CAI 4104/6108: Machine Learning Engineering
Project Report: Customer Churn Prediction in Banking System

June 15, 2024

1 Introduction

One of the major issues that businesses face is customer churn, which refers to customers abandoning a company or service. For banks, customer retention is crucial for maintaining a stable customer base and ensuring long-term profitability.

This project aims to build a reliable prediction model that can effectively classify customers as churned or not churned based on their historical data and behavioral patterns. The motivation behind this project is to mitigate the significant impact of customer attrition on the bank's profitability and market position. By identifying the factors contributing to customer churn and accurately predicting which customers are at risk of leaving, banks can take proactive measures to enhance customer retention strategies and improve overall customer satisfaction.

In this report, we outlined our methodology, implementation of our prediction model, evaluation of model performance, and discussed our findings.

2 Approach: Dataset(s) & Pipeline(s)

Our approach involved the following steps:

- Data Cleaning: We performed null and duplicate value checks on the dataset to ensure data quality.
- Class Balance: We analysed the class values (1 - churned, and 0 - non-churned customers) to identify potential class imbalance issues.
- Data Analysis: We visualized the data distributions for numerical features and analyzed the correlation between features to gain insights into the data.
- Feature Engineering: Categorical features were one-hot encoded, and numerical features were scaled using the MinMaxScaler to ensure consistent feature scaling.
- Data Split: We split the dataset into train, validation and test sets in the ratio 80, 10, 10 respectively.
- Model Development: We implemented various models for binary classification, including Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, and XGBoost.
- Hyperparameter Tuning: We used Grid Search to tune the hyper parameters of each model, to find the parameters for best performing model.
- Evaluation Metrics: We evaluated each model using the F1 score and area under ROC curve (ROC-AUC score).
- Model Comparison: Finally, we compared all the models based on the evaluation metrics to identify the best model.

2.1 Dataset

The dataset[1] comprised information on 10,000 account holders from ABC Multinational Bank, sourced from Kaggle. It included 12 features per account, encompassing demographic details and account activity, which are instrumental in predicting customer churn.

[1] Dataset: https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset.

2.2 Data Analysis & Prepossessing

2.2.1 Overview

This analysis explores various factors contributing to customer churn at the bank, utilizing heatmaps, bar charts, and box plots for a comprehensive examination of the correlations and distributions of customer attributes.

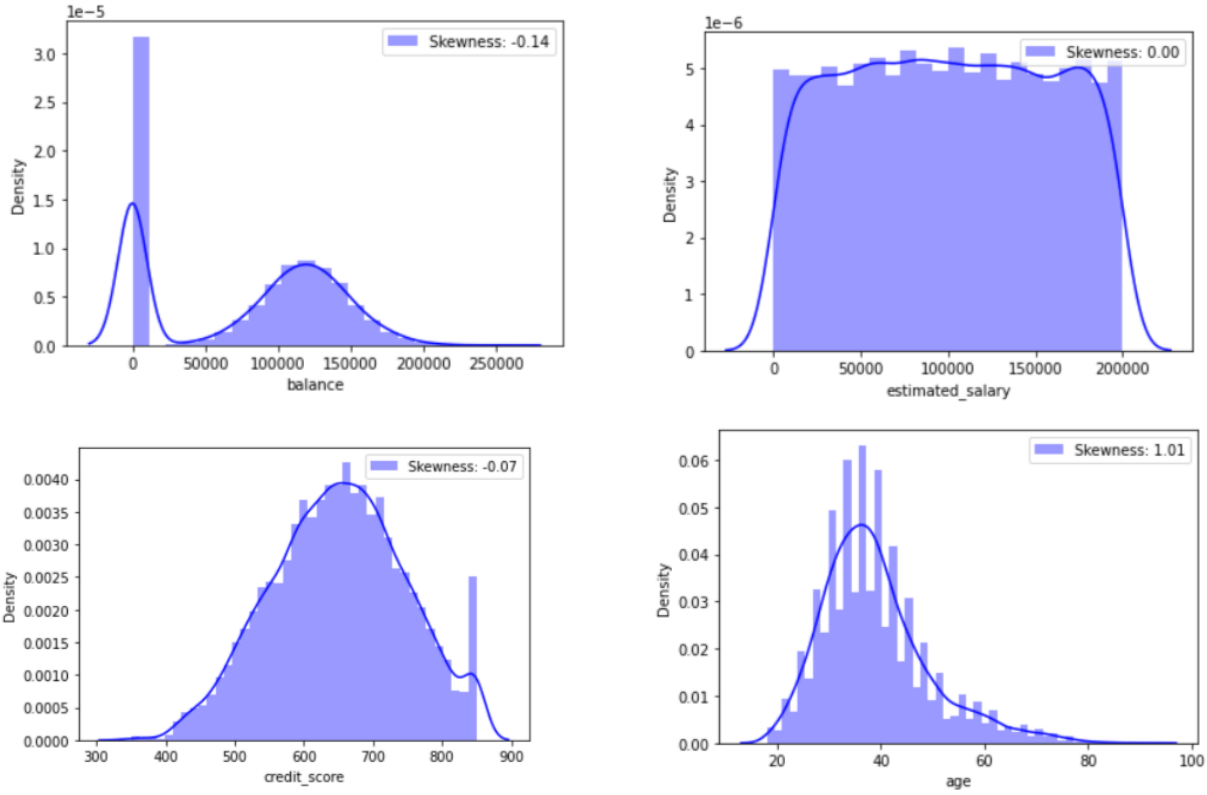

Figure 1: Density plots showing right-skewed customer balance, uniformly distributed salaries, left-skewed credit scores, and right-skewed age.

2.2.2 Detailed Correlation Analysis

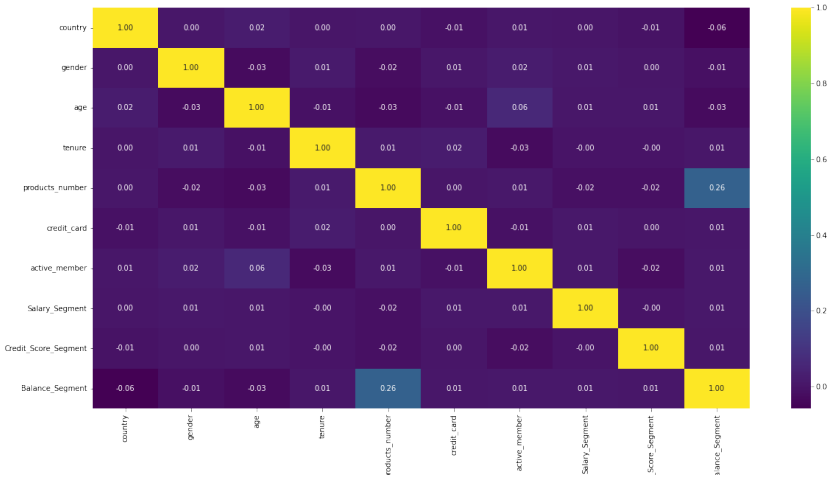

Figure 2: Heatmap illustrating the correlation between customer attributes.

The heatmap (Figure 2) and Bar chart (Figure 3) clearly illustrates key correlations:

- **Age**: There is a consistent positive correlation of 0.29 with churn, suggesting that the likelihood of churn increases with customer age. This is the strongest positive correlation observed in the dataset.
- **Balance**: Demonstrates a negative correlation of 0.3 with the number of products held by customers, indicating that customers with higher balances often have fewer products.
- **Active Membership**: Exhibits a moderately negative correlation of -0.16 with churn, signifying that inactive members are more prone to discontinuing their services.
- **Correlation among attributes**: There is no much correlation among the attributes, so no redundancy is observed among the attributes, which implies no columns need to be dropped.

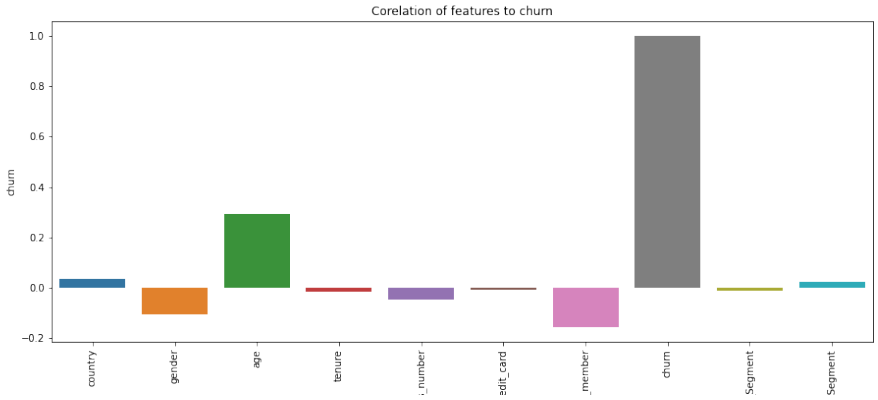

Figure 3: Bar chart illustrating the strength of correlation between various factors and customer churn

2.2.3 Visual Data Insights

- **Age Distribution**: The age distribution among customers who churned shows a higher concentration in older age groups, as depicted in the box plots (Figure 4). This supports the correlation data, highlighting age as a crucial factor in churn risk.
- **Balance Trends**: Box plots (Figure 4) also reveal that customers who churned generally have higher balances compared to those who remained, confirming the negative correlation with product holdings.
- **Gender and Churn**: A slight negative correlation (-0.11) with churn suggests minor difference in loyalty between genders.
- **Tenure and Churn**: Negative correlations observed suggest that customers with longer tenure are less likely to churn, emphasizing the importance of customer loyalty and long-term relationships.

2.2.4 Implications

Based on our analysis, we have identified age and balance as the primary indicators of potential churn, with older customers and those with higher balances at greater risk. Additionally, the active membership status significantly influences churn, suggesting that enhancing customer engagement and satisfaction could effectively reduce churn rates.

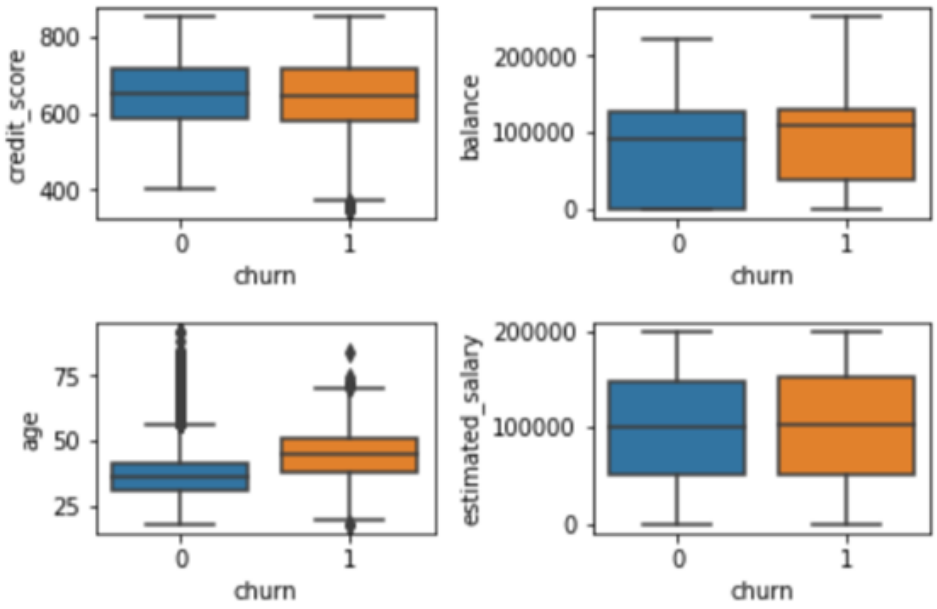

Figure 4: Box plots showing distributions of credit score, age, estimated salary and balance among churned and retained customers. *blue - no churn, orange - churn.*

2.2.5 Data Prepossessing

- **Handling Categorical Data**
  We first categorized the dataset into numerical and categorical data types. Following this, we utilized one-hot encoding to transform the categorical data into numerical format. The categorical features that we had (gender, country) has no ordinal significance and hence we assume that one hot encoding is the better choice here.

- **Handling Data Imbalance**
  Initially, we identified a substantial imbalance in the dataset between 'churn' and 'no churn' data points, with a disproportionate ratio of 80% to 20%. To address this, we strategically selected 1000 samples from each class, merging the data from both classes to establish a balanced sub-dataset. This sub-dataset was then partitioned into validation and test sets using stratified sampling such that the distribution of samples remain same for validation and test. The leftover data, not included in the validation and test sets, was merged to form the training set. However, since training set is still imbalanced. To rectify this, we utilized the Synthetic Minority Oversampling Technique (SMOTE) to perform oversampling on the minority class within the training set. This technique synthesizes additional examples for the minority class, aiding in balancing the class distribution and enhancing the model's performance on imbalanced datasets.

- **Dimensionality Reduction**
  The correlation matrix analysis reveals no redundant features, indicating each feature provides unique information. Additionally, the feature "Customer ID" exhibits no discernible impact on the target variable and is therefore excluded from the dataset to streamline model training.

- **Feature Scaling**
  After splitting the data into train, test and validation set, we performed Standardized Scaling on each sets to address the issue of different scales and distributions of feature data and fit to a normal distribution. We did this Standardized Scaling separately to each set as doing this for a combined data might be leaking some info to test and validation data. To avoid this the standardization was done after the split.

3 Evaluation Methodology

We implemented and evaluated various classification models suitable for the binary classification task of customer churn prediction, which include Logistic Regression, k-Nearest Neighbors (KNN) with grid search for optimal k value, Random Forests as an ensemble of Decision Trees, and XGBoost.

For each model, we performed training on the training set, hyperparameter tuning through grid search on the validation set, evaluation on the test set using F1 score and area under the ROC curve (ROC-AUC score) as metrics. The performances of these models were compared to select the most suitable one for predicting customer churn on our dataset.

To evaluate the performance of our models, we used the following metrics:

- F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance in terms minimizing both false positives and false negatives.
- Area Under the Curve (AUC) Score: This metric measures the area under the ROC curve, which provides an aggregate measure of the model's performance in classifying churned and non-churned customers.

We compared our model's performance against the following baselines:

- Min Baseline: Our minimum baseline was the performance of a simple k-nearest neighbors (KNN) model, which predicts the output based on the closest matching points.
- Max Baseline: According to the paper [1], the benchmark ROC-AUC score for this task is 0.84, achieved using ensemble models like XGBoost and neural networks. This served as our maximum baseline.

4 Results

We have achieved promising results in predicting customer churn for our dataset. The best-performing model was the Random Forest classifier, which achieved an ROC-AUC score of 0.80 and an F1 score of 0.72. These scores surpassed the minimum baseline, but was slightly below the maximum baseline indicating the effectiveness of our model.

The performance of various models on the validation and test sets is tabulated in Table 1. This includes F1 and ROC-AUC scores for each model.

Model	F1 Score		ROC-AUC Score	
	Validation	Test	Validation	Test
KNN	0.47	0.48	0.67	0.69
Logistic Regression	0.62	0.65	0.76	0.76
Decision Trees	0.67	0.67	0.53	0.54
Random Forest	0.72	0.70	0.80	0.79
XG Boost	0.67	0.67	0.63	0.62

Table 1: F1 Score and ROC-AUC Score for Validation and Test Data

4.1 Performance Analysis

Out of the models that we have trained, we found that Random Forest performed the best both in terms of F1 score and ROC-AUC Score. The Logistic Regression model performed well in terms of ROC-AUC score but has a lower F1 score, indicating that model is good at distinguishing between classes across different thresholds, while it struggle to balance precision and recall effectively. KNN did not show notable enhancements compared to Random Forests, highlighting its limited effectiveness in this scenario.

5 Conclusions

This project has demonstrated the potential of machine learning techniques, particularly ensemble learning model like Random Forest in predicting customer churn. By leveraging historical data and behavioral patterns, we were

able to develop a robust model that outperformed the minimum baseline in terms of both ROC-AUC and F1 scores.

Future work could involve employing more advanced models like Neural Networks, exploring more advanced feature engineering techniques like deriving new features from existing features.

References

[1] Louis Geiler, Séverine Affeldt, and Mohamed Nadif. A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, 14(3):217–242, 2022.