Assignment 8 Final Report

Enhancing Fairness in AI-driven Credit Approval Systems

Group 9

Yiwen Ma, Yuetong Jiang, Chenyang Huan

Faculty of Information, University of Toronto

INF 2207: Practical Elements of Responsible AI Development

Prof. Tegan Maharaj

April 13, 2024

Introduction

A slowdown in the Australian credit card market was witnessed between 2018 and 2021 due to the global pandemic but still with an expected modest resurgence in 2022. This decline can be explained by debit cards' relative cost-effectiveness over credit cards, resulting in a lower preference throughout the time. According to a MarketLine report (2023), the industry experienced a growth of 8.9% last year with total credit card transactions up to 3,169.2 million. In addition, the entire worth of the credit card market in Australia expanded by just 0.4% over 2022 and is now worth $22.9 billion. Looking ahead, forecasts indicate a promising future for the market. By 2027, the sector is projected to increase 33.3% from 2022 with approximately 4,223.2 million credit transactions. The forecast reflects the view that the credit card industry would be rejuvenated due to changing consumer preferences and, at the same time, possibly more favorable economic conditions.

With a brighter future on the horizon for the Australian credit card market, we would like to shift our focus to the exploration of credit approval AI systems employed by Australian banking firms, especially to assess their fairness and impartiality. Anchored by The Equal Credit Opportunity Act - ECOA (2011), 15 U.S.C. 1691, the law prohibits creditors from considering demographic factors, such as gender, race, color, marital status, religion, etc., to determine whether to give credit to an applicant, when they apply for a credit card. Our analysis will focus on the mirage element to understand whether any applicants have been unjustly disadvantaged. Permutation feature importance model will be used to evaluate the significance of each feature in the model regarding its impact on the model's performance. Through this method, we want to improve the system's interpretability and explainability while also confirming the favorable influence of the AI credit approval system on the rejuvenation of the Australian credit card industry as a whole.

Dataset

The Credit Card Approval dataset available on Kaggle concerns Australia credit card applications, with variables that describe the personal, financial and demographics of the applicants. The dataset has been cleaned with missing values replaced and categorical names inferred. The variables include binary indicators for gender, marital status, bank customer status, employment, prior loan defaults and possession of a driver's license. Also, it contains numerical values for age, debt, years employed, credit score, zip code and income. Additionally, it features categorical data on industry, ethnicity and citizenship status. The objective of examing this dataset is to analyze the features to predict the outcomes of the application, indicated by the "Approved" column, where approved is categorized by 1 and not approved is 0.

Methodology

Our aim remains the same as what we suggested in the proposal: to investigate if there are biases in this dataset that directly impact its decision to authorize a credit card or not, particularly in terms of the demographic variables included in the dataset, such as gender, race and marriage status. After training, validating, and testing the dataset to check if it contains demographic biases, we discovered that there are substantial disparities between persons who are married and those who are not married in terms of credit card approval. Statistical parity difference and differential impact methods are used to determine if this dataset contains biases. The permutation feature Importance is utilized after debiasing to comprehend and explain the difference in prediction power between the variables in the model before and after debiasing.

To predict which people can get approval in applying for credit, Debt, YearsEmployed, Income, CreditScore, Gender, Married, BankCustomer, Employed, Citizen, DriversLicense, Race, and Age were used for analysis and classification with a Neural Network Model, where the target variable is 'Approved'. The model parameters, such as number of hidden layers and activation functions will be tested iteratively to find the model with best performance and highest accuracy. For assessing model performance, 70% of the data is used as training set, 15% as validation set, and 15% as testing set. The model is trained, validated, and tuned iteratively on the training and validation set, and the finalized model is tested on the testing set. Accuracy metrics, for example, balanced accuracy, precision, recall, confusion matrix, are calculated to assess model performance.

In order to check if the dataset contains bias and discrimination between privileged group and unprivileged group in protected demographic attributes, two metrics are calculated. The first one is Statistical Parity Difference, which is a metric used in fairness evaluation that measures the difference in the proportion of favorable outcomes between privileged and unprivileged groups, where 0 is ideal which represents equal probability. Another one is Disparate Impact, which refers to the unequal or disproportionate effects and disparities in outcomes between different groups, signaling potential systemic biases. The ideal value is 1, where less than 1 indicates that the unprivileged group is under disadvantages. If the bias exists, a debiasing method, reweighting preprocessing approach, could be used, in which non-equal weights are assigned to each data point to decrease bias and discrimination in the original dataset to 0. After reweighting, the sensitive features should not have independent impacts on the final decisions.

Permutation feature Importance is a method to understand the prediction power of each variable used in the model. It randomly shuffles the values of each variable one at a time and then calculates the model's prediction error, which is the difference between the accuracy score of the shuffled model and the original one. The feature is important if it has a positive result, and the larger the result the more important it is to the model. The effect of a negative score in permuted feature importance shows that the model's performance improved after shuffling the values of each feature. It suggests the original values might exist biased and could mislead the decisions of the model. After the debiasing method, a shift from positive score to 0 or negative score is expected as it indicates the successful removal of bias in the model.

Result and Discussion

After computing the fairness metrics for Gender, Race, and Marital Status, a significant bias between married and unmarried individuals has been found, with a Statistical Parity Difference of -0.2326 and a Disparate Impact of 0.5136. These values indicated that the unprivileged group (unmarried people) was less likely to be assigned with favorable labels (e.g., approval) compared to the privileged group (married people).

After reweighting, the Statistical Parity Difference between married and unmarried groups within the training set was effectively balanced to 0, with the Disparate Impact increased to 1. The adjustment suggests that after assigning weights to each data point, the distribution of favorable outcomes became more equitable across marital status groups. Besides, the Neural Network model was trained where the model which maximized prediction accuracy contained 3 hidden layers, with 'logistic' activation function, 'adam' solver function for gradient descent, and regulation and penalty rate 0.01. The accuracy of the Neural Network model stood at 0.79 on the training set, both before and after reweighting.

Before debiasing, the testing set's marital status had a Statistical Parity Difference of -0.1955 and a Disparate Impact of 0.646. Both metrics have been successfully moved in the desired direction after debiasing attempts, where Statistical Parity Difference increased to 0.09 and Disparate Impact reached 1.19. In other words, the bias within the dataset was effectively reduced. As expected, the non-equal weighting of the data points caused the accuracy on the testing set to drop from 0.83 to 0.77 after reweighting. However, the model's overall performance demonstrated little variation before and after reweighting. This result implies that the system's predictive efficacy was retained even after bias was mitigated.

Table 1: Feature Importance Scores

Feature	**Before Debiasing**	**After Debiasing**
Employed	0.087	0.034
Credit Score	0.077	0.101
age_group_>40	0.029	0.010
Income	0.019	0.019
Years Employed	0.019	0.010
Married	0.010	-0.014
age_group_20-30	0.010	0.014
age_group_30-40	-0.019	-0.000
Citizen By Other Means	-0.019	-0.010
Debt	-0.010	0.010
Drivers License	-0.010	0.014
Race_White	-0.010	0.000
Gender	0.000	-0.014
Bank Customer	0.000	-0.014
Citizen Temporary	0.000	-0.005

Following the application of debias on the dataset, the importance score for several demographic features has a noticeable decrease, which aligns with the objectives of the project.The feature "employed" shows a -0.05 decrease, as its predictive power was overestimated before the debias method. Notably, the feature "married" initially exhibited bias, but after implementing debias strategies, the influence was successfully neutralized as the score shifts from 0.10 to -0.14. This result underscores the improvement of fairness in the model. Also, age and ethnicity experience a decrease in the importance score. The effect of potentially biased demographic features has been minimized and the model now relies on more objective and accountable data such as credit score and income.

Limitation

A notable limitation is that the model relies on historical data, which could reveal inherent past biases and inequalities within social norms and financial practices. Even though several debiasing methods were carried out aimed at minimizing those, the potential for residual bias might not be eliminated thoroughly. Since the data collection process is unknown, the historical dataset remains non-transparent and the remaining bias could perpetuate the decision making system, which might lead to harm to marginalized groups. Additionally, even with Permuted feature importance analysis, it is still challenging to fully explain the neural network model's process for making decisions. Neural network model operates through layers of nodes and the decision making process is not linear which makes it difficult to be explained to stakeholders.

Future Direction

There are variety of pathways for further research on fairness in AI-based credit card approval systems, expanding on the insights and limits discussed in the above sections. One potential direction is to expand the dataset's present demographic variables. For example, include educational background, kind of work, religion, and more complex subcategories of race and ethnicity in the dataset. This would allow for a more in-depth analysis of the relevance of demographic biases and their intersections, potentially revealing hidden patterns that are not visible with the current variable collection. Furthermore, in real-world applications, using more advanced debiasing algorithms on the side may improve the fairness of predictive modelling. Adversarial training is a strategy in which one model is taught to predict the result while the second model tries to anticipate the protected characteristics based on the first model's outputs. This strategy might assist to reduce the potential that protected attributes have an indirect impact on decision-making.

Conclusion

To conclude, after reweighting, the influence of Gender, Marriage status, and Race are effectively neutralized. Fairness and ethics in data science address the need for unbiased, equitable outcomes, guarding against discrimination in data algorithms, and decision-making. Debiasing techniques, such as reweighting, can help avoid biases in datasets and enhance the reliability and inclusivity of data-driven systems, fostering trust and responsible deployment of AI applications. Prioritizing fairness ensures that algorithms and models serve diverse populations without perpetuating or exacerbating existing societal inequalities. This commitment to fairness is vital for building AI systems that are both ethical and effective.

Reference:

Credit Cards Industry Profile: Australia. (2023). *MarketLine, a Progressive Digital Media business.*

Cortinhas, S. (n.d.). *Credit card approval clean data*. Kaggle. Retrieved January 27, 2024 from https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data/code

Moldovan, D. (2022). *A benchmark study on methods to ensure fair algorithmic decisions for credit scoring.* Faculty of Economics and Business Administration, Babes-Bolyai University.Retrieved April 13, 2024, from https://www.researchgate.net/publication/363652161_A_benchmark_study_on_methods_to_ensure_fair_algorithmic_decisions_for_credit_scoring

Molnar, C. (n.d.). *Feature importance. Interpretable Machine Learning*. Retrieved April 12, 2024, from https://christophm.github.io/interpretable-ml-book/feature-importance.html

Reimers, C., Bodesheim, P., Runge, J., & Denzler, J. (2021). *Towards Learning an Unbiased Classifier from Biased Data via Conditional Adversarial Debiasing*. Retrieved April 13, 2024, from https://doi.org/10.48550/arxiv.2103.06179

Scikit-learn developers. (n.d.). *Permutation importance. scikit-learn.* Retrieved April 12, 2024, from https://scikit-learn.org/stable/modules/permutation_importance.html

U.S. Government Publishing Office. (2011). *15 U.S.C. 1691 et seq (2011) - Equal Credit Opportunity.* Retrieved from April 12, 2024 from https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm