Analysing Data Features on Algorithmic Fairness in Machine Learning

Comparing the sensitivity of data features under fairness properties between different sectors

Pavlos Markesinis

Supervisor: Anna Lukina

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Pavlos Markesinis
Final project course: CSE3000 Research Project
Thesis committee: Anna Lukina, Christoph Lofi

Abstract

Fairness in machine learning is an increasingly important yet complex issue, especially as these algorithms are integrated into critical decision-making processes across various sectors. This research focuses on the impact of features under fairness properties across multiple sectors. The primary research question addressed is: "Which data features are the most sensitive when monitoring fairness properties on criminal data, and how do these features perform when monitoring fairness properties on data from different sectors?" The study examines features such as age, $race$, $gender$, and $educational\ level$ across datasets from criminal justice, healthcare, finance, and education sectors. Utilizing logistic regression models and a proposed dynamic monitoring algorithm, sensitivity of features to fairness violations is assessed for Demographic Parity and Equal Opportunity properties. The findings indicate that age is the most sensitive feature in almost all sectors, highlighting inherent biases and the necessity for sector-specific fairness considerations. However, statistical analysis revealed that these differences in sensitivity values across sectors were not statistically significant, suggesting that the observed patterns are not strong enough to be deemed conclusive.

1 Introduction

Fairness in computer science has been both crucial and a difficult topic for everyone to agree on. On the one hand, with machine learning algorithms being used more and more in decision-making algorithms such as approving bank loans, or recommending criminal sentencing, it is extremely important to incorporate fairness to avoid bias and discrimination. On the other hand, fairness is still not an easy problem to solve, given that there are no objective definitions of it. Most definitions of fairness rely on interest [6] and not surprisingly some of them contradict each other [12]. Therefore, the need for investigating fairness in computer science continues to grow. As technology advances and embeds even more automated decision-making algorithms in everyday life, their social importance is increasing as well.

An approach of incorporating fairness into a machine learning algorithms is to set a specific fairness properties in the beginning and examine if they are satisfied at the end. To provide a straightforward example, let a machine learning algorithm X, which predicts whether a person is likely to default on a loan, used commonly in the financial industry, and a fairness property y which ensures that there is no discrimination against minorities, known as Demographic Parity. Assume that the majority and minority groups in this case are $male$ and $female$ respectively. After obtaining the predictions of X, the property y could be evaluated by comparing the conditional probabilities between the minority and the majority group, and establishing a threshold c as the boundary of this fairness property. In other words, expressed mathematically [1]:

$$\frac{\text{P[loan approval} \mid \text{gender = female]}}{\text{P[loan approval} \mid \text{gender = male]}} \geq c \qquad (1)$$

However, an innovative method involves not examining the fairness properties at the end, but continuously monitoring them during runtime of the machine learning algorithm. Dynamic monitoring of fairness is indeed an active area of research within machine-learning fairness[9]. This paper focuses on the dynamic approach of monitoring fairness properties during runtime of machine learning algorithms. It also investigates whether dataset features have an impact on fairness in machine learning. This is achieved by comparing the performance of features across different sectors. Thus, the research question addressed is:

> "Which data features are the most sensitive when monitoring fairness properties on criminal data, and how do these features perform when monitoring fairness properties on data from different sectors?"

To provide some definitions on this context, data features refer to common attributes across datasets, such as race, age, etc. Fairness properties are rules defined to ensure fairness in an automated decision-making algorithms, like parameter y in the previous example. A fairness property violation implies that there is bias in the algorithm. For example, in Inequality 1, a violation would be a ratio that is less than the threshold c. Sensitivity refers to the extend of violating the defined fairness properties. Thus, the most sensitive feature would violate a fairness property the most. Finally, we examine features on four different sectors: firstly on criminal justice, and then on healthcare, finance and education.

Furthermore, to address the research question, we pose two main sub-questions:

1. Which data features are the most sensitive when monitoring fairness properties on criminal data?
2. How do these features perform when monitoring the same fairness properties on data from healthcare, finance and education?

[1] This expression is not definitive and can vary depending on the fairness properties; this is merely an example.

By addressing these sub-questions, this research aims to provide a detailed understanding of the most sensitive features in the chosen criminal dataset and how these features perform in different datasets. Comparing the sensitivity of each feature across different domains can help identify patterns. This analysis identified age as the most influential feature, followed by $gender$ with $education\ level$, when establishing fairness in each dataset. Lastly, for dynamically monitoring fairness properties we propose an algorithm[2] written in Python, inspired from Verifair[3][2], which is further analysed in subsection 3.3.

2 Related Work

There has been previous work on establishing fairness in machine learning algorithms. Machine learning models on a crowd sourced platform exhibit bias and the critical need for fairness in machine learning applications has been verified [4]. Many top-rated predictive models from Kaggle[4] are biased inadvertently by optimization techniques [4]. This supports the motive of this research based on fairness. Furthermore, another relevant topic is fairness cost. Some of the trade-offs associated with incorporating fairness into machine learning models includes compromising on other critical performance metrics [17]. This highlights the complexity of bias in machine learning and the need to explore data features to mitigate it.

Moreover, related work on dynamic monitoring fairness shows that embedding fairness directly into the code and using runtime monitoring can significantly reduce biases in automated decision-making [1]. Also, studies about the long-term effects of fairness criteria in machine learning used a similar one-step feedback model to understand how these criteria impact the well-being of disadvantaged groups over time [11]. The findings revealed that while fairness constraints are intended to protect vulnerable groups, they can sometimes lead to unintended negative consequences, especially when applied without careful consideration of temporal effects and measurement accuracy [11]. Furthermore, simulations revealed that static fairness assessments are insufficient for understanding the true impacts of machine learning decisions, highlighting the need for a nuanced approach that considers the evolving nature of policies and their interaction with the system's state in real-world environments [5]. A study examined dataset features and concluded that using correlated non-private features can effectively reduce bias and ensure fairness in machine learning models without the need for private attributes. [17]. Lastly, the importance of evaluating fairness in real-time has been highlighted, as systems interacting with humans can develop biases over time [9].

As shown in the previous paragraph, this topic has been researched thoroughly. Our paper builds on the findings of the highlighted related work. It aligns with the highlighted trade-offs associated with incorporating fairness into machine learning models [17] by examining fairness properties dynamically. Furthermore, it follows the method of embedding fairness directly into code [5] by continuously monitoring data over time, and aligns with the analysis of the long-term effects of fairness criteria [11]. This proactive and dynamic monitoring approach addresses the limitations mentioned above, emphasizing the importance of evolving policies that interact with the system's state [5]. Finally, this research expands on the paper highlighting that incorporating fairness into machine learning models can compromise other metrics [17] by examining the impact of various dataset features on fairness.

3 Methodology

This section outlines the methodology we employed in our experiments to address the research question. Firstly, each dataset was prepared for the machine learning algorithm by splitting the data, one-hot encoding categorical variables, filling missing values, and creating age ranges including all the values. Secondly, the fairness properties were Demographic Parity and Equal Opportunity, both defined with mathematical inequalities from which sensitivity values were calculated. Lastly, the proposed algorithm received a number of iterations, based on which the split data are divided into batches, and for each iteration the sensitivity values are calculated for both fairness properties.

3.1 Dataset Acquisition and Preparation

Our research utilizes four public datasets, all from distinct domains: criminal justice [8], healthcare [3], finance [13], and education [7] [14]. Each dataset addresses a specific societal issue: recidivism[5] within three years for criminal justice, lung cancer diagnosis for healthcare, loan repayment for finance, and student dropout rates for education. The machine learning algorithms were trained to predict the outcome represented by the societal issue in each dataset.

The specific features examined under the fairness properties are age, $race$, $gender$ and $educational\ level$. These features are fully present in the criminal justice dataset and partially in the other sectors. For example, the education dataset includes age, $gender$ and $educational$

[2] This algorithm is available on this GitLab repository .

[3] VeriFair is a tool implemented for probabilistic verification of fairness properties in machine learning models.

[4] An online community of data scientists and machine learning engineers.

[5] Recidivism is the tendency of a convicted criminal to re-offend. The machine learning algorithm predicting recidivism returns either low or high risk.

$level$, but does not include $race$. This variation allows for a comparative analysis of how each feature impacts fairness across different domains.

The preparation of each dataset before it was input to the machine learning model for prediction was the following:

- Split the data into training and test sets, X and y respectively, by excluding the target variable column from X and assigning it to y.
- One-hot encode categorical variables. Transform categorical data into a numerical format that machine learning algorithms can process. It ensured that the categorical variables contribute appropriately to the model's predictions without introducing unintended biases.
- Identify numerical and categorical columns.
- Fill missing values: for numerical columns, use the mean to maintain consistency and avoid bias; for categorical columns, use the mode to represent the most frequent category.
- Create age ranges for all age values in the datasets. The ranges were aimed to be roughly equal in size, but they varied for each dataset. For example, the age ranges for the students in the student dropout dataset were "17-18", "19-20" and "21 or older".

3.2 Fairness Properties and Sensitivity

In this research, the experiments continuously monitor two fairness properties. The first one is a version of Demographic Parity, commonly known as the "80% rule,". This rule assesses Demographic Parity by checking whether the rate of a minority group is at least 80% of the rate of a majority group. For example, in the context of recidivism in criminal justice, this rule mandates that the proportion of minority individuals deemed low risk should be at least 80% of the proportion of majority individuals deemed low risk. By reusing Inequality 1, with threshold $c = 0.8$ and $male$ and $female$ as majority and minority groups respectively, the fairness property is:

$$\frac{\text{P[low risk} \mid \text{gender = female]}}{\text{P[low risk} \mid \text{gender = male]}} \geq 0.8 \tag{2}$$

Furthermore, the second fairness property is Equal Opportunity. Equal Opportunity ensures that individuals who qualify for a positive outcome have an equal chance of being selected, irrespective of their demographic group. This principle is crucial in preventing discrimination against qualified individuals based on the features such as age, gender, etc. For instance, in the context of healthcare, Equal Opportunity requires that the probability of diagnosing lung cancer, given that an individual actually has lung cancer, should be approximately the same for all demographic groups. Mathematically, this can be expressed as:

$$\begin{gathered}|\text{P[diagnosed} \mid \text{lung cancer, age-group = A]}- \\ \text{P[diagnosed} \mid \text{lung cancer, age-group = B]}| \leq \epsilon\end{gathered} \tag{3}$$

Since it is impractical to expect two probabilities to be exactly equal to verify fairness, the absolute difference between the two groups is calculated and a threshold ϵ is set as the boundary, demonstrated in Inequality 3. In this research, we used $\epsilon = 0.05$ for assessing Equal Opportunity.

Those two properties are established as the fairness criteria for the experiments of this research. The sensitivity of each issue is calculated by using the appropriate inequality for each property, like Inequality 2 and Inequality 3, and taking the extend of the violation into account. For example, assume a Demographic Parity violation by using Inequality 2, and $\frac{\text{P[low risk|gender = female]}}{\text{P[low risk|gender = male]}} = 0.6$. Since the inequality does not hold, there is a violation. Moreover, the sensitivity in this case is $0.8 - 0.6 = 0.2$. Accordingly, in an Equal Opportunity violation, take Inequality 3 and assume $|\text{P[diagnosed} \mid \text{lung cancer, age-group = A]}- \text{P[diagnosed} \mid \text{lung cancer, age-group = B]}| = 0.15$. With $\epsilon = 0.05$ then Inequality 3 does not hold. In this case, sensitivity is $0.05 - 0.15 = -0.1$. The sensitivity of each feature in Equal Opportunity is considered with opposite signs. So in this example, sensitivity would be estimated as $-(-0.1) = 0.1$. The reason for this choice is to avoid confusion and establish a general rule for the sensitivity in our experiments; *if the sensitivity is positive, there is a fairness violation*. Albeit, it is important to note that the features' sensitivity values between fairness properties are not compared, since they come from different inequalities with different thresholds as boundaries (c and ϵ). The comparison of sensitivities involves only the same feature and fairness property, but different sectors. For example, comparing the sensitivity of age under Equal Opportunity in criminal justice, with the sensitivity of age under Equal Opportunity in healthcare.

3.3 Machine Learning Model and Fairness Assessment Implementation

This subsection addresses the main implementation of the algorithm proposed, in addition to the machine learning model used for each dataset.

All the datasets were treated the same. For the machine learning model, we used Logistic Regression to predict the target variable for each dataset. Furthermore, we created a class, $LogisticModel$, which is initialized with a Logistic Regression model, the test data of the dataset, and a number representing the iterations. The purpose of iterations is to simulate dynamic monitoring of fairness over time. Then, the $LogisticModel$ class splits the test data into batches, equal to the number of iterations. For each iteration, it used the Logistic Regression model to predict one batch of the test data.

Moreover, the training and test data were separated based on the values of each feature. For instance, if the feature was $gender$, the data were divided into separate training and test sets for $male$ and $female$. Then, these datasets were standardized to ensure uniformity in feature scales before used by the logistic regression model for initialization and fitting.

Our proposed algorithm calculates and stores the sensitivity of the selected feature under both fairness properties (Demographic Parity and Equal Opportunity), on each iteration. As the iterations progress, the algorithm receives the predictions of the new batch from the $LogisticModel$ and adds them to the total predictions accumulated so far. This approach ensures that, with each iteration, the conditional probabilities used to evaluate definitions like Inequality 2 and Inequality 3 are continuously updated, and new sensitivity values are stored. Thus, the algorithm's results include all the sensitivity values, for each feature, from the first to the last iteration.

As briefly mentioned in the introduction, this algorithm was inspired from Verifair [2]. The initial plan was to utilize that tool. However, despite the similarities between the goals of the research paper in which it was employed and our research, certain differences compelled us to develop our own algorithm inspired by it. The main difference is that fairness properties established on Verifair are different, something which made it difficult for adapting it to our experiments. Also, Verifair does not use real datasets, instead it generates samples randomly. This research examines how various features influence datasets to identify impactful patterns, unlike random data generation, which yields unreliable results.

4 Results

In this section we address the experiments of the research and their findings. Then, we answer the two research sub-questions by using the findings of the experiments. There are two experiments in total, each aiming to reveal findings for the two research sub-questions raised in the Introduction.

4.1 Experiments

The first experiment reveals the sensitivity values of the selected features from the criminal justice dataset. Sensitivity for each feature is calculated for both Demographic Parity and Equal Opportunity, as described in subsection 3.2.

The second experiment extends the sensitivity analysis of the features in the other datasets: healthcare, finance, and education. Again, the sensitivity of each feature in these datasets is calculated twice , one for each fairness property. Lastly, the number of iterations is $n = 100$, and the threshold values for the boundaries of Demographic Parity and Equal Opportunity are $c = 0.8$ and $\epsilon = 0.05$ respectively.

Experiment 1: Sensitivity of Features in the Criminal Justice dataset

In this experiment, we evaluated the sensitivity of four selected features: $race$, age, $gender$, and $educational\ level$. Recall that the criminal justice dataset used in this study focused on predicting recidivism within three years. The results for Demographic Parity are depicted in Figure 1. Each subplot in the figure represents the sensitivity values in 100 iterations for different pairs of minority and majority groups. All pairs in Demographic Parity graphs, have the minority group first and the majority group second. For example, in the top right subplot of Figure 1, the only pair examined is $white\ vs\ black$. As explained, $white$ is the minority group and $black$ is the majority group. This means that there are more rows in the dataset where $race = black$ than $race = white$. Furthermore, take the bottom left subplot of Figure 1, which demonstrates the sensitivity values of age. This feature has three different pairs being examined. As we observe, the pair $43\_or\_older\ vs\ 18\_to\_27$ is the most sensitive with its value converging just over 0.2 after the last iteration. In order to evaluate these findings, and answer the posed research question, we must have one sensitivity value per subplot below. Thus, in cases where multiple pairs are examined, such as age, the pair with the highest sensitivity is chosen as the representative. Lastly, from each subplot we choose the sensitivity of the last iteration. This approach is chosen because each sensitivity value in the iteration accounts for the preceding iterations. Therefore, using the last value is more meaningful than using the mean, as it represents all the data and previous iterations.

Furthermore, the results for Equal Opportunity are depicted in Figure 2. Similarly with Demographic Parity, each subplot in the figure represents the sensitivity of the possible pairs between the values of each feature over 100 iterations. The difference between Figure 1 and Figure 2 is the fairness properties defined by Inequality 2 and Inequality 3. Observing once more one of the subplots, the bottom right one is demonstrating the three different pairs of $educational\ level$. In this case, the representative pair chosen is $college\ vs\ hs\_diploma$ since it ends up with the highest sensitivity among the rest, approximately 0.15

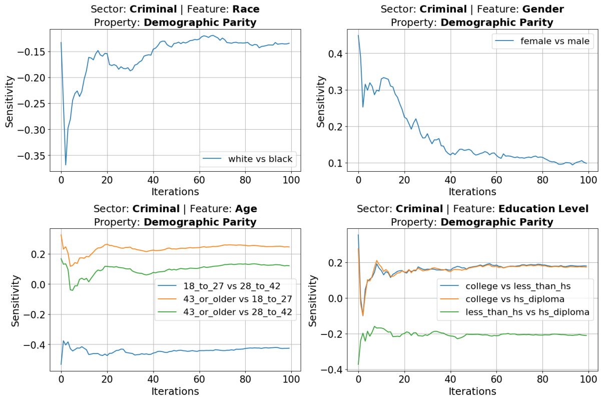

Figure 1: Demographic Parity - Features in the criminal justice. Top left is *race*, top right is *gender*, bottom left is *age* and bottom right is *education level*. Each line in the subplots represents the sensitivity values over the iterations for each pair.

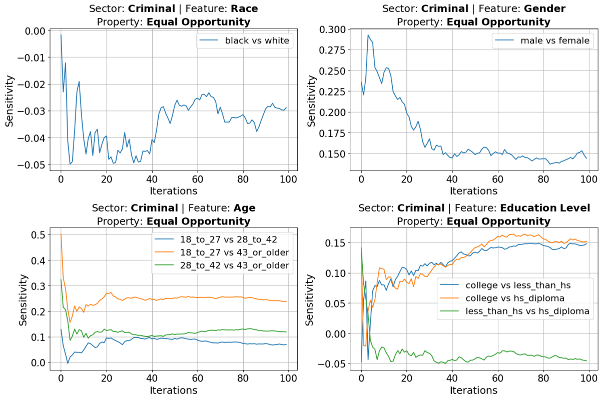

Figure 2: Equal Opportunity - Features in the criminal justice dataset. Top left is *race*, top right is *gender*, bottom left is *age* and bottom right is *education level*. Each line in the subplots represents the sensitivity values over the iterations for each pair.

Experiment 2: Sensitivity of Features in the other datasets

The second experiment extended the sensitivity analysis to healthcare, finance, and education datasets. The same features (*age*, *gender*, and *educational level*) were evaluated, with sensitivity calculated for both fairness properties in 100 iterations. To recall the objective of each dataset: the healthcare dataset aimed to predict whether a person has lung cancer, the finance dataset focused on predicting whether an individual will repay their loan, and the education dataset aimed to determine whether a student will drop out or not. This experiment holds a crucial position in our research. It can help identify patterns in the features across the datasets. All the graphs are included in Appendices B to D, from Figure 8 to Figure 13. The detailed results are analysed in the next subsections.

4.2 Which data features are the most sensitive when monitoring fairness properties on criminal data?

To address which data features are the most sensitive under the criminal justice sector, we evaluate the final sensitivity of the feature pairs in the 100 iterations. Recall that, for both fairness properties, if the sensitivity is above zero, then there is a fairness violation. Table 1 demonstrates the final sensitivity of each feature for Demographic Parity and Equal Opportunity from the subplots of Figure 1 and Figure 2. As explained, the sensitivity of *age* in the criminal dataset is 0.2467 for Demographic Parity and 0.2379 for Equal Opportunity, since the those are the highest final sensitivity values among all pairs.

Feature	**Demographic Parity**		**Equal Opportunity**	
	Pair	**Sensitivity**	**Pair**	**Sensitivity**
Race	white **vs** black	**-0.1346**	black **vs** white	**-0.0289**
Gender	female **vs** male	**0.0987**	male **vs** female	**0.1441**
Age	18 to 27 **vs** 28 to 42	-0.4259	18 to 27 **vs** 28 to 42	0.0693
	43 or older **vs** 18 to 27	**0.2467**	18 to 27 **vs** 43 or older	**0.2379**
	43 or older **vs** 28 to 42	0.1217	28 to 42 **vs** 43 or older	0.1186
Education Level	college **vs** less than hs	**0.1794**	college **vs** less than hs	0.1478
	college **vs** hs diploma	0.1730	college **vs** hs diploma	**0.1520**
	less than hs **vs** hs diploma	-0.2103	less than hs **vs** hs diploma	-0.0458

Table 1: Criminal Justice - Sensitivities for each feature pair for both fairness properties

To directly answer the first research sub-question, we compare the representative sensitivity values of each feature from Table 1. The results are illustrated in a feature ranking, in Figure 3. For Demographic Parity, *age* emerges as the most sensitive feature, followed by *education level*, *gender*,and finally *race*. Similarly, for Equal Opportunity, *age* remains the most sensitive feature, followed by *gender*, *education level*, and *race* in the same order. This consistent ranking of features by sensitivity across both fairness properties suggests that these features play a significant role and indicate an inherent bias within the dataset. Moreover, the only feature that satisfies the fairness properties is *race*. All of the rest features violate both fairness properties, since their sensitivity values are positive.

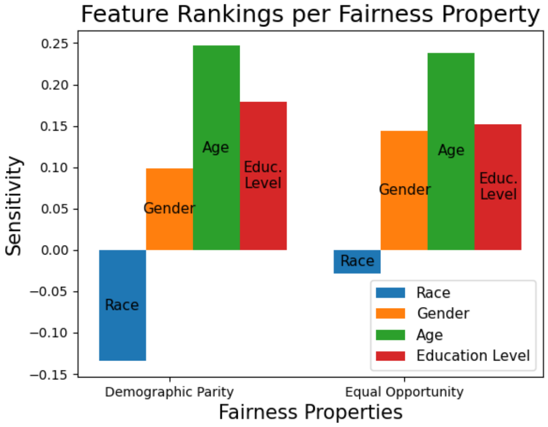

Figure 3: Feature rankings for Demographic Parity and Equal Opportunity in Criminal Justice.

4.3 How do these features perform when monitoring the same fairness properties on data from healthcare, finance and education?

Similarly with subsection 4.2, Table 2 demonstrates the final sensitivity of each feature across healthcare, finance, and education. For each feature, the highest sensitivity value is highlighted in bold. It should be noted that when comparing to other sectors, some features are not included. For example, *education level* is only examined in Finance. To the extend of this issue, *race* is not examined in any of the other sectors. This problem is thoroughly discussed in section 6, Limitations.

To evaluate the meaning of those numbers, we refer to the feature rankings across all sectors. Figure 4 and Figure 5 demonstrate the feature rankings of Demographic Parity and Equal Opportunity respectively. To begin with Demographic Parity, in healthcare, *gender* is the most sensitive feature, followed by *age*. However, both of them have a negative sensitivity, which means that there are no fairness violations. Education has the opposite feature ranking of healthcare, so *age* is the most sensitive with *gender* following with a major difference in sensitivity. In this sector though, in *age* there is a fairness violation of 0.4921. Lastly, in Finance, the ranking of the features includes *age* as the most sensitive, followed by *education level* and *gender*. However, the whole sector of Finance in Demographic Parity does not include any fairness violation, which means that the fairness property is satisfied across all features, and thus no bias might be inherited in the algorithm. Last but not least, comparing these three sectors with the criminal one, we observe certain facts. Even though healthcare, education and finance have only negative sensitivity values with only one exception (*age* in education), the most sensitive feature has remained the same in almost all sectors (except healthcare), for Demographic Parity, and that is *age*.

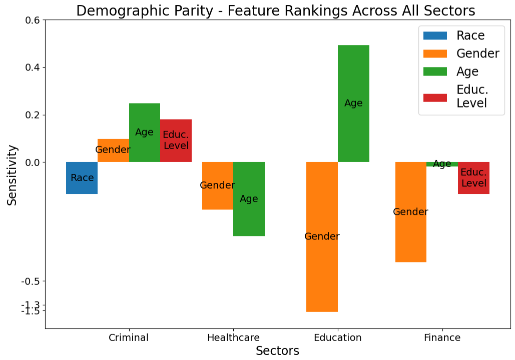

Figure 4: Demographic Parity - Feature rankings across all sectors.

Focusing accordingly in Equal Opportunity, the results are depicted in Figure 5. In healthcare, *age* is the most sensitive feature, and second comes *gender* with a major difference in sensitivity. In education, we observe once more that *age* is the most sensitive feature, and then again *gender* comes next, this time with a smaller difference in sensitivity. Lastly in finance, we notice the only case where *gender* is the most sensitive feature, followed by *age* and *education level*. For Equal Opportunity, each feature except from *race* violates the fairness property across all sectors. Again, *age* is the most sensitive feature in almost all sectors (except finance).

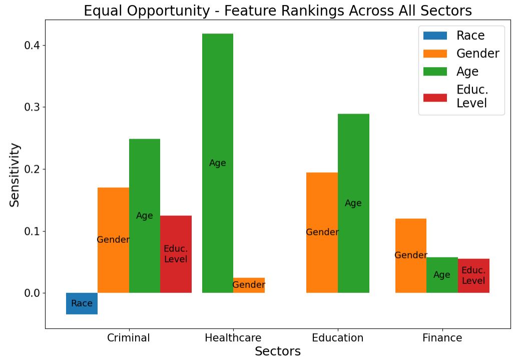

Figure 5: Equal Opportunity - Feature rankings across all sectors.

The overall conclusion from the experiment's findings can be drawn from the feature rankings. There is some evidence suggesting that *age* has higher impact than the

Sector	Feature	Demographic Parity		Equal Opportunity	
		Pair	Sensitivity	Pair	Sensitivity
Healthcare	Gender	female **vs** male	**-0.2000**	male **vs** female	**0.0167**
	Age	21 to 57 **vs** 63 to 69	**-0.3111**	21 to 57 **vs** 58 to 62	-0.0083
		58 to 62 **vs** 63 to 69	-0.3667	21 to 57 **vs** 63 to 69	-0.0500
		70 or older **vs** 21 to 57	-0.4353	21 to 57 **vs** 70 or older	**0.2357**
		70 or older **vs** 58 to 62	-0.3667	58 to 62 **vs** 63 to 69	-0.0083
		70 or older **vs** 63 to 69	-0.5125	58 to 62 **vs** 70 or older	0.1881
				63 to 69 **vs** 70 or older	0.1881
Education	Gender	male **vs** female	**-1.5357**	male **vs** female	**0.2380**
	Age	17 to 18 **vs** 19 to 20	0.1109	17 to 18 **vs** 19 to 20	0.0181
		17 to 18 **vs** 21 or older	**0.4921**	17 to 18 **vs** 21 or older	**0.3067**
		19 to 20 **vs** 21 or older	0.3531	19 to 20 **vs** 21 or older	0.2386
Finance	Gender	female **vs** male	**-0.4222**	male **vs** female	**0.1167**
	Age	18 to 28 **vs** 29 to 39	**-0.0182**	18 to 28 **vs** 29 to 39	0.0879
		40 or older **vs** 18 to 28	-0.4500	18 to 28 **vs** 40 or older	**0.0929**
		40 or older **vs** 29 to 39	-0.2000	29 to 39 **vs** 40 or older	-0.0500
	Education Level	hs or below **vs** college	**-0.1348**	hs or below **vs** bachelor or above	**0.0864**
		bachelor or above **vs** hs or below	-0.2588	hs or below **vs** college	0.0294
		bachelor or above **vs** college	-0.2000	college **vs** bachelor or above	-0.0500

Table 2: Sensitivities for Each Feature Pair Under Demographic Parity and Equal Opportunity in Healthcare, Education, and Finance.

other features.

5 Evaluation

In this section we evaluate the findings from section 4 and examine whether there is a statistical significance, between the different sensitivities across sectors for each feature. By using a Mann-Whitney U Test, we ultimately reveal that the differences in sensitivity values are not statistically significant.

The rationale for using Mann-Whitney U Test is based on the context of this research. Firstly, Mann-Whitney U Test is a non-parametric statistical test to compare differences between two independent groups on an ordinal outcome [10]. Non-parametric statistical test is ideal for this research because it does not assume a normal distribution of the data. In our case, the data were at most four sensitivity values per feature, one for each dataset that was utilized. Thus, with only four values we neither tested normal distribution nor assumed it, since the results would be misleading and unreliable. So, statistical tests like t-test or ANOVA were not considered, since they are parametric and assume normal distribution of the data. Furthermore, Mann-Whitney U Test can be used with ordinal data[6] [16] which completely aligns with the feature rankings used in subsections 4.2 and 4.3. Lastly, it is suitable for small sample sizes [15].

To statistically test the patterns revealed from the feature rankings, we initially set a null and an alternative hypothesis:

- Null Hypothesis (H_0): The sensitivity of data features (age, $gender$, $educational\ level$) when monitoring fairness properties does not significantly differ across sectors.
- Alternative Hypothesis (H_1): The sensitivity of data features (age, gender, educational level) when monitoring fairness properties significantly differs across sectors.

Then, we collected the sensitivity values of each feature from Figure 3, Figure 4 and Figure 5. Once more, this was done for both fairness properties separately. For example, for $gender$, the sensitivity values collected were 0.0987, -0.200, -1.5357 and -0.4222 for Demographic Parity. Additionally, for Equal Opportunity, the collected sensitivity values were 0.1441, 0.0167, 0.2380 and 0.1167. These values can be found highlighted in Table 1 and Table 2.

After collecting all the sensitivity values for all features in both fairness properties, we used the Mann-Whitney U Test to calculate the u_stat and p value of all possible feature pairs. The u_stat is a statistic which represents the number of times a value from one group precedes a value from another group in the ranked data. The u_stat ranges from 0 to $n_1 * n_2$ where n_1 and n_2 are the sample sizes of the two groups respectively. A u_stat value near 0 suggests a large difference between the groups, while a value near the maximum value suggests little to no difference. The p value indicates the probability of observing the test results under the null hypothesis. As typically used, the significance level was set to 5%. Thus, if p was less than 0.05 then the null hypothesis H_0 could be rejected. The results of the Mann-Whitney U Test are demonstrated in Table 3.

[6] Data that can be ranked.

Feature Pair	**Demographic Parity**		**Equal Opportunity**	
	u-stat	**p-value**	**u-stat**	**p-value**
Age & Gender	13.0	0.20	11.0	0.49
Age & Education Level	5.0	0.80	7.0	0.26
Gender & Education Level	1.0	0.26	4.0	1.00

Table 3: Statistical test - Mann Whitney U Test between all the feature pairs.

The results of the Mann-Whitney U Test for Demographic Parity reveal that the sensitivity values of age, $gender$, and $educational\ level$ do not significantly differ across the sectors. Specifically, the u_stat of 13.0 and p value of 0.20 for Age & Gender suggest a medium level of overlap in the ranks of sensitivities, indicating no statistically significant difference. Similarly, for Age & Education Level, the u_stat of 5.0 and p value of 0.80 show a higher degree of overlap, reaffirming the lack of significant difference. For Gender & Education Level, the u_stat of 1.0 and p value of 0.26 indicate the greatest overlap among the pairs, further supporting the null hypothesis H_0. Therefore, for Demographic Parity, we conclude that the differences in sensitivity values of these features across sectors are not statistically significant.

The analysis for Equal Opportunity also supports the null hypothesis H_0 that there is no significant difference in the sensitivity values of age, $gender$, and $educational\ level$ across the different sectors. The u_stat of 11.0 and p value of 0.49 for Age & Gender indicate a medium overlap in ranks, suggesting no significant difference in sensitivities. For Age & Education Level, the u_stat of 7.0 and p value of 0.26 reflect a substantial overlap, further indicating no significant difference. Lastly, the u_stat of 4.0 and p value of 1.00 for Gender & Education Level show a high degree of overlap, reaffirming the absence of significant differences. Thus, for Equal Opportunity, we also conclude that the observed differences in sensitivity values across sectors are not statistically significant, indicating that the sensitivities of these features are consistent across the various domains studied.

6 Limitations

This section refers to the main limitations we faced during the research. Those were the datasets, which played a vast role in this paper, and the limited values of certain features within those datasets.

To begin with the datasets, they were the biggest limitation of this research. As one can observe, feature $race$ was not examined in any dataset other than the criminal justice. The reason was due to the lack of real-world datasets including all the needed features and the column with a societal issue (like recidivism, lung cancer diagnosis, etc.) together. Most datasets included either a societal issue with limited features, for example only $gender$, or a plethora of features but without a societal issue to be combined with a machine learning algorithm and a fairness property. This was the reason that $education\ level$ was also not examined in the healthcare and education sectors as well.

Moreover, another limitation was that within the chosen datasets, certain feature values were unbalanced. For example, in the finance sector, under the feature $gender$, there are two values: $male$ and $female$. However, the $male$ rows are equal to 423 while the $female$ rows are equal to 77. This caused problems on monitoring fairness over time and for certain models splitting the data into batches for the $LogisticModel$ became problematic.

7 Conclusions and Future Work

This research analyzed the sensitivity of data features under fairness properties in machine learning algorithms across multiple sectors. By dynamically monitoring these properties, it was found that age emerged as the most sensitive feature in almost all sectors, underscoring its significant impact on fairness. $Gender$ and $educational\ level$ also showed considerable sensitivity, though to a lesser extent. The consistent ranking of these features across sectors highlights their importance when striving for fairness in machine learning models. However, statistical analysis revealed that the differences in sensitivity values across sectors were not statistically significant, suggesting that while there may be observable patterns, they are not strong enough to be deemed conclusive. These insights contribute to a better understanding of how different features influence fairness, providing a foundation for considering them when setting fairness properties, while also emphasizing the need for further investigation to confirm these patterns.

Future research should address the limitations encountered in this study, particularly the availability and completeness of datasets. Expanding the scope to include a more diverse range of datasets and features, such as incorporating $race$ in sectors beyond criminal justice, would provide a more comprehensive analysis. Another potential direction is the application of real-time fairness interventions based on continuous monitoring, which could dynamically adjust the algorithm to mitigate bias as it arises. Lastly, investigating the impact of feature interactions on fairness and considering it in fairness properties could be useful for advancing this field.

8 Responsible Research

In this section, we reflect on the ethical aspects of this research, and discuss the reproducibility of the proposed algorithm.

8.1 Resources

The resources used for this research are the datasets from the different sectors, which all are publicly available [14] [7] [3] [6] [13]. The algorithm was inspired from Verifair [2], which is publicly available as well. The datasets were not intentionally modified to influence the results. Standard pre-processing procedures, such as filling missing values with the column's mean, were applied. All pre-processing steps have been reported in subsection 3.1.

Furthermore, ChatGPT from OpenAI was employed to enhance the stylistic elements of this paper and to perform simple tasks within the code, such as writing comments, without contributing to problem-solving.

8.2 Reproducibility

The repository of this research is publicly available on GitLab[7]. The reproducibility of our research is ensured through the use of publicly available datasets and a clear description of the proposed algorithm. The preprocessing steps, fairness properties, and sensitivity calculations are explicitly outlined in section 3, allowing other researchers to replicate our experiments. All datasets are included in the $data$ folder of the repository on GitLab, ensuring easy access for verification and further studies.

To further support reproducibility, the repository on GitLab includes a README file that specifies all dependencies and provides step-by-step instructions to reproduce the results. This documentation ensures that other researchers can set up their environment correctly and replicate the findings of this study without ambiguity.

References

[1] Aws Albarghouthi and Samuel Vinitsky. Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 211–219, New York, NY, USA, 2019. Association for Computing Machinery.

[2] Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. Verifying fairness properties via concentration. *CoRR*, abs/1812.02573, 2018.

[3] Mysar Ahmad Bhat. Lung cancer, 2021. Dataset available on Kaggle. Last updated: October 1, 2021.

[4] Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. *CoRR*, abs/2005.12379, 2020.

[5] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 525–534, New York, NY, USA, 2020. Association for Computing Machinery.

[6] Tonni Das Jui and Pablo Rivas. Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, pages 1–31, January 2024.

[7] The Devastator. Higher education predictors of student retention, 2021. Dataset available on Kaggle.

[8] Georgia Crime Information Center Georgia Department of Community Supervision. NIJ's Recidivism Challenge Full Dataset. https://data.ojp.usdoj.gov/Courts/NIJ-s-Recidivism-Challenge-Full-Dataset/ynf5-u8nk/about_data, 2021. Data provided by Georgia Department of Community Supervision, Georgia Crime Information Center. Dataset owner: Joel Hunt (joel.hunt@usdoj.gov). Last updated: July 15, 2021. Public access level: Public. Bureau code: 011: Office of Justice Programs.

[9] Thomas A. Henzinger, Mahyar Karimi, Konstantin Kueffner, and Kaushik Mallik. Runtime Monitoring of Dynamic Fairness Properties. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 604–614, June 2023. arXiv:2305.04699 [cs].

[10] LibreTexts. Mann-whitney u test, 2023.

[11] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158. PMLR, 10–15 Jul 2018.

[12] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[13] N. Meraihi. Loan payments data, 2021. Dataset available on GitHub. Last updated: July 15, 2021.

[14] Valentim Realinho, Jorge Machado, Luís Baptista, and Mónica V. Martins. Predict students' dropout and academic success, December 2021.

[7] https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Lukina/pmarkesinis-Dynamic-Algorithmic-Fairness-in-Machine-Learning

[15] D.J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 3rd edition, 2003.

[16] Statology. Mann-whitney u test, 2023.

[17] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. You can still achieve fairness without sensitive attributes: Exploring biases in non-sensitive features. *CoRR*, abs/2104.14537, 2021.

A Results in Criminal Justice

The graphs demonstrating the sensitivity values of each feature. There are two main subsections, one for each fairness property.

A.1 Demographic Parity

The graphs demonstrating the sensitivity values of each feature in Demographic Parity.

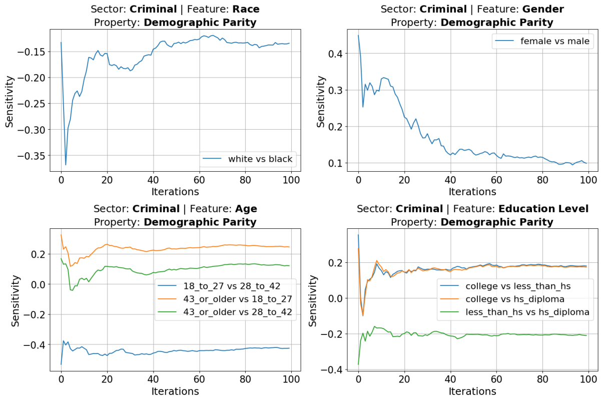

Figure 6: Monitoring Demographic Parity on features from the criminal justice dataset. At the top left is Race, top right is Gender, bottom left is Age and bottom right is Education Level. Each line in the subplots represents the sensitivity values over the iterations for different pairs of minority and majority groups.

A.2 Equal Opportunity

The graphs demonstrating the sensitivity values of each feature in Equal Opportunity.

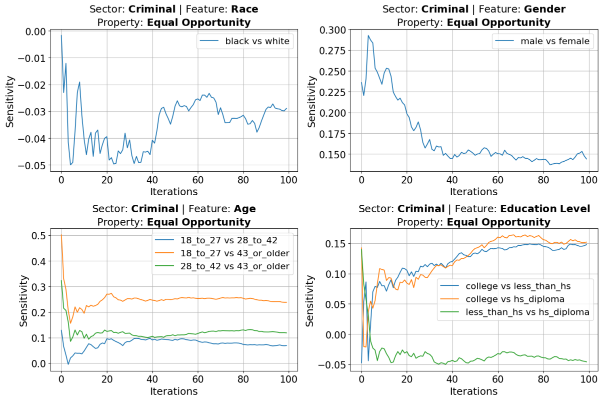

Figure 7: Monitoring Equal Opportunity on features from the criminal justice dataset. At the top left is Race, top right is Gender, bottom left is Age and bottom right is Education Level. Each line in the subplots represents the sensitivity values over the iterations for different pairs in the group.

B Results in Healthcare

The graphs demonstrating the sensitivity values of each feature. There are two main subsections, one for each fairness property.

B.1 Demographic Parity

The graphs demonstrating the sensitivity values of each feature in Demographic Parity.

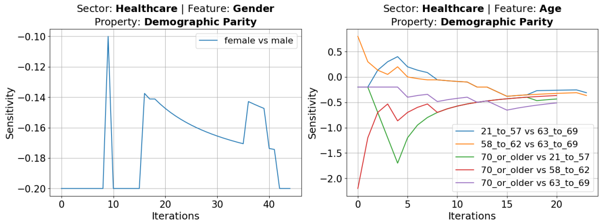

Figure 8: Monitoring Demographic Parity on features from the healthcare dataset. On the left is Gender, and on the right is Age. Each line in the subplots represents the sensitivity values over the iterations for different pairs of minority and majority groups.

B.2 Equal Opportunity

The graphs demonstrating the sensitivity values of each feature in Equal Opportunity.

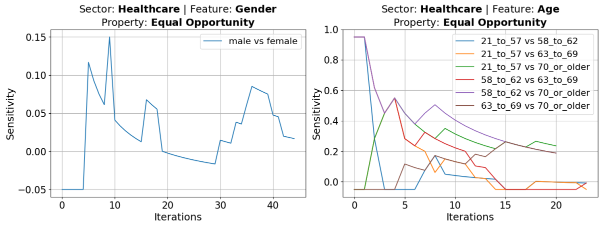

Figure 9: Monitoring Equal Opportunity on features from the healthcare dataset. On the left is Gender, and on the right is Age. Each line in the subplots represents the sensitivity values over the iterations for different pairs in the group.

C Results in Education

The graphs demonstrating the sensitivity values of each feature. There are two main subsections, one for each fairness property.

C.1 Demographic Parity

The graphs demonstrating the sensitivity values of each feature in Demographic Parity.

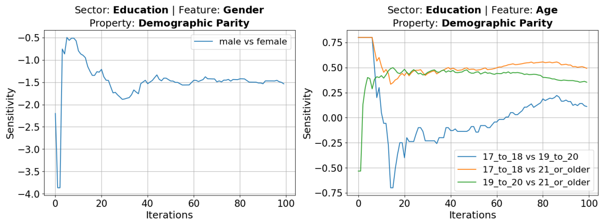

Figure 10: Monitoring Demographic Parity on features from the education dataset. On the left is Gender, and on the right is Age. Each line in the subplots represents the sensitivity values over the iterations for different pairs of minority and majority groups.

C.2 Equal Opportunity

The graphs demonstrating the sensitivity values of each feature in Equal Opportunity.

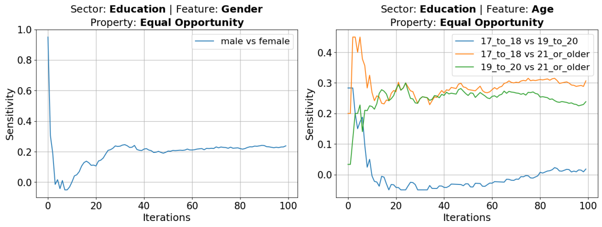

Figure 11: Monitoring Equal Opportunity on features from the education dataset. On the left is Gender, and on the right is Age. Each line in the subplots represents the sensitivity values over the iterations for different pairs in the group.

D Results in Finance

The graphs demonstrating the sensitivity values of each feature. There are two main subsections, one for each fairness property.

D.1 Demographic Parity

The graphs demonstrating the sensitivity values of each feature in Demographic Parity.

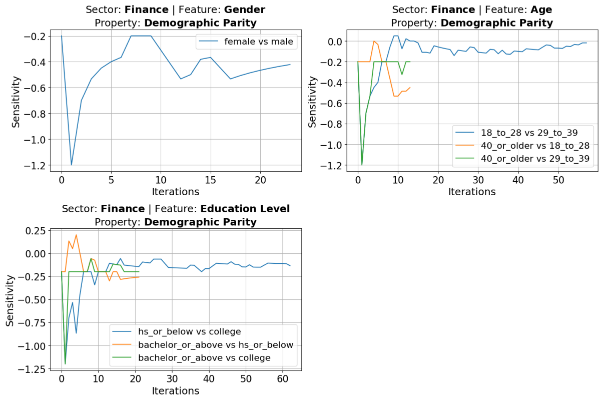

Figure 12: Monitoring Demographic Parity on features from the finance dataset. At the top left is Gender, top right is Age, and at the bottom left is Education Level. Each line in the subplots represents the sensitivity values over the iterations for different pairs of minority and majority groups.

D.2 Equal Opportunity

The graphs demonstrating the sensitivity values of each feature in Equal Opportunity.

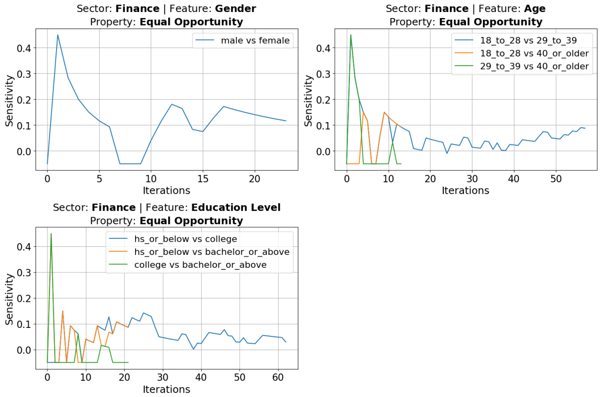

Figure 13: Monitoring Equal Opportunity on features from the finance dataset. At the top left is Gender, top right is Age, and at the bottom left is Education Level. Each line in the subplots represents the sensitivity values over the iterations for different pairs in the group.