**CONTENTS**

[**1.** **BUSINESS UNDERSTANDING** 2](#_Toc160365188)

[1.1 STAKEHOLDERS: 3](#_Toc160365189)

[**2.** **DATA UNDERSTANDING & PREPARATION** 4](#_Toc160365190)

[2.1 DATA DESCRIPTION 4](#_Toc160365191)

[2.2 EXPLORATORY DATA ANALYSIS (EDA) 5](#_Toc160365192)

[**3.** **MODELING** 10](#_Toc160365193)

[3.1 LOGISTIC REGRESSION 10](#_Toc160365194)

[3.2 CHI SQUARE 11](#_Toc160365195)

[3.3 TWO SAMPLED T-TEST 12](#_Toc160365196)

[3.4 N-WAY ANOVA 13](#_Toc160365197)

[**4.** **RESULTS AND EVALUATION** 14](#_Toc160365198)

[**5.** **RECOMMENDATIONS:** 16](#_Toc160365199)

[**6.** **FAITH AND ETHICS IMPLICATIONS:** 17](#_Toc160365200)

[**REFERENCES:** 18](#_Toc160365201)

# **BUSINESS UNDERSTANDING**

Customer churn can have significant financial implications for a bank, as acquiring new customers is generally more expensive than retaining existing ones. Harvard Business Review published a study indicating that acquiring a new customer is anywhere from five to 25 times more expensive than retaining an existing one. This emphasizes the financial significance of customer retention and highlights the cost-effectiveness of retaining customers compared to the substantial expenses incurred in acquiring new ones. The banking sector, specifically in North America, grapples with an annual attrition rate of approximately 11%. This high number forces banks to spend considerable marketing dollars to sign up new customers to keep the customer count flat. The effects of bank customer churn go well beyond losing individual clients. They impact the institution's income sources, profitability, and overall organizational viability. A bank's financial viability depends heavily on its ability to keep a loyal client base. Therefore, customer retention is an essential component of strategic planning.

Considering this information, implementing predictive models for customer churn can help banks to manage and minimize the loss of clients more effectively. These models allow for the early detection of possible churn issues by utilizing historical data and advanced analytics. To properly use these models, businesses must invest in a strong data infrastructure, analytical knowledge, and technological integration. Furthermore, building a culture that values data-driven decision-making and assuring continual model development over time are critical components of an organization's ability to successfully handle customer churn using predictive analytics.

**Business goal:** This will be to reduce Customer Churn by lowering the rate at which clients leave the bank. This is important for maintaining a beneficial client base and ensuring consistent revenue and long-term financial sustainability.

**Data mining goal**: use predictive analytics to find trends and factors contributing to client attrition. The goal is to explore different models and parameters to identify clients most likely to leave based on different variables.

**Success Criteria:** The prediction models' accuracy in detecting possible churners will be used to assess the data mining goal's success. Precision, recall, and model correctness will be the key for performance measures.

## 1.1 STAKEHOLDERS:

**Bank Executives and Management**: This group is concerned with overall business performance, including customer retention and profitability. They are interested in understanding churn rates, identifying reasons for churn, and implementing strategies to reduce it.

**Marketing Department**: This department is responsible for customer acquisition and retention strategies. They analyze churn patterns to improve marketing campaigns, customer targeting, and messaging aimed at retaining customers.

Customer Service Department: Customer service representatives interact directly with customers and may be the first to notice signs of dissatisfaction or intent to churn. They play a crucial role in addressing customer concerns and retaining them.

**Data Analysts and Data Scientists:** This group analyzes customer data to identify churn patterns, predict potential churners, and develop strategies to prevent churn. They use various statistical and machine-learning techniques to derive insights from data.

**Product Development Team:** Understanding customer churn helps product development teams improve existing products or services and develop new ones that better meet customer needs and preferences.

**Investors and Shareholders**: Investors and shareholders are interested in the bank's financial performance, including its ability to retain customers. High churn rates can signal potential problems with the bank's business model or customer satisfaction.

**Customers:** While customers themselves are not typically considered stakeholders in the traditional sense, their churn directly impacts the bank and its stakeholders. Understanding customer needs and addressing their concerns is essential for reducing churn and maintaining a loyal customer base.

# **DATA UNDERSTANDING & PREPARATION**

2.1 DATA DESCRIPTION

The dataset contains 10,000 rows (data points) and 17 columns (features). The data set is displayed below. Each row represents a customer, and each column represents a feature, such as Customer Id, Surname, Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Has Credit Card, Is Active Member, Estimated Salary, Exited, Complain, Satisfaction Score, Card Type, Points Earned.

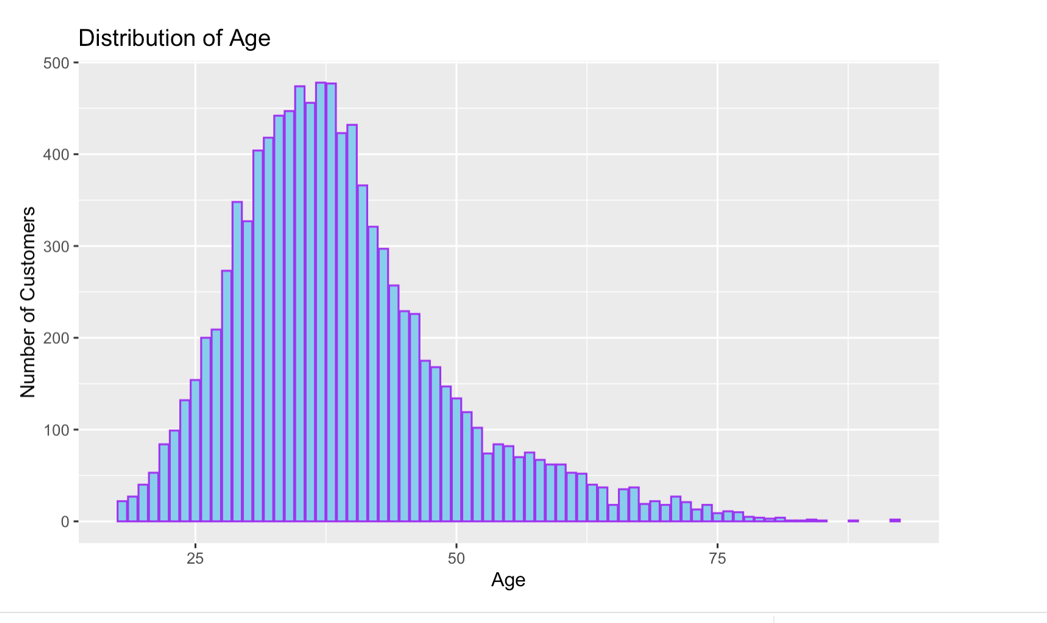
A screenshot of a screen

Description automatically generated  
*Figure 2.1 The Dataset Display Table*

As it is shown in Figure 2.1, there are multiple descriptive futures such as the credit score of the customer, where is the customer located, the customer’s gender , their age, the number of years that the customer has been a client of the bank, the balance of the account, the number of products that a customer has purchased through the bank, whether or not a customer has a credit card; this column is relevant, since people with a credit card are less likely to leave the bank; the estimated salary, people with lower salaries are more likely to leave the bank compared to those with higher salaries, if the customer has a complaint or not, a satisfaction score provided by the customer for their complaint resolution, type of card hold by the customer and the points earned by the customer for using a credit card. For the predictive analysis, the target feature will be the variable Exited, which determines whether the customer has left the bank.

## 2.2 EXPLORATORY DATA ANALYSIS (EDA)

In the first three graphs, we can observe that most of the customers are between 30 and 60 years old, and the majority are men, even though there is also a significant percentage of women. This age group covers the working population and those likely to require banking services such as loans, mortgages, and investments. For the third graph (*Figure 2.2.3*), we used an aggregate function to generate five groups of age 18-30, 31-40, 41-50, 51-60, and 60+, and then we created a bar chart for the age group and the geography of customers. As a result, most of the customers are from France, and they are between 31 and 40 years old.



*Figure 2.2.1 Distribution of Customer’s age*

A graph with a blue and purple squares

Description automatically generated

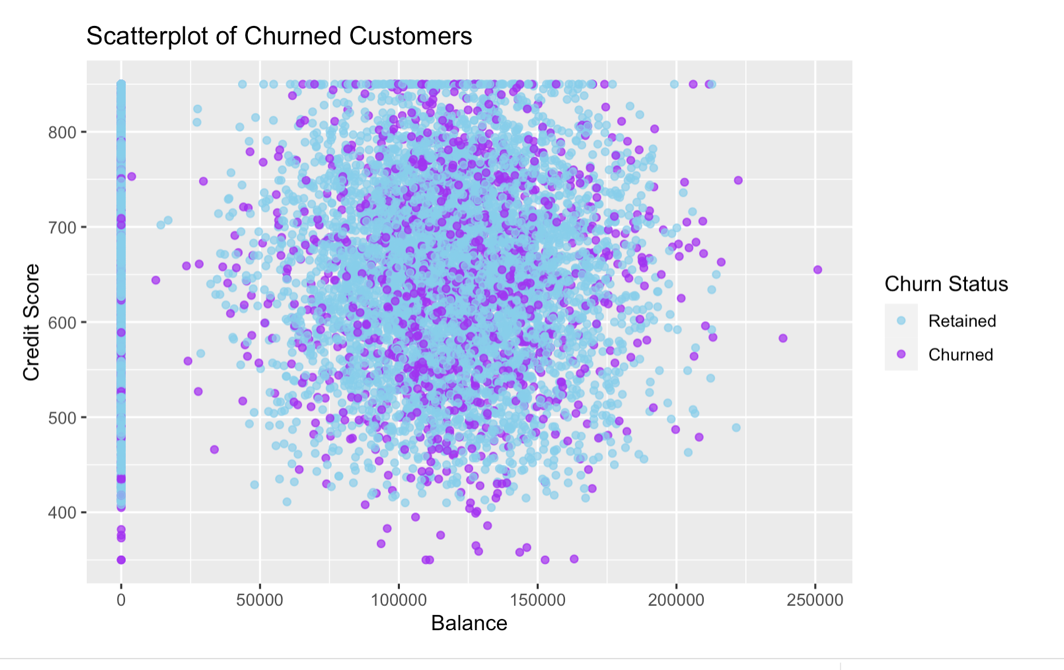
*Figure 2.2.2 Distribution of Customer’s Gender*

A graph of customers by age group and geography

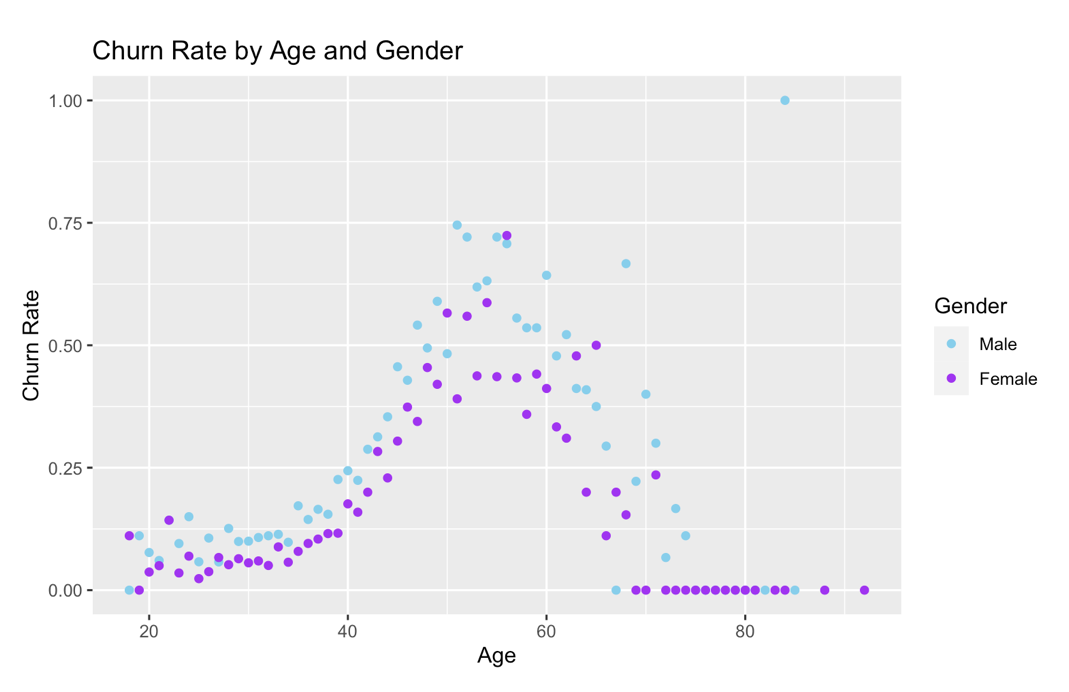
Description automatically generated

*Figure 2.2.3 Number of Customers by Group Age and Geography*

In the following scatter plot (*Figure 2.2.4*), all customers with a credit score of less than 400 are churned, and most of the clients whose balance is more than 200k are also churned. Furthermore, Figure 2.2.5 shows that customers aged between 45 and 70 have a higher tendency to leave the bank, and males have a higher churn rate than females. To be able to create this graph, we used the following expression for ‘‘churn rate’’ : churn\_rate <- aggregate(Exited ~ Age + Gender, data = df, FUN = function(x) sum(x) / length(x))



*Figure 2.2.4 Customer’s Churned Based on Credit Score and Balance*



*Figure 2.2.5 Customer’s Churn Rate by Age and Gender.*

For the next section, it was necessary to mutate the variable ‘‘Exited’’ to ‘‘churned’’ and ‘‘non-churned’’ where 1 was equal to churned, and 0 was equal to non-churned. Then, we created four different data frames with three variables each. 1) Number of products, Churned, Exited Count 2) Satisfaction Score, Churned, Exited Count 3) Complain, Churned, Exit Count 4) Tenure, Churned, Exited count.

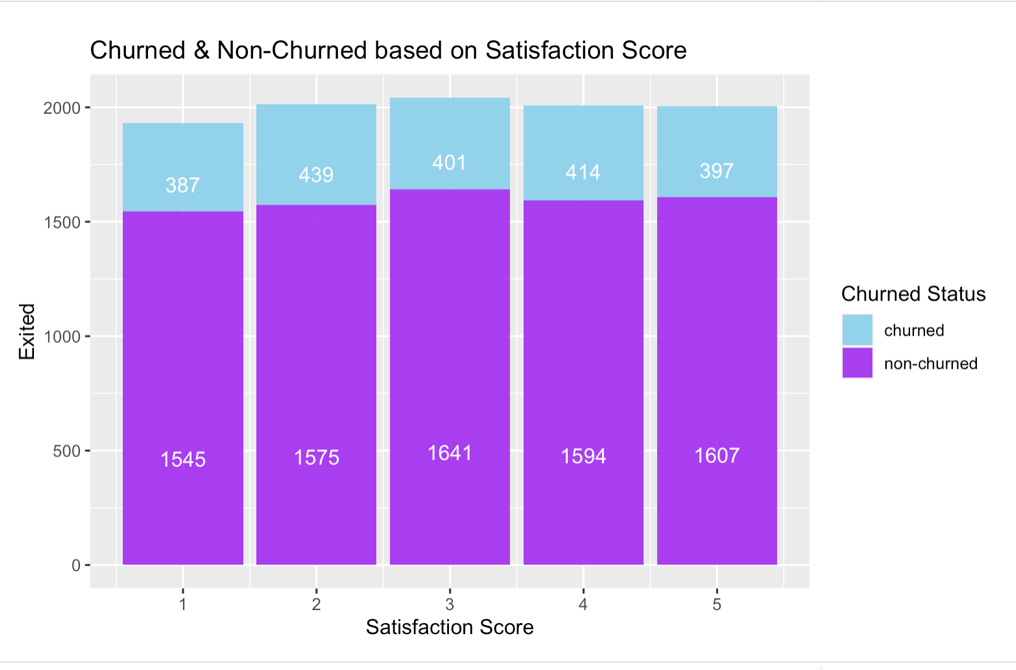
The following bar chart (*Figure 2.2.6)*  represents churned and non-churned customers based on the number of products purchased by the customer. Clients who purchased one or two products were likelier to stay in the bank than those who purchased three products, which exited at 82.8%, while those who purchased four exited utterly.

A graph with purple and blue bars

Description automatically generated

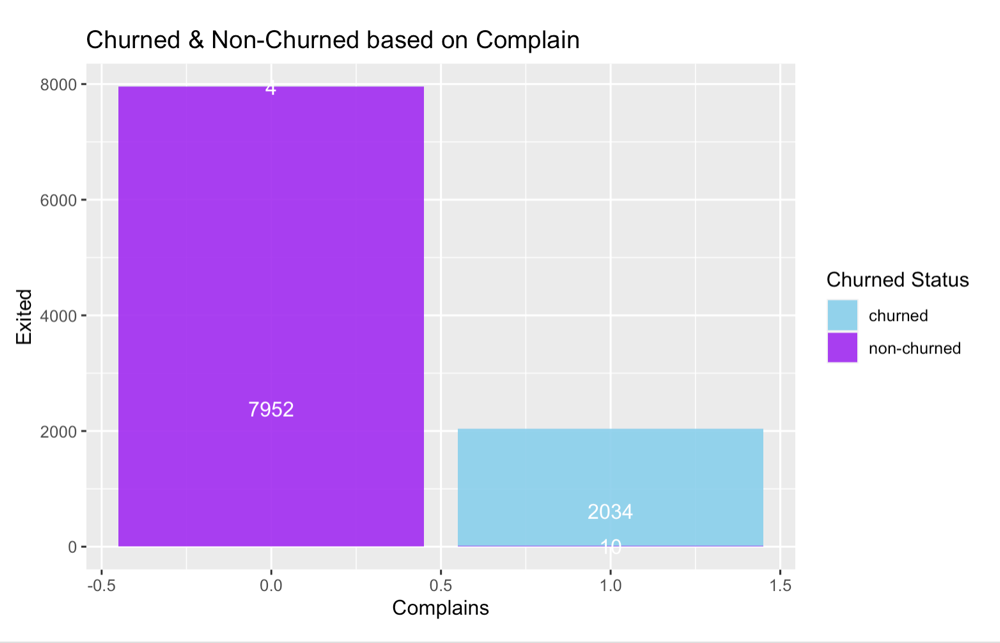
*Figure 2.2.6 Customer’s Churn Based on Number of Purchased Products.*

Figure 2.2.7  represents churned and non-churned customers based on the satisfaction score number. 397 customers who were extremely happy with the replies to their complaints left, but 1547 customers who were unhappy with their experience stayed. This proved that the ranking did not accurately represent the facts.



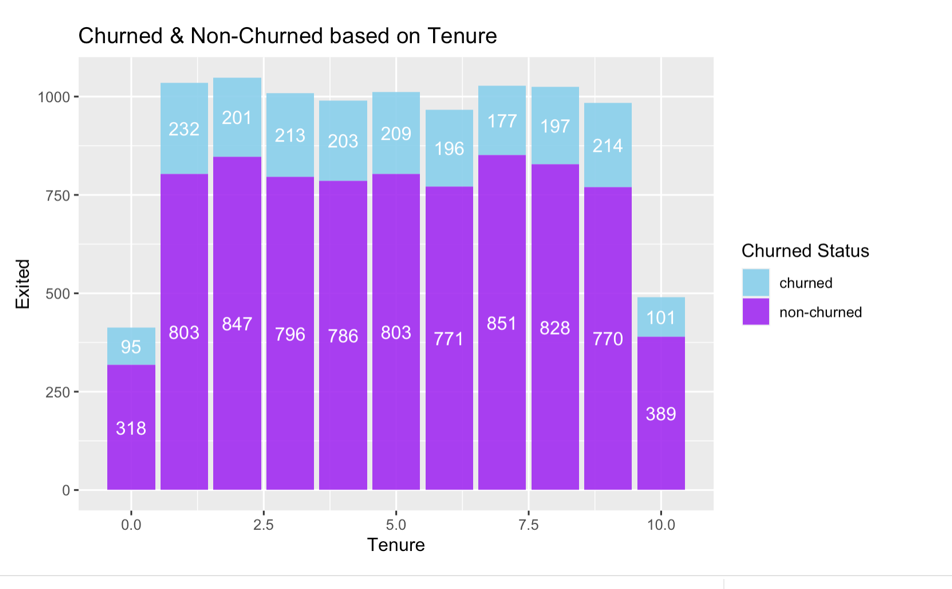
*Figure 2.2.7 Customer’s Churn Based on Satisfaction Score.*

In Figure 2.2.8 only 10 clients who complained didn't exit, and 2034 of the clients who complained exited.



*Figure 2.2.8 Customer’s Churn Based on Complains*

Finally, in Figure 2.2.9, we can observe there is a similar behavior of customers that have been in the bank for 2 to 9 years. Older clients are more loyal and less likely to leave a bank.



*Figure 2.2.9 Customer’s Churn Based on Tenure*

# **MODELING**

## 3.1 LOGISTIC REGRESSION

We created a logistic regression model using the glm() function, which stands for Generalized Linear Models. This is Used for classification where the Target variable is categorical (usually binary) instead of a continuous numeric variable (. In this data, Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Has Credit Card, Is Active Member, Estimated Salary, Satisfaction Score, Card Type, Points Earned are continuous predictors (x) and a binary outcome (y) would be Exited/Complain.

Then we defined the formula for the logistic regression model using the glm() function. The formula follows the pattern Target ~ predictor1 + predictor2 + .... where Target would be the binary outcome variable indicating whether the customer exited the bank/has a complaint, and predictor1, predictor2, etc., would be the other predictor variables.

1. **Logistic Regression to predict whether a customer is likely to exit the bank**

For the first model we used the logistics Regression to predict customer churn (Exited) based on various predictors available in the dataset, while excluding unnecessary or redundant variables. For this we used Exited is the binary outcome variable indicating whether the customer exited the bank. “ . ” indicates that all other variables in the dataset are used as predictors. “ – “ RowNumber - CustomerId - Surname - age\_group specifies the variables that are excluded from the predictors because RowNumber, CustomerId, and Surname are likely to be unique identifiers and not informative for prediction, while age\_group might already captured information about age.

For the Family: binomial(link='logit'), binomial specifies that the outcome variable follows a binomial distribution, suitable for binary outcomes and link='logit' specifies the logistic link function, which is appropriate for binary logistic regression.

1. **Logistic Regression to predict whether a customer is likely to file a complaint**

For the Second model we used the logistics Regression to predict customer Complaints based on various predictors available in the dataset, while excluding unnecessary or redundant variables. For this we used Complaints is the binary outcome variable indicating whether the customer exited the bank. We used the same family and link parameters as above because we are predicting wether a given customer is likely to complain or not,

Other parameters include: weights specified as “Null” because it is an an optional vector of ‘prior weights’ to be used in the fitting process. NA is a function which indicates what should happen when the data contain NAs. For this function it is set to default : na.action.

## 3.2 CHI SQUARE

1. **Chi-Square Test to find the association between Geographical Location and Customer Churn at 95% confidence level**

The Chi-Square Test for Independence is a statistical test used to determine whether there is a significant association between two categorical variables. In this context the Chi-Square Test was used to determine whether there is a significant association between Geographical Location (such as France, Germany, or Spain) and Customer Churn (Exited or Non-Exited).

*Null Hypothesis (H0):* There is no association between Geographical Location and Customer Churn. In other words, the two variables are independent of each other.

*Alternative Hypothesis (H1):* There is an association between Geographical Location and Customer Churn. In other words, the two variables are dependent on each other.

For this test, we had to provide the two categorical variables to test for association. The chisq.test() function calculates expected frequencies assuming that the variables are independent. However, we can adjust this behavior by specifying expected frequencies or probabilities. Adjusting the significance level may be necessary based on the specific requirements of the analysis or the level of significance desired. (chisq.test Function - RDocumentation, n.d.)

However, we adjusted the significance level (alpha) to 0.05. If simulate.p.value is FALSE, the p-value is computed from the asymptotic chi-squared distribution of the test statistic; continuity correction is only used in the 2-by-2 case (if correct is TRUE, the default). Otherwise the p-value is computed for a Monte Carlo test (Hope, 1968) with B replicates.

## 3.3 TWO SAMPLED T-TEST

1. **Two sampled t-test to compare the Mean customer churn of Males and Females at 95% confidence level.**

The Two-Sample T-Test is a statistical test used to compare the means of two independent groups to determine if there is a statistically significant difference between them. In this context, we have used the Two-Sample T-Test to compare the mean customer churn of males and females.

*Null Hypothesis (H0):* Mean Customer Churn for males is equal to Mean Customer Churn for females

*Alternative Hypothesis (Ha):* Mean Customer Churn for males is not equal to Mean Customer Churn for females

For this test, we provided a character string specifying the alternative hypothesis, (t.test Function - RDocumentation, n.d.) this must be one of "two.sided" (default), "greater" or "less". We specified “two.sided” because we are testing the difference in means for males and females. We also used the default to getOption("na.action") function which indicates what should happen when the data contain NAs. If paired is TRUE then both x and y must be specified and they must be the same length. Missing values are silently removed (in pairs if paired is TRUE), hence we decided to proceed with paired as default “FALSE”. We set the confidence level of the interval as 0.95.

## 3.4 N-WAY ANOVA

1. **Test whether age,tenure, credit score, balance, products purchased have an effect on Exited Customers**

The N-Way ANOVA (Analysis of Variance) test is a statistical test used to determine whether there are any statistically significant differences between the means of three or more independent groups. In this context, the N-Way ANOVA test can be used to test whether age, tenure, credit score, balance, and products purchased have an effect on whether customers have exited or not.

*Null Hypothesis (H0):* There is no significant difference in the means of exited customers across different levels of age, tenure, credit score, balance, and products purchased.

*Alternative Hypothesis (Ha):* There is a significant difference in the means of exited customers across different levels of age, tenure, credit score, balance, and products purchased.

In this test, we used “Exited” is the binary outcome variable indicating whether the customer exited the bank. “Age, Tenure, CreditScore, Balance, and NumOfProducts” are the predictor variables. “results” contains the results of the N-Way ANOVA test, including the F-statistic, degrees of freedom, and p-value. The default settings of the “aov()” function are suitable for most applications of the N-Way ANOVA test. To explain further, (Anova Function - RDocumentation, n.d.) Sum of Squares (Sum Sq) represents the total variability in the response variable that is explained by each predictor variable individually and collectively. Mean Square (Mean Sq) is the sum of squares divided by its degrees of freedom and represents the average variability in the response variable explained by each predictor variable. F-value is the ratio of the variability explained by the model to the variability not explained by the model. It is used to test the overall significance of the model.

# **RESULTS AND EVALUATION**

The performance metrics used in logistic regression models are concerned with the probability (P) that an instance will be classified as Customer exited or not. The significance codes (\*\*\*, \*\*, \*) indicate the level of significance for each predictor variable based on their p-values. Lower p-values (e.g., < 0.05) indicate greater significance. From the Logistic model we determined the intercept term (-8.544), it represents the log odds of the base category (Exited) when all predictor variables are zero. We were able to determine that the variable "Complain" is highly significant with a p-value (< 2e-16) close to zero. This suggests that customers who lodged a complaint are strongly associated with exiting the bank. The implications of this finding indicate that addressing customer complaints and improving customer satisfaction could potentially reduce the likelihood of churn. It also highlights the importance of a robust customer service strategy and effective complaint resolution processes to retain customers.

Although not as extreme as the Complain variable, Age also has a very low p-value (0.000118), indicating high significance. We can say that age plays a significant role in predicting customer churn. The variable IsActiveMember (p-value: 0.005607) has a p-value slightly higher than 0.001 but still falls within the range of moderate significance. This implies that customers who are active members are associated with a lower likelihood of churn compared to inactive members. The implications may include, Encouraging and incentivizing customers to actively engage with the bank's services and offerings to improve customer retention.

In the logistic regression model provided to predict the likelihood of customer complaints, the significance of variables is again determined by the associated p-values. The variable "Exited" is highly significant with a very low p-value. This indicates that customer churn (whether a customer exited the bank) is strongly associated with the likelihood of complaints.

In The performance of two sample test we found that there is a statistically significant difference in customer churn between males and females. The null hypothesis (H0) is rejected in favor of the alternative hypothesis (Ha), indicating that the mean customer churn for males is not equal to the mean customer churn for females. Based on the results, The calculated t-value is 10.536, which represents the difference between the mean customer churn for males and females in units of standard error. A higher absolute t-value indicates a stronger evidence against the null hypothesis. The 95% confidence interval for the difference in means between female and male customer churn is (0.06997802, 0.10196769). This interval provides a range of plausible values for the true difference in mean churn rates. Finally, we can observe that The p-value is less than 2.2e-16, which is extremely small. This indicates strong evidence against the null hypothesis. Therefore, The differences in churn rates may also reflect varying preferences, satisfaction levels, or experiences with the product or service offered.

Based on the results of chi-square test to determine the association between geographical location and customer churn. The null hypothesis (H0) suggests no association, while the alternative hypothesis (H1) suggests an association. With a very low p-value (< 2.2e-16), we have concluded that there is indeed an association between geographical location and customer churn. The expected frequencies represent the frequencies that would be expected in each category (churned or not churned) . The observed frequencies represent the actual frequencies observed in the data. From the results we can see that in the "France" category, the expected frequency of customers who churned (1) is 1021.8532, but the observed frequency is 811. This indicates that there are fewer churned customers in France than would be expected. X-squared values represents the chi-square statistic, in this test (X-squared) is 300.63 with 2 degrees of freedom which measures the extent of the association between the two categorical variables (geographical location and customer churn). Higher values indicate stronger associations.

Based on the results of ANOVA test where we test the impact of different predictor variables (Age, Tenure, CreditScore, Balance, NumOfProducts) on customer churn, we can reject Null hypothesis because age, credit score, and balance have significant impacts on customer churn, while tenure and the number of products purchased do not have significant impact. We also found that The predictor variable "Balance" has the highest Mean Square value (19.85), indicating that it explains the largest amount of variability in the response variable compared to the other predictor variables. The predictor variable "Age" has the highest F-value (898.271), indicating that it has the greatest overall significance in explaining the variability in the response variable among all the predictor variables. These variables have the highest explanatory power on the response variable, suggesting that they are important factors influencing the outcome being measured. For example, in a financial context, customers with higher balances or older customers might exhibit different behaviors such as higher spending or lower churn rates. The predictor variable "CreditScore" is also statistically significant, with a moderately low p-value (0.0071) and a moderate F-value (7.251). This suggests that creditworthiness is also an important factor affecting the outcome being measured, such as loan defaults or customer attrition.

# **RECOMMENDATIONS:**

* According to (Hult, 2023) the following principles can effectively address and resolve customer complaints, ultimately leading to reduced churn and improved overall customer satisfaction.
* Appreciating complaints as constructive criticism can lead to stronger customer loyalty and a more competitive brand.
* Allocate resources and prioritize complaint resolution to ensure a positive outcome for the customer.
* Establish accessible channels for complaints and ensure that customers feel heard and valued throughout the resolution process.
* Understand that it is often better to have customers complain than not complain and simply leave.
* Based on the results of Logistic Regression, we have determined that isActiveMember is statistically significant to customer churn. To address this, the Bank can Implement strategies to enhance the perceived value and utility of membership benefits to encourage customer activity.
* Tailoring marketing and retention strategies based on different age groups like offering age-specific promotions or services to retain older customers who may have different needs and preferences. For example, Gen X consumers prioritize financial stability and long-term planning. A survey by Fiserv revealed that 73% of Gen Xers believe that financial stability is more important than material wealth (Fiserv Consumer Trends Survey).
* Understanding the life stage of customers to anticipate potential churn triggers. For instance, younger customers might be more sensitive to changes in financial circumstances, while older customers might be more concerned about retirement planning.
* Understanding the factors contributing to higher churn rates among a particular gender can help in designing targeted interventions to reduce churn and improve customer loyalty. For example, there can be differences in financial behaviors between genders, such as risk tolerance, investment preferences, and spending habits.
* T test results suggests that Businesses can segment their customer base by gender and customize their communications, promotions, and support services to address the unique needs of each segment, potentially leading to improved customer satisfaction and retention.
* Based on Chi square: Businesses can tailor their strategies based on geographical regions to better address churn and improve customer retention. For example, Since the expected churn is lesser than observed churn in France, Germany and Spain, the bank can implement region specific goals .

# **FAITH AND ETHICS IMPLICATIONS:**

* We must ensure that customers are informed about the data collection and analysis processes. Also, explicit consent must be obtained, especially regarding personal data.
* Given the sensitivity of financial information, it is necessary to implement strong security measures to protect client data from unauthorized access or compromise.
* We must understand the customer's needs, values, and beliefs and regularly treat them with respect and dignity. This entails being more conscious of our biases and preconceptions so that we do not unintentionally project our own beliefs onto them. For example, cultural and language distinctions must be addressed with respect, encouraging to explore effective modes of communication that transcend any differences in background or experiences among participants.
* As data analysts we must avoid discrimination or bias in the treatment of different customer groups, ensuring that the model is fair across demographic and socioeconomic categories. Furthermore, we need to implement strategies to identify and mitigate biases in the data and algorithms to prevent unfair treatment of certain customer segments.
* We must regularly validate and update the models to ensure they remain accurate and relevant. Inaccurate predictions may lead to unwarranted consequences for customers.
* We must consider the broader societal implications of predicting and acting on customer churn. Strive to create positive outcomes and avoid negative consequences.
* To avoid legal and ethical issues, it is crucial to adhere to relevant data protection and privacy regulations, such as GDPR, HIPAA, or other regional standards.
* As Christians, we emphasize stewardship and responsibility for the well-being of people so we may prioritize harmony. We must ensure information is used ethically to avoid harm to individuals or communities. In Philippians 2:4 ‘‘Everyone should look out not only for his interests but also for the interests of others’’, and 1 Corinthians 4:2, "Now it is required that those who have been given a trust must prove faithful."

# **REFERENCES:**

Hult, G. T. M. (2023, March 14). *8 Best Practices for Creating a Compelling Customer Experience*. Harvard Business Review. <https://hbr.org/2023/03/8-best-practices-for-creating-a-compelling-customer-experience>

*More Alike Than Different? Millennials, Baby Boomers and Financial Services*. (n.d.). <https://fiserv.com/en/about-fiserv/the-point/more-alike-than-different-millennials-baby-boomers-and-financial-services.html>

Hope, A. C. A. (1968). *A simplified Monte Carlo significance test procedure*. Journal of the Royal Statistical Society Series B, 30, 582--598.

Anova function *- RDocumentation*. (n.d.). <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/anova>

*t.test function - RDocumentation*. (n.d.). <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test>

*chisq.test function - RDocumentation*. (n.d.). <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/chisq.test>