ORIE 5741: Bank Customer Churn Prediction

Chang Chen, Ryan Ren, Jason Pan

1 Abstract

Our project aims to predict the bank customer churn using various machine learning techniques such as logistic regression, random forest, and support vector machine. We believe that it is important for banks to know the factors that lead a client toward the decision to leave the company. For the banks, it is more expensive to sign in a client than keeping an existing one, so this analysis will help them develop churn prevention strategies that would help them keep as many customers as possible. The preliminary results of our analysis indicate that the attribute complaint dominates the prediction as complaints often lead to irreversible customer churn. In order to improve our models, we re-implemented the three models without the complaint feature, and observed significant decrease in F1 scores. Consequently, we implemented measures such as adjusting threshold, hyperparameter tuning, and XG boost, which brought the F1 score back to the 0.6 range. These model improvement strategies are more robust as it reduces the inherent bias within the feature.

2 Introduction

In an increasingly competitive banking landscape, retaining customers became paramount for sustained profitability and growth. Customer churn, which is the phenomenon of customers discontinuing their services with a bank, poses a significant challenge for financial institutions worldwide. Factors contributing to churn are multifaceted and may include dissatisfaction with services, competitive offerings from rival banks, life events, or economic factors. Understanding the factors influencing customer churn is crucial for devising effective retention strategies.

Traditionally, banks have relied on heuristic rules and demographic segmentation to identify at-risk customers. However, these approaches often lack accuracy and fail to capture the complex interplay of factors influencing churn behavior. Machine learning offers a promising solution to this challenge. Through analysis of historical customer data, machine learning models like logistic regression and random forest can uncover subtle indicators of churn. They will also offer insights into identifying at-risk customers preemptively and implementing targeted retention initiatives.

3 Data Description

The dataset used is from the Kaggle website. It contains customer data of account holders at an anonymous multinational bank. The entire dataset consists of 10,000 rows and each row represents a customer with a unique identifier. It has a total of

18 columns, but we will drop three columns (row number, customer ID, and surname) during the modeling stage. We have summarized the remaining columns below.

- **Credit Score**: An integer that represents the customer's social security credit score. Normally, customers with higher credit scores are less likely to leave the bank.
- **Geography**: A customer's location in one of the three categories: France, Spain, or Germany.
- **Gender**: A customer's gender (Male or Female).
- **Age**: An integer representing the customer's age. The value ranges from 19 to 92.
- **Tenure**: An integer representing the number of years the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
- **Balance**: A float point number representing the customer's remaining balance at the bank.
- **NumOfProducts**: An integer representing the number of products that a customer has purchased through the bank. The value ranges from 1 to 4.
- **HasCrCard**: A binary integer (1 or 0) indicating whether a customer has a credit card. Normally, people with a credit card are less likely to leave the bank.
- **IsActiveMember**: A binary integer (1 or 0) indicating whether a customer is active. Normally, active customers are less likely to leave the bank.
- **EstimatedSalary**: A float point number representing the customer's estimated salary.
- **Exited**: A binary integer (1 or 0) indicating whether a customer left the bank or not.
- **Complain**: A binary integer (1 or 0) indicating whether a customer has a complaint or not.
- **Satisfaction Score**: An integer provided by the customer for their complaint resolution. The value ranges from 2 to 5.
- **Car Type**: A string representing the type of card held by the customer. The card type is represented by one of the four categories: DIAMOND, GOLD, SILVER, or PLATINUM.
- **Points Earned**: An integer representing the points earned by the customer for using a credit card.

This dataset contains a large volume of client data and a comprehensive set of customer attributes that are highly relevant in predicting customer churn in a banking context. For instance, economic metrics like CreditScore, Balance, and EstimatedSalary are highly correlated with customer loyalty. Customers with higher credit score, balance, and salary level are less likely to leave due to their financial stability. Furthermore, demographic attributes such as age, tenure, and geography are also good indicators for churn rate predictions. Older customers and long tenure typically indicate loyalty or lower churn rate. As for geography, regional economic conditions and cultural factors can influence customer behavior. The dataset also includes nuance metrics such as Satisfaction Score and Complain that capture the intricacies of customer sentiment. Low satisfaction scores coupled with previous complaints may serve as early warning signs of potential churn. By incorporating these features in predictive modeling, banks can proactively identify at-risk customers and tailor retention strategies.

3.1 Feature Engineering

We performed several transformations on the data to convert all features into numerical format. First, we applied one-hot encoding to the 'Geography' feature since it is categorical. For the other categorical feature, 'Card Type', we converted it to numerical data by assigning values based on its ordinal nature: 'Silver' to 1 and 'Diamonds' to 4, as we believe the different card types are ranked in order. Lastly, we normalized the remaining numerical features to ensure they all fall within the same range.

4 Methodology and Model Performance

4.1 Evaluation Metric

To assess the prediction accuracy, the F1 score was implemented. The F1 score is considered suitable for our accuracy metric because it provides a comprehensive evaluation of a classifier's performance, especially in our scenario involving class imbalance where false positives and false negatives have different implications.

$$F_1 \text{ Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

4.2 Model Performance

Model	F_1 Score
Logistic Regression	0.997
Random Forest	0.996
SVM	0.997

Table 1: Model Performance Summary Table

4.2.1 Logistic Regression

Given the binary classification characteristic of the outcome in the dataset, logistic regression was first performed to model the features. After fitting the model, the accuracy was surprisingly high, with an F1 score of 0.997. According to Figure 1, the logistic regression model correctly predicted 1999 of the 2000 testing samples, underscoring a low misclassification rate.

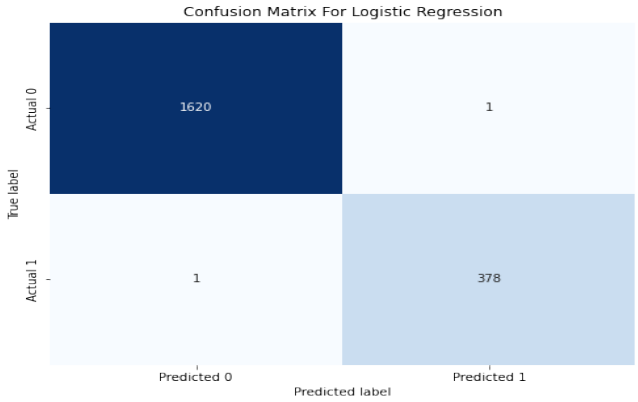

Figure 1: Confusion Matrix For Logistic Regression

By plotting the distribution of coefficients in the logistic regression model, it was observed that the "Complain" feature has an extremely high magnitude as compared with the rest of the features. This indicates that the "Complain" feature significantly alters the model prediction. The correlation between the "Complain" feature and the outcome and its potential detriments is further analyzed in the subsequent section.

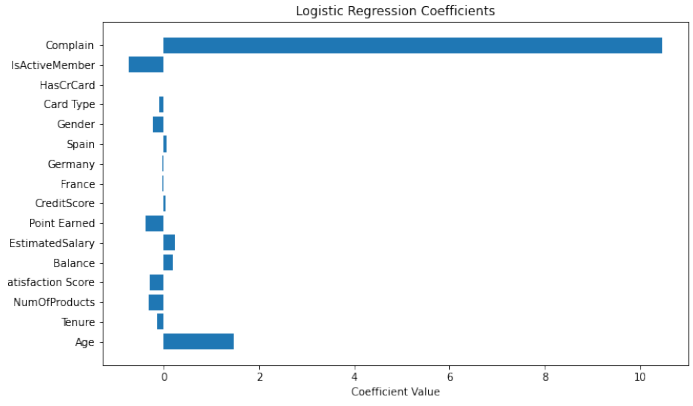

Figure 2: Confusion Matrix For Logistic Regression

4.2.2 Random Forest

After observing the accurate performance of the logistic model, we then employed a tree-ensemble model. In this model, each decision tree makes splits at nodes based

on feature thresholds, where the features are subsampled from the entire dataset. By employing this feature subsampling technique, we aim to reduce the correlation between features, thereby decreasing the overall model variance. To address the Bias-Variance Tradeoff and prevent overfitting, we constrained the depth of the forest to 4 and limited the number of estimators to 10. The model achieved an F1 score of 0.996, demonstrating its efficacy in predicting churn rates once again.

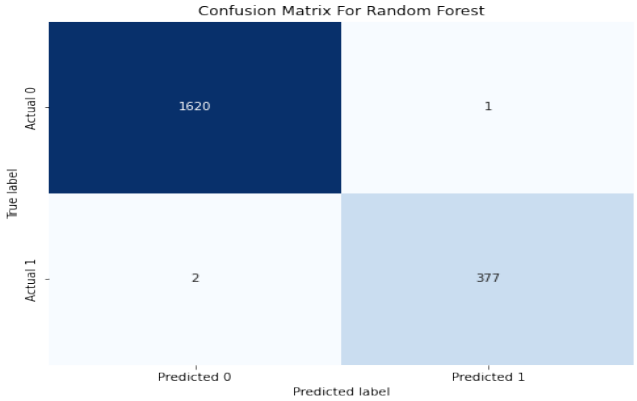

Figure 3: Confusion Matrix For Random Forest

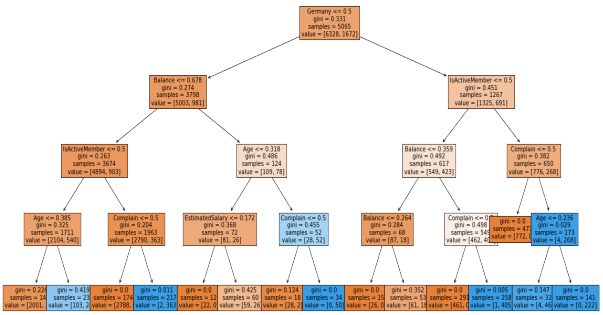

Figure 4: Random Forest Tree Plot

4.2.3 Support Vector Machine

Lastly, we implemented a powerful supervised learning tool known as the Kernelized Support Vector Machine (SVM). This method excels in finding a hyperplane in an N-dimensional space to effectively classify data points, even when the underlying relationships are non-linear. Remarkably, our model achieved outstanding predictive performance, as evidenced by an impressive F1 score of 0.997.

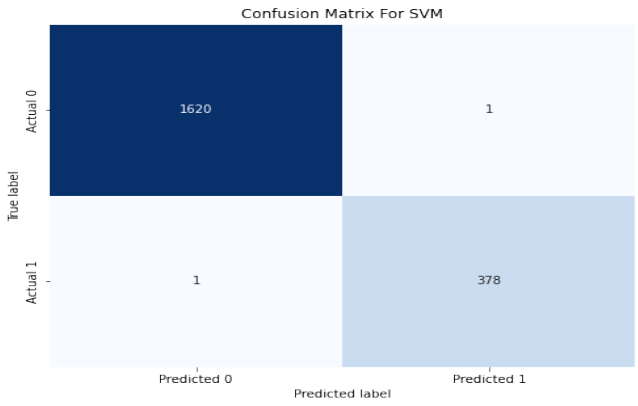

Figure 5: Confusion Matrix For Support Vector Machine

5 Model Results Interpretation

5.1 Logistic Regression

After building and testing our logistic regression model, we aimed to interpret the results and understand which features are most important in predicting whether a customer is likely to leave the bank. By examining the coefficients of the logistic regression, we found that the magnitude of the feature "complain" is significantly higher than the others. As a result, it has a greater impact on the final prediction.

To further interpret the coefficients, we can use the logistic regression formulas to transform them into probabilities, providing a clearer understanding of how each feature influences the likelihood of customer churn.

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta \cdot \text{Complain}+...)}}$$

$$\frac{P(y = 1|x = \text{Complain})}{P(y = 1|x = \text{NoComplain})} = \frac{1 + e^{0}}{1 + e^{-\beta_{Complain}}}$$

Using this formula, we can calculate the probability ratio between customers who have complained in the past and those who haven't, which is approximately 2. This means that the probability of a customer who had a complaint leaving the bank is twice that of a customer who did not have a complaint. Note that when calculating these probabilities, we assume all other features, except for 'Complain', remain unchanged.

5.2 Random Forest

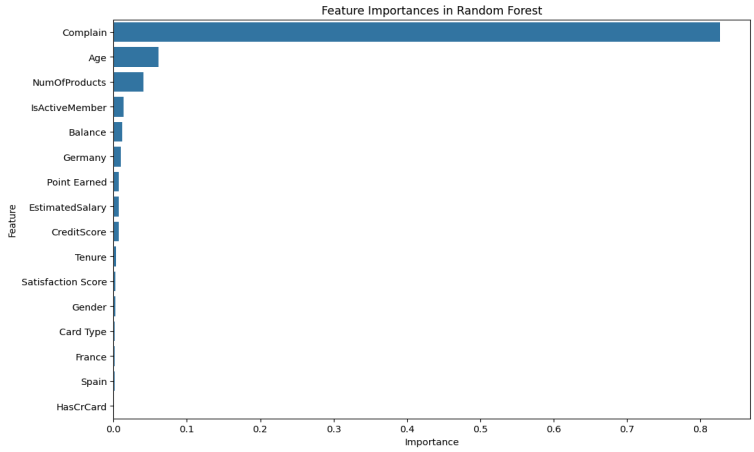

Figure 6: Features Importance of Random Forest Model

We used the built-in function from scikit-learn to show the importance of each feature in our Random Forest model. Similar to logistic regression, the feature 'Complain' dominates the prediction. Additionally, the features 'Age' and 'NumofProducts' are the features ranking just behind 'Complain' in importance.

6 Potential Issue

Our model proved that complaints are a critical issue for the bank because if a customer has a complaint, it is nearly impossible to retain that customer.

However, this feature raised concerns about the effectiveness of our model. The business objective of our project is to predict customer behavior and intervene in advance to avoid customer churn. Since we know that if a customer has a complaint, there is nothing we can do to retain that customer at that point, and our model becomes ineffective. As a result, we want to drop the complaint feature and use other features to make predictions.

7 Model Improvements

In this section we will re-implement the models we used in previous sections on the data without 'Complain' feature, and we will do some optimization to enhance model performance.

7.1 Logistic Regression with Adjusted Thresholds

By naively implementing a logistic regression model on the data without the 'Complain' feature, we observed poor performance, with an F1 score of 0.3. To optimize the model's performance, we first checked the AUC score, which was 0.77. This indicated that there was still room for improvement by adjusting the threshold (the default threshold for prediction is 0.5). After experimenting with different thresholds, we found that the

optimal threshold for this model is 0.26, which resulted in a significantly improved F1 score of 0.51.

7.2 Random Forest With Hyperparameters Tuning

Similarly, we initially implemented a Random Forest model on the data without the 'Complain' feature. The performance was better than logistic regression, with an F1 score of approximately 0.595. Next, we conducted a randomized search to optimize the hyperparameters of our model. This process resulted in a final F1 score of 0.62.

7.3 XGBoost

To furthur improve the F1 score, we implemented XGBoost. The reason we chose XGBoost is because of its robust performance and ability to handle imbalanced classes effectively. In addition, it incorporates various regularization techniques, which help prevent overfitting, especially when considering the small dataset we have. Lastly, the model achieved a F1 score of 0.625.

8 Weapon of Math Destruction Discussion

- **Outcome Not Easily Measurable**: No. Our predictive models aim to forecast whether a bank customer will churn. It is possible to tell whether a customer left the bank or not by consulting the bank record.
- **Predictions with Negative Consequences**: Yes. While the primary goal of churn prediction is to empower banks to implement targeted retention strategies and improve customer satisfaction, there is a risk of misclassification or bias in the predictive models that could lead to negative consequences such as incorrectly identifying loyal customers as at-risk churners.
- **Self-fulfilling (or Defeating) Feedback Loops**: Yes. The deployment of churn prediction models has the potential to create self-fulfilling or self-defeating feedback loops. If banks rely solely on model predictions to determine customer treatment strategies, they may inadvertently reinforce patterns of behavior predicted by the model.

In essence, our model is at risk of producing a WMD. To mitigate such risks, it is vital to carefully consider the ethical and social implications of the model's predictions, validate the model's performance against real-world outcomes, and implement safeguards to prevent the model from perpetuating harmful feedback loops or biases.

9 Conclusion

In our analysis, we initially built logistic regression, Random Forest, and SVM models to predict bank customer churn. The feature 'Complain' was found to be highly influential in all models. However, recognizing that complaints indicate irreversible churn, we decided to exclude this feature and focus on optimizing our models without it. Future directions for this project could involve acquiring larger datasets with more features, enabling more extensive feature engineering to further improve our models.

Member Contribution

- **Chang Chen**: Background, Data Description, WMD Discussion
- **Ryan Ren**: Model Interpretation, Model Improvements, Conclusion
- **Jason Pan**: Methodology and Model Performance

References

[1] Kollipara R. (2023) Bank customer churn. Kaggle, https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn.

[2] Github Link: https://github.com/chenchang00/ORIE_5741_Final_Project