# Investigating Bias in Predictive Fairness Using Bank Customer Churn

**ARTIFICIAL INTELLIGENCE ETHICS AND APPLICATIONS**

**CIS4057-N**

**SCHOOL OF COMPUTING, ENGINEERING & DIGITAL TECHNOLOGIES**

**C2160212**

**Kehinde Paul Akinbile**

**SUBMISSION DATE 01 – 05 - 2024**

# Introduction

In retail banking, accurately forecasting customer churn is a critical business imperative. Churn, which occurs when customers switch from one bank to another, has been identified as a more financially pressing issue than acquiring new customers (Becker and Ostrom, 1993). The traditional approach to predicting churn relies heavily on statistical analysis. Still, the advent of artificial intelligence (AI) and data mining offers a promising avenue to enhance prediction accuracy through scalable models capable of revealing novel data patterns (Barocas et al., 2019). However, the black-box nature of these AI models presents ethical conundrums regarding accountability and interpretability (Rudin, 2019). An underexplored facet of this ethical debate is the fairness of the model predictions, particularly when examined through the lens of gender—a critical protective characteristic in machine learning (Buolamwini and Gebru, 2018).

Elkan (2001) highlighted that the goal of learning machines is to optimize performance criteria through training data sets. However, this process could inadvertently perpetuate existing societal biases, manifesting as gender bias in AI predictions—a form of unintentional differential treatment based on sex. This phenomenon is reflective of broader cultural biases embedded within the training data (Mehrabi et al., 2019). Gender bias in AI not only echoes but also amplifies existing societal prejudices.

To combat such biases, a rigorous evaluation of model fairness is essential (Selbst et al., 2019). It entails comparing subgroup error rates—in this case, between male and female customers—to ensure equitable treatment. Current practices rarely undertake this comparison, although it is crucial for establishing the nondiscriminatory nature of predictive algorithms (Suresh and Guttag, 2021).

Furthermore, the surge in women's active participation in professional roles underscores the necessity for banks to understand and address gender-specific customer behaviors (Wachter et al., 2017). Insights into gender differences are vital for formulating customer retention strategies and can significantly enhance a bank’s ability to maintain customer loyalty and reputation in the long term. As banks increasingly turn to AI for churn prediction, ensuring the fairness of these models, particularly concerning gender, is not just an ethical mandate but a strategic one. Addressing gender-specific customer behaviors and preemptively mitigating gender biases in AI will be instrumental in building robust, equitable, and trustworthy predictive systems within the banking sector.

# Data Preprocessing

In our study, we took great care in preparing the CHURN dataset for effective churn prediction modeling. First , we checked for missing values and were relieved to find that the dataset was complete—no missing values i n sight, which is a dream scenario for any data analyst. This spared us the need for imputation strategies, ens uring the dataset's integrity.

Next, we turned our attention to outliers, as they can skew analysis and lead to biased predictions. Using the Interquartile Range (IQR) method, we identified potential outliers in numerical columns like 'CreditScore', 'Ag e', and 'NumOfProducts'. To address this issue without losing valuable data, we applied winsorization, adjust ing outliers to nearby acceptable values and minimizing their impact.

We also eliminated unnecessary variables like 'RowNumber', 'CustomerId', and 'Surname' from the dataset, as they didn't contribute to predictive modeling. This streamlined the dataset, focusing only on features relev ant to churn prediction.

Normalization was another crucial step, ensuring numerical features were scaled to a standard range for eas ier processing by machine learning models.

# Exploratory Data Analysis

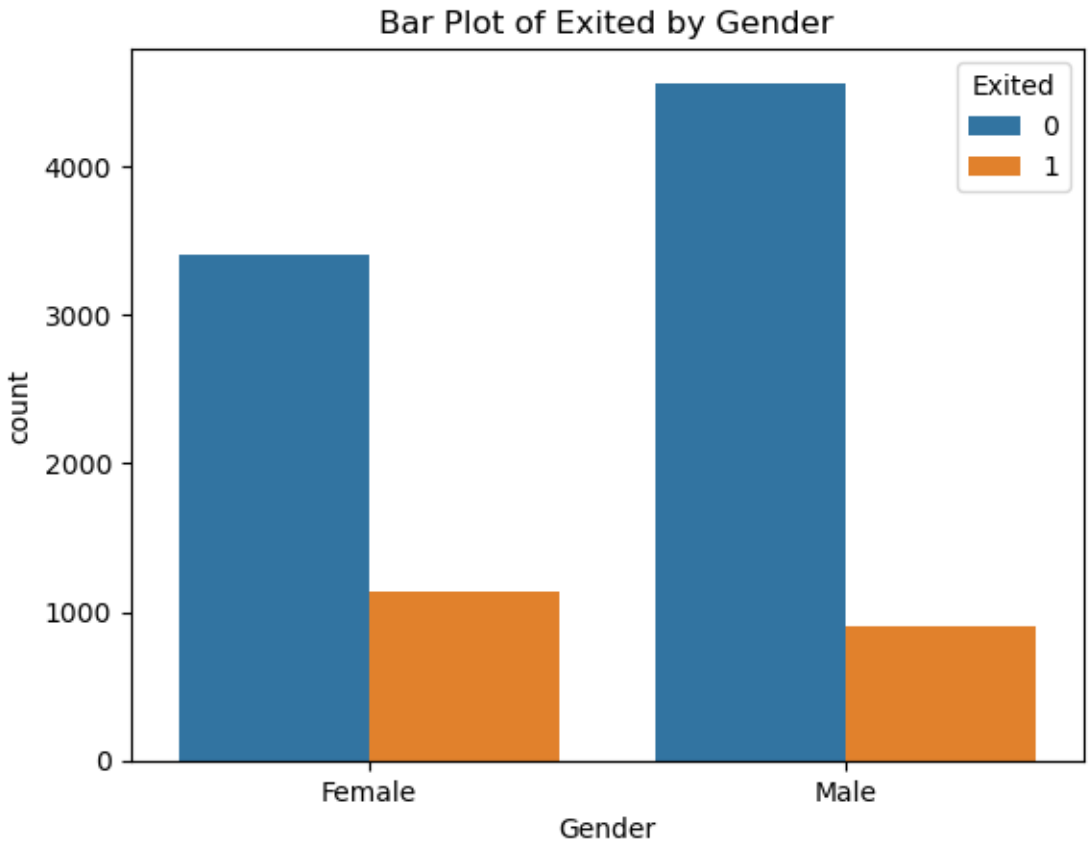
Our exploratory data analysis (EDA) of the CHURN dataset leveraged visualization tools to uncover patterns i n customer churn within a banking context. Starting with histograms, we differentiated between customers w ho stayed and those who left the bank, setting the stage for more detailed examinations.

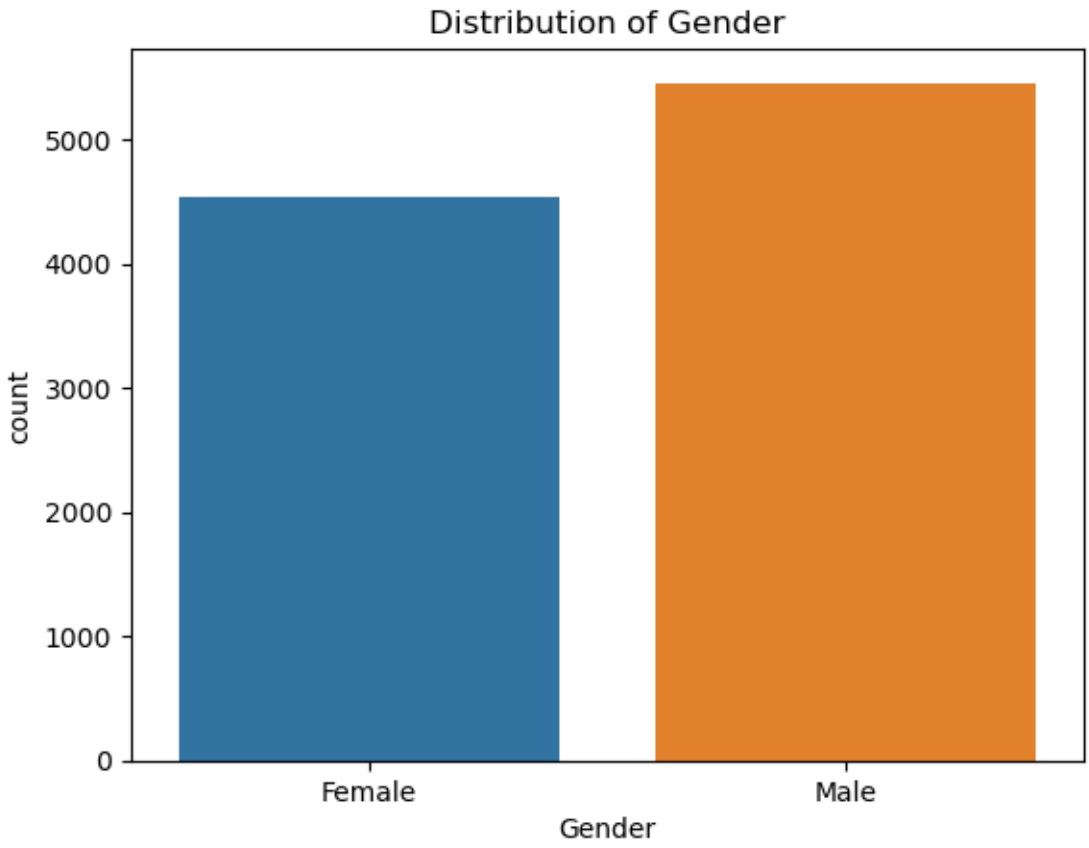
Violin plots for 'EstimatedSalary' and 'Age' revealed how these factors might influence churn, showing distrib utions and pinpointing age groups or salary ranges with higher churn tendencies. Bar plots were then used to assess the impact of gender on churn, identifying any gender-specific patterns that could inform targeted ret ention strategies.

We also analyzed financial data through box plots of customer balances, which helped identify outliers and t he spread of balances associated with churn risk. A scatter plot of 'Age' versus 'Balance', colored by churn st atus, allowed us to observe potential correlations and trends.

Furthermore, bar plots provided insights into the geographical and gender distribution within the dataset, ess ential for understanding demographic influences on customer behavior.

Overall, our EDA aimed to provide a comprehensive understanding of the factors affecting churn, using visual tools to highlight key trends and anomalies that could impact subsequent modeling efforts. This approach en sured a data-driven foundation for developing effective machine-learning models to predict customer churn.


Bar Plot of Exited by Gender
Exited
0
1
4000
3000
2000
1000
0
count
Female
Male
Gender




## Model Development

We started by taking great care in tackling the task and understanding the need for a thorough approach that includes data prep, algorithm choice, and performance evaluation.

Our journey kicked off with data preprocessing, a vital step in getting the dataset ready for training. We used one-hot encoding to turn categorical variables like 'Gender' and 'Country' into binary ones, making it easier to use them in the model.

Moving on to model training, we opted for a Support Vector Machine with a radial basis function (RBF) kernel. This choice was informed by SVM's capability to handle complex, non-linear relationships within the data. Additionally, we configured the model with balanced class weights to address the inherent class imbalance in churn observations. By doing so, we aimed to bolster the model's predictive prowess for both minority and majority classes, ensuring equitable treatment of all outcomes.

In essence, our approach entailed algorithmic selection, guided by a commitment to fairness and performance optimization. Through these endeavors, we sought to develop a robust and effective predictive model tailored to the twist of the CHURN dataset.

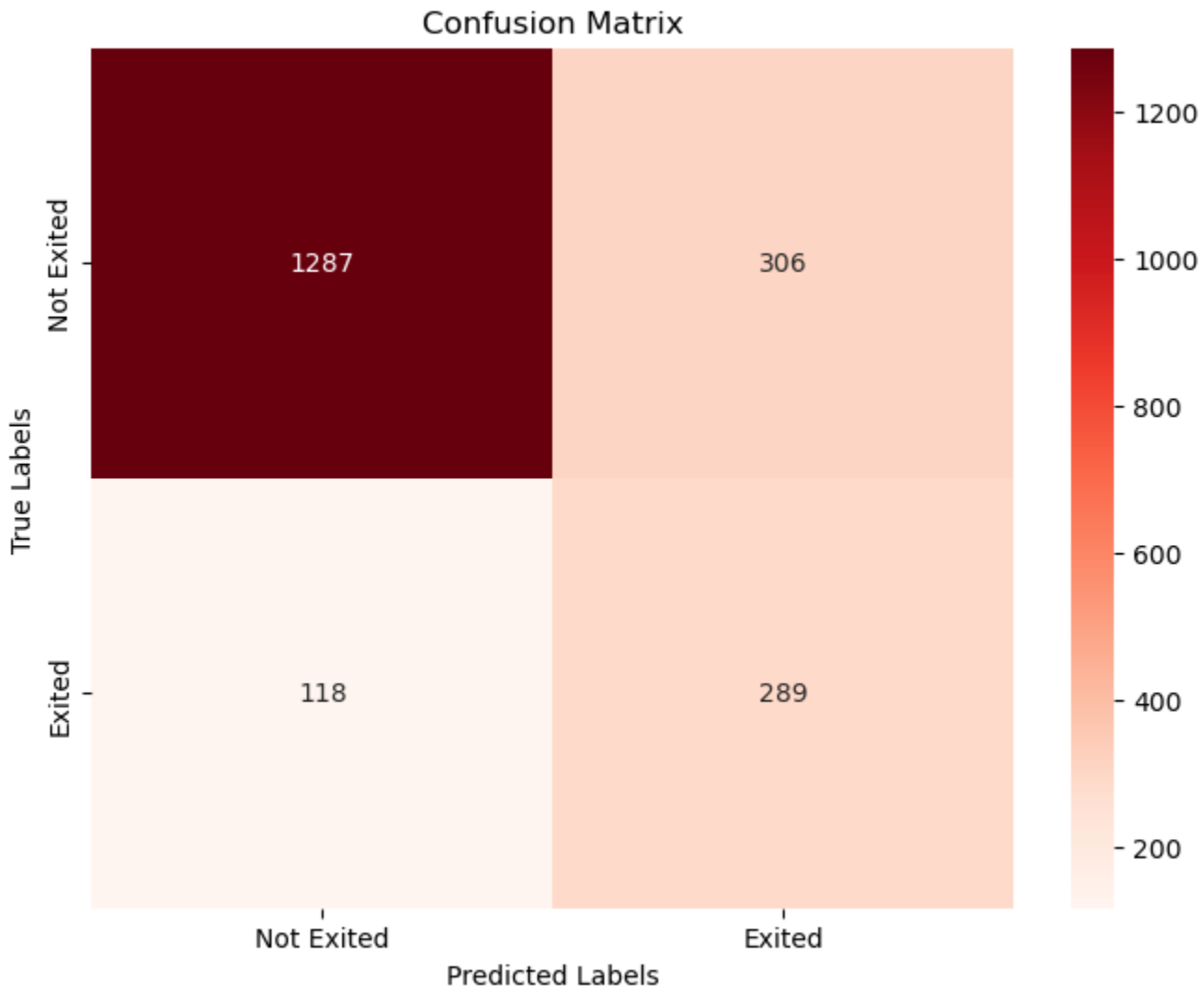
## Evaluation Metrics

For performance metric evaluation, we employed standard metrics: accuracy, precision, recall, and the F1 score. The model exhibited an accuracy of approximately 78.8% on the test data, reflecting its general ability to correctly classify customers' churn status. Precision, which quantifies the accuracy of positive predictions, stood at 82.84%, and recall, the model's ability to detect all positive instances, was also at 78.8%. The F1

score, which balances precision and recall, was 80.12%, indicating a robust model that fairly weighs false positives and negatives.

A confusion matrix visualized the model's performance, offering a breakdown of true positives, false positives, true negatives, and false negatives. This insight is critical for understanding the model's strengths and potential areas of improvement.

Overall, the SVM model's development process and the subsequent performance metrics have shown it to be a reliable tool for predicting customer churn, with a strong potential to guide strategic decisions in customer retention for the banking sector.




The confusion matrix is a cornerstone in evaluating the accuracy of a classification model. It shows the model's predictions against the actual outcomes. Here's the breakdown:

**(TP): 289 -** These are cases where the model correctly predicted that customers would not churn. It signifies the model's success in identifying the positive class (no churn).

**(TN): 1287** - This number represents the customers the model correctly identified as churn risks. It indicates how well the model can recognize the negative class (churn).

**(FP): 306 -** This denotes the customers who were incorrectly predicted to stay (no churn) when they churned. It is a type of error, where a positive result is wrongly observed for a negative condition.

**(FN): 118 -** Here, the model predicted churn where there was none. This is a type of error, where a negative result is incorrectly observed for a positive condition.

**Performance Metrics Explained**

- **Accuracy** is the proportion of true results among the total number of cases examined. It is expressed as:

  Accuracy = (TP + TN) / (TP + TN + FP + FN) = (289 + 1287) / (289 + 1287 + 306 + 118)

- **Precision** is the fraction of relevant instances among the retrieved instances. It focuses on the model's performance in predicting the no-churn cases:

  Precision = TP / (TP + FP) = 289 / (289 + 306)

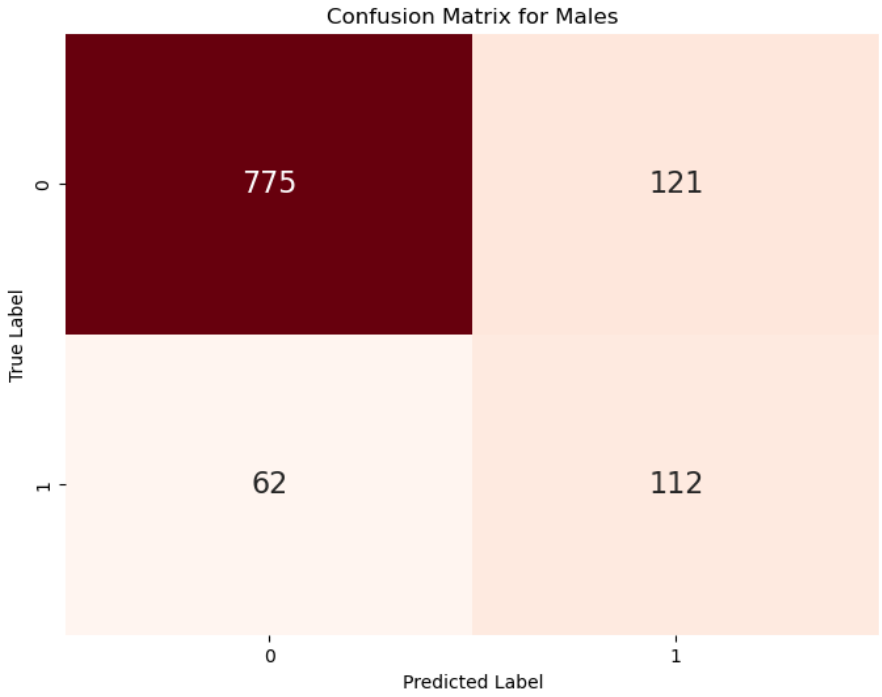
- **Recall**, also known as sensitivity or the True Positive Rate, is the fraction of relevant instances that were retrieved. It evaluates how well the model identifies customers who will not churn:
  Recall = TP / (TP + FN) = 289 / (289 + 118)
- **Positive Rate** is defined as the proportion of true positive predictions relative to the total number of positive predictions made by the model. It reflects how many of the instances predicted as positive are positive.
  Positive Rate = (TP + FP) / (TP + FP + TN +FN)

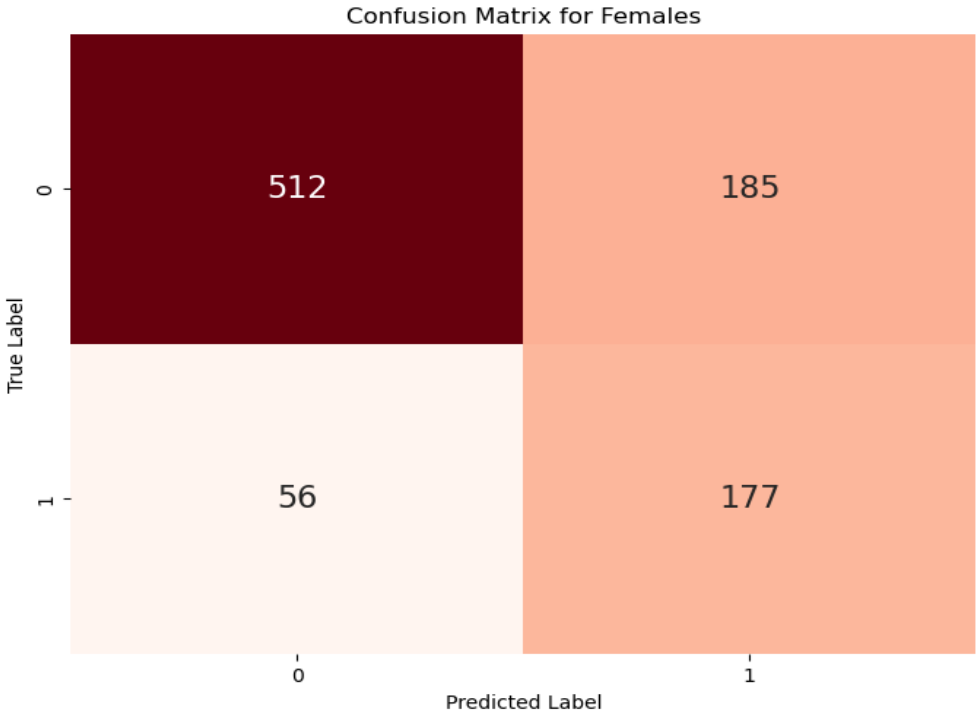
# Fairness Metrics by Group:

The model's performance was assessed using gender-segregated data. Fairness was evaluated using Equal Accuracy, Demographic Parity, and Equal Opportunity metrics, and traditional performance metrics were calculated using standard formulas:

**Confusion Matrix Counts:**

- Males: TP = 112, FP = 121, TN = 775, FN = 62
- Females: TP = 177, FP = 185, TN = 512, FN = 56




**Performance Metrics:**

**Male Performance:**

- Accuracy: Accuracy = (112 + 775) / (112 + 121 + 775 + 62) = 0.8289719626168224

- Precision: Precision = 112 / (112 + 121) = 0.48
- Recall: Recall = 112 / (112 + 62) = 0.64
- F1-Score: F1-Score = 2 * (0.48 * 0.64) / (0.48 + 0.64) = 0.55

**Female Performance:**

- Accuracy: Accuracy = (177 + 512) / (177 + 185 + 512 + 56) = 0.7408602150537634
- Precision: Precision = 177 / (177 + 185) = 0.49
- Recall: Recall = 177 / (177 + 56) = 0.76
- F1-Score: F1-Score = 2 * (0.49 * 0.76) / (0.49 + 0.76) = 0.59

# Fairness Metrics:

## ➢ Equal Accuracy

Equal Accuracy Difference measures the disparity in accuracy between different groups. It is calculated by taking the absolute difference between the accuracies of the two groups:

**Equal Accuracy** = |0.8289719626168224 - 0.7408602150537634|

Difference = 0.088111747563059

## ➢ Demographic Parity

Demographic Parity Difference compares the proportions of positive outcomes (both true and false positives) across different groups. It is calculated by dividing the difference in positive outcome rates by the total population size.

**Demographic Parity:**

Male: Demographic Parity = (112 + 121) / (112 + 121 + 775 + 62) = 0.21

Female: Demographic Parity = (177 + 185) / (177 + 185 + 512 + 56) = 0.38

Difference: |0.21 - 0.38| = 0.17

## ➢ Equal Opportunity

Equal Opportunity Difference focuses on the true positive rates between groups. It is calculated by taking the absolute difference between the true positive rates of the two groups.

**Equal Opportunity**:

Male: Equal Opportunity = 112 / (112 + 62) = 0.64

Female: Equal Opportunity = 177 / (177 + 56) = 0.76

Difference: |0.64 - 0.76| = 0.12

# Findings

The analysis of the model's performance and fairness metrics reveals a significant presence of bias, specifically gender bias, in predicting customer churn. The fairness metrics employed were Equal Accuracy, Demographic Parity, and Equal Opportunity, which highlighted discrepancies between the outcomes for male and female customers.

The result is an **Equal Accuracy Difference** of roughly 0.0881, which measures the disparity in accuracy between genders. A non-zero value suggests that the model's predictive performance is not uniform across male and female groups.

For **Demographic Parity**, which evaluates the equality of outcome across different groups, the model showed a difference of 17% (0.17) between males and females. Specifically, the model predicted males to churn at a rate of 21% and females at a higher rate of 38%. This suggests that the model is more likely to predict churn for females than males under similar circumstances, indicating a potential overestimation of churn risk among female customers.

**Equal Opportunity**, which focuses on the equality of true positive rates, also indicated a bias with a 12% (0.12) difference. It showed that females are more likely to be correctly identified as churners compared to males (76% vs. 64%), which could lead to disproportionate targeting of retention efforts towards female customers, potentially neglecting males who are also at risk of churning.

These findings underscore the necessity to address and correct gender bias in the churn prediction model. Not only does this bias lead to potential unfair treatment of customers based on gender, but it may also result in inefficient allocation of resources and missed opportunities for customer retention strategies. The evidence demonstrates that the model does not perform equitably across gender lines, necessitating further investigation and adjustment to ensure fairness and accuracy in predictive outcomes.

# Limitations

These metrics give us a glimpse into possible biases, but they have their limitations. They focus on simple outcomes and gender classifications, which don't fully capture the complexity of gender diversity. Also, they overlook other factors that could contribute to churn beyond what the model knows.

To grasp the biases and factors involved, we need to dig deeper into how the model makes decisions and the data it relies on. This requires a more comprehensive analysis that considers a wider range of features and factors, not just gender and binary outcomes. By gathering more diverse data, we can better understand what influences churn predictions.

In short, while these metrics are a good starting point, they're just the tip of the iceberg. A more thorough examination is needed to truly understand fairness and accuracy, ensuring our models reflect real-world complexity and treat everyone fairly.

# References


1. Becker, G. S., & Ostrom, E. (1993). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

2. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Abstraction in Sociotechnical Systems. *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 59-68.

3. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.

4. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency*, 77-91.

5. Elkan, C. (2001). The paradoxical success of fuzzy logic. *IEEE Expert, 16*(3), 11-11.

6. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *arXiv preprint arXiv:1908.09635*.

7. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *ACM Conference on Fairness, Accountability, and Transparency*, 59-68.