Customer Churn Analysis

Martin Liang

1 INTRODUCTION

Customer churn refers to customers terminating their association with a company, resulting in a discontinuation of the customer-company relationship. High churn rates indicate a business losing many customers, leading to reduced revenue and profitability. Conducting a customer churn analysis can aid in identifying customers who are at a higher risk of termination and enable retention strategies to be implemented. This research paper aims to analyze the bank customer churn dataset, answer six business questions related to customer churn, and discover the impact on bank profitability. Each of the questions will recommendations on how to solve the associate issue. Predictive analysis will be included to provide informative insights that can assist bank management in making good business decisions.

- What is the proportion of customers still using the banking services compared to those that have left in the period covered in the dataset? Is there a significant difference in the proportion that the bank authority should be worried about?
- What is the relationship between the number of complaints received by the bank authorities and the number of exited customers?
- What are the characteristics and statistics (in terms of gender, age groups, tenure etc,) of the customers that are more likely to complain?
- Is there a significant difference between the credit scores of all the customers complaining and those not in the period covered in the dataset?
- Bank has a reward system for associate bank card ranks; Diamond, Gold, Silver, and Platinum. Is there a significant difference in the average points earned by the different groups of customers?

These mentioned questions are not chosen randomly. The questions are aimed at addressing important business concerns related to customer satisfaction, retention, and complaint management in the banking industry. The first question seeks to determine the proportion of customers who have left the bank and whether there is a difference that should concern the bank authority. The second question investigates the relationship between the number of complaints received by the bank and the number of exited customers. The third question focuses on the characteristics and statistics of customers who are more likely to complain. The fourth question investigates whether there is a significant difference in credit scores between customers complaining and those not in the period covered in the dataset. The last question explores the reward system to see if there are differences in average points earned by the different rank groups. These are the following business questions that will be answered;

2 DATASET DESCRIPTION

The purpose of this analysis is to assist bank businesses in identifying the reasons for customer churn. Two datasets have been used for this analysis: the "Main Sample" dataset, which will be used to answer business questions, and the "New Sample" dataset, which will be used to predict customer churn in the associated bank system. The "Main Sample" dataset has a size of 78.1 KiB and contains 9,980 observations. These are the following columns;

- CustomerId
- CreditScore
- Location
- Gender
- Age
- Tenure
- Balance
- NumOfProducts
- HasCreditCard
- IsAcativeMember
- EstimatedSalary
- Exited
- Complain
- Satisfaction Score
- Card Type
- Point Earned

the second one is New Sample. It has a size of 288.0 B and 20 observations. The New Sample contains the following columns;

- CustomerId
- CreditScore
- Location
- Gender
- Age
- Tenure
- Balance
- NumOfProducts
- HasCreditCard
- IsAcativeMember
- EstimatedSalary
- Exited
- Satisfaction Score
- Card Type
- Point Earned

3 DATA ANALYSIS

3.1 Question 1: What is the proportion of the customers that are still using the banking services compared to those that have left in the period covered in the dataset? Is there a significant difference in the proportion that the bank authority should be worried about?

The analysis was conducted consisting of three variables: "exited", "complaining", and "gender". The variable "exited" was used to differentiate between customers who have exited and those who have continued their engagement with the service provider. The variable "complaining" is used to distinguish customers who have registered a complaint from those who have not. Furthermore, gender was incorporated into the analysis to provide a better understanding of the analysis.

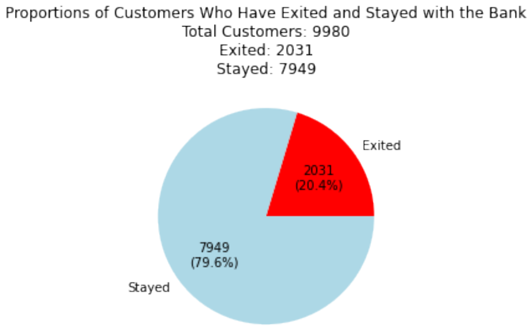

Figure 1: Pie-chart of customers who have exited and stayed

According to the data presented in Figure 1, it can be observed that the total number of customers is 9980. Among these customers, 7949 customers have continued their engagement with the bank service, while 2031 customers have decided to discontinue their association. This implies that approximately 79.6 percent of the customers have chosen to stay with the bank service, while the remaining 20.4 percent have exited.

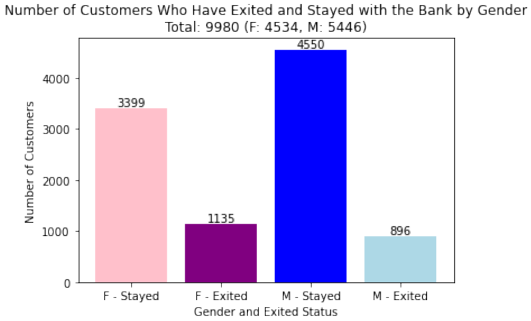

Figure 2: Number of Customers genders who have exited and stayed with the bank

Based on the data presented in Figure 2, it can be observed that the total number of customers is 9980, comprising 5446 males and 4534 females. Among the male customers, 4550 have decided to continue their association with the bank, while 896 males have exited. Similarly, among the female customers, 3399 females have chosen to stay with the bank, while 1135 females have discontinued their engagement. These findings reveal a higher proportion of male customers are inclined towards staying with the bank, whereas a greater number of female customers are exiting the system.

In conclusion, the analysis of the data presented in Figure 1 and Figure 2 reveals customer retention and attrition rates in the context of bank service. While a significant proportion of the customers have chosen to continue their engagement with the bank service, a notable number have discontinued the bank association. Given the differences in the proportions of customers leaving and staying with the bank, the bank authority may not have a reason to be concerned about the number of customers exiting the system. If a proportion issue arises for this, a recommendation of conducting surveys and gathering feedback from customers who have chosen to discontinue the bank service will mitigate the issue.

3.2 Question 2: What is the relationship between the number of complaints received by the bank authorities and the number of exited customers?

The Analysis uses the "Complain" column to indicate whether a customer has made a complaint (1) or not (0), while the "Exited" does the same but indicates whether a customer has exited the company. Given the nature of these variables, it is reasonably related, with customers who have complained potentially being more likely to exit the company.

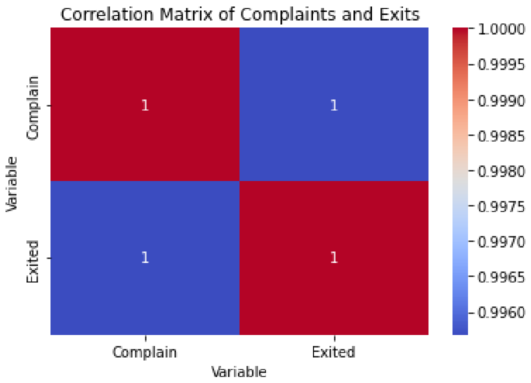

Figure 3: Correlation Matrix of Complaints and Exits

The correlation analysis that has been generated in Figure 3 revealed a strong positive correlation between the number of complaints, and the number of customers who have exited the service with a correlation coefficient of 1. This indicates that as the number of complaints received by the bank increases, it will also increase the number of customers who decide to discontinue the association with the bank service. However, if the correlation coefficient resulted in -1, it would indicate a strong negative correlation, which is an indication of the desired outcome. To further validate this finding, a linear correlation plot (figure 4) was generated.

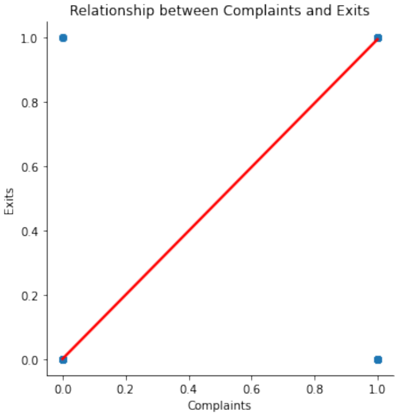

Figure 4: Relationship between complaints and exits

Based on the figure 4, this is a strong positive correlation. The two variables are strongly related and tend to move in the same direction. This indicates that there is a direct relationship between the two variables. In conclusion, Customer complaints have a significant impact on customer retention rates, as an increase in complaints tends to result in a corresponding increase in customer attrition rate. To mitigate this issue, improving customer satisfaction can reduce the number of complaints.

3.3 Question 3: What are the characteristics and statistics (in terms of gender, age groups, and tenure etc,) of the customers that are more likely to complain?

The analysis of the research question was performed utilizing three variables. "Age", "Complain", and "HasCreditcard". The variable "Complain" was used as a filter to obtain information associated with specific characteristics. The variable "Age" was utilized to identify the age group that complains the most, and "HasCreditcard" was analyzed to whether individuals with or without credit cards complain more. "Gender" is used for analyzing balance complaint distribution between genders.

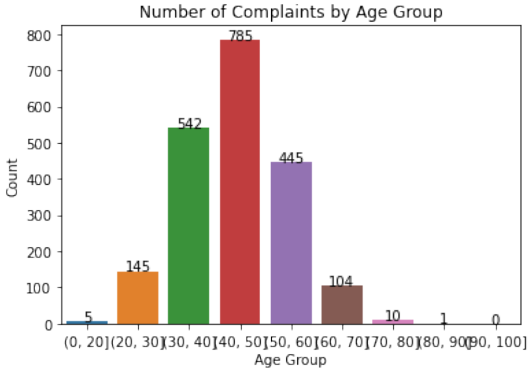

Figure 5: Number of complaints by age group

Based on the analysis presented in Figure 5, it can be observed that the age group of 40-50 exhibits the highest number of complaints, with a value of 785. Following this group, the age group of 30-40 had 542 complaints, while the age group of 50-60 had 445 complaints. The findings suggest that customers in their 40-50 age group tend to lodge more complaints as compared to customers in other age groups. In conclusion, the analysis conducted highlights that the 40-50 age group has the highest number of complaints among the selected age groups. This finding indicates that there may be certain factors or experiences that are more favored among customers in this age group that lead to dissatisfaction with bank service.

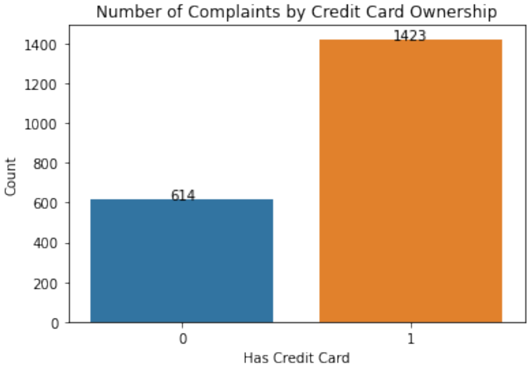

Figure 6: Genders who have exited and stayed

The dataset values used in the analysis are binary, represented by 0 and 1, where 0 signifies individuals who do not possess a credit card, and 1 indicates those who do. The findings presented in Figure 6 indicate that customers who hold credit cards tend to complain more frequently, with a value of 1423, compared to those who do not use credit cards, with a value of 614.

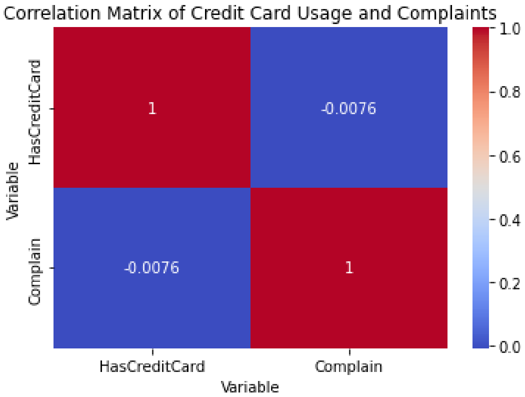

Figure 7: Correlation Matrix of Credit Card Usage and Complaints

To further research this, According to Figure 7, there is a positive correlation between credit card usage and complaints, as indicated by the correlation coefficient value of 0.0024. This suggests that customers who hold credit cards tend to complain more frequently than those who do not have credit cards. However, the correlation coefficient is relatively weak, meaning that credit card usage does not have a strong impact on the frequency of complaints.

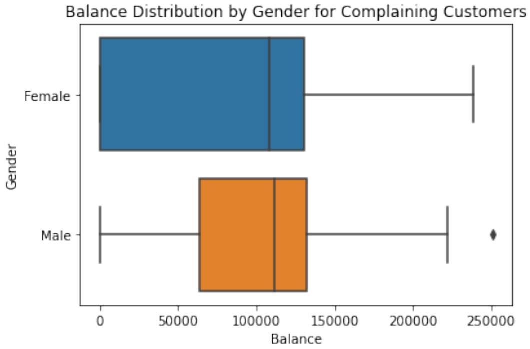

Figure 8: Balance Distribution by Gender for complaining Customers

Percentage of complaining customers by gender:	
Female	25.099250
Male	16.507528
Average age of complaining customers	44.78
The average tenure of complaining customers by gender:	
Female	4.932337
Male	4.936596

Table 1: Percentage of Customer Complaining .

According to Figure 8, it can be observed that female customers spend more than their male counterparts, as the height of the bar representing female customers is greater than that of male customers. Additionally, Table 1 provides further insights into the data, including the percentage of complaining customers by gender, the average age of complaining customers, and the average tenure of complaining customers by gender. The data indicates that the percentage of female customers who filed complaints (25.1%) is higher than that of male customers (16.5%), suggesting that female customers are more likely to file complaints. Moreover, the average age of complaining customers, regardless of gender, is average of 44.78 years. The average tenure of complaining female and male customers is very similar, with female customers having an average tenure of 4.932337 years and male customers having an average tenure of 4.936596 years. In, conclusion, individuals in the age group of 40-50, with a combination of possessing a credit card, and being female are more likely to file a complaint. To mitigate this issue, the bank can perform a survey of customers to better understand their experiences with the bank and identify any issues that could lead to customer churn.

3.4 Question 4: Is there a significant difference between the credit scores of all the customers complaining and those not in the period covered in the dataset?

The research question is aimed at investigating the relationship between two variables, namely "Complain" and "CreditScore," to answer the research question of whether there is a significant difference between the credit score of customers who have lodged complaints and those who have not.

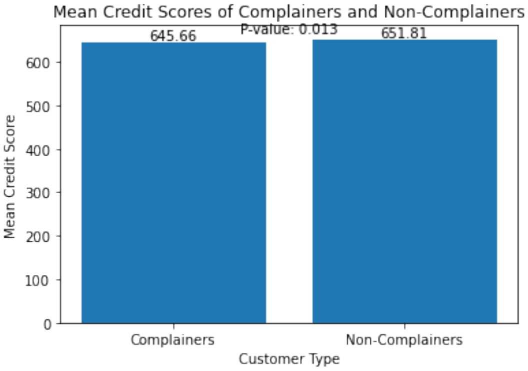

Figure 9: Mean Credit Scores of Complainers and non-complainers

Based on the results of the generated analysis for Figure 9, it was observed that customers who complained had a mean credit score of 645.66, while those who did not complain had a mean credit score of 651.81. As presented in Figure 8, the calculated p-value for the two-sample t-test was found to be 0.013. The p-value indicates that there is a statistically significant difference in the mean credit score between complainers and non-complainers. In conclusion, there is a difference between the credit scores of customers who complained and those who did not during the period covered by the dataset. To mitigate this issue, investigating other factors about credits could solve this issue. A potential issue is that the customers starting with low balances influenced the decision to file a complaint.

3.5 Question 6: Bank has a reward system for associate bank card ranks; Diamond, Gold, Silver, and Platinum. Is there a significant difference in the average points earned by the different groups of customers?

Columns from the dataset; ’Card Type’ and ’Point Earned’ are used to answer the question to find necessary information to group the data by card type and calculate the mean and standard deviation of points earned for each group. The ’Card Type’ column is used to group the data, and the ’Point Earned’ contains the numerical values that are being compared. The ANOVA test is then performed using the ’Point Earned’ column to determine if there is a significant difference in the means of the different groups.

	mean	std
DIAMOND	606.158210	226.362039
GOLD	606.924309	225.490165
PLATINUM	608.947833	226.630065
SILVER	604.078778	225.058461
one-way ANOVA test:		
F-statistic:	0.1976248954357644	
P-value	0.8980588339589598	

Table 2: Mean and standard deviation of points by each card type.

Based on Table 2, the one-way Analysis of the Variance test, a method to analyze the difference between the means value of more than two groups, indicates that there is no significant difference in the average points earned by the different groups of customers. The p-value is greater than 0.05. This means it does not reject the null hypothesis that there is no difference in the means of the populations. In conclusion, it appears that the reward system does not favor one card type over another in terms of points earned. To fix this issue, perform surveying customers to gather feedback on the reward system and their satisfaction with it.

4 REGRESSION ANALYSIS FOR CUSTOMER COMPLAIN PREDICTION MODEL

4.1 Logistic Regression model prediction

Classification Report:				
	precision	recall	f1-score	support
0	0.82	0.97	0.89	2418
1	0.41	0.09	0.15	576
accuracy			0.80	2994
macro avg	0.61	0.53	0.52	2994
weighted avg	0.74	0.80	0.74	2994

Table 3: Classification report for Logistic Regression model.

The classification report (Table 3) indicates that the model's precision for predicting no complaints was high at 0.82, while the precision for predicting complaints was low at 0.41. The recall rate for predicting complaints was also low at 0.09, which means that the model failed to identify many of the customers who complained. The f1-score for predicting complaints was also low at 0.15. The accuracy of the model was 0.80, which indicates that it correctly predicted the outcome for 80% of the bank customers in the New Sample dataset.

Index	CustomerId	Predicted_Complain
0	15710408	0
1	15598695	0
2	15649354	0
3	15737556	0
4	15671610	0
5	15625092	1
6	15741032	0
7	15750014	0
8	15784761	0
9	15768359	0
10	15805769	0
11	15719508	0
12	15609011	0
13	15703106	0
14	15626795	0
15	15773731	0
16	15756196	0
17	15687903	0
18	15777599	0
19	15754577	0

Table 4: Result report for Logistic Regression model.

The table shows the results of the Logistic Regression model's predicted complaints for a list of customers identified by their CustomerId. The model predicted that index 5 with an ID of 15625092 will likely complain.

4.2 Decision Tree Classifier Model

Classification Report:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	2418
1	0.99	0.99	0.99	576
accuracy			1.00	2994
macro avg	0.99	0.99	0.99	2994
weighted avg	1.00	1.00	1.00	2994

Table 5: Classification report for Decision Tree Classifier Model.

According to Table 5, the results of using The decision Tree Classifier Model achieved high precision, and recall, for both the "complain" and "no complain" classes. This indicates that it was accurate in identifying customers who would and would not complain. The accuracy of the model was also high at 1.00, indicating that it correctly predicted the outcome for all bank customers in the dataset. Overall, the decision tree classifier model appears to be effective in predicting customer complaints based on credit history.

Index	CustomerId	Predicted_Complain
0	15710408	0
1	15598695	0
2	15649354	0
3	15737556	1
4	15671610	0
5	15625092	1
6	15741032	0
7	15750014	0
8	15784761	1
9	15768359	0
10	15805769	0
11	15719508	1
12	15609011	1
13	15703106	0
14	15626795	1
15	15773731	0
16	15756196	0
17	15687903	0
18	15777599	0
19	15754577	1

Table 6: Result report for Decision Tree Classifier Model.

The table shows the results of the Decision Tree Classifier Model's predicted complaints for a list of customers identified by their CustomerId. 1 indicating a predicted complaint and 0 indicating no predicted complaint. The model predicted that 7 customers will likely complain.

5 PREDICTION MODEL DISCUSSION

The reason why the Decision Tree Classifier Model and Logistic Regression Model were picked for predicting customers who will likely complain is due to effective algorithms that solve binary problems. Both models are considered easy to interpret, which makes the model useful for businesses to understand the results. Some machine learning models are better at regression problems, for example, Linear regression and polynomial regression to predict larger numbers.

Based on the results of the classification report, the decision tree classifier model outperformed the Logistic Regression model in terms of accuracy, precision, and recall score The decision tree model achieved an accuracy of 1.00, indicating that it correctly predicted the outcome for all customers in the dataset, while the Logistic Regression model achieved an accuracy of 0.80, indicating that it correctly predicted the outcome for 80 percent of the customers. In terms of precision, the decision tree model achieved high precision for both the "complain" and "no complain" classes, while the Logistic Regression model achieved high precision only for the "no complain" class and low precision for the "complain" class. This means that the decision tree model was better at identifying both customers who would and would not complain, while the Logistic Regression model was better at identifying customers who would not complain. The recall rate for predicting complaints was also low for the Logistic Regression model, indicating that it failed to identify many of the customers who complained. On the other hand, the decision tree model achieved a high recall rate for predicting complaints, meaning that it was able to identify most of the customers who complained.

In conclusion, the decision tree classifier model appears to be a more effective model for predicting customer complaints based on credit history, as it achieved higher accuracy, precision, recall, and f1-score than the Logistic Regression model.

6 CONCLUSION

The analysis has provided valuable insight and has provided 6 questions with recommendations on how to mitigate different factors of custom The data sets were predicted on two predictive models, the Logistic Regression model and Decision Tree Classifier Model. The Tree Classifier Model predicted the data set close to perfect in comparison to the logistic regression model only acquiring high accuracy on non-complain.