Inside the Suspicious Machine

Ana Azevedo
FEUP & FCUP
up202108654@fe.up.pt

Diogo Silva
FEUP & FCUP
up202104794@fe.up.pt

Félix Martins
FEUP & FCUP
up202108837@fe.up.pt

Francisco Campos
FEUP & FCUP
up202108735@fe.up.pt

João Figueiredo
FEUP & FCUP
up202108829@fe.up.pt

Abstract

All over the world, countless welfare recipients are being investigated after automated systems flag them as potential fraud risks. What many do not realize is that these systems closely monitor and assess their lives, reducing them to a score based on various factors. A journalistic investigation into one of these systems, in Rotterdam, uncovered that it unfairly targets individuals based on their gender and racial background. The investigation found that the system problematically views signs of vulnerability as potential indicators of suspicious activity.

1 Introduction - Context

Every year, thousands of people in Rotterdam receive welfare benefits to help cover essential expenses. As a result, thousands of these beneficiaries are investigated annually under suspicion of committing benefits fraud.

In 2017, the city implemented a machine learning algorithm built by Accenture Consulting. This algorithm was trained to generate a risk score for all welfare-seeking individuals, using data about people who had already been investigated for fraud.

To evaluate the value of the risk score, factors such as age, gender, and Dutch proficiency were taken into account. When introduced, Accenture praised the system as a model for other cities. However, in 2021, Rotterdam halted its use following a critical ethical review requested by the Dutch government. The review uncovered that the system exhibited discriminatory behavior based on ethnicity and gender. The algorithm functions as a "suspicion machine," flagging individuals based on characteristics beyond their control, such as gender or ethnicity. Vulnerabilities, such as low self-esteem, are treated as suspicious when entered into the system.

Despite processing vast amounts of data, the algorithm's performance is only slightly better than random chance. Risk-score algorithms, like the one used in Rotterdam, are often accused of embedding human bias, though proving this can be challenging.

2 AI System Details

This section provides an overview of the technical and operational aspects of the welfare fraud prediction system used in Rotterdam. It examines the technology behind the model, the data and methodology employed, the input variables considered, and its performance, transparency, and limitations.

2.1 Technology and Development

The welfare fraud prediction system in Rotterdam is a machine learning model, specifically a gradient boosting machine. The building blocks of this type of model are decision trees. Gradient boosting machines build upon decision trees by stacking them sequentially, with each tree aiming to correct the errors of its predecessor.

In this case, the model uses 500 decision trees. To classify each instance, the values assigned to each person at the leaves of the trees are then combined and adjusted to produce the final risk score.

2.2 Training Data and Methodology

The model uses supervised machine learning, trained on a labelled dataset of previously investigated cases. The training data consists of individuals who have been investigated, with labels of 'yes fraud' or 'no fraud'[8]. This approach aims to identify patterns that distinguish fraudulent cases from non-fraudulent ones.

2.3 Input Variables

The system processes 315 input variables for each welfare recipient. These inputs include:

- **Demographic information:** age, gender, marital status
- **Personal characteristics:** language skills, mental health history
- **Socioeconomic factors:** neighbourhood, education level
- **Interaction history:** types of appointments, number of emails to the city
- **Subjective assessments:** caseworker comments on physical appearance, ability to convince others
- **Behavioural data:** hobbies, sports participation
- **Relationship information:** length of last romantic relationship

2.4 Algorithm and Scoring

The algorithm generates a risk score between 0 and 1 for each welfare recipient. Higher scores indicate a higher perceived risk of fraud. The system then ranks all recipients, with those in the top 10% (on average, about 3,000 individuals) being referred for investigation.[9]

2.5 Performance and Accuracy

Internal evaluations by Rotterdam found that the model is only 50% more accurate at predicting fraud than random selection. This raises significant concerns about the system's effectiveness and the potential for false positives.

The city shared the model's ROC curve, which looks at the tradeoff between the share of individuals correctly predicted as high risk amongst those labelled high risk (TP/(TP+FN)) and the share of individuals wrongly predicted as high risk amongst those labelled low risk (FP/(FP+TN)).

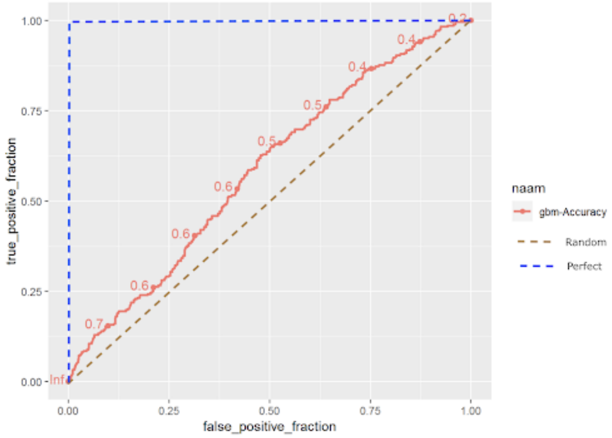

Figure 1: Model performance

The model only marginally improves on the diagonal line, which denotes random selection. Given the ROC curve, experts state that it is "essentially random guessing". The model completely fails to meet the performance criteria for a real world application.

2.6 Transparency and Interpretability

Since the code and decision-making process are not publicly available, it's difficult to understand how the algorithm arrives at its conclusions, raising several concerns regards **transparency** - which will be addressed on the next section.

2.7 Data Quality and Sources

The algorithm uses a combination of data sources, including government databases and external data providers. However, the quality and accuracy of this data are not publicly disclosed, which raises questions about the reliability of the system's outputs.

Given this type of information, it would be possible to verify, for example, that the model performs equally well across subgroups. For example, checking whether women and men at the same risk score commit fraud at equal rates, or that the error across groups is not substantially different. However, even if this holds true, significantly different trends for different groups need to be carefully analysed. Fraud rates may be higher for certain groups, just because those groups were highly focused during past investigations.

Even without direct access to the data, we can confidently state that the data quality is unequivocally poor. Rotterdam's reliance on random selection, anonymous tips, and shifting category checks creates a highly inconsistent and biased foundation for training the algorithm. This approach introduces systemic biases that inevitably become embedded in the algorithm's outcomes.

A clear example of this flawed design is the treatment of younger welfare recipients. Although around 880 young people should have been included in the data to reflect their true proportion in the welfare system, only 52 were represented. Despite this under representation, age emerged as the most significant factor in raising risk scores, leading the algorithm to falsely associate youth with a higher likelihood of committing welfare fraud.

In addition, there are 54 variables based on subjective assessments made by caseworkers, and any bias by the caseworker would reflect in the data. Another problem with some of the subjective features is that they basically do not make any sense. There are several comment fields, where caseworkers record general observations, treated as binary. Each comment, regardless of its content, is converted to a 1. This means both positive and negative comments influence the risk score in the same way - for example, a comment like "shows desire to achieve results" has the same impact as "shows no desire to achieve results", since it translates to "comment on motivation = 1".

3 Outcomes

This section examines the real-world implications of the AI system, highlighting its inefficiencies, discriminatory practices, lack of transparency, and absence of accountability. Despite its goal to detect fraud, the system's failures led to significant societal and ethical concerns, raising questions about its design and implementation.

3.1 Efficiency and Decision-making

This system did not lead to any improvements in efficiency or decision-making. Although it was designed to detect welfare fraud by cross-referencing data from multiple sources, the AI failed to achieve its main goal. Over its five years of operation, only two out of five projects requested by the government were executed.[2]

In addition, the complexity of the AI's data analysis capabilities did not translate into meaningful results, since later research conducted on this algorithm showed that it can not identify new forms of fraud. This just proves that this system is ineffective and inefficient in practice.

SyRI was also ineffective in identifying fraudulent cases from honest mistakes, or just missing information someone forgot to fill in. For example, in some cases, even trivial errors, such as a missing signature, resulted in accusations of fraud. This extrapolated even further the problem of the algorithm, since these slips could easily be fixed, without directly involving governmental authorities accusing someone of fraudulent activity.

3.2 Discrimination

The system was predominantly used in areas already deemed high-risk, which intensified existing biases. Such focus on vulnerable households resulted in disproportionate scrutiny of low-income individuals. In the end, the imbalance in how the system was applied increased massively societal inequalities, since most aid received from government institutions could be stripped away from flagged individuals in a heartbeat (in some cases, even demanding paying back the funds issued).[7]

Although the data did not explicitly include a column for race or ethnicity, it contained clear proxy variables that enable discriminatory practices - one of the most significant proxies being Dutch fluency, serving as as a surrogate for racial or ethnic identity. This issue was compounded by the dataset's failure to differentiate between orchestrated fraud cases and honest mistakes. Since immigrants are more likely to misinterpret Dutch forms due to language barriers, they were particularly vulnerable to this lack of nuance, as the system would disproportionally penalize innocent accidents.

3.3 Transparency

The logic behind this AI decision-making process was not clear to the users being tested against it. The model operated as a "black box," meaning that the algorithms and data it used to assess risk were kept secret. As a result, individuals flagged for further investigation were left without any explanation of why they were chosen or the criteria the system relied upon.

This lack of transparency made it impossible for individuals to challenge or understand the rationale behind their inclusion in the risk pool, heightening the despair of being flagged as fraudulent. T

Furthermore, since the algorithm's reasoning was also unknown to prosecutors of the fraudulent cases, municipal authorities often relied heavily on the algorithm's outputs, leading to a lack of detailed investigations into individual circumstances. Moreover, this fact implies a blind trust of the system by the government on cases which have a lot of nuance.

This poses a significant issue, as individuals have a legal right to understand why they are being investigated. Providing an explanation such as "the machine said so" is wholly inadequate and fails to meet basic standards of transparency. In some cases, if the model's choices were accurately determined, a more accurate explanation for why someone was flagged would be as outrageous as "you are a single mother in financial difficulties.". However, the blind reliance on machine-driven decisions and the perceived infallibility of algorithms has created a facade of legitimacy, allowing such absurd and unjust situations to persist unchallenged.

3.4 Accountability

Accountability was another major issue with this system. When it failed to deliver its intended outcomes or when it flagged individuals unjustly, there was no clear party or person taking responsibility. Although the system was eventually halted by a court ruling in 2020, which declared it in violation of privacy rights, the Dutch government did not hold itself accountable for the negative impacts of SyRI. Civil rights groups played a crucial role in bringing the issue to light, but no direct actions were taken by the government to address the flaws or the discrimination caused by the system before it was legally stopped.

In addition, Accenture, the company responsible for developing the SyRI system, didn't explicitly acknowledge accountability for the discriminatory issues in the Rotterdam welfare fraud algorithm.

Accountability in situations like these remains a legal gray area, and it is crucial for authorities to establish more detailed and comprehensive laws to address the recent surge in AI adoption. As AI becomes more and more prevalent in society, clear legal frameworks are necessary to ensure that these technologies are used responsibly and ethically.

4 Societal Response

There were notable societal responses, including protests and legal actions led by civil rights groups against the algorithm. Activists raised issues about privacy and discrimination, resulting in a key court decision in 2020 that ruled the algorithm unconstitutional.

The court ordered the immediate halt of the algorithm because it violated Article 8 of the European Convention on Human Rights, which safeguards individuals' privacy. The judges found that the legislation lacked sufficient transparency and safeguards against privacy violations, failing to balance social interests with citizens' rights. This ruling prompted national discussions about the ethical use of algorithms in public policy and highlighted the necessity for greater accountability and reform in welfare technologies.

Media coverage played a crucial role in raising awareness, influencing public opinion, and pushing for policy changes regarding similar systems 6. The algorithm's disproportionate impact on vulnerable communities was brought to light in reports, which raised concerns about automated decision-making in welfare systems. The government is considering enacting legislative measures to guarantee that algorithms comply with human rights norms because of the public discussion surrounding this issue. [1]

Overall, the combination of civil society activism and media attention catalyzed significant political dialogue about the future of algorithmic governance in the Netherlands. These developments reflect the growing recognition of the importance of transparency and ethical considerations in the deployment of technology within public services.

5 Sara and Yusef

In order to help the reader visualize the gravity of the problem, the reporters created two hypothetical Rotterdam residents: Sara and Yusef—both of whom would have been flagged for fraud by the algorithm.

Sara represents a single mother of two who recently separated from her partner and, despite financial difficulties, quit her job to care for her sick child and its brother.

To understand the impact of specific variables on the algorithm's risk ranking, the reporters contrasted Sara's profile with Jan, an average Rotterdam male resident who has a partner, no children, and is financially comfortable. Jan's initial ranking by the algorithm placed him at position 16,815 out of 30,000 individuals, far from the threshold for investigation (position 27,000).

Let's describe Jan by the following attributes:

- **Gender** = Man
- **HasPartner** = True
- **NumChildren** = 0
- **FinancialDifficulties** = False

5.1 Becoming Sara

The journalists, having access to the model's predictions, incrementally altered Jan's profile to match Sara's, observing how these changes affected his risk ranking. Each variable added contributed significantly to his placement on the risk list:

- Changing **Gender** from Male to Woman pushed their ranking up by 4,542 spots.

- Adding **NumChildren** = 2 increased their rank by another 2,134 spots.
- Modifying **HasPartner** to False elevated their rank by an additional 3,267 spots.
- Finally, setting **FinancialDifficulties** = True placed them at rank 28,717 — well within the group selected for investigation.

5.2 Becoming Yusef

Now, consider Jan in a different scenario. Suppose Jan speaks Arabic and did not pass his language proficiency requirement, has three roommates (instead of none), lives in a social housing neighborhood, and his caseworker expresses skepticism about his employment prospects. With these modifications, Jan's ranking skyrockets to 28,746, making him a clear target for investigation. Through these incremental changes, Jan effectively "becomes" Yusef — a hypothetical individual embodying traits commonly associated with minority or immigrant communities.

Yusef's situation demonstrates how proxies for ethnicity, such as language skills, neighborhood, and living arrangements, contribute disproportionately to the algorithm's risk evaluation. These factors are often unrelated to actual fraudulent behavior but are instead reflective of socio-economic or cultural realities, amplifying the discriminatory nature of the algorithm.

5.3 Counterfactual Fairness

The reporters' analysis of Sara and Yusef aligns with the concept of counterfactual fairness, as outlined by Kusner et al. (2017) [5]. Counterfactual fairness considers a model fair if its predictions remain unchanged when sensitive attributes, such as gender or ethnicity, are altered in a hypothetical scenario, provided all other factors remain constant.

By progressively transforming an average welfare recipient into two vulnerable individuals flagged as fraudulent by the model, the reporters were effectively evaluating whether the algorithm was counterfactually fair.

Despite its simplicity, counterfactual fairness is a powerful tool for assessing and improving fairness in AI models. It not only ensures that models make unbiased predictions but also enhances explainability, since the concept is straightforward to apply. It compares the outcomes for an individual in hypothetical scenarios where only sensitive attributes, like gender or ethnicity, are altered while other factors remain constant. This makes it accessible and easily understood by people who may lack technical expertise in machine learning - and that is the main goal of explainable AI.

5.3.1 Wired Interface

Wired [3] provided an interactive interface that allows users to adjust various variables and observe how they influence an individual's risk prediction. This tool offers an accessible way to understand the impact of specific features on the algorithm's output, making the concept of algorithmic bias more tangible.

One of the most striking observations from the interface is how age affects rankings. Being under 35 years old can increase an individual's ranking by as many as 10,000 spots, making nearly all young people disproportionately susceptible to investigation. This highlights the previously mentioned significant imbalance in the treatment of younger individuals, demonstrating how poor data quality (underrepresentation of subgroups) can adversely affect entire demographic groups.

Despite the limited number of variables available on the interface, we encourage readers to experiment with it to gain a deeper understanding of how specific variables contribute to the model's predictions.

6 Dataset

This section outlines the selection and preprocessing of the data used for the MBA admission model. Due to the unavailability of the original dataset, an alternative dataset was used, focusing on fairness and explainability in admission decisions. Key preprocessing steps, including handling missing values, removing irrelevant columns, and encoding categorical variables were applied to prepare the data for modeling.

6.1 Dataset Selection

Since the original dataset referenced in the algorithm was not accessible, all analyses were conducted using a different dataset that may be associated with the one described in the case study.

We chose the **MBA Admission Dataset, Class of 2025**, as it provides a relevant context for examining fairness and explainability in decision-making processes. This dataset presents a classification problem, containing multiple records described by nine features, with the goal of predicting the admission outcome: Admit, Waitlist, or Deny. Although the dataset is fairly balanced, it may reflect biases linked to attributes such as gender and race, which could impact the equity of the admissions decisions.

Both the MBA Admission Dataset and the Rotterdam system share concerns regarding fairness and explainability, albeit in different contexts. While the Rotterdam system is punitive in nature, denying essential benefits, the MBA admission process is not punitive but still has a significant impact on candidates' future professional opportunities.

Just as the Rotterdam system faced criticism for its lack of fairness and transparency, the MBA admission process also raises similar concerns. Factors like gender, race, academic performance, and work experience may inadvertently introduce bias, potentially leading to unfair outcomes for certain groups. Improving explainability is essential to ensure that the decision-making criteria are transparent and well-understood by all stakeholders. By doing so, we can ensure that admissions decisions are based on merit, free from discriminatory influences, and aligned with principles of fairness, transparency, and accountability.

Addressing these concerns in the MBA Admission Dataset will not only provide insights into reducing bias but also contribute to developing more equitable frameworks for other high-stakes decision-making systems.

6.2 Feature Engineering and Preprocessing

During the preprocessing stage, several steps were applied to prepare the MBA Admission Dataset for machine learning modeling. These transformations ensured data consistency, addressed missing values, and encoded categorical features for algorithm compatibility. Below, we detail the key steps and justifications:

6.2.1 *Handling Unnecessary Columns*

The application_id column, which served as a unique identifier, was removed as it does not provide predictive value for the model.

6.2.2 *Filling Missing Values*

- The target variable, admission, contained missing values. These were interpreted as cases of no admission and replaced with the label "No admit" to standardize the dataset.
- The race column had missing values for applicants marked as international (international = True). These missing values were replaced with the label "International". Following this, the international column was removed, as its information was already incorporated into the race column.

After these adjustments, all missing values were resolved, and the dataset was confirmed to have no null entries.

6.2.3 *Encoding Categorical Features*

To prepare categorical features for modeling, the following transformations were applied:

- The admission column (target variable) was mapped to integers: "Admit" was assigned 1, while "Waitlist" and "No admit" were both assigned 0.
- Other categorical features, such as gender, major, race, and work_industry, were encoded using mapping dictionaries and Scikit-learn's LabelEncoder. The gender column was converted to binary format, while the remaining categorical features were numerically encoded.

6.2.4 *Preserving Original Data*

To retain the raw, unmodified data for future visualization and analysis, two datasets were created:

- X_display and y_display contain the original input features and target variable, respectively.
- X and y contain the processed, encoded versions for use in modeling.

6.2.5 *Identifying Admission Patterns*

To explore potential disparities, the average admission rate for each racial group was calculated. This analysis revealed variations in admission probabilities across groups. For instance:

- White applicants had an average admission rate of approximately 16.8%.
- Black applicants had a lower admission rate of 8.7%.

Such findings highlight the need for further fairness analysis during model evaluation.

6.2.6 *Summary*

Through these preprocessing steps, the dataset was fully prepared for machine learning modeling. All missing values were addressed, irrelevant columns were removed, and categorical variables were encoded. These transformations ensured consistency, interpretability, and compatibility with the algorithms to be applied.

6.3 Modeling

Now that we have preprocessed the data, we can proceed with building a machine learning model to predict the admission status based on the applicant's features. For this task, we will use the Gradient Boosting Machine (GBM) model, the same algorithm used in the Rotterdam case. The following steps outline the process we followed:

6.3.1 *Splitting the Data*

We split the data into training and testing sets using an 80-20 split. This allows us to train the model on a subset of the data and evaluate its performance on unseen data.

6.3.2 *Using BL-SMOTE*

To address class imbalance in the dataset, we applied the Borderline-SMOTE (BL-SMOTE) technique. BL-SMOTE is an extension of the Synthetic Minority Over-sampling Technique (SMOTE) that targets borderline instances, which are minority class samples near the decision boundary or surrounded by majority class samples. These instances are at higher risk of misclassification. By oversampling these borderline instances, BL-SMOTE improves the representation of the minority class in challenging regions, enhancing the model's ability to correctly classify these cases while minimizing the risk of overfitting.

6.3.3 *Training the Gradient Boosting Machine (GBM)*

We trained the Gradient Boosting Machine (GBM) model on the resampled training data. The model learned the relationships between the applicant's features and their admission status.

6.4 Model Evaluation

The classification report evaluates the model's ability to predict MBA admission, where 0 represents "not admitted" and 1 represents "admitted."

6.4.1 *Class 0 (Not Admitted)*

- Precision (0.92): The model accurately identifies most candidates who should not be admitted.
- Recall (0.81): It correctly detects 81% of actual "not admitted" cases.
- F1-score (0.86): Indicates strong overall performance for this class.

6.4.2 *Class 1 (Admitted)*

- Precision (0.38): Only 38% of predicted "admitted" cases are correct, indicating many false positives.
- Recall (0.61): The model captures 61% of actual "admitted" cases but misses some true positives.
- F1-score (0.47): Highlights weaker performance for this class.

6.4.3 *Overall Metrics*

- Accuracy (0.78): 78% of instances are correctly classified, though this favors the majority class.
- Macro Avg (F1-score 0.67): Reflects weaker performance on class 1.
- Weighted Avg (F1-score 0.80): Skewed by the dominance of class 0.

6.4.4 *Conclusion*

Since we used the BL-SMOTE technique to address class imbalance, we observed some improvement in the model's performance, particularly in terms of recall for the minority class. However, the precision for the minority class remains low, indicating a high rate of false positives. This suggests that the model may be incorrectly classifying some applicants as admitted when they should not be - which is not as bad as rejecting a deserving applicant.

Classification	Precision	Recall	F1-Score	Support
Not Admitted	0.92	0.82	0.87	1043
Admitted	0.39	0.61	0.48	196
Accuracy			0.79	1239
Macro Avg	0.66	0.72	0.67	1239
Weighted Avg	0.83	0.79	0.81	1239

Table 1: Classification Report

6.5 Transparency and Explainability

According to the Trustworthy AI guidelines [4], AI systems should be transparent and fair, among other requirements. Decisions from these systems need to be explained in an understandable manner to humans.

From the analysis of this report on the Rotterdam Welfare Fraud case, we can identify that the decisions to investigate fraud were not clearly justified. In the dataset we chose, we show how explaining model decisions can help in understanding them and even identifying any potential biases. To explain model predictions, we mainly used SHAP (SHapley Additive Explanations) [6] and LIME (Local Interpretable Model-Agnostic Explanations) [10].

6.5.1 LIME

LIME is an explainability tool designed to explain a specific local prediction of a model by approximating it within a local neighborhood with a much simpler model. This simpler model is interpretable by itself, allowing us to extract the contribution of each feature's value to the overall prediction. LIME's explanations may vary according to the defined neighborhood, which brings some instability to this method.

Figure 2 shows each feature's contribution obtained by LIME in an example instance that was not accepted. Negative values (red) mean that the feature contributed negatively to the prediction (not accepted), while positive values (green) represent the opposite. This example reveals a fairness concern that should be further investigated, since the gender contributed considerably to the prediction.

6.5.2 SHAP

SHAP is based on cooperative game theory. It distributes the result of the prediction (like the result of a game) among the features used (like the players of the game). This is done using Shapley values. They provide a way to attribute a prediction of the model to each feature. Basically, each feature's value will have an associated Shapley value, which indicates if it contributed positively or negatively to the overall prediction.

For tree-based models, like the one we used, Shapley values can be efficiently calculated. However, for the general case, it is an NP-hard problem. For a small dataset, it may still be feasible, but generally it requires approximations.

SHAP can be used to explain local predictions, just as with LIME, but it can also explain the global reasoning of the model. We will see examples of both.

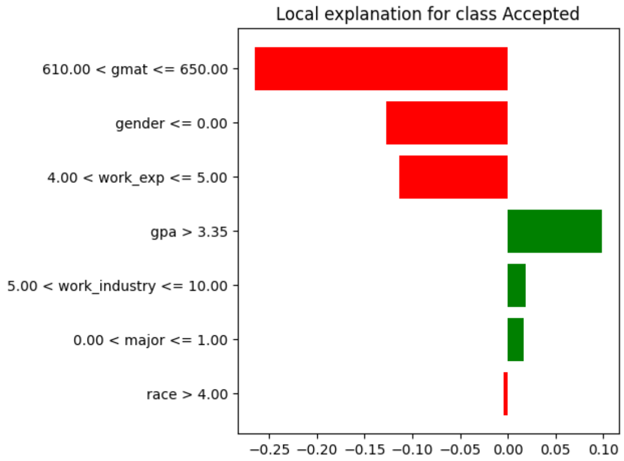

Figure 2: Feature contribution using LIME in a not accepted instance

Figure 3 shows a similar plot to the one obtained with LIME in Figure 2. These plots are mainly consistent, but there are differences in the GPA contribution. LIME's explanation shows that it contributes positively, but SHAP's shows that it is neutral. This could be attributed to the fact that LIME's explanations vary depending on the selected neighborhood, meaning that with a different neighborhood the results may be consistent with SHAP.

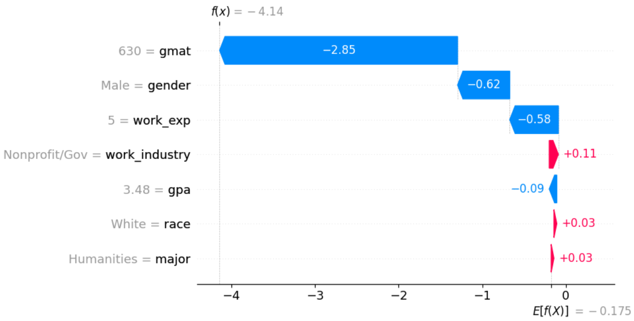

Figure 3: Feature contribution (Shapley values) using SHAP in a not accepted instance

As mentioned above, SHAP also provides global explanations. These are useful to understand the overall reasoning of the model. Figure 4 shows the Shapley values for each feature depending on their value. For categorical variables, it may be harder to understand their feature values, since they need to be encoded into integers. In this type of variables, different colors in the plot represent different categories. For this model, we can conclude that instances with higher grades (GPA and GMAT) have an higher likelihood of being accepted. However, for sensitive attributes (gender and race), it shows a bias towards certain groups.

SHAP also provides dependence plots to visualise the relationships between features, as well as the target. These plots are useful to understand how different features affect the target. The most

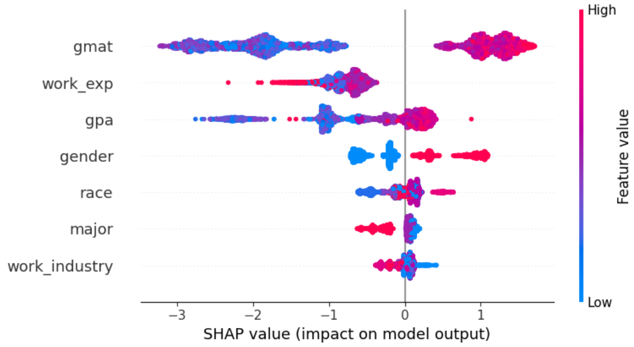

Figure 4: SHAP Global Explanations

interesting ones regard the sensitive attributes. Figure 5 shows the dependence plot for race with respect to the GMAT score. When interpreting this plot, it is important to note that it only represents the Shapley values of the race attribute, with respect to the GMAT scores, and not at all the direct contribution of GMAT scores. In the given plot, not all races have a neutral contribution, which raises fairness concerns for the trained model.

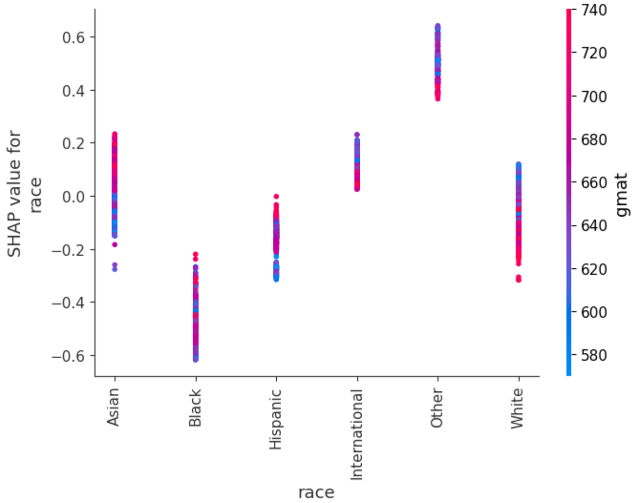

Figure 5: SHAP Dependence Plot for race (and GMAT score)

6.6 Counterfactual Fairness

Besides SHAP and Lime, we also experimented with counterfactual fairness on the chosen dataset. Our unique take on it involved selecting groups of individuals that met specific conditions and observing how altering sensitive variables impacted the model's outcome.

For the first test case, we focused on a subset of **black** individuals who had been misclassified as false negatives — those who were unjustly denied access to the MBA program by the model. We then applied counterfactual fairness by **switching their race to white**, and observed that 39% (14) of them would have been accepted under the model's prediction.

Next, we explored the opposite scenario. We selected a subset of correctly accepted (true positives) **white** individuals and **switched their race to black.** The model then predicted that 33% (54) of these individuals should be rejected.

In the third and final test case, we focused on a subset of Asian **women** who had been wrongfully rejected. We chose this test case based on observations from the SHAP analysis, which suggested that gender exhibited a bias towards accepting men within the Asian racial group. By **changing the gender to male**, 2 (7%) of the Asian women would have been accepted by the model.

Although we know that our model is not the best, these experiments still emphasize the critical role of counterfactual fairness in assessing bias within machine learning systems. The key takeaway from this analysis is how straightforward it is to check if a model is or not counterfactually fair. Systematically altering sensitive attributes by simple pandas operations can easily display how the model's predictions change based on sensitive attributes.

7 Conclusions

The model employed in the Rotterdam Welfare Fraud detection system had significant fairness issues and biases. Moreover, it lacked transparency, as no explanations were provided for decisions made about specific individuals.

Currently, AI systems are required to adhere to the Trustworthy AI Guidelines, which mandate that such biases and transparency deficiencies be addressed. The techniques presented in Sections 6.5 and 6.6 help follow these guidelines by providing methods to understand the model's decisions, while simultaneously showing any potential issues with the model.

Our analysis reveals that the model we trained falls short of meeting the prescribed standards. Furthermore, applying this analysis to the Rotterdam Welfare Fraud detection system would have highlighted several issues, raising concerns about its deployment in the real world.

We also consider relevant to note that Rotterdam is under scrutiny primarily because they had the courage to share their system with journalists. Many similar systems are currently in use, which may be even less accurate and potentially more biased, ultimately representing significant risks to welfare recipients all around the globe.

Ultimately, it's important to understand that, even though AI systems are extremely powerful and are a great tool to enhance productivity and increase efficiency of many systems, they are also machines built by humans, prone to human error and biases. By recognizing how these systems can encompass prejudices through their data and machine learning models, we can be more conscious of their shortcomings. To fully harness AI's potential to benefit society, we must critically evaluate AI predictions, create and follow ethical guidelines such as the Ethic Guidelines for Trustworthy AI[4] and uphold high data quality standards.

References

[1] Oxford Academic. 2024. Human Rights Law Review: Algorithmic Discrimination and Welfare. *Human Rights Law Review* 22, 2 (2024). https://academic.oup.com/hrlr/article/22/2/ngac010/6568079?login=false

[2] AlgorithmWatch. 2024. *SYRI and the Netherlands Algorithm.* https://algorithmwatch.org/en/syri-netherlands-algorithm/

[3] Eva Constantaras, Gabriel Geiger, Justin-Casimir Braun, Dhruv Mehrotra, and Htet Aung. 2023. The Welfare State Is Turning to Algorithms. *Wired* (March 2023). https://www.wired.com/story/welfare-state-algorithms/ Accessed: 2024/10/20.

[4] High-Level Expert Group on AI. 2019. *Ethics guidelines for trustworthy AI.* Report. European Commission, Brussels. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[5] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf Accessed: 2024/11/25.

[6] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[7] Politico. 2024. *A Dutch Algorithm Scandal Serves as a Warning to Europe.* https://www.politico.eu/newsletter/ai-decoded/a-dutch-algorithm-scandal-serves-a-warning-to-europe-the-ai-act-wont-save-us-2/

[8] Lighthouse Reports. 2024. *Suspicion Machine.* https://www.lighthousereports.com/methodology/suspicion-machine/

[9] Lighthouse Reports. 2024. *Suspicion Machines.* https://www.lighthousereports.com/investigation/suspicion-machines/

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016.* 1135–1144.