Optimising flight pass prices

A Kenya Airways case study

M.M.D. Gillis

Technische Universiteit Delft

Optimising flight pass prices

A Kenya Airways case study

by

M.M.D. Gillis

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday March 13, 2018 at 10:00 AM.

Student number:	4176367	
Project duration:	August 13, 2017 – March 13, 2018	
Thesis committee:	Dr. ir. B.F. Santos,	TU Delft, supervisor
	Dr. ir. H.G Visser,	TU Delft
	Dr. ir. E. van Kampen ,	TU Delft
	Marco van Vliet,	Kenya Airways, Head of Global Sales, supervisor
	Thomas Omondi,	Kenya Airways, Director Strategy and Performance mgmt.

Preface

After 9 months of hard work, I can proudly present my thesis. I was lucky to find a subject I am very passionate about. This gave me a lot of enthusiasm and curiosity to finish my research in time. I am very grateful that I received the opportunity to perform my research at Kenya Airways. I learned a lot about the airline business and the Kenyan culture. Living in Kenya comes with some challenges and difficulties but the hospitality and kindness of the Kenyans make you forget these challenges. This project was not possible without the help of certain persons. Therefore, I would like to thank everyone who supported me during the project.

First of all I want to thank my supervisor, Bruno Santos for his guidance during my research. He taught me how to think creatively and how to perform research. Our discussions and his critical feedback were both valuable and necessary.

Next, I want to thank my Kenyan supervisors and colleagues. First of all, Marco van Vliet, who helped a lot with the research and always made time for me in his busy schedule and I want to thank Marco for the warm welcome he and his family gave me in Kenya.

I want to thank Thomas Omondi who gave me the opportunity to do my project at Kenya airways and I want to thank Thomas for our lunches and for always being there if I needed something.

Many thanks as well to my colleagues from global sales, Anne, Peter, Carol, Eunice and Richard for the amazing time in Kenya. I am grateful to everyone I met in Kenya: you all contributed to my wonderful experience.

To Jef, thank you for the support, patience during the times I was in Nairobi and the help with structuring my thoughts regarding my thesis.

I want to thank my parents, who always supported me with everything during my time as a student and for their endless confidence in me even when I did not believe in it myself.

To Erik, thank you for your advise on my thesis and for help me out whenever I needed it. Without your knowledge about machine learning models, the result would not have been the same.

Thanks Bram, for your help on building my web application and for your time and support whenever it failed.

To Jef, Alexander and Marijke for proofreading every word of my thesis.

To all my friends, thank you for the support and the amazing time we had in Delft. You are the reason for my daily smile!

M.M.D. Gillis
Delft, February 2018

Abstract

Motivation and Problem statement

Airlines encounter challenges with respect to satisfying both the customer satisfaction and customer fidelity. Therefore, the flight pass is brought to life with the goal of increasing both satisfaction and fidelity of customers. The concept of the flight pass is based on the fact that customers pre-purchase a number of flights for a flat fee. This flat flee can be customised by choosing different options, such as the number of flights, the travel period, how early the flight can be booked, which cabin is used,..etc.

The industry has the urging request to scientifically support the prices for the pass and the extra discounts and fees for the different options.

A model is designed to fill this research gap. The requirement from the industry is to avoid dilution. Since the flight pass is a new concept, there is no useful data available yet. However, there is historical booking data from revenue management available. This data includes the effects of price discrimination, market segmentation, product differentiation and inventory control in order to avoid dilution, which is the practice of customers paying less for their tickets than the price they are actually willing to offer. Furthermore, this booking data includes additional information such as: (1) Month of flying; (2) Day (of week) of flying; (3) Time (of day) of flying, (4) Length of stay, (5) Ticketing lead time, (6) Cabin, (7) Point of sale,... This additional information is called, "features".

Methodology

The booking data from revenue management and additional booking information (month of flying, day of week of flying, time of day of flying, length of stay, ticketing lead time, cabin, point of sale, ...) can be used to predict the ticket price of an individual flight. Three models are compared based on evaluation metrics to predict the price of a ticket: (1) A multiple linear regression model; (2) A random forest regression; (3) A multilayer perceptron neural network. These models are compared in terms of (1) Fit; (2) Root mean squared error and (3) Mean absolute error.

Since 46 routes are hard to analyse, clusters are made to determine the importance of the different features in predicting the flight price. This is done with the use of the k-means clustering model.

Results

The random forest regression performs best for every of the 46 routes in terms of fit, root mean squared error and mean absolute error. The mean of all the routes is shown in Table 1. As can be seen the fit (R) is significantly better for the random forest regression and the errors are significantly smaller (RMSE and MAE).

Table 1: Mean of final results

	R^2	RMSE	MAE
MLR	0,304	0,268	0,194
RFR	0,569	0,210	0,130
NN	0,370	0,254	0,179

The features are clustered, to analyse which features are mainly determining the price. The routes are grouped in 5 clusters. The result of the clusters can be seen in Figure 1 where Table 2 gives the relative feature importance for each cluster.

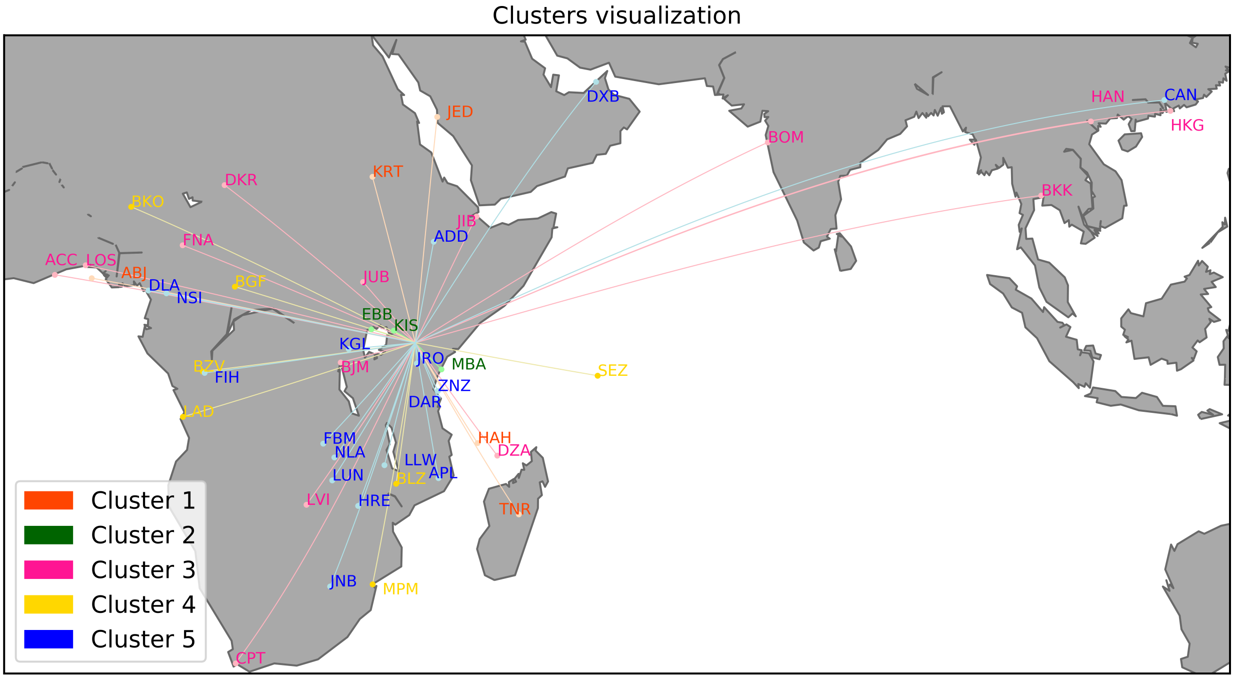


Figure 1: Clusters presented on the KQ network

Table 2: Relative feature importance for each cluster

	RET	**SUN + LOS**	**TOD**	**DOW**	**Month**	**TLT**
Cluster 1	- -	-	- -	+/-	++	-
Cluster 2	- -	- -	++	+	-	+/-
Cluster 3	-	+/-	-	+/-	+/-	+
Cluster 4	++	++	- -	- -	-	+/-
Cluster 5	+	-	+	+/-	+/-	+/-

Simulation

After predicting the value of a specific individual flight, the flight pass price can be determined. This is based on every flight that can be booked within the selected bounds of the flight pass. Therefore, a simulation based on the flying behaviour of a passenger is made to determine the weighted flight price. A Monte-Carlo simulation is used to simulate the behaviour of a passenger. The result of this simulation can be seen in a web-application built to visualise the model. This can be accessed via: `https://ec2-35-158-56-26.eu-central-1.compute.amazonaws.com:8888/notebooks/Run.ipynb?dashboard`

Conclusion

The requirement from the industry was to avoid dilution. This requirement has been translated to a model where additional revenue management booking data is used to predict the value of a flight. Three models are compared and the random forest regression performs significantly better than the multiple linear regression and the multilayer perceptron neural network. With a fit varying from 34 percent till 77 percent and an average of 57 percent the model can adequately predict the flight price. As a surplus, the model gives insights in the importance of features in predicting the price route dependent. It can be concluded that with the model, Kenya airways can start with offering flight pass prices that are data-drive. Since there is no flight pass booking data available yet, no valid validation can be done and, therefore the first recommendation would be to validate the results with flight pass data.

Contents

List of Figures

List of Tables

List of Equations

Nomenclature

List of Abbreviations

AA	American Airlines
ADAM	Adaptive Moment Estimation
BA	British Airways
CAB	Civil Aviation Board
CART	Classification and Regression Tree
CRM	Customer relationship management
DINAMO	Dynamic Inventory Optimisation and Maintenance Optimiser
DOW	Day Of Week
FFP	Frequent-Flyer Program
FSC	Full Service Carriers
GDS	Global distribution system
KLM	Koninklijke Luchtvaart Maatschappij
KMC	k-Means Cluster
KQ	Kenya Airways
LBFGS	Limited-memory Broyden–Fletcher–Goldfarb–Shanno
LCC	Low Cost Carriers
LF	Load Factor
LF	Load Factor
LOS	Length Of Stay
MAE	Mean Absolute Error
MLR	Multiple Linear Regression

NN	Neural Network
OD	Origin - Destination
POS	Point Of Sale
RELU	Rectified Linear Unit
RELU	Stochastic Gradient Descent
RET	Return flight
RFR	Random Forest Regression
RM	Revenue management
RMSE	Root Mean Squared Error
SUN	Sunday Stay
TLT	Ticketing Lead Time
TOD	Time Of Day

Glossary

Cabin	The discrimination between economy and business class
Destination	The city from from where the inbound flight departs
DOW	The day of the week the flight takes place
Flight	The number of the flight
LOS	The amount of days between the outbound and return flight
Month	The Month the flight takes place
Origin	The city from where the outbound flight departs
POS	The country from where someone buys a flight ticket
RET	Whether the flight is a one way or a return trip
Subclass	The fare class of the booking
SUN	Whether there is a Sunday stay between the outbound and return flight
TLT	The amount of days between the purchase of a ticket and the flight
TOD	The departure time of the flight

List of Symbols

β_0	Intercept
β_p	Slopes or coefficients
ϵ	Residuals
$\hat{y}_i$	The predicted value of the dependent value
μ_i	Centroids
D_i	Cook's Distance
h_i	Leverage
R^2	Coefficient of determination
S_i	Different clusters
x_i	Independent variable
y_i	The real value of the dependent value
d	Distance
DW	Durbin-Watson coefficient
e	Error
k	The number of clusters
m	Dimensionality
MSE	Mean Squared Error
N	The number of observations
r	Pearson correlation coefficient
s	Silhouette coefficient
SSE	The error sum of squares
SSR	The regression sum of squares
SST	The total sum of squares
t	Studentised deleted residual
VIF	Variance inflation factor
Y	Dependent variable

1 Introduction

1.1. Motivation

From the start of commercial flying, airlines have attempted to maximise their revenue. As the current airline market is highly competitive, there is a constant need to innovate and gain market share. That is why, over the years, airlines have come up with several new innovative business models and implemented many of those.

Another current trend, is that service-oriented business move towards flat fee prices like: monthly data plan for a smart-phone, public transport pass, Netflix, Spotify and many more,..

Combining the state of art of the airline industry and the current trends in the service-oriented business, brings the airline industry to a logical new product: **The flight pass**.

1.2. The flight pass

The structure of the flight pass is based on the fact that customers pre-purchase a number of flights for a flat fee. This flat flee can be customised by choosing different options, such as the number of flights, the travel period, how early the flight can be booked, which cabin is used,..etc. The passenger's total price depends on the options chosen. By using this structure, the airline company wants to give extra freedom to its passenger. The structure and use of the flight pass is shown in Figure 1.1.

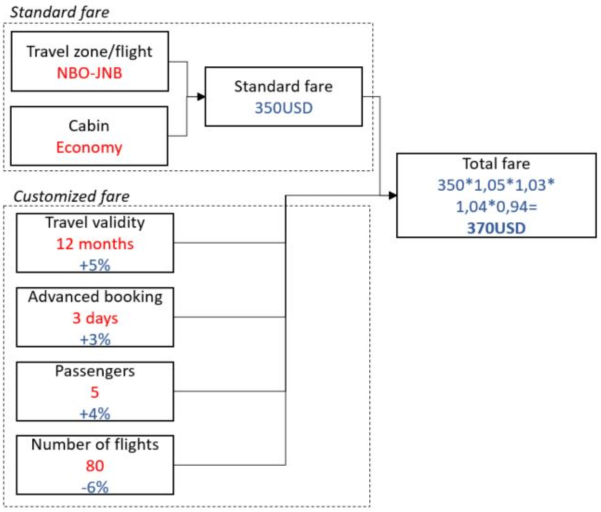


Figure 1.1: The structure of the flight pass

As can be seen in Figure 1.1, the standard fare is determined by the route and the cabin. In contrast, the customised (flexible/endogenous) fare depends on several options depicted in red on Figure 1.1, whereas fictive examples of prices are rendered in blue. The total price shown is the total price per one-way flight.

The options selected (red) are defined as follows:

- A travel validity of 12 months refers to the duration of the travel period wherein the flight pass is valid
- Advanced booking of 3 days refers to how many days before departure the flights need to be booked.
- 5 passengers refers to the maximum number of passengers who can book and fly using the flight pass.
- 80 number of flights refers to the total number of flights that can be book with the flight pass.

The flight pass as described above is a rather new pricing structure. The practice started in 2015 and more and more airlines such as: British Airways, Oman Air, Air Asia X, Vietnam Airlines, Air France and the Koninklijke Luchtvaart Maatschappij (KLM) are experimenting with this structure. This means that they are selling the passes based on the structure described in Figure 1.1. However every airline has the freedom to choose which options they want to offer to customise the flight pass.

Kenya Airways (KQ), who started to sell the flight pass one year ago, provided the context and the data to perform this research.

The main **goal** of the flight pass concept is to increase market share by attracting new customers and increase customer loyalty. Currently, the flight pass pricing is based on experience and market knowledge instead of having a scientific tool supporting the pricing. Therefore, the **objective** of the research is to determine a methodology to define the settings for an ideal flight pass price.

Before diving into the research, some additional benefits of the flight pass are listed below to give some more context and background information about the advantages of the flight pass.

- Customer satisfaction: The flight pass is a new product to better serve frequent flyers. It provides more comfort and (the perception of) a good deal.
- Price transparency: The prices of airline tickets do not have a good reputation among costumers [84]. The fares of the tickets are complex and can differ for two people sitting next to each other on the same flight. Cheap prices can be available today and can change extremely tomorrow. This has led to a certain antipathy and incomprehension towards the method airlines use to price their tickets. There is a lack of transparency in the prices from a customer perspective. Miao and Mattila studied the effect of price transparency on customers perception. The result showed that the price fairness perception and willingness-to-pay grows when the information on pricing is highly transparent [69]. Ferguson and Ellen, concluded in their research that if the perceived price is fair, the satisfaction of the customer will be positively strengthened and satisfaction strengthens loyalty [43]. The flight pass price are managed by the customer itself. Every customised option has an effect on the price of the pass. The customer gets insights in the price structure.
- Customer data and profiling: The data obtained can be used for customer relationship management (CRM) purposes. The flight pass gives insights in the flying behaviour of passengers, A database with customer data can determine the value of each individual customer and, eventually, the data can help making customised offers.
- Save on distribution costs: Most airline tickets are sold using third party distributors like online travel agencies. Those parties use the Global Distribution System (GDS). GDS is an access point for reserving seats at an airline for (online) travel agents [38]. According to an expert in the airline sales field, the costs of using the GDS are around 10 dollars per ticket. The flight pass does not need the GDS and therefore those 10 dollars per flight are saved.
- Cash flow: When the flight pass is bought, the full pass has to be paid. This results in more cash in advance for the airline.

1.3. Report structure

The report starts with a literature study about airline revenue management and predictive models. This is described in Chapter 2. In Chapter 3, the research plan including the research goal, objective, scope and impact is discussed. The methodology used to answer the research questions, is described in Chapter 4. The case study performed at Kenya airways including the data treatment and the use of the models, is described in Chapter 5. The results of the case study are discussed in Chapter 6. A simulation of the flight pass passenger flying behaviour is illustrated in Chapter 7 and a validation and verification are described in Chapter 8. Finally, conclusions, research recommendation and practical recommendation for Kenya Airways are given in Chapter 9.

2
Literature Study

The main **goal** of the flight pass concept is to increase market share by attracting new customers and increase customer loyalty. However, it might be difficult to avoid dilution of loyal customers, which is the practice of customers paying less for their tickets than the price they are actually willing to offer. This means that dilution can be avoided when the price of the ticket equals the maximum price a customer is willing to pay for the ticket.

Avoiding dilution might, therefore, sound simple in practice, but in reality it is rather difficult. If preventing dilution is difficult when pricing the flight pass, Revenue Management (RM) could offer some help. RM systems are built with the goal to prevent dilution. Understanding how the different systems work and how research is conducted, is crucial to determine how the pricing of the flight pass can benefit from RM research. However, one should keep in mind that the flight pass does not replace the RM system, but is an additional system next to the RM system. Therefore, the two systems should be considered in conjunction, meaning that the flight pass price is influenced by RM prices and vice versa.

2.1. Airline revenue management

As described in the introduction, understanding how the different RM systems work and how research is conducted, can be beneficial in determining the pricing of the flight pass.

Nowadays, there are many definitions of RM, making the definition of RM difficult to understand due to its inconsistent usage. McGill and van Ryzins definition of RM includes pricing [71] while Belobaba defined RM as the subsequent process after pricing [18]. To avoid misapprehension, this literature study defines revenue management as the overall process determining ticket prices. RM exists of both pricing and inventory control.

First, the origin of the mature systems of RM is outlined in Chapter 2.1.1. Second, in Chapter 2.1.2 the different sub-systems of the RM system are explained and the (recent) trends are listed in Chapter 2.1.3.

2.1.1. History of revenue management

In 1972, British Airways (BA) sold discounted early bird tickets for seats that would otherwise be empty. This experimental offering of differentiated fare products could be considered as the start of RM. The first one to analyze a simple two fare class model (discount and normal price) of capacity allocations on a single flight leg was Littlewood (working at BA at that moment) in 1972 [64].

In 1977, American Airlines (AA) with Bob Crandall as CEO, introduced a new discount fare named the super saver fares. The tickets could only be bought a long time in advance, reserving thirty percent of the seats on the plane for the super saver fares. What Bob Crandall quickly realized was that the super saver fares were not optimally allocated. This is because the demand was varying by route depending on the time of day, the day of week and season, making it clear that more investment was needed to uncover underlying demand patterns [37]. That is why AA started with forecasting and monitoring passenger demand by constructing large databases and developing computer systems. Bob Crandall is credited with giving this integrated set of people, process and systems a name. He called it ‘Yield Management’.

With the airline deregulation act in 1978, the U.S. Civil Aviation Board (CAB) stopped with strictly managing airline prices, where the regulation was based on profitability targets and fixed price settings. Consequently, a rapid change and a stream of innovation was seen in the industry. This resulted in Full service carriers (FSC) being free to change schedules, services and prices and new LCC could enter the market with lower prices that the FSC.

Bob Crandall was forced to enhance the current systems and invested millions in a new generation system called DINAMO (Dynamic Inventory Optimisation and Maintenance Optimiser). In 1985, DINAMO was fully implemented together with the new Ultimate Super Saver Fares. The fares where the lowest discount fares available in the American market. The Yield Management system carefully targeted those discounts to only those situations where they had ‘surplus seats’ that they could use to outmanoeuvre the competition. AA was the pioneer in ’Yield management’ and more airlines followed. United Airlines was the first to develop and implement an Origin-Destination(OD)-based system instead of leg based. Today ’Yield management’, now named Revenue Management, is applied in almost every airline and in other industries like hotels and car rentals [81].

2.1.2. Revenue management concepts

RM aims to sell the right product to the right customer at the right price. To achieve this, RM is divided in two main parts: pricing and inventory control. Pricing is the practice of deciding on fare levels and the different restrictions and service amenities that belong to the different fare levels. Inventory control is the next step and is the subsequent process of determining how many seats to make available at each fare level. Both steps are essential if the goal is to sell the right product to the right customer at the right time for the right price. A schematic overview of RM is given in Figure 2.1. The overview shows the process of determining the ticket price. In the flowchart, ① refers to pricing, explained in Chapter 2.1.2.1 and ② refers to inventory control, explained in Chapter 2.1.2.2.

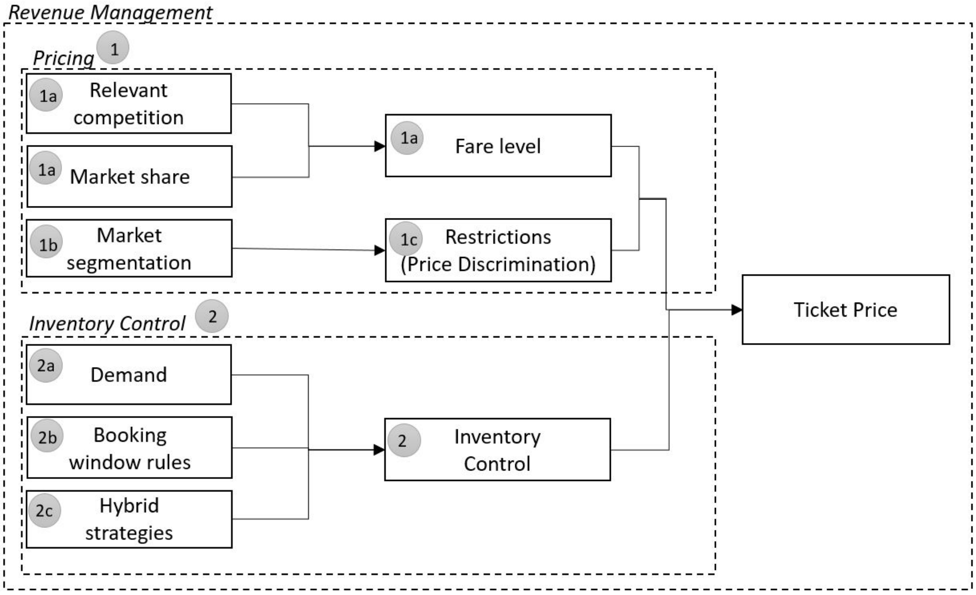


Figure 2.1: Overview of revenue management

2.1.2.1 Pricing

A fare level needs to be determined based on the relevant competition and the market share 1a. Next, the service amenities and restrictions are determined for each fare level. This can be done by **Price Discrimination** 1c. Price discrimination refers to the difference paid by customers for a ticket where the difference paid is not justified by the difference in costs. The discrimination is based on different consumers' "willingness to pay" for the specific ticket. Botimer and Belobaba, were the first to research price discrimination considering the structure of airline fare products offered in an OD market [23].

Another option is to offer different products. This is done by clearly offering identifiable and different products. For example by giving passengers the opportunity to cancel or change a flight. This concept is commonly known as **Product Differentiation** (1d). Botimer and Belobaba have done research in this field, examining and identifying many fundamental relationships between consumer purchases and the differentiated fares on products. Here, the focus lays on the behavioural motivation of passengers under fare product differentiation [23].

Price discrimination and product differentiation can only be successful if an airline is able to differentiate passengers in several demand groups. This concept is known as **Market segmentation** (1b), the strategy of dividing the market in smaller groups of customers, which have the same characteristics and needs. The aim of market segmentation is to distinguish high paying travellers from low paying customers, thereby keeping the high paying travellers away from purchasing the low fares [53]. Gallego uses a simple deterministic model to examine pricing and market segmentation decisions in the airline industry [46].

Gallego and Ryzin concluded that inventory control is relatively ineffective when pricing decisions are made correctly. This is because when pricing is not optimally determined. Only when pricing is not optimal, inventory control systems can be useful [47].

2.1.2.2 Inventory control

After setting price levels combined with restrictions, an airline has to decide on the number of seats to make available for each fare level. Seat inventory control is the practice of finding the correct balance in the number of discounted bookings and full-fare bookings. The goal is not to ensure that the aircraft has a 100 percent load factor (LF), but to maximise the total passenger revenue. The LF will increase when more discounted seats are made available, while selling too many seats at a discount will result in a decrease in total revenue. By opening and closing different classes (the different fare levels combined with the restrictions), airlines control the final price of a ticket. Deciding how many seats to make available for the different fare level is based on three main points:

1. The demand: there are variations in demand on daily, weekly and annual basis, resulting in peak periods at popular times. Cheaper classes close when demand is rising. This prediction of the demand is determined by using forecast methods. Demand forecasting is one of the most crucial parts of inventory control (2a).

   Van Ryzin described forecasting systems as everything that is required to turn raw data into actionable market information including [87]:

   (a) Data sources
   (b) Information technology for collecting and storing data
   (c) Various statistical estimation models
   (d) Algorithms used to process and analyse the data
   (e) Infrastructure for deploying model outputs

   Weatherford highlighted the challenge of the dimensions of forecasting. The increasing amount of data makes it challenging to use accurate forecasting methods that are feasible in the restricted computational time [89].

2. The booking window: X days before departure, the lowest booking class will close (2b).

3. Hybrid strategies: Classes close due to specific booking trends (2c).

Seat inventory control research started with Littlewood, who controlled each flight leg independently and only took into account two classes [64]. He proposed a rule that stated that discount fare bookings should be accepted as long as the revenue value exceeded the expected revenue of future full fare bookings.

Belobaba extended the problem for multiple classes, but his model, which he called Expected Marginal Seat Revenue (EMSR), still leaned on the assumption of independent demand between the different classes. This seemed to be a problem for the industry because of the fares being almost the same for some classes and resulted in Belobaba extending his own model (EMSR) [17]. The new model, EMRSb, is able to aggregate the demand between the different classes. Today, the enhanced model of Belobaba is still widely used in the industry [19].

Brumelle and McGill introduced the use of a joint demand probability distribution, which allowed the authors to improve Belobaba's, EMSRb model by 0.5% [28].

In addition, Curry tried to couple the Belobaba model with mathematical formulas that incorporated an OD system. Curry reduces the influence of the assumption that no network effect is considered. Nevertheless, his model cannot be considered as a full OD system as the inventory is considered not shared between the OD pairs [35].

Williamson was the first to research the network inventory controlling process taking into account the interaction of flight legs and the flow of traffic across a network. In her paper, she also introduced the bid price and defined it as: the marginal value of the network, understood as the value of the last seat available on a given flight leg. The bid price is a value, which an airline can use when deciding to accept an OD request for a specific itinerary. The revenue could be improved by using her model with an average of two percent [93].

RM is practised in all mayor and most of the smaller airlines. In this literature study when referring about RM, RM systems of FSC (American Airlines, British Airways, Qantas, KLM,...) are addressed and not the LCC (Ryanair, EasyJet, Southwest Airlines,...). LCC RM systems have different aims and are based on different assumptions. LCC RM systems do not work with sub-classes and are not segmenting customers, because they are segmenting on the type of flight: business or leisure.

Talluri and Ryzin highlight the complexity of the airline's RM problem. A major airline deals with hundreds of flights each day with around 20 fare classes and the flights can take place between hundreds of different OD pairs. The size of the RM problem makes it impossible for a human to make those decisions. A computer system can take into account a combination of mathematical models and thus provide the best prices in an automated way [81].

Nevertheless, mathematical models still do not take into account, special events, the dynamics of competition prices and the dependency of demand between flights. It is therefore essential to incorporate human market experience into the models.

As the airline industry is a highly competitive industry, airlines are not willing to share their forecasts, pricing and inventory control models. Those models are the basis of RM, which is an airline's only significant revenue stream. Hence, if airlines would make their models public, other airlines could profit from that knowledge and thereby dampen the profitability of the first airline company. As a result, airlines perform their own research. Talluri and Ryzin mention the proprietary nature of airline RM research. They highlight the fact that there are only a few published reports that document the performance of the different RM methods, because key details are deleted [81].

2.1.3. Revenue trends

The practice of RM in the airline industry is very mature but at the same time the airline market has changed to a highly competitive market where airlines need to be innovative to get more market share and higher profits.

Therefore, new trends are constantly developing, the most significant and comparable one to the flight pass being the frequent flyer program. But there are also other innovative trends like auctions, corporate deals and unlimited flying.

2.1.3.1 Frequent flyer program

A frequent-flyer program (FFP) is a loyalty program offered by airlines. Loyalty programs are a marketing strategy based on offering an incentive with the aim of securing customer loyalty [48].

Gwinner described the frequent flyer program as a reward from either, functional, relational, sociological, economical, informational or hedonist nature. The main assumptions in frequent flyer programs is that loyal customers are more profitable [52].

The idea behind a frequent-flyer loyalty program is simple. The loyalty of the members of the frequent flyer program is rewarded by receiving credits generally in the form of miles, points, or segments after they have flown with a specific carrier or its partners. The class and distance travelled determines the amount of miles that are awarded. Recently, members of frequent flyer programs do not only receive miles by flying. They can also obtain miles by other products and services. Next to flying, the customer can earn miles by spending money on car rental, overnight stays, banks or retail stores. In addition, miles can be changed for upgrades, free travel, other services and goods. Members of frequent flyer programs also have other benefits:

- Priority boarding, check-in and baggage handling
- Lounge access
- Additional checked luggage
- Seat preference
- Free upgrades
- Guarantee seating of flights that are overbooked.

Hess demonstrated that customers within the higher business segment are willing to pay 125 dollar more for a ticket that has an elite frequent flyer account [78]. The threshold for members of a frequent flyer program to switch to a competitive airline is higher because this results in the loss of points and miles [48]. Butcher estimated that frequent flyer programs of Air France-KLM, BA and Lufthansa are worth around 20 to 30 percent of their total revenues [29].

2.1.3.2 Corporate deals

Large companies, with employees that often travel, generally negotiate with one or more airlines to receive significant discounts or other incentives. This segment is called managed business travel [33]. Granados claims that the majority of the passengers are corporate passengers accounting for on average 55% of the airline passengers world wide [51].

However, there are no robust decision support systems for determining corporate deals. This is because in most airlines, corporate deals are part of the sales department where a manager is assigned to different accounts. The manager's task is to negotiate contracts, to maintain the relationship and to track the performance. The negotiations are about discounts or incentives the airline offers for the combination of route, cabin, and class that the corporate will be flying. The account managers do not have the right support systems that can strengthen the strategies in these negotiations.

Pachon, Erkoc and Iakovou were the first researchers to develop a model that designs optimal corporate contracts. Their results showed that an airline can significantly increase revenue if they would use their non-linear programming tool that determines the discounts for routes, cabin and classes. The profit of the airline is modelled using a multi-nominal logit function [75].

2.1.3.3 Unlimited flying

Unlimited flying is offered by a subscription based airline, which offers unlimited flights for a (monthly) fee. One could refer to it as to the 'Netflix' style of flying. This concept differs from the flight pass, which is a multi-flight ticket.

Flying unlimited reminds people of the 'huge disaster' of AA back in 1981. For only $250,000 per pass and for an additional $150,000, a passenger and 'its companion', could step on any flight from anywhere and anytime and that by using first class only. It is no surprise that customers took their advantage and flew extravagantly, resulting in insurmountable costs. Some users were costing AA more than a million a year. This was the first experiment using a unlimited flying approach and proved that unlimited flying could not work properly in this setting. [11]. Lately, however, some new experiments are conducted in this field, giving rise to new forms of unlimited flying, which can be divided into 3 main categories:

- **New airlines, with their own fleet, offering unlimited flights for a (monthly) subscription fee.**
  - In 2013, SurfAir started to offer private flights between business destinations in the US for a monthly fee [9].
  - In 2014, La Compagnie, a French start-up started to fly only between Charles de Gaulle and Newark Liberty airports. Consumers can buy a “L'Unlimited” annual passes and hop on the flights whenever they want [7].
  - In 2015, the start-up Rise followed the same business model as SurfAir for different business destinations in the US [4].
  - In July 2017, Surf Air started operations in Europe, Surf Air in Europe operates under the same subscription model and started connecting London-Zurich in weekdays and London-Ibiza in weekends [9].
- **Existing airlines experimenting with unlimited flying programs next to their normal ticket selling.**
  - In 2009, jet blue offered All-You-Can-Jet fare for a fixed fee, the customer could travel unlimited over jet blues network. The experiment was declared unsuccessful [42].
  - In 2012, Alaska Airlines came with the Air Pass program, another "flexible" flying deal where leisure passengers were encouraged to see multiple destinations [5].
  - Wideroe is offering unlimited flight within Norway for tourists. Ticket is valid for unlimited travel in Norway for 2 weeks [10].
  - Air Canada uses a similar concept as the flight pass with the difference there is an option to buy in bulk tickets or unlimited [6].
- **A third party offering unlimited flying by offering seats on flights of already existing airlines.**
  - In February 2016, the start-up OneGo started with offering unlimited, non-stop, economy-class flights to more than 75 airports on the main airlines in the US for a monthly fee [8].

While there are a lot of ongoing experiments with unlimited flying there is not a single research paper supporting this business model.

2.1.3.4 Auction

Auction is a concept that is not frequently used in the airline industry. However research has proven that auctions could improve revenue when selling last minute tickets. Baker and Murty did research on the benefits of auction in revenue management and found that auctions for last minute tickets can increase revenues up to 16 percent. Next to attracting more revenue, auctions can also provide better insights into willingness-to-pay of customers when comparing it to classic RM models [75]. Eso also investigated the matter and discussed the auction process for the extra-capacity flights of the airlines. He proposed an integer programming model and concluded that his model could be profitable for the travel industry [39].

Some airlines are experimenting with auctions and each of them is applying a roughly similar concept. All flights have a starting price (lowest 1 dollar), which will rise as long as there is a customer willing to pay. The one who will get the ticket, will have to put the highest bid. In 2012, Royal Jordanian debuted by using auctions on its website [3] and in 2015, Iberia started with selling last minutes flights for auction [1].

Ticket sales are not the only type of auctions that airlines are testing. Since 2011, several airlines, including KLM, Lufthansa, Qantas, and Etihad, have offered travellers who have bought economy-class tickets the chance to bid for a seat in a higher class of service.

2.1.4. Conclusions and further direction

This chapter will highlight the important points of Chapter 2.1. First, airline RM has had a turbulent history which was always driven by optimising revenue. This resulted in very mature systems. At the same time the airline market changed into a highly competitive market where airlines need to be innovative to gain market share and additional profits.

As mentioned in the introduction, the prices of the flight pass are dependent on the RM ticket prices. The flight pass pricing model, has to find a way to minimise dilution and mimic the effects of price discrimination, market segmentation, product differentiation and inventory control. Gallego ad Ryzin, concluded that when pricing decisions are made correctly, inventory control is relatively ineffective [47]. This is rather interesting and valuable for pricing the flight pass. This means that the pricing of the flight pass should not be dependent on the inventory of the seats.

Interesting trends discussed are:

- Frequent flyer: The FFP has the same main goal as the flight pass, namely: Increase market share by increasing customer fidelity. Research demonstrated that the threshold for members of a frequent flyer program to switch to a competitive airline is higher [48]. This result is of great value for the flight pass since passengers will not switch to a competitor when the incentives are sufficiently high. There is however a challenge concerning the FFP, because the flight pass cannot undermine the FFP and should create a new market segment.

- Corporate: The flight pass model could be used as a basis for cooperate deals since there is a lack of decision support systems.

- Unlimited flying: Unlimited flying is recent trend and the concept is very similar to the flight pass but not the same. The flight pass is a multi-flight ticket whereas unlimited flying is subscription based flying based on a (monthly) fee. There is no research on unlimited flying so far. With a significant amount of airlines experimenting with unlimited flying or the flight pass, more research is necessary in order to determine a sound methodology to define the ideal settings for a flight pass price.

Since the flight pass is a rather new concept, there is not enough booking data from the flight pass available to define the settings for an ideal flight pass price. There is, however, historical booking data from RM available. This data includes the effects of price discrimination, market segmentation, product differentiation and inventory control in order to avoid dilution. The booking data includes additional information as: month of flying, day of week of flying, time of day of flying, length of stay, ticketing lead time, cabin, point of sale,... However, this information is not directly used to predict a ticket price in RM. As explained in Chapter 2.1.2, ticket prices are determined by different systems and historical booking data is not used to find trends to directly determine the ticket price. Only the sub system, forecasting is using booking data to predict the demand. Most forecast methods predict demand using only 1 parameter, namely the amount of bookings. Nevertheless, additional booking information is available but still unused.

Hence, it would be interesting to combine the booking data from RM and additional booking information (month of flying, day of week of flying, time of day of flying, length of stay, ticketing lead time, cabin, point of sale, ...) to better predict ticket prices. This is often done by using Machine learning techniques.

Machine learning is one of the fastest growing areas of computer science, with far-reaching applications. Machine learning detects patterns in data that are meaningful. It is a powerful tool which combines statistics with computer science. The computers and its computational power can handle the huge amount of data which a statistician cannot. To understand the benefits and disadvantages and the usability for the flight pass problem, machine learning is studied in Chapter 2.2.

2.2. Predictive models using machine learning techniques

Machine learning is everywhere: Anti- spam programs can learn how to filter emails, search engines like Google know how to show the most relevant results and digital cameras are able to learn how to detect faces. Using machine learning , patterns can be found in complex data-sets in a way that a human would not be able to do due to complexity or time constraints. Machine learning models have been around for a while but the reason that they gains more attention these days is because of the recent development in storing data. Over the past years, obtaining data and storing big data sets has become possible, allowing people to use machine learning techniques to analyse these data-sets faster and more precise. Before diving into the research papers, some machine learning definitions need to be explained [80].

- The set of attributes that predicts the price (month of flying, day of week of flying, time of day of flying, length of stay, ticketing lead time, cabin, point of sale,...), is called **features** in machine learning terms.

- **Labelled data** is defined as values or categories assigned to examples, in this research the price of the flight pass.

- **Supervised learning** is the process of using the labelled data to classify the unlabelled data. A supervised learning algorithm generally avoids over-fitting of the data.

- Imagine, there is only unlabelled data. Supervised learning is not possible and **unsupervised learning** should be applied. The data is clustered where there are similarities. The algorithm finds hidden structures in unlabelled data.

- **Semi-supervised learning** is using both labelled and unlabelled data to categorise. Semi-supervised learning tries to obtain better accuracy's using both labelled and unlabelled data.

- **Classification** is the process of splitting data into separate classes. Classification can be seen as identifying a group membership.

- **Regression** is the process where the output variables take continuous values. Regression estimates or predicts a response.

machine learning techniques should be placed in context. **Data mining** is the process to extract and transform information from a data set to a clear and understandable structure for further use. This transformation can be done using **machine learning** techniques. Machine learning techniques *predict* outcomes. **Statistical methods** for **pattern recognition** are the basis for the further elaborated machine learning techniques. **Artificial intelligence** is a broader concept and is defined as anything that concerns intelligence in computers. Machine learning is an application of artificial intelligence. **Big data** is a large data set. Big data sets are outgrowing the simple databases.

It can be concluded that predictive analysis contains a variety of techniques from statistics, pattern recognition, machine learning and data mining that analyse big data to make predictions about the unknown.

The pricing of the flight pass is a **Supervised regression algorithm** so this is the focus in this chapter. Demand forecasting models, described in Chapter 2.1.2.2, are already using machine learning algorithms to predict demand using booking data, therefore it would be interesting to study this in depth since this already happens in the airline industry. The research is described in Chapter 2.2.1

2.2.1. Prediction models in demand forecasting

Airline demand forecasting is a different problem than the flight pass pricing. It predicts demand by using only 1 parameter namely, the amount of bookings. Nevertheless forecasting models are predictive models used throughout the airline industry and there are many techniques and methods that address the problem of forecasting, such as machine learning , which is the topic of this chapter.

Some basics are explained, first. Demand forecasting can be divided in three categories that describe methods based on a data set [92] [63] [90]:

- **The historical booking model** uses historical data (data from already departed flights). The historical booking model solves the demand forecasting problem by using a time series model.
- **The advanced booking model** uses reservation data and the concept of 'Pick-Up'. Pick-up is the increment from now to a day T of N bookings in the future. At day T, there are already K reservations. This results in an occupation forecast of K + N for day T.
- **The combination of both**

2.2.1.1 Airline demand forecasting

Machine learning techniques are quite new in forecasting demand in the airline industry. Most researchers do agree on the fact that the pick-up method is (at this moment) the most recognised way to forecast demand [63] [91] [82]. The pick-up method identifies an increase of bookings in different periods and then accumulates it into a total demand figure that is expected in the future. Since the pick-up is not a machine learning technique, the focus is on other research papers that describe the use of machine learning models.

For instance, in the research of L. R. Weatherford, Gentry, and Wilamowski, the use of a neural network (NN) in airline context is described. A neural network is able to recognise patterns by learning from input data which is called training.

neural network got its inspiration from the human brain. It is a computational method that imitates the cognitive ability of the brain with artificial neurons to solve complex problems. The features are the input variables. In neural network terms this is called, an input layer. There is also an output layer which presents the output. The hidden layer allows for the combination of input data in an almost infinite number of ways [49]. These manners are all variations on the neural network, although the general architecture can be seen in Figure 2.2.

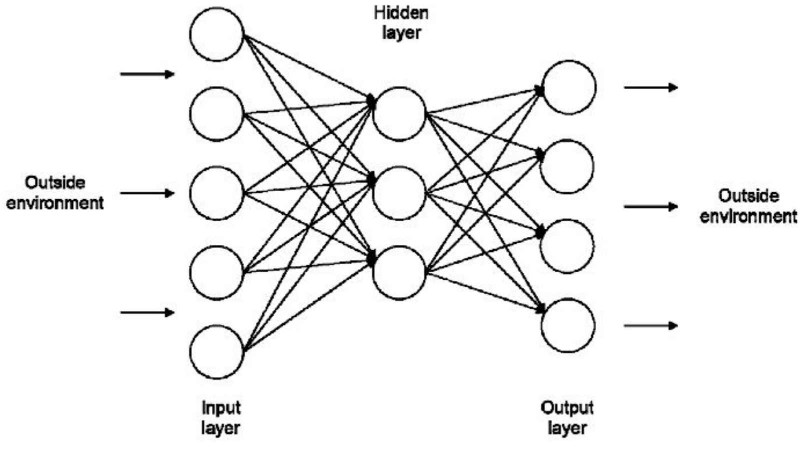


Figure 2.2: The illustration of the neural network functional structure [68]

L. R. Weatherford, Gentry, and Wilamowski use two types of neural network methods in their research:

- The single-layer feed-forward network: The network has one input layer, with 8 different inputs for 8 different historical points, and one output layer. When there is no layer between the input and output layer, only linear combinations can be produced. Figure 2.3 shows the single-layer feed-forward network.
- A functional link network: The functional links produce non-linear combinations of input. Figure 2.4 shows the functional link network.

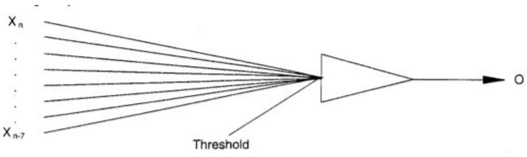


Figure 2.3: A single-layer feed-forward network [88]

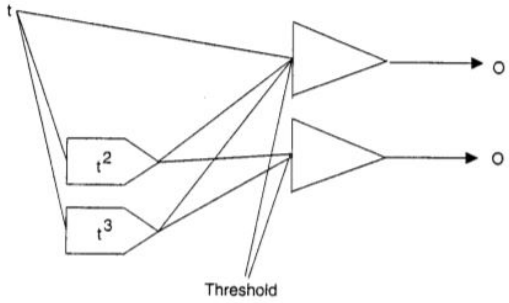


Figure 2.4: A functional link network [88]

The researchers compared the neural network approach to more traditional methods (Simple moving averages, weighted moving averages, cubic regression, simple exponential smoothing,...) used in the airline industry [88]. The single-layer network outperformed the functional link network. For a short-term forecast the single-layer network performed better than the traditional forecasting methods. For the long-term forecast the cubic regression and single-layer network performed similarly and better than the other methods.

L. R. Weatherford, Gentry, and Wilamowski conclude that a neural network is a promising area of research. Until now, the neural network approach has not yet been proposed in other papers that concern bookings forecasting for airline revenue management. The reason for this is that:

- A neural network needs more data storage and computational time [88].
- The determination of outliers, the selection of training data and establishing confidence intervals for the forecast can be difficult when using a neural network [13].

2.2.1.2 Hotel demand forecasting

Hotel demand forecasting operates under the same principle as airline demand forecasting. Caicedo-Torres, compared several machine learning techniques for demand forecasting in the hospitality industry. The author compared Ridge Regression, Kernel Ridge Regression, multilayer Perceptron neural network and Radial Basis Function neural network with the use of three different data-sets: occupation time series data, occupation times series data plus additional variables, and reservations data. The Ridge regression model with quadratic features trained on the reservations data set, outperformed the other models [30].

2.2.1.3 Tourist demand forecasting

In 1999, the researchers, Law and Au, introduced a new approach for forecasting tourism demand which was at the time largely dominated by time series, regression techniques and pickup methods. They proposed a supervised feed-forward neural network model to forecast tourism demand and in particular Japanese tourist arrivals in Hong Kong. The results showed that the neural network model outperformed the multiple regression-, moving average-, and exponent smoothing models [62].

Laws neural feed forward network is composed of three distinct layers, an input layer, one or more hidden layers and an output layer. Each of these layers contains nodes, and they are connected to nodes on the adjacent layer. For them, each node of a neural network is a processing unit that contains a weight and a sum function. A weight returns a mathematical value for the relative strength of connections to transfer data from one layer to another layer, whereas a sum function computes the weighted sum of the input elements entering a processing unit. The nodes in the input layer represent independent problem variables, the hidden layer is used to add an internal representation of handling non-linear data and the output of a neural network is the solution to a problem [62].

In 2000, Law introduced a back propagation learning method to improve the accuracy of the neural network when forecasting tourism demand. He concluded that using a back propagation algorithm to train neural network outperforms regression models and time-series models in terms of forecasting accuracy [61].

In 2016, Constantino, Fernandes and Teixeira, modelled the tourism demand for Mozambique by using the same model as Law and Au, where they concluded that the model was performing as expected with a high forecast accuracy [34].

Research in forecasting methods using machine learning techniques is limited to the use of a neural network. Nevertheless researchers are stating that neural networks are a promising area of research.

In chapter 2.2.2, a more similar problem to price the flight pass is researched. Predicting prices in the real estate industry, where there is a new product (specific house) that needs to be valued. The analogy with the flight pass is illustrated in Figure 2.5. Real estate prices depend on parameters such as the number of bedrooms, living area, location, the number of bathrooms, the presence of a garden, the presence of a garage... The flight pass price depends on: month of flying, day of week of flying, time of day of flying, length of stay, ticketing lead time, cabin, point of sale,...

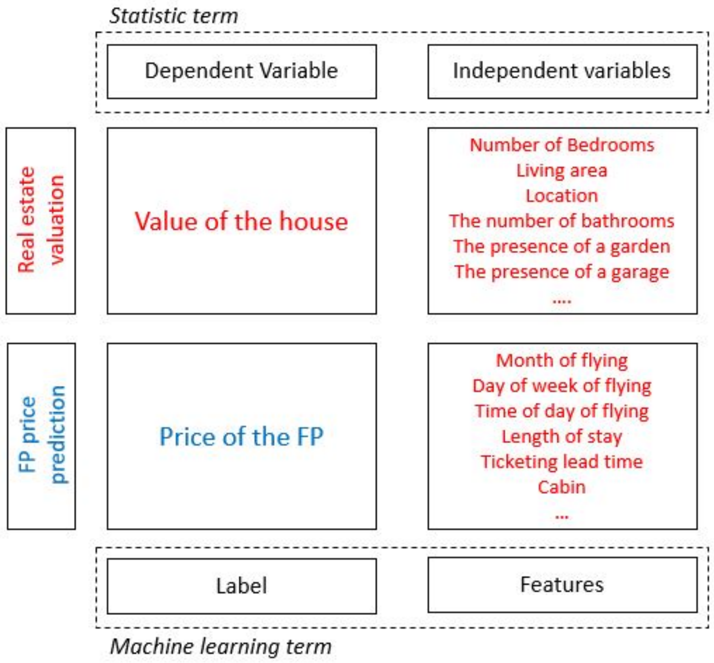


Figure 2.5: Analogy of real estate valuation and the flight pass pricing

2.2.2. Valuation in real estate industry

Traditionally, the sales comparison approach has been used to justify the value of a house. This method compares a piece of property to other properties with similar characteristics were sold recently. The sales comparison approach looks at local properties to determine what they have in common. This allows appraisers to determine values for property features, such as fireplaces or two car garages, and requires less sophisticated statistical methods than checking on sales of properties in a wider geographic area [94].

The downsides of the comparison model is that it compares only one or a few properties that are similar to the the property that needs to be valuated, it is subjective and there is no methodological structure [70].

With the significant increase in the amount of data being stored in databases, accelerated by the success of the relational model for storing data and the development of data retrieval and manipulation technologies, machine learning techniques arose when determining the value of houses.

2.2.2.1 Multiple linear regression

Multiple linear regression (MLR) is not only comparing a few properties but it takes into account all the properties, which is an improvement compared to the sales comparison approach [20].

In 2001, Lusht suggested that multiple linear regression algorithm can be used to quickly value a large amount of houses. This is why the multiple linear regression grows in popularity with tax assessors [66].

In addition, in 2011, Zurada, Levitan and Guan described multiple linear regression as an established method, used in many applications and the method has a large acceptance degree by academics [97].

In 2013, Božić furthermore concluded that multiple linear regression offers a reliable tool to get accurate value for any property [24].

Next to the advantages compared to the sales comparison approach, some researchers highlight the risks of the multiple linear regression. Benjamin, Guttery and Sirmans warn for human judgement on the fit of the relationship between the variables, the linearity and the normality of the inputs [20].

Kang describes the risk of multicollinearity in multiple linear regression. This is the phenomenon where independent variables in a multiple linear regression model are correlated to each other, resulting in misleading outcomes [57].

Multiple linear regression is a technique where a relationship between one of the input variables and the output value can be seen as a straight line. Multiple linear regression has more than one input variable, therefore the straight line can be thought of as a (hyper)plane.

The different input values are used to predict the output value. Each of the input features (x_i), where i is the amount of independent/predictor features, is weighted using a coefficient (β). The goal of the learning algorithm is to obtain a combination of coefficients (β) that results in optimal predictions of (Y) [14]. The Mathematical Formulation is given in Equation 2.1.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_i + \epsilon \tag{2.1}$$

2.2.2.2 Neural networks

In 1992, Do and Grudnitski performed one of the first researches on the valuation of real estate modelled by a neural network. Do and Grudnitski used one hidden layer perceptron model trained by back-propagation. Their motivation was that neural network techniques do not have the same shortcomings that most of the other valuation techniques such as the sales comparison approach and multiple linear regression have [36]. The authors compared multiple linear regression and a neural network using the data of individual houses in San Diego. They concluded that the neural network performed better than the multiple linear regression algorithm. The working mechanism of the neural network is explained in Chapter 2.2.1.1.

In 1995, Borst agreed with Do and Grudnitski and shared the opinion that a neural network is the logical follow-up of the multiple linear regression [22]. Borst achieved similar results compared to Do and Grudnitski when comparing multiple linear regression and a neural network.

Tay achieved the same results when using a data set of residential apartment properties in Singapore to test the predictive performance of a neural network and the multiple linear regression. Tay argued that real estate valuation is especially a problem of “pattern recognition” where a neural network is a tool that can learn from historical sales [85].

In 2002, Kauko, Hooimeijer, and Hakfoort examined a neural network with an application to the housing market in Helsinki, Finland. They concluded that various dimensions of housing sub-market formation could be identified by uncovering patterns in the data set [58].

In 2011, Kontrimas and Verikas compared multiple linear regression, Support vector regression and multilayer perceptron neural network. They concluded that a non-linear model is required in the appraisal of real estate [60].

However, neural networks are not always described as successful or favourable. Worzala tested the previous studies with similar data and her results were not promising [96]. Allen illustrates that neural networks are not necessarily superior. The author highlights that small changes in the operational parameters of the same neural network can result in very different output findings, when the model is presented with new information [12].

Mora-Esperanza discussed in his research, artificial intelligence applied to real estate valuation and the advantages and disadvantage of the neural network in real estate valuation. The main advantages, compared to previous models, are higher precision and increased capacity to estimate the value of special properties. However, the main disadvantage of the neural network is that there is no way to explain how the model predicts the prices of the houses, also known as the *black box* fallacy. The complexity of multiple neuron connections and the iterative weighting correction process makes it quite impossible to explain the working. Therefore, a neural network must be used very carefully for the valuation of real estate, because different model settings can generate opposite results and the accuracy of results can depend on the specifics of the data set [74].

It is obvious, therefore, that multiple linear regression compared to a neural network shows some advantages against each other depending on the quality, amount of the data and the correlations between the variables.

2.2.2.3 k-Nearest neighbour

In 1999, McCluskey and Anand introduced the use of k-nearest neighbours. They concluded that the application of the k-nearest neighbours is more transparent than the use of a neural network and that the k-nearest neighbours facilitates the retrieval of comparables on which the house price is based [70].

In 2014, Pow compared regression methods to predict real estate property prices in Montreal. He compared multiple linear regression, support vector regression, k-nearest neighbours, classification- and regression trees (CART) and Random Forest Regression (RFR). The algorithms for CART and random forest regression are explained in Chapter 2.2.2.4 and Chapter 2.2.2.4. Pow concluded that random forest regression and k-nearest neighbours performed significantly better than the support vector regression and multiple linear regression. The author argued that the reason for this are the nonlinear interactions between the features and the real estate price [77].

In 2017, Borde compared -like Pow- various methods to predict the real estate prices for Mumbai. Borde compared multiple linear regression, linear regression using gradient descent, k-nearest neighbours and random forest regression. Borde concluded that the multiple linear regression outperformed the linear regression using gradient descent but the k-nearest neighbours and the random forest regression have significantly better results than the linear variants, with random forest regression showing slightly better results [21].

Moosavi predicted the real estate rental and sales prices in Switzerland. The author compared random forest regression, k-nearest neighbours, Bayesian regularised regression, multiple linear regression, and local regression. Moosavi concluded the random forest regression and the k-nearest neighbours out-preformed the other models, with random forest regression showing slightly better results [73].

k-nearest neighbours is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure, which is the distance equation shown in Equation 2.2. The first researchers to use k-nearest neighbours were Fix and Hodges in 1951 for the Air-force of the united states [44]. The equation calculates the distance between x and w where m is the dimensionalty of the feature space.

$$d(x,w) = \sqrt{\sum_{i=1}^{m}(x_i - w_i)^2} \tag{2.2}$$

2.2.2.4 Decision tree, bootstrap aggregation and random Forest

In 2006, Fan used a tree-based, CART, approach to predict the house prices in the Singapore resale public housing market. CART is explained in Chapter 2.2.2.4. CART is an important statistical pattern recognition tool that can examine the relationship between house prices and housing characteristics. Fan concludes that CART can provide an effective approach to identify the most important determinants of public housing resale prices. [41].

Antipov and Pokryshevskaya are the first and the only researches who compared multiple linear regression, radial basis function neural network, multilayer perceptron neural network, k-nearest neighbours, CART, bootstrap aggregation algorithm, explained in Chapter 2.2.2.4, and random forest regression. They predicted the valuation of residential apartments in Saint-Petersburg and concluded that random forest regression, bootstrap aggregation algorithm and k-nearest neighbours were outperforming the other methods. With the bootstrap aggregation algorithm performing slightly better than k-nearest neighbours and random forest regression performing the best of all the methods. Antipov and Pokryshevskaya discussed the reason of the random forest regression being the most promising method for the valuation of housing prices. Random forest regression is stable concerning outliers, random forest regression can work properly with missing values and random forest regression can handle categorical variables with many levels [15].

As mentioned in Chapter 2.2.2.3, Moosavi, Paw and Borde concluded that the random forest regression performs slightly better than any of the other used methods [73] [21] [77].

Classification and regression decision tree

Breiman introduced in 1984, the CART learning algorithm, which is capable of performing regression as well as classification [27]. In a CART, an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. A regression tree is built by recursively partitioning the sample (= the "root node") into more and more homogeneous groups, so-called nodes, down to the "terminal nodes." The partitioning is called a split. The goal is to choose a split among all the possible splits at each node so that the resulting new nodes (child nodes) are the "purest". Each split is based on the values of one variable and is selected according to a splitting criterion. The model takes into account all possible splits for each feature and choose the feature and the split that minimises the MSE (regression).

Bootstrap aggregation

The bootstrap aggregation algorithm uses different decision trees on different subsets of the data. To create some variability between the different decision trees and to reduce the risk of over-fitting, the subsets are drawn at random from the full training set. The different decision trees are combined and averaged [25].

Random forest

In 2001, Breiman added a second parameter next to the number of trees, which is unique for the random forest algorithm [26]. The second parameter captures how many features to search over to find the best feature. During tree creation, a random number of features is chosen from all available features and the best feature that splits the data is chosen.

In short, the random forest algorithm followed up the CART and the bootstrap aggregation algorithm algorithm. A very complex decision tree that grows very deep, learns irregular patters, which means the tree is over fitting the training set. The goal of the random forest regression is to avoid the over fitting of one decision tree model by training trees on different subsets of the data and by randomly selecting the amount of features to consider in a split.

2.2.3. Conclusions and chosen direction

This chapter focused on machine learning techniques. Firstly the demand forecasting has been studied in the airline, hotel and tourism industry. It is safe to conclude that the machine learning approaches are quite limited to the usage of a neural network. Researchers are concluding that neural networks are a promising area of research.

Secondly, the valuation of the real estate industry is discussed. Machine learning is widely used in a significant amount of researches. Multiple linear regression, is considered a reliable tool and established method. When comparing multiple linear regression with non-linear machine learning techniques, the non-linear techniques are outperforming multiple linear regression, which is because of the non linear behaviour between the features and the housing price.

The difficulty in comparing these studies is that they use different data, and their results may be data dependent. Nevertheless, most researchers agree that a neural network outperforms the linear regression techniques.

The main disadvantage of the neural network, the black-box, is a significant disadvantage to consider for the pricing of the flight pass. The model should be used in the airline industry, where there is insufficient knowledge of neural networks and methods which are less a black-box are more favourable to use.

Some researchers use a random forest regression and in the papers studied, random forest regression outperforms all the other machine learning techniques including the neural network [15] [73] [21] [77]. This combined with the fact that, in general, random forest regression is the best performing machine learning technique in high dimensions [32] [31], makes it a suitable method to predict housing prices.

The relationship of the features that predicts the flight pass price is not known to be linear or non-linear. Therefore, in this research a linear machine learning algorithm is used first, namely multiple linear regression and compare it with the two most promising non-linear machine learning algorithms. It can be concluded that random forest regression and neural network are suitable methods to predict houses. And that the use of the neural network is promising in the demand prediction for airlines. Therefore, random forest regression and a neural network are used in this research and compared to multiple linear regression.

3 Research Plan

By reviewing the literature, knowledge gaps within the industry were identified which should be clarified. In order to solve these, a structured research plan should be in place. This Chapter will describe all aspects of this research plan. First, in Chapter 3.1 the problem statement is summarised . The research goal and objective that will be achieved during the thesis are described in Chapter 3.2. To achieve the objective, it is important to define a scope which sets the constraints in which this research will be done. The Scope is discussed in Chapter 3.3. Finally to conclude, the research impact on both academic and industry level are explained in Chapter 3.4.

3.1. Problem statement

The problem is directly derived from a challenge in the industry. Airlines encounter challenges with respect to the satisfaction and fidelity of customers. Therefore a new concept, called the flight pass, is brought to life with the goal to increase both satisfaction and fidelity of customers. Besides that, industry has also the urging request to support the prices for the pass with data driven methods.

A model should be designed to fill this research gap. The requirement from the industry is to avoid dilution. However, there is no useful data available as the flight pass is a new concept. Nevertheless, there is historical booking data from RM available. This data includes the effects of price discrimination, market segmentation, product differentiation and inventory control in order to avoid dilution. The booking data includes additional information such as: month of flying, day of week of flying, time of day of flying, length of stay, ticketing lead time, cabin, point of sale,... Hence, it would be interesting to combine the booking data from RM and additional booking information (month of flying, day of week of flying, time of day of flying, length of stay, ticketing lead time, cabin, point of sale, ...) to predict the value of a ticket.

3.2. Research goal and objective

Based on the literature study, the following research question has been established.

How to price the flight pass prices depending on (categorical) features with data-driven machine learning techniques?

In order to answer this research question, the research question is divided into four sub-questions. These sub-questions collectively contribute to the research question and split the research into the four phases; every successive phase requires the expertise achieved in the previous phase.

- ***Sub question 1:*** *What features to select to predict the price of the flight pass?*

  The features that predict the price should be **selected**. The importance of the features is a result of the model. This is a feedback system in which, after analysing the importance, features can be replaced or eliminated. The features are limited to the available information linked to a booking. Features which can be used as input are the following:

- Month of departure
- Load factor (LF)
- Day of week (DOW) of departure
- Time of day (TOD) of departure
- Ticketing lead time (TLT)
- Length of stay (LOS)
- Sunday Stay (SUN)
- Point of Sale (POS)
- Return flight (RET)
- ...

After selection, it should be tested how the model performs if the options of a feature are **grouped** or categorised. Results could be more meaningful if for example months are grouped into seasons or if TLT is divided in categories instead of a continuous variable.

The features are selected and grouped based on the best result of the models. The best result is obtained when dilution is avoided.

- ***Sub question 2:*** *Which model to use to predict the flight pass prices?*

  The flight pass prices are predicted by three different models. These models are subsequently compared. The general characteristics of the predictive models are addressed. The random forest regression, neural network and multiple linear regression are compared in terms of accuracy according to the selected metrics:

  - The adjusted coefficient of determination (R^2)
  - The root mean squared error (RMSE)
  - The mean absolute error (MAE)

  After comparison, the model that performs best, based on accuracy, is selected to predict the prices for every individual route. Keep in mind that this might mean that all models are used because the performance is evaluated for each individual route separately.

- ***Sub question 3:*** *How to cluster the different routes depending on the importance of the features?*

  The price prediction and importance of the features are different for each route. This makes the results quite complicated since an airline has more routes that could make use of the flight pass. Therefore to simplify the analysis of the results, the different routes could be clustered based on the feature importance of each route.

- ***Sub question 4:*** *How to simulate the use of the flight pass?*

  The result of the described models is the prediction of the price of an individual flight. The next step is to predict the price of a flight pass. Since a customer can book different flights within one pass, the total flight pass price based on individual flight prices has to be estimated. This estimation should be done by simulating the flying behaviour of a passenger.

This can be summarised as the objective:

To develop a model that ***predicts*** *an individual flight price depending on (categorical)* ***features*** *by minimising the difference with revenue management prices using machine learning techniques and subsequently* ***simulates*** *the use of the flight pass with a Monte-Carlo simulation in order to determine the suitable flight pass prices.*

3.3. Research scope

The research objective stated in Chapter 3.2 can be summarised in a research framework, which can be seen in Figure 3.1.

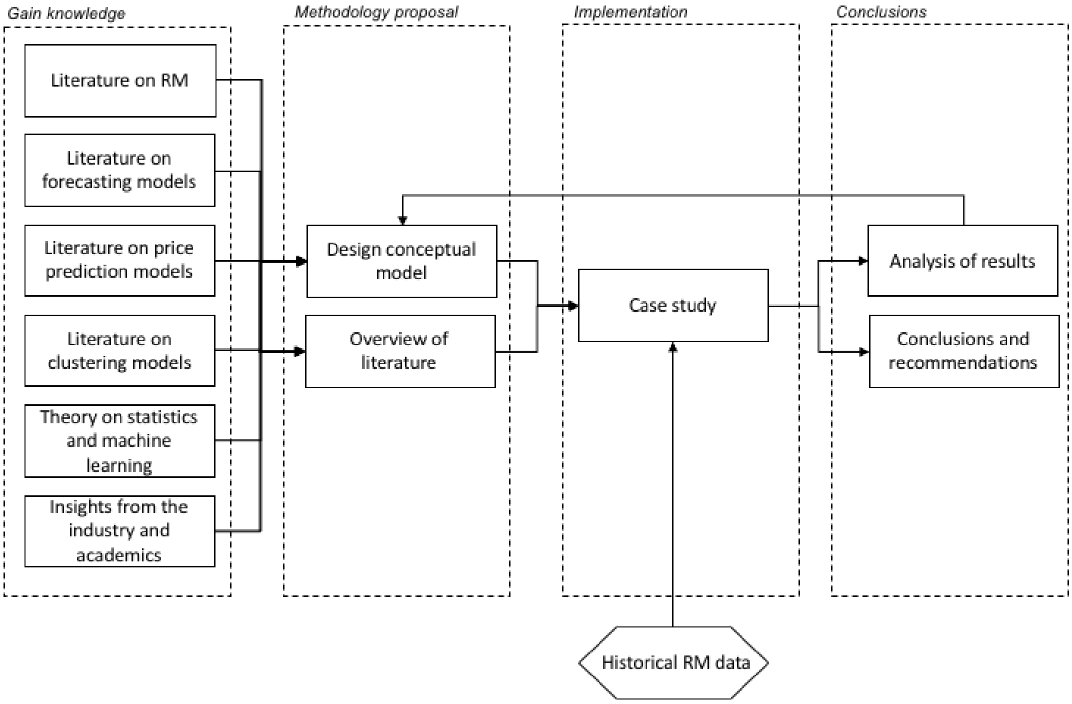
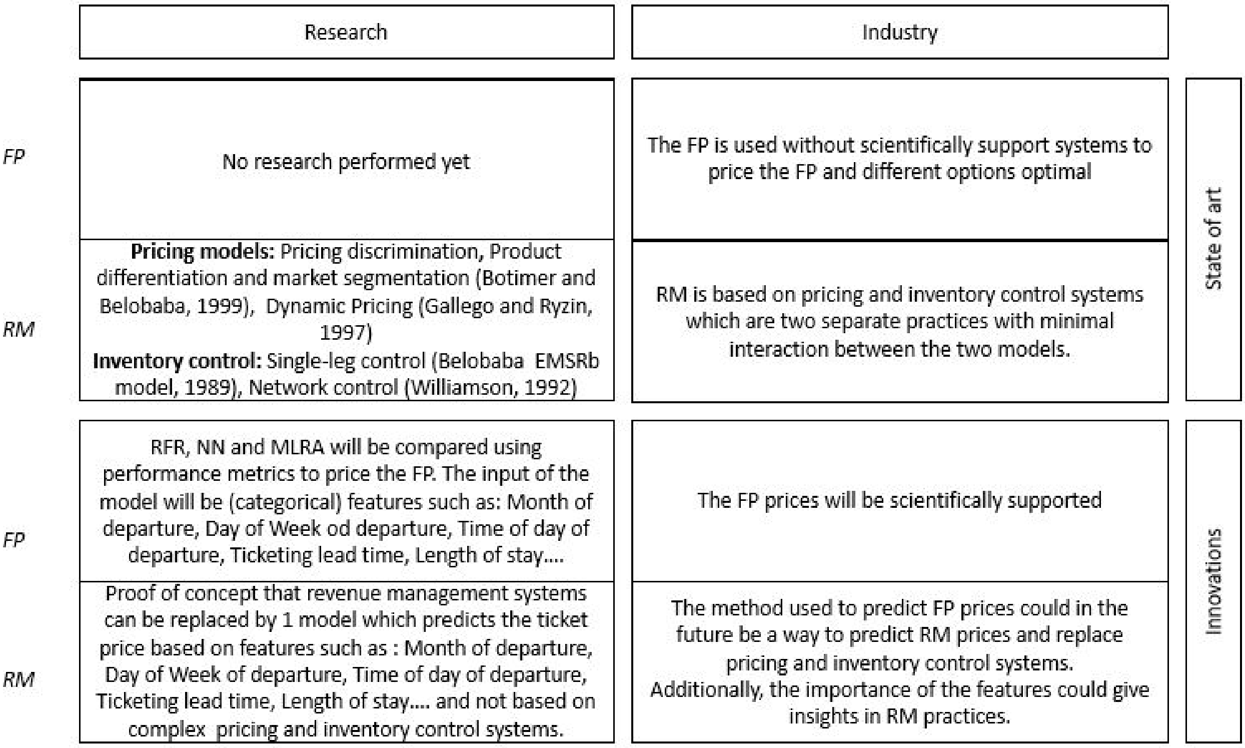

Gain knowledge
Methodology proposal
Implementation
Conclusions
Literature on RM
Literature on forecasting models
Literature on price prediction models
Literature on clustering models
Theory on statistics and machine learning
Insights from the industry and academics
Design conceptual model
Overview of literature
Case study
Analysis of results
Conclusions and recommendations
Historical RM data

Figure 3.1: Research framework

To achieve the objective, it is important to define a scope that limits the research. This research is limited to first gain knowledge about RM, forecasting models, price prediction models and clustering models. Then this knowledge is translated into a conceptual model. Subsequently the model is verified with a case study and then the results are analysed and finally, conclusions are drawn.

3.4. Research impact

Figure 3.2 gives an overview of the state of art and the innovations in the industry and research. The main innovations are the scientific support of the pricing of the flight pass. However, the model used for pricing the flight pass could also be beneficial for RM practice and replace the current systems.



	Research	Industry	
FP	No research performed yet	The FP is used without scientifically support systems to price the FP and different options optimal	State of art
RM	**Pricing models:** Pricing discrimination, Product differentiation and market segmentation (Botimer and Belobaba, 1999), Dynamic Pricing (Gallego and Ryzin, 1997) **Inventory control:** Single-leg control (Belobaba EMSRb model, 1989), Network control (Williamson, 1992)	RM is based on pricing and inventory control systems which are two separate practices with minimal interaction between the two models.	
FP	RFR, NN and MLRA will be compared using performance metrics to price the FP. The input of the model will be (categorical) features such as: Month of departure, Day of Week od departure, Time of day of departure, Ticketing lead time, Length of stay....	The FP prices will be scientifically supported	Innovations
RM	Proof of concept that revenue management systems can be replaced by 1 model which predicts the ticket price based on features such as : Month of departure, Day of Week of departure, Time of day of departure, Ticketing lead time, Length of stay.... and not based on complex pricing and inventory control systems.	The method used to predict FP prices could in the future be a way to predict RM prices and replace pricing and inventory control systems. Additionally, the importance of the features could give insights in RM practices.	

Figure 3.2: Impact and innovative elements

4 Methodology

As described in the research plan, this thesis will describe a data-driven method to determine the price of the flight pass while minimising dilution. The following chapter will describe in detail how this is done. Based on the information below, one should be able to reproduce the research assuming the data is available.

As the flight pass price will be based on a combination of multiple individual tickets prices, the methodology is split into two parts.

- First, a specific individual flight price is predicted before the total flight pass price can be estimated. A specific individual flight prediction is defined as the prediction of the revenue per passenger for one flight depending on the different features. These features used for the prediction are listed in Chapter 5.2.

  Three models are compared to predict the price of a specific individual flight. These three models: multiple linear regression, random forest regression and multilayer perceptron neural network are chosen as a result of the literature review. Since there are 46 routes, different model calibrations are used for every route. It is not efficient to analyse 46 routes, therefore a k-means cluster model is used to cluster the different routes based on feature importance in the price prediction. The prediction models are described in the following Chapters 4.2, 4.3 and 4.4. The Evaluation metrics used to compare the models are discussed in Chapter 4.5 and the k-means cluster model is described in Chapter 4.6.

- Second, when knowing all the individual flight ticket prices, the flight pass prices can be determined. As a passenger can book different specific individual flights within one flight pass, the flying behaviour needs to be simulated. By simulating the passengers behaviour a correct estimation can be made for the final flight pass price while satisfying the requirement. A Monte Carlo simulation is used for this estimation. This is described in Chapter 4.7.

The translation of the requirement to the different models is illustrated in Figure 4.1.

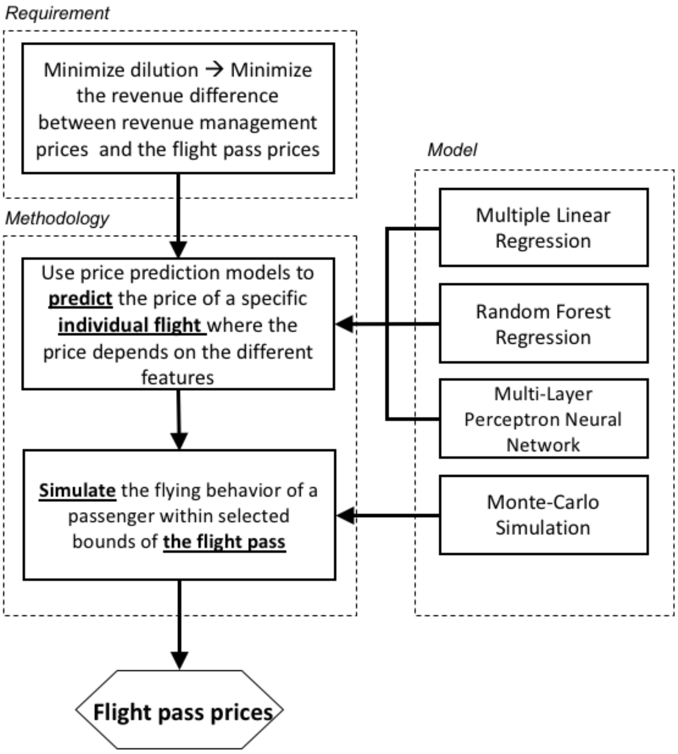

Requirement
Minimize dilution → Minimize the revenue difference between revenue management prices and the flight pass prices
Methodology
Use price prediction models to predict the price of a specific individual flight where the price depends on the different features
Simulate the flying behavior of a passenger within selected bounds of the flight pass
Model
Multiple Linear Regression
Random Forest Regression
Multi-Layer Perceptron Neural Network
Monte-Carlo Simulation
Flight pass prices

Figure 4.1: From requirements to model

The multiple linear regression, random forest regression and the multilayer perceptron neural network are machine learning methods. Machine learning methods make predictions based on data therefore some clarifications about the methodology regarding the data are explained in the following section.

4.1. Data sets and cross-validation

The multiple linear regression, random forest regression and multilayer perceptron neural network use three data sets (training, validation and test set) that are used in the different stages of the creation of the model. Training and validation data sets are used to minimise the error, which is evaluated using a test data set.

- **Training set**: Used to train the model.
- **Validation set**: Used to tune the hyper-parameters.
- **Test set**: Used to assess the performance of the model. The test data set is not used in the training of the model. The only purpose of the test data set is to evaluate the final model.

A random sample of training, validating and testing data set is used. In the data sets, there is no explicit validation set but cross-validation is used, which creates a number of temporary validation sets from the training set.

Cross-validation partitions the data set into multiple blocks used for either training or validation, as illustrated by Figure 4.2. The algorithm is fitted on each training set separately an afterwards tested against the test set. In this research a 10-fold cross validation is used, where the process is repeated 10 times. Cross validation is a powerful tool when data is scarce since all data is used for validating and training and the data should not be split in separate validating and training data sets.

Figure 4.2 shows the example of a 10-fold cross-validation scheme, where the white boxes indicate the training data used for modelling in each fold, the black box corresponds to the validation set.

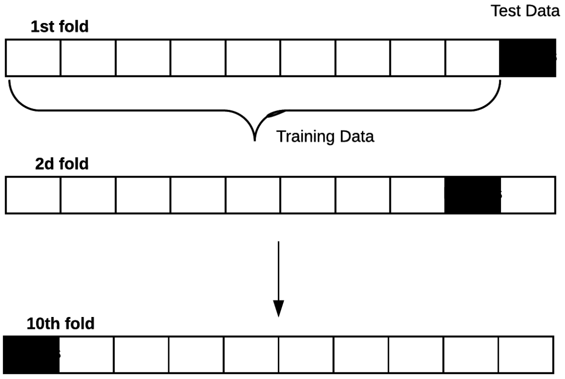


Figure 4.2: 10-fold cross-validation scheme

4.2. Multiple linear regression

The first algorithm that can predict the price of a specific individual flight is the multiple linear regression. Multiple linear regression is a technique where a relationship between one of the input variables and the output value can be seen as a straight line. Multiple linear regression has more than one input variable, therefore the straight line can be thought of as a (hyper)plane.

The different input values are used to predict the output value. Each of the input features (x_i), where i is the amount of independent/predictor features, is weighted using a coefficient (β). The goal of the learning algorithm is to obtain a combination of coefficients (β) that results in optimal predictions of (Y) [14]. The Mathematical Formulation is given in Equation 2.1.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon \tag{4.1}$$

A critical part of the multiple linear regression process involves ensuring that the data to be analysed can actually be analysed by multiple linear regression. Therefore, eight assumptions have to be checked. This repetitive process of checking the assumptions is based on a step-by-step manual for multiple linear regression using SPSS by Laerd statistics [83].

1. **Assumption 1: The dependent variable is continuous.**

2. **Assumption 2: There are two or more independent variables, which can be either continuous or categorical.**

3. **Assumption 3: There has to be independence of observation.**
   The independence of observations is statistically tested using the Durbin-Watson test. The test is used to detect the presence of auto-correlation (relationship between values by a given time lag) in the prediction errors. For this reason, it is important to check the Durbin-Watson variable since booking data is collected over time. The Durbin-Watson statistic is always between 0 and 4. A value of 2 means that there is no auto-correlation in the sample. Values close to 0 indicate positive auto-correlation and values close to 4 indicate negative auto-correlation. Values close to 0 and 4 are correlated across time. Values under 1 and above 3 are a cause for concern. The Durbin-Watson coefficient is calculated using Equation 4.2

$$\mathrm{DW} = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \tag{4.2}$$

4. **Assumption 4: There needs to be a linear relationship between (a) the dependent variable and each of the independent variables, and (b) the dependent variable and the independent variables collectively.**

5. **Assumption 5: The data needs to show homoscedasticity of residuals (equal error variances).** To check for homoscedasticity, one can plot the studentised residuals against the unstandardised predicted values. If there is homoscedasticity, the spread of the residuals will not increase or decrease as one moves across the predicted values, meaning there is an approximate constant spread. A studentised residual is the result of dividing the residual by the standard error of the residual, where the residual is the difference between a predicted value and the observed value.

6. **Assumption 6: The data must not show multicollinearity.** If there are two or more independent variables highly linearly related with each other, multicollinearity occurs. Multicollinearity gives problems on understanding which independent variable contributes to the output of a multiple regression model. There are two manners to identify multicollinearity: inspection of correlation coefficients and VIF values (Variance Inflation Factor).

   - Correlation coefficients: Measure the strength of the association between two variables. The Pearson correlation coefficient, r, can take a range of values from +1 to -1, where 0 is no dependence between variables. None of the independent variables have correlations greater than 0.7. The Equation is given in 4.3.

   $$r = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}} \tag{4.3}$$

   - VIF: Calculates an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity. A VIF greater than 10 means that there might be a collinearity problem. Equation 4.4 demonstrates how VIF is calculated, where R_k^2 is the R^2-value obtained by regressing the k^{th} predictor on the remaining predictors. R^2 is the coefficient of determination and is explained in Chapter 4.5.1.

   $$VIF_k = \frac{1}{1-R_k^2} \tag{4.4}$$

   Every route showed correlations between the input variables Sunday stay and length of stay. These two parameters are assuming that there is a return flight, which means there is already a correlation. Secondly, when the length of stay exceeds 7 days, there is definitely a Sunday stay. Therefore one of these parameters is chosen for the prediction of the individual flight price. This selection process is explained further on in Chapter 5.6.

7. **Assumption 7: There should be no significant outliers, high leverage points or highly influential points.**

   Outliers, leverage and influential points are different terms for unusual points. This classification reflects the impact the point has on the regression lines. An observation can be classified as different types of unusual point. Outliers, leverage and influential points can have a negative effect on the multiple linear regression equation. Therefore, if an unusual point has an influence on the he predictive accuracy of the model, the point can be deleted or an independent variable can be transformed.

   - Outliers: An outlier is an observation (data point) that does not follow the usual pattern of points. The observation is far away from their predicted value. Outliers can be checked by use of the studentised deleted residual. Equation 4.5 gives the studentised deleted residual where e_i is the ordinary residual, $MSE_{(i)}$ is the mean square error based on the estimated model with the i_{th} observation deleted, and the leverage, h_{ii}. An outlier can be an error in the data, in that case the data point should be removed and the assumptions should be checked again. If there is no clear indication of an error, the point should be evaluated together with the leverage and influence.

   $$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \tag{4.5}$$

- High leverage points: Leverage is a way of finding out how far away the independent variable values of an observation are from those of the other observations. To determine whether any cases has high leverage, a general rule of thumb is used. Leverage values less than 0.2 are considered as safe, leverage values between 0.2 to 0.5 are considered as risky, and values of or higher 0.5 are considered as dangerous. Equation 4.6 shows how leverage is calculated with N, the number of observations. If there are high leverage values that are of concern, a record of those points needs to made and there should be checked whether the points also have a high influence.

$$h_i i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum\limits_{j=1}^{n} (x_j - \bar{x})^2} \tag{4.6}$$

- Highly influential points: Cook's Distance is a measure of influence, measuring the effect of deleting a given observation. If there are Cook's Distance values above 1, than these values need to be investigated. Cook's Distance, D_i is shown in Equation 4.7 and depends on the residual e_i, the leverage, $h_i i$ and p, the number of parameters including the intercept. If there are values of concern, there are 2 options: Delete the specific cases or transform an independent variables.

$$D_i = \frac{(y_i - \hat{y}_i)^2}{p \times MSE} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right] \tag{4.7}$$

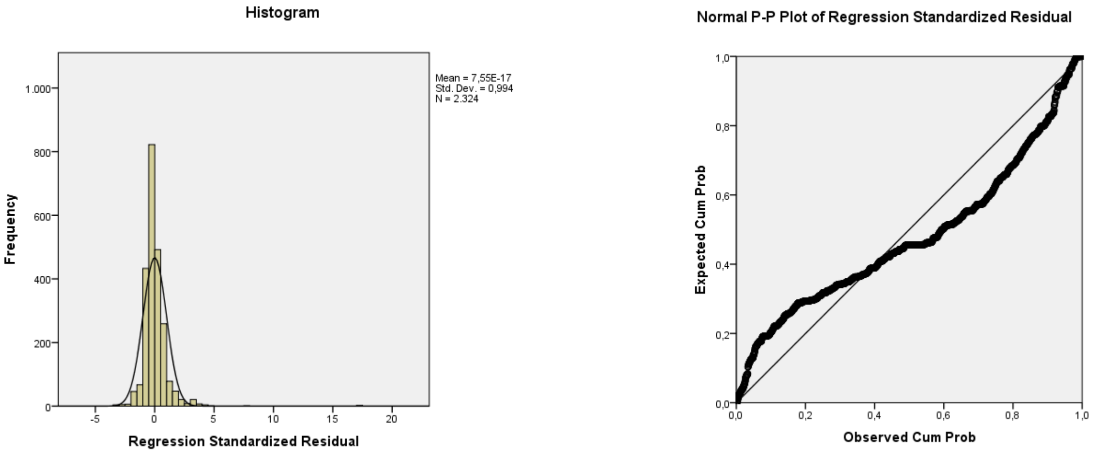
The data showed unusual points. Some points were clearly errors and could be deleted but others needed to be investigated. After investigation, some independent variables were transformed resulting in less unusual points and a better linear fit.

8. **Assumption 8: The residuals (errors) are approximately normally distributed.**
The errors in the prediction (the residuals) need to be normally distributed to determine statistical significance. The methods used to check the assumption of normality of the residuals are:

    (a) P-P Plot with the use of standardised residuals

    (b) Normal Q-Q Plot of the studentised residuals

The P-P histogram and Q-Q plot for the the route NBO-JED are shown in Figure 4.3a and 4.3b.



(a) Normal Q-Q Plot for JED

(b) P-P Plot for JED

Figure 4.3: Assumption 8: The residuals (errors) are approximately normally distributed

As can be seen from the histogram in Figure 4.3a, the standardised residuals appear to be approximately normally distributed. The mean and standard deviation are shown in the top-right and should have values of approximately 0 and 1, respectively. The distribution is somewhat peaked but normally distributed enough to indicate that the residuals are close enough to normal. To confirm that the residuals are approximately normally distributed, one should analyse the P-P plot. The points should be approximately aligned with the diagonal line. Regression analysis is fairly robust to deviations from normality so therefore approximate normality is good enough. As one can see from the P-P Plot above, the data is not perfectly aligned along the diagonal line. Although, the data is close enough to indicate that the residuals are close enough to normal. As multiple regression analysis is fairly robust against deviations from normality, this result can be accepted. This means that no transformation needs to take place.

These assumptions are checked for every route. Assumption 1,2,3 are correct for every route. From assumption 4 some data points had to be changed or deleted and transformation of variables took place. This process was an iterative for each route separately. After changing a value in the data, every assumption was checked again. Finally all assumptions were confirmed and the model could be used for prediction.

4.3. Random forest regression

The second model chosen to predict the ticket price of a flight is the random forest regression. The random forest starts with a standard machine learning technique called a "decision tree". Breiman et al. [27] introduced in 1984 the CART decision tree learning algorithm, which is capable of performing regression as well as classification. As the names suggests, regression trees have a continuous response variable and classification trees have a discrete response variable. This research makes use of the CART algorithm for a regression predictive modelling problem.

Unlike linear regression, decision trees are non-parametric models. Trees are represented by a string of choices and decisions. A regression tree is built by recursively partitioning the sample (= the "root node") into more and more homogeneous groups, so-called nodes, down to the "terminal nodes". The partitioning is called a split. Each split is based on the values of one variable and is selected according to a splitting criterion. The splitting criterion used for a regression tree is the mean squared error (MSE). The model takes into account all possible splits for each feature and choose the feature and the split that minimises the MSE. The following general steps are required to model a decision tree:

1. Find the best splitting point for each feature. The best split is found by minimising the square errors in the two separate parts, using Equation 4.8.

$$min = \sum_{i:x_{ij}>s} (y - y_i)^2 + min = \sum_{i:x_{ij}\leq s} (y - y_i)^2 \tag{4.8}$$

2. Find the feature that maximises the performance of the algorithm using the splitting point calculated in step 1. This is the same as minimising the error, this error is in a regression tree calculated by the MSE.

3. Repeat for each new node (child) until the stopping condition is met.

The stopping condition is a pre-defined parameter, which can be either the minimum samples per split or the minimum samples per node.

To illustrate the decision tree algorithm, the flight pass problem is simplified, since the real problem has more than 20 features resulting in huge trees. Figure 4.4 shows the decision tree that predicts the price of a flight depending on only 2 input features for the route from Nairobi to Jeddah (JED). The first feature is point of sale and the second feature is ticketing lead time. Point of sale is a dummy variable and therefore split into 3 variables with a value of 0 or 1. If the point of sale is Kenya, the first dummy variable is 1 and the others 0. If the ticket is bought in Saudi-Arabia the second dummy variable is 1 and if the ticket is bought in neither Kenya or Saudi-Arbabia the third dummy variable is 1. Ticketing lead time is a continuous variable and values differ from 1 day until 60 days. Eventually four features (1 continuous and 3 dummies) predict the price of the flight.

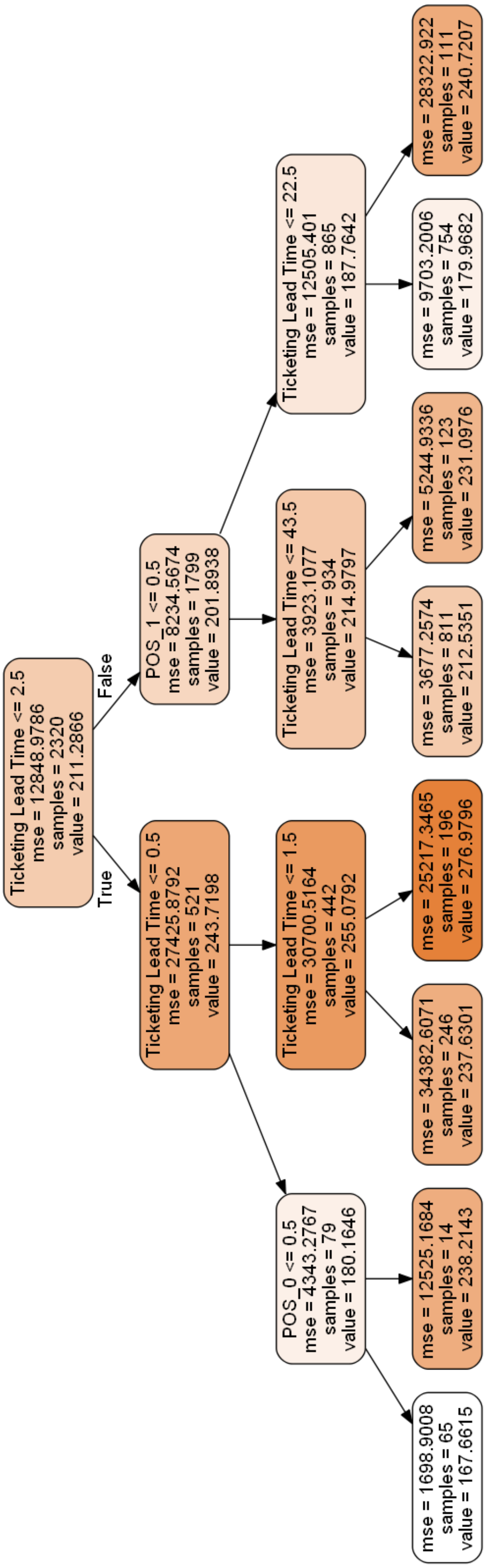


Figure 4.4: Decision Tree: Predicting the ticket price using ticketing lead time and point of sale

The result of the first step is the splitting value for each of the 4 features. Step 2 results in ticketing lead time smaller than 3 days providing the best first split. Repeating this for the child nodes, the better splits for the right and left branches are respectively the ticketing lead time and point of sale. Important and significant features are split near the top of the tree. The terminal nodes represent the predicted value for the followed path in the tree. One can see that the final prices differ. The cheapest ticket price prediction is 167 euro while the most expensive is 267 euro.

Like in real life, a forest is an ensemble of trees. Breiman introduced the general concept of a random forest in 2001 [26]. A random forest model operates by constructing a lot of parallel decision trees with different subsets of data. To create some variability between the different decision trees and to reduce the risk of over-fitting, the subsets are drawn at random from the full training set. The different decision trees will be combined and averaged. Breiman added a second parameter next to the number of trees which is unique for the random forest algorithm. The second parameter is how many features to search for to find the best feature split. During tree creation, a random number of features is chosen from all available features and the best feature that splits the data is chosen.

Summarised, the random forest algorithm is a follow up from a decision tree. A very complex decision tree that grows very deep, learns irregular patters, which means the tree is over-fitting the training set. The goal of the random forest model is to avoid the over-fitting of one decision tree model by training trees on different subsets of the data (bootstrap aggregation) and by randomly selecting the amount of features to consider in a split.

Since the random forest regression is not a probabilistic model, but a binary split, there are no assumptions to make like the multiple linear regression. However only one common assumptions should be made and that is that the sampling is representative. For example, if one feature consists of two components and one component represents 99,9% of the sample and the other component only represents 0,1% of the sample then most individual decision trees only notice the first component and the random forest regression misclassify the second component.

Every route has different hyper-parameters which results in 46 different variations of the random forest model. For each route the following hyper-parameters, which are parameters that need a value before the start of the machine learning process, are optimised:

- The number of trees in the forest
- The number of features to consider when looking for the best split
- The maximum depth of the tree
- The minimum number of samples required to split an internal node
- The minimum number of samples required to be at a leaf node

The values for those parameters are derived when the model is trained. The concept of optimising the hyper-parameters is called tuning and defined as the problem of choosing a set of optimal hyper-parameters for the machine learning algorithm. Hyper-parameters have to be tuned so that the model can solve the problem in an optimal way. The performance and accuracy of the different combinations of the hyper-parameters is measured by an evaluation metric. Finally the best combination with the highest performance is chosen.

Grid search is the traditional way of tuning. Grid search exhaustively considers all parameter combinations that are manually specified as sets and bounds. In this research the coefficient of determination is used to evaluate the performance. Grid search can be limited due to the curse of the high amount of dimensions but typically the optimal value of a hyper-parameter is independent of the other hyper-parameters.

4.4. Multilayer perceptron neural network

There are different variations of the neural network, as discussed in the literature study. The model used in the research is the Multilayer Perceptron also commonly known as a feed-forward neural network. When referring to a neural network in this research the multilayer perceptron neural network is addressed.

A multiple linear regression or a random forest regression only considers the basic inputs feed in the algorithm, while a neural network combines the basic inputs into a higher dimension concept. A neural network mimics connections in the brain. When learning how to read, individual letters need to be recognised, these letters combined are words and the words form sentences. Eventually it is easy to recognise words without explicitly thinking about the letters.

The building blocks for neural networks are neurons, which can be described as units that have input signals that are weighted and produce an output signal while using an activation function. Figure 4.5 shows the architecture of a neuron [79]. The simplification of the flight pass problem used to explain the methodology of the random forest regression is used again to explain the working principle of the neural network.

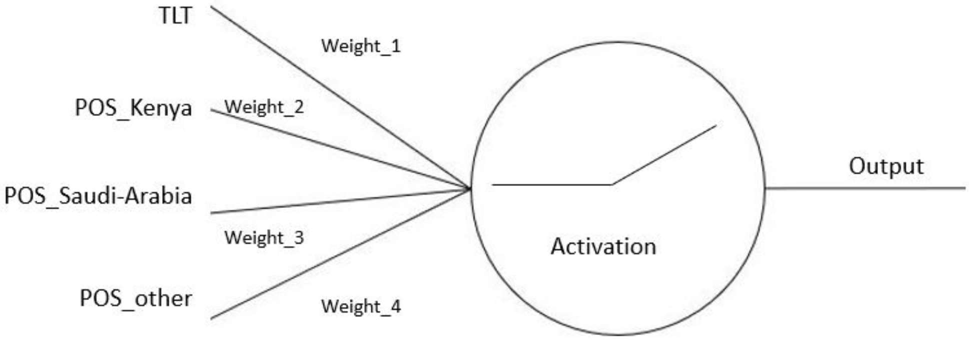


Figure 4.5: Active neuron

As can be seen in Figure 4.5, the inputs (ticketing lead time and point of sale) are weighted. One can compare it with the coefficients in an multiple linear regression. Like a regression, each neuron also has a bias. A neuron may have three inputs in which case it needs four weights. Each input and the bias requires a weight. The weighted inputs are summed and passed through an activation function.

The activation function

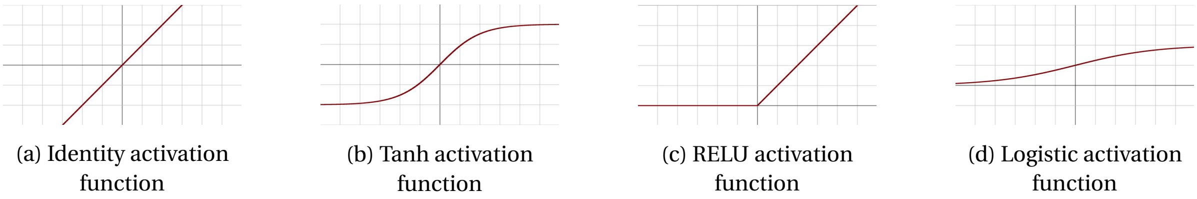
An activation function calculates a 'weighted sum' of its input to determine the output of the neuron. The activation functions used in the model are: the identity, tanh, rectified linear unit (RELU) and logistic function, which can be seen in Equation 4.9, 4.10, 4.11 and 4.12. The graphs are shown in Figure 4.6. These activation functions are the most popular activation functions because of their high accuracy and performance. Within one model, only one type of activation function can occur but since there are models for every route, the activation function that gives the best performance is chosen route dependent.

$$f(x) = x \tag{4.9}$$

$$f(x) == \frac{2}{1+e^{-2x}} - 1 \tag{4.10}$$

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \tag{4.11}$$

$$f(x) = \frac{1}{1+e^{-x}} \tag{4.12}$$



(a) Identity activation function

(b) Tanh activation function

(c) RELU activation function

(d) Logistic activation function

Figure 4.6: Function neural network activation function

Neurons are organised into a network of neurons. The architecture is shown in Figure 4.7. A row of neurons is called a layer. There is an input layer, a hidden layer and an output layer. The input layer consists of the input feature while every neuron in the hidden layer has the neuron structure, with an activation function as shown in Figure 4.5. The orange box corresponds to Figure 4.5. This network has 4 inputs and one hidden layer of 7 neurons with one output variable. The complexity of the network increases exponential when adding extra hidden layers. In this research only one hidden layer is considered because of time limits.

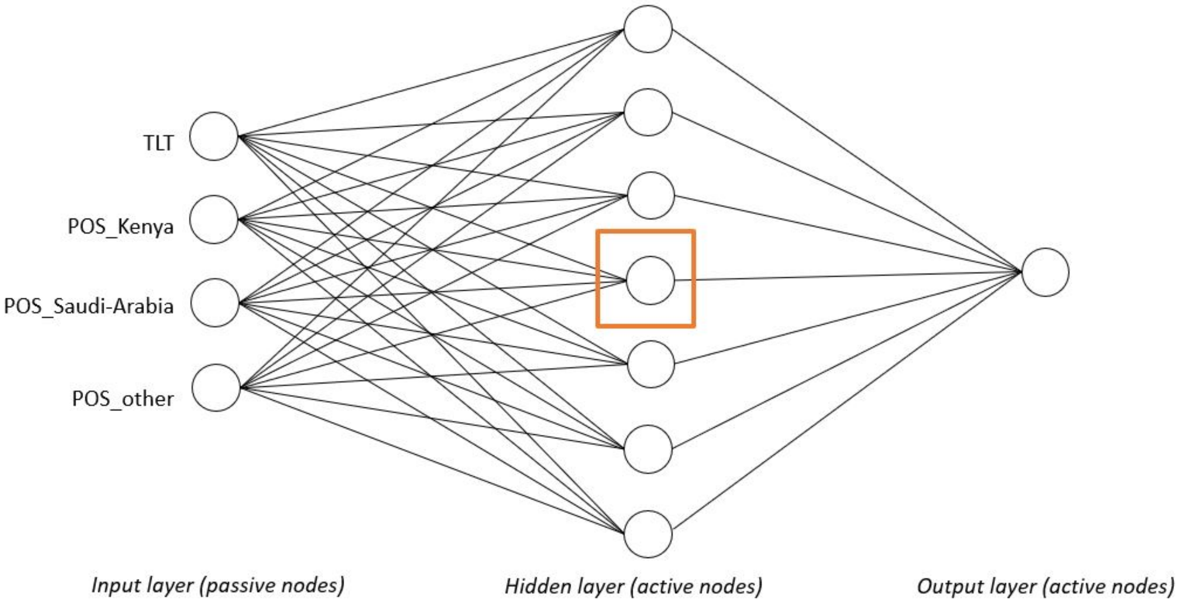

TLT
POS_Kenya
POS_Saudi-Arabia
POS_other
Input layer (passive nodes)
Hidden layer (active nodes)
Output layer (active nodes)

Figure 4.7: The illustration of the NN functional structure

The training of the network

After configuration, the neural network is trained. There are different training algorithms used depending on the route:

- Stochastic gradient descent (SGD) [55]: One row of data at a time is presented as an input. The input is processed upward activating neurons as it eventually produces an output value. This is described as a forward pass on the network. Next, the error is calculated by comparing the output of the network to the expected output. This error propagates back through the network, one layer at a time, and the weights are updated according to the amount that they contributed to the error. This is called the back-propagation algorithm.

- Adaptive Moment Estimation (ADAM) [59]: The ADAM algorithm is an extension and a variant to the basic SGD. The ADAM algorithm stores an exponentially decaying average of past squared gradients.
- Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) [45]: The LBFGS is a quasi-newton method where the Hessian in Newton's method is replaced with an approximate Hessian. LBFGS is more complicated to implement and requires more storing.

The learning rate of the network

The amount of weights that are updated is controlled by a parameter, called the learning rate. It controls the change made to the network weight.

There are different learning rates considered in time:

- Constant: A constant learning rate.
- Invscaling: Gradually decreases the learning at each time step 't' using an inverse scaling exponent.
- Adaptive: Keeps the learning rate constant as long as training loss keeps decreasing. Each time two consecutive epochs fail to decrease training loss, the current learning rate is divided by 5.

Every model has different characteristics. For each route the following hyper-parameters are optimised:

- The number of neurons in the hidden layer
- How the network is trained (Solver)
- Which activation function is used
- How the algorithm learn (Learning rate)
- The maximum iterations

Every route has different values for the above named parameters. For example the model for Jeddah has: (1) 15 neurons in the hidden layer; (2) it is trained with an LBFGS; (3) the RELU activation function is used; (4) the algorithm learns adaptive and (5) there are 100 iterations. While the model for Mombassa has 5 different values resulting in a complete different neural network model. In order to select the optimal values for these hyper parameters, a grid search is used which is also used for the random forest model

4.5. Evaluation metrics

To evaluate the performance of the multiple linear regression, the random forest regression model and the neural network, evaluation metrics are chosen.

The main objective of the thesis is to predict the price of the flight pass. Since the predictor variable is continuous, this is a regression problem. The main evaluation metrics used for regression are the root mean squared error (RMSE), the mean absolute error (MAE) and the Coefficient of determination.

4.5.1. Coefficient of determination

The coefficient of determination, which is also known as R^2, is a standard way of measuring how well the model fits the data. It can be interpreted as the proportion of variability in a data set that can be explained by the statistical model. R^2 can take 0 as minimum, and 1 as maximum. Values close to 0 present a low fit and values close to 1 are 'good' fits. Without knowing the nature of the problem there are no guidelines for a suitable R^2. For a physical process an R^2 of 0,9 could be good but when predicting human behaviour this value would be way to high and would insinuate over-fitting. Predicting human behaviour, will usually result in R^2 values lower than 0,5 [95] [86].

Revenue management prices consist of mathematical models incorporated with human market experience. Therefore, an expected 'good' fit for this problem would have R^2 values between 0,4 and 0,8. The equation explaining the coefficient of determination is given in Equations 4.13, 4.14 and 4.15.

$$\mathrm{SST}_{\mathrm{otal}} = \mathrm{SSE}_{\mathrm{xplained}} + \mathrm{SSR}_{\mathrm{esidual}} \tag{4.13}$$

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 \tag{4.14}$$

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSR}_{\mathrm{esidual}}}{\mathrm{SST}_{\mathrm{otal}}} \tag{4.15}$$

- $\hat{y}_i$: The predicted value of the dependent value.
- y_i: The real value of the dependent value.
- $\bar{y}$: The mean of the observed data

$\mathrm{SSR}_{\mathrm{esidual}}$ is the sum of the squares of residuals, which is the deviations predicted from actual empirical values of data. $\mathrm{SST}_{\mathrm{otal}}$ is the total sum of squares, which is the sum, over all observations, of the squared differences of each observation from the overall mean. So if the model explained all the variation, $\mathrm{SSR}_{\mathrm{esidual}} = \sum(y_i - \hat{y}_i)^2 = 0$, and $\mathbf{R^2 = 1}$.

4.5.2. Root mean squared error

The root mean squared error is the difference between values predicted by a model and the values actually observed. RMSE is not the average error but RMSE is the square root of the average of squared differences between the prediction and the actual observation. The formula for RMSE is shown in Equation 4.16.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.16}$$

- n: The number of elements in the sample.

RMSE punishes larger errors giving relative high weights on those errors. When large errors are particularly undesirable, RMSE is very useful.

4.5.3. Mean absolute error

The mean absolute error measures the average magnitude of the errors in a set of predictions. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The formula for MAE is shown in Equation 4.17.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{4.17}$$

4.6. K-means cluster

After selecting the best performing model for each specific route, the different routes are clustered. The feature importance output of the best performing prediction model is used to cluster the routes.

Clustering is defined as the grouping of a particular set of objects based on their characteristics and group them based on their similarities. Jose describes a cluster as the collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters [56]. Clustering is considered as one of the most used unsupervised classification problems. The goal of clustering is to group unlabelled data. Literature has been studied to compare the different cluster algorithms. Researchers agree that there is no objectively ”correct” clustering algorithm [72] [40].

Therefore, the choice was made to start with one of the most simple and widely-used methods, namely, the k-means clustering algorithm (KMC). k-means clustering is based on the vector distance between the different data entries. Distance is measured by subtracting one point from another and squaring. If $X = \{x_1, \cdots x_n\}$ is a set of feature vectors, then the k-means clustering algorithm tries to minimise the objective function, in order to cluster n feature vectors into k clusters, namely $S_1 \cdots S_k$. μ_i is the centroid of cluster S_i [67]. The objective function is shown in Equation 4.18.

$$J = \sum_{i=1}^{k}\sum_{x\in S_i}||x-\mu_i||^2 \tag{4.18}$$

Figure 4.8, shows a clustered scatter plot. As can be seen there are $k = 5$ clusters. The blue lines show the different cluster S_i. The blue dots are the centroids, μ_i. The black dots are the set of feature vectors, X.

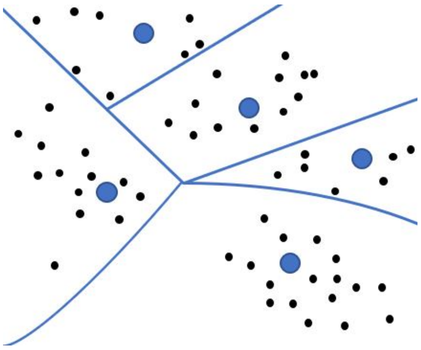


Figure 4.8: A clustered scatter plot.

The k-means clustering algorithm is composed by the different steps, explained in Figure 4.9.

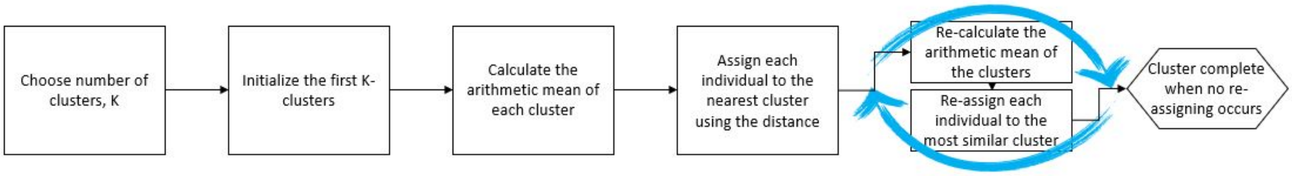

Choose number of clusters, K
Initialize the first K-clusters
Calculate the arithmetic mean of each cluster
Assign each individual to the nearest cluster using the distance
Re-calculate the arithmetic mean of the clusters
Re-assign each individual to the most similar cluster
Cluster complete when no re-assigning occurs

Figure 4.9: The k-means algorithm

As can be seen in Figure 4.9 the model has to be initialised. The methods used for initialisation are:

- Random: The initial centers are picked randomly
- K-means++: The initial centers are allocated with an even spread. This way convergence is speed up [16].

The cluster model is only one model (not depending on the route), where the following hyper-parameters are optimised.

- Method for initialisation
- Number of clusters
- Number of time the k-means algorithm runs with different centroid seeds
- The maximum of iterations

Grid search is used to select the optimal values for the parameters.

Silhouette coefficient

The Silhouette Coefficient is used to evaluate the cluster algorithm. A higher Silhouette Coefficient score means that the model has better defined clusters. The minimum score is -1 for incorrect clusters and the maximum +1 for highly dense clusters. Scores around zero are for overlapping clusters. The Silhouette Coefficient s for a single sample is then given in Equation 4.19. The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample.

$$\mathrm{s} = \frac{b-a}{max(a,b)} \tag{4.19}$$

- a: The mean distance between a sample and all other points in the same class.
- b: The mean distance between a sample and all other points in the next nearest cluster.

4.7. Monte-Carlo simulation

Since a customer can book different flights within one pass, the total flight pass value has to be estimated. A Monte-Carlo simulation is used to simulate the behaviour of a passenger. The set of samples created is called an iteration. This Monte-Carlo simulation calculates a result 10 000 times using a different set of random values. The result of the Monte-Carlo simulation is a probability distribution of the outcomes.

The goal of the Monte-Carlo simulation is to provide a comprehensive view of what can happen when someone buys the flight pass. The distribution does not only show what can happen but the likelihood whether it will happen.

The prediction of the price of these 10 000 flights is plotted in a histogram where the y-axis represents the amount of times the prices occur. The x-axis represents the price of the flight. The price of a flight which is at 80 percent of the 10 000 flights is chosen for the final value of a flight within the flight pass bounds. The 80% is an assumption made with an expert.

To successfully simulate the flying behaviour within the bounds of a pass, bounds need to be selected. Assumptions on these bounds need to made. Therefore a user case is explained in Chapter 7 to fully explain the estimation of the final flight pass price.

5 Case Study

To come up with the price of a flight pass, first the individual specific tickets prices are determined using a multiple linear regression, a random forest regression or a multilayer perceptron neural network model after which a Monte Carlo simulation is performed to simulate the customer behaviour. Nevertheless, in order to do this data sets are needed. This chapter will describe the data which is used in this research. First by explaining the data treatments needed in order to be able to use the data and afterwards by listing some practical implementations and limitations of the methodology used.

This research uses data of Kenya airways and the concept of the flight pass of Optiontown, therefore the background information of the companies is described in Chapter 5.1. The data treatment is described in Chapter 5.2. Chapter 5.3, 5.4 and 5.5 present the implementation of respectively the multiple linear regression, random forest regression and neural network. Different tests on the features are performed to optimise the models, these tests are discussed in Chapter 5.6 and finally, the practical implementation of the k-means clustering is explained in Chapter 5.7.

5.1. Background information on Kenya Airways and Optiontown

Kenya Airways (KQ), the pride of Africa, is the national carrier of Kenya and part of the SkyTeam alliance. It is the fourth largest African airline after South African Airways, Egyptian and Ethiopian Airlines. Kenya Airways is operating a fleet of 30 aircraft's consisting of Embraer 190 (15), Boeing 737-800 (8) and 700 (2) and Boeing 787-8 (5).

It is operating a hub and spoke network with Jomo Kenyatta International Airport (JKIA or NBO) as its main hub, which is located in Nairobi. Figure 5.1 shows the flight pass network of Kenya Airways. The flight pass is available for every flight except for the flights to Europe (CDG, AMS and LHR). The company that offers the IT structure and platform for the flight pass is Optiontown. Optiontown develops models that are unique and innovative to the travel industry. They want to benefit as well travellers as travel providers. The optimisation techniques are extensively researched at Massachusetts Institute of Technology (MIT), Boston [2].

The flight pass is one of the innovative travel options that Optiontown offers. Several major airlines like KLM, Alitalia, Garuada Indonesia and British Airways have a partnership with Optiontown.

Kenya Airways has the freedom to choose the parameters that determine the final price of the flight pass. Kenya Airways can choose to add parameters as day of week or time of day of the flight. One could access the website via: `https://www.optiontown.com/home_page.do?processAction=SelectProduct`

At this moment the prices are determined with a linear model and the model is not route dependent. This means that for every route, booking 1 day in advance or 14 days in advance, the price of the flight pass increases with the same percentage.

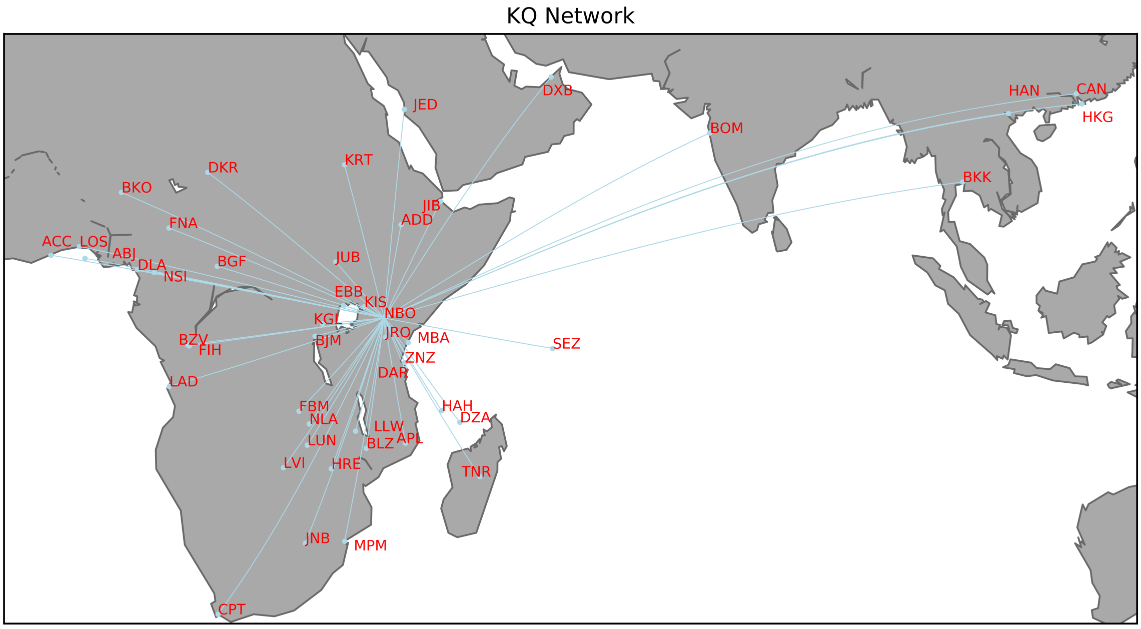


Figure 5.1: The network of Kenya Airways

5.2. Description and treatment of data sets

The data used for this thesis is obtained from the revenue management systems at Kenya Airways. The data is extracted from *Monet*. The system is originally developed by the KLM.

Monet contains revenue management data which contains flight data of already realised flights. Every month, *Monet* is updated with data of the previous month. The data in *Monet* is based on flight details and not on passenger details. This means that if a passenger flies from Amsterdam to Madagascar, the passengers details are available in the flight leg Amsterdam-Nairobi and Nairobi-Madagascar. However, it is not possible to identify that the two flights belong to the same booking. Therefore, only point to point flights are extracted from Monet. *Monet* does not make predictions for the future but only stores the revenue information about completed flights. The data format as extracted from *Monet* is shown in Table 5.1. The data is already aggregated for the different fields extracted. This means that the revenue is the **total** revenue for the selected fields.

Table 5.1: Format of data extracted from *Monet*

Flight	Cls	Month	DOW	Org	Dest	POS	Sunday stay	LOS	TLT	Revenue	PaxKm
KQ0887	E	201608	Fri	CAN	NBO	China	Sunstay Yes	'7->	'8	20131,6	459631
KQ0783	T	201608	Wed	LVI	NBO	China	Unknown	'Unknown	'39	19864,57	92401
KQ0764	Q	201609	Mon	JNB	NBO	South Africa	Sunstay Yes	'3	'3	19719,3	218374
KQ0765	Q	201609	Fri	JNB	NBO	South Africa	Sunstay Yes	'3	'4	19719,3	218374
KQ0886	N	201610	Thu	CAN	NBO	Kenya	Sunstay Yes	'7->	'3	19225,11	442286
KQ0762	C	201703	Sat	JNB	NBO	Kenya	Unknown	'Unknown	'23	18563,37	32028
KQ0871	L	201608	Sat	CAN	NBO	China	Sunstay Yes	'7->	'1	18470,9	260168
KQ0256	L	201612	Sat	NBO	TNR	China	Unknown	'Unknown	'15	18329,38	65509
KQ0793	N	201705	Sat	CPT	NBO	South Africa	Sunstay Yes	'7->	'1	18309,75	311634
KQ0740	L	201611	Mon	MPM	NBO	Kenya	Unknown	'Unknown	'6	18062,48	94351
KQ0887	T	201608	Fri	CAN	NBO	China	Sunstay Yes	'7->	'10	17940,66	364236
KQ0448	B	201608	Fri	KGL	NBO	Kenya	Sunstay Yes	'7->	'3	17922,64	31089
KQ0871	Z	201707	Tue	CAN	NBO	China	Sunstay Yes	'7->	'48	17920,05	138756
KQ0870	Z	201608	Wed	CAN	NBO	China	Sunstay No	'7->	'21	17784,8	173446
KQ0740	L	201611	Wed	MPM	NBO	South Africa	Unknown	'Unknown	'3	17750,14	77701
KQ0351	N	201608	Sat	JUB	NBO	Kenya	Unknown	'Unknown	'6	17594,51	52634
KQ0577	L	201701	Sun	BGF	NBO	Kenya	Unknown	'Unknown	'2	17221,88	53548
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The different parameters extracted from *Monet* are explained below:

1. **Flight**: the flight number of the flight.
2. **Subclass (Cls)**: the fare class in which the reservation is made.
3. **Month**: the month the flight took place.
4. **Day of Week (DOW)**: the day of the week the flight took place.
5. **True Origin (Org)**: the International Air Transport Association (IATA) airport code of the origin of the passenger.
6. **True Destination (Dest)**: the International Air Transport Association (IATA) airport code of the destination of the passenger.
7. **Point of Sale (POS)**: the country where the ticket is bought.
8. **Sunday Stay (SUN)**: is there a Sunday between the flight and return flight.
9. **Length of Stay (LOS)**: the amount of days between the flight and return flight.
10. **Ticketing Lead Time (TLT)**: the amount of days before departure the ticket is bought.
11. **Revenue**: the total revenue of all the passengers that have fields 1 to 10 in common.
12. **PaxKm**: the amount of passengers that have fields 1 to 10 in common.

Some data specifics need to be mentioned:

- The data extracted is from exactly 1 year from August 2016 till July 2017.
- Only point to point passengers are considered. This means that a passenger flying from Amsterdam to Mombasa (AMS-NBO and NBO-MBA) is not extracted from *Monet*.
- Only flights which offer a flight pass are considered. These are 46 point to point routes departing or arriving in NBO and one of the airports listed in Appendix B.

Different manipulations are used to make sure that the data is ready for calibration of the multiple linear regression, random forest model and multilayer perceptron neural network. The overview of the manipulation process is shown in Figure 5.2.

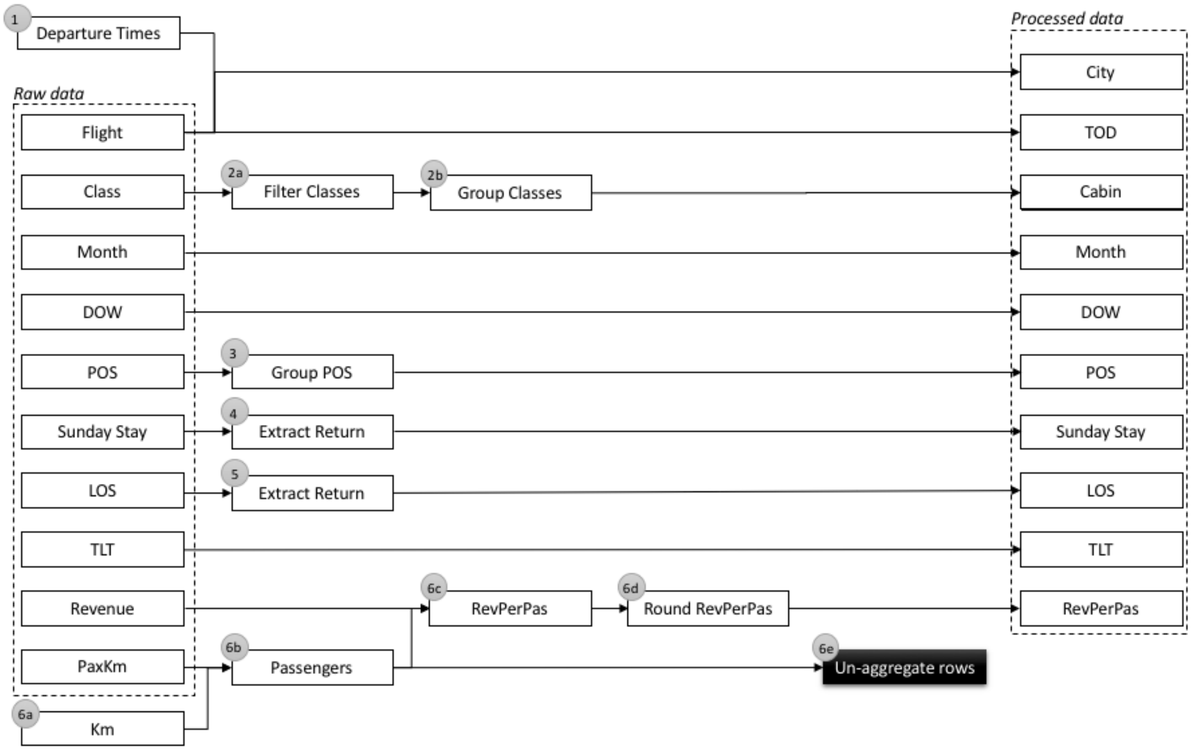


Figure 5.2: Pre-processing data

1. **Raw data departure times**: It is useful to include time of departure of the flight as an input variable of the model. The data which has the flight number and the time of departure of the flight is merged with the raw data of the bookings. The variable is defined as Time of Day (TOD).

2. **Classes**: A ticket can be booked in 20 different (sub)classes. Since the goal of the model is to be able to predict the price without explicitly modelling the inventory control, the class should not be an input parameter of the model.

   (a) **Filter classes**: There are several classes that should not be considered in the model. An overview of the classes is given in Appendix A. The flights booked in classes that are not of interest where removed from the data.

   - Tickets booked with award miles: O and X, these are not of interest because there is no money paid for the ticket.
   - Group bookings: G, group bookings are not of interest in predicting prices because; (1) Large group bookings can dominate the prices and (2) The pricing structure for group bookings is different than the other classes and would weaken the performance of the models.

   (b) **Group classes**: As discussed, the class should not be an input variable for the model but there should be made a distinction between Economy and Business class. This variable is defined as Cabin.

3. **Group POS**: There are 145 different point of sales in the booking data. This is too much to include in the model. Therefore the point of sale is grouped for each route in:

   - Kenya
   - Country of origin/destination
   - Other countries

   This means there are three categories for point of sale for each route.

4. **Extract return**: If the value for Sunday stay is unknown in the raw data, one can assume a one way flight.

5. **Extract return**: If the value for length of stay is unknown in the raw data, one can assume a one way flight. (same situation as Sunday stay). Length of stay is grouped 1,2,3,...> 7.

6. **RevPerPas**: The final goal is to convert the data to booking level.

   (a) **Raw data Km**: In the raw data file, the PaxKm is a variable. Another data file with the distance of the different flights is merged with the data file. So when the amount of km of the flight is known, the amount of passengers can be calculated.

   (b) **Passengers**: If the amount of km and the PaxKm of the flight is known, the amount of passengers can be calculated.

   (c) **RevPerPas**: Eventually the RevPerPas can be calculated using the revenue and the amount of passengers. Equation 5.1 shows the calculations.

$$\text{RevPerPas} = \frac{\text{Revenue}}{\frac{\text{PaxKm}}{\text{Km}}} \tag{5.1}$$

   (d) **Round RevPerPas**: The value of the dependent variable in the random forest model can not be unique, meaning the frequency of the value of the dependent value should be at least 2. Therefore the RevPerPas is rounded to 5 dollars for economy class en 10 dollars for business class.

   (e) **Un-aggregate rows**: One row/entry is still aggregated for an x amount of passengers. The input of the model should be un-aggregated meaning that one row should be multiplied by the amount of passengers that represent that row.

After the previous steps the final matrix with processed data is presented in Table 5.2.

Table 5.2: Processed data

City	TOD	Cabin	Month	DOW	POS	Sunday stay	LOS	TLT	RevPerPas
CAN	21:40	Economy	8	Fri	Org/Dest	Yes	7	8	380
LVI	18.15	Economy	8	Wed	Other	Single Ticket	Single Ticket	39	470
JNB	21.15	Economy	9	Mon	Org/Dest	Yes	3	3	260
JNB	01:15	Economy	9	Fri	Org/Dest	Yes	3	4	260
CAN	23:59	Economy	10	Thu	Kenya	Yes	7	3	380
JNB	12:55	Business	3	Sat	Kenya	Single Ticket	Single Ticket	23	1680
CAN	22:10	Economy	8	Sat	Org/Dest	Yes	7	1	615
TNR	11:10	Economy	12	Sat	Other	Single Ticket	Single Ticket	15	630
CPT	14:35	Economy	5	Sat	Org/Dest	Yes	7	1	240
MPM	11:05	Economy	11	Mon	Kenya	Single Ticket	Single Ticket	6	530
CAN	21:40	Economy	8	Fri	Org/Dest	Yes	7	10	430
KGL	07:35	Economy	8	Fri	Kenya	Yes	7	3	440
CAN	22:10	Business	7	Tue	Org/Dest	Yes	7	48	1120
CAN	23:45	Business	8	Wed	Org/Dest	No	7	21	890
MPM	11:05	Economy	11	Wed	Other	Single Ticket	Single Ticket	3	630
JUB	10:20	Economy	8	Sat	Kenya	Single Ticket	Single Ticket	6	305
BGF	14:45	Economy	1	Sun	Kenya	Single Ticket	Single Ticket	2	690
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

After the initial processing of the data, it became apparent that there existed some erroneous data.

- RevPerPas can theoretically not drop under 0 but even a RevPerPas under 50 dollars is highly unlikely. Therefore data entries that have a RevPerPas < 50 are deleted from the data set assuming the data is incorrect.

- Flights can have the same flight number for a different origin and destination. An example is shown in Figure 5.3. Flight KQ 0408 departs in Nairobi at 14:50 and flies directly to Djibouti where it arrives at 17.25. At 18.05 the flight goes back to Nairobi with a stop over in Addis Abeba. The flight arrives in Addis Abeba at 19.20 an takes off again at 20.00, finally it arrives in Nairobi at 22.10. When merging the departure time data file with the raw date, erroneous combinations of flights and time of departures where noted.

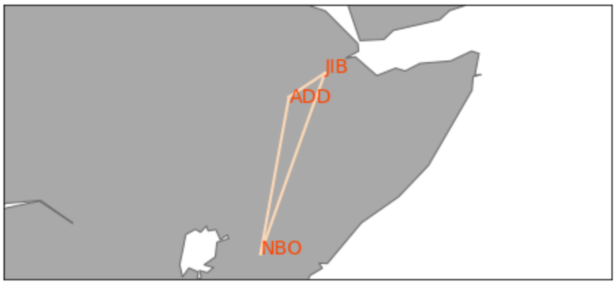


Figure 5.3: Flight KQ 0408: NBO-JIB-ADD-NBO

The choice was made to make separate models for each **Route** and for each **Cabin**. This means there are two models for each route (Business and Economy) When combining this with all the 46 routes, this results in 92 (46 routes x 2 cabins) models. Nevertheless, these models are used in all three different methods (the multiple linear regression, random forest regression and neural network). This means that there are 276 (3 (MLR+RFR+NN) x 46 routes x 2 cabins) models made in total.

With the completion of the steps described, the data is now ready to calibrate. 70% of the data is used as training and data and 30% is used for testing the model.

The sample size for the different models differs. One can find the different sample sizes in Appendix C.

The software used to model the problem is Scikit-learn. Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems [76].

5.3. Multiple linear regression model

At this point the data of KQ is known and discussed, therefore the models are discussed with the limitations and insights of the data. First the features that are extracted from the data are evaluated: (1) TOD; (2) Month; (3) DOW; (4) POS; (5) Sunday stay; (6) LOS and (7) TLT. These features predict the revenue per passenger for that specific individual flight.

The features can be used in different forms such as categorical, linear or cyclic. Therefore, tests are performed and the best combination of different forms is used for the final result. This is discussed in Chapter 5.6. For every combination, every assumption, should be checked again. To account for the categorical nature of some parameters, dummy variables need to be created.

5.4. Random forest regression model

The first step in the analysis of the random forest regression is to tune the hyper-parameters. The random forest regression model has the following parameters to tune: (1) The number of trees in the forest; (2) The number of features to consider when looking for the best split; (3) The maximum depth of the tree; (4) The minimum number of samples required to split an internal node and (5) The minimum number of samples required to be at a leaf node [76].

A range for hyper-parameters is given as input. Grid search can select the optimal combinations of values for the hyper-parameters. The ranges are listed below.

1. The number of trees in the forest: 10, 50, 100, 200
2. The number of features to consider when looking for the best split: auto, sqrt, log2

3. The maximum depth of the tree: None, 10, 15, 20, 25
4. The minimum number of samples required to split an internal node: 2, 4, 6, 8, 10
5. The minimum number of samples required to be at a leaf node: 1, 2, 5, 10

As the sample size of the data depends on the route, a wide range of parameters is considered. For some low sample size routes a maximum depth of 25 would be too much. But for high sample sizes, the maximum depth could be 25.

The number of features to consider when looking for the best split has 3 possible values and are the most typically used [65]:

- auto → max_features = n_features
- sqrt → max_features = $\sqrt{n_{features}}$
- log2 → max_features = $\log^2\left(n_{features}\right)$

,

The route depended results of the hyper-parameter tuning of the random forest regression are shown in Appendix D.

5.5. Neural network model

The hyper parameters that need to be tuned to get to the optimal neural network are: (1) The number of neurons in the hidden layer; (2) How the network is trained (Solver); (3) Which activation function is used; (4) How the algorithm learn (Learning rate) and (5) The maximum iterations [76].

This selection of the solvers, activation functions and learning rates is based on the most popular methods [54] [50].

1. The number of neurons in a hidden layer: 5, 7, 10, 15
2. Solver: SGD, ADAM, LBFGS
3. Activation function: Identity, logistic, tanh, RELU
4. Learning rate: Constant, invscaling, adaptive
5. The maximum iterations: 100, 200

The route depended results of the hyper-parameter tuning of the neural network are shown in Appendix E.

5.6. Feature tests

First, four tests are performed to optimise the models. The goal of the tests is to optimise the input features to get the optimal final results.

The first part consists of selecting the features. For multiple linear regression it is important to select features that are independent and not correlated. The only parameters which are correlated, are Sunday stay and length of stay. These two parameters are assuming that there is a return flight, which means there is already a correlation. Secondly, when the length of stay exceeds 7 days, there is definitely a Sunday stay. To comply with the assumptions stated, only one variable should be selected. The parameter which gives eventually the best performance of the model are chosen. Results are discussed in Chapter 6.1. This test is referred to as test Ⓘ. The random forest regression model is insensitive for correlations so correlated features can be used as an input. Therefore there is no test needed for which features to select. But the feature inputs can be manipulated or transformed.

Figure 5.4 shows an overview of the different 7 input features and types.

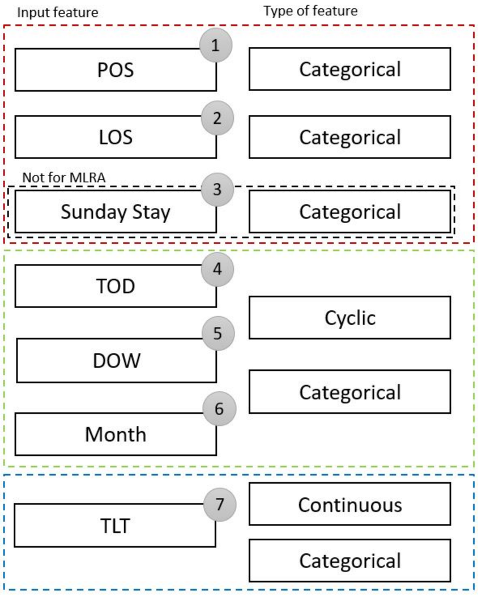


Figure 5.4: The type of feature

For the red box, the features have a categorical behaviour.

1. Point of sale is divided in 3 categories:
   - Kenya
   - The other country of origin or destination
   - Other countries
2. LOS is also divided in 3 categories:
   - Less than 3 days
   - 3 to 7 days
   - More than 7 days
3. Sunday stay has 2 categories:
   - Yes, there is a Sunday stay.
   - No, there is no Sunday stay.

The green box has 2 options to interpret the features. The features can be divided in categories or they can be transformed to circular coordinates. The categories are determined together with an expert and grouped as follows:

4. TOD:
   - Morning: 6AM - 10AM
   - Day: 10AM -18PM
   - Evening: 18PM - 24PM
   - Night: 24PM - 6AM

5. DOW:

   - Monday
   - Tuesday
   - Wednesday
   - Thursday
   - Friday
   - Saturday
   - Sunday

6. Month:

   - January
   - February
   - March
   - April
   - May
   - June
   - July
   - August
   - September
   - October
   - November
   - December

A year, week or day has a cyclic behaviour: 23h in the evening is as close to 2h in the morning and 21h in the evening. Therefore one could assume that the features month, day of week and time of day behave circular, that is why these features are translated into two continuous variables. Equations 5.2 and 5.3 illustrate the translations for time of day as its position varies smoothly as it travels around the unit circle. The same translation can be done for the months or different days of week.

$$x = \sin\left(2\pi \frac{hour}{24}\right) \tag{5.2}$$

$$y = \cos\left(2\pi \frac{hour}{24}\right) \tag{5.3}$$

Test II compares the different results for the features seen as a cyclic input or as a categorical input.

In the blue box there is only one feature, namely the ticketing lead time. Ticketing lead time is a linear input however, test III compares the linear input with the ticketing lead time grouped in categories. These categories are based on the options which are now in use for the flight pass. These options are defined as follows:

7. TLT:

   - 1 day before travel date
   - 2 days before travel date
   - 3 days before travel date
   - 4 days before travel date
   - 5 days before travel date
   - 6-7 days before travel date
   - 7-14 days before travel date
   - 14-30 days before travel date
   - 30-60 days before travel date
   - > 60 days before travel date

Different tests are performed to check the best performance of the combination of grouped features. Results are compared when values for a feature are grouped.

- Months are grouped into seasons.
- Day of week is grouped into weekend and weekday.
- Length of stay is grouped by longer or shorter than 7 days.

Test IV compares the features grouped as described above with the non grouped features.

5.7. k-means model

As discussed, the reason for clustering the routes is to analyse groups instead of the 46 routes separately.

The k-means cluster has parameters that have to be tuned: (1) Method for initialisation; (2) Number of clusters; (3) Number of time the k-means algorithm will be run with different centroid seeds and (4) The maximum of iterations

1. Method for initialisation: k-means++, random
2. Number of clusters: 1,2,3,4,5
3. Number of initialisation: 100, 200, 500
4. The maximum iterations: 200, 500, 1000

The result of the hyper-parameter tuning of the k-means models is shown in Table 5.3.

Table 5.3: Result of hyper-parameter tuning k-means clustering

Hyperparameter	**Value**
Method for initialisation	k-means ++
Number of clusters	5
Number of initialisation	500
The maximum iterations	500

6

Results and Discussion

After running all models, results were simulated. This chapter will present and discuss these results. First, test results are presented in Chapter 6.1. The final specific flight prediction results are presented and discussed in Chapter 6.2. Finally, the output of the best scoring model is used to cluster the routes so a meaningful analysis of the results can be performed. The output of the cluster model is described and discussed in Chapter 6.4.

6.1. Feature tests

In order to determine which features need to be used for the different models, the models are evaluated using the evaluation metrics: the fit, the root mean squared and the absolute error. Every combination of route and class is modelled, however only the economy results are shown in this chapter.

To summarise, an overview of the different tests is given.

(i) Test whether Sunday stay or length of stay should be used in the prediction for the multiple linear regression

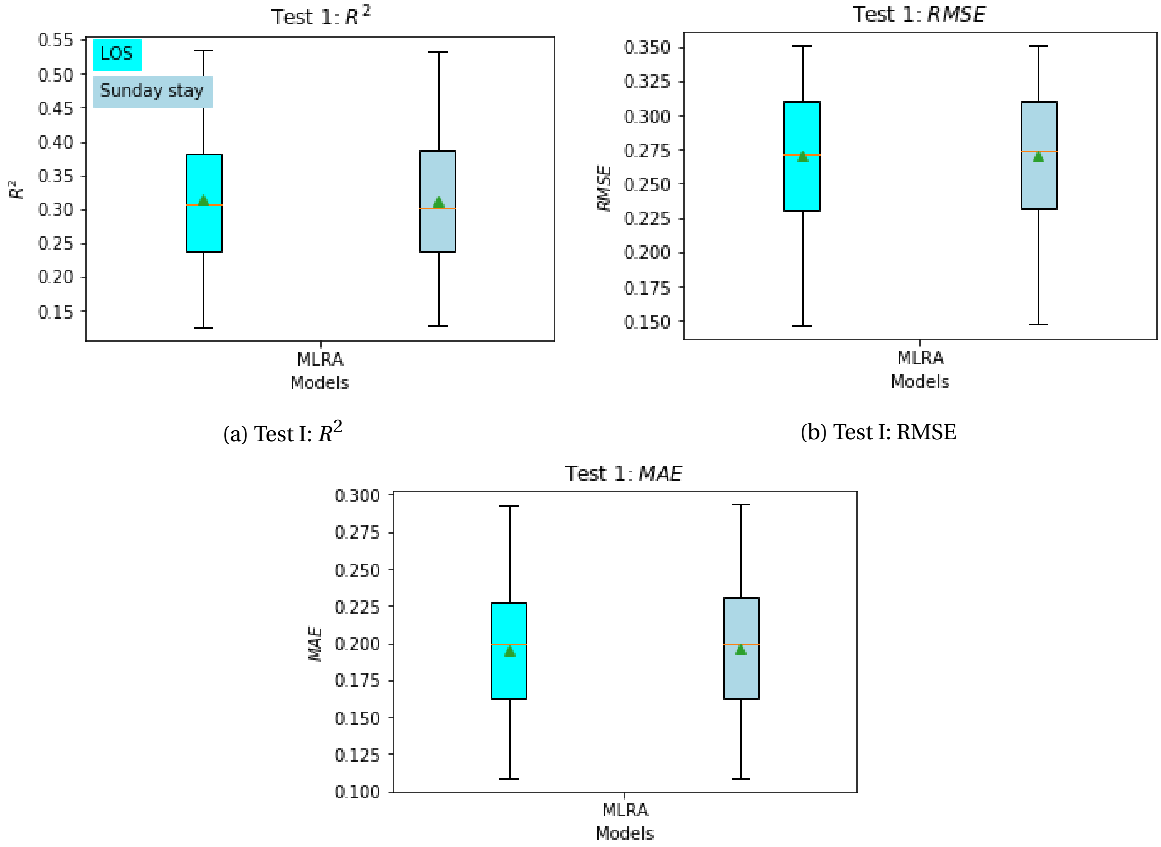
(ii) Test whether day of week, Month and time of day should be used as cyclic or categorical inputs

(iii) Test whether ticketing lead time should be used as continuous or categorical input

(iv) Test whether Months, day of week and LOS should be grouped or not

Test Ⅰ⃝ is only performed using the multiple linear regression while test Ⅱ⃝, Ⅲ⃝ and Ⅳ⃝ are performed using the multiple linear regression, random forest regression and neural network.

Test I Test Ⓘ tests whether Sunday stay or length of stay gives the best prediction. The complete result set for every route in economy can be found in Appendix F. The spread of the results of the evaluation metrics is shown in Figures 6.1a, 6.1b and 6.1c.



(a) Test I: R^2

(b) Test I: RMSE

(c) Test I: MAE

Figure 6.1: Evaluation metrics for test I

The box-plot shows the spread of the result set. The length of stay is shown in cyan and Sunday stay is shown in light-blue. The green triangle represents the mean of the result set and the red line represents the median. RMSE and MAE are errors, therefore the better performing model shows smaller values. R^2 represents the fit, which means that the better performing model shows bigger values.

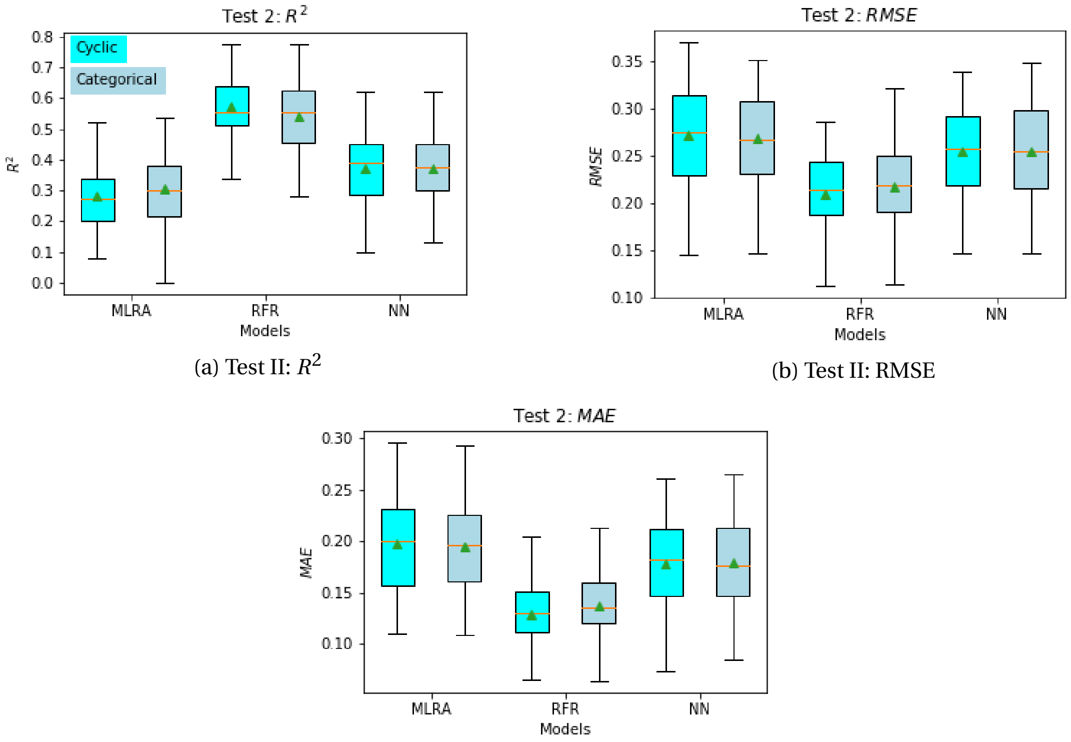
As one can see, the results for using length of stay or Sunday stay are almost identical. Length of stay performs slightly better compared to Sunday stay, nevertheless the difference is almost negligible.

As can be seen in Table 6.1, the mean and median of the fit (R^2) is better for length of stay compared to Sunday stay. The same observation can be made for the mean and median of RMSE and MAE were both of the mean and medium show smaller errors for length of Stay compared to Sunday stay. Therefore, the length of stay is used in the prediction when using the multiple linear regression model.

Table 6.1: Summary of test I

	Length of stay			**Sunday stay**		
	R^2	*RMSE*	*MAE*	R^2	*RMSE*	*MAE*
Mean	0,315	0,269	0,195	0,311	0,27	0,197
Median	0,308	0,272	0,199	0,302	0,274	0,200

Test II Test (II) refers to the test where circular and categorical features are compared for time of day, day of week and month. This is tested for the three different models (multiple linear regression, random forest regression and neural network). The complete result set can be found in Appendix G. The spread of the results of the evaluation metrics is shown in Figures 6.2a, 6.2b and 6.2c.



(a) Test II: R^2

(b) Test II: RMSE

(c) Test II: MAE

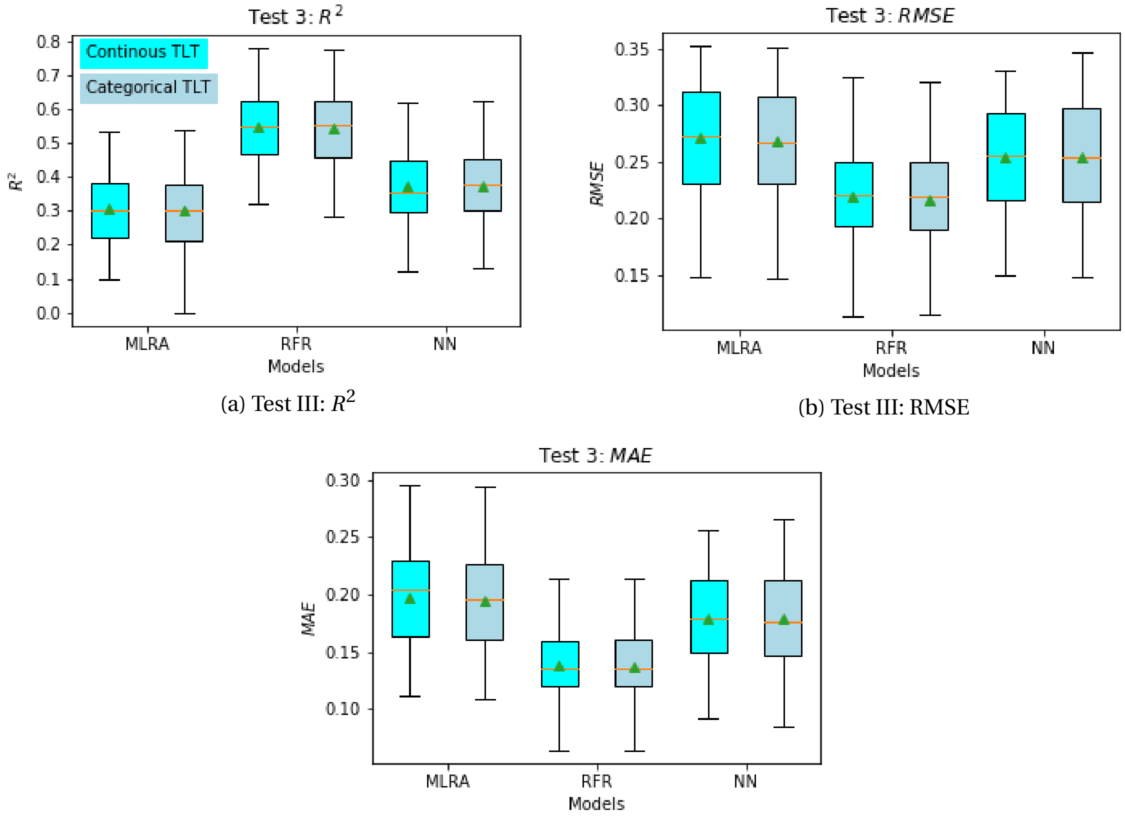
Figure 6.2: Evaluation metrics for test II

As one can see, the results for using circular or categorical inputs differ depending on the prediction model. For the multiple linear regression, the categorical inputs have a better fit while the fit for the random forest regression is significantly better with circular inputs. Negative values for R^2 are transformed to 0. When a negative value occurs, the mean of the data provides a better fit to the outcomes than do the fitted function values. The errors for the multiple linear regression model are the lowest with categorical inputs and the random forest regression model has the lowest errors using circular inputs. The results for the fit for the neural network are quite identical when comparing the mean, but the categorical inputs perform slightly better but almost negligible. When comparing the median, the circular performs better than the categorical. However when taking into account the errors, as well the mean as median of both the mean absolute and the root square error are higher for the circular inputs. From test (II), it can be concluded that the random forest regression will use circular inputs and the multiple linear regression and neural network will use the categorical inputs for: day of week, time of day and month. The categories are described in Chapter 5.6. The summary of the final results of the evaluation metrics for the three different models is shown in Table 6.2.

Table 6.2: Summary of test II

		Circular			Categorical		
		R^2	RMSE	MAE	R^2	RMSE	MAE
Mean	MLR	0,281	0,272	0,197	0,304	0,268	0,194
	RFR	0,574	0,209	0,129	0,543	0,217	0,137
	NN	0,369	0,253	0,179	0,370	0,254	0,178
Median	MLR	0,272	0,275	0,201	0,300	0,267	0,196
	RFR	0,557	0,215	0,130	0,553	0,219	0,135
	NN	0,389	0,258	0,183	0,375	0,255	0,177

Test III Test III refers to the test where a continuous and a categorical ticketing lead time are compared. The complete result set can be found in Appendix H. The spread of the results of the evaluation metrics is shown in Figures 6.3a, 6.3b and 6.3c.



(a) Test III: R^2

(b) Test III: RMSE

(c) Test III: MAE

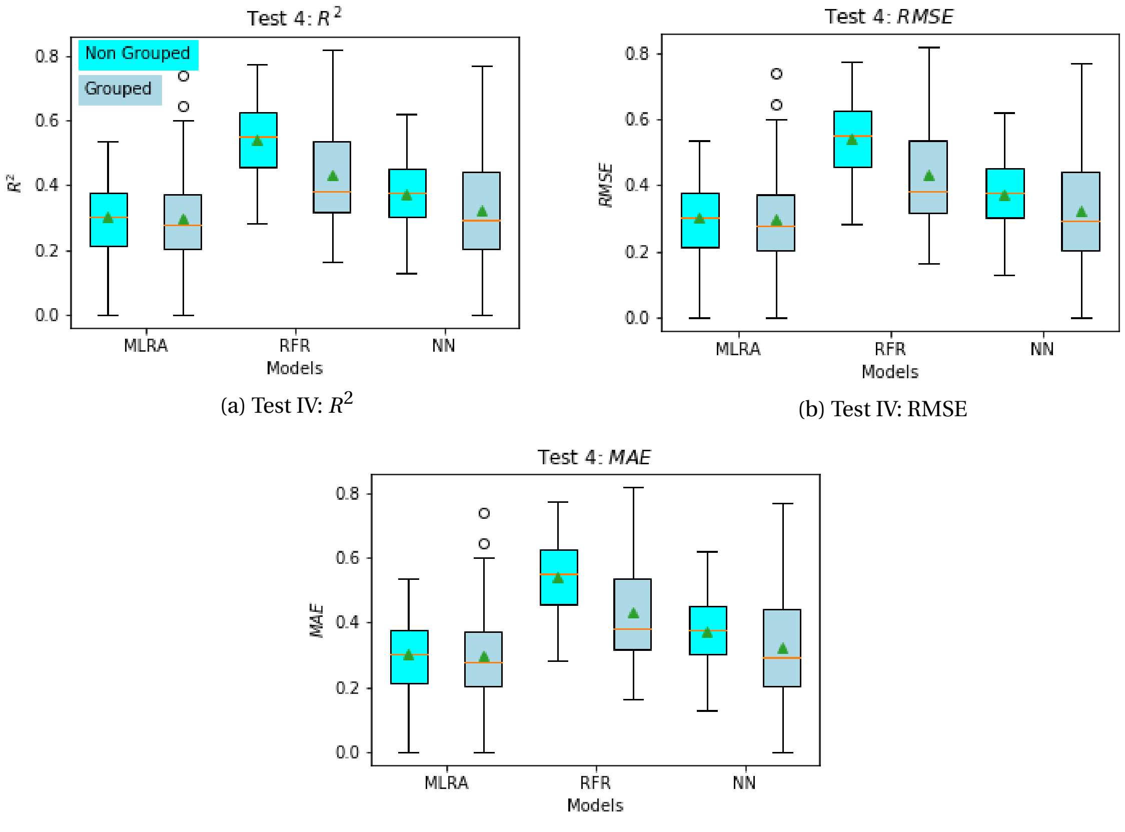
Figure 6.3: Evaluation metrics for test III

The summary of the final results of the evaluation metrics for the three different models is shown in Table 6.3. The results are again different, depending on the model. In general, the choice for using a continuous or categorical ticketing lead time is depending on very subtle differences. The results of multiple linear regression show exactly the same fit but the errors are bigger for a continuous ticketing lead time, therefore a categorical ticketing lead time performs better. The results of the random forest regression show a better fit but bigger errors for a continuous ticketing lead time input. Finally, the fit of the neural network has the same mean but a better median for a categorical input of ticketing lead time and the errors are smaller or the same for both the mean and the median. It can be concluded that the multiple linear regression and neural network will use a categorical input and the random forest regression will use a continuous input.

Table 6.3: Summary of test III

		Continuous			Categorical		
		R^2	RMSE	MAE	R^2	RMSE	MAE
Mean	MLR	0,304	0,271	0,197	0,304	0,268	0,194
	RFR	0,546	0,219	0,137	0,543	0,217	0,137
	NN	0,370	0,254	0,179	0,370	0,254	0,178
Median	MLR	0,300	0,273	0,205	0,300	0,267	0,196
	RFR	0,554	0,221	0,136	0,553	0,219	0,135
	NN	0,353	0,256	0,179	0,375	0,255	0,177

Test IV Test (IV) refers to the test whether months should be grouped into season, day of week in weekday or weekend and length of stay in larger or smaller than a week. The complete result set can be found in Appendix I. The spread of the results of the evaluation metrics is shown in Figures 6.4a, 6.4b and 6.4c.



(a) Test IV: R^2

(b) Test IV: RMSE

(c) Test IV: MAE

Figure 6.4: Evaluation metrics for test IV

The summary of the final results of the evaluation metrics for the three different models is shown in Table 6.4. The results are the same for as well multiple linear regression, random forest regression as neural network. Grouping parameters does not increase the prediction ability of any of the models. It can be concluded that all models perform better when no parameter is grouped.

Table 6.4: Summary of test IV

		Grouped			Non Grouped		
		R^2	RMSE	MAE	R^2	RMSE	MAE
Mean	MLR	0,296	0,271	0,198	0,304	0,268	0,194
	RFR	0,433	0,239	0,164	0,543	0,217	0,137
	NN	0,323	0,264	0,187	0,370	0,254	0,178
Median	MLR	0,275	0,268	0,196	0,300	0,267	0,196
	RFR	0,385	0,228	0,157	0,553	0,219	0,135
	NN	0,294	0,261	0,182	0,375	0,255	0,177

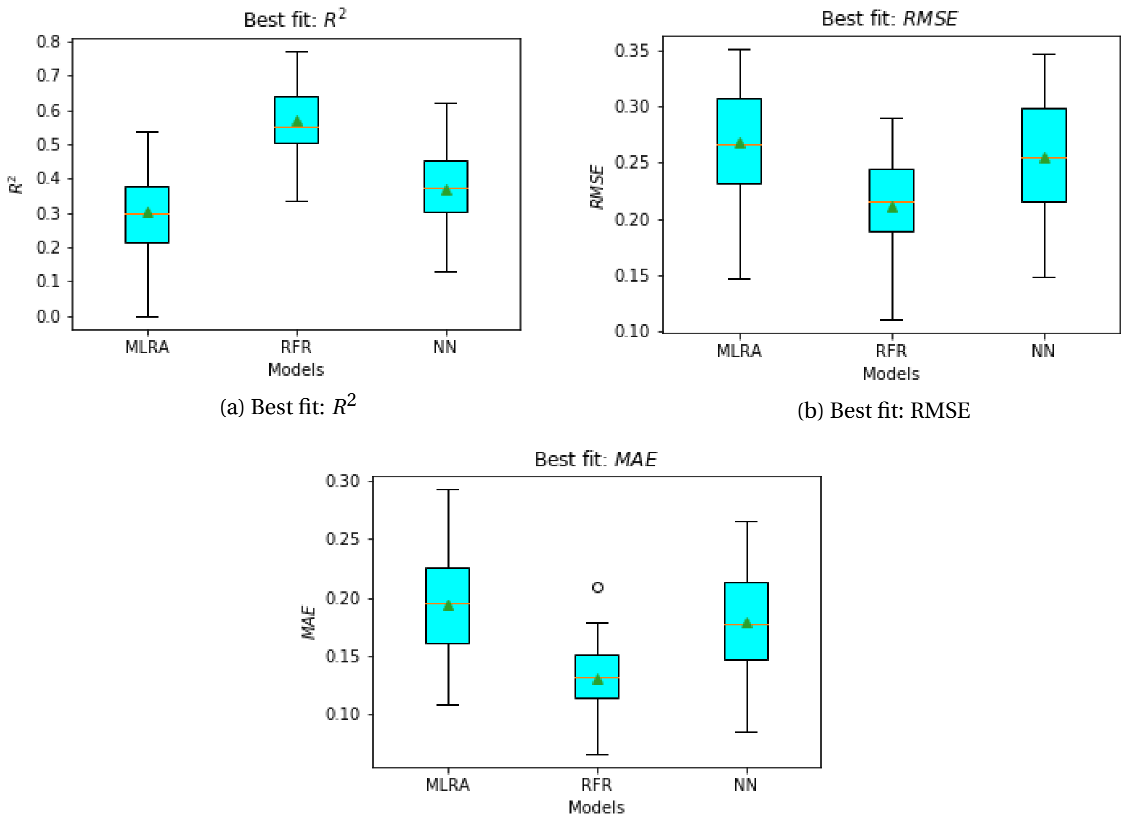
6.2. Final prediction results

To make a final prediction, the inputs for the three model, presented in Table 6.5, are used.

Table 6.5: Input features

	MLR	**RFR**	**NN**
POS	Categorical		
LOS			
Sunday stay	—	Categorical	
TOD	Categorical	Cyclic	Categorical
DOW			
Month			
TLT		Continuous	

For each of the 3 models, the results of the best fit are presented in Figures 6.5a, 6.5b and 6.5c.



(a) Best fit: R^2

(b) Best fit: RMSE

(c) Best fit: MAE

Figure 6.5: Evaluation metrics for best fit

The summary of the final results of the evaluation metrics for the three different models is shown in Table 6.6.

Table 6.6: Summary of final results

		R^2	RMSE	MAE
Mean	MLR	0,304	0,268	0,194
	RFR	0,569	0,210	0,130
	NN	0,370	0,254	0,179
Median	MLR	0,300	0,267	0,196
	RFR	0,553	0,216	0,132
	NN	0,375	0,255	0,177

As one can see, **random forest regression** performs better than the **multiple linear regression** and the **neural network**. The multiple linear regression can predict 30,4% of prices correctly. The neural network performs better than the multiple linear regression and can predict on average 37,0% of the prices. However, the random forest regression significantly outperforms the other two models with a fit of 56,9%. There are no guidelines for a 'good' R^2 but taking into account the human input in the determination of prices and the dependency on demand, one could say that an expected 'good' fit for this problem would have R^2 values between 0,4 and 0,8. (Chapter 4.5.1) The prediction results when using the random forest regression are for most routes within this range. Therefore, one can conclude that the random forest regression is a suitable prediction model for this problem.

RMSE gives high weights for bigger errors which is especially helpful if big errors are undesirable. As can be seen, random forest regression has less and smaller errors than the other two models. The MAE of the random forest regression is significantly lower than the other two models.

One of the reasons that the random forest regression outperforms the neural network can be that there was not enough data to train the neural network since it needs significantly more data than the random forest regression. When data-sets are quite small and very structured the chances are high that a random forest regression model outperforms the neural network. Appendix J shows the results for every route separately, one could see that for every route the random forest regression performs best.

Another reason that the random forest regression outperforms the neural network can be the fact that a random forest performs very well in high dimensions when comparing to the neural network [31].

It can be concluded that the data behaves non linear and that multiple linear regression is not the most suitable method to predict flight prices. Next, the random forest regression significantly outperforms the neural network, therefore the random forest regression predicts the flight prices best compared to the multiple linear regression and neural network. The choice was made to go only further with the results of the random forest regression for further analysis of the features and routes.

6.3. Feature importance

Next to the prediction, the random forest regression shows the importance of the features in predicting the price. The 7 features used for prediction: (1) TLT; (2) Month; (3) DOW; (4) TOD; (5) POS; (6) LOS and (7) SUN are ordered based on their contribution in prediction the price.

Figure 6.6 shows the average feature importance of all routes. As can be seen, the addition of the different importance's adds to 1 in total. This means that on average ticketing lead time contributes to 26,2% of the price prediction, Month to 23,9%, day of week to 11,4% and so on. It can be concluded that the most important feature in predicting the ticket price is ticketing lead time.

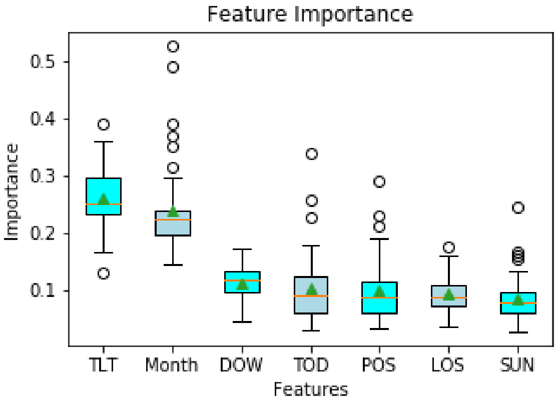


Figure 6.6: The average feature importance

6.4. Cluster model

To make an in depth analysis of the importance of the features for the different routes, the results are clustered. Instead of analysing 46 routes, 5 clusters representing different 'route types' are analysed. In this way the analysis is easier and more efficient.

The results of the random forest regression are used to cluster routes for the analysis of the routes and features. The feature importance for each route is used to cluster the different routes.

The k-means clustering algorithm is used to group the different routes. The algorithm is evaluated using the Silhouette Coefficient. The score can be between -1 and 1 where -1 indicates incorrect clusters and 1 perfectly dense and separable clusters. The result of the k-means clustering is 0,272 which indicates that clusters are identifiable but somehow slightly overlapping.

The decision was made to use 5 clusters since the Silhouette Coefficient grows with 5,2% from 4 to 5 clusters. When using 6 clusters the growth was only 1,4%.

Seven features are used for prediction: (1) POS; (2) LOS; (3) Sunday stay; (4) TOD; (5) DOW; (6) Month and (7) TLT. These features are not directly used to cluster the different routes

1. Point of Sale is not used as an input to cluster routes since the strategy behind the sale location can be very different and is difficult and undesirable to analyse.
2. Length of stay and Sunday stay can be seen as one feature and they are summed with each other since the impact of the two features is low and the two features are correlated.
3. Return is extracted out of length of stay and Sunday stay because this is an important feature to describe clusters.

This results in six new features: (1) Return (RET) ; (2) LOS+SUN; (3) TOD; (4) DOW; (5) Month and (6) TLT.

The results of the clusters are shown in Table 6.7 and displayed on the map. This map can be seen in Figure 6.7.

Table 6.7: Results of clustering

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
JED	EBB	BOM	BKO	DLA
TNR	KIS	DKR	BGF	NSI
HAH	MBA	FNA	BZV	FIH
ABJ		ACC	LAD	KGL
KRT		HKG	BLZ	ADD
		HAN	MPM	JRO
		BKK	SEZ	ZNZ
		DZA		DAR
		CPT		FBM
		LVI		NLA
		BJM		LUN
		JIB		HRE
		LOS		JNB
		JUB		LLW
				APL
				CAN
				DXB

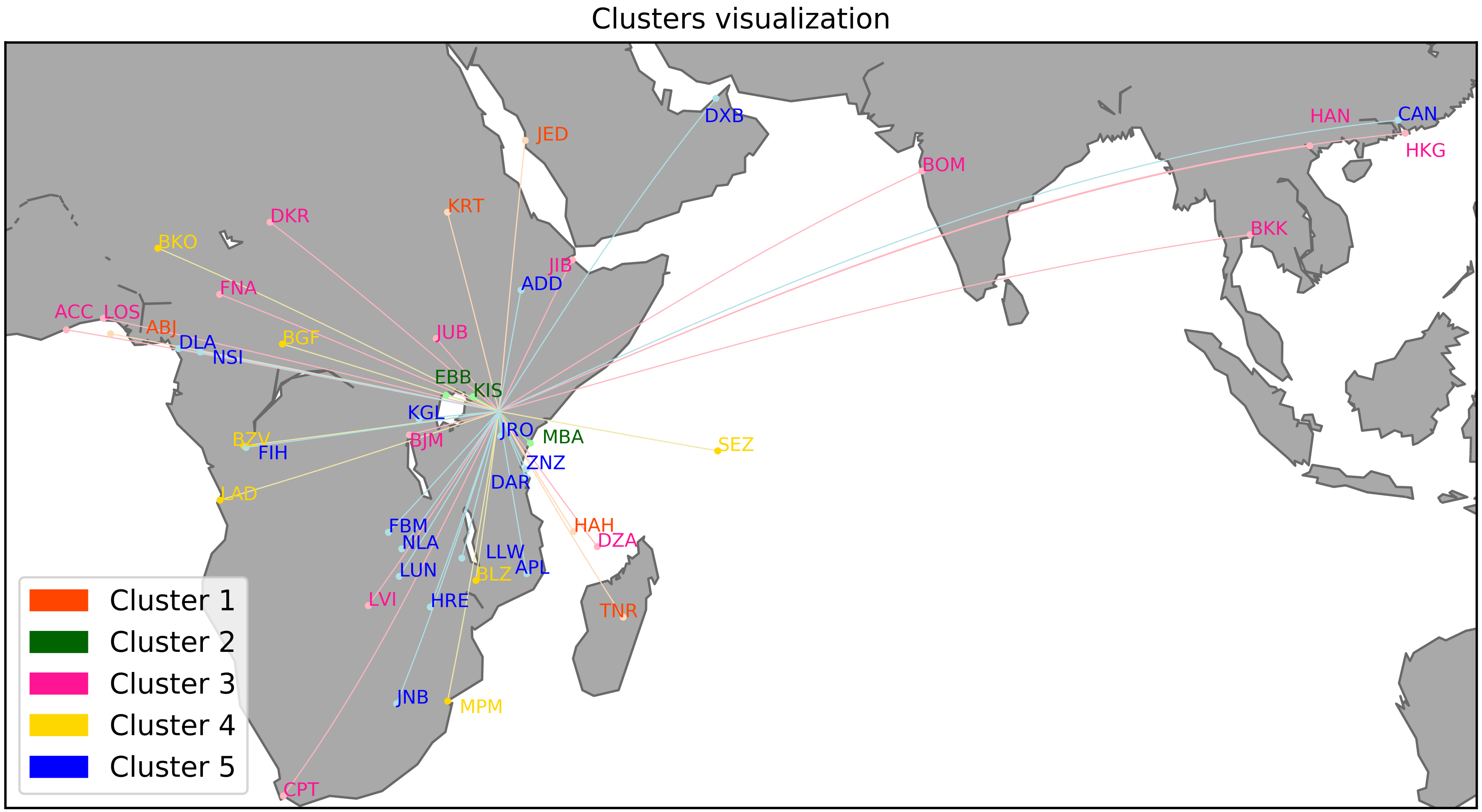


Figure 6.7: Clusters presented on the KQ network

Before going into the analysis of the clusters some important facts can be noted from the map shown in Figure 6.7.

1. EBB, KIS and NBA belong to one cluster and are very close to NBO (Cluster 2).
2. The destinations that are furthest like CPT, BOM, HAN, BKK, HKG, ACC and length of stay also belong to the same cluster while distance is never used as a criteria to cluster the routes. (Cluster 3)

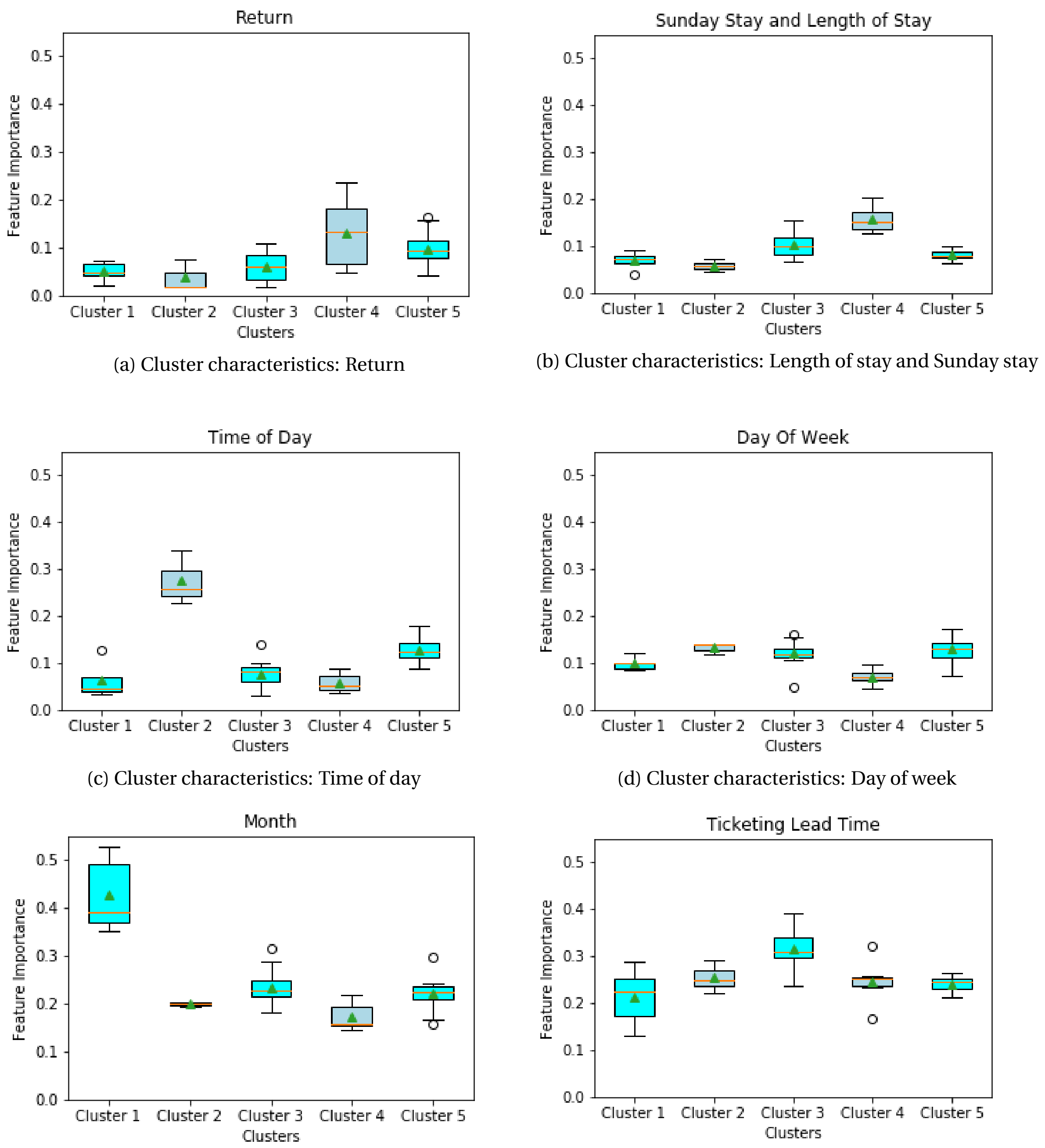
The five different clusters are compared in terms of the six features. The averages of the six features are shown in Table 6.8. Figure 6.8a shows the feature return flight, it shows that cluster 4 and 5 are above average. Sunday stay and length of stay are more important for cluster 3 and 4, as shown in Figure 6.8b. Figure 6.8c shows the importance of time of day, it shows that there is a significant difference of almost 20% between the average time of day importance and the importance of cluster 2. The day of week is illustrated in Figure 6.8d. Day of week is somehow constant however, cluster 4 has a relatively low importance of day of week. Cluster 1 has a significantly bigger, more than 20% importance in month than average, as can be seen in Figure 6.8e. Finally, Figure 6.8f represents the importance of the ticketing lead time. Ticketing lead time is on average the most important feature in predicting the price. Ticketing lead time is significantly more important for cluster 3.

Table 6.8: Summary of feature importance

	RET	SUN + LOS	TOD	DOW	Month	TLT
Mean	0,083	0,096	0,104	0,115	0,239	0,262
Median	0,075	0,085	0,092	0,118	0,224	0,253

The next step is describing the clusters with the information seen in Figure 6.8. Firstly, the meaning of these features is described and then summarised for each cluster.

- **RET:** When the importance of return is high, it means that the route price is strongly driven by the fact whether the passenger already knows when to fly back to the origin. Holidays are planned in advance and passengers book (mostly return tickets) while business passengers often do not know yet when to return or fly to another destination afterwards.
- **SUN + LOS:** When the importance of Sunday stay and length of stay is high, it means that the route price is strongly driven by the fact if someone stays a week or only a day and if there is a Sunday in a stay. For some routes it is hard to tell if the route is mainly business or leisure oriented therefore length and Sunday stay can help to segregate business from leisure passengers. When there is a high Sunday and length of stay importance there is a high need to segregate the business from leisure passengers.
- **TOD:** When the importance of time of day is high, it means that the route price is strongly driven by the time someone flies. If there is only 1 flight a day, the importance will be low but for high frequency routes the importance can grow if there are significant price differences between morning, afternoon, evening and night flights.
- **DOW:** When the importance of day of week is high, it means that the route price is strongly driven by which day to fly. This is the case if KQ want to make the flights more expensive on Friday and Monday to segregate business passengers from leisure passengers.
- **Month:** When the importance of month is high, it means that the route price is strongly driven by the season.
- **TLT:** When the importance of ticketing lead time is high, it means that the route price is strongly driven by when the ticket is bought and thus strongly dependent on demand. Ticketing lead time is an important feature in every case but for some routes more important. It can be because of price changes in the competition prices or demand changes over time.



(a) Cluster characteristics: Return

(b) Cluster characteristics: Length of stay and Sunday stay

(c) Cluster characteristics: Time of day

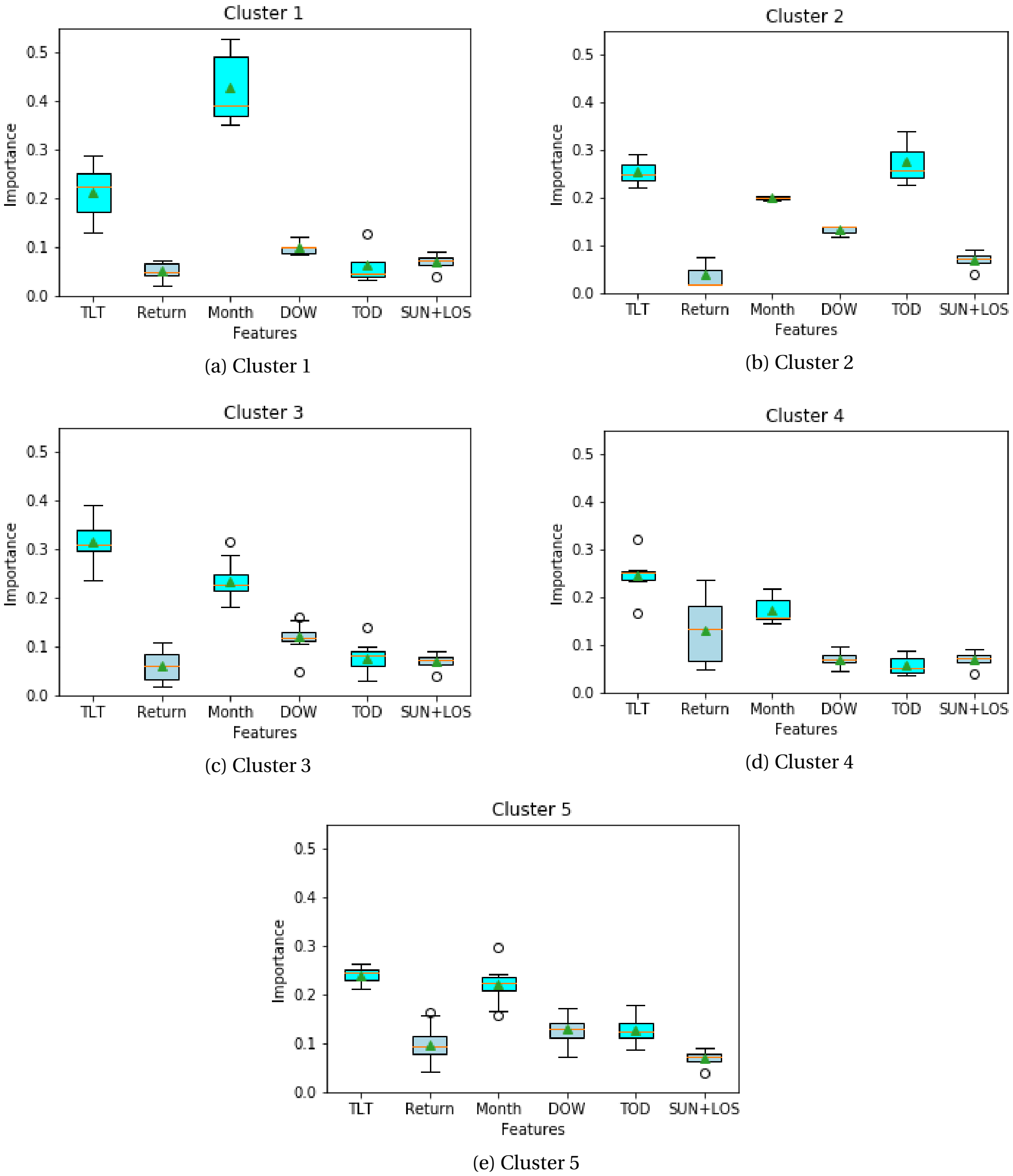
(d) Cluster characteristics: Day of week

(e) Cluster characteristics: Month

(f) Cluster characteristics: Ticketing lead time

Figure 6.8: Cluster characteristics

The next step is to describe the clusters in terms of feature importance. Figure 6.9 summarises the feature importance for each cluster. It can be noticed that cluster 3 and 5 show similar characteristics. However, ticketing lead time is around 7 percent more important in predicting the price in cluster 3 than 5 and one can see that the spread of return differs when comparing cluster 3 and 5.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

Figure 6.9: Characteristics for each cluster

The following step is to help describing the clusters. One can see what the important features are compared to the average for each cluster. For each cluster, the features are divided in five categories: (−−) less than 30% important than average; (−) between less than 30% and 15% important than average; (+/−) between 15% less and 15% more important than average; (+) between more than 15% and 30% important than average and (++) more than 30% important than average. The results are shown in Table 6.9.

Table 6.9: Relative feature importance for each cluster

	RET	SUN + LOS	TOD	DOW	Month	TLT
Cluster 1	- -	-	- -	+/-	++	-
Cluster 2	- -	- -	++	+	-	+/-
Cluster 3	-	+/-	-	+/-	+/-	+
Cluster 4	++	++	- -	- -	-	+/-
Cluster 5	+	-	+	+/-	+/-	+/-

- **Cluster 1:** The Month is important for this cluster which means that the prices are seasonally determined. Return and time of day are not important at all and all the others are relatively not so important. The routes JED, TNR, HAH, ABJ and KRT belong to this cluster. This means that seasonality is mainly driving the price for these routes.
- **Cluster 2:** The time of day is very important, the flights in this cluster are short, high frequency flights, meaning that there are price differences between an afternoon and morning flight. The day of the flight is quite important. This makes sense because of the short distance, a business passengers prefers to fly on Monday morning and not on Saturday evening. The short haul high frequency routes EBB, KIS and MBA belong to this cluster.
- **Cluster 3:** There are no extreme specifics for cluster 3 but it is the only cluster where the feature ticketing lead time is above average important. Return and time of day are below average. One of the reasons why ticketing lead time has a higher importance is due to the fact that the major part of these routes have competition of other airlines on the flight. When the competition drops their prices, KQ has to change their strategies. As this happens often, those strategies change constantly over time.

  Secondly, the routes in this cluster, are relatively far in distance from the main hub (Nairobi). Therefore, one could argue that leisure passengers book tickets in advance which means that ticketing lead time plays a bigger role than short flights. The routes belonging to this cluster are: BOM, DKR, FNA, ACC, HKG, HAN, BKK, DZA, CPT, LVI, BJM, JIB, LOS and JUB.
- **Cluster 4:** Cluster 4 is driven by the fact whether there is a return flight, Sunday stay and the length of stay. One-way passengers are much more likely to have to travel, and therefore they are tolerant to higher prices. Routes in this cluster are mainly business oriented. When a passengers stays for the weekend, the passenger is 'classified' as a leisure passenger and will get cheaper flights. The roues BKO, BGF, BZV, LAD, BLZ, MPM and SEZ belong to this cluster.
- **Cluster 5:** Time of day and return flight are relatively important while Sunday stay and length of stay do not influence the flight price. Most of these flights are scheduled more than once a day and could be described as short/medium haul flights. Therefore time of day is more important compared to cluster 1,3 or 4. DLA, NSI, FIH, KGL, ADD, JRO, ZNZ, DAR, FBM, NLA, LUN, HRE, JNB, LLW, APL, CAN and DXB belong to this cluster.

To conclude, the 5 different clusters represent 5 different route types. These clusters are used in the verification process. If the results of every cluster can be verified, one can assume that results of every route are verified.

At this moment, the individual flight ticket prices are known. As a passenger can book different specific individual flights within one flight pass, the flying behaviour needs to be simulated. This simulation is explained in the next chapter. By simulating the passengers behaviour a correct estimation can be made for the final flight pass price.

7

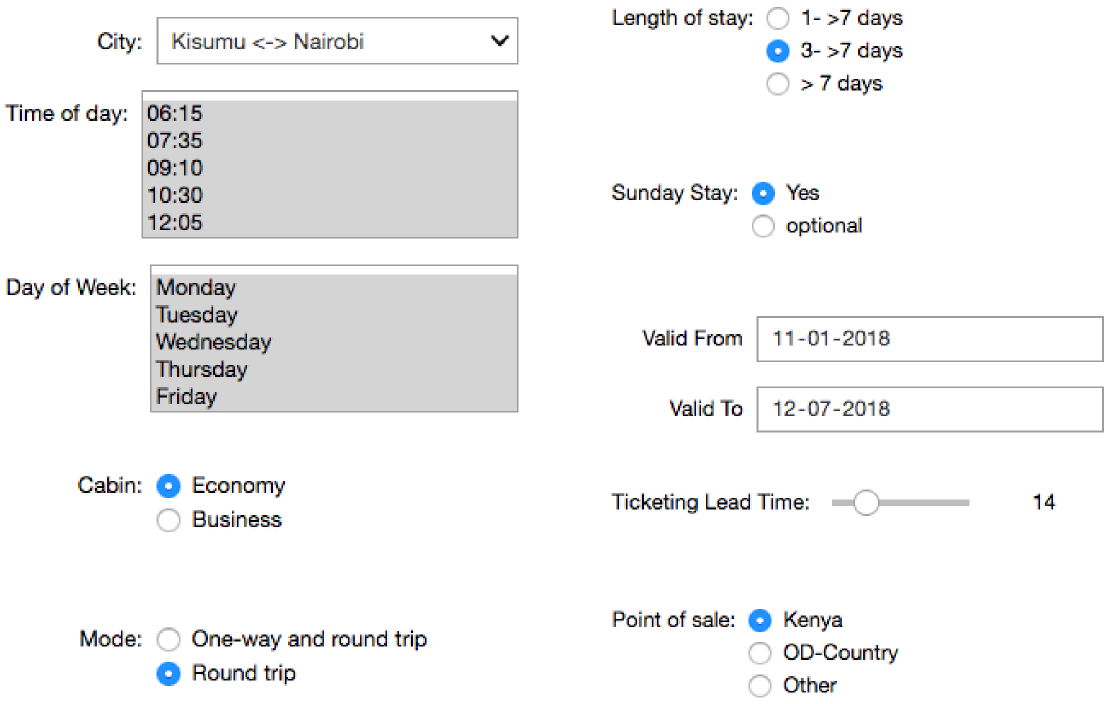
Simulation: Use Case

The result of the random forest regression is the prediction of the price of a specific flight. The next step is to predict the price of a flight pass. Since a customer can book different flights within one pass, the total flight pass value has to be estimated. This estimation is done by simulating the flying behaviour of a passenger. To support this simulation, a web application is built. This can be accessed via: `https://ec2-35-158-56-26.eu-central-1.compute.amazonaws.com:8888/notebooks/Run.ipynb?dashboard`

Chapter 7.1 is an introduction in the use of the self-made web application to visualise the model. The distributions of the input variables are discussed in Chapter 7.2. In Chapter 7.3, the results of the Monte-Carlo simulation are shown.

7.1. Introduction to the web application

Appendix K elaborates on how to prepare the website for use. When the tool is loaded the interface shown in Figure 7.1 should be visible.



City: Kisumu <-> Nairobi

Time of day:
06:15
07:35
09:10
10:30
12:05

Day of Week:
Monday
Tuesday
Wednesday
Thursday
Friday

Cabin: ● Economy ○ Business

Mode: ○ One-way and round trip ● Round trip

Length of stay: ○ 1- >7 days ● 3- >7 days ○ > 7 days

Sunday Stay: ● Yes ○ optional

Valid From 11-01-2018

Valid To 12-07-2018

Ticketing Lead Time: 14

Point of sale: ● Kenya ○ OD-Country ○ Other

Figure 7.1: Website interface

When a customer buys a flight pass the inputs (1) route; (2) Flight(s); (3) Day of week(s); (4) Cabin; (5) Round trip; (6) Travel period (7) Ticketing lead time and (8) Point of sale should be selected. When round trip is selected, the customer can choose for a (9) Length of stay and (10) Sunday stay.

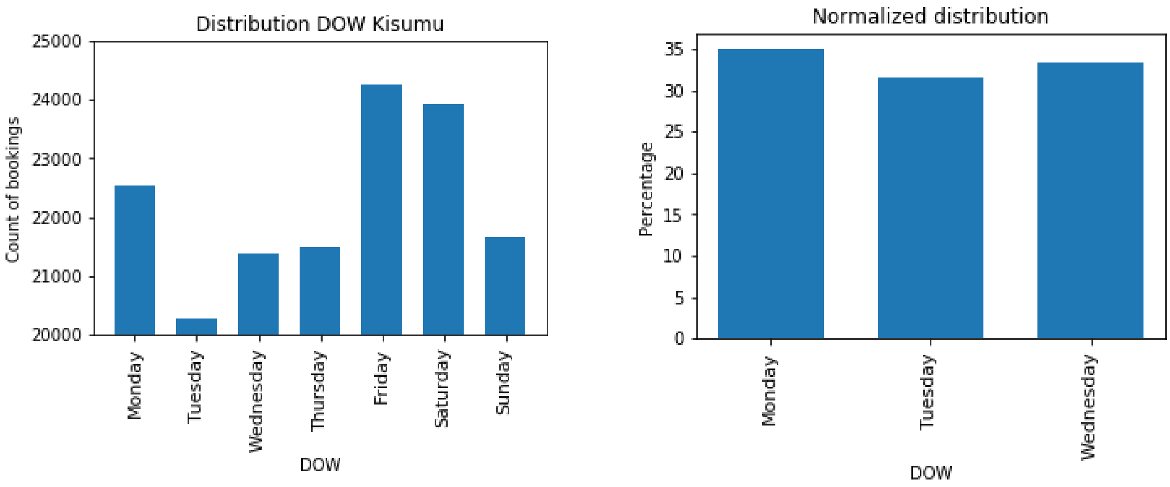
With the example visible in Figure 7.1, the flight pass is valid for:

1. The route: Nairobi-Kisumu
2. Every flight on every time of day
3. Every flight on every day of week
4. Economy class
5. Only round trips
6. Six months between January 11 and July 10
7. The flights need to be reserved at least 14 days in advance
8. The flight pass should be bought in Kenya
9. The length of stay is at least 3 days
10. There is a Sunday in the trip

7.2. Assumptions on the input distributions

A prediction is made for every possible flight that can occur within these bounds of the flight pass. Some assumptions are made regarding the input variables and the flying behaviour of the passengers.

- **City**: A customer can only choose one city to fly from/to. The city determines together with the cabin which model variation of the random forest is used to predict the price of the single specific flight.
- **Day of Week**: A customer can select on which day of week the flight pass should be valid. When a customer selects to fly on Monday, Tuesday and Wednesday, a distribution which reflects the probability of flying on one of these days is made. This distribution is deviated from the raw booking data. Depending on the route, the amount of bookings for each day is normalised. A non normalised distribution for all day of weeks for Kisumu is shown in Figure 7.2a. When this is normalised depending on the days selected, it looks like Figure 7.2b.



(a) Distribution of DOW, KIS

(b) Normalised distribution, KIS

Figure 7.2: Distribution DOW: KIS

- **Time of day**: The same strategy as day of week is applied. A customer can select on which flights (The time of departure) the flight pas should be valid. When a customer selects more flights on different time of day, a distribution is deviated from the flying behaviour presented in the raw booking data.
- **Valid from to**: The same strategy as day of week and time of day is applied. A customer can choose to use the flight pass from a certain date to another date. When choosing a validity of 6 months from January till July, the normalised distribution from January till July is used.
- **Cabin**: When Economy is selected the prediction model variation of the random forest regression for economy and the city selected is used to predict the value of a single flight.
- **Mode**: When one way and round trip is selected only one way flights are used to simulate the customers behaviour, while a customer could use it for return flights too. This assumption is made because it is hard to assume whether a customer would use it for return flights and if so how many times. Therefore the decision was made to simulate the passenger in the most price strict behaviour. When a round trip is chosen, the flight pass is only valid for a round trip booking. To simulate the customer behaviour, only round trips are used.
- **Length of stay**: Length of stay will only appear when a round trip is selected. There are 3 options: (1) At least 7 days; (2) At least 3 days and (3) At least 1 day. When option 1 is selected, all lengths of stay with more than 7 days are used. When option 2 is selected, only flights with length of stay between 3 and 7 days are used and when option 3 is selected only flights with length of stay between 1 and 3 days are used to simulate the customers behaviour. This assumption is (again) assuming the most strict behaviour and is the consequence of the assumption that when customers choose option 1, they will probably stay between 1 till 3 days.
- **Sunday stay**: Sunday stay will only appear when a round trip and option 1 or 2 in length of stay are selected. When 'Sunday stay optional' is selected, only flights where there is no Sunday stay are used to simulate the behaviour of the customers. When selecting Sunday stay yes, only flights with a Sunday stay are used to simulate the customer behaviour.
- **Ticketing Lead Time**: The options for ticketing lead time go from 1 day before departure till 60 days before departure. When the option 14 days before departure is selected, a customer has to book at least 14 days before departure but 60 days is also possible. After discussing this with an expert, the decision was made to assume that the customer will book the ticket 14 days before departure is set to 50%, every day thereafter the probability is divided by 2. This means that for 14 days the probability is 50%, 15 days before departure, the probability at 14 days is 50%, 15 days before departure the probability drops to 25%, 16 days before departure the probability drops to 12,5% and so on. The distribution is shown in Figure 7.3.

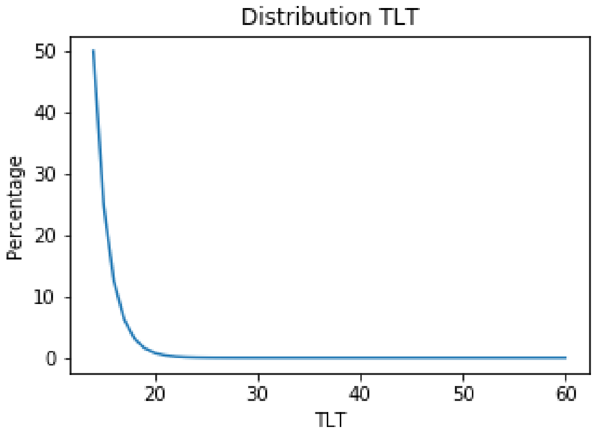


Figure 7.3: Example distribution of TLT, 14 days

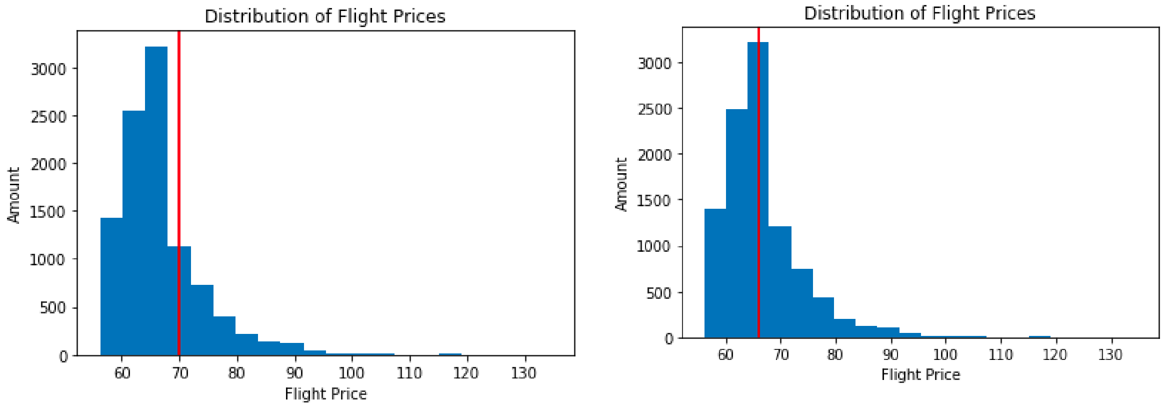
Since this assumption is based on hard numbers, it is wise to analyse how sensitive the model is to these numbers. Therefore a sensitivity analysis is performed in Appendix L.

- **Point of sale**: This option is not available when booking a flight pass but should be implemented in the website and should be selected automatically depending on the country where the flight pass is bought.

7.3. Monte-Carlo simulation

A Monte-Carlo simulation is used to simulate the behaviour of a passenger taking into account the distribution of the input variables discussed above.

The prediction of the price of these 10 000 flights is plotted in a histogram. The price predictions of the example shown in Figure 7.2a, are illustrated in Figure 7.4a. The y-axis represents the amount of times the prices occur. The x-axis represents the price of the flight. The red line is the price of a flight which is at 80 percent of the 10 000 flights. In this example, the red line, and thus the value of a single flight with the selected bounds for Kisumu, is 70 euros. The 80% is an assumption made with an expert but it is possible to change the red line when the business wants to. This can be done by selecting the change percentage button, when the industry wants the price at 50%, the weighted price for a single flight is 66 euros. This is shown in Figure 7.4b.



(a) Weighted price per flight for the example shown in Figure 7.1, 80%

(b) Weighted price per flight for the example shown in Figure 7.1, 50%

Figure 7.4: Results of Monte-Carlo simulation

The final result of the simulation is the weighted single flight price for the specific options/bounds selected. After doing the simulation several times with the same input variables, one can note that the weighted single flight price will not change.

The simulation made, still has some limitations. These limitations are listed below:

- **The number of flights**: The main goal of the flight pass is to increase fidelity by offering more flights for a discount. At this moment, the customer can choose between 6 and 500 flights when using the Optiontown model. The thesis simulation does not take into account the number of flights since the discount given for the amount of flights is a business decision which is not included in the model/simulation.

- **The amount of passengers**: The maximum number of passengers who can book and fly using the flight pass can variate from 1 to 200 people when using the Optiontown model. The thesis model/simulation does not take into account the number of passengers that use the same flight pass. This is again a business decision that has to be made.

- **Validity time**: The airline can give extra discount when the validity of the pass is only a month compared to a year. The model does not take into account the validity of the pass as a variable to determine the price. This is again a business decision that has to be made.

- **Discount based on passenger use**: The airline can decide to give an extra discount assuming that the customer will not always fully use the flight pass. Sometimes a customer will use 8 out of the 10 flights. The model/simulation does not take this effect into account.
- **Strategic discounts**: The airline can give some extra discounts in some routes. If there is a competitor on a price sensitive route, an airline can decide to make the flight pass cheaper than the equivalent normal ticket price to gain customers for the flight pass from the competitive airline. This is again a business decision and is not implemented in the simulation.

8

Verification and Validation

After simulating the flight pass use, the model can be verified and validated. The verification process is described in Chapter 8.1. This verification process is done by example with the use of the simulation tool. The validation process is described in Chapter 8.2, where the results of the simulation tool are compared to the current flight pass prices of Optiontown.

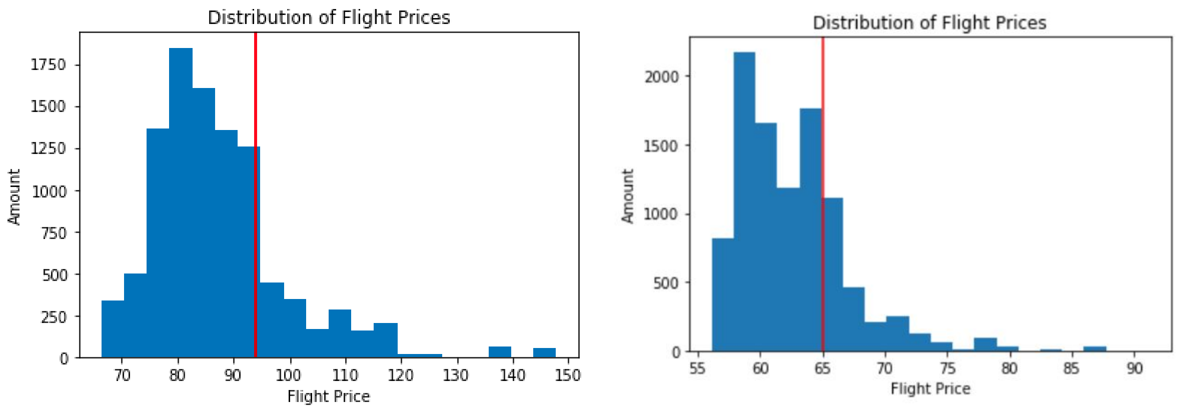
8.1. Verification

The model is verified by example. Clustering the routes helped in organising the structure of the verification process. Five different routes belonging to the five different clusters are verified. To verify the model some test are performed and the expectations of the results are described. The tests are done by comparing two different values from the same feature and reason if the difference makes sense. The standard input parameters used in the examples are the following:

- Every month
- Every time of day
- Every day of week
- One way and round trip
- Ticketing lead time of 14
- Valid from 11/01/2018 till 10-07-2018
- Point of sale is Kenya

8.1.1. Kisumu - TOD

The first example is the route Kisumu. Kisumu belongs to cluster 2, where time of day mainly defines the price of the flight. (See Figure 6.9b) Therefore, when only a peak time flight is selected, the flight should be more expensive than when an off peak flight is selected. The flight of 18:20 is defined as peak flight and the flight of 10:30 as off peak flight. The result is shown in Figure 8.1a and Figure 8.1b.



(a) KIS: Result of peak flight

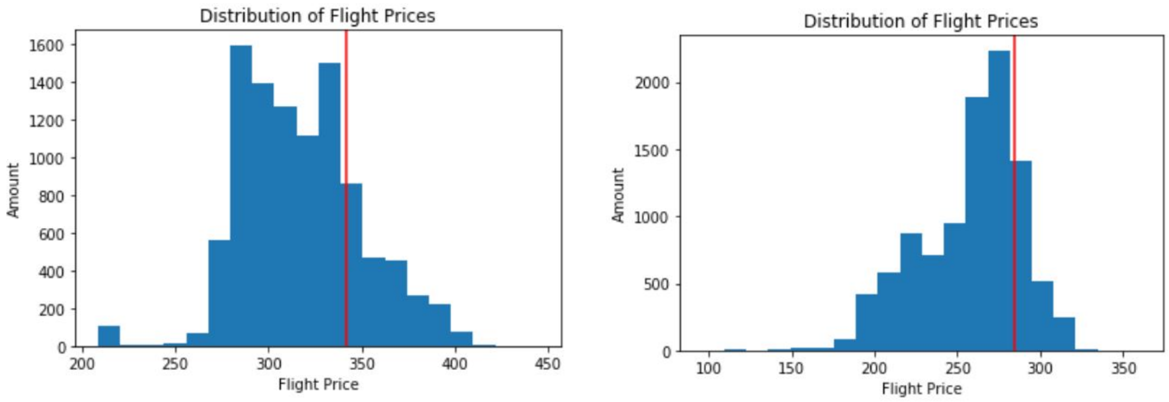
(b) KIS: Result of off-peak flight

Figure 8.1: Verification TOD: KIS

As shown in the Figures 8.1a and 8.1b, the hypothesis is correct and the tickets are more expensive for the peak flight. The price of the peak flight is on average 94 euros whereas the price of the off-peak flight is on average 71 euros.

8.1.2. Mahe Island - Return

The next example is one from a different cluster, namely Mahe Island on the Seychelles. This destination belongs to cluster 4, where the feature return ticket is relatively important. (See Figure 6.9d) Therefore, when a one way trip is selected, the flight should be more expensive than when a return trip is selected. The result is shown in Figure 8.2a and Figure 8.2b.



(a) SEZ: Result of one-way

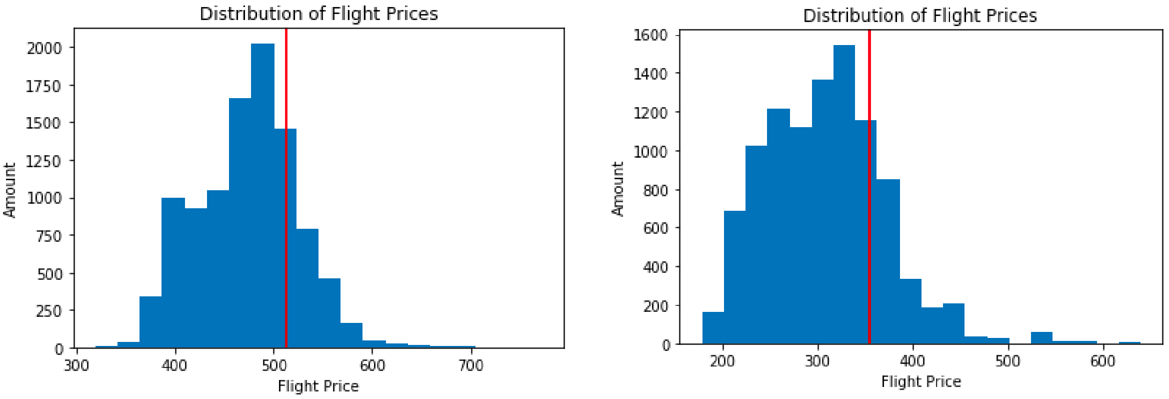
(b) SEZ: Result of return flight

Figure 8.2: Verification mode: SEZ

As shown in the Figures 8.2a and 8.2b, the hypothesis is correct and the tickets are more expensive for the one way trip. The price for the one way trip is on average 352 euros whereas the price of a return trip is on average 290 euros.

8.1.3. Mumbai - TLT

The third example is an example from cluster 3, namely Mumbai, where ticketing lead time is the most important feature to determine the flight price. (See Figure 6.9c) Therefore, when ticketing lead time of 1 day is selected, the flight should be more expensive than when a ticketing lead time of 60 days is selected. The result is shown in Figure 8.3a and Figure 8.3b.



(a) BOM: Result of TLT 1 day before departure

(b) BOM: Result of TLT 60 days before departure

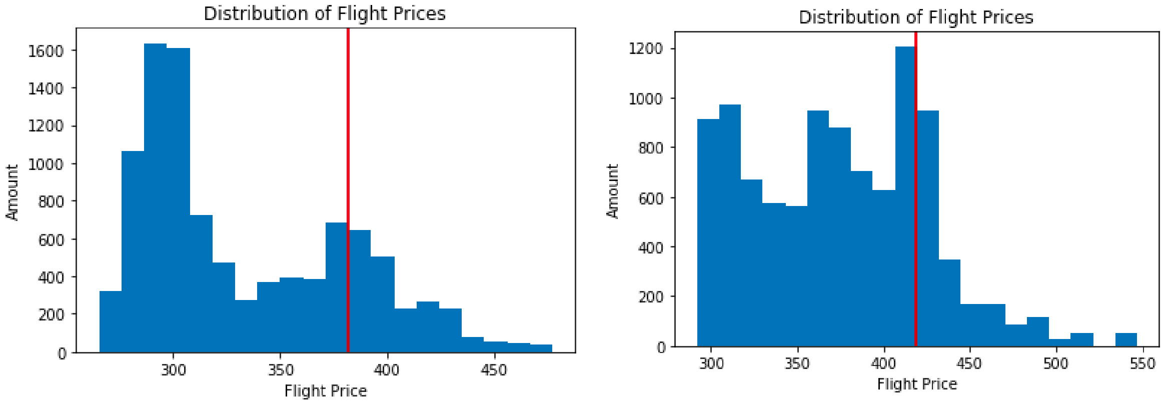
Figure 8.3: Verification TLT: BOM

As shown in the Figures 8.3a and 8.3b, the hypothesis is correct and the tickets are more expensive for a ticketing lead time of 1 day. The price for trip booked 1 day in advance is 513 euros whereas the price for a trip booked 60 days in advance is 355 euros.

8.1.4. Khartoum - Month

The fourth example is an example from cluster 1, namely Khartoum, where month of flying is relatively the most important feature to determine the flight price. (See Figure 6.9a)

Therefore, when high season months are selected, the flight should be more expensive than when low season months are selected. The months: December, January and February are defined as low season flights and the flights in September, October and November are defined as high season flights. The result is shown in Figure 8.4a and Figure 8.4b.



(a) KRT: Result of months: Dec, Jan and Feb

(b) KRT: Result of months: Sep, Oct, Nov

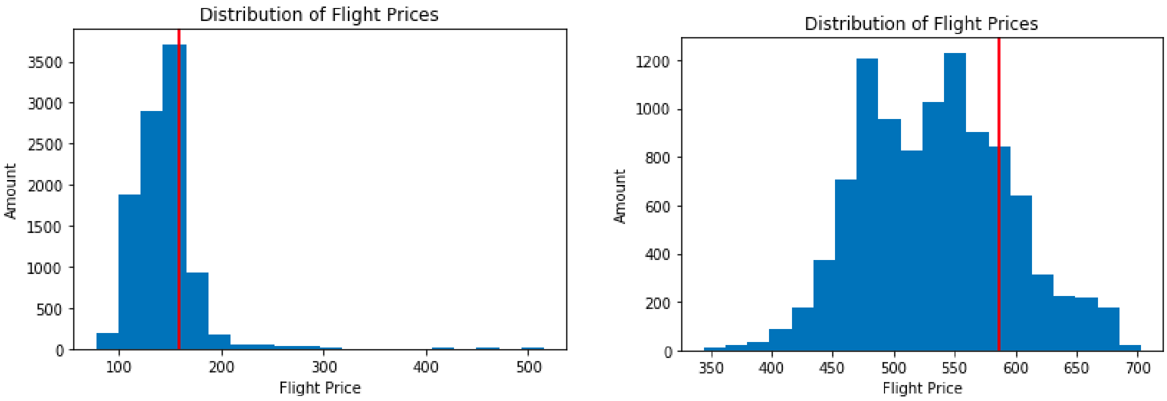
Figure 8.4: Verification months: KRT

As shown in the Figures 8.4a and 8.4b, the hypothesis is correct and the tickets are more expensive for high season. The price for high season flights is around 419 euros whereas the price for flight in low season is around 382 euros.

8.1.5. Addis Abeba - Cabin

The last example is an example from cluster 5, namely Addis Abeba, where nothing in particular is relatively important. (See Figure 6.9e)

Therefore, the prices of business and economy flights are compared. There is expected that a business flight is around 4 times as expensive as an economy flight. The result is shown in Figure 8.5a and 8.5b.



(a) ADD: Result of economy

(b) ADD: Result of business

Figure 8.5: Verification cabin: ADD

As shown in the Figures 8.5a and 8.5b, the hypothesis is correct and the tickets are more expensive for a business flight. The business flights are 3.7 times as expensive as the economy flights. The price for a business flight is around 586 euros whereas the price for an economy flight is around 159 euros.

8.2. Validation

To validate the results of the random forest regression and the simulation tool, the prices are compared to the prices now available on the Optiontown website. Two models can be distinguished.

- The model proposed in the thesis with its limitations of the number of flights, the amount of passengers, the validity, discount based on passenger use and strategic discounts. This model will be referred to as **Thesis model**.
- The model available on the Optiontown website with the limitation of linearity, same model for every route and the missing input variables like time of day, day of week, month, length of stay and Sunday stay. This model will be referred to as **Optiontown model**

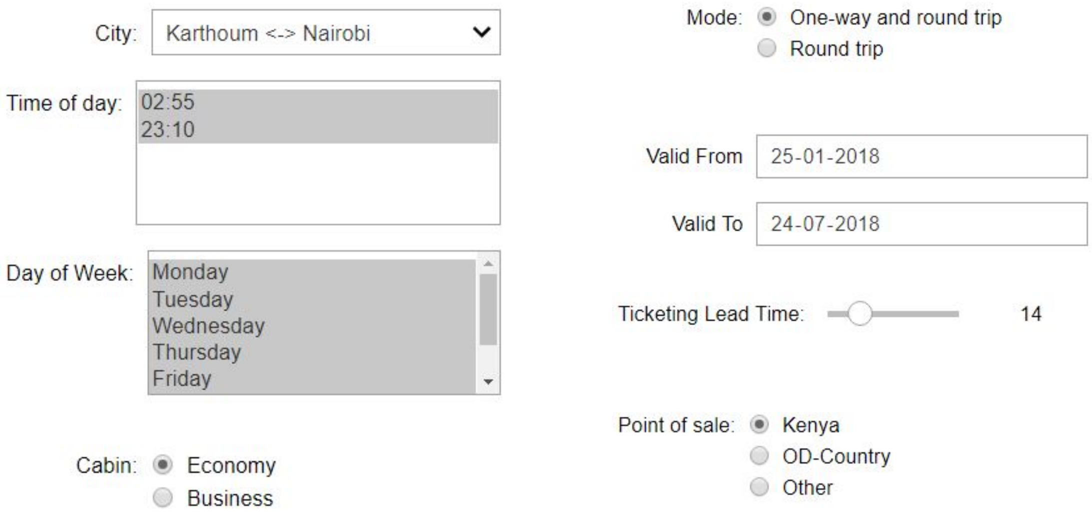
To validate the results, only the thesis model and Optiontown model can be compared in prices. To compare those prices some examples are chosen. These prices can be accessed via the Optiontown website: `https://kenya-airways.optiontown.com/)`

Since the Optiontown model, is the same model for every route and the prices are not determined on a data-driven basis, it is hard to validate the obtained results of the thesis model. But it is possible to check whether the results of the thesis model are in the same range as the prices of the Optiontown model. This is done by means of an example. The Optiontown model already gives discounts for the total number of flights that can be booked with the flight pass. The expectation is that the price of the Optiontown model will be on average a bit lower than the price determined by the thesis model.

Before diving into the validation, the Optiontown model is explained in detail.

- The price differences are in percentage the same for every route when changing the same parameter.
- Every parameter change is independent of the other parameters. For example: the price difference of ticketing lead time 1 and 14 days will be the same for a return or a one way trip.

First, the difference in input variables is discussed. Figure 8.6 shows the input variables for the thesis model and Figure 8.7 shows the input variables for the Optiontown model.



City: Karthoum <-> Nairobi

Time of day: 02:55, 23:10

Day of Week: Monday, Tuesday, Wednesday, Thursday, Friday

Cabin: ● Economy ○ Business

Mode: ● One-way and round trip ○ Round trip

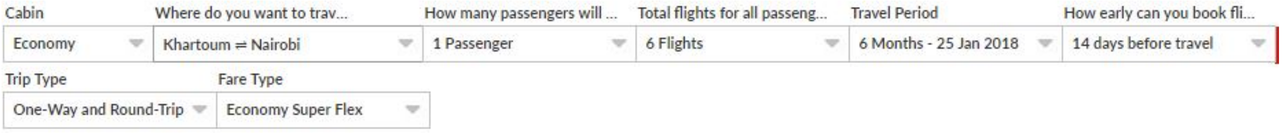
Valid From: 25-01-2018

Valid To: 24-07-2018

Ticketing Lead Time: 14

Point of sale: ● Kenya ○ OD-Country ○ Other

Figure 8.6: Validation: Prices of the thesis model



Cabin	Where do you want to trav...	How many passengers will ...	Total flights for all passeng...	Travel Period	How early can you book fli...
Economy	Khartoum ⇌ Nairobi	1 Passenger	6 Flights	6 Months - 25 Jan 2018	14 days before travel

Trip Type	Fare Type
One-Way and Round-Trip	Economy Super Flex

Figure 8.7: Validation: Prices of Optiontown

- The Optiontown model does not give the possibility to choose a time of day or a day of week. The assumption was made to select all times of day and all days of week in the thesis model.
- The Optiontown model does not give the possibility to choose the point of sale. The assumption was made that the point of sale is Kenya
- The website of Optiontown gives the option to choose the number of passengers that can use the pass. The assumption was made that the number of passengers that can use the pass is 1.
- The Optiontown model gives the option to choose the total number of flights that can be booked with the pass. The minimum that can be selected is 6 flights, therefore 6 flights are assumed.
- The Optiontown model gives the option to choose a fare type, there a four options: (1) Best buy; (2) Economy; (3) Economy Flex and (4) Economy super flex. The assumption was made that the economy super flex is the fare type, this because every class is taken into account in the thesis model.

The prices for the flight pass are compared with the inputs above. The result for the thesis model is 403 euros while the result of the Optiontown model is 373 euros. This is definitely in the range considering that there are no extra discounts in the thesis model for the amount of flights given.

To validate further the inputs are independently changed to:

1. Business
2. Round trip
3. Ticketing lead time of 1 day of 60 days
4. Ticketing lead time of 1 day
5. Mumbai

The results are shown in Table 8.1. When prices are within 15% bounds, one could say that the prices are in the same range as the prices of the Optiontown model.

Table 8.1: The results of the validation of the inputs shown in Figure 8.6 and 8.7

	Thesis model [€]	**Optiontown model[€]**	**Percentage difference**
Standard	403	373	7,7 %
Business	953	822	14,7%
Round trip	248	319	25%
TLT 60 days	298	335	11,7%
TLT 1 day	425	547	25,1%

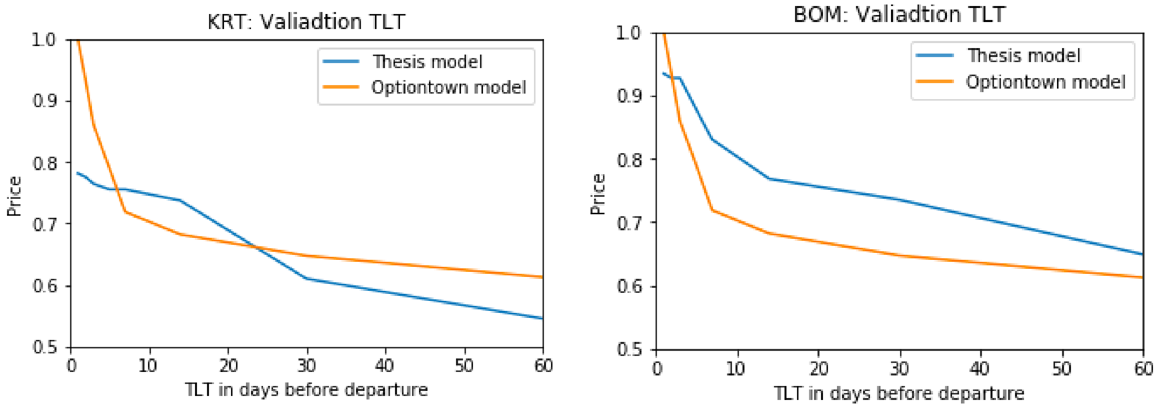
As discussed, the price difference for the standard inputs, shown in Figure 8.6 and 8.7, is within the 15% bound which is defined as reasonable. For a business flight the difference is still proportional and within the 15 % bounds.

For a round trip the difference is quite remarkable, since one would expect that the Optiontown model would be cheaper and the price difference is more than 15%. The result of the Optiontown model is not in line with the thesis model.

The ticketing lead time is again somehow out of bounds, as well as the price of the ticket for 60 days as 1 day before departure is cheaper than the Optiontown model while the price for booking 14 days in advance is more expensive. It would be interesting to see how the different curves behave. It is important to note that the Optiontown model uses a ticketing lead time curve which behaves the same for every route while the curve of the thesis model differs for each route. Figure 8.8a shows the price for the different ticketing lead time values. The prices are scaled so one can compare different routes, 1 on the y-axis corresponds to the price for the specific route 1 day before departure.

As shown in Figure 8.8a, the flight price of the Optiontown model (orange line), is 1 to 5 days before departure significantly higher and from 5 days onward, the curve drops gently. The small differences between 1 and 60 days for the thesis model (blue line) corresponds to the characteristics of cluster 1, where Khartoum (KRT) is a part of.

It would be interesting to check the ticketing lead time for a route of cluster 3 (ticketing lead time more important). Figure 8.8b shows the ticketing lead time behaviour for Mumbai (BOM).



(a) Khartoum (KRT)

(b) Mumbai (BOM)

Figure 8.8: Validation: TLT curve from 1 to 60 days for the Optiontown model (orange) and the thesis model (blue)

As shown in Figure 8.8b, the ticketing lead time curve from the Optiontown model has the same shape for Khartoum (KRT) and Mumbai (BOM), while the shape of the curve of the thesis model is completely different. This corresponds to the characteristics of cluster 3, where ticketing lead time is the most important feature.

The final validation is done by checking the price differences of the Optiontown model and the thesis model for each cluster. The same routes are chosen as for the verification process. The same definition is used for reasonable bounds. The standard inputs, shown in Figure 8.6 and 8.7 are used for the following cities: KIS, SEZ, BOM, KRT and ADD. The results are shown in Table 8.2.

Table 8.2: The results of the validation comparing clusters

	Thesis model [€]	**Optiontown model[€]**	**Percentage difference**
KIS	74	75	1,3%
SEZ	349	389	10,8%
BOM	420	335	22,5%
KRT	403	373	7,7%
ADD	159	183	14,0%

As shown in Table 8.2, the price differences differ strongly dependent on the route. The price difference for Kisumu is negligible. The prices for Mahe Island (SEZ) are 10,8% more expensive with the Optiontown model. This is defined as reasonable bounds but still remarkable that the Optiontown model prices are more expensive. The prices for Mumbai (BOM) are more expensive when using the thesis model, which would be expected but the price difference is defined as 22,5% which is significantly big. The price difference for Khartoum (KRT) is already discussed and the price difference for Addis Abeba (ADD) is within reasonable bounds (14%) but again the Optiontown model is more expensive which is not expected.

There are some possible explanations for the remarkable differences.

- The assumptions made regarding the flying behaviour are not the representing the flying behaviour in real life.
- It is possible that the Optiontown model already offered extra strategic discounts on some routes.

To conclude, it is hard to validate the results since the prices of the Optiontown model are not based on a data-driven model and therefore it is hard to estimate the ’correctness’ of the Optiontown model prices. The first recommendation would be to validate the thesis model with flight pass data, when this data is available.

9 Conclusion

An overview of the conclusions is presented in Chapter 9.1. The limitations and recommendations for future work are covered in Chapter 9.2. Finally, a list of practical recommendations for Kenya Airways is provided in Chapter 9.3.

9.1. Conclusions and research contributions

The requirement from the industry was to avoid dilution. This requirement has been translated to a model where additional RM booking data is used to predict the value of a flight.

To conclude the research, the sub-questions are separately answered below.

Sub question 1: *What features to select to predict the price of the flight pass?*

Seven features are used to predict the flight pass price for each cabin and route combination: (1) Month of flying; (2) Day of week of flying; (3) Time of day of flying; (4) Length of stay; (5) Sunday stay; (6) Ticketing lead time and (7) Point of sale.

Sub question 2: *Which model to use to predict the flight pass prices?*

Three models are compared and the random forest regression significantly performs better than the multiple linear regression and the multilayer perceptron neural network. With a fit varying from 34 percent till 77 percent and an average of 57 percent the model can adequately predict what the flight price will be. It is hard to determine what is a good fit but taking into account how the price of a ticket is normally determined with pricing and inventory control (see Chapter 2.1), it is quite remarkable that one static model can mimic both the effects of pricing and inventory control and predict the correct flight price in, on average 57% of the cases.

Sub question 3: *How to cluster the different routes depending on the importance of the features?*

A k-means clustering algorithm is used to cluster the 46 different routes. Five clusters are made based on the importance of the above discussed features.

Sub question 4: *How to simulate the use of the flight pass?*

The goal of the research is to support the pricing of the flight pass based on data driven machine learning models. It can be concluded that the price of a single ticket can be predicted in an adequately manner but this does not mean that the flight pass is predicted in a fair way.

Therefore, to answer the fourth sub-question a simulation is built where the use of the flight pass is simulated. The results are accessible via the web application. The model together with the simulation of the use of the flight pass is verified and validated together with an expert in the field. Although no specific data is available to proof this, it can be stated that, with the industry experience of the expert, the simulation gives a fair representation of the real life scenario.

As a surplus the model gives insight in the importance of features in predicting the price, route dependent. This can be beneficial for the flight pass pricing as well as for revenue management.

It can be concluded that with the model, Kenya Airways can start with offering flight pass prices that are based on data driven machine learning models.

The fact that an easier model could eventually replace the RM process is a proof of concept of high scientifically value. At this moment a lot of research goes to the demand forecasting which is the driving input for inventory control. With this research one could say that, with the huge amount of RM data available, the time has come to start researching the use of predictive machine learning models in revenue management.

9.2. Limitations and recommendations for future research

This chapter provides a discussion of the limitations of the proposed modelling frameworks. Some of the limitations could be easily translated into recommendations.

Limitations and assumptions

1. The main limitation of the model is that it is not possible to use flight pass data to validate the model due to the limited amount of data.
2. Only one year of revenue management data is used to train the model.
3. At this moment every route is calibrated individual for the optimal hyper-parameters but one could change the input parameters accordingly to the best fit. The results of the different tests had different results dependent on route but at this moment the best average test results are used.
4. The flight pass can only be used for one route. When a passenger frequently flies two routes, two different passes need to be bought.
5. The neural network has only 1 layer.
6. To simulate the flying behaviour, assumptions are made regarding the probabilities of the flying behaviour. These assumptions are listed in Chapter 7.
7. The model does not take into account:
    (a) The number of flights that can be booked with one pass: The amount of passengers
    (b) The maximum number of passengers who can book and fly using the flight pass
    (c) The validity (time period) of the pass
    (d) Strategic discounts
    (e) Discounts based on not fully using the pass

Recommendations

1. Validate the results with flight pass data. The expectation is that this is possible after one year.
2. Train the models with data gathered over multiple years to be able to extract the effect of seasons.
3. The results of the tests were different depending on the route. The optimal input parameters could be chosen for each route independent.
4. It would be interesting to make a model which can predict a price for a flight pass, which can be used for more routes.

5. Tuning a neural network requires knowledge from an expert. Given the time, only networks with one layer are considered in the tuning process. It would be recommended that an expert in neural networks tunes the model again with more layers. This expert could also look into different architectures of neural networks.

6. Machine learning is a promising field to replace the complete process of revenue management. This was a proof of concept where prediction is only based on 7 features but when someone would investigate more features, results would probably even be more promising. It would be interesting to add more features to predict prices of flights. One could think about connecting flights, competition on routes, load factor, etc...

9.3. Practical recommendations for Kenya Airways

In this chapter practical recommendations for Kenya Airways are given. If the system of Optiontown allows a non linear model, the RF model would be beneficial for Kenya Airways. If the systems does not offer the option for a non linear model the following practical changes are recommended:

- Make ’month’ a parameter to predict the prices for the following routes: ABJ, TNR, KRT, JED and HAH.
- Change the curve for TLT depending on the route.
- Add TOD as parameter for the following routes: KIS, MBA and EBB.
- RM could benefit from the feature importance known for every route.
- Validate the flying behaviour and especially the buying behaviour as soon as possible with the flight pass data.

A

Appendix - (Sub)classes of KQ

Table A.1: Order of booking classes (High to Low)

Economy	Business
Y	J
B	C
M	D
U	I
K	Z
H	O*
L	
Q	
T	
R	
E	
N	
V	
W	
X*	
G**	

* Frequent Flyer Program
** Group Booking

B

Appendix - Airport Codes

Code	City	Country	Airport
NBO	Nairobi	Kenya	Jomo Kenyatta International Airport
BOM	Mumbai	India	Chhatrapati Shivaji International Airport
SEZ	Mahé	Seychelles	"Seychelles International Airport
DZA	Dzaoudzi	Mayotte	Dzaoudzi–Pamandzi International Airport
HAH	Hahaya,	Comoros	Prince Said Ibrahim International Airport
TNR	Antananarivo	Madagascar	Ivato International Airport
DXB	Dubai,	United Arab Emirates	Dubai International Airport
JED	Jeddah	Saudi Arabia	King Abdulaziz International Airport
KRT	Khartoum	Sudan	Khartoum International Airport
JUB	Juba,	South Sudan	Juba International Airport
ADD	Addis Ababa	Ethiopia	"Addis Ababa Bole International Airport
JIB	Ambouli	Djibouti	Djibouti–Ambouli International Airport
EBB	Entebbe	Uganda	Entebbe International Airport
HRE	Harare	Zimbabwe	Robert Gabriel Mugabe International Airport
LUN	Lusaka	Zambia	Kenneth Kaunda International Airport
JRO	Hai District	Tanzania	Kilimanjaro International Airport
ZNZ	Stone Town, Zanzibar	Tanzania	Abeid Amani Karume International Airport
DAR	Dar es Salaam	Tanzania	Julius Nyerere International Airport
BJM	Bujumbura	Burundi	Bujumbura international airport
KGL	Kigali	Rwanda	Kigali international airport
ACC	Accra	Ghana	Kotoka international airport
FNA	Freetown	Sierra Leone	Lungi international airport
BKO	Bamako	Mali	Bamako–Sénou International Airport
DKR	Dakar	Senegal	Leopold Sedar Senghor International airport
ABJ	Port-Bouët	Côte d'Ivoire	Port Bouet Airport
DLA	Douala	Cameroon	Douala internaltional airport
NSI	Yaoundé	Cameroon	Yaoundé Nsimalen International Airport
LOS	Lagos	Nigeria	Murtala Muhammed International Airport
BZV	Brazzaville	Republic of the Congo	Maya Maya
FIH	Kinshasa	Democratic republic of the Congo	N'djili Airport
BGF	Bangui	Central African Republic	Bangui M'Poko International Airport
FBM	Lubumbashi	Democratic Republic of the Congo	Lubumbashi International Airport
NLA	Ndola	Zambia	Simon Mwansa Kapwepwe International Airport
MBA	Mombasa	Kenya	Moi international airport
KIS	Kisumu	Kenya	Kisumu International Airport
LLW	Lilongwe	Malawi	Kamuzu International Airport
APL	Nampula	Mozambique	Nampula Airport
MPM	Maputo	Mozambique	Maputo International Airport
BLZ	Blantyre	Malawi	Chileka International Airport
JNB	Johannesburg	South Africa	Johannesburg
LAD	Luanda	Angola	"Quatro de Fevereiro Airport
CPT	Cape Town	South Africa	Cape Town International Airport
LVI	Livingstone	Zambia	Harry Mwanga Nkumbula International Airport
BKK	Bangkok	Thailand	Suvarnabhumi Airport
HKG	Hong Kong	Hong Kong	Hong Kong International airport
CAN	Guangzhou	China	Guangzhou Baiyun International Airport
HAN	Hanoi	Vietnam	Noi Bai International Airport

C

Appendix - Samples for each route

Table C.1: Sample size

Code	Sample size
MBA	299532
KIS	163230
DAR	99653
EBB	87514
JNB	44578
JUB	37318
BOM	33189
DXB	32932
ADD	28580
KGL	27585
BJM	15522
ACC	15229
HRE	13510
LUN	11380
LOS	10201
KRT	9657
ZNZ	9424
LLW	7943
SEZ	7373
TNR	7093
JIB	6443
JRO	6080
MPM	5677
CPT	4110
DKR	3293
ABJ	2992
FIH	2866
HAH	2724
JED	2327
LVI	1839
NLA	1594
DLA	1575
BKK	1522
BKO	1463
BLZ	1385
FBM	1374
NSI	1352
BGF	1303
CAN	1137
APL	903
BZV	874
DZA	872
FNA	672
LAD	659
HKG	418

D

Appendix - Best hyper-parameters for the random forest regression

Table D.1: Best hyper-parameters for the random forest regression

Code	# of trees	of features	depth	samples required in a split	samples required in a leaf
BOM	200	sqrt	25	5	2
SEZ	200	sqrt	15	4	2
DZA	100	sqrt	20	5	2
HAH	100	log2	25	6	4
TNR	100	sqrt	25	6	2
DXB	100	log2	none	5	1
JED	100	sqrt	20	2	1
KRT	200	auto	15	6	2
JUB	100	sqrt	15	4	1
ADD	100	sqrt	15	2	2
JIB	100	log2	15	4	1
EBB	200	sqrt	20	4	2
HRE	100	auto	15	5	2
LUN	100	sqrt	15	4	1
JRO	200	auto	25	4	1
ZNZ	200	auto	20	4	2
DAR	200	sqrt	20	4	1
BJM	100	sqrt	20	2	1
KGL	100	sqrt	15	2	1
ACC	100	auto	15	6	2
FNA	100	sqrt	20	6	2
BKO	100	auto	15	5	1
DKR	100	log2	15	6	1
ABJ	100	sqrt	15	6	2
DLA	100	sqrt	20	6	2
NSI	100	log2	20	4	2
LOS	100	sqrt	25	5	2
BZV	100	log2	20	4	2
FIH	100	log2	15	5	2
BGF	100	sqrt	20	5	2
FBM	100	sqrt	20	4	2
NLA	100	log2	15	5	2
MBA	200	sqrt	25	4	1
KIS	200	log2	25	4	2
LLW	100	sqrt	25	6	2
APL	100	auto	none	4	2
MPM	200	sqrt	15	5	2
BLZ	100	sqrt	35	4	1
JNB	200	log2	15	2	1
LAD	100	sqrt	25	2	1
CPT	200	sqrt	20	4	1
LVI	200	sqrt	25	5	2
BKK	100	sqrt	none	4	2
HKG	100	sqrt	20	2	1
CAN	100	sqrt	25	5	2
HAN	100	log2	none	4	1

E

Appendix - Best hyper-parameters for the neural network

Table E.1: Best hyper-parameters for the neural network

Code	# of neurons	Solver	Activation function	Learning rate	Iteration
BOM	15	sgd	identity	adaptive	100
SEZ	15	sgd	logistic	adaptive	200
DZA	5	lbfgs	logistic	invscaling	100
HAH	7	lbfgs	relu	adaptive	200
TNR	7	sgd	logistic	adaptive	200
DXB	15	sgd	identity	invscaling	100
JED	15	lbfgs	tanh	adaptive	100
KRT	15	lbfgs	logistic	adaptive	200
JUB	7	sgd	logistic	constant	100
ADD	15	lbfgs	logistic	adaptive	200
JIB	15	lbfgs	relu	invscaling	200
EBB	15	lbfgs	logistic	adaptive	200
HRE	15	sgd	tanh	constant	200
LUN	15	sgd	tanh	adaptive	100
JRO	7	lbfgs	logistic	constant	200
ZNZ	15	sgd	tanh	adaptive	100
DAR	15	lbfgs	logistic	adaptive	200
BJM	15	lbfgs	logistic	invscaling	200
KGL	15	lbfgs	logistic	adaptive	200
ACC	15	sgd	logistic	adaptive	200
FNA	15	lbfgs	logistic	invscaling	100
BKO	5	lbfgs	logistic	constant	100
DKR	15	lbfgs	tanh	constant	100
ABJ	15	lbfgs	tanh	adaptive	100
DLA	15	lbfgs	relu	invscaling	100
NSI	15	lbfgs	logistic	invscaling	200
LOS	15	sgd	logistic	adaptive	100
BZV	15	lbfgs	tanh	adaptive	100
FIH	5	lbfgs	relu	invscaling	200
BGF	7	lbfgs	logistic	invscaling	100
FBM	5	lbfgs	relu	adaptive	100
NLA	7	lbfgs	tanh	adaptive	100
MBA	15	lbfgs	logistic	adaptive	200
KIS	15	lbfgs	logistic	adaptive	200
LLW	10	sgd	logistic	invscaling	100
APL	15	lbfgs	logistic	constant	100
MPM	15	sgd	logistic	adaptive	100
BLZ	15	lbfgs	logistic	adaptive	100
JNB	15	sgd	tanh	invscaling	200
LAD	5	lbfgs	tanh	invscaling	100
CPT	15	lbfgs	tanh	invscaling	200
LVI	15	lbfgs	logistic	constant	200
BKK	15	lbfgs	logistic	adaptive	200
HKG	5	lbfgs	logistic	adaptive	100
CAN	15	lbfgs	relu	invscaling	100
HAN	15	lbfgs	identity	invscaling	100

F

Appendix - Results of test I

Table F.1: Complete set of results test 1, part 1/3

		LOS	SUN
KRT	R^2	0,398	0,386
	MAE	0,229	0,232
	RMSE	0,298	0,301
ADD	R^2	0,356	0,355
	MAE	0,179	0,180
	RMSE	0,257	0,257
JIB	R^2	0,208	0,206
	MAE	0,163	0,162
	RMSE	0,303	0,303
DAR	R^2	0,186	0,186
	MAE	0,243	0,243
	RMSE	0,329	0,329
JRO	R^2	0,243	0,241
	MAE	0,245	0,246
	RMSE	0,328	0,328
BJM	R^2	0,127	0,129
	MAE	0,200	0,199
	RMSE	0,260	0,260
EBB	R^2	0,261	0,262
	MAE	0,245	0,245
	RMSE	0,314	0,313
KGL	R^2	0,199	0,198
	MAE	0,233	0,234
	RMSE	0,331	0,331
ZNZ	R^2	0,304	0,312
	MAE	0,256	0,256
	RMSE	0,350	0,348

Table F.2: Complete set of results test 1, part 2/3

BZV	R^2	0,324	0,456
	MAE	0,143	0,141
	RMSE	0,217	0,195
MBA	R^2	0,150	0,152
	MAE	0,252	0,251
	RMSE	0,311	0,311
HRE	R^2	0,236	0,236
	MAE	0,211	0,211
	RMSE	0,276	0,276
LLW	R^2	0,330	0,332
	MAE	0,173	0,174
	RMSE	0,237	0,237
LUN	R^2	0,216	0,214
	MAE	0,185	0,186
	RMSE	0,266	0,266
JNB	R^2	0,260	0,255
	MAE	0,205	0,206
	RMSE	0,278	0,279
BKK	R^2	0,354	0,335
	MAE	0,211	0,214
	RMSE	0,284	0,288
HAN	R^2	0,254	0,248
	MAE	0,149	0,150
	RMSE	0,221	0,222
JUB	R^2	0,296	0,288
	MAE	0,167	0,168
	RMSE	0,231	0,232
ACC	R^2	0,499	0,453
	MAE	0,182	0,191
	RMSE	0,227	0,237
BKO	R^2	0,200	0,198
	MAE	0,293	0,294
	RMSE	0,351	0,351
DKR	R^2	0,379	0,385
	MAE	0,245	0,245
	RMSE	0,331	0,329
ABJ	R^2	0,319	0,305
	MAE	0,225	0,227
	RMSE	0,291	0,294
DLA	R^2	0,292	0,282
	MAE	0,203	0,205
	RMSE	0,285	0,287
LOS	R^2	0,526	0,519
	MAE	0,174	0,175
	RMSE	0,256	0,258
BGF	R^2	0,326	0,325
	MAE	0,162	0,162
	RMSE	0,211	0,212
KIS	R^2	0,388	0,384
	MAE	0,193	0,194
	RMSE	0,263	0,264
CPT	R^2	0,289	0,272
	MAE	0,142	0,144
	RMSE	0,204	0,207
LVI	R^2	0,137	0,133
	MAE	0,140	0,138
	RMSE	0,232	0,232
FIH	R^2	0,283	0,269
	MAE	0,143	0,145
	RMSE	0,213	0,216

Table F.3: Complete set of results test 1, part 3/3

FBM	R^2	0,283	0,273
	MAE	0,109	0,109
	RMSE	0,147	0,148
NLA	R^2	0,472	0,477
	MAE	0,224	0,223
	RMSE	0,317	0,315
MPM	R^2	0,373	0,387
	MAE	0,241	0,239
	RMSE	0,334	0,330
APL	R^2	0,442	0,397
	MAE	0,142	0,149
	RMSE	0,193	0,200
BLZ	R^2	0,326	0,318
	MAE	0,224	0,225
	RMSE	0,300	0,301
BOM	R^2	0,424	0,407
	MAE	0,146	0,148
	RMSE	0,200	0,203
SEZ	R^2	0,493	0,473
	MAE	0,178	0,184
	RMSE	0,232	0,237
DZA	R^2	0,311	0,300
	MAE	0,215	0,218
	RMSE	0,300	0,302
HAH	R^2	0,536	0,534
	MAE	0,140	0,140
	RMSE	0,212	0,212
TNR	R^2	0,381	0,380
	MAE	0,233	0,233
	RMSE	0,307	0,307
DXB	R^2	0,170	0,188
	MAE	0,214	0,220
	RMSE	0,331	0,327
FNA	R^2	0,456	0,456
	MAE	0,144	0,147
	RMSE	0,233	0,233
LAD	R^2	0,205	0,182
	MAE	0,199	0,201
	RMSE	0,268	0,271
HKG	R^2	0,000	0,577
	MAE	0,159	0,081
	RMSE	0,232	0,146
CAN	R^2	0,205	0,557
	MAE	0,199	0,106
	RMSE	0,268	0,200

G

Appendix - Results of test II

Table G.1: Complete set of results test 2, part 1/3

		Cyclic			Categorical		
		MLR	RFR	NN	MLR	RFR	NN
KRT	R^2	0,344	0,648	0,535	0,398	0,652	0,543
	MAE	0,238	0,143	0,186	0,229	0,141	0,186
	RMSE	0,312	0,228	0,262	0,298	0,227	0,260
ADD	R^2	0,337	0,573	0,448	0,356	0,557	0,451
	MAE	0,184	0,131	0,160	0,179	0,132	0,162
	RMSE	0,261	0,209	0,238	0,257	0,213	0,237
JIB	R^2	0,202	0,472	0,201	0,208	0,449	0,207
	MAE	0,162	0,110	0,162	0,163	0,121	0,163
	RMSE	0,304	0,247	0,304	0,303	0,253	0,303
DAR	R^2	0,119	0,454	0,264	0,186	0,367	0,257
	MAE	0,255	0,178	0,229	0,243	0,202	0,229
	RMSE	0,342	0,269	0,313	0,329	0,290	0,314
JRO	R^2	0,200	0,547	0,257	0,243	0,524	0,292
	MAE	0,248	0,162	0,238	0,245	0,172	0,230
	RMSE	0,337	0,253	0,325	0,328	0,260	0,317
BJM	R^2	0,108	0,516	0,288	0,127	0,513	0,316
	MAE	0,203	0,120	0,174	0,200	0,121	0,168
	RMSE	0,263	0,194	0,235	0,260	0,195	0,231
EBB	R^2	0,250	0,554	0,389	0,261	0,433	0,334
	MAE	0,246	0,162	0,212	0,245	0,201	0,227
	RMSE	0,316	0,244	0,285	0,314	0,275	0,298
KGL	R^2	0,184	0,497	0,336	0,199	0,477	0,338
	MAE	0,235	0,166	0,211	0,233	0,174	0,211
	RMSE	0,334	0,262	0,301	0,331	0,267	0,300
ZNZ	R^2	0,226	0,703	0,416	0,304	0,693	0,420
	MAE	0,275	0,134	0,236	0,256	0,136	0,235
	RMSE	0,370	0,229	0,321	0,350	0,233	0,320

Table G.2: Complete set of results test 2, part 2/3

BZV	R^2	0,335	0,665	0,391	0,324	0,617	0,400
	MAE	0,140	0,084	0,141	0,143	0,097	0,143
	RMSE	0,216	0,153	0,206	0,217	0,163	0,205
MBA	R^2	0,090	0,476	0,231	0,150	0,332	0,233
	MAE	0,263	0,177	0,236	0,252	0,213	0,236
	RMSE	0,322	0,244	0,296	0,311	0,276	0,296
HRE	R^2	0,196	0,507	0,335	0,236	0,453	0,355
	MAE	0,216	0,141	0,191	0,211	0,157	0,188
	RMSE	0,283	0,222	0,257	0,276	0,233	0,253
LLW	R^2	0,317	0,535	0,389	0,330	0,485	0,403
	MAE	0,176	0,122	0,158	0,173	0,127	0,156
	RMSE	0,240	0,198	0,227	0,237	0,208	0,224
LUN	R^2	0,200	0,512	0,341	0,216	0,474	0,357
	MAE	0,188	0,130	0,165	0,185	0,133	0,164
	RMSE	0,269	0,210	0,244	0,266	0,218	0,241
JNB	R^2	0,245	0,502	0,341	0,260	0,458	0,344
	MAE	0,208	0,142	0,187	0,205	0,159	0,186
	RMSE	0,281	0,228	0,262	0,278	0,238	0,262
BKK	R^2	0,321	0,605	0,319	0,354	0,607	0,350
	MAE	0,212	0,138	0,212	0,211	0,139	0,212
	RMSE	0,291	0,222	0,291	0,284	0,221	0,284
HAN	R^2	0,287	0,678	0,099	0,154	0,653	0,367
	MAE	0,151	0,076	0,118	0,187	0,080	0,097
	RMSE	0,195	0,131	0,219	0,247	0,136	0,183
JUB	R^2	0,211	0,439	0,359	0,254	0,380	0,337
	MAE	0,153	0,118	0,136	0,149	0,132	0,140
	RMSE	0,228	0,192	0,205	0,221	0,202	0,209
ACC	R^2	0,278	0,525	0,381	0,296	0,521	0,393
	MAE	0,171	0,123	0,154	0,167	0,123	0,153
	RMSE	0,234	0,190	0,217	0,231	0,191	0,215
BKO	R^2	0,457	0,618	0,435	0,499	0,629	0,477
	MAE	0,192	0,151	0,196	0,182	0,145	0,185
	RMSE	0,237	0,199	0,241	0,227	0,196	0,232
DKR	R^2	0,190	0,468	0,257	0,200	0,468	0,217
	MAE	0,296	0,205	0,261	0,293	0,204	0,265
	RMSE	0,353	0,286	0,338	0,351	0,286	0,347
ABJ	R^2	0,339	0,634	0,524	0,379	0,624	0,388
	MAE	0,256	0,164	0,207	0,245	0,167	0,244
	RMSE	0,341	0,254	0,289	0,331	0,257	0,328
DLA	R^2	0,317	0,554	0,302	0,319	0,560	0,306
	MAE	0,226	0,156	0,228	0,225	0,164	0,227
	RMSE	0,291	0,235	0,294	0,291	0,234	0,294
LOS	R^2	0,258	0,547	0,428	0,292	0,547	0,431
	MAE	0,209	0,138	0,175	0,203	0,135	0,176
	RMSE	0,292	0,228	0,256	0,285	0,228	0,256
BGF	R^2	0,520	0,636	0,517	0,526	0,613	0,523
	MAE	0,172	0,129	0,173	0,174	0,137	0,174
	RMSE	0,258	0,225	0,259	0,256	0,232	0,257
KIS	R^2	0,219	0,516	0,411	0,326	0,458	0,412
	MAE	0,178	0,122	0,148	0,162	0,136	0,147
	RMSE	0,228	0,179	0,198	0,211	0,190	0,198
CPT	R^2	0,296	0,624	0,390	0,388	0,627	0,467
	MAE	0,205	0,124	0,188	0,193	0,122	0,176
	RMSE	0,282	0,206	0,262	0,263	0,205	0,245
LVI	R^2	0,206	0,774	0,465	0,289	0,776	0,409
	MAE	0,147	0,065	0,114	0,142	0,064	0,105
	RMSE	0,216	0,115	0,177	0,204	0,115	0,186
FIH	R^2	0,140	0,339	0,135	0,137	0,283	0,131
	MAE	0,140	0,113	0,140	0,140	0,122	0,140
	RMSE	0,231	0,203	0,232	0,232	0,211	0,232

Table G.3: Complete set of results test 2, part 3/3

FBM	R^2	0,266	0,476	0,259	0,283	0,448	0,270
	MAE	0,145	0,118	0,148	0,143	0,115	0,146
	RMSE	0,216	0,182	0,217	0,213	0,187	0,215
NLA	R^2	0,298	0,573	0,286	0,283	0,522	0,272
	MAE	0,110	0,073	0,110	0,109	0,079	0,109
	RMSE	0,145	0,113	0,146	0,147	0,120	0,148
MPM	R^2	0,485	0,685	0,563	0,472	0,618	0,513
	MAE	0,221	0,139	0,194	0,224	0,166	0,206
	RMSE	0,313	0,244	0,288	0,317	0,269	0,304
APL	R^2	0,407	0,549	0,415	0,373	0,419	0,383
	MAE	0,230	0,151	0,228	0,241	0,179	0,239
	RMSE	0,324	0,283	0,322	0,334	0,321	0,331
BLZ	R^2	0,445	0,681	0,439	0,442	0,658	0,436
	MAE	0,140	0,088	0,142	0,142	0,092	0,143
	RMSE	0,192	0,146	0,193	0,193	0,151	0,194
BOM	R^2	0,280	0,653	0,459	0,326	0,602	0,456
	MAE	0,234	0,138	0,200	0,224	0,157	0,200
	RMSE	0,310	0,215	0,268	0,300	0,230	0,269
SEZ	R^2	0,413	0,738	0,525	0,424	0,696	0,517
	MAE	0,148	0,077	0,126	0,146	0,088	0,128
	RMSE	0,202	0,135	0,182	0,200	0,145	0,183
DZA	R^2	0,443	0,715	0,465	0,493	0,675	0,507
	MAE	0,181	0,115	0,179	0,178	0,121	0,177
	RMSE	0,243	0,174	0,238	0,232	0,186	0,229
HAH	R^2	0,264	0,619	0,307	0,311	0,558	0,310
	MAE	0,228	0,139	0,217	0,215	0,157	0,217
	RMSE	0,309	0,223	0,300	0,300	0,240	0,300
TNR	R^2	0,470	0,761	0,619	0,536	0,737	0,623
	MAE	0,159	0,084	0,126	0,140	0,089	0,126
	RMSE	0,226	0,152	0,192	0,212	0,159	0,191
DXB	R^2	0,329	0,596	0,468	0,381	0,590	0,485
	MAE	0,242	0,170	0,209	0,233	0,169	0,206
	RMSE	0,319	0,248	0,284	0,307	0,250	0,280
FNA	R^2	0,117	0,403	0,128	0,170	0,390	0,170
	MAE	0,216	0,154	0,222	0,214	0,157	0,222
	RMSE	0,341	0,281	0,339	0,331	0,284	0,331
LAD	R^2	0,456	0,538	0,451	0,456	0,553	0,456
	MAE	0,141	0,108	0,144	0,144	0,110	0,147
	RMSE	0,233	0,214	0,234	0,233	0,211	0,233
HKG	R^2	0,077	0,605	0,524	0,000	0,630	0,246
	MAE	0,142	0,075	0,073	0,159	0,077	0,085
	RMSE	0,215	0,141	0,154	0,232	0,136	0,194
CAN	R^2	0,208	0,560	0,138	0,205	0,554	0,187
	MAE	0,199	0,105	0,207	0,199	0,109	0,200
	RMSE	0,267	0,199	0,279	0,268	0,200	0,271

H

Appendix - Results of test III

Table H.1: Complete set of results test 3, part 1/3

		Continous			Categorical		
		MLR	RFR	NN	MLR	RFR	NN
KRT	R^2	0,397	0,648	0,557	0,398	0,652	0,543
	MAE	0,228	0,142	0,182	0,229	0,141	0,186
	RMSE	0,299	0,228	0,256	0,298	0,227	0,260
ADD	R^2	0,350	0,559	0,453	0,356	0,557	0,451
	MAE	0,179	0,132	0,161	0,179	0,132	0,162
	RMSE	0,258	0,213	0,237	0,257	0,213	0,237
JIB	R^2	0,209	0,463	0,283	0,208	0,449	0,207
	MAE	0,162	0,114	0,166	0,163	0,121	0,163
	RMSE	0,303	0,250	0,288	0,303	0,253	0,303
DAR	R^2	0,163	0,362	0,266	0,186	0,367	0,257
	MAE	0,245	0,203	0,228	0,243	0,202	0,229
	RMSE	0,333	0,291	0,312	0,329	0,290	0,314
JRO	R^2	0,246	0,516	0,322	0,243	0,524	0,292
	MAE	0,245	0,173	0,227	0,245	0,172	0,230
	RMSE	0,327	0,262	0,310	0,328	0,260	0,317
BJM	R^2	0,100	0,510	0,296	0,127	0,513	0,316
	MAE	0,204	0,121	0,172	0,200	0,121	0,168
	RMSE	0,264	0,195	0,234	0,260	0,195	0,231
EBB	R^2	0,226	0,431	0,338	0,261	0,433	0,334
	MAE	0,252	0,201	0,225	0,245	0,201	0,227
	RMSE	0,321	0,275	0,297	0,314	0,275	0,298
KGL	R^2	0,185	0,476	0,330	0,199	0,477	0,338
	MAE	0,236	0,175	0,214	0,233	0,174	0,211
	RMSE	0,334	0,267	0,303	0,331	0,267	0,300
ZNZ	R^2	0,302	0,693	0,445	0,304	0,693	0,420
	MAE	0,257	0,137	0,227	0,256	0,136	0,235
	RMSE	0,351	0,233	0,313	0,350	0,233	0,320

Table H.2: Complete set of results test 3, part 2/3

BZV	R^2	0,311	0,624	0,411	0,324	0,617	0,400
	MAE	0,143	0,094	0,142	0,143	0,097	0,143
	RMSE	0,219	0,162	0,203	0,217	0,163	0,205
MBA	R^2	0,140	0,331	0,224	0,150	0,332	0,233
	MAE	0,254	0,213	0,238	0,252	0,213	0,236
	RMSE	0,313	0,276	0,297	0,311	0,276	0,296
HRE	R^2	0,236	0,450	0,347	0,236	0,453	0,355
	MAE	0,211	0,157	0,189	0,211	0,157	0,188
	RMSE	0,276	0,234	0,255	0,276	0,233	0,253
LLW	R^2	0,320	0,483	0,393	0,330	0,485	0,403
	MAE	0,175	0,127	0,156	0,173	0,127	0,156
	RMSE	0,239	0,208	0,226	0,237	0,208	0,224
LUN	R^2	0,210	0,475	0,356	0,216	0,474	0,357
	MAE	0,186	0,133	0,163	0,185	0,133	0,164
	RMSE	0,267	0,217	0,241	0,266	0,218	0,241
JNB	R^2	0,257	0,457	0,330	0,260	0,458	0,344
	MAE	0,207	0,159	0,190	0,205	0,159	0,186
	RMSE	0,278	0,238	0,264	0,278	0,238	0,262
BKK	R^2	0,342	0,596	0,344	0,354	0,607	0,350
	MAE	0,212	0,141	0,213	0,211	0,139	0,212
	RMSE	0,286	0,224	0,286	0,284	0,221	0,284
HAN	R^2	0,209	0,653	0,437	0,154	0,653	0,367
	MAE	0,153	0,081	0,093	0,187	0,080	0,097
	RMSE	0,205	0,136	0,173	0,247	0,136	0,183
JUB	R^2	0,251	0,378	0,340	0,254	0,380	0,337
	MAE	0,149	0,132	0,139	0,149	0,132	0,140
	RMSE	0,222	0,202	0,208	0,221	0,202	0,209
ACC	R^2	0,286	0,521	0,405	0,296	0,521	0,393
	MAE	0,168	0,123	0,150	0,167	0,123	0,153
	RMSE	0,233	0,191	0,212	0,231	0,191	0,215
BKO	R^2	0,478	0,626	0,444	0,499	0,629	0,477
	MAE	0,187	0,146	0,192	0,182	0,145	0,185
	RMSE	0,232	0,196	0,239	0,227	0,196	0,232
DKR	R^2	0,193	0,462	0,287	0,200	0,468	0,217
	MAE	0,295	0,205	0,257	0,293	0,204	0,265
	RMSE	0,353	0,288	0,331	0,351	0,286	0,347
ABJ	R^2	0,378	0,606	0,544	0,379	0,624	0,388
	MAE	0,245	0,181	0,202	0,245	0,167	0,244
	RMSE	0,331	0,263	0,283	0,331	0,257	0,328
DLA	R^2	0,327	0,555	0,312	0,319	0,560	0,306
	MAE	0,225	0,159	0,227	0,225	0,164	0,227
	RMSE	0,289	0,235	0,292	0,291	0,234	0,294
LOS	R^2	0,286	0,547	0,427	0,292	0,547	0,431
	MAE	0,205	0,136	0,179	0,203	0,135	0,176
	RMSE	0,286	0,228	0,256	0,285	0,228	0,256
BGF	R^2	0,499	0,612	0,497	0,526	0,613	0,523
	MAE	0,179	0,137	0,178	0,174	0,137	0,174
	RMSE	0,264	0,232	0,264	0,256	0,232	0,257
KIS	R^2	0,301	0,458	0,411	0,326	0,458	0,412
	MAE	0,166	0,136	0,147	0,162	0,136	0,147
	RMSE	0,215	0,190	0,198	0,211	0,190	0,198
CPT	R^2	0,386	0,624	0,481	0,388	0,627	0,467
	MAE	0,194	0,126	0,173	0,193	0,122	0,176
	RMSE	0,264	0,206	0,242	0,263	0,205	0,245
LVI	R^2	0,317	0,779	0,296	0,289	0,776	0,409
	MAE	0,143	0,064	0,147	0,142	0,064	0,105
	RMSE	0,200	0,114	0,203	0,204	0,115	0,186
FIH	R^2	0,126	0,279	0,121	0,137	0,283	0,131
	MAE	0,139	0,122	0,140	0,140	0,122	0,140
	RMSE	0,233	0,212	0,234	0,232	0,211	0,232

Table H.3: Complete set of results test 3, part 3/3

FBM	R^2	0,269	0,412	0,259	0,283	0,448	0,270
	MAE	0,143	0,126	0,147	0,143	0,115	0,146
	RMSE	0,215	0,193	0,217	0,213	0,187	0,215
NLA	R^2	0,268	0,532	0,252	0,283	0,522	0,272
	MAE	0,111	0,078	0,111	0,109	0,079	0,109
	RMSE	0,148	0,119	0,150	0,147	0,120	0,148
MPM	R^2	0,448	0,617	0,512	0,472	0,618	0,513
	MAE	0,230	0,167	0,208	0,224	0,166	0,206
	RMSE	0,324	0,270	0,304	0,317	0,269	0,304
APL	R^2	0,394	0,404	0,398	0,373	0,419	0,383
	MAE	0,236	0,177	0,232	0,241	0,179	0,239
	RMSE	0,328	0,325	0,327	0,334	0,321	0,331
BLZ	R^2	0,414	0,651	0,399	0,442	0,658	0,436
	MAE	0,148	0,094	0,149	0,142	0,092	0,143
	RMSE	0,198	0,153	0,200	0,193	0,151	0,194
BOM	R^2	0,299	0,599	0,456	0,326	0,602	0,456
	MAE	0,230	0,157	0,198	0,224	0,157	0,200
	RMSE	0,305	0,231	0,269	0,300	0,230	0,269
SEZ	R^2	0,428	0,697	0,556	0,424	0,696	0,517
	MAE	0,146	0,088	0,122	0,146	0,088	0,128
	RMSE	0,199	0,145	0,175	0,200	0,145	0,183
DZA	R^2	0,502	0,669	0,486	0,493	0,675	0,507
	MAE	0,174	0,122	0,179	0,178	0,121	0,177
	RMSE	0,230	0,187	0,234	0,232	0,186	0,229
HAH	R^2	0,316	0,564	0,315	0,311	0,558	0,310
	MAE	0,214	0,155	0,216	0,215	0,157	0,217
	RMSE	0,298	0,238	0,299	0,300	0,240	0,300
TNR	R^2	0,531	0,743	0,619	0,536	0,737	0,623
	MAE	0,141	0,088	0,125	0,140	0,089	0,126
	RMSE	0,213	0,158	0,192	0,212	0,159	0,191
DXB	R^2	0,374	0,588	0,489	0,381	0,590	0,485
	MAE	0,234	0,170	0,204	0,233	0,169	0,206
	RMSE	0,308	0,250	0,279	0,307	0,250	0,280
FNA	R^2	0,388	0,768	0,188	0,170	0,390	0,170
	MAE	0,221	0,087	0,216	0,214	0,157	0,222
	RMSE	0,312	0,192	0,327	0,331	0,284	0,331
LAD	R^2	0,195	0,352	0,351	0,456	0,553	0,456
	MAE	0,211	0,154	0,163	0,144	0,110	0,147
	RMSE	0,326	0,292	0,254	0,233	0,211	0,233
HKG	R^2	0,347	0,530	0,190	0,000	0,630	0,246
	MAE	0,163	0,109	0,092	0,159	0,077	0,085
	RMSE	0,255	0,216	0,201	0,232	0,136	0,194
CAN	R^2	0,194	0,544	0,173	0,205	0,554	0,187
	MAE	0,206	0,108	0,209	0,199	0,109	0,200
	RMSE	0,269	0,203	0,273	0,268	0,200	0,271

I

Appendix - Results of test IV

Table I.1: Complete set of results test 4, part 1/3

		Grouped			Non Grouped		
		MLR	RFR	NN	MLR	RFR	NN
KRT	R^2	0,282	0,386	0,470	0,398	0,652	0,543
	MAE	0,233	0,212	0,206	0,229	0,141	0,186
	RMSE	0,307	0,284	0,280	0,298	0,227	0,260
ADD	R^2	0,384	0,490	0,452	0,356	0,557	0,451
	MAE	0,150	0,131	0,163	0,179	0,132	0,162
	RMSE	0,205	0,186	0,250	0,257	0,213	0,237
JIB	R^2	0,404	0,576	0,214	0,208	0,449	0,207
	MAE	0,194	0,152	0,165	0,163	0,121	0,163
	RMSE	0,249	0,210	0,318	0,303	0,253	0,303
DAR	R^2	0,251	0,382	0,255	0,186	0,367	0,257
	MAE	0,230	0,206	0,182	0,243	0,202	0,229
	RMSE	0,310	0,282	0,261	0,329	0,290	0,314
JRO	R^2	0,413	0,558	0,296	0,243	0,524	0,292
	MAE	0,165	0,142	0,321	0,245	0,172	0,230
	RMSE	0,235	0,204	0,392	0,328	0,260	0,317
BJM	R^2	0,287	0,383	0,250	0,127	0,513	0,316
	MAE	0,255	0,226	0,193	0,200	0,121	0,168
	RMSE	0,328	0,306	0,261	0,260	0,195	0,231
EBB	R^2	0,293	0,697	0,241	0,261	0,433	0,334
	MAE	0,249	0,123	0,247	0,245	0,201	0,227
	RMSE	0,328	0,215	0,318	0,314	0,275	0,298
KGL	R^2	0,314	0,476	0,195	0,199	0,477	0,338
	MAE	0,250	0,201	0,239	0,233	0,174	0,211
	RMSE	0,315	0,275	0,337	0,331	0,267	0,300
ZNZ	R^2	0,190	0,306	0,367	0,304	0,693	0,420
	MAE	0,141	0,128	0,228	0,256	0,136	0,235
	RMSE	0,210	0,194	0,319	0,350	0,233	0,320

Table I.2: Complete set of results test 4, part 2/3

BZV	R^2	0,372	0,487	0,772	0,324	0,617	0,400
	MAE	0,180	0,157	0,086	0,143	0,097	0,143
	RMSE	0,265	0,239	0,138	0,217	0,163	0,205
MBA	R^2	0,221	0,300	0,157	0,150	0,332	0,233
	MAE	0,157	0,133	0,256	0,252	0,213	0,236
	RMSE	0,309	0,293	0,311	0,311	0,276	0,296
HRE	R^2	0,195	0,239	0,264	0,236	0,453	0,355
	MAE	0,256	0,247	0,230	0,211	0,157	0,188
	RMSE	0,326	0,317	0,293	0,276	0,233	0,253
LLW	R^2	0,262	0,369	0,273	0,330	0,485	0,403
	MAE	0,332	0,291	0,166	0,173	0,127	0,156
	RMSE	0,401	0,371	0,241	0,237	0,208	0,224
LUN	R^2	0,278	0,473	0,196	0,216	0,474	0,357
	MAE	0,235	0,199	0,227	0,185	0,133	0,164
	RMSE	0,336	0,288	0,310	0,266	0,218	0,241
JNB	R^2	0,093	0,331	0,334	0,260	0,458	0,344
	MAE	0,228	0,175	0,188	0,205	0,159	0,186
	RMSE	0,285	0,245	0,257	0,278	0,238	0,262
BKK	R^2	0,144	0,227	0,382	0,354	0,607	0,350
	MAE	0,245	0,230	0,172	0,211	0,139	0,212
	RMSE	0,346	0,329	0,231	0,284	0,221	0,284
HAN	R^2	0,212	0,261	0,000	0,154	0,653	0,367
	MAE	0,187	0,178	0,210	0,187	0,080	0,097
	RMSE	0,267	0,258	0,373	0,247	0,136	0,183
JUB	R^2	0,203	0,281	0,282	0,254	0,380	0,337
	MAE	0,158	0,146	0,131	0,149	0,132	0,140
	RMSE	0,223	0,212	0,200	0,221	0,202	0,209
ACC	R^2	0,184	0,184	0,254	0,296	0,521	0,393
	MAE	0,205	0,172	0,154	0,167	0,123	0,153
	RMSE	0,324	0,324	0,219	0,231	0,191	0,215
BKO	R^2	0,441	0,537	0,353	0,499	0,629	0,477
	MAE	0,196	0,177	0,209	0,182	0,145	0,185
	RMSE	0,250	0,228	0,268	0,227	0,196	0,232
DKR	R^2	0,271	0,406	0,161	0,200	0,468	0,217
	MAE	0,314	0,266	0,332	0,293	0,204	0,265
	RMSE	0,371	0,335	0,393	0,351	0,286	0,347
ABJ	R^2	0,103	0,346	0,146	0,379	0,624	0,388
	MAE	0,284	0,220	0,276	0,245	0,167	0,244
	RMSE	0,362	0,309	0,346	0,331	0,257	0,328
DLA	R^2	0,303	0,632	0,184	0,319	0,560	0,306
	MAE	0,097	0,074	0,096	0,225	0,164	0,227
	RMSE	0,152	0,110	0,126	0,291	0,234	0,294
LOS	R^2	0,738	0,820	0,191	0,292	0,547	0,431
	MAE	0,120	0,089	0,219	0,203	0,135	0,176
	RMSE	0,148	0,123	0,322	0,285	0,228	0,256
BGF	R^2	0,213	0,396	0,650	0,526	0,613	0,523
	MAE	0,159	0,137	0,238	0,174	0,137	0,174
	RMSE	0,245	0,215	0,289	0,256	0,232	0,257
KIS	R^2	0,644	0,784	0,310	0,326	0,458	0,412
	MAE	0,239	0,163	0,148	0,162	0,136	0,147
	RMSE	0,291	0,227	0,198	0,211	0,190	0,198
CPT	R^2	0,213	0,354	0,441	0,388	0,627	0,467
	MAE	0,153	0,126	0,168	0,193	0,122	0,176
	RMSE	0,244	0,221	0,244	0,263	0,205	0,245
LVI	R^2	0,282	0,374	0,452	0,289	0,776	0,409
	MAE	0,122	0,104	0,111	0,142	0,064	0,105
	RMSE	0,159	0,148	0,186	0,204	0,115	0,186
FIH	R^2	0,115	0,162	0,073	0,137	0,283	0,131
	MAE	0,265	0,254	0,170	0,140	0,122	0,140
	RMSE	0,319	0,310	0,256	0,232	0,211	0,232

Table I.3: Complete set of results test 4, part 3/3

FBM	R^2	0,267	0,309	0,202	0,283	0,448	0,270
	MAE	0,155	0,148	0,158	0,143	0,115	0,146
	RMSE	0,205	0,199	0,245	0,213	0,187	0,215
NLA	R^2	0,156	0,269	0,292	0,283	0,522	0,272
	MAE	0,237	0,210	0,120	0,109	0,079	0,109
	RMSE	0,317	0,295	0,158	0,147	0,120	0,148
MPM	R^2	0,198	0,319	0,388	0,472	0,618	0,513
	MAE	0,247	0,221	0,196	0,224	0,166	0,206
	RMSE	0,306	0,282	0,269	0,317	0,269	0,304
APL	R^2	0,272	0,362	0,608	0,373	0,419	0,383
	MAE	0,166	0,150	0,162	0,241	0,179	0,239
	RMSE	0,241	0,226	0,216	0,334	0,321	0,331
BLZ	R^2	0,361	0,480	0,514	0,442	0,658	0,436
	MAE	0,195	0,170	0,096	0,142	0,092	0,143
	RMSE	0,275	0,248	0,155	0,193	0,151	0,194
BOM	R^2	0,532	0,707	0,375	0,326	0,602	0,456
	MAE	0,117	0,087	0,218	0,224	0,157	0,200
	RMSE	0,152	0,121	0,288	0,300	0,230	0,269
SEZ	R^2	0,599	0,686	0,453	0,424	0,696	0,517
	MAE	0,163	0,130	0,137	0,146	0,088	0,128
	RMSE	0,218	0,193	0,195	0,200	0,145	0,183
DZA	R^2	0,279	0,350	0,417	0,493	0,675	0,507
	MAE	0,201	0,188	0,183	0,178	0,121	0,177
	RMSE	0,267	0,254	0,249	0,232	0,186	0,229
HAH	R^2	0,499	0,603	0,233	0,311	0,558	0,310
	MAE	0,131	0,108	0,235	0,215	0,157	0,217
	RMSE	0,217	0,193	0,316	0,300	0,240	0,300
TNR	R^2	0,376	0,511	0,497	0,536	0,737	0,623
	MAE	0,186	0,156	0,151	0,140	0,089	0,126
	RMSE	0,258	0,229	0,220	0,212	0,159	0,191
DXB	R^2	0,333	0,601	0,369	0,381	0,590	0,485
	MAE	0,142	0,102	0,234	0,233	0,169	0,206
	RMSE	0,205	0,158	0,310	0,307	0,250	0,280
FNA	R^2	0,507	0,536	0,125	0,170	0,390	0,170
	MAE	0,122	0,086	0,218	0,214	0,157	0,222
	RMSE	0,196	0,190	0,340	0,331	0,284	0,331
LAD	R^2	0,135	0,372	0,502	0,456	0,553	0,456
	MAE	0,219	0,167	0,130	0,144	0,110	0,147
	RMSE	0,280	0,239	0,216	0,233	0,211	0,233
HKG	R^2	0,000	0,240	0,445	0,000	0,630	0,246
	MAE	0,238	0,130	0,091	0,159	0,077	0,085
	RMSE	0,389	0,217	0,208	0,232	0,136	0,194
CAN	R^2	0,205	0,492	0,175	0,205	0,554	0,187
	MAE	0,199	0,124	0,179	0,199	0,109	0,200
	RMSE	0,268	0,214	0,274	0,268	0,200	0,271

J

Appendix - Best fit results

Table J.1: Complete set of final results, part 1/3

		MLR	RFR	NN
KRT	R^2	0,398	0,652	0,543
	MAE	0,229	0,143	0,186
	RMSE	0,298	0,227	0,260
ADD	R^2	0,356	0,570	0,451
	MAE	0,179	0,132	0,162
	RMSE	0,257	0,210	0,237
JIB	R^2	0,208	0,465	0,207
	MAE	0,163	0,113	0,163
	RMSE	0,303	0,249	0,303
DAR	R^2	0,186	0,451	0,257
	MAE	0,243	0,178	0,229
	RMSE	0,329	0,270	0,314
JRO	R^2	0,243	0,529	0,292
	MAE	0,245	0,170	0,230
	RMSE	0,328	0,258	0,317
BJM	R^2	0,127	0,512	0,316
	MAE	0,200	0,122	0,168
	RMSE	0,260	0,195	0,231
EBB	R^2	0,261	0,550	0,334
	MAE	0,245	0,163	0,227
	RMSE	0,314	0,245	0,298
KGL	R^2	0,199	0,494	0,338
	MAE	0,233	0,167	0,211
	RMSE	0,331	0,263	0,300
ZNZ	R^2	0,304	0,700	0,420
	MAE	0,256	0,135	0,235
	RMSE	0,350	0,230	0,320

Table J.2: Complete set of final results, part 1/3

BZV	R^2	0,324	0,657	0,400
	MAE	0,143	0,087	0,143
	RMSE	0,217	0,155	0,205
MBA	R^2	0,150	0,473	0,233
	MAE	0,252	0,178	0,236
	RMSE	0,311	0,245	0,296
HRE	R^2	0,236	0,503	0,355
	MAE	0,211	0,142	0,188
	RMSE	0,276	0,223	0,253
LLW	R^2	0,330	0,538	0,403
	MAE	0,173	0,121	0,156
	RMSE	0,237	0,197	0,224
LUN	R^2	0,216	0,507	0,357
	MAE	0,185	0,131	0,164
	RMSE	0,266	0,211	0,241
JNB	R^2	0,260	0,497	0,344
	MAE	0,205	0,143	0,186
	RMSE	0,278	0,229	0,262
BKK	R^2	0,354	0,603	0,350
	MAE	0,211	0,140	0,212
	RMSE	0,284	0,222	0,284
HAN	R^2	0,154	0,674	0,367
	MAE	0,187	0,075	0,097
	RMSE	0,247	0,132	0,183
JUB	R^2	0,254	0,439	0,337
	MAE	0,149	0,119	0,140
	RMSE	0,221	0,192	0,209
ACC	R^2	0,296	0,521	0,393
	MAE	0,167	0,124	0,153
	RMSE	0,231	0,191	0,215
BKO	R^2	0,499	0,620	0,477
	MAE	0,182	0,151	0,185
	RMSE	0,227	0,198	0,232
DKR	R^2	0,200	0,460	0,217
	MAE	0,293	0,210	0,265
	RMSE	0,351	0,288	0,347
ABJ	R^2	0,379	0,634	0,388
	MAE	0,245	0,163	0,244
	RMSE	0,331	0,254	0,328
DLA	R^2	0,319	0,540	0,306
	MAE	0,225	0,157	0,227
	RMSE	0,291	0,239	0,294
LOS	R^2	0,292	0,544	0,431
	MAE	0,203	0,139	0,176
	RMSE	0,285	0,229	0,256
BGF	R^2	0,526	0,626	0,523
	MAE	0,174	0,132	0,174
	RMSE	0,256	0,228	0,257
KIS	R^2	0,326	0,515	0,412
	MAE	0,162	0,123	0,147
	RMSE	0,211	0,179	0,198
CPT	R^2	0,388	0,614	0,467
	MAE	0,193	0,129	0,176
	RMSE	0,263	0,209	0,245
LVI	R^2	0,289	0,771	0,409
	MAE	0,142	0,065	0,105
	RMSE	0,204	0,116	0,186

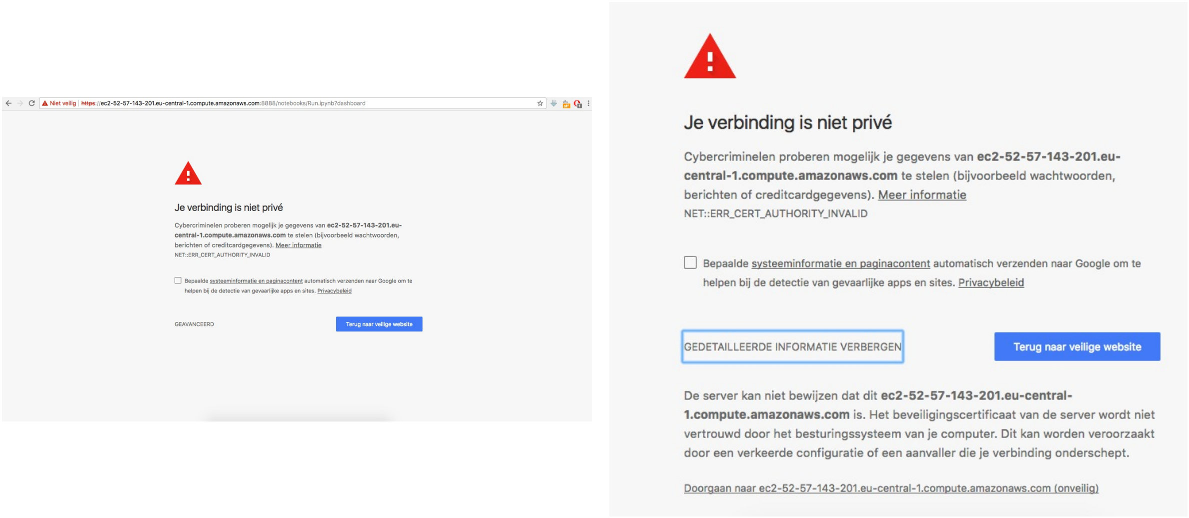
Table J.3: Complete set of final results, part 1/3

FIH	R^2	0,137	0,336	0,131
	MAE	0,140	0,112	0,140
	RMSE	0,232	0,203	0,232
FBM	R^2	0,283	0,473	0,270
	MAE	0,143	0,117	0,146
	RMSE	0,213	0,183	0,215
NLA	R^2	0,283	0,591	0,272
	MAE	0,109	0,071	0,109
	RMSE	0,147	0,111	0,148
MPM	R^2	0,472	0,677	0,513
	MAE	0,224	0,139	0,206
	RMSE	0,317	0,247	0,304
APL	R^2	0,373	0,525	0,383
	MAE	0,241	0,154	0,239
	RMSE	0,334	0,290	0,331
BLZ	R^2	0,442	0,684	0,436
	MAE	0,142	0,088	0,143
	RMSE	0,193	0,145	0,194
BOM	R^2	0,326	0,651	0,456
	MAE	0,224	0,139	0,200
	RMSE	0,300	0,216	0,269
SEZ	R^2	0,424	0,736	0,517
	MAE	0,146	0,077	0,128
	RMSE	0,200	0,135	0,183
DZA	R^2	0,493	0,716	0,507
	MAE	0,178	0,116	0,177
	RMSE	0,232	0,174	0,229
HAH	R^2	0,311	0,617	0,310
	MAE	0,215	0,140	0,217
	RMSE	0,300	0,223	0,300
TNR	R^2	0,536	0,756	0,623
	MAE	0,140	0,085	0,126
	RMSE	0,212	0,153	0,191
DXB	R^2	0,381	0,593	0,485
	MAE	0,233	0,171	0,206
	RMSE	0,307	0,249	0,280
FNA	R^2	0,170	0,371	0,170
	MAE	0,214	0,158	0,222
	RMSE	0,331	0,288	0,331
LAD	R^2	0,456	0,530	0,456
	MAE	0,144	0,109	0,147
	RMSE	0,233	0,216	0,233
HKG	R^2	0,000	0,577	0,246
	MAE	0,159	0,081	0,085
	RMSE	0,232	0,146	0,194
CAN	R^2	0,205	0,557	0,187
	MAE	0,199	0,106	0,200
	RMSE	0,268	0,200	0,271

K

Appendix - Web-application AWS

To support the use of the prediction model, a web application is built. This can be accessed via: `https://ec2-35-158-56-26.eu-central-1.compute.amazonaws.com:8888/notebooks/Run.ipynb?dashboard` The application can only be used with a **Chrome browser**. After clicking on the link, there will pop up a warning that the connection is not private. Next, you should click on advanced, and there you can choose to proceed anyway. (See Figure K.1)



Niet veilig | ~~https~~://ec2-52-57-143-201.eu-central-1.compute.amazonaws.com:8888/notebooks/Run.ipynb?dashboard

Je verbinding is niet privé

Cybercriminelen proberen mogelijk je gegevens van **ec2-52-57-143-201.eu-central-1.compute.amazonaws.com** te stelen (bijvoorbeeld wachtwoorden, berichten of creditcardgegevens). Meer informatie

NET::ERR_CERT_AUTHORITY_INVALID

□ Bepaalde systeeminformatie en paginacontent automatisch verzenden naar Google om te helpen bij de detectie van gevaarlijke apps en sites. Privacybeleid

Je verbinding is niet privé

Cybercriminelen proberen mogelijk je gegevens van **ec2-52-57-143-201.eu-central-1.compute.amazonaws.com** te stelen (bijvoorbeeld wachtwoorden, berichten of creditcardgegevens). Meer informatie

NET::ERR_CERT_AUTHORITY_INVALID

□ Bepaalde systeeminformatie en paginacontent automatisch verzenden naar Google om te helpen bij de detectie van gevaarlijke apps en sites. Privacybeleid

GEDETAILLEERDE INFORMATIE VERBERGEN
Terug naar veilige website

De server kan niet bewijzen dat dit **ec2-52-57-143-201.eu-central-1.compute.amazonaws.com** is. Het beveiligingscertificaat van de server wordt niet vertrouwd door het besturingssysteem van je computer. Dit kan worden veroorzaakt door een verkeerde configuratie of een aanvaller die je verbinding onderschept.

Doorgaan naar ec2-52-57-143-201.eu-central-1.compute.amazonaws.com (onveilig)

Figure K.1: Opening website

After proceeding to the website, a password is asked. The password is, **tfpkqsim008**. Just wait 10 sec until the website is finished loading and than it is ready to use!

L

Appendix - Sensitivity analysis of the ticketing lead time

One of the features which determines the actual price of the flight pass is the ticketing lead time (TLT). This ticketing lead time is defined as the minimum amount of time (in days) which needs to be between booking the flight and the departure of the flight. For example this means, when the option 1 day before departure is selected, a customer has to book at least 1 day before departure but 60 days before departure is also possible. After discussing this with an expert, the decision was made to assume that the customer will book the ticket 1 day before departure is set to 50%, every day thereafter the probability is divided by 2. This means that for 1 day the probability is 50%, 2 days before departure the probability drops to 25%, 3 days before departure the probability drops to 12,5% and so on. Since this assumption is based on hard numbers, it is wise to analyse how sensitive the model is to these numbers. Therefore a sensitivity analysis is performed.

The sensitivity of the ticketing lead time is checked by use of a one-factor-at-a-time method. This method changes only one variable (Ticketing lead time) to see what effect this produces on the output. Different values for the ticketing lead time are assumed. These assumed values for when the option 1 day before departure is selected, are listed below:

- **Assumption 1:** The probability is 100 percent that someone buys 1 day in advance
- **Assumption 2:** The probability is the same for buying 1 day, 2 days till 60 days in advance
- **Assumption 3:** The probability depends on the buying behaviour of passengers for normal tickets.

Figure L.1 illustrates the different input distributions for the model. There are 4 distributions where the first one is the used distribution in the model. The other 3 are used to check how sensitive the model is to the ticketing lead time.

As one can see, the behaviour of the distribution of the first assumption is peaked at a ticketing lead time of 1 days (yellow line). The green line represents the distribution of assumption 2 where there is an evenly spread distribution. The red line is based on the buying behaviour of passengers for normal tickets. The peak at 60 days ticketing lead time is because every ticket sold from 350 days till 60 days of departure is summed at that point. The blue line is the distribution used in the model.

The route Mumbai (BOM) - Nairobi (NBO) is used for the buying behaviour of passengers for normal tickets. Mumbai is part of cluster 3 where ticketing lead time mainly drives the ticket price, therefore one can assume that ticketing lead time will have less impact on the ticket price for other routes from other clusters.

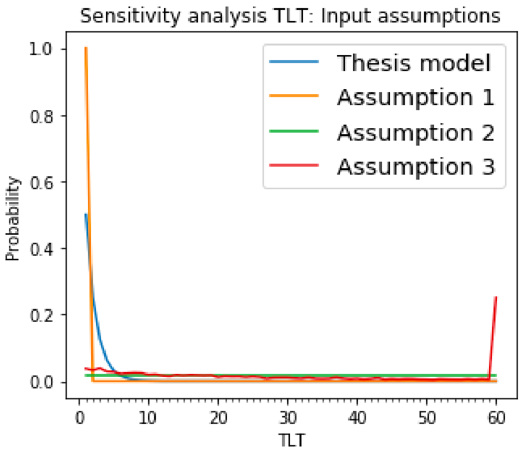


Figure L.1: Sensitivity analysis TLT: Input assumptions'

One would expect the most expensive prices for the first assumption because there the probability is 100 percent that someone buys 1 day in advance. Assumption 3 should be the cheapest since the distributions shows a peak at 60 days before departure. The thesis model is expected to be more expensive than the prices for assumption 2 since the center of gravity of the input distribution of the ticketing lead time for the thesis model is close to 1 day ticketing lead time.

The different prices are predicted for the 4 different input distributions. When someone selects ticketing lead time 60 days the final price will be the same for the 4 assumptions and when someone selects 1 day before departure, the price differences will be bigger since a customer can now book 1 day till 60 days in advance. Therefore, the price differences assuming the passengers selects 1 day as ticketing lead time is calculated and shown in Figure L.2.

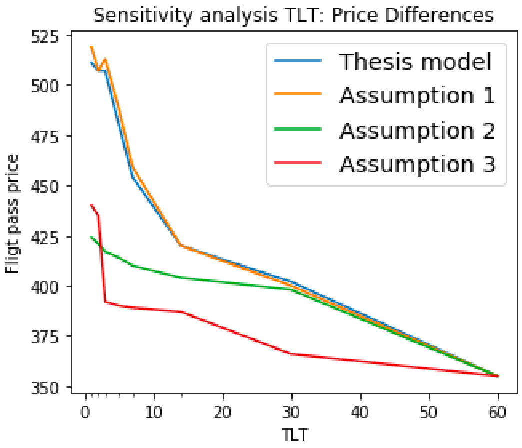


Figure L.2: Sensitivity analysis TLT: Price differences'

As one can see, the assumption of when people buy their ticket has a significant impact on the final price. A maximum price differences of 25% is observed at 3 days ticketing lead time. Therefore is is important to validate this ticketing lead time input distribution as soon as possible when there is data of the flight pass available.

Bibliography

[1] Auction iberia. `subastas.iberia.com`.

[2] Optiontown. `https://www.optiontown.com/`.

[3] Auction royal jordanian. `http://www.rj.com/en/auctions`.

[4] Rise - homepage. `https://www.iflyrise.com/`.

[5] Alaska air - homepage. `https://www.alaskaair.com/`.

[6] Air canada- flight pass. `https://www.aircanada.com/ca/en/aco/home/book/manage-bookings/flight-pass.html`.

[7] La compagnie - homepage. `https://www.lacompagnie.com/en`.

[8] Onego - homepage. `https://www.onego.com/`.

[9] Surfair - homepage. `https://www.surfair.com/eu/`.

[10] Wideroe - homepage. `http://www.wideroe.no/en`.

[11] The frequent fliers who flew too much.

[12] WC Allen and JK Zumwalt. Neural networks: a word of caution. *Unpublished Working Paper, Colorado State University*, pages 127–145, 1994.

[13] Héctor Allende, Claudio Moraga, and Rodrigo Salas. Artificial neural networks in time series forecasting: A comparative analysis. *Kybernetika*, 38(6):685–707, 2002.

[14] Paul D Allison. *Multiple regression: A primer*. Pine Forge Press, 1999.

[15] Evgeny A Antipov and Elena B Pokryshevskaya. Mass appraisal of residential apartments: An application of random forest for valuation and a cart-based approach for model diagnostics. *Expert Systems with Applications*, 39(2):1772–1778, 2012.

[16] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[17] Peter Belobaba. *Air travel demand and airline seat inventory management*. PhD thesis, Massachusetts Institute of Technology, 1987.

[18] Peter Belobaba, Amedeo Odoni, and Cynthia Barnhart. *The global airline industry*. John Wiley & Sons, 2015.

[19] Peter P Belobaba. Or practice—application of a probabilistic decision model to airline seat inventory control. *Operations Research*, 37(2):183–197, 1989.

[20] Guttery Benjamin, Randall and Sirmans. Mass Appraisal : An Introduction to Multiple Regression Analysis for Real Estate Valuation. 2017.

[21] Swapna Borde, Aniket Rane, Gautam Shende, and Sampath Shetty. Real estate investment advising using machine learning. 2017.

[22] R.A. Borst. Artificial neural networks in mass appraisal. *Journal of Property Tax Assessment Administration*, 1(2):5–15, 1995.

[23] T C Botimer and P P Belobaba. Airline pricing and fare product differentiation: A new theoretical framework. *Journal of the Operational Research Society*, 50(11):1085–1097, Nov 1999.

[24] Branko Božić, Dragana Milićević, Marko Pejić, and Stevan Marošan. The use of multiple linear regression in property valuation. *Geonauka*, 1(1):41–45, 2013.

[25] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[26] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[27] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[28] S. L. Brumelle and J. I. McGill. Airline seat allocation with multiple nested fare classes. *Operations Research*, 41(1):127–137, 1993.

[29] Ken Butcher, Beverley Sparks, and Frances O'Callaghan. Evaluative and relational influences on service loyalty. *International Journal of Service Industry Management*, 12(4):310–327, 2001.

[30] William Caicedo-Torres and Fabián Payares. *A Machine Learning Model for Occupancy Rates and Demand Forecasting in the Hospitality Industry*, pages 201–211. Springer International Publishing, Cham, 2016.

[31] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[32] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 96–103. ACM, 2008.

[33] Eric K. Clemons and Michael C. Row. Sustaining it advantage: The role of structural differences. *MIS Quarterly*, 15(3):275–292, 1991.

[34] H.A. Constantino, P.O. Fernandes, and J.P. Teixeira. Tourism demand modelling and forecasting with artificial neural network models: The mozambique case study. *Tékhne*, 14(2):113 – 124, 2016.

[35] Renwick E Curry. Optimal airline seat allocation with fare classes nested by origins and destinations. *transportation science*, 24(3):193–204, 1990.

[36] A Quang Do and Gary Grudnitski. A neural network approach to residential property appraisal. *The Real Estate Appraiser*, 58(3):38–45, 1992.

[37] Anthony W Donovan. Yield management in the airline industry. *Journal of Aviation/Aerospace Education & Research*, 14(3), 2005.

[38] Rita Marie Emmer, Chuck Tauck, Scott Wilkinson, and Richard G Moore. Using global distribution systems. *The Cornell Hotel and Restaurant Administration Quarterly*, 34(6):80–89, 1993.

[39] Marta Eso. An iterative online auction for airline seats. *IMA Volumes In Mathematics And Its Applications*, 127:45–58, 2001.

[40] Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75, 2002.

[41] Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh. Determinants of house price: A decision tree approach. *Urban Studies*, 43(12):2301–2315, 2006.

[42] Amy Farley. Jetblue announces 599 unlimited flight pass, 2009.

[43] Jodie L Ferguson, Pam Scholder Ellen, Jodie L Ferguson, and Pam Scholder Ellen. Pricing strategy & practice Transparency in pricing and its effect on perceived price fairness. 2014.

[44] E. Fix and J. Hodges. Discriminatory analysis. Non-parametric discrimination: Consistency properties. *USAF School of Aviation Medicine: Technical Report*, 4, 1951.

[45] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[46] G Gallego. A demand model for yield management. *Work-ing paper, Department of Industrial Engineering and Operations Research, Columbia University, New York*, 1996.

[47] Guillermo Gallego and Garrett van Ryzin. A multiproduct dynamic pricing problem and its applications to network yield management. *Oper. Res.*, 45(1):24–41, February 1997.

[48] Blanca García Gómez, Ana Gutiérrez Arranz, and Jesús Gutiérrez Cillán. The role of loyalty programs in behavioral and affective loyalty. *Journal of Consumer Marketing*, 23(7):387–396, 2006.

[49] Carlos Gershenson. Artificial neural networks for beginners. *arXiv preprint cs/0308031*, 2003.

[50] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[51] Nelson F Granados, Alok Gupta, and Robert J Kauffman. The impact of it on market information and transparency: A unified theoretical framework. *Journal of the Association for Information Systems*, 7(1): 7, 2006.

[52] Kevin P. Gwinner, Dwayne D. Gremler, and Mary Jo Bitner. Relational benefits in services industries: The customer's perspective. *Journal of the Academy of Marketing Science*, 26(2):101, Mar 1998.

[53] Russell I. Haley. Benefit segmentation: A decision-oriented research tool. *Journal of Marketing*, 32(3): 30–35, 1968.

[54] Geoffrey E Hinton. Connectionist learning procedures. In *Machine Learning, Volume III*, pages 555–610. Elsevier, 1990.

[55] Stay Foolish Isaac Changhau Stay Hungry. Activation functions in artificial neural networks, Jan 2018. URL `https://isaacchanghau.github.io/2017/05/22/Activation-Functions-in-Artificial-Neural-Networks/`.

[56] Jeeva Jose. *Customer Payment Trend Analysis Based on Clustering for Predicting the Financial Risk of Business Organizations*. Anchor Academic Publishing, 2017.

[57] Han-Bin Kang and Alan K Reichert. An empirical analysis of hedonic regression and grid-adjustment techniques in real estate appraisal. *Real Estate Economics*, 19(1):70–91, 1991.

[58] Tom Kauko, Pieter Hooimeijer, and Jacco Hakfoort. Capturing housing market segmentation: An alternative approach based on neural network modelling. *Housing Studies*, 17(6):875–894, 2002.

[59] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[60] Vilius Kontrimas and Antanas Verikas. The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, 11(1):443–448, 2011.

[61] Rob Law. Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management*, 21(4):331–340, 2000.

[62] Rob Law and Norman Au. A neural network model to forecast japanese demand for travel to hong kong. *Tourism Management*, 20(1):89 – 97, 1999.

[63] Anthony Owen Lee. Airline reservations forecasting: Probabilistic and statistical models of the booking process. Technical report, Cambridge, Mass.: Flight Transportation Laboratory, Dept. of Aeronautics and Astronautics, Massachusetts Institute of Technology, 1990.

[64] Kenneth Littlewood. Forecasting and control of passenger bookings. *Airline Group International Federation of Operational Research Societies Proceedings, 1972*, 12:95–117, 1972.

[65] Gilles Louppe. Understanding random forests: From theory to practice, 10 2014. arXiv:1407.7502.

[66] Kenneth M Lusht. *Real estate valuation: principles and applications*. KML publishing, 2001.

[67] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[68] Pietro Manzi and Paolo Barbini. From creativity to artificial neural networks: Problem-solving methodologies in hospitals. In Alfonso J. Rodriguez-Morales, editor, *Current Topics in Public Health*, chapter 05. InTech, Rijeka, 2013. doi: 10.5772/52582. URL `http://dx.doi.org/10.5772/52582`.

[69] Anna S Mattila. How and how much to reveal? The effects of price transparency on consumers' price perceptions. 31(4):530–545, 2007.

[70] William McCluskey and Sarabjot Anand. The application of intelligent hybrid techniques for the mass appraisal of residential properties. *Journal of Property Investment & Finance*, 17(3):218–239, 1999.

[71] Jeffrey I McGill and Garrett J Van Ryzin. Revenue management: Research overview and prospects. *Transportation science*, 33(2):233–256, 1999.

[72] Richard Merrell and David Diaz. Comparison of data mining methods on different applications: Clustering and classification methods. *Information Sciences Letters*, 4(2):61, 2015.

[73] Vahid Moosavi. Urban Data Streams and Machine Learning : A Case of Swiss Real Estate Market. *arXiv preprint arXiv:1704.04979*, 2017.

[74] Julio Gallego Mora-Esperanza. Artificial intelligence applied to real estate valuation: An example for the appraisal of madrid. *CATASTRO.*, 2004.

[75] Julian Pachon, Murat Erkoc, and Eleftherios Iakovou. Contract optimization with front-end fare discounts for airline corporate deals. *Transportation Research Part E: Logistics and Transportation Review*, 43(4):425–441, 2007.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[77] Nissan Pow, Emil Janulewicz, and L Liu. Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal. 2014.

[78] Transportation Review, Thomas Jay, Adler Rsg, and Resource Systems Group. Modelling airport and airline choice behaviour with the use of stated preference survey data Universities of Leeds , Sheffield and York. 2007.

[79] Raul Rojas. The backpropagation algorithm. In *Neural networks*, pages 149–182. Springer, 1996.

[80] Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25:27, 1995.

[81] Garrett J Van Ryzin. *An Introduction to Revenue Management*. 2005.

[82] D. K Skwarek. Competitive impacts of yield management system components: Forecasting and sell-up models. Technical report, Cambridge, Mass.: Flight Transportation Laboratory, Dept. of Aeronautics and Astronautics, Massachusetts Institute of Technology, 1996.

[83] Laerd Statistics. Multiple regression analysis using spss statistics. *Laerd Research Ltd*, 2013.

[84] Kalyan T Talluri and Garrett J Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Science & Business Media, 2006.

[85] Danny P.H. Tay and David K.H. Ho. Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation and Investment*, 10(2):525–540, 1992.

[86] Richard Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39, 1990.

[87] Garrett J van Ryzin. Future of revenue management: Models of demand. *Journal of Revenue and Pricing Management*, 4(2):204–210, 2005.

[88] L R Weatherford, T W Gentry, and B. Wilamowski. Neural network forecasting for airlines: A comparative analysis. *Journal of Revenue and Pricing Management*, 1(4):319–331, 2003.

[89] Larry Weatherford. The history of forecasting models in revenue management. *Journal of Revenue and Pricing Management*, 15(3):212–221, 2016.

[90] Larry R Weatherford and Sheryl E Kimes. A Comparison of Forecasting Methods for Hotel Revenue Management. 2003.

[91] R. R. Wickham. Evaluation of forecasting techniques for short-term demand of air transportation. Technical report, Cambridge, Mass.: Flight Transportation Laboratory, Dept. of Aeronautics and Astronautics, Massachusetts Institute of Technology, 1995.

[92] Richard Robert Wickham. *Evaluation of forecasting techniques for short-term demand of air transportation*. PhD thesis, Massachusetts Institute of Technology, 1995.

[93] Elizabeth Louise Williamson. *Airline network seat inventory control: Methodologies and revenue impacts*. PhD thesis, Massachusetts Institute of Technology, 1992.

[94] DG Wiltshaw. Valuation by comparable sales and linear algebra. *Journal of Property Research*, 8(1):3–19, 1991.

[95] Jeffrey M Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2015.

[96] Elaine Worzala. Currency risk and international property investments. *Journal of Property Valuation and Investment*, 13(5):23–38, 1995.

[97] Jozef Zurada, Alan Levitan, and Jian Guan. A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research*, 33(3):349–387, 2011.