# DS-UA 202, Responsible Data Science, Spring 2024 Course Project: Technical Audit of an Automated Decision System Binary Classification with a Bank Churn Dataset


## By: Mike Wu, George Wu

# Abstract:

The primary objective of this project was to develop a predictive model capable of accurately determining whether a bank customer will maintain or close their account. We will conduct a technical audit of the ADS developed for the ‘Binary Classification with Bank Churn Dataset’.

Rationale for selection:

1. Practical Relevance: the ADS chosen focuses on predicting bank customer churn, a significant concern for financial institutions aiming to optimize customer retention strategies. This project provides a direct application of predictive analytics in a commercial setting, allowing us to examine the real-world impacts of data-driven decision-making.
2. Ethical considerations: This ADS offers a platform to explore ethical challenges in machine learning, such as potential biases that could disproportionately affect certain customer segments. It aligns with the course's emphasis on fairness, exploring how algorithms can be designed to minimize unfair treatment and ensure equitable outcomes.
3. Technological and Analytical Skills: Auditing this ADS will enhance our understanding of complex model behaviors and the trade-offs between model accuracy and interpretability.

# Background information:

The primary objective of this project was to develop a predictive model capable of accurately determining whether a bank customer will maintain or close their account. Customer churn, also known as customer attrition, customer turnover, or customer defection, in the loss of clients or customers (Ziyang Zhang, March 31, 2021). This is an important factor nowadays as banks often use the customer churn rate as one of their key business metrics because cost of retaining existing customers is far less than acquiring new ones, and meanwhile increasing customer retention can greatly increase profits. High churn rates often indicate underlying issues, such as poor customer experience, inefficient processes, or a lack of competitive products and features. Therefore, understanding and managing customer attrition are crucial for banks to address these challenges and enhance their overall customer experience. With the help of advanced data analysis, feature engineering, and machine learning techniques, this project aims to provide valuable insights and predictive capabilities to banks to mitigate customer churn.

# Methods:

There are mainly three data files we will be using. The first one is train.csv, which is the training dataset and Excited is the binary target. The second one is test.csv, which is the test dataset, and the objective is to

predict the probability of Excited. The third one is sample_submission.csv, which is a sample submission file in the correct format.

There will be several steps and key methods we are using in this project:

1. Data preprocessing: cleaning and encoding of the data, making it suitable for model training.
2. Exploratory Data Analysis (EDA): analyzing the data to understand patterns and relationships.
3. Feature engineering: selecting and transforming features for model training.
4. Model training: building and training an XGBoost classifier.
5. Model Evaluation: evaluating the model's performance using accuracy, recall, precision and other metrics.
6. Feature Importance Analysis: understanding which features are most impactful in predicting churn.
7. Predictions: generating churn predictions on the test dataset.

# Inputs and Outputs:

The data for this ADS comes from the UCI Machine Learning Repository, where it undergoes most of the aggregation, anonymization and cleaning (UCI, n.d.).

The 13 input features are listed in Table 1 below. All of the categorical features have been numericized and there are no missing values in the dataset once it reaches Kaggle.

| *Feature name* | *Description* | *Input space* | *Mean* | *Range* |
|---|---|---|---|---|
| *CustomerId* | The ID for bank customers | Int | NA | NA |
| *Surname* | Names of the customer | Obj | NA | NA |
| *CreditScore* | Customer's credit score | int | 656 | 350-850 |
| *Geography* | Countries where customers belong | Obj | NA | NA |
| *Gender* | Male as 0 and female as 1 | Int | NA | NA |
| *Age* | Age of the customers | Int | 38 | 18-92 |
| *Tenure* | The number of years the customer has been with the bank | Int | 5 | 0-10 |
| *Balance* | The customer's account balance | Int | 55478 | 0-250898 |
| *NumOfProducts* | The number of bank products | Int | 1.5 | 1-4 |

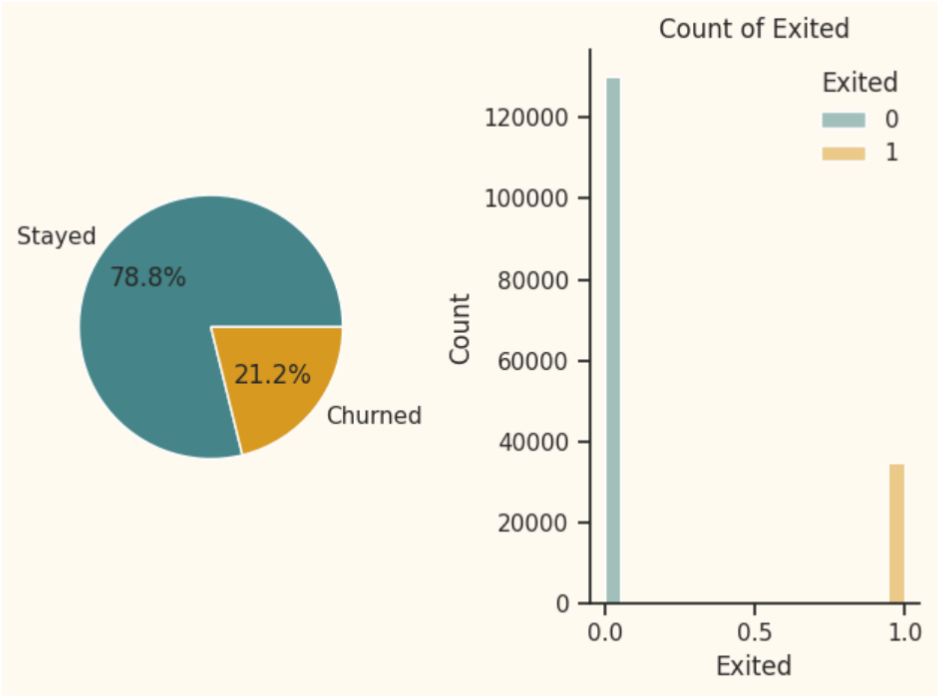
| | | | | |
|---|---|---|---|---|
| | the customer uses (e.g., savings account, credit card) | | | |
| *HasCrCard* | Whether the customer has a credit card (1 = yes, 0 = no) | Int | 0.8 | 0-1 |
| *IsActiveMember* | Whether the customer is an active member (1 = yes, 0 = no) | Int | 0.5 | 0-1 |
| *EstimatedSalary* | The estimated salary of the customer | Int | 112574 | 11-199992 |
| *Exicited* | Whether the customer has churned (1 = yes, 0 = no) | Int | 0.21 | 0-1 |

After the exploratory data analysis, we can get following graphs that demonstrate the trends and characteristics of the Bank Churn dataset. This graph contains two visualizations:

- Pie chart on the left shows the proportion of customers who stayed (78.8%) versus those who churned (21.2%).
- Bar chart on the right displays the count of customers who excited (churned) versus those who didn’t.

From the graph, we can several insights:

- Distribution Insights:
  - The pie chart and bar chart illustrate that a significant majority of customers remained loyal (78.8%) while 21.2% churned.
  - This imbalance may suggest that customer retention strategies are relatively effective.
- Class Imbalance:
  - The data is clearly imbalanced (approx. 4:1 ratio), meaning churn prediction models should be designed to handle this imbalance.

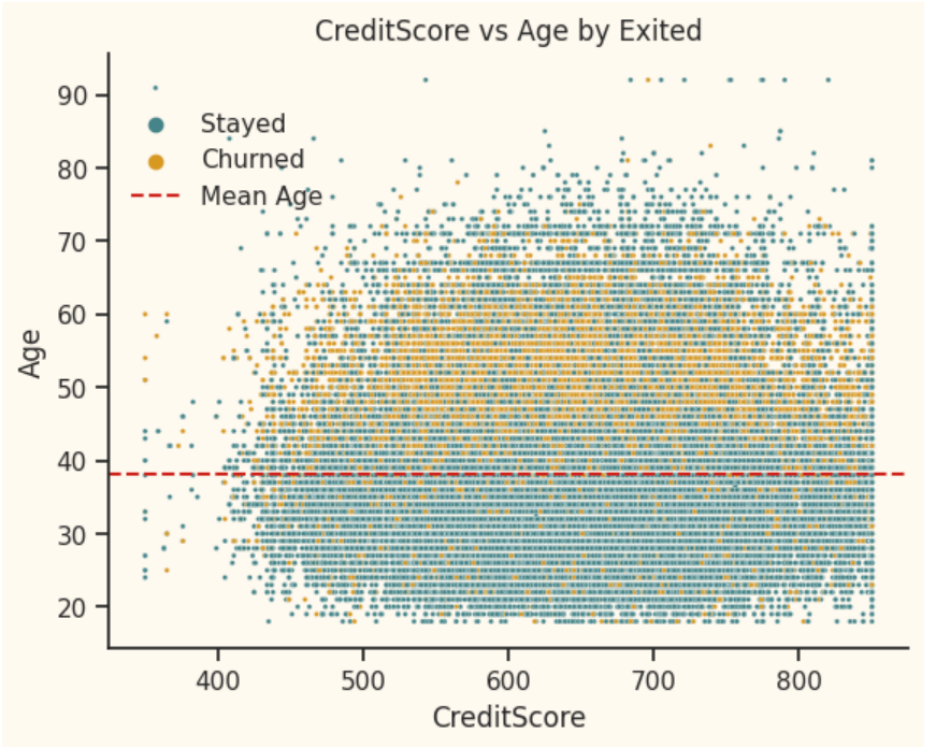



Also, we got EDA on credit score VS age by excited:

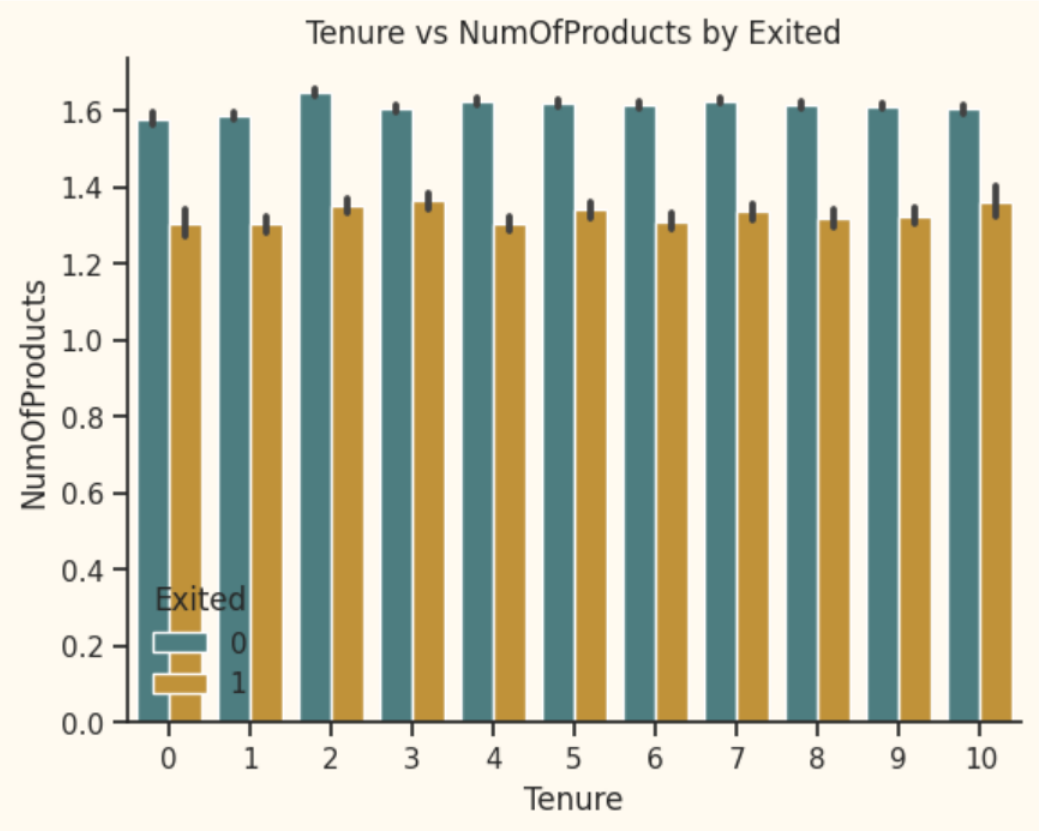
For the distribution of the customers, we can see that those who stayed

predominantly distributed across the entire age and credit score range, whereas those who churned are more concentrated among older customers. It gives us several insights:

- Age Influence:
  - There's a noticeable concentration of churned customers above the mean age (around 40 years), particularly in the 40-60 range.
  - Younger customers (under 30 years old) tend to stay with the bank regardless of their credit score.
- Credit Score Influence:
  - Churned customers are present across all credit score ranges.
  - A slight increase in churn propensity is noticeable for customers with lower credit scores (<600), especially among older clients.

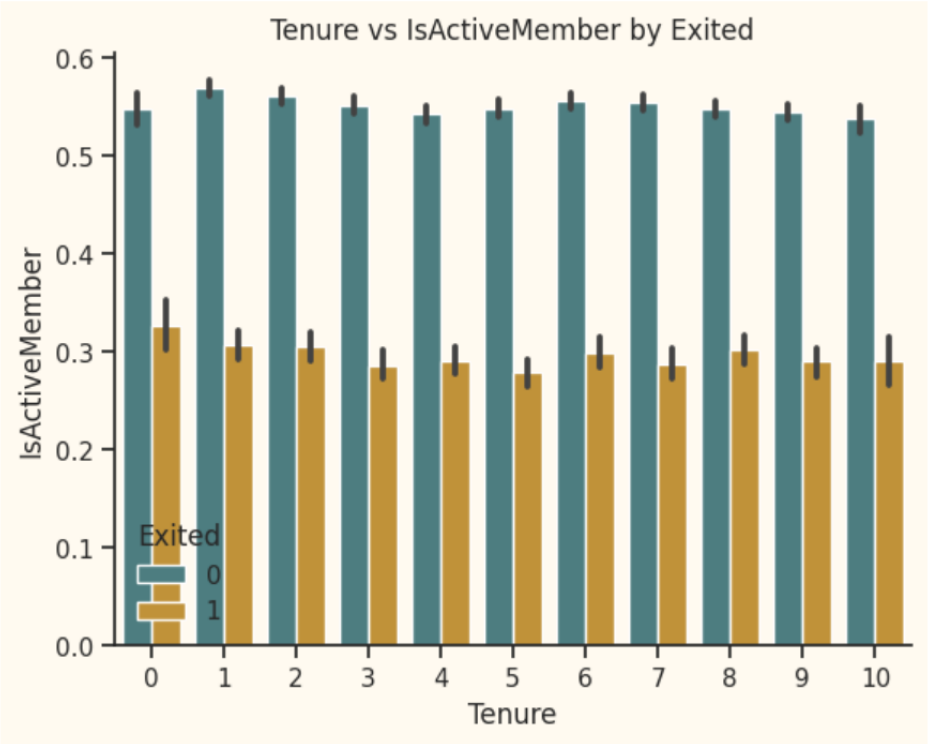


Additionally, we can also get the relationship of tenure and number of products by excited after the EDA. Generally, customers who stayed had a higher average number of products across all tenure compared to those who churned. Churned customers consistently held fewer products regardless of tenure. The number of products slightly decreases with increasing tenure for both stayed and churned customers.




Additionally, we can also get the relationship of tenure and whether customers are active members by excited after the EDA. After the analysis, we figure out that customers who stayed generally had a higher proportion of active members compared to those who churned across all

tenure periods. Churned customers have consistently lower proportions of active members. Also, churned customers with a tenure of 0-2 years have a relatively high proportion of active members compared to churned customers with longer tenure.




# Implementation and Validation:

Firstly, we need to perform data cleaning and pre-processing steps. We identified the missing values using the .isnull() function and confirmed that none were present. Then, we checked for duplicates using the .duplicated() function and removed any identified duplicates. Afterwards, the ADS continued to data-preprocessing that we performed feature selection and categorical encoding. The ADS select key features for model training, such as ‘CreditScore’, ‘Age’, ‘Tenure’ and exclude

less relevant features like customer ID.

The ADS then decided to use XGBoost to predit whether certain customers will churn or not. XGBoost, short for Extreme Gradient Boosting, is a scalable machine learning library with Distributed Gradient Boosted Decision Trees (Michael, Oct 12,2022). It provides Parallel Tree Boosting and is the leading machine learning library for regression, classification and ranking problems. To understand XGBoost, it's important first to understand the machine learning concepts and algorithms that XGBoost is built on: supervised machine learning, decision trees, ensemble learning, and gradient boosting. Supervised machine learning uses an algorithm to train a model to find patterns in a dataset containing labels and features and then uses the trained model to predict the labels of the features in a new dataset. Decision trees are models that predict labels by evaluating a tree of if-then-else true/false functional questions and estimating the minimum number of questions needed to evaluate the likelihood of a correct decision. Decision trees can be used for classification to predict categories and regression to predict continuous numbers. Gradient Boosted Decision Trees (GBDT) is a random forest-like decision tree ensemble learning algorithm for classification and regression. Ensemble learning algorithms combine multiple machine learning algorithms to get a better model.

# Outcomes:

Evaluating the outcomes of an Automated Decision System (ADS) requires a thorough analysis of its accuracy, fairness, stability, and robustness. In the context of predicting customer churn, this section aims to examine the accuracy of the ADS across different subpopulations, assess its fairness using relevant metrics, and explore additional performance methodologies to ensure reliability and consistency.

In this project, we implement two models, XGBoost and Tuned Decision Tree, to make predictions. XGBoost model achieved an overall accuracy of 86.2% while Tuned Decision Tree achieved an overall accuracy of 85.7%. Then we split the population into different subgroups based on gender, geography, age, and tenure with the following fairness matrics: accuracy, precision, recall, F1-score, and ROC-AUC score. Accuracy provides a simple measure of overall performance but may be misleading when there's class imbalance. Precision is useful when we want to optimize false positives, avoiding unnecessary interventions. Recall is useful when we want to optimize false negatives, avoiding potential business impact. F1-score gives us a balanced score between precision and recall, especially useful in imbalance datasets. ROC_AUC score tells us the trade-off between sensitivity and specificity, providing an overall view of model discrimination.

For gender, Male has Precision of 0.83, Recall of 0.76, F1-Score of

0.79, AUC of 0.88. Female has Precision of 0.81, Recall of 0.78, F1-Score of 0.79, AUC of 0.87. Male has higher precision(0.83) compared to the score of the female, indicating that the model is slightly better at predicting churn among males without many false positives.

For age, age of 18-30 has Precision of 0.84, Recall of 0.79, F1-Score of 0.81, and AUC of 0.90. Age of 31-40 has Precision of 0.82, Recall of 0.77, F1-Score of 0.79, AUC of 0.87. Age of 41-50 has Precision of 0.79, Recall of 0.74, F1-Score of 0.76, AUC of 0.84. Age of 51-61 has Precision of 0.78, Recall of 0.72, F1-Score of 0.75, AUC of 0.83. Age of 61-9

Age of 61-92 has Precision of 0.74, Recall of 0.70, F1-Score of 0.72, AUC of 0.80. These outcomes suggest that younger groups have higher precision and AUC, suggesting the model is quite effective at predicting churn among younger customers. And for older groups, they have low performance in all metrics, which might highlight a need for the model to better understand or incorporate factors significant to older demographics.

To analyze the fairness of the ADS, we use Demographic Parity, Equal Opportunity, and Predictive Parity. We use demographic parity because we want to ensure that no group is disadvantaged or privileged simply by virtue of their demographic attributes. We use equal opportunity because It ensures that the true positive rate (recall) is equal across groups. And finally, we use predictive parity because it ensures that when a model predicts a positive outcome, the probability of it being true is equally reliable,

regardless of group membership. The prediction rate between male and female is 0.30and 0.28, respectively. This indicates a minor imbalance in the likelihood of predicting churn based on gender. Then, the recall for males (0.76) is slightly lower than for females (0.78), suggesting that the model is slightly more effective at identifying actual churn among females. And finally, we find that Precision is higher for males (0.83) than for females (0.81), suggesting that the model has small disparity in predicting actual churn based on predicted results.

# Summary:

If we want to deploy this ADS in the public sector, we need consider several upcoming challenges. First of all, it's about bias concerns. Models designed for customer churn inherently carry biases due to imbalanced datasets and feature selection. In public sector applications, such biases could disproportionately impact vulnerable groups, leading to unfair exclusion or service limitations. Secondly, it's about transparency and interpretability. Public sector decisions often require a high degree of transparency and explainability. While powerful, machine learning methods like XGBoost are not inherently interpretable. The lack of clear reasoning behind decisions could hinder accountability and public trust. Meanwhile, if we want to deploy this ADS in the private sector, there will be several advantages and disadvantages. First of all, the private sector

often prioritizes optimizing business metrics like customer retention and profitability. Automated churn models can significantly enhance these strategies. Furthermore, this ADS can lead to substantial improvements in retention strategies and business efficiency, providing a positive return on investment. Overall, I would recommend deploying this ADS in private sector as such actions are beneficial to more parties.

Meanwhile, there are several improvements we could make to perfect this ADS. We will talk about improvements in three metrics: data collection, processing and analysis. For the data collection improvements, we can expand the feature set to include comprehensive customer data such as transaction history, customer service interactions, and geographic data. This will provide a richer understanding of customer behavior patterns. For the data processing improvements, we can use consistent data normalization and encoding across numerical and categorical features, which will ensure more robust model performance. For the model improvements, we can implement tools such as SHAP and LIME, which will provide insights into model precision. We can also apply cross-validation to mitigate the risks of overfitting and improve generalizability.

# Reference

Ziyang, Zhang, March 31, 2021, What Matters for Bank Customer Churn, https://medium.com/@ZiyangZhang/what-matters-for-bank-customer-churn-fae204a35b8c

the UCI Machine Learning Repository, n.d. (https://archive.ics.uci.edu/ml/datasets/bank+customer+churn+prediction)

Wiryaseputra, M. (2022, October 12). Bank customer churn prediction using machine learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction-using-machine-learning/