Analytics in Banking: An Examination of Customer Churn and Retention Rates

Kristofer DeYoung

1 ABSTRACT

This report focused on analyzing customer churn rates within the banking sector using predictive and descriptive analytics to evaluate various customer attributes. The report incorporated geographical locations, customer data, and measures of satisfaction, utilizing various techniques, including machine learning, to forecast customer attrition. Key findings spotlighted a strong correlation between complaints and customer exits, regional differences, and the influence of distinct factors impacting customer retention rate.

2 INTRODUCTION

In his seminal 1996 Harvard Business Review article, Reichheld posited that American corporations were losing half of their customers every five years, a phenomenon with profound implications for business's profitability and longevity [6]. However, the causes of customer defection remained elusive and differed markedly across sectors. Reichheld further noted that long-term customers demanded less time investment and tended to purchase more products. This observation underscores the significance of customer retention, especially considering the expenses related to acquiring new customers and initial startup costs.

Subsequently, Reichheld argued in a follow-up paper that a modest 5% increase in customer retention could boost profits by an impressive 25% to 95% [5]. Despite the ubiquity of this assertion in both academic discourse and business practice, the original research underpinning these statistics remains somewhat obscure, spanning the period from 1990 to 2001 while Reichheld was at Bain Company. The validity of these widely accepted statistics notwithstanding, they certainly serve as a compelling foundation for analyzing bank churn rate data. The significance of customer retention extends far beyond the direct impact on profit margins. Businesses often only appreciate the crucial role of customer loyalty once dwindling profit margins become unavoidable. Such myopia can result in hastily implemented solutions that merely address symptoms while neglecting the fundamentals of value creation. Therefore, insights into the factors driving customer loss could be invaluable, enabling businesses to tailor value-added services to specific customer segments and boost profitability.

The concept of fractional reserve banking, which prevails in contemporary banking practices, imposes a statutory obligation on banks to maintain a specific ratio of liquid assets to debt [2]. Banks often borrow additional funds from central reserves to enhance their investment potential. However, this strategy hinges on specific customers' (depositors) readiness to hold their money in the bank, incentivizing banks to offer attractive products and interest rates. If a bank's debt-to-liquidity ratio falls beneath the legally prescribed limit, it risks insolvency. Moreover, a crisis of confidence can precipitate a bank run, a situation where customers simultaneously attempt to withdraw their funds, potentially causing a liquidity crisis [1]. This scenario is becoming increasingly relevant in today's uncertain financial climate. Rising interest rates pressure banks with high exposure to the government bond market are being forced to sell their investments at a loss. Consequently, depositors may question the bank's ability to meet its fiduciary responsibilities, potentially triggering a domino effect, eventually leading to bankruptcy. A recent case in point was the unprecedentedly rapid bank run on Silicon Valley Bank (SVB), fueled by the power of social media to disseminate fear contagiously [3].

Furthermore, a study by Gur Ali et al. (2014) revealed intriguing insights into the churn rate among customers in diverse sectors, including finance [4]. Their research demonstrated that independent binary classification models outperformed other techniques, such as survival-based regression models, even using standard data balancing methods like oversampling. Employing a dynamic churn prediction methodology comprising multiple binary classification models, they assessed the accuracy between models and used SMOTE to balance the dataset. This strategy resulted in substantially higher accuracy than single observation methods, highlighting the importance of innovative approaches in understanding and predicting customer churn.

3 DATA DESCRIPTION

This report utilized a dataset comprising 9980 distinct records, each featuring 16 different attributes. These attributes were evaluated in three categories: customer descriptions, current status, and dynamic variables that might predict customer activity. Customer descriptions included identifiers such as CustomerId, Location, Gender, and Age. The current status of the customer was represented through attributes such as Tenure, IsActiveMember, Exited, Complain, and Satisfaction Score. Dynamic variables, potentially predictive of customer activity, included HasCrCard, Card Type, Points Earned, NumOfProducts, Balance, CreditScore, and Estimated Salary. An initial exploration of the dataset was conducted to understand the attributes and reformat respective values for compatibility with various statistical techniques. The data types for each attribute were modified to suit the proposed questions and are presented in Table 1.

Table 1: Data Types of the Dataset

Column	Dtype
CustomerId	int64
CreditScore	int64
Location	category
Gender	object
Age	int64
Tenure	int64
Balance	float64
NumOfProducts	int64
HasCreditCard	bool
IsActiveMember	bool
EstimatedSalary	float64
Exited	bool
Complain	bool
Satisfaction Score	int64
Card Type	category
Point Earned	int64

Considerations were also made to convert NumOfproducts to categorical; this could potentially benefit such as; preserving the order of values benefiting ML algorithms.

3.1 Data Preprocessing and Outlier Detection

The data preprocessing stage included evaluating all attributes in the dataset for potential outliers using standard deviation Z-scores and the Interquartile Range (IQR). The normal threshold for what is considered an outlier is a z-score of 3; this indicates a maximum positive or negative deviation of 3 from the mean. Instead, the normal threshold for IQR is 1.5 resulting in anything outside 99.72% data that is within three deviations of the mean being flagged as an outlier. In this case, for a matching comparison between z-score and IQR, the IQR threshold was adjusted 1.7, resulting in fewer potential outliers. The underlying reasoning was the initial analysis through boxplots and histograms revealed an unusually clean and normally distributed data set, giving little reason to suspect any significant impact from outliers. Using both methods, only three attributes were flagged using the method previously outlined, specifically CreditScore, Age, and NumOfProducts. Table 2 shows the results of outlier detection with the extremes on both ends of the spectrum.

Table 2: Number of outliers for the specified attributes

	Z-score	IQR	Min (Z-score)	Max (Z-score)	Min (IQR)	Max (IQR)
CreditScore	8	6	350.0	359.0	350.0	351.0
Age	133	281	71.0	92.0	65.0	92.0
NumOfProducts	59	59	4.0	4.0	4.0	4.0

3.1.1 Credit Score. Totaled eight outliers, all on lower the lower end of the spectrum; after further observations, none seem to qualify as outliers, and no further action is taken. It is, however, noteworthy that there was a deviation from the normal distribution of 233 customers with a max credit score of 850, leading to the observation that if higher tiers of credit scores existed, there is a high probability that many customers qualify for higher scores.

3.1.2 Age. Had a larger difference between the two chosen methods of detection; further analysis of 281 records of people flagged as outliers shows that they were all over the age of 65, totaling 2,8% of all customers in the data set, when cross-referencing this to global averages in 2019 from the United Nations[7]which was 9%. It clearly shows these data points cannot be considered outliers, providing insight into the bank's target audience.

3.1.3 NumOfProducts. This attribute resulted in all 59 customers with four products flagged as outliers; clearly, however, not the case, as this is still relevant data to explore; it is noteworthy that only 0.6% of customers fully engaged in all bank offerings.

3.2 Data Conversion

An alternate dataset was transformed into a numerical format to enhance compatibility with machine learning algorithms. This transformation involved converting categorical and boolean data into integer values, resulting in a purely numerical data frame. Furthermore "binning" or grouping of data such as ages, balance, and Income was considered to provide more clarity when answering specific questions; however, the loss of information was considered to negatively impact the results.

3.3 Summary

In summary, the data set was carefully explored and pre-processed, ensuring its compatibility with the predictive model. Though outliers were identified, these data points were not removed in CreditScore, Age, and NumOfProducts. Their inclusion was deemed potentially valuable for the analysis, and it was noted that the chosen predictive model, logistic regression, is not overly sensitive to outliers. The decision to refrain from further expanding the outlier threshold was driven by the understanding that these observations could still hold value within the context of this research. Future studies with more domain knowledge could examine these outliers' impact on predicting customer behavior in banking.

4 DESCRIPTIVE ANALYSIS

4.1 Customer churn rate

4.1.1 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 the proportion that the bank authority should be worried about?

4.1.2 Analysis. During initial data exploration, disparities were identified among the branches, prompting a more detailed examination of the data set segmented by location (France, Germany, and Spain). We grouped the data by the 'Location' attribute, and due to the boolean nature of the 'Exited' attribute, we categorized the customers into two groups: 'Remained' (False) and 'Exited' (True). We then counted the total number of customers and calculated the mean for each group, as summarized in Table 3.

Table 3: Number of customers who exited and remained for each country

Location	Remained	Exited	Total	Exited %
France	4197	808	5005	16.14
Germany	1692	812	2504	32.43
Spain	2060	411	2471	16.63
Average	2649	677	3326	20.35

As evident from Table 3, on average, 20.35% of customers exited the bank during the period covered by the dataset. The French and Spanish branches experienced an exit rate of roughly 16%, while the German branch had a considerably higher rate of 32.43%. This trend of customer churn, particularly in the German branch, is more strikingly illustrated in Figure 1.

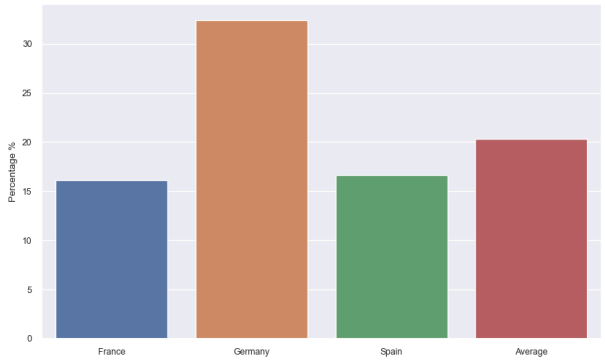

Figure 1: Percentage of Customers Exited by Location

The presented analysis underlines the potential benefits of evaluating customer retention strategies. Furthermore, it points to the need for targeted interventions, especially in locations with high customer churn rates.

4.1.3 In conclusion. , our findings suggest that the German branch, in particular, might benefit from further investigation and a revised customer retention approach. While the overall average customer exit rate of 20.35% may appear high, it is important to benchmark this against sector averages. Further investigation into industry standards would provide a more nuanced understanding of these findings.

4.2 Relationship between Complaints and customer churn rate

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

4.2.2 Analysis. Grouping the data by the attribute 'Complain' and 'Exited' into each of their respective boolean categories: True (complaint lodged) and False (no complaint lodged). Given that both attributes are boolean and therefore non-continuous, neither Pearson nor Spearman correlation coefficient may seem ideal initially. This is because the former measures linear relationships between continuous variables, while the latter determines any monotonic relationship that don't necessarily move in the same direction.

However, initial observations showed that the Boolean values for 'Complain' and 'Exited' mirrored each other closely. Therefore, using either correlation coefficient could still provide insights into the relationship between the two variables. For simplicity, the Pearson correlation coefficient was employed as per the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The analysis was focused on the "Exited" and 'Complain" attributes, as shown in Figure2. The correlation matrix reveals a correlation of 0.996, indicating an almost perfect relationship between these two variables. This analysis suggests that customers who lodged a complaint were highly likely to have exited the bank later.

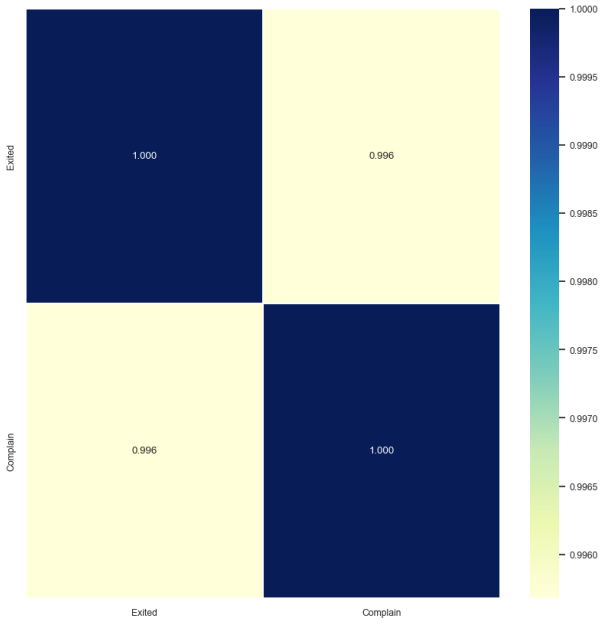

Figure 2: Correlation Matix Complained vs. Exited

A further inspection into Table 4 reveals that out of the 2037 customers who lodged a complaint, only ten remained with the bank. This corresponds to a retention rate of 0.49% post-complaint or 79.69% overall, pointing to a significant opportunity for improvement in the bank's complaint-handling process.

Table 4: Number of customers who have exited and have complained vs. complained and not exited

Complain	Exited	Count
False	False	7939
False	True	4
True	False	10
True	True	2027

As highlighted earlier, 20.4% of the total customer base lodged a complaint. While it is unclear whether a retention rate of 0.49% post-complaint aligns with industry standards, this area merits further investigation.

4.2.3 In conclusion. , the data suggests that the bank did not anticipate a correlation of this magnitude between complaint lodging and customer exit, indicating the need for measures to handle better and resolve customer complaints, thereby improving customer retention rates.

4.3 Profiling complaint-prone customers

4.3.1 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?

4.3.2 Analysis. Our analysis aims to delineate the defining characteristics of customers prone to lodging complaints. Given most data's categorical or boolean nature, we converted the dataset into a numerical format to effectively utilize the mean for our examination. Data was bifurcated into 'exited' and 'not exited' categories. By comparing the percentage difference between these categories, we identified attributes with a deviation larger than 5% as significant contributors to the profile, as depicted in table 5.

Table 5: Mean of each feature for exited(True) and not exited(False) customers

Complain	Gender	Age	Balance	IsActiveMember
False	0.57	37.41	72758.51	0.55
True	0.44	44.78	91203.84	0.36
Percentage	77.10	119.70	125.35	65.39

While most attributes exhibited minimal differences between customers who exited and those who stayed, significant discrepancies were observed in four chosen attributes shown in table 5: "Gender", "Age", "Balance", and "IsActiveMember". In our numerical dataset, where 'Male' and 'Female' were encoded, the likelihood of a complainer being female was 56%. The average age was higher (around 45), and they had notably larger account balances (approximately 91000, which was 125% greater). Also, there was a 64% probability that the complainer was not an active member.

The data was further segmented by branches for deeper regional insights. As shown in table 6, the German branch had almost double the account balance for complaining customers compared to other branches while spontaneously ascertaining that Germans are double as likely to complain. Indicating unique customer expectations in the German market that might need attention.

Table 6: Regional characteristics

Location	Complain	Balance
France	0.16	62127.51
Germany	0.33	119726.18
Spain	0.17	61902.28

4.3.3 In conclusion. , while complaint-prone customers generally resemble the average customer, specific traits like larger account balances and regional trends set them apart. Addressing these specific needs could enhance complaint management and customer retention. Additionally, it's recommended to analyze further using simple binning or, preferably, determining the min-max of each attribute within the first Inter Quartile Range for a more nuanced customer profile.

4.4 Credit scores relationship with complaints

4.4.1 Question 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 by the dataset?

4.4.2 Analysis. Our dataset indicates a normal distribution of credit scores, with 2.3% of scores at the maximum limit of 850. An in-depth analysis of the remaining scores shows minimal deviation from the overall distribution. Hence, it seems that a large number of scores at 850 is not due to outliers but rather represents customers who have achieved the maximum credit score during the dataset's time frame. In this context, further examination of credit scores focuses on discerning any significant differences between scores. The exploration of related data shows the normal distribution of credits score, showing little necessity for other methods in calculating averages than the standard mean; the distribution is further illustrated in Figure 3.

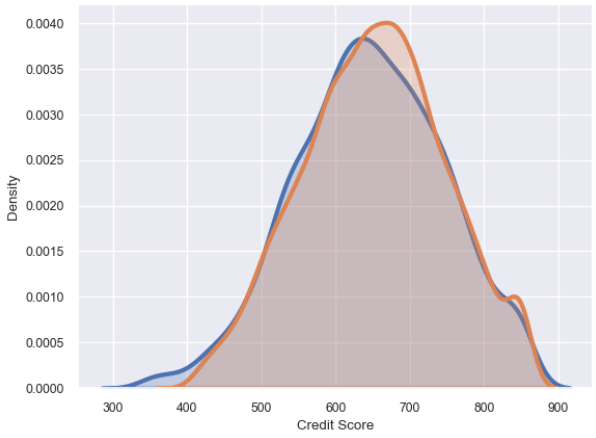

Figure 3: Distribution of Credit Scores

Table 7 summarizes the average credit score for both the customers who have lodged complaints(True) and those who haven't(False).

Table 7: Average credit score complained vs. not

Complained	CreditScore
False	651.81
True	645.66

Even though the mean scores in table 7 differ by a mere 0.04%, we will conduct a t-test for independent samples to investigate whether this difference is statistically significant. The null hypothesis (H0) states no difference in the credit scores of the two customer groups, while the alternative hypothesis (H1) proposes a difference. The t-test involves computing a t-statistic using the following formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

- t is the t-statistic.
- $\bar{X}_1$ and $\bar{X}_2$ are the means of the two samples.
- ${s_1}^2$ and ${s_2}^2$ are the variances of the two samples
- n_1 and n_2 are the sizes of the two samples.

The degrees of freedom are given by:

$$df = n_1 + n_2 - 2$$

Applying these calculations, we get the following:

- T-statistic: -2.561
- P-value: 0.0104

As shown in Figure 4, our calculated t-statistic exceeds the critical value, leading us to reject the null hypothesis. This suggests a statistically significant difference in the credit scores of the customers who have complained and those who haven't, at a significance level of 0.05.

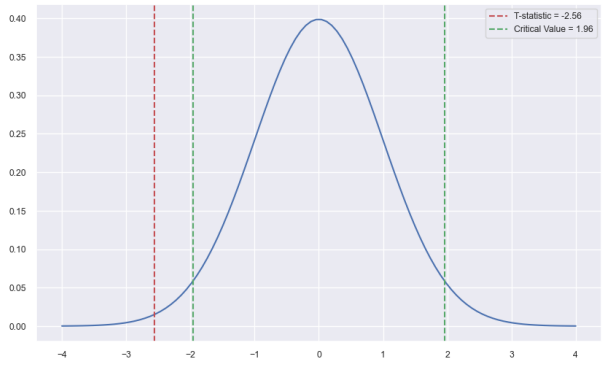

Figure 4: T-Distribution

4.4.3 In conclusion. , our analysis indicates a likely difference in credit scores between the two customer groups within the given period. Further investigation could provide insights into the factors contributing to this difference and its implications for customer complaints.

4.5 Customer satisfaction

4.5.1 Question 5: Do the satisfaction scores on complaint resolution provide an indication of the customers' likelihood of exiting the bank?

4.5.2 Analysis. The dataset provides evidence suggesting a correlation between customer attrition(churn rate) and lower satisfaction scores. However, caution should be exercised in interpreting these findings due to the small sample size of customers who lodged a complaint but chose to remain with the bank. Out of 2037 customers who registered complaints, only ten (0.495%) continued their relationship with the bank; this was previously analyzed in Question 2 and table 4. Table 8 presents an overview of the satisfaction scores relative to the customers who have complained. It shows that the average satisfaction score of customers who complained but decided to stay is, on average, 110% higher than the customer that has left. However, it is noteworthy that the number of customers who chose to leave was 20270% higher than those who chose to stay, table 8.

Table 8: Mean satisfaction score for customers who exited(True) and not exited(False)

Exited	Satisfaction Score	Amount
False	3.30	10.00
True	3.00	2027.00
Percentage	110%	20270%

However, this satisfaction score may not accurately reflect the general sentiment among customers who have lodged complaints. Since only ten customers have remained with the bank after lodging complaints, this high satisfaction score may be subject to a high degree of variance. It may not generalize to a broader population of customers.

Furthermore, due to the small population size (N=10), the statistical power of any inferences made based on this score is likely to be low. This means the chance of detecting a true difference (if one exists) is reduced, and the likelihood of obtaining false-negative results increases. Therefore, while the satisfaction score could be a relevant metric for assessing customer sentiment, the current dataset does not offer a robust basis for such an analysis.

4.5.3 In conclusion. , while the data suggest a potential link between low satisfaction scores and customer attrition, the limited sample size restricts the conclusiveness of this finding. Further research with a larger sample size of customers who have complained but remained with the bank would provide more robust insights into this relationship.

4.6 Analysis of Rewards System

4.6.1 Question 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.

4.6.2 Analysis. Figure 5 illustrates that there is no clear distinction in the median points earned across the four card types. This observation suggests a low variance in the earned points irrespective of

the card type. Furthermore, the dataset does not provide explicit criteria determining the assignment of different card types to customers, making it challenging to infer potential reasons for these similarities.

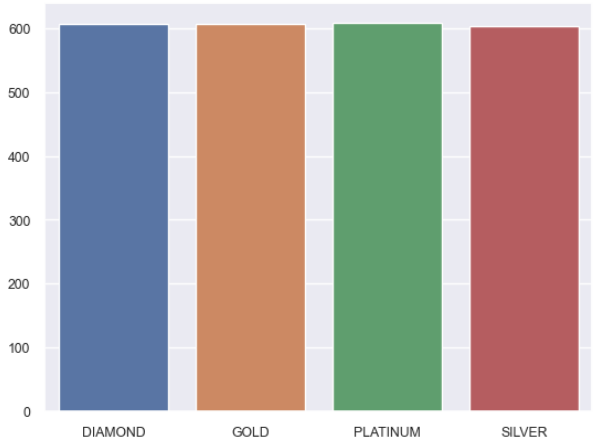

Figure 5: Median points earned by card type

To further explore the distribution and density of each card type in relation to the points earned as illustrated in figure 6, the findings only coincide with the aforementioned conclusion.

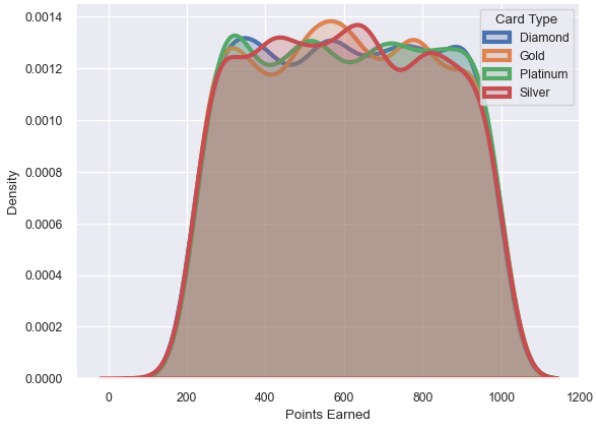

Figure 6: Distribution of points earned across different card types

To formally examine if the similarities observed in the figure translate to statistical significance, we conduct a one-way Analysis of Variance (ANOVA) test. The null hypothesis (H0) is that there is no significant difference in the mean points earned among the four card types. The alternative hypothesis (H1) is that there is a significant difference.

The ANOVA test yields an F-statistic of 0.198 and a p-value of 0.898. Given a significance level (alpha) of 0.05, the p-value is significantly greater than the alpha. Thus, we fail to reject the null hypothesis, suggesting that there is no significant difference in the mean points earned by customers holding different types of cards.

Table 9 further supports this conclusion:

Table 9: ANOVA table for the card types

Source of Variation	SS	df	MS	F
Between Groups	30251.32	3	10083.77	0.20
Within Groups	509023589.82	9976	51024.82	NaN
Total	509053841.14	9979	NaN	NaN

4.6.3 In conclusion. , table 9 reveals that the sum of squares within groups (representing the variability within each card type) is vastly greater than the sum of squares between groups (representing the variability between different card types). This result further supports our conclusion: the type of card does not significantly affect the points a customer earns in the bank's reward system.

5 PREDICTIVE ANALYSIS

5.1 Introduction

5.1.1 Objective. Develop a model to predict whether a customer will complain or not given the historical customer records.

5.1.2 Data description. The initial data exploration and literature review indicate that this problem can be framed as a binary classification task, where we use multiple attributes to predict one of two outcomes. To effectively use these attributes, it's essential to balance the dataset and normalize all data points prior to applying a prediction model. During the descriptive analysis for this report, a strong correlation of 0.996 was observed between "Exited" and "Complain," as discussed in Question 2. This correlation can serve as an extremely accurate predictor. However, considering the real-world context, it is assumed that complaints usually precede a customer's exit. Thus, a prediction model based on the "Exited" attribute wouldn't provide much practical value. Instead, we aim to predict future "complainers," enabling the bank to take preventative measures, thus averting unfavorable outcomes for both the business and the customer.

5.2 Regression

5.2.1 Logistic Regression. Logistic regression presents an optimal model for this study, primarily due to the binary classification problem. The lack of multicollinearity(independent variables) among the attributes and the presence of categorical data makes the model ideal for our dataset. Moreover, the abundance of categorical data, represented as boolean values or discrete categories like country names, aligns well with the logistic regression model's assumptions. Unlike linear regression, which assumes a continuous relationship, logistic regression can handle these categorical inputs effectively.

It's also worth noting that logistic regression is robust to outliers, as it models the outcome's probability in a non-linear manner, reducing the impact of extreme values. Although there might be other potential candidates for predictive analysis, such as GaussianNB, these are beyond the scope of this study and the current expertise

of the researcher. The logistic regression prediction utilized the following equations:

The linear combination of features and weights:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Logistic function (sigmoids function):

$$p = \frac{1}{1 + e^{-z}}$$

Logistic regression:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

5.3 Data preparation

5.3.1 Balancing. The dataset must be balanced to ensure unbiased predictions, meaning that the model doesn't favor attributes appearing more frequently. The distribution of customers who have complained versus those who haven't, as depicted in Figure 7, is unbalanced. The number of non-complainers is four times larger than the number of complainers. Therefore, data augmentation or under-sampling is necessary, resulting in the random removal of Hasen't complained to 2037 records on each. Over-sampling was also considered, as the research into literature Introduction did give insights into libraries such as smote. However, reproducing or multiplying 2037 records to roughly 8000 on customer bank data needed more research into the documentation and methods to maintain data integrity.

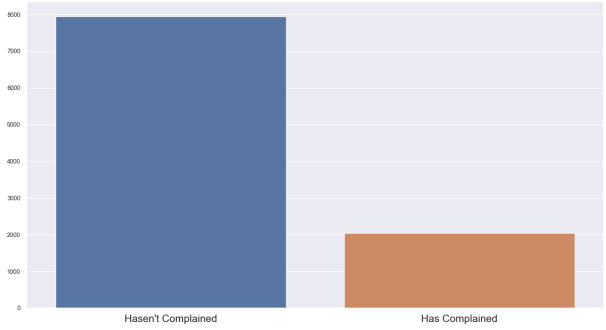

Figure 7: Histogram of number customer complaints

5.3.2 Normalizing data. Given the dataset's non-uniform nature, with attributes like 'balance' varying from zero to six figures and other attributes being boolean (0 and 1), normalization is required. All relevant values are scaled between 0-1 using a Python library that utilizes percentiles. The "ExtraTreesClassifier" Python function was used to rank the importance of each attribute. This function splits the data into decision trees clustered in a forest, with each tree voting for a class. The class with the most votes is deemed the most important. The results, depicted in Figure 8, show that the top six features are 'CreditScore', 'Age', 'Balance', 'NumOfProducts', 'EstimatedSalary', and 'Point Earned'.

5.3.3 Ranking attribute for Importance. The "ExtraTreesClassifier" Python function was used to rank the importance of each attribute(feature). This function splits the data into decision trees clustered in a forest, each tree voting for a class. Essentially ranking each attribute for relational importance with the chosen feature we wish to predict, in this case, "Complain". The results, depicted in Figure 8, show that the top six features are 'CreditScore', 'Age', 'Balance', 'NumOfProducts', 'EstimatedSalary', and 'Point Earned'.

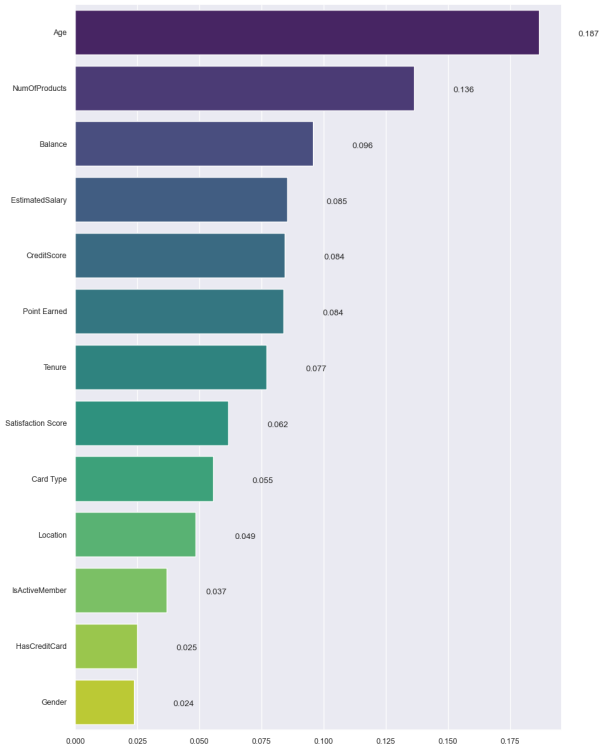

Figure 8: Importance of each feature(attribute)

5.3.4 Training and test data. Splitting the testing and training data allows us to have separate data-set to retain data points to test our model; after all modifications, the data was randomly split was used 80% for training data and 20% for testing data(the same random split was used for all iterations).

5.4 Predictions

The results are seen in table 10:

Table 10: Classification report for the logistic regression model

	precision	recall	f1-score	support
Will Not Complain	0.66	0.72	0.68	400.00
Will Complain	0.70	0.64	0.67	415.00
accuracy	0.68	0.68	0.68	0.68
macro avg	0.68	0.68	0.68	815.00
weighted avg	0.68	0.68	0.68	815.00

In the case of the bank sector, specificity is the priority, as the False positives will not significantly impact of overall business other than perhaps a slight increase in respective customer's benefits; however, on the loss, a current customer will most likely have higher cost tied to it a reviewed in the Introduction. It noted the model performed with an accuracy of 0.68; however, the precision is seen in the classification report table 10 of "will complain" 0.70 is the most relevant measure for this prediction, which is further visualized in the confusion matrix figure 9.

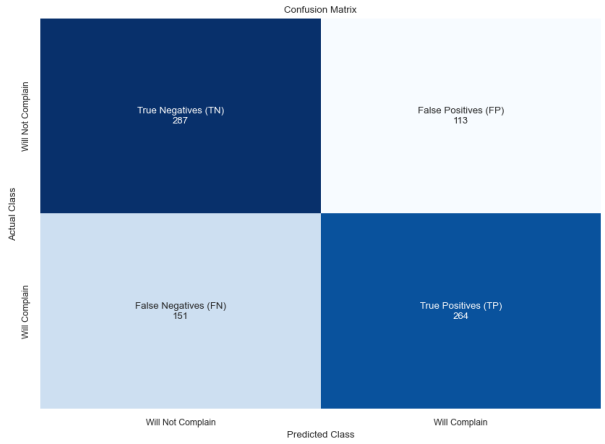

Figure 9: Confustion matrix complaint predictions

The model that proved useful in prediction delivered the highest number of True positives. The best model within the scope of this project. The model was used to predict complaints in the sample set prediction, and results are tabularized in Appendix.

5.5 Feature Engineering and Model Tuning

Feature engineering and model tuning are important aspects of building an optimal machine-learning model. The creation of new features or the transformation of existing ones can significantly impact a model's performance. Similarly, tuning the model's hyperparameters can help achieve the best possible performance. This analysis did do many iterations experimenting with different features; little tuning was done to hyperparameters using defaults for a logistic regression model. Future analyses should explore feature engineering and model tuning for improved performance.

5.6 Conclusion

The logistic regression model provided insights into the factors influencing whether a customer complains. The model achieved a precision of 0.70 and an accuracy of 0.68, indicating its reliability in predicting customer complaints. However, further improvements can be made through feature engineering, model tuning, and testing alternative models. The findings from this analysis can guide the bank in identifying customers who are likely to complain and taking preemptive actions to address their concerns. By doing so, the bank can enhance customer satisfaction and reduce the likelihood of losing valuable customers.

6 DISCUSSION

To begin, we thoroughly examined the 16 attributes in the dataset. This included attributes tied to customer demographics, current status, and dynamic variables that may help predict customer behavior. The research analyzed 9980 unique records. In order to effectively utilize descriptive and predictive statistics, it was necessary to carefully evaluate specific attributes and convert the data types to ensure seamless compatibility.

Potential outliers within the dataset were pinpointed using Z-score and IQR methods. For outlier detection, a detailed analysis was conducted on the "CreditScore", "Age", and "NumOfProducts" attributes. Despite identifying specific data points as outliers, their potential value for subsequent analysis led to their retention. Given that the predictive analysis utilized logistic regression-a model resistant to the impact of the outliers-this decision was justified.

The study emphasized the crucial role of the correlation between customer complaints and attrition. With a striking correlation coefficient of 0.996, the data suggested that customer complaints were a significant precursor to customer exits, accentuating the necessity for enhanced complaint resolution strategies.

A key observation pinpointed discrepancies among customers lodging complaints at the German branch, indicating potential regional variations in customer expectations or perceptions of service warranting a more detailed examination.

Hypothesis tests were conducted, and one key t-test comparing credit scores between two customer groups to identify unique characteristics of customers predisposed to complaints. Despite a negligible mean difference of 0.04%, the test indicated a statistically significant difference, prompting further exploration to understand the contributing factors and implications for customer complaints.

An analysis of customer satisfaction scores showed a high average score of 110% among customers who lodged complaints but opted to stay with the bank. However, due to the limited sample size, this finding's generalizability is restricted, necessitating further research with larger samples for more conclusive insights.

An assessment of the connection between the type of card and the points earned in the bank's reward system revealed that the card type did not significantly influence the reward points, suggesting either the presence of other influential factors or the potential to improve befit package.

The logistic regression model, trained to predict customer complaints, emerged as a tool that could enhance a bank's ability for proactive customer relationship management. The model's performance evaluation using test data demonstrated promising results in differentiating between customers who would and would not complain.

7 SUMMATIVE CONCLUSION

The report underscored the critical role of efficient customer complaint management in retaining customers, particularly within the banking sector. The insights derived from the analysis, which included potential regional variations in customer behavior and the importance of rigorous statistical analysis in deciphering customer data, could provide insights into targeting strategies to boost customer satisfaction and retention. Developing a logistic regression

model capable of predicting customer complaints based on historical data emphasizes the potential of data-driven approaches in understanding and managing customer behavior. Despite limitations in knowledge regarding Machine Learning, the model delivered adequate performance, suggesting a similar or modified approach could be used for banks to address customer complaints, thereby enhancing satisfaction and retention proactively. Furthermore, future research should consider other predictive models and feature selection methods to improve prediction accuracy. Implementing such models in real-world scenarios can yield insights into their effectiveness in customer management strategies. The study pointed out avenues for further research, including an in-depth exploration of the factors influencing customer complaints and satisfaction scores and their impact on customer retention. This study reinforced the necessity for meticulous data exploration and preprocessing to achieve reliable predictive modeling results.

REFERENCES

[1] Viral V. Acharya and Nada Mora. [n. d.]. A Crisis of Banks as Liquidity Providers. https://doi.org/10.2139/ssrn.2022301

[2] Philipp Bagus and David Howden. [n. d.]. Fractional Reserve Free Banking: Some Quibbles. 4 ([n. d.]).

[3] J. Anthony Cookson, Corbin Fox, Javier Gil-Bazo, Juan Felipe Imbet, and Christoph Schiller. [n. d.]. Social Media as a Bank Run Catalyst. https://doi.org/10.2139/ssrn.4422754

[4] Ozden Gur Ali and Umut Arturk. [n. d.]. Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. 41, 17 ([n. d.]), 7889–7903. https://doi.org/10.1016/j.eswa.2014.06.018

[5] Fred Reichheld. [n. d.]. Prescription for Cutting Costs. ([n. d.]).

[6] Frederick F. Reichheld. [n. d.]. Learning from Customer Defections. ([n. d.]). https://hbr.org/1996/03/learning-from-customer-defections Section: Market research.

[7] nations United. [n. d.]. *Ageing*. https://www.un.org/en/global-issues/ageing

	CustomerId	CreditScore	Location	Gender	Age	Tenure	Balance	NumOfProducts	HasCreditCard	IsActiveMember	EstimatedSalary	Exited	Satisfaction Score	Card Type	Point Earned	Predicted Complain
0	15710408	584	Spain	Female	38	3	0.00	2	True	True	4525.40	False	2	GOLD	941	False
1	15598695	834	Germany	Female	68	9	130169.27	2	False	True	93112.20	False	5	GOLD	882	True
2	15649354	754	Spain	Male	35	4	0.00	2	True	True	9658.41	False	1	SILVER	474	False
3	15737556	590	France	Male	43	7	81076.80	2	True	True	182627.25	True	1	DIAMOND	253	True
4	15671610	740	France	Male	36	7	0.00	1	True	True	13177.40	False	5	SILVER	466	False
5	15625092	502	Germany	Female	57	3	101465.31	1	True	False	43568.31	True	5	SILVER	882	True
6	15741032	733	France	Male	48	5	0.00	1	False	True	117830.57	False	1	DIAMOND	674	True
7	15750014	755	Germany	Female	37	0	113865.23	2	True	True	117396.25	False	3	GOLD	589	False
8	15784761	554	Spain	Female	46	7	87603.35	3	False	True	96929.24	True	4	PLATINUM	818	True
9	15768359	534	France	Male	36	4	120037.96	1	True	False	36275.94	False	4	PLATINUM	488	False
10	15805769	656	Spain	Male	33	4	0.00	2	True	False	116706.00	False	4	SILVER	994	False
11	15719508	575	Germany	Male	49	7	121205.15	4	True	True	168080.53	True	2	DIAMOND	227	True
12	15609011	480	Spain	Male	47	8	75408.33	1	True	False	25887.89	True	4	SILVER	556	True
13	15703106	575	France	Male	40	5	0.00	2	True	True	122488.59	False	3	PLATINUM	251	False
14	15626795	672	France	Female	40	3	0.00	1	True	False	113171.61	True	5	GOLD	755	False
15	15773731	758	Spain	Female	35	5	0.00	2	False	False	100365.51	False	1	DIAMOND	833	False
16	15756196	682	France	Male	50	6	121818.84	2	False	True	124151.37	False	1	SILVER	813	True
17	15687903	501	France	Female	29	8	0.00	2	True	False	112664.24	False	5	SILVER	222	False
18	15777599	746	Germany	Male	34	6	141806.00	2	True	True	183494.87	False	3	SILVER	236	False
19	15754577	556	France	Female	51	8	61354.14	1	True	False	198810.65	True	4	GOLD	647	True