Final Project – Predicting Bank Customer Churn: A Machine learning approach.

Sruthi Kondra

College of Professional Studies, Northeastern University, Boston

ALY 6015: Intermediate Analytics Winter A 2024 (CRN: 20613)

Prof. Zhi (Richard) He

February 14, 2024

Introduction

In the fast-paced banking industry, accurately predicting customer churn is key to sustaining profitability and fostering growth. This project delves deep into analyzing customer data to forecast churn effectively. Utilizing three sophisticated machine learning models—Logistic Regression, Decision Tree, and SVM—we explore a comprehensive dataset that includes customer demographics, account specifics, and financial activities. Our detailed approach aims not only to pinpoint the most reliable churn predictor but also to derive meaningful insights that can inform strategic decision-making. Through this report, we aim to provide a detailed comparative analysis of three models.

Methods:

Our approach incorporates Logistic Regression, Decision Tree, and SVM models, complemented by Chi-Square tests, to navigate the intricacies of customer churn. The choice of these models allows us to explore both linear and non-linear patterns in customer behavior, with Logistic Regression assessing the impact of various predictors, Decision Trees uncovering the decision-making pathways, and SVM identifying complex patterns in high-dimensional data. Chi-Square tests further aid our analysis by evaluating the strength of association between categorical variables and churn, ensuring a holistic understanding of the factors at play.

This multifaceted methodological framework is designed to provide a comprehensive view of customer churn, integrating predictive modeling with statistical testing to uncover the myriad factors influencing customer decisions. Through this approach, we aim to not only predict churn more accurately but also offer actionable insights for strategic customer retention efforts, thereby contributing to enhanced sustainability and growth in the banking sector.

Analysis: Exploratory Data Analysis

Our exploratory data analysis (EDA) began with a comprehensive review of the dataset, focusing on uncovering patterns and insights into customer churn. Through a series of visualizations, including pie charts, bar charts, and box plots, we have understood the demographic and financial characteristics of the bank's customers. This initial investigation highlighted key areas for deeper analysis, setting the stage for applying machine learning models to predict churn.

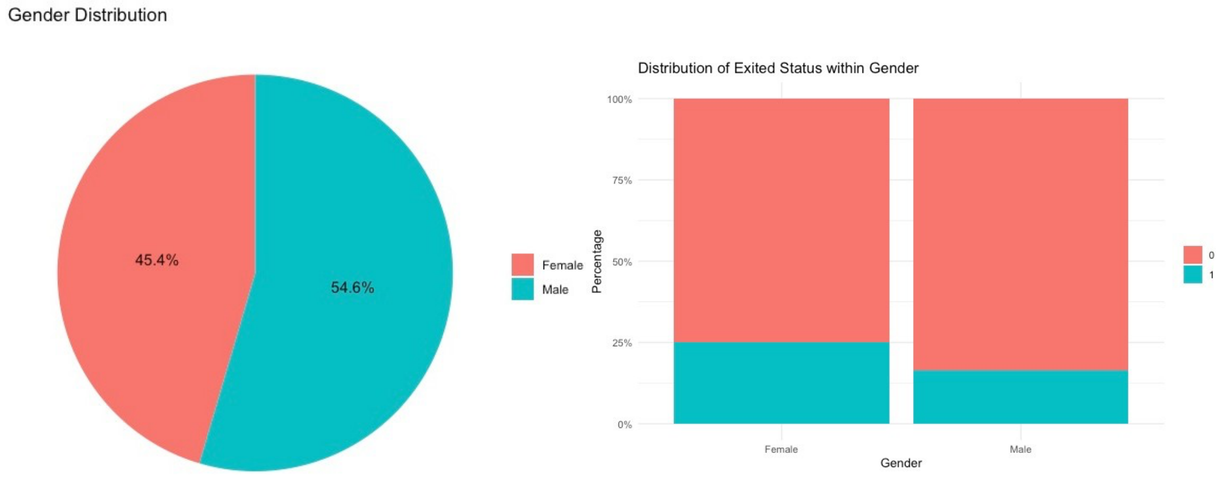


Figure 1. Distribution of Gender and their corresponding Exit Status

The pie chart indicates a slight male majority in the dataset, with males comprising 54.6% and females 45.4%. The stacked bar chart reveals a gender-wise distribution of churn, with slightly higher likelihood of female customers compared to male customers.

The pie chart as shown in Figure 2 illustrates the customer distribution across three geographies, showing that half are from France, while the rest are almost evenly split between Germany and Spain. The stacked bar chart compares the churn rates within these regions, highlighting variations that may indicate geographical influence on customer retention.

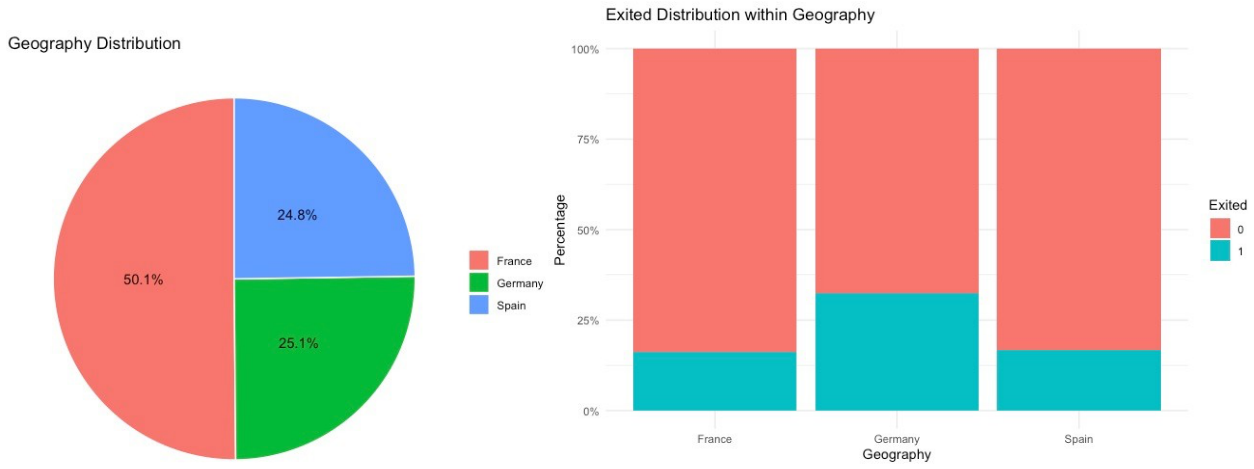


Figure 2. Distribution of customer across geographies and their exit status

The boxplots as shown below compare the age and credit score distributions between customers who have stayed with the bank and those who have left. Customers who have exited tend to be older, as indicated by the higher median age in the 'Exited' category. Credit scores do not show a distinct difference between the two groups. The estimated salary boxplot indicates that salary does not significantly differ between customers who stayed and those who exited, with similar medians and spreads across both categories.

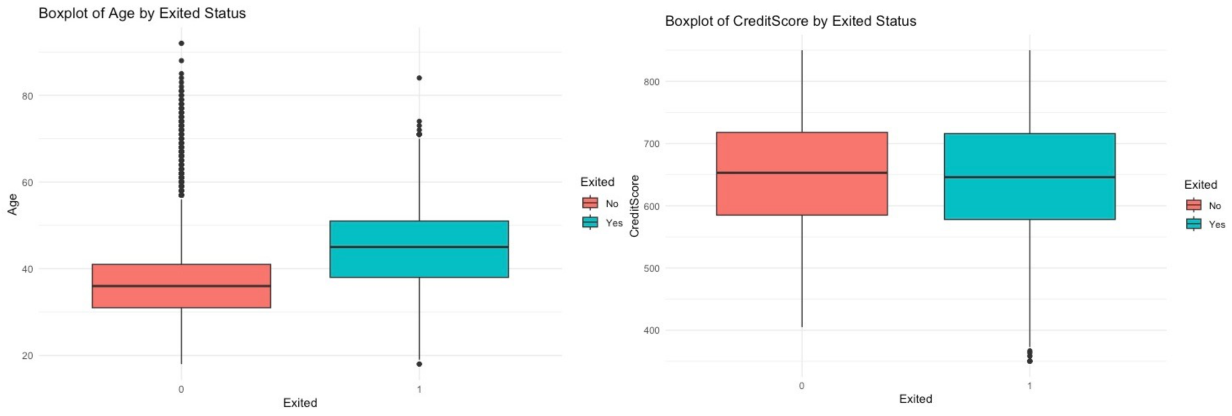


Figure 3. Boxplot of Age and CreditScore by Exited Status

The stacked bar chart below for the number of products shows that customers with only one or two products have lower churn rates, while nearly all customers with four products have left. Those with three products present an interesting contrast, as they have a noticeably higher churn rate compared to those with fewer products. This suggests that the number of products a customer has could be a significant factor in their decision to stay with or leave the bank.

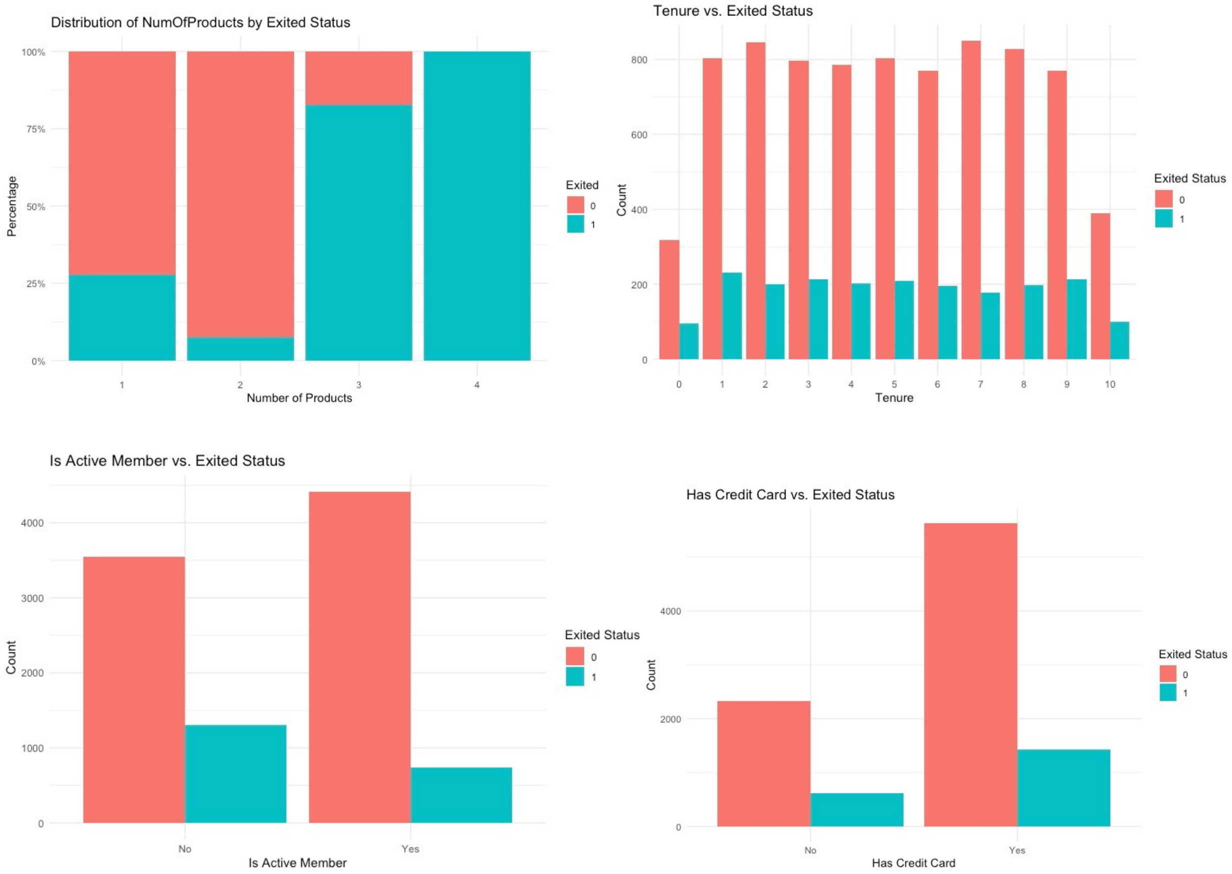


Figure 4. Distribution of Exited status of customers based on NumOfProducts.

The grouped bar charts as shown above explore the relationship between tenure, active membership, credit card possession, and customer churn. The tenure chart shows churn distributed across various lengths of tenure, with no clear trend suggesting tenure alone doesn't predict churn. Active membership status presents a more pronounced difference, with non-active members showing higher churn. Similarly, credit card ownership appears less influential, as the churn rates among those without a card are proportionally similar to those with a card, indicating other factors might play more significant roles in predicting churn.

Correlation Matrix

In the data preparation stage, we converted categorical variables into a format suitable for modeling and normalized numerical variables. Dummy variables were created for categorical features 'Geography' and 'Gender'. We then applied log transformation to skewed numerical data to improve model accuracy and interpretability. Finally, we scaled the features to a consistent range and visualized a correlation matrix to identify any strong correlations between variables, informing our subsequent modeling efforts.

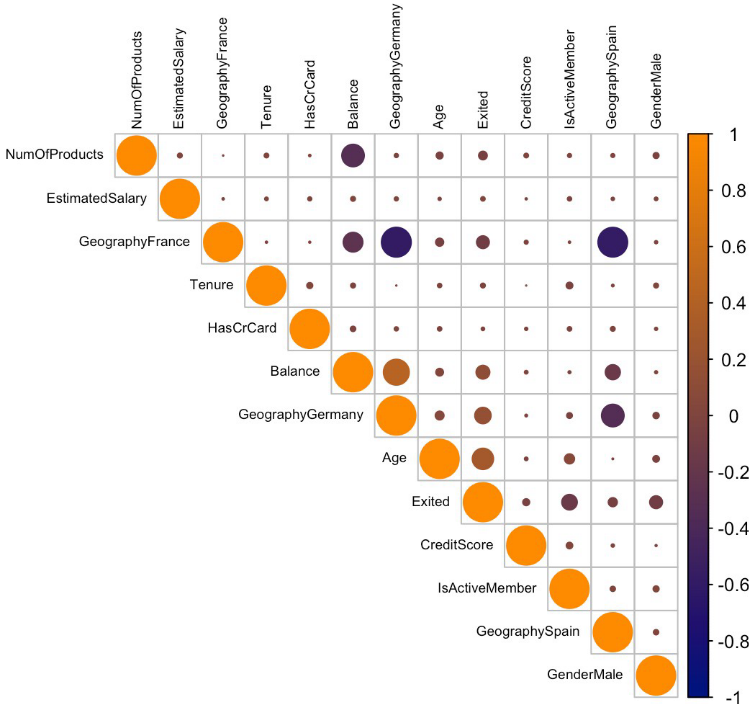


Figure 5. Correlation Matrix

Chi – Square Tests

The Chi-Square tests applied to various categorical variables against customer exit status yielded significant insights. To perform these tests below are the null and alternative hypothesis assumed for variables Gender, Credit Card possession, Active Membership status, Number of Products, and Geography

Null Hypothesis (H0): No association exists between categorical variable and exit status.

Alternative Hypothesis (H1): There exists association between chosen categorical variable and exit status.

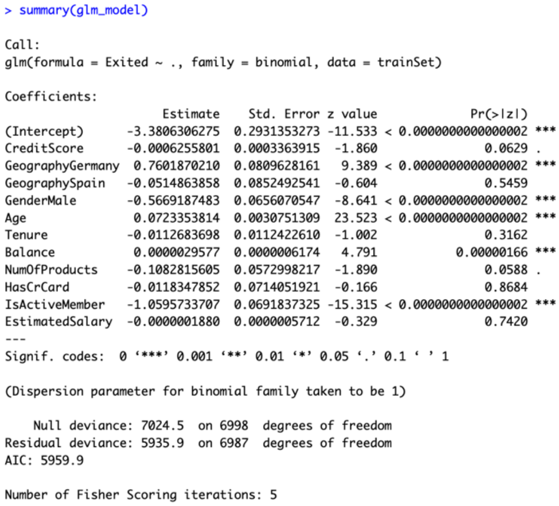
The strong chi-square statistic for 'Gender' indicates a significant association between gender and churn. Conversely, the 'HasCrCard' variable shows no such association due to its high p-value. 'IsActiveMember' and 'NumOfProducts' both show very strong associations with churn, as evidenced by their high chi-square statistics and low p-values. Similarly, 'Geography' also exhibits a significant relationship with customer churn. These results suggest that gender, active membership status, the number of products used, and geography are important factors in predicting churn, while having a credit card is not.

Variable	Chi-Square Statistic	Degrees of Freedom (df)	p-value	Significant Association?
Gender	112.2	1	< 0.0001	Yes
HasCrCard	0.4556	1	0.4997	No
IsActiveMember	242.76	1	< 0.0001	Yes
NumOfProducts	1501.3	3	< 0.0001	Yes
Geography	301.92	2	< 0.0001	Yes

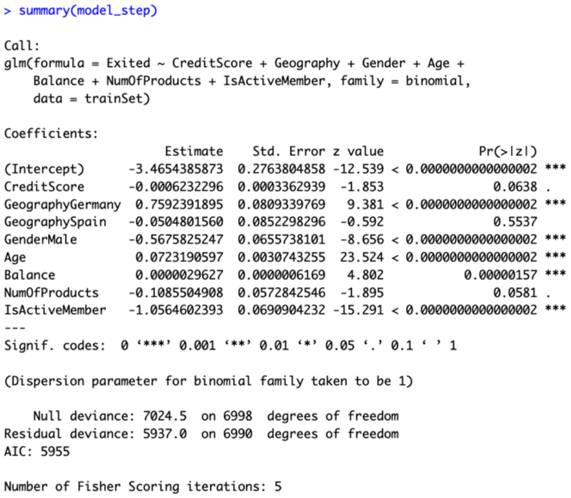
Table 1. Summary of Chi – Square Tests

Logistic Regression Model

The logistic regression model's summary points to several significant predictors of customer churn, with geography (particularly Germany), gender (male), age, account balance, and active membership status emerging as key factors. Credit score and the number of products, though part of the final model, showed marginal significance. We refined the model using Stepwise Regression, employing the stepAIC function, which optimizes the model by including variables that contribute the most explanatory power to churn prediction while removing less impactful ones.

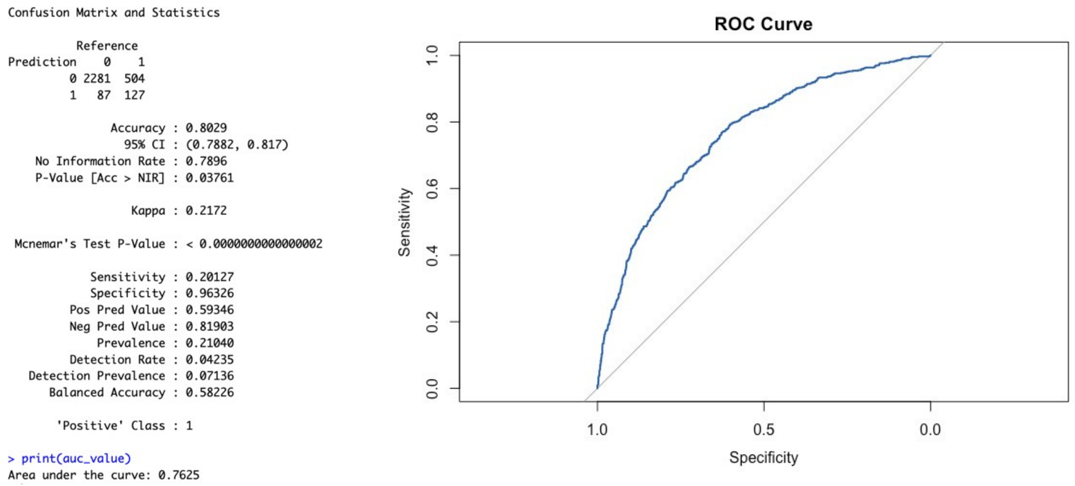


```
> summary(glm_model)

Call:
glm(formula = Exited ~ ., family = binomial, data = trainSet)

Coefficients:
                   Estimate    Std. Error z value               Pr(>|z|)    
(Intercept)     -3.3806306275  0.2931353273 -11.533 < 0.0000000000000002 ***
CreditScore     -0.0006255801  0.0003363915  -1.860               0.0629 .  
GeographyGermany 0.7601870210  0.0809628161   9.389 < 0.0000000000000002 ***
GeographySpain  -0.0514863858  0.0852492541  -0.604               0.5459    
GenderMale      -0.5669187483  0.0656070547  -8.641 < 0.0000000000000002 ***
Age              0.0723353814  0.0030751309  23.523 < 0.0000000000000002 ***
Tenure          -0.0112683698  0.0112422610  -1.002               0.3162    
Balance          0.0000029577  0.0000006174   4.791           0.00000166 ***
NumOfProducts   -0.1082815605  0.0572998217  -1.890               0.0588 .  
HasCrCard       -0.0118347852  0.0714051921  -0.166               0.8684    
IsActiveMember  -1.0595733707  0.0691837325 -15.315 < 0.0000000000000002 ***
EstimatedSalary -0.0000001880  0.0000005712  -0.329               0.7420    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7024.5  on 6998  degrees of freedom
Residual deviance: 5935.9  on 6987  degrees of freedom
AIC: 5959.9

Number of Fisher Scoring iterations: 5
```



```
> summary(model_step)

Call:
glm(formula = Exited ~ CreditScore + Geography + Gender + Age + 
    Balance + NumOfProducts + IsActiveMember, family = binomial, 
    data = trainSet)

Coefficients:
                   Estimate    Std. Error z value               Pr(>|z|)    
(Intercept)     -3.4654385873  0.2763804858 -12.539 < 0.0000000000000002 ***
CreditScore     -0.0006232296  0.0003362939  -1.853               0.0638 .  
GeographyGermany 0.7592391895  0.0809339769   9.381 < 0.0000000000000002 ***
GeographySpain  -0.0504801560  0.0852298296  -0.592               0.5537    
GenderMale      -0.5675825247  0.0655738101  -8.656 < 0.0000000000000002 ***
Age              0.0723190597  0.0030743255  23.524 < 0.0000000000000002 ***
Balance          0.0000029627  0.0000006169   4.802           0.00000157 ***
NumOfProducts   -0.1085504908  0.0572842546  -1.895               0.0581 .  
IsActiveMember  -1.0564602393  0.0690904232 -15.291 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7024.5  on 6998  degrees of freedom
Residual deviance: 5937.0  on 6990  degrees of freedom
AIC: 5955

Number of Fisher Scoring iterations: 5
```

This feature selection process confirmed the importance of variables such as CreditScore, Geography, Gender, Age, Balance, NumOfProducts, and IsActiveMember, and excluded non-contributing predictors. Our model evaluation through a confusion matrix on the test set demonstrated an accuracy of 80.29% with moderate sensitivity and high specificity. Furthermore, the VIF analysis indicated no serious multicollinearity among the predictors, ensuring that the model's predictive power is reliable. The ROC curve as shown below, with an area under the curve of 0.7625, underscores the model's decent capability to distinguish between churned and retained customers.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2281  504
         1   87  127

               Accuracy : 0.8029
                 95% CI : (0.7882, 0.817)
    No Information Rate : 0.7896
    P-Value [Acc > NIR] : 0.03761

                  Kappa : 0.2172

 Mcnemar's Test P-Value : < 0.0000000000000002

            Sensitivity : 0.20127
            Specificity : 0.96326
         Pos Pred Value : 0.59346
         Neg Pred Value : 0.81903
             Prevalence : 0.21040
         Detection Rate : 0.04235
   Detection Prevalence : 0.07136
      Balanced Accuracy : 0.58226

       'Positive' Class : 1

> print(auc_value)
Area under the curve: 0.7625
```

Figure 6. Confusion Matrix and ROC Curve – Logistic Regression Model

Decision Tree – Classification Model

For the Decision Tree model, we initiated the process by fitting the model to the training data, employing `rpart` with classification settings. To fine-tune our model, we conducted cross-validation, setting specific parameters for complexity (`cp`) and minimum splits, ensuring we derived a model that balances complexity with predictive power. The subsequent pruning of the tree, based on minimizing cross-validation error, led to a more generalized model, less prone to overfitting.

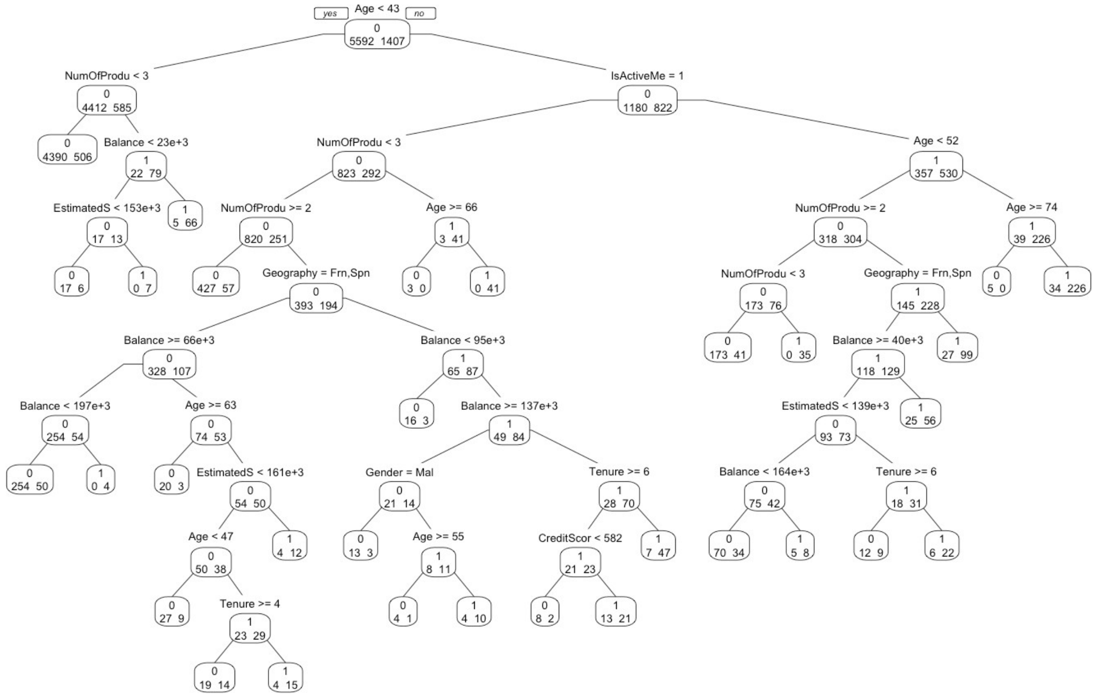
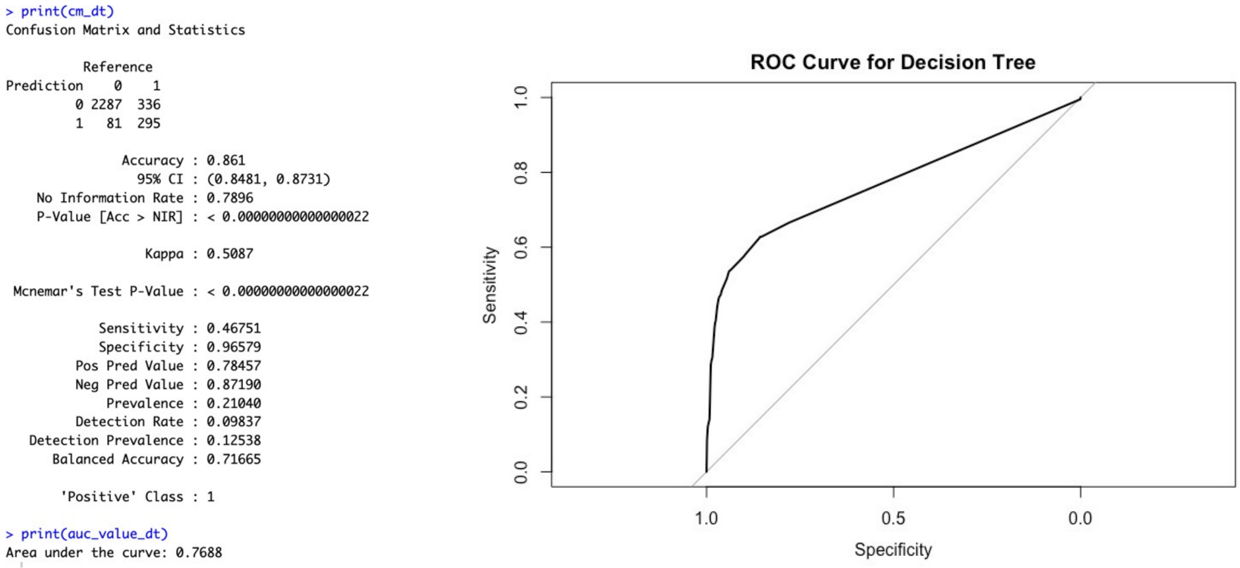


Figure 7. Decision Tree Model



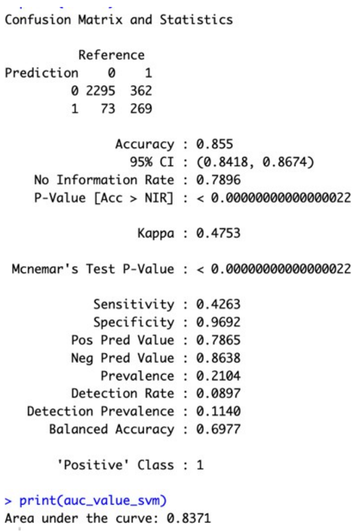
```
> print(cm_dt)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2287  336
         1   81  295

               Accuracy : 0.861
                 95% CI : (0.8481, 0.8731)
    No Information Rate : 0.7896
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5087

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.46751
            Specificity : 0.96579
         Pos Pred Value : 0.78457
         Neg Pred Value : 0.87190
             Prevalence : 0.21040
         Detection Rate : 0.09837
   Detection Prevalence : 0.12538
      Balanced Accuracy : 0.71665

       'Positive' Class : 1

> print(auc_value_dt)
Area under the curve: 0.7688
```

Figure 8. Confusion Matrix and ROC Curve – Decision Tree Model

Upon testing, the pruned Decision Tree model exhibited promising results. The confusion matrix highlighted an accuracy of 86.1%, indicating a high level of correct predictions. Sensitivity and specificity were 46.75% and 96.579% respectively, showcasing the model's effectiveness in identifying true positives and true negatives. Notably, the positive predictive value and the balanced accuracy stood out, reinforcing the model's reliability. The AUC of 0.7688 further attests to the model's capability to predict customer churn.

Support Vector Machine (SVM) Model

In the final phase of our analysis, we explored the Support Vector Machine (SVM) model, leveraging its capacity for high-dimensional data. Utilizing the `e1071` package, we trained the SVM with a radial basis function kernel, enabling probability estimates to enhance our understanding of churn likelihood. This approach is particularly suited for complex datasets where linear boundaries are insufficient to separate the classes effectively.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2295  362
         1   73  269

               Accuracy : 0.855
                 95% CI : (0.8418, 0.8674)
    No Information Rate : 0.7896
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.4753

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.4263
            Specificity : 0.9692
         Pos Pred Value : 0.7865
         Neg Pred Value : 0.8638
             Prevalence : 0.2104
         Detection Rate : 0.0897
   Detection Prevalence : 0.1140
      Balanced Accuracy : 0.6977

       'Positive' Class : 1

> print(auc_value_svm)
Area under the curve: 0.8371
```

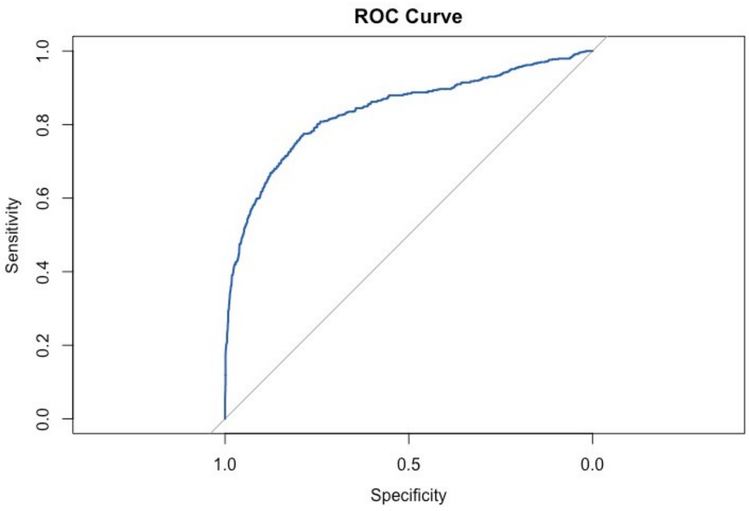


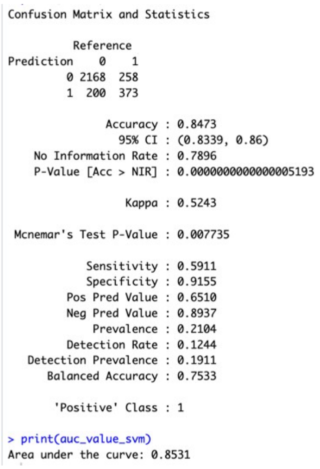
Figure 9. Confusion Matrix and ROC Curve – SVM Model

The SVM model demonstrated strong performance, as evidenced by the confusion matrix and ROC analysis. With an accuracy of 85.5%, the model significantly outperformed the no-information rate, indicating its effectiveness in predicting customer churn. The sensitivity and specificity rates of 42.63% and 96.92%, respectively, alongside a positive predictive value of 78.65%, highlight the model's precision in identifying churn. The AUC of 0.8371 further underscores the SVM model's superior discriminative ability compared to the previously discussed models, making it a compelling choice for tackling customer churn prediction in banking.

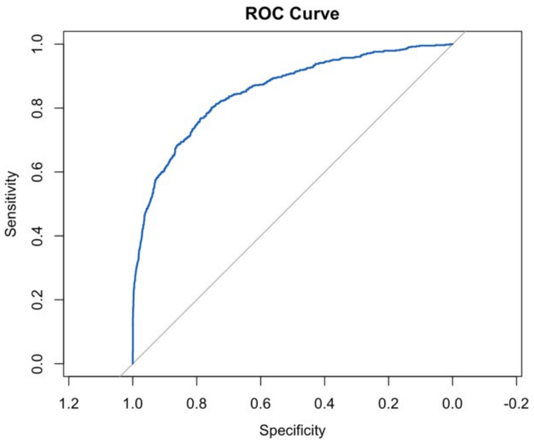
Improving sensitivity of model

Following feedback on our draft project, we initiated a effort to enhance our predictive model's sensitivity in predicting customer churn. This involved sophisticated feature engineering and experimentation with classification thresholds across various models, including Logistic Regression, Decision Trees, and Support Vector Machines (SVM). It was the SVM model that exhibited a notably superior enhancement in performance. This improvement was primarily attributed to adjusting the classification threshold, a strategy that significantly increased the model's sensitivity to accurately identifying customers at risk of churn.

In refining our SVM model with a polynomial kernel for churn prediction, we focused on adjusting the classification threshold to improve sensitivity. Initially set at 0.5, lowering the threshold to 0.25 resulted in significant performance enhancements. This adjustment increased the model's sensitivity from 42.63% to 59.11%, enabling better detection of customers likely to churn. Although specificity slightly decreased from 96.92% to 91.51%, and accuracy dipped marginally from 85.6% to 84.69%, these changes were strategically acceptable to prioritize early churn identification. The Positive Predictive Value (PPV) was observed at 64.98%, and the Area Under the Curve (AUC) improved to 0.8531 from 0.8361, affirming the model's robust discriminative ability.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2168  258
         1  200  373

               Accuracy : 0.8473
                 95% CI : (0.8339, 0.86)
    No Information Rate : 0.7896
    P-Value [Acc > NIR] : 0.0000000000000005193

                  Kappa : 0.5243

 Mcnemar's Test P-Value : 0.007735

            Sensitivity : 0.5911
            Specificity : 0.9155
         Pos Pred Value : 0.6510
         Neg Pred Value : 0.8937
             Prevalence : 0.2104
         Detection Rate : 0.1244
   Detection Prevalence : 0.1911
      Balanced Accuracy : 0.7533

       'Positive' Class : 1

> print(auc_value_svm)
Area under the curve: 0.8531
```



This targeted approach of kernel selection and threshold adjustment underscores our analytical strategy to enhance model precision for churn prediction. The polynomial kernel SVM, with its adjusted threshold, demonstrated a balanced trade-off between detecting true positives and maintaining overall model performance. These efforts reflect our objective to develop a predictive model that effectively identifies at-risk customers, supporting informed and timely retention strategies.

Comparative Analysis

Upon evaluation of performance of the Logistic Regression, Decision Tree, and Support Vector Machine (SVM) models on a customer churn prediction dataset, a nuanced understanding of each model's efficacy emerges. The Logistic Regression model manifested an accuracy of approximately 80.29%, which, while respectable, was surpassed by the other models in comparison. Its Area Under the Curve (AUC) of roughly 0.7625 reflects a competent, albeit not exceptional, capability in distinguishing between churned and retained customers.

In contrast, the Decision Tree model showcased a notable improvement, achieving an accuracy of 86.1% and an AUC of 0.7688. This increment in performance metrics indicates a better grasp of the dataset's underlying patterns, likely attributable to the model's ability to capture non-linear relationships more effectively than its logistic counterpart. Additionally, the Decision Tree model provided a sensitivity of 46.75% and a specificity of 96.579%, with a positive predictive value (PPV) of 78.65% and a negative predictive value (NPV) of 86.38%, presenting a balanced approach to predicting churn.

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	80.29%	20.13%	96.33	0.7625
Decision Tree	86.1%	46.75%	96.579%	0.7688
SVM (Radial)	85.6%	42.63%	96.92%	0.8371
SVM (Polynomial)	84.73%	59.11%	91.55%	0.8531

Table 2. Summary of Models Performance

The SVM model, leveraging a polynomial function kernel, emerged as the superior model with an accuracy of 84.73% and an outstanding AUC of 0.8531, indicating its superior discriminative power. Its sensitivity stood at 59.11% with a specificity of 91.51%, though improving from earlier model it has scope for improvement. The higher AUC signifies SVM's enhanced capability in effectively separating the churned customers from those who remained, making it the most robust model among the three.

Conclusion

In concluding our exploration of predictive modeling for customer churn within the banking sector, it's evident that the application of machine learning models, particularly the Support Vector Machine (SVM), has markedly enhanced our ability to identify customers at risk of leaving. The SVM model, with its standout AUC performance, underscores the potential of advanced analytics in transforming customer retention strategies. This project not only highlights the critical role of predictive modeling in understanding customer behavior but also paves the way for banks to proactively address churn. Armed with these insights, banks can refine their approaches to customer engagement, deploy targeted interventions, and ultimately cultivate a more stable and loyal customer base. This endeavor reaffirms the value of leveraging sophisticated data analysis techniques to drive business decisions and foster enduring customer relationships in today's competitive banking landscape.

Reference:

1. A. Bluman, *Elementary Statistics* 10th Edition, McGraw Hill ISBN 978-130-7494-327

2. R. Kabacoff, *R in Action* 2nd Edition, Manning Publisher ISBN 978-161-7291-388

3. Dhakadd, S.2022. *Bank customer churn prediction [Data set]*. Kaggle. https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction?resource=download

4. Wiryaseputra, M. (2022, October 12). *Bank customer churn prediction using machine learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction-using-machine-learning