Artificial Intelligence Ethics and Applications

Module Code: CIS4057-N

IDENTIFICATION OF BIAS IN AI MODEL FOR PREDICTING BANK CHURN

CHEDE, EDWARD (STUDENT)

D3333789

01-05-2024

Word count 2160

1.0 Introduction

Large corporations strive to both attract new customers and retain existing ones. Analysing customer churn is vital for identifying long-term customers who may leave, enabling the development of retention strategies. In banking, detecting customer churn assists management in predicting potential churners early and targeting them with promotions. Additionally, it provides insights into factors influencing customer retention strategies.

Customer churn prediction involves forecasting whether a customer will stop using a company's services. Inspired by the annual increase in bank churn customers, estimated at 1.5 million (**Tekouabou et al., 2022),** and the cost-effectiveness of retaining existing customers compared to acquiring new ones (Hasraddin & Yerdelen, 2021), this study underscores the value of prolonged customer relationships in banking. Loyal customers not only contribute to bank profitability but also attract new customers through positive referrals **Ozden & Umut, (2014).**

However, while machine learning models have become increasingly employed for churn prediction, concerns about fairness have emerged. Such algorithms, though powerful, may inadvertently perpetuate discrimination, especially in predicting customer churn. The quest for fairness in machine learning aims to design. algorithms that make equitable predictions, free from bias and discrimination. This study focuses on group fairness and equal opportunity, emphasizing the importance of mitigating bias and promoting equity in predictive modelling **Pratik & Mykola, (2017**).

Gender serves as the protected attribute in this study, prompting an investigation into the concept of "fairness through unawareness," However, achieving fairness through unawareness alone may not suffice to prevent discrimination when other background knowledge is available (**Pratik & Mykola, 2017).** Through an examination of performance metrics, the study assesses the impact of gender on fairness and equal opportunity across classes, aiming to achieve demographic parity where the positive rate is consistent across all groups **Pratik & Mykola, (2017).**

1.1 Related Study

According to **(Buolamwini and Gebru, 2018)** Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women. In their study, they showed that the AI model exhibited a huge bias against dark skin women, as against light skinned men.

Julia Dressel and Hany Farid, (2018) demonstrated that the COMPAS algorithms, utilized for forecasting recidivism among criminal defendants, exhibited no greater accuracy or fairness compared to judgments made by

individuals lacking expertise in criminal justice. Notably, COMPAS yielded a false-positive rate of 40.4% for black defendants and 25.4% for white defendants, a statistically significant discrepancy. Similarly, the false-negative rate for black defendants stood at 30.9%, while it was 47.9% for white defendants, also significantly different. These findings suggest a lack of fairness toward black defendants, paralleling the unfairness observed in judgments made by the participants in the study and the COMPAS algorithms.

AI models have been shown to be bias and discriminate against individuals and groups and has consequently inflicted different levels of harm to the affected group. Asimov's Law of Robotics states that robot should not cause harm to human either directly or indirectly, however a biased model is ultimately causing harm and this leaves with a huge responsibility on us to ensure that every AI model deployed is devoid of bias.

2.0 Exploratory Data Analysis (EDA)

We carried out Exploratory Data Analysis (EDA) in order to examine the structure of the dataset, its quality, and relationships to inform subsequent analysis. The EDA helps us identify trends, patterns, and potential issues like outliers or missing values. The EDA aids hypothesis formulation, feature selection, and communication of findings. It's foundational, ensuring data reliability and guiding effective analysis strategies for informed decision-making.

Fig 1: Bar chart showing the distribution of 'Exited' and Retained customers.

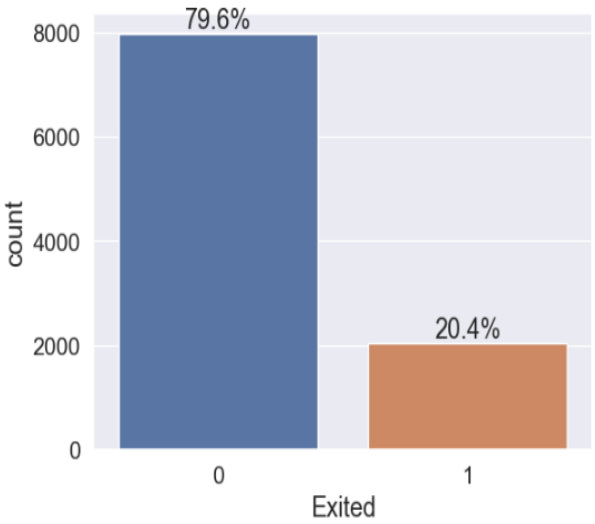


Fig 2: Gender distribution showing that there were more male in the distribution.

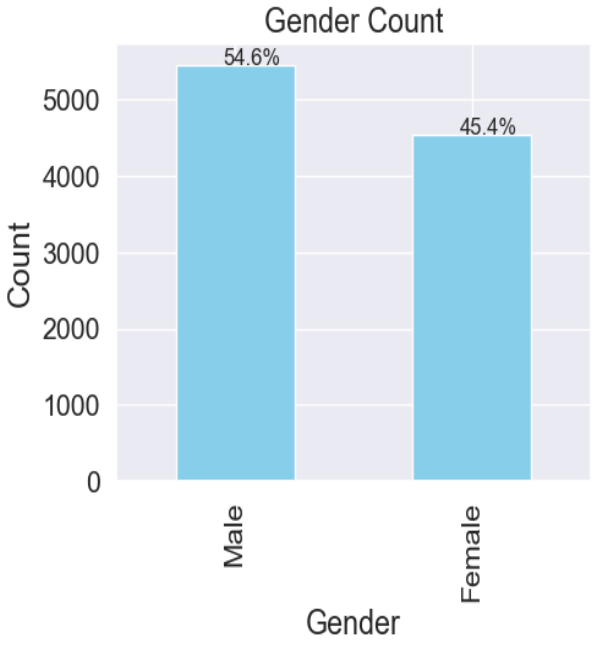


Fig 3: Gender with respected to those that Exited.

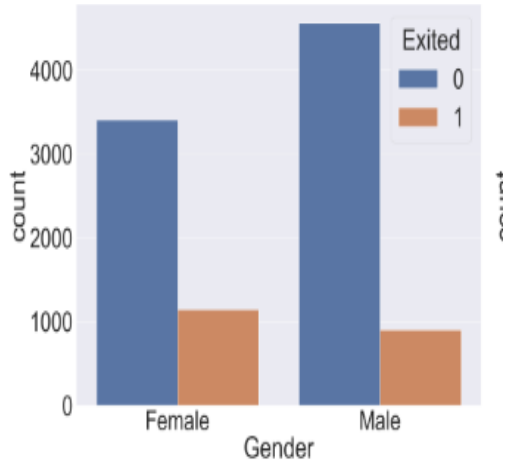


Fig 4: Number of products.

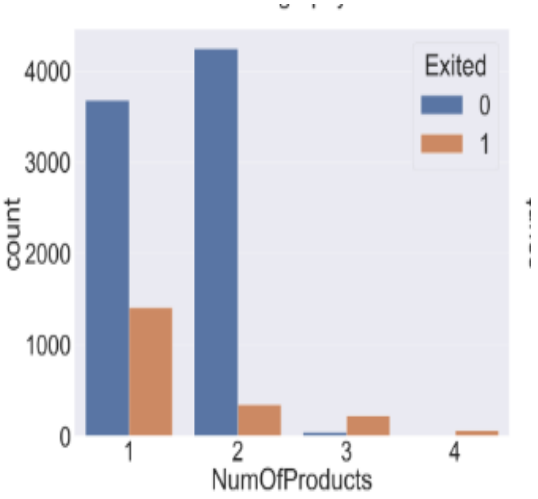


Fig 5: Balance Estimated Salary

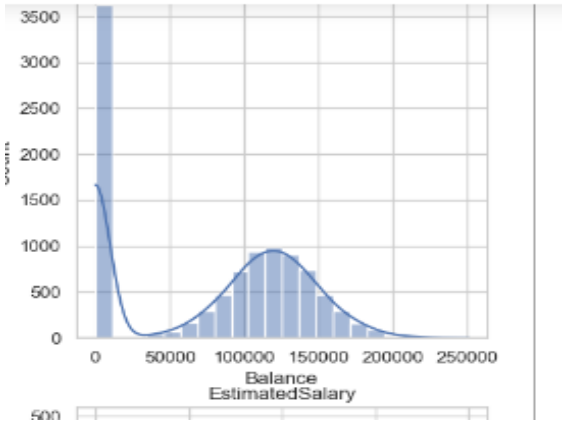


Fig 6: Credit score

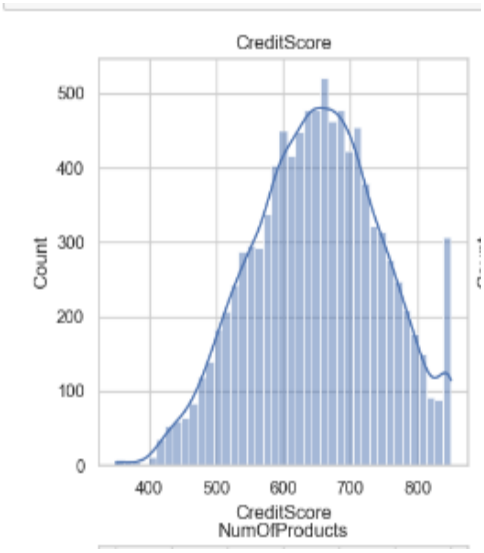
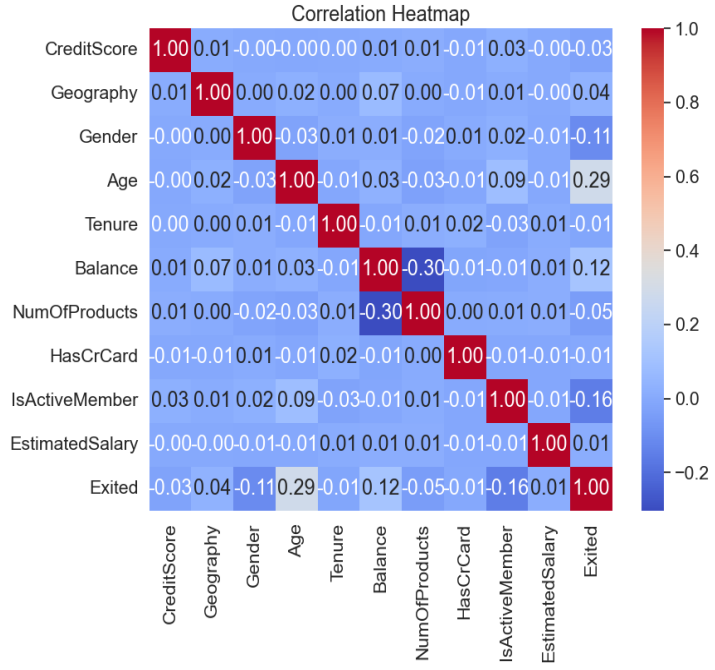


Fig 7: Heatmap for correlation between features



The correlation heatmap serves as a visual tool for identifying patterns and relationships between variables within a dataset. Stronger correlations are represented by brighter or darker colours,

while weaker correlations appear lighter or closer to neutral shades. In this context, the heatmap reveals that none of the features exhibit high correlation, indicating a lack of strong linear relationships among them.

3.0 Methodology

3.10 Data Preprocessing

3.11 The Dataset

The original data used in this work was sourced from Kaggle. The dataset showed details of the bank customers, such as their credit score, geographic location, gender, age, tenure, balance, number of products, credit card status, activity status, estimated salary, and whether the customer exited or not. There are 10,000 rows and 13 variables in the dataset. In order to prepare the dataset for the machine learning model, we cleaned the data and carried out feature engineering. This step of cleaning the data played a vital role in optimizing the performance of the models by making the dataset suitable for the study. The dataset is accessible through the link ***https://www.kaggle.com/code/nasirislamsujan/bank-customer-churn-prediction***

Table 1: Variables and description

Variable Name	Variable Description
Row number	Row numbers from 1 to 10000.
Customer Id	Unique Ids for bank customer identification
Surname	Customer's last name
Credit Score	Credit score of the customer
Geography	The country from which the customer belongs.
Gender	Male or Female
Age	Age of the customer
Tenure	Number of years for which the customer has been with the bank.
Balance	Bank balance of the customer
Num of Products	Number of bank products the customer is utilizing(savings account, mobile banking, internet banking etc.).
Has Cr Card	Binary flag for whether the customer holds a credit card with the bank or not.
Is Active Member	Binary flag for whether the customer is an active member with the bank or not.
Estimated Salary	Estimated salary of the customer in Dollars.
Exited	Binary flag 1 if the customer closed account with bank and 0 if the customer is retained

3.12 Overview of Variables

3.13 Dependent Variable

The dependent variable is influenced and affected by changes in the independent variables. The dataset contained one dependent variable 'Exited' which explores whether or not a customer of a bank will abandon the bank's product/ services. The variable is assigned '1'. If the customer leaves the bank and assigned '0' if he is retained.

The distribution of the target variable was heavily biased to the extent of 79.6% to 20.4% in favour of the customers retained. To address it, we utilized SMOTE for up sampling the target variable. SMOTE generates synthetic data points by interpolating between existing minority class samples, effectively balancing the class distribution. This method ensures

that the model is trained on a more representative dataset, reducing the risk of over-fitting and minimizing the potential introduction of noise and bias.

3.14 Independent Variables

These are the predicators and are not affected by the changes in any other variable. The independent variable explores its effect on the dependent variable. The dataset contains 12 independent features representing the characteristics of each customer.

3.20 Feature Engineering

In a bid to transform the raw data into features that better represent the underlying problem to the predictive models and improve their performance, we carried out feature selection, Encoding and standardization.

3.21 Feature Selection

Feature selection was conducted by removing irrelevant features (variable). This was done to improve our model's prediction performance and reduce computational complexities. The 'Customer id' and 'Surname' variables were dropped, while the 'Row number' variable, was the serial numbering was indexed.

3.22 One-Hot-Encoding

The categorical variables were encoded using the One-hot-encoder. Unlike the Ordinal encoder and Label encoder which are ordinal and suitable where there are hierarchies in the categories that need to be maintained. The One-hot-encoder is suitable for nominal relationships. We used the One-hot encoding to prevent the model from assuming a false ordinal relationship between categories and thereby introducing bias.

3.23 Scaling

We used the Robust Scaler to transform the values of numerical features and the encoded categorical features to a specific range. The purpose of the scaling was to ensure that all features have similar ranges of values, which can be crucial for the machine learning algorithms because they are sensitive to the scale of features. Scaling helps to normalize the data distribution and prevent features with larger magnitudes from dominating those with smaller magnitudes during model training.

4.0 Model Implementation

In this study, Support Vector Machine (SVM) classifiers was employed for supervised learning, given the presence of a target variable in the dataset. Since the target variable consisted of binary responses, signifying two distinct classes, classification methodology was chosen. The SVM classifier was specifically selected for model training and testing due to its robustness in handling classification tasks.

The primary objective of the study was to identify bias in the ML model with respect to the gender distribution.

The target variable, denoted as 'Exited', exhibited a significant imbalance, with the '0' class comprising approximately 79.6% of the samples and the '1' class representing about 20.4%. Concurrently, the distribution of gender within the dataset revealed a predominance of

males, accounting for 54.6%, and females constituting 45.5%.
To address the imbalance within the target variable, Synthetic Minority Over-sampling Technique (SMOTE) was applied, a method commonly used for handling imbalanced datasets. However, it was noted that the SMOTE procedure led to alterations in the original gender distribution. SMOTE, by generating synthetic data points through interpolation of existing minority class samples, resulted in an increased proportion of female instances. Specifically, after SMOTE, the gender distribution shifted to 59% females, comprising 1879 instances, and 41% males, totalling 1307 instances.
Subsequent to the data preprocessing, the SVM model was trained and tested, and performance metrics were evaluated to gauge its effectiveness in predicting the target variable. Subsequently, the model's predictions for both male and female groups were extracted, and performance metrics were computed for the male and female class. This was done to investigate if there was any bias towards either of the gender class. The values of the accuracy, Precision, Recall and Positive rates were determined and used to test the fairness criteria of the AI model.
This approach facilitated a comprehensive evaluation of the model's performance and its impact on different demographic groups, thereby enhancing the study's robustness and validity.

5.0 Findings

5.1 Trade-off Between the FP and FN (Type 1 and Type 2 Error)

In this study, it is important that the model effectively identifies the customers that are likely to churn before they leave. The False Negative (FN) predictions are instances where the model wrongly predict that a customer will stay (not leave), while a False Positive prediction is when the model wrongly predicts that a customer will churn (not stay). The FN has more adverse impact for the bank in terms of the customer base and overall profit, consequently, we focused on the FN. However, the False Positives may lead to unnecessary retention efforts or interventions for customers who were not actually at risk of leaving. But the actual threat of customers churn lies within the False Negatives. In analysing the performance of the model, we considered the Accuracy, Recall and Positive Rate, but the Recall will give a better view of the model's performance.

Table 2: Performance Metrics

Table 2: Performance Metrics

Performance indicators	Performance metric of the model with number disparity of the male and female classes	Performance metrics for the male class with disparity in gender size	Performance metrics for the female class with disparity in gender size
Calculated Accuracy	**0.76**	**0.77**	**0.76**
Calculated Positive Rate	**0.51**	**0.22**	**0.71**
Calculated Recall	**0.77**	**0.49**	**0.86**

5.2 Equal Accuracy

The accuracy of the male class and female class are 0.77 and 0.76 respectively. The accuracy metric measures the percentage of the correct classification by the model. Judging from the accuracy we could not identify any bias because the values are very closely related. However, the

accuracy is limited and cannot be relied upon alone to predict fairness.

5.3 Positive Rate

The positive rate measures the statistical/ demographic parity. It represents the total number of instances that the model predicted as positive outcome regardless of whether they were actually positive or not. It is independent of the ground-truth. The positive rate for the male class is significantly lower than that of the female, standing at 0.22 for the male class and 0.71 for the female class. This is suggestive of a bias in against the male class.

5.4 Equal Opportunity

For the model to satisfy equal opportunity among the group, the True Positive Rate (Recall) must be the same across the male class and the female class. Again, there is a significant difference in the values. The male class has a Recall value of 0.49, while the female class has a recall value of 0.86. This confirms the existence of bias against the male group.

Fig 8: Confusion matrix for the male class

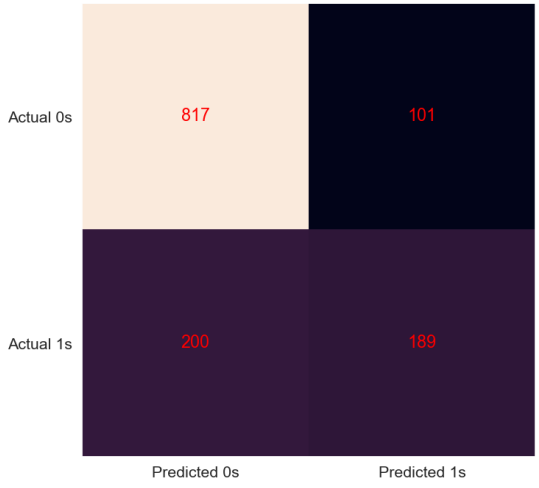
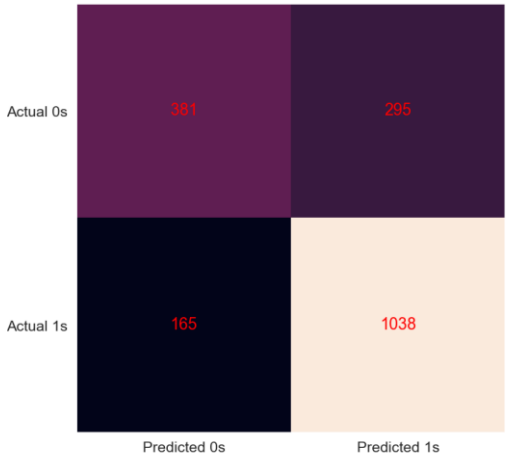


Fig 9: Confusion matrix for the male class



5.5 Conclusion and Implication of Bias in the Model

From the result of the performance metrics, it has been established that there is a significant bias against the male class. The principle of equal opportunity between the two gender groups could bot be attained by the AI model, which was evident in the significant difference in the 'Recall values' of the two groups.

It is imperative that AI models are tested to ensure that they are not biased on the grounds of the protected attributes (age, disability, gender reassignment, marriage or civil partnership, pregnancy and maternity, race, religion or belief, sex and sexual orientation, as it would be a violation of the Equality Act of 2010, and consequently cause harm to the affected individual or group

Effectively identifying male customers likely to leave the bank could prompts management's strategic efforts to retain them, with activities such as targeted promotions and personalized services. However, a failure to retain this group due to a biased AI model's oversight is deemed harmful to them, because it neglects their

concerns to accessing equal banking opportunities.

Such bias may not only undermine the equitable treatment of male group but may also erode trust and confidence among male customers, and this prejudice can make them feel mistreated.

More so, the perpetuation of bias in predictive models perpetuates systemic inequalities, making some groups more disadvantaged and excluded from opportunities, which can keep them marginalized.

References

[1] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability, and Transparency, Proceedings of Machine Learning Research, 81, 1–15.

[2] Dressel, J., & Farid, H. (2018). The COMPAS Risk Assessment: A Case Study of Algorithmic Fairness. arXiv preprint arXiv:1610.07524.

[3] Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184.

[4] Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in the banking sector: Evidence from explainable machine learning models. JAME, 1(2), 85–93.

[5] Özden Gür Ali, U., & Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. Expert Systems with Applications, 41(17), 7889–7903. https://doi.org/10.1016/j.eswa.2014.06.018

[6] Tekouabou, S. C. K., Gherghina, Ș. C., Toulni, H., Mata, P. N., & Martins, J. M. (2022). Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. Mathematics, 10(14), 2379. https://doi.org/10.3390/math10142379