Understanding Customer Churn in Banking : A Data-Driven Exploration

MATH 40024/50024: Computational Statistics

December 10, 2023

Introduction

Customer churn, the departure of clients from a bank's services, presents a significant challenge in the financial sector. This study aims to discern the underlying factors contributing to customer attrition by employing analytical methods on the "Customer_Churn_Records.csv" dataset. The objective is to leverage data analytics to comprehend why customers discontinue their relationship with the bank and potentially predict future churn instances.

Dataset Overview

The dataset "Customer_Churn_Records.csv," comprises various variables pertaining to 10,000 customers of a bank with no missing values. It is taken from the website "Kaggle".The attributes such as CustomerID, Surname, and RowNumber were removed out of 18 variables as they did not contribute significantly to the analysis. The remaining attributes provide insights into customer demographics, financial details, and tenure.

Churn Status: The target variable, 'Exited,' represents customer churn, providing a binary classification for analysis.

By applying data-driven computational methods to this dataset, I aim to uncover patterns, relationships, and influential factors contributing to customer churn, enabling informed decision-making and strategies for customer retention.

Research Objectives:

1. Determining Churn Drivers

This analysis likely aims to explore and analyze factors contributing to customer churn, encompassing diverse aspects such as demographics, financial behavior, product engagement, satisfaction levels, and engagement with services. The dataset could be utilized to identify patterns, correlations, and predictive indicators related to customer churn, ultimately aiding in the development of strategies to enhance customer retention and reduce churn rates within the specific context of the service or product offered.

2. Predictive Modeling

The goal is to construct predictive model utilizing available demographic and banking variables. This model aims to forecast potential churn cases, empowering proactive intervention strategies.

Why Data-Driven computational Approach?

The data-driven, computational approach is highly advantageous for analyzing the "Customer_Churn_Records.csv" dataset due to its diverse variables and the objective of understanding customer attrition. Computational methods excel in uncovering complex relationships among demographics, financial behaviors, satisfaction scores, and churn indicators present in the dataset. Leveraging these techniques allows the development of predictive models that forecast churn based on attributes like age, geography, and product usage.

Furthermore, computational analysis enables the identification of critical factors influencing churn, such as customer satisfaction and tenure, essential for tailoring effective retention strategies. With its ability to efficiently handle large datasets and provide quantifiable insights, this approach ensures the creation of robust models and iterative improvements, supporting informed decision-making to mitigate churn and enhance customer retention efforts.

Computational Methods

The analytical process involves the following steps:

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:

Exploratory Data Analysis (EDA):

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

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.

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.

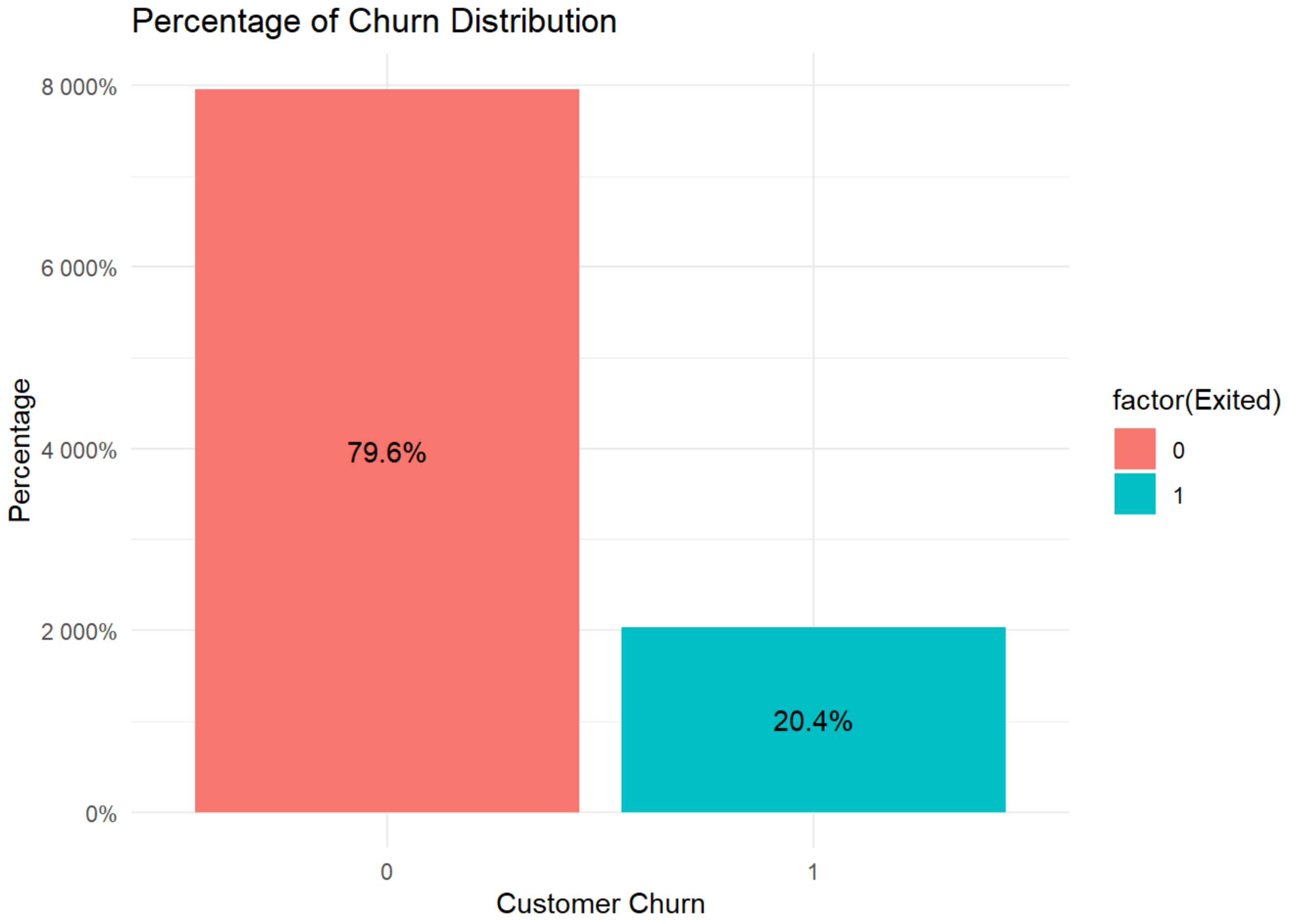
ROC-AUC: Evaluating the model's ability to distinguish between churn and non-churn instances.

Confusion Matrix.

Data Analysis and Results

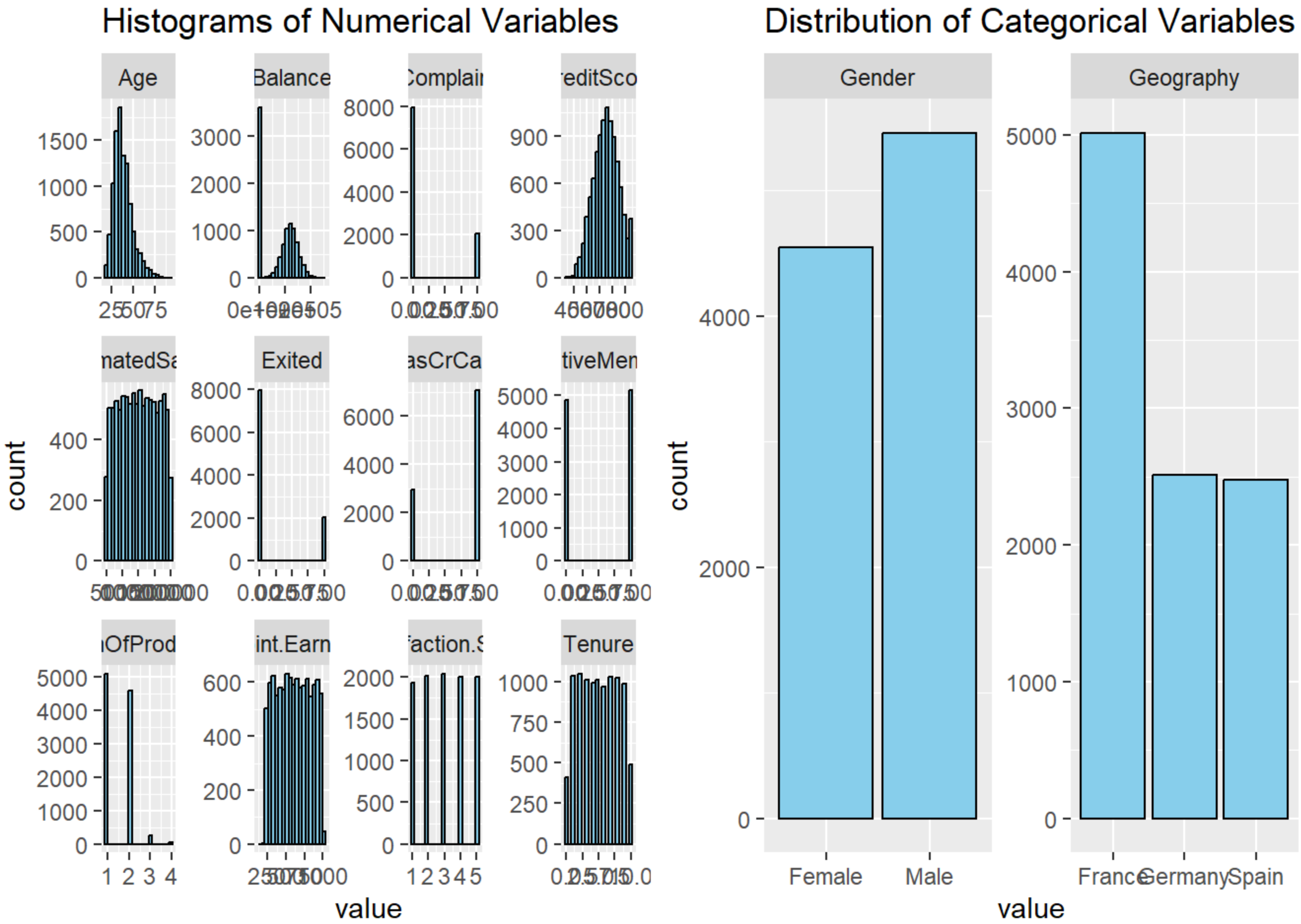
Exploratory Data Analysis: Understanding Variables

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



In the total dataset, approximately 20.4% of customers experienced churn, 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

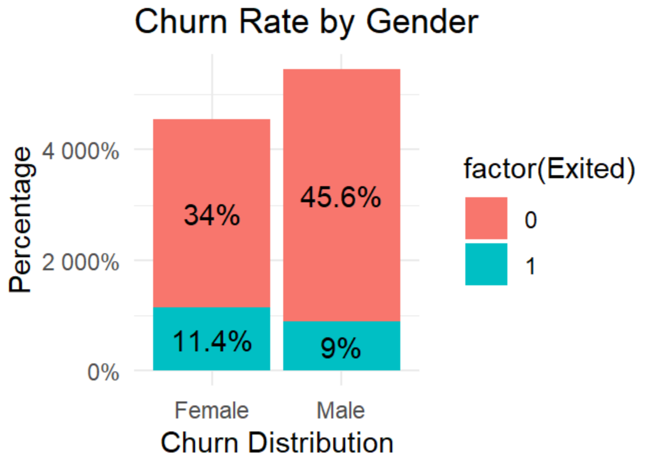


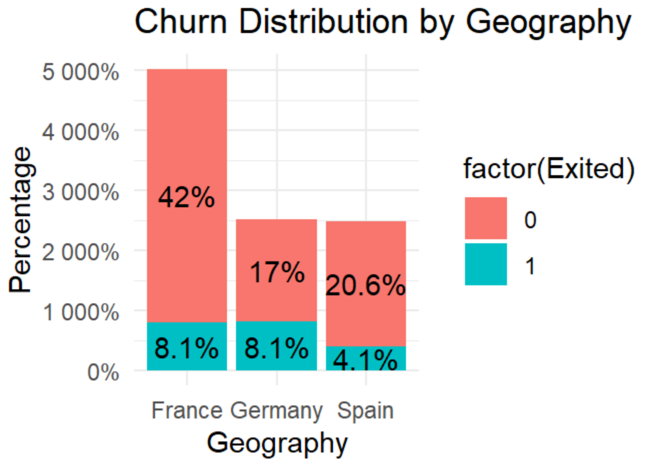
These visualizations provide a comprehensive overview of the key attributes, aiding in identifying trends and potential patterns across various customer demographics and behaviors.

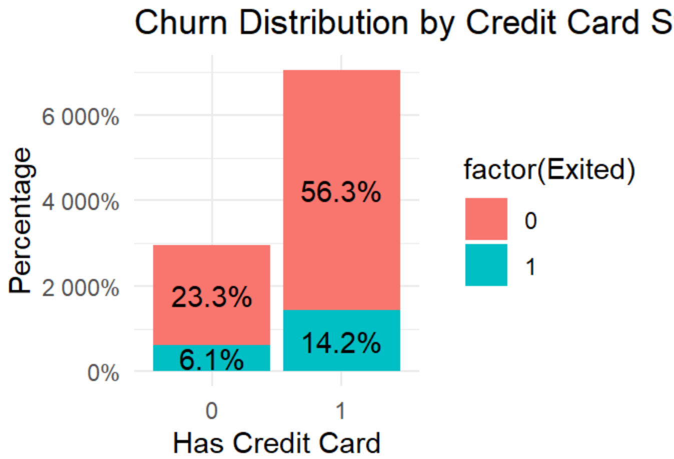
Inference

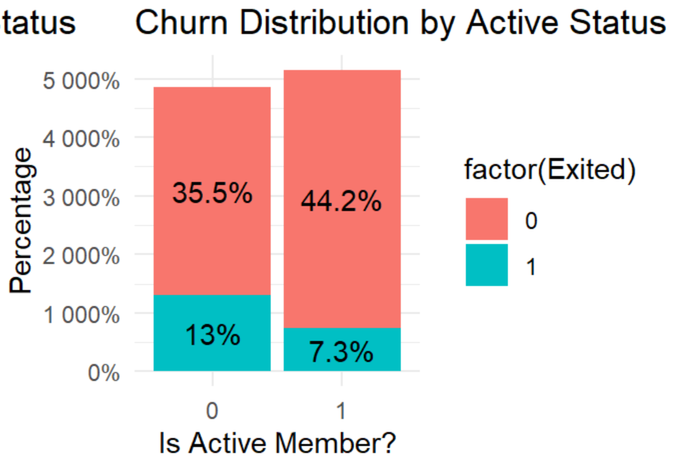
CreditScore displays a somewhat normal spread, while Age and Tenure show uniform distributions. Balance skews right, indicating more customers with lower balances. NumOfProducts mostly involves 1 or 2 products, and HasCrCard represents credit card ownership. IsActiveMember indicates customer activity. Exited shows churned vs. retained customers, and Complain highlights complaint instances. Satisfaction Score ranges from 1 to 5, Points Earned tracks loyalty points, and Estimated Salary shows income distribution. Customer distributions across Geography, indicating varying populations across France, Germany, and Spain. Gender distribution highlights the representation of Female and Male customers within the dataset.

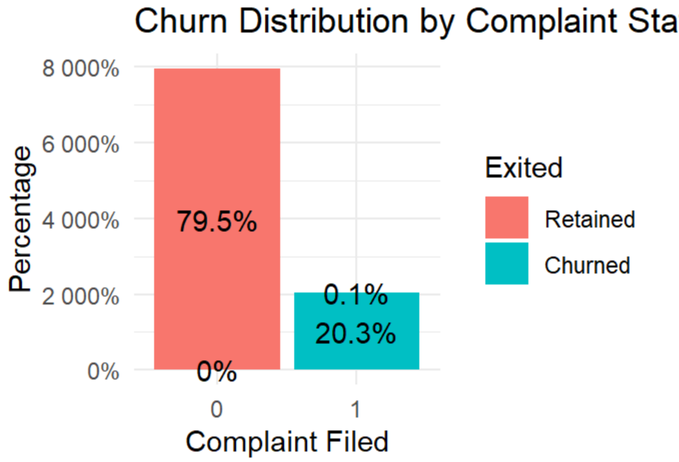
Customer Churn Patterns

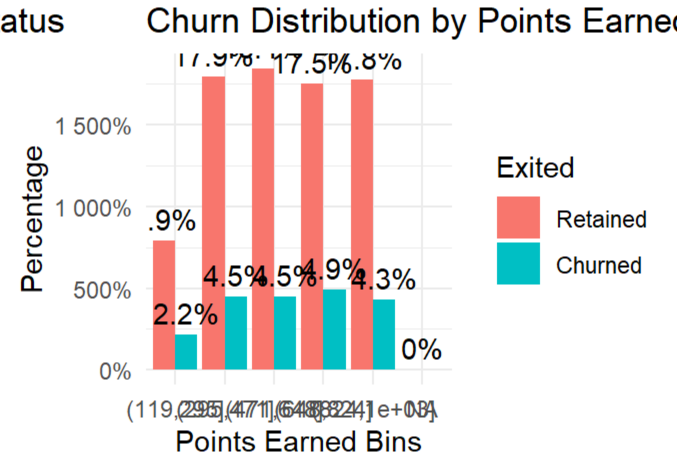

Churn Rate by Gender

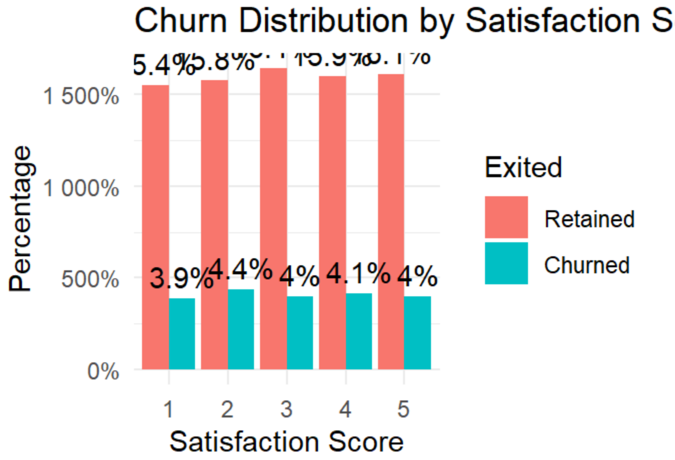

Churn Distribution by Geography

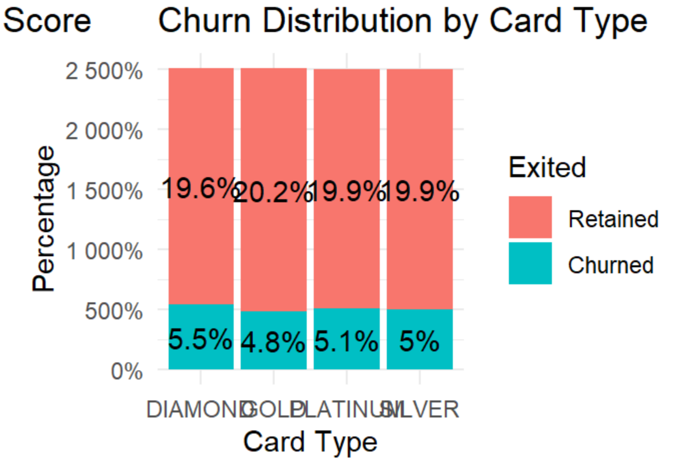

Churn Distribution by Credit Card Status


Churn Distribution by Active Status


Churn Distribution by Complaint Status

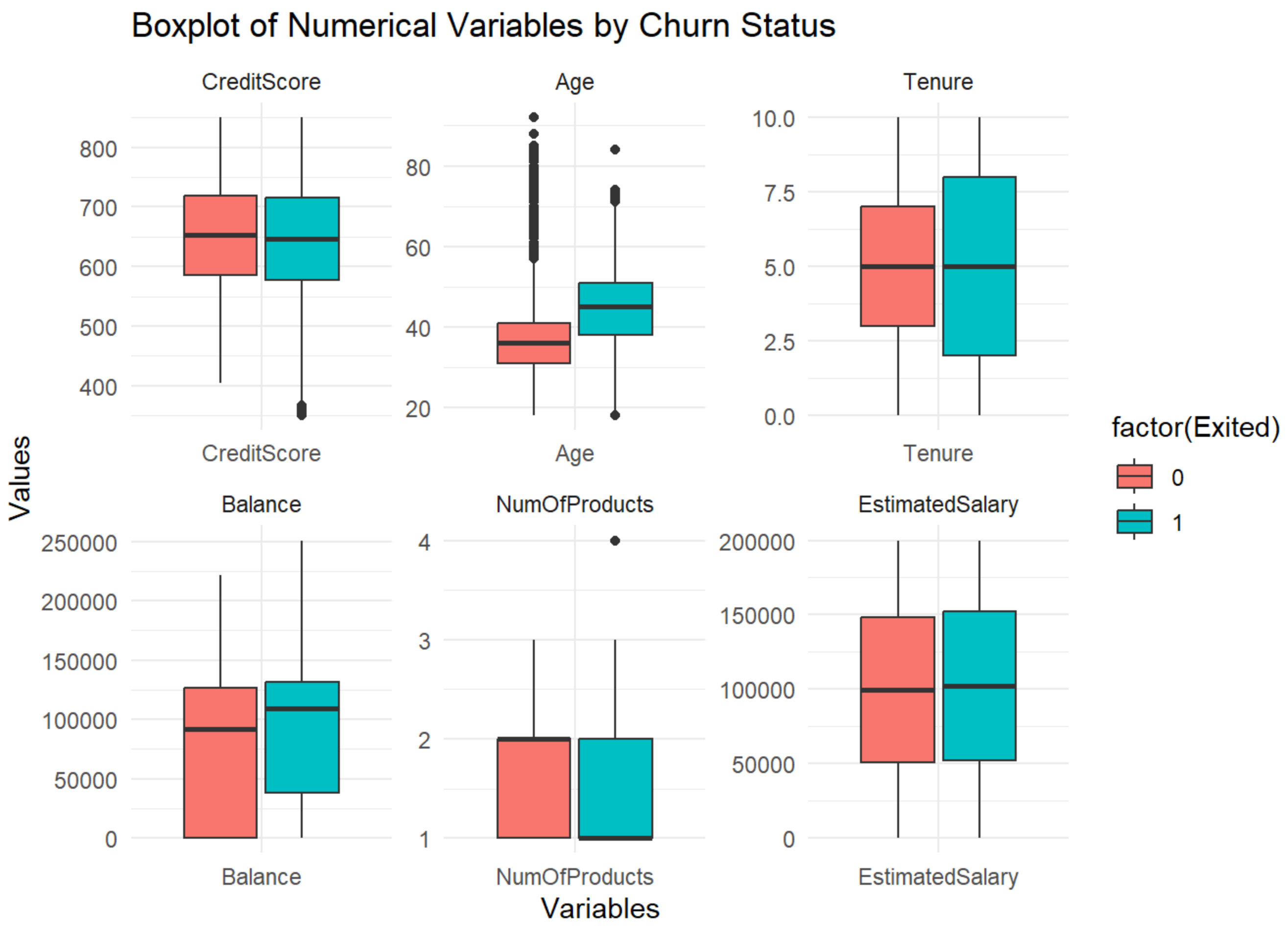

Churn Distribution by Points Earne


Churn Distribution by Satisfaction Score


Churn Distribution by Card Type

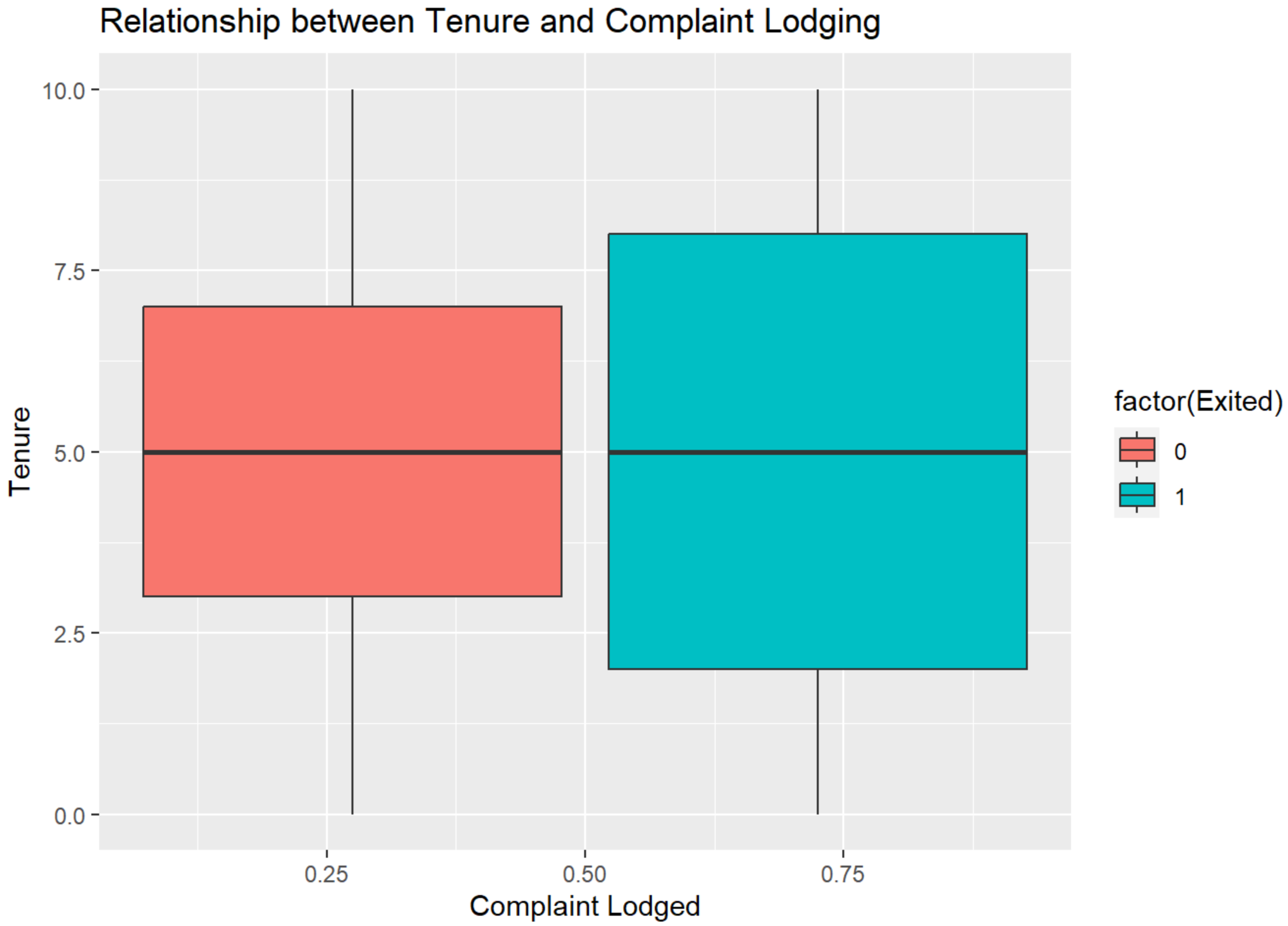
Pivotal insights due to multiple facets affecting customer churn:

Geographically, France dominates the customer base, but regions with fewer customers experience higher churn, suggesting potential service issues. Gender-wise, females exhibit higher churn rates than males, indicating possible gender-related factors impacting churn behavior. While credit card ownership among churned customers is notable, further investigation is required to understand its true effect on churn. Inactive members contribute significantly to high churn rates, emphasizing the need for targeted engagement strategies. Additionally, lodged complaints coincide with increased churn, pointing to a correlation between complaints and customer attrition. Interestingly, points earned and card types show varied impacts on churn, highlighting the complexity of factors influencing customer retention. However, higher satisfaction scores are consistently linked with lower churn rates, underscoring the importance of customer satisfaction in fostering loyalty and reducing churn.



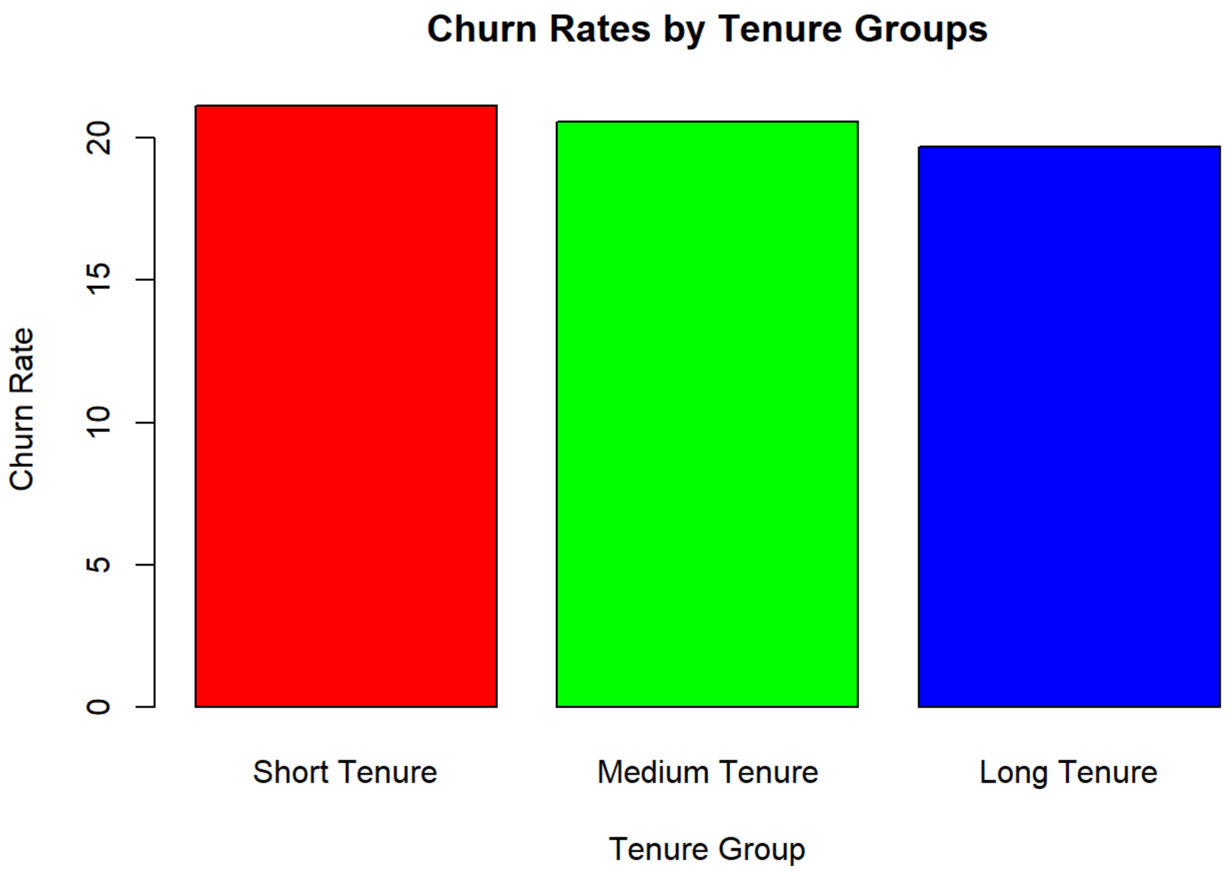
Analysis of credit scores didn’t reveal distinct differences between churned and retained customers. In contrast, age emerged as a significant factor, with older customers showing a higher tendency to churn compared to younger ones. Additionally, tenure played a crucial role, indicating that both short and lengthy tenures relate to increased churn, emphasizing the need to focus on retaining customers at extreme tenure periods. Surprisingly, customers with substantial bank balances exhibited a probability to churn, posing potential risks to the bank’s available capital. However, factors like the number of products held or estimated salary did not seem to notably impact the likelihood of churn, indicating their lesser influence on customer attrition within the dataset.

Impact of Complaint Lodging on Customer Tenure in Churn Analysis



Customers not filing complaints (Complain = 0) show consistent tenure regardless of churn. Churned customers who complained (Complain = 1, Exited = 1) exhibit slightly lower median tenure. Those churned with complaints display a wider tenure range than non-churned complainers.

Churn Analysis by Tenure Group



Churn rates show a slight decline with longer tenure, indicating lower churn among long-term customers. Shorter tenure correlates with slightly higher churn. Pearson’s Chi-squared test (p-value = 0.4077) suggests no significant relationship between churn and tenure categories.

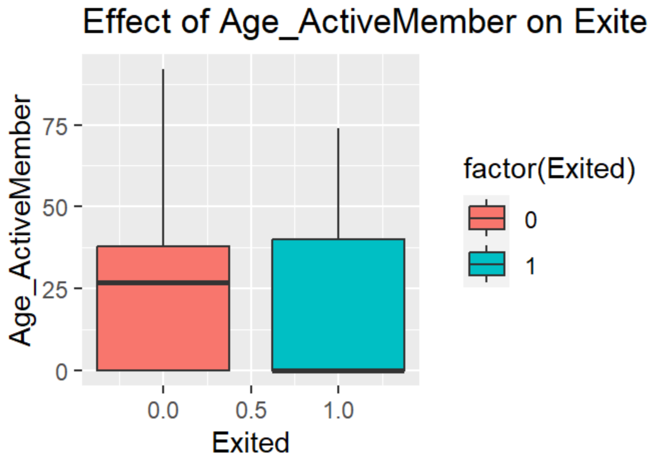
Comparing Tenure Across Customer Complaints and Churn Status

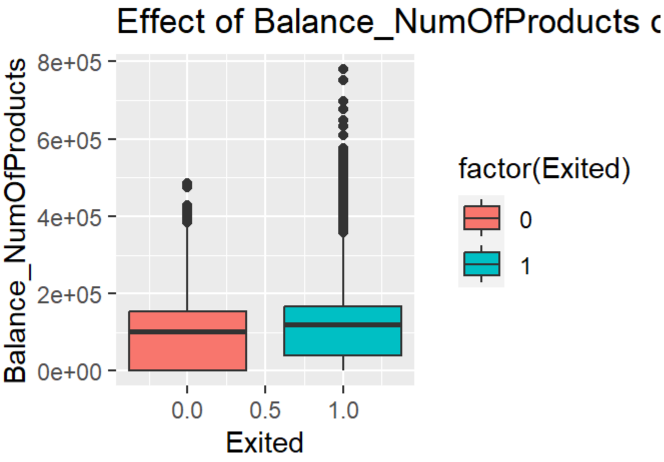
The conducted t-tests comparing tenure based on complaint filing and churn status suggest that there might not be a substantial difference in tenure duration concerning customers who lodged a complaint or those who churned

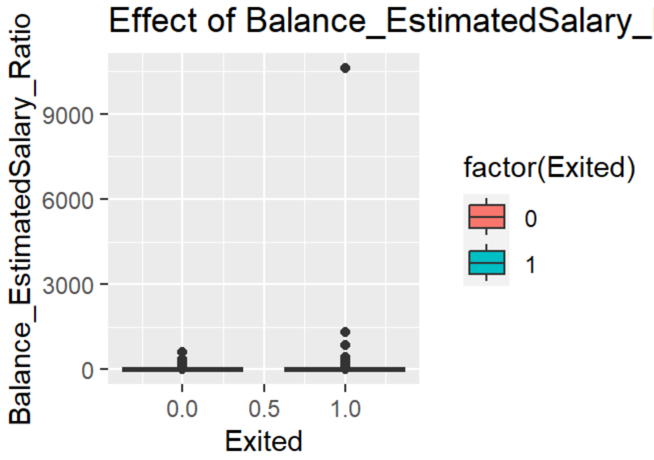
Interaction Effects and Feature Engineering

The following interaction terms were created to see their effect on churning

Age_ActiveMember, Balance_NumOfProducts, Balance_EstimatedSalary_Ratio



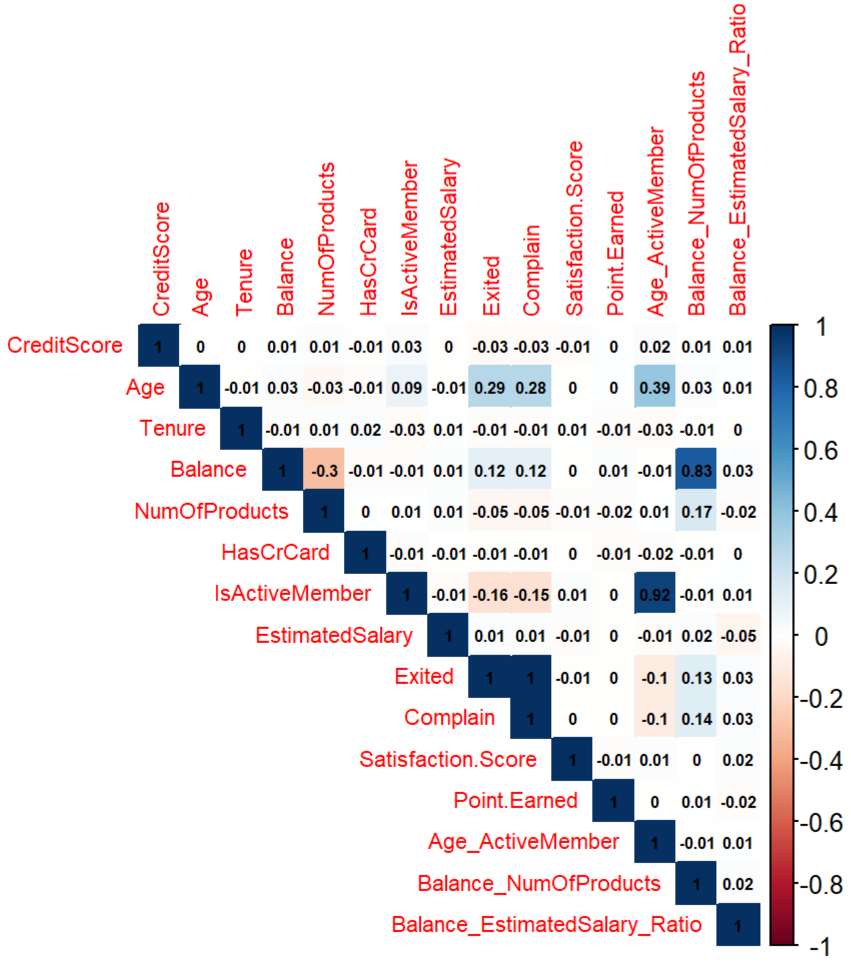




Inference on Interaction term’s Effects for Churn

The derived interaction variables, particularly Age_ActiveMember and Balance_NumOfProducts, suggest potential in distinguishing between churned and retained customers. These variables could influence churn prediction, hinting that customers with higher values in these derived features might exhibit a slightly elevated propensity to churn.

Correlation Analysis for Churn Prediction



Age exhibits a moderate positive correlation with the likelihood of churning. Variables such as IsActiveMember and NumOfProducts display moderate negative correlations, indicating that active members and customers with more products are less inclined to churn. A stronger correlation is observed between the Complain variable and Exited, requiring further investigation to ascertain its significance regarding churn. Engineered features like Balance_NumOfProducts and Balance_EstimatedSalary_Ratio show 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
```

A significant relationship exists between ‘Geography’ and ‘Exited’, suggesting that Geography notably influences customer churn. ‘Complain’ and ‘Exited’ showcase a highly significant relationship, indicating that customers who lodged complaints tend to churn significantly more than those who haven’t. ‘Number of Products’ held by a customer displays a significant relationship with ‘Exited’, indicating an impact on customer churn based on the number of products. No significant relationship is found between ‘Points Earned’ and ‘Exited’, implying a lack of substantial influence on churn.

Modelling

Significant Predictors

The 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.

```
## [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"
```

Performance Metrics of Logistic Regression Model

The logistic regression model exhibits high performance across various evaluation metrics: Accuracy of 99.8%, Precision of 99.75%, Recall of 100%, F1-Score of 99.87%, and an impressive ROC-AUC of 99.92%. These metrics collectively indicate the model’s robustness and effectiveness in predicting customer churn.

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

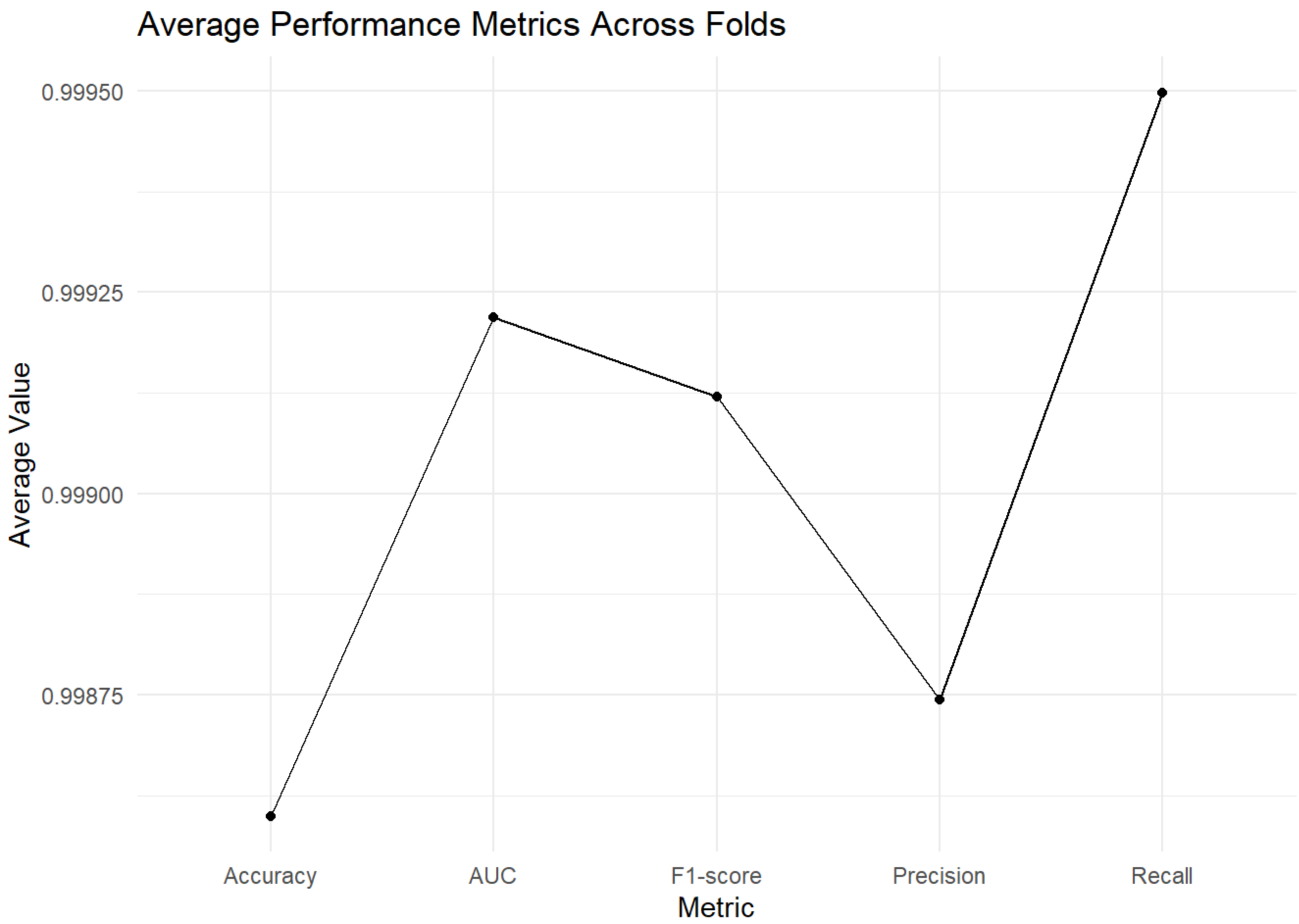
```
## Average AUC: 0.9992189
```

```
## Average Precision: 0.998744
```

```
## Average Recall: 0.9994978
```

```
## Average F1-score: 0.9991203
```

```
## Average Accuracy: 0.9985998
```



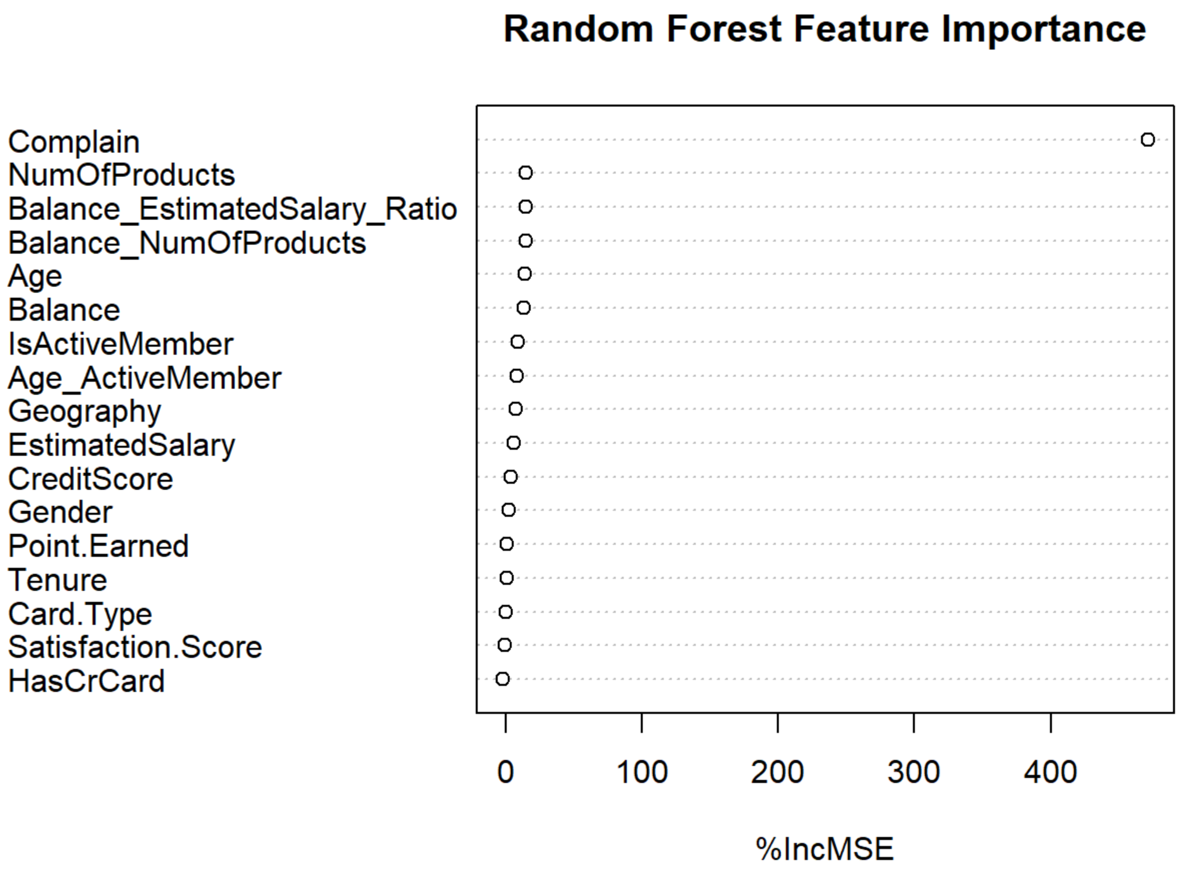
Cross-Validated Logistic Regression Model Evaluation

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.

Random Forest Feature Importance Analysis for Churn Prediction



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.

Conclusion

In identifying influential factors driving customer churn within the banking context, this study conducted an in-depth exploration. Visualizations meticulously dissected product engagement, card usage, geographic distribution, customer complaints, and tenure, serving as integral components linked to customer attrition. Leveraging these insights, predictive models were constructed utilizing demographic and banking variables using sophisticated modeling techniques like logistic regression and random forest. Analyzing customer churn in banking unveils vital insights essential for strengthening customer loyalty. By understanding customer behavior across demographics and regions, tailored retention strategies can be developed, addressing the specific needs of diverse customer groups. Prioritizing quick complaint resolution and delivering personalized services based on identified satisfaction patterns can significantly reduce churn rates and foster lasting customer loyalty.

Scope and Generalizability

This study provides nuanced insights valuable for devising effective customer retention strategies in industries dealing with similar challenges. However, to broaden the study’s applicability and ensure its reliability across diverse business landscapes, further validation across different datasets or industry domains is recommended. Strengthening feature engineering methodologies and refining the models through advanced techniques could significantly enhance their predictive accuracy and relevance across varied business scenarios, thereby reinforcing the study’s credibility and practical usefulness.

Limitations and Possibilities for improvement.

While the analysis provided valuable insights into customer churn within the banking sector, certain limitations warrant consideration. The scope of the analysis primarily focused on demographic and banking variables, potentially overlooking broader external factors influencing churn. Additionally, constraints within the dataset and the application of conventional modeling techniques like logistic regression and random forest might limit the depth of insights and predictive accuracy. To enhance the analysis, incorporating more diverse datasets, exploring advanced modeling approaches beyond traditional techniques, and considering external factors such as economic conditions could offer a more comprehensive understanding of customer churn behavior. These improvements could potentially enrich the depth and robustness of the analysis, enabling a more holistic view of factors driving customer attrition.