**Machine Learning II Final Project**

Group 1: Allison Ko, Yuxuan Chen, Balajigowda H S

**Introduction**

In the dynamic realm of competitive banking, retaining customers is crucial, as their departure can lead to revenue loss from both lending activities and financial transactions. Traditional methods of detecting churn may not suffice in effectively tackling this challenge. Thus, our proposed solution involves employing predictive analysis techniques to address customer churn.

By delving into historical banking data, we aim to develop models that forecast the likelihood of a customer departing. This proactive approach allows us to identify high-risk customers early and implement targeted retention strategies. These strategies could involve personalized interactions and incentives, all designed to persuade customers to remain loyal to the bank. Through the adoption of predictive analytics for churn management, the bank can reinforce its revenue streams and bolster long-term profitability.

**Dataset Summary:**

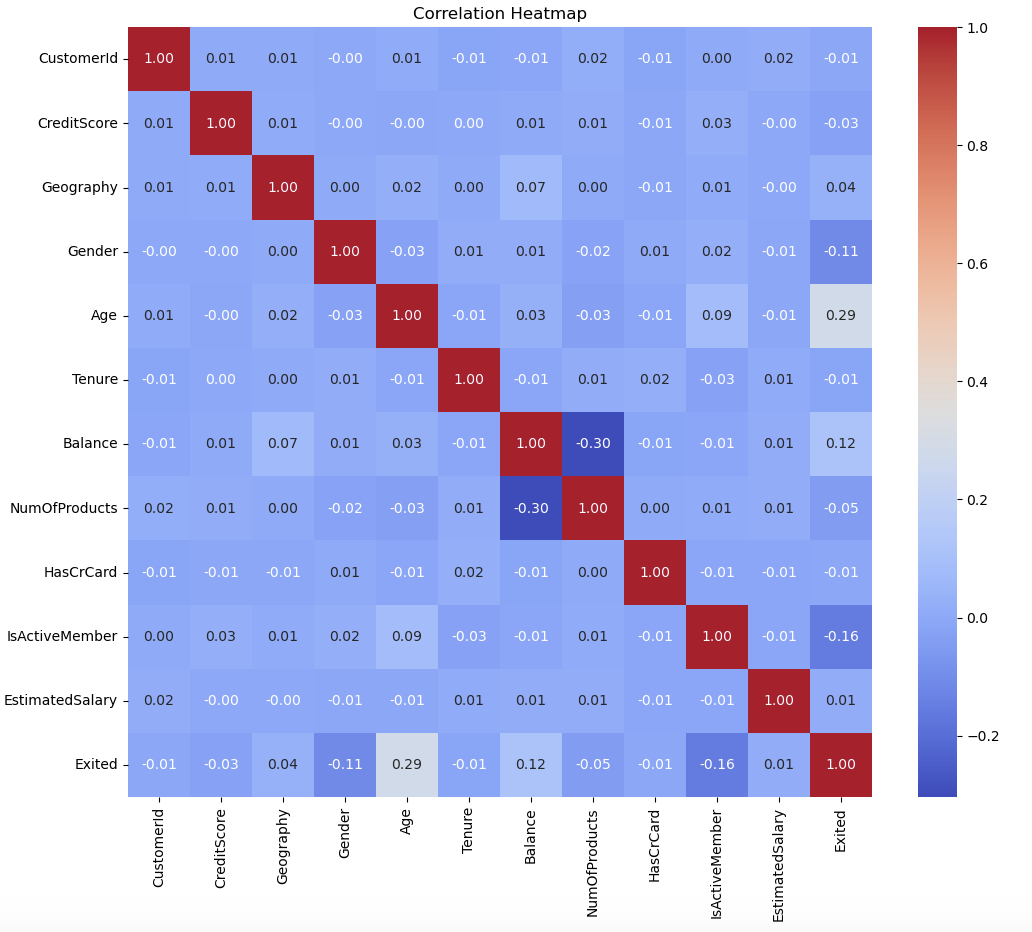
We obtained the dataset from Kaggle on the [Churn Problem for Bank Customers](https://www.kaggle.com/code/mathchi/churn-problem-for-bank-customer/input). It consists of 10,000 observations with 12 independent variables specified about customer information. The dependent variable is “Exited” which refers to whether the customer has left the bank or not.

**Table 1. Data Dictionary**

| **Column Name** | **Description** |
| --- | --- |
| **RowNumber** | Corresponds to the record (row) number and has no effect on the output. |
| **CustomerId** | Contains random values and has no effect on customers leaving the bank. |
| **Surname** | The surname of a customer has no impact on their decision to leave the bank. |
| **CreditScore** | Can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank. |
| **Geography** | A customer’s location can affect their decision to leave the bank. |
| **Gender** | It’s interesting to explore whether gender plays a role in a customer leaving the bank. |
| **Age** | This is certainly relevant since older customers are less likely to leave their bank than younger ones. |
| **Tenure** | Refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank. |
| **Balance** | Also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances. |
| **NumOfProducts** | Refers to the number of products that a customer has purchased through the bank. |
| **HasCrCard** | Denotes whether or not a customer has a credit card. This column is also relevant since people with credit cards are less likely to leave the bank. |
| **IsActiveMember** | Active customers are less likely to leave the bank. |
| **EstimatedSalary** | As with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries. |
| **Exited** | Whether or not the customer left the bank. |

**Data Prepossessing**

We've confirmed that our dataset doesn't contain any missing values. By converting the categorical variables “Geography” and “Gender” into numerical variables, we will be able to use them for conducting a correlation heatmap, showing the correlation between each variable before we proceed with modeling.



**Figure 1. Correlation Heatmap**

The correlation heatmap provides a visual representation of the relationships between various customer attributes and their churn status. Notably, “Age” and “Balance” show a moderately positive correlation with the likelihood of customer churn, as indicated by their correlation coefficients with the “Exited” variable. Negative correlations are observed between the number of products a customer uses and their churn status, suggesting that customers with more products are less likely to leave. The heatmap also highlights the lack of strong correlations between variables such as “Geography” or “CustomerId” with churn, leading to the decision to drop these from the modeling process. The clear blocks of high correlation among identical variables across the diagonal confirm the heatmap's accuracy in depicting these relationships.

**Modeling**

We analyzed the performance of logistic regression, decision tree, random forest, and XGBoost models using precision, recall, and F1-score metrics. The analysis focuses on their ability to identify churn (Class 1) versus non-churn (Class 0). Logistic regression performs exceptionally well in identifying non-churn customers, achieving high precision (0.86) and excellent recall (0.95), resulting in a strong F1-score of 0.90. However, its performance in detecting churn customers is considerably weaker, with a precision of 0.65 and a low recall of 0.37, leading to a modest F1-score of 0.47. This indicates that while logistic regression is effective at recognizing non-churn instances, it struggles with accurately identifying churn cases.

The decision tree model shows balanced metrics for non-churn predictions, with both precision and recall at 0.86, mirrored by an F1-score of 0.86. However, for churn predictions, the model's performance declines sharply, evidenced by both precision and recall at 0.45 and an equally low F1-score of 0.45. This demonstrates the model's limitations in effectively distinguishing churn cases. Random forest stands out with the highest precision (0.88) and recall (0.96) for non-churn predictions, reaching the highest F1-score of 0.92. In churn predictions, it also leads with a precision of 0.73 and a recall of 0.45, resulting in the highest F1-score for churn at 0.56. Random forest shows a relatively better capability at handling churn predictions compared to the other models. XGBoost closely follows random forest in performance for non-churn predictions, with a precision of 0.87 and a recall of 0.95, leading to an F1-score of 0.91. For churn predictions, it shows a moderate precision of 0.67 and a recall of 0.44, resulting in an F1-score of 0.53. XGBoost demonstrates a reasonable ability to classify churn cases effectively.

Overall, the random forest model and XGBoost emerge as the most effective across all metrics, particularly in the challenging task of churn prediction. It shows robust performance in classifying both churn and non-churn instances, proving to be the superior model in this analysis.

**Table 2. Accuracy Metrics**

| **Model** | **Class** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic | 0 | 0.86 | 0.95 | 0.90 |
| 1 | 0.65 | 0.37 | 0.47 |
| Decision Tree | 0 | 0.86 | 0.86 | 0.86 |
| 1 | 0.45 | 0.45 | 0.45 |
| Random Forest | 0 | 0.88 | 0.96 | 0.92 |
| 1 | 0.73 | 0.45 | 0.56 |
| XGBoost | 0 | 0.87 | 0.95 | 0.91 |
| 1 | 0.67 | 0.44 | 0.53 |

Next, we evaluated the performance of the four models using receiver operating characteristic (ROC) curves. Each model's effectiveness is assessed by the area under the curve (AUC), which measures the ability to distinguish between churn and non-churn classes.

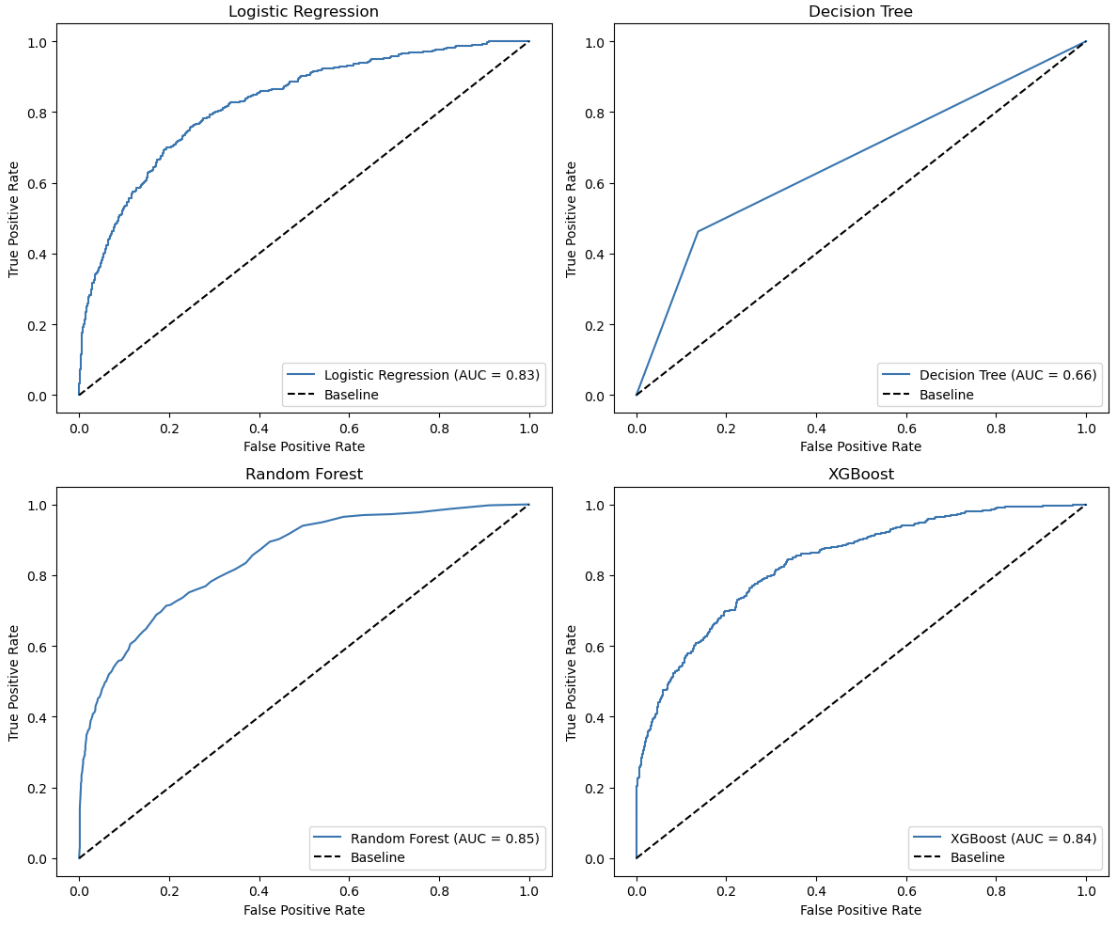
The logistic regression model demonstrates a commendable balance between sensitivity and specificity, achieving an AUC of 0.8253. This indicates that the model performs significantly better than random chance at distinguishing between the classes, as evidenced by an ROC curve that stays well above the baseline. On the other hand, the decision tree model underperforms, with an AUC closely mirroring 0.6. This suggests that it has limited ability to differentiate between classes, performing only slightly better than a random guess. In contrast, the random forest model exhibits superior classification abilities, achieving the highest AUC of 0.8468 among the models evaluated. Its ROC curve approaches the ideal top left corner, indicative of a high true positive rate and a low false positive rate, demonstrating its efficacy in correctly classifying churn. Similarly, the XGBoost model performs competitively, with a robust AUC of 0.8364. Its ROC curve closely resembles that of the random forest, indicating strong classification capabilities.

While the decision tree shows limitations, logistic regression, though adequate, does not reach the predictive power of the ensemble methods, random forest and XGBoost emerge as the most effective models for predicting customer churn, underscored by their higher AUC scores and favorable positions on the ROC curves.

**Table 3. Area Under the Curve (AUC)**

| **Model** | **AUC (OVO)** | **AUC (OVR)** |
| --- | --- | --- |
| Logistic Regression | 0.8253 | 0.8253 |
| Decision Tree | 0.6625 | 0.6625 |
| Random Forest | 0.8468 | 0.8468 |
| XGBoost | 0.8364 | 0.8364 |

**Figure 2. ROC Curve**



Finally, in evaluating the effectiveness of the models, we focus on the true positive rates, particularly when the business strategy emphasizes interventions such as offering incentives, which have associated costs.

Logistic regression identified 148 true positives but had a significant number of false positives (1,524). This indicates moderate effectiveness in predicting churn, coupled with a high likelihood of misclassifying non-churn customers as churn risks. The decision tree model improved the true positive count to 178, reducing false positives to 1,380, and showcasing better precision compared to logistic regression. Random forest slightly outperformed the decision tree model in detecting true positives (180) but with the highest false positives (1,535). This might suggest potential overfitting. XGBoost exhibited the best balance, with 176 true positives and the lowest count of false positives (1,517). This balance makes it particularly effective for churn prediction, as it maintains a strong identification rate for actual churn cases while minimizing the risk of false identification.

While XGBoost and random forest have very similar ROC, AUC, and accuracy metrics, we chose XGBoost over random forest due to its superior balance in the confusion matrix. XGBoost provides a more favorable ratio of false positives to true positives, reducing unnecessary spending on customers who are unlikely to leave while still accurately retaining valuable customers at risk. This balance allows the bank to target incentives effectively, maximizing revenue and improving customer retention.

**Table 4. Confusion Matrix**

| **Model** | **Confusion Matrix** | | | |
| --- | --- | --- | --- | --- |
| **TP** | **TN** | **FP** | **FN** |
| Logistic Regression | 148 | 78 | 1524 | 250 |
| Decision Tree | 178 | 222 | 1380 | 220 |
| Random Forest | 180 | 67 | 1535 | 218 |
| XGBoost | 176 | 85 | 1517 | 222 |

Overall, random forest and XGBoost consistently outperform the other models in predicting customer churn. While logistic regression is effective at recognizing non-churn cases, it struggles with churn detection. Decision tree demonstrates limitations in distinguishing between churn and non-churn classes. As a result, we conclude that XGBoost emerges as the most reliable model, combining high accuracy with robust classification capabilities.

In this analysis, we interpret the feature importance plots for the XGBoost model used in predicting customer churn. Understanding which features contribute most to churn predictions can help the bank develop targeted strategies to reduce customer attrition.

XGBoost highlights estimated salary and credit score as the most significant predictors, with higher salaries and lower credit scores increasing the likelihood of churn. Balance and age also contribute to churn predictions, alongside active membership status and the number of products.

**DECISION PERFORMANCE EVALUATION**

**Bank Revenue Streams**

Banks generate revenue through various channels, primarily leveraging funds from user deposits. These funds are utilized in lending and investments, which constitute the core revenue-generating activities. Let's delve into the breakdown of these revenue sources:

1. **Lending Activities:**
   * **Consumer Loans:** These encompass personal loans, auto loans, and credit card loans tailored to individual customers, typically yielding interest rates ranging from 4% to 15%, contributing to overall returns.
   * **Mortgages:** Providing loans for purchasing residential and commercial properties, generating interest income over the loan tenure, with average returns ranging from 3% to 6%.
   * **Business Loans:** Offering financial assistance to enterprises for expansion, operations, or other business purposes, with average returns varying from 5% to 10%.
   * **Corporate Loans:** Extending credit facilities to large corporations for diverse financial needs, yielding returns typically between 3% to 8%.
2. **Investment Activities:**
   * **Stock Investments:** Banks may invest in equities, aiming for capital appreciation and dividends, with average returns historically averaging around 25% to 30% annually.
   * **Mutual Funds:** Investing in diversified portfolios of stocks, bonds, or other assets, managed by professional fund managers, with average returns ranging from 5% to 8%.
   * **Government Bonds:** Purchasing bonds issued by governmental entities, generating interest income, with average returns varying from 1% to 4%.
   * **Corporate Bonds:** Investing in debt securities issued by corporations, providing fixed interest payments, with average returns typically ranging from 3% to 6%.

In a typical quarter, banks aim for robust returns from their investments and lending activities. With a strategic blend of consumer loans, mortgages, business loans, corporate financing, and diverse investments including stocks, mutual funds, government and corporate bonds, banks anticipate earning substantial returns. On average, banks expect to generate around 15% to 18% returns on their investments per quarter. This encompasses interest income from loans, dividends and capital appreciation from equities, fixed income returns from bonds, and various service charges. However, it's important to note that these returns are subject to market fluctuations, economic conditions, and the efficacy of risk management strategies employed by the bank. Despite these variables, banks maintain a consistent focus on optimizing returns to sustain profitability and meet stakeholder expectations. (*Note: All the numbers and the returns which are mentioned are taken from the Investopedia*)

**Banks' Revenue Before Using the Machine Learning Model:**

**Revenue =** ((Balance\*0.7) + Estimated Salary\*(0.3/12) \*3)\*0.15) − ((Balance\*0.7 + Estimated Salary\*((0.3/12)​)\*3)×0.03) + 50

This formula predicts the quarterly revenue generated by an account, factoring in the account holder's saving strategy and the bank's investment approach over a three-month period. Initially, it divides the account balance and monthly salary into savings, allocating 70% of the balance and 30% of the salary to this purpose. These savings are then utilized by the bank for investments, targeting an average return of 15%. Additionally, recognizing the account holder's commitment to saving, the bank pays them a 3% interest rate on their saved funds for the quarter, and an annual maintenance charge of $50.

In essence, this formula provides a short-term outlook, forecasting the quarterly revenue based on the symbiotic relationship between the account holder's savings behavior and the bank's investment activities. By considering these factors, it offers insights into the potential revenue generation for both the account holder and the bank within the specific timeframe of one quarter.

Utilizing the provided formula, we conducted revenue projections for a single quarter for the bank, based on data from the ‘X\_Test’ dataset, encompassing approximately 2,000 account holders. The calculated revenue amounted to approximately $11.43 million. This revenue estimation was derived by individually assessing each account holder's financial profile, including their account balances and estimated monthly salaries, and factoring in the bank's investment strategy and interest payments on savings. The resultant figure provides valuable insights into the bank's anticipated revenue generation over the specified quarter, aiding in financial planning and performance evaluation.

**Banks Revenue After Using the Machine Learning model:**

After testing various machine learning models, XGBoost emerged as the most accurate in predicting revenue using the 'X\_test' dataset. Analyzing the results through a confusion matrix, we gained insights into revenue implications for different prediction scenarios:

**Revenue2** = ((Balance\*0.7) + Estimated Salary\*(0.3/12)\*3)\*0.15) − ((Balance\*0.7 + Estimated Salary\*((0.5/12)​)\*3) × 0.03) + 50

As an incentive for account holders predicted to leave the bank according to the model, the interest rate on their savings account will be increased from 3% to 5% annually. This adjustment aims to encourage these customers to reconsider their decision and remain with the bank. By offering a higher interest rate, the bank seeks to provide added value to these customers, potentially enhancing their loyalty and retention. This proactive measure aligns with the bank's goal of prioritizing customer satisfaction and retention, ultimately contributing to long-term business success.

* **False Negatives (y\_pred = 0) & (y\_test = 1)**: These cases, where churn wasn't predicted accurately, led to an estimated revenue loss of about $1.37 million. It highlights missed opportunities to retain potentially leaving customers.
* **False Positives (y\_pred = 1) & (y\_test = 0)**: Instances where churn was predicted incorrectly resulted in a revenue of approximately $1.73 million. This represents unnecessary spending to retain customers who were unlikely to leave.
* **Non-Churned Customers**: Accurate predictions allowed us to retain existing customers, generating revenue of around $9.67 million.
* **Churned Customers**: Despite efforts, some customers left, resulting in a revenue loss of about $2.01 million, which we aim to recover by applying the XGBoost model. By implementing targeted business strategies to retain these at-risk customers, we expect to increase revenue and improve overall customer retention.

**Total Revenue** = Churned Customers + Non-Churned Customers + False Positives

*(We are not considering the False Negative because of the wrong prediction of the model we cannot give any incentives for the people.)*

The total revenue for one quarter is $13.42 million after employing the machine learning model. Before using the model, the revenue was $11.43 million. So, the difference in revenue after using the model is $1.98 million. This difference highlights the impact of the machine learning model in optimizing revenue. Specifically, revenue from customers who were predicted to churn (but stayed), and those who were predicted to stay (but left), contributed significantly to this difference. Overall, the model helped in retaining valuable customers and maximizing revenue, showcasing the benefits of leveraging data-driven insights in business decision-making.

**Conclusion**

The implementation of XGBoost model significantly optimized the bank’s customer retention strategy, resulting in a $1.98 million revenue increase per quarter. This emphasizes the importance of data-driven decision making in identifying at risk customers and applying our targeted incentives to retain customers. While predictive accuracy can still be improved to minimize revenue losses from false negatives, the overall strategy demonstrates the benefit of machine learning models in enhancing profitability.

**Recommendations**

We're looking to develop an incentive plan to prevent customers from leaving the bank. We'll be sending out mail, email, and text messages to those customers who are at risk of leaving, offering them an increased interest rate on their debit card savings accounts. Mailing campaigns are a traditional form of communication. We will use it to target inactive customers as well as customers who are older than 60 years old. Email campaigns will directly towards the younger generation who are more responsive to online communication. Text messages promotion will target the younger generation and middle-aged customers who are younger than 60 years old, assuming these people are adapted to using smartphones.

In addition, we'll be encouraging these potentially departing customers to refer a friend to open a new bank account. If their friend keeps a $500 balance for one month, both the current bank customer and their friend will receive a bonus $100 cash deposit into their bank accounts. This approach will not only help us retain our current customers but also attract new ones, ultimately resulting in increased deposits and revenue.

If certain customers are considering leaving the bank due to monthly maintenance fees, the bank might entice them to stay by offering a waiver of these fees. This offer could be contingent upon the customer agreeing to remain with the bank for an additional year, increase their deposit amount, and maintain their account balance for a specified period.

**Teamwork Contribution**

Allison Ko, Yuxuan Chen, and Balajigowda H S implemented the ML models; evaluated the accuracy and decision performances; and wrote the final report.