Bank Customer Churn Prediction

Sridipta Roy, Syeda Tooba Ali
Northeastern University
roy.sr@northeastern.edu, ali.syeda@northeastern.edu

Abstract

For businesses to succeed, and especially in the banking sector, it is very important to retain existing customers as acquiring new ones is more costly and difficult than retaining existing ones. Therefore, it is important to identify early in time which customers are at risk of churn. This project aims at tackling the problem of customer churn by using machine learning models to predict the customers at churn risk given certain factors. We evaluated a bank customer dataset, which included account related information as well as customer's demographic information. We started off with preprocessing and cleaning the dataset and then identified key features responsible for customer churn. We then trained 15 machine learning models on the dataset to predict customer churn. Our solution was comprehensive, and included preprocessing, exploratory data analysis and application of advanced machine learning models for prediction. The best performing model was found to be the CatBoost algorithm which achieved a high AUC score of 0.8847 on the test set. The age of the customer, whether the customer is a Credit Card owner or not, and the status of the customer (i.e., if they are active or inactive) were found to be the most important features in predicting customer churn.

Dataset: https://www.kaggle.com/competitions/playground-series-s4e1/data

1. Introduction

Strategies to retain customers in the banking sector have become a priority as acquiring new customers is very difficult and costly, and losing the existing customers can have a significant negative impact on the banking industry. Predicting churn early in time not only helps banks to retain customers but also helps them in revising their policies and strategies for the future to provide better and more personalized services, gain customer satisfaction, and build better relationships with their customers.

This project aims at solving the problem of customer churn at banking industry by using machine learning models to predict customer churn early in time; giving the banks a chance to improve their services, finding the reasons that lead to customer churn, and thereby offering solutions that can help retain customers at the risk of churn. The dataset includes customer's account related information like credit score, number of bank products they use, tenure, balance activity status etc, as well as customer demographics like gender, age and geography. Our goal is to use this dataset to identify if customers are at the risk of churn given these features and to identify which features are the most significant in predicting churn.

To solve this problem, we trained 15 machine learning models with different hyperparameters that can predict the customers at risk churn. We also identified the key features in the dataset that are significant for predicting customer churn. We discuss the results in this paper.

2. Background

Bank customer churn prediction is inherently a binary classification problem aimed at identifying customers likely to leave a bank. In this project, we explored a varied range of classification algorithms, which included Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), tree-based algorithms like Decision Tree (DT), Random Forest (RF), boosting algorithms (like XGBoost, AdaBoost, CatBoost) and neural networks like Multilayer Perceptron (MLP). These traditional machine learning methods were chosen for their proven effectiveness in classification tasks and were evaluated using standard performance metrics. They are computationally efficient, ensuring faster training and real-time inference capabilities. Our aim was to find the best model which not only ensures high prediction accuracy on training dataset but generalizes well on unseen test dataset.

In this project we explored exploratory data analysis (EDA) techniques to identify relationships between consumer attributes and churn behaviour. The complete EDA process can be helpful for banks to find the underlying trends and features of their client base by way of sharp examination of the variables producing customer turnover.

We also used the model's feature importance feature to identify key variables contributing to model's predictive behaviour thus making the analysis highly interpretable for banks to focus on strategies that focus on high-risk customers.

3. Related Work

Customer churn prediction has become a pivotal research area in the banking sector, driven by the need to retain clients and reduce operating costs. The adoption of machine learning (ML) techniques has led to diverse modeling strategies and system architectures aimed at improving prediction accuracy. Vu [1] introduced a stacking model, based on the stacked generalization method, to augment the performance of the predictive model. The authors of [2] presented new hybrid model called the logit leaf model (LLM) based on DT and LR to overcome the limitations by combining the strengths of both algorithms. Similarly, Huseyinov and Okocha [3] highlighted the significance of optimizing feature engineering processes. Soni and Nelson [4] proposed a churn prediction framework that integrates profit-aligned metrics to enhance decision-making by identifying not only likely churners but also high-retention customers crucial to profitability.

4. Project Description

4.1 Project Workflow

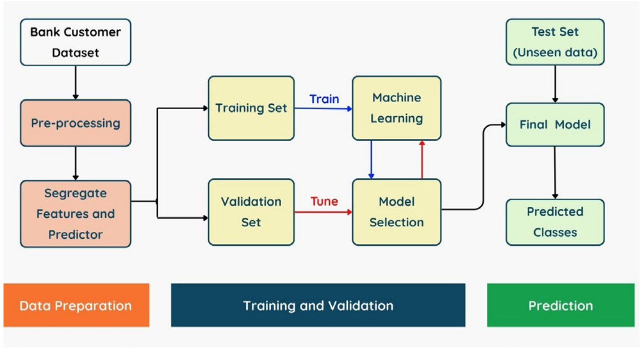
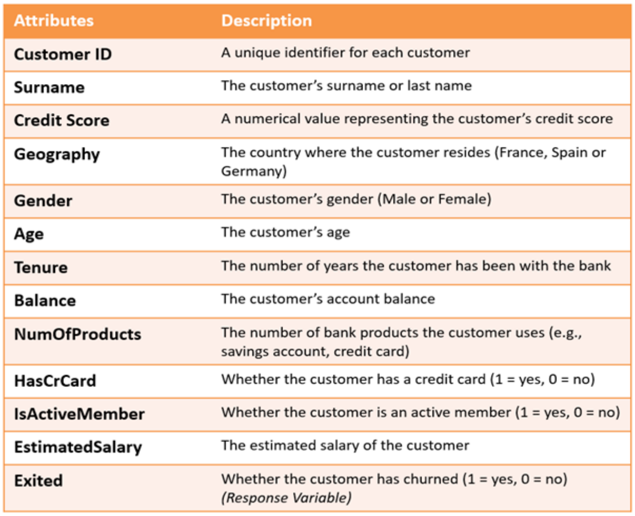


Fig 4.1.1: Overall Project Workflow

The overall prediction workflow of our project is shown in Figure 4.1.1. We started with data ingestion, followed by data exploration and pre-processing. Features and the target variable are then segregated, followed by a split into training and validation sets. The training set is used to train multiple classification models, while the validation set helps in hyperparameter tuning and selecting the best model through thorough evaluation. Once the best model is chosen, it is then used to make predictions on an unseen test set.

4.2 Dataset Description

The dataset used in our project was sourced from Kaggle and it is part of a playbook competition. The dataset contains separate train and test files. The train data set has about 165k rows and 14 columns. We used this dataset for model training and evaluation. Once the final model was selected, we ran the model on test set to analyze its generalization ability. Figure 4.2.1 shows the variables used in our analysis. The column "Exited" is our predictor variable, where 1 represents "Churned" and 0 represents "Retained" customers.



Attributes	Description
Customer ID	A unique identifier for each customer
Surname	The customer's surname or last name
Credit Score	A numerical value representing the customer's credit score
Geography	The country where the customer resides (France, Spain or Germany)
Gender	The customer's gender (Male or Female)
Age	The customer's age
Tenure	The number of years the customer has been with the bank
Balance	The customer's account balance
NumOfProducts	The number of bank products the customer uses (e.g., savings account, credit card)
HasCrCard	Whether the customer has a credit card (1 = yes, 0 = no)
IsActiveMember	Whether the customer is an active member (1 = yes, 0 = no)
EstimatedSalary	The estimated salary of the customer
Exited	Whether the customer has churned (1 = yes, 0 = no) *(Response Variable)*

Fig 4.2.1: Attributes used in our analysis

4.3 Exploratory Data Analysis

4.3.1 Univariate Analysis

We started with univariate analysis of the dataset, focusing on the statistical and visual exploration of individual variables. Numerical features are analyzed using histograms, while categorical variables are examined with bar charts (as shown in Figure 4.3.2). Key insights include age distribution and demographic trends in geography and gender. Figure 4.3.1 shows the distribution of churned and retained customers in our dataset.

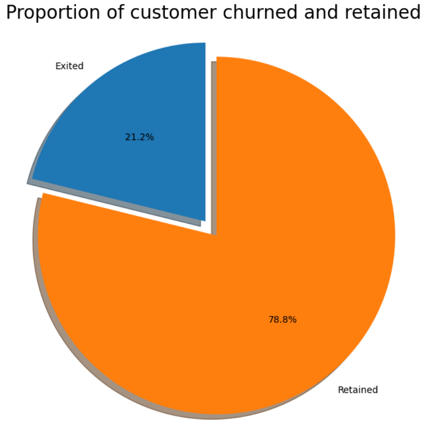


Fig 4.3.1: Distribution of target variable

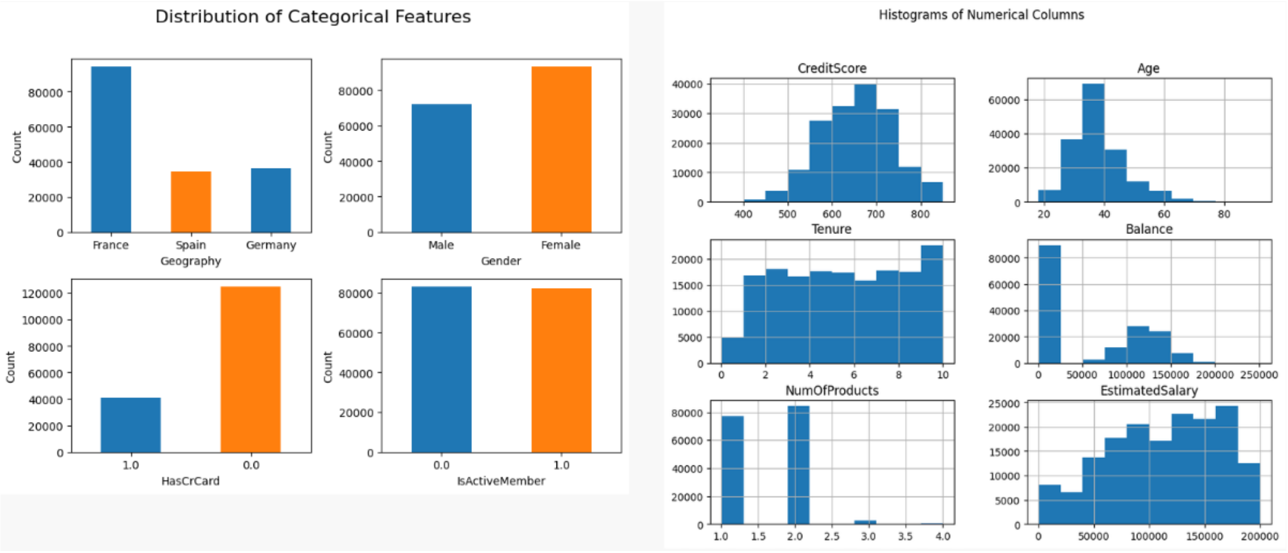


Fig 4.3.2: Univariate Analysis

4.3.2 Bivariate Analysis

For bivariate analysis that examines relationships between customer attrition and other variables, we used visual tools such as boxplots and stacked bar graphs to compare exited and non-exited customers across various features.

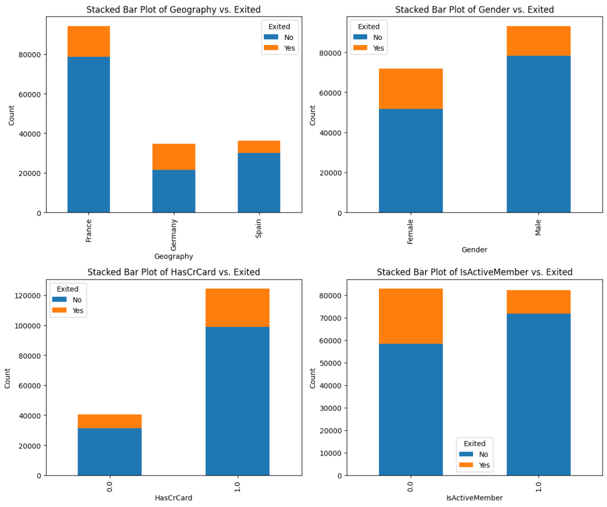


Fig 4.3.3: Bivariate Analysis of categorical features

From Figure 4.3.3 we can infer that churn rate is high among female gender compared to men. Active members are less likely to churn than non-active members. Figure 4.3.4 shows the boxplot analysis. As we can see, Age and Credit Score columns showed some outliers.

For better data understanding, we categorized the continuous predictor variable "Age" into distinct groups as shown in Figure 4.3.5. This revealed that customers aged between 35–45 and 45–60 exhibit the highest churn rates, while churn is relatively lower among the youngest and oldest age groups. Figure 4.3.6 illustrates the grouping of the "Credit Score" variable to analyze its impact on customer churn.

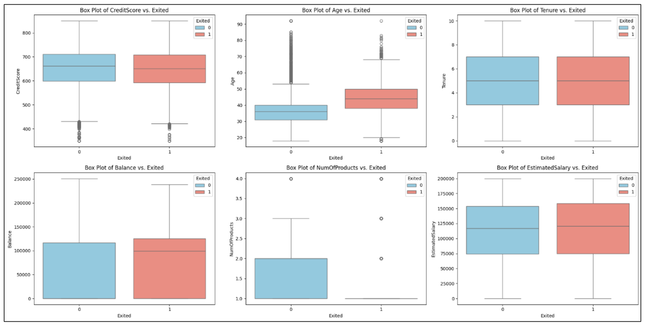


Fig 4.3.4: Bivariate Analysis of numerical features

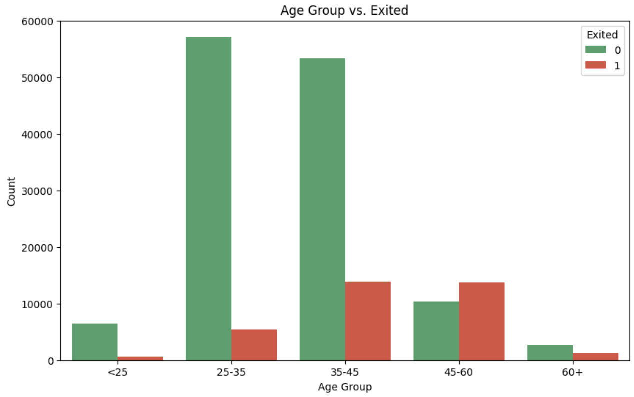


Fig 4.3.5: Churn Rate by Age Group

4.3.3 Correlation Analysis

We developed a correlation matrix for the dataset to assess the interactions of the numerical variables (columns of type float64 and int64). The correlation matrix aides in determining the degree of linear relationship between pairs of features. Figure 4.3.7 shows the resultant heat map where colors suggest the strength and direction of correlation. This study aids in the identification of likely strongly connected characteristics both with one another and with regards to the predictor variable, Exited.

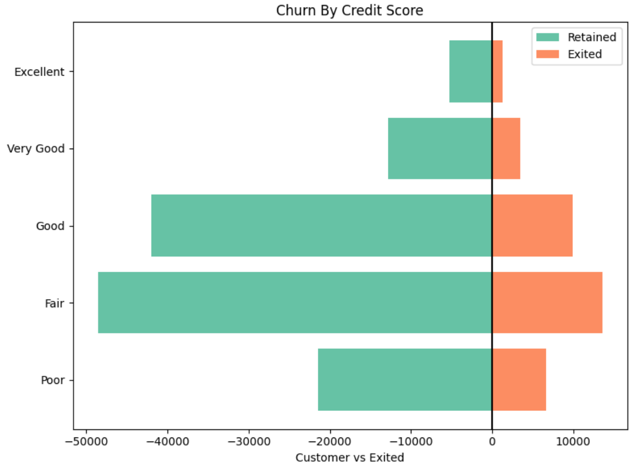


Fig 4.3.6: Churn Rate by Credit Score

The heatmap reveals that “Age” has highest positive correlation coefficient with “Exited” column. “IsActiveMember” and “NumOfProducts” have high negative correlation, indicating that active members and customers using a greater number of products are less likely to churn.

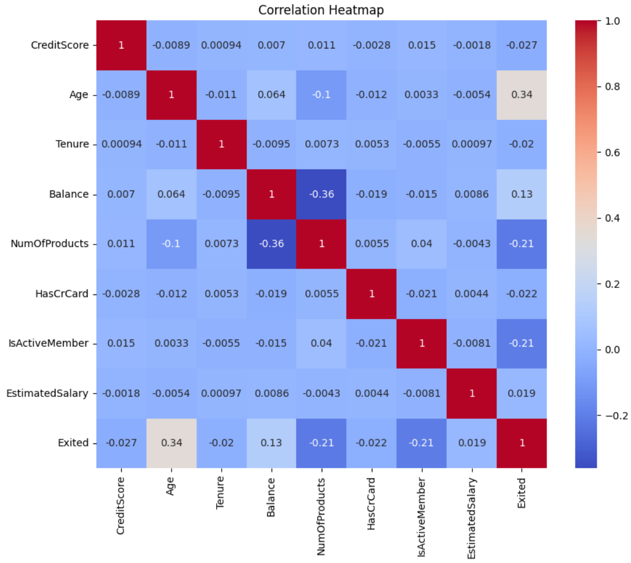


Fig 4.3.7: Correlation Heatmap

4.5 Data Pre-processing

The following steps were performed to pre-process the data:

- Checked for duplicates in the dataset
- Checked for null values among the columns – our dataset didn’t have any null column values.
- Excluded columns with low predictive relevance – such as CustomerId, id (row id), Surname.
- Converted categorical features to numerical format using One-Hot encoding, with the first category dropped (drop='first') to avoid multicollinearity
- Standardized numerical columns using StandardScalar() from “sklearn.preprocessing”, Standardizing or scaling the data ensures that each feature contributes equally to the distance metrics used by many machine learning algorithms.

Some new features created through feature engineering were excluded from final variable selection as they didn’t show better correlation with “Exited” response variable column compared to the original features. We also analyzed some outliers present for “Age” and “Credit Score” column. The outliers were retained, as their impact was minimal due to low standard deviation observed in statistical analysis (used describe() function in pandas is used to generate summary statistics).

4.6 Data Partitioning and Preparation

To prepare the data for model training, we segregated the data into independent (X) and dependent (y) variables. The dataset was then split into training and validation sets using 80:20 ratio, as a separate test set was already available. The “train_test_split” function from the “sklearn.model_selection” module was used for this purpose, with stratified sampling (stratify=y) applied to maintain the original class distribution in both subsets.

As reflected by target distribution, our data is highly imbalanced. SMOTE (Synthetic Minority Over-sampling Technique) was used to generate synthetic samples for the minority class and achieve a more balanced dataset. The class distributions before and after applying SMOTE are shown in Figures 4.6.1 and 4.6.2.

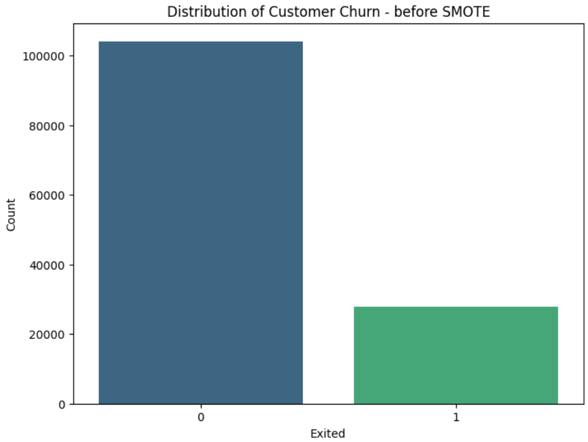


Fig 4.6.1: Initial Target Distribution of training set

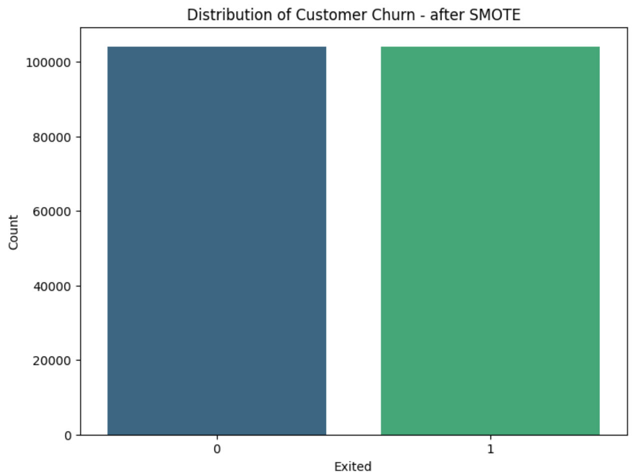


Fig 4.6.2: Target Distribution on training set after SMOTE

5. Empirical Results

Once the preprocessing was completed, we conducted a comprehensive empirical analysis to find out the most effective supervised learning model for predicting customer churn based on our dataset. After preprocessing, we worked on model selection, hyperparameter tuning, and finally evaluation using a metric that is reliable and suitable for imbalance problems like this one.

5.1 Evaluation Metric – AUC:

As our primary evaluation metric, we chose Area under the Curve (AUC) of a Receiver Operating Characteristic (ROC) curve. AUC-ROC helps measure the overall performance of a binary classifier across all possible threshold values. Since our dataset demonstrated class imbalance, AUC was able to evaluate the model's capability of churn and non-churn classes, across all threshold values.

Given our dataset is imbalanced with a lower number of churn cases, accuracy can be misleading in demonstrating the model's prediction ability. AUC on the other hand measures the trade-off between the True Positive Rate and False Positive Rate thereby giving an overall performance of the model.

5.2 Model Comparison and AUC Scores

We experimented with 15 machine learning algorithms including basic and advanced models and evaluated their performance based on AUC scores.

Some of these models were:

1. Logistic Regression
2. Random Forest
3. Multi-Layer Perceptron (MLP) Classifier
4. XGBoost
5. CatBoost

The AUC scores on the validation set were as follows: (Fig 5.2.1 shows the AUC Score Comparison for all 15 models)

1. Logistic Regression ~ 0.799
2. Random Forest ~ 0.8790
3. Multi-Layer Perceptron (MLP) Classifier ~ 0.8792
4. XGBoost ~ 0.8864
5. CatBoost ~ 0.8886

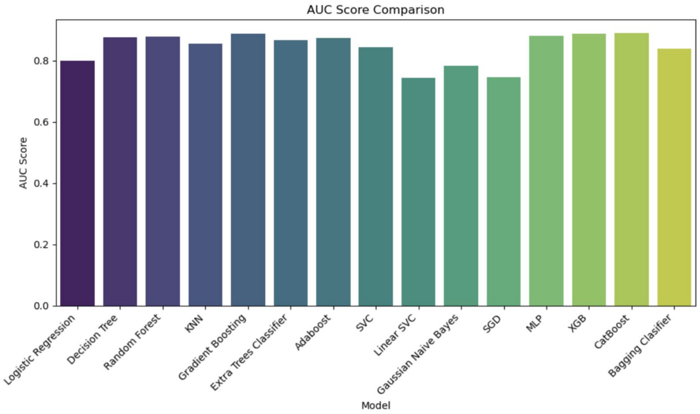


Fig 5.2.1: AUC Score Comparison of all Models

Based on the AUC Score, the best five performing models are shown in Figure 5.2.2 and Figure 5.2.3.

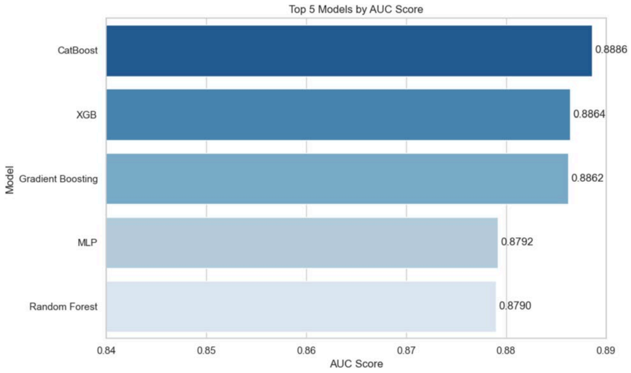


Fig 5.2.2: Top 5 Models by AUC Score

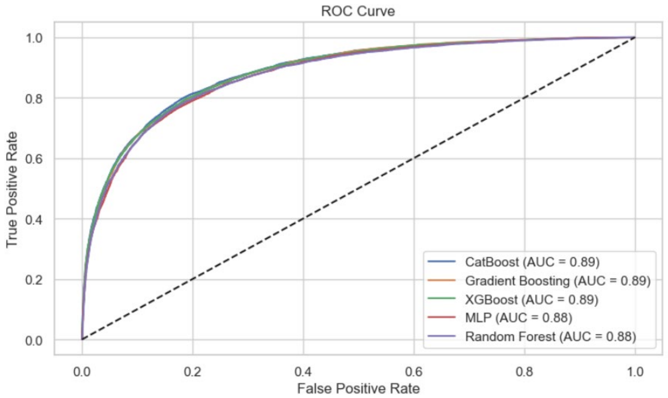


Fig 5.2.3: ROC Curve for Top 5 Models

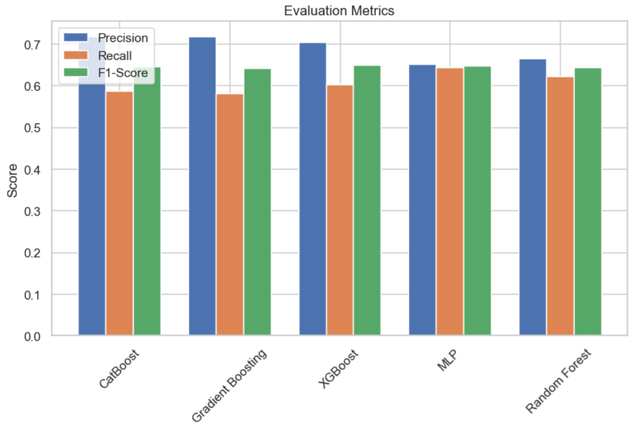


Fig 5.2.4: Evaluation Metrics for Top 5 Models

Among all the machine learning models, **CatBoost** gave the best results with the highest AUC score of 0.8886. Therefore, we selected CatBoost as the final model for evaluation on Kaggle's test set.

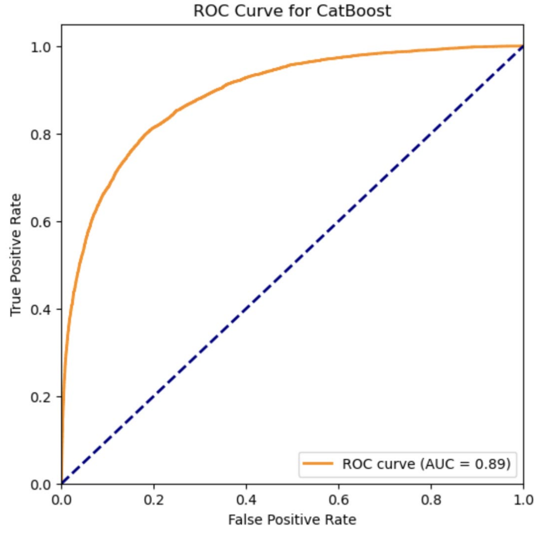


Fig 5.2.5: ROC Curve for CatBoost

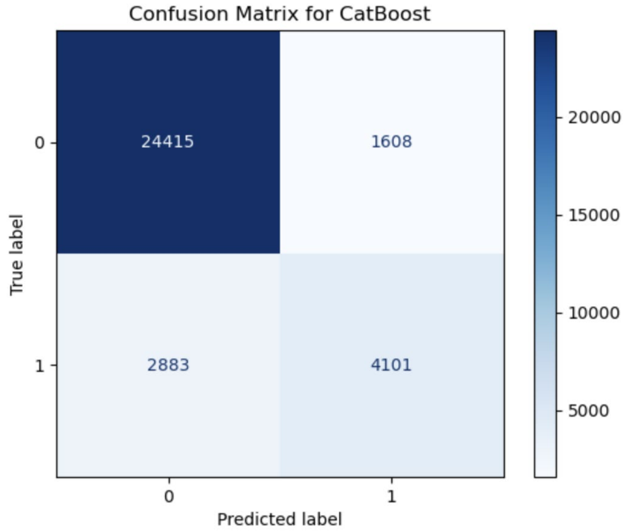


Fig 5.2.6: Confusion Matrix for CatBoost

5.3 Hyperparameter Tuning with Grid Search

The above results are based on the model's evaluation after hyperparameter tuning and on the best selected hyperparameters. In all cases, the models gave better performance after hyperparameter tuning was completed and the best hyperparameters were chosen. We conducted hyperparameter tuning using GridSearchCV, with 5-fold cross-validation using "roc_auc" as scoring metric. Grid Search performs an exhaustive search on the specified parameter values for a model.

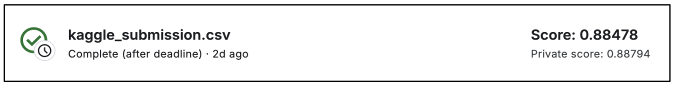
The best parameters for CatBoost were found to be:

Hyperparameter	Value
Depth	6
Number of Iterations	200
Learning Rate	0.1

The depth of the tree controls the model complexity, the number of iterations determines the number of trees in the model, and the learning rate the rate at which a step is taken towards the minimum of the loss function.

5.4 Final Evaluation on Kaggle Test Set

We submitted the predictions on Kaggle's test dataset using our selected tuned Catboost model and achieved an AUC score of 0.88478 which shows that the model generalized well and exhibited good performance on test set.



kaggle_submission.csv
Complete (after deadline) · 2d ago

Score: 0.88478
Private score: 0.88794

Fig 5.4.1: Kaggle Submission Score

5.5 Feature Importance

The CatBoost model also provided insights into the most important features for predicting customer churn.

- IsActiveMember was the top predictor for customer churn. Customers who are inactive are at a higher risk of churn.

- Age was the second most predictor for predicting customer churn which shows that the customer's age is highly correlated with the churn prediction.

- HasCreditCard was the third most important predictor for predicting customer churn. People with a credit card were less likely to churn.

- NoOfProducts, Balance and Geography showed a moderate impact on prediction.

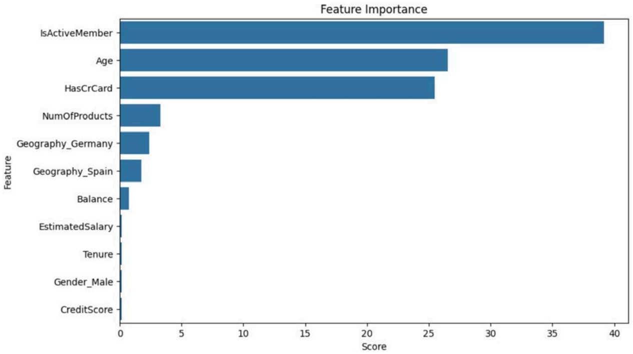


Fig 5.5.1: Feature Importance

6. Future Work

In the future iterations of this project, we plan to explore other more advanced class balancing techniques like Adaptive Synthetic Sampling (ADASYN) and RandomUnderSampling techniques and evaluate our model based on them.

Moreover, we will perform additional more advanced feature engineering to find any complex interactions between the variables and the target variable.

Future iterations will also include experimenting with hybrid approaches for ensemble methods that will combine multiple classifiers and will involve using techniques like stacking and blending with the goal of further improving the predictive performance.

Hydrid algorithm approach can be implemented, as experimented in [2], where the model operates in two stages: a segmentation stage followed by a prediction stage. During segmentation, customers are grouped into distinct segments based on decision rules. In the subsequent prediction stage, a separate predictive model is built for each segment

Conclusion

In this project we have successfully demonstrated the application of machine learning techniques to predict customer churn in the banking sector using a customer dataset. Through comprehensive preprocessing, exploratory data analysis, and model evaluation, the CatBoost classifier emerged as the most effective model, achieving a high AUC score of 0.8847 on the unseen test set.

Key churn factors include the number of bank products used, customer activity status, and age. These insights highlight the importance of interpretable models in helping banks proactively identify at-risk customers. Key takeaways for financial stakeholders include focusing on engagement indicators and considering strategies like product bundling, personalized offers, and re-engagement campaigns, to enhance customer retention.

References

1. Vu, VH. An Efficient Customer Churn Prediction Technique Using Combined Machine Learning in Commercial Banks. Oper. Res. Forum 5, 66 (2024). *https://doi.org/10.1007/s43069-024-00345-5*

2. De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760–772. *https://doi.org/10.1016/j.ejor.2018.02.009*

3. I. Huseyinov and O. Okocha, "A Machine Learning Approach To The Prediction Of Bank Customer Churn Problem," 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2022, pp. 1-5, doi: 10.1109/IISEC56263.2022.9998299.

4. Pranshu Kumar Soni and Leema Nelson. PCP: Profit-Driven Churn Prediction using Machine Learning Techniques in Banking Sector [J]. Int J Performability Eng, 2023, 19(5): 303-311

5. Bhuria, R., & Gupta, S., et al. (2025). Ensemble-based customer churn prediction in banking: A voting classifier approach for improved client retention using demographic and behavioral data. Discover Sustainability, 6. *https://doi.org/10.1007/s43621-025-00807-8*