BANK CUSTOMER CHURN ANALYSIS

By – Sayan Das

MAY 1, 2023

Contents

Introduction

In today's highly competitive banking industry, customer retention is of paramount importance. Acquiring new customers can be costly and time-consuming, making it crucial for banks to focus on retaining their existing customer base. Customer churn, the phenomenon where customers discontinue their relationship with a bank, poses a significant challenge for financial institutions. Understanding the factors that drive customer churn and developing effective strategies to predict and prevent it have become critical tasks for banks seeking to maintain a strong market position.

The Bank Customer Churn Analysis and Prediction Project aims to address this challenge by utilizing advanced data analytics and machine learning techniques to identify patterns and trends that contribute to customer churn. By analyzing historical customer data, including transactional records, demographics, account activities, and customer interactions, this project seeks to uncover valuable insights that can guide proactive customer retention efforts.

The objectives of this project are two-fold: first, to develop a comprehensive understanding of the factors that influence customer churn in the banking industry, and second, to build an accurate predictive model that can anticipate and flag potential churners. Armed with this knowledge, banks can take proactive measures to engage at-risk customers, tailor personalized retention strategies, and ultimately reduce churn rates.

Dataset

In this project we are going to use the bank customer churn dataset found in Kaggle. The link to the dataset is given below –

https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn

This is a dataset of the account details of 10000 customers of an anonymous multinational bank. The dataset contains 18 columns describing different features of the customers. The columns are described below.

The columns, used in the dataset, are described here, one by one –

1. **RowNumber**—corresponds to the record (row) numbe.
2. **CustomerId**—contains random values.
3. **Surname**—the surname of a customer.
4. **CreditScore**—refers to the credit score of the customers.
5. **Geography**—a customer's location.
6. **Gender**—refers to the gender of the customers.
7. **Age**—Age of the customer.
8. **Tenure**—refers to the number of years that the customer has been a client of the bank.
9. **Balance**—account balance of the customer.
10. **NumOfProducts**—refers to the number of products that a customer has purchased through the bank.
11. **HasCrCard**—denotes whether or not a customer has a credit card..
12. **IsActiveMember**—indicates whether the customer in question is an active one or not.
13. **EstimatedSalary**—as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.

14. **Exited**—whether or not the customer left the bank.
15. **Complain**—customer has complaint or not.
16. **Satisfaction Score**—Score provided by the customer for their complaint resolution.
17. **Card Type**—type of card hold by the customer.
18. **Points Earned**—the points earned by the customer for using credit card.

Among these variables the following variables are of categorical type- **Surname, Geography, Gender, HasCrCard, IsActiveMember, Exited, Complain, Card Type.**

The rest of the variables are of numerical nature.

Objective

Objective of this project is given by following -

1. To analyse historical customer data from a bank and identify key factors that contribute to customer churn.
2. To gain a comprehensive understanding of the patterns and trends associated with customer churn in the banking industry.
3. To develop a predictive model using advanced data analytics and machine learning techniques to accurately predict customer churn.
4. To assess the performance of the predictive model and validate its accuracy in identifying potential churners.
5. To provide insights into customer behaviour and preferences that can guide the development of targeted retention strategies.
6. To enable banks to proactively identify at-risk customers and implement personalized interventions to prevent churn.

Here the variables **Exited** is the study variable on the other hand the other variables are predictors. We want to study how the response or the study variable changes when the predictors are changing.

Visualisations

We have made a lot of visualisation to understand the intrinsic nature of the dataset. We have shown then one by one below-

Absence of any missing values

The dataset has no missing values. So do not have to bother about imputing the missing values in the dataset.

Imbalance in the dataset

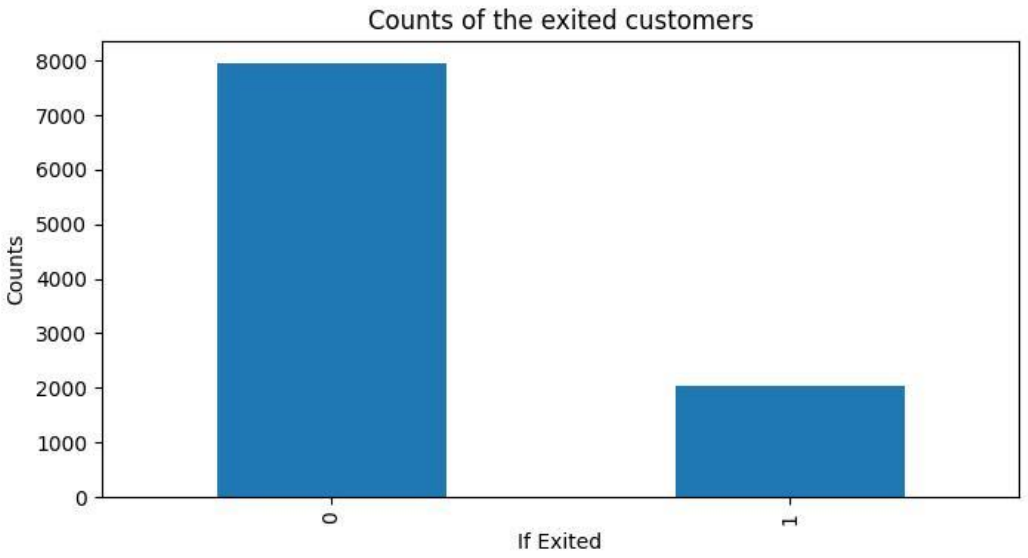
We have found the frequency distribution of the response variable **Exited.** The frequency distribution is shown in the following table-

Table 1: Frequency Distribution of the response

Exited	Frequency	Proportion
0	7962	0.7962
1	2038	0.2038

Around 20% of the customers have exited the bank and remaining 80% choose to stay with the bank. Clearly there is an imbalanced into the dataset. Based on the frequency count we have plotted a bar diagram-

Figure 1: Frequency count of the response: "Exited"

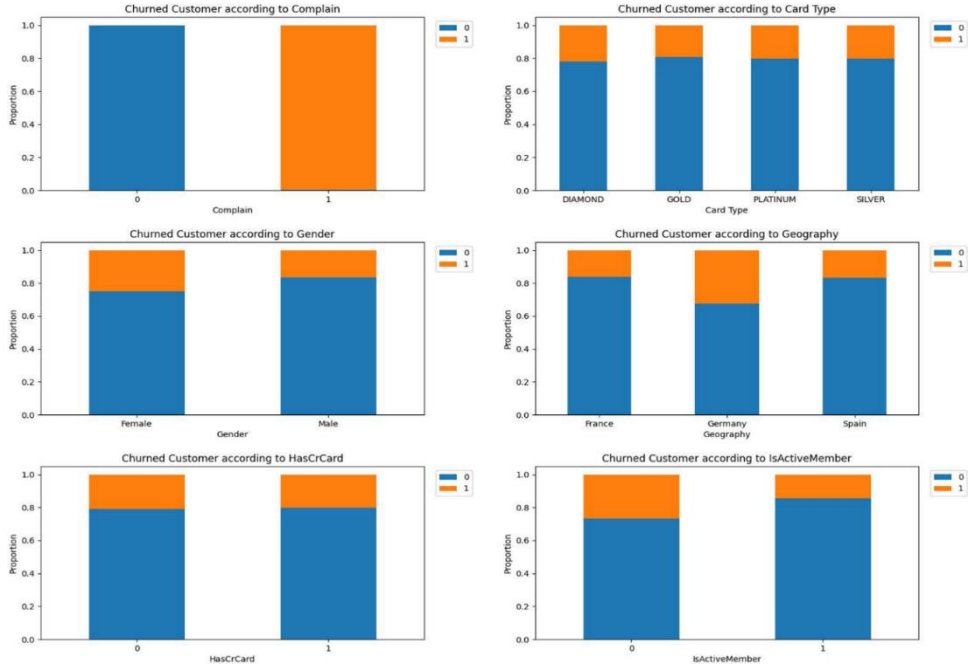


From the visualisation its clear that the observation of customers leaving the bank is very low with respect to that of customers not leaving the bank.

Plots of response according to Categorical Variables

The bar-plot of the **Exited** variable according to all the categorical variables are given below.

Figure 2: Bar-plot of the response according to categorical variables



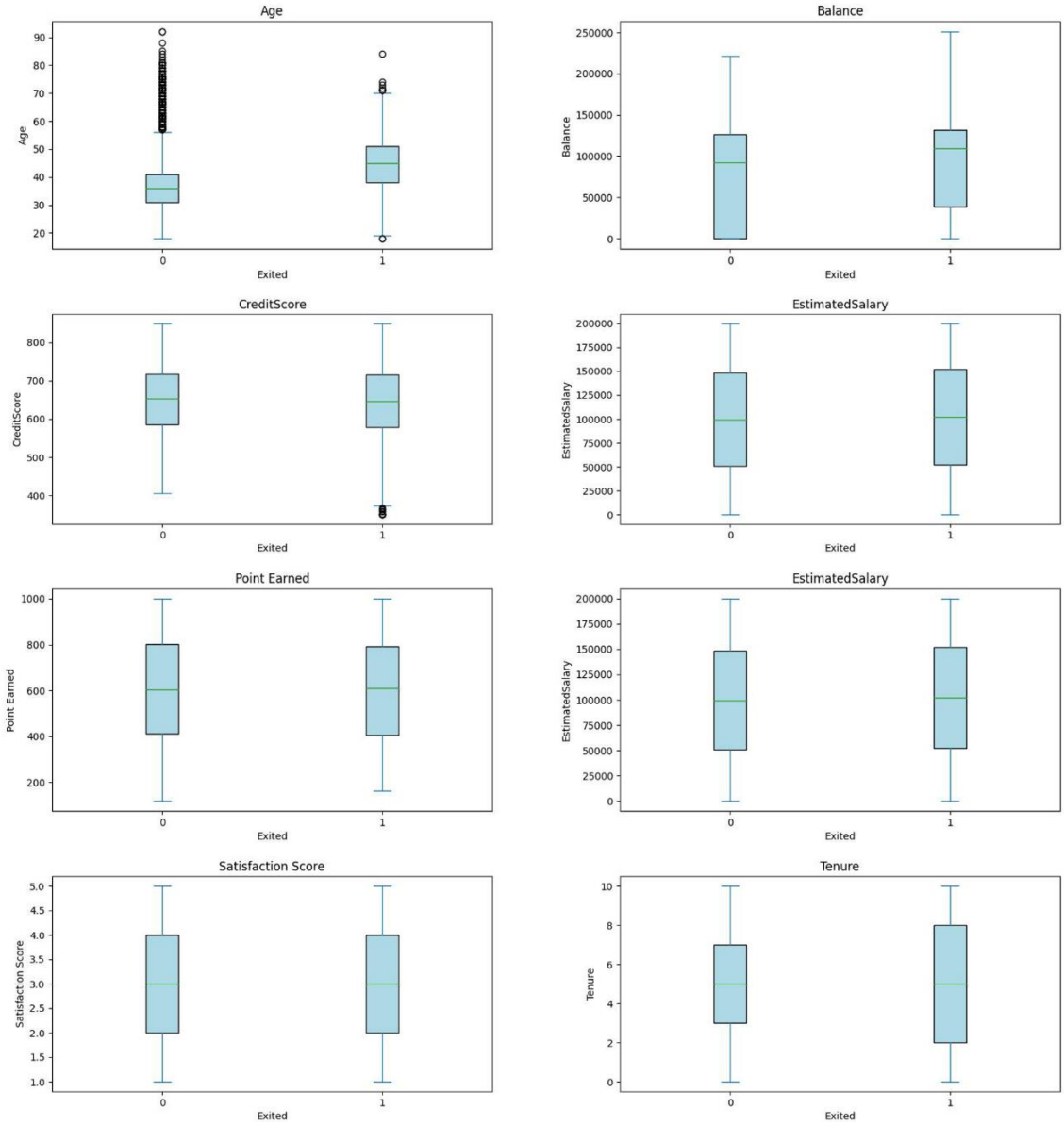
From the plot it seem that the variables **IsActiveMember**, **Complain, Gender** and **Geography** is affecting whether a customer will leave the bank or not.

1. Almost all the customers who have made a complain to the bank have exited the bank as well.
2. Females are more likely to leave the bank than Male.
3. Customers who are not an active member of the bank is more likely to leave the bank.
4. Customers from the German Branches have shown maximum occurrence of leaving the bank.

Plots of Numerical Variables according to response

The box-plots of the numerical variables divided according to the levels of the response is shown below-

Figure 3: Numerical variables according to response



According to boxplot of the numerical variables it seems-

1. Most of the customers who left the bank are of higher ages.
2. Customers with the lesser account balance are more likely to stay with the bank

Presence of outliers

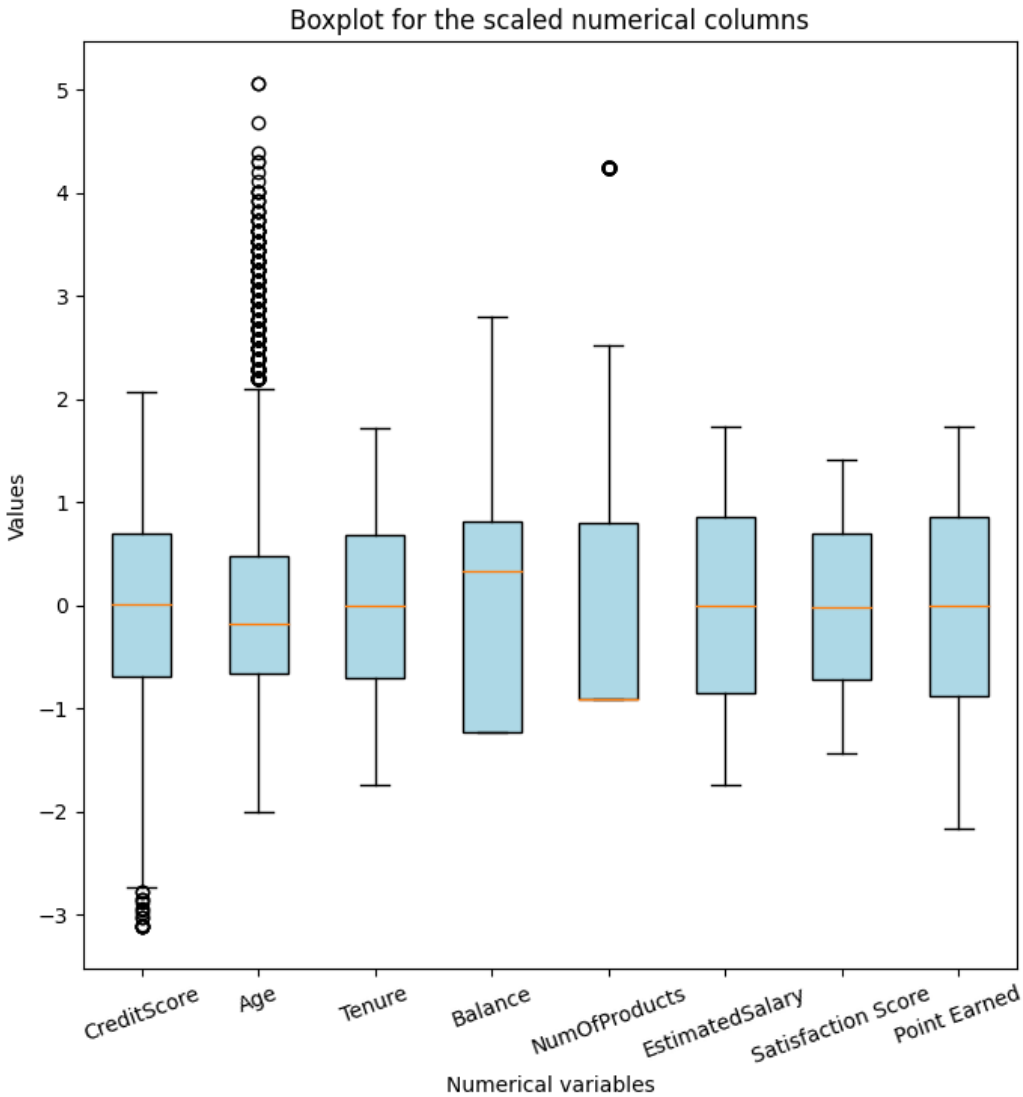
To detect the presence of the outliers we have-

1. Standardised the numerical columns using the z scoring method. Here to standardise the variable x we use the following formula-

$$x_{std} = \frac{x - \bar{x}}{sd(x)}$$

2. In the next step we had plotted the standardise variables in the boxplots. The boxplot is shown below-

Figure 4: Boxplot of the numerical columns



From the given plot its is clear that the columns **Age, CreditScore** and **NumOfProducts** have outliers.

Data Pre-processing

From the previous data exploration, we had realised that the dataset is not perfect. The data contains, outliers, all the numerical variables are not of the same scale. The categorical variables are not good for applying any statistical treatment on it/ Therefore this dataset needs to be processed before proceeding further.

Scaling the numerical columns

The first step of pre-processing is scaling the numerical columns. We converted the numerical columns using the formula-

$$x_{std} = \frac{x - \bar{x}}{sd(x)}$$

After scaling the numerical columns, the variability seems to boil down into a same level.

Detecting and deleting the outliers

From Figure 4 we have seen that the columns **CreditScore**, **Age** and **NumOfProduct** have outliers. To remove the outliers, we have computed two bounds for each of the numerical columns. For the variable x the upper bounds and the lower bounds are given by-

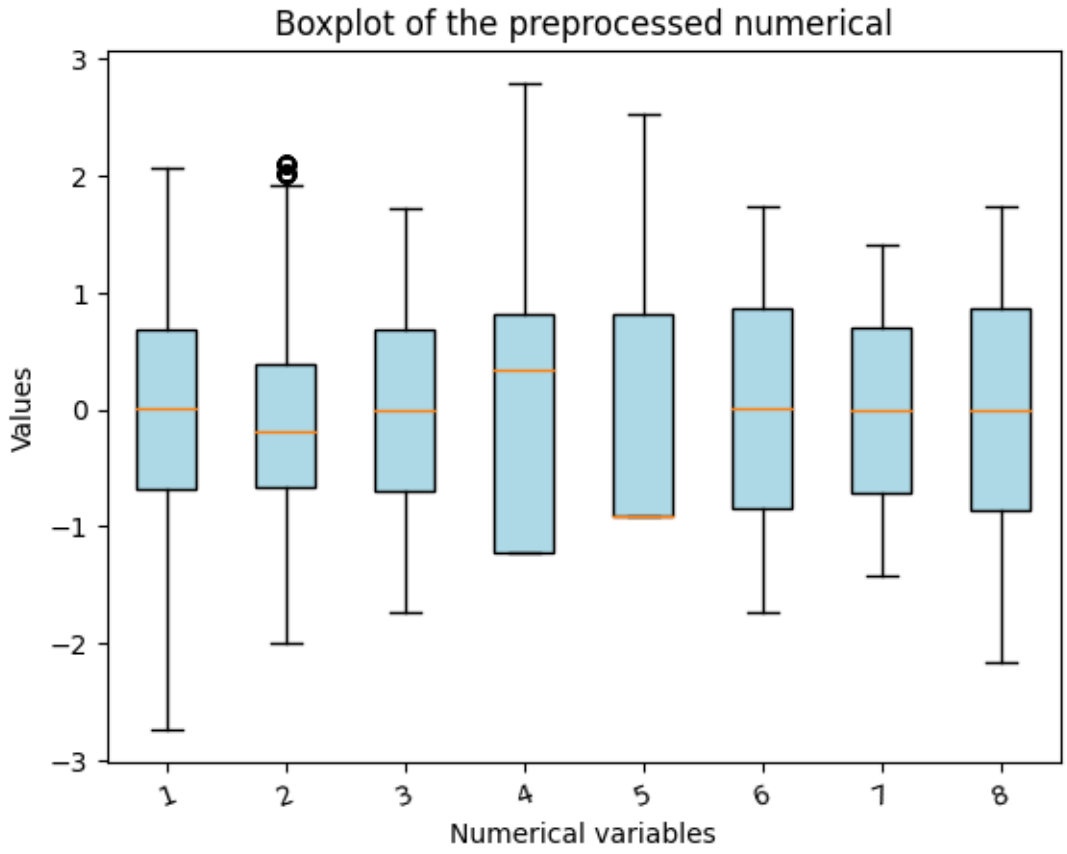
$$(UpperBound, LowerBound) = (Q_1(x) - 1.5.IQR(x), Q_3(x) + 1.5.IQR(x))$$

Where $Q_1(x)$ and $Q_3(x)$ are the 1st and 3rd quartile of the variable x. and IQR(x) is the Inter Quartile Range of x.

Values of x lesser than LowerBound or greater than UpperBound is detected as the outliers and rows of the dataset satisfying this, are deleted.

After deleting the outliers, the boxplots of the numerical columns is given by the following-

Figure 5: Boxplot of the numerical columns after removing the outliers



Encoding the categorical columns

In the next step, we have encoded the categorical columns, using the one-hot encoding technique. Here we represent the categorical columns using sequence of 0 and 1. Suppose we have a variable X that takes value "A", "B" or "C". Then after applying the one hot encoding technique the column X will be represented by 2 columns X_A and X_B. Whenever X takes value "A" we map (X_A, X_B) = (1, 0), similarly if X takes value B then (X_A, X_B) = (0, 1) and finally if X is "C" then (X_A, X_B) = (0, 0).

Handling imbalance in the dataset

To handle imbalance in the dataset we will use the over sampling technique SMOTE (Synthetic Minority Over-sampling Technique) to resample observations of customers who has actually exited the bank. By the end of applying this technique we will have a dataset that has equal number of observations for customers who has exited the bank and the customers who does not have exited the bank.

Diagnostic Analysis

We will now try to understand how the predictors contribute regarding the fact whether the customer will leave the bank or not. To find this out we have applied the logistic regression on the dataset, taking the variable **Exited** as response and the other variables as predictors.

Logistic Model

Consider Y as response variable.

$$Y = \begin{cases} 1\ if\ \ exited \\ 0\ if\ not\ exited \end{cases}$$

and $x_1, x_2, \ldots x_p$ are the predictors. Then under Logistic regression we model the log-odd of the Y=1 as linear function. Like below-

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + ß_1 x_1 + \cdots + ß_p x_p$$

or alternatively we can write

$$\pi = \frac{e^{\alpha + ß_1 x_1 + \cdots + ß_p x_p}}{1 + e^{\alpha + ß_1 x_1 + \cdots + ß_p x_p}}$$

We used the Maximum Likelihood Estimation (MLE) technique to estimate the parameters α and $ß_i$'s. The MLE estimates of those parameters have no closed form. So, we have to depend upon the iterative techniques to solve out the value of the estimates.

Estimated Coefficients

After fitting the model, we got the following value of the coefficients that are shown in the following table. Also, we have shown the p-value associated to the estimates to know whether the estimate value is significant or not-

Table 2: Coefficients of the fitted logistic regression model

features	coefficients	p-value
Constant	-6.4636	0.000
Credit Score	0.1188	0.682
Age	1.5458	0.000
Tenure	-0.1941	0.486
Balance	0.1918	0.550
Number of Product	-0.4355	0.051
Estimated Salary	0.0179	0.947
Satisfaction Score	-0.3589	0.195
Point Earned	-0.6975	0.021
Geography Germany	-1.1005	0.138
Geography Spain	0.4207	0.571
Male	-0.4977	0.379
Has Credit Card	0.2615	0.683
Is Active Member	-2.2929	0.001
Complain	15.8196	0.000
Card Type Gold	-1.1229	0.184
Card Type Platinum	-1.3901	0.089
Card Type Silver	0.1127	0.889

Interpretation of the Coefficients

Constant: when all the other values are set to 0 then value of the log-odds in favour of a customer leaving the bank is -6.4636, which is significant.

Credit Score: keeping the other values of the Predictors fixed the log-odds in favour of the customers leaving the bank increases 0.1188 unit for unit increment in the Credit Score. But this change is not significant.

Age: For 1 unit increase in the age, keeping the other predictors fixed, the log-odds of customer leaving the bank significantly increases 1.5458 unit.

Tenure: Keeping the other predictors constant 1 unit increment in the tenure-ship results in 0.1941 unit decrease in the log-odds of customer leaving the bank. But this change is not significant.

Balance: Keeping the other predictors constant 1 unit increase in the balance will result in increment of the log-odds of customer leaving the bank by 0.1918 unit. However, the change is not significant.

Number of Products: 1 unit increment in the number of purchase of the product (keeping other predictors constant) will result in decrease in the log-odds of customer leaving the bank by 0.4355 unit. And we consider this change as significant as the p-value is so close to 0.05.

Estimated Salary: 1 unit increment in the number of estimated salary (keeping other predictors constant) will result in decrease in the log-odds of customer leaving the bank by 0.0179 unit. But the effect of the change is not significant.

Satisfaction Score: Keeping the other predictors constant 1 unit increment in the satisfaction score will result in an insignificant decrease in the log-odds of customer leaving the bank by 0.3589 unit.

Points Earned: Keeping the other predictors constant 1 unit increment in the points earned will result in a significant decrease in the log-odds of customer leaving the bank by 0.6975 unit.

Geography: On an average, keeping the other predictors constant the log-odds of customer leaving the bank is 1.1005 unit less for customers in Germany than the customers in France and 0.4207 unit

more for the customers in Spain than the customers in France. But none of the difference is significant.

Gender: On an average, keeping the other predictors unaltered, the log-odds of customer leaving the bank is 0.4977 units less for the male customers than the female customers. But this difference is not significant.

Having Credit Card: On an average, keeping other thing constant, the log-odds of customer leaving the bank is 0.2615 unit more for the people having credit card than those who does not have credit card. However, this difference is not significant.

If an active member: On an average, keeping the other predictors constant, the log-odds of the customers leaving the bank is 2.2929 unit for the active members than the non-active members. The difference is significant.

Complain: On an average, keeping the other predictors constant, the log-odds of customer leaving the bank is 15.8196 unit significantly more for the people having complained than the people who have not complained.

Card Type: On an average, keeping the other predictors constant, the log-odds of customer leaving the bank is 1.1229 unit less for the gold credit card type that diamond credit card type, 1.3901 less for the platinum card type than for the diamond card type and 0.1127 unit more for the silver card type than the diamond card type.

From the value of the estimated coefficients we can find out that the features **Age**, **Number of Products Purchased**, **Points Earned**, **whether the customers is an active** member and **whether the customer have made any complain** are important for determining whether the customer will stay with the bank or leave the bank because these features had a significant effect on the log-odd of customers leaving the bank.

Prognostic Analysis

Here our primary target is to find a model that would efficiently predict whether the customers going to leave the bank given the customers information. We measure the level of efficiency of the model from the testing accuracy of the model. For this reason we first split the dataset into training and testing dataset. We will train the model based on the training dataset and then calculate the accuracy of the model based on it prediction for the testing dataset.

We first applied the Logistic regression in the data set and it is giving quite a good result in the testing dataset. So, we are going to use this model for the prediction.

Choosing the best threshold and prediction of the Customers

From the form of the model of the logistic regression, we can understand that the model returns us the probability(π) or the log-odds($\log(\frac{\pi}{1-\pi})$) of the customer leaving the bank. We are going to decide whether the customer is going to leaving the bank or not based on π. We have to decide a threshold (t) such that if P(π > t) then we predict customer will leave the bank ($\hat{Y} = 1$) otherwise we will decide ($\hat{Y} = 0$).

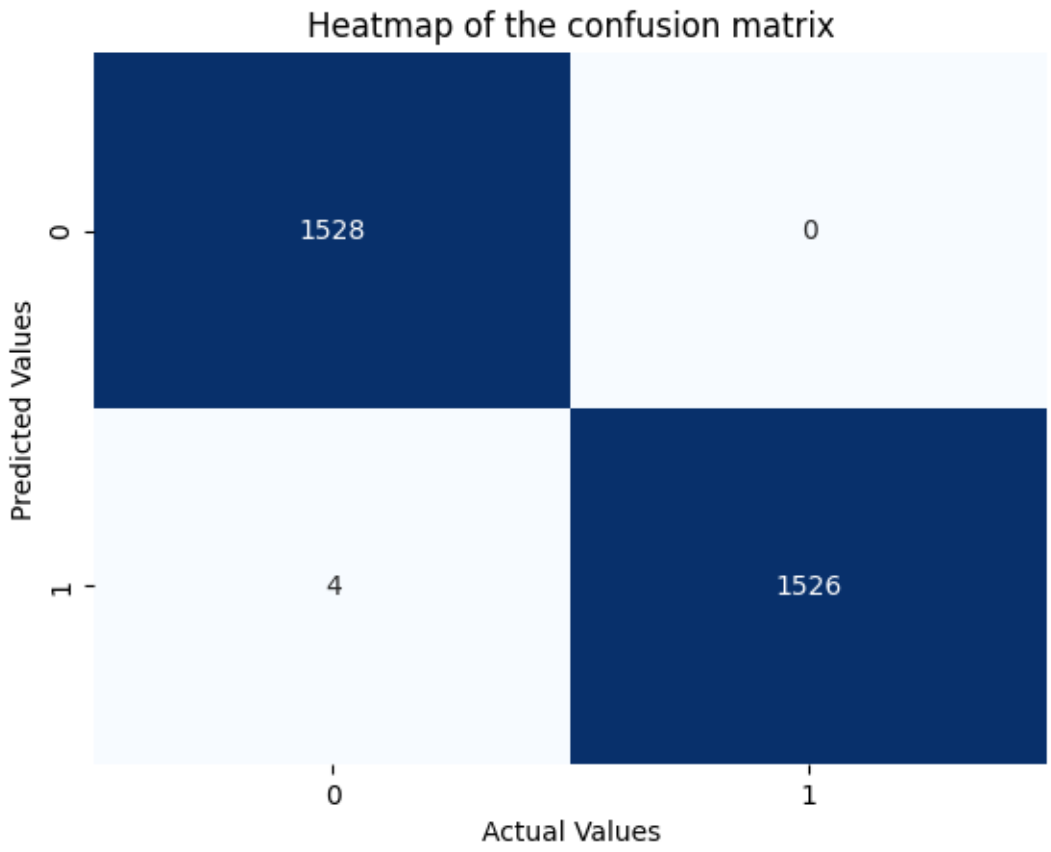
To choose the best threshold t we will one by one consider all the predicted value of π as threshold compute the predicted value of Y ($\hat{Y}$) and then the True Positive Rate(TPR) given by P($\hat{Y} = 1$|Y=1) and False Positive Rate(FPR) given by P($\hat{Y} = 1$|Y=0). Then we choose that π as t for which TPR*(1-FPR) is maximum. Here according to our data, the best threshold is coming out to be t = 0.8588.

Based on this best threshold, we compute the $\hat{Y}$.

Testing accuracy

We first calculated the confusion matrix of the predictions of the testing dataset. The confusion matrix is shown below in the form of the heatmap diagram.

Figure 6: The confusion matrix



Here we can see that the proportion of the wrong classification is very small as compared to the total observations.

We have used a 10-fold cross-validation technique to get an estimate of the testing accuracy (the proportion of mis-classification in the testing dataset). In this method the average testing accuracy is coming out to be 0.9983 i.e. around 99.83% of the customers are correctly classified by the logistic regression model.

This much testing accuracy is quite good for the model. It is clear that using this logistic regression model for the prediction purpose is quite logical.

Conclusion

In conclusion, this project aimed to analyze and predict customer churn in a bank using logistic regression. We worked with a customer dataset, performing various data processing and analysis techniques to gain insights into the factors influencing customer attrition.

We began by pre-processing the dataset, performing outlier removal, scaling and encoding to ensure the data's quality and integrity. We also addressed the issue of class imbalance, which is common in churn prediction tasks, by applying appropriate techniques such as oversampling.

Using a diagnostic approach, we conducted exploratory data analysis to understand the relationships between the features and the churn outcome.

Based on the insights gained from the data analysis, we built a logistic regression model. This model utilized the customer's details, such as demographic information, transaction history, and engagement metrics, to predict whether a customer is likely to leave the bank or not. The model provided valuable coefficients, which indicated the significance and direction of the impact of each feature on churn prediction.

The project's findings and the predictive model can provide significant value to the bank. By accurately identifying customers who are likely to leave, the bank can optimize their customer retention strategies, reduce churn rates, and ultimately improve customer satisfaction and profitability.

Overall, this project demonstrates the importance of data analysis and predictive modeling in understanding and addressing customer churn in the banking industry. The insights gained from this project can guide strategic decision-making and help businesses optimize their customer retention efforts.

We have also built a simple web app based on python flask. The app codes amd all the other code are given in the following link.

References

1. Dataset Link: https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn
2. Project Link: https://github.com/sayandas1302/Bank-Customer-Churn-Prediction