# BA 723 – Business Analytics Capstone

## Documentation

# Mitigating Customer Churn: Leveraging Historical Data Analysis and Predictive Modelling to Formulate Effective Customer Retention Strategies

**August 15, 2024**

**Alester Joshua D Costa**

**301321745**

## Executive Summary

In 2020, a Statista market research study reported a 25% churn rate in the financial industry of the United States, the joint-highest churn rate across all sectors. Nova Apex Bank, a prominent American financial institution headquartered in Seattle, Washington, has similarly experienced a 20% churn rate in European operations, constituting France, Germany, and Spain. If not addressed, customer attrition could lead to significant revenue losses, increased costs for acquiring new customers and potential reputational damage to Nova Apex Bank. This project aims to decrease customer churn at Nova Apex Bank through comprehensive historical data analysis and predictive modeling. By identifying the primary factors contributing to customers ceasing to do business with Nova Apex Bank, the company can develop targeted retention strategies to enhance customer loyalty and retention. The analysis, conducted using Python and visualized through a Tableau dashboard, provides clear insights into customer behavior. Among the predictive models tested, the random forest model emerged as the most reliable, with a ROC-AUC score of 0.79, an accuracy rate of 0.76 and an F1 score of 0.57. The results highlight that age, the number of products held and account balances are the key drivers of customer churn at Nova Apex Bank. Based on these insights, specific retention strategies have been proposed to decrease churn at Nova Apex Bank and strengthen the company's customer base.

# Table of Contents

# 1. Introduction

## 1.1. Background

In the global economy, banking plays a critical role by carrying out a variety of essential functions that foster growth, stability and financial inclusion (Utah Community Credit Union, 2024, The Backbone section, para. 1). Lake and Strohm (2024) highlight that any banking firm, whether in a brick-and-mortar or a digital format, maintains the flow of money between people and organizations (para. 10). As the principal supplier of credit, banking firms provide money for individuals for numerous reasons, from purchasing cars to homes, as well as enterprises to buy equipment, enlarge their operations and meet their payrolls (Hall, 2023, para. 9). The various financial services offered by banks include checking accounts, savings accounts, credit cards, loans and many more services (Morawski, 2023, para. 5).

Nova Apex Bank, a leading American financial institution within the ever-growing banking industry, offers its customers a wide range of banking services. With a mission that strongly focuses on a client-centric approach, Nova Apex Bank has become a trusted partner for customers due to its exceptional products and services. First established in Seattle, Washington, Nova Apex Bank has expanded its services to the European countries of France, Germany and Spain, offering an extensive portfolio of financial services, such as checking accounts, savings accounts, credit cards and personal loans. Through its commitment to innovation and seamless digital incorporation, Nova Apex Bank has positioned itself as a forward-thinking leader within the financial sector.


*Figure 1: Logo of Nova Apex Bank*

### 1.2. Problem Statement

With the banking industry forming a vital part of the economies worldwide, customer retention is crucial for preserving any banking enterprise's growth and financial stability. According to Rodgers (2023), customer retention is the capacity of banking companies to retain their existing customers over time (FAQs About Banking Customer Retention Strategies section, para. 2). An effective customer retention strategy is centered on creating a solid relationship with customers, offering exceptional services and providing value for individuals according to their behaviors and preferences (Glassbox, 2023, para. 4). Nonetheless, the levels at which an organization loses its customers, known as churn rate, can be alarming to any banking firm, including Nova Apex Bank (Investopedia, 2024, para. 1). In 2020, the financial industry, including banking companies, displayed a customer churn rate of 25%, representing a joint-high (the other being the cable industry) level of customers exiting a company within any industry in the United States. Conversely, the big-box electronics industry exhibited a churn rate of only 11%, a substantially lower figure than banking firms in the United States (Statista, 2022).

Despite its strong market presence, Nova Apex Bank has faced the obstacle of customer churn within its European operations, including France, Germany and Spain. The diverse customer base in Europe presents different challenges when contemplating why customers quit their association with Nova Apex Bank. If not tackled, customer churn may lead to several consequences for Nova Apex Bank, such as revenue loss, increased cost of acquiring new customers compared to existing customers and a damaged corporate reputation (Mihup, 2024, para. 6). Addressing the causes of customers ceasing to do business with Nova Apex Bank necessitates a deeper investigation of customer behavior to reduce churn and foster long-term customer loyalty.

### 1.3. Objectives and Measurement

This project aims to utilize historical customer data of Nova Apex Bank to understand customer behaviors concerning churn at the company. Based on the historical data analysis carried out through descriptive analytics, predictive models will be developed to forecast the likelihood of customers quitting their association with Nova Apex Bank using essential predictors, with the best prediction model being chosen. By analyzing key customer metrics and selecting the optimal predictive model, the overall goal is to discover the factors contributing to customer churn at Nova Apex Bank and formulate retention strategies to keep customers. The key objectives to be addressed in this project are stated by the following points:

1. Develop and implement an interactive customer churn analysis dashboard that effectively aggregates and visualizes customer behavior from the historical customer data of Nova Apex Bank through historical data analysis

2. Build multiple predictive models by leveraging critical attributes from the historical customer data to forecast the likelihood of customer churn at Nova Apex Bank

3. Determine the predictive model that generates the highest reliability in predicting customer churn at Nova Apex Bank using different evaluation metrics

4. Extract actionable insights from the most reliable predictive model, supported by the historical data analysis, to identify key factors contributing to customer churn at Nova Apex Bank and develop targeted retention strategies for customers at Nova Apex Bank

To ensure a simplistic yet effective analytical process, the performance of the predictive models will be appraised using three key components: Receiver Operating Characteristic - Area under the Curve (ROC-AUC), accuracy and F1-Score. Using these metrics, the most reliable model that predicts customer churn at Nova Apex Bank can be ascertained.

### 1.4. Assumptions and Limitations

As part of this project, it is assumed that the historical customer data of Nova Apex Bank, due to the lack of information on the time period, reflects the company's customer base in 2023 and accurately represents customer behaviors and interactions. Moreover, it is also deduced that the available data is comprehensive and captures all appropriate information that affects customer churn.

From the perspective of the project limitations for forecasting customer churn at Nova Apex Bank, it can be emphasized that the predictive models may not consider unforeseen changes in customer behaviors or market conditions. Additionally, the evaluation metrics of the predictive models are constrained by the quality of the available customer data.

# 2. Data Source

## 2.1. Dataset Introduction

Within this project centered on customer churn prediction at Nova Apex Bank, the chosen dataset has been obtained from Kaggle, a popular platform for datasets and data science competitions. This dataset constitutes various customer attributes relevant to churns, such as geography, gender, credit score, tenure and many more components (Zangari, 2024). Utilizing the available historical customer data of Nova Apex Bank offers a strong foundation for analyzing customer behavior and building several predictive models through descriptive analytics and predictive analytics, respectively.

## 2.2. Data Dictionary

An important point of consideration is that before the data pre-processing and feature engineering stages, the original dataset consisted of 14 variables. However, after carrying out the data pre-processing facet, the dataset comprises 13 columns and is further increased to 20 columns upon the conclusion of the feature engineering process. To ensure an easier understanding of the attributes present in the historical customer dataset of Nova Apex Bank, the data dictionary describes the dataset produced specifically after the data pre-processing phase of this project.

| Variable | Description | Data Type | Values |
|---|---|---|---|
| *Gender* | Genders of customers | Categorical | • 'Female': Female customers<br>• 'Male': Male customers |
| *Age* | Ages of customers, in years | Numerical | Customers aged between 18 and 92 years |
| *CategorizedAge* | Categorized ages of customers based on an assigned range; newly added variable | Categorical | • 'Young Adults': Customers aged between 18 years and 24 years<br>• 'Early Career Adults': Customers aged between 25 years and 34 years<br>• 'Mid-Career Adults': Customers aged between 35 years and 44 years<br>• 'Late Career Adults': Customers aged between 45 years and 54 years<br>• 'Near-Retirement Adults': Customers aged between 55 years and 64 years |

| Variable | Description | Data Type | Values |
|---|---|---|---|
| | | | • 'Retired Adults': Customers aged 65 years and above |
| *Geography* | Country where the customers are in | Categorical | • 'France': Customers are in France<br>• 'Germany': Customers are in Germany<br>• 'Spain': Customers are in Spain |
| *Balance* | Bank balances of customers, in Euros (€) | Numerical | Customers' bank balance between €0 and €250,898.09 |
| *EstimatedSalary* | Estimated salary of customers, in Euros (€) | Numerical | Customers' estimated salary between €11.58 and €199,992.48 |
| *CreditScore* | Credit score of customers using the Fair Isaac Corporation (FICO) score | Numerical | Customers' credit score between 300 to 850 using a FICO score |
| *CategorizedCreditScore* | Categorized credit score of customers based on an assigned range using the Fair Isaac Corporation (FICO) score; newly added variable | Categorical | • 'Poor': Customers' credit score below 580<br>• 'Fair': Customers' credit score between 580 and 669<br>• 'Good': Customers' credit score between 670 and 739 |

| Variable | Description | Data Type | Values |
|---|---|---|---|
| | | | • 'Very Good': Customers' credit score between 740 and 799<br>• 'Exceptional': Customers' credit score above 800 |
| *NumOfProducts* | Number of products possessed by customers with Nova Apex Bank | Categorical | • 1: Customers hold 1 product<br>• 2: Customers hold 2 products<br>• 3: Customers hold 3 products<br>• 4: Customers hold 4 products |
| *HasCrCard* | Holding of a credit card by customers with Nova Apex Bank | Categorical | • 'No': Customers do not hold a credit card<br>• 'Yes': Customers hold a credit card |
| *Tenure* | Number of years customers have been with Nova Apex Bank | Numerical | Customers' tenure between 0 and 10 years |
| *IsActiveMember* | Activeness of customers as members of Nova Apex Bank | Categorical | • 'No': Customers are not an active member of Nova Apex Bank<br>• 'Yes': Customers are an active member of Nova Apex Bank |

| Variable | Description | Data Type | Values |
|---|---|---|---|
| *Exited* | Customers have churned at Nova Apex Bank | Categorical | • 'No': Customers have not churned<br>• 'Yes': Customers have churned |

# 3. Data Pre-processing

After identifying the appropriate dataset, the historical customer data of Nova Apex Bank should be properly prepared by accounting for missing values, rectifying errors and converting variables into the relevant format. This will ensure that the historical data analysis and prediction process will lead to precise and dependable outcomes due to the refined quality and integrity of the data.

The data pre-processing component of this project has been carried out in Python using various techniques. These methods ensure the necessary data quality before investigating historical customer data and predicting customer churn at Nova Apex Bank.

## 3.1. Identification of Missing Values

To evaluate the completeness of the historical data concerning customer churn at Nova Apex Bank, any potential missing values present within the dataset have been investigated. In this scenario, no missing data is present within the historical customer data of Nova Apex Bank without any errors, as stated in Figure 2.

```
# Identify any missing values in the dataset
churn_df.isna().sum()
```

| | 0 |
|---|---|
| RowNumber | 0 |
| CustomerId | 0 |
| Surname | 0 |
| CreditScore | 0 |
| Geography | 0 |
| Gender | 0 |
| Age | 0 |
| Tenure | 0 |
| Balance | 0 |
| NumOfProducts | 0 |
| HasCrCard | 0 |
| IsActiveMember | 0 |
| EstimatedSalary | 0 |
| Exited | 0 |

dtype: int64

*Figure 2: Identification of Missing Values*

### 3.2. Detection of Outliers

While the historical customer data of Nova Apex Bank does not contain missing data, it is critical to grasp any possible outliers present in the data to ensure the integrity of the overall analysis. An essential point of consideration is that outliers have been identified before proceeding to the following stages of data preprocessing. This process focuses solely on numerical variables, as identifying outliers depends on measuring the distance from a central value, which cannot be done with non-numerical data.

From the computed descriptive statistics (Figure 3) and developed boxplots (Figure 4) for each numerical variable of the historical customer data of Nova Apex Bank, it has been revealed that several variables, including *CreditScore*, *Age*, *Tenure* and *NumOfProducts,* demonstrate a narrow interquartile range (IQR) due to the presence of outliers towards the end of each distribution. Nevertheless, the *Balance* variable displays a wider interquartile range (IQR); this represents substantial variability among customer balances. Numerous customers have no balances, with notable outliers at the higher end highlighting that some customers maintain large account balances. Similarly, the variable *EstimatedSalary* has a wider interquartile range, which describes a wide range of customer salaries.

```
# Generate descriptive statistics to identify the distribution, central tendency and spread of numerical variables
churn_df[numerical_variables].describe()
```

| | CreditScore | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|
| count | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.00000 | 10000.000000 | 10000.000000 | 10000.000000 |
| mean | 650.528800 | 38.921800 | 5.012800 | 76485.889288 | 1.530200 | 0.70550 | 0.515100 | 100090.239881 | 0.203700 |
| std | 96.653299 | 10.487806 | 2.892174 | 62397.405202 | 0.581654 | 0.45584 | 0.499797 | 57510.492818 | 0.402769 |
| min | 350.000000 | 18.000000 | 0.000000 | 0.000000 | 1.000000 | 0.00000 | 0.000000 | 11.580000 | 0.000000 |
| 25% | 584.000000 | 32.000000 | 3.000000 | 0.000000 | 1.000000 | 0.00000 | 0.000000 | 51002.110000 | 0.000000 |
| 50% | 652.000000 | 37.000000 | 5.000000 | 97198.540000 | 1.000000 | 1.00000 | 1.000000 | 100193.915000 | 0.000000 |
| 75% | 718.000000 | 44.000000 | 7.000000 | 127644.240000 | 2.000000 | 1.00000 | 1.000000 | 149388.247500 | 0.000000 |
| max | 850.000000 | 92.000000 | 10.000000 | 250898.090000 | 4.000000 | 1.00000 | 1.000000 | 199992.480000 | 1.000000 |

*Figure 3: Descriptive Statistics of Numerical Variables*

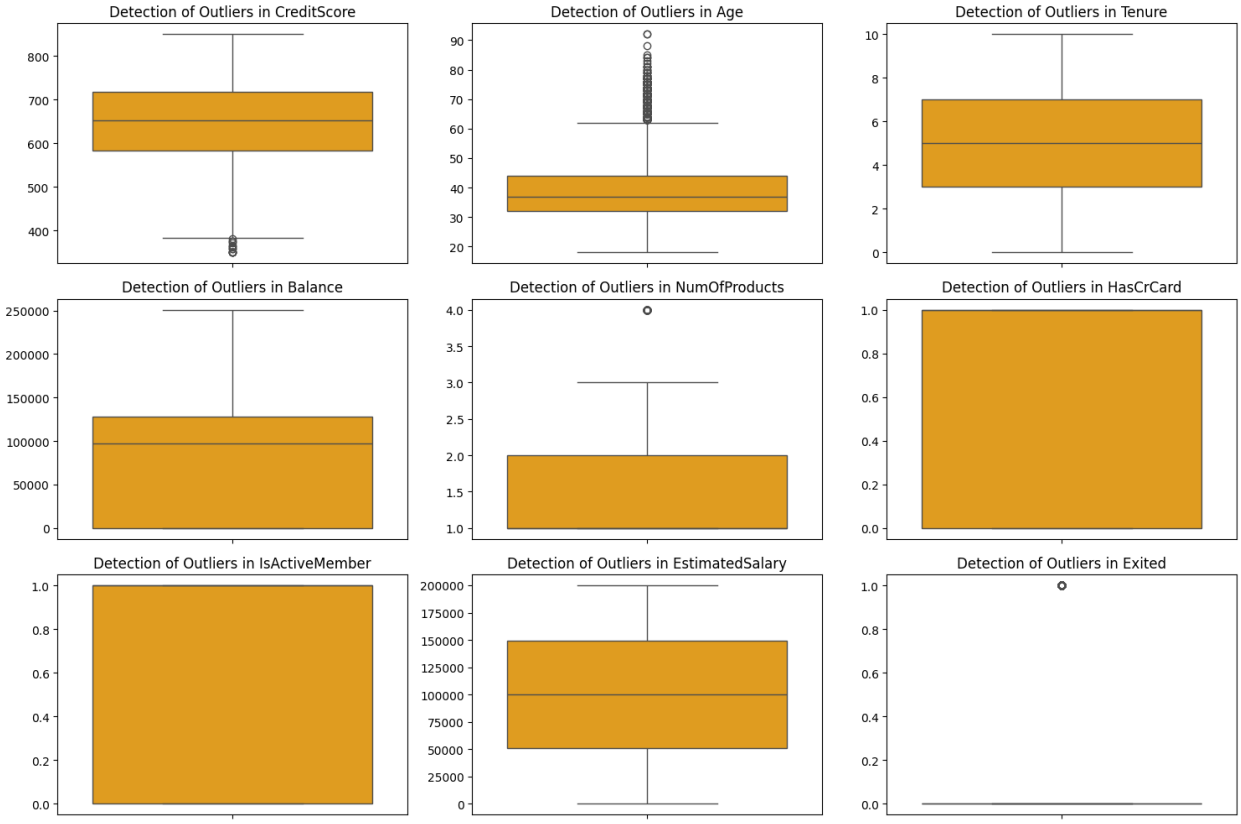


*Figure 4: Box Plots of Numerical Variables*

### 3.3. Correlation Analysis between Numerical Variables

A correlation matrix in the form of a correlation heatmap has been constructed to identify the correlation between all numerical variables present in historical customer data of Nova Apex Bank. By demonstrating relationships between numerical features, it enables acertaining vital drivers of churn, such as tenure, number of products and more.

Like discovering outliers in the dataset, the correlation analysis has been carried out before the later data pre-processing stages because it only works with numerical variables. In this context, a few variables, constituting *HasCrCard*, *IsActiveMember* and *Exited*, are assumed to be numerical variables for the correlation analysis and have been transformed into categorical variables at an upcoming stage.

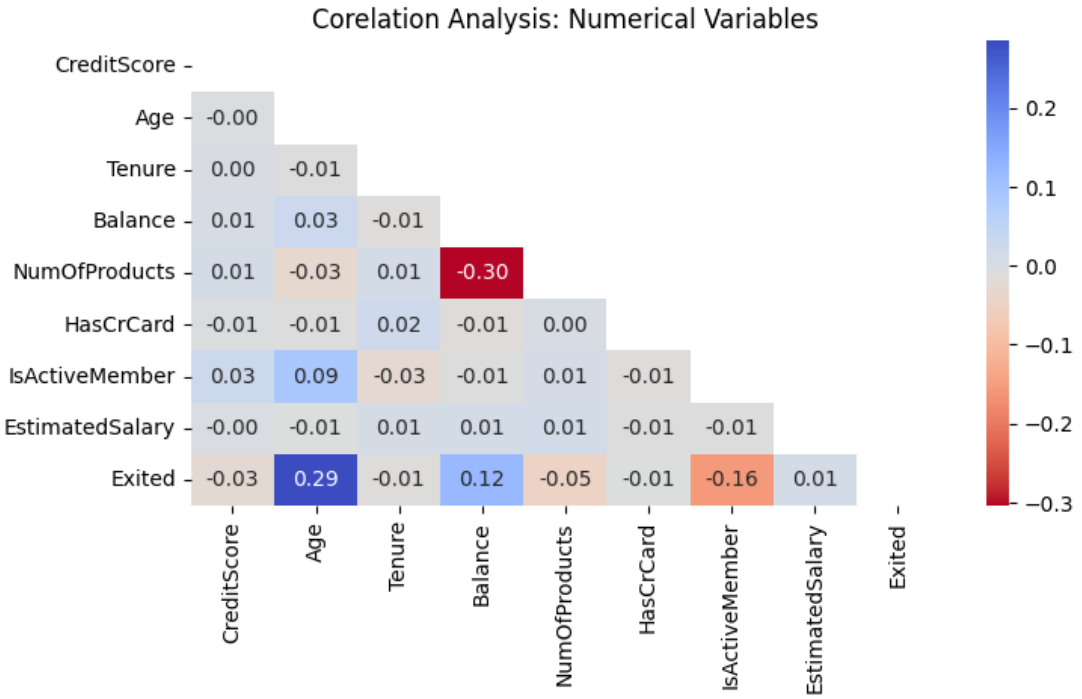


*Figure 5: Correlation Analysis between Numerical Variables*

From the developed correlation heatmap (Figure 5) concerning the identification of correlations between numerical variables present in the historical customer data of Nova Apex Bank, various inferences can be made, as explained by the following points:

- **Age and churn (0.29):** At a 0.29 correlation coefficient, a weak positive correlation exists between the customer's age (*Age*) and churn (*Exited*), hinting that young-aged customers are less likely to churn as opposed to old-aged customers.

- **Account balance and number of products (0.30):** With a -0.30 correlation coefficient, there is a weak negative correlation between the customer's account balance (*Balance*) and the number of products held by them (*NumOfProducts)*. In this regard, it can be deduced that consumers with more products have lower account balances.

- **Account balance and churn (0.12):** Due to a 0.12 correlation coefficient, there is a weak positive correlation between the customer's account balance (*Balance)* and churn (*Exited*). In this context, customers with a larger account balance are much likelier to quit their association with Nova Apex Bank.

- **Activeness as a member and churn (-0.16):** At a -0.16 correlation coefficient, a weak negative correlation exists between customers' being active members (*IsActiveMember*) and churn (*Exited*). From this information, it can be suggested that customers who are not highly engaged are more likely to cease doing business with Nova Apex Bank.

Unlike the other correlations between variables that have been described above, the other correlations are closer to 0, demonstrating that there is no correlation between them. Nonetheless, the probability of non-linear relationships between variables not scored by the correlation coefficient implies that the lack of correlation cannot be ruled out.

### 3.4. Categorization of Relevant Variables

Within the historical customer data of Vova Apex Bank, it has been discovered that two variables, comprising *CreditScore* and *Age*, could be individually grouped to effectively understand the classification of customers' credit score and their ages, respectively. The categorization of *CreditScore* and *Age* simplifies historical data analysis for identifying patterns and refining the performances of predictive modeling by decreasing inconsistencies and focusing on key segments of customer behaviors related to churn.

- **Credit Score of Customers:** From the perspective of customers' credit score (*CreditScore*), each customer's credit score has been classified based on a FICO score rating, a popularly used element within the banking industry. Developed by the Fair Isaac Corporation (FICO), an analytical software company, the firm uses information in each customer's credit reports to compute their credit scores. Eventually, these scores are then utilized by lenders, in this scenario Nova Apex Bank, to evaluate its customers' creditworthiness, and establish whether to accept customers' applications for loans, credit cards and other borrowings (Hayes, 2024, para. 3; Lake, 2023, para. 2). The table below provide vital information about each category of customers' FICO credit score ranges, varying from poor to exceptional (Fair Isaac Corporation, 2024, FICO Scores by Percent of Scorable Population section):

| FICO Score Ranges | Classification | Description |
|---|---|---|
| Below 580 | Poor | A customer has a well below-average credit score, indicating to lenders that a customer is a risky borrower |
| 580 to 669 | Fair | A customer has a below-average credit score, highlighting that many lenders will approve loans |
| 670 to 739 | Good | A customer has a near or slightly above average credit score, with most lenders considering this a good score |

| FICO Score Ranges | Classification | Description |
|---|---|---|
| 740 to 799 | Very Good | A customer has an above-average credit score, indicating to lenders that a customer is a very reliable borrower |
| Above 800 | Exceptional | A customer has a well above-average credit score, indicating to lenders that a customer is an exceptional borrower |

- **Ages of Customers:** Another variable grouped to ensure a practical understanding of customer churn is the age of customers at Nova Apex Bank. Consumers at Nova Apex Bank are between 18 and 92 years old, and this variable has been grouped from young adults to retired adults, as highlighted by the table below:

| Age Ranges | Classification | Description |
|---|---|---|
| 18 to 24 | Young Adults | Customers between the age group of 18 to 24 years |
| 25 to 34 | Early Career Adults | Customers between the age group of 25 to 34 years |
| 35 to 44 | Mid-Career Adults | Customers between the age group of 35 to 44 years |
| 45 to 54 | Late Career Adults | Customers between the age group of 45 to 54 years |
| 55 to 64 | Near-Retirement Adults | Customers between the age group of 55 to 64 years |
| Above 65 | Retired Adults | Customers above the age of 65 years |

### 3.5. Conversion of Relevant Variables into Categorical Data Types

With multiple variables present in the historical customer data of Nova Apex Bank, several variables originally represented as object and integer data types have been reclassified into categorical data types, as illustrated by Figure 6. By doing so, the historical data analysis would be easier to understand, and the efficiency of predictive models would improve as predictive models process categorical data efficiently. In this regard, the following variables have been converted into categories:

- *Gender*
- *CategorizedAge*
- *Geography*
- *CategorizedCreditScore*
- *NumOfProducts*
- *HasCrCard*
- *IsActiveMember*
- *Exited*

```
# Convert relevant variables, including Gender, CategorizedAge, Geography, CategorizedCreditScore, NumOfProducts, HasCrCard, IsActiveMember and Exited, to categories

churn_df['Gender'] = churn_df['Gender'].astype('category')
churn_df['CategorizedAge'] = churn_df['CategorizedAge'].astype('category')
churn_df['Geography'] = churn_df['Geography'].astype('category')
churn_df['CategorizedCreditScore'] = churn_df['CategorizedCreditScore'].astype('category')
churn_df['NumOfProducts'] = churn_df['NumOfProducts'].astype('category')
churn_df['HasCrCard'] = churn_df['HasCrCard'].astype('category')
churn_df['IsActiveMember'] = churn_df['IsActiveMember'].astype('category')
churn_df['Exited'] = churn_df['Exited'].astype('category')
```

*Figure 6: Conversion of Relevant Variables into Categorical Data Types*

### 3.6. Replacement of Binary Variables for Improved Data Interpretation

While various variables of the historical customer data of Nova Apex Bank have been converted into categories, a few of these variables, including *HasCrCard*, *IsActiveMember* and *Exited*, have binary categorical values of 0 and 1, which have been replaced with No and Yes, respectively, as depicted in Figure 7. In this way, the historical customer data of Nova Apex Bank would be easier to interpret, with the new values providing a simple understanding of the variables' meaning. Additionally, this replacement of values would enhance the apprehension of insights using visualizations.

```
# Replace categorical variables, including HasCrCard, IsActiveMember and Exited, with 0 and 1 to No and Yes, respectively

churn_df[['HasCrCard', 'IsActiveMember', 'Exited']] = churn_df[['HasCrCard', 'IsActiveMember', 'Exited']].replace({0: 'No', 1: 'Yes'})
```

*Figure 7: Replacement of Binary Variables for Improved Data Interpretation*

### 3.7. Removal of Irrelevant Variables

As part of the data pre-processing facet carried out in this project, the variables *RowNumber*, *CustomerID* and *Surname* have been eliminated from the historical customer data of Nova Apex Bank (Figure 8) since they represent customer identifiers and do not provide essential information for analysis. This way, the dataset is simplified to intensely focus on customer characteristics and generate valuable insights from the historical data analysis and predictive models.

```
# Remove irrelvant variables, including RowNumber, CustomerID and Surname

churn_df = churn_df.drop(['RowNumber', 'CustomerId', 'Surname'], axis = 1)
```

*Figure 8: Removal of Irrelevant Variables*

### 3.8. Organization of All Variables

Another component carried out within pre-processing the historical customer data of Nova Apex Bank is arranging variables into two facets: customer demographics and financial information. The primary reason behind organizing variables is to improve the clarity of analytical efforts, simplifying insights on how customer demographics correspond to financial behaviors.

The following table highlights the components present under the customer demographics and financial information of the historical customer data of Nova Apex Bank:

| Dataset Component | Variable |
|---|---|
| Customer Demographics | *Gender* |
| | *Age* |
| | *CategorizedAge* |
| | *Geography* |
| Financial Information | *Balance* |
| | *EstimatedSalary* |
| | *CreditScore* |
| | *CategorizedCreditScore* |
| | *NumOfProducts* |
| | *HasCrCard* |
| | *Tenure* |
| | *IsActiveMember* |
| | *Exited* |

### 3.9. Exporting of Pre-processed Data

The final stage of the data pre-processing stage for the historical customer data of Nova Apex Bank involves exporting the modified data as a comma-separated value (CSV) file, as highlighted by Figure 9. This has been performed to develop a customer churn analysis dashboard on Tableau, as stated in section 4.3. of this report.

```
# Export the pre-processed dataset for creating a dashboard in Tableau

churn_df.to_csv('Bank_CustomerChurn_New.csv')
```

*Figure 9: Exporting of Pre-processed Data*

# 4. Data Exploration

With the historical customer data of Nova Apex Bank rectified for any inconsistencies in data formats and organizations, the available data has been explored through descriptive analytics by developing several visualizations and interpreting their relevant insights. Through this, the historical customer data can be analyzed effectively and the factors influencing customer churn at Nova Apex Bank can be ascertained before developing predictive models based on the historical customer data.

Both Python and Tableau have been utilized to perform the data exploration stage of this project. This allows for a clear comprehension of the historical customer data before proceeding to the other stages of the project.

## 4.1. Univariate Analysis: Single Variable Analysis

To effectively understand the historical customer data of Nova Apex Bank, each variable has been analyzed in Python to ensure familiarity with the data and their characteristics. In this regard, the univariate analysis is broken down into two components based on customer information at Nova Apex Bank: demographics and financial information of customers, as previously stated in section 3.9 of this report.

### 4.1.1. Demographics

#### 4.1.1.1. Gender

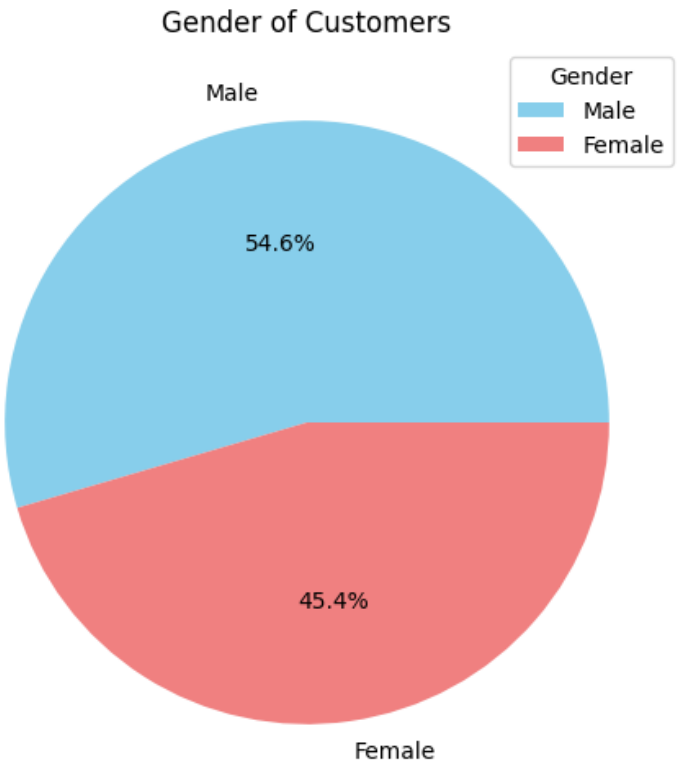


*Figure 10: Gender of Customers*

| Gender | Number of Customers |
|---|---|
| Female | 4,543 |
| Male | 5,457 |
| **Total** | **10,000** |

| Gender | Percentage of Customers |
|---|---|
| Female | 45.43% |
| Male | 54.57% |
| **Total** | **100%** |

#### 4.1.1.2. Age

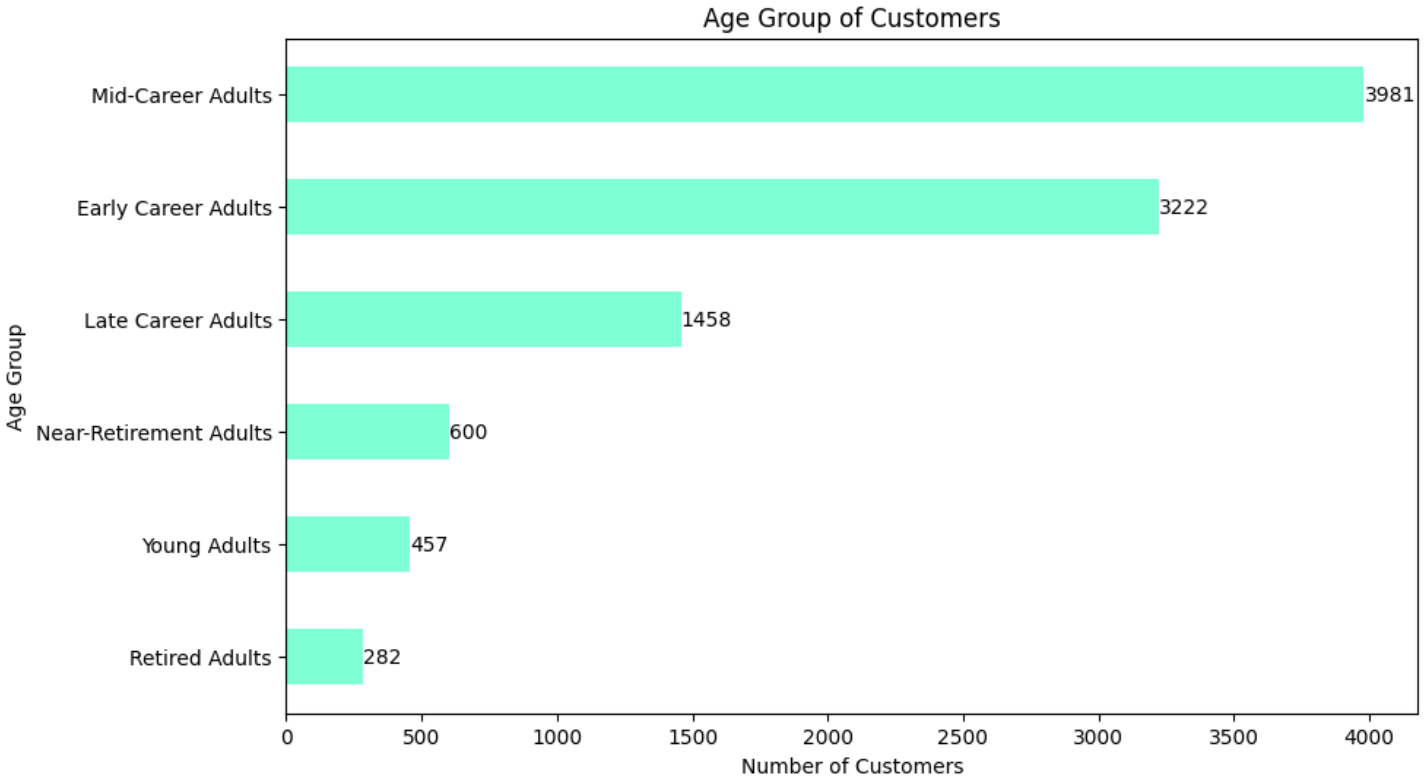


*Figure 11: Age Group of Customers*

| Age Group | Number of Customers |
|---|---|
| Young Adults | 457 |
| Early Career Adults | 3,222 |
| Mid-Career Adults | 3,981 |
| Late Career Adults | 1,458 |
| Near Retirement Adults | 600 |
| Retired Adults | 282 |
| **Total** | **10,000** |

| Age Group | Percentage of Customers |
|---|---|
| Young Adults | 4.57% |
| Early Career Adults | 32.22% |
| Mid-Career Adults | 39.81% |
| Late Career Adults | 14.58% |
| Near Retirement Adults | 6.00% |
| Retired Adults | 2.82% |
| **Total** | **10,000** |

#### 4.1.1.3. Geography

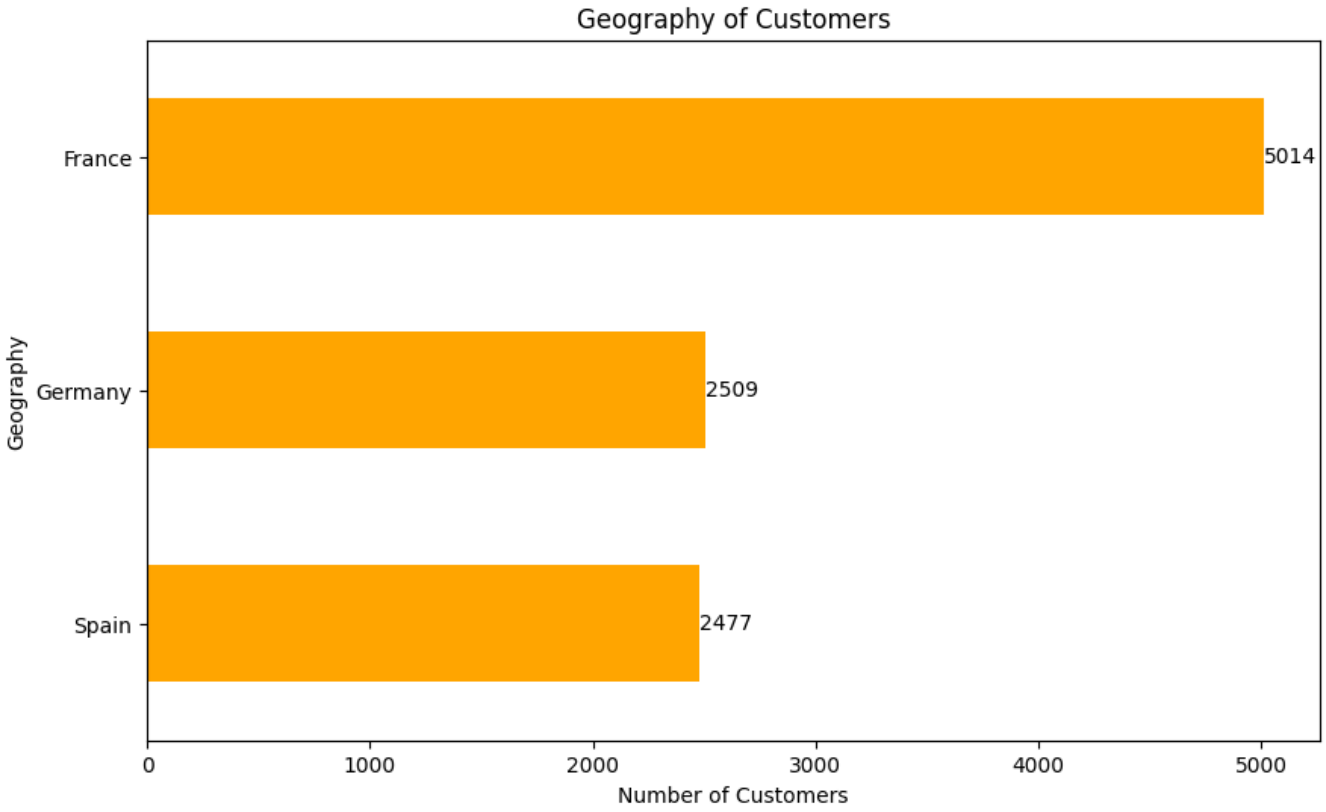


*Figure 12: Geography of Customers*

| Geography | Number of Customers |
|---|---|
| France | 5,014 |
| Germany | 2,509 |
| Spain | 2,477 |
| **Total** | **10,000** |

| Geography | Percentage of Customers |
|---|---|
| France | 50.14% |
| Germany | 25.09% |
| Spain | 24.77% |
| **Total** | **100%** |

### 4.1.2. Financial Information

#### 4.1.2.1. Balance

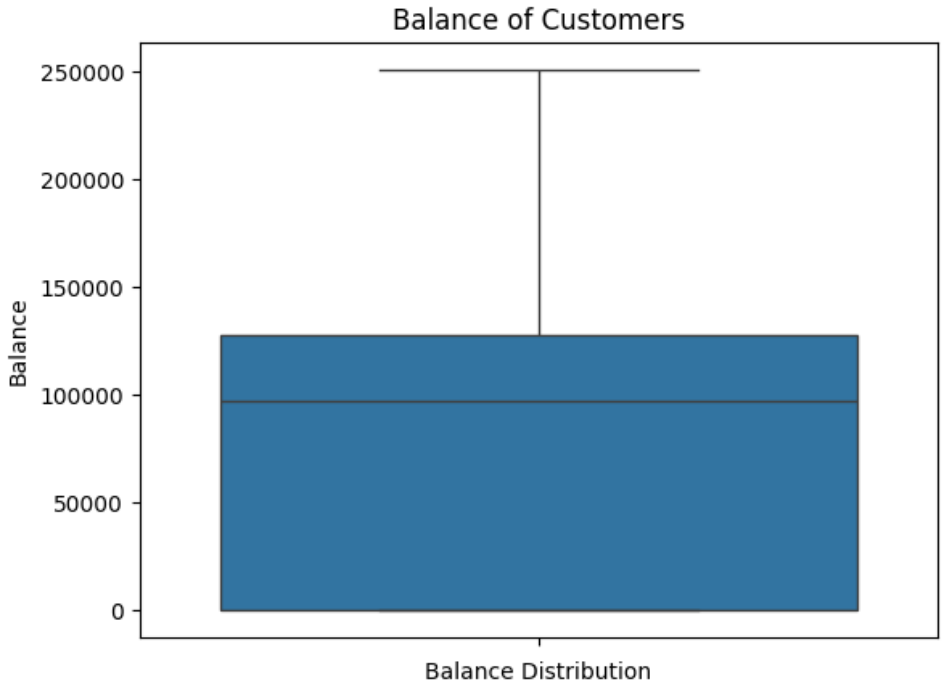


*Figure 13: Balance of Customers of Customers*

| Descriptive Statistic | Number |
|---|---|
| Number of Observations | 10,000 |
| Mean | 76,485.89 |
| Standard Deviation | 62,397.41 |
| Minimum | 0 |
| 25% | 0 |
| Median | 97,198.54 |
| 75% | 127,644.24 |
| Maximum | 250,898.09 |

#### 4.1.2.2. Estimated Salary

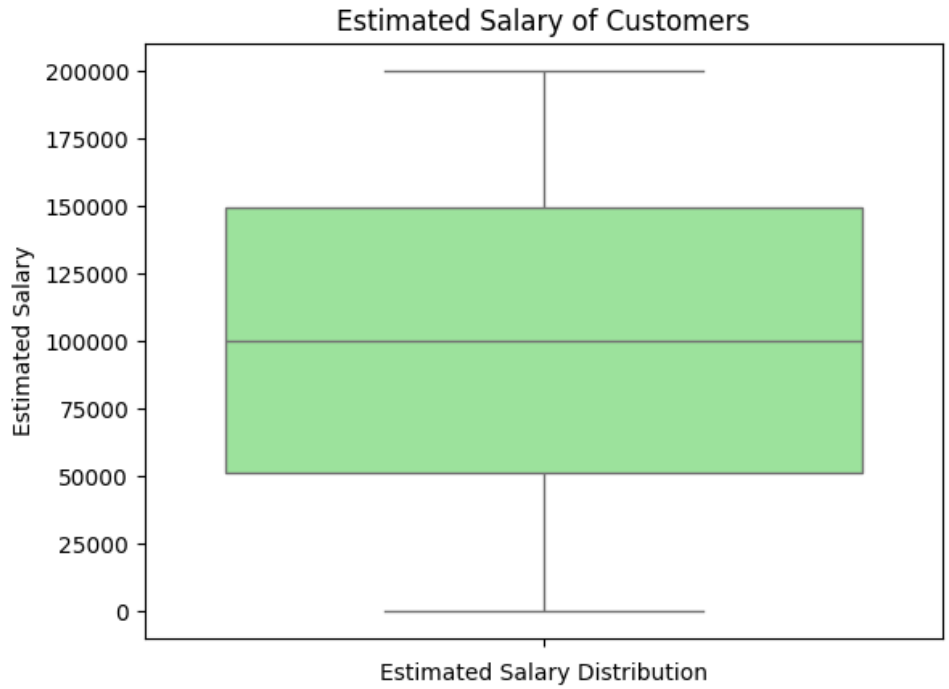


*Figure 14: Estimated Salary of Customers*

| Descriptive Statistic | Number |
|---|---|
| Number of Observations | 10,000 |
| Mean | 100,090.24 |
| Standard Deviation | 57,510.49 |
| Minimum | 11.58 |
| 25% | 51,002.11 |
| Median | 100,193.92 |
| 75% | 149,388.25 |
| Maximum | 199,992.48 |

#### 4.1.2.3. Credit Score Rating

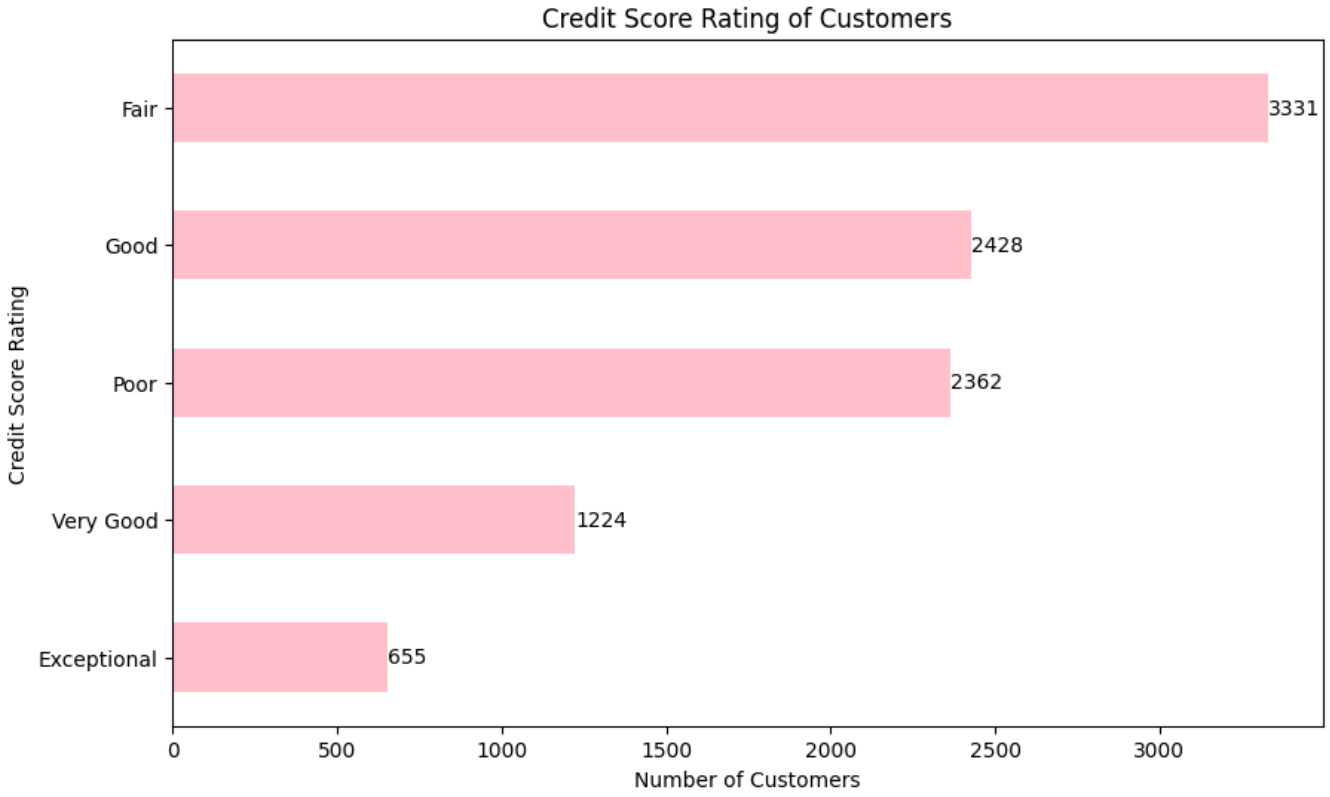


*Figure 15: Credit Score Rating of Customers*

| Credit Score Rating | Number of Customers |
|---|---|
| Poor | 2,362 |
| Fair | 3,331 |
| Good | 2,428 |
| Very Good | 1,224 |
| Exceptional | 655 |
| **Total** | **10,000** |

| Credit Score Rating | Percentage of Customers |
|---|---|
| Poor | 23.62% |
| Fair | 33.31% |
| Good | 24.28% |
| Very Good | 12.24% |
| Exceptional | 6.55% |
| **Total** | **100%** |

#### 4.1.2.4. Number of Products Held

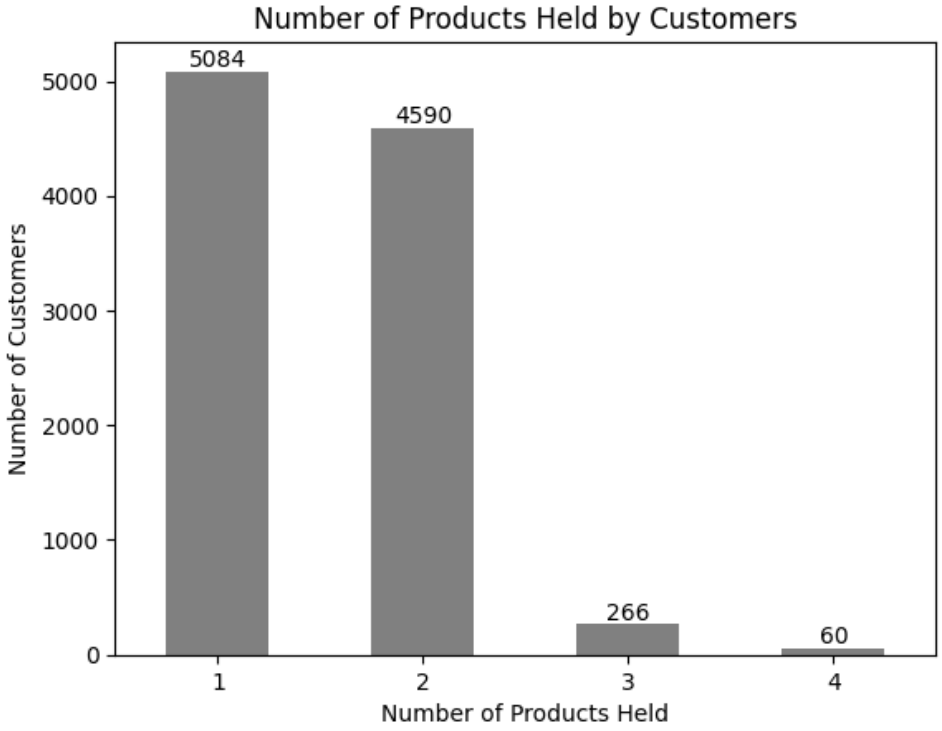


*Figure 16: Number of Products Held by Customers*

| Number of Products Held | Number of Customers |
|---|---|
| 1 | 5,084 |
| 2 | 4,590 |
| 3 | 266 |
| 4 | 60 |
| **Total** | **10,000** |

| Number of Products Held | Percentage of Customers |
|---|---|
| 1 | 50.84% |
| 2 | 45.90% |
| 3 | 2.66% |
| 4 | 0.60% |
| **Total** | **100%** |

#### 4.1.2.5. Credit Card Holder

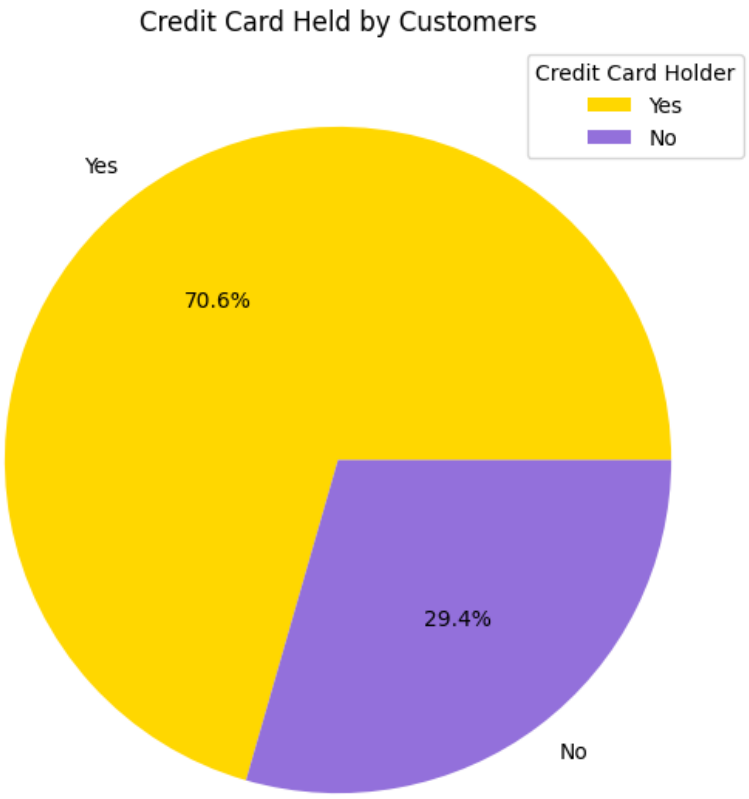


*Figure 17: Credit Card Held by Customers*

| Credit Card Holder | Number of Customers |
|---|---|
| No | 2,945 |
| Yes | 7,055 |
| **Total** | **10,000** |

| Credit Card Holder | Percentage of Customers |
|---|---|
| No | 29.45% |
| Yes | 70.55% |
| **Total** | **100%** |

#### 4.1.2.6. Tenure

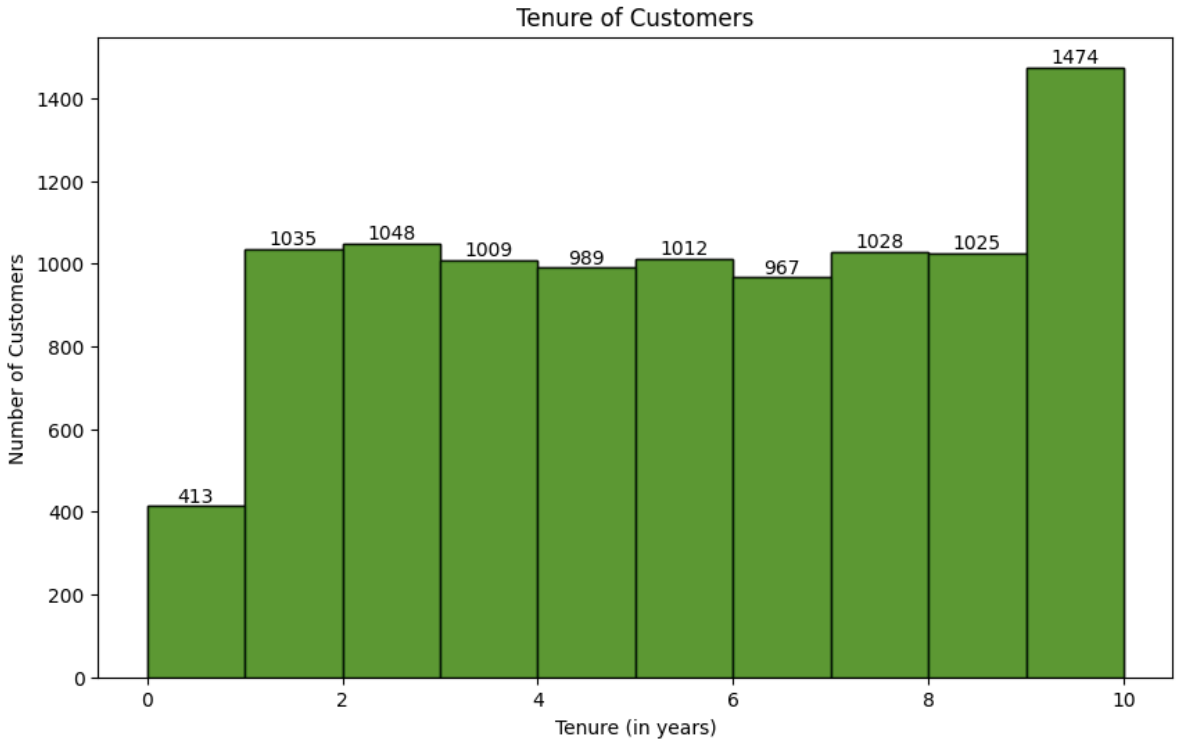


*Figure 18: Tenure of Customers*

| Tenure | Number of Customers |
|---|---|
| 0 | 413 |
| 1 | 1,035 |
| 2 | 1,048 |
| 3 | 1,009 |
| 4 | 989 |
| 5 | 1,012 |
| 6 | 967 |
| 7 | 1,028 |
| 8 | 1,025 |
| 9 | 984 |
| 10 | 490 |
| **Total** | **10,000** |

| Tenure | Percentage of Customers |
|---|---|
| 0 | 4.13% |
| 1 | 10.35% |
| 2 | 10.48% |
| 3 | 10.09% |
| 4 | 9.89% |
| 5 | 10.12% |
| 6 | 9.67% |
| 7 | 10.28% |
| 8 | 10.25% |
| 9 | 9.84% |
| 10 | 4.90% |
| **Total** | **100%** |

#### 4.1.2.7. Activeness as a Member

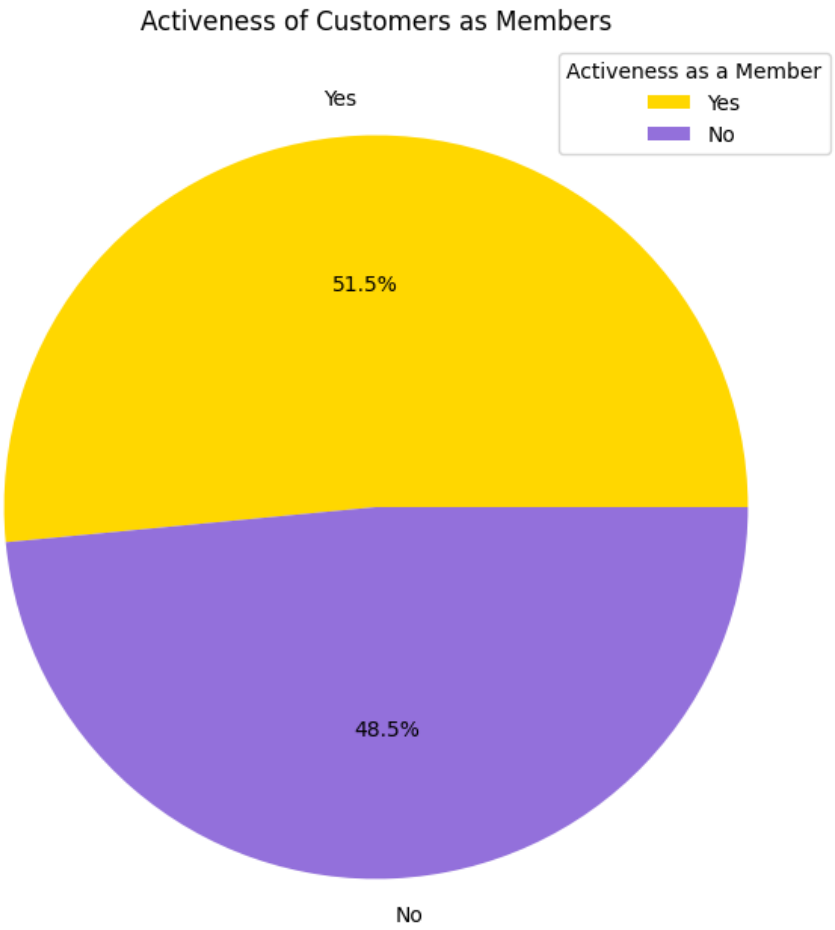


*Figure 19: Activeness of Customers as Members*

| Activeness as a Member | Number of Customers |
|---|---|
| No | 4,849 |
| Yes | 5,151 |
| **Total** | **10,000** |

| Activeness as a Member | Percentage of Customers |
|---|---|
| No | 48.49% |
| Yes | 51.51% |
| **Total** | **100%** |

#### 4.1.2.8. Churn

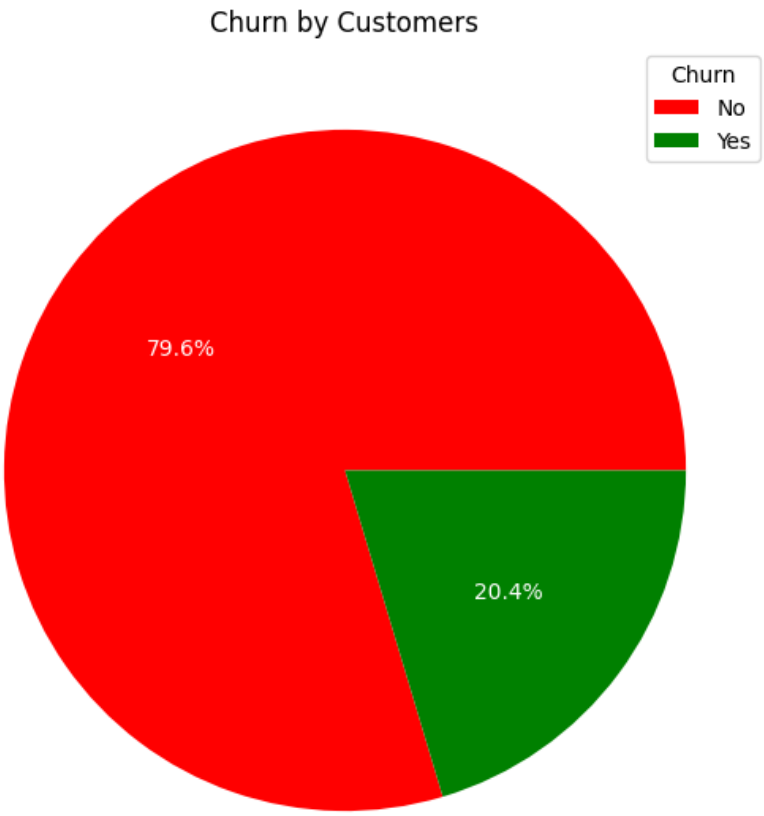


*Figure 20: Churn by Customers*

| Churn | Number of Customers |
|---|---|
| No | 7,963 |
| Yes | 2,037 |
| **Total** | **10,000** |

| Churn | Percentage of Customers |
|---|---|
| No | 79.63% |
| Yes | 20.37% |
| **Total** | **100%** |

### 4.2. Bivariate Analysis: Multiple Variable Analysis

To further identify the reasons behind customer churn at Nova Apex Bank, a bivariate analysis has been conducted by comparing the relationship between churn, i.e., *Exited,* and another customer attribute within demographics and financial information. Exploring the relationship between two variables would enable the identification of factors behind customer churn at Nova Apex Bank and taking measurable steps to limit churn even before such circumstances occur at the organization.

### 4.2.1. Relationship between Churn and Demographics

#### 4.2.1.1. Churn and Gender

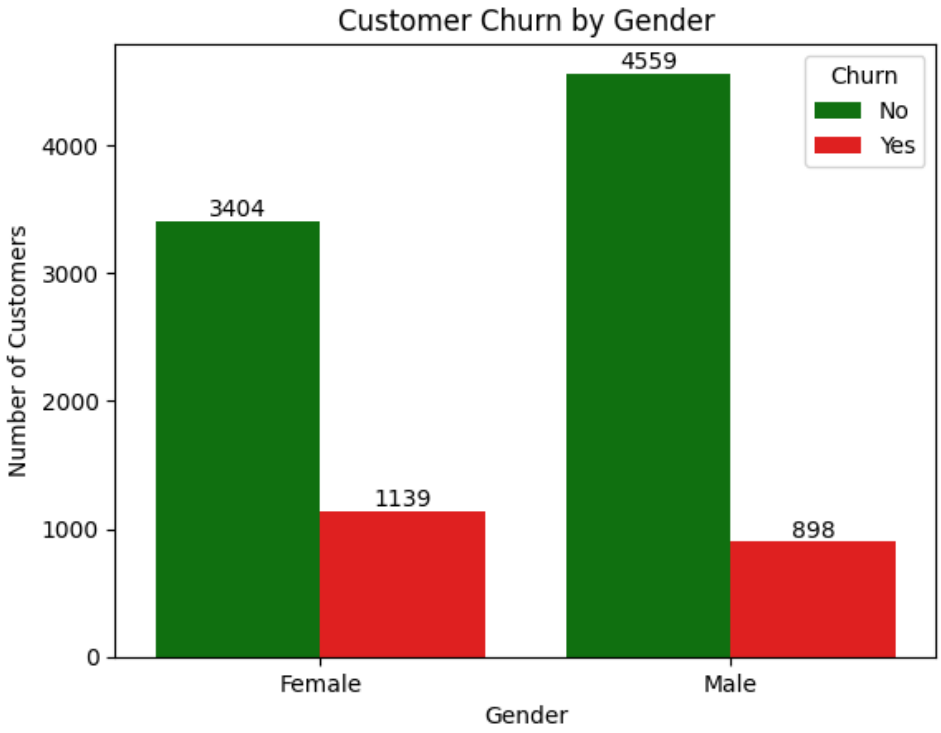


*Figure 21: Relationship between Customer Churn and Gender*

| | Churn | | |
|---|---|---|---|
| **Gender** | **No** | **Yes** | **Total** |
| Female | 3,404 | 1,139 | 4,543 |
| Male | 4,559 | 898 | 5,457 |
| **Total** | **7,963** | **2,037** | **10,000** |

As stated in Figure 21, the relationship between customer churn (*Exited*) and gender (*Gender*) reveals that out of 4,543 female consumers that have been part of Nova Apex Bank's operations, 1,139 female customers have discontinued their association with the firm, while 3,404 female customers have not churned. In contrast, 898 individuals out of 5,457 male customers have ceased to do business with Nova Apex Bank, with the remaining 4,559 customers continuing to be a part of the company. This information shows that female customers represent the largest gender, which has churned at Nova Apex Bank, as opposed to male consumers, despite the company having fewer female customers.

#### 4.2.1.2. Churn and Age

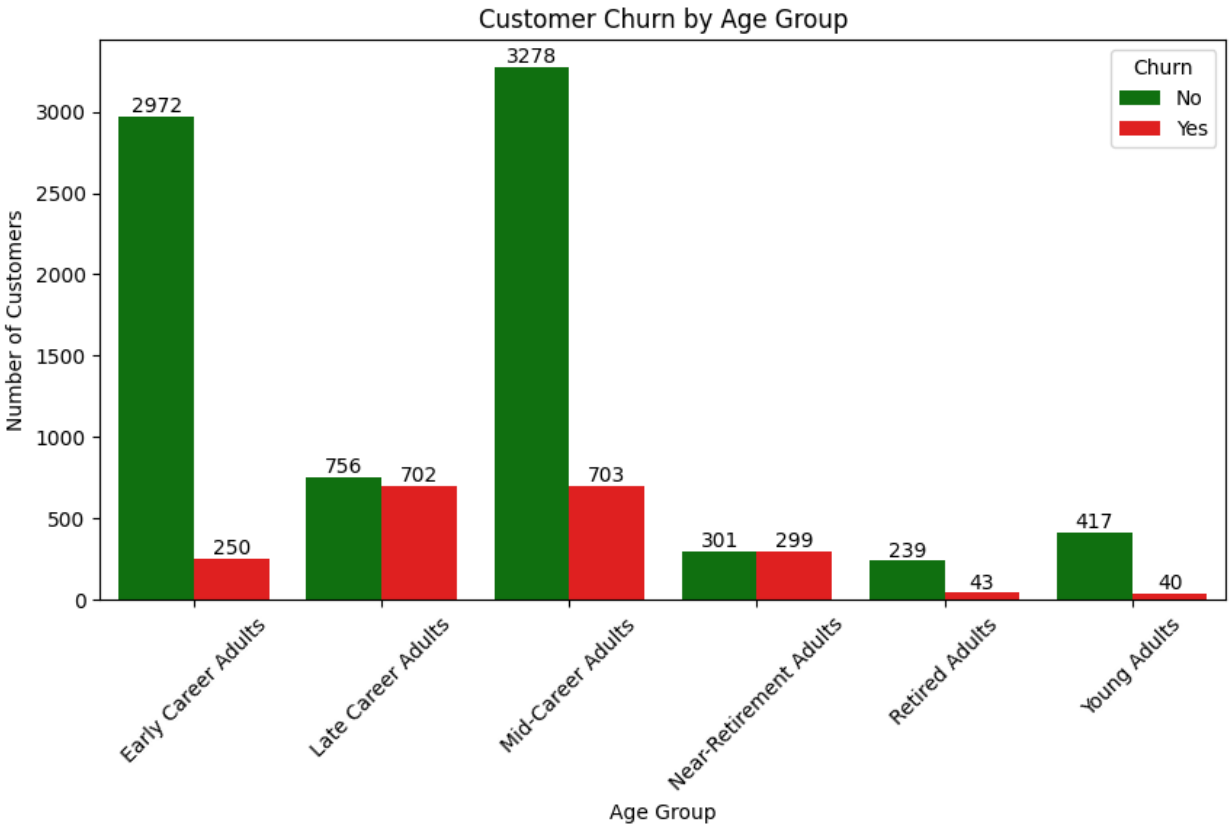


*Figure 22: Relationship between Customer Churn and Age*

| | Churn | | |
|---|---|---|---|
| **Age Category** | **No** | **Yes** | **Total** |
| Young Adults | 417 | 40 | 457 |
| Early Career Adults | 2,972 | 250 | 3,222 |
| Mid-Career Adults | 3,278 | 703 | 3,981 |
| Late Career Adults | 756 | 702 | 1,458 |
| Near-Retirement Adults | 301 | 299 | 600 |
| Retired Adults | 239 | 43 | 282 |
| **Total** | **7,963** | **2,037** | **10,000** |

According to Figure 22, the relationship between customer churn (*Exited*) and age groups (*CategorizedAgeGroup*) highlights that those customers belonging to the age groups of mid-career adults, i.e., between 35 and 44 years old, and late career adults, i.e., 45 and 54 years, represent the most significant number of individuals that have stopped doing business with Nova Apex Bank. While 703 individuals have churned from the mid-career adults category, 702 customers belong to the late career adults group, implying near-identical figures of churned customers. In contrast, retired adults (65 years and above) and young adults (18 to 24 years) demonstrate the lowest customer churns, with 43 individuals and 40 individuals, respectively. In this regard, it can be stated that churn is prevalent among customers who are in the middle of their careers but much before their retirement.

#### 4.2.1.3. Churn and Geography

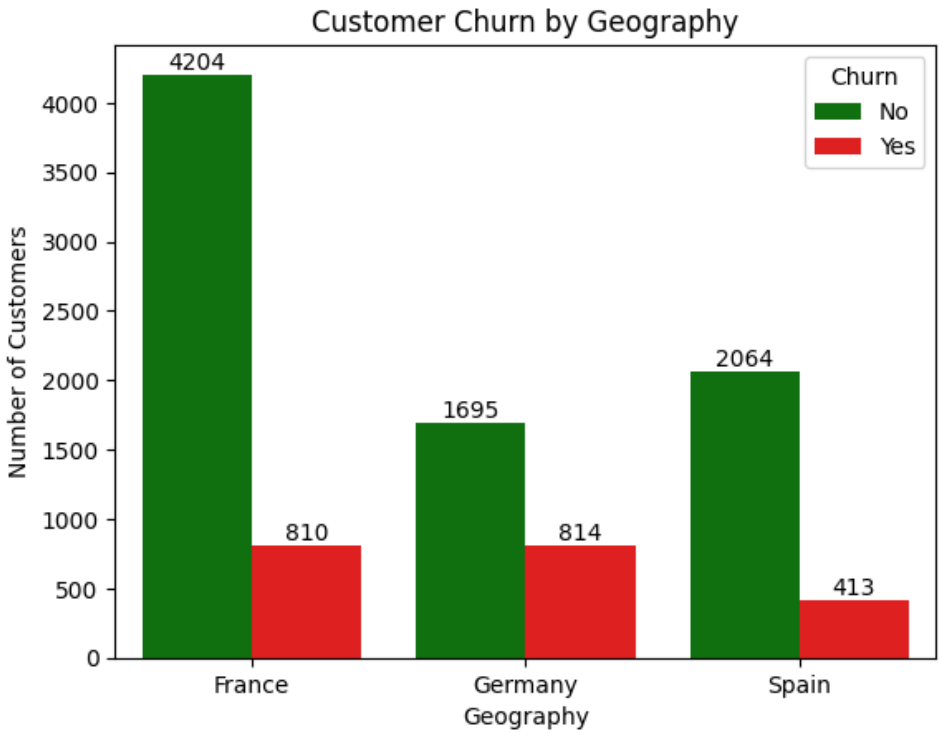


*Figure 23: Relationship between Customer Churn and Geography*

| | Churn | | |
|---|---|---|---|
| **Geography** | **No** | **Yes** | **Total** |
| France | 4,204 | 810 | 5,014 |
| Germany | 1,695 | 814 | 2,509 |
| Spain | 2,064 | 413 | 2,477 |
| **Total** | **7,963** | **2,037** | **10,000** |

Upon investigating the relationship between customer churn (*Exited*) and geography (*Geography)* in Figure 23, the data indicates that customers from Germany and France have nearly identical figures on customers quitting their association with Nova Apex Bank. While 814 individuals out of a total customer base of 10,000 have stopped doing business with Nova Apex Bank in Germany, 810 customers in France have chosen not to continue doing business with the company. On the contrary, only 413 customers from Spain have ceased to do business with Nova Apex Bank. This information shows that most customers in Germany and France have churned compared to customers in Spain.

### 4.2.2. Relationship between Churn and Financial Information

#### 4.2.2.1. Churn and Balance

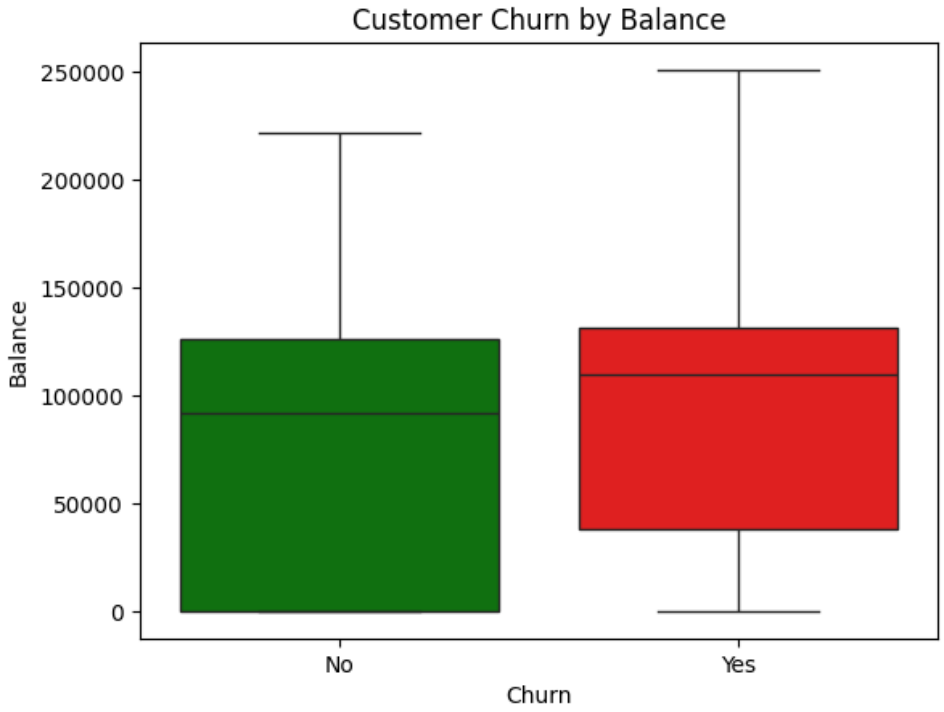


*Figure 24: Relationship between Customer Churn and Balance*

| Descriptive Statistic (No) | Number |
|---|---|
| Number of Observations | 7,963 |
| Mean | 72,745.30 |
| Standard Deviation | 62,848.04 |
| Minimum | 0 |
| 25% | 0 |
| Median | 92,072.68 |
| 75% | 126,410.28 |
| Maximum | 221,532.80 |

| Descriptive Statistic (Yes) | Number |
|---|---|
| Number of Observations | 2,037 |
| Mean | 91,108.54 |
| Standard Deviation | 58,360.79 |
| Minimum | 0 |
| 25% | 38,340.02 |
| Median | 109,349.29 |
| 75% | 131,433.33 |
| Maximum | 250,898.09 |

When examining the relationship between customer churn (*Exited*) and account balances (*Balance*), as stated in Figure 24, the data reveals that customers who have stopped doing business with Nova Apex Bank demonstrate a higher median balance of nearly €109,349 as opposed to a median balance of €92,072 that continue to remain customers of the firm. Moreover, the interquartile range of account balances of churned customers is wider, from nearly €38,340 to €131,433; this represents more considerable variability. On the contrary, the interquartile range of account balances is narrower for customers still part of Nova Apex Bank, ranging from €0 to €126,410, with a large portion having no balance. Additionally, the maximum balance for churned customers stands at approximately €250,898 as compared to a maximum balance of €221,533 for non-churned customers. This information suggests that high-balance customers are more likely to churn at Nova Apex Bank.

#### 4.2.2.2. Churn and Estimated Salary

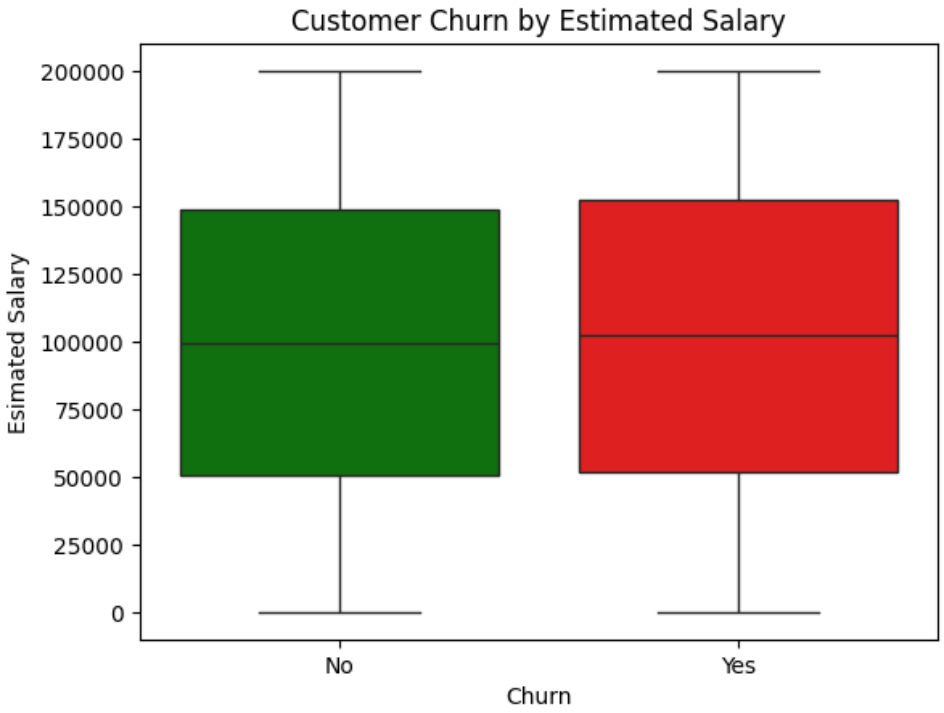


*Figure 25: Relationship between Customer churn and Estimated Salary*

| Descriptive Statistic (No) | Number |
|---|---|
| Number of Observations | 7,963 |
| Mean | 99,738.39 |
| Standard Deviation | 57,405.59 |
| Minimum | 90.07 |
| 25% | 50,783.49 |
| Median | 99,645.04 |
| 75% | 148,609.96 |
| Maximum | 199,992.48 |

| Descriptive Statistic (Yes) | Number |
|---|---|
| Number of Observations | 2,037 |
| Mean | 101,465.68 |
| Standard Deviation | 57,912.42 |
| Minimum | 11.58 |
| 25% | 51,907.72 |
| Median | 102,460.84 |
| 75% | 152,422.91 |
| Maximum | 199,808.10 |

As part of identifying the relationship between customer churn (*Exited*) and estimated salary (*EstimatedSalary*) in Figure 25, the insights reveal that the estimated median salary of Nova Apex Bank customers is comparable among individuals that have churned and not churned, at figures of approximately €102,461 and €99,645, respectively. The mean of the estimated salary of Nova Apex Bank consumers is marginally higher among individuals who do not do business with the firm, at a value of nearly €101,465 as opposed to a mean of €99,738 for those who are still customers. Furthermore, there are identical interquartile ranges among churned and non-churned consumers of Nova Apex Bank; the $25^{th}$ percentile for churned and non-churned individuals stands at €51,907 and €50,783, respectively. Similarly, the $75^{th}$ percentile for customers still part of Nova Apex Bank stands at €152,423, while the $75^{th}$ percentile figure of €148,610 is among customers who have ceased business with the company. With the standard deviations, as well as the maximum and minimum estimated salaries of customers, also being identical, it can be inferred that estimated salary distributions are nearly comparable for churned and non-churned customers at Nova Apex Bank.

#### 4.2.2.3. Churn and Credit Score Rating

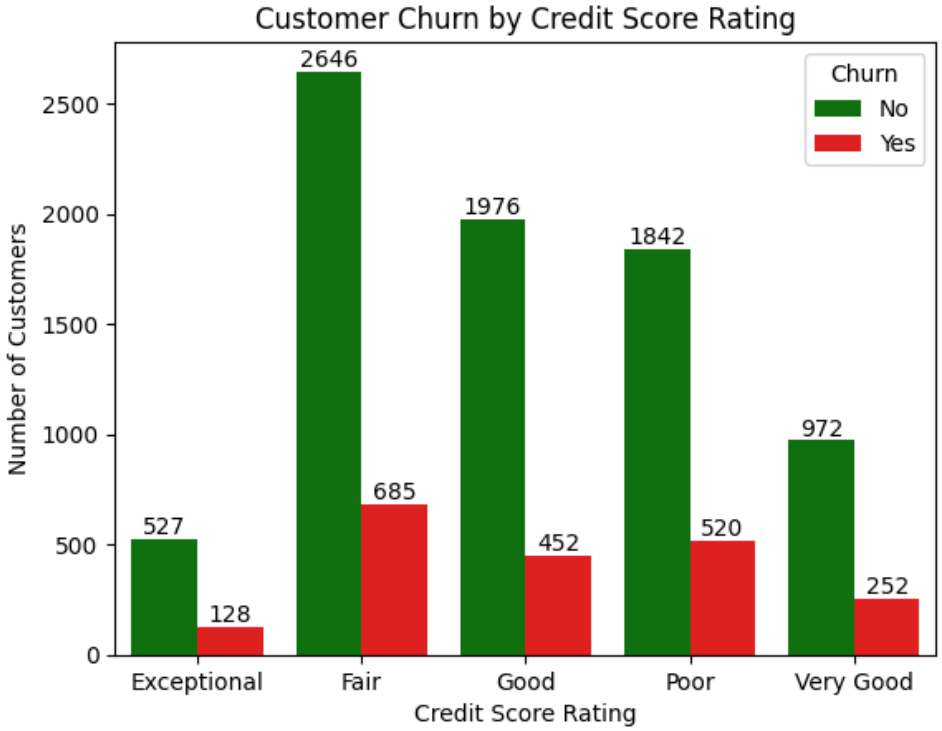


*Figure 26: Relationship between Customer Churn and Credit Score Rating*

| | Churn | | |
|---|---|---|---|
| **Credit Score Category** | **No** | **Yes** | **Total** |
| Poor | 1,842 | 520 | 2,362 |
| Fair | 2,646 | 685 | 3,331 |
| Good | 1,976 | 452 | 2,428 |
| Very Good | 972 | 252 | 1,224 |
| Exceptional | 527 | 128 | 655 |
| **Total** | **7,963** | **2,037** | **10,000** |

When exploring the relationship between customer churn (*Exited*) and credit score rating based on the FICO score ratings used in the financial industry (*CategorizedCreditScore*), as illustrated in Figure 26, numerous customers possessing a FICO credit score rating of fair or poor have decided to stop doing business with Nova Apex Bank. The data highlights that a high figure of 685 customers has obtained a fair credit score rating, i.e., a FICO score between the range of 580 and 669, as opposed to 2,646 individuals who have obtained the same FICO score rating and continue their association with Nova Apex Bank. Furthermore, 520 customers have received a poor FICO credit score rating, representing a credit score rating below 580. Nonetheless, customers with above-standard credit score ratings (good, very good, exceptional) continue associating with Nova Apex Bank.

#### 4.2.2.4. Churn and Number of Products Held

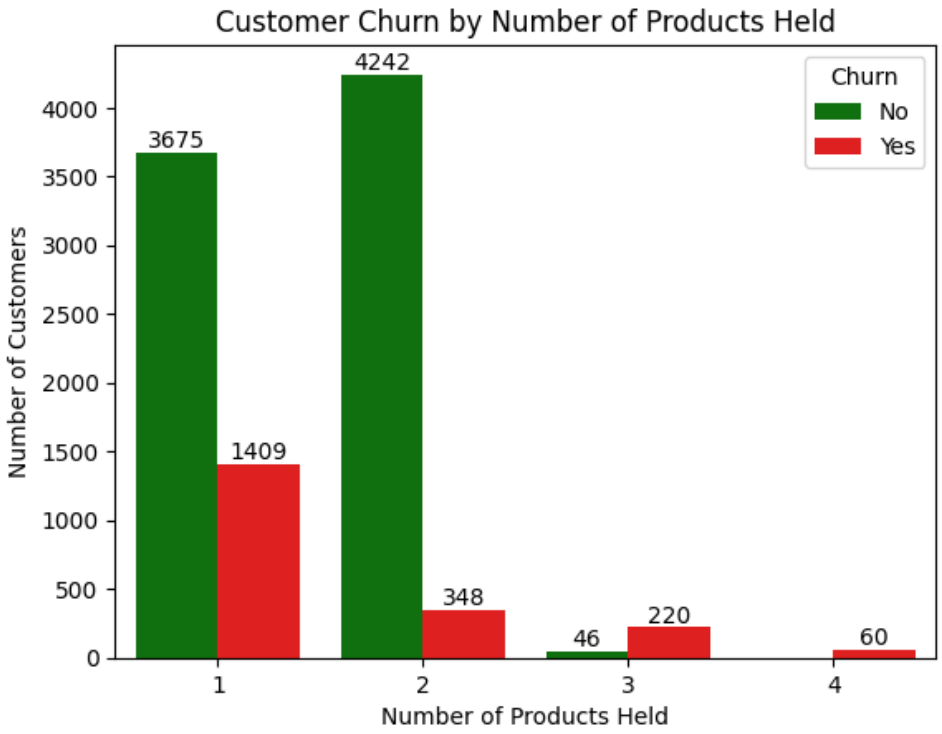


*Figure 27: Relationship between Customer Churn and Number of Products Held*

| | Churn | | |
|---|---|---|---|
| **Number of Products/** | **No** | **Yes** | **Total** |
| 1 | 3,675 | 1,409 | 5,084 |
| 2 | 4,242 | 348 | 4,590 |
| 3 | 46 | 220 | 266 |
| 4 | 0 | 60 | 60 |
| **Total** | **7,963** | **2,037** | **10,000** |

When analyzing the relationship between customer churn (*Exited*) and the number of Nova Apex Bank products (*NumOfProducts*) held, as highlighted in Figure 27, the data highlights that 1,409 customers possessing one product have decided not to maintain their association with the organization, representing the highest number among all customers. As customers own more products offered by Nova Apex Bank, the number of consumers discontinuing their association with the company becomes smaller. For instance, 348 customers owning two products have decided not to preserve their relationship with Nova Apex Bank, followed by 220 customers and 60 customers who own three and four products, respectively. In contrast, there are no customer churns among individuals who own four products provided by Nova Apex Bank, while there are 4,242 individuals who own twoproducts and continue to remain customers of Nova Apex Bank; the latter represents the most significant number of non-churned customers at Nova Apex Bank.

#### 4.2.2.5. Churn and Credit Card Holder

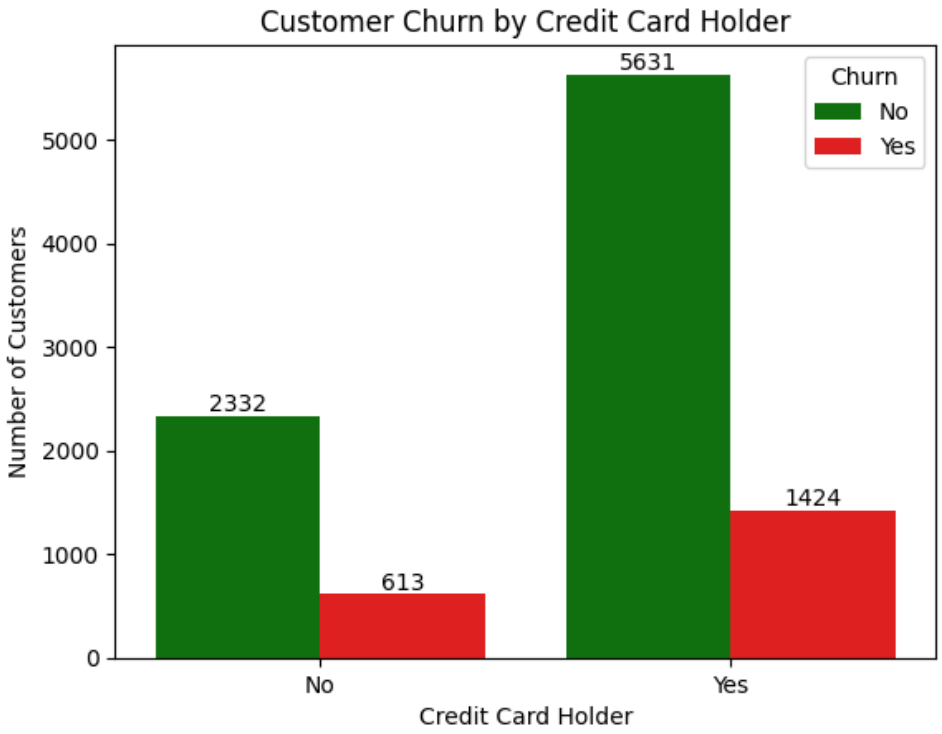


*Figure 28: Relationship between Customer Churn and Credit Card Holder*

| | Churn | | |
|---|---|---|---|
| **Credit Card Holder** | **No** | **Yes** | **Total** |
| No | 2,332 | 613 | 2,945 |
| Yes | 5,631 | 1,424 | 7,055 |
| **Total** | **7,963** | **2,037** | **10,000** |

According to Figure 28, the relationship between customer churn (*Exited*) and possession of a credit card (*HasCrCard)* exhibits that out of 7,055 consumers who own a credit card, 1,424 individuals have stopped doing business with Nova Apex Bank. On the contrary, out of the same number of 7,055 individuals having a credit card, a figure of 5,631 individuals continue to remain customers of Nova Apex Bank, representing the largest group of customers in this category. Furthermore, out of 2,945 customers at Nova Apex Bank, 613 individuals do not own a credit card and have chosen not to continue their association with the firm. In contrast, 2,332 individuals remain customers of Nova Apex Bank despite lacking a credit card. From this information, it can be discerned that most customers holding a credit card have churned the most at the company as opposed to those who do not hold a credit card.

#### 4.2.2.6. Churn and Tenure

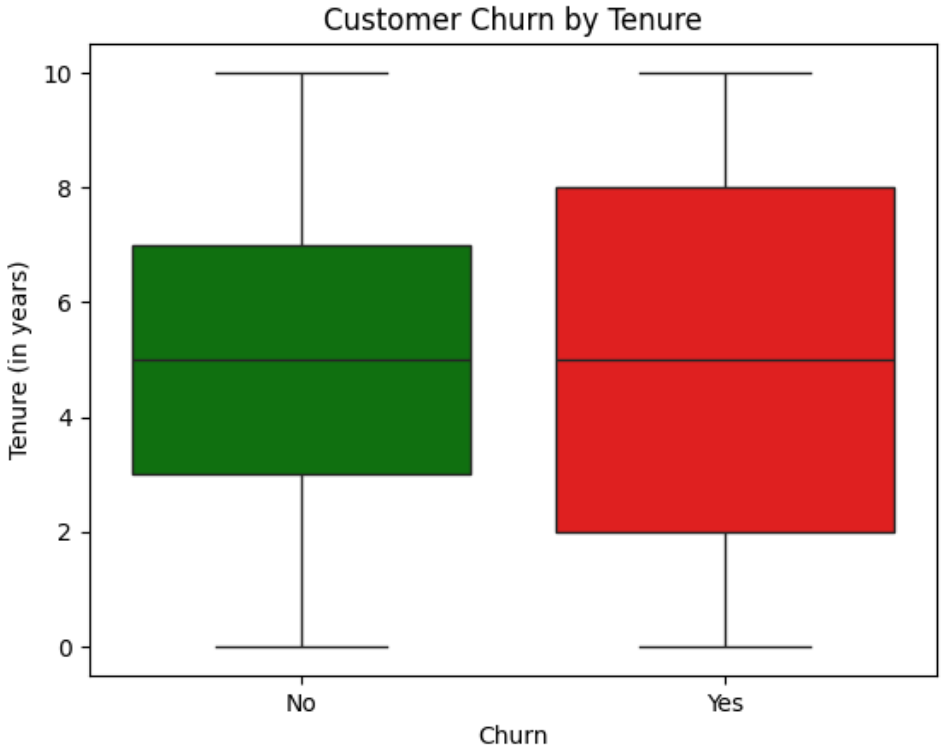


*Figure 29: Relationship between Customer Churn and Tenure*

| Descriptive Statistic (No) | Number |
|---|---|
| Number of Observations | 7,963 |
| Mean | 5.03 |
| Standard Deviation | 2.88 |
| Minimum | 0 |
| 25% | 3 |
| Median | 5 |
| 75% | 7 |
| Maximum | 10 |

| Descriptive Statistic (Yes) | Number |
|---|---|
| Number of Observations | 2,037 |
| Mean | 4.93 |
| Standard Deviation | 2.94 |
| Minimum | 0 |
| 25% | 2 |
| Median | 5 |
| 75% | 8 |
| Maximum | 10 |

While comprehending the relationship between customer churn (*Exited*) and tenure (*Tenure*) using Figure 29, it has been ascertained that the median tenure of customers, irrespective of whether they stop doing business with Nova Apex Bank or not, stands at 5 years, along with a similar range of 5 to 10 years. From the perspective of mean tenure, non-churned customers show a better performance of nearly 5.03 years as opposed to 4.93 years for consumers who have ceased to do business with Nova Apex Bank, representing a slightly larger tenure for individuals who continue to be customers. Furthermore, the interquartile range for customers who discontinued their association with Nova Apex Bank is 3 years compared to 4 years for those who have not churned, showing a marginally broader spread among non-churned customers. The standard deviations between the churned and non-churned consumers show similar outputs, at 2.94 years and 2.88 years, respectively. In this regard, it can be emphasized that tenure distributions are identical for customers that have churned or not.

#### 4.2.2.7. Churn and Activeness as a Member

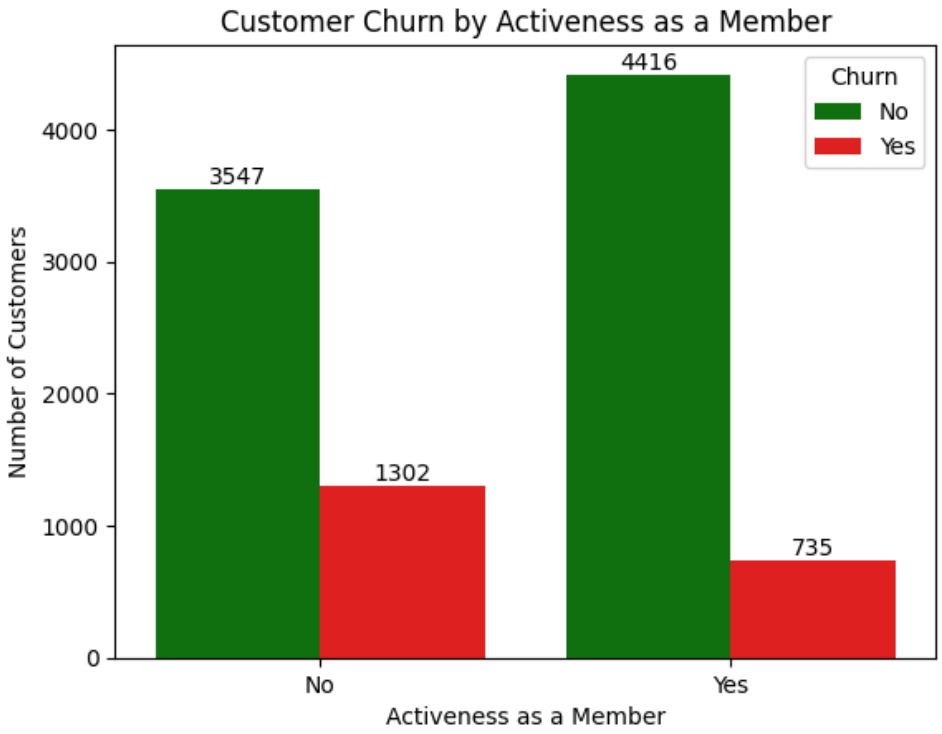


*Figure 30: Relationship between Customer Churn and Activeness as a Member*

| | Churn | | |
|---|---|---|---|
| **Activeness** | **No** | **Yes** | **Total** |
| No | 3,547 | 1,302 | 4,849 |
| Yes | 4,416 | 735 | 5,151 |
| **Total** | **7,963** | **2,037** | **10,000** |

Upon discerning the relationship between customer churn (*Exited*) and the activeness of customers as members of Nova Apex Bank (*IsActiveMember*) in Figure 30, the data indicates that most customers, i.e., 4,416 individuals, are engaged members of the company and continue their association with the firm. Similarly, 3,547 customers continue to do business with Nova Apex Bank, yet they are not active members of the firm. On the other hand, the data emphasizes that 1,302 individuals are not active members and have chosen to quit their association with Nova Apex Bank. In comparison, a low number of 735 engaged members have decided to stop their association with the company. This information highlights that customers who are not active members of Nova Apex Bank have churned the most at the company as opposed to those who are regularly engaged members.

### 4.3. Customer Churn Analysis Dashboard

Within the previous section of the report, the past historical customer data of Nova Apex Bank has been analyzed using bivariate and multivariate analysis to comprehend better why customers quit their association with the company. While essential insights have been attained from the bivariate and multivariate analyses of customer churn, it may be beneficial to provide a tool that can help interpret what is happening with the historical customer data in a more accessible and actionable format through pre-developed calculations and graphics. To do so, an interactive dashboard on customer churn analysis has been developed on Tableau; this serves as an extension to the historical data analysis carried out on Python by incorporating all customer attributes on demographics and financial information in the form of aggregations and visualizations. The following points highlight the benefits of utilizing the customer churn analysis dashboard:

- Dynamic aggregations and visualizations to quickly understand the state of customer churn at Nova Apex Bank

- Real-time monitoring of customer churn at Nova Apex Bank by replacing the existing historical customer data utilized in the dashboard with a new dataset at any period

- Simplification of historical data analysis to quickly identify trends and patterns in customer churn without manual data processing


*Figure 31: Relationship between customer churn and credit card holder*

As illustrated in Figure 31, various aggregations and visualizations have been created within the customer churn analysis dashboard. Furthermore, any individual accessing the dashboard can utilize the provided churn filter based on the *Exited* variable containing the values of 'No' and 'Yes,' representing churned and non-churned customers, respectively. This would enable any user to either understand Nova Apex Bank's entire customer base or a specific segment, whether churned or non-churned customers.

The following points explain more about the aggregations and visualizations present in the customer churn analysis dashboard:

- **Aggregations:** The customer churn analysis dashboard comprises three aggregations that represent key metrics by providing high-level information about the historical customer data of Nova Apex Bank. In this context, each aggregation, constituting total customers, churned customers and churn rate of customers, have been explained in the table below:

| **Aggregations** | |
|---|---|
| **Aggregation** | **Description** |
| Total Customers | The total customers aggregation counts the number of customers present within the historical customer data of Nova Apex Bank. |
| Churned Customers | The churned customers aggregation counts the number of customers that churned out of the total customers at Nova Apex Bank. |
| Churn Rate | The churn rate aggregation calculates the proportion of customers that churned out of the total customers in the form of a percentage at Nova Apex Bank.<br><br>The churn rate aggregation is based on the total customers and churned customers aggregations and is computed by the following formula:<br><br>$$\text{Churn Rate} = \frac{\text{Churned Customers}}{\text{Total Customers}}$$ |

- **Visualizations:** Several visualizations, including pie charts, bar charts (horizontal, vertical), histograms, a line chart and a bubble chart, have been developed on the customer churn analysis dashboard of Nova Apex Bank. A few of these visualizations loosely resemble the bivariate analysis visualizations explained in section 4.2 of this report. The table below highlights the visualizations that have been created on the customer churn analysis dashboard based on several variables of the historical customer data of Nova Apex Bank.

| **Visualizations** | | |
|---|---|---|
| **Visualization** | **Variable** | **Description** |
| Gender | *Gender* | The gender pie chart highlights the proportion of customers for the *Gender* variable, enabling the identification of the largest and smallest churned or non-churned gender at Nova Apex Bank. |
| Age | *CategorizedAgeGroup* | The age horizontal bar chart illustrates the breakdown of customers for the *CategorizedAgeGroup* variable, allowing the comprehension of the most and least number of churned or non-churned age groups, from young adults to retired adults, at Nova Apex Bank. |
| Geography | *Geography* | The geography bubble chart outlines the number of customers for the *Geography* variable, enabling to discover the most or least number of churned or non-churned customers of Nova Apex Bank in France, Germany and Spain. |
| Credit Score Rating | *CategorizedCreditScore* | The credit score rating horizontal bar chart describes the breakdown of customers for |

| Visualizations | | |
|---|---|---|
| **Visualization** | **Variable** | **Description** |
| | | the *CategorizedCreditScore*, emphasizing the largest and smallest number of churned or non-churned customers at Nova Apex Bank according to their FICO credit score rating. |
| Number of Products Held | *NumOfProducts* | The vertical bar chart on the number of products held shows the number of customers under the *NumOfProducts* variable, representing the number of products possessed by most or least number of churned or non-churned customers at Nova Apex Bank, with a minimum of 1 product and a maximum of 4 products. |
| Credit Card Holder | *CreditCardHolder* | The pie chart on credit card holders displays the proportion of customers for the *CreditCardHolder* variable, exhibiting the highest and lowest number of churned or non-churned customers at Nova Apex Bank who have a credit card and those who do not. |
| Tenure | *Tenure* | The tenure line chart illustrates the number of customers for the *Tenure* variable, describing the distribution of churned and non-churned customers for the number of years that a customer of Nova Apex Bank has been at the company. |
| Activeness as a Member | *IsActiveMember* | The pie chart on activeness as a member shows the proportion of customers for the |

| Visualizations | | |
|---|---|---|
| **Visualization** | **Variable** | **Description** |
| | | *IsActiveMember* variable, emphasizing the highest and lowest number of churned or non-churned customers at Nova Apex Bank, whether they are active members of the organization or not. |
| Estimated Salary | *EstimatedSalary* | The estimated salary histogram exhibits the distribution of customers based on the *EstimatedSalary* variable, enabling to discover whether churned or non-churned customers are concentrated in specific salary ranges or not. |
| Balance | *Balance* | The balance histogram displays the distribution of customers based on the *Balance* variable, allowing the identification of whether churned or non-churned customers are present under specific balance ranges or not. |

# 5. Feature Engineering

After the historical data analysis carried out on the historical customer data of Nova Apex Bank through data exploration, the next stage involves feature engineering. Unlike data pre-processing, which primarily deals with fundamental tasks such as handling missing data, modifying errors and formatting columns to ensure data quality before analyzing the historical customer data of Nova Apex Bank, feature engineering focuses on complex tasks. These tasks constitute the creation and transformation of features to significantly enhance the prediction power of the models for forecasting customer churn at Nova Apex Bank.

Like the data pre-processing stage, the feature engineering process has been performed on Python through multiple procedures. By doing so, the historical customer data of Nova Apex Bank would be ready to carry out the essential stage of the predictive analytics process on the same data.

## 5.1. Selection of Features and Target Variable

The first step of the feature engineering aspect of customer churn prediction at Nova Apex Bank is concerned with picking the vital pieces of information, i.e., features for forecasting an outcome, i.e., target variable. In the context of this project, the target variable is *Exited* (represented as y), while the features include the other fifteen columns (represented as X), as highlighted in Figure 32.

```
# Separate the features (X) and the target variable (y) from the dataset

X = model_churn_df.drop(columns = 'Exited')  # Features: all columns except 'Exited'
y = model_churn_df['Exited']  # Target variable: 'Exited' column
```

*Figure 32: Selection of Features and Target Variable*

### 5.2. Creation of Dummy Variables

As part of the feature engineering component of forecasting customer churn at Nova Apex Bank, building dummy variables transforms categorical data (Figure 33) into columns with binary values (Figure 34), enabling predictive models to process categorical data effectively. For example, within the historical customer data of Nova Apex Bank, the variable *Gender* is converted into 'Gender_Male,' with values of 'True' or 'False' that highlight whether a customer is Male.

```
# Specify the categorical variables that need to be converted into dummy/indicator variables

categorical_variables = ['Gender', 'Geography', 'CategorizedAge', 'CategorizedCreditScore', 'HasCrCard', 'IsActiveMember']
```

*Figure 33: Specification of categorical variables for creating dummy variables*

```
# Convert categorical variables into dummy variables

X = pd.get_dummies(X, columns = categorical_variables, drop_first = True)
```

*Figure 34: Creation of Dummy Variables*

### 5.3. Standardization of Numerical Features

Within the feature engineering component of predicting customer churn at Nova Apex Bank, standardization involves adjusting numerical features to have a mean of 0 and a standard deviation of 1, allowing for easier comparison (Figure 35). By standardizing numerical features, the accuracy and performance of predictive models are enhanced, resulting in more reliable business insights.

```
# Initialize the StandardScaler

scaler = StandardScaler()

# Select the numerical columns for scaling

numerical_features = ['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'EstimatedSalary']

# Apply scaling to the numerical features

X[numerical_features] = scaler.fit_transform(X[numerical_features])
```

*Figure 35: Standardization of Numerical Features*

### 5.4. Splitting of Data into Training and Testing Sets

The last step of the feature engineering aspect in forecasting customer churn at Nova Apex Bank involves splitting the historical customer data into two portions: training the predictive models and testing them (Figure 36). While the training data is utilized for developing and optimizing the predictive models, the testing data evaluates their performance on new, unseen data. In this way, each predictive model performs well in real-world applications, giving rise to sound business decisions.

For this project, the historical customer data of Nova Apex Bank has been divided into a 70-30 percent ratio split, with 70% used for training and 30% for testing. The 70-30 percent ratio ensures that the predictive models are trained with sufficient data while reserving enough unseen data to evaluate their performance effectively and generalize well to real-world scenarios.

```
# Split the dataset into training and testing sets with a 70 (train) to 30 (test) ratio

train_X, test_X, train_y, test_y = train_test_split(X, y, test_size = 0.3, random_state = 1)
```

*Figure 36: Splitting of Data into Training and Testing Sets*

# 6. Model Exploration

Various predictive models have been developed to predict customer churn at Nova Apex Bank, including logistic regression (full, forward, backward, and stepwise), decision tree and random forest. Each model's performance was enhanced through hyperparameter tuning, which involved systematically altering key settings, i.e., parameters. Specifically, a grid search method was employed to explore different combinations of these parameters, ensuring that each model's optimal configuration was determined to maximize predictive accuracy.

Since multiple predictive models have been developed in this project, several evaluation metrics have been incorporated to comprehend the performance of every model. In this regard, the evaluation metrics utilized throughout the project, in order of their prioritization, constitute the receiver operating characteristic - area under the curve (ROC-AUC), accuracy and F1-score. The following points highlight each of these evaluation metrics:

- **ROC-AUC:** The ROC-AUC metric enables the assessment of the effectiveness of each predictive model to differentiate between customers likely to churn and those not likely to churn at Nova Apex Bank. The higher the ROC-AUC value, the more reliable and accurate each predictive model is for forecasting customer churn at Nova Apex Bank. The following table provides information about the classified ROC-AUC score ranges:

| ROC-AUC Score Range | Classification | Description |
|---|---|---|
| 0.50 | Random | The predictive model cannot distinguish between churned and non-churned customers. |
| 0.50 to 0.59 | Poor | The predictive model is not effective in predicting customer churn, similar to random guessing. |

| ROC-AUC Score Range | Classification | Description |
|---|---|---|
| 0.60 to 0.69 | Fair | The predictive model has limited predictive ability for predicting customer churn and requires improvements to be useful. |
| 0.70 to 0.79 | Good | The predictive model performs well in predicting customer churn and is suitable for simple analysis and retention plans. |
| 0.80 to 0.89 | Very Good | The predictive model is effective for predicting customer churn, supporting detailed segmentation and targeted efforts. |
| 0.90 to 1.00 | Excellent | The predictive model is extremely effective for predicting customer churn, allowing precise retention strategies. |

- **Accuracy:** The accuracy metric computes the proportion of correctly forecasted instances (both churn and non-churn) out of the total cases. A higher accuracy highlights that a predictive model reliably detects customers who will churn and those who will not at Nova Apex Bank. The table below highlights the classification of accuracy value ranges:

| Accuracy Score Range | Classification | Description |
|---|---|---|
| Below 0.50 | Very Poor | The predictive model is ineffective for predicting customer churn due to many errors compared to correct predictions and is unsuitable for decision-making. |

| Accuracy Score Range | Classification | Description |
|---|---|---|
| 0.50 to 0.59 | Poor | The predictive model is unreliable for predicting customer churn due to many errors, possibly resulting in ineffective retention strategies. |
| 0.60 to 0.69 | Fair | The predictive model has moderate accuracy due to modest errors and would need refinements to predict better customers that will churn. |
| 0.70 to 0.79 | Good | The predictive model performs well in predicting customer churn with a reasonable number of errors and is suited for general analysis and retention strategies. |
| 0.80 to 0.89 | Very Good | The predictive model precisely predicts customer churn with few errors, guiding detailed and targeted retention efforts. |
| 0.90 to 1.00 | Excellent | The predictive model is exceptionally reliable for predicting customer churn with few errors, resulting in substantially accurate and effective retention strategies. |

- **F1-Score:** The F1-Score metric balances precision (preciseness of customer churn predictions) and recall (capacity to discover real churn cases) into a single value. While forecasting customer churn at Nova Apex Bank, a bigger F1-score emphasizes that a predictive model is effective in accurately discovering customers that will churn, i.e., recall, while ensuring that most of the predicted churn cases are correct, i.e., precision. In the context of this project, only the class of customers predicted to churn, i.e., class "Yes" from the target variable of *Exited*, are considered. The table below highlights the F1-Score ranges along with their classifications:

| F1-Score Range (Class "Yes") | Classification | Description |
|---|---|---|
| Below 0.50 | Very Poor | The predictive model is ineffective at discovering churned customers. |
| 0.50 to 0.59 | Poor | The predictive model struggles to discover churned customers effectively. |
| 0.60 to 0.69 | Fair | The predictive model is moderately effective in discovering churned customers. |
| 0.70 to 0.79 | Good | The predictive model performs well in discovering customer churn. |
| 0.80 to 0.89 | Very Good | The predictive model reliably discovers customer churn. |
| 0.90 to 1.00 | Excellent | The predictive model is exceptionally effective at discovering customer churn. |

In addition to the evaluation metrics of ROC-AUC, accuracy, and F1-score, feature importance has been utilized within each predictive model to discover the top three factors influencing a customer's decision to stop doing business with Nova Apex Bank. As part of understanding the feature importance of the predictive models, Gini Impurity has been used for the decision tree and random forest. At the same time, coefficients were utilized to determine the importance of all logistic regression models' features. The following points explain the meaning of coefficients and Gini Impurity:

- **Coefficients:** In logistic regression, coefficients represent the strength and direction of the relationship between each predictor variable and the probability of customer churn at Nova Apex Bank. For instance, a positive coefficient for the *Age* variable may hint that older customers are more likely to churn. On the other hand, a negative coefficient for *Balance* could mean that more significant balances decrease the possibility of churn. Any big absolute values indicate more vital predictors.

- **Gini Impurity:** The Gini Impurity estimates how often a randomly selected element would be wrongly labeled if it was randomly labeled according to the distribution of labels within a dataset. From the perspective of customer churn, the Gini Impurity discerns the most notable factors in the decision tree and random forest models by assessing how well each feature, for instance, *Age* and *Balance*, splits the data to forecast churn at Nova Apex Bank.

### 6.1. Logistic Regression

A logistic regression model is a statistical method used for binary classification. It is ideal for forecasting customer churn at Nova Apex Bank by computing the possibility that a customer will leave based on several predictor variables. This project has deployed four types of logistic regression models: full logistic regression, forward logistic regression, backward logistic regression, and stepwise logistic regression. Each model uses a different method to choose and improve predictor variables, improving the accuracy and reliability of the churn predictions.

#### 6.1.1. Hyperparameter Tuning

Even though the grid search was carried out separately for each logistic regression model (full, forward, backward, and stepwise), a common set of parameters was used across all models, as illustrated in Figure 37. The standard parameters tested for all logistic regression models include the regularization strength (C) and the solver algorithm (solver). This ensured consistency in the optimization process while allowing each logistic regression model's performance to be enhanced based on the same criteria.

```
# Define the hyperparameter grid for tuning all logistic regression models

param_grid_logistic_regression = {
    'C': [0.001, 0.05, 0.5, 5, 50],  # Regularization strength
    'solver': ['liblinear', 'saga']  # Algorithms to use in the optimization problem
}
```

*Figure 37: Hyperparameter Grid for Tuning All Logistic Regression Models*

### 6.1.2. Full Logistic Regression

A full logistic regression model uses all available predictor variables to evaluate their collective impact on the outcome, in this case, *Exited*. This comprehensive approach removes no potentially essential variable from the analysis, offering a thorough understanding of the factors influencing churn at Nova Apex Bank.

#### 6.1.2.1. Model Performance

| Model | ROC-AUC | Accuracy | F1-score (Class "Yes") |
|---|---|---|---|
| Full Logistic Regression | 0.79 | 0.82 | 0.44 |

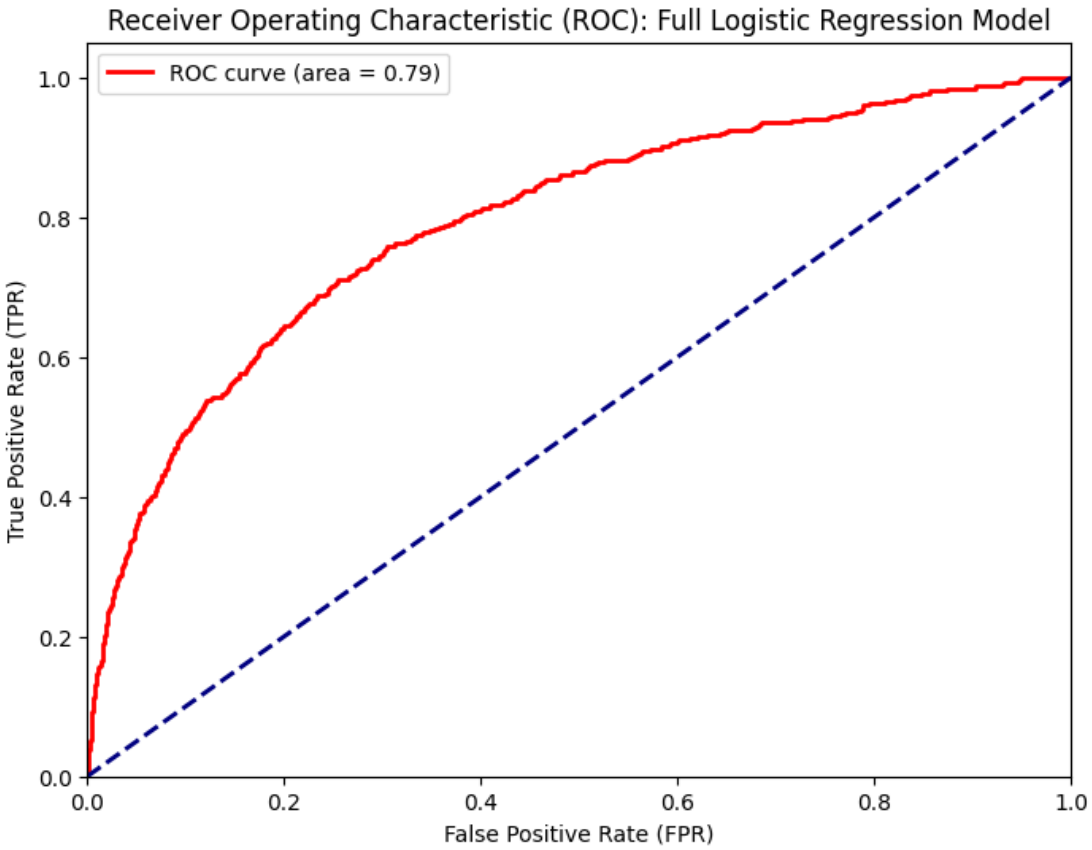


*Figure 38: ROC-AUC chart of the Full Logistic Regression Model*

When assessing the performance of the full logistic regression model, a good ROC-AUC score of 0.79 (Figure 38) and a very good accuracy value of 0.82 have been achieved. Nonetheless, the full logistic regression model has obtained an F1-Score of 0.44 for the "Yes" class of the *Exited* variable, representing a very poor figure in that category.

| | **Predicted Churn** | |
|---|---|---|
| **Actual Churn** | No | Yes |
| No | 2,272 | 101 |
| Yes | 424 | 203 |

The precision score of 67% indicates the accuracy of the full logistic regression model in forecasting customer churn and a recall figure of 32% reflects its capacity to identify actual churn scenarios.

$$\text{Precision ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$= \frac{203}{203 + 101} = \frac{203}{304} = 0.67$$

$$\text{Recall ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FP)}}$$

$$= \frac{203}{203 + 624} = \frac{203}{627} = 0.32$$

### 6.1.2.2. Feature Importance

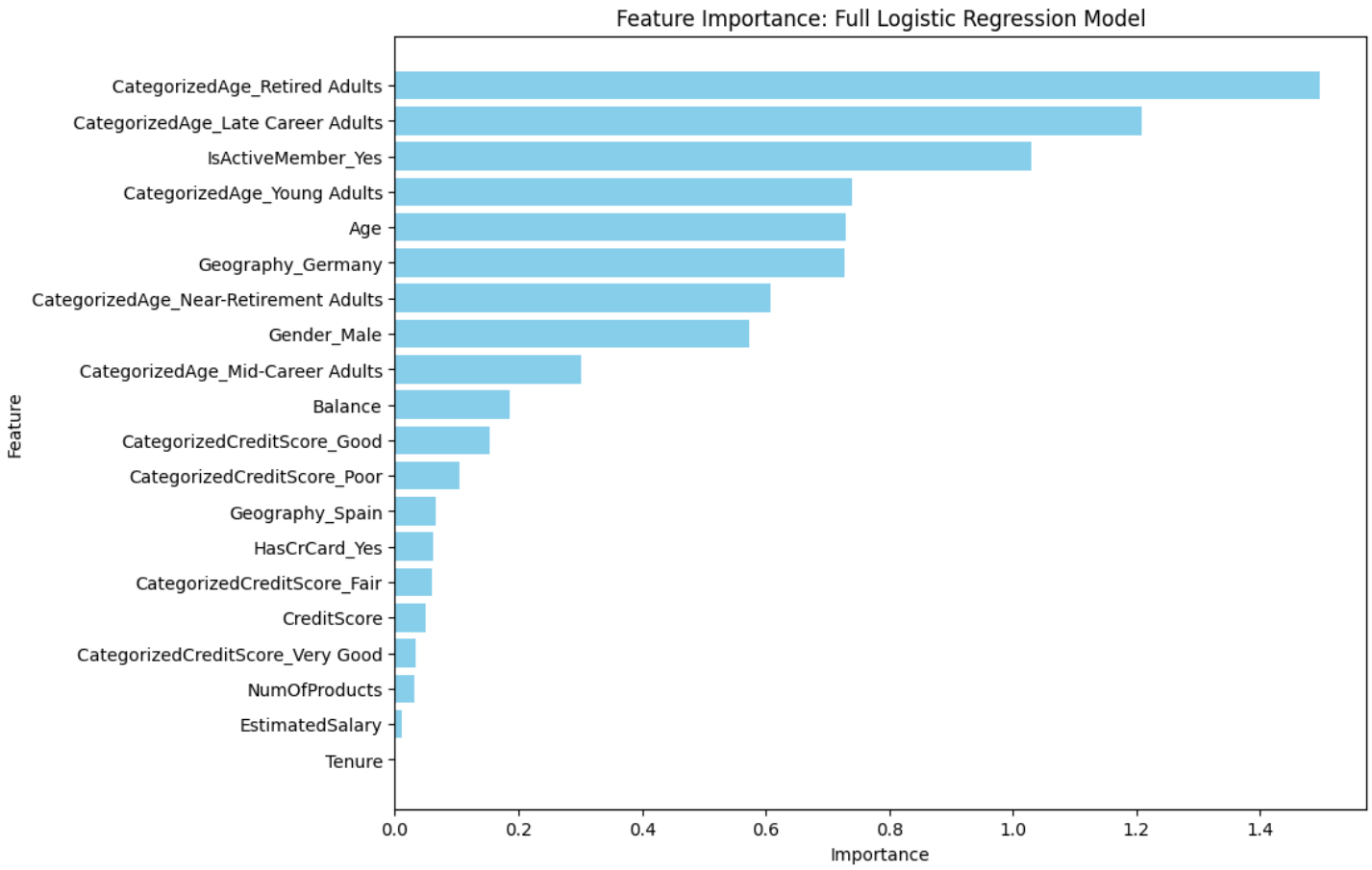


*Figure 39: Feature Importance Chart of the Full Logistic Regression Model*

While considering the feature importance of the full logistic regression model, the top three essential features obtained from the model, based on the coefficients (Figure 39), include:

- Retired adult customers, containing individuals above the age of 65 years
- Late career adults, constituting individuals between 55 to 64 years, and
- Customers that are active members

### 6.1.3. Forward Logistic Regression

A forward logistic regression model begins with no predictors and adds variables one at a time according to their statistical significance. The process continues until no further variables refine the forward logistic regression model, ensuring that only the most influential predictors are considered for forecasting customer churn at Nova Apex Bank.

#### 6.1.3.1. Model Performance

| Model | ROC-AUC | Accuracy | F1-score (Class "Yes") |
|---|---|---|---|
| Forward Logistic Regression | 0.76 | 0.82 | 0.42 |

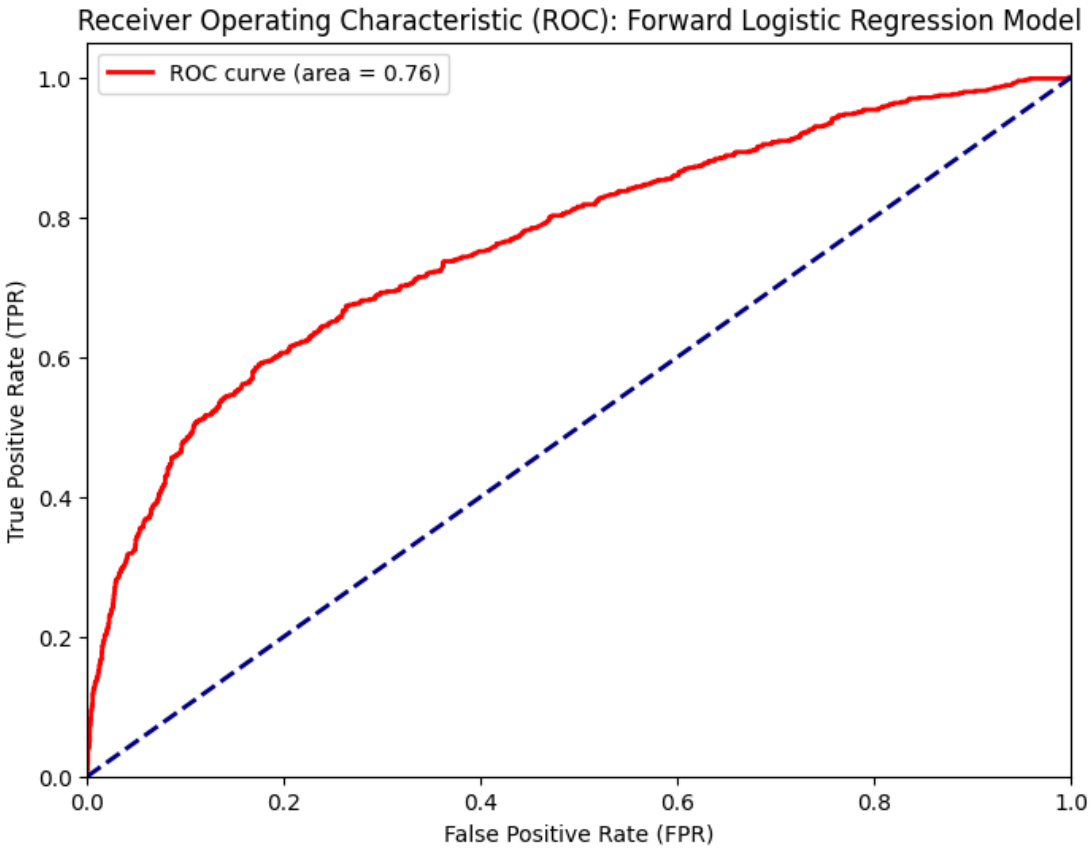


*Figure 40: ROC-AUC chart of the Forward Logistic Regression Model*

The performance of the forward logistic regression model reveals that a good AUC-ROC score of 0.76 (Figure 40) has been attained. Moreover, its accuracy stands at 0.82, implying that it has achieved a very good accuracy yet represents the lowest accuracy among all models developed. However, the F1-score of the forward logistic regression model stands at 0.42 for the "Yes" class of the *Exited* variable, indicating a very poor performance within this component.

| | **Predicted Churn** | |
|---|---|---|
| **Actual Churn** | No | Yes |
| No | 2,283 | 90 |
| Yes | 438 | 189 |

Within the forward logistic regression model, a precision value of 68% emphasizes how precisely the model predicts customer churn. In comparison, a 32% recall score emphasizes the model's proficiency in detecting real churn cases.

$$\text{Precision ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$= \frac{189}{189 + 90} = \frac{189}{279} = 0.68$$

$$\text{Recall ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FP)}}$$

$$= \frac{189}{189 + 438} = \frac{189}{627} = 0.30$$

### 6.1.3.2. Feature Importance

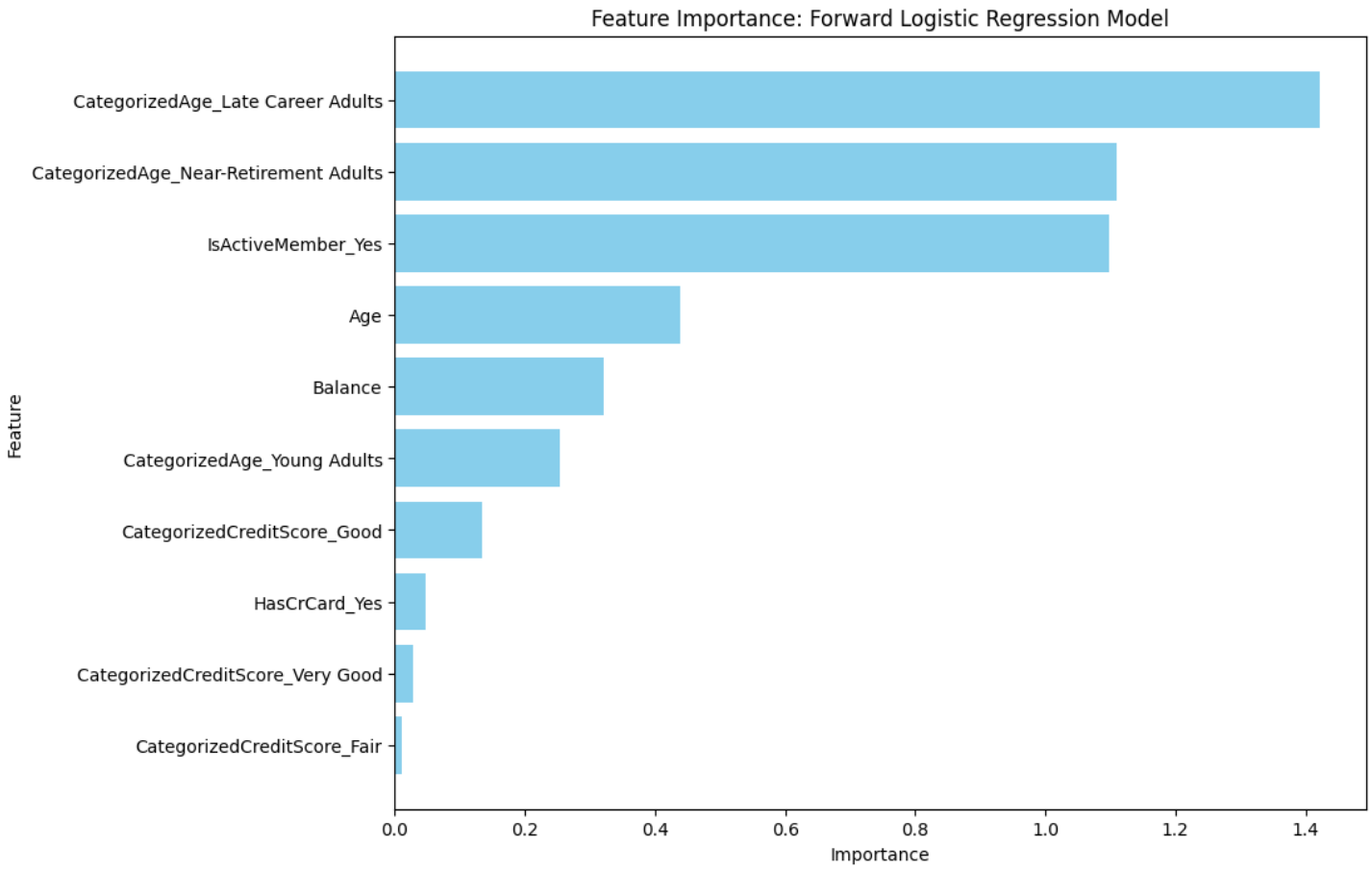


*Figure 41: Feature Importance Chart of the Forward Logistic Regression Model*

The top three features derived from the feature importance of the forward logistic regression model using the coefficients (Figure 41) are:

- Late career adults, where individuals are aged are between 55 years to 64 years
- Near retirement adults, containing individuals aged between 65 years and above, and
- Customers that are engaged members

### 6.1.4. Backward Logistic Regression

Unlike a forward logistic regression model, a backward logistic regression model starts with all predictors and omits the least significant variables one by one. This continues until only the statistically significant variables remain, with the model only including the most appropriate predictors for forecasting customer churn at Nova Apex Bank.

#### 6.1.4.1. Model Performance

| Model | ROC-AUC | Accuracy | F1-score (Class "Yes") |
|---|---|---|---|
| Backward Logistic Regression | 0.79 | 0.83 | 0.45 |

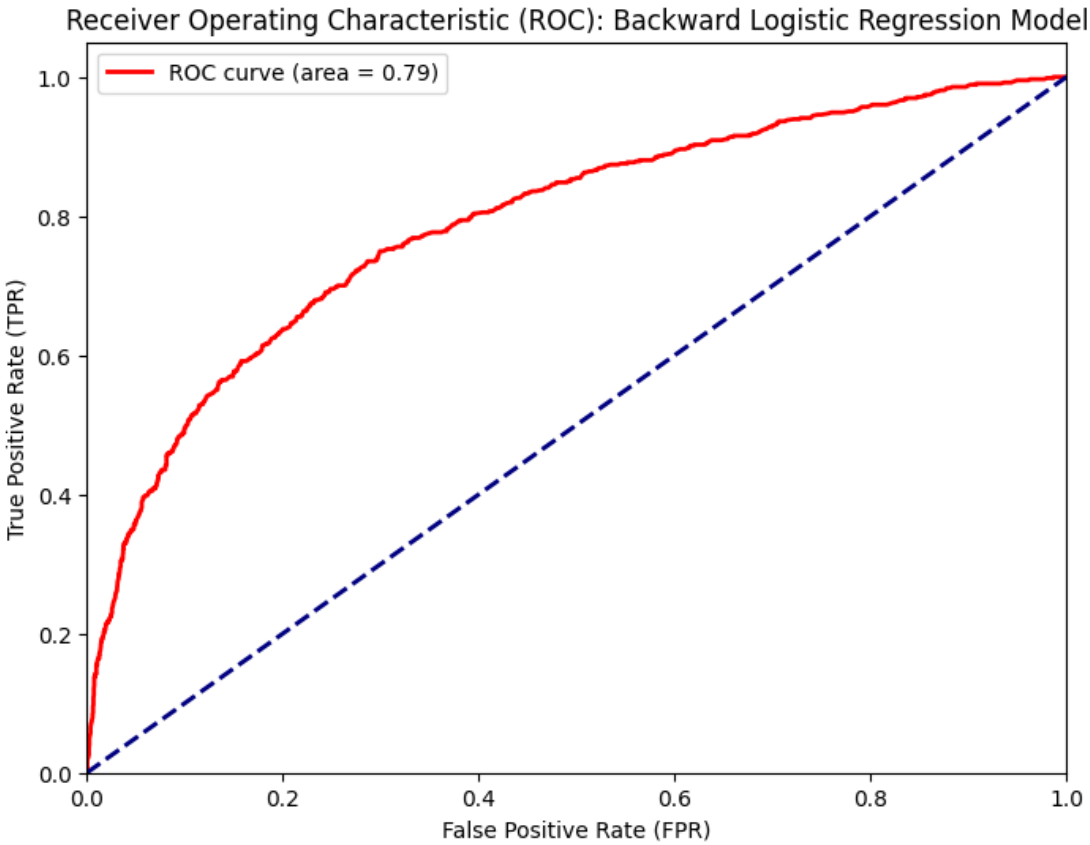


*Figure 42: ROC-AUC Chart of the Backward Logistic Regression Model*

The performance assessment of the backward logistic regression model highlights that it has achieved a good ROC-AUC score of 0.79 (Figure 42) and a very good accuracy of 0.83. However, the backward logistic regression model has generated a very poor F1-Score of 0.45 for the "Yes" class of the *Exited* variable despite being the highest F1-Score among any logistic regression model.

| | **Predicted Churn** | |
|---|---|---|
| **Actual Churn** | No | Yes |
| No | 2,282 | 91 |
| Yes | 421 | 206 |

As part of the backward logistic regression model, a precision value of 69% exhibits the model's effectiveness in forecasting customer churn, whereas a 33% recall score demonstrates the model's ability to ascertain actual churn cases.

$$\text{Precision ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$= \frac{206}{206 + 91} = \frac{206}{297} = 0.69$$

$$\text{Recall ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FP)}}$$

$$= \frac{206}{206 + 421} = \frac{206}{627} = 0.33$$

### 6.1.4.2. Feature Importance

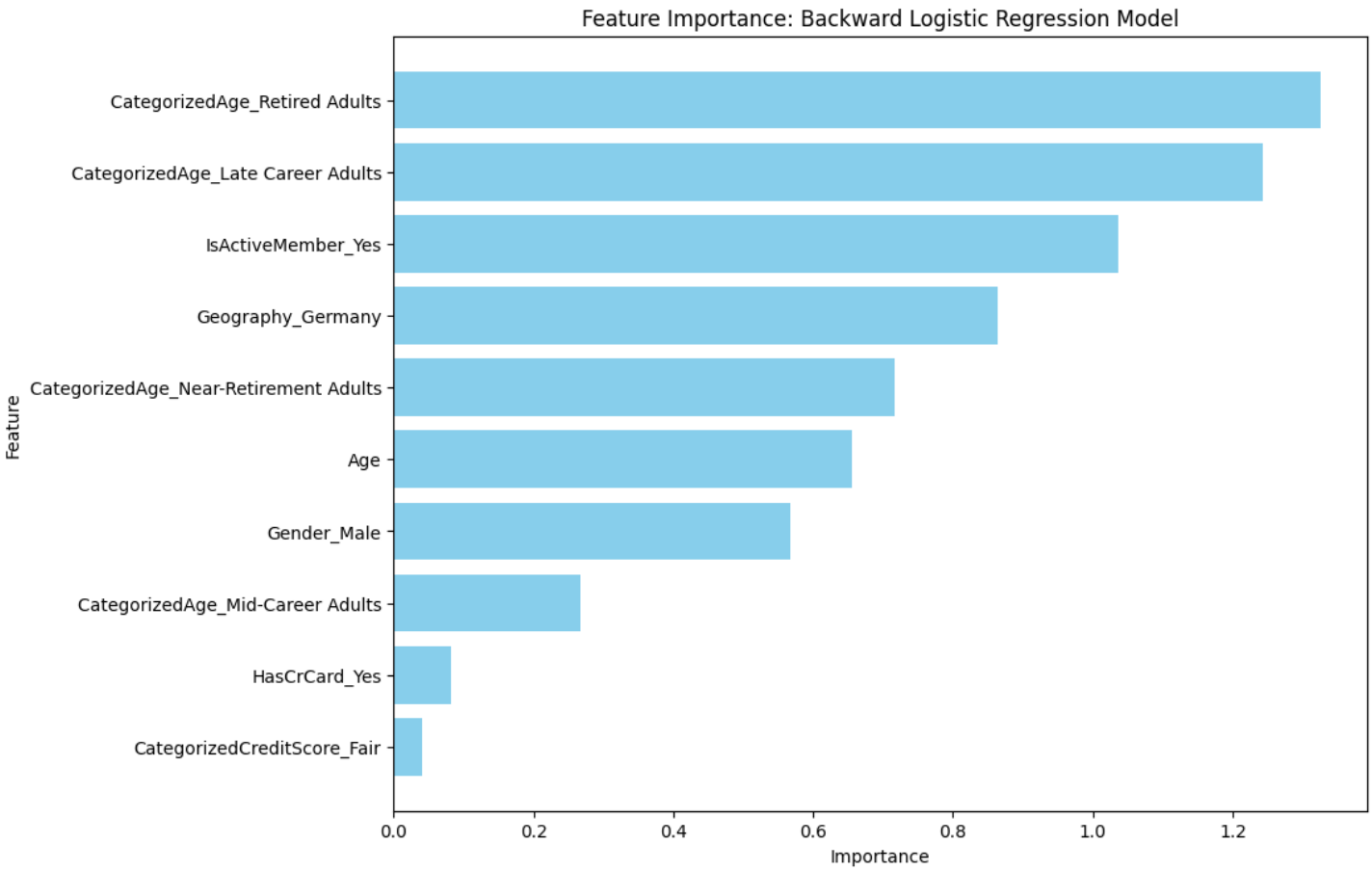


*Figure 43: Feature Importance Chart of the Backward Logistic Regression Model*

When comprehending the feature importance of the backward logistic regression model through the coefficients (Figure 43), the primary three features of the model are:

- Retired adults, i.e. customers aged 65 years and above
- Late career adults, comprising customers aged between 45 years and 54 years, and
- Customers that are active members

### 6.1.5. Stepwise Logistic Regression

Within a stepwise logistic regression model, both forward selection and backward elimination methods used in the forward logistic regression model and the backward logistic regression model, respectively, are merged. Variables are included or eliminated based on their significance until only the most relevant predictors remain, improving the stepwise logistic regression model iteratively for predicting customer churn at Nova Apex Bank.

#### 6.1.5.1. Model Performance

| Model | ROC-AUC | Accuracy | F1-score (Class "Yes") |
|---|---|---|---|
| Stepwise Logistic Regression | 0.76 | 0.83 | 0.42 |

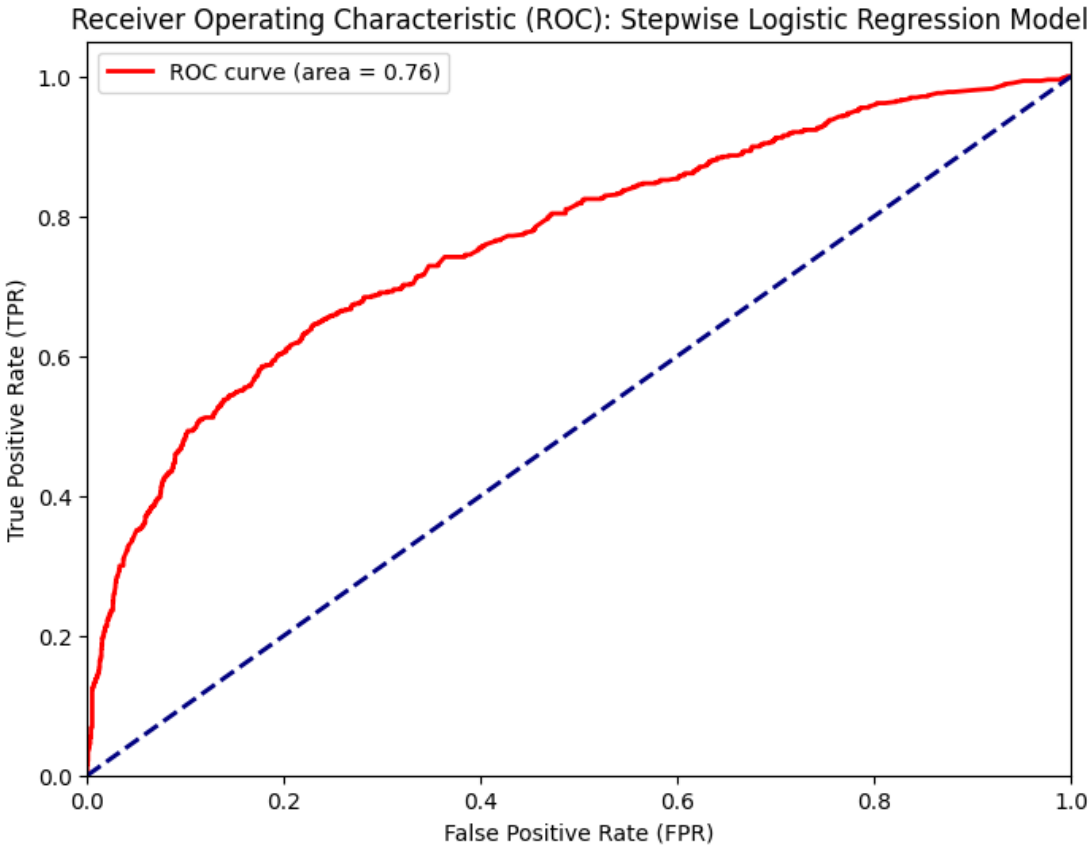


*Figure 44: ROC-AUC Chart of the Stepwise Logistic Regression Model*

Like the other logistic regression models, the stepwise logistic regression model has obtained a good ROC-AUC score of 0.76 (Figure 44), followed by a very good accuracy value of 0.83, the latter being the highest accuracy of any logistic regression model. Nonetheless, the stepwise logistic regression also produces a very poor F1-score of 0.42 for the "Yes" class for the *Exited* variable, just like the other logistic regression models.

| | **Predicted Churn** | |
|---|---|---|
| **Actual Churn** | No | Yes |
| No | 2,290 | 83 |
| Yes | 439 | 188 |

The stepwise logistic regression model has a computed precision figure of 69%, highlighting its ability to predict customer churn precisely. A 30% recall score illustrates the model's capability to capture real churn scenarios.

$$\text{Precision ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$= \frac{188}{188 + 83} = \frac{188}{271} = 0.69$$

$$\text{Recall ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FP)}}$$

$$= \frac{188}{188 + 439} = \frac{206}{627} = 0.30$$

### 6.1.5.2. Feature Importance

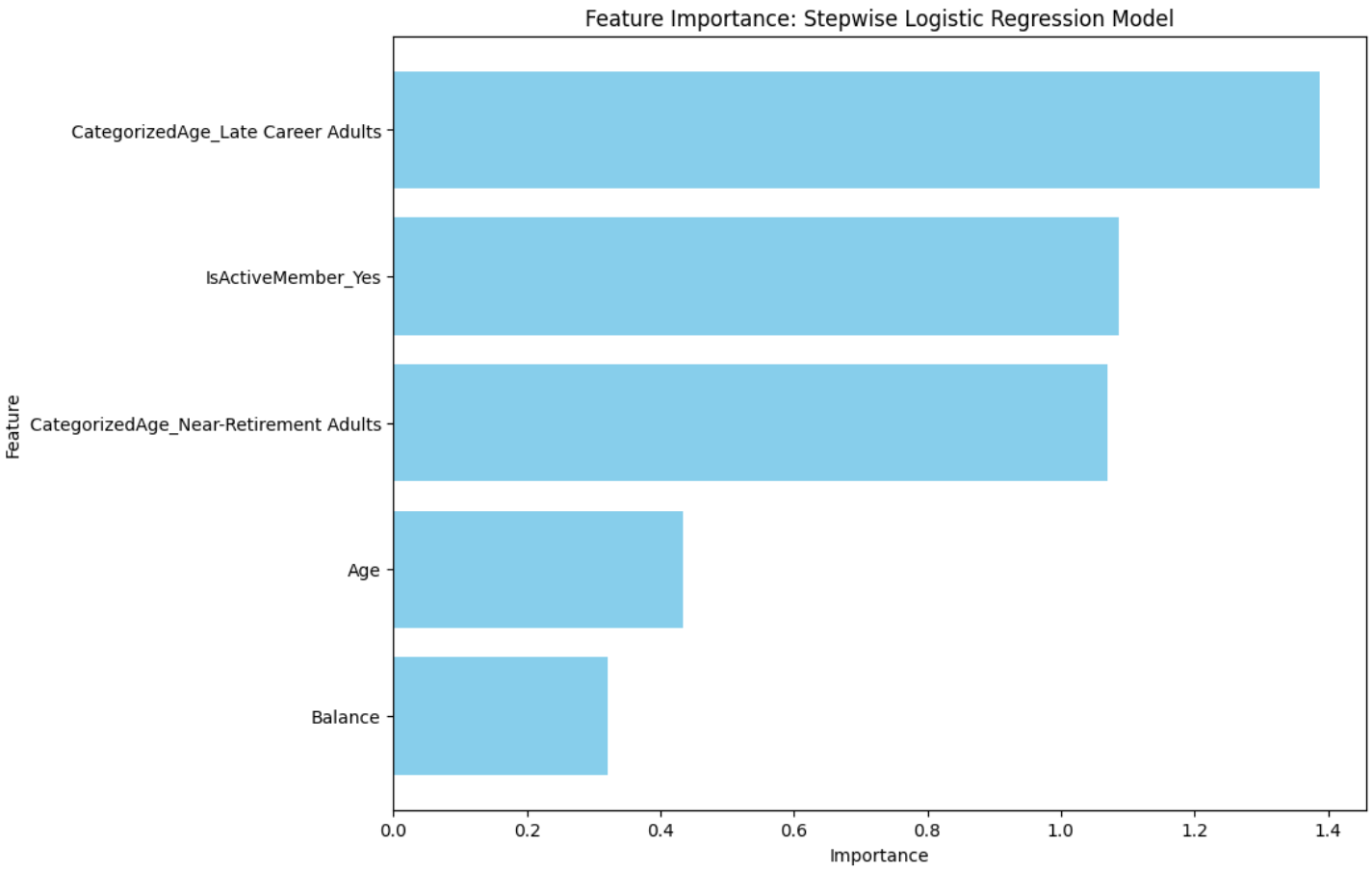


*Figure 45: Feature Importance Chart of the Stepwise Logistic Regression Model*

Based on the coefficients, the analysis of the feature importance of the stepwise logistic regression model reveals three key features (Figure 45), including:

- Late career adults, comprising customers aged between 45 years and 54 years
- Engaged customers as members, and
- Near retirement adults, constituting customers aged between 55 years and 64 years

### 6.2. Decision Tree

A decision tree is a predictive model that forecasts customer churn at Nova Apex Bank by splitting data into different branches based on features. A decision tree discerns the vital factors causing customer churn at Nova Apex Bank by showing decision paths and outcomes, without necessarily using visual representation.

#### 6.2.1. Hyperparameter Tuning

For the decision tree model, a grid search has been used to fine-tune the model's performance. As highlighted in Figure 46, this process involved testing different combinations of key settings, i.e. parameters, such as how deep the tree can grow (max_depth), how many customers need to be in a group before it's split (min_samples_split), and the minimum number of customers required at the end of a branch (min_samples_leaf).

```
# Define the hyperparameter grid for tuning the decision tree model

param_grid_decision_tree = {
    'max_depth': [2, 3, 6, 9],  # Maximum depth of the tree
    'min_samples_split': [3, 4, 6],  # Minimum number of samples required to split an internal node
    'min_samples_leaf': [2, 9, 10]  # Minimum number of samples required to be at a leaf node
}
```

*Figure 46: Hyperparameter Grid for Tuning the Decision Tree Model*

### 6.2.2. Model Performance

| Model | ROC-AUC | Accuracy | F1-score (Class “Yes”) |
|---|---|---|---|
| Decision Tree | 0.85 | 0.86 | 0.58 |

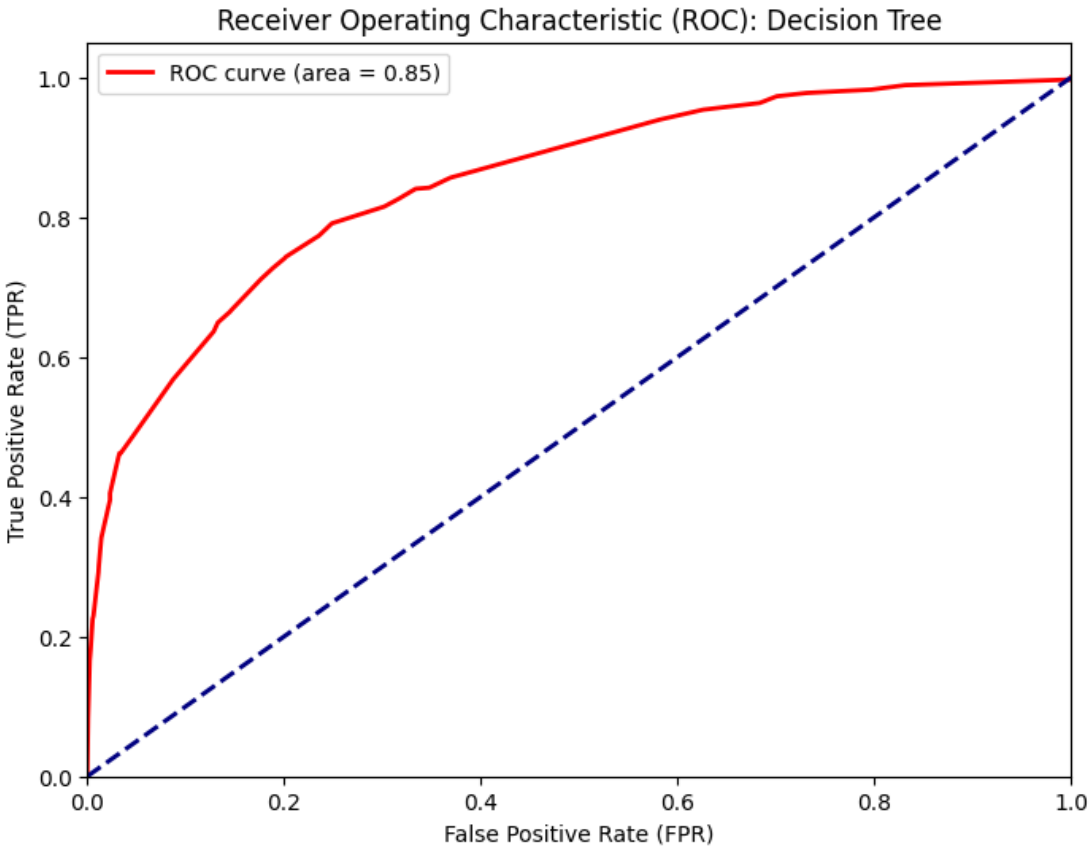


*Figure 47: ROC-AUC Chart of the Decision Tree Model*

The analysis of the performance of the decision tree model reveals that an ROC-AUC score of 0.85 (Figure 47) has been achieved, representing a very good ROC-AUC score. Furthermore, an accuracy value of 0.86 has been generated by the decision tree model; this represents a joint-highest accuracy score among all predictive models. Unlike the logistic regression models, the decision tree shows a marginally improved performance concerning its F1-Score with a poor score of 0.58 for the "Yes" class of the *Exited* variable, the highest value among all prediction models despite a below-satisfactory score.

| | Predicted Churn | |
|---|---|---|
| **Actual Churn** | No | Yes |
| No | 2,295 | 78 |
| Yes | 337 | 290 |

The decision tree model has obtained a precision score of 79%, which shows its capability to predict customer churn precisely. On the other hand, a 46% recall score describes the decision tree model's proficiency in discerning actual churn cases.

$$\text{Precision ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP) + False Positives (FP)}}$$

$$= \frac{290}{290 + 78} = \frac{290}{368} = 0.79$$

$$\text{Recall ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP) + False Negatives (FP)}}$$

$$= \frac{290}{290 + 337} = \frac{290}{627} = 0.46$$

### 6.2.3. Feature Importance

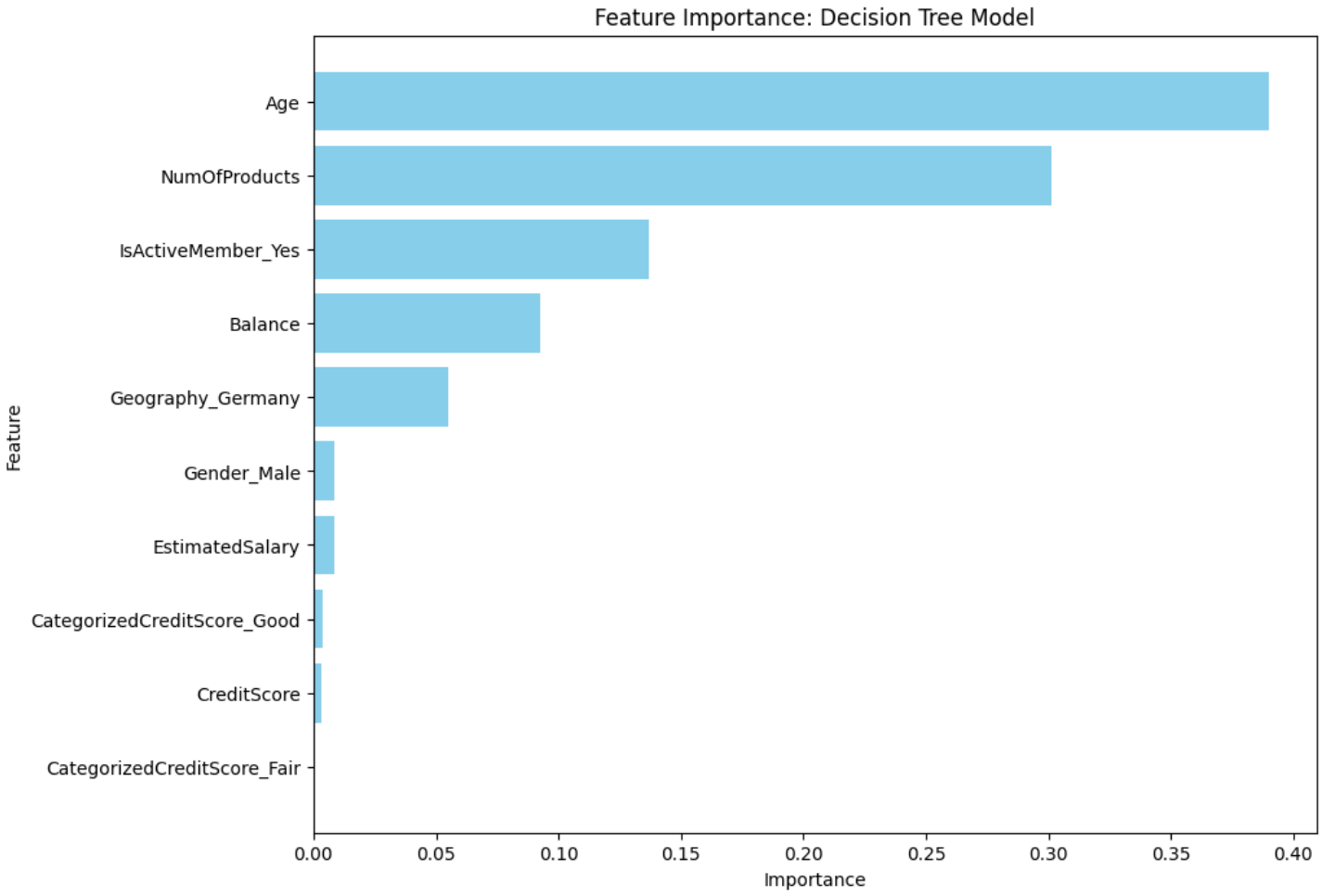


*Figure 48: Feature Importance Chart of the Decision Tree Model*

As part of the decision tree model, the top three features identified from the feature importance of the model based on the Gini Impurity (Figure 48) include:

- Ages of customers, regardless of the age range they belong to
- Number of products held by customers, and
- Customers that are active members

### 6.3. Random Forest

A random forest model forecasts customer churn at Nova Apex Bank using multiple decision trees. Each tree uses a subset of data, and the final model averages their results for a more precise prediction. This enables the discovery of key factors influencing customer churn at Nova Apex Bank.

#### 6.3.1. Hyperparameter Tuning

A grid search has been used within the random forest model to optimize its performance by experimenting with various parameters, as depicted in Figure 49. In this regard, the various combinations of the parameters that were experimented with constitute the total number of trees in the model (n_estimators), the maximum levels the trees could reach (max_depth), the minimum customer count needed to create a split (min_samples_split), and the smallest group size allowed to form a final branch (min_samples_leaf).

```
# Define the hyperparameter grid for tuning the random forest model

param_grid_random_forest = {
    'n_estimators': [50, 150, 175],  # Number of trees in the forest
    'max_depth': [None, 15, 19],  # Maximum depth of the tree
    'min_samples_split': [4, 6, 9],  # Minimum number of samples required to split an internal node
    'min_samples_leaf': [1, 3, 10]  # Minimum number of samples required to be at a leaf node
}
```

*Figure 49: Hyperparameter Grid for Tuning the Random Forest Model*

### 6.3.2. Model Performance

| Model | ROC-AUC | Accuracy | F1-score (Class “Yes”) |
|---|---|---|---|
| Random forest | 0.86 | 0.86 | 0.57 |

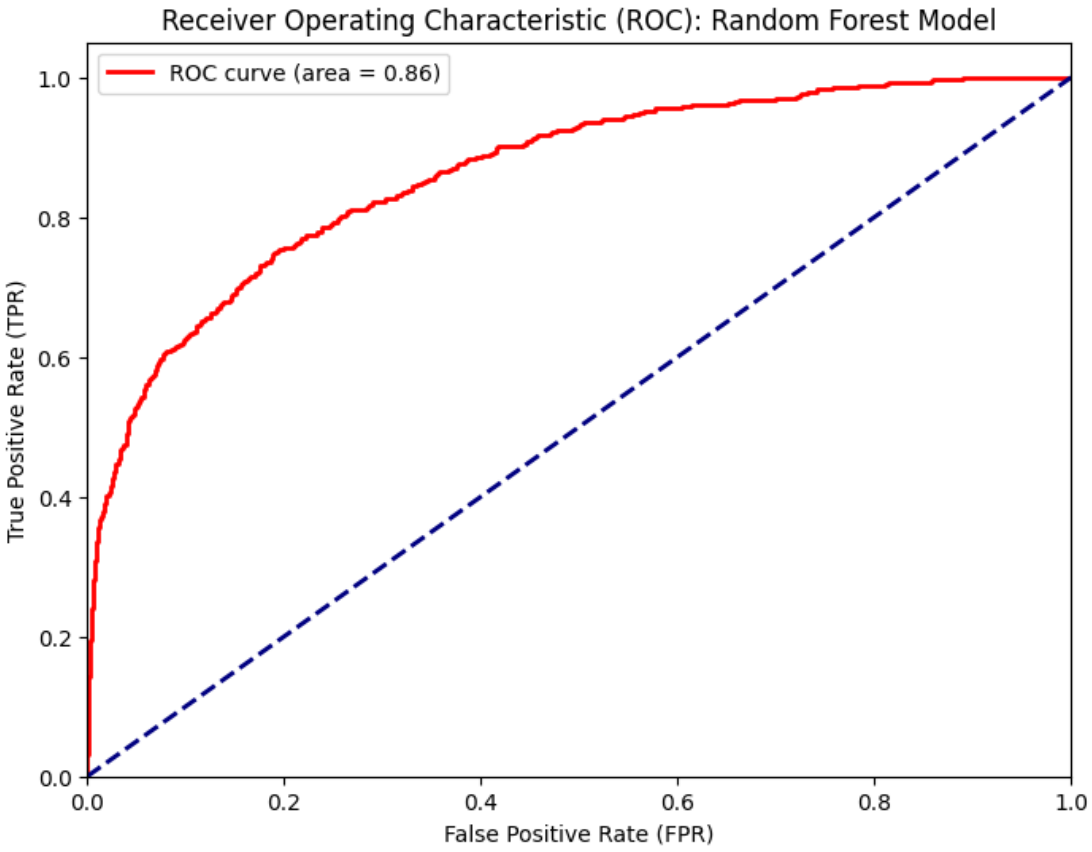


*Figure 50: ROC-AUC Chart of the Random Forest Model*

Upon evaluating the performance of the random forest model, the ROC-AUC score is found out to be a very good value of 0.86 (Figure 50); this represents the largest value than any other predictive model in that component. Similarly, a joint-highest very good accuracy of 0.86 has been achieved by the random forest model. Nevertheless, like every other predictive model demonstrating a very poor or poor in its F1-Score, the random forest model has obtained the latter classification due to a 0.57 F1-Score for the "Yes" class for the *Exited* variable, just falling short of the F1-Score of 0.58 produced by the decision tree model.

| | Predicted Churn | |
|---|---|---|
| **Actual Churn** | No | Yes |
| No | 2,300 | 73 |
| Yes | 346 | 281 |

As part of the random forest model, a precision figure of 79% has been obtained, which shows the preciseness of the model in accurately forecasting customer churn. Conversely, a 45% recall score explains the capacity of the random forest model to detect real churn cases.

$$\text{Precision ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$= \frac{281}{281 + 73} = \frac{281}{354} = 0.79$$

$$\text{Recall ("Yes" class)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FP)}}$$

$$= \frac{281}{281 + 346} = \frac{281}{627} = 0.45$$

### 6.3.3. Feature Importance

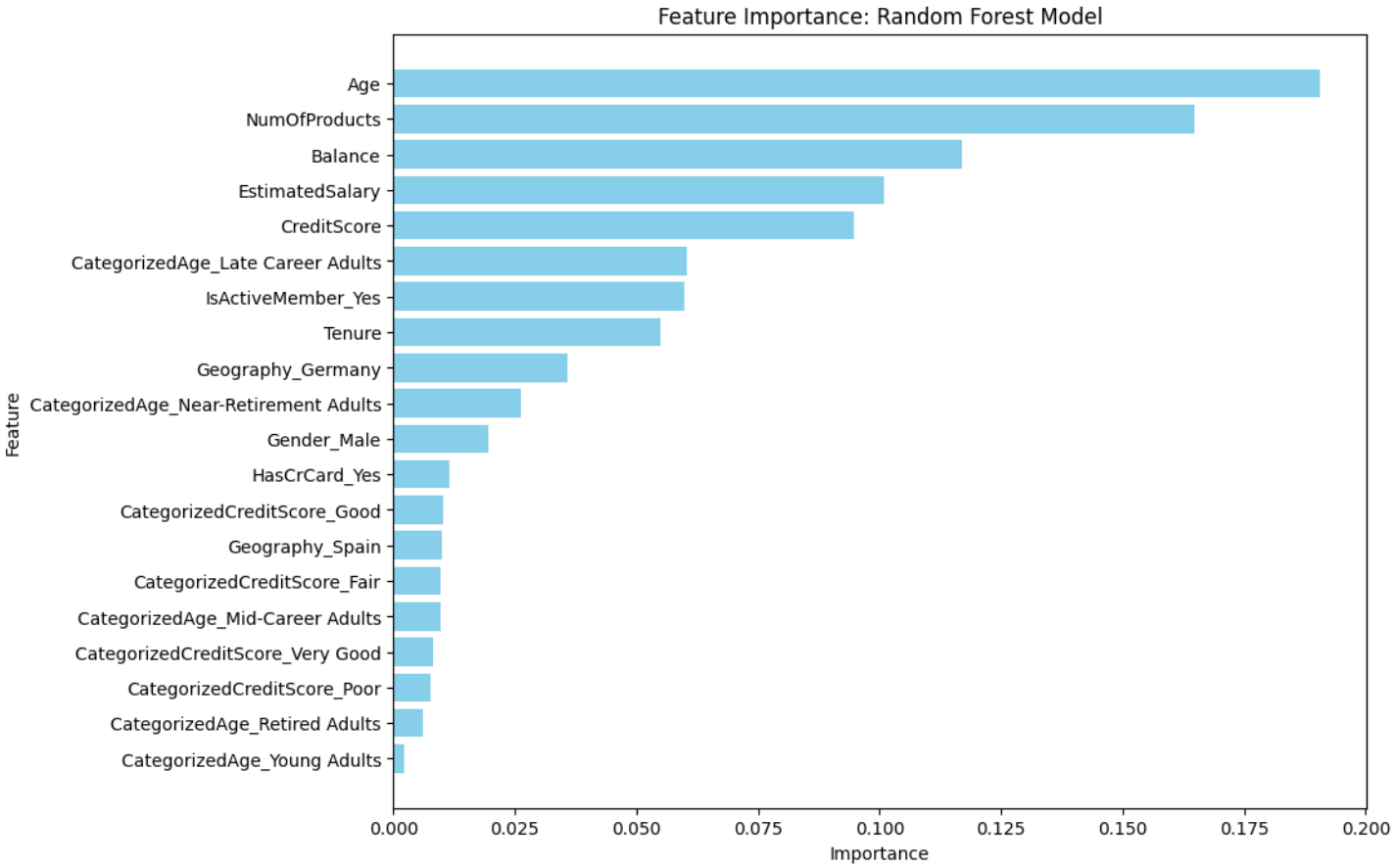


*Figure 51: Feature Importance Chart of the Random Forest Model*

Based on the Gini Impurity, the top three features extracted from the feature importance of the random forest model (Figure 51) contain:

- Age of customers, irrespective of the age range that an individual belongs to
- Number of products possessed by customers, and
- Balance held by customers in their accounts

# 7. Model Comparison

To identify the best-performing predictive model among logistic regression (full, forward, backward, stepwise), decision tree, and random forest for forecasting customer churn at Nova Apex Bank, three pre-established evaluation metrics were utilized in order of prioritization: ROC-AUC score, accuracy, and F1-score. These criteria were previously highlighted at the start of the model exploration stage.

| Model | ROC-AUC | Accuracy | F1-score (Class "Yes") |
|---|---|---|---|
| Random forest | 0.86 | 0.86 | 0.57 |
| Decision Tree | 0.85 | 0.86 | 0.58 |
| Backward Logistic Regression | 0.79 | 0.83 | 0.45 |
| Full Logistic Regression | 0.79 | 0.82 | 0.44 |
| Stepwise Logistic Regression | 0.76 | 0.83 | 0.42 |
| Forward Logistic Regression | 0.76 | 0.82 | 0.42 |

### 7.1. Analysis of Best Model

Upon applying the pre-established criteria to the formulated prediction models, the random forest model has emerged as the best-performing predictive model for predicting customer churn at Nova Apex Bank. The following points summarize the overall performance of the random forest model:

- Firstly, the random forest model has obtained a ROC-AUC score of 0.86, demonstrating its effectiveness in distinguishing between churned and non-churned customers at Nova Apex Bank as opposed to the other predictive models.

- Secondly, an accuracy of 0.86 has been generated by the random forest model, highlighting its soundness in forecasting correct predictions. A larger accuracy enables the random forest model to effectively detect churned and non-churned customers of Nova Apex Bank, thereby supporting the firm in making informed decisions.

- Lastly, the random forest model has produced an F1-score of 0.57 for predicting customer churn at Nova Apex Bank. From this information, it can be deduced that the random forest model achieves a moderate balance between precision (0.79) and recall (0.45), meaning it is somewhat effective at correctly identifying customers who will churn, i.e., recall while ensuring that a fair proportion of its churn predictions are accurate, i.e., precision.

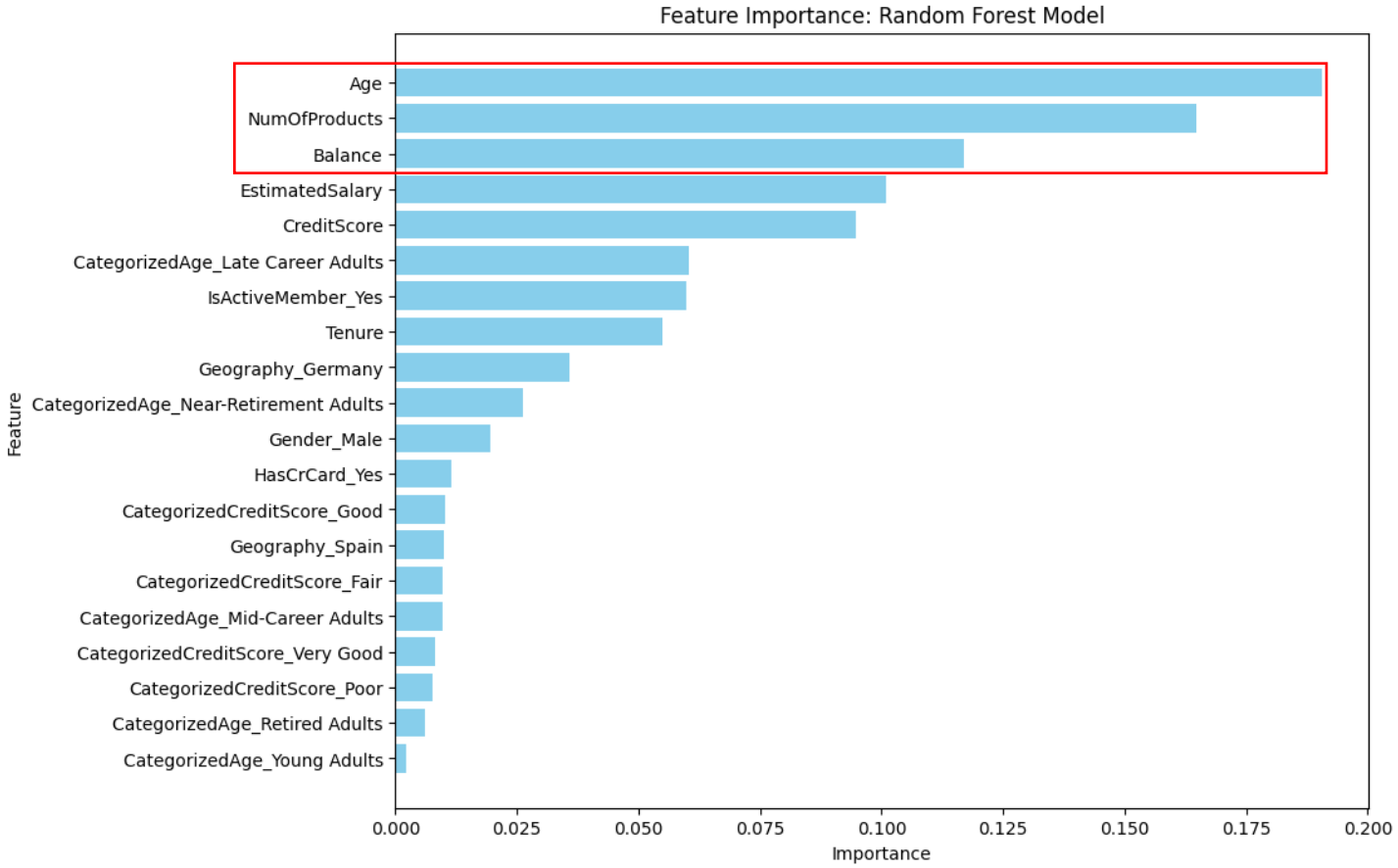


*Figure 52: Feature Importance Chart of the Random Forest Model*

| | Feature | Importance (Gini Impurity) |
|---|---|---|
| 0 | Age | 0.190698 |
| 4 | NumOfProducts | 0.164777 |
| 1 | Balance | 0.117053 |
| 2 | EstimatedSalary | 0.100888 |
| 3 | CreditScore | 0.094838 |
| 9 | CategorizedAge_Late Career Adults | 0.060458 |
| 19 | IsActiveMember_Yes | 0.059714 |
| 5 | Tenure | 0.054995 |
| 7 | Geography_Germany | 0.035899 |
| 11 | CategorizedAge_Near-Retirement Adults | 0.026268 |
| 6 | Gender_Male | 0.019433 |
| 18 | HasCrCard_Yes | 0.011651 |
| 15 | CategorizedCreditScore_Good | 0.010125 |
| 8 | Geography_Spain | 0.010078 |
| 14 | CategorizedCreditScore_Fair | 0.009665 |
| 10 | CategorizedAge_Mid-Career Adults | 0.009597 |
| 17 | CategorizedCreditScore_Very Good | 0.008060 |
| 16 | CategorizedCreditScore_Poor | 0.007543 |
| 12 | CategorizedAge_Retired Adults | 0.006026 |
| 13 | CategorizedAge_Young Adults | 0.002232 |

*Figure 53: Feature Importance Values of the Random Forest Model using the Gini Impurity*

As illustrated in Figure 52 and Figure 53, the random forest model has detected the age of customers, number of products held by customers and balance of customers in their bank accounts as the top three features affecting customer churn at Nova Apex Bank using their Gini Impurity scores. These features are vital to comprehend and predict the likelihood of customers ceasing to do business with Nova Apex Bank and have been explained by the following points:

- **Age of customers (0.19):** The age of customers stands out as the most influential factor for forecasting customer churn at Nova Apex Bank because of the highest Gini Impurity value of nearly 0.19 within the random forest model. In this regard, the likelihood of customers quitting their association with Nova Apex Bank is strongly correlated with their age, representing a critical factor for the random forest model's decision-making process.

- **Number of products held by customers (0.16):** After age, the number of products possessed by customers of Nova Apex Bank indicates the second essential feature of the random forest model for forecasting the customer churn at Nova Apex Bank due to a Gini Impurity score of approximately 0.16. Customers that have different levels of products held have differing risks of quitting their relationship with Nova Apex Bank, entailing the importance of this feature within the random forest model.

- **Balance of customers in their accounts (0.11):** The balance of customers held in their accounts portrays a third essential feature of the random forest model for predicting customer churn at Nova Apex Bank. From this information, it can be inferred that the amount of money maintained by a customer is also a vital indicator of their possibility to stop doing business with Nova Apex Bank despite being less influential than the other features of age and balance.

# 8. Conclusion and Recommendations

To address the business problem of tackling customer churn at Nova Apex Bank, past historical data of the organization has been analyzed to comprehend customer behavior at the company through descriptive analytics. Based on the historical data analysis, predictive analytics has been carried out by developing several predictive models, including logistic regression (full, forward, backward, stepwise), decision tree and random forest, to understand further what is likely to cause customers to cease their association with Nova Apex Bank.

Upon evaluation of all predictive models, the random forest model has been identified as the best model out of any predictive model due to larger scores in various evaluation metrics, constituting a ROC-AUC score and accuracy score of 0.86 each for the overall model, along with an F1-score specifically for churned customers (0.57). Based on the feature importance of the random forest model that utilizes Gini Impurity, the essential features that have been identified as predictors of customer churn are age, number of products held and the bank balance of an individual.

## 8.1. Impacts on the Business Problem

Various inferences have been identified by considering three aspects of customer behavior, including age, number of products held and bank balance of customers, as revealed by the random forest and supported by the historical data analysis that may influence customer decision to cease their association with Nova Apex Bank. The following points highlight each of these interpretations:

- **Age of customers:** The age of customers represents the most significant driver behind customer churn at Nova Apex Bank. The analysis indicates that customers in the middle to late stages of their careers are the largest customer churn group at Nova Apex Bank. While mid-career adults, aged between 35 years and 44 years, have the largest number of churns, this is followed by late career adults, aged between 45 to 54 years.

- **Number of products held by customers:** Another essential facet that affects customer churn at Nova Apex Bank is the number of products possessed by an individual, ranging from 1 to 4 products. From the data, it can be suggested that customers holding 1 product or 2 products are more likely to cease their association with Nova Apex Bank than those possessing 3 or 4 products.

- **Balance of customers in their accounts:** Besides the customers' ages and number of products held by the customer, the balance of customers within their Nova Apex Bank account represents another factor behind customer churn at the company. Based on the analysis, it has been discerned that customers with larger bank balances are more prone to stop doing business with Nova Apex Bank.

### 8.2. Recommended Next Steps

From the historical data analysis and predictions carried out, various retention strategies may be recommended to minimize customer churn at Nova Apex Bank based on the key churn drivers of the customers' ages, the number of products held by customers, and the balance of customers. In this way the overall business performance and customer satisfaction at Nova Apex Bank would substantially increase. The following points have emphasized each of these proposed retention strategies:

- **Tailored offerings for different age groups:** At one end, targeted marketing campaigns focusing on innovative and technology-driven banking solutions, such as cashback on digital transactions, student loan discounts and access to financial management applications, may be developed for younger customers of Nova Apex Bank. On the other hand, older customers of Nova Apex Bank may rely on stability and personalized services, implying that the company could offer devoted customer service, individualized financial guidance and products focused on retirement planning and wealth management.

- **Cross-selling through bundling for customers with limited products:** To address the risk of churn among customers holding a minimal number of products, Nova Apex Bank may implement cross-selling strategies to encourage customers to obtain multiple products by providing bundled packages with appealing discounts. For example, customers' savings accounts may be unified with investment products or insurance plans. In this regard, Nova Apex Bank should clearly state the benefits to its customers of utilizing multiple products, leading to deeper customer engagement with the institution.

- **Exclusive benefits for customers with higher balances:** To decrease churn among customers with higher bank balances, a variety of targeted incentive programs, including preferential interest rates and exclusive access to premium banking services may be offered to such individuals. It is vital for Nova Apex Bank to recognize the loyalty of its customers by rewarding them with special benefits, enabling the customers to feel valued and appreciated.

# 9. References


Fair Isaac Corporation. (2024). *What is a FICO® Score?*. https://www.myfico.com/credit-education/what-is-a-fico-score

Glassbox. (2023, September 27). *Customer retention for banks: 5 ways to boost loyalty and lifetime value*. https://www.glassbox.com/blog/customer-retention-in-banking

Hall, M. (2023, June 1). *How the Banking Sector Impacts Our Economy*. Investopedia. https://www.investopedia.com/ask/answers/032315/what-banking-sector.asp

Hayes, A. (2024, July 21). *What is a FICO Score?*. Investopedia. https://www.investopedia.com/terms/f/ficoscore.asp

Investopedia. (2024, March 21). *Churn Rate: What It Means, Examples, and Calculations*. https://www.investopedia.com/terms/c/churnrate.asp

Lake, R. (2023, March 20). *FICO Credit Scores Explained*. https://www.investopedia.com/fico-credit-scores-explained-5072985

Lake, R. and Strohm, M. (2024, July 31). What Is A Bank And How Does It Work?. *Forbes*. https://www.forbes.com/advisor/in/banking/how-do-banks-work

Mihup (2024, February 15). *Maximizing Retention: The Power of Bank Customer Churn Prediction*. https://mihup.ai/maximizing-retention-bank-customer-churn-prediction

Morawski, B.R. (2023, August 14). How Do Banks Work?. *U.S. News & World Report*. https://www.usnews.com/banking/articles/how-do-banks-work

Rodgers, E. (2023, October 24). Banking Customer Retention Strategies to Implement in 2024. *Drive Research*. https://www.driveresearch.com/market-research-company-blog/banking-customer-retention-strategies-that-work-in-2023

Statista. (2022, July 6). *Customer churn rate in the United States in 2020, by industry*. https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us

Utah Community Credit Union. (2024, June 24). *Banking: The Backbone of the Economy*. https://www.uccu.com/banking-the-backbone-of-the-economy

Zangari, M. (2024). *Bank Customer Churn Prediction*. Kaggle. https://www.kaggle.com/datasets/murilozangari/customer-churn-from-a-bank/data