Report On

Bank Customer Churn Prediction

Prepared
By

Tarun - 211000059 - tarun21100@iiitnr.edu.in
Vipul Guru - 211000061 – vipul21100@iiitnr.edu.in

Submitted

To
Mallikharjuna Rao k
Assistant Professor, IIIT-NR

Course Name: Statistical Data Analysis

Dr. Shyama Prasad Mukherjee

International Institute of Information Technology, Naya Raipur

(A Joint Initiative of Govt. of Chhattisgarh and NTPC)

Email: iiitnr@iiitnr.edu.in Tel: (0771)2474040 Web: www.iiitnr.ac.in

Bank Customer Churn Prediction

Tarun
tarun21100@iiitnr.edu.in
Computer Science
International Institute of Information Technology
Naya Raipur

Vipul Guru
vipul21100@iiitnr.edu.in
Computer Science
International Institute of Information Technology
Naya Raipur

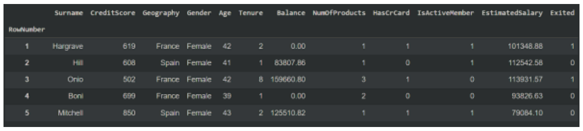
***Abstract*—This report presents a study on the prediction of customer churn in a bank using machine learning techniques. The data used in this study includes various customer attributes, such as Credit Score, Age, Tenure etc. Two machine learning algorithms, K-Nearest Neighbors (KNN) and Logistic Regression, were used to build predictive models for identifying customers at risk of churning. The findings of this study demonstrate the potential of machine learning techniques, particularly KNN and Logistic Regression, in predicting customer churn in the bank. These predictive models can be used by banks to develop targeted retention strategies and reduce customer churn.**

I Introduction

In the world of banking, customer churn is a significant concern for financial institutions. Losing customers can result in decreased revenue, reduced market share, and decreased brand loyalty. Therefore, it is essential for banks to predict customer churn and take proactive measures to retain customers. In this machine learning project, we have utilized the KNN and Logistic Regression algorithms to predict bank customer churn. By analyzing customer data such as age, tenure, and credit scores, we aim to identify the factors that contribute to customer churn and develop an accurate model to predict future customer behavior. Our project has the potential to help financial institutions identify customers who are at risk of leaving, allowing them to take the necessary actions to retain them.

II Dataset

The dataset used in this project contains 10,000 rows and consists of 13 columns, including



	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
RowNumber												
1	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Image 1: Dataset

'CustomerID': The unique identifier for each customer.
'Surname': The customer's surname.
'CreditScore': The customer's credit score.
'Geography': The customer's country of origin.
'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 has.
'HasCrCard': Whether the customer has a credit card or not (1 = Yes, 0 = No).
'IsActiveMember': Whether the customer is an active member or not (1 = Yes, 0 = No).
'EstimatedSalary': The estimated salary of the customer.
'Exited': Whether the customer has left the bank or not (1 = Yes, 0 = No).
To evaluate the performance of our machine learning model, we divided the dataset into a training set and a testing set in an 80:20 ratio.

III Methodology

Data Cleaning : We performs various data cleaning tasks such as removing duplicates and missing values, converting data types, removing outliers and invalid values, and saving the cleaned data to a new file. These tasks help to ensure the data is accurate, consistent, and ready for analysis.

Balancing and Scaling : We used the SMOTE algorithm to balance the data by oversampling the minority class. Then we applied MinMax scaling to ensure that all the features are on a similar scale. Balancing helps to improve precision and F1 score with a compromised accuracy.

Data Visualization : Visualization is an important tool in exploratory data analysis and helps to understand patterns and relationships in the data. Types of visualizations used are scatter plots and heat maps to visualize the data and relationships between variables.

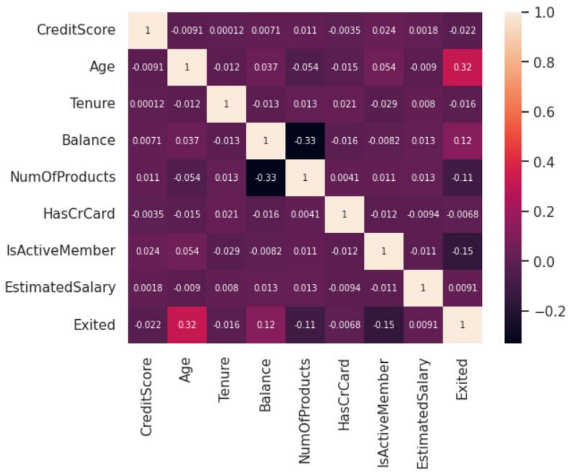


Image 2 : Heatmap

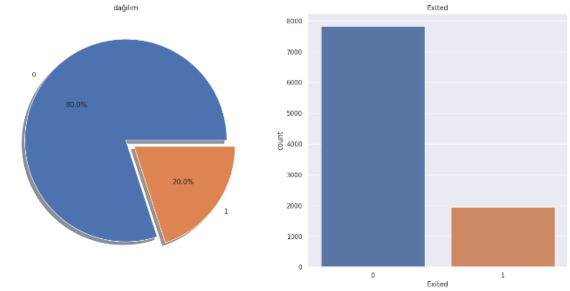


Image 3 : Existed column binary classification

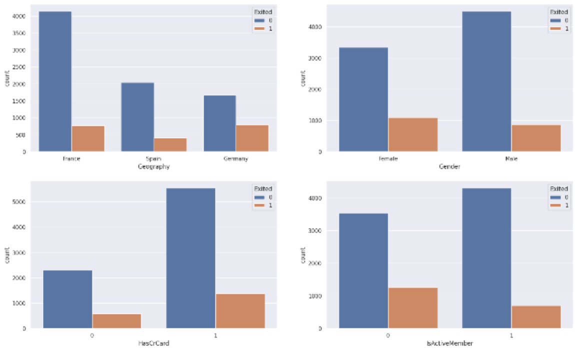


Image 4 : Exited column relation with other columns

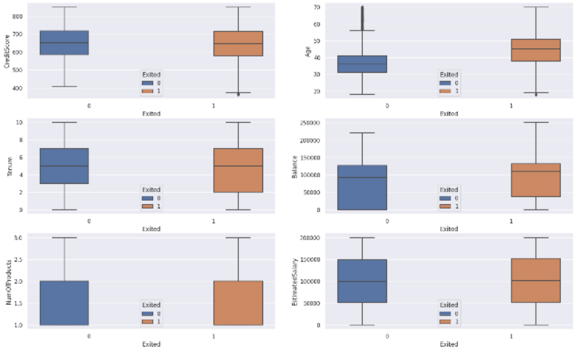


Image 5 : Boxplot

K-Nearest Neighbors Algorithm :

KNN is a type of classification algorithm that makes predictions based on the most similar training examples in the feature space. First a grid search is performed to find the optimal number of neighbors, and the best parameter and score are printed. The KNN model is then trained on the training data using the optimal number of neighbors and tested on the test data. The accuracy and confusion matrix of the KNN model are printed as the performance metrics.

Logistic Regression : Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. Logistic regression model is trained on the x_train and y_train data, and then predictions are made on the x_test data. The performance of the model is evaluated using accuracy, precision, F1 score, and confusion matrix.

IV EVALUATION MATRICS

Evaluation metric refers to a measure that We use it to evaluate different models. We have used the following evaluation metrics:

1) Confusion Matrix: A confusion matrix is a table thatis often used to evaluate the performance of a classification model. It shows the number of correct and incorrect predic- tions made by the model compared to the actual outcomesin the test data, and is a useful tool for evaluating the effectiveness of a model’s predictions.

2) Precision: Precision is a metric that measures the pro- portion of true positive results among the total positive results predicted by a model.

3) Accuracy: Accuracy is a commonly used evaluation metric in classification tasks, which measures the proportion of correctly classified samples to the total number of samples.

4) Recall: Recall is a metric that measures the proportion of true positive results among the total actual positive resultsin the dataset.

5) F1 Score: F1 score is a harmonic mean of precision and recall that gives an overall measure of a model's accuracy in binary classification tasks.

6) ROC Curve: The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. It is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds.

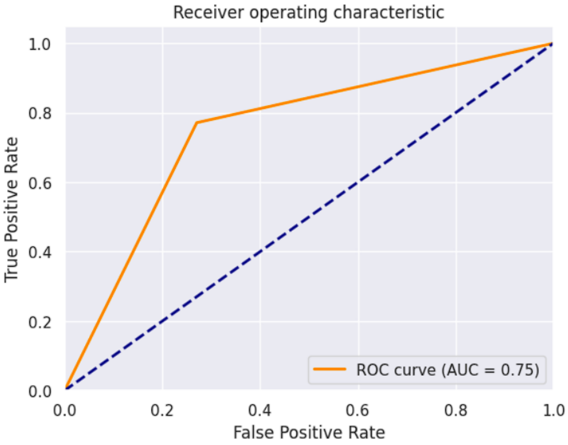


Image 6 : ROC curve for Logistic Regression Model

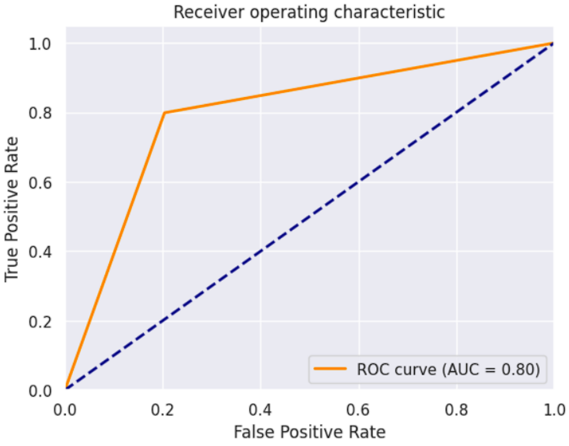
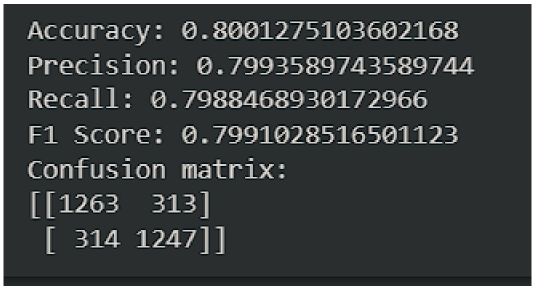


Image 7 : ROC curve for KNN Model

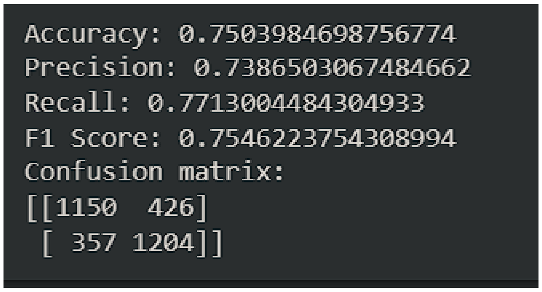
V. RESULTS/OBSERVATIONS

For benchmarking, we have compared the confusion matrix,accuracy, and the F1-Score. All the results are shown in the next page.



```
Accuracy: 0.8001275103602168
Precision: 0.7993589743589744
Recall: 0.7988468930172966
F1 Score: 0.7991028516501123
Confusion matrix:
[[1263  313]
 [ 314 1247]]
```

Image 8 : Results for KNN model



```
Accuracy: 0.7503984698756774
Precision: 0.7386503067484662
Recall: 0.7713004484304933
F1 Score: 0.7546223754308994
Confusion matrix:
[[1150  426]
 [ 357 1204]]
```

Image 9 : Result for Logistic Regression Model

VI. CONCLUSION

In conclusion, predicting customer churn in the banking industry is a crucial task that can help banks retain customers and improve customer satisfaction. In this project, we used various machine learning algorithms such as logistic regression, KNN, and decision trees to predict customer churn. We also utilized techniques such as SMOTE balancing and MinMax scaling to improve the performance of our models. The logistic regression model gave the best results in terms of accuracy, precision, and F1 score. This project can be further improved by collecting more data and experimenting with other techniques such as feature engineering and ensemble methods. Overall, this project demonstrates the potential of machine learning in predicting customer churn and helping businesses make data-driven decisions..

VII. REFERENCES

Dataset

https://www.kaggle.com/datasets/adammaus/predicting-churn-for-bank-customers