Customer Churn Analysis

Neco Darian

1 ABSTRACT

This report presents an extensive analysis of a bank's customer data, which we aim to predict the likelihood of customer complaints and to understand several factors influencing customer churn. This research employs exploratory data analysis, correlation analysis, Chi-Square tests, and Analysis of Variance (ANOVA) to identify significant predictors of customer complaints. The identified key variables include age, balance, tenure, and card type. A Random Forest Classifier model is developed and tested, displaying high performance in predicting customer complaints, which is validated by strong accuracy, precision, recall and F1-score metrics. Despite the model's good performance, the report emphasizes the need for continuous monitoring to ensure its ongoing effectiveness, given context and data variations. This study provides valuable insights for the bank to proactively address customer complaints, mitigate churn, and enhance customer retention.

2 INTRODUCTION

The purpose of this project is to analyze customer churn in a banking context. Customer churn, which occurs when customers leave a company, is an important metric for businesses to track because it can be more costly to acquire new customers than to retain existing ones. In the banking industry, customer churn can result in lower growth rates and have a significant impact on sales and profits. Therefore, it is essential for banks to understand and predict customer churn to take targeted actions to improve customer satisfaction and retention.

To achieve this goal, machine learning algorithms have emerged as a powerful tool for predicting customer churn. By analyzing large amounts of customer data, these algorithms can identify patterns and behaviors that may indicate that a customer is at risk of leaving. Several research papers such as [1, 4, 5] have explored the application of machine learning algorithms in predicting customer churn in the banking industry and found out that XGBoost, RandomForest Classifier and Logistic regression have shown good accuracy of prediction.

The focus of this project is to conduct a comprehensive analysis of a bank customer churn dataset to gain insights into customer behavior and help the bank management make informed business decisions. The analysis will be aimed at answering several key questions related to customer information such as the proportion of customers still using the banking services, characteristics of customers more likely to complain, and the relationship between complaints and customer exit. Additionally, machine learning techniques will be used to develop a model that can predict whether a customer is likely to complain based on their historical records. By identifying customers who are at risk of leaving and improving customer satisfaction and retention, the bank can increase profitability and ensure long-term success.

3 DATA DESCRIPTION

In this project, we have access to a dataset collected by a bank, which includes various explanatory variables as detailed in Table 1. The original data is thoughtfully divided into two parts: the 'Main Sample' and the 'New Sample' to facilitate comprehensive analysis. With this dataset at our disposal, we are well-equipped to conduct a thorough exploration of the data and extract meaningful insights.

Variable	DataType	Description
CustomerId	int64	Contains random values
CreditScore	int64	Represents the credit score of the customer
Location	object	The country where the customer is located
Gender	object	Male or female
Age	int64	Represents the age of the customer
Tenure	int64	Number of years as member
Balance	float64	Balance in the customer account
NumOfProducts	int64	Number of products purchased
HasCreditCard	int64	customer has a credit card 1 or not 0
IsActiveMember	int64	Whether a customer is active (1) or not (0)
EstimatedSalary	float64	The salary of the customer
Exited	int64	(1) or not (0)customer left the bank
Complain	int64	customer has complaint (1) or not (0)
Satisfaction Score	int64	Score for complaint resolution (1-5)
Card Type	object	Type of card held by the customer
Point Earned	int64	Points earned by customer using credit card

Table 1: CUSTOMER FEATURES IN DATASET

4 DATA ANALYSIS

4.1 Task A.1

To answer the first question regarding the proportion of customers still using the banking services compared to those who have left, we start by obtaining an overview of the data. Table 2 provides the count of customers who have exited and those who have not. With these numbers available, we can compute the proportions of customers in each category.
Figure 1 shows the visualization of the distribution. The bar chart displays the count of customers who have exited and those who have not, with each bar annotated with the corresponding proportions. The pie chart presents another view of these proportions.

There are customers that have Exited and are still active. A reason for that can be that the bank provides several services to these customers. For sake of simplicity, we will not focus on that part. Based on Josh Howarth's research [3], the average churn rate for financial institutions in 2023 is 25%. This allows us to conclude that our result of a 20% churn rate falls below the sector's average.

Exited	Count
No	7949
Yes	2031
Total	**9980**

Table 2: Customer Count

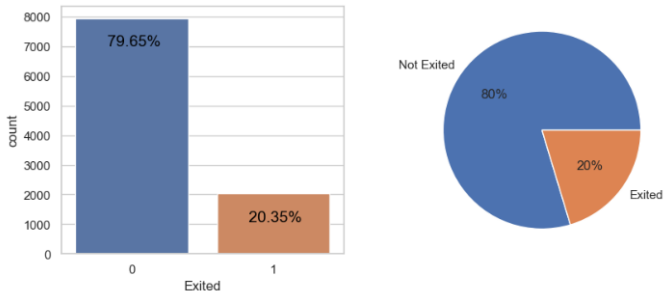

Figure 1: Visualizing the distribution

To figure out whether the difference in proportions is significant, we can conduct a hypothesis test. In this case the null hypothesis would be that there is no significant difference in the proportion of customers who have left and those customers that are still using the services of the bank. Alternative hypothesis would then be that there is a significant difference between the two proportions.

A chi-squared test would be suitable to test this hypothesis by comparing the observed frequencies of the two variables (customers who have left and those who have not) to the expected frequencies, assuming no association between the variables. We will reject the null hypothesis if the p-value is less than the significance level, and we will fail to reject the null hypothesis if p-value is greater than the significance level.

Since we already know the proportion of customers who have left and those who have not, we can conclude that there is a significant difference in the proportions without conducting a hypothesis test. A proportion of 20.35% of customers leaving the bank may be a cause of concern for the bank authority, as it may lead to a loss of revenue and market share. Therefore, it may be important for the bank to understand why customers are leaving and take appropriate actions to retain customers.

4.2 Task A.2

In question two, we want to find the relationship between the number of complaints received by the bank authorities and the number of exited customers.

We start by examining the mean values of relevant features for exited and non-exited customers, to figure out any noticeable differences in the numbers of complaints between these two groups. To explore the relationship further, we calculate the ratio of exited customers 'Table 3' for each complaint category to determine the likelihood of customer churn based on the number of complaints. Looking at the proportion of exited customers within each category will give us valuable information on the potential influence of complaints on customer retention.

To assess the correlation between the number of complaints and customer churn, we will use the correlation coefficient, table 4. It will provide a quantifiable measure of the strength and direction of the relationship between the variables. As shown in figure 2, the Complain distribution between Exited categories is almost identical to the Exited distribution in figure 1. From the findings we can conclude that almost 100% of customers who have complaint actually Exited and vice versa, those that not Exited have no Complain- which indicate a very strong correlation.

Complain	Exited	ratio
0	0	0.999496
0	1	0.000504
1	1	0.995091
1	0	0.004909

Table 3: Complain-Exited category ratio.

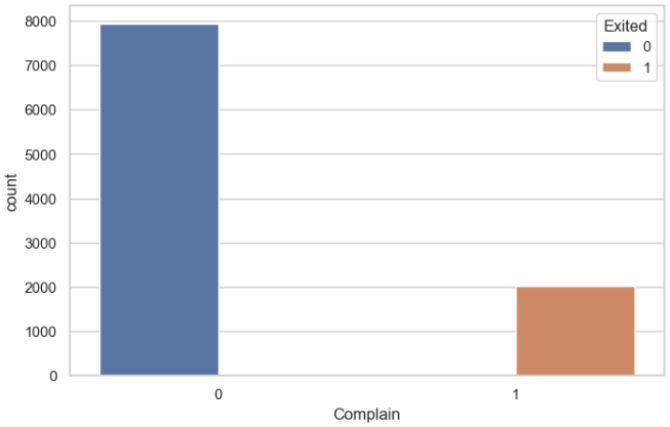

Figure 2: Distribution of Complaints among exited and non-exited

	Exited	Complain
Exited	1.000000	0.995679
Complain	0.995679	1.000000

Table 4: Correlation between Exited and Complain variables.

4.3 Task A.3

Next, we are asked to find the characteristics and statistics (in terms of gender, age groups, and tenure etc) of the customers that are more likely to complain.

We will first look at numerical features, by calculating the mean values for each group. From these values we see that there are some differences for those who have complained and those who have not.
To go more in detail and gain further insights, we will generate box plots and histograms for each numerical feature. These visualizations provide a clear understanding of the distributions and comparison between customers who complain and those who have not complained. Especially among different age groups 'figure 3', we find a significant difference in the number of complaints. Based on descriptive statistics in table 5, we find out that customers of age between 38 and 51 are more likely to complain than customers of age from 31 to 41.

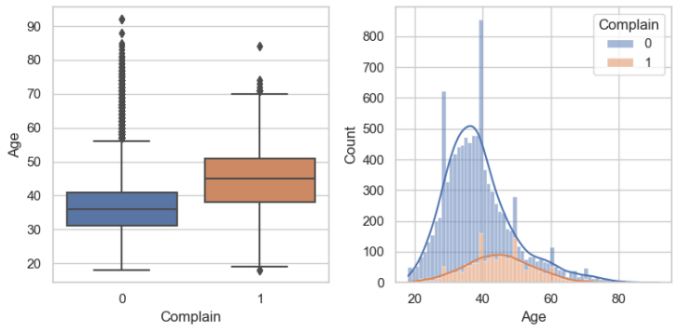

Figure 3: Distribution of Complaints among exited and non-exited

Statistic	Complain	No-Complain
Count	7943.000000	2037.000000
Mean	37.410550	44.778596
Standard Deviation	10.131782	9.775906
Minimum	18.000000	18.000000
25% Percentile	31.000000	38.000000
Median	36.000000	45.000000
75% Percentile	41.000000	51.000000
Maximum	92.000000	84.000000

Table 5: Descriptive statistics for the Age variable.

When it comes to categorical features, to explore the impact on customer complaints, we calculate the ratio of complaints for each category within these features. In this way we will determine the likelihood of complaining for customers belonging to each category. Based on the overall analysis, we can identify the following characteristics of customers more likely to complain:

- Customers located in Germany are more likely to complain, compared to customers in France and Spain 'figure 4'.
- Female customers had a higher likelihood of complaining compared to male customers 'table 6'.
- Customers with a tenure of 1 or 2 years were more likely to complain.
- Customers with more than two products (specially 3 or 4) were more likely to complain.
- It is more likely that those customers that are not active members will complain.
- There is no substantial difference in complaint likelihood based on the type of card or satisfaction score.

Figure 4: Distribution of Complaints among exited and non-exited

Gender	Complain	Ratio
Female	0	0.749007
Female	1	0.250993
Male	0	0.834925
Male	1	0.165075

Table 6: Complain ration based on gender.

4.4 Task A.4

Is there a significant difference between the credit scores of all the customers that have complained and those who have not in the period covered in the dataset?

To answer this question, we will first take a look at the chart using a boxplot 'figure 5'. From the first glance we can not say if there is a significant difference in Credit score of those who have complained and those who have not. We investigate further using statistical methods. First we need to figure out the distribution. If the distribution is normal then we will perform a t-test to compare two groups of observations.
To check the data for normality we will check if the distribution of the data is bell-curved on the histogram plot 'figure 6' or along the probability line on the Q-Q plot. In addition to that we will also perform the Shapiro-Wilk test for normality. From the Shapiro-Wilk test we find out that the CreditScore values are not normally distributed.

In this case we can try to perform a non-parametric test used to compare the distribution of two independent groups. For that, we will use the Mann-Whitney U test.

- By Null Hypothesis we state that there are no significant differences in the credit scores between two observed groups.
- If the p-value is less than or equal to 0.05, we will reject our Null Hypothesis.

From the results we can conclude that there is a significant difference between the credit scores of customers who left and those who are still active (Mann-Whitney U test).

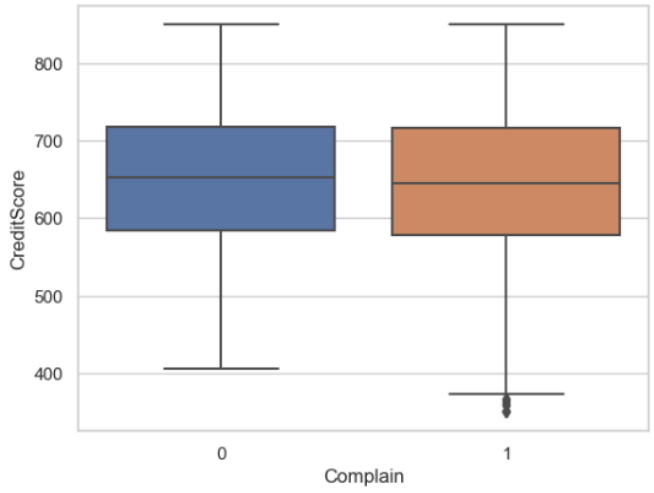

Figure 5: Distribution of Credit score.

4.5 Task A.5

Do the satisfaction scores on complaint resolution provide indication of the customers' likelihood of exiting the bank?
The satisfaction Score has 5 levels. Our target variable is Exited, which has two values - 1 and 0. Since we have categorical data, we

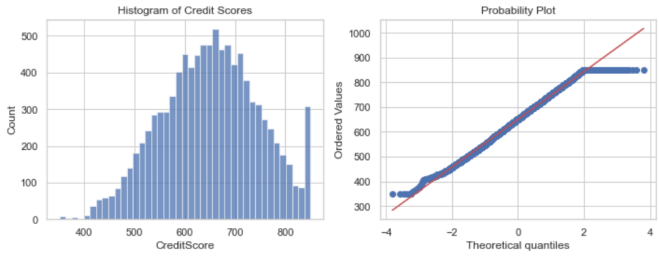

Figure 6: Checking the distribution using Histogram and Q-Q plot.

can use Chi Squared test to check if there is a significant difference between the frequencies in categories of a contingency table 7. It is a non-parametric test that is performed on categorical data to analyze the relationship between the observed and expected values.

- Null hypothesis is that there is no association between the variables- We expect to see equal frequencies of Exited customers in all satisfaction score groups.
- The alternative hypothesis is that there is an association of some type, meaning that a certain satisfaction score gives us information that the customer will likely leave the bank.

Satisfaction Score	Not-Exited	Exited
1	1541	386
2	1574	438
3	1638	401
4	1592	411
5	1604	395

Table 7: Contingency table

We will get the p-value by performing a Chi square test. If the p-value is less than 0.05, we can reject the Null hypothesis and accept the Alternative hypothesis.
In this case the p-value is 0.4519 and is greater than 0.05, therefore we do not reject the Null hypothesis at 95% level of confidence and it means that the Satisfaction Score and Exited variables are independent.
In Summary the satisfaction scores on complaint resolution do not provide indication of the customer's likelihood of exiting the bank.

4.6 Task A.6

The bank has a reward system where the customers earn points when they use their Diamond, Gold, Silver, and PLatinum bank card. Determine if there is a significant difference in the average points earned by the different groups of customers.
To answer this question, we will firstly calculate the average points earned for each card type, as shown in table 8 and visualize the data distribution by treatments using boxplot 'figure 7'.

Card Type	Mean
DIAMOND	606.158210
GOLD	606.924309
PLATINUM	608.947833
SILVER	604.078778

Table 8: Average Points Earned by Card Type.

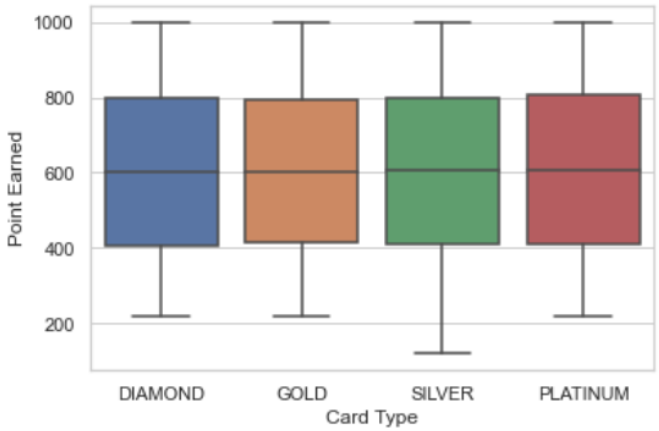

Figure 7: Data distribution using box plot.

Additionally we need to compare 4 groups in terms of the means of values. To Formally test the difference between two or more means, we will use the statistical method called One-Way Analysis Of Variance (ANOVA).

- The Null Hypothesis is that there is no significant difference in group means.
- Alternative Hypothesis is that there is a significant difference.

From the results provided in table 9, we see that the F-statistics is 0.198 and p-value is 0.898. Since the p-value is greater than the significance level of 0.05, we fail to reject the Null hypothesis. This will mean that there is no significant difference in the average Points earned between the different Card Type groups of customers.

	Sum of Squares	Deg_of_free	F-Statistic	p-value
Card Type	3.025132e+04	3.0	0.197625	0.898059
Residual	5.090236e+08	9976.0	NaN	NaN

Table 9: ANOVA Results for Card Type.

5 TASK B

The objective of this section is to develop a model that predicts whether a customer will complain or not based on the historical data.

To begin, we will examine the correlation with heatmap which reveals that the "Exited" variable has the strongest correlation with the target variable "Complain". Since our goal is to predict and prevent complaints, we will remove the "Exited" variable from the model features due to the logic.
As a customer leaves the bank after they have a complaint, not vice versa, our target will be to avoid the complaint, due to not wanting to lose customers. Similarly with the "Satisfaction Score" variable, this is a score of the "Complain" solution or result, therefore this can not be used to predict the Complain and will also be removed.

For model creation we will choose the variables with the strongest correlation except "Exited" as shown in table 10. Next, we will pre-process the dataset by encoding the categorical variables as numeric dummy variables to be able to use them in the model. We will also standardize or normalize the numerical features to ensure they are of a similar scale.

Feature	Correlation
Age	0.283140
Balance	0.119136
EstimatedSalary	0.011952
Point_Earned	-0.002812
HasCreditCard	-0.007592
Tenure	-0.013602
CreditScore	-0.025634
NumOfProducts	-0.047023
IsActiveMember	-0.154658

Table 10: Correlation of Features with 'Complain'

To address the issue related to class imbalance, there are more "No Complain" customers than "Complain" customers. To overcome this issue we will use oversampling through "SMOTE" which stands for Synthetic Minority Over-sampling Technique, to balance the classes.
After preparing our dataset, we split the data into two sets: the training set and the test set. We will use 80% of the data for training and other 20% to evaluate the accuracy of the model. Next, we will use the LazyPredict package to rank (table 11) the machine learning models that will most likely be suitable. This package contains both lazy Classifier and lazy Regressor which allow us to predict binary and continuous variables.

Model	Accuracy	Balanced	ROC AUC	F1 Score	Time
ExtraTreesCls	0.91	0.91	0.91	0.91	2.01
XGBClassifier	0.91	0.91	0.91	0.91	2.12
LGBMClassifier	0.91	0.91	0.91	0.91	0.53
RandomForestCls	0.90	0.90	0.90	0.90	2.30
BaggingCls	0.88	0.88	0.88	0.88	0.70
LabelSpreading	0.87	0.87	0.87	0.87	18.70
LabelPropagation	0.87	0.87	0.87	0.87	15.44
AdaBoostCls	0.84	0.84	0.84	0.84	0.78
DecisionTree	0.83	0.83	0.83	0.83	0.15
KNeighborsCls	0.81	0.81	0.81	0.81	0.30
ExtraTreeCls	0.80	0.80	0.80	0.80	0.05
ExtraTreeCls	0.80	0.80	0.80	0.80	0.05
SVC	0.81	0.81	0.81	0.81	13.99
NuSVC	0.81	0.81	0.81	0.81	17.69
ExtraTreeCls	0.80	0.80	0.80	0.80	0.05
LogisticReg	0.71	0.71	0.71	0.71	0.09
CalibratedCls	0.71	0.71	0.71	0.71	6.52

Table 11: Model Performances

5.1 Random Forest Classifier

From all tested models in table 11, we will choose RandomForestClassifier as our main model to use for this task.
The random forest classifier is an ensemble learning method that combines multiple individual decision trees to produce accurate predictions. It works by creating a large number of decision trees, and each is trained on a random subset of the training data and using a random subset of the features. Every tree will independently generate a prediction, and the final prediction is based on majority voting (see figure 8).

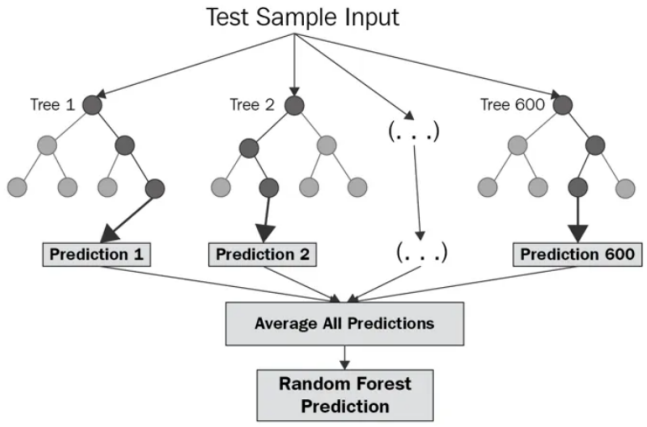

Figure 8: Random Forest Classifier

The strength of random forest can be found in its ability to produce highly accurate predictions by leveraging the diversity and independence of the decision trees. Due to differences in the trained subset of data, they are able to capture different aspects of the pattern in the data. Lastly, the predictions made by each decision tree should have low correlation with each other. In this way it ensures that errors made by some of the trees will not have a dominant effect on the overall prediction.
To optimize the performance of the model, we will fine-tune the hyperparameters of the model. This is performed to either enhance the performance and predictive power of models or to make the model faster[2].
We evaluate the developed model using the test dataset and calculate the overall accuracy score "0.899", which indicates the accuracy of the model in predicting customer complaints.

5.2 Feature and Result analysis

Furthermore, we will analyze the most important features of the model to identify the most influential features in our prediction. We can see from "figure 9" that the most influential features are Age, NumOfProducts and Balances.

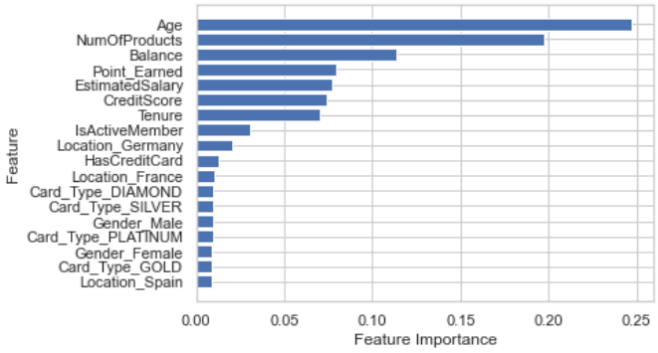

Figure 9: Feature Importance

Before analyzing the classification report, it is important to understand what each score means for the model.
There are four ways where we can check if our prediction is right or wrong:

- 1. TN / True Negative: the case is negative and prediction is negative
- 2. TP / True Positive: the case is positive and prediction is positive
- 3. FN / False Negative: the case is positive but prediction is negative
- 4. FP / False Positive: the case is negative but prediction is positive

Precision can be described as the ability of the classifier not to label an instance positive, if the case is negative. It is defined as the ratio of true positives to the sum of true positive and false positives.

- Precision is the accuracy of positive predictions.
- Precision = TP/(TP + FP)

Recall- is the ability of the classifier to find all the positive instances. It is defined as the ratio of true positives to the sum of true positives and false negatives.

- Recall is the fraction of positives that are correctly identified.
- Recall = TP/(TP + FN)

F1 score is described as a weighted harmonic mean of precision and recall, whereas the best score is 1.0 and the worst score is 0.0. F1 scores are lower than accuracy measures because they embed precision and recall into their computation.

- F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Support is the number of actual occurrences of the class in the specified dataset. If the training dataset is imbalanced, it may indicate structural weaknesses in the report scores of the classifier.

The ROC curve (figure 10) shows us the trade-off between sensitivity (TPR) and specificity (1-FPR). The closer to top-left corner is an indication of a better performance. The closer the curve comes to the 45-degree diagonal of the ROC curve, is an indication of a less accurate test.

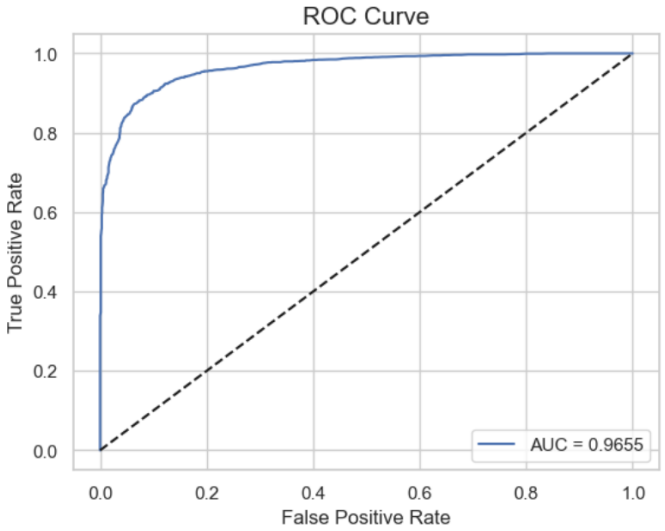

Figure 10: ROC Curve

Now, looking at the numbers provided in the classification report (table 12, we see that precision, recall, and f1-score were all around 0.90, which is a good indication of an overall good performance of the model.

	Precision	Recall	F1-Score	Support
0	0.91	0.89	0.90	1593
1	0.89	0.91	0.90	1585
Accuracy			0.90	3178
Macro Avg	0.90	0.90	0.90	3178
Weighted Avg	0.90	0.90	0.90	3178

Table 12: Classification Report

Based on the confusion matrix (figure 11), it is evident that the number of false predictions is low. Also, we see that the number of false negative predictions is lower than the false positive predictions. This outcome aligns with our objective when it comes to bank customer complaints. It is preferable to prioritize detecting potential complaints, even if it means paying more attention to certain customers, rather than missing someone who would raise a complaint.

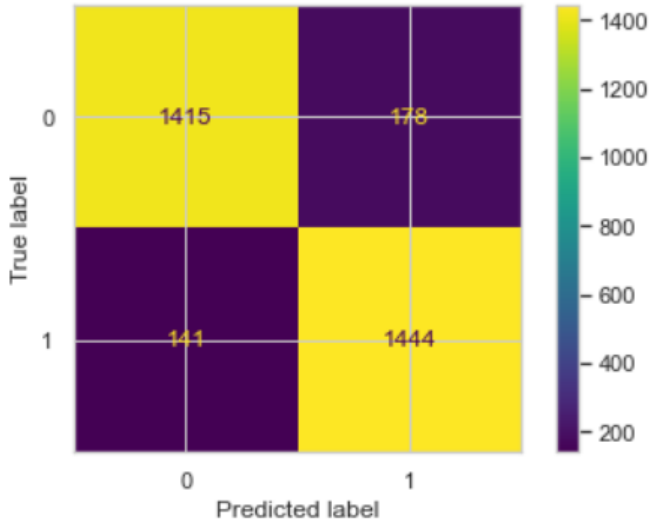

Figure 11: Confusion Matrix

6 TASK C

Once the model is finalized, use it to predict whether the bank customers included in New Sample file, will complain about the banking services based on their profile information in the dataset.

To run the model on the new dataset, we first need to load this New Sample dataset and make the same preprocessing steps as we did for the Main Sample. We will run our trained model, once the New Sample dataset is ready to predict whether the bank customers in the dataset will complain about the banking services. The model predicted the following outcomes for each customer as presented in table 13. The prediction indicates whether each customer is expected to complain (1) or not complain (0) based on their profile information in the dataset.

Index	Prediction
0	0
1	0
2	0
3	0
4	0
5	1
6	1
7	0
8	1
9	0
10	0
11	1
12	1
13	0
14	1
15	0
16	0
17	0
18	0
19	1

Table 13: Predictions on New Sample dataset

7 DISCUSSION

This statistical analysis of the bank customer provides us with valuable insights into different factors influencing customer complaints and the ability to predict customer behavior. Thorough exploration of the data, including examination of data distributions and finding patterns and trends, has given us an overview about the customer behavior.

We have through exploratory analysis, statistical tests such as Chi-Square test for categorical variables and ANOVA for numerical variable, correlation analysis, and feature selection managed to identify significant variables as predictors of customer complaints.The results from the correlation analysis conducted has shown a nearly perfect correlation, which highlights the significance of addressing customer concerns in order to mitigate churn and improve customer retention.

The variables such as age, balance, tenure and card type, were used to develop a Random Forest Classifier model. Our model performed very well in predicting customer complaints, as evidenced by high accuracy, precision, recall and F1-score metrics.

Important to mention that while our model has achieved high performance on the given dataset, its prediction can vary based on different contexts or with new data. Therefore it is crucial that the model stays monitored to ensure its continued good performance and accuracy.

8 CONCLUSION

The findings of this analysis have significant implications for bank management. The developed model can be utilized by the bank to proactively identify those customers who are likely to complain and take appropriate actions to deal with the negative outcome. By doing so, the bank can implement different targeted strategies to enhance customer retention, mitigate customer complaints and ultimately improve customer satisfaction and loyalty. Lastly, The ability to develop predictive models and extract meaningful insights from data is crucial for making informed business decisions and optimizing customer relationship and management strategies.

REFERENCES

[1] Matthias Bogaert and Lex Delaere. “Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art”. In: *Mathematics* 11.5 (2023), p. 1137. URL: https://www.mdpi.com/2227-7390/11/5/1137.

[2] Sruthi E.R. *Understand Random Forest Algorithms With Examples (Updated 2023).* URL: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. (accessed: 10.5.2023).

[3] Josh Howarth. *Customer Retntion Rates.* URL: https://explodingtopics.com/blog/customer-retention-rates. (accessed: 10.5.2023).

[4] Hoang Tran, Ngoc Le, and Van-Ho Nguyen. “CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR USING MACHINE LEARNING-BASED CLASSIFICATION MODELS.” In: *Interdisciplinary Journal of Information, Knowledge & Management* 18 (2023). URL: https://www.informingscience.org/Publications/5086.

[5] Irfan Ullah et al. “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector”. In: *IEEE Access* 7 (2019), pp. 60134–60149. DOI: 10.1109/ACCESS.2019.2914999.