Bias in Lending and Credit Systems in Banking Sector

Sharvari Kalgutkar and Vansh Jain

University of Southern California

I. Abstract

A strong increase has been observed in the utilisation of AI in the finance industry to improve efficiency. However, this has also led to a significant increase in biased financial decision made by institutions. This study aims to access fairness in AI of systems such as loan approval prediction and credit card default prediction. Eight methods including oversampling, reweighing, penalty matrix, class weights are utilised to minimize bias. Fairness metrics such as Equal Opportunity and Statistical Parity are used as a measure to access bias mitigation methodologies used.

II. Introduction

Companies around the world have strongly shifted their focus towards automation and digitization. This has increased the adaptation of smart technologies and incorporation of AI to avoid human errors. Especially in the Banking sector, AI is being increasingly used in many use cases like loan approval, predicting credit card defaulters, etc. The shift from human decisions to AI was mainly to mitigate any individual bias towards race, gender etc. and to make the process fast and more efficient.

However, it is important to ensure that AI systems used are fair and do not rely on information like the customers race or gender to determine the customers eligibility to receive a loan or a credit card. For example, a person shouldn't be denied loan because of his race or a credit card default shouldn't be based on a person's gender. These models should not consider a person's personal information and should only rely on variables such as income, credit score, past financial history, etc.

For example, the infamous Apple card by Goldman Sachs [1] was accused of having a gender biased AI algorithm for credit limit approval. As per a Forbes article [2], 80% Black mortgage applicants and 40% Latino applicants have a higher probability of being rejected by AI algorithms, .

This project aims to analyse the feature importance of such machine learning models and whether the model gives high importance to personal information of the applicant when predicting loan approvals and credit card defaulters. Secondly, understanding if the error rates of the model are similar across gender, race, marital status etc. in-order to determine any potential bias present in the AI system. Further, exploring whether people with similar financial capabilities like similar income group (eg. less than 50k) are denied loans or credit cards by the AI system based on personal information.

III. Literature Review

Bias in credit scoring systems is a huge issue in Banking sector. Kozodoi, N., Jacob, J., Lessmann, S. [3] attempt to study how fair ML can be utilised for unbaised credit scoring. They aim to mitigate age bias between customer less than 25 years of age and greater than 25 years of age. The study benchmarks eight fairness processors across pre-processing, in-processing and post-processing fair ml methods. The following metrics were used for comparison: AUC, Profit, Independence, Separation and Sufficiency. The study trains independently using these fairness processors and calculates average performance in comparison to benchmark model gain. The average performance gain is calculated using four models namely logistic regression, ANN, XGBoost and Random forest. The study concludes that trade-offs exist between profitability and fairness. Post-processors and pre-processors perform better in terms of fairness but can cost banking systems profitability.

Fairness in Finance industry can exist in the form of AI systems used to predict whether a customer will default their credit card payment or not. Zhang, Yukun, and Longsheng Zhou [4] focused on exploring fairness in Banking System using credit card default dataset. The study compared four approaches of mitigating gender bias in credit card default system. In case 1, the study trained a LightGBM model on the biased data to create a baseline for comparison. In case 2, since the dataset is imbalanced the study used SMOTE to balance the dataset such that there are equal number of instances of credit card defaults. In case 3, the study used AI 360 on the original imbalanced dataset and used aif360 where they changed the training weights such that more importance is given to the unprivileged class. In case 4, the study combined methods 2 and 3 and performed reweighing on Synthetically balanced data. The study observed that, Case 4 had best performance in terms of accuracy and reduced false negative rate. The study also concluded that synthetic balancing of data had a stronger effect on accuracy than bias mitigation using reweighing for the given dataset.

Credit scores are also used by Peer-to-Peer (P2P) lending financial institutes, which require a higher grading of the customer's ability to repay a loan compared to binary classification in banking, where we classify a customer as either eligible to receive a loan or not. J. -Y. Shih and Z.

-H. Chin [5] study fairness in context of P2P lending. The study grades the customer into seven classes from A-G where A and B are the positive class in normal banking scenario. The study uses four modeling approaches. First approach uses cost sensitive models without reweighing. Second approach uses reweighing. Third approach uses cost sensitive models. Fourth approach uses cost sensitive models combined with reweighing. Each method is then tested on three variations of data. First variation of data does not use unemployment, zip code and uninsured as part of data. Second variation does not include unemployment and uninsured but includes zip code to introduce some unfairness in geographical context. Third variation does not include zip code but includes all other variables. The study utilizes metrics like Overall Fairness and Accuracy to measure model fairness and performance. The study concludes that financial institutes must be cautious about including geographical data as it can introduces bias. It also concludes that reweighing can potentially reduce classification but can lead to overbalancing.

Generally, the bias in credit or loan system is unintentional and may originate from bias present in the training data or the model. Christophe Hurlin Christophe Pérignon Sébastien Saurin [6] concentrate on algorithm fairness testing of models and also find the variable causing the bias. They follow a three-step methodology. First, they use the null hypothesis to evaluate the algorithm fairness. Second, they develop a new technique , Fairness Partial Dependence Plot (FPDP), to find variables that cause the bias. Finally, with the help of bias causing variables, they try to mitigate the bias. To test their methodology, they develop models such as logistic regression, tree based models, XGBoost, SVM and artificial neural networks on the German Credit Dataset. Using their methodology, they conclude that there is bias present in the dataset and model does make decision based on the gender variable. To externally validate, they also applied their methodology on the Taiwan Dataset and found no correlation with the gender variable.

In a different study, authors Arashdeep Singh, Jashandeep Singh, Ariba Khan and Amar Gupta [7], focused on removing the bias from the dataset instead of the model. The study tries to develop a fair-loan classifier and they use the Home Mortgage Disclosure Act dataset. They proposed a new methodology DualFair, that removes selection bias and label bias from the dataset. They split their datasets into sub-datasets based on the protected variables. For example if there are 3 protected variables : sex , race and ethnicity and each variable has 3 values, then based on the combination, there would be 27 datasets. For each dataset, they balanced the predictor variable using SMOTE or undersampling. This removed the selection bias and to remove the label bias, they used Situation testing and pre-trained ML model to remove data points that showed a change in prediction if the protected variable changes. DualFair creates a dataset free of bias and removes the accuracy-fairness tradeoff issue. Once the dataset is free of bias, they trained a basic Logistic Regression model using 5 fold cross validation and reported the median of the evaluation metrics. To evaluate the model, they proposed a new metric : Alternate World Index (AWI) which calculates ratio of the number of bias points in dataset to the total number of points in the dataset. Biased points refer to data instances where altering the value of a protected variable leads to a change in prediction, while all other features remain constant. The study concludes that they get the lower AWI values and same accuracy after debiasing the dataset using DualFair. However, one limitation of their methodology is the requirement for a sufficiently large dataset. During the splitting of the datasets into sub-dataset, each sub-dataset must contain both y=1 and y=0 data instances to perform SMOTE or undersampling.

IV. Dataset

The study plans to use the ”Default of Credit card clients” from UCI Machine Learning Institute [8]. This dataset contains information about credit defaults of customers in Taiwan from UCI Machine Learning Institute and has been commonly used as a dataset to study bias in finance. For example Authors Zhang, Yukun, and Longsheng Zhou used this dataset for fairness assessment [4]. This dataset consist of protected variable gender as shown in Fig 4. The dataset has in total 30,000 rows and 24 columns and the predictor variable is the binary variable credit_payment_default. The dataset is imbalanced with 23364 non-defaulters and 6636 defaulters as shown in Fig 5. The dataset consists of 18112 Female entries and 11888 entries for Male as shown in Fig 6. Fig 7 further shows how the count of female and male varies across default and non-default credit card cases. Females have more number of entries in both cases. Fig. 9 shows how the percentage of females and males changes with credit limit and loan status. Below the median credit limit of 140000 percentage of female defaulters are lower compared to male. Above median credit limit, percentage of Male defaults is slightly higher.

Another dataset that the study plans to explore is the ”Loan Approval Prediction” dataset from Pracertcbs GitHub [9] repository which focuses on property loan approval status. This dataset consist of protected variable gender as shown in Fig 1. The dataset has in total 614 rows and 13 columns and the predictor variable is the binary variable Loan_Status. This dataset consist of missing values and we plan to impute those, except for gender. The dataset is imbalanced with 422 yes and 192 no as shown in Fig 2. Additionally, there is imbalance present in the gender column as well with 489 males and 112 females as shown in Fig 3. Fig 8 further shows that the count of females for both approved and denied loan status is lower than that of Males. Fig. 10 shows how the count of female and male loan approval differs by 3% below 5000 with females getting lower approvals. However above the median income of 5000, approval rates are comparable for females and males. Both datasets have multiple protected variables like marriage, age, etc. which can lead to bias in prediction but we are focusing on gender bias.

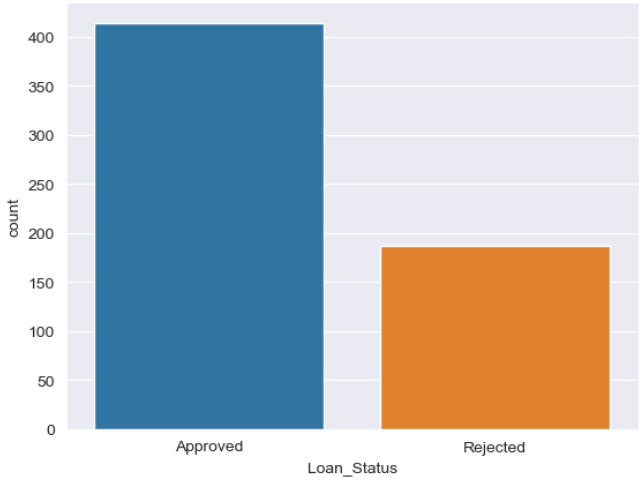

Fig. 1. Imbalance in the Predictor Variable for Loan Dataset

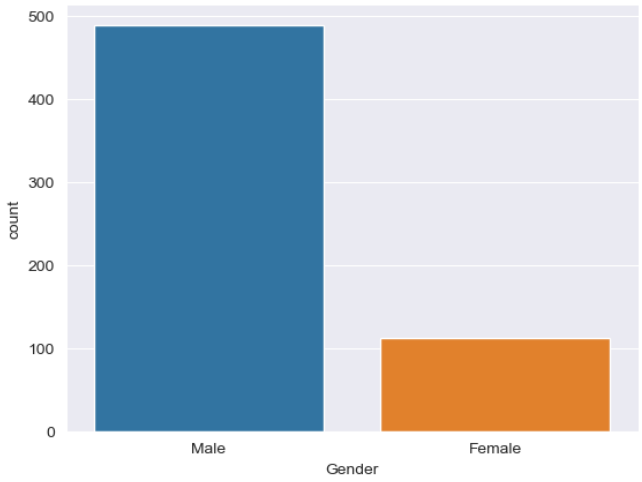

Fig. 2. Imbalance in the Gender Variable for Loan Dataset

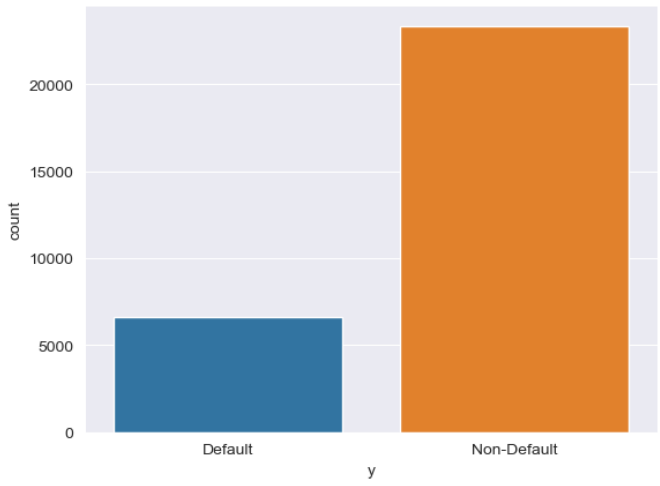

Fig. 3. Imbalance in the Predictor Variable for Credit card default dataset

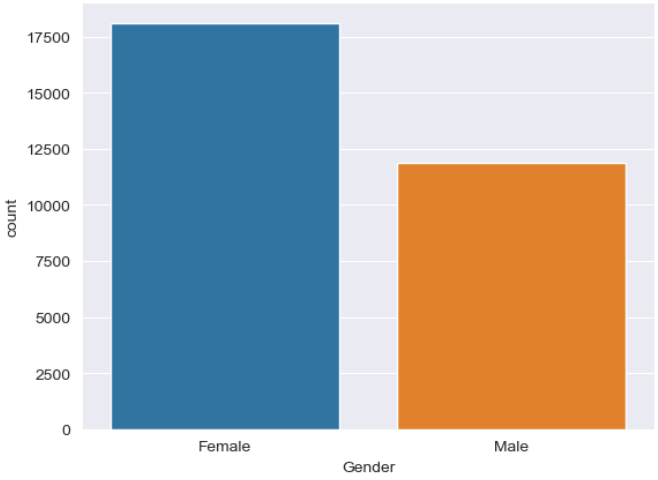

Fig. 4. Imbalance in the Gender Variable for Credit card default dataset

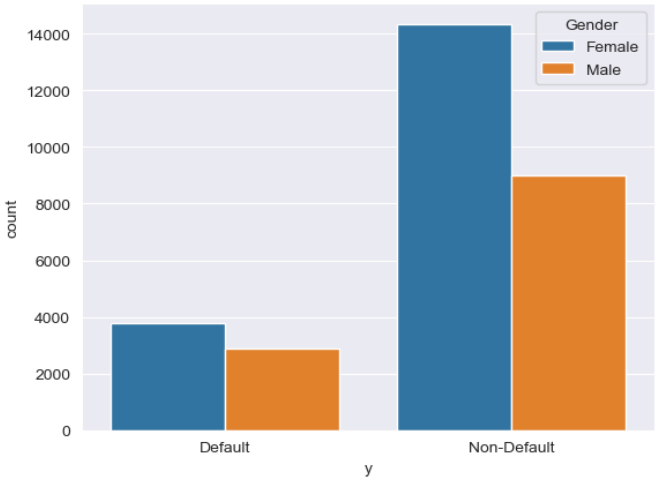

Fig. 5. Data Count of Default and Non-Default for each gender

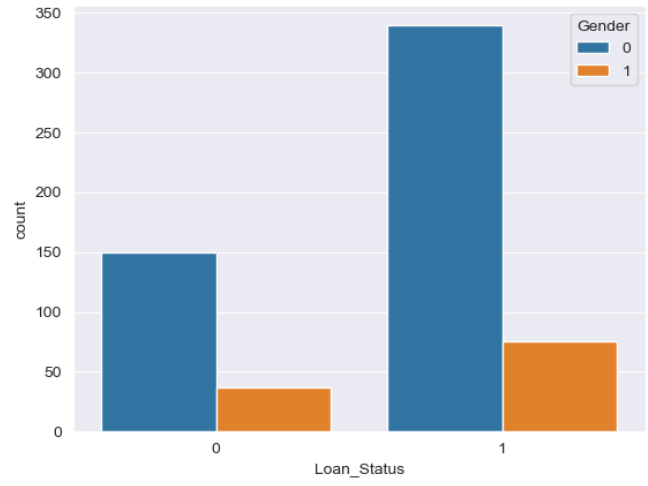

Fig. 6. Data Count of Loan Approval Status for each gender

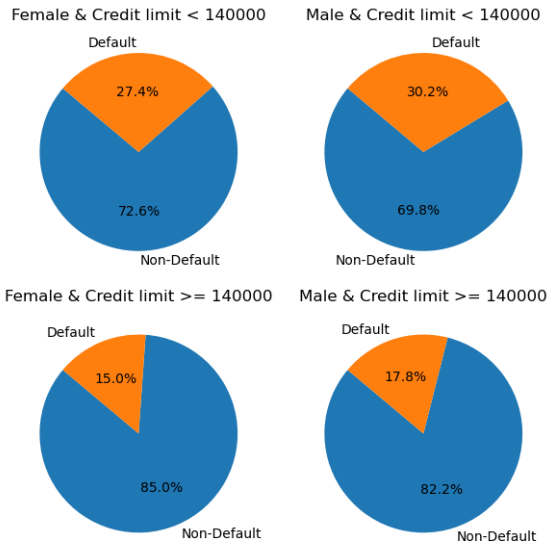

Fig. 7. Credit Card Default Based on Gender and Credit Limit

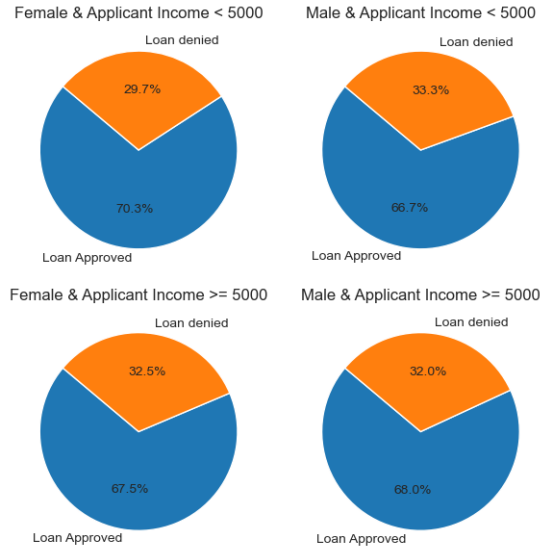

Fig. 8. Loan Approval Status Based on Gender and Applicant Income

V. Methodology

A. Data Cleaning

The Loan Approval Dataset has missing values in columns such as Marital Status, Dependent information, Self Employment, Loan amount taken, the term of loan and credit history. The missing values are imputed with mean of the respective column for numerical data. Missing values are filled with mode of the respective column for categorical data. The categorical columns are further processed using encoding to convert them into numerical columns for further analysis. For the Credit default dataset Gender was encoded such that females were represented using 1 and Males 0.

TABLE I
Fairness Metrics obtained for Credit Card Default

Method	Diff Statistical Parity	Equal Opportunity
LightGBM	0.024	0.048
SMOTE + LightGBM	0.129	0.134
BorderLineSMOTE+ LightGBM	0.138	0.152
Reweighing + LightGBM	**0.014**	0.033
Class Weights + LightGBM	0.020	0.041
Xgboost with Penalty (XGB)	0.010	0.015
SMOTE + XGB	0.130	0.159
Reweighing + XGB	**0.010**	0.015

B. Modeling

We implemented different modeling approaches to overcome bias. In Case 1 we trained a base LightGBM model for comparison. Since our dataset had some imbalance, to overcome data imbalance in Case 3 we are using oversampling techniques like SMOTE. In case 4, we also compared the performance of SMOTE with BorderLine SMOTE, another oversampling method. In case 5, we computed the class weights as another method to overcome the effect of gender bias and class imbalance. Class weights enforce that more during model training more weightage is given to the underrepresented female class. In case 6, to overcome the effect of bias we compute reweighing using aif360 and use the weights to train the model. Reweighing is a method used to make the statistical distribution of the sensitive attribute i.e gender similar to reduce the effect of bias while training. In case 7, we uniquely introduce a penalty matrix. A penalty matrix assigns penalties to protected variables. For example, if a model is biased against females, then we will assign a higher penalty to the females. So if the model gives a wrong prediction for females, it will be penalized more while training. Thus, this will help to mitigate the bias while training.

Since we are trying to mitigate gender bias i.e aiming to achieve group fairness, we plan to use metrics like Equality of Opportunity and Statistical Parity as fairness metrics to measure the effectiveness of bias mitigation. We intend to prioritize measuring the false negative rate in our loan approval model and the false positive rate in our credit default prediction model. This approach is crucial because it minimizes the risk of denying loans to qualified applicants and incorrectly labeling financially responsible customers as defaulters.

VI. Results

Table I. shows the fairness metrics and Table II shows obtained for the Credit card default data. In total eight experiments were conducted. The LightGBM model is used as

TABLE II
RESULTS OBTAINED FOR CREDIT CARD DEFAULT

Method	False Positive Rate	Accuracy
LightGBM	0.047	0.82
SMOTE + LightGBM	0.146	0.77
BorderLineSMOTE+ LightGBM	0.149	0.77
Reweighing + LightGBM	0.043	0.82
Class Weights + LightGBM	0.039	0.82
Xgboost with Penalty (XGB)	0.053	0.82
SMOTE + XGB	0.143	0.77
Reweighing + XGB	0.053	0.082

TABLE III
FAIRNESS METRICS OBTAINED FOR LOAD APPROVAL

Method	Diff Statistical Parity	Equal Opportunity
LightGBM	0.0019	0.022
SMOTE + LightGBM	0.024	0.075
BorderLineSMOTE+ LightGBM	0.035	0.008
Reweighing + LightGBM	0.058	0.091
Class Weights + LightGBM	0.007	0.041
Xgboost with Penalty (XGB)	0.0019	0.019
SMOTE + XGB	0.033	0.075
Reweighing + XGB	0.0019	0.019

a base model for comparison. On performing SMOTE and BorderLine SMOTE for overcoming dataset imbalance we saw that equal opportunity improved however statistical parity difference increased. It showed an increase in False positive rate and a decrease in accuracy. Reweighing showed the lowest statistical parity when trained in combination with LightGBM and XGBoost model, however we saw that equal opportunity slightly decreased. Class weights showed a slight decrease and a comparable performance with respect to the base model of LightGBM. When Penalty was applied to XGBoost, we noticed a decrease in statistical parity difference but Equal opportunity also saw a dip. In terms of performance there was a slight increase in false positive rate, however accuracy performance remained intact. Smote with XGB showed a increase in equal opportunity but a decrease in statistical parity and showed a drop in terms of performance metrics namely accuracy and false positive rate.

TABLE IV
RESULTS OBTAINED FOR LOAN APPROVAL

Method	False Negative Rate	Accuracy
LightGBM	0.114	0.79
SMOTE + LightGBM	0.195	0.74
BorderLineSMOTE+ LightGBM	0.126	0.79
Reweighing + LightGBM	0.057	0.82
Class Weights + LightGBM	0.034	0.82
Xgboost with Penalty (XGB)	0.149	0.74
SMOTE + XGB	0.195	0.74
Reweighing + XGB	0.149	0.74

Table III. shows the fairness metrics and Table IV shows obtained for the Loan approval data. In total eight experiments were conducted for this dataset similar to the Credit Card default data. The LightGBM model is used as a base model for comparison. On performing SMOTE and BorderLine SMOTE for overcoming dataset imbalance we saw that equal opportunity improved however statistical parity difference increased similar to the previous results. It showed an increase in False positive rate and a decrease in accuracy for SMOTE. For BorderLine SMOTE there was similar accuracy comparable to the base LightGBM model. Reweighing was used in combination with LightGBM and XGBoost model. Reweighing with XGBoost showed comparable results with the base model. However it showed a increase in statistical parity but an overall performance improvement for equal opportunity, False Negative Rate and Accuracy is seen. Class weights showed a slight increase in statistical parity and a improved performance in terms of other metrics with respect to the base model of LightGBM. When Penalty was applied to XGBoost, we noticed similar statistical parity difference but Equal opportunity also saw a dip. In terms of performance there was a slight increase in false positive rate, however accuracy performance also decreased. Smote with XGBoost showed a increase in equal opportunity but a decrease in statistical parity and showed a drop in terms of performance metrics.

VII. Discussions

The study observes similar trends in overcoming bias across both credit card default and loan status data. However, the eight bias mitigation experiments also showed slight variations between the two data, which shows the need for tailoring the approach according to the use case.

Overall it was observed that oversampling methods like SMOTE were more effective in increasing equal opportunity, while slightly impacting model performance. Whereas, Reweighing and Penalty matrix were efficient in decreasing Statistical Parity, while showing a slight decrease in model performance. This shows the potential tradeoff between fairness metrics and model performance such as accuracy and a need to carefully find a balance between the two. Especially in the financial sector, where profit making is the key, it is also important to balance fairness to overcome significant bias within the system. However, it is important to to focus on bias mitigation in financial sector since the implications of these decision made by the institutes can have significant impact on the lives of people.

VIII. Conclusion and Future Work

The study conducted eight experiments across two datasets focusing on mitigating gender bias in Banking sector applications. Overall results suggest that methods often show a trade-off between mitigating bias and model performance. Reweighing showed the best performance in terms of reducing bias as well as model accuracy and performance.

Future work could include accessing the effect of other biases like ageism, racial bias etc. and examine how these methods perform on overcoming different types of biases that we may encounter. Alternate methodologies like adversarial learning, causal learning etc. can be compared with the baseline model for better overall performance comparison.

IX. Github Code Link

Link : Github Link

References

[1] Vigdor, Neil. "Apple Card Investigated after Gender Discrimination Complaints." The New York Times, 10 Nov. 2019, www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

[2] Hale, Kori. "A.I. Bias Caused 80% of Black Mortgage Applicants to Be Denied." Forbes, www.forbes.com/sites/korihale/2021/09/02/ai-bias-caused-80-of-black-mortgage-applicants-to-be-denied/?sh=469c35936feb. Accessed 31 Jan. 2024.

[3] Kozodoi, N., Jacob, J., Lessmann, S. (2021). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research. doi:10.1016/j.ejor.2021.06.023

[4] Zhang, Yukun, and Longsheng Zhou. Fairness Assessment for Artificial Intelligence in Financial Industry. 2019.https://doi.org/10.1007/s13204-021-02063-4

[5] J. -Y. Shih and Z. -H. Chin, "A Fairness Approach to Mitigating Racial Bias of Credit Scoring Models by Decision Tree and the Reweighing Fairness Algorithm," 2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), Taichung, Taiwan, 2023, pp. 100-105, doi: 10.1109/ICEIB57887.2023.10170339. keywords: Measurement;Costs;Big Data;Data models;Peer-to-peer computing;Decision trees;Internet of Things;credit scoring;decision tree;fairness algorithm;p2p lending

[6] Hurlin, Christophe, et al. "The Fairness of Credit Scoring Models." ArXiv.org, 20 May 2022, arxiv.org/abs/2205.10200.

[7] Singh, A.; Singh, J.; Khan, A.; Gupta, A. Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair. Mach. Learn. Knowl. Extr. 2022, 4, 240-253. https://doi.org/10.3390/make4010011

[8] Yeh,I-Cheng. (2016). Default of credit card clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H.

[9] "Basic-Dataset/Loan-Approval-Prediction.csv at Master · Prasertcbs/Basic-Dataset." GitHub, github.com/prasertcbs/basic-dataset/blob/master/Loan-Approval-Prediction.csv. Accessed 31 Jan. 2024.

[10] Weber, M., Yurochkin, M., Botros, S., Markov, V. (2020). Black Loans Matter: Distributionally Robust Fairness for Fighting Subgroup Discrimination. arXiv [Cs.CY]. Retrieved from http://arxiv.org/abs/2012.01193

[11] Cristina, Ana, et al. "Algorithmic Discrimination in the Credit Domain: What Do We Know about It?" AI Society, 17 May 2023, https://doi.org/10.1007/s00146-023-01676-3.