**Project Report**

**Utilizing Logistics Regression and Decision Tree Model**

**to Predict Customer Churn for a Financial Services Company**

# **Executive Summary**

Companies in the financial services industry can have a median annual churn rate to be as high as 19% (Tessitore). In this industry, preventing churn can save millions of dollars, and hence predicting churn before it happens is crucial. Developing a solid model to predict churn will generate insight into the customer profile that has a higher possibility to churn so that the company will be able to create quick prevention measures.

The target of this project is to create a logistic regression model and the decision tree model to predict customer churn. It will start by preparing the data and conducting a thorough exploratory data analysis using graphs to visually depict the variables’ influence on customer churn and remove outliers.

The logistic regression and decision tree model will be established using a provided training dataset. Another dataset will be utilized to test the model’s accuracy as a testing dataset. The project will compare the two models and determine which model gave the more accurate result. It will show the most important variables that can affect customer churn.

Both models show an accuracy higher than 80% which is sufficient accuracy in predicting customer churn. The model shows that there are five most important variables that significantly affect churn: customer’s age, gender, and country, as well as the number of products used and whether the customer is an active member. Based on the result, the project recommends the company to analyze its customer relationship program in terms of its service suitability for the customers' demographic and the activity of the customers. It specifically recommends the company to improve its operation in Germany to decrease its churn rate, as well as create programs for older customers, female customers, and inactive members.

Keywords: churn, predictive model, logistic regression, decision tree

# **Introduction**

Businesses rely heavily on their ability to retain customers. Losing customers would mean a reduction in overall customer lifetime value, and it directly affects profitability. Moreover, the loss would be translated into higher customer acquisition costs since more new customers should be acquired to balance off the number of leaving customers. It is imperative that this issue has raised major concerns in various industries.

While the issue is relevant in every industry, it is particularly vital in the financial services industry. Companies in this industry can have a median annual churn rate to be as high as 19% (Tessitore). Preventing churn can save millions of dollars, and hence predicting churn before it happens is crucial. Developing a solid model to predict churn will generate insight into the customer profile that has a higher possibility to exit the service, so quick prevention measures can be made. The objective of this project is to create a logistic regression model and the decision tree model to predict customer churn based on the churn data of a financial service firm.

# **Methodology**

The project will start by examining the dataset to see whether a cleaning process would be necessary. After the data is ready to be processed, it will continue to the exploratory data analysis, which will analyze each variable and its relationship to whether the customer has exited the service or has not. It will use graphs to show the visual summary of the relationship. It will also try to eliminate outliers from numerical variables.

After the dataset is cleaned, the project will use it to create a logistic regression model and decision tree model. The model will be used to predict customer churn from another dataset that consists of 1000 observations. This testing dataset will be used to test the model’s accuracy by comparing the model’s result with the actual result.

# **Data Overview**

## **Variables, Data Types, and Pre-processing**

The training data consists of 9,000 observations with 12 variables, ranging from customers’ profiles, such as ID, age, and gender, to the profile of their interaction with the company, such as tenures, balance, and estimated salary. There is also the target variable Exited, which will be valued as yes when the customer has churned or no when the customer has not This target variable will be made as a factor for the predictive model.

Variables Gender and Geography will be converted from character variables to factors. Binary variables and the number of products will also be converted to factors so that the program understand the possible values of the variable. After the conversion, the data type for each variable in the dataset can be shown as follows.

|  |  |
| --- | --- |
| The data type for every variable obtained from R | Type |
| $ CustomerId : int 15634602 15647311 15619304 15701354...  $ CreditScore : int 619 608 502 699 850 645 822 376 501 ...  $ Geography : Factor w/ 3 levels "France","Germany",..: 1 3  1 1 3 3 1 2 1 1 ...  $ Gender : Factor w/ 2 levels "Female","Male": 1 1 1 1 1  2 2 1 2 2 ...  $ Age : int 42 41 42 39 43 44 50 29 44 27 ...  $ Tenure : int 2 1 8 1 2 8 7 4 4 2 ...  $ Balance : num 0 83808 159661 0 125511 ...  $ NumOfProducts : Factor w/ 4 levels "1","2","3","4": 1 1 3 2 1  2 2 4 2 1 ...  $ HasCrCard : Factor w/ 2 levels "no","yes": 2 1 2 1 2 2 2 2  1 2 ...  $ IsActiveMember : Factor w/ 2 levels "no","yes": 2 2 1 1 2 1 ...  $ EstimatedSalary: num 101349 112543 113932 93827 79084 ...  $ Exited : Factor w/ 2 levels "no","yes": 2 1 2 1 1 2 ... | Numerical  Numerical  Nominal  Nominal  Numerical  Numerical  Numerical  Nominal  Nominal  Nominal  Numerical  Nominal |

Tabel 1 Data type for every variable after the conversion into factors

To ensure data quality, the program should check whether there are missing values and eliminate unnecessary variables that will not be used in the model. A preliminary check has found that there are no missing values in the dataset. The training dataset can also eliminate variable CustomerID, since it will not be used to determine whether a customer will churn or not. The program can also check if there is a correlation between numeric variables. If there is a strong correlation between two numeric variables, the analysis can eliminate one of them in the logistic regression analysis. However, this project will include all numeric variables first before making adjustments. Thus, the project will analyze all 9,000 observations, using 10 predictor variables to predict the target variable Exited.

## **Exploratory Data Analysis (EDA)**

To better understand the dataset, the project will execute a thorough exploratory data analysis (EDA). The EDA will be executed for categorical variables first and then numerical variables. For categorical variables, the EDA will provide bar graphs for each variable and with an overlay to target variable Exited. The bar graph will also provide the normalized view to better understand how a variable may be useful to predict Exited. Contingency tables will also be provided to better visualize the categorical variables and how it relates to Churn.

For numerical variables, descriptive statistics will be provided along with a histogram that shows how the variables may relate to Churn. The EDA will look for outliers using boxplot. The project will also examine the correlation between these numerical variables.

### **EDA for categorical variables**

The bar graphs that show how each categorical variable relates to the target variable Exited are shown in **Appendix 1**, along with the normalized bar graphs and contingency tables. There are multiple variables that may prove to be significant predictors of Exited variable. For instance, in the variable Geography, customers from Germany may have a higher chance of churning, compared to customers from France and Spain. It can be shown by the overlay bar graph and contingency in Appendix 1. While the dataset has more observations for France, the normalized bar graph shows that more than 25% of the customers from Germany have exited. It can be an initial signal that the company’s operation in Germany may have to improve. The company may have to analyze the differences in these countries that may affect a higher churn rate especially when the deeper analysis shows that these are not because of random chances.

Executing the same analysis for each variable, the project obtained a bar graph overlay for each variable related to Exited, the normalized bar graph, and the contingency table, which are shown in **Appendix 1**. Using these graphs, EDA can see some potential variables to predict Exited and some that are not. For instance, the variable HasCrCard may not provide significant influences to Exited. The logistic regression will have a deeper analysis of how these variables may or may not be useful to predict Churn.

### **EDA for numerical variables**

The descriptive statistics for the numerical variables of the dataset are shown in **Appendix 2**, along with the histogram overlay and its normalized version. The summary of the numerical variables and boxplot shows that there are outliers in CreditScore and Age. The outliers are then excluded from the analysis. This is important because it can shift the data. For instance, before and after outliers are excluded, the histogram for the Age variable is shown in Appendix 2. After the outliers are excluded, the Age variable seemed to show that old customers have a higher churn rate, which is not clear in the upper plot where outliers are still included.

The histogram overlay and the normalized version can provide a preliminary analysis of how each numerical variable may influence Exited. By default, R divides the histogram into 30 bins. While there are a different number of observations in each bin, the normalized version of the histogram shows that churning customers proportion to those who have not. Histograms for the other numerical variables are shown in **Appendix 2**, along with the summary of the data. Aside from the variable Age, the plots for other variable shows an unclear effect of the variables to Exited. Customers with low credit scores have a high churn rate, but it does not show a linear effect when the credit score increases. The variables Balance, Tenure, and EstimatedSalary seemed to have no linear influence. The logistic regression will have a deeper analysis of how each of these numerical variables may or may not be useful to predict Exited.

Aside from how each numerical variable relates to churn, a correlation between these numerical variables can also be examined to better understand the data. The correlation for each variable can be shown through the correlation matrix below, which shows that there is no strong relationship between the variables. Hence, the logistic regression can be directly performed.

# **Results and Analysis**

The model to predict churn (Exited variable) will be developed using logistic regression and a decision tree. Both are also the most common technique used by data analysts to model binary response prediction since both are well-developed and well-studied methods. In this case, the binary response would be Yes or No in the variable Exited, describing whether a customer has churned or not. The training and testing dataset has been provided.

## **Logistic Regression**

The analysis will provide three logistic regression models. The first model will use all of the variables to predict Exited and evaluate which variables have a high significance in the model. The second model will use variables with higher significance that was found in the first model to predict Churn. The third model will not use any variables to predict Exited and let it be decided by random. The analysis will create a comparison of the three models using the analysis of variance (ANOVA). It will also compare the accuracy of the model using the test dataset.

Using the *glm* function in R with the family set to binomial, the logistic regression summary can be obtained and it is shown below in Appendix 3. The coefficients table shows the coefficient for each variable, as well as its standard error, z-value, and p-value. The stars at the end of each row depict the significance of the variables. Variables with three stars have higher significance than variables with two or lower numbers of stars. The analysis found that there are various significant variables, such as Geography, Gender, Age, NumOfProducts, and IsActiveMember. These variables have a significant coefficient estimate value and a lower than 0.05 p-value which means that it is statistically significant.

Logistic regression will provide a linear model to predict the log odds of churn, with the estimated value above being the coefficient value of the variable. By solving the logit, the model can provide the churn odds equation. Since the calculation will involve numerous variables with complicated numbers, the prediction will be calculated using a built-in function in R.

The same process can be executed to create the second model. The model will use five variables that are the most significant from the first model: Geography, Gender, Age, NumOfProducts, and IsActiveMember. The logistic regression summary result is shown on Appendix 3-2. Notice that most of the variables are statistically significant with a p-value less than 0.05. Using the same analysis as the previous model, the linear logit equation can be written using the estimated value above as the coefficient for each variable, but the calculation will be performed by the program.

### **ANOVA Test**

ANOVA for the first model and second model shows a p-value of 0.4157 which is higher than 0.05. It means that the difference in variance between the first and second models is not statistically significant. However, the ANOVA of the first and second models shows a significant p-value when each of the models is compared to a random model without predictor variables. It means that both are significantly better models than random chances. The ANOVA will provide the model with less variance when modeling the training dataset. However, there is still a possibility of overfitting, in which a model can not be consistently good in predicting the target variable. In supervised methods, the model should also be tested using a test dataset to determine the accuracy and see whether overfitting exists.

### **Model Accuracy Test**

The accuracy of the first and second models above can be tested using the test dataset that is provided in the file Churn\_score\_data.csv (contains the variable to be predicted) and Churn\_score\_true\_data.csv (contains the actual Exited variable). Using R, the test will use variable values of each customer in the test dataset and predict whether the customer will churn or not. Then, the program will compare the prediction and the actual data to determine the accuracy level of the model.

The program will calculate the churn odds for the first customer. If the customer’s churn odd is higher than 0.5 then the customer will be classified as “yes” in the Exited prediction. On the contrary, if the churn odd is lower than 0.5, then the customer is predicted to stay. Iterating this process for each customer in the testing dataset, the program will get the overall accuracy of the model. For the first and second models, the accuracy is 0.83 which shows that the difference in both models is insignificant. The confusion matrix for both models is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| Model 1 Confusion matrix | | Predicted data | |
| No | Yes |
| Actual data | No | 752 | 69 |
| Yes | 101 | 78 |

|  |  |  |  |
| --- | --- | --- | --- |
| Model 2 Confusion matrix | | Predicted data | |
| No | Yes |
| Actual data | No | 752 | 69 |
| Yes | 102 | 77 |

*Table 2. The confusion matrix of the first and second models.*

The matrix shows that the first and second model accurately predicts 752 cases of not churning and 78 cases of churning customers (77 for the second model). The accuracy calculated earlier can be obtained by dividing the number of accurate predictions by the number of inaccurate predictions. For instance, the first model accuracy is 0.83 which is obtained from the correct prediction (752+78) divided by all predictions (752+78+101+69). The result shows that the difference between the first and second models is insignificant. Hence, the project may choose either model since both have a similar level of accuracy.

## **Decision Tree**

Using R, the decision tree can be established from the training dataset as a comparative to the logistic regression model. The result of the decision tree diagram can be shown in **Appendix 4**. The root node used the variable Age, which means that the variable is the most important one in predicting customer churn. The model consists of 7 leaf nodes and 5 branch nodes. Note that the model only uses 4 variables out of the 10 variables in the dataset. Here, the most important variables to predict churn are Age, NumOfProducts, IsActiveMember, and Geography. This result is slightly consistent with the second model of the logistic regression model which only uses the same 4 most important variables with the addition of the Gender variable.

|  |  |  |  |
| --- | --- | --- | --- |
| Decision Tree Confusion matrix | | Predicted data | |
| No | Yes |
| Actual data | No | 784 | 106 |
| Yes | 37 | 73 |

*Table 3. Confusion matrix of how the prediction accuracy of the model in R.*

The confusion matrix for the decision tree is shown above. The matrix shows that the decision tree model accurately predicts 784 cases of not churning and 73 cases of churning customers. The accuracy of this model can be obtained by dividing the number of accurate predictions by the number of inaccurate predictions. Using the same calculation as above, the accuracy of this model is obtained as 0.857. The decision tree model is proven to have a better accuracy level than the logistic regression model. Both methods reach more than 80% accuracy which is a good sign that the models are consistent.

# **Discussion: Conclusion and Recommendation**

The application of logistic regression and decision tree technique primarily found that several variables are the most important to predicting customer churn. Those variables are Age, NumOfProducts, IsActiveMember, Geography, and Gender. It is important to find logical reasoning for how these variables significantly affect and determine churn.

## **Demographics: Age and Gender**

The second logistic regression model shows that a one-year increase in age will increase the logit churn odds by 0.11, which means that the older customers will have higher churn odds. Similarly, the decision tree uses the Age variable as the root node with 45 years old as the criterion. The bar graph also shows that older customers have a higher churn rate. The company should analyze the reason behind this data. The company should try to understand the need of older customers that may differ from younger customers. Then, the company can make efforts to improve the services to be more appealing to older customers.

While the decision tree does not use Gender as a predictor variable, the logistic regression shows a statistically significant influence of the variable. The result shows that female customers have a higher churn rate. The company should analyze the difference in needs of male and female customers that may lead to a higher rate of churning of female customers. The company can create customer relationship programs that specifically target female customers.

The first recommendation is for the company to give more focus on the customers' needs as they relate to the customers' age and gender. The company can make the services easier to be delivered and more appealing to older customers while also answering the dynamic nature of the younger demographic. The company can develop a special product that answers the need of older customers to retain them within the company’s services. The customer relationship program should also be created to target specifically female customers and inactive members. The company can create female-oriented programs that can entice female customers to stay longer in the company.

## **NumOfProducts**

Intuitively, if a customer uses a higher number of products from one company, the customer will have a lower chance of churning. However, it is not the case in this dataset. The training data set shows that all customers that have the highest number of products of 4 have been churned. The company should analyze why customers that use more products of the company have a higher churn rate. The company should ask whether there are no additional benefits of using more products for the customers or whether the switching cost of churning from a higher number of the company’s products is insignificant.

The company should create more engaging customer relationship programs that provide enticing benefits and incentives for customers that use more than 1 of the company’s products. It can launch a strong marketing campaign that can encourage customers to use the company’s product and stay in the company’s services for a longer period of time. The company should strengthen its competitive advantage so that it can have a strong position in the industry, instilling the perception of higher switching costs.

## **Geography**

The logistic regression model shows that a customer in Germany has a higher probability of churning, which is also shown in the bar graph in **Appendix 1**. It implies that the operation in Germany needs to be improved to decrease its churn rate. The company can start by analyzing the reason for the higher churn rate in Germany and addressing the issues found.

Therefore, the next recommendation is to analyze the company’s operation in Germany to see the reason behind the higher churn rate in the country. The company should strengthen its operation in Germany to decrease the churn rate. It can also compare the operation in Germany to two other countries: France and Spain.

## **IsActiveMember**

The logistic regression and decision tree show that an active member has a lower probability of churning compared to a less active member. The company can use this insight to create programs that will increase the customers’ activity within the company’s services. The company can develop a program that reaches less active members to increase their activity by providing incentives or opportunities to win benefits.

## **Predictive Model to Be Selected**

The decision to choose which predictive model to use is determined by various criteria, such as the model’s accuracy, bias and variance, and ease of use. The first and second logistic regression models as well as the decision tree model have similar accuracy of about 83% to 85%. The ANOVA shows that the difference in the first and second logistic regression models is not statistically significant. The second model is favorable because it uses fewer more important variables. Hence, the option is only between the second regression model and the decision tree.

The decision tree is relatively easier to understand and easier to execute. However, there are several limitations of the decision tree. The first is that it tends to oversimplify the impact of a variable. For instance, in this case, a young customer (<45 years old) with 1 or 2 products will be directly categorized as a churning customer. In reality, several other factors may have been involved in affecting the probability of churning. The second limitation is that the decision tree may have potential misinterpretation, especially for variables with high correlation.

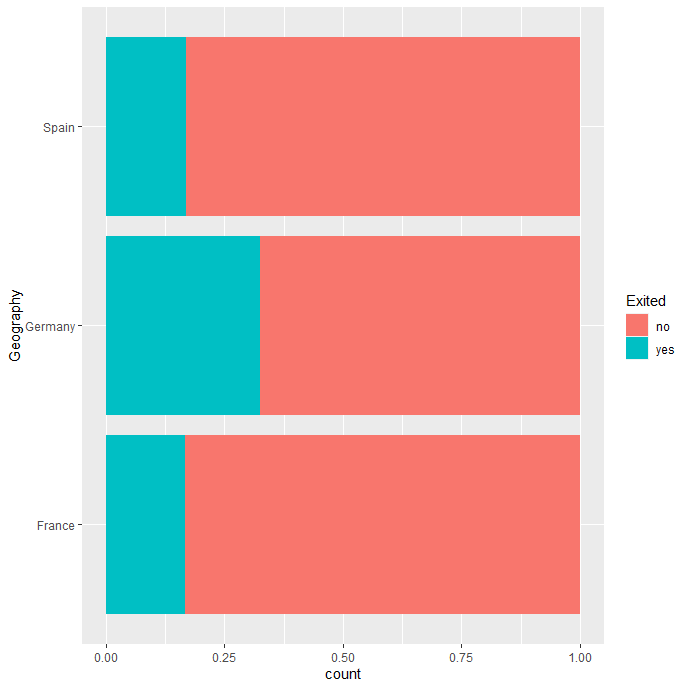
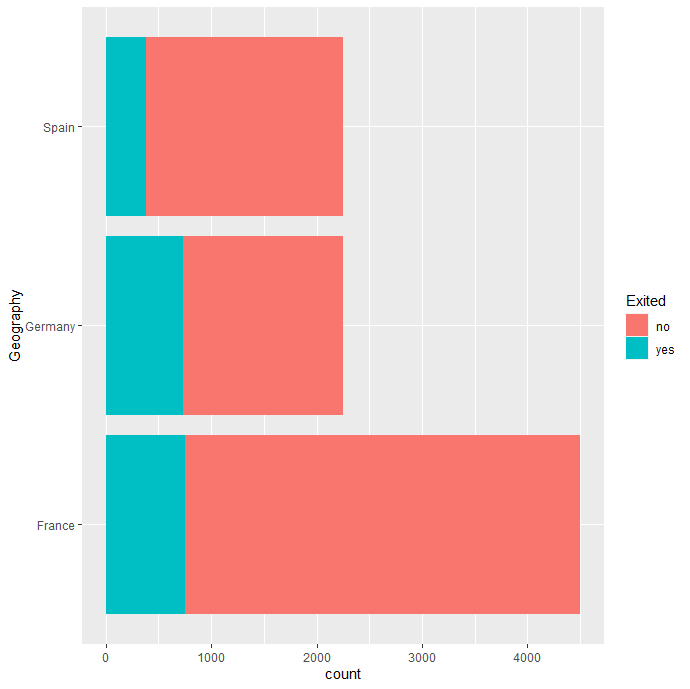
Meanwhile, the logistic regression is more difficult to understand and it requires some statistical and mathematical knowledge to correctly interpret the result. It can lead to wrong interpretation and causes an overall wrong model application. However, if it can be properly performed, it usually provides a better depiction of how a variable influences the target variable.

Research shows that for an academic study aiming to provide solid conclusions about a predictive model, logistic regression is typically better than a decision tree, but, if the goal is to describe the data or solely to create a prediction, then a decision tree is usually better. In this case, logistic regression can provide more insights for further study but the decision tree has higher model accuracy. Thus, the recommendation is to use the decision tree model for its simplicity but to use the logistic regression model for further insights.

# **Appendix 1**

**Exploratory Data Analysis for Categorical Variables**

1. **Geography**



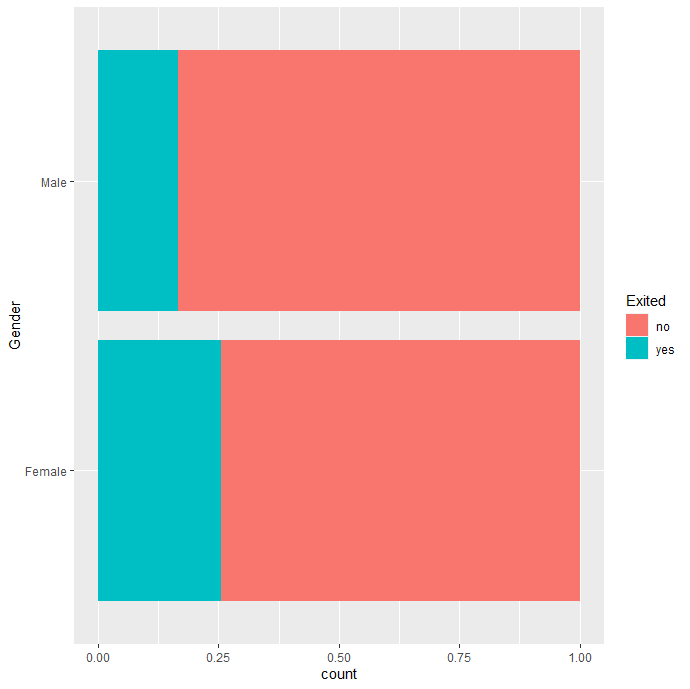
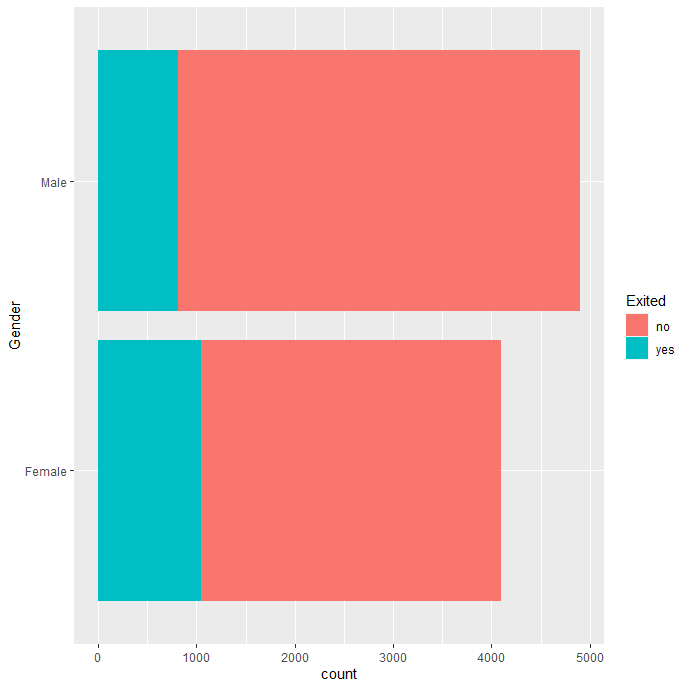
Exited France Germany Spain total

no 3750 1519 1873 7142

yes 748 730 380 1858

total 4498 2249 2253 9000

1. **Gender**



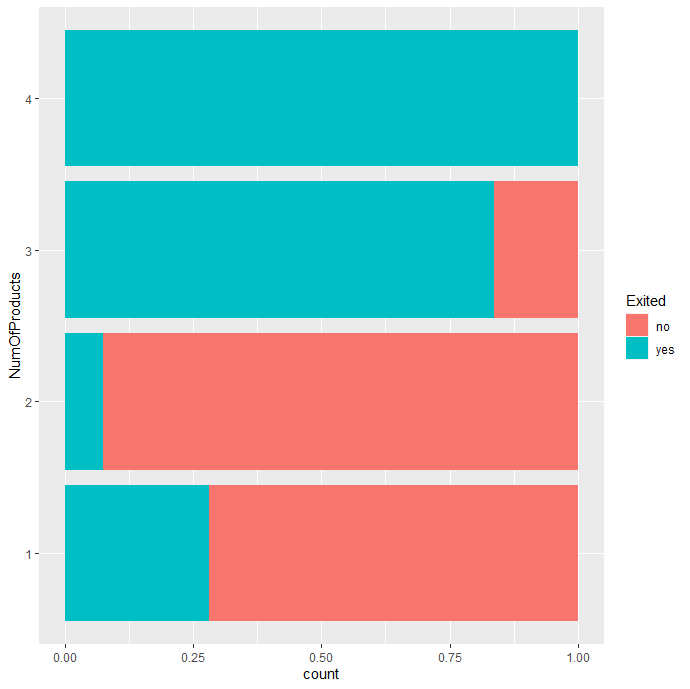
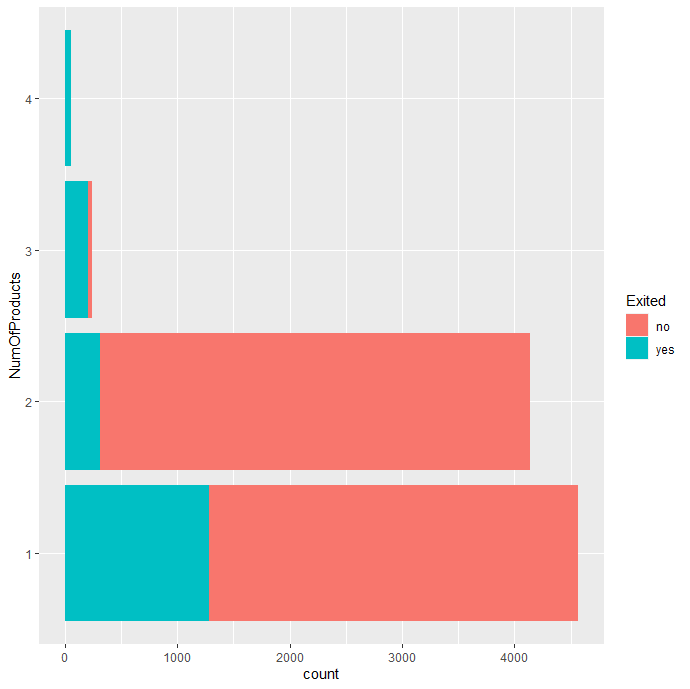
Female Male total

no 3058 4084 7142

yes 1045 813 1858

total 4103 4897 9000

1. **NumOfProducts**



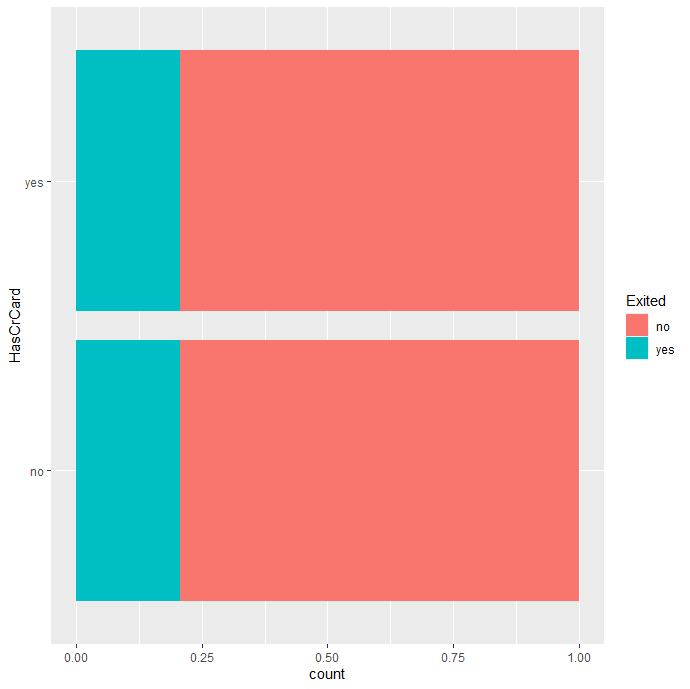
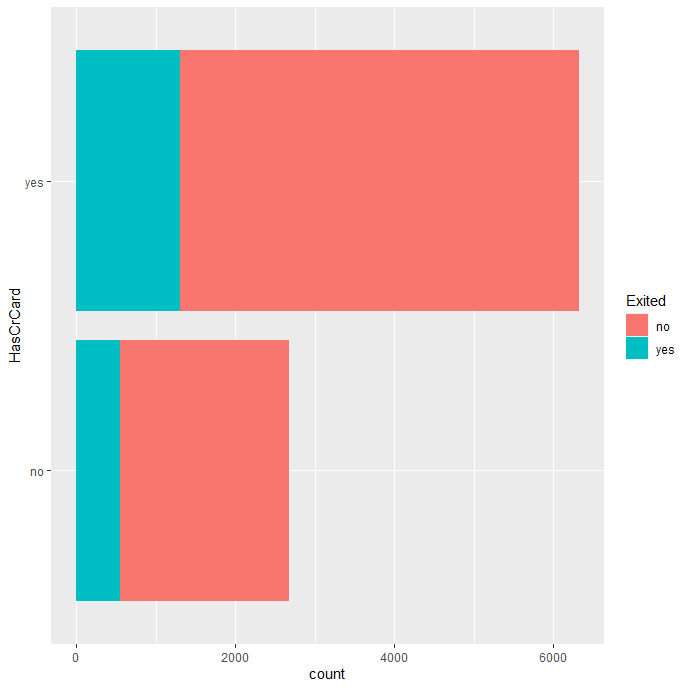
1 2 3 4 total

no 3278 3824 40 0 7142

yes 1288 312 205 53 1858

total 4566 4136 245 53 9000

1. **HasCrCard**



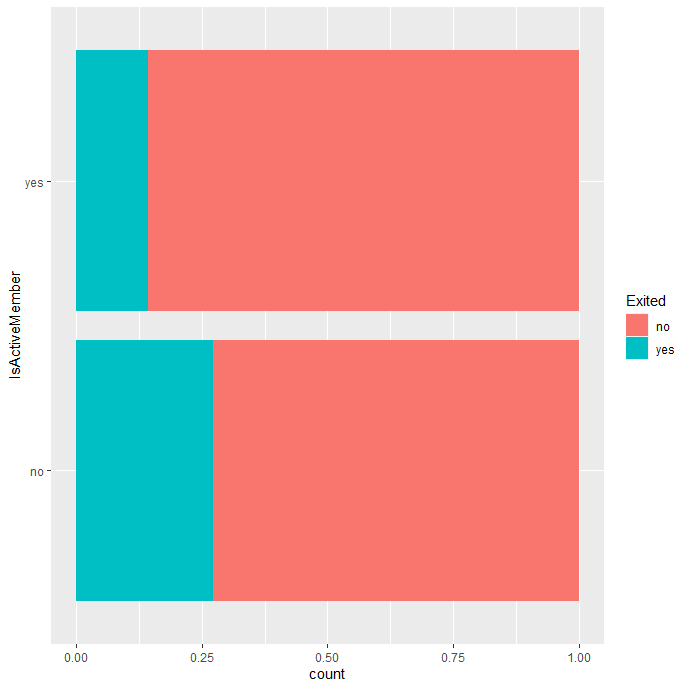
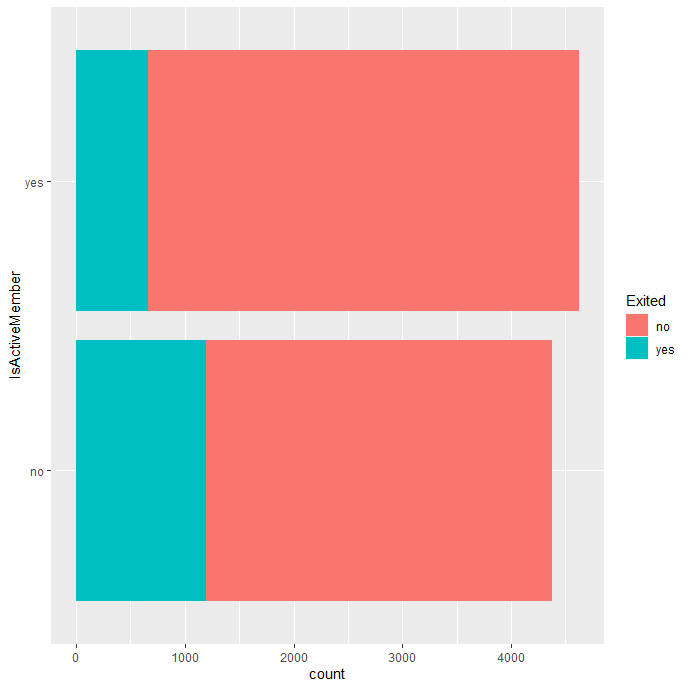
no yes total

no 2123 5019 7142

yes 556 1302 1858

total 2679 6321 9000

1. **IsActiveMember**



no yes total

no 3182 3960 7142

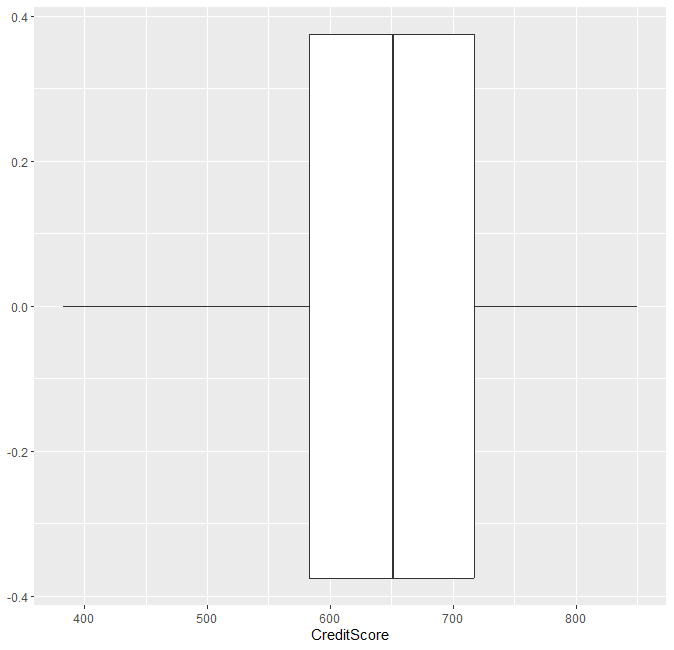
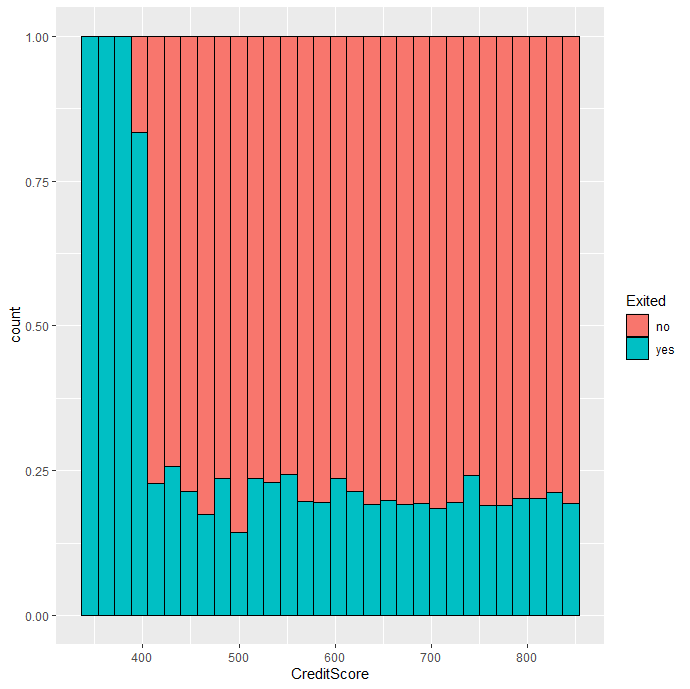
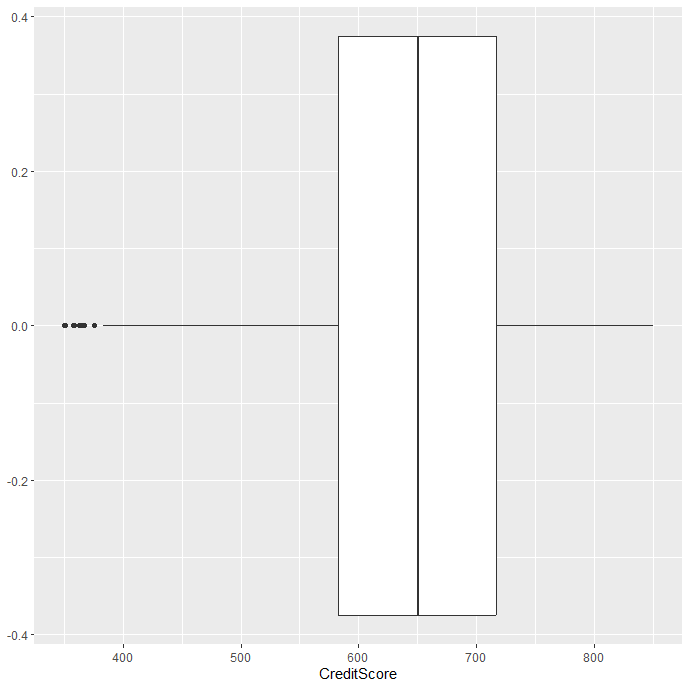
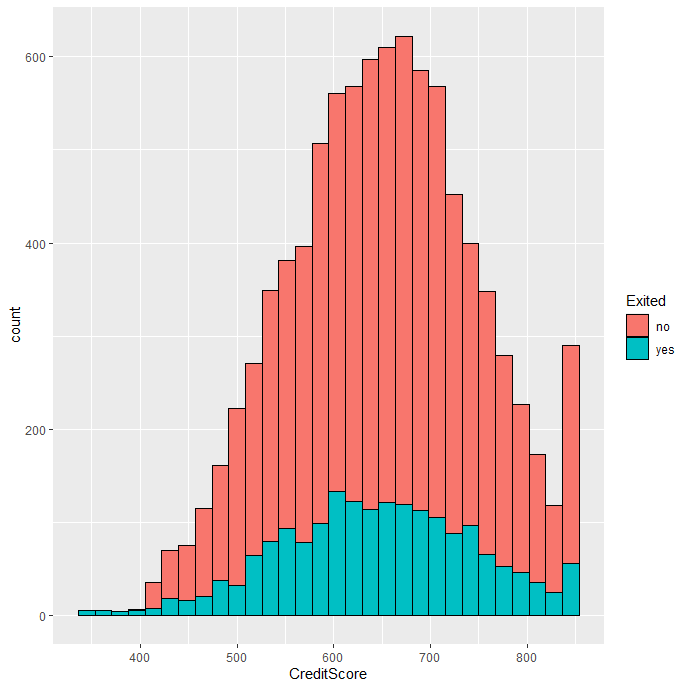
yes 1196 662 1858

total 4378 4622 9000

# **Appendix 2**

**Exploratory Data Analysis for Categorical Variables**

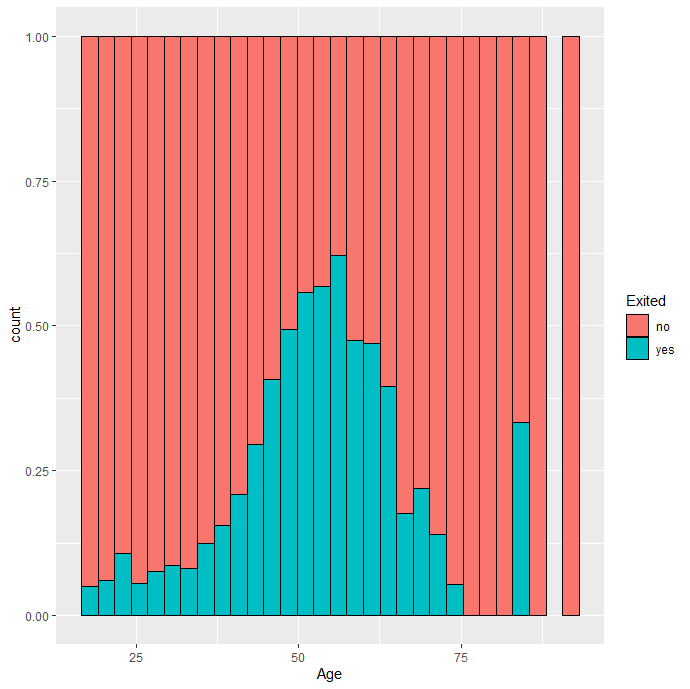
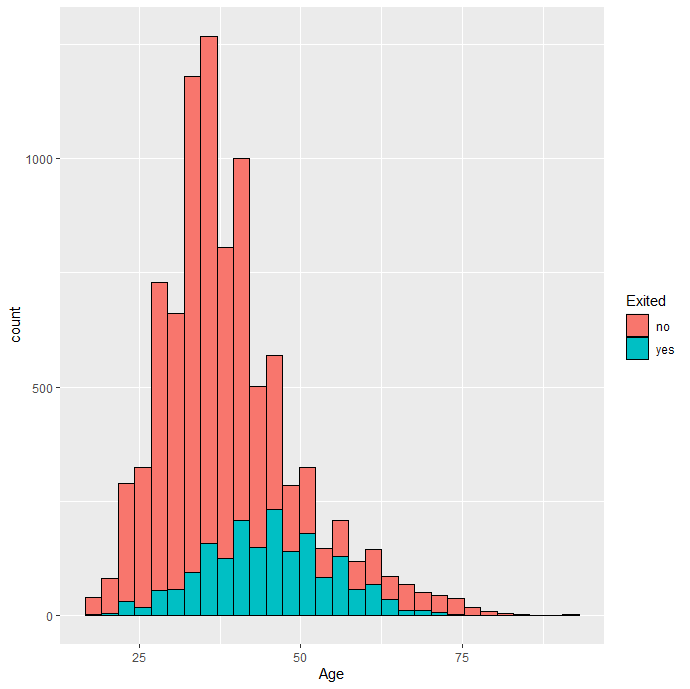
1. **CreditScore**

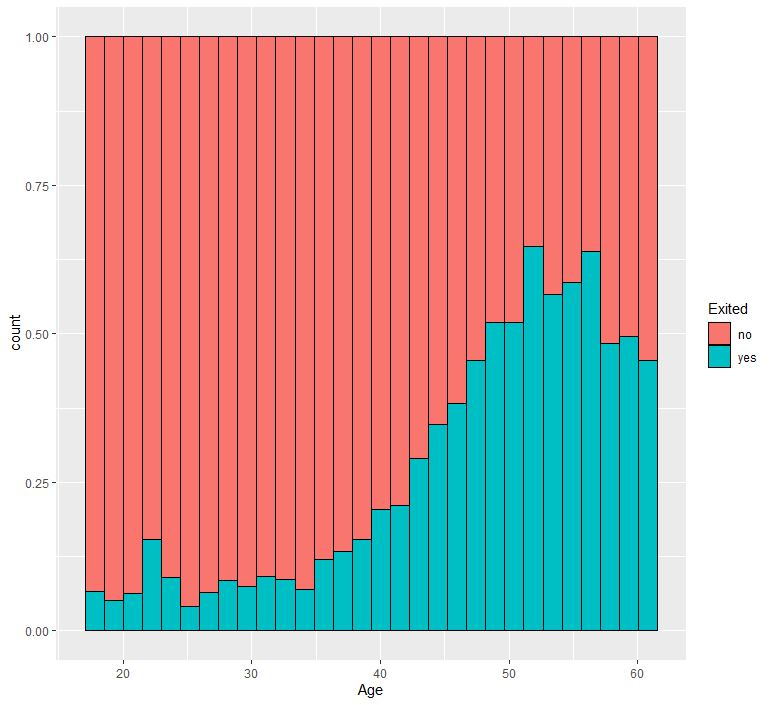
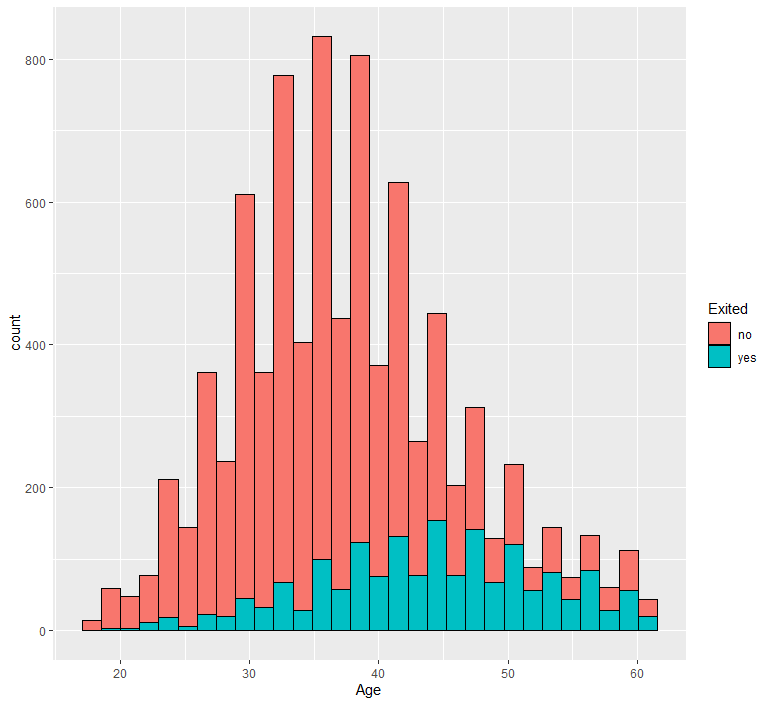


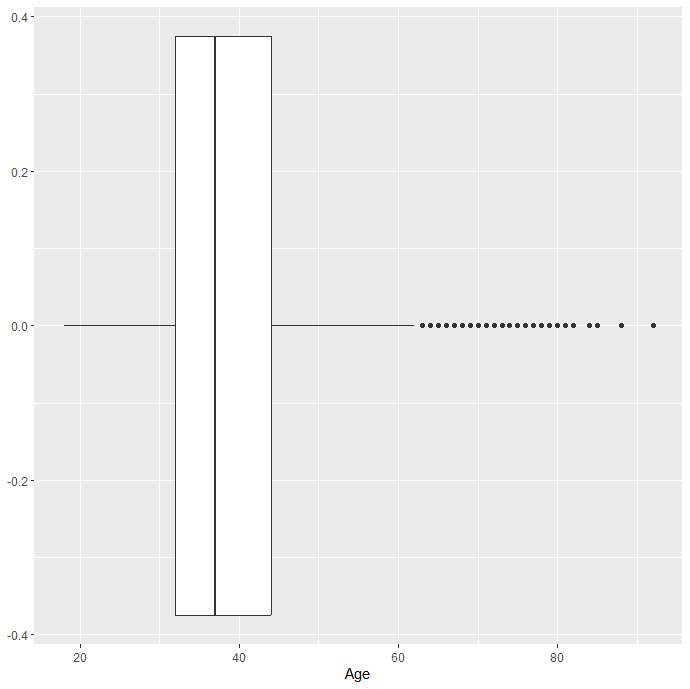
Min. 1st Qu. Median Mean 3rd Qu. Max.

350.0 583.0 651.0 650.1 717.0 850.0

1. **Age**



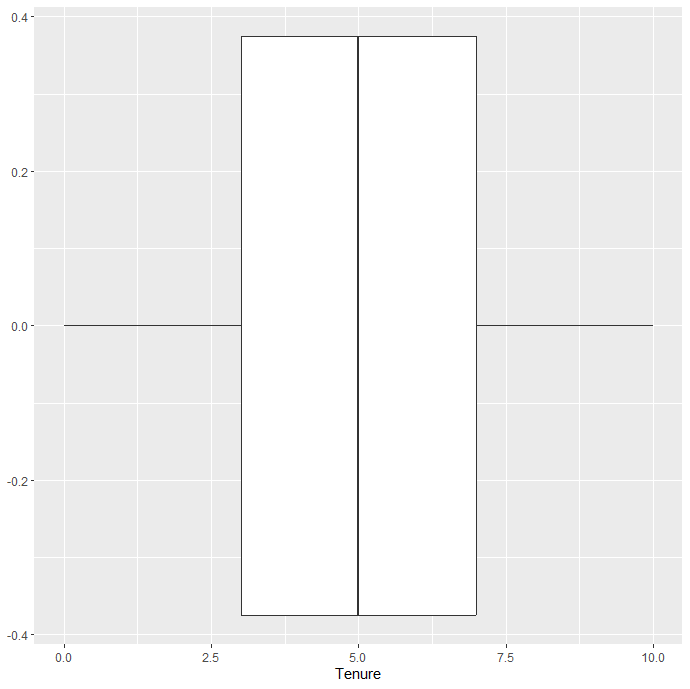
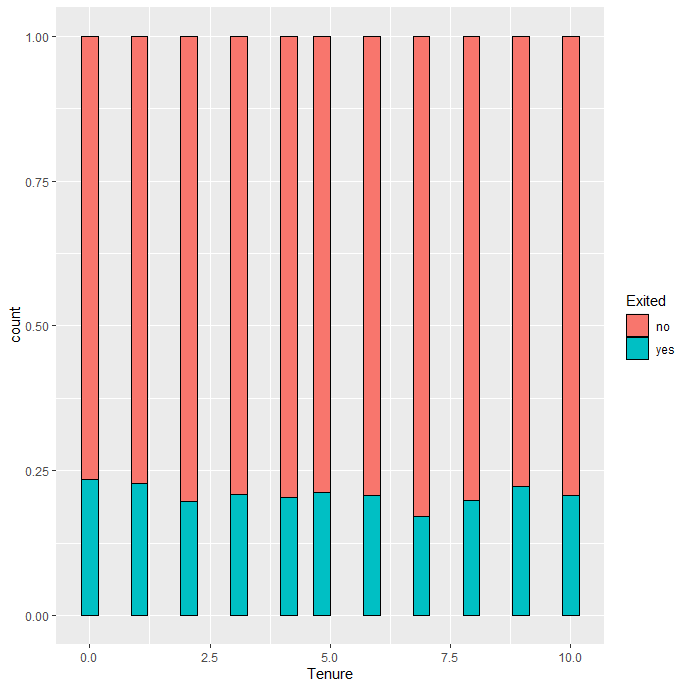
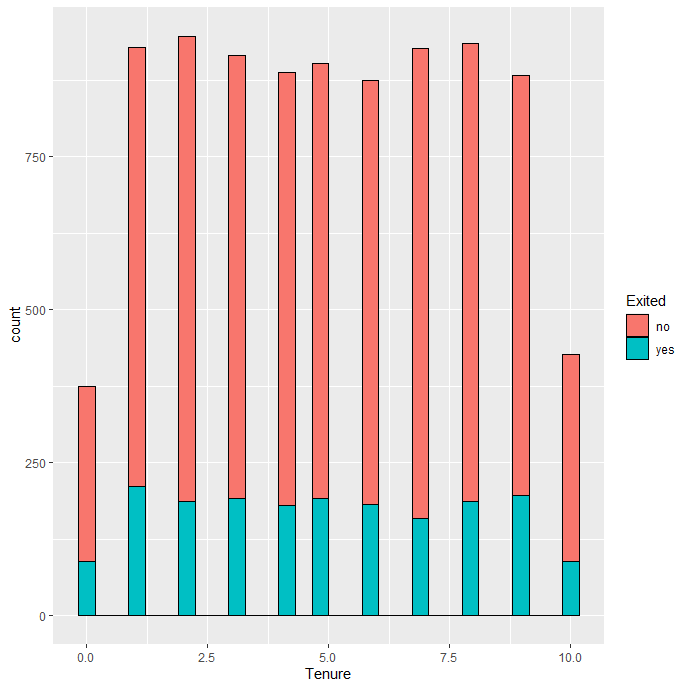




Min. 1st Qu. Median Mean 3rd Qu. Max.

18.00 32.00 37.00 38.93 44.00 92.00

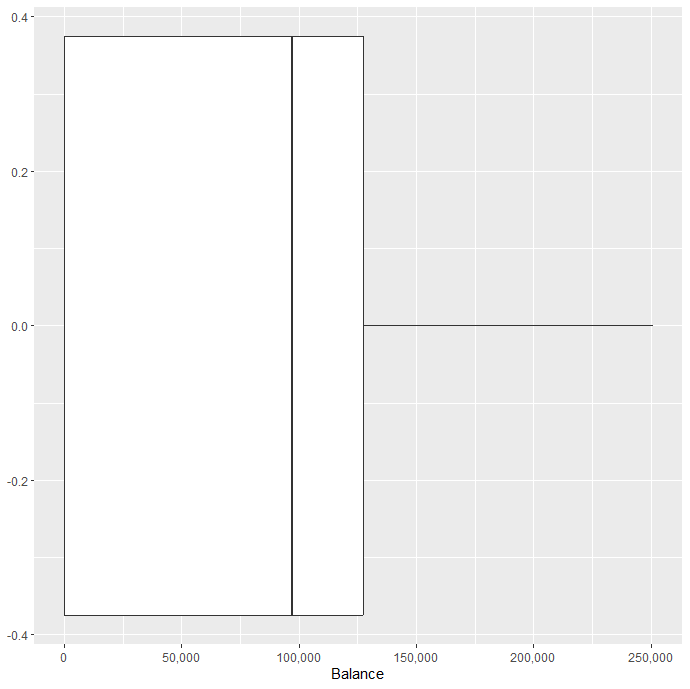
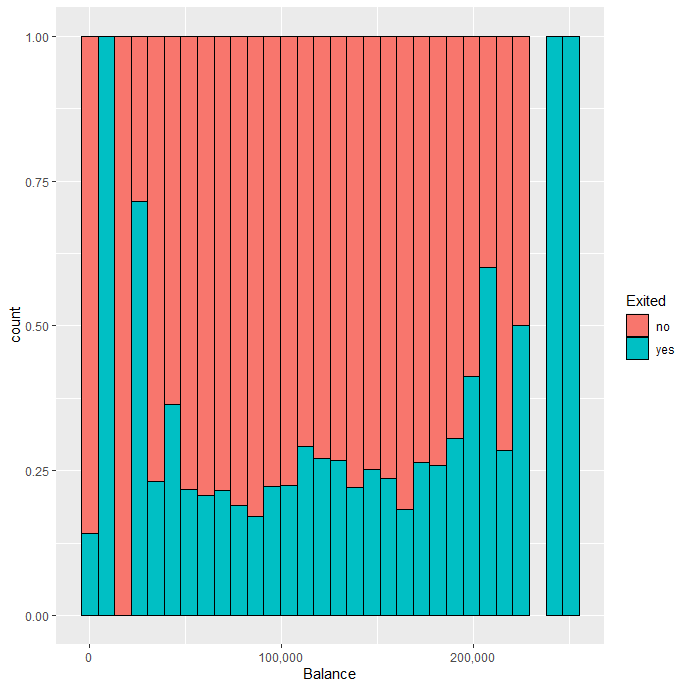
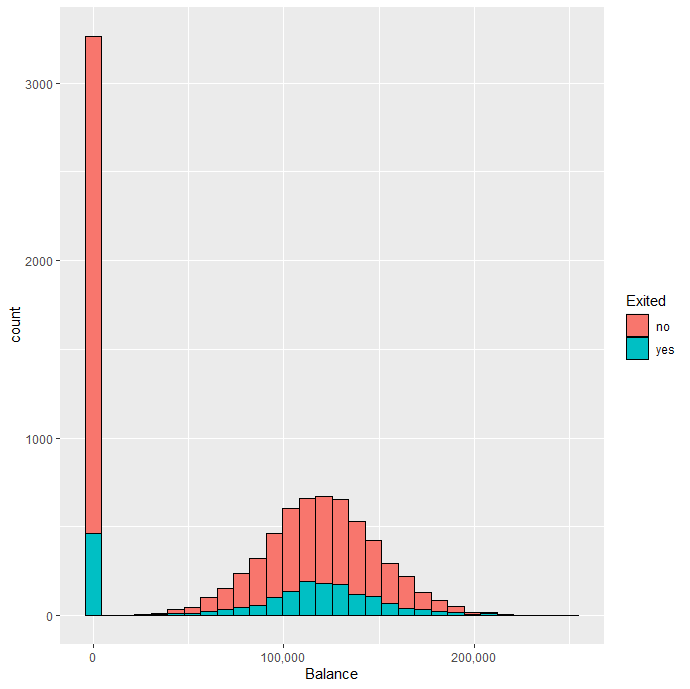
1. **Tenure**



Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 3.000 5.000 5.006 7.000 10.000

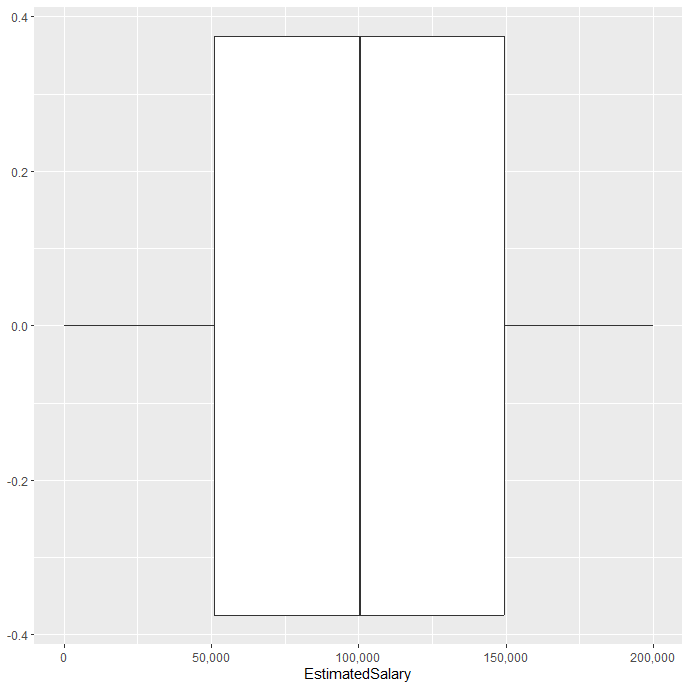
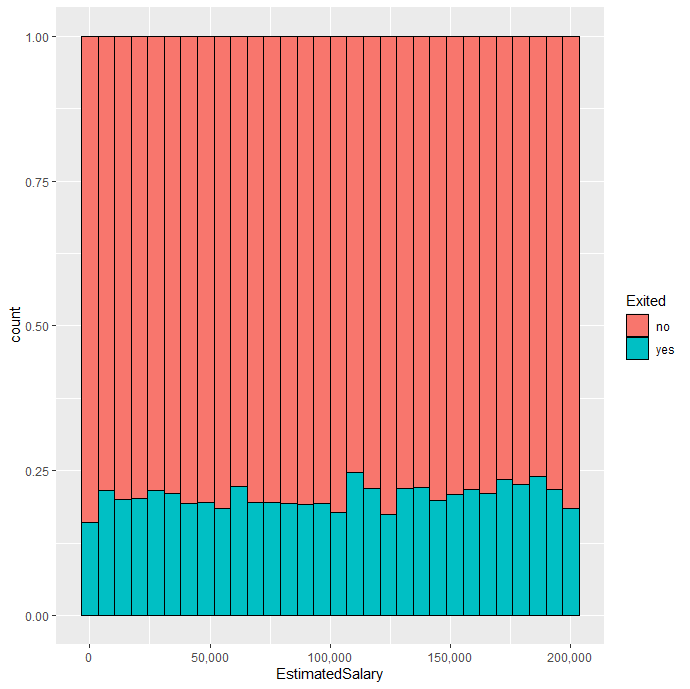
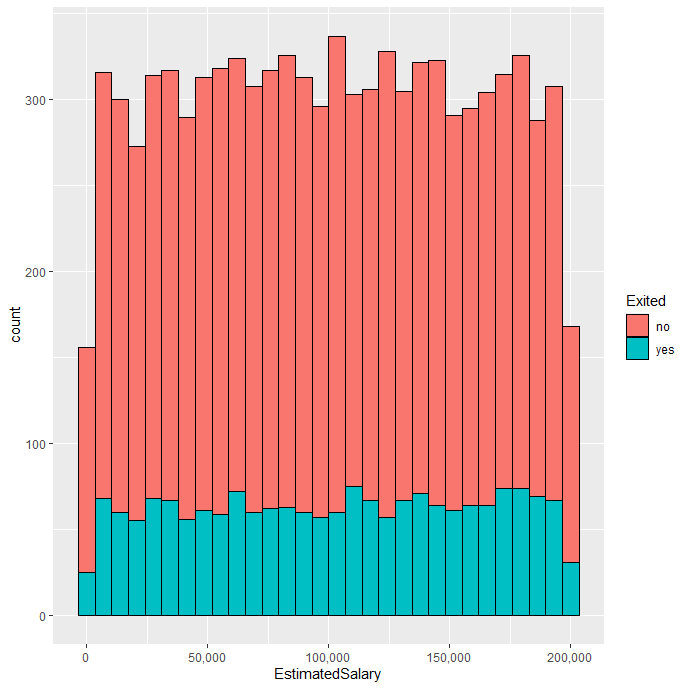
1. **Balance**



Min. 1st Qu. Median Mean 3rd Qu. Max.

0 0 97335 76427 127651 250898

1. **Estimated Salary**



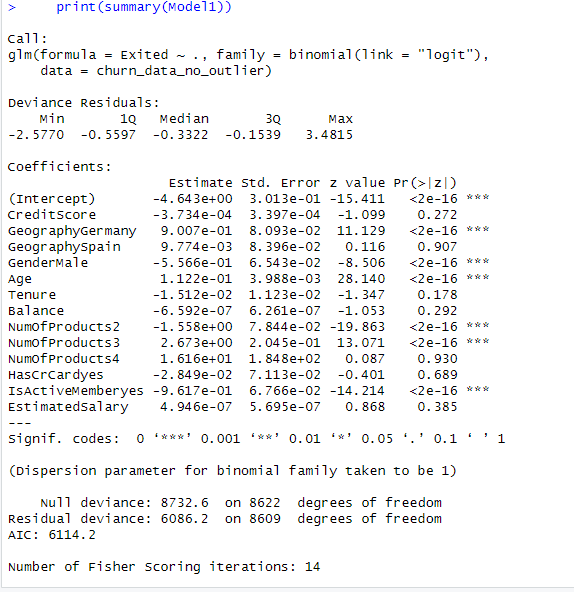
Min. 1st Qu. Median Mean 3rd Qu. Max.

11.58 51014.84 100438.02 100201.48 149405.60 199992.48

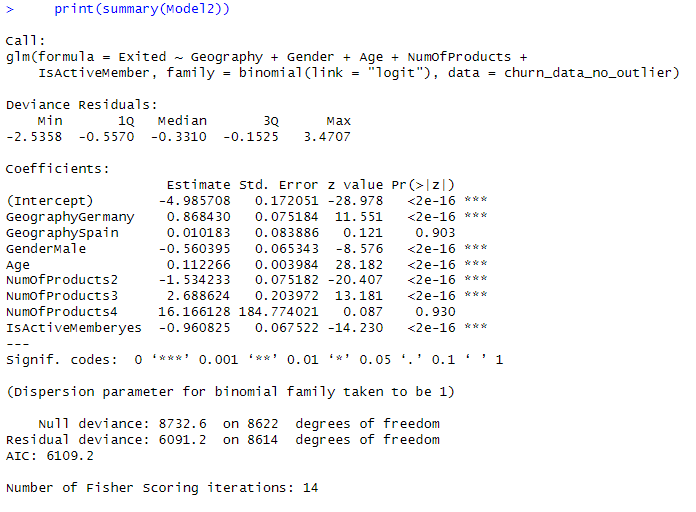
# **Appendix 3**

**Logistic Regression Summary**

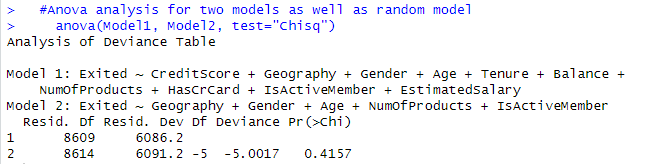
1.

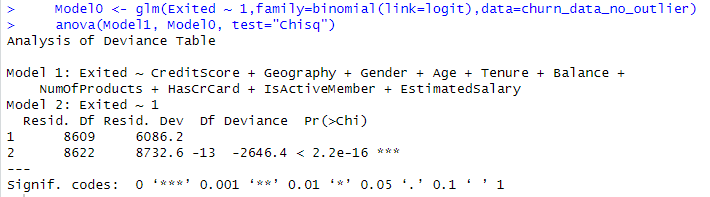


2.



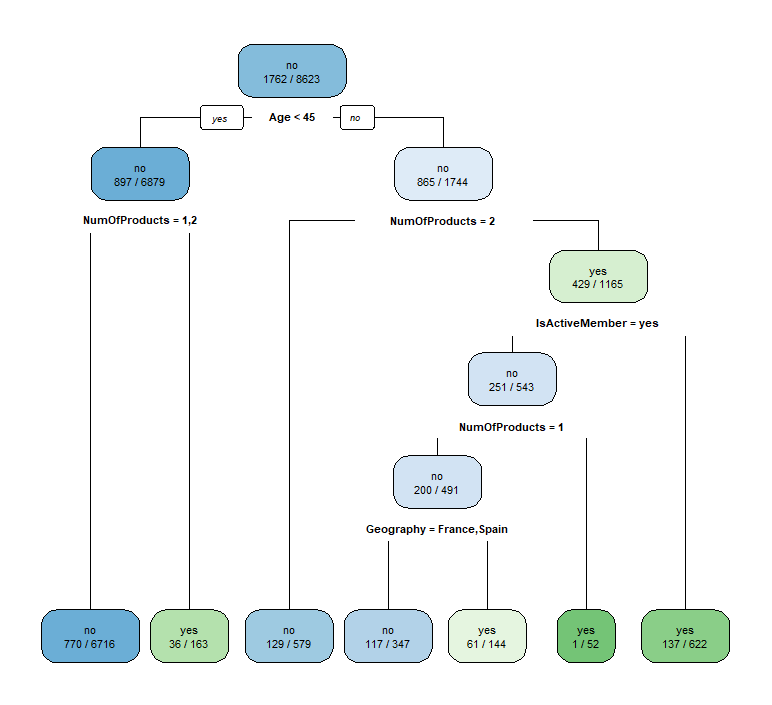
3. Anova Results





# **Appendix 4**

**Decision Tree**

 talhlyCharges 733nd Tota is sufficient.t servicgthe cha