (pulled) part of the vehicle. The inspectors were asked:

> "Lease vehicle violations come on the name of the lease company, how does this have an impact?"
>
> "Does a tractor usually drive with it's one equipment or from a different owner?"
>
> "Are there combinations of Dutch trucks and foreign tractors? What other combinations of vehicle and owner are there?"

For vehicle maintenance, we focused on shrinking the list of vehicle maintenance features. Here we asked the inspectors:

[1] https://www.mural.co/

"Is vehicle examination data relevant for the risk profile?"

"Which categories of vehicle examination are of importance?"

For company structure the following questions were asked:

"International transport organisation with Dutch parent company, how does it have an impact and which countries are involved?"

"How does being an own carrier(with own transport vehicles) have an impact?"

Besides these specific guiding questions we also gave room to deviate and brainstorm about ideas that the inspectors brought up. The results of the answers that the inspectors gave were noted down, extracting the important main points of their discussion. The results of the interview are given in chapter 7. Following the interview, us as the data scientists discussed our notes from the session and converged on the main points from the session for each topic. These then lead to improvements of the model explained in chapter 8.

5.5.3. Analysis

An Analysis session was held with 3 inspectors and 4 data scientists. The session was held in the form of an interview. The inspectors were presented with 8 data instances in total. The feature for the model were selected by the data scientists out of the set of high contributing features based on the SHAP feature contributions. We note that we made choices in what features to showcase, which means that we are at risk of introducing some of our own biases, still we presented mostly only the highly contributing features. We made the feature values more comprehensible for the Experts. This means that as an example we would not present them with: "The mass of the vehicle was 10256 kg", but rather: "The mass of the vehicle was in the category of large and heavy vehicles (supported by the distribution of the vehicles)". Lastly, as the working practice of these inspectors is highly based on visual characteristics, we provided the predictions with a visual image of the vehicles that the instances were based on. Giving this visual representation can potentially skew the judgement of the inspector. Either way, it can also help in the process of understanding the sessions for the inspectors. In figure 5.3a we show an example of a case that was presented, you can see the image of the vehicle on the right side.

The data instances consisted of 5 instances which were high confidence errors and 3 correct predictions. Since our analysis aims to reduce high confidence errors, we focused on presenting these mostly. However, as we want to give the experts a more balanced view, and not give them the impression that the model is always erroneous, we also presented them with 3 correct predictions. For each instance we asked the experts judgement on whether they think it is a compliant or non-compliant vehicle, we also asked them how high of a risk they would deem the vehicle. After this, we presented them with the actual decision of the model and the confidence score. We discussed the similarities and differences in judgement. In figure 5.3b we present the view that the inspectors would see after they made their predictions, on the right hand side it contains the SHAP feature contributions of the instance and on the left hand side the model score and whether it was an error or not.

In this Chapter we covered our Sessions Methodology for the Iterative Domain Expert Sessions. We gave our motivation for the structure of these sessions by giving an overview of some of the relevant literature and previous work on this field. We gave a general outline of the aim of the sessions and how they should be structured. Lastly we presented our case study that we used to validate the session model with the inspectors of the ILT.

(a) Truck and Variable View

(b) Model Prediction View

Figure 5.3: Example of case during the Analysis Session in (a) we present the view that the inspectors are presented with to make their judgement on. After the inspectors make their decision and explain their reasoning they are shown view (b) where the model's decision is shown accompanied by the confidence score. To explain the decision of the model we show a SHAP feature contribution plot for the data instance.

6

Model Domain

In order to improve the model we need to understand the underlying domain of the model. This means that we have to understand what the model bases its choices on in order to improve upon these. This section aims to understand everything ranging from the performance of the model, those variables that are most important in achieving the final prediction and to give an understanding of the errors that the model produces. We performed an analysis of the High Confidence errors and built a new model using the errors as a label, such that we can try to predict which instances are most likely to be in error. This helps us in better characterizing the nature of the High Confidence errors. Thus, this section answers the research question RQ1: *What instances characterize system knowledge.*

6.1. System Knowledge

As explained in chapter 4 a train/test split was made on the dataset. We chose a split of 80/20 percent for train and test sets respectively. This choice was made weighing the size (8584 data instances) of the set as well as the need for a representative test set.

The total set of instances consists of a set of 8584 vehicle inspections. With the split of these we have a Training set of 6868 instances and a Test set of 1716 instances. In the set we have a total of 2830 vehicles that had a violation during the inspection. This gives us a baseline precision for this set of 0.3297.

We use the Random Forest model which is trained on a train set consisting of 6868 vehicles. For the model we used a total of 1500 trees and we tuned, using a tunegrid, the mtry to be around the optimal value of 18. We do this by choosing the best performing mtry in the range of $\sqrt{FeatureCount}-2$ and $\sqrt{FeatureCount}+8$. The square root of the amount of features has shown to be a good optimal for the mtry variable as presented in section 2.2.2. For the amount of trees to aggregate over(i.e. ntree) a default value would be 500 trees, we found that a tree count of 1500 gives the best performance for our model for this reason we set ntree to 1500. The model was trained using a 10 fold cross-validation. As discussed in section 2.4 the model was optimized for Precision. The final average precision of all folds is 0.5679999. The AUC score is 0.4496914.We have an Out-of-Bag(OOB) error estimate of 32.25%. For the High Confidence errors we saw that the values in the top 20% had a total average magnitude of 0.70435, whereas for the top 5% this was 0.78047.

Table 6.1 gives a summary of all metrics and their values.

Metric	Value
Precision	0.56800
AUC	0.44969
OOB error	32.25
HCE 20	0.70435
HCE 5	0.78047

Table 6.1: Table displaying the performance metric values for the First Model

6.1.1. Variable Importance

We performed an analysis of the model of the first iteration to get an insight into what variables of the model contribute to the prediction score.

We used the trained model to make new predictions on the test set of 1716 vehicle instances. Figure 6.1 shows the distribution of the probability scores of all the predictions made.

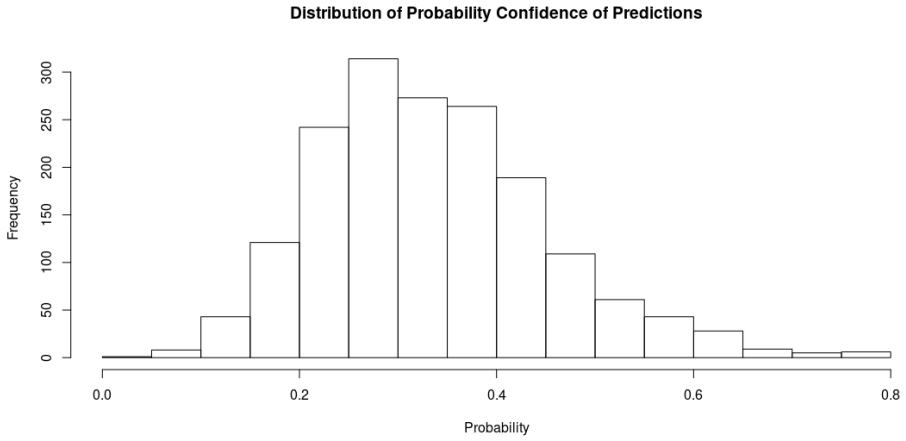


Figure 6.1: Distribution of Probability Scores

The entire model is trained on a set of 137 variables, of these we present the top 100 variable importances in figure 6.2. From this figure, we can see that there is a set of variables with a large variable importance, after which there is a long tail with variables that are only contributing weakly. For this reason, we choose to focus in on only the top 20 most important variables, these are displayed in figure 6.3.

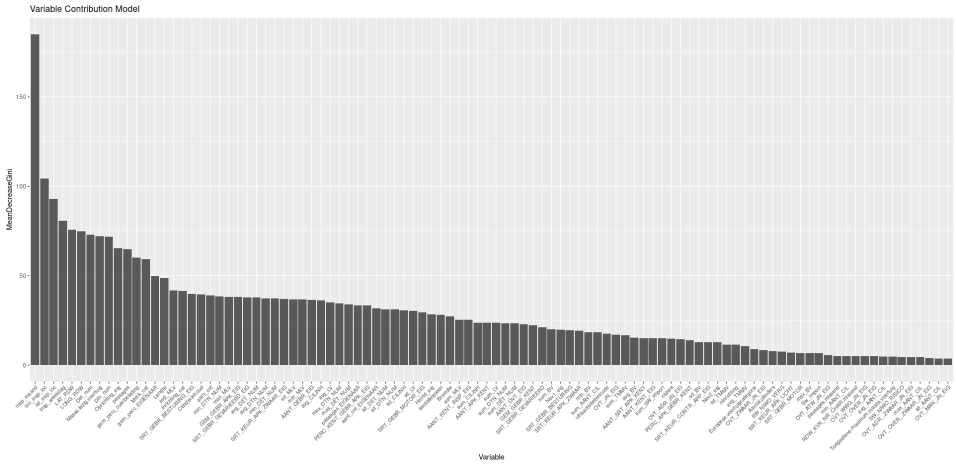


Figure 6.2: Variable Importance as a mean decrease in the Gini index of each variable of the model

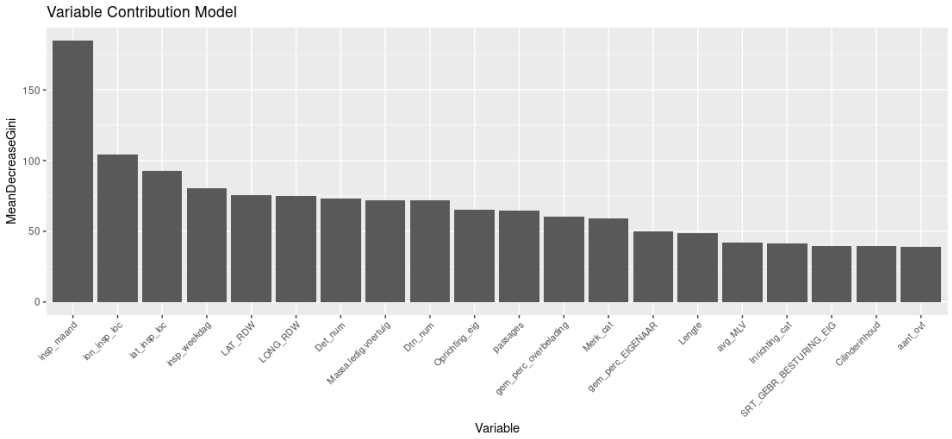


Figure 6.3: Subset of the top 20 most Important Variables

The variables are given in their Dutch and abbreviated names. A description of these variables is given in table 6.2. These descriptions are listed in order of variable importance. From the variable importances we can see a trend of the model putting a strong emphasis on the location and time that the inspections were performed. Additionally there is also a significant amount of variables related to the mass of the vehicle and the percentage overweight of the owner/vehicle. These variables are thus good candidates for a more thorough analysis.

Variable Name	Description
insp_maand	The month in which the inspection was performed
lon_insp_loc	Longitude coordinate of the inspection location
lat_insp_loc	Latitude coordinate of the inspection location
insp_weekdag	The day of the week on which the inspection was performed
LAT_RDW	Latitude coordinate of company at which the vehicle is registered
LONG_RDW	Longitude coordinate of company at which the vehicle is registered
Det_num	Years from first accession of vehicle to NL to the inspection date
Massa.ledig.voertuig	The mass of the empty vehicle
Dtn_num	Years from registration of vehicle to RDW owner to the inspection date
Oprichting_eig	The age of the company at the time of inspection
passages	Amount of passages of the vehicle over a WIM location
gem_perc_overbelading	The average percentage of overweight measured for the vehicle on the WIM location
Merk_cat	The Brand of the vehicle
gem_perc_EIGENAAR	The average of overweight measured for the owner of all vehicles on the WIM location
Lengte	The length of the vehicle
avg_MLV	The average mass of all vehicles on the name of the RDW owner
inrichting_cat	The vehicle type
SRT_GEBR_BESTURING_EIG	Vehicle maintenance problem
Cilinderinhoud	Cylinder capacity of the vehicle
aant_ovt	Amount of measured violations for the vehicle on passage of the WIM location

Table 6.2: Description of the top 20 most important variables, listed in order of importance

6.1.2. Model Errors

As we know the true values for the test set, we can see whether the predictions that the model made were correct or not. We perform a check for all predictions to compare the predicted label of violator/compliant to the true label. All predicted labels that differ from the true label will be classified as being in error.

On the Test set we have a total of 534 Errors out of the 1716 predicted instances. We plot the confusion matrix for the test set in figure 6.4. We can observe a preference for the model for predicting instances as being compliant (Dutch: nalever) rather than violator (Dutch: overtreder).

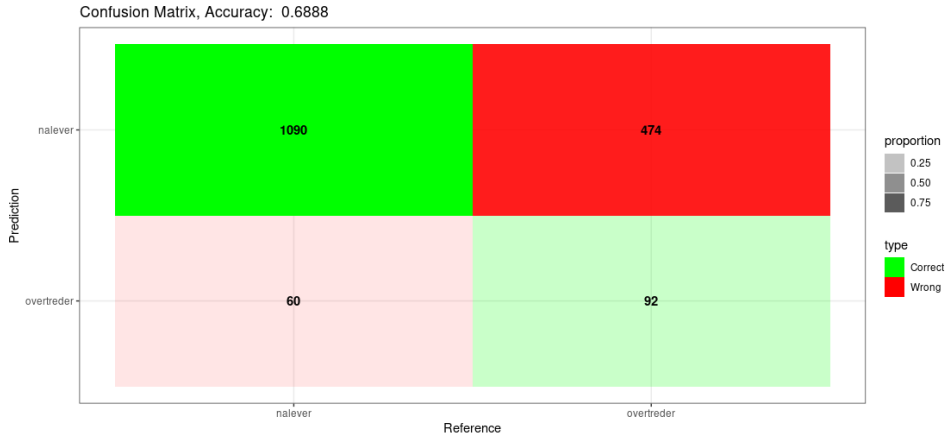


Figure 6.4: Confusion Matrix for Predictions made by the model on the Test Set. Prediction are shown for (Dutch: nalever) and Violator (Dutch: overtreder) vs. the True label values

We want to analyse what variables contribute to the occurences of these errors and how these relate to the Test set and the set of violators within the set.

We focus in on some of the most important variables featured in table 6.2. We first look at the variable that is ranked the highest by a significant amount: The Month of the inspection. In figure 6.5 we show a distribution of the months of the instances. We show both the months for the entire test set as well as the errors overlaid in red and the violators in blue. We see that the distributions for both the errors as well as the violators follow quite closely the entire set, with only some months with small deviations. When we look at the fractions of the errors vs the total set in figure 6.6 this becomes even more clear. From the plot you can see that most bars are even with only slight deviations for months such as April, June and September.

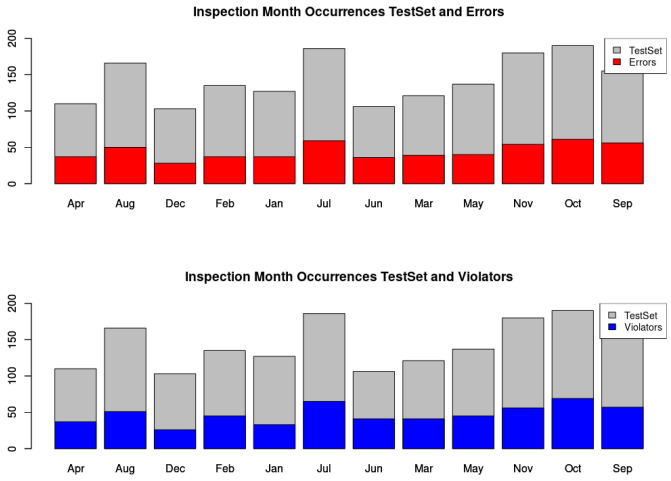


Figure 6.5: Plots for Month Errors (Red) and Violators (Blue)

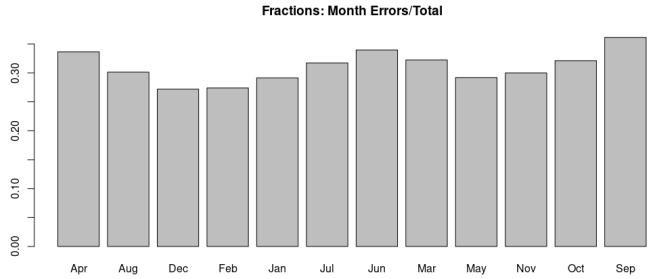


Figure 6.6: Plot for Fractions of Month Errors vs Test set

Doing the same for the other temporal variable: the Day of inspection, we get the distribution in figure 6.7. Plotting the fractions in figure 6.8 shows a clear outlier for inspections made on Saturday for errors to occur. Looking at the full plot, this gives us the insight that there is a very low number of inspections on the weekend days, which can likely contribute to the uncertainty of the model for these days.

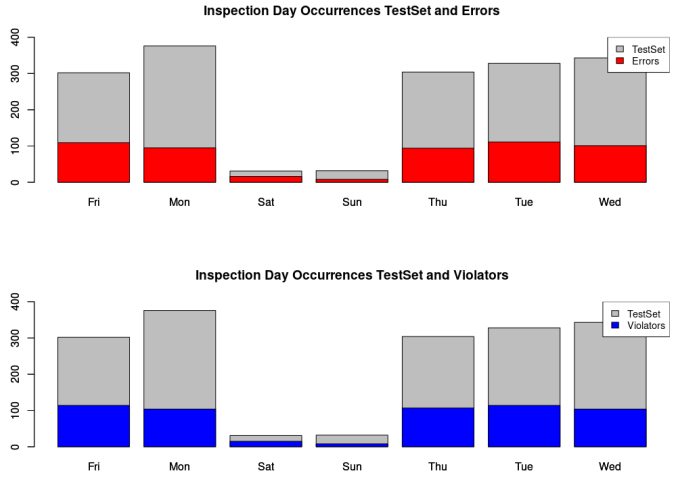


Figure 6.7: Plots for Day Errors (Red) and Violators (Blue)

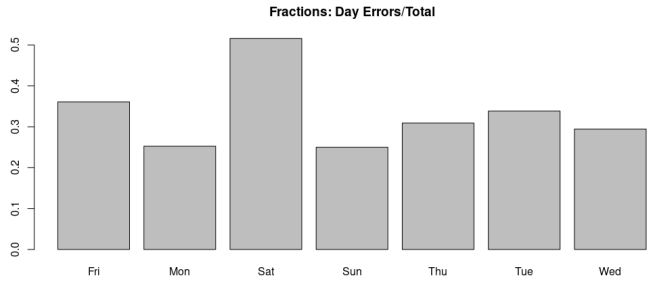


Figure 6.8: Plot for Fractions of Day Errors vs Test set

Next, we want to look at the location aspects which have also been ranked as highly important for the prediction. In figure 6.10 we plot the locations of where the inspections were performed. The plot is in the form of a heatmap showcasing the latitude and longitude information. As a reference, to get a better sense of what locations the plots are showing, we present again in figure 6.9 the map of the Netherlands with the WIM locations, which often also coincide with important inspection locations.

From the plot of the inspection locations we can see that there are two hot spots for inspection locations, one around the Zwolle area and one around the West/Randstad area. The latter is likely also caused due to the larger density of WIM locations in the west/Randstad area as can be observed in figure 6.9. From the density map showing the errors in the top right, we can see that the errors follow closely also the distribution of the test set, while we can see that for the violators (as seen in the bottom right) there is a slightly larger amount of occurrences in the eastern Zwolle hotspot. In figure 6.11 we show a Ceteris Paribus plot, further highlighting the relation between the prediction score and the hotspot locations. This plot shows the effect on the prediction score by only changing the values of the latitude and longitude of the instances.

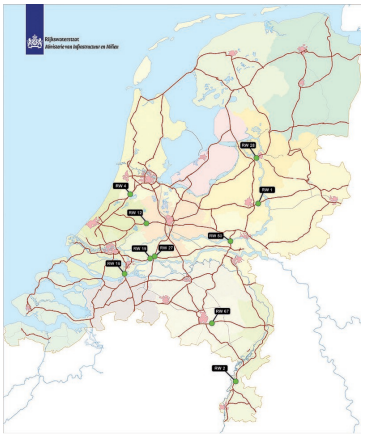


Figure 6.9: Map of The Netherlands showing locations of WIM passages

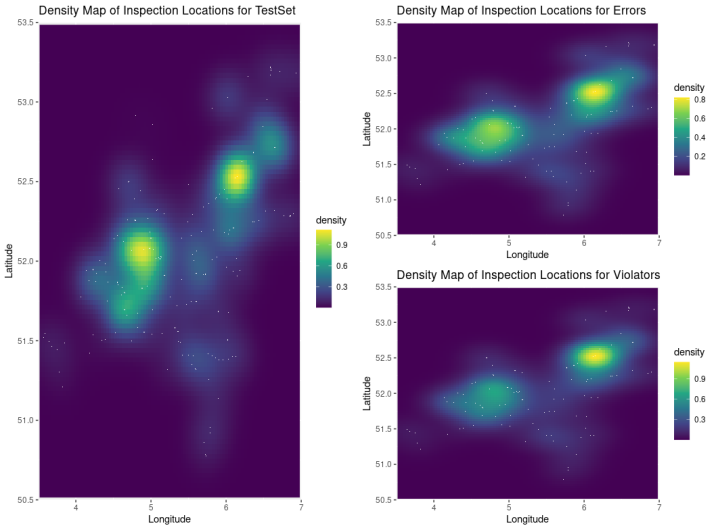


Figure 6.10: Plots for Inspection Locations Errors (Top Right) and Violators (Bottom Right)

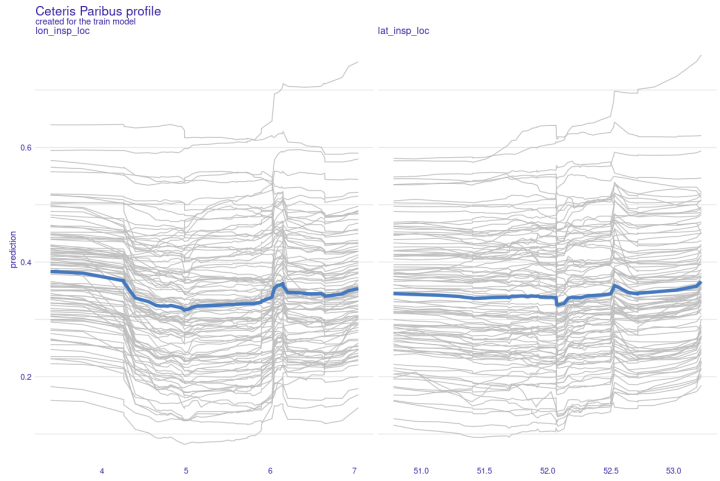


Figure 6.11: Plot showing the risk prediction score relative to the longitude and latitude of the inspection location

We ran the same analyses for most of the remaining variables in the top 20 from table 6.2. These can be found in appendix A

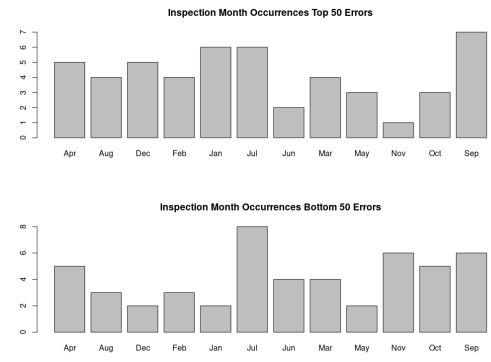
6.2. High Confidence Analysis

We have shown some of the distributions of errors and how they relate to the distribution of the set as a whole. However, as our aim is to characterize and mitigate *High Confidence* Errors, this is what we spend this next section on.

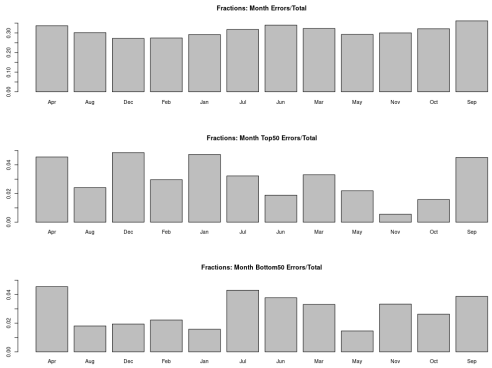
For the selection of our high confidence errors, we want to take a subset of the errors which are produced with a sufficiently large confidence score such that it would be troublesome for the trust of the model if these occur in a real scenario. We show the distribution of the confidence scores for all errors In figure 6.12. We see that the distribution is slightly more of a negatively skewed normal distribution than the positive skew that we saw in 6.1. Since we are dealing with a relatively small set of errors (534), we also have a relatively small number of High Confidence Errors. We define our high confidence errors as being on both sides of the distribution, meaning both the False Positive (High probabilities) instances as well as the False Negative instances (Low probabilities). Since our model is naturally more inclined to predict vehicles as compliant, we have a larger amount of High Confidence errors on the low probability end. Taking into account the low amount of HCEs we found that the (double sided) Top 50 and Bottom 50 errors to be an appropriate cut off. This cutoff accounts for all errors that occur with a confidence score of more than 1 standard deviation from the mean. We highlight these parts as red in figure 6.12. Where the left hand side are the Bottom 50 errors and the right hand side the Top 50 errors. *These* are the High Confidence Errors that we analyse.

Figure 6.12: Distribution of Probability Scores for all Errors made by the model. The red areas denote the High Confidence Errors that we want to mitigate.

Similar to the approach in the previous section, we will start by analyzing some of the most important variables. In figure 6.13a we plot the top 50 and bottom 50 high confidence errors for the month variable. We also plot the fractions of the errors vs the total test set in 6.13b. From the plots we can clearly see a bias for high confidence errors to occur more often for certain months such as April and September. This is in agreement with the findings that we also found in the previous section. We make a similar plots for the Day variable in figures 6.14a,6.14b we note an increased occurrence of errors for inspections on weekend days.

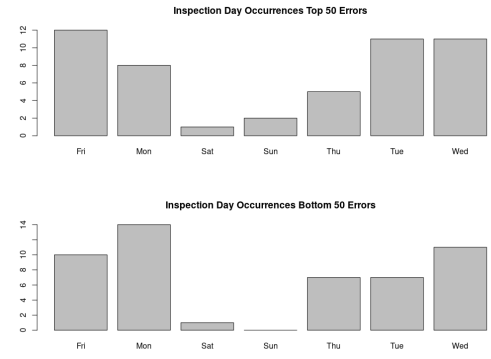


(a) Plots for Top 50 and Bottom 50 Errors

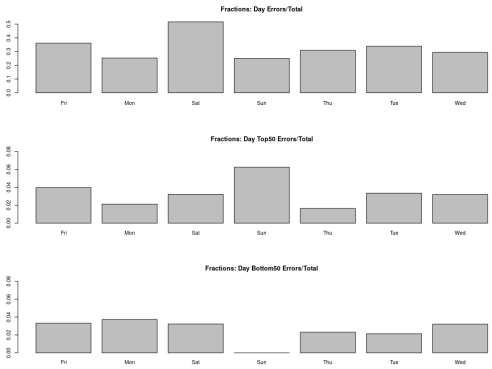


(b) Plots for Month Fractions

Figure 6.13: Plots for Month High Confidence Errors



(a) Plots for Top 50 and Bottom 50 Errors



(b) Plots for Day Fractions

Figure 6.14: Plots for Day High Confidence Errors

We next make the heatmap plots for the top 50 and bottom 50 errors of the inspection locations. These plots show an interesting finding that the top 50 errors are mostly made in the hotspot around Zwolle where also the occurrence of violators was the highest, while the bottom 50 errors show the opposite pattern, where it shows mostly all the spots not in that area. This shows a clear bias for the model to predict on the locations where a high number of violators has been found.

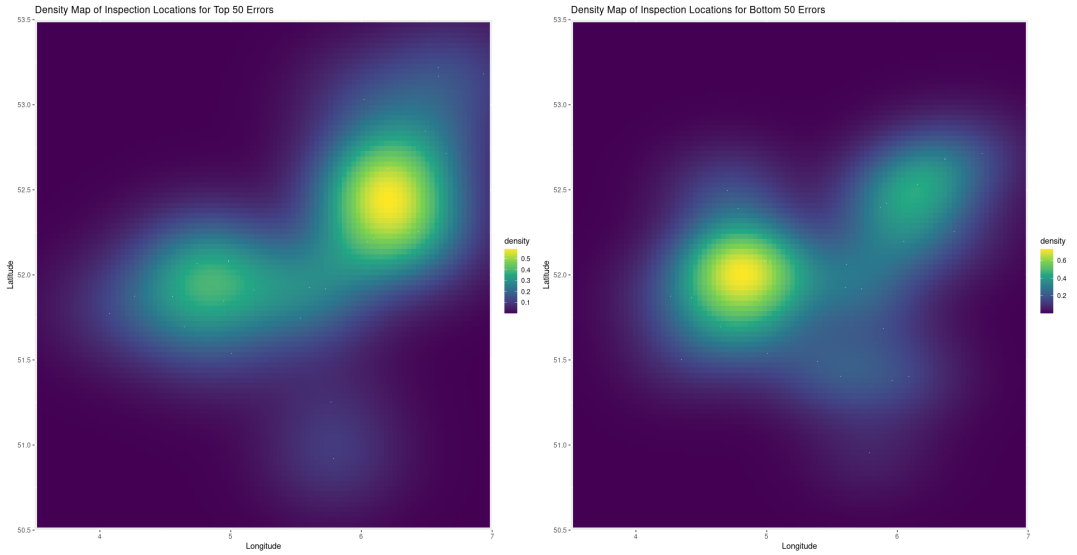
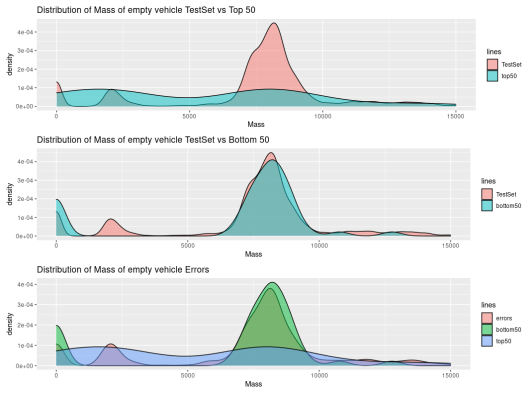


Figure 6.15: Plots for Inspection Locations High Confidence Errors, Top 50 and Bottom 50 Errors

Lastly we want to look at some of the continuous variables. We will focus on the variables *Massa.ledig.voertuig* indicating the empty vehicle mass and *Oprichting_eig* indicating the age of the company. We see an interesting observation for the vehicle mass in figure 6.16a, here the top 50 errors show a mostly even distribution over all vehicle masses, while for the bottom 50 errors the errors follow the test set. The total set of errors also follows the density of the test set. For the age of the company that owns the inspected vehicle in figure 6.16b we see that the top 50 errors happen more often for the younger companies and the bottom 50 errors have a drift towards the midrange, for older vehicles. This indicates the models preference to predict younger vehicles as being violators more often.



(a) Density for Mass of empty Vehicle High Confidence Errors, Top 50 and Bottom 50 Errors

(b) Density for age of the company High Confidence Errors, Top 50 and Bottom 50 Errors

Figure 6.16: High Confidence Error Densities for the Mass of the empty vehicle and the Age of the company

From these analyses we can see that the most important variables to the model are often also the cause of most of the high confidence errors. The model overrelies on these few variables when making predictions, meaning that when the reality is different, the model will wrongly predict these instances with a high confidence. The most clear examples are those of the locations and the age of the vehicle as well as some of the time and weight related variables that we have shown. Analyses for the remaining top 20 variables can again be found in the appendix A.

6.2.1. Predicting Errors

As a last step to finalize our search for those instances that are most likely to cause errors, we want to use a more data driven approach. For this, we want to try to predict, using our found errors, what instances

would be likely to be erroneously predicted. As explained in chapter 4, we relabeled all instances that were in error to have the label "Error" while we labeled all other instances as "Correct" instances. We then trained a new Random Forest model using the relabeled Test Set data as the training data. We do note that because of the size of the Test set, the number of instances that the model has been trained on is relatively small. The hyperparameters for this model were chosen to be mtry of 20 as being the optimal mtry and ntree of 1500, calculated similarly to the original model. The model was trained using 10 fold cross-validation. The original Training set is used as a Test set to evaluate the trained model on. The model has a precision of 0.70225 and AUC of 0.76154. The OOB error estimate is 32.46%. These are summarized in table 6.3.

Metric	Value
Precision	0.70225
AUC	0.76154
OOB error	32.46%

Table 6.3: Table displaying the performance metric values for the Error Prediction Model

We analyse the top 20 variables that contribute to the classification of the errors. These are presented in figure 6.17. We observe that for the prediction of errors, those same variables that were important for the original classification are also the most contributing to classifying errors. This is in concordance with our earlier findings that the model relies on a set of important variables for its predictions, which causes these variables to produce more errors when predictions are made. We see that the impact of the inspection month variable is even more exaggerated in this model. Further, we can observe that the latitude coordinate of the location is deemed less important than in the risk model and also the mass of vehicle is deemed less important than before.

Having created and analyzed our model we now want to make some predictions. We perform our predictions on our original train set of the risk model. This gives us a distribution of probabilities given in figure 6.18. As we are interested in characterizing High Confidence Errors we will compare both those cases predicted as being in error as well as those predicted with a high confidence (26 data instances in total). For this, we choose the cutoff of 0.7 probability confidence of being in error. We will compare for a set of important features the effect of all errors vs those predicted with a high chance of being in error.

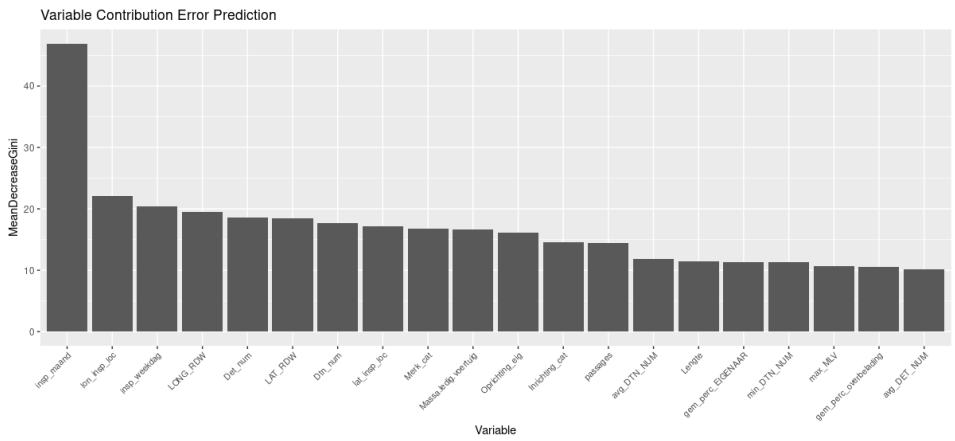


Figure 6.17: Top 20 Most important Variables for predicting Errors

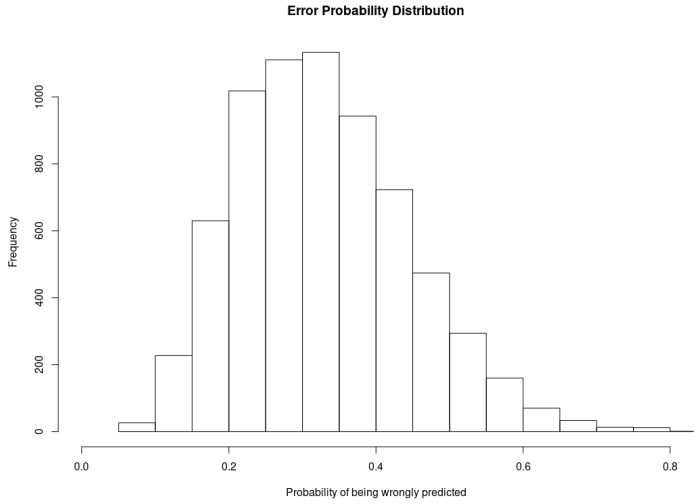


Figure 6.18: Distribution of Probabilities of predicted instances for the Error Prediction model.

We first look at by far the strongest predictor, namely the Month feature. this feature is plotted in figure 6.19. Where we show both the error predictions as well as the high chance predictions. We see a clear spike at the July month, whereas contrary to what we have seen before, the months of April and September do not occur often in the set of high error chance instances. In figure 6.20 we show the weekdays plotted for the error instances. Here we note no large differences in the occurrences, however, whereas previously we saw a large impact on those instances that were on weekends, here this effect is less pronounced. We also looked at the effect of the Brand of the vehicle on whether the model would deem it an error. This is visible in the plot in figure 6.21. Here we observe that mostly SCANIA vehicle and vehicles of which the brand is unknown (Dutch: onbekend) are highly contributing to the errors.

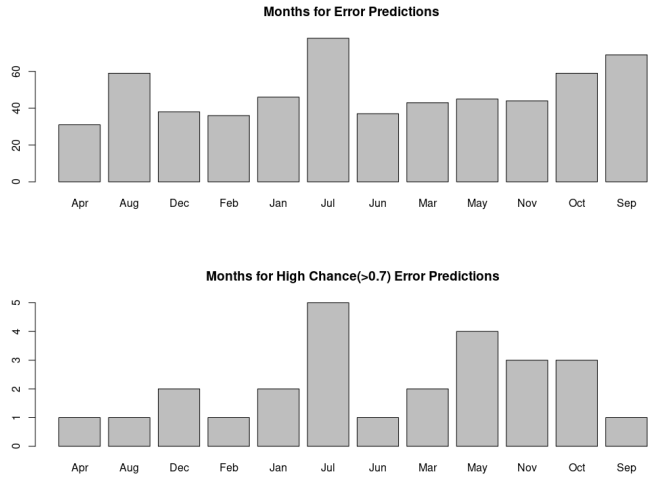


Figure 6.19: Plots for Month distribution for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

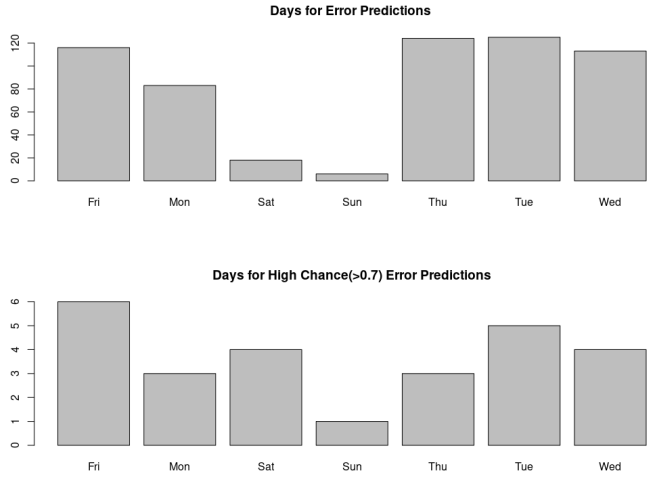


Figure 6.20: Plots for Day distribution for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

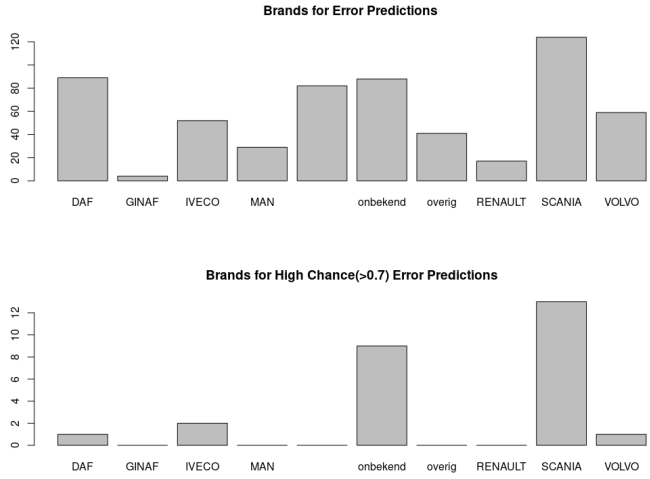


Figure 6.21: Plots for Brand distribution for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

Next, plotting the location variables in figure 6.22 we note that the hotspots that we previously observed are again seen as occurring most often for the error instances. An interesting observation that we did make is for the RDW owner locations in figure 6.23. Previously, in figures A.12, A.13 we noted no real impact from this variable on whether an instance is in error or not, both for the normal errors as well as the High Confidence errors. However, in this model we see that there is a large hotspot for the High chance errors in the eastern North-Brabant/Limburg area.

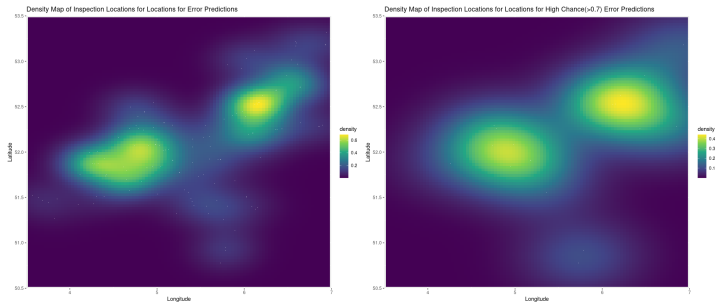


Figure 6.22: Heatmap Density Plot for Location distribution for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

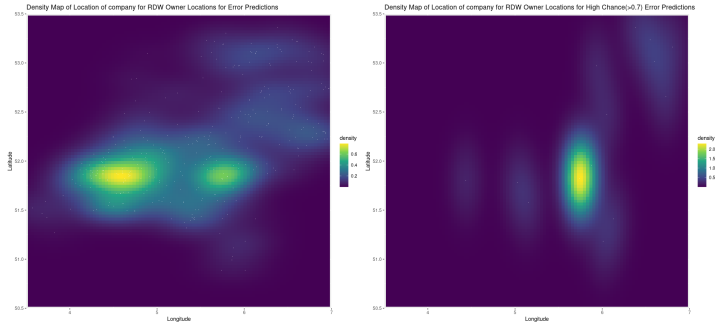


Figure 6.23: Heatmap Density Plot for RDW owner locations for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

Lastly, we present two density plots, namely of the mass of the empty vehicle in figure 6.24 as well as the age of the company in figure 6.25. The Mass of the empty vehicle seems to follow the same pattern for both the high chance errors as well as the normal errors. For the age of the companies we see a clear peak for the high confidence errors for the older aged companies.

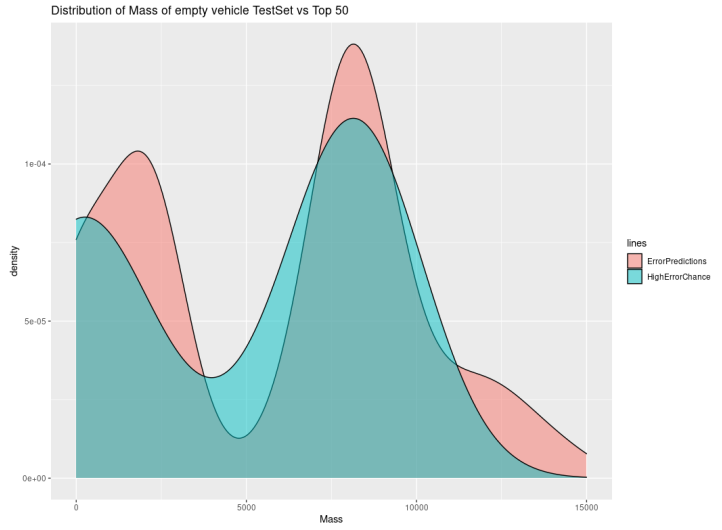


Figure 6.24: Density Plot for Mass empty vehicle for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

Figure 6.25: Density for age of the company for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

The plots of all remaining error predictions can be found in the appendix B. In this chapter we have seen that there is a large reliance on a set of a few essential variables for the model. These most important variables are also one of the main causes for the occurrence of High Confidence errors. Common variables that have shown bias for the model are those of time, place and weight. The Month feature especially has shown a strong bias for predicting whether a vehicle is in violation. This can likely be due to some of the seasonality of the work that the inspectors do, but this does not exclude other factors that can be at play for this bias to occur. Furthermore, we note a strong overreliance on some hotspot locations where a large number of violators are found. Especially the Zwolle area and the Randstad area show two large hotspots, where the large number of violators in the Zwolle area is particularly prone to cause High Confidence errors. These variables essentially tell us more about the practice of the inspectors rather than the actual risk of the vehicle being in violation. These findings give a sense of what areas in the data need more attention and potential changes to the feature set. More on this is outlined in section 8.

6.3. Further Domain Analysis

Repeat Violations An interesting question in the context of risk assessments is the question of do the vehicles/owners perform subsequent violations, meaning that some are at an increased risk of relapse into violating behaviour. To see if there was any effect we first wanted to see whether the data contained a sufficient amount of repeat occurrences of inspections. We plotted the number of repeat occurrences of vehicles in the dataset, as can be seen in figure 6.26. We see that only a total of 676 out of the 8584(7%) have more than one inspection occurrence and of this only 118(1%) are repeat violations. Less than 100(1%) vehicles have been inspected more than 2 times. This means that checking for repeat violators will not have an impact on the model output.

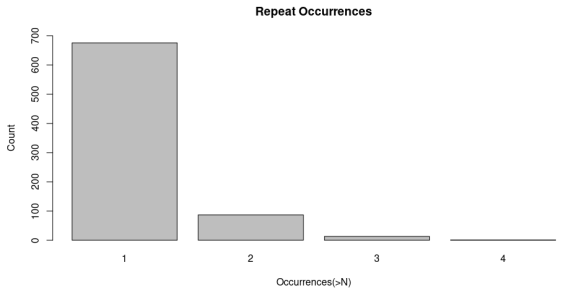
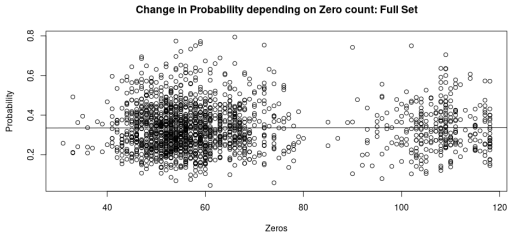


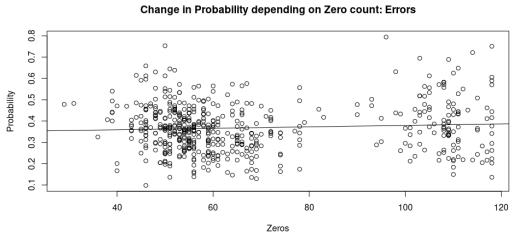
Figure 6.26: The amount of Vehicles that have repeat occurrences, meaning more than N occurrences

Zero Values The annotation of the data instances is for the most part done by the inspectors. Besides this, other sources are also used to gather the data for the model. A more detailed description of all these sources can be found in section 2.5. Due to the human factor of annotation and the disparate data sets that are joined together, there are still many instances that have missing values. Due to the way Random Forest models are designed, this is not inherently an issue, however, the missingness of data may still have an impact on the performance.

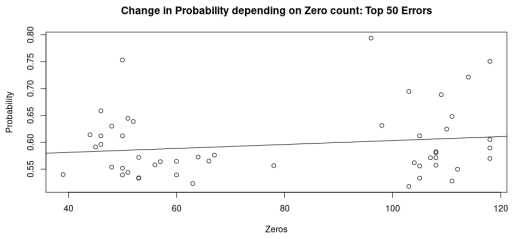
We analyze the effect of missing variable values on the model errors. Missing values can be either numeric, as a 0 or null for other variable types. As the majority of variables in use are of a numeric nature we will only focus on the zero values. To check for the effect of data missingness, we count the amount of zeros for each instance of the full dataset and of the errors, and observe whether the outcomes of these instances' predictions show a significant difference when looking at the total zero count. In figure 6.27 we plot the probability chance of a violation vs. the total amount of zero's for an instance. From a visual we can see that in figures 6.27a and 6.27b for the full set and the entire set of errors there does not appear to be a significant linear effect. For the top 50 and bottom 50 High confidence errors in figures 6.27c and 6.27d there does appear to be a small effect of that the zero count increases the probability of a vehicle being classified as violator.



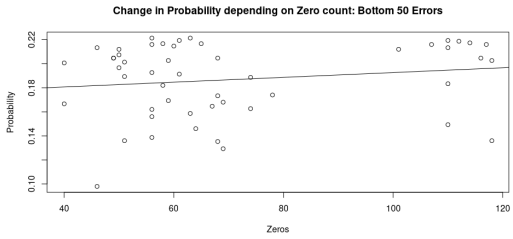
(a) Change in Probability depending on the total amount of Zero Values for an instance: Full Set



(b) Change in Probability depending on the total amount of Zero Values for an instance: Errors



(c) Change in Probability depending on the total amount of Zero Values for an instance: Top 50 Errors



(d) Change in Probability depending on the total amount of Zero Values for an instance: Bottom 50 Errors

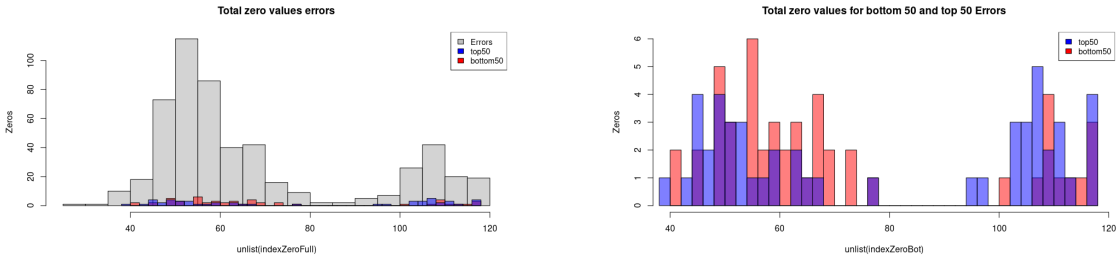
Figure 6.27: Plots showing effect of total amount of Zero Values on the prediction probabilities

To see how significant this effect really is we want to perform a statistical test to see if the distributions differ. We perform a Student's t-test to essentially check whether the errors come from the same distribution as the whole set, by checking whether they have the same mean. This gives us the p-values shown in table 6.4. A p-value of < 0.05 is significant, showing that the errors come from a different distribution, showing that there is a significant effect.

Set	Mean Zeros	p-value ttest
Full Set	65.2	NA
Errors	68.0	0.0065
Top 50 Errors	78.9	0.0015
Bottom 50 Errors	71.2	0.0949

Table 6.4: Means of total amount of zeros per set with the p-values for the t-tests

From the results we see that there is a significant effect for the entire set of errors as well as the top 50 errors. However, here it is good to note the key assumption of the t-test, which is that the data should be normally distributed. To check for this we first plot the distributions in figure 6.28.



(a) Distribution of Zero Values for Errors

(b) Distribution of Zero Values for Top and Bottom 50 Errors

Figure 6.28: Distributions of Zero Values for Errors

When looking at the distributions in figures 6.28 we can see that these are most likely not normally distributed, which is an assumption for the t-test to work. To confirm this, we plot a Quantile-Quantile(QQ) plot. For it to be distributed normally, the quantiles should show approximately a straight line. We plot the QQ quantiles in figure 6.29.

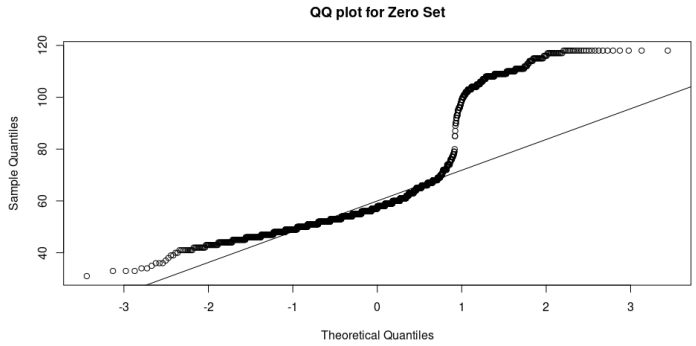


Figure 6.29: Quantile-Quantile plot showing the quantiles for the data vs the normal distribution quantiles

The quantiles show a large deviation at the tails. Additionally, performing a Shapiro-Wilk normality test gives a p-value of 2.2e-16, meaning we can reject the Null-hypothesis of the set being normally distributed. Since the zero values are not normally distributed we need to perform a non-parametric test to see if the effect is significant. For this we perform a Wilcoxon rank sum test(or Mann-Whitney). This test is non-parametric and checks essentially whether two independent sample groups originate from the same distribution. The p-values for these tests are shown in table 6.5

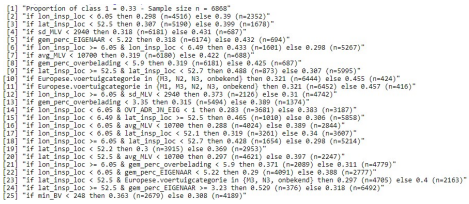
Set	Mean Zeros	p-value wilcox
Full Set	65.2	NA
Errors	68.0	0.0634
Top 50 Errors	78.9	0.0204
Bottom 50 Errors	71.2	0.0342

Table 6.5: Means of total amount of zeros per set with the p-values for the Wilcoxon tests

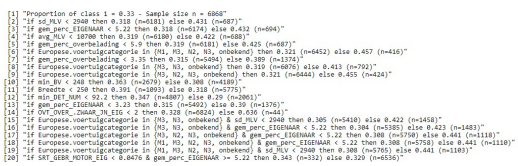
Here we see that the p-values for the Top and Bottom errors are sufficiently low for us to conclude that they are from different distributions. This means that the presence of zeros or missingness of data does have an effect on instances being high confidence errors.

SIRUS Decision Rules In section 3.4.3 we mentioned SIRUS as a method that can be used to help interpret the choices of a model. SIRUS provides the user with a set of decision rules which are extracted from the nodes of the Random Forest. We created a SIRUS model and ran this on the dataset. We use SIRUS's built in cross validation approach to come to the optimal set of decision rules. This gives us a set of decision rules displayed in figure 6.30a. We can see that an often repeating rule is one based on the location of

the vehicle. We have seen the location features to be important in characterizing errors. To have a clearer picture of what other variables are essential, we chose to remove the location variables and rerun the optimal SIRUS model on this set. This gives us the set of decision rules in figure 6.30b. Looking at this set of decision rules we make some notable observations that differ from the findings in section 6.1.1. We see that the SIRUS model puts a greater emphasis on the overweight related variables such as *sd_MLV*(standard deviation),*avg_MLV* and *gem_perc_EIGENAAR*, *gem_perc_overbelading*. Further we note the inclusion of *Europese.voertuigscategorie*(English: European.vehiclecategory) in the set of rules. This variable was not previously seen as being highly influential, but in the SIRUS model it repeats many times, also combined with some of the weight-related variables, indicating some relation between vehicle categories and overweight violations.



(a) Optimal set of SIRUS decision rules with location



(b) Optimal set of SIRUS decision rules with location variable removed

Figure 6.30: Optimal set of SIRUS decision rules

7

Expert Session Findings

In this chapter we present the results from the sessions that were performed as part of the the case study as outlined in section 5.5. The significance of these findings for the model and the changes that were made to the model as a result, are discussed in chapter 8.

Prior to the sessions, an initial Team Review meeting was held with inspectors. Here we already noticed some points to pay attention to, namely that the inspectors can have some triggers. Examples given were trucks that keep their windows open can hint at the driver working over-hours. Another example given was trucks that have valuable cargo, such as transport of cars are potentially less at risk for being overweight than for example sand trucks. With this idea of inspector triggers in mind we went into the Expert sessions.

7.1. Exploratory Session Findings

Here we present the results from the exploratory sessions. In total, 2 exploratory sessions were held. 1 day-in-the-life sessions with 4 ADR inspectors, during which we followed the inspectors in their evening session where we attended from the afternoon until the early evening. For the second exploratory session, one 3-hour long interview session was held with 7 Road and Transportation inspectors. For the first Exploratory session we had a specific goal. The goal being to assess if a risk model is applicable to the inspectors' working practice. For the second Exploratory session we performed a semi-structured interview session with the Road and Transportation inspectors. Here our goal was to better understand what indicators the inspectors look at when performing their selection of vehicles for inspection. A more elaborate explanation of the sessions was given in section 5.5.

7.1.1. Day-in-the-life session: ADR Inspectors

During the day-in-the-life session we, as the data scientists of the ILT, spent the day with the ADR inspectors on their location in Geldrop. Here we could see how inspections are performed. The selection of the vehicles on the highway was not observed due to corona measures that were in place at the time. These measures restricted multiple people from being in a vehicle at the same time. This is a pity as the model's main focus is on aiding the selection of vehicles.

During the inspection of the vehicles we were able to ask questions on what the inspectors look at when making their selections while also giving the inspectors ample opportunity to tell us what they find important during their inspection of the vehicle.

During the day the inspectors inspected around a dozen or so vehicles. The majority of the vehicles that were inspected were compliant vehicles with only some remarks or minor issues for some. There was only 1 vehicle in violation on multiple levels. The nationality of the vehicles was well distributed in the amount of Dutch and foreign vehicles. Some examples of issues that were noted: One vehicle had an issue where the freight was not properly secured. Another vehicle had small containers that were not properly secured on the outside of the vehicle. One vehicle had two tankers close to each other, which made reading the tanker information difficult.

The vehicle that was in violation had many issues. The vehicle was transporting a container with an old and rusty appearance. Paperwork of the driver was not in order, with no correct paperwork of what freight was

in the container. Furthermore other safety requirements, such as the necessity for a working fire extinguisher were not met. The container had to be opened and the freight inspected, the driver was issued a fine.

General Observations From the session we learned that the ADR inspections and the other Road and Transport inspections work separately from each other. The ADR inspectors do not put their focus on violations regarding laws such as rest and driving times manipulation or vehicle overweight. The ADR inspectors are concerned with the cargo that the vehicle is transporting. There is no way to precisely know the contents of a vehicle until after the vehicle has been stopped. This means that in the selection process of the vehicles the inspectors mostly look at whether the vehicle has an orange Kemler ADR indicator sign, as well as if it has any stickers indicating that dangerous substances are being transported. A description and example[1] of such signs is given in figure 7.1. The inspectors explain that they go from their experience to decide on which vehicle they select for inspection. When asked if they can pinpoint certain aspects that they focus on they said it is difficult for them to say.

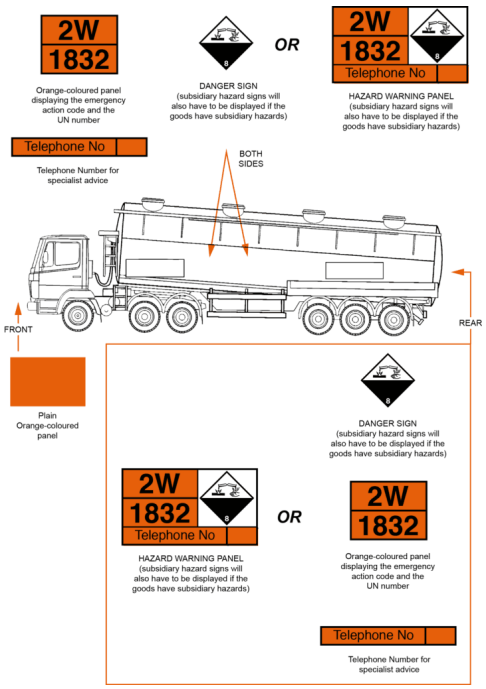


Figure 7.1: Graphic displaying the various ADR signs that a vehicle transporting dangerous goods should have.

It is important for the inspectors to know what type of cargo the vehicles are transporting. This is usually easier to see for tank container vehicles (which have clear stickers and details on them) as opposed to regular logistics trucks. At the same time, the tank containers usually carry more potentially dangerous freight, which may be a bigger risk. Regular trucks, that only show their orange ADR sign, do not always show what kind of substances they are carrying. If there is an indication for further inspection (such as leaking fluids) this is performed. The inspectors do not necessarily annotate many details about the appearance of the vehicle. However, if there are problems or violations of the vehicle, there will also be checks for further issues. This means that there is more data generally available for those vehicles that were in violation, which may give a skewed prediction when trying to model these. The responsibility for the freight may sometimes be for the producer only and sometimes for both the driver and producer Some aspects of the vehicle that are of importance are country specific, such as French vehicles that require stickers for blindspots.

Selection of vehicles happens a lot on sight and experience. No real rules were specified by the inspectors when queried on this. Some aspects the inspectors do pay attention to are:

- Driving behavior: rushed driving or slow driving
- Differing ownership of container or trailer

[1]Source: https://www.hibiscus-plc.co.uk/adr-plate-adr-panel-hazchem/

- Companies known to have had violations in the past

Some amount of random selection is also at play since the inspectors pick those vehicles that drive by at the moment that they are driving on the road.

The findings of these sessions are ofcourse only a small slice of the entirety of the work and focus of the inspectors, there is simply no way to capture all this knowledge from one session, either way the session helped us in getting a better grasp of the practice through the findings explained above.

7.1.2. Interview Session with Inspectors

The core part of the sessions were the interview sessions held with the Road and Transportation inspectors. As part of the first Exploratory session, we wanted to get a deeper understanding of the reality of the inspectors and what aspects they pay much attention to during their inspection practice. For the interview sessions we chose to focus on those inspection types that are most often violated on and which the inspectors check the most for. These inspections we focused on were the following (for an explanation of these see section 2.1):

- **Rest and Driving Times/Tachograph Manipulation**
- **Vehicle Overload**
- **Cabotage**[2]

The Road and Transportation inspectors have a tough job, where with little information they have to make judgements about vehicles and have to spend their time and resources in inspecting these. For this they largely rely on their expertise in the field of what locations they should be present, when they should be present there and what to look for in a potential violator.

For each inspection type we asked the inspectors to brainstorm to give their indicators and to afterwards rank these indicators by their importance. The inspectors gave a plethora of indicators that could be seen as indicative of a non-compliant vehicle. The assumption here is that even though there can be personal bias and experience involved in each of these indicators, the overall sum of inspectors and their indications provide a reasonable insight into the reality of a non-compliant vehicle

The following is a list (in no specific order of importance), for each inspection type, of what indicators are of interest. This is by no means an exhaustive list, merely a set of indicators that the inspectors deemed important enough to mention during the session.

- Rest and Driving Times
  - Combination of type of freight with inspection location and time/date
    - Foreign cars on Monday near Flower auction on Monday (how do they get there so fast)
    - Ferry transport to Scandinavia/England
    - Near the Eastern border transport
  - Known non-compliant companies
  - The layout of the vehicle, e.g. advertisements or blank trailers.
  - Type of tachograph: analog or digital
  - Container transport (changing containers wrongly seen as rest)
  - Semi trailers with no owner/owner unclear
  - The appearance of the vehicle. e.g., rusty or well kept, but also different personalizations/modifications made to the vehicles
- Vehicle Overload
  - Number of axles: less axles means less chance of overload
  - LHV(longer heavier vehicle), vehicle with all axles down

[2]Cabotage is defined as a restriction of the operation of transport services within a particular country. In essence this means that if a foreign EU-member truck comes into the Netherlands loaded with freight, and transports freight between two points in the border, it is known as cabotage. After delivering its freight it can do a maximum total of 3 routes within the Netherlands after this the 4th route has to be an international route. This can be checked through the Tachograph data. Source: https://www.ilent.nl/onderwerpen/cabotage

  - How does the load appear to be distributed
  - Visual aspects indicating potential overload e.g. sagging vehicle, deformation of tires
  - Sector of transport, agriculture and construction
  - Seasonality of certain types of freight

- Cabotage
  - Monitoring of parking lots with people overstaying
  - Location of the transport: the port area can have higher risk
  - Truck and trailer with distinct types of license plates
  - Type of transport (container/ferry/packages)
  - Foreign carriers
  - Trailers of known unaccompanied (RoRo) vehicles
  - Dutch company with foreign location
  - Visual aspects: Blank truck, Is the truck the owner of the trailer/transport? i.e how well does the trailer fit the tractor truck

7.2. In Depth Session Results

Following the Exploratory sessions we held the In-Depth sessions. The aim of the In Depth Sessions is to further narrow in on certain points of interest for the model and to find those features from the data that are the most useful and contribute to the prediction the most. Based on the information and feedback that we have received we question the inspectors on certain areas. Here we present the results obtained from the session performed with the inspectors.

7.2.1. Interview Session with Inspectors

During the In-Depth interview session we questioned in total 4 Road and Transportation inspectors on the addition of data from 5 areas. These areas are:

- **Location data of the vehicles** (in particular WIM[3] passage location data)

  Vehicles pass through WIM locations which collect lots of data on how many times the vehicles pass through which locations, this can give indications of riskful driving patterns, as such give a risk profile for the individual vehicle.

- **Driving Times and Freight**

  The Time and Freight factor was already seen to be important in the Exploratory session. We want to see how to use the data to better represent these in our model.

- **Vehicle Ownership**

  The ownership of the vehicle and the violations can be different for vehicles. All violations are put under the name of the owner, also when a vehicle is leased the violation is put on the lease company's name. Also the ownership of the truck and trailer can differ.

- **Corporate Group Structures**

  Different corporate structures can be created especially for the purposes of saving costs, such as hiring foreign workforces which can provide cheaper labor. Moreover, license plates and vehicle licenses can belong to different companies with some structures.

- **Vehicle Maintenance** (APK)

  We want to know whether the maintenance data of the vehicles can be relevant for the risk profile, and if so, which aspects of vehicle maintenance are most indicative.

[3] a collection of 10 locations[36] on the Dutch roadways where automated weighing systems are used.For more info see 2.5.2

We want to learn from the inspectors how this data can contribute to the classification. The following is a list of points of interest resulting from the In-Depth session:

- Location
  - There are only 10 WIM passage locations, which means that there are many roadways not covered, which can lead to tunnel vision on these WIM locations
  - Locations need to be expanded, especially in the north end of the country, where there is little coverage.
  - Important to look at the effective use of smart tachograph[4]
- Driving Times and Freight
  - Certain specific periods and holidays combined with freight type.
  - Certain Time intervals of the day can prove interesting for the model.
  - Freight companies that store information about their journeys give potential for more data. Hard to get specific information
- Vehicle Ownership
  - Leasing of vehicles happens often 40% of vehicles however it is not necessarily indication of risk.
  - Temporary renting of vehicles can indicate risk.
  - Important for data collection is to also annotate the license plate at the back.
  - Older companies have more capital to have their own vehicles, starters do not usually.
  - Lack of data about leasing companies.
  - Vehicle combinations change per sector.
- Corporate Group
  - No differences in risk for specific nationalities of foreign enterprises.
  - The nationality of a vehicle does not necessarily have to be the same as that of a company.
  - The existence of foreign parent companies is not of interest for the inspectors.
  - Dutch companies are under Dutch government jurisdiction, meaning sometimes stricter control.
- Vehicle Maintenance(APK)
  - Vehicle maintenance is not always an indicator of risk.
  - This indicator can be bound to the reputation of the company.
  - Maintenance problems of the bodywork of the vehicle can indicate risk.
  - Structure of the vehicle is of importance

This is by no means an exhaustive list, rather a collection of everything we could cover in the short time of the interview session. How these findings were used to make improvements to the model will be further covered in chapter 8.

7.3. Analysis Session Results

Following the Exploratory sessions and the In-Depth session, we held the Analysis Session. During the session, 8 cases of data instances predicted by the model were presented, 5 High Confidence Errors and 3 correct predictions. These cases were the following shown in table 7.1, shown in the order that they were presented to the inspectors:

The 3 Road and Transportation inspectors were first presented with a view such as in figure 5.3a, showing the data of the vehicle. After they gave their judgement on the data instance they were presented with a

[4]Smart tachograph is an improved version of the digital tachograph that is able to automatically store and send location data using short-range communications technology, such that roadside inspection authorities have an easier job finding non-compliant vehicles

Type	Confidence	Prediction	True Value
Correct Prediction	0.684	Violation	Violation
High Confidence Error	0.045	Compliant	Violation
High Confidence Error	0.914	Violation	Compliant
Correct Prediction	0.014	Compliant	Compliant
High Confidence Error	0.783	Violation	Compliant
High Confidence Error	0.516	Violation	Compliant
High Confidence Error	0.092	Compliant	Violation
Correct Prediction	0.85	Violation	Violation

Table 7.1: 8 cases presented during the analysis session in order of appearance to the inspectors. The inspectors were shown 3 correct predictions and 5 high confidence errors.

view such as in figure 5.3b, showing the model's prediction score and feature contribution. Both views were accompanied with a discussion between the data scientists and the inspectors.

There are certain variables and combinations of variables that the model deems as being high risk. But this does not always reflect the actual risk of a vehicle. We want to discern those variables that are biased from variables that truly indicate risk We look at certain variables that have a large effect on these predictions. We look at what the experts would think and with what reasons. Would they make the same error? This indicates a bias that better reflects reality.

During the session we found that the inspectors gave the same judgements of the vehicles as those of the model for *all 8 data instances*. We also asked them to give their relative confidence of how certain they are. These were surprisingly also similar to those confidence scores given by the model.

From the Analysis sessions some additional points of action could also be extracted as a result of the discussions following the data instances and feature contributions shown. This is not an exhaustive list, but some of those most worth mentioning.

- There are certain dates and years where regulations or devices used, such as Tachographs were changed. These can lead to interesting data shifts.
- More information on the calibration date of tachographs could be included.
- A feature indicating the type of tachograph can prove useful.
- More useful information can be extracted from WIM data, e.g. the maximum speed when passing these locations. As speed violations can indicate a profile of riskful behaviour
- Variable values that cause misinterpretation by the model can be changed, e.g. -1 (or -100) default value instead of 0.

This concludes the findings of the final Analysis session and as such also the chapter on the findings of the Expert sessions. The choice of what findings to include into the model for improvement depends on different factors such as the availability of data and the feasibility of the feature. These results will be further discussed in chapter 8 together with the findings obtained in chapter 6.

8

Model Improvements

In this chapter we discuss the improvements that were made as a result of the analyses done in chapter 6 as well as the findings from the sessions outlined in chapter 7. We started off with a biased model that mainly predicted based on inspector practice. From a simple accuracy perspective, this would be sufficient, but when the model is to be used in the real world, from a practical point of view this over reliance on biased indicators would be undesirable. In this chapter we discuss the findings of the Expert sessions and their impact on the model. We also cover some of the problems encountered during the sessions and how these could have impacted the results. Furthermore, we outline the improvements for each Model iteration. We cover our Intermediate model which was made as a consequence of the findings of the data analysis part in section 6 as well as our findings during the Exploratory sessions and In-Depth sessions. We outline the model performance as a result of the changes and to what degree this helped in our aim of mitigating High Confidence errors.

8.1. Session Discussion

In this section we discuss findings from the sessions with the inspectors. We cover some of the conclusions we took from these sessions and what pitfalls we encountered.

ADR Session During the Day-in-the-life ADR session we quickly discovered that the ADR inspections happen separate from the other inspection types. The risk model of the ILT is a more general model that covers all law violations, this means that since the risks are different for the ADR it is wise to remove the ADR as a target for the classification model.

Regarding risk for ADR vehicles, perhaps risk analysis can be of more use for logistics trucks as these vehicles do not display well from their outside appearance what cargo they are carrying. As the inspectors have no need for a predictive risk model, the model will no longer predict on the ADR violations. Older mixed inspections with ADR violations will still be used but single ADR violations will not be used.

Exploratory Interview Session During the Exploratory interview session it became clear that for the inspectors there are many visual characteristics that they pay attention to when selecting vehicles for inspection. They look at the state of the vehicle itself, the state of the driver, behaviours and modifications. They also pay attention to aspects such as seasonality, transport sectors, and locations of the transport. Many of the visual aspects we are not able to capture well in data, thus we should leave these aspects to the judgement and experience of the inspectors. We should rather focus on those aspects that we are able to capture in data that can give a real support in the decision making of the inspectors. This means using more features of data that are, at the moment of selection, unknown to the inspectors.

With regard to Rest and Driving Times, we can capture the state and appearance of vehicles by focusing on vehicle maintenance data. We can capture more company and vehicle owner related features such as previous violations and company registration data. Further, with regard to Vehicle Overload, we can use extensive data from vehicle weighing points (WIM) to support and give a substitute for the visual aspects mentioned, such as number of axles, sagging vehicles or tire deformations. The WIM data can give more information into both the individual vehicle as well as the owner's patterns of overload percentages. Besides

this, the WIM data can also give more information into the time and location patterns of a vehicle. This can mean whether the vehicle drives more at night, or during the weekends, or on specific days or specific times of the day or even if it drives often on specific holidays.

In-Depth Question Session Following the Exploratory session we had many areas in the data to question the Experts further on. The In-Depth session gave us the opportunity to go deeper in on these. To capture more aspects such as location data and driving patterns as well as information about overweight, the expansion of the WIM features mentioned above was the best approach to take. The information from these WIM locations give not only a view on the overweight risk of drivers but also provide a broader insight into where and when and how often these violations happens, helping in creating a general risk profile. However, some coverage of these points is still lacking, especially in the northern parts of the country, meaning that for drivers operating mostly in this area, the model can give a skewed view. Further, adding better integration of the historical owner of the vehicle can give more information about the owner's company. Combining this with vehicle type and combinations information, we can get a broader view of the transport sector that the vehicle works in, which from the sessions has shown to indicate risk, particularly when combined with seasonality aspects. Corporate information does not seem to be of much importance to the risk of violation according to the experts. Lastly, vehicle maintenance information is vital to capture latent indicators of violation risk, however, not all maintenance indicators are equally important, picking the right ones is essential. As a result of the knowledge gained from the session, more focus will go on the vehicle problems concerning the structure and bodywork of the vehicle.

The sample size of inspectors was smaller in the In-Depth session compared to the Exploratory session, 4 vs. 7, which may have influenced some of the significance. Either way, much valuable information was gathered. The more open setting of the In-Depth interview gave more room for discussion. The Exploratory session was more constrained due to the large number of indicators that had to be covered in a short amount of time.

Analysis Showcase Session During the Analysis session the inspectors were presented with 8 instances. For all 8 instances the inspectors agreed with the prediction as well as the confidence of the prediction. For the lower confidence values the inspectors were also slightly more hesitant to make a judgement, while for the highly confident predictions the inspectors showed the same confidence. This is beneficial for the trust in the model from the inspectors side. However, we should still be cautious to make any hard conclusions. The sample size of vehicles chosen was small, on top of this the vehicles that we picked were partially known to the inspectors. This is both because these were known vehicles/companies but also because the data instances are based on the set of inspections from the inspectors themselves. In a real world scenario, the inspectors will pick new vehicles to inspect and will not have the same prior knowledge. Either way the results are positive for the aim that we had with the sessions: building trust in the model.

From the session we learned that there are different types of tachographs which can create different profiles of violations. Analog tachographs make it easier to violate rest and driving time laws than the new generation of Smartachs. Using data of tachograph installation and important dates for tachograph changes should give more insight into these risk profiles. During the sessions the inspectors also noted that violators are more prone to commit speed violations. WIM location data also measures the maximum speed and whether it was violated, this is a valuable feature to add to the risk model. Lastly, during the session some instances had a company age value of 0, this could cause confusion for both the inspectors as well as the model. To deal with missing values in the data we should set these to more distinct values such as -1 (or -100 for more of a distinction on a continuous scale).

In retrospect of the Analysis session we found that The inclusion of the visual picture of the vehicles presumably had a big impact on the judgement of the inspectors. In hindsight, we note that this may have impacted their judgement as in their practice they prefer visuals over data. Additionally, the experts had the option to look up the vehicle in their own system, even though for most cases they responded quite quickly, before having the chance to look up anything, but either way this could have impacted their judgement.

8.2. Model Changes

Following the sessions and the data-driven analysis, changes were made to the model in a sequential manner. In this section we give a description of these changes and their impact on the performance both of the model in general, as well as on the mitigation of the High Confidence errors. The entire model improvement process

is displayed graphically in figure 8.1. Here the green areas show the concrete changes to the data and consequently the model. The blue area shows the expert-driven changes through the iterative expert sessions, while the red area shows the data-driven part and the changes as a result.

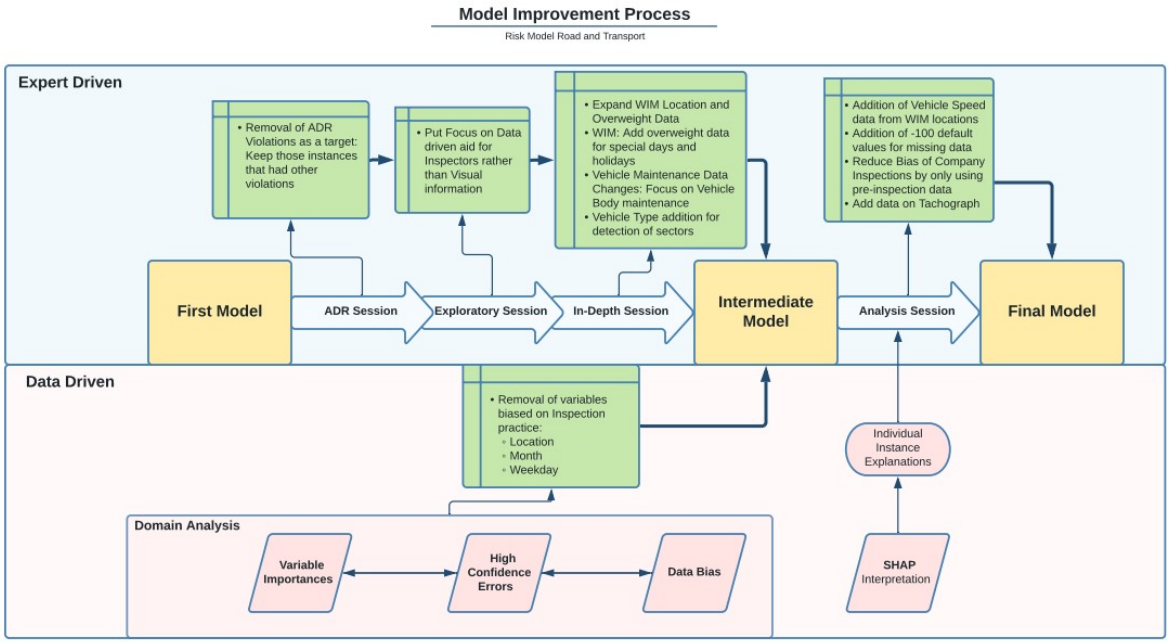


Figure 8.1: Graphical Representation of Model Improvement Process. The green squares show the improvements made as a result of both the sessions and the data analysis. The blue area shows the Expert Sessions and the red part the Data driven approaches.

8.2.1. Intermediate Model Iteration

Analysis of the model in chapter 6 found that an overreliance on certain variables, such as inspection month and location caused bias in the model. This, in a sense, modeled the practice of an inspector rather than the actual risk of a vehicle. For instance, certain locations where inspectors performed many inspections caused these locations to be weighed as overly important in the model, as these increased inspections also caused an increased number of violations. These locations do not necessarily indicate risk by themselves, rather the inspectors choose those locations where they know that violations are more prone to be found. This indicates that the model mirrors the inspectors' practice, rather than predicting the risk of the vehicles. Thus this knowledge would not add value to the model when used in a real world scenario.

This led to the removal of these variables, namely: Location (Latitude, Longitude), inspection Month and inspection Day. Although the time variables are useful as seasonality indicators, these seasonality factors are now instead captured by the WIM location data. There were additional changes to the model as a result of the ADR, Exploratory and In-Depth sessions, as discussed in the previous section. These are also visible in figure 8.1.

The changes made led to the Intermediate model. This model was created in the same way as the First Model as covered in chapter 6. A 80/20 stratified Train-Test split was used. The optimal mtry parameter with the best performance was found to be and mtry of 20. The ntree was kept at 1500 similar as for the first model. The performance metrics of the new Intermediate model are displayed and compared with the First model in table 8.1. We can see a reduction in most of metrics in absolute terms, with only the Out-of-bag Error improving. However, we have to look at these metrics in a relative sense due to the removal of the ADR vehicle violations as a model target. Because we removed the violations based on ADR the overall baseline precision of the model has significantly decreased, from 0.32968 in the first model to 0.23473 in the new model. This means that taken relative to the baseline, our model performance has increased. This is a 25.1% increase in precision considering the reduced baseline. We do note an increase in High Confidence Errors in the top 20% and top 5% as a result of the removal of some of the high contributing variables that were overfitting on the inspection set, increasing some of the uncertainty in the model.

Parameter	**First Model**	**Intermediate Model**
mtry	18	20
ntree	1500	1500
Metric	**First Model**	**Intermediate Model**
Baseline Precision	0.32968	0.23473
Precision	0.56800	0.50605
AUC	0.44969	0.38702
OOB error	32.25%	23.53%
HCE 20	0.70435	0.75914
HCE 5	0.78047	0.84073

Table 8.1: Table displaying the performance metric values for the Intermediate model and comparing these with the First Model

The Intermediate model uses a total set of 195 features, of these features we calculated the variable importances similarly to chapter 6. Of this we display the top 100 variables importances in figure 8.2. Because this set of variables is so large and most variables near the end are only weakly contributing, we want to focus in on only the top 20 variables. We show the calculated top 20 variable importances for the Intermediate model in figure 8.3. The description of these variables is give in table 8.2, in order of importance. We can see that the majority of highly contributing variables are now based on the WIM data. Interestingly, now the *Oprichting_eig* variable, or the age of the company, is the most highly contributing variable. Where younger companies are more often seen to be violators by the model. We also note that despite the removal of the weekday variable, the WIM data has now in essence taken the place of this variable as many of the highly contributing variables are based on the WIM passages on certain days or certain times of the day. This is in agreement with the earlier found fact in the Exploratory session that the day of the week has a real effect on the risk.

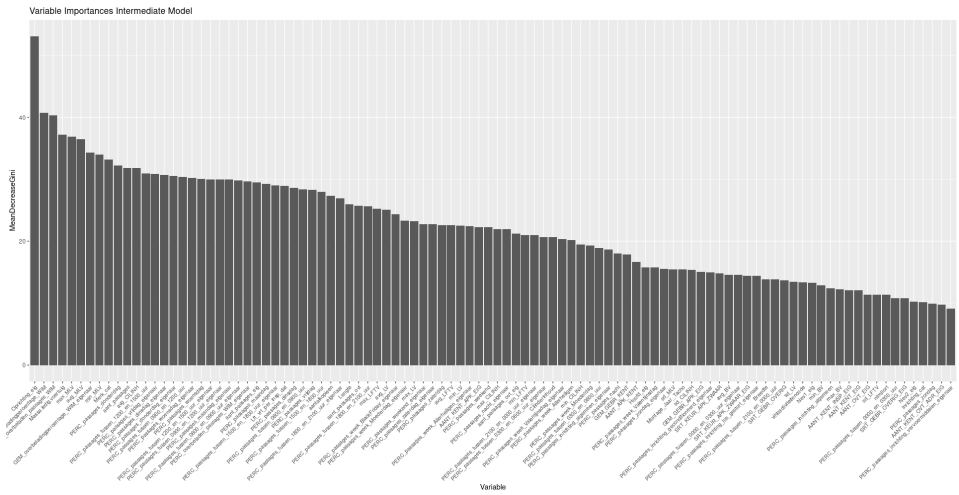


Figure 8.2: Top 100 Most important Variables for the Intermediate Model

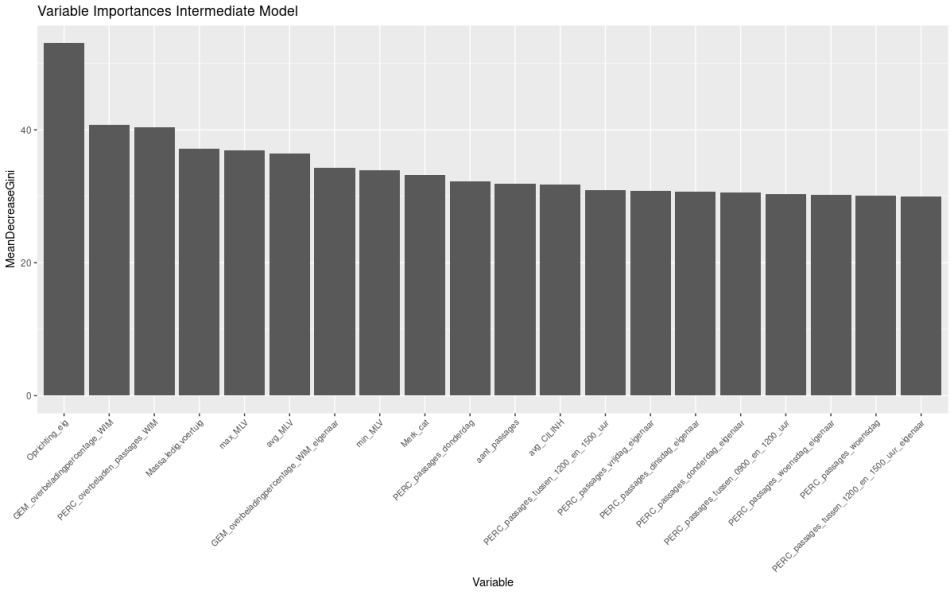
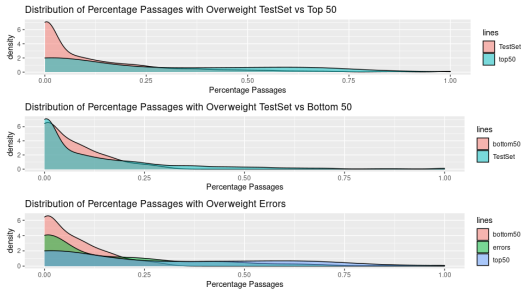


Figure 8.3: Top 20 Most important Variables for the Intermediate Model

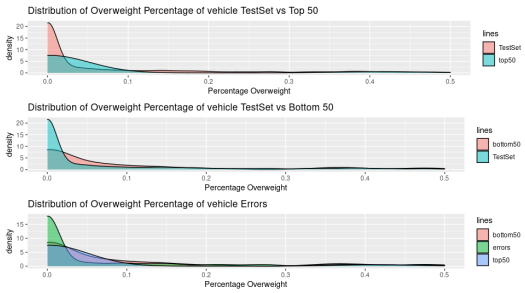
Variable Name	Description
Oprichting_eig	The age of the company at the time of inspection.
GEM_overbeladingpercentage_WIM	The average percentage of overweight of the vehicle.
PERC_overbeladen_passages_WIM	The percentage of passages over the WIM location of the vehicle with overweight detected.
Massa.ledig.voertuig	The mass of the empty. vehicle
max_MLV	The maximum mass of an empty vehicle on the name of the RDW owner.
avg_MLV	The average mass of all vehicles on the name of the RDW owner.
GEM_overbeladingpercentage_WIM_eigenaar	The average percentage of overweight of all vehicles of an owner.
min_MLV	The minimum mass of an empty vehicle on the name of the RDW owner.
Merk_cat	The Brand of the vehicle.
PERC_passages_donderdag	The percentage of passages over the WIM location done on a Thursday.
aant_passages	The total amount of passages over the WIM location.
avg_CILINH	The average cylinder capacity of the vehicles on the name of the RDW owner.
PERC_passages_tussen_1200_en_1500_uur	The percentage of WIM passages done between the hours of 12 and 15.
PERC_passages_vrijdag_eigenaar	The percentage of passages over the WIM location done on a Friday of all vehicles of an owner.
PERC_passages_dinsdag_eigenaar	The percentage of passages over the WIM location done on a Tuesday of all vehicles of an owner.
PERC_passages_donderdag_eigenaar	The percentage of passages over the WIM location done on a Thursday of all vehicles of an owner.
PERC_passages_tussen_0900_en_1200_uur	The percentage of WIM passages done between the hours of 9 and 12 in the morning.
PERC_passages_woensdag_eigenaar	The percentage of passages over the WIM location done on a Wednesday of all vehicles of an owner.
PERC_passages_woensdag	The percentage of passages over the WIM location done on a Wednesday.
PERC_passages_tussen_1200_en_1500_uur_eigenaar	The percentage of WIM passages done between the hours of 12 and 15 of all vehicles of an owner.

Table 8.2: Description of the top 20 most important variables for Intermediate Model, listed in order of importance

From this we highlight plots of 4 of the most important variables. In figures 8.4a and 8.4b we show respectively the number of overweight passages on the WIM locations, as well as the average percentage of actual overweight on these passages. We can see that the bottom 50 errors are usually occurring for lower percentages of overweight while for the top 50 it is much more spread out over the entire range. This indicates that the model is more prone to miss violations when the vehicle displays no larger pattern of overweight risk. In figure 8.5a we display the plot of the age of the company, which according to the variable importance is the strongest predicting feature. We see here that younger companies are more often classified as violators than older companies, which also causes the top 50 errors to more often occur on younger companies. Lastly, we show an interesting new important feature, namely the amount of passage over the WIM on a Thursday. In figure 8.5b we see that the bottom 50 errors closely follow the Test set distribution, while the top 50 errors seem to be more spread out, indicating some risk for vehicles that deviate from the norm in terms of how much they drive on Thursday, where vehicles that drive less on Thursday seem to be more often wrongly classified as violator.

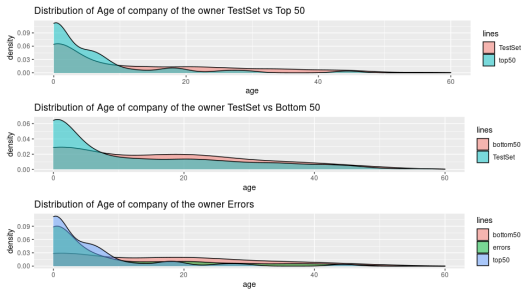


(a) Density Plot for the amount of overweight passages on the WIM locations for the Intermediate Model

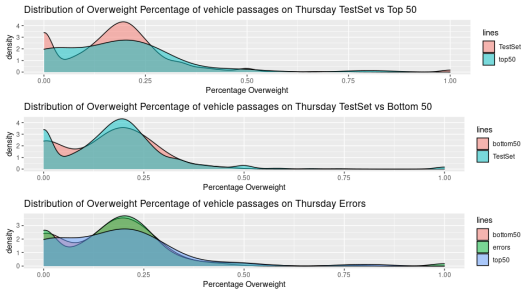


(b) Density Plot for the average percentage of overweight for the passages on the WIM locations.

Figure 8.4: Density Plots for the Intermediate model for PERC_overbeladen_passage_WIM & GEM_overbeladingpercentage_WIM



(a) Density Plot for the age of the company of the owner for the Intermediate Model



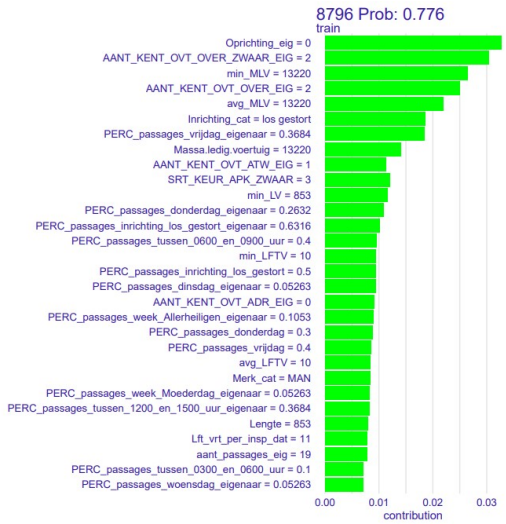
(b) Density Plot for the percentage of passages on the WIM locations on a Thursday.

Figure 8.5: Density Plots for the Intermediate model for Oprichting_eig & PERC_passages_donderdag

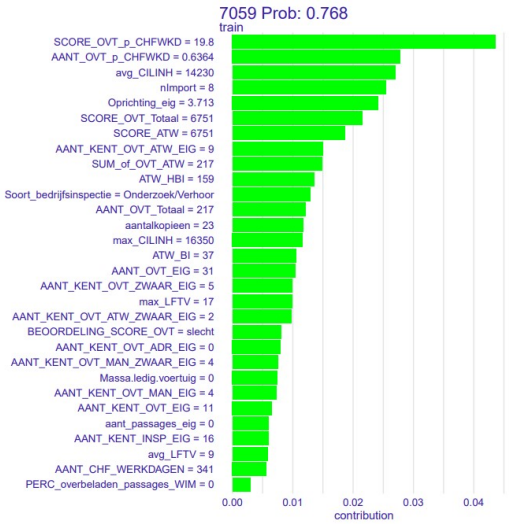
8.2.2. SHAP Interpretation

For the purpose of the Analysis sessions mentioned in section 5.5.3, we need to have interpretable instances of High Confidence errors, such that we can present these to the inspectors. These instances help us not only by showing what variables are causing these errors to happen but also what variable values are more likely to cause errors in the model. We created plots for the top 5 High Confidence Errors for the violator/top side as well as the top 5 for the compliant/bottom side. In figure 8.6 we show the plots for the top 2 and In figure 8.7 for the bottom 2 errors. Here we can see that for the 2 top errors there is a high contribution of features indicating previous violations. The model this way indicates that having had previous violations will indicate a risk of repeat violation. We also see that the company of the vehicle owner being a young company contributes. However, in plot 8.6a the age of the company (Oprichting_eig) is 0 meaning the data was unavailable. Interestingly, for the bottom 2 errors in plot 8.7a we can see that the lowest probability was given for a vehicle where no data was available, as all feature values are 0. This shows the model's tendency to predict more vehicles as compliant due to the distribution of compliant vs violator in the dataset. Further, in 8.7b we can see for the age of the company that we are dealing with an older company (27.34 years), this makes the model more prone to classify it as compliant, however, in this case the vehicle was in violation.

The plots for all 10 instances are displayed in appendix C we also display the top 10 correct predictions. For the Analysis session a different set of instances were used where 5 high confidence errors and also 3 correct instances were shown. This is covered in section 5.5.3.

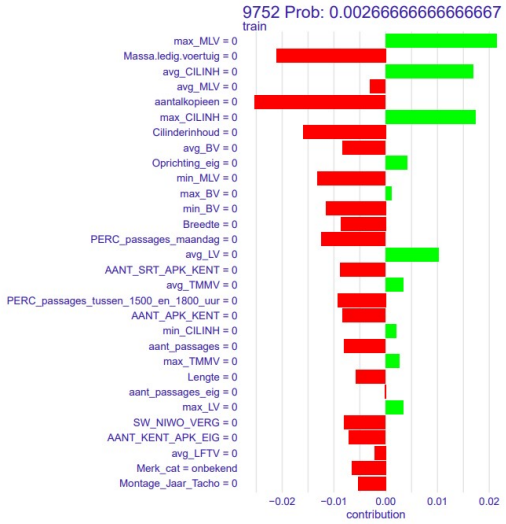


(a) SHAP contribution plots for top 1 High Confidence Error

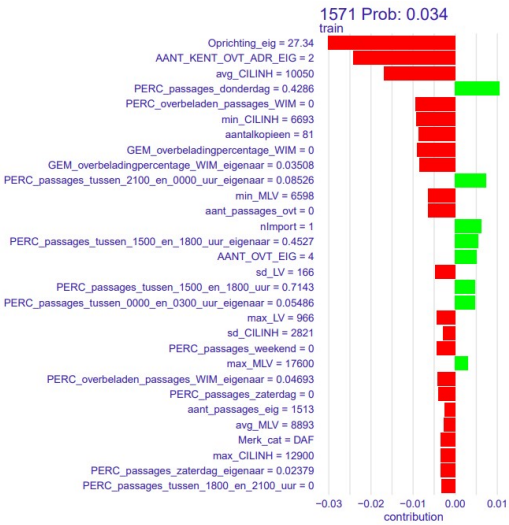


(b) SHAP contribution plots for top 2 High Confidence Error

Figure 8.6: SHAP contribution plots for the top High Confidence Error instances



(a) SHAP contribution plots for bottom 1 High Confidence Error



(b) SHAP contribution plots for bottom 2 High Confidence Error

Figure 8.7: SHAP contribution plots for the bottom High Confidence Error instances

8.2.3. Final Model

Taking into account the results of the Analysis session and all previous sessions and analyses, we come to the Final model. As stated earlier, the most concrete changes were the addition of the feature of vehicle maximum speed violations, the addition of -100 default values for missing data, additional date information on tachographs and lastly we also changed the violations of company inspections to only be counted for those that happened before the inspection date, else the data was biased with company inspection data that happened after the inspection was done, sometimes causing an overly negative risk view for a vehicle. The final model was run in a similar fashion as the previous models, this time using an optimally calculated mtry of 17. In table 8.3 we show the performance results of the Final model, comparing them with the previous models. The table shows that the precision of the model went up from 0.50605 in the previous model to 0.52077 in the final model. We also note a large decrease in the magnitude of High Confidence errors in the top 20% of errors, namely from 0.75914 in the previous model to 0.70465 in the new model.

The improvements made after the insights from the analysis sessions show to have an increase in the model performance as well as a significant decrease in magnitude of High Confidence errors.

Parameter	**First Model**	**Intermediate Model**	**Final Model**
mtry	18	20	17
ntree	1500	1500	1500
Metric	**First Model**	**Intermediate Model**	**Final Model**
Precision	0.56800	0.50605	**0.52077**
Baseline Precision	0.32968	0.23473	0.23473
AUC	0.44969	0.38702	0.39282
OOB error	32.25%	23.53%	**23.37%**
HCE 20%	0.70435	0.75914	**0.70465**
HCE 5%	0.78047	0.84073	**0.82857**

Table 8.3: Table Comparing the parameters and performance metrics for all model iterations.

The top 20 variable importances for the final model are displayed in figure 8.8. The variables that the model uses are close to the same as those for the Intermediate model. We do see the addition of a new important variable, namely *PERC_gem_snelh_overschrijding_WIM* in position 5, which is the average percentage of vehicle passages of the owner where driving speed violations were committed. This is one of the newly added variables as a result of the analysis session.

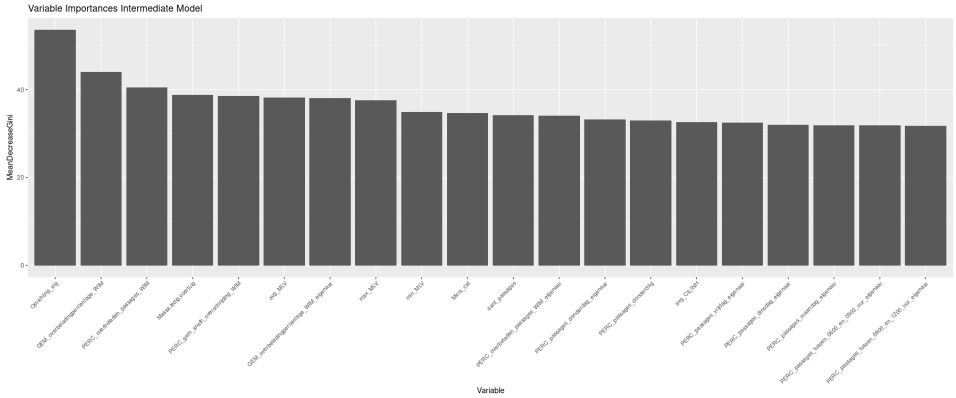


Figure 8.8: Top 20 Most important Variables for the Final Model

8.2.4. Error Predictions

To compare this model's errors with the first model, we run the Error Prediction model on the new data as well. The model has a precision of 0.76905 and AUC of 0.83701. The OOB error estimate is 24.28%. These are summarized in table 8.4.

Metric	Value
Precision	0.76905
AUC	0.83701
OOB error	24.28%

Table 8.4: Table displaying the performance metric values for the Error Prediction Model for the Final Model

We display the top 20 most important variables for the error prediction model in figure 8.9. We see many of the deciding variables from the original final model also returning in the top end of the Error Prediction model. However, it is interesting to note that in the top 3 we have two time of day related variables from WIM, while in the original model these were ranked only in the bottom end of the top contributing variables.

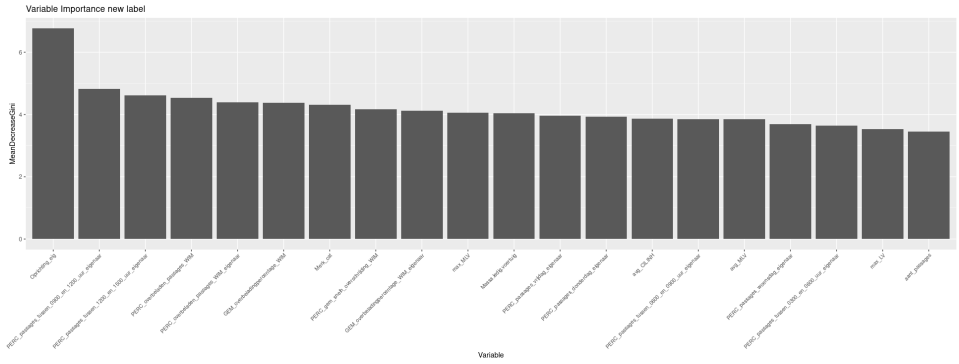


Figure 8.9: Top 20 Most important Variables for the Final Error Prediction Model

9

Conclusion

With this thesis we set out to tackle the issue of model accuracy and end-user trust by exploring the nature of High Confidence Model errors and the knowledge acquisition gap between Data Scientists and Domain Experts and working out methods to keep these at a minimum.

This thesis had the aim of answering the following question:

> *How can we best characterize, and mitigate predictive errors that are produced with a high model confidence?*

In order to answer this question we have explored both data-driven as well as human-centered ways to come to an answer. Namely, we have shown that a detailed analysis of the dataset used can highlight certain biases in the data, which can help in characterizing the nature of High Confidence Errors. In order to mitigate these errors, we have proposed a Human-In-The-Loop based methodology to leverage the knowledge of domain experts to improve the model, while also increasing end-user trust. we used this methodology to answer the following sub-research questions:

RQ 1: *What Instances Best Characterize System Knowledge?*

RQ 2: *How to best interpret what the model has learned?*

RQ 3: *What knowledge do we want the model to have?*

RQ 4: *What does the model not know?*

9.1. Summary

To summarise the thesis, we provided a background on previous methods of reducing and characterizing Unknown-Unknowns. We outlined what previous work has gone into Human-Computer interaction methods. To facilitate proper Human-In-The-Loop interactions a good foundation of interpretable machine learning was needed. For this we also performed a study of the core aspects of interpretable machine learning and outlined some of the most valuable interpretation methods. We then gave a data-driven analysis of the model in use by the ILT. We explored areas of data bias and outlined on what the model bases its predictions. From this we showed areas of the data that lead to High Confidence Errors. We gave a methodology of Iterative Expert Sessions and validated this methodology by applying it to Data Scientist/Domain Expert sessions with the Road and Transportation Inspectors of the ILT in a series of sessions. The first session being the Exploratory session where our aim was to get a general view of the working practice of the inspectors. Following this was the In-Depth session, where we dived deeper into the specific data that is required for a good risk profile of the vehicle. Lastly we performed an Analysis session with the inspectors, where we showcased several instances and compared the model's results to the judgement of the inspectors. We presented the results from these sessions and outlined how these results lead to which model changes with the aim of improving the model and reducing High Confidence Errors.

9.2. Conclusions

To answer our first research question: *What Instances Best Characterize System Knowledge?* a data driven approach was required. A data-driven analysis of the model found that an over-reliance on certain variables, such as inspection month and location caused bias in the model. Through our prediction of errors in section 6.2.1 we find that those features that are highly contributing to the prediction are also most likely to cause High Confidence errors to occur. Certain variable values, such as highly occurring inspection locations or months with many violations cause the model to over rely on these features and predict based on these features with high confidence.

These features in a sense modeled the practice of an inspector rather than the actual risk of an individual vehicle or vehicle owner to be in violation. The inspectors, through their planning, experience and expertise, have certain preferences for important locations and times of inspections,

This connects to our question of *How to best interpret what the model has learned?*. A study of model interpretability shows SHAP to be the most suitable method for our use of interpreting (High Confidence) individual instances. These interpretations can be used for both the Data Scientists to analyze the instances as well as being a tool for the Analysis sessions with the Domain Experts. These sessions help all participants to better understand the decisions of the model, and SHAP serves as a tool for this.

Based on the Iterative Session methodology that we proposed, interview and in-person sessions were held with the experts and data scientists of the ILT. Through these we answer our research question of: *What knowledge do we want the model to have?* From the results of the sessions we saw that we could validate the iterative session model as being valuable for the process of bridging the knowledge gap between Data Scientist and Expert. The In-person Exploratory session gave us the insight of removing ADR inspections from the model and the Interview session gave us clear indicators that the experts focus on. Much of the inspectors intuitions are based on visual characteristics of the vehicle, but they combine this with previous known violators, as well as what type of freight they are dealing with. The In-Depth session helps us to narrow down those features that are valuable for our classification goal of assessing vehicle risk. The session lead to a broadening in the use of WIM location passage data as well as a clearer view of what vehicle maintenance aspects are important to the assessment of risk. Lastly, the Analysis session proved to be a good method of familiarizing the experts with the model. The experts found the decisions of the model to be reasonable and in alignment with their own judgement. Despite being in error the model preserves the trust of the experts. The session also sparked discussion on certain features that we might have missed such as vehicle speed violations which can serve as an indicator for a pattern of riskful behaviour. These features found to be missing during the sessions also connect to answering the research question of: *What does the model not know?*. The model essentially does not learn enough of what it means to be a violating vehicle, rather it prefers to rely on approximations such as locations or periods with many violations.

9.2.1. Model Improvements

As a result of the data-driven analysis and the Exploratory and In-Depth sessions, improvements were made to the initial model. The initial model had a precision of 0.56800 with a baseline of 0.32968 with an average confidence score of 0.70435 in the top 20% of errors. After the improvements of the first two sessions, the removal of the ADR inspections from the prediction target, as well as the removal of biased variables on location and time of inspection, the models' precision went to 0.50605 with a baseline of 0.23473 with an average confidence score of 0.75914 in the top 20% of errors. This is a 25.1% increase in precision considering the reduced baseline due to the removal of the ADR targets. The data bias was removed but at the cost of an increase in High Confidence errors in the top 20 %. Following the Analysis session, new improvements were made which led to the final model. The final model has a precision of 0.52077 with a baseline of 0.23473. The average High Confidence score in the top 20% of errors is 0.70465. This is an improvement of 28.8% in precision when compared to the initial model as well as a 2.9% increase compared to the intermediate model. We also see an improvement in the magnitude of High Confidence Errors when compared to the previously improved Intermediate model. All metrics are compared and shown in table 8.3.

From the sessions we have found that the continuous integration style development helps bridging the gap between domain experts and data scientists in the context of road and transportation law violations while also fostering trust in the model. It helps aid in reducing data bias issues that cause High Confidence Errors to be made. The improvements made during the Exploratory and In-Depth sessions provided the inspectors with a more coherent prediction. Data biases were filtered out and important new opportunities for data variables that can indicate risk were discovered. Finally we observed that some HCEs reflect a nature of risk of the vehicle, even though they are errors at the moment of inspection. Future inspections of these can still

identify irregularities, making it difficult to distinguish real errors from model errors due to the temporal and transient nature of being a violator.

In conclusion we can validate that the Iterative Session Methodology proposed shows to be a good method for domain expert interaction. These sessions combined with the analysis of the data shows a reduction in overall sample bias. We show an improvement in terms of model performance as a result of the changes made. Finally, we have presented a better characterization of High Confidence errors and a decrease in magnitude of these errors after the removal of data biases based on inspection practice.

9.3. Contributions

Some of the core contributions(also see section 1.6 for full list) that come from this thesis are the following:

- A novel approach to Domain Expert interaction through and Iterative Session Model, validated in a real life scenario setting with inspectors of Road and Transporation of the ILT
- A Direct improvement in performance for the Road and Transportation risk assessment model.
- Data-driven method for characterizing High Confidence errors and a reduction of magnitude of High Confidence errors, following the removal of data biases. This allows us to better characterize High Confidence Errors that are predicted both as False Positives and as False Negatives.
- Through the direct involvement of domain experts in the process of building the model, we provide an interactive modeling method to reduce the presence of data issues and biases, as well as recommendations and lessons learned on how to elicit Expert Feedback for model development.

9.4. Limitations and Recommendations

During the thesis some limitations of the research were encountered. One of the primary is of course the issue with human-centered studies, which is the availability of participants. It is a difficult task to have a significant amount of Domain Experts available at the same moment for the interview sessions. The fact that these sessions are quite time consuming means that much prior preparation and scheduling is needed, leading to only a three interview sessions being able to be held, while more sessions is always better for the quality of the results. Optimally you would like to have as much input as possible from many Experts, during the first Exploratory session this was on a relatively good level. However, with only 3 inspectors for our Analysis Session especially, the lack of domain experts could potentially be an issue for the significance of the findings.

Due to the small sample size of experts it can be difficult to get solid quantitative results from the sessions. Instead much of the focus is on those qualitative results that we did get. A combination of interview sessions, surveys and questionnaires can improve the quantitative results.

Furthermore, we sometimes found it difficult to make experts focus on the data and the model rather than their expertise during the sessions, for this we would recommend to guide the experts as much as possible to step out of their comfort zone and focus their attention on the data. This connects to the fact that how the sessions are set up and what you show (e.g. showing vehicles and showing SHAP plots) can have an impact on the results of the sessions and what you are measuring. This means that you have to be careful with these choices.

Finally, another limitation to consider is that of the thesis research being done as part of the ILT internship. The ILT has its own goals and interests and it can be a challenge to align these with the goals of the scientific research pursuits.

9.5. Future Work

As future work, further research is to be done into the optimal setting for each session. As has been stated before in the Introduction in chapter 1and Background in chapter 3, the research into Human-Centered Machine Learning is still in its early stages, so a lot is still unknown as to the optimal workflow. This thesis provides a good general methodology to work further on. Having the session model deployed in other contexts can further validate the methodology.

More research needs to go into the field of High Confidence errors as well. The question of what is "high confidence" and in how far these errors are simply a consequence of the natural distribution of prediction

probabilities is still to be further researched. The lessons learned from the prediction of errors can be used to develop methods to find and predict changes in the data.

Regarding the risk model itself, we plan to improve the effectiveness of the model by using a hybrid approach consisting of a predicted risk score, as well as providing feature contribution values and decision rules, based on what the model has learned. This should help the inspectors in their practice, as they will see not only the risk score but also some indicators that they find of interest.

A

Variable Plots First Model

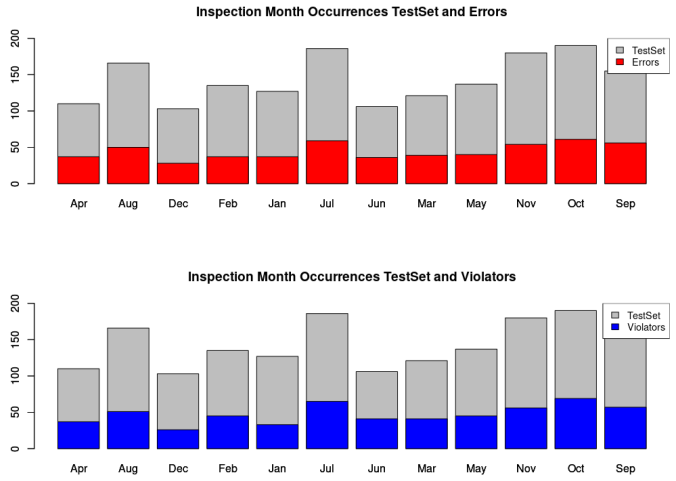


Figure A.1: Plots for Month Errors and Violators

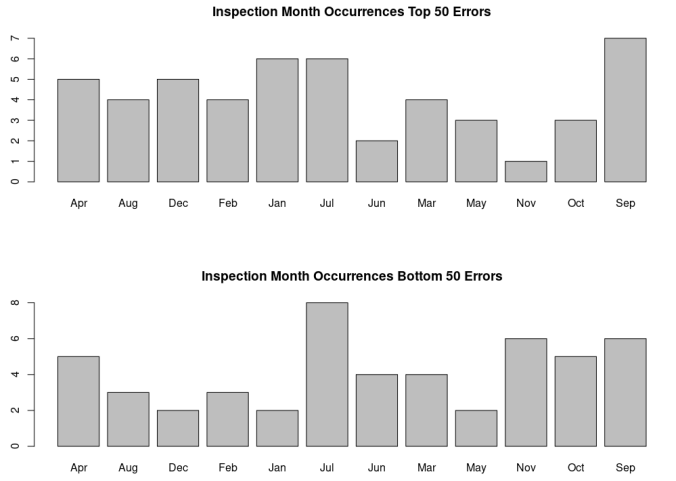


Figure A.2: Plots for Month High Confidence Errors, Top 50 and Bottom 50 Errors

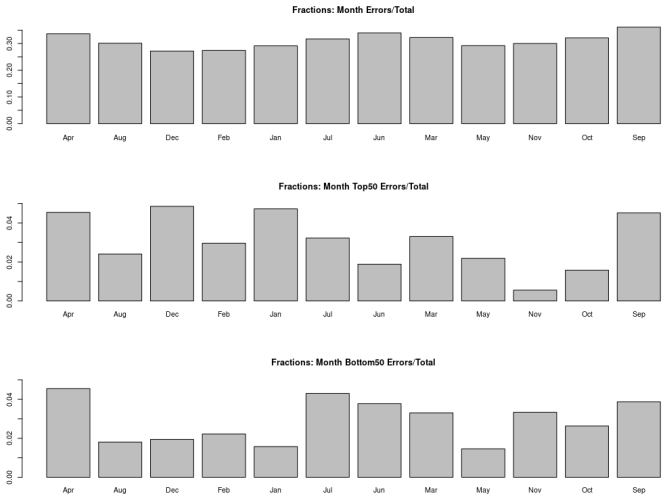


Figure A.3: Plots for Month Fractions

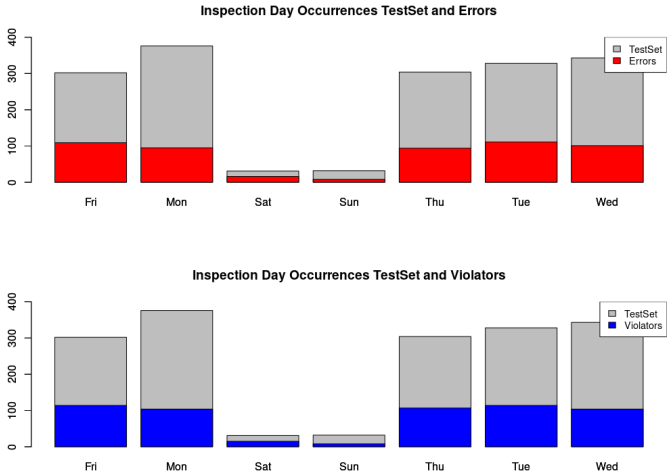


Figure A.4: Plots for Day Errors and Violators

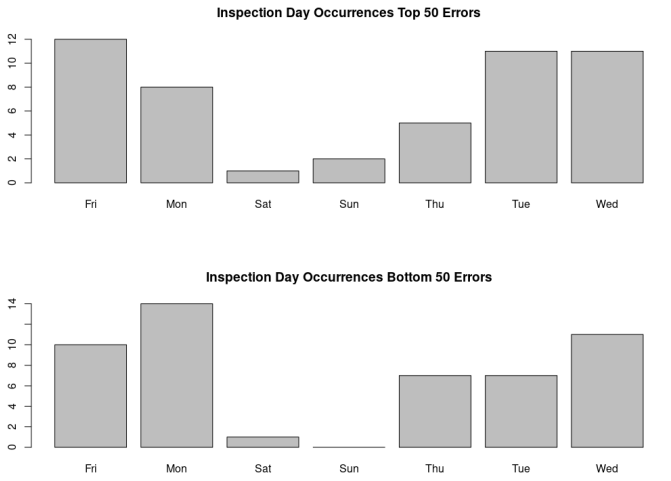


Figure A.5: Plots for Day High Confidence Errors, Top 50 and Bottom 50 Errors

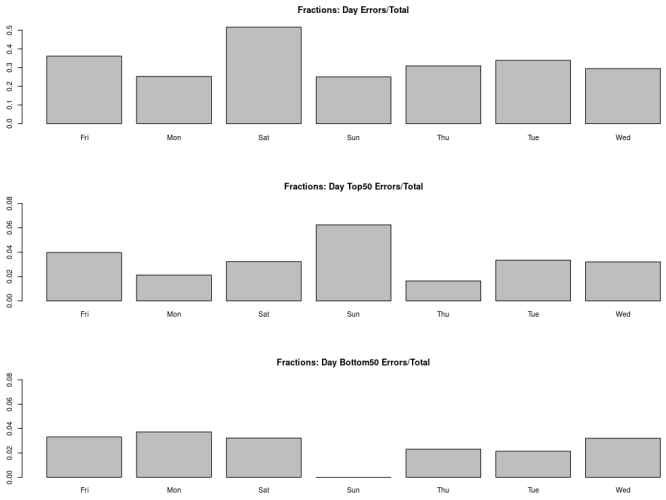


Figure A.6: Plots for Day Fractions

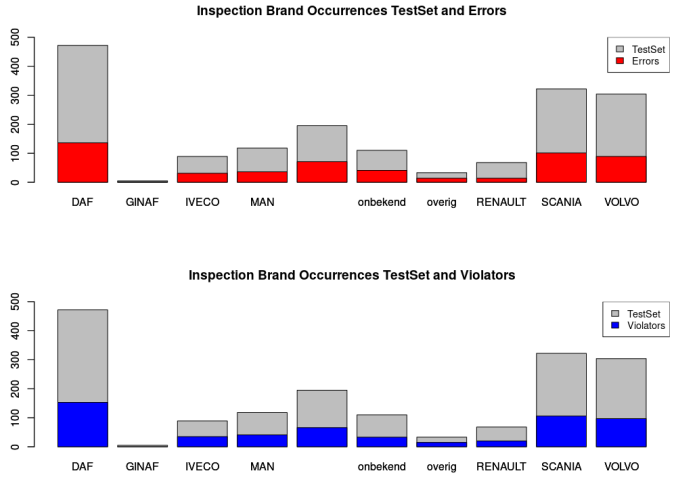


Figure A.7: Plots for Brand Errors and Violators

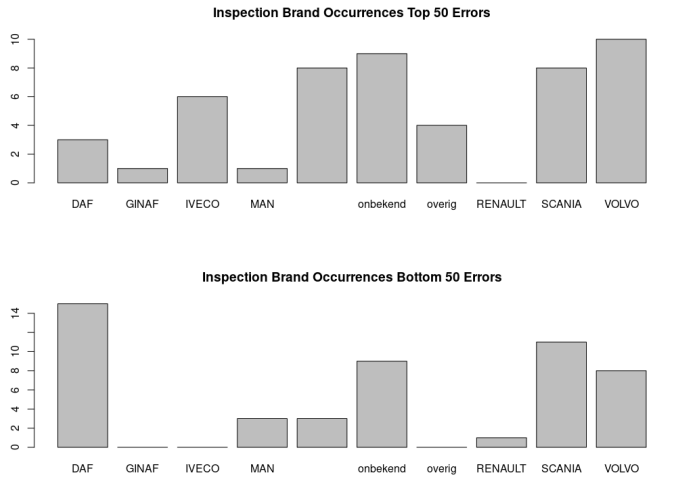


Figure A.8: Plots for Brand High Confidence Errors, Top 50 and Bottom 50 Errors

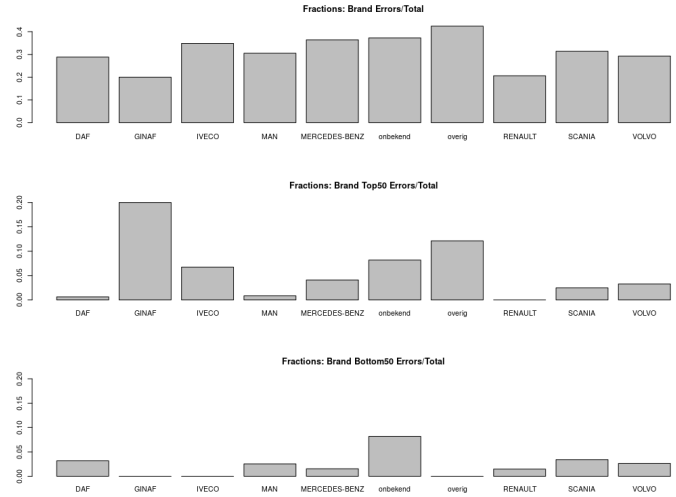


Figure A.9: Plots for Brand Fractions

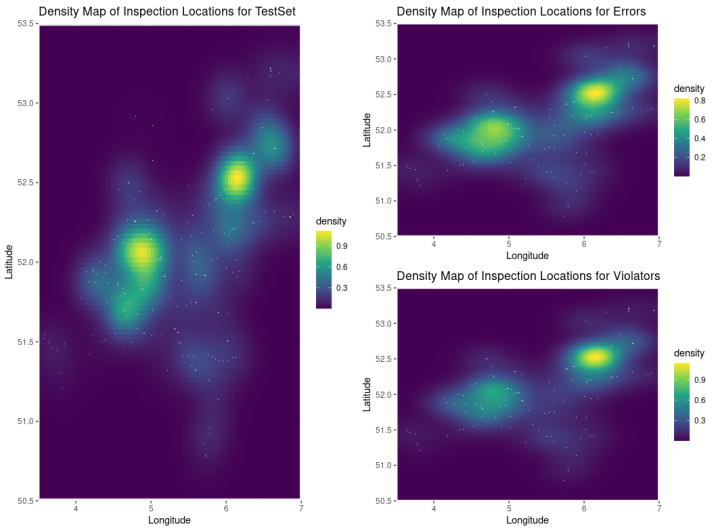


Figure A.10: Plots for Inspection Locations Errors and Violators

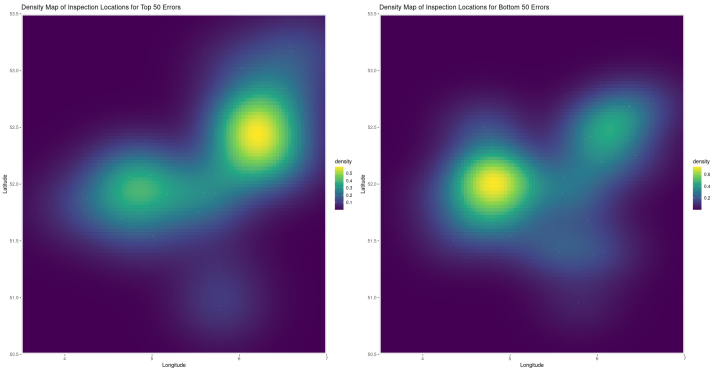


Figure A.11: Plots for Inspection Locations High Confidence Errors, Top 50 and Bottom 50 Errors

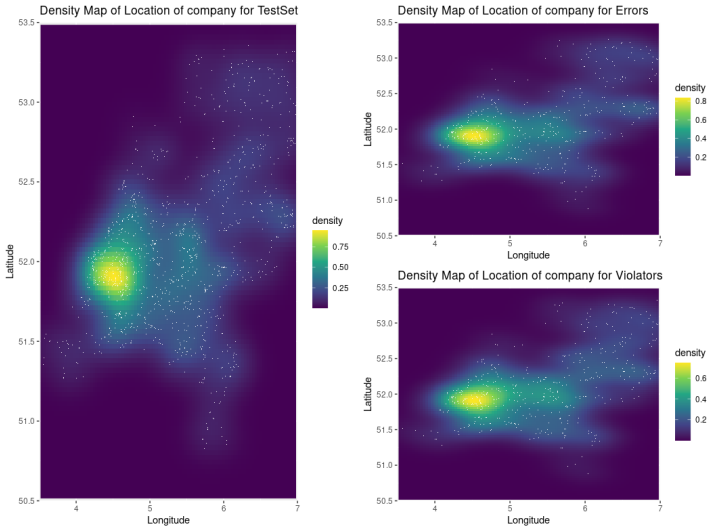


Figure A.12: Plots for RDW Owner Locations Errors and Violators

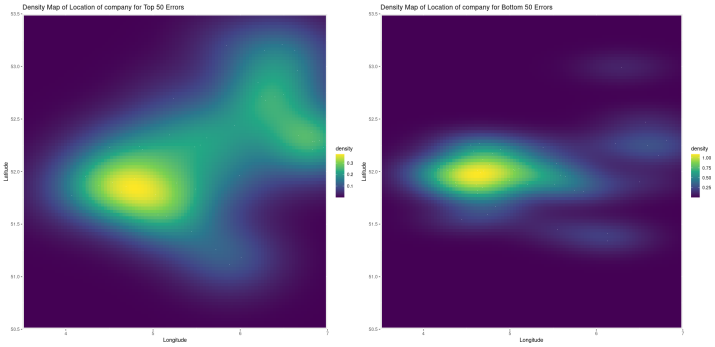


Figure A.13: Plots for RDW Owner Locations High Confidence Errors, Top 50 and Bottom 50 Errors

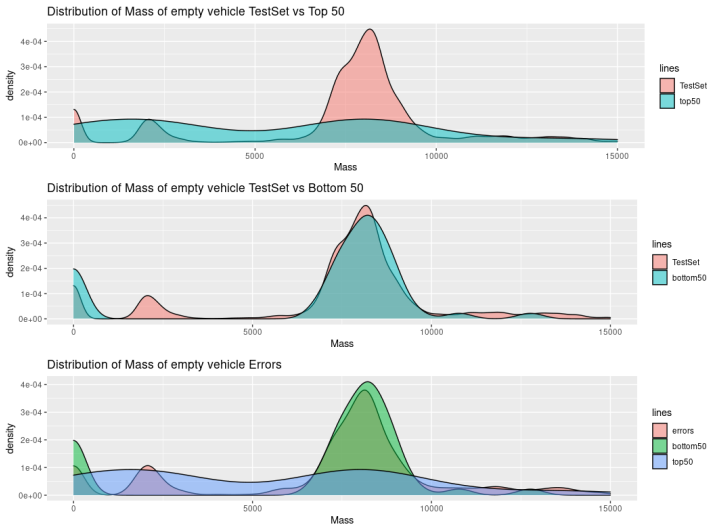


Figure A.14: Density for Mass of empty Vehicle High Confidence Errors, Top 50 and Bottom 50 Errors

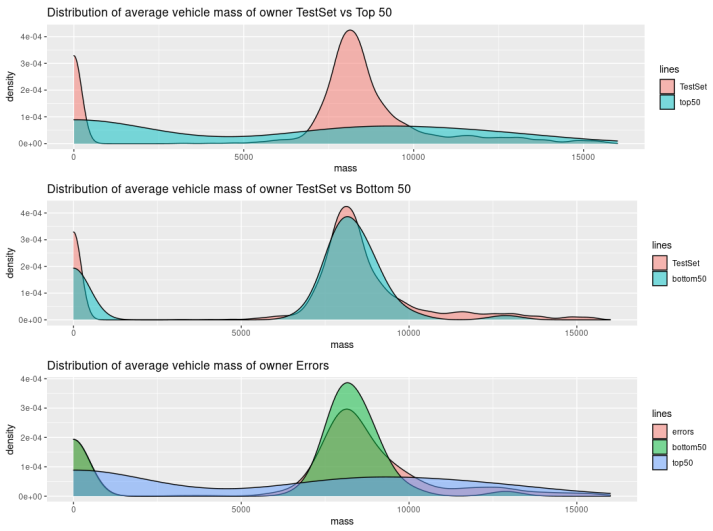


Figure A.15: Density for Average Mass of Vehicles owner High Confidence Errors, Top 50 and Bottom 50 Errors

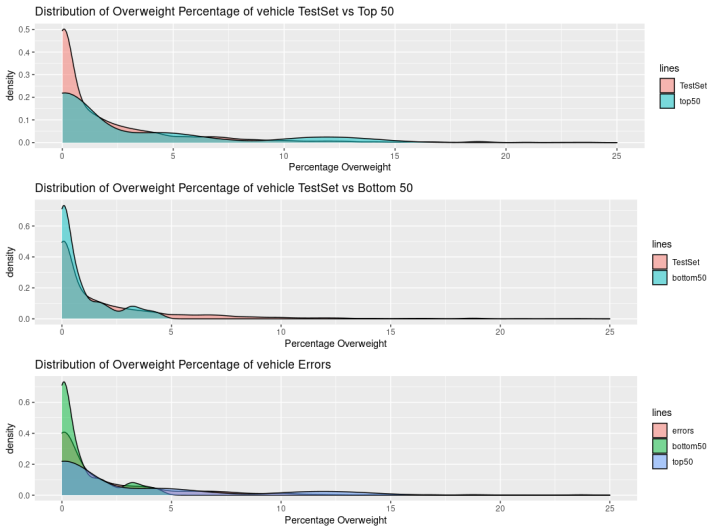


Figure A.16: Density for Percentage Overweight of Vehicle High Confidence Errors, Top 50 and Bottom 50 Errors

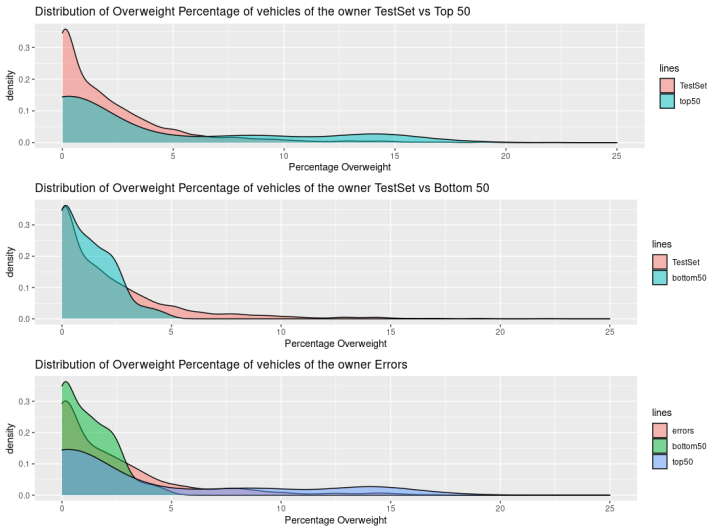


Figure A.17: Density for Percentage Overweight of Owner High Confidence Errors, Top 50 and Bottom 50 Errors

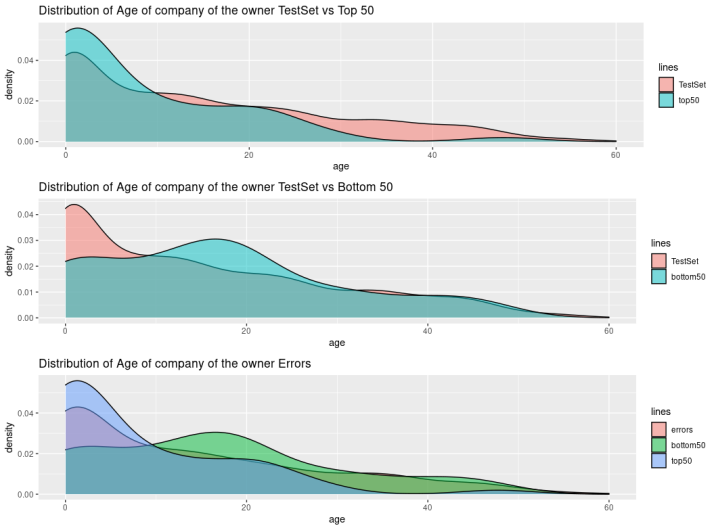


Figure A.18: Density for age of the company High Confidence Errors, Top 50 and Bottom 50 Errors

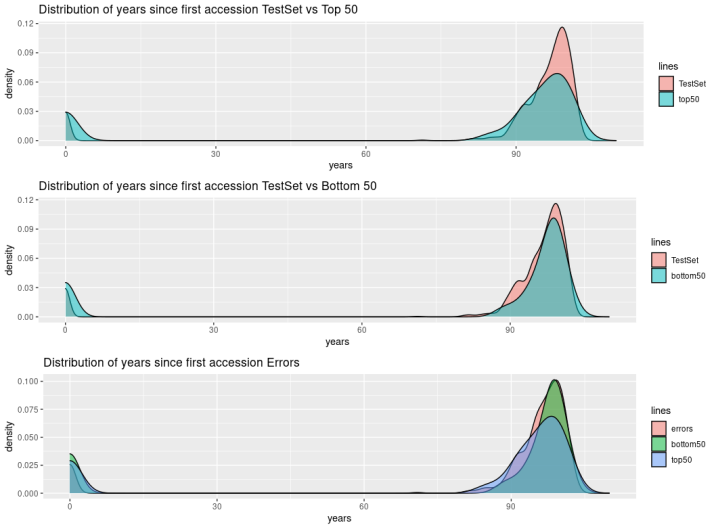


Figure A.19: Density for years since first accession High Confidence Errors, Top 50 and Bottom 50 Errors

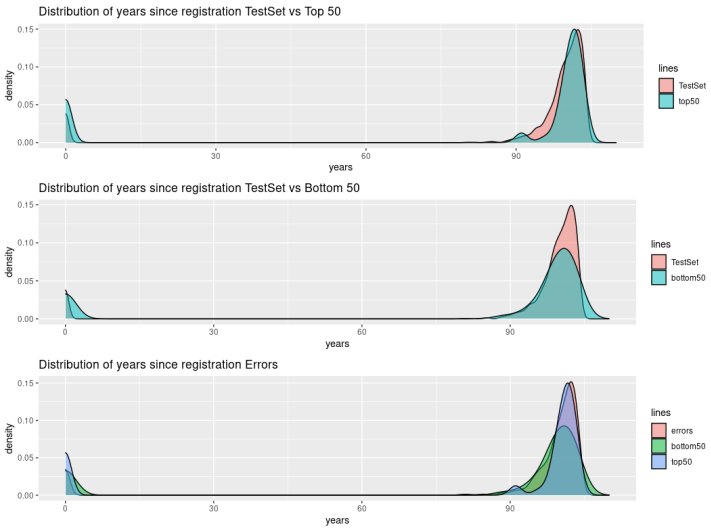


Figure A.20: Density for years since registration High Confidence Errors, Top 50 and Bottom 50 Errors

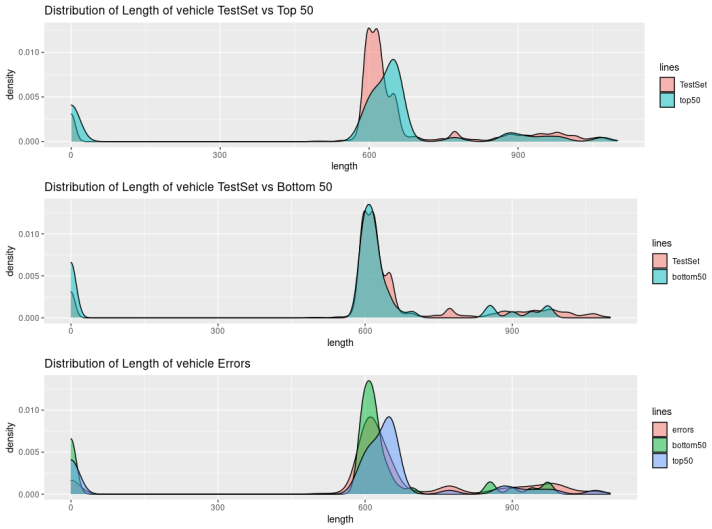


Figure A.21: Density for Length of Vehicle High Confidence Errors, Top 50 and Bottom 50 Errors

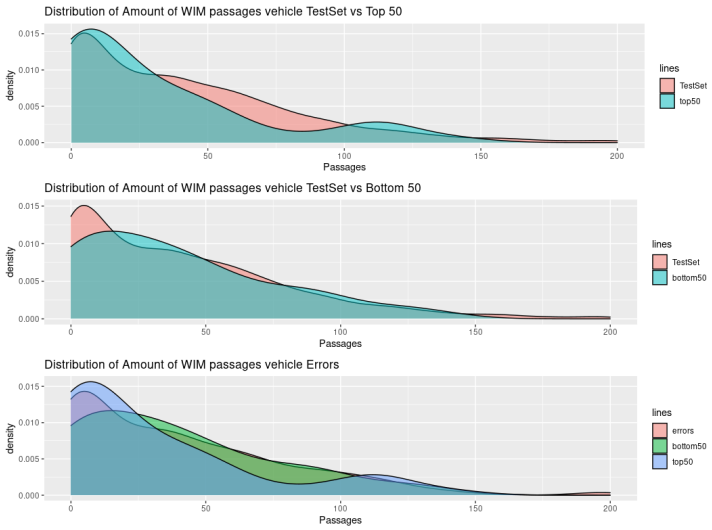


Figure A.22: Density for amount of WIM passages High Confidence Errors, Top 50 and Bottom 50 Errors

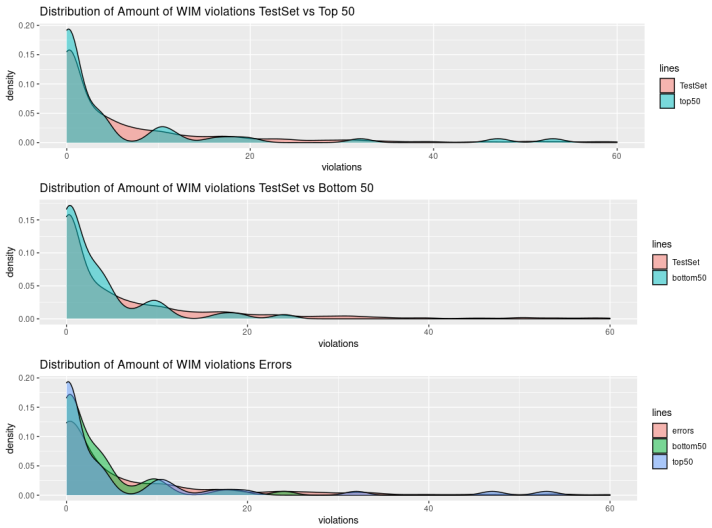


Figure A.23: Density for violations on WIM passages High Confidence Errors, Top 50 and Bottom 50 Errors

B

Variable Plots For First Error Prediction Model

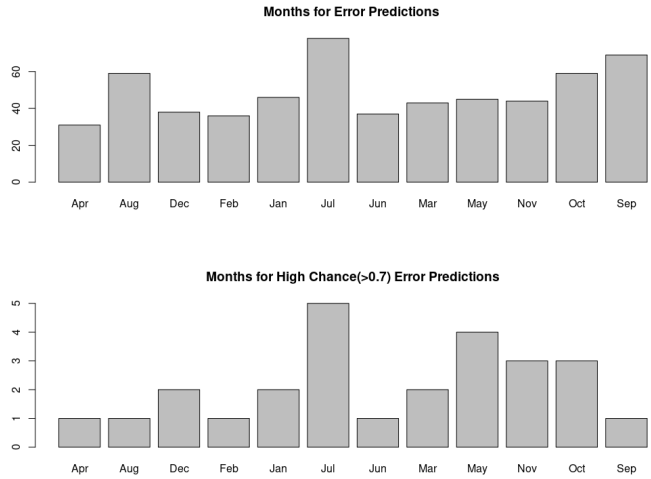


Figure B.1: Plots for Month distribution for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

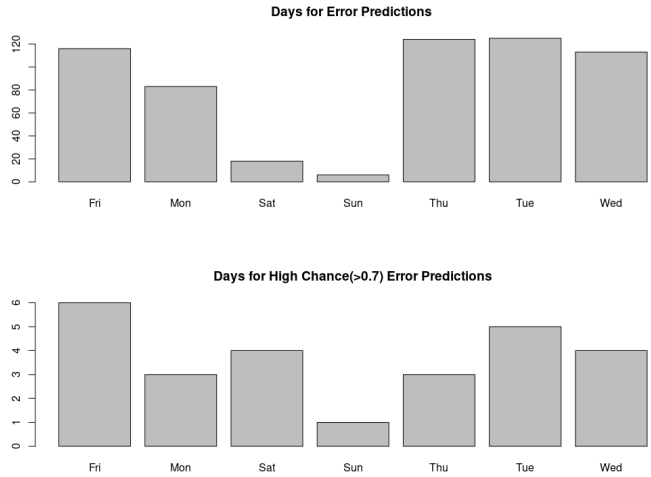


Figure B.2: Plots for Day distribution for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

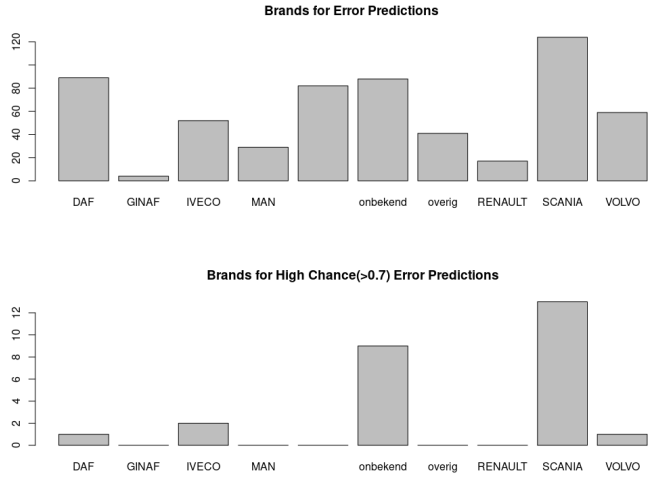


Figure B.3: Plots for Brand distribution for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

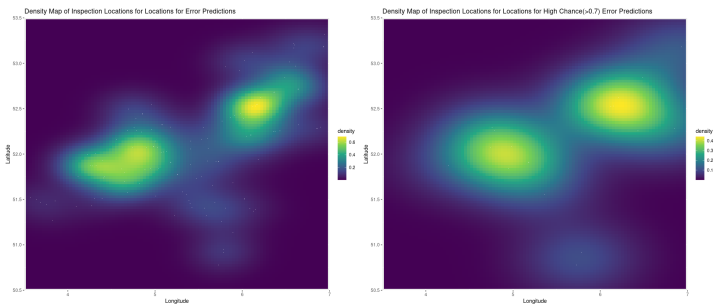


Figure B.4: Heatmap Density Plot for Location distribution for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

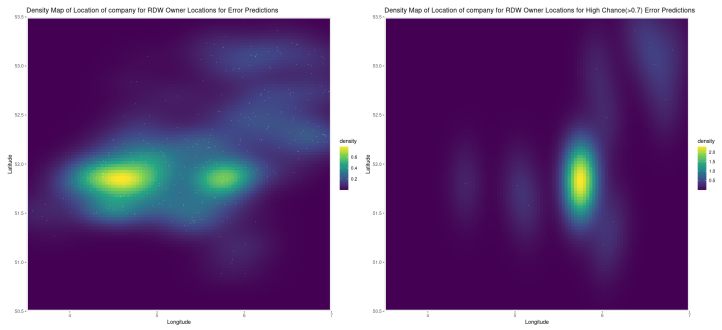


Figure B.5: Heatmap Density Plot for RDW owner locations for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

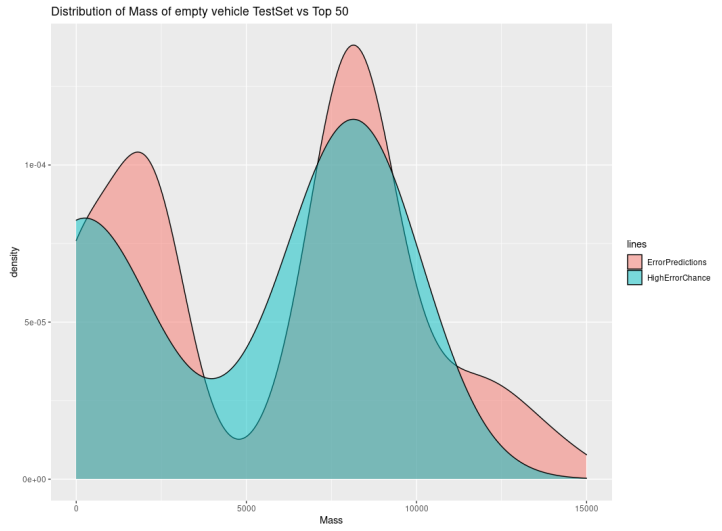


Figure B.6: Density Plot for Mass empty vehicle for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

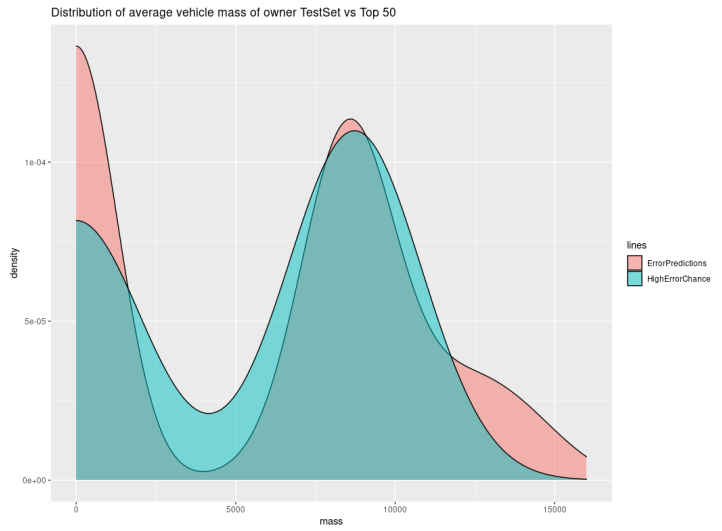


Figure B.7: Density for Average Mass of Vehicles owner for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

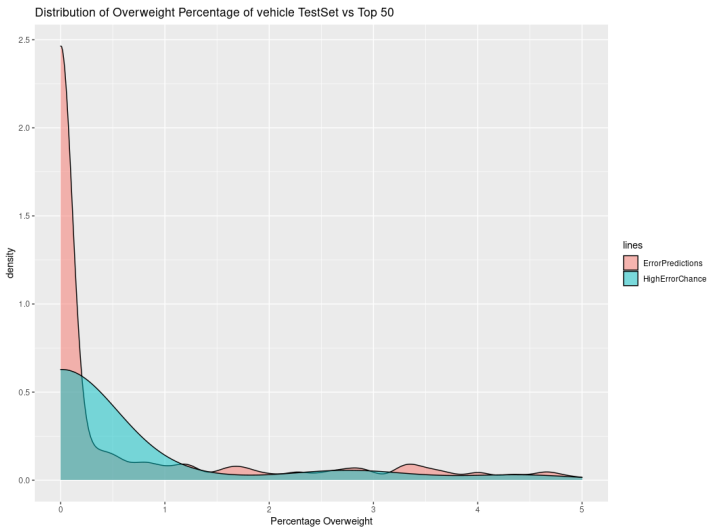


Figure B.8: Density for Percentage Overweight of Vehicle for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

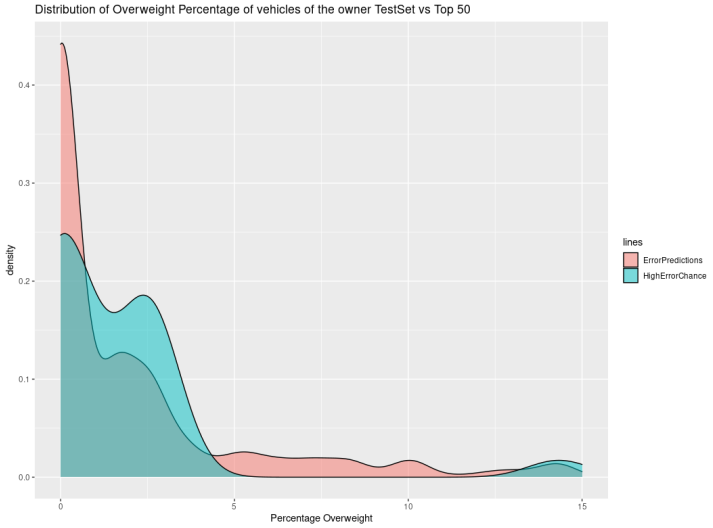


Figure B.9: Density for Percentage Overweight of Owner for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

Figure B.10: Density for age of the company for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

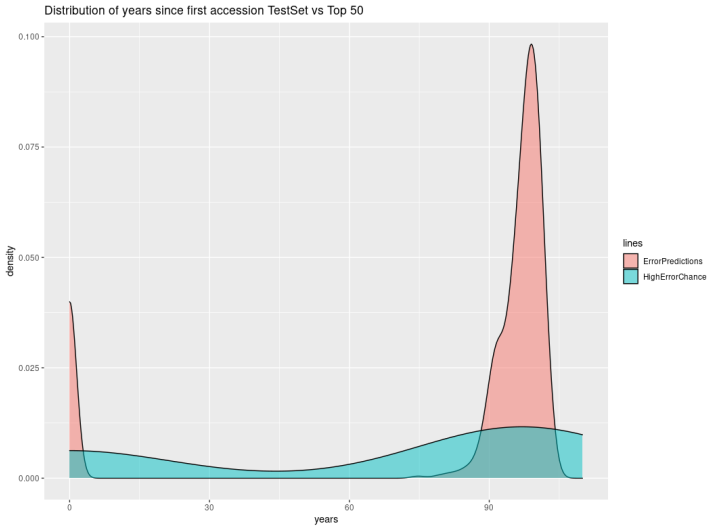


Figure B.11: Density for years since first accession for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

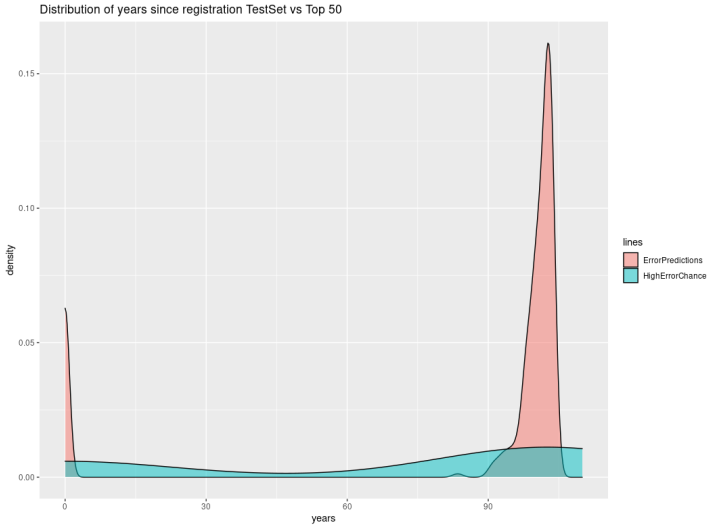


Figure B.12: Density for years since registration for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

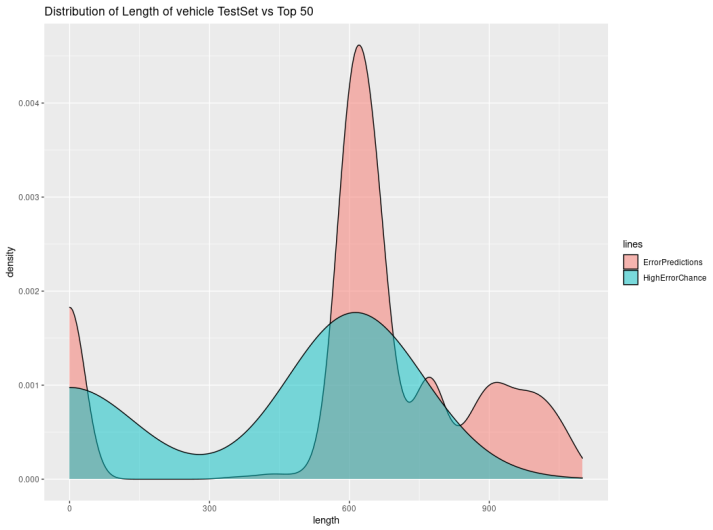


Figure B.13: Density for Length of Vehicle for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

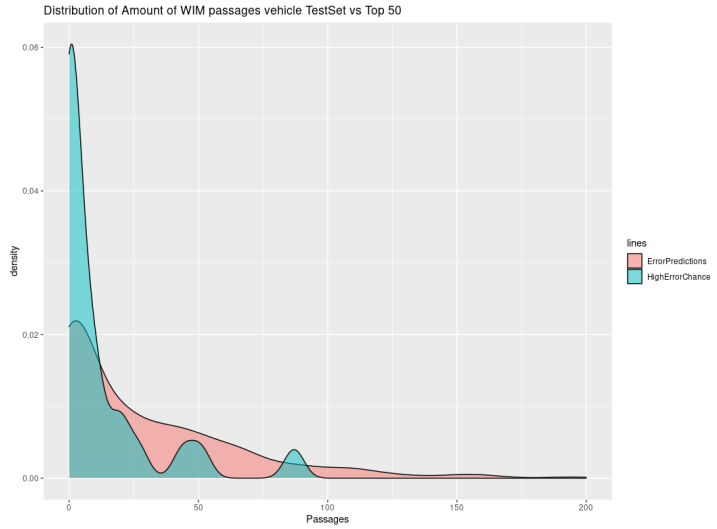


Figure B.14: Density for amount of WIM passages for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

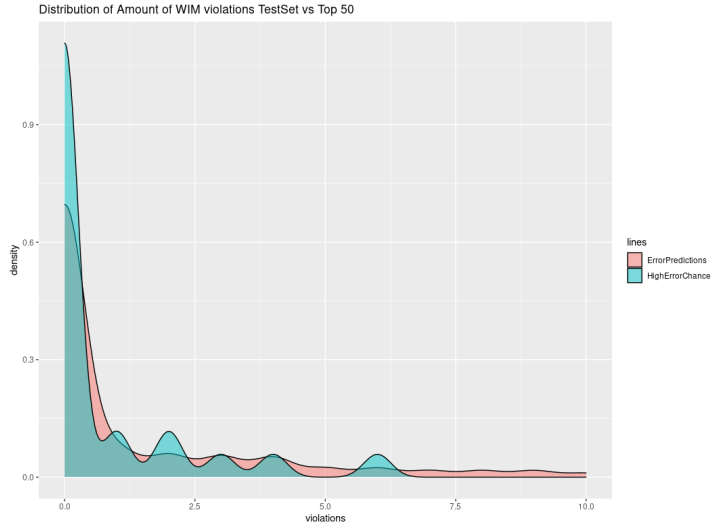
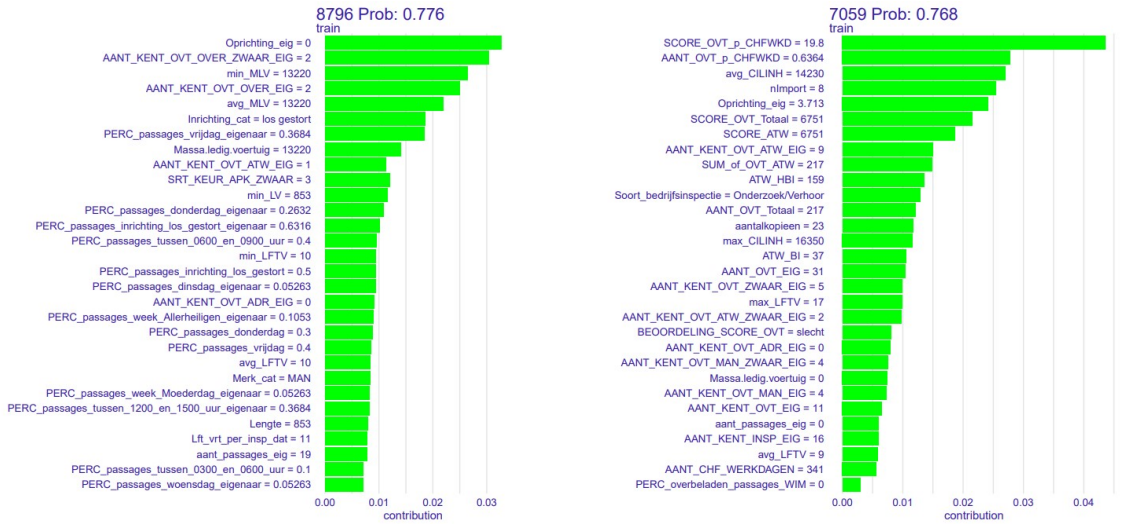


Figure B.15: Density for violations on WIM passages for Predicted Errors, All Errors vs. High Error Chance(probability > 0.7)

C

SHAP Plots: High Confidence Error Instances

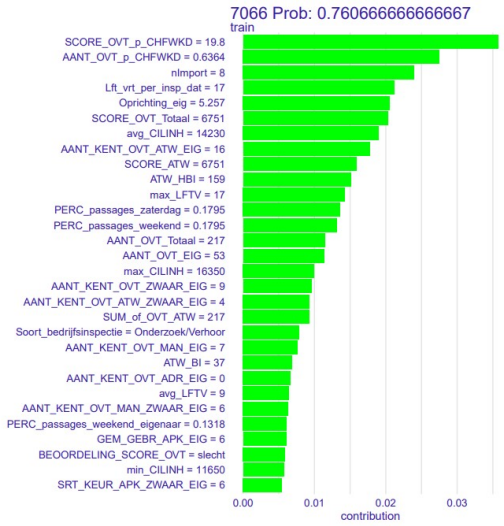
C.1. Top High Confidence Error SHAP plots

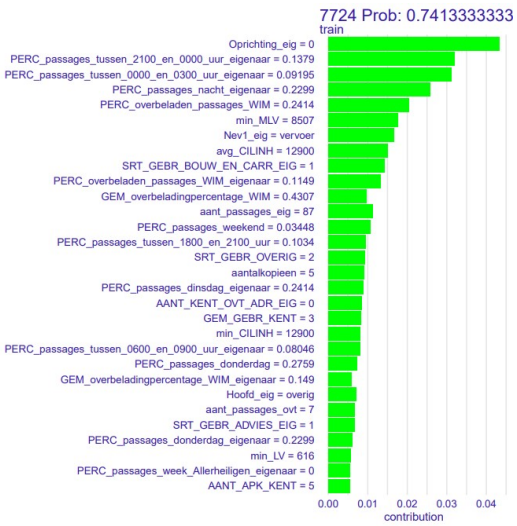


(a) SHAP contribution plots for top 1 High Confidence Error

(b) SHAP contribution plots for top 2 High Confidence Error

Figure C.1: SHAP contribution plots for the top High Confidence Error instances

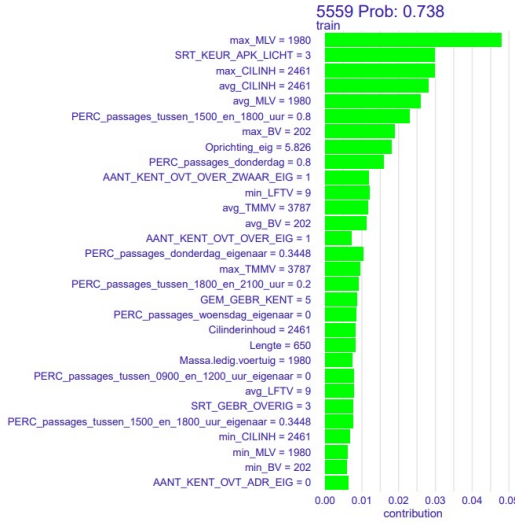


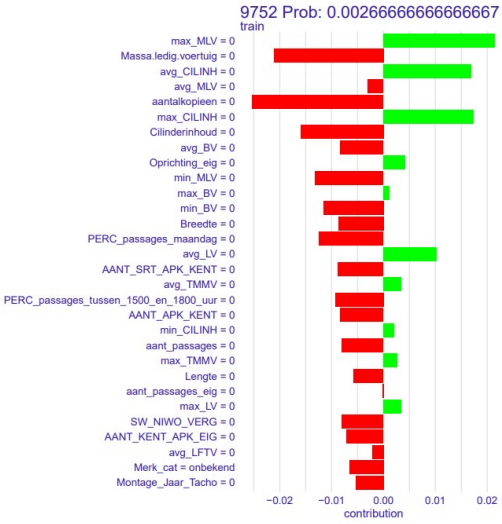


(a) SHAP contribution plots for top 3 High Confidence Error

(b) SHAP contribution plots for top 4 High Confidence Error

Figure C.2: SHAP contribution plots for the top High Confidence Error instances

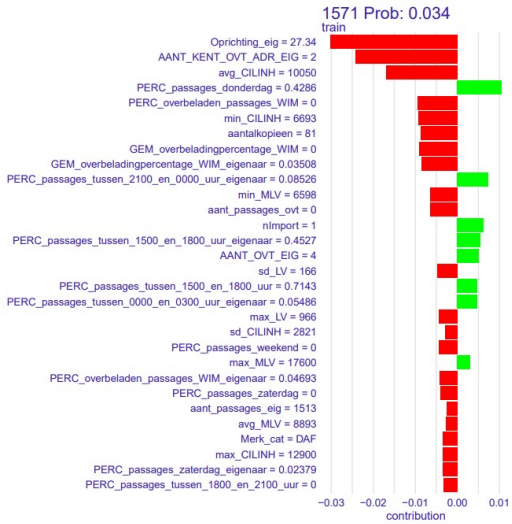




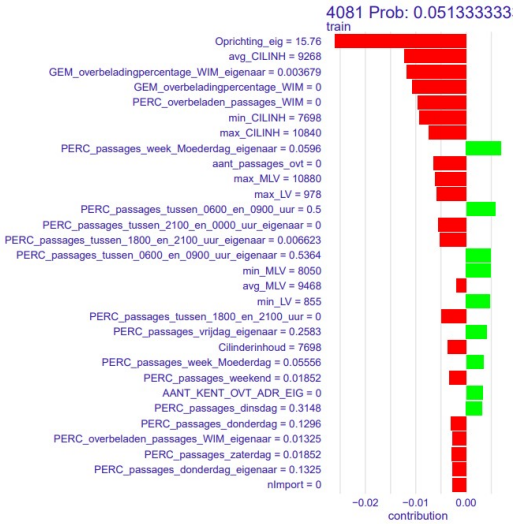
(a) SHAP contribution plots for top 5 High Confidence Error

(b) SHAP contribution plots for bottom 1 High Confidence Error

Figure C.3: SHAP contribution plots for the top and bottom High Confidence Error instances

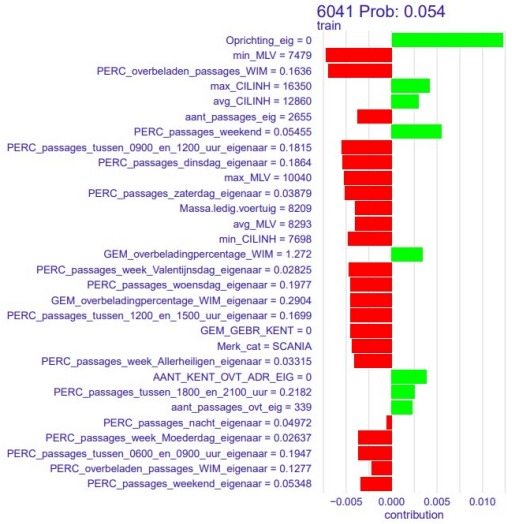


(a) SHAP contribution plots for top 3 High Confidence Error

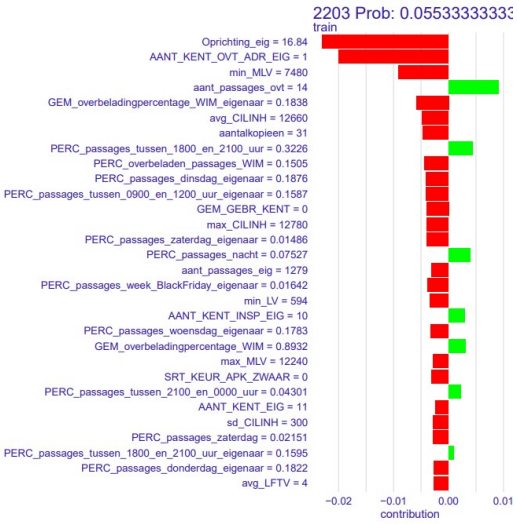


(b) SHAP contribution plots for top 4 High Confidence Error

Figure C.4: SHAP contribution plots for the bottom High Confidence Error instances



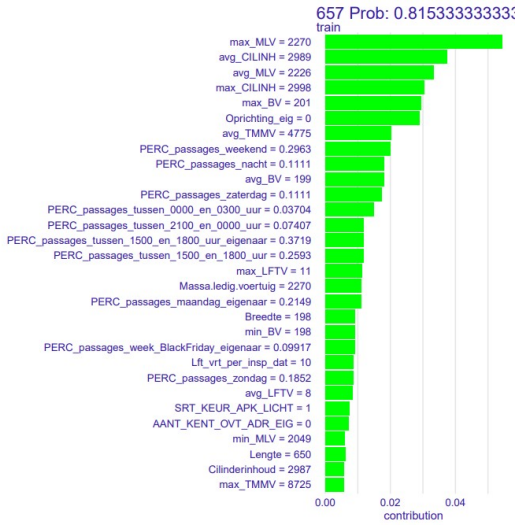
(a) SHAP contribution plots for top 4 High Confidence Error



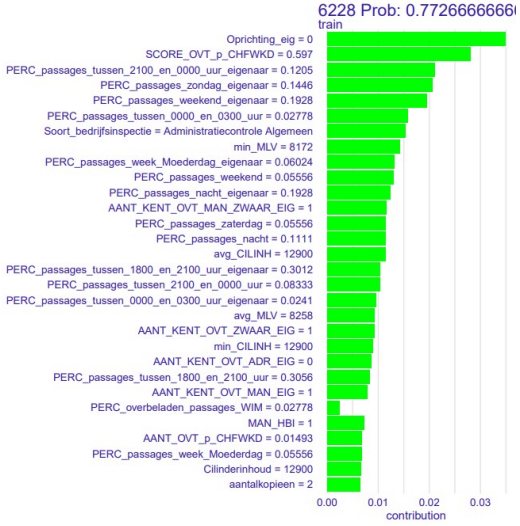
(b) SHAP contribution plots for top 5 High Confidence Error

Figure C.5: SHAP contribution plots for the bottom High Confidence Error instances

C.2. Top Correct Predictions SHAP plots

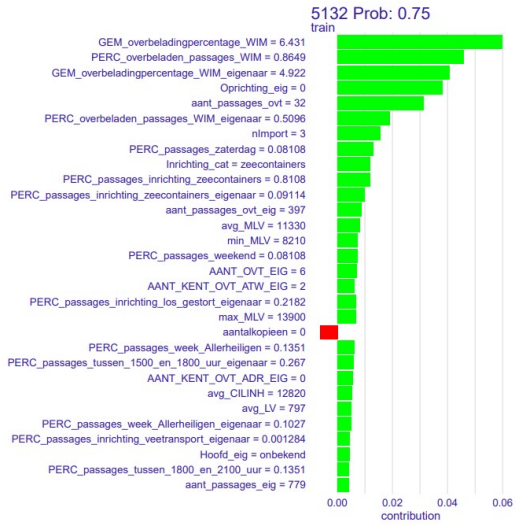


(a) SHAP contribution plots for top 1 Correct Prediction

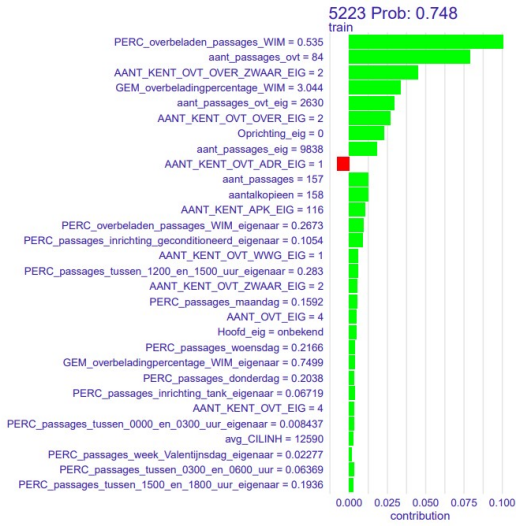


(b) SHAP contribution plots for top 2 Correct Prediction

Figure C.6: SHAP contribution plots for the top Correct Predictions

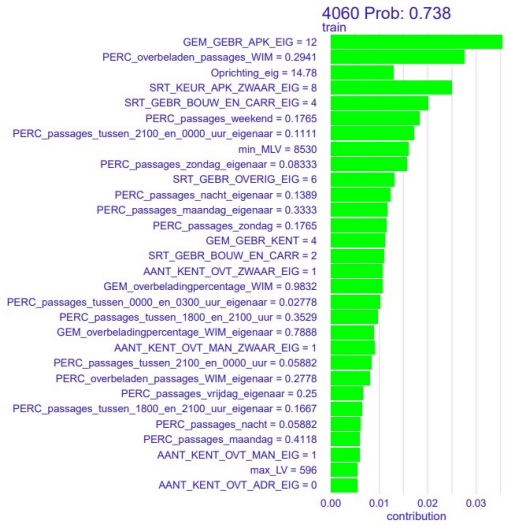


(a) SHAP contribution plots for top 3 Correct Prediction

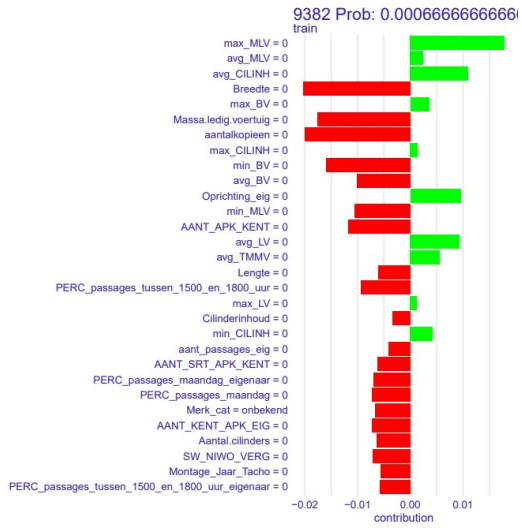


(b) SHAP contribution plots for top 4 Correct Prediction

Figure C.7: SHAP contribution plots for the top Correct Predictions



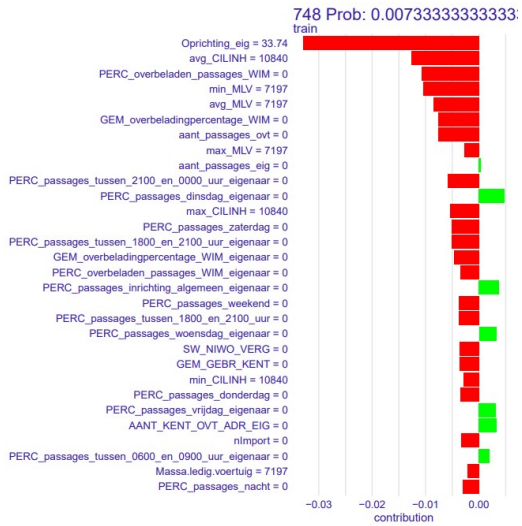
(a) SHAP contribution plots for top 5 Correct Prediction



(b) SHAP contribution plots for bottom 1 Correct Prediction

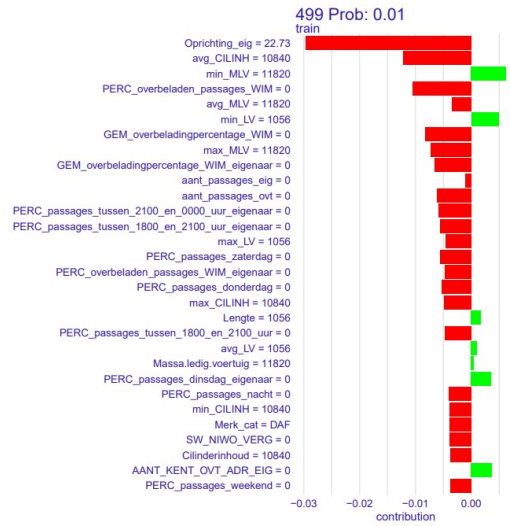
Figure C.8: SHAP contribution plots for the top and bottom Correct Predictions

(a) SHAP contribution plots for bottom 2 Correct Prediction

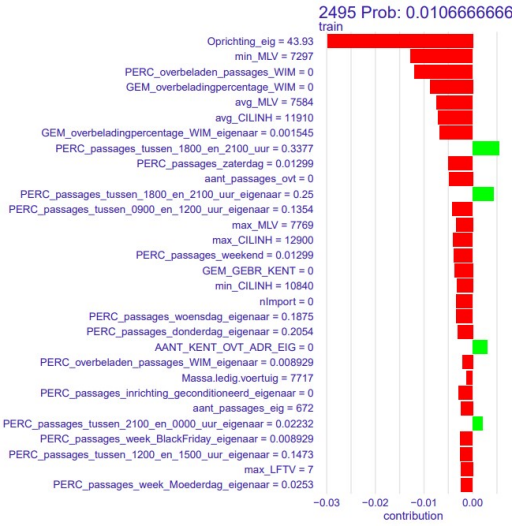


(b) SHAP contribution plots for bottom 3 Correct Prediction

Figure C.9: SHAP contribution plots for the bottom Correct Predictions



(a) SHAP contribution plots for bottom 4 Correct Prediction



(b) SHAP contribution plots for bottom 5 Correct Prediction

Figure C.10: SHAP contribution plots for the bottom Correct Predictions

D

HCOMP 2021 Paper

Exploring the Role of Domain Experts in Characterizing and Mitigating Machine Learning Errors

Pavel Hoogland[1], **Margje Schuur**[2], **Piet van Hoffen**[2], **Jasper van Vliet**[2], **Oana Inel**[1], **Jie Yang**[1]

[1] Delft University of Technology

[2] Inspectie Leefomgeving en Transport, Dutch Ministry of Infrastructure and Water Management

P.Hoogland@student.tudelft.nl, {O.Inel, J.Yang-3}@tudelft.nl, {margje.schuur, Piet.van.Hoffen, jasper.van.vliet}@ilent.nl

Abstract

In the use of machine learning systems, end-users' trust can often be hard to attain as many state-of-the-art systems operate as black-boxes. Errors produced by these systems, without further explanation as to why the decisions are made, will deteriorate trust. This effect is especially strong when these erroneous decisions are generated with a high confidence. This paper presents a human-in-the-loop methodology to characterize and mitigate high-confidence errors by engaging domain experts through a series of interaction sessions. We study the problem in the context of Road and Transportation law violations, by engaging inspectors in day-in-the-life and in-house interview sessions. We show that by bridging the knowledge gap between domain experts and data scientists through these iterative expert sessions, we can improve the model predictions and achieve increased user trust.

Introduction

The problem of a disconnect between *data scientists* and *domain experts* is often present when using machine learning (ML) systems (Viaene 2013). This disconnect can be present on different levels, i.e., the concept or the process (how-to) level (Mao et al. 2019; Convertino et al. 2008, 2009) and can lead to decreased *domain experts* trust in the system. Providing model outcomes without insights into why the predictions were made, could harm user trust even further and inhibit user-developer interactions. This is particularly the case for erroneous outcomes. Thus, understanding why certain predictions are made, is key to user engagement, system adoption, and sustainability (Sousa, Lamas, and Dias 2014).

Errors produced by machine learning systems fall into two broad categories. Errors near the decision boundaries of a model are more understandable as they are caused by the inherent variances within the data. However, when an erroneous decision is made far from the decision boundary it can hint at inherent issues with the data used to train the model, whether it be a lack or under-representation of data points. These errors are produced with a high model confidence, namely the High-Confidence Errors (HCEs). These should be avoided and kept to a minimum to preserve trust.

Data selection and feature construction can be seen as the main crux of classical machine learning. Previous work (Fails and Olsen 2003) has shown that for certain applications, a continuously interactive way of machine learning can lead to improvements. A study into development tools for statistical ML (Patel et al. 2008) has shown that there is a need for an exploratory and iterative process in the process of data, and feature selection. In the natural language processing domain, (Park et al. 2021) proposed interactive tools that enable sharing domain knowledge through domain concept extraction and label justifications. Nonetheless, the role of the *domain experts* themselves is often overlooked in the development of trustworthy machine learning systems.

This paper explores the effect of iterative experts' engagement in a series of interaction sessions on understanding the requirements, challenges, and characterizing system errors. We show that these sessions are a promising method to facilitate model development and build trust with the model.

Methodology

We start by describing our proposed *Iterative Expert Sessions* to bridge the knowledge gap between *domain experts* and *data scientists*. It consists of three types of sessions, namely *exploratory*, *in-depth*, and *analysis*, aiming to close the knowledge gap and help data scientists improve the model in an iterative manner. A graphical summary of the sessions is displayed in Figure 1.

Exploratory sessions aim to get a deeper understanding of the working practice of the *domain experts*; learn which are the indicators or features that experts use to ground their decisions on, and how they compare to how the machine learning model is built. The driving question of these sessions is *"Where do the domain experts put their focus when making their decision?"*. To assess this, we propose either a day-in-the-life style session where the *data scientists* join the experts in their practice, or an interview setting where *data scientists* choose topics of expertise for which they need more insight. In this session, *domain experts* provide *data scientists* with insights in their reasoning, to help them better understand what features and data are essential for predictions.

In-Depth sessions aim to focus on those points from the exploratory session(s) that require further study. The focus lies on understanding what data is useful for the model and how much it contributes to the prediction – extracting and ranking features that are vital to the model's goal. The driv-

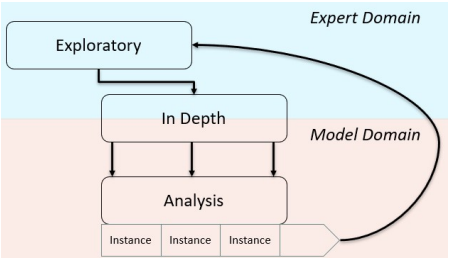


Figure 1: Graphical representation of Expert Sessions

.

ing question of this session is *"How does the available data contribute to the classification goal?"*. Answering this question sometimes requires further elaboration of information gathered from *domain experts* during the previous session. The in-depth sessions are performed as interviews, to offer space for discussions. The data scientists choose those areas in the data that they still have doubts in and ask from the *domain experts* to assess their relevance to the classification.

Analysis sessions serve two purposes: (1) present experts with actual use case data instances and model classifications and (2) provide hands-on experience with the model. By studying experts decisions and model classifications on actual data instances, we can better compare the model performance with how experts make decisions in real-world scenarios. Furthermore, we want to see how much the experts trust these model classifications. The driving question of this session is *"Do the domain experts trust the model predictions?"*. By accompanying model predictions with appropriate interpretation methods (we use a SHAP value variable contribution analysis (Lundberg and Lee 2017)), we offer *domain experts* hands-on experience with the model. The analysis sessions are performed as interviews. The *data scientists* present pre-selected data instances for which experts' input is informative, such as HCEs. The *domain experts* are to first give their assessment and classification of the instance. Then the model prediction is revealed and experts discuss similarities or differences in their judgement.

Case Study

We conducted this study in collaboration with the Netherlands Human Environment and Transport Inspectorate (ILT)[1]. The case study consists of assessing the risk of non-compliance for transportation vehicles on the Dutch roadways. The aim is to help road inspectors (*i.e.*, *domain experts*) to identify vehicles that are potentially non-compliant with road and transportation laws.

Data scientists at ILT developed a random forest based classifier which provides a risk score for potentially non-compliant vehicles, and thus can help select vehicles for examination. The classifier is trained on historical inspection data of Dutch vehicles. The historical inspection data is combined with data from other data sources such as vehicle and company registration.

[1]Inspectie Leefomgeving en Transport (ILT), Ministerie van Infrastructuur en Waterstaat: https://www.ilent.nl/

An *exploratory session* was held with 7 *inspectors*, 4 *data scientists* and 2 discussion coordinators. During this session the *data scientists* proposed 3 areas of exploration to the *inspectors*: overload, rest and driving times and cabotage. The *inspectors* brainstormed to identify and rank the most important indicators, while collaboratively taking notes. At this stage, *data scientists* can also ask clarifying questions. An *in-depth interview* session was held with 4 *inspectors* and 4 *data scientists*. Inspectors were queried on the following areas of interest regarding model data: moment and location of inspection, cargo, and maintenance, vehicle-ownership and company structures. Finally, an *analysis session* was held with 3 *inspectors* and 4 *data scientists*. The inspectors were presented with 8 data instances, the feature contributions (SHAP plot) and outcome of the model: 5 HCEs and 3 correct predictions. An image of the vehicle was also provided. Their judgement was used to assess the model performance.

Results

In the *exploratory session*, the inspectors concluded that vehicle violations are more prone to happen for certain types of freight/vehicles and on certain periods or days. This lead to revisiting of the feature set to better represent these aspects. The *in-depth sessions* led to conclude that more information about the vehicle maintenance and overload is needed. The vehicle maintenance information can hint at problems with the vehicle signaling a risk of error, and vehicle overload information can hint at overload risk. After making these changes, we saw a slight model performance improvement, but also less overfitting on biased historical data. In the *analysis session*, we found that experts agreed with all shown model predictions, disregarding the correctness of the prediction. Thus, at least for these cases, the model's judgement closely mirrors that of the inspectors and the presented HCEs do not deteriorate trust, since they still do indicate a potential risk.

In addition, we found that on-site visual characteristics of vehicles are extremely insightful for inspectors, but difficult to capture in the data. Nevertheless, even with a lack of visual features, the model still provides strong support for inspectors in the decision-making process.

Conclusion

We show that our proposed iterative session model can bridge the knowledge gap between *data scientists* and *domain experts*, in the context of road and transportation law violations. The model improvements made during the *exploratory* and *in-depth* sessions provided the inspectors with a more coherent prediction. We observed that some HCEs, even though an error at the time of inspection, still reflect a nature of risk of the vehicle - which helps in maintaining user trust in the system. Future inspections of these can still identify irregularities, making it difficult to distinguish real errors from model errors due to the temporal aspect. As future work, we plan to improve the effectiveness of the model by using a hybrid approach consisting of a predicted risk score, as well as providing feature importance values and decision rules, based on what the model has learned.

Acknowledgements

We would like to thank and acknowledge the Road Transportation inspectors that participated together with us during the case study sessions. Gert Bosman, Willem van Dijk, Jan van der Laarse, Govert Pelkmans, Erik Rozenbrand, Bert van Voorthuizen, Björn Weekhout. We also want to thank Mark van der Ham, teamleader of ILT Roadtransport, for helping schedule the sessions and select the inspectors. Special thanks to Karima Azzouz and Jacques Niehof, who helped coordinate the first exploratory session. Finally thanks to Victor Ciulei and Paul Ozkohen who helped provide feedback and insights during the development of the sessions.

References

Convertino, G.; Mentis, H.; Rosson, M. B.; Slavkovic, A.; and Carroll, J. 2009. Supporting content and process common ground in computer-supported teamwork. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009*, 2339–2348. doi: 10.1145/1518701.1519059.

Convertino, G.; Mentis, H. M.; Rosson, M. B.; Carroll, J. M.; Slavkovic, A.; and Ganoe, C. H. 2008. Articulating Common Ground in Cooperative Work: Content and Process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 1637–1646. New York, NY, USA: Association for Computing Machinery. ISBN 9781605580111. doi:10.1145/1357054.1357310. URL https://doi.org/10.1145/1357054.1357310.

Fails, J. A.; and Olsen, D. R. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, 39–45. New York, NY, USA: Association for Computing Machinery. ISBN 1581135866. doi:10.1145/604045.604056. URL https://doi.org/10.1145/604045.604056.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4768–4777. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Mao, Y.; Wang, D.; Muller, M.; Varshney, K. R.; Baldini, I.; Dugan, C.; and Mojsilović, A. 2019. How Data ScientistsWork Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? *Proc. ACM Hum.-Comput. Interact.* 3(GROUP). doi:10.1145/3361118. URL https://doi.org/10.1145/3361118.

Park, S.; Wang, A. Y.; Kawas, B.; Liao, Q. V.; Piorkowski, D.; and Danilevsky, M. 2021. Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models. In *26th International Conference on Intelligent User Interfaces*, IUI '21, 585–596. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380171. doi:10.1145/3397481.3450637. URL https://doi.org/10.1145/3397481.3450637.

Patel, K.; Fogarty, J.; Landay, J. A.; and Harrison, B. 2008. Investigating Statistical Machine Learning as a Tool for Software Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 667–676. New York, NY, USA: Association for Computing Machinery. ISBN 9781605580111. doi:10.1145/1357054.1357160. URL https://doi.org/10.1145/1357054.1357160.

Sousa, S.; Lamas, D.; and Dias, P. 2014. A Model for Human-Computer Trust. In Zaphiris, P.; and Ioannou, A., eds., *Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences*, 128–137. Cham: Springer International Publishing. ISBN 978-3-319-07482-5.

Viaene, S. 2013. Data Scientists Aren't Domain Experts. *IT Professional* 15(6): 12–17. doi:10.1109/MITP.2013.93.

Bibliography

[1] 2016. URL: https://ec.europa.eu/transport/modes/road/news/2016-03-18-classification-road-infringements-and-tachograph_en.

[2] *2018 reform of EU data protection rules*. European Commission. May 25, 2018. URL: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.

[3] William Adams. "Conducting Semi-Structured Interviews". In: Aug. 2015. DOI: 10.1002/9781119171386.ch19.

[4] Cristina M Alberini. *Long-term memories: The good, the bad, and the ugly*. Sept. 2010. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3574792/.

[5] David Alvarez-Melis and Tommi S. Jaakkola. *On the Robustness of Interpretability Methods*. 2018. arXiv: 1806.08049 [cs.LG].

[6] Joshua Attenberg, Panos Ipeirotis, and Foster Provost. "Beat the Machine: Challenging Humans to Find a Predictive Model's “Unknown Unknowns”" in: *Journal of Data and Information Quality* 6.1 (Mar. 4, 2015), 1:1–1:17. ISSN: 1936-1955. DOI: 10.1145/2700832. URL: https://doi.org/10.1145/2700832 (visited on 03/23/2021).

[7] Paul Bakker. *Determine and explain confidence in predicting violations on inland ships in the Netherlands*. 2020.

[8] Pablo Barceló et al. *Model Interpretability through the Lens of Computational Complexity*. 2020. arXiv: 2010.12265 [cs.AI].

[9] Thomas Beardsworth and Nishant Kumar. *Who to sue when a robot loses your fortune*. May 2019. URL: https://www.bloomberg.com/news/articles/2019-05-06/who-to-sue-when-a-robot-loses-your-fortune..

[10] Pieter Beers et al. "Common Ground, Complex Problems and Decision Making". In: *Group Decision and Negotiation* 15 (Nov. 2006), pp. 529–556. DOI: 10.1007/s10726-006-9030-1.

[11] Clément Bénard et al. *Interpretable Random Forests via Rule Extraction*. 2021. arXiv: 2004.14841 [stat.ML].

[12] Leo Breiman. "OUT-OF-BAG ESTIMATION". In: 1996.

[13] Leo Breiman. "Random Forests". English. In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A%3A1010933404324.

[14] Leo Breiman et al. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[15] John T. Cacioppo, Stephanie Cacioppo, and Jackie K. Gollan. "The negativity bias: Conceptualization, quantification, and individual differences". In: *Behavioral and Brain Sciences* 37.3 (2014), pp. 309–310. DOI: 10.1017/S0140525X13002537.

[16] Robyn Carston. "Herbert H. Clark, Using language. Cambridge: Cambridge University Press, 1996. Pp. xi+432." In: *Journal of Linguistics* 35 (Mar. 1999), pp. 167–222. DOI: 10.1017/S0022226798217361.

[17] Rich Caruana and Alexandru Niculescu-Mizil. "An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics". In: *In Proc. 23 rd Intl. Conf. Machine learning (ICML'06*. 2005, pp. 161–168.

[18] Chelsea Chandler, Peter W Foltz, and Brita Elvevåg. "Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness". In: *Schizophrenia Bulletin* 46.1 (Nov. 2019), pp. 11–14. ISSN: 0586-7614. DOI: 10.1093/schbul/sbz105. eprint: https://academic.oup.com/schizophreniabulletin/article-pdf/46/1/11/32927507/sbz105.pdf. URL: https://doi.org/10.1093/schbul/sbz105.

[19] Gregorio Convertino et al. “Articulating Common Ground in Cooperative Work: Content and Process”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’08. Florence, Italy: Association for Computing Machinery, 2008, pp. 1637–1646. ISBN: 9781605580111. DOI: 10.1145/1357054.1357310. URL: https://doi.org/10.1145/1357054.1357310.

[20] Gregorio Convertino et al. “How Does Common Ground Increase?” In: *Proceedings of the 2007 International ACM Conference on Supporting Group Work*. GROUP ’07. Sanibel Island, Florida, USA: Association for Computing Machinery, 2007, pp. 225–228. ISBN: 9781595938459. DOI: 10.1145/1316624.1316657. URL: https://doi.org/10.1145/1316624.1316657.

[21] Gregorio Convertino et al. “Supporting content and process common ground in computer-supported teamwork”. In: *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009*. Apr. 2009, pp. 2339–2348. DOI: 10.1145/1518701.1519059.

[22] Jeffrey Dastin. *Amazon scraps secret AI recruiting tool that showed bias against women*. Oct. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[23] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[24] Jerry Alan Fails and Dan R. Olsen. “Interactive Machine Learning”. In: *Proceedings of the 8th International Conference on Intelligent User Interfaces*. IUI ’03. Miami, Florida, USA: Association for Computing Machinery, 2003, pp. 39–45. ISBN: 1581135866. DOI: 10.1145/604045.604056. URL: https://doi.org/10.1145/604045.604056.

[25] Jerome H. Friedman and Bogdan E. Popescu. “Predictive learning via rule ensembles”. In: *The Annals of Applied Statistics* 2.3 (Sept. 2008). ISSN: 1932-6157. DOI: 10.1214/07-aoas148. URL: http://dx.doi.org/10.1214/07-AOAS148.

[26] Mouzhi Ge, Carla Delgado, and Dietmar Jannach. “Beyond accuracy: Evaluating recommender systems by coverage and serendipity”. In: Jan. 2010, pp. 257–260. DOI: 10.1145/1864708.1864761.

[27] Leilani H. Gilpin et al. “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018, pp. 80–89. DOI: 10.1109/DSAA.2018.00018.

[28] Bryce Goodman and Seth Flaxman. “European Union regulations on algorithmic decision-making and a “right to explanation””. In: *AI magazine* 38.3 (2017), pp. 50–57.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[30] Monika Hengstler, Ellen Enkel, and Selina Duelli. “Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices”. In: *Technological Forecasting and Social Change* 105 (2016), pp. 105–120. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2015.12.014. URL: https://www.sciencedirect.com/science/article/pii/S0040162515004187.

[31] Robert R. Hoffman et al. “Eliciting Knowledge from Experts: A Methodological Analysis”. In: *Organizational Behavior and Human Decision Processes* 62.2 (1995), pp. 129–158. ISSN: 0749-5978. DOI: https://doi.org/10.1006/obhd.1995.1039. URL: https://www.sciencedirect.com/science/article/pii/S0749597885710394.

[32] Fred Hohman et al. “Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19: CHI Conference on Human Factors in Computing Systems. Glasgow Scotland Uk: ACM, May 2, 2019, pp. 1–13. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300809. URL: https://dl.acm.org/doi/10.1145/3290605.3300809 (visited on 04/19/2021).

[33] Fred Hohman et al. *Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers*. 2018. arXiv: 1801.06889 [cs.HC].

[34] Andreas Holzinger. “Interactive machine learning for health informatics: when do we need the human-in-the-loop?” In: *Brain Informatics* 3.2 (2016), pp. 119–131. DOI: 10.1007/s40708-016-0042-6.

[35] Pavel Hoogland et al. “Exploring the Role of Domain Experts in Characterizing and Mitigating Machine Learning Errors”. In: *The Ninth AAAI Conference on Human Computation and Crowdsourcing HCOMP 2021* (Nov. 2021), p. 3.

[36] ILT. 2013. URL: https://www.ilent.nl/documenten/publicaties/2013/11/08/overzicht-weegpunten-in-nederland.

[37] Max Kuhn. *caret: Classification and Regression Training*. R package version 6.0-86. 2020. URL: https://CRAN.R-project.org/package=caret.

[38] Himabindu Lakkaraju et al. "Identifying unknown unknowns in the open world: representations and policies for guided exploration". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, Feb. 4, 2017, pp. 2124–2132. (Visited on 03/23/2021).

[39] Jeff Larson et al. *Machine bias*. May 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[40] Sam Levin and Julia Carrie Wong. *Self-driving uber kills Arizona woman in first fatal crash involving pedestrian*. Mar. 2018. URL: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

[41] Chunyang Li. "Probability Estimation in Random Forests". In: (), p. 35.

[42] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

[43] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1 (2020), p. 18. DOI: 10.3390/e23010018.

[44] Zachary C. Lipton. *The Mythos of Model Interpretability*. 2017. arXiv: 1606.03490 [cs.LG].

[45] Anthony Liu et al. "Towards Hybrid Human-AI Workflows for Unknown Unknown Detection". In: *Proceedings of The Web Conference 2020*. WWW '20. New York, NY, USA: Association for Computing Machinery, Apr. 20, 2020, pp. 2432–2442. ISBN: 978-1-4503-7023-3. DOI: 10.1145/3366423.3380306. URL: https://doi.org/10.1145/3366423.3380306 (visited on 03/23/2021).

[46] Junhua Lu et al. "Recent progress and trends in predictive visual analytics". English (US). In: *Frontiers of Computer Science* 11.2 (Apr. 2017). Publisher Copyright: © 2016, Higher Education Press and Springer-Verlag Berlin Heidelberg., pp. 192–207. ISSN: 2095-2228. DOI: 10.1007/s11704-016-6028-y.

[47] Yafeng Lu et al. "The State-of-the-Art in Predictive Visual Analytics". In: *Computer Graphics Forum* 36.3 (2017), pp. 539–562. DOI: https://doi.org/10.1111/cgf.13210. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13210. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13210.

[48] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. *Consistent Individualized Feature Attribution for Tree Ensembles*. 2019. arXiv: 1802.03888 [cs.LG].

[49] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964.

[50] James D. Malley et al. "Probability machines: consistent probability estimation using nonparametric learning machines." In: *Methods of information in medicine* 51 1 (2012), pp. 74–81.

[51] Yaoli Mao et al. "How Data ScientistsWork Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question?" In: *Proc. ACM Hum.-Comput. Interact.* 3.GROUP (Dec. 2019). DOI: 10.1145/3361118. URL: https://doi.org/10.1145/3361118.

[52] Ričards Marcinkevičs and Julia E. Vogt. *Interpretability and Explainability: A Machine Learning Zoo Mini-tour*. 2020. arXiv: 2012.01805 [cs.LG].

[53] European Commission. Directorate General for Mobility and Transport. *Next Steps Towards 'Vision Zero': EU Road Safety Policy Framework 2021-2030*. Publications Office of the European Union, 2020. ISBN: 9789276132196. URL: https://books.google.nl/books?id=pISBzQEACAAJ.

[54] Andrew Monk. "Common Ground in Electronically Mediated Communication: Clark's Theory of Language Use". In: (Dec. 2003). DOI: 10.1016/B978-155860808-5/50010-1.

[55] Soya Park et al. "Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models". In: *26th International Conference on Intelligent User Interfaces*. IUI '21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 585–596. ISBN: 9781450380171. DOI: 10.1145/3397481.3450637. URL: https://doi.org/10.1145/3397481.3450637.

[56] Kayur Patel et al. “Investigating Statistical Machine Learning as a Tool for Software Development”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’08. Florence, Italy: Association for Computing Machinery, 2008, pp. 667–676. ISBN: 9781605580111. DOI: 10.1145/1357054.1357160. URL: https://doi.org/10.1145/1357054.1357160.

[57] *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. European Commission. Apr. 21, 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%5C%3A52021PC0206.

[58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. arXiv: 1602.04938 [cs.LG].

[59] Cynthia Rudin. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. 2019. arXiv: 1811.10154 [stat.ML].

[60] Tom Seymoens et al. “A Methodology to Involve Domain Experts and Machine Learning Techniques in the Design of Human-Centered Algorithms”. In: *Human Work Interaction Design. Designing Engaging Automation*. Ed. by Barbara Rita Barricelli et al. Cham: Springer International Publishing, 2019, pp. 200–214. ISBN: 978-3-030-05297-3.

[61] L. S. Shapley. *17. A Value for n-Person Games*. Dec. 1953. DOI: 10.1515/9781400881970-018. URL: http://dx.doi.org/10.1515/9781400881970-018.

[62] Sonia Sousa, David Lamas, and Paulo Dias. “A Model for Human-Computer Trust”. In: *Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences*. Ed. by Panayiotis Zaphiris and Andri Ioannou. Cham: Springer International Publishing, 2014, pp. 128–137. ISBN: 978-3-319-07482-5.

[63] Simone Stumpf et al. “Interacting meaningfully with machine learning systems: Three experiments”. In: *International Journal of Human-Computer Studies* 67.8 (Aug. 1, 2009), pp. 639–662. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2009.03.004. URL: https://doi.org/10.1016/j.ijhcs.2009.03.004 (visited on 04/22/2021).

[64] Colin Vandenhof and Edith Law. “Contradict the Machine: A Hybrid Approach to Identifying Unknown Unknowns”. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS ’19. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, May 8, 2019, pp. 2238–2240. ISBN: 978-1-4503-6309-9. (Visited on 03/23/2021).

[65] Jennifer Wortman Vaughan and Hanna Wallach. “A Human-Centered Agenda for Intelligible Machine Learning”. In: (Aug. 31, 2020). URL: https://www.microsoft.com/en-us/research/publication/a-human-centered-agenda-for-intelligible-machine-learning/ (visited on 03/23/2021).

[66] Stijn Viaene. “Data Scientists Aren’t Domain Experts”. In: *IT Professional* 15.6 (2013), pp. 12–17. DOI: 10.1109/MITP.2013.93.

[67] Sandra Wachter, Brent Mittelstadt, and Chris Russell. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. 2018. arXiv: 1711.00399 [cs.AI].

[68] Doris Xin et al. “Accelerating Human-in-the-Loop Machine Learning: Challenges and Opportunities”. In: *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*. DEEM’18. Houston, TX, USA: Association for Computing Machinery, 2018. ISBN: 9781450358286. DOI: 10.1145/3209889.3209897. URL: https://doi.org/10.1145/3209889.3209897.

[69] Pulei Xiong et al. *Towards a Robust and Trustworthy Machine Learning System Development*. 2021. arXiv: 2101.03042 [cs.LG].