Customer Churn Analysis in Banking

Jeevana Dova

Statistical Methods

May 7, 2025

**Table of Contents**

[**1. Introduction 3**](#_xme56jjbgago)

[**2. The Data 3**](#_4jbzh92hp7kd)

[2.1. Data Description 3](#_nise0uab8dxz)

[2.2. Summary Statistics 4](#_8y797sso5ep2)

[2.3. Data Visualization 5](#_472r48qqeqjl)

[2.3.1 Histograms 5](#_avnylfc6b9z1)

[2.3.2 Boxplots 6](#_8opmtmq8g48m)

[2.3.3 Bar Charts 7](#_sk5gsjua6k9j)

[**3. Distributions 8**](#_ul4xkt3xhmn)

[3.1 Fitted Histograms 8](#_9wdkls6f1d7b)

[3.2 QQ Plots 9](#_6jm0il7nrq8r)

[**4. Hypothesis Testing 10**](#_chkkjqma6sq5)

[4.1 Levene’s Test for Equality of Variances 10](#_hd28qs1gem4w)

[4.2 Independent Two-Sample T-Test (Welch's T-Test) 10](#_7j2arn7zvyr)

[4.3 Chi-square test of independence 11](#_alzvxfovyyyo)

[**5. Correlation Analysis 13**](#_mmqkiqcpt3ir)

[**6. ANOVA 14**](#_q3kl8lxi9p3g)

[6.1 Balance ~ Geography 14](#_97g8x0ld9il0)

[6.2 Age ~ Geography 15](#_hs3hv9pmcqog)

[6.3 CreditScore ~ Geography 16](#_oxkv92fh7bj5)

[**7. Regression Models 17**](#_crw4sjl6r8a7)

[7.1 Multivariate Linear Regression 17](#_s73bjloel30b)

[7.1.1 Predicting Balance 17](#_40zo30pqtwgv)

[7.1.2 Predicting Age 18](#_5agnu7wqmlgn)

[7.2 Logistic Regression 19](#_628dufhkr3lv)

[7.2.1 Model Evaluation & Confusion Matrix 20](#_yxpopypbo4cl)

[7.2.2 Classification Report 20](#_oeemwifv0i49)

[7.2.3 ROC Curve and AUC Score 21](#_gnov0n6f6zlq)

[7.3 Lasso Regression 21](#_eoq99g3znkw4)

[**8. Conclusion 23**](#_trbf3arjjn)

# 

# 

# 1. Introduction

This study examines the relationship between different variables of a customer dataset from a financial institution. The data includes various features about the customer’s account, their personal information, and whether or not the customer has exited the bank. The aim is to perform a comprehensive analysis to provide insight on customer behaviors, which are beneficial for companies to understand when making decisions.

The analysis begins with examining the types and distributions of the variables in the dataset, estimating relevant parameters and assessing their normality for continuous numerical features. This is followed by hypothesis testing to evaluate differences and determine statistical significance of features contributing to the churn of a customer. For analysis between categorical and numerical variables, T-Tests and one-way ANOVA Tests will be applied to evaluate variable differences and associations. Analysis between two categorical variables will utilize a Chi-Squared Test for Independence. Correlation analysis is conducted to understand the strength and direction of relationships between numerical variables, providing a foundation for selecting significant features.

To further understand the data, this study will explore predictive modeling. For modeling continuous outcomes, such as account balance and customer age, we will implement Multivariate Linear Regression. For modeling a binary classification that determines the churn of customers, we will implement Logistic Regression.

The primary objective of this study is to analyze financial data associated with customer demographics and behaviors, using statistical methods to identify significant patterns, correlations, and predictors of key outcomes.

# 2. The Data

## 2.1 Data Description

The dataset we are utilizing is titled Banking Customer Churn Prediction Dataset, and it is sourced from Kaggle at this [link](https://www.kaggle.com/datasets/saurabhbadole/bank-customer-churn-prediction-dataset). There are 10,000 observations recorded, and 14 variables that describe the customer, their account, and their churn status. There are no missing values or duplicate rows present in the dataset, so data cleaning is not a necessary step. Identifiers for each observation are not considered in statistical analysis, so the following features were removed from consideration: RowNumber, CustomerId, Surname.

The quantitative variables are:

* **CreditScore:** Credit score of the customer
* **Age:** Age of the customer
* **Balance:** Account balance of the customer
* **Tenure:** Number of years the customer has been with the bank
* **NumOfProducts:** Number of bank products the customer has
* **EstimatedSalary:** Estimated salary of the customer

The categorical variables are:

* **Geography:** Country of the customer
* **Gender:** Male or female
* **HasCrCard:** Indicates whether the customer has a credit card
* **IsActiveMember:** Indicates whether the customer is an active member
* **Exited:** Indicates whether the customer has exited the bank

## 2.2 Summary Statistics

The following section provides an overview of the descriptive statistics for each numerical feature. This includes the mean, median, mode, standard deviation, 25% and 75% quartiles, range, and minimum/maximum values.

CreditScore ranges from 350 to 850, with a mean of approximately 650.528 and a standard deviation of 96.653. The distribution appears fairly wide, given a range of 500, indicating varied creditworthiness among customers. The median of 652 is slightly higher than the mean. The mode is 850, meaning that the most common individual credit score is the maximum value.

Age varies between 18 and 92 years, with an average of around 38.921 years and a standard deviation of 10.488. The middle 50% of customers fall between the ages of 32 and 44 years. The most commonly occurring age of a customer is 37 years, which is also the median of the customers.

Tenure ranges from 0 to 10 years at the bank. The average tenure is about 5 years, which is equivalent to the median. The 25% quartile is 3 and the 75% quartile is 7 years, indicating a symmetrical interquartile range since the median is the average of the two quartiles. The most common length of time spent at the bank is 2 years.

Balance displays considerable variation, ranging from $0 to approximately $250,898.09. The mean account balance is around $76,486.89, with a high standard deviation of $62,397.41. The 25% quartile is 0, indicating that at least 25% of customers have a balance of zero. The most common balance across all the customers is also zero. This suggests that a significant portion of customers do not hold deposits in their accounts.

NumOfProducts held by the customers ranges from 1 to 4, with an average of 1.53 and standard deviation of 0.581. The median is 1 product, so at least 50% of customers only hold a single product with the bank. The mode is also 1 product, further indicating that the majority of customers only have 1 product.

HasCrCard is a binary variable, with 0 representing customers without a credit card and 1 representing those with one. The most commonly occurring value is 1, and based on the average, about 70.55% of customers have a credit card.

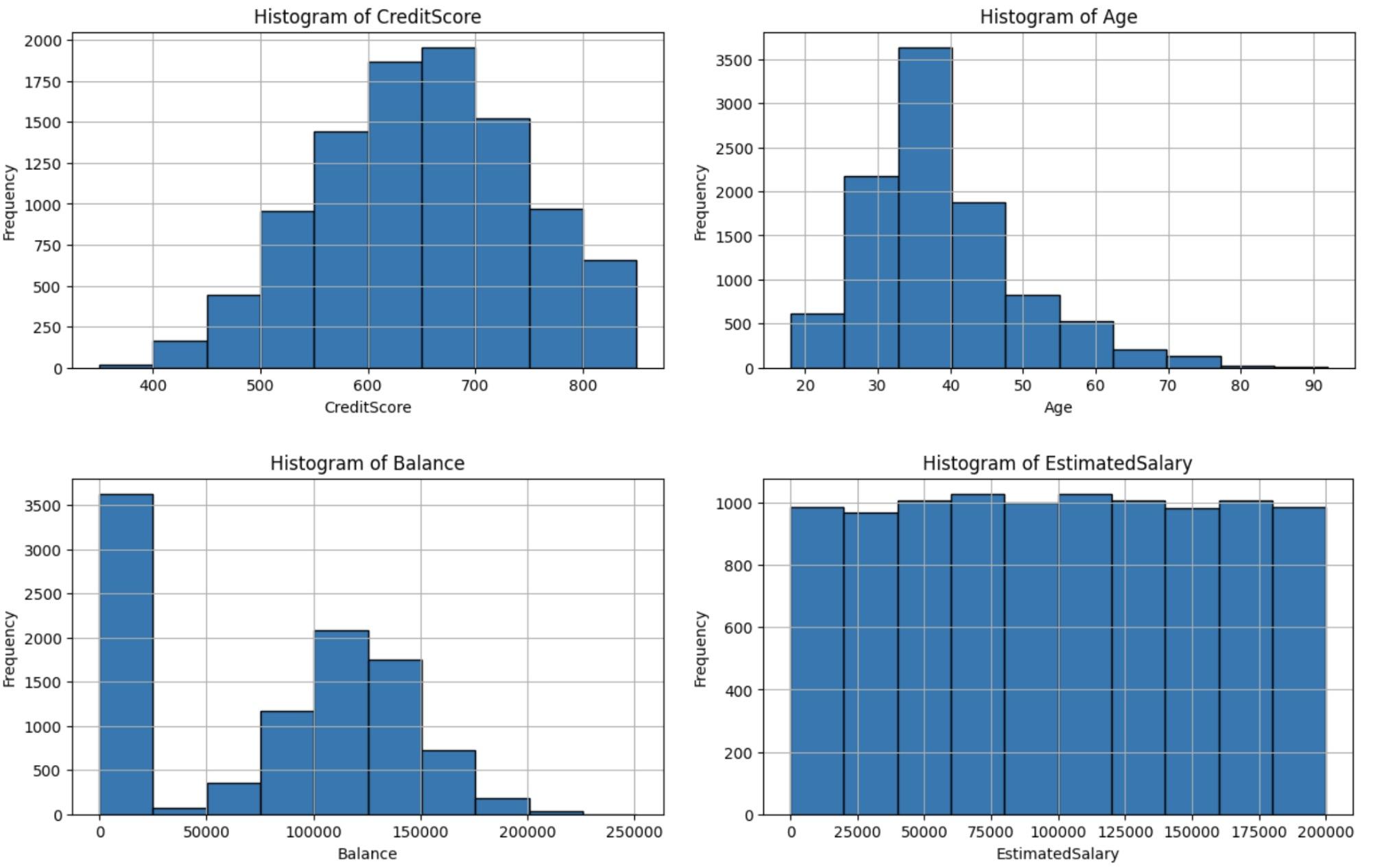
IsActiveMember is a binary variable, with 0 representing customers not actively a member of the bank and 1 representing those who are active members. The most common value is 1 and based on the average, roughly 51.51% of customers are classified as active members. This indicates a fairly balanced split in the level of engagement with the bank across all the customers.

Estimated Salary ranges from $11.58 to approximately $199,992.48, with a mean salary of $100,090.24 and a standard deviation of $57,510.49. The mode of the salary is $24,924.92. The median is about $100,193.92, which is just slightly higher than the mean. The 25% quartile is $51,002.11 and the 75% quartile is $149388.25, which are about the same distance from the median, indicating a roughly symmetric interquartile range.

Exited is a binary variable, with 0 representing customers who remain at the bank and 1 representing those who have left (churned). The most common value is 0, and based on the average, only 20.37% of customers have exited. This means that the majority of customers, 79.63%, have remained, indicating a class imbalance in the dataset.

## 2.3. Data Visualization

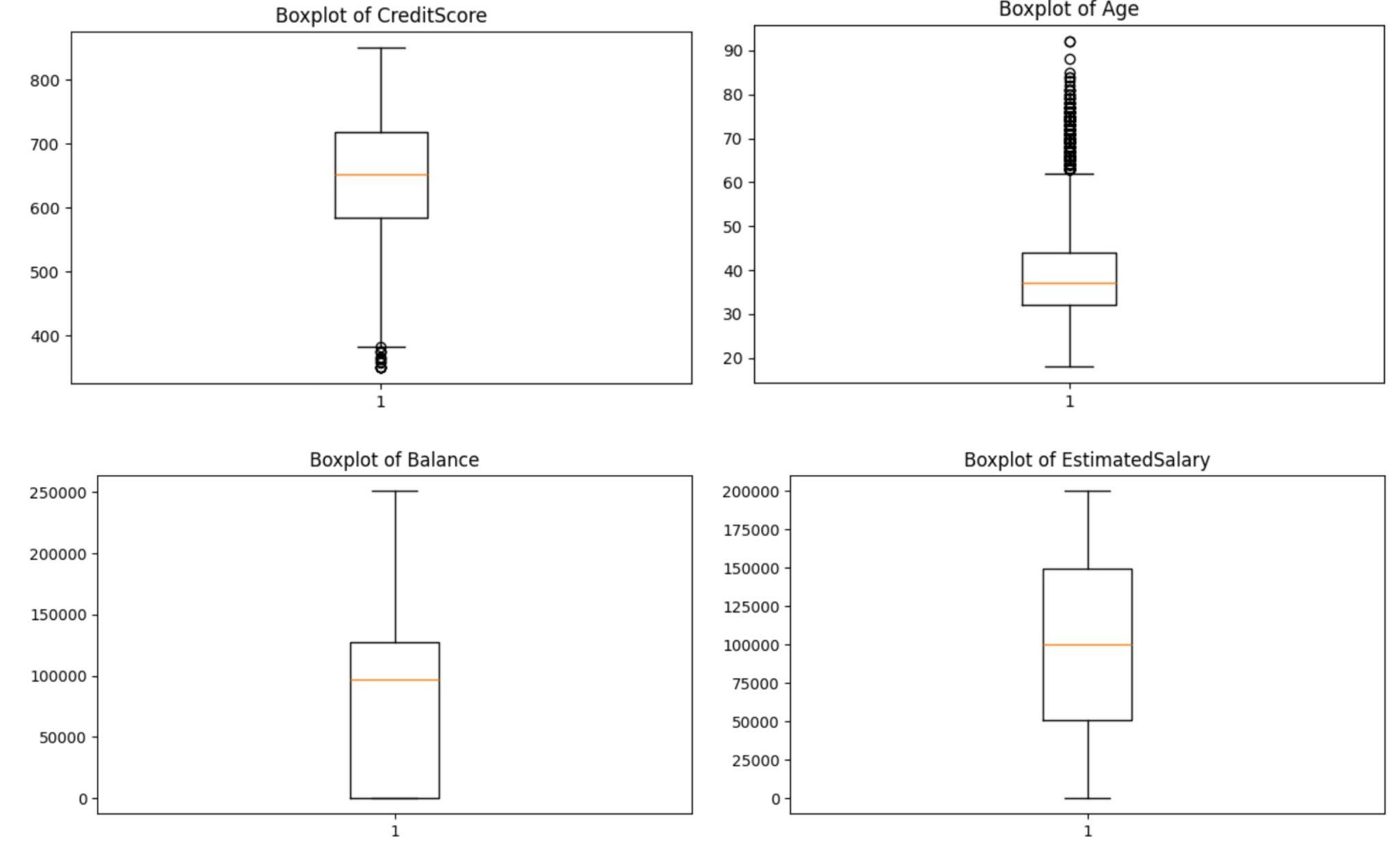
### 2.3.1 Histograms



**Fig.2.1 Histograms of Continuous Numerical Variables**

Based on the histograms for the continuous variables of CreditScore, Age, Balance, and EstimatedSalary, we can visualize their distributions. The features CreditScore and Age are unimodal, but each has some slight skewing. CreditScore is relatively symmetric with some slight skewing to the left, while Age has a more visual right skew. Balance shows a bimodal histogram, where the first peak occurs at 0 and the second occurs around 100,000. The first peak is from the significant number of customers that have no deposits in their account and hold a balance of $0. The second peak appears to be symmetrical and without any noticeable skewing. The EstimatedSalary histogram shows the frequency within each bin is roughly the same, which results in a nearly flat appearance.

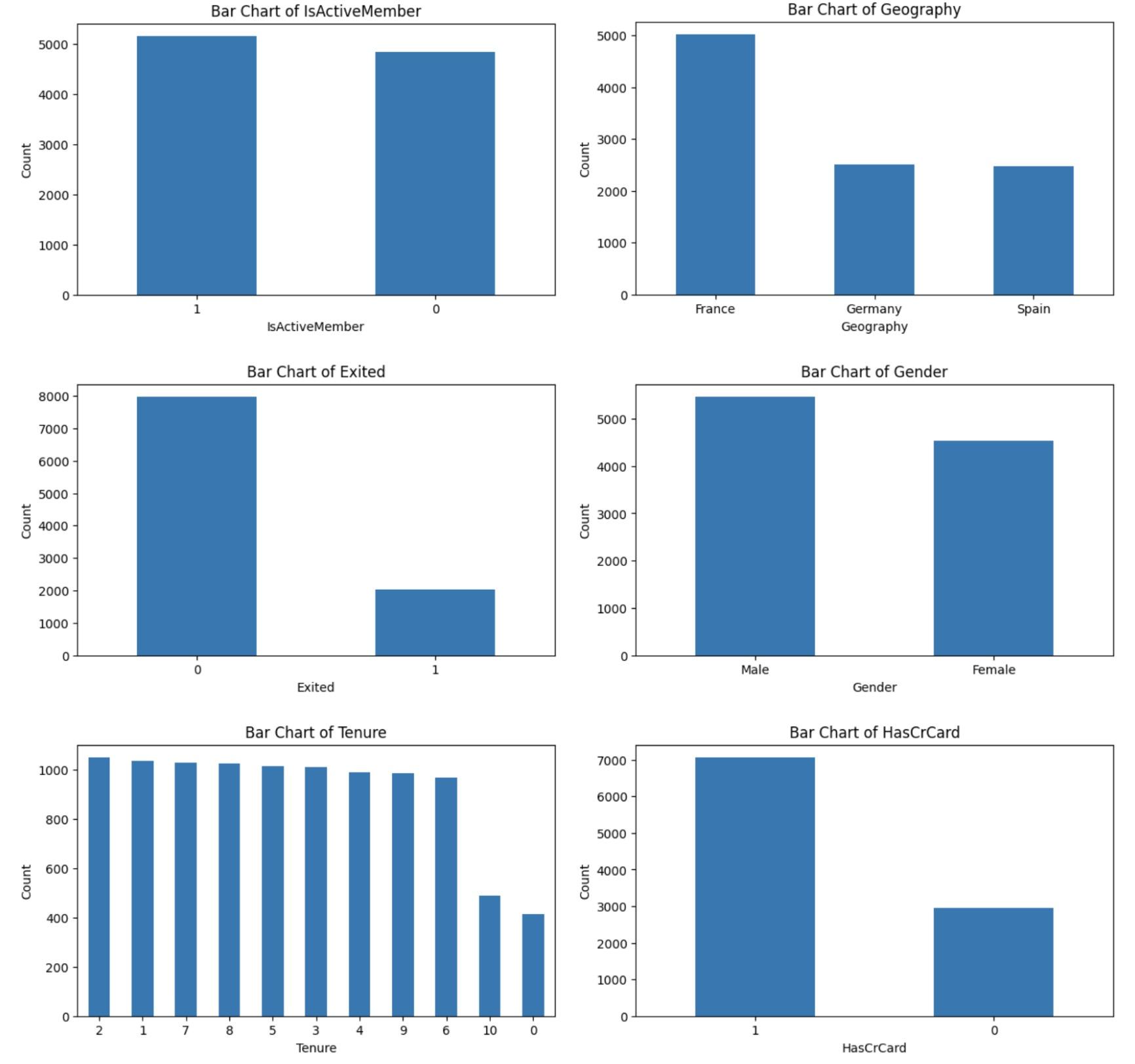
### 2.3.2 Boxplots



**Fig. 2.2 Boxplots of Continuous Numerical Variables**

The continuous variables of CreditScore, Age, Balance, and EstimatedSalary can also be visualized as boxplots. For CreditScore, we can see the presence of outliers on the lower end that are below a score of 400. This indicates the slightly negative skewing in the distribution. The boxplot of Age displays a significant number of outliers above 60 years, which represents a small number of much older individuals when compared to the other customers existing at the bank. The number of higher outliers, demonstrates a stronger positive skew in the data. The Balance boxplot has no tail on the bottom end and a noticeable amount of values near zero, suggesting many customers have no account balance. EstimatedSalary displays a uniform distributed boxplot with no outliers or skewing.

### 2.3.3 Bar Charts



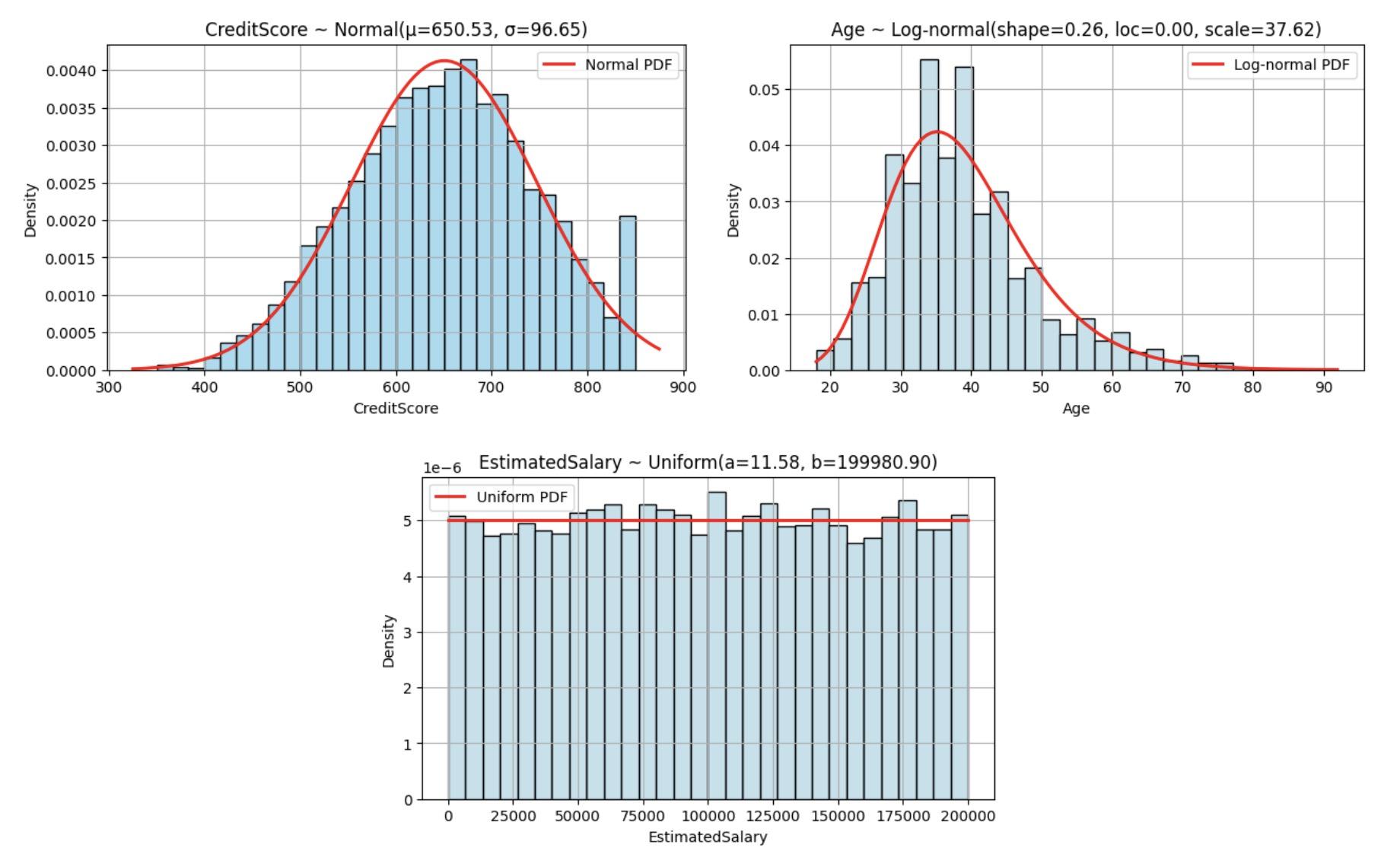
**Fig. 2.3 Bar Charts of Categorical and Discrete Numerical Variables**

The counts of categorical and discrete numerical values can be visualized with bar charts. We can see a relatively balanced distribution between active and inactive members at the bank, since both categories have similar counts. There is a notable imbalance in the geography of the customers, since France has over 2,000 more customers than both Germany and Spain. The count of customers who remain at the bank is significantly higher than the number of customers who have left, which indicates a low churn rate overall. The gender distribution is skewed towards males than females, but there is no significant difference between either counts. Tenure ranges from 0 to 10 years, with 2 years being the most common length a customer has been at the bank and 0 being the least common length. The tenure lengths of 1 to 9 years have a relatively uniform frequency. There is a strong imbalance in the attribute describing credit card ownership of customers. There is a significant number of customers who have a credit card compared to those who do not.

# 3. Distributions

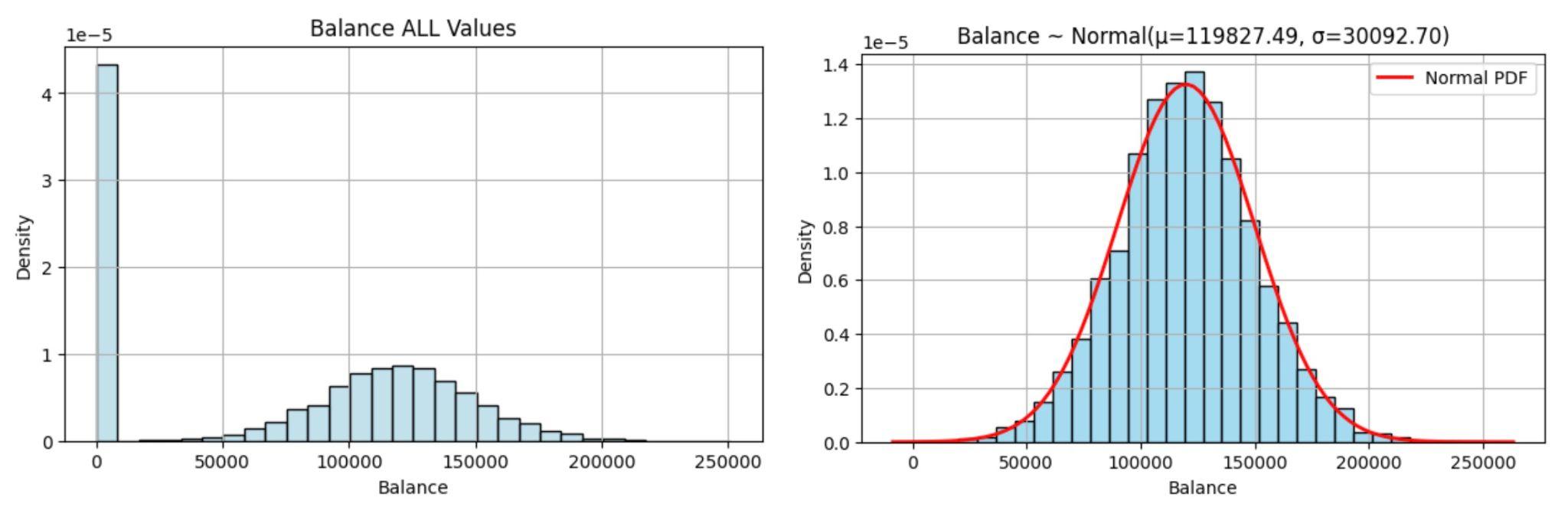
Further analysis into the histograms and QQ plots can reveal information about the distributions of each quantitative random variable.

## 3.1 Fitted Histograms



**Fig. 3.1 Fitted Histograms of Quantitative Variables with Distribution Curves**

The CreditScore feature approximately follows a normal distribution since the histogram is fairly symmetric and closely follows the fit Normal(μ = 650.53, σ = 96.65). Age of customers follows a log-normal distribution, meaning if the log of the age values are taken, the resulting values are normally distributed. The histogram demonstrates a right-skewed distribution that fits the log-normal(shape = 0.26, loc = 0.00, scale = 37.62). EstimatedSalary appears to fit the distribution Uniform(a = 11.58, b = 199980.90), since the histogram is roughly flat.

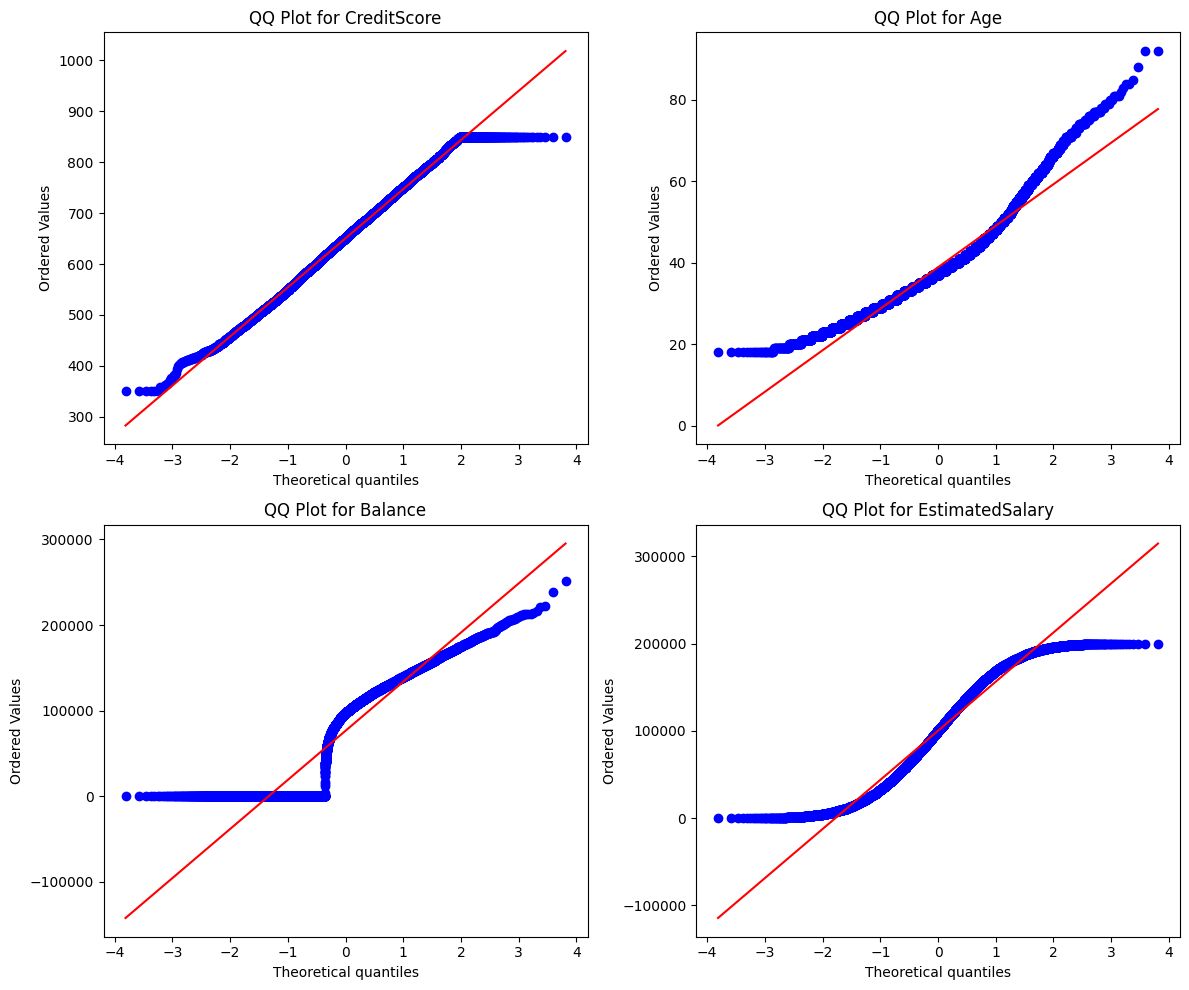


**Fig. 3.2 Fitted Histograms of Balance Variables with Distribution Curves**

Balance follows the bimodal distribution where the first peak are customers with a zero balance and the other peak are the remaining non-zero balances. Analyzing the first peak reveals the standard deviation and variance are zero because there is no spread or variability in the data, since all observations are zero balances. Analyzing the second peak reveals a normal distribution that follows the fit Normal(μ = 119827.49, σ = 30092.70).

## 3.2 QQ Plots

In addition to the fitted histograms, QQ plots were also created to further investigate the normality of the continuous features. The red lines signify a perfectly normal distribution to compare against the true quantiles in blue.



**Fig. 3.3 QQ Plots of Quantitative Variables with Distribution Curves**

The CreditScore feature follows the red line fairly well, but flattens out at the upper limit of 850. This indicates that the feature is roughly normal, but there is a truncation effect due to the maximum possible credit score value. The Age characteristic has a noticeable curve, where the data starts above the red line, curves below, and then continues above the red line. This indicates a right-skewed, non-normal distribution. This pattern aligns with the previous Age boxplot, which showed a large number of outliers above the age of 60 years. The Balance plot displays a heavy deviation from the red line at the lower end and then a steep incline. This divergence suggests a non-normal distribution with a significant spike at zero, which is heavily influenced by the majority of customers having no account balance. The EstimatedSalary has a noticeable S-shaped curve, indicating a heavier tail than expected on both sides. This trend is consistent with a uniform distribution, which aligns with the previous fitted histogram of EstimatedSalary.

# 4. Hypothesis Testing

## 4.1 Levene’s Test for Equality of Variances

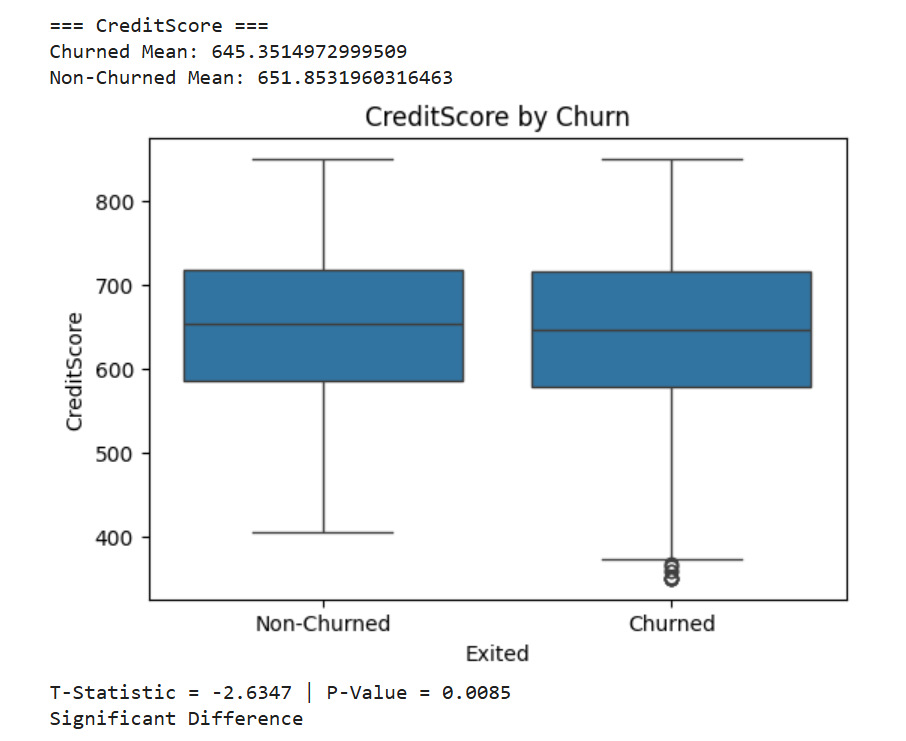
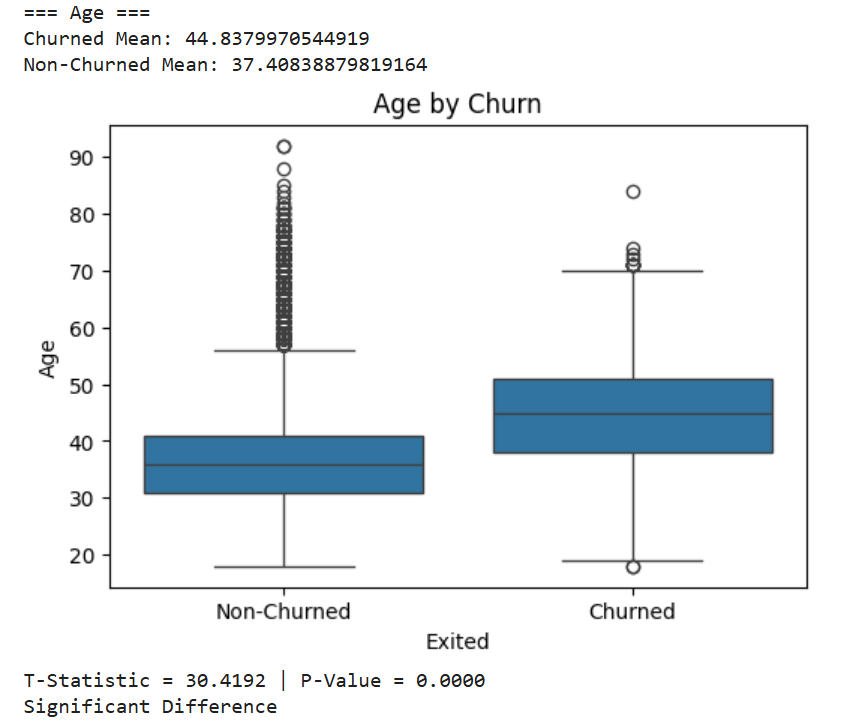
Prior to performing the Independent T-Test, Levene’s Test was conducted to evaluate the assumption of equal variances between the churned and non-churned customer groups across key numerical variables.

The variables tested include CreditScore, Age, Balance, EstimatedSalary, Tenure.

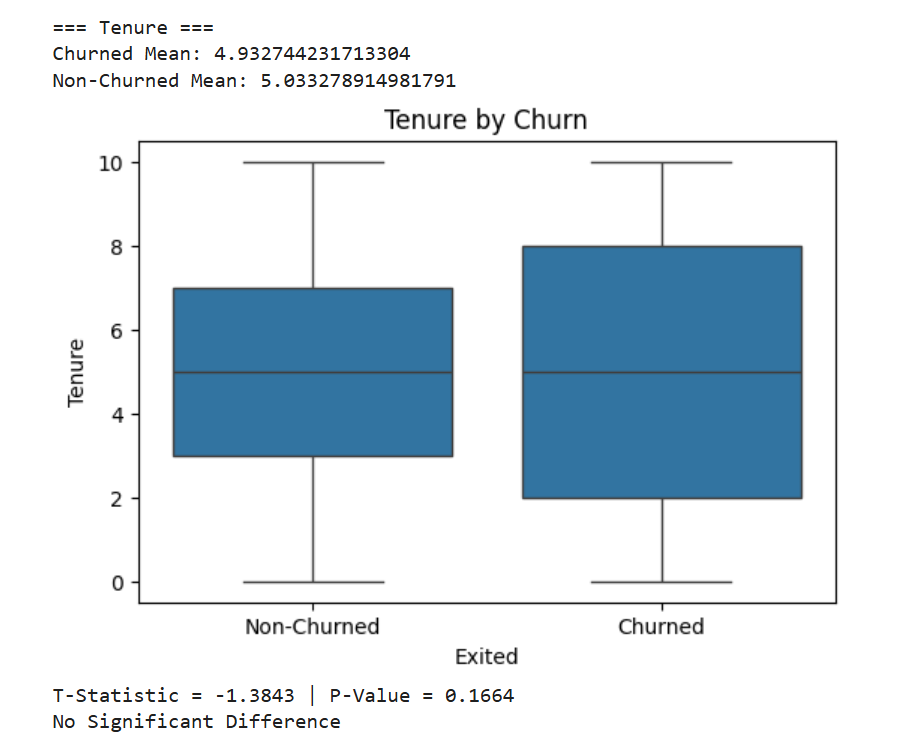
The results of Levene’s Test indicated that the variables CreditScore, Age, and Balance exhibited unequal variances between churned and non-churned customer groups (p-value < 0.05). Therefore, Welch’s T-Test, which is appropriate for comparing means when variances are unequal, was applied to these variables to ensure the validity of the results.

## 4.2 Independent Two-Sample T-Test (Welch's T-Test)

An Independent Two-Sample T-Test was performed to assess whether there are statistically significant differences in the mean values of the numerical variables between churned and non-churned customers. The results of the T-Test are summarized below:

## 



**Fig. 4.1 Boxplots of Numerical Features by Churn Status (Welch’s T-Test Results)**

Based on the results of the analysis, CreditScore, Age, and Balance were identified as statistically significant factors impacting customer churn. Conversely, EstimatedSalary and Tenure did not exhibit a significant effect on churn behavior.

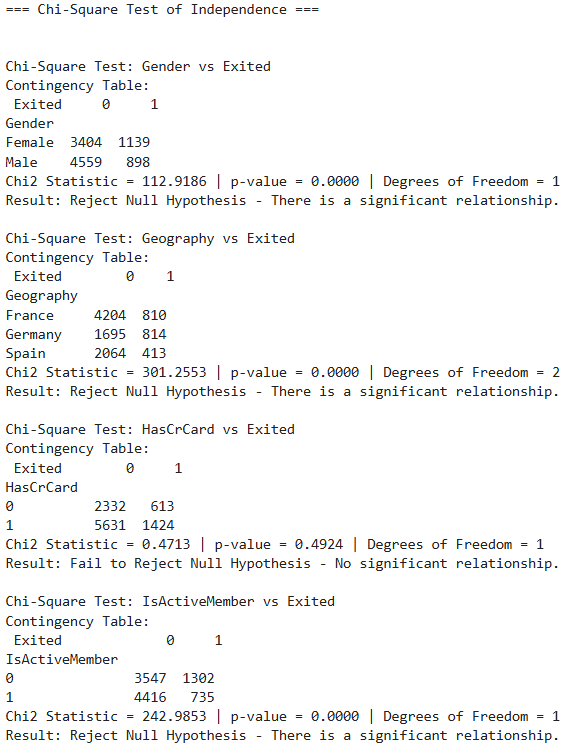
## 4.3 Chi-square Test of Independence

To assess whether certain categorical variables have a statistically significant association with customer churn (represented by the variable Exited), a Chi-Square Test of Independence was performed. This test is used to evaluate whether the distribution of customers who churn versus those who remain at the bank is dependent on the category of another variable.

The test was applied to the following categorical features: Gender, Geography, HasCrCard, and IsActiveMember. A contingency table was constructed for each variable against churn status, and the chi-square statistic, degrees of freedom, and p-value were calculated.

The results showed statistically significant relationships (p-value < 0.05) between churn and the variables Gender, Geography, and IsActiveMember. This indicates that churn behavior is not uniformly distributed across the levels of these categorical features. For example, the proportion of customers leaving the bank differs significantly between males and females, among different countries, and between active and inactive members. These findings suggest that these variables may have predictive value in churn modeling and should be considered in customer retention strategies.

In contrast, HasCrCard showed a non-significant relationship with churn (p-value = 0.4924), suggesting that credit card ownership does not significantly influence whether a customer exits the bank. Therefore, this variable may be less important for churn prediction and could be excluded in feature selection for modeling.



**Fig. 4.2 Chi-Square Test of Independence**

These results provide insights into customer behaviors and preferences and help identify specific segments where churn rates are higher, enabling more targeted and effective retention interventions.

# 5. Correlation Analysis

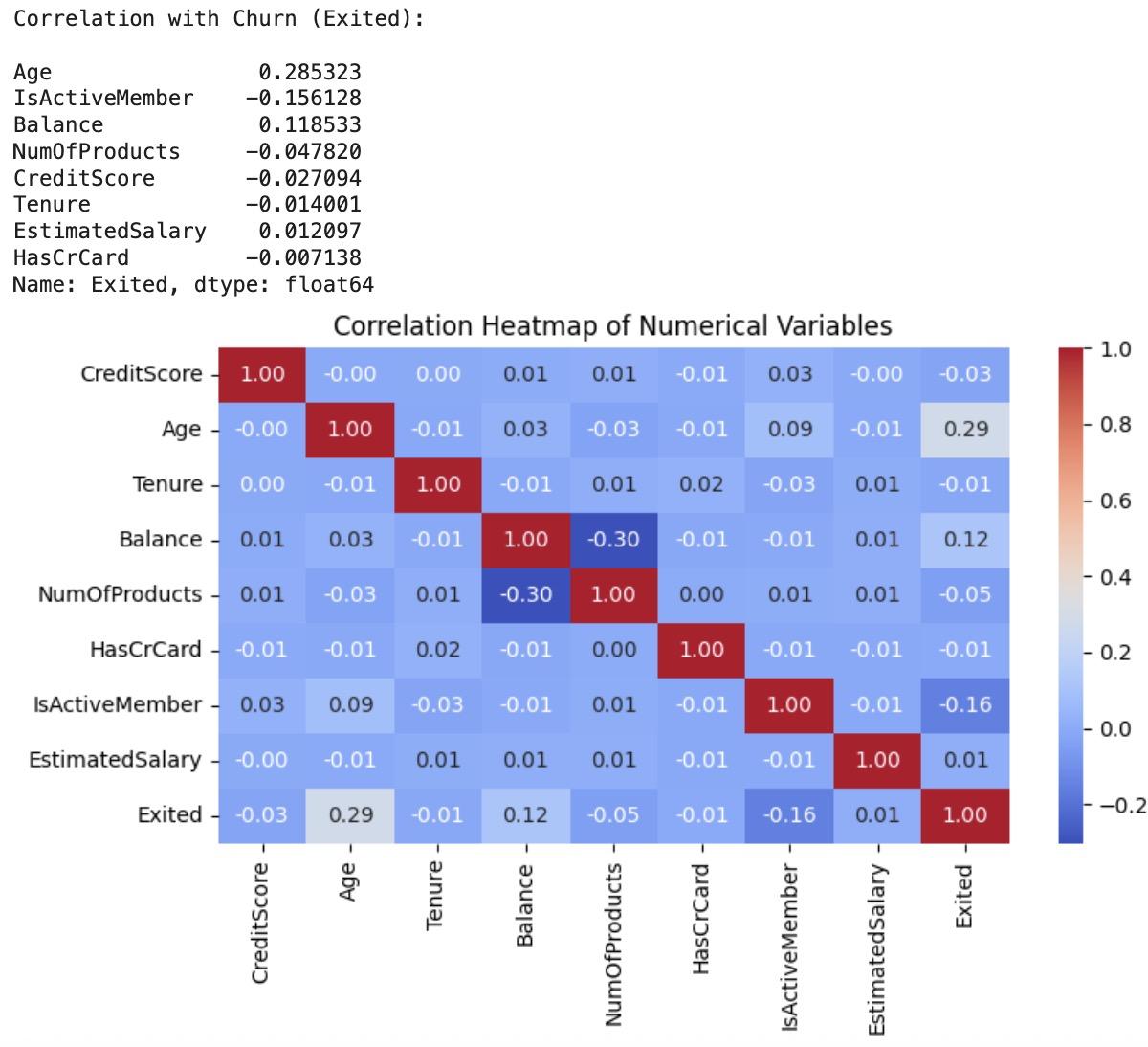
In this section, we explored the relationships between numerical variables and customer churn (Exited) using Pearson correlation. The goal was to identify which variables are most associated with a customer's likelihood to leave the bank.

To achieve this, we first converted the target variable Exited into a float to compute correlation values accurately. We then selected all numerical variables from the dataset and calculated their correlation coefficients with churn. A heatmap was also generated to visualize the relationships among all numerical features.

The correlation values ranged between -1 and +1, where:

* A positive value indicates a direct relationship (as one increases, the other tends to increase).
* A negative value indicates an inverse relationship.
* Values close to 0 suggest little to no linear relationship.

The correlation results are shown below:



**Fig. 5.1 Correlation Heatmap of Numerical Variables**

The above correlation heatmap of numerical variables shows the strength and direction of linear relationships among the dataset's features, particularly in relation to customer churn (Exited). From the matrix, Age exhibits the strongest positive correlation with churn (0.285), suggesting that older customers are more likely to exit the bank. This makes age a key variable for churn prediction.

IsActiveMember shows a moderate negative correlation (-0.156), indicating that active customers are less likely to leave, underlining the importance of engagement. Balance (0.118) also has a slight positive correlation, meaning high-balance customers may be marginally more inclined to churn, potentially due to attractive offers from competitors.

Other variables like NumOfProducts, CreditScore, Tenure, EstimatedSalary, and HasCrCard all have correlation coefficients close to zero, implying minimal influence on churn behavior. Their weak relationships indicate they may not be strong standalone predictors of customer exit behavior.

Overall, this correlation analysis highlights Age and IsActiveMember as the most relevant features for modeling churn, helping guide feature selection in predictive modeling tasks.

# 6. ANOVA

To investigate whether the geography of a customer has a statistically significant impact on key financial features, a one-way ANOVA was conducted for three chosen continuous variables: Balance, Age, and CreditScore. These variables were analyzed because they have been identified as statistically significant factors that influence customer churn. The categorical variable, Geography, has three groups: France, Germany, and Spain.

Since ANOVA assumes equal variances across the groups, the variance for each country was calculated and compared for the different features. For Balance, the variance for Germany appears much smaller than France or Spain. For Age and CreditScore, the variances appear relatively the same across the countries. To confirm a difference in variances, we performed a Levene’s Test for each feature. Balance resulted in a p-value < 0.05 and Age and CreditScore resulted in p-values > 0.05, so Balance was determined to have unequal variances across the countries.

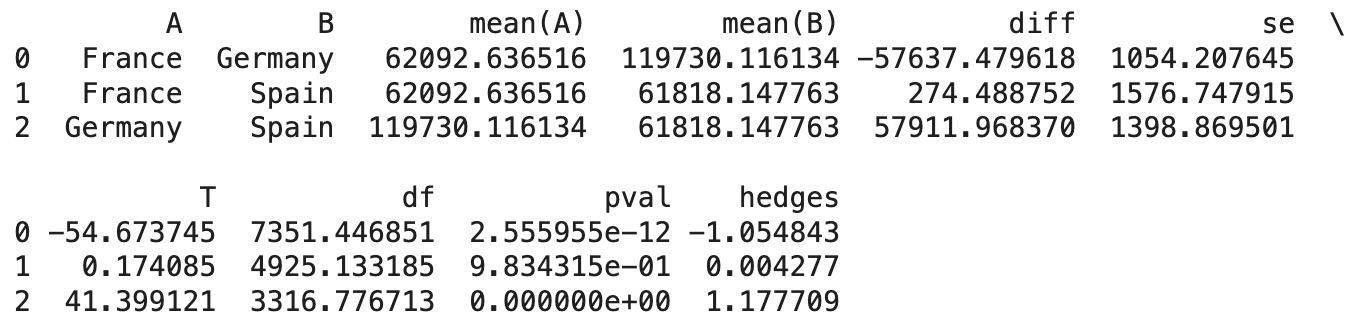
## 6.1 Balance ~ Geography

Because the financial feature Balance was identified to have unequal variances across the different countries, we could not utilize the typical one-way ANOVA test due to its assumption of homogenous variances. Instead, we used a Welch's ANOVA Test, which does not assume equal variances.



**Fig. 6.1 Welch's ANOVA Table for Balance by Geography**

The results indicated that there was a difference within the means of customer account balances across different countries, as the p-value of 0 < 0.05. To further identify which country was different, we used a post hoc test that does not assume equal variances or equal sample sizes: Games-Howell. This method is similar to Tukey’s test in its formulation and uses Tukey’s studentized range distribution.

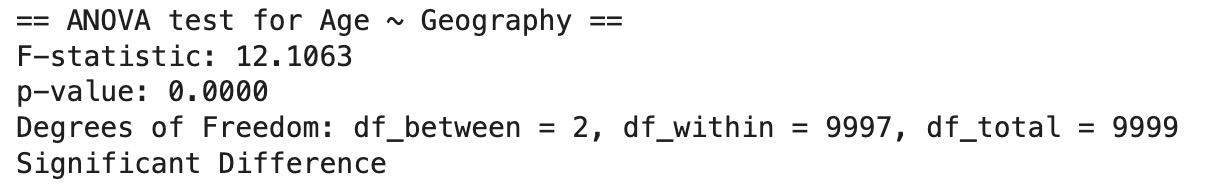


**Fig. 6.2 Games-Howell Test Results for Balance by Geography**

From the results, it is clear that Germany is statistically significant, since both pairs including Germany have p-values of approximately zero (p<0.05), indicating that Germany’s mean is different from both France and Spain’s mean. The pair of France and Spain has a p-value greater than 0.05, so it does not indicate a statistical significance between the means of France and Spain.

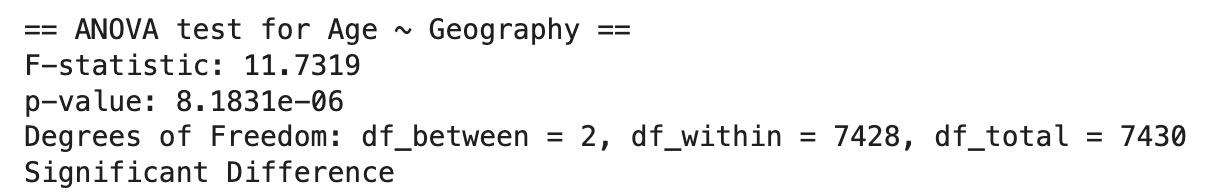
## 6.2 Age ~ Geography

Since Age was determined to have equal variances across countries, we used a typical one-way ANOVA Test. We completed this test on all the observations in the dataset.



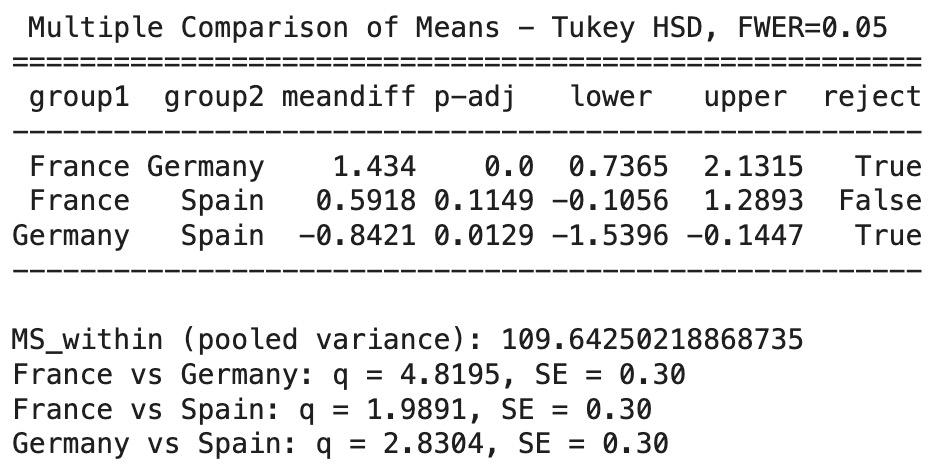
**Fig. 6.3 One-Way ANOVA for Age by Geography (Original Sample)**

The results indicate that there is at least one mean of customer age that is significantly different within the countries, since the p-value < 0.05. Further analysis can be completed with Tukey's test, since this test assumes homogeneity of variances. However, Tukey’s test requires equal sample sizes for each group, which does not occur in our dataset. The number of customers in France is much larger than the number of customers in Spain or Germany. To properly conduct Tukey’s Test, we decided to resample the data so that there was an equal number of customers for each country. The new sample size for each country is the minimum size across the initial groups, which was 2,477. With these samples of equal size, we conducted the one-way ANOVA again and came to the same conclusion that at least one mean was different.

****

**Fig. 6.4 One-Way ANOVA for Age by Geography (Resampled Equal Groups)**

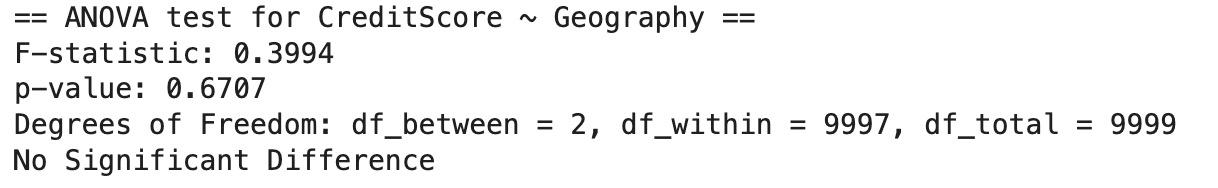
To identify which mean was different, we now performed the Tukey’s Test on the sampled data of equal size. It shows Germany is the significantly different group, since all pairs including this country have p-values less than 0.05.



**Fig. 6.5 Tukey’s HSD Test for Age by Geography (Equal Sample Size)**

## 6.3 CreditScore ~ Geography

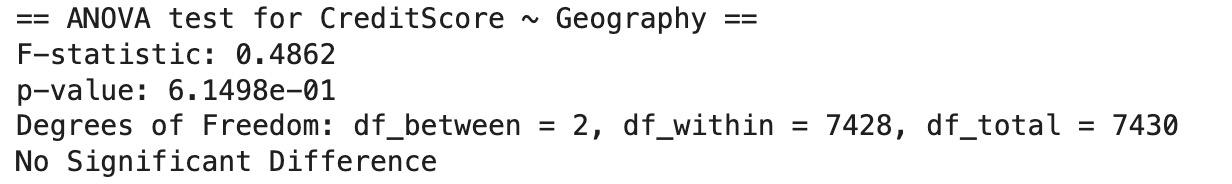
Since CreditScore was determined to have equal variances, a typical one-way ANOVA Test was applied to all the observations.



**Fig. 6.6 ANOVA Results for CreditScore by Geography (Full Data)**

The results indicate that there is no significant difference between means of credit score for customers in different countries, since the p-value is much greater than 0.05. No further analysis is needed with Tukey’s Test because there is no difference between means.

To verify that the unbalanced sample sizes do not impact the results of the ANOVA test, the same method was applied to the sampled data with equal sizes for each country. The results provide the same conclusion that there is no difference between the average credit score for customers in different countries, since the p-value is much larger than 0.05.



**Fig. 6.7 ANOVA Results for CreditScore by Geography (Sampled Data)**

The analysis of categorical variables with ANOVA helps to identify which geographical locations may impact certain financial factors or customer demographics. This offers insight into potential business strategies or region-specific marketing that the bank can implement.

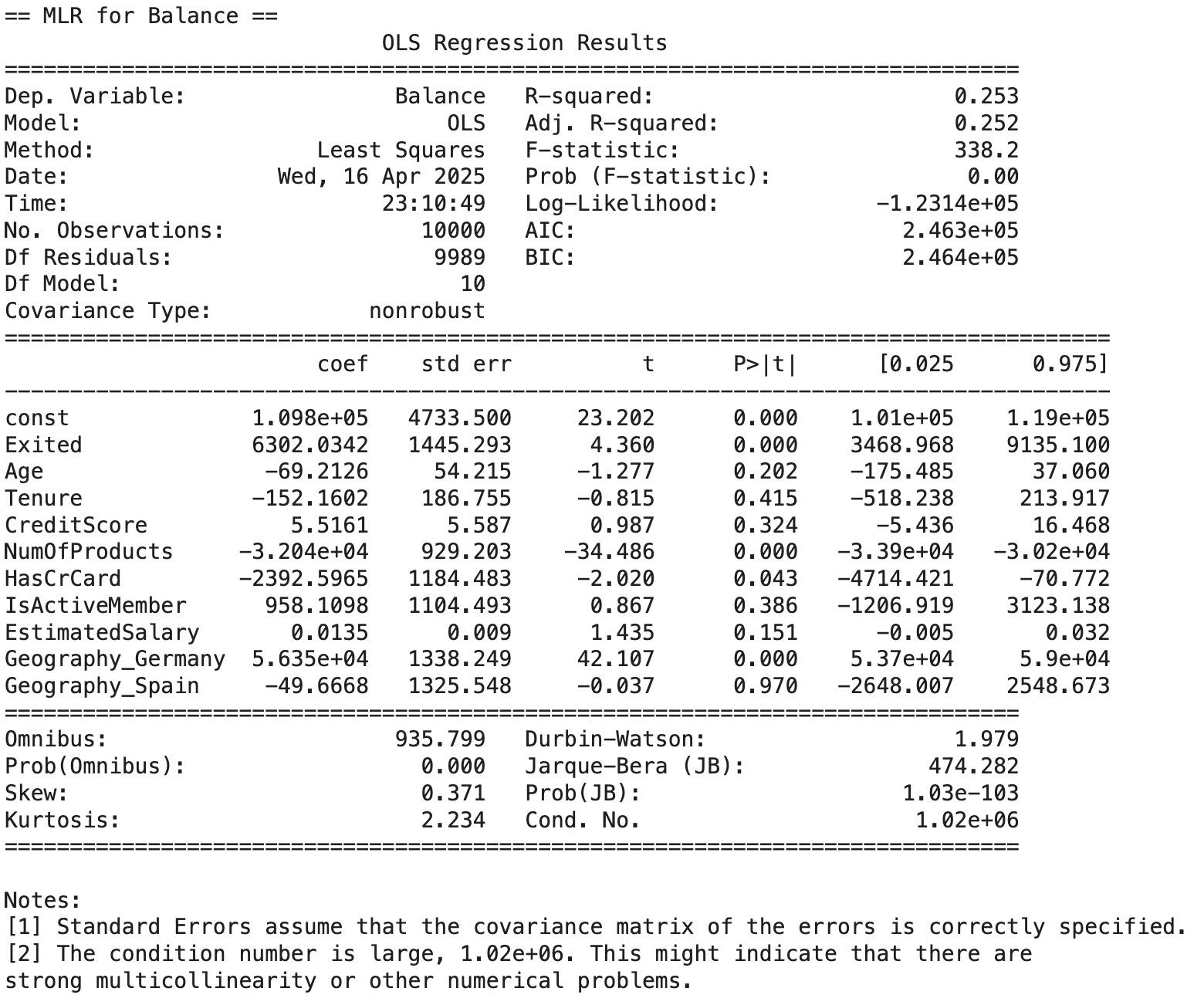
# 7. Regression Models

## 7.1 Multivariate Linear Regression

To investigate the effect of multiple customer characteristics on different features like customer finances and age demographics, two multivariate linear regressions (MLR) were performed with Balance and Age as the dependent variables. Independent variables included all other quantitative variables and the geographical variables represented as dummy variables: Geography\_Germany and Geography\_Spain.

### 7.1.1 Predicting Balance

The MLR model was initially created using all other features, however it was noted that there was strong multicollinearity between the variables and that multiple of the features were not significant to the prediction of a customer’s account balance.



**Fig. 7.1 Multivariate Linear Regression Results for Predicting Customer Balance**

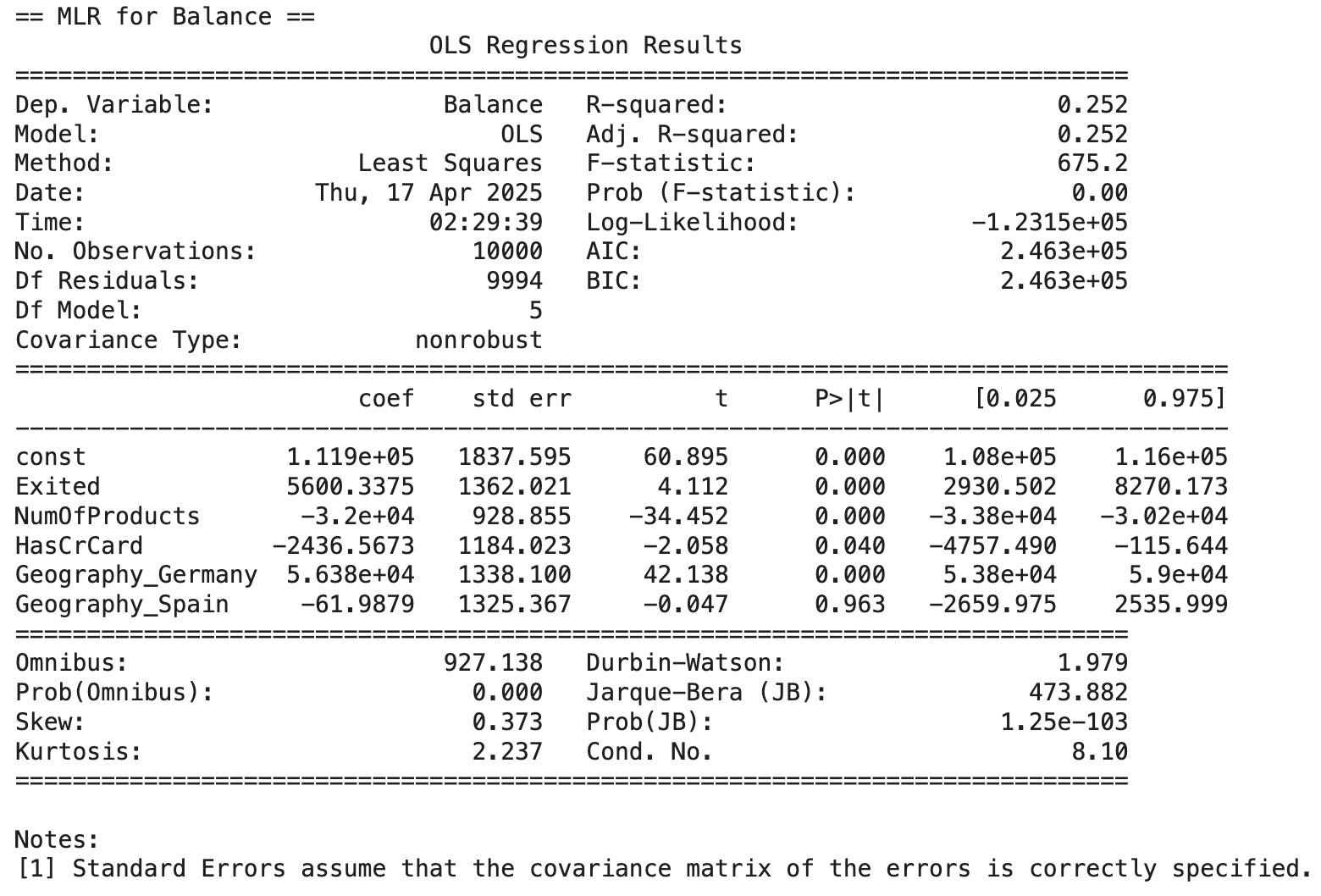
Looking at the F-test for overall significance, there is strong evidence that the model as a whole explains a portion of the variation in Balance. This means at least one of the independent variables contributes meaningfully to predicting the outcome. Looking at the T-test for individual significance, Exited, NumOfProducts, HasCrCard, and Geography\_Germany were found to be statistically significant predictors of Balance at an alpha level of 0.05.

The coefficients of these significant predictors can provide details about their impact on the customer’s balance. The Geography\_Germany variable has an extremely large positive value, suggesting that customers located in Germany held much higher average balances than those in other countries. In contrast, NumOfProducts has a large negative coefficient, indicating an inverse relationship with Balance: the more products a customer holds with the bank, the lower their account balance. The variable Exited is positively associated with Balance, which suggests customers who left the bank had higher account balances.

The model resulted in an adjusted R-squared value of 0.252, indicating that approximately 25.2% of the variance in Balance can be explained by the predictors. This is very low and likely indicates that the features captured in the dataset are not strong predictors of balance.

The model was run again with only the significant predictors, and the same conclusion was drawn about the adjusted R-squared. The model to predict balance is:

Balance = 111,900 – 32,000(NumOfProducts) – 2436.5673(HasCrCard) + 5600.3375(Exited) + 56,380(Geography\_Germany) – 61.9879(Geography\_Spain)



**Fig. 7.2 Regression Results for Balance Model – Significant Predictors Only**

### 7.1.2 Predicting Age

Similarly to predicting Balance, the MLR model for Age was initially created using all features, but there was a strong multicollinearity between the many variables.

### 

**Fig. 7.3 Multivariate Linear Regression Results for Age**

Using the F-test for overall significance offers significant evidence that the model in general explains a portion of the variation in Age. The T-test for individual significance, reveals that Exited, NumOfProducts, and IsActiveMember were found to be statistically significant predictors of Age at an alpha level of 0.05. The variables Exited and IsActiveMember are positively associated with Age, which indicates that customers who left the bank or are actively engaged with the bank are generally older.

The NumOfProducts is negatively associated with Age, which suggests that the more products a customer has at the bank, the younger their age. It is interesting to note that none of the coefficients of these significant predictors have a large magnitude. The adjusted R-squared value of 0.099, indicates that only 9.9% of the variance in Age can be explained by the predictors. The customer's financial and geographical attributes are not strong predictors of age.

The model was run again with only the significant predictors, and the same conclusion was drawn about the adjusted R-squared. The model to predict age is:

Age = 36.3405 − 0.3130(NumOfProducts) + 2.7973(IsActiveMember) + 7.9499(Exited)

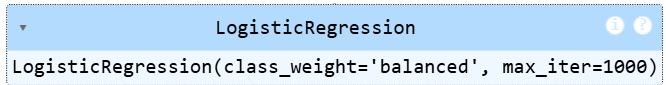


**Fig. 7.4 Reduced MLR Output for Age**

## 7.2 Logistic Regression

Logistic Regression was employed to model the probability of customer churn based on a set of independent variables. The predictors used in the model included CreditScore, Age, Balance, EstimatedSalary, and Tenure, while the dependent variable was Exited, representing whether a customer had churned (1) or not (0).

Due to the observed class imbalance in the dataset, with a higher proportion of non-churned customers, the model was configured using class\_weight='balanced' to penalize misclassification of the minority class more heavily and mitigate bias.

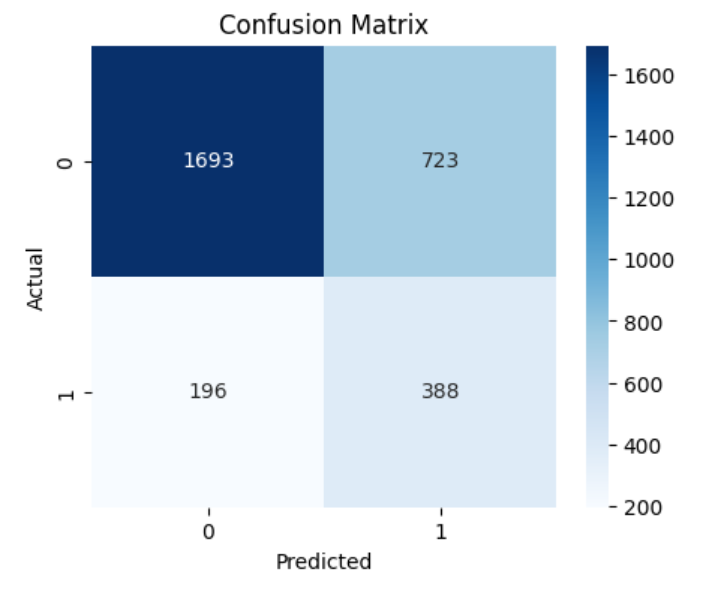


**Fig. 7.5 Logistic Regression Model Setup**

The dataset was partitioned into 70% training and 30% testing subsets. The logistic regression model was trained using the training data and evaluated on the test data to assess its generalization performance.

### 7.2.1 Model Evaluation & Confusion Matrix

The performance of the model was assessed using standard classification metrics. The model achieved an overall accuracy of 69% on the test set. The confusion matrix is shown below:



**Fig. 7.6 Confusion Matrix for Logistic Regression Model**

**True Negatives (TN):** 1693 → Correctly predicted non-churned customers

**False Positives (FP):** 723 → Predicted churn, but actually stayed

**False Negatives (FN):** 196 → Predicted stay, but actually churned

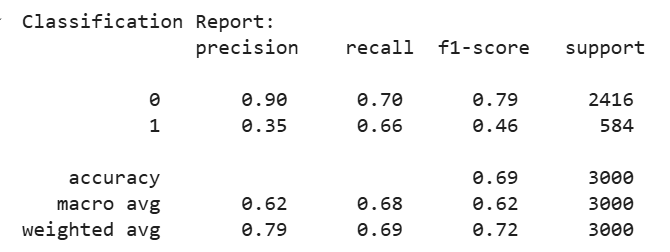
**True Positives (TP):** 388 → Correctly predicted churned customers

### 7.2.2 Classification Report

The classification report provides a comprehensive breakdown of the logistic regression model’s performance across both classes—churned and non-churned customers—using precision, recall, and F1-score metrics.

* For the non-churned class (0), the model achieved high precision (90%) and reasonable recall (70%), indicating that it correctly identified the majority of customers who did not churn, with minimal false positives.
* For the churned class (1), the model exhibited lower precision (35%) but a moderate recall (66%), suggesting that while it captured a considerable portion of actual churners, it also misclassified a substantial number of non-churned customers as churned.
* The F1-score of 0.46 for churned customers reflects a moderate balance between precision and recall in identifying the minority class, which is a key focus in churn prediction tasks.

Overall, the classification report highlights that the model is highly reliable in predicting non-churned customers and moderately effective in identifying churners. While recall for the churn class is acceptable, the relatively low precision indicates potential for further improvement through more advanced modeling techniques or feature engineering.



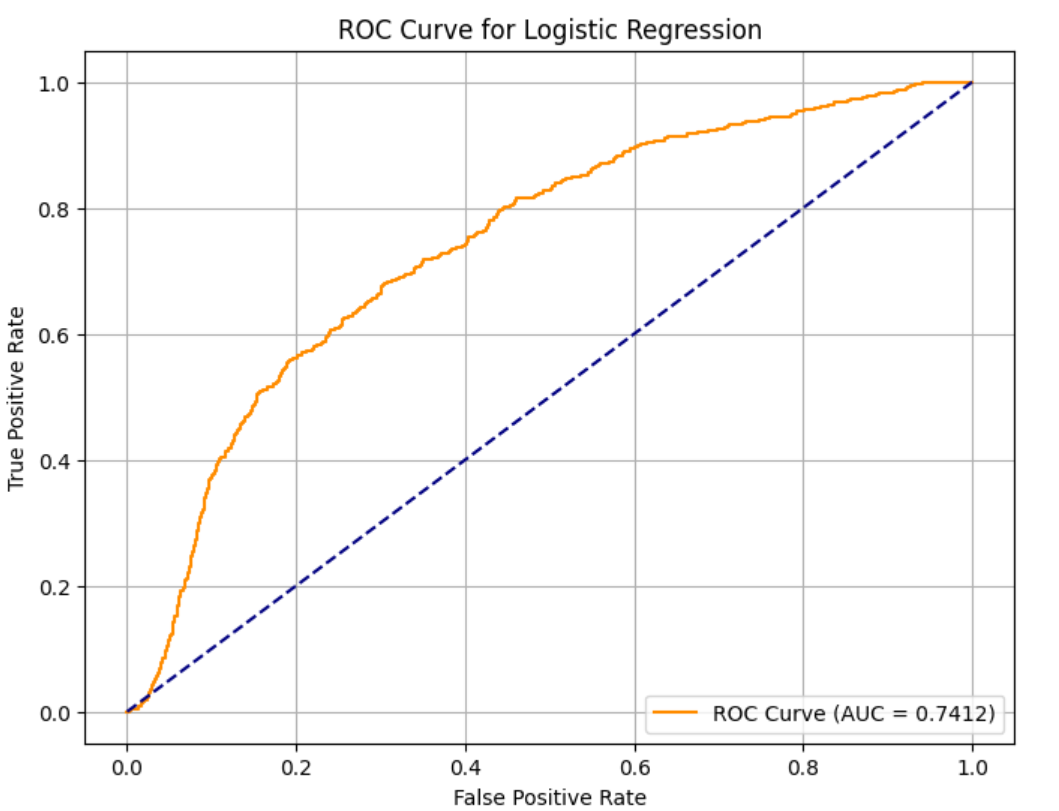
**Fig. 7.7 Classification Report for Logistic Regression Model**

### 7.2.3 ROC Curve and AUC Score

The ROC curve provided a visual representation of the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate for various classification thresholds. The curve for the logistic regression model was positioned well above the diagonal (random classifier baseline), indicating that the model was effective in distinguishing between churned and non-churned customers.

The model achieved an **Area Under the Curve (AUC) score of 0.7412**, which is considered a moderate to strong indicator of classification performance. An AUC greater than 0.7 generally signifies that the model has good discriminative ability, i.e., it is able to correctly rank and classify a large proportion of the churn vs. non-churn observations.

Therefore, the ROC-AUC evaluation confirms that the logistic regression model is reasonably effective at separating churned from non-churned customers and provides a strong baseline for further predictive modeling and business decision-making.



**Fig. 7.8 ROC Curve for Logistic Regression**

## 7.3 Lasso Regression

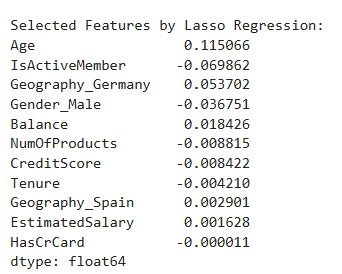
To improve model performance and reduce overfitting, Lasso Regression was employed to identify the most important predictors of customer churn. Lasso (Least Absolute Shrinkage and Selection Operator) is a linear regression technique that includes an L1 penalty, which encourages sparsity by shrinking less relevant feature coefficients toward zero. This helps in both regularization and automatic feature selection.

The dataset was first preprocessed by converting categorical variables into dummy variables using one-hot encoding. Features were standardized using StandardScaler() to ensure equal contribution to the model, as Lasso is sensitive to feature scale. The dataset was then split into training and testing subsets, and LassoCV was applied with five-fold cross-validation to determine the optimal regularization strength.

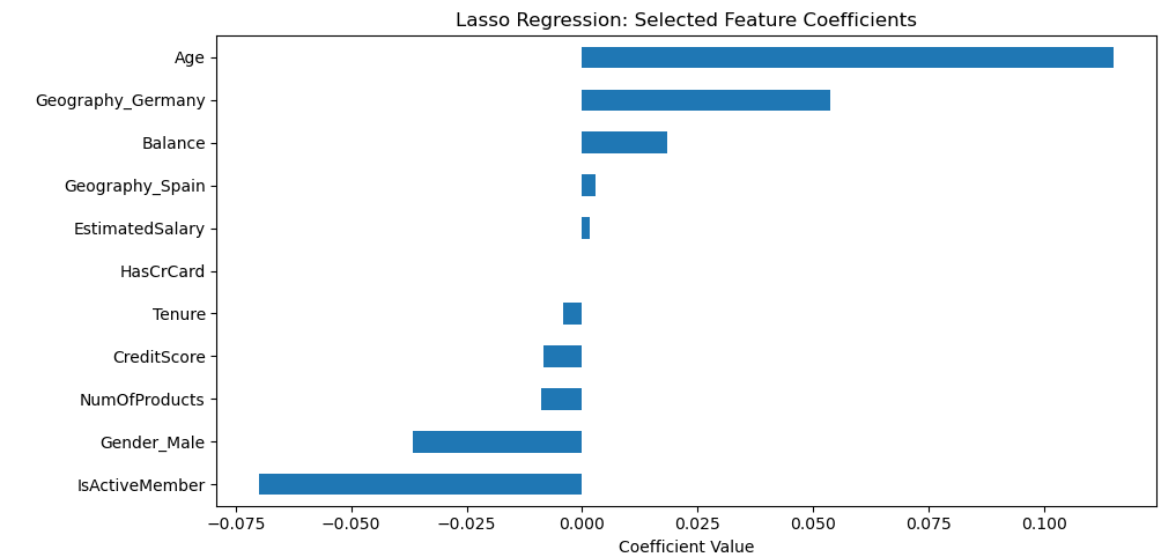
The regression results revealed several variables with non-zero coefficients, indicating their relative importance in predicting customer churn. Among these, Age was identified as the most influential feature, with a positive coefficient suggesting that older customers are more likely to churn. Geography\_Germany also had a notable positive coefficient, meaning customers from Germany have a higher tendency to leave the bank. In contrast, IsActiveMember had a strong negative coefficient, implying that active members are less likely to churn and tend to remain loyal.

Other features such as Gender\_Male, NumOfProducts, and CreditScore contributed moderately, while features like HasCrCard, Tenure, and EstimatedSalary had near-zero coefficients, indicating negligible influence on churn. These results were further supported by a bar chart visualization, which clearly highlighted the most impactful variables.

Overall, Lasso Regression simplified the feature set and improved model interpretability by isolating the variables that matter most in customer churn prediction. This technique is valuable for building efficient predictive models and informing strategic decision-making.



**Fig 7.9 Lasso-selected feature coefficients ranked by importance**



**Fig 7.10 Bar chart showing magnitudes of Lasso regression coefficients**

# 8. Conclusion

This study set out to identify and understand the main factors contributing to customer churn in the banking sector by applying both statistical methods and machine learning techniques. The analysis began with a thorough exploratory data examination, revealing the distribution, range, and outliers in both numerical and categorical variables. Clear visualizations such as histograms, box plots, and pairplots helped identify initial patterns and potential predictors of churn.

Inferential statistics provided further insight into the behavior of churned versus retained customers. Welch’s T-tests and ANOVA highlighted that Age, Balance, and Number of Products had statistically significant differences between churned and non-churned customers. The Chi-Square Test of Independence showed that Geography, Gender, and IsActiveMember were significantly associated with churn behavior, making them critical categorical variables for predictive modeling.

The Correlation Analysis demonstrated that Age and IsActiveMember had the strongest linear relationships with churn. However, other features like EstimatedSalary, Tenure, and HasCrCard showed little to no correlation, indicating that not all numerical features carry predictive value.

To enhance prediction and reduce model complexity, Lasso Regression was used for feature selection. The results confirmed that Age, Geography\_Germany, IsActiveMember, and Balance are among the most informative features. These were identified through non-zero coefficients and were visualized to support interpretation. Lasso’s ability to shrink irrelevant coefficients to zero enabled the construction of a more efficient, interpretable model.

In summary, this analysis successfully combined statistical theory with practical modeling to uncover the most influential factors driving churn. The findings provide actionable insights for banks to improve customer retention—particularly by focusing on older, inactive customers and tailoring strategies for specific geographic segments. This data-driven approach can inform customer engagement policies and improve long-term loyalty.