Power of Predictive Analytics: A Statistical Learning Approach for Customer Churn Prediction in Banking

MATH 40028/50028: Statistical Learning

March 13, 2024

Introduction

In today's competitive landscape, customer churn poses a significant obstacle for banks, impacting both financial stability and customer satisfaction. This study embarks on a journey to unravel the intricacies of customer attrition using the "Customer_Churn_Records" dataset. By leveraging sophisticated analytical methods, I would like to uncover the nuanced factors driving customers to discontinue their relationship with the bank.In the endeavor to understand and predict customer behavior, primary objective is to forecast the "Exited" field, leveraging various factors available in our dataset. Through the application of robust statistical models and methods, my aim is to discern patterns and relationships that elucidate the drivers behind customer churn. Through this multifaceted approach, we can aspire to empower banks with actionable insights that enable proactive retention strategies, ultimately fostering long-term customer loyalty and organizational growth.

DATASET DESCRIPTION:

The "Customer_Churn_Records.csv" dataset consists of detailed information on 10,000 customers from a bank, with no missing values. This dataset, sourced from Kaggle, has been carefully curated for analysis, with certain variables like CustomerID, Surname, and RowNumber removed due to their negligible contribution to the analysis.

Details into the Dataset:

The dataset comprises a total of 18 variables, offering insights into various aspects of customer demographics, financial behavior, and tenure with the bank. These attributes provide a comprehensive view of the customer base, enabling a thorough exploration of factors that may influence churn.

At the heart of our analysis lies the target variable, "Exited," which signifies customer churn. This binary variable classifies customers into two categories: those who have exited the bank's services and those who have not. Understanding and predicting churn is crucial for financial institutions to retain customers and sustain business growth.

By applying data-driven computational methods to this dataset, we aim to uncover hidden patterns, identify significant relationships between variables, and pinpoint influential factors contributing to customer churn. Armed with these insights, we can develop informed strategies and initiatives aimed at enhancing customer retention, thereby bolstering the bank's performance and competitiveness in the market.

SAMPLING STRATEGIES AND POTENTIAL BIAS:

Response Bias: Customers who responded to surveys or provided feedback may have different characteristics compared to those who did not respond, leading to biased results. Class Imbalance Bias: If there is a significant imbalance between the number of churned and non-churned customers in the dataset, predictive models may be biased towards the majority class.

PROBLEMS IDENTIFIED

- Prediction Problem: To predict the churn of bank customers based on several factors by leverage advanced statistical learning techniques to develop predictive models.
- Analyze extensive customer churn data, to formulate a comprehensive understanding of the underlying patterns and drivers of churn behavior.

DATA SEGMENTATION AND ASSESSMENT STANDARDS

The dataset was first identified and checked for missing values. Then categorical varaibles such as Gender and Geography were identified and factorized. The data was then systematically split into training and test sets in the 80:20 ratio following a rigorous approach to ensure the integrity and efficacy of the subsequent machine learning models. Utilizing randomization techniques, the dataset was shuffled before partitioning, thereby preventing biases stemming from the data's original order. This ensured that data points were randomly distributed across the training and test sets, enhancing the generalizability of the models.

The training set served as the basis for Exploratory Data Analysis and model training, where various machine learning algorithms, including classification was employed to learn patterns and relationships from the input features. The test set, comprising unseen data, facilitated the evaluation of model performance by measuring metrics such as accuracy, precision, recall, etc. This evaluation provided insights into how well the trained models generalized to new, unseen data.

Supervised learning algorithms such as decision trees, logistic regression, Lasso Logistic, Ridge Logistic were used to build classification models based on the predictor variables.Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are used to evaluate the performance of classification models. These metrics measure the model's ability to correctly classify the customers into their respective category("Exited", "Not Exited").

Statistical learning strategies and Methods

Data Preparation:

Data Exploration: Initial examination and understanding of the dataset's structure and contents.

Data Cleaning: Addressing inconsistencies, handling missing values, and ensuring data integrity.

Feature Engineering: Deriving new insights by transforming and enhancing existing variables.

Analytical Techniques:

Visualizations: Utilization of graphical representations to uncover patterns and insights.

Correlation Analysis: To Examine relationships between various factors to unveil potential churn influencers.

Chi-square test: To check if there exists a statistically significant association between categorical variables (e.g., geography, card type, gender) and the occurrence of churn.

t-tests: to evaluate whether there exist statistically significant differences in numerical variables (e.g., Tenure) between churn and non-churn groups.

Predictive Modeling with K-Fold Cross-Validation and Bootstrapping

Model Development: Utilizing Logistic Regression to construct predictive models for customer churn based on available attributes.

K-Fold Cross-Validation: Implementing k-fold cross-validation (k=10) to assess the model's performance robustness.

Bootstrapping: Bootstrapping was used to construct confidence intervals for parameters or statistics. By computing the percentile intervals from the bootstrap distribution of the estimator, we obtain a range of plausible values for the parameter with a specified level of confidence. This is particularly used to assess the performance and stability of predictive models. By resampling from the dataset multiple times we fit the model to each bootstrap sample, and evaluate the variability in model performance metrics such as accuracy, precision, recall, and area under the ROC curve.

Evaluation Metrics:

To assess the efficacy of analyses and predictive models, the following metrics are employed:

Accuracy and Precision: Gauging overall correctness and reliability of predictive models.

Recall: Assessing the ratio of correctly predicted churn cases among actual churn instances.

F1-Score: Providing a balanced measure of model performance via harmonic mean of precision and recall.

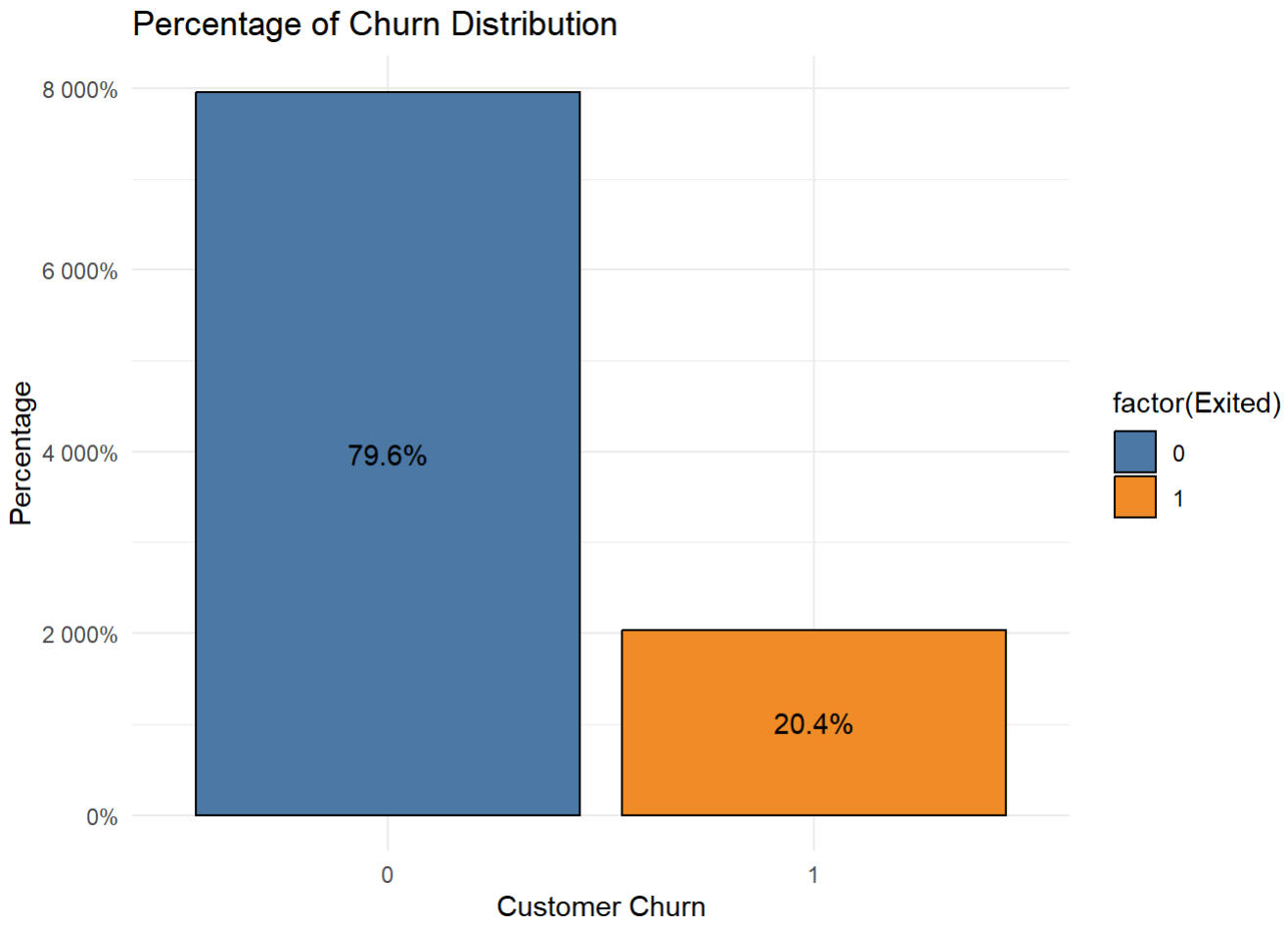
ROC-AUC: Evaluating the model's ability to distinguish between churn and non-churn instances.

Confusion Matrix: To identify the number of correct predictions.

Data Analysis on the Train Data

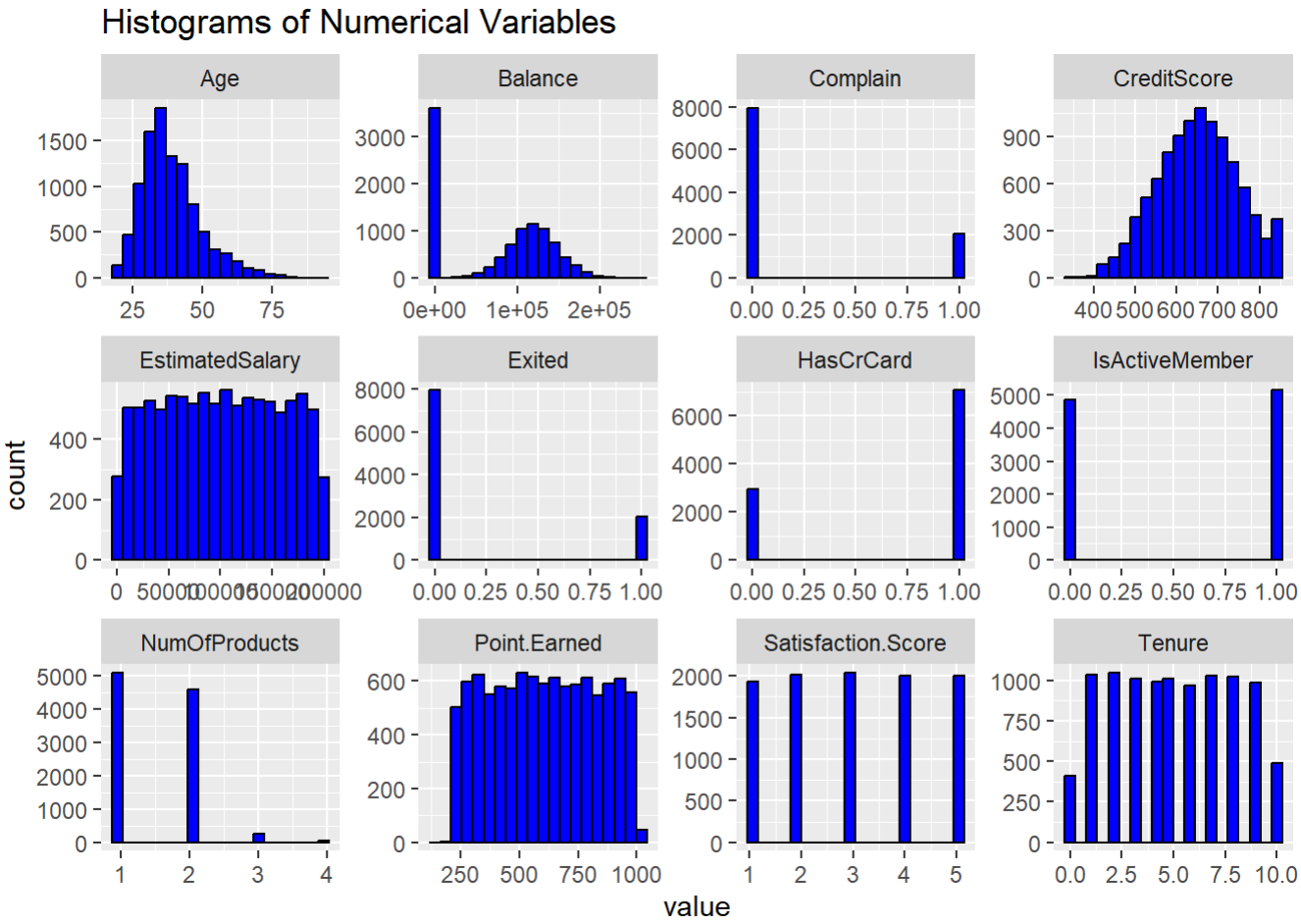
EXPLORATORY DATA ANALYSIS

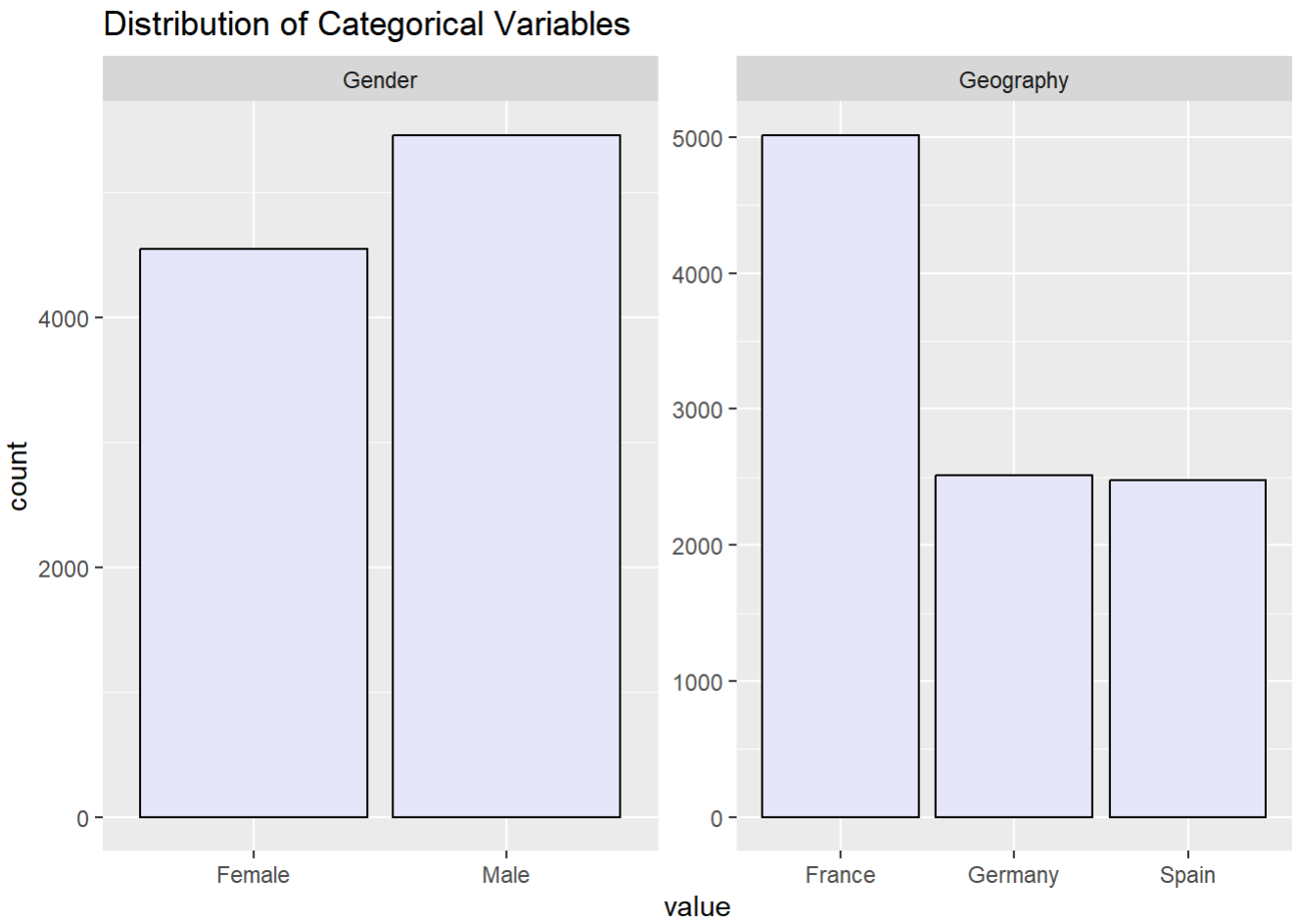
Distribution of Customer Churn in Dataset - “Exited” - Target Variable



In the total dataset, approximately 20.4% of customers experienced churn or exit, while the remaining 79.6% did not. This disparity in percentages indicates an imbalance within the dataset regarding customer churn, with a significantly higher proportion of customers not churning compared to those who did.

Visualizing Numerical and Categorical Variables


Histograms of Numerical Variables
Age
Balance
Complain
CreditScore
EstimatedSalary
Exited
HasCrCard
IsActiveMember
NumOfProducts
Point.Earned
Satisfaction.Score
Tenure
count
value



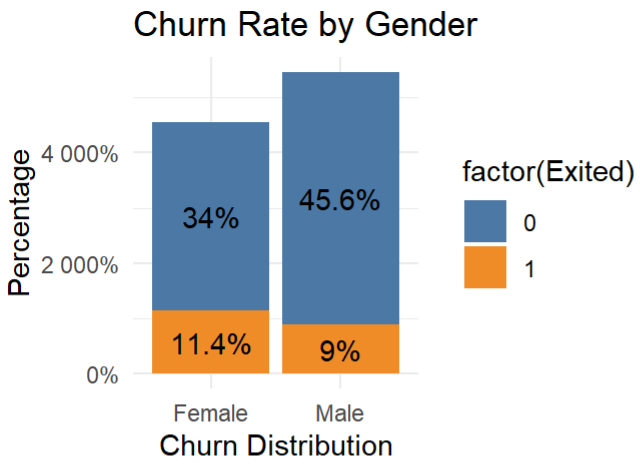
These visualizations provide a comprehensive overview of the key attributes, aiding in identifying trends and potential patterns across various customer demographics and behaviors.

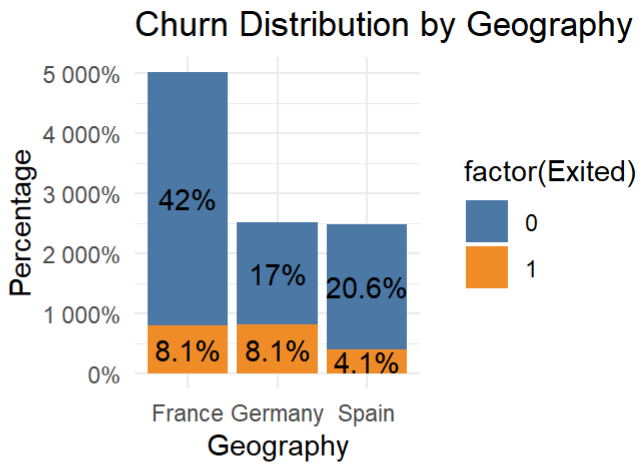
Inference:

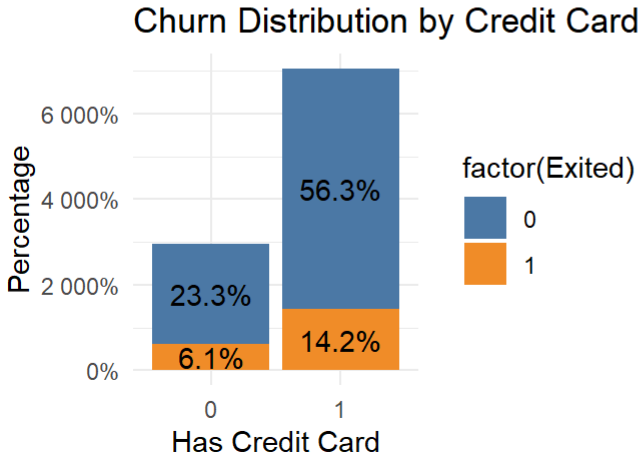
The CreditScore demonstrates a typical distribution, while Age and Tenure exhibit uniform patterns. Balance skews towards the right, indicating a higher number of customers with lower balances. NumOfProducts predominantly involves 1 or 2 products, and HasCrCard denotes ownership of a credit card. IsActiveMember reflects customer engagement levels. Exited distinguishes between churned and retained customers, while Complain identifies instances of complaints. Satisfaction Score ranges from 1 to 5, Points Earned tracks loyalty program points, and Estimated Salary illustrates income distribution. Geographic distribution showcases variations in population across France, Germany, and Spain. Gender distribution outlines the representation of female and male customers within the dataset.

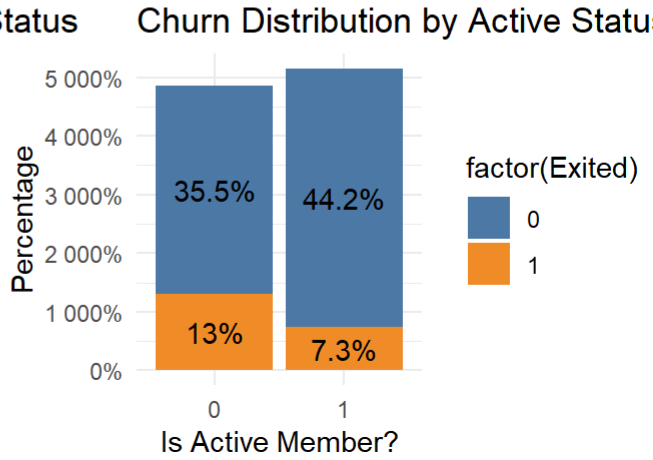
Customer Churn Patterns

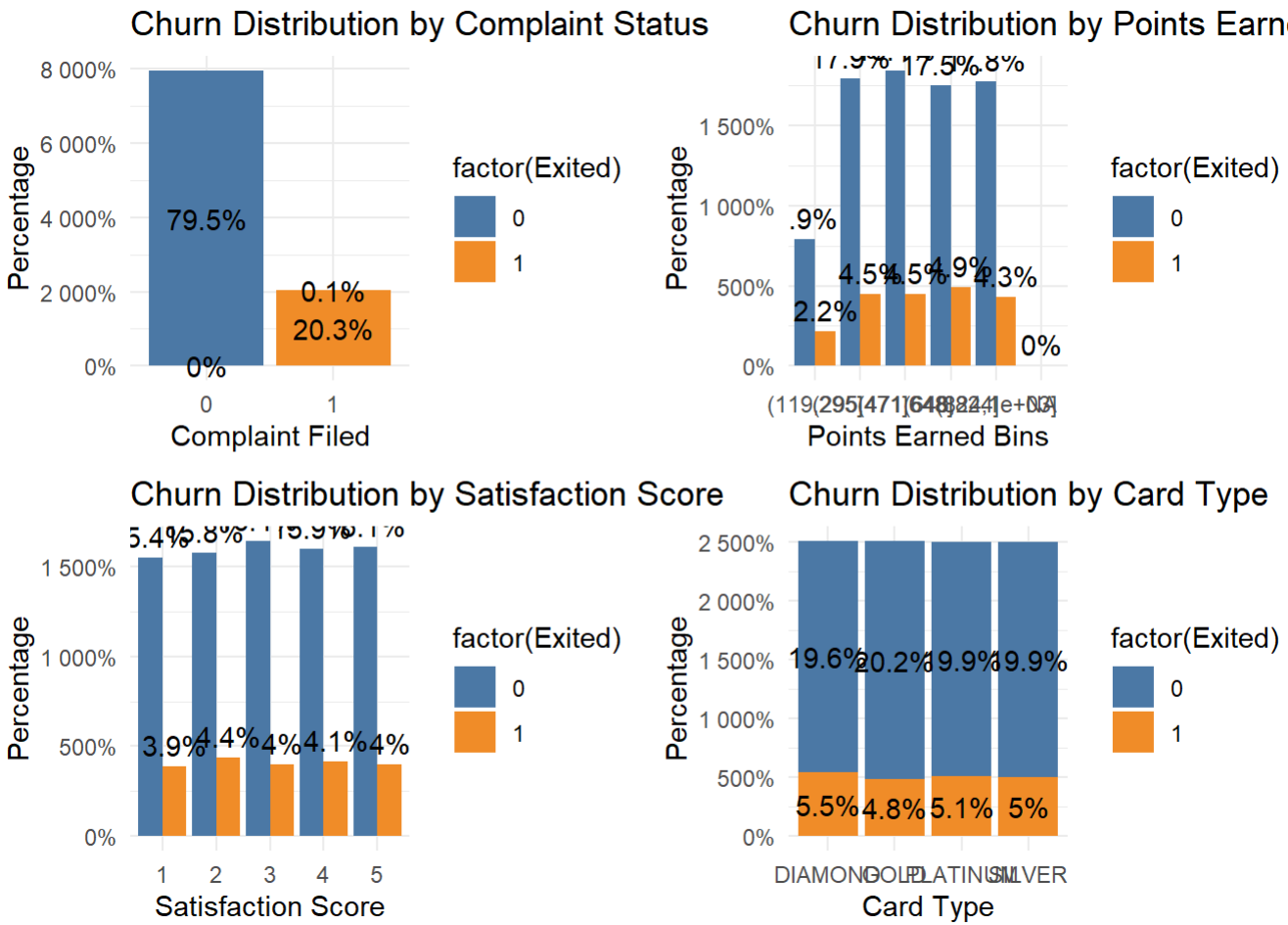
```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```


Churn Rate by Gender
Percentage
4 000%
2 000%
0%
34%
45.6%
11.4%
9%
Female
Male
Churn Distribution
factor(Exited)
0
1


Churn Distribution by Geography
Percentage
5 000%
4 000%
3 000%
2 000%
1 000%
0%
42%
17%
20.6%
8.1%
8.1%
4.1%
France
Germany
Spain
Geography
factor(Exited)
0
1

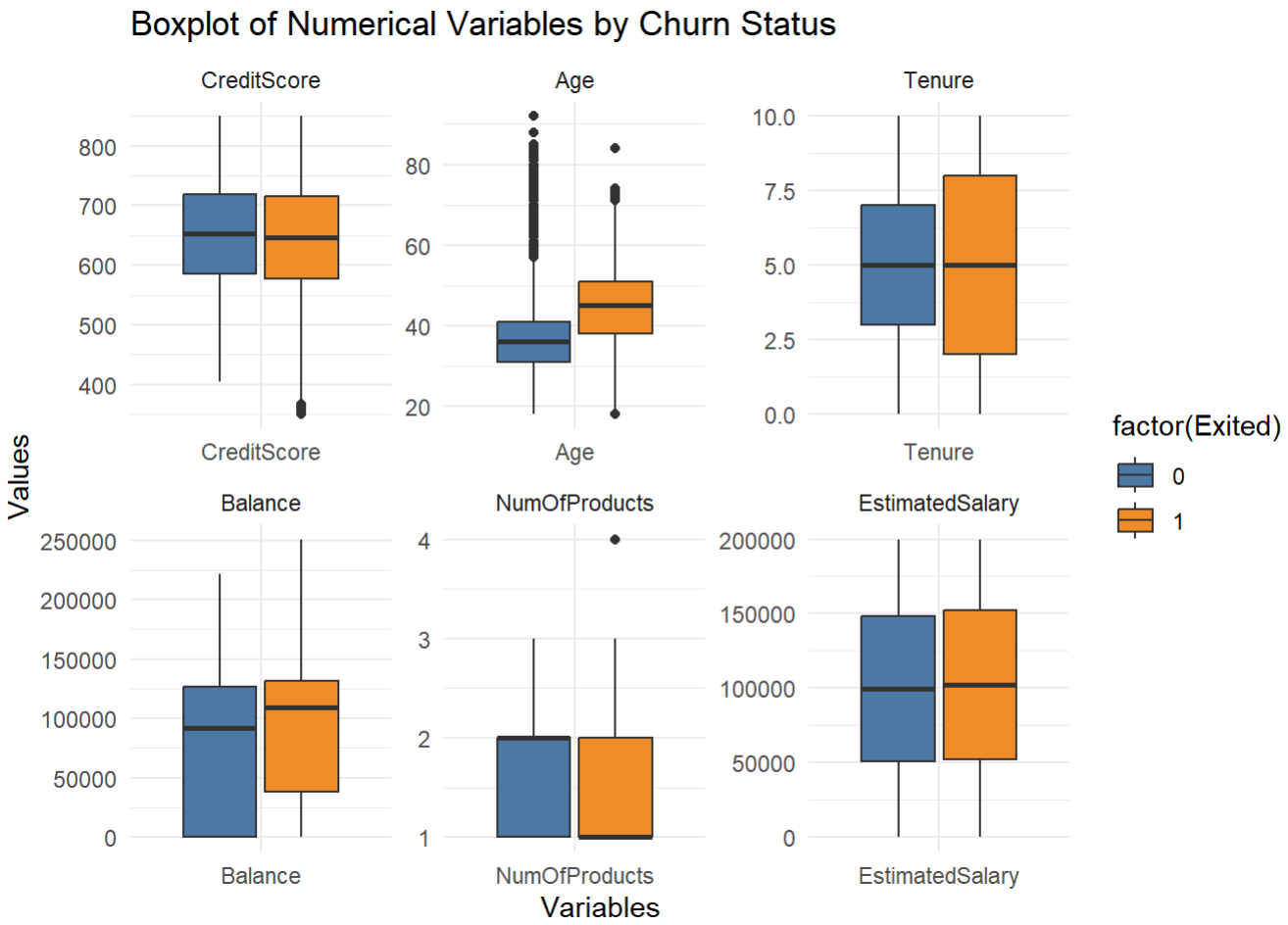

Churn Distribution by Credit Card Status
Percentage
6 000%
4 000%
2 000%
0%
23.3%
56.3%
6.1%
14.2%
0
1
Has Credit Card
factor(Exited)
0
1


Churn Distribution by Active Status
Percentage
5 000%
4 000%
3 000%
2 000%
1 000%
0%
35.5%
44.2%
13%
7.3%
0
1
Is Active Member?
factor(Exited)
0
1



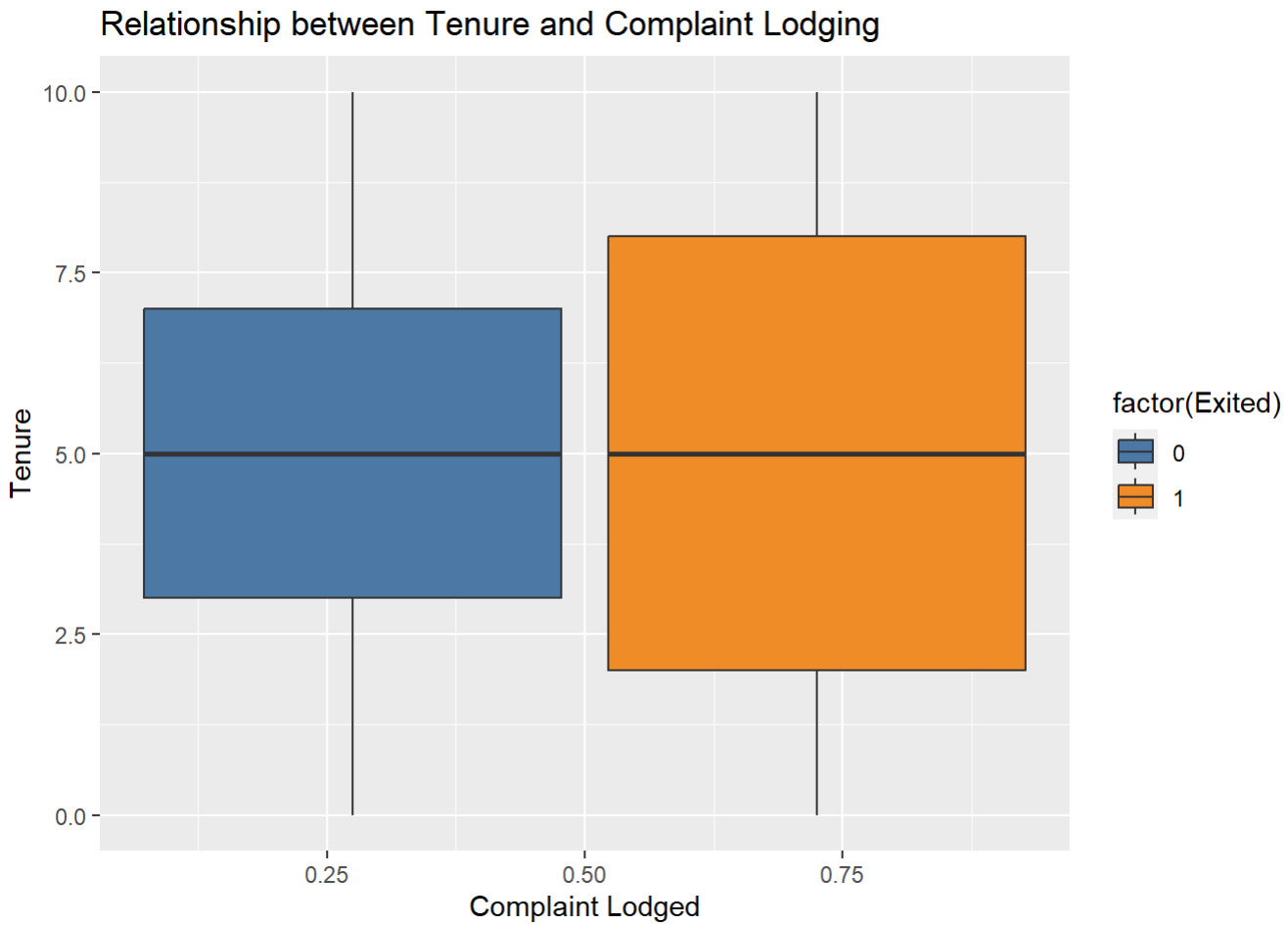
Key Observations Arising from Various Factors Influencing Customer Churn:

- From a geographic perspective, France stands out as the dominant contributor to the customer base, yet regions with fewer customers demonstrate elevated churn rates, hinting at potential service-related issues.
- Gender-wise, females display higher churn rates compared to males, suggesting the presence of gender-specific factors influencing churn behavior. Although credit card ownership is prevalent among churned customers, a deeper investigation is necessary to ascertain its actual impact on churn.
- The presence of inactive members significantly contributes to heightened churn rates, emphasizing the necessity for tailored engagement strategies. Moreover, the occurrence of lodged complaints aligns with increased churn, indicating a correlation between complaints and customer attrition.
- Notably, the impact of points earned and card types on churn varies, underscoring the intricate nature of factors influencing customer retention. Nevertheless, consistently higher satisfaction scores are associated with lower churn rates, highlighting the critical role of customer satisfaction in fostering loyalty and mitigating churn.



Examination of credit scores did not uncover notable disparities between churned and retained customers. Conversely, age emerged as a significant determinant, with older customers displaying a greater propensity to churn compared to younger ones. Moreover, tenure played a pivotal role, suggesting that both short and extended tenures were associated with heightened churn rates, highlighting the importance of customer retention efforts during extreme tenure periods. Interestingly, customers with substantial bank balances demonstrated a propensity to churn, presenting potential risks to the bank's available capital. However, factors such as the number of products held or estimated salary appeared to have minimal impact on the likelihood of churn, indicating their limited influence on customer attrition within the dataset.

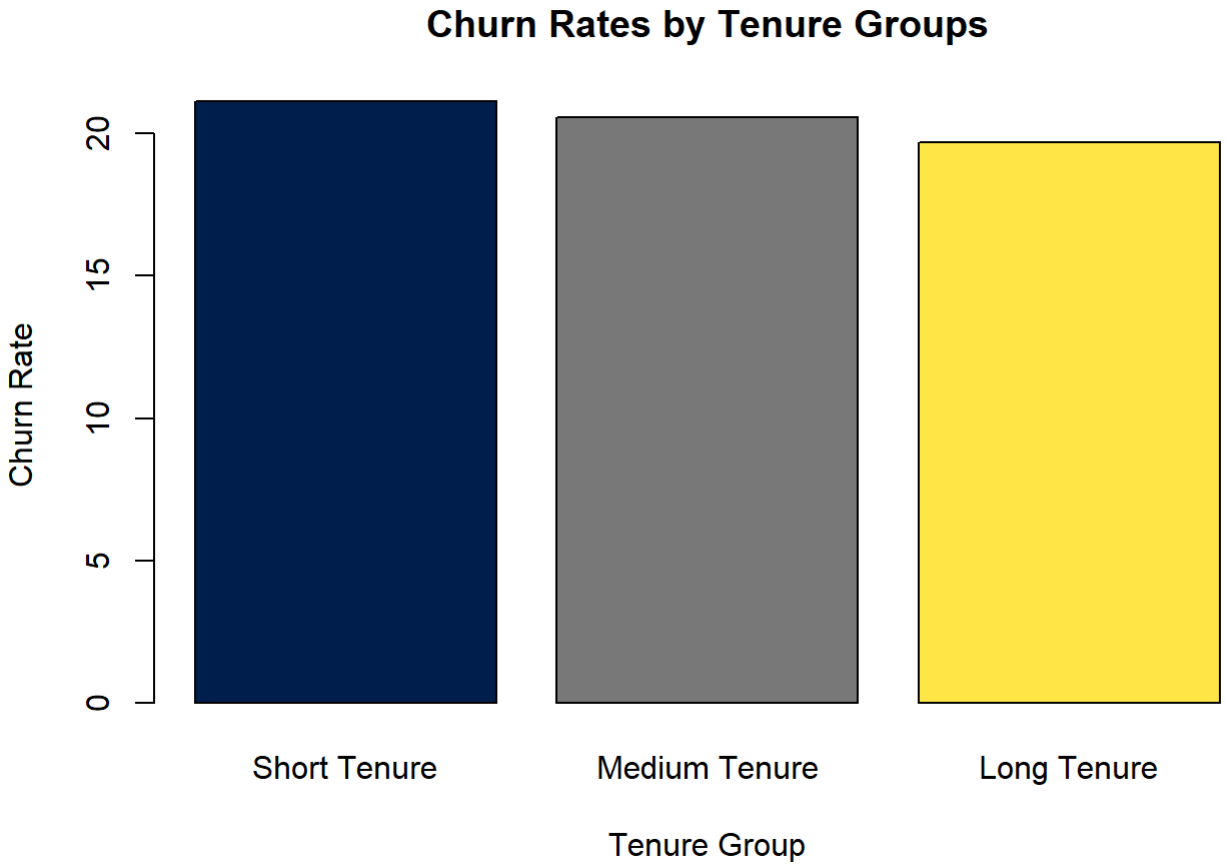
Impact of Complaint Lodging on Customer Tenure in Churn Analysis



Customers who did not file complaints (Complain = 0) demonstrate consistent tenure regardless of churn status. On the other hand, churned customers who lodged complaints (Complain = 1, Exited = 1) show a slightly lower median tenure. Churned customers with complaints exhibit a broader range of tenures compared to non-churned customers who filed complaints.

Churn Analysis by Tenure Group

```
## Loading required package: viridisLite
```



The churn rates exhibit a slight decrease as tenure lengthens, suggesting a reduction in churn among customers with longer-term relationships. Conversely, shorter tenures are associated with a slightly higher churn rate. Pearson's Chi-squared test (p-value = 0.4077) indicates no substantial association between churn and tenure categories.

Analysis of Tenure in Relation to Customer Complaints and Churn Status

The t-tests conducted to compare tenure duration among customers who filed complaints and those who churned indicate that there might not be a significant disparity in tenure length between these two groups.

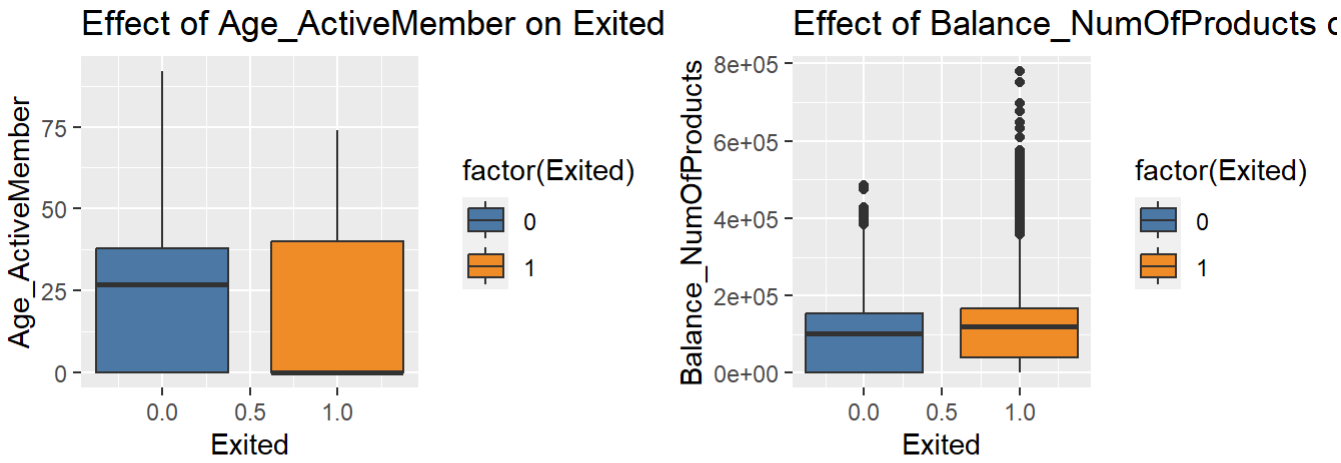
Exploration of Interaction Effects and Feature Engineering

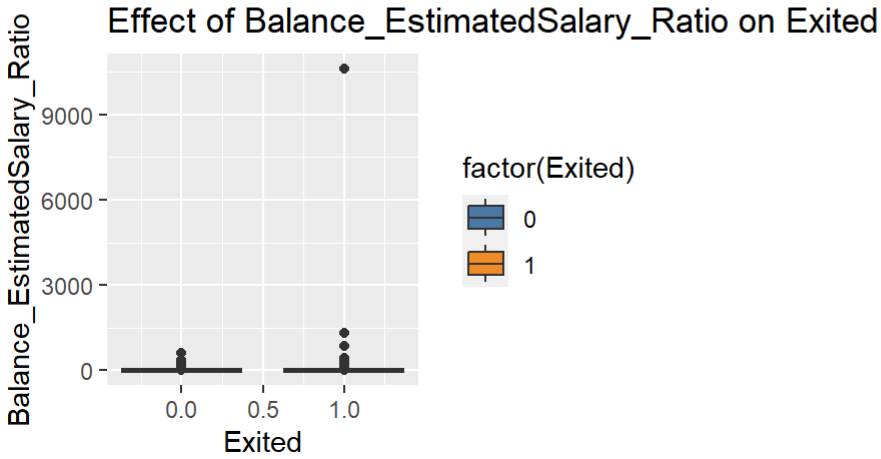
To explore their impact on churn behavior, the following interaction terms were generated:

1. Age_ActiveMember
2. Balance_NumOfProducts

3. Balance_EstimatedSalary_Ratio

These interaction terms were created to assess their potential influence on the likelihood of churning.

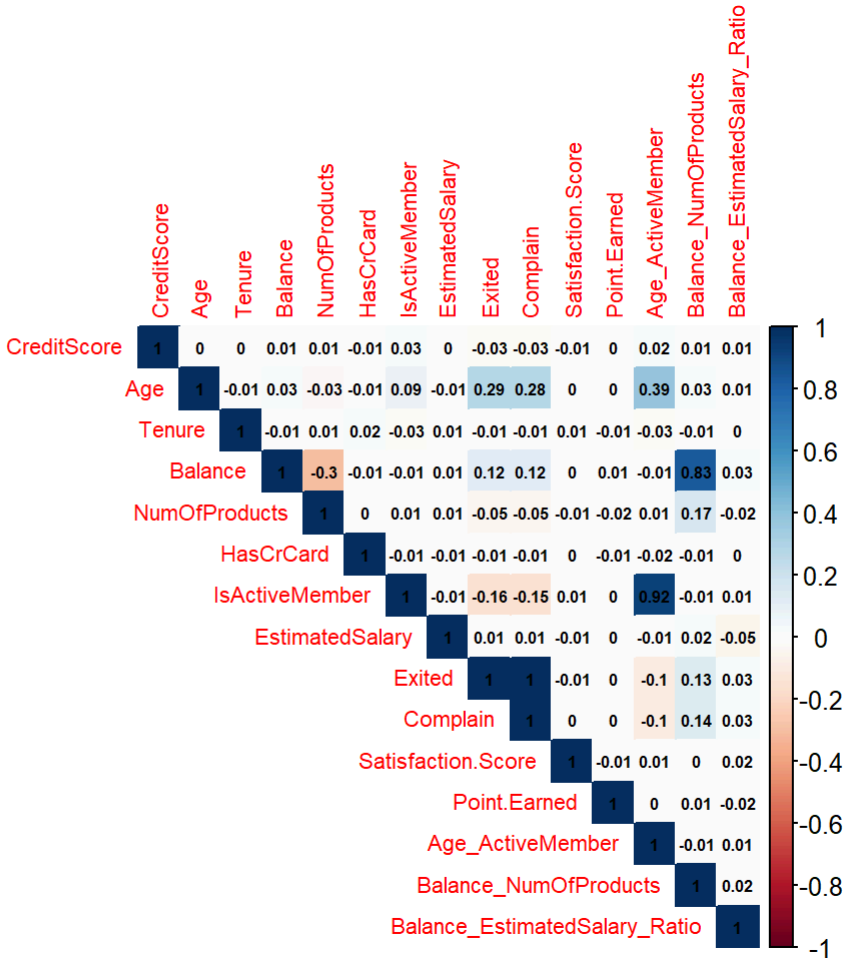




Inference on Interaction term's Effects for Churn

The interaction variables created, notably Age_ActiveMember and Balance_NumOfProducts, show promise in differentiating between churned and retained customers. These variables may impact churn prediction, suggesting that customers with higher values in these derived features might demonstrate a slightly increased likelihood of churning.

Correlation Analysis for Churn Prediction



Age demonstrates a moderate positive correlation with the probability of churning. Variables such as IsActiveMember and NumOfProducts exhibit moderate negative correlations, suggesting that active members and customers with a higher number of products are less likely to churn. A more pronounced correlation is noted between the Complain variable and Exited, warranting further investigation to determine its significance in relation to churn. Engineered features like Balance_NumOfProducts and Balance_EstimatedSalary_Ratio demonstrate slight positive correlations with Exited.

Inference on Chi-Square Tests for Churn Analysis

```
##
##  Pearson's Chi-squared test
##
## data:  table(bank$Card.Type, bank$Exited)
## X-squared = 5.0532, df = 3, p-value = 0.1679
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(bank$Geography, bank$Exited)
## X-squared = 300.63, df = 2, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(bank$Complain, bank$Exited)
## X-squared = 9907.9, df = 1, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(bank$NumOfProducts, bank$Exited)
## X-squared = 1501.5, df = 3, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(bank$Point.Earned, bank$Exited)
## X-squared = 797.23, df = 784, p-value = 0.3636
```

There is a notable association between ‘Geography’ and ‘Exited’, indicating that Geography significantly affects customer churn. ‘Complain’ and ‘Exited’ demonstrate a highly significant relationship, suggesting that customers who filed complaints tend to churn significantly more than those who did not. The ‘Number of Products’ held by a customer exhibits a significant relationship with ‘Exited’, implying an impact on customer churn based on the number of products. However, no significant relationship is detected between ‘Points Earned’ and ‘Exited’, suggesting a lack of substantial influence on churn.

STATISTICAL LEARNING APPROACHES AND STRATEGIES FOR FEATURE ENGINEERING

Logistic Regression:

Logistic regression is a statistical method used for predicting the probability of a binary outcome, such as customer churn (‘Exited’ or ‘Not Exited’), based on one or more predictor variables. It’s particularly suitable for customer churn prediction because it provides interpretable coefficients that indicate the strength and direction of the relationship between each predictor variable and the likelihood of churn. Logistic regression is efficient, easy to understand, and implement, making it a popular choice for churn analysis. However, it assumes a linear relationship between predictor variables and the log odds of the outcome, which limits its ability to capture complex interactions or nonlinear relationships.

Decision Trees:

Decision trees are a versatile machine learning method used for classification and regression tasks. In the context of customer churn prediction, decision trees are advantageous because they can capture complex interactions between predictor variables and identify nonlinear relationships, which are prevalent in churn analysis. Decision trees are easy to interpret and visualize, allowing analysts to gain insights into the factors driving churn. They are also robust to outliers and can handle both numerical and categorical variables without preprocessing. However, decision trees tend to overfit the training data, especially when the tree is deep, which may lead to poor generalization performance on unseen data.

Lasso Regression:

Lasso regression, also known as L1 regularization, is a linear regression technique that adds a penalty term to the regression coefficients, forcing some coefficients to shrink to zero. In the context of customer churn prediction, Lasso regression can be beneficial for feature selection, as it automatically selects the most important features while discarding irrelevant ones. This can help improve model interpretability and generalization performance by reducing overfitting. Lasso regression is particularly useful when dealing with high-dimensional datasets with many predictor variables, as it can effectively handle multicollinearity and prevent model overfitting.

Ridge Regression:

Ridge regression, also known as L2 regularization, is another linear regression technique that adds a penalty term to the regression coefficients. Unlike Lasso regression, Ridge regression does not force coefficients to shrink to zero but rather shrinks them towards zero. In the context of customer churn prediction, Ridge regression can help improve model robustness and generalization performance by reducing the variance of the coefficient estimates. It is particularly useful when dealing with multicollinear predictor variables, as it can mitigate the issue of high variance in the coefficient estimates. Ridge regression is suitable for situations where feature selection is not a primary concern, but rather the goal is to improve the overall predictive accuracy of the model.

Bootstrapping:

Bootstrapping is a resampling technique used to assess the variability of a statistical estimate by repeatedly sampling from the observed data with replacement. In the context of customer churn prediction, bootstrapping was utilized to estimate the variability of model performance metrics such as accuracy, precision, recall, and F1-score. By resampling the dataset multiple times and evaluating the model on each resampled dataset, bootstrapping provides more reliable estimates of the model’s performance and helps assess the stability of the results. This technique improves the robustness of the analysis and provides a more accurate assessment of the model’s predictive capability.

Feature Importance:

Feature importance analysis was conducted to identify the most influential predictors of customer churn. For logistic regression and Lasso regression, the coefficients of the predictor variables were examined to determine their impact on the likelihood of churn. Feature importance analysis through random forest helped identify the most influential predictors of customer churn. This information can guide strategic decision-making by highlighting the key drivers of churn and informing targeted intervention efforts to reduce churn risk effectively. Additionally, feature importance analysis enhances the interpretability of the random forest model by providing insights into the underlying factors driving customer attrition.

K-Fold Cross Validation:

Using 5-fold cross-validation (CV) alongside random forest for feature importance analysis enhanced the reliability and robustness of the results.

1.Model Training with 5-Fold Cross-Validation: In 5-fold cross-validation, the dataset was divided into five equally sized folds. The random forest model is trained five times, each time using four folds as training data and one fold as validation data. This process is repeated five times, ensuring that each fold serves as validation data exactly once. By training the model on multiple subsets of the data and validating it on independent subsets, 5-fold CV provides a more accurate estimate of the model's performance and generalization ability.

2. Feature Importance Calculation: During each iteration of 5-fold CV, feature importance is computed based on metrics such as Gini impurity or Mean Decrease in Accuracy (MDA). For each fold, the random forest model is trained using the training data, and feature importance is assessed based on the predictive performance of the model on the validation data. This process is repeated for all five folds, resulting in five sets of feature importance scores.

3. Aggregation of Feature Importance Scores: Once feature importance scores are computed for each fold, they are aggregated across all folds to obtain a comprehensive estimate of feature importance. This aggregation process typically involves averaging the feature importance scores across all folds. By combining the results from multiple iterations of 5-fold CV, a more robust and stable estimate of feature importance was obtained, reducing the influence of random variation and ensuring the reliability of the results. Using 5-fold cross-validation alongside random forest for feature importance analysis ensured that the identified important features are consistent across different subsets of the data, enhancing the credibility and generalizability of the results. Additionally, it provides valuable insights into the stability of feature importance rankings and helps identify potential sources of variability in the model's predictive performance.

Predictive Analysis and Results

LOGISTIC REGRESSION:

```
##
## Call:
## glm(formula = Exited ~ . + Tenure:Complain, family = "binomial",
##     data = bank)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.067e+01  3.146e+00  -3.392 0.000694 ***
## CreditScore                    9.792e-04  2.836e-03   0.345 0.729882
## GeographyGermany              -7.812e-01  7.137e-01  -1.095 0.273714
## GeographySpain                 3.702e-01  7.936e-01   0.466 0.640872
## GenderMale                    -3.213e-01  5.446e-01  -0.590 0.555202
## Age                            1.447e-01  3.420e-02   4.231 2.32e-05 ***
## Tenure                        -1.124e-01  1.783e-01  -0.630 0.528645
## Balance                        4.295e-06  1.150e-05   0.373 0.708803
## NumOfProducts                 -2.417e-01  6.562e-01  -0.368 0.712637
## HasCrCard                     -5.630e-02  6.046e-01  -0.093 0.925818
## IsActiveMember                 1.378e+00  1.880e+00   0.733 0.463565
## EstimatedSalary                6.277e-07  4.767e-06   0.132 0.895238
## Complain                       1.396e+01  1.335e+00  10.456  < 2e-16 ***
## Satisfaction.Score            -2.193e-01  1.995e-01  -1.099 0.271785
## Card.TypeGOLD                 -5.598e-01  7.865e-01  -0.712 0.476586
## Card.TypePLATINUM             -6.872e-01  7.741e-01  -0.888 0.374682
## Card.TypeSILVER                2.982e-01  8.138e-01   0.366 0.714040
## Point.Earned                  -2.511e-03  1.336e-03  -1.879 0.060293 .
## Age_ActiveMember              -7.696e-02  4.456e-02  -1.727 0.084127 .
## Balance_NumOfProducts         -8.894e-07  5.959e-06  -0.149 0.881353
## Balance_EstimatedSalary_Ratio  5.213e-03  9.544e-03   0.546 0.584913
## Tenure:Complain                6.447e-02  2.109e-01   0.306 0.759781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10112.51  on 9999  degrees of freedom
## Residual deviance:   162.38  on 9978  degrees of freedom
## AIC: 206.38
```

```
##
## Number of Fisher Scoring iterations: 11
```

Significant Predictors

The logistic regression model reveals ‘Age’ and ‘Complain’ as statistically significant predictors ($p < 0.05$), indicating their substantial impact on the likelihood of a customer exiting the bank. Moreover, neither the tenure of a customer nor the interaction between tenure and lodging complaints seem to significantly predict or explain customer churn.

LOGISTIC REGRESSION - Standard Method

```
## [1] "Evaluation Metrics of Logistic Regression Model are: "
```

```
## [1] "Accuracy: 0.998"
```

```
## [1] "Precision: 0.997495303694427"
```

```
## [1] "Recall: 1"
```

```
## [1] "F1-Score: 0.998746081504702"
```

```
## [1] "ROC-AUC: 0.999215340177224"
```

The logistic regression model appears to perform exceptionally well based on the evaluation metrics provided:

- The accuracy of 99.8% indicates that the model correctly predicts the outcome for the vast majority of cases.
- With a precision of 99.75%, the model demonstrates a high proportion of correct positive predictions among all positive predictions made.
- A recall score of 100% suggests that the model identifies all actual positive cases correctly, indicating no false negatives.
- The F1-score, which combines precision and recall, is 99.87%, indicating a robust balance between precision and recall.
- The ROC-AUC score of 99.92% indicates a high discriminatory power of the model in distinguishing between positive and negative cases.

Overall, these metrics suggest that the logistic regression model performs excellently in predicting customer churn, with high accuracy, precision, recall, F1-score, and ROC-AUC score.

LOGISTIC LASSO AND RIDGE REGRESSION prediction metrics:

```
## [1] "Evaluation Metrics of Logistic Regression Model with Lasso:"
```

```
## [1] "Accuracy: 0.998"
```

```
## [1] "Precision: 0.997495303694427"
```

```
## [1] "Recall: 1"
```

```
## [1] "F1-Score: 0.998746081504702"
```

```
## [1] "ROC-AUC: 0.999401793996498"
```

```
## [1] "Evaluation Metrics of Logistic Regression Model with Ridge:"
```

```
## [1] "Accuracy: 0.998"
```

```
## [1] "Precision: 0.997495303694427"
```

```
## [1] "Recall: 1"
```

```
## [1] "F1-Score: 0.998746081504702"
```

```
## [1] "ROC-AUC: 0.999249523377424"
```

From the evaluation metrics provided for the logistic regression models with Lasso and Ridge regularization, we can infer the following:

High Accuracy: Both models achieve a high accuracy of 99.8%, indicating that they are effective at correctly classifying instances of both classes (churned and retained customers) in the dataset.

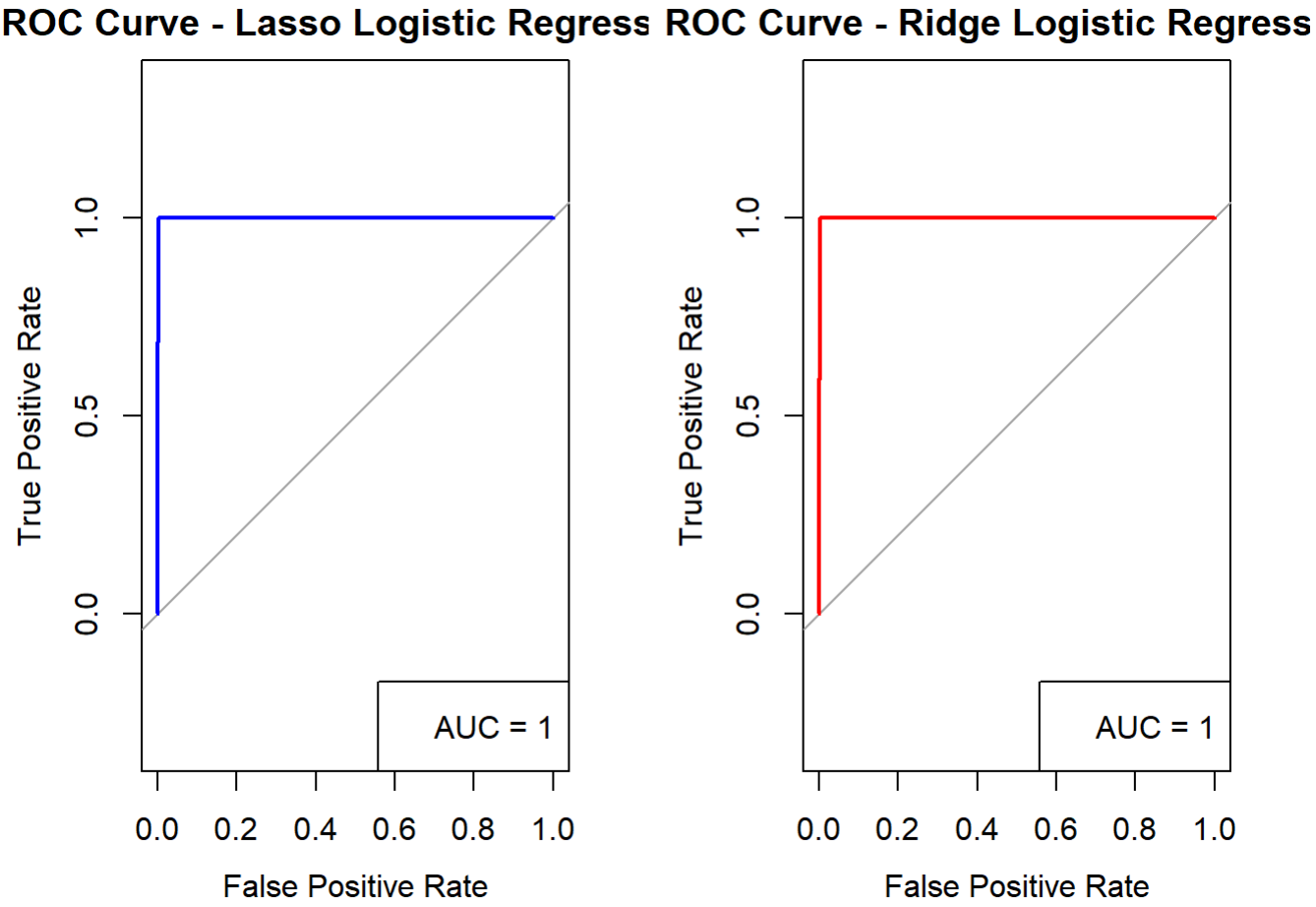
High Precision and Recall: The precision and recall values are also high, with precision and recall both at 99.75%. This suggests that the models have a high level of precision in predicting positive cases (churned customers) and are able to identify a high proportion of true positive cases among all predicted positive cases.

High F1-Score: The F1-score, which is the harmonic mean of precision and recall, is also high at 99.87%. This indicates a balance between precision and recall, demonstrating that the models are performing well in terms of both false positives and false negatives.

High ROC-AUC Score: The ROC-AUC score, which measures the area under the receiver operating characteristic (ROC) curve, is very high at 99.94% for Lasso and 99.92% for Ridge. This suggests that the models have excellent discriminative ability and are able to distinguish between positive and negative cases with high accuracy.

Overall, based on these evaluation metrics, we can conclude that both the logistic regression models with Lasso and Ridge regularization performed exceptionally well in predicting customer churn. They demonstrate high accuracy, precision, recall, F1-score, and ROC-AUC score, indicating their effectiveness in identifying potential churners and retaining valuable customers.

LOGISTIC LASSO AND RIDGE REGRESSION prediction plots:



LOGISTIC LASSO AND RIDGE REGRESSION with BOOTSTRAPPING:

The results show that the Accuracy, Recall, Precision and F1 score are same with and without bootstrapping.

LOGISTIC REGRESSION model predictions on train_data using Bootstrapping:

The logistic regression model trained with bootstrapping on the training data yielded the following evaluation metrics:

Average Accuracy: 99.87% with a confidence interval of [99.84%, 99.88%] Average Precision: 99.93% with a confidence interval of [99.91%, 99.94%] Average Recall: 99.90% with a confidence interval of [99.87%, 99.91%] Average F1 Score: 99.92% with a confidence interval of [99.90%, 99.92%] Average AUC (Area Under the Curve): 99.93% with a confidence interval of [99.87%, 99.97%]

These metrics provide insights into the model's performance and confidence in its predictions, indicating high accuracy, precision, recall, F1 score, and AUC on the training data is pretty much same as that on test data indicating a perfect fit.

Lasso Logistic Regression model with the Selected Features

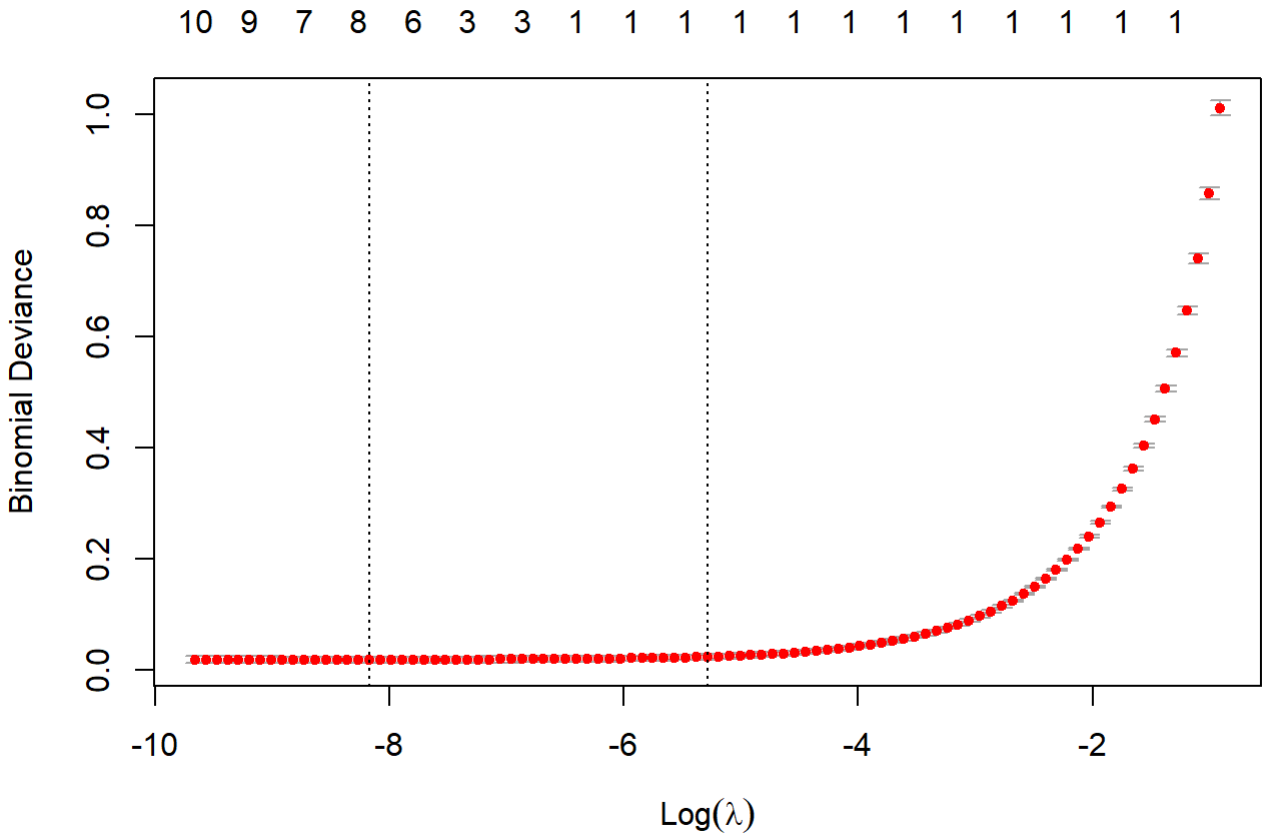
Selected Features found using Lasso Logistic Regression are

```
#Lasso Feature Selection
# Fit a logistic regression model with Lasso (L1 regularization)
lasso_model <- cv.glmnet(train_matrix, train_data$Exited, family = "binomial", alpha = 1)

# Extract selected features
selected_features <- which(coef(lasso_model, s = "lambda.min")[-1, ] != 0)


# Print selected feature names
selected_feature_names <- colnames(train_matrix)[selected_features]
print(selected_feature_names)
```

```
## [1] "Age"              "NumOfProducts"    "IsActiveMember"   "Complain"
## [5] "Point.Earned"     "Age_ActiveMember"
```



OPTIMAL LAMBDA INFERENCE

The optimal lambda value suggests that a moderate level of regularization is applied to the model. This value was determined through cross-validation to strike a balance between bias and variance, resulting in a model that is neither too complex (overfit) nor too simple (underfit), effectively capturing the underlying patterns in the data while avoiding over-reliance on any particular set of features.

Logistic Regression model with the Selected Features

```
##
## Call:
## glm(formula = Exited ~ ., family = "binomial", data = train_data[,
##     selected_features])
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.893960   0.119363  -32.62   <2e-16 ***
## Age             0.073995   0.002833   26.12   <2e-16 ***
## IsActiveMember -1.095566   0.063344  -17.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8102.5  on 7999  degrees of freedom
## Residual deviance: 7160.0  on 7997  degrees of freedom
## AIC: 7166
##
## Number of Fisher Scoring iterations: 5
```

```
## Confusion Matrix:
##  1550 349 47 54
```

```
## Accuracy: 0.802
```

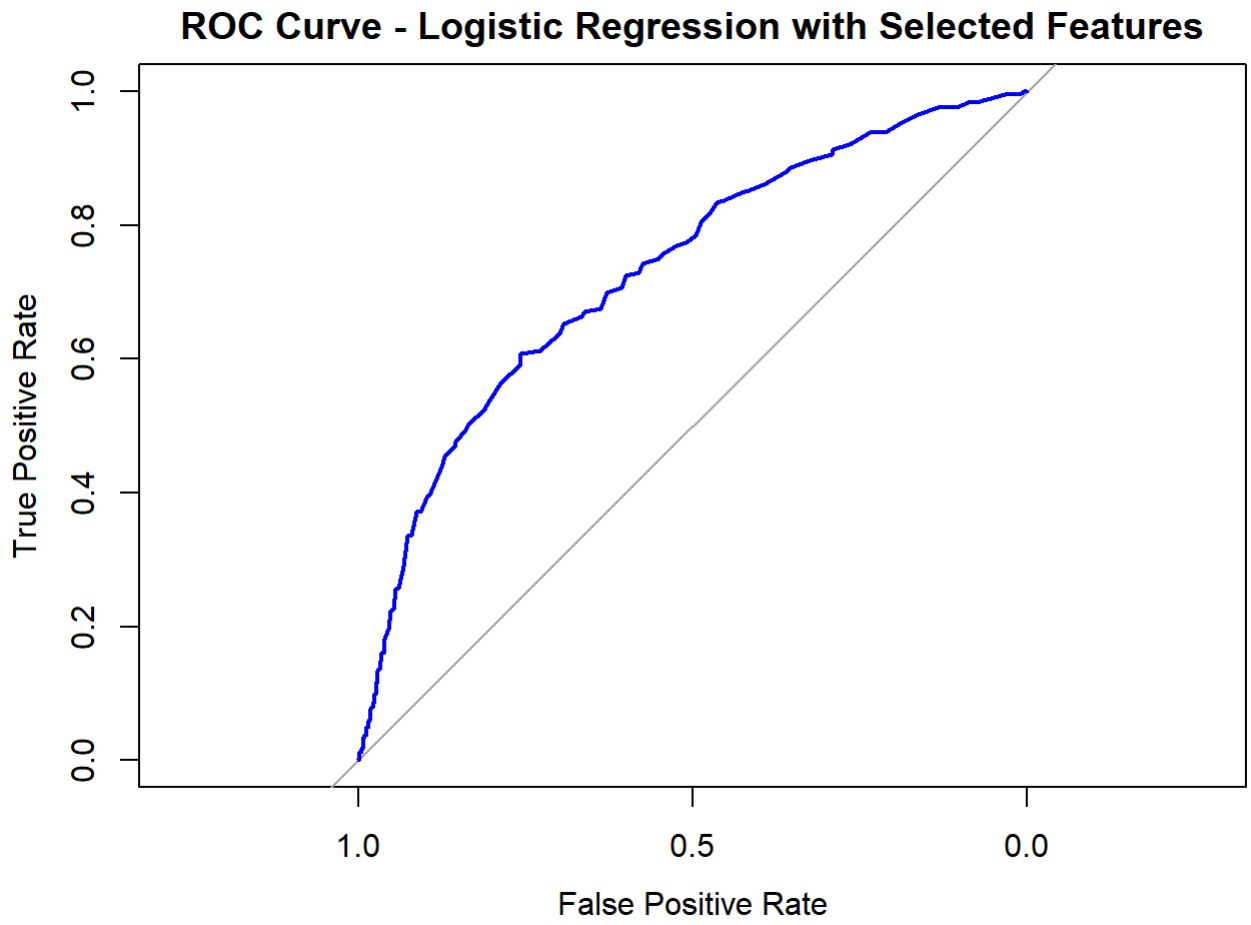
```
## Precision: 0.5346535
```

```
## Recall: 0.133995
```

```
## F1 Score: 0.2142857
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Inference:

Age has a significant positive impact on the likelihood of exiting the bank, implying that older customers are more likely to churn. IsActiveMember has a significant negative impact on the likelihood of exiting the bank, suggesting that active members are less likely to churn.

The model achieved an accuracy of 80.2%, indicating that it correctly classified approximately 80.2% of the observations. Precision, recall, and F1 score provide additional insights into the model's performance, with precision indicating the proportion of correctly predicted positive cases(i.e customers who have churned (Exited = 1)), recall indicating the proportion of actual positive cases that were correctly classified, and F1 score representing the balance between precision and recall.

Performance Metrics of Cross Validated Logistic Regression Model

```
## [1] "Evaluation Metrics of Cross-Validated Logistic Regression Model are: "
```

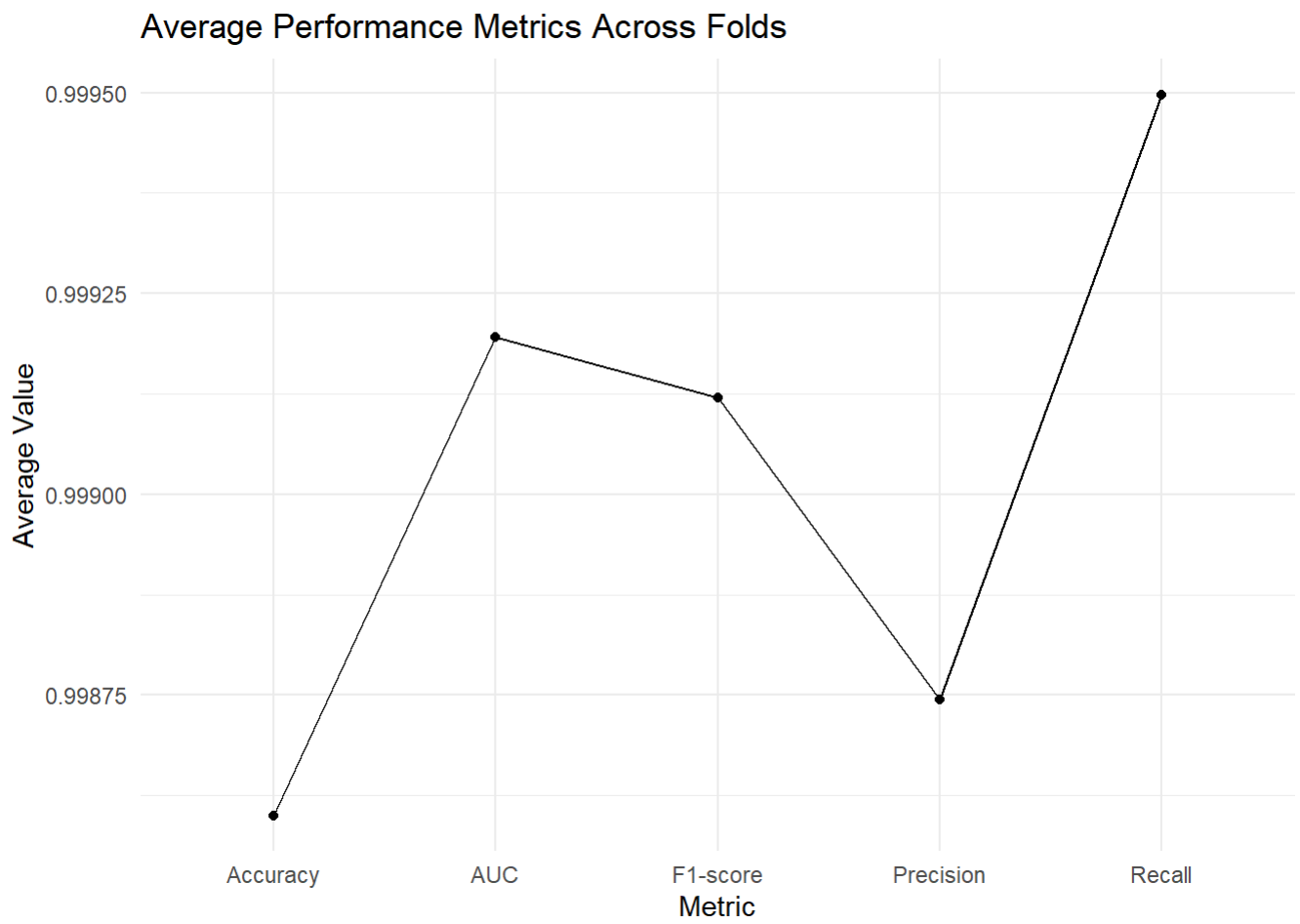
```
## Average AUC: 0.9991961
```

```
## Average Precision: 0.9987439
```

```
## Average Recall: 0.9994976
```

```
## Average F1-score: 0.9991202
```

```
## Average Accuracy: 0.9986003
```



Cross-Validated Logistic Regression Model Evaluation

The cross-validated logistic regression model shows consistent high performance across multiple evaluation metrics, average AUC (Area Under the Curve) values for each fold are consistently high, indicating excellent model performance in distinguishing between churn and non-churn instances. Precision, Recall, F1-score, and Accuracy metrics are exceptionally high across folds, indicating a robust model performance in predicting both churn and non-churn cases.

Comparison: Cross-Validated Logistic Regression Model Vs. logistic regression model

Comparing the cross-validated model's metrics with the earlier logistic regression model, both models exhibit exceptional performance. The cross-validated model demonstrates slightly improved average metrics compared to the single logistic regression model, showcasing its robustness and reliability across multiple folds. This consistency strengthens confidence in the model's predictive capability and highlights its stability when applied to different subsets of the data.

DECISION TREE model predictions on test_data:

```r
# Decision Tree Model
# Install and load the rpart package
library(rpart)

# Decision Tree Model
dt_model <- rpart(Exited ~ ., data = train_data, method = "class")

# Make predictions on the test set
dt_preds <- predict(dt_model, newdata = test_data, type = "class")

# Confusion matrix
dt_conf_matrix <- confusionMatrix(factor(test_data$Exited), factor(dt_preds))

# Traditional metrics
dt_accuracy <- dt_conf_matrix$overall["Accuracy"]
dt_precision <- dt_conf_matrix$byClass["Precision"]
dt_recall <- dt_conf_matrix$byClass["Recall"]
dt_f1_score <- dt_conf_matrix$byClass["F1"]

# Print or use the metrics as needed
print("Evaluation Metrics of Decision Tree Model:")
```

```
## [1] "Evaluation Metrics of Decision Tree Model:"
```

```r
print(paste("Accuracy:", dt_accuracy))
```

```
## [1] "Accuracy: 0.998"
```

```r
print(paste("Precision:", dt_precision))
```

```
## [1] "Precision: 0.997495303694427"
```

```r
print(paste("Recall:", dt_recall))
```

```
## [1] "Recall: 1"
```

```
print(paste("F1-Score:", dt_f1_score))
```

```
## [1] "F1-Score: 0.998746081504702"
```

The Decision Tree model demonstrates excellent performance with high accuracy (99.8%), precision (99.75%), recall (100%), and F1-Score (99.87%). These metrics indicate that the model performs well in correctly classifying both positive and negative cases.

DECISION TREE model predictions on test_data with Bootstrapping:

Evaluation Metrics of Decision Tree Model with bootstrapping are

Average Accuracy: 0.998

Confidence Interval (Accuracy): [0.998 , 0.998]

Average Precision: 1

Confidence Interval (Precision): [1 , 1]

Average Recall: 0.997495303694427

Confidence Interval (Recall): [0.997495303694427 , 0.997495303694427]

Average F1 Score: 0.998746081504702

Confidence Interval (F1 Score): [0.998746081504702 , 0.998746081504702]

Average AUC: 0.998747651847214

Confidence Interval (AUC): [0.998747651847214 , 0.998747651847214]”

Inference:

The Decision Tree model with bootstrapping shows similar performance to the standard Decision Tree model. Bootstrapped model’s confidence intervals for accuracy, precision, recall, F1 Score, and AUC are very narrow, indicating high confidence in the model’s performance estimates.

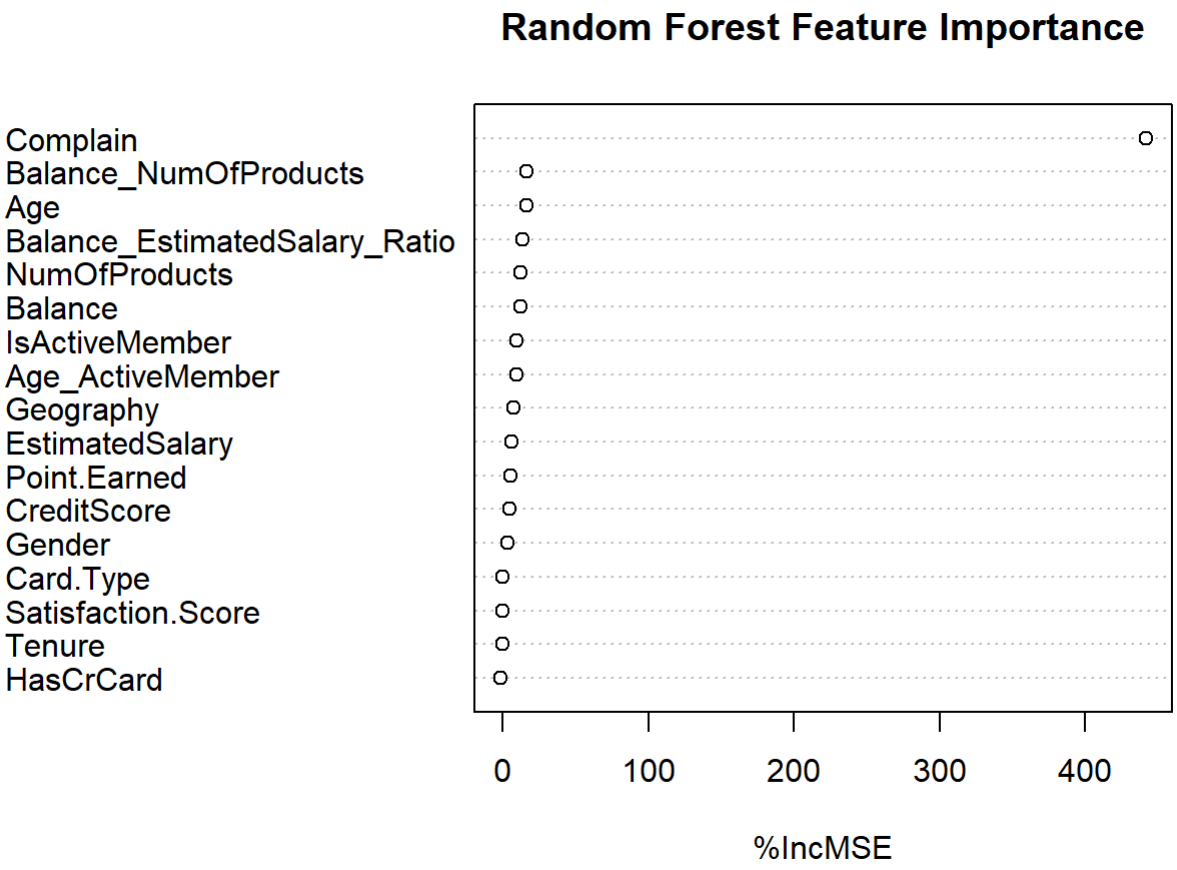
Random Forest Feature Importance Analysis for Churn Prediction

```r
# Load required library
library(randomForest)


# Create Random Forest model
rf_model <- randomForest(Exited ~ ., data = analysis(fold), importance = TRUE)

# Get feature importance scores
importance_scores <- importance(rf_model)
#importance_scores

# Plot feature importance
varImpPlot(rf_model, type = 1, main = "Random Forest Feature Importance")
```


Random Forest Feature Importance
Complain
Balance_NumOfProducts
Age
Balance_EstimatedSalary_Ratio
NumOfProducts
Balance
IsActiveMember
Age_ActiveMember
Geography
EstimatedSalary
Point.Earned
CreditScore
Gender
Card.Type
Satisfaction.Score
Tenure
HasCrCard
0
100
200
300
400
%IncMSE

```
varImpPlot
```

```
## function (x, sort = TRUE, n.var = min(30, nrow(x$importance)),
##     type = NULL, class = NULL, scale = TRUE, main = deparse(substitute(x)),
##     ...)
## {
##     if (!inherits(x, "randomForest"))
##         stop("This function only works for objects of class `randomForest'")
##     imp <- importance(x, class = class, scale = scale, type = type,
##         ...)
##     if (ncol(imp) > 2)
##         imp <- imp[, -(1:(ncol(imp) - 2))]
##     nmeas <- ncol(imp)
##     if (nmeas > 1) {
##         op <- par(mfrow = c(1, 2), mar = c(4, 5, 4, 1), mgp = c(2,
##             0.8, 0), oma = c(0, 0, 2, 0), no.readonly = TRUE)
##         on.exit(par(op))
##     }
##     for (i in 1:nmeas) {
##         ord <- if (sort)
##             rev(order(imp[, i], decreasing = TRUE)[1:n.var])
##         else 1:n.var
##         xmin <- if (colnames(imp)[i] %in% c("IncNodePurity",
##             "MeanDecreaseGini"))
##             0
##         else min(imp[ord, i])
##         dotchart(imp[ord, i], xlab = colnames(imp)[i], ylab = "",
##             main = if (nmeas == 1)
##                 main
##             else NULL, xlim = c(xmin, max(imp[, i])), ...)
##     }
##     if (nmeas > 1)
##         mtext(outer = TRUE, side = 3, text = main, cex = 1.2)
##     invisible(imp)
## }
## <bytecode: 0x00000215a9889388>
## <environment: namespace:randomForest>
```

The Random Forest model’s feature importance analysis reveals key predictors influencing customer churn-

Age emerges as a significant factor, indicating its strong influence on churn likelihood. Additionally, Balance, NumOfProducts, and the ‘Complain’ variable surprisingly hold substantial importance in predicting churn. Interaction terms like Balance_NumOfProducts, Age_ActiveMember, and Balance_EstimatedSalary_Ratio also contribute significantly to churn prediction, showcasing the complexity of factors impacting customer attrition within the dataset. Moreover, IsActiveMember appears as a crucial factor influencing churn rates, collectively highlighting the pivotal predictors considered within the model’s evaluation of churn.

RESULT DISCUSSION:

For Logistic Regression we observe that the test results do not vary much despite including Lasso and Ridge models while feature selection does impact the accuracy indicating that reducing features does eliminate the dense relationships between the varaibles.

The results are constant with the resampling techniques empoloyed - Bootstrapping and 5-fold Cross Validation.

Feature importance revealed ‘Age’ as the important feature in impacting the model’s performance.

Precision-Recall Trade-Off for Decision Tree:

Both standard and bootstrapped model achieved high precision and recall scores, with precision being slightly lower in the standard model compared to the bootstrapped model. This indicates that the standard model may have slightly more false positives than the bootstrapped model, while the bootstrapped model may have slightly more false negatives. The consistency of performance across both models, as indicated by the narrow confidence intervals in the bootstrapped model, suggests that the Decision Tree model is robust and reliable. This consistency is essential for deploying the model in real-world applications where consistent performance is desired.

Conclusion

Through detailed predictive analysis, we dissected product engagement, card usage, geographic distribution, customer complaints, and tenure, identifying crucial components linked to customer attrition. Utilizing these insights, we constructed predictive models employing demographic and banking variables, employing sophisticated techniques like logistic regression and random forest.

Analyzing customer churn in banking reveals essential insights crucial for bolstering customer loyalty. By comprehending customer behavior across demographics and regions, tailored retention strategies can be developed to address the specific needs of diverse customer segments. Prioritizing prompt complaint resolution and delivering personalized services based on identified satisfaction patterns can significantly reduce churn rates and foster enduring customer loyalty.

SCOPE AND GENERALIZABILITY OF THE PREDICTIVE ANALYSIS

While this study offers nuanced insights beneficial for crafting effective customer retention strategies, it is imperative to validate its findings across industries facing similar challenges. Further validation across diverse datasets or industry domains can enhance the study's applicability and reliability. Strengthening feature engineering methodologies and refining the models through advanced techniques could significantly improve their predictive accuracy and relevance across various business scenarios, thereby enhancing the study's credibility and practical utility.

LIMITATIONS AND POSSIBILITIES FOR IMPROVEMENT

While the analysis provided valuable insights into customer churn within the banking sector, it is essential to acknowledge certain limitations. The analysis primarily focused on demographic and banking variables, potentially overlooking broader external factors that could influence churn. Additionally, limitations within the dataset and the utilization of conventional modeling techniques such as logistic regression and random forest may restrict the depth of insights and predictive accuracy. To augment the analysis, integrating more diverse datasets, exploring advanced modeling approaches beyond conventional techniques, and considering external factors such as economic conditions could offer a more comprehensive understanding of customer churn behavior. These enhancements have the potential to enrich the depth and robustness of the analysis, facilitating a more holistic perspective on the factors driving customer attrition.