Image 1

Customer Churn Analysis – Banking Sector

This project was completed as part of a team. I contributed to all phases — from data preparation and exploratory analysis to model development, evaluation, and final reporting.

Abstract:

This paper explores the key drivers of bank customer churn by applying multiple data mining techniques to a dataset of over 10,000 customer records. Our objective was to determine which demographic, behavioral, and financial attributes strongly influence whether a customer stays with or exits a bank. Using a range of models—logistic regression, linear regression, CART analysis, k-nearest neighbors (kNN), and Naive Bayes—we evaluated each model's effectiveness at church prediction and identified the most informative predictors.

Overall, several models resulted in accuracy in the mid-to-high 80th percentile, indicating the models show promise for fine-tuning and eventual deployment as useful predictors to help banks prevent customer churn. The machine learning models identified tenure and account balance as important retention indicators, though they struggled with sensitivity. Most models were highly successful in predicting what makes customers loyal but struggled to identify customers who left, telling us our models have valuable information to give to the banking industries but have room for improvement and fine-tuning when specifically looking at why customers churn.

Our findings suggest that tenure, account balance, and customer activity levels are consistently important in predicting churn. We recommend that banks lean towards using predictive models that balance sensitivity and specificity to find appropriate retention strategies to keep customers, especially focusing on new and low-engagement customers. This research paper demonstrates how applied machine learning can provide valuable insights for customer retention and strategic decision-making in the banking sector. Below, you will find detailed explanations of our models, accompanied by graphs and findings.

Introduction:

In today's increasingly competitive banking industry, customer retention has become a top priority. Even minor improvements in retention can lead to substantial profit gains. Reichheld and Sasser (1990) found that a 5% increase in customer retention could boost profits by 25% to 95%. Understanding why customers leave and identifying the warning signs in advance is crucial for designing effective strategies to reduce churn and build long-term loyalty within financial institutions. Not only does this allow banks to better serve their clientele, but also reach their target base better.

Although churn prediction has been explored in industries like telecommunications and online retail, its application within the banking industry remains relatively under-discussed. While variables like credit score and tenure are often cited as indicators of customer loyalty, there is limited exploration of how a combination of financial, behavioral, and demographic variables influences a customer's decision to leave. Our project aims to address this gap by applying classification models and statistical analysis to uncover the most significant predictors of churn and evaluate how they relate to one another.

The main objective of this project is to determine the factors most strongly influencing whether a customer chooses to stay with or leave their bank. We aim to identify the relationships between demographic information, customer behavior, and financial indicators and use that analysis to build models that can accurately predict churn. Through this process, we aim to offer insights that assist banks in developing data-driven customer retention strategies, reducing the cost of customer acquisition and increasing long-term profitability.

This topic is particularly relevant because customer churn is not just a financial concern.

It also reflects the quality of a bank's services and relationships with its clients. Banks can respond proactively by better understanding the behaviors and characteristics of customers who leave, offering targeted solutions and personalized experiences. Finally, this project provides our team with the opportunity to apply data mining techniques learned throughout the semester to a practical, real-world problem, thereby reinforcing our skills in data exploration, feature selection, model development, and interpretation.

Methodology and Results:

The dataset for our project was sourced from Kaggle.com and provides a rich basis for predictive modeling due to its combination of customer profile attributes and financial metrics. The dataset contains the customer data of account holders at an anonymous multinational bank, and the data centers on predicting churn based on attributes such as *Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Active Membership, Estimated Salary, Complain status, Satisfaction Score, and Card Type*.

We began preprocessing by removing non-predictive variables *(Row Number, Customer ID, and Surname*). Since there were no missing values, we moved directly to converting categorical variables to factors and normalizing the numeric features. To further improve our predictive power, we created several engineered features, including *Age Groups, HasBalance* (binary indicator), *Credit_Age* (Credit Score * Age), *Active_Products* (Active Member * Number of Products), and a *HighRiskProfile* marker based on risky customer behavior. Additional variables included *BalanceToSalary* ratio, grouped *Tenure* categories ("New," "Mid-Term," "Loyal"), and a *LoyaltyScore* combining tenure, satisfaction, activity, and complaints. After feature engineering, we performed a stratified 70/30 train-test split and set up 10-fold cross-validation to ensure a robust dataset for application to our chosen models.

We also briefly explored the data through visualizations, such as churn distribution by Loyalty Score, Balance-to-Salary ratio, and Tenure Group (see Appendix A, Figure A1).

With the data prepared, we proceeded to apply our models to predict churn outcomes: Logistic Regression, Linear Regression, CART Analysis, K-Nearest Neighbor, and Naive Bayes.

Logistic Regression:

This analysis applied logistic regression to estimate the likelihood of customer churn based on a bank dataset containing ten predictors: CreditScore, Geography, Gender, Age, Balance, EstimatedSalary, Tenure, IsActiveMember, NumOfProducts, and Exited. These features were selected due to their expected influence on customer behavior and financial engagement. Similar variables are effective in prior churn prediction models (Verbeke et al., 2012).

Preprocessing involved converting categorical variables to factors and splitting the data into training and test sets (70/30) using the *caret* package in R. A logistic regression model was trained with the *glm()* function using a binomial link. Predictions on the test set were converted to binary outcomes and evaluated using various classification metrics.

The model reached 81.49% accuracy, showing strong overall performance. However, sensitivity was low at 22.59%, indicating poor detection of churned customers, while specificity was high at 96.57%, effectively identifying retained customers. The AUC was around 0.77, reflecting good discriminative ability.

Confusion Matrix:

```
Accuracy: 81.49%

Kappa: 0.2514
```

```
Sensitivity (True Positive Rate): 22.59%

Specificity (True Negative Rate): 96.57%

AUC (Area Under the Curve): ~0.77

Reference

Prediction   no  yes

       no  2306  473

       yes   82  138
```

The confusion matrix showed 2,306 true negatives and 138 true positives, with 473 false negatives and 82 false positives. This highlights the model's strength in predicting non-churners but weakness in detecting actual churners, reflected in its low sensitivity (22.59%).

Appendix A, Figure A2 shows that customers from Germany and Spain are more likely to churn, as indicated by positive coefficients for Geography_Germany and Geography_Spain. In contrast, negative coefficients for Gender_Male, IsActiveMember, and NumOfProducts suggest that male, active, and more engaged customers are less likely to leave. These patterns highlight the influence of both demographic and behavioral factors on churn.

Overall, the model was better at identifying customers likely to stay. This imbalance may stem from logistic regression's limitations in capturing complex patterns. Future models like Naive Bayes or decision trees could better handle such complexity and improve churn detection.

$\text{Model}=\log(1-P(\text{Exited}=1)P(\text{Exited}=1))=\beta0+\beta1\cdot\text{CreditScore}+\beta2\cdot\text{Geography}+\beta3\cdot\text{Gender}+\beta4\cdot\text{Age}+\beta5\cdot\text{Balance}+\beta6\cdot\text{EstimatedSalary}+\beta7\cdot\text{Tenure}+\beta8\cdot\text{IsActiveMember}+\beta9\cdot\text{NumOfProducts}$

Linear Regression:

To prepare the data, unnecessary identifier columns were removed, and key variables like age, gender, geography, balance, and credit score were retained. The machine learning technique applied was Linear Regression. The aim was to predict a customer's credit score using five specific variables: Age, Gender, Geography, Balance, and IsActiveMember. These were selected based on their potential influence on creditworthiness. Prior research supports the use of regression models for financial predictions. For instance, Khandani, Kim, and Lo (2010) demonstrated that borrower attributes and transaction data can be used to develop accurate credit-risk models.

Data preprocessing involved converting categorical variables to factors and engineering several new variables. While these additional features were created for exploratory purposes, the final model used only the original five predictors to maintain simplicity and interpretability.

Model = CreditScore ~ Age + Gender + Geography + Balance + IsActiveMember + LoyaltyScore + BalanceToSalary + TenureGroup + Active_Products

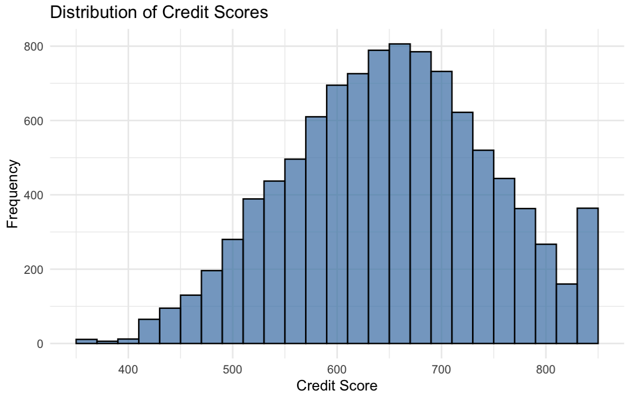
Predictions were generated on the holdout set and compared to actual values. The final model had a Root Mean Squared Error of 96.4, indicating that predictions were, on average, about 96 points away from the actual credit score. The Mean Absolute Error was 78.23, which similarly suggests considerable error dispersion. The R-squared value was just 0.0016, meaning less than 0.2% of the variation in credit scores was explained by the model. Nonetheless, based on RMSE relative to the mean credit score, an informal pseudo-accuracy was computed as 85.18%. While this figure seems high, it reflects closeness to the mean rather than strong explanatory power.

These results suggest that the chosen predictors were not particularly effective in explaining variations in credit score. The poor performance may be due to weak linear relationships or unaccounted interactions between variables. More complex models or enhanced feature selection could improve performance in future iterations.

A small sample of the model's predictions is presented below:

- Predicted: 651.45, Actual: 850, Residual: -198.55
- Predicted: 650.12, Actual: 645, Residual: 5.12
- Predicted: 648.59, Actual: 822, Residual: -173.41

Charts were created to visually support the analysis:



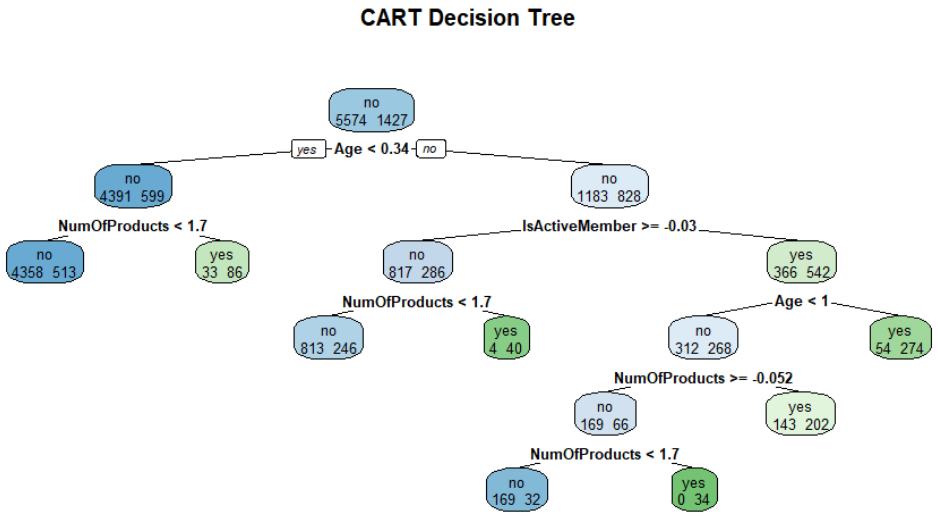
This histogram revealed that most scores clustered around the mid-600s, with a roughly normal distribution. This visualization helped confirm the appropriateness of linear regression for this dataset and provided insight into how scores are distributed across the customer base.

The scatterplot (see Appendix A, figure A3) shows a neutral relationship between age and credit score. Although the trend line slightly declines, the wide spread of data points indicates that age is not a strong predictor of credit score.

In conclusion, this section demonstrated the application of linear regression to a financial dataset and highlighted the challenges of modeling credit scores based on limited features. While the model's performance was weak, the exercise offered valuable practice in data cleaning, feature selection, and interpreting regression metrics.

CART:

In this analysis, a Classification and Regression Tree (CART) model was utilized to predict customer churn. The target variable for this analysis was, once again, Exited, a binary outcome indicating if the customer left the bank (1 = yes) or not (0 = no). To build the model, the selected predictor variables used were: CreditScore, Age, Balance, NumOfProducts, IsActiveMember, and EstimatedSalary. These were chosen to minimize overfitting by excluding highly correlated or irrelevant variables (Breiman et al., 1986). The CART model was trained using the training dataset, and predictions were generated for both the training and holdout datasets in order to compare performance and accuracy measures. The finished tree is displayed below:



Performance was evaluated using confusion matrices and related classification metrics. For the training set, the accuracy was 85.4%, the sensitivity (correctly identifying customers who did not exit) was 95.8%, the specificity (correctly identifying customers who did exit) was 44.6%, and the balanced accuracy was 70.2% (see Appendix A, Figure A4). The positive predictive value (precision) for 'no' was 87.1%, and for 'yes' it was 73.1% (Breiman et al., 1986).

For the holdout set, the accuracy was 85.3%, the sensitivity was 95.6%, the specificity was 44.8%, and the balanced accuracy again came to 70.2%. The positive predictive value for 'no' was 87.1%, and for 'yes' it was 72.5% (See Appendix A, Figure A5).

These results indicate that the model is quite effective at identifying customers who are likely to stay with the bank, but less effective at predicting which customers will leave, which is likely due to class imbalance, where the majority class dominates prediction performance (Breiman et al., 1986). Still, the tree provides clear, interpretable rules (see above tree) and highlights key churn-related predictors, such as Age, Balance, and IsActiveMember, which can inform customer retention strategies.

kNN:

For this analysis, we applied the k-Nearest Neighbors (kNN) algorithm to predict customer churn, using a dataset of 7,001 samples. The outcome variable was Exited, a binary variable indicating whether a customer left the bank’s service (1 = yes) or remained (0 = no). The dataset included 8 predictor variables related to customer demographics and account information: CreditScore, Age, Balance, NumOfProducts, IsActiveMember, EstimatedSalary, LoyaltyScore, and BalanceToSalary. Before modeling, we centered and scaled the predictors to standardize their values, which is a recommended preprocessing step for distance-based algorithms like kNN (Halder et al., 2024, p. 3).

We trained the kNN model using 10-fold cross-validation and evaluated model performance across values of k ranging from 1 to 16. The model was tuned based on the ROC (Receiver Operating Characteristic) score which measures the balance between true positive rate and false positive rate across all classification thresholds, selecting the value of k that maximized the area under the ROC curve.

As seen from Figure A6 in Appendix A, k = 16 produced the highest ROC score of 0.8245, making it the best-performing model. While sensitivity (true positive rate) remained high throughout, specificity (true negative rate) was relatively lower, suggesting the model is better at identifying customers who stayed than those who left, as the model exhibited a tendency to misclassify exiting customers as loyal ones.

After training the model with optimal parameters (k = 16), we tested it on a held-out test dataset. The final performance metrics were:

- Accuracy: 85.4%
- Kappa: 0.4434
- Sensitivity (True Positive Rate): 97.40%
- Specificity (True Negative Rate): 38.46%
- AUC (Area Under the Curve): 0.8274

The model performed well overall but showed a low specificity. This means while it accurately predicts those who will remain, it struggles to flag customers who are likely to churn, which is an important insight for retention strategy design. The confusion matrix below shows that the model correctly predicted 2,326 customers who stayed and 235 who left, but also misclassified 376 churned customers as stayers.

Confusion Matrix:

```
            Reference
Prediction   no     yes
        no   2326   376
       yes   62     235
```

Visualizations:

The ROC curve in Appendix A, Figure A7 shows the AUC improving with higher values of *k*, peaking at *k = 16*, which justifies its selection for the final model. An AUC near 0.83 indicates strong classification performance, well above random guessing (AUC = 0.5). Appendix A, Figure A8 displays the ROC curve for the k-Nearest Neighbors (kNN) model, which arcs above the diagonal line, highlighting effective classification with a good balance between sensitivity and specificity. The curve's shape, bowing toward the top-left, suggests an AUC of approximately 0.82–0.83. The model shows high sensitivity, accurately detecting most churners, but lower specificity, leading to some false positives. This trade-off is often acceptable in business, where prioritizing the identification of churners outweighs the risk of misclassifying loyal customers (Halder et al., 2024, p. 29).

The kNN model demonstrated strong predictive power with an AUC of 0.8274. However, the low specificity indicates a tendency to falsely classify some exiting customers as staying, which could correlate to negative implications in customer retention strategies. From a business perspective, this model could be valuable in identifying loyal customers with high certainty, but additional refinement or a hybrid model may be necessary to better flag at-risk customers before they churn (Halder et al., 2024, p. 14).

Naive Bayes:

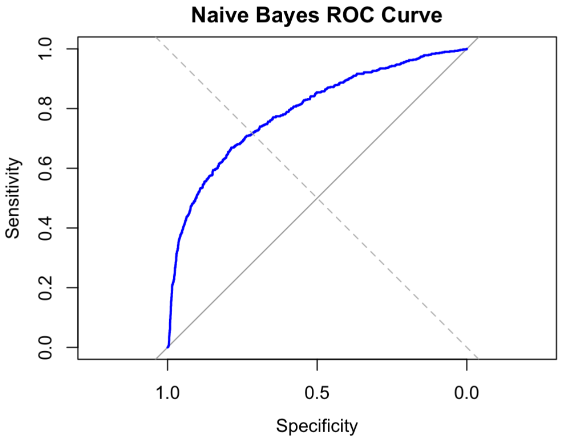
A Naive Bayes model was trained, hinging again on the target variable - “Exited.” We handpicked several variables (for their relevance to customer behavior and financial engagement) to use to avoid overfitting and multicollinearity: *Age*, *NumOfProducts*, *IsActiveMember*, *EstimatedSalary*, *Balance*, *CreditScore*, *BalanceToSalary*, *TenureGroup* (categorical), and *LoyaltyScore*

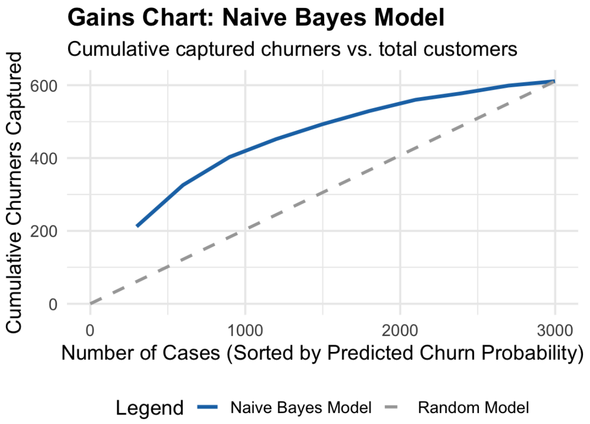
Rish (2001) demonstrated that the Naive Bayes model performs best in two cases: (1) when features are completely independent, and (2) when features are functionally dependent, meaning one feature can be somewhat determined by the other (Rish, n.d., 46). By contrast, model performance suffers when the dataset falls somewhere in the middle. In our project’s customer churn dataset, data feature dependencies and non-uniform distributions (i.e., *BalanceToSalary*, *IsActiveMember*) may have violated these optimal Naive Bayes conditions, which explains the model’s comparatively subpar performance (see Appendix A, Figure A9).

After training, the model was applied to a holdout set. Its classification performance was evaluated using several key metrics:

- Accuracy: 79.96%
- Kappa: 0.0513
- Sensitivity (True Positive Rate): 99.41%
- Specificity (True Negative Rate): 3.93%
- AUC (Area Under the Curve): ~0.52

The model achieved very high sensitivity, correctly identifying nearly all non-churned customers. However, it displayed noticeably low specificity, misclassifying the majority of customers who churned. This suggests a heavy bias toward predicting the dominant class (non-churners), potentially due to an imbalance in the dataset or other issues requiring further exploration. Given the very low specificity and lower accuracy than other models, we would not suggest implementing or deploying a Naive Bayes model to accurately determine whether or not a customer exited the bank system.





The ROC curve displayed on the left above arcs only slightly above the diagonal reference line, with an estimated AUC near 0.52. This implies that the model performs only marginally better than random guessing in distinguishing churners from non-churners. Again, confirming our decision to discard Naive Bayes as a usable model to predict bank churn. The gains chart on the right shows the model struggles to rank customers by churn risk. The cumulative capture of true churners is only slightly better than the baseline, which is what a random model would predict. Furthermore, the confusion matrix shows the severe skew in predictions, with the model overwhelmingly favoring the “no churn” class (see Appendix A, Figure A10).

While Naive Bayes is a simple and efficient model, its performance in this project was not ideal. The model's decent accuracy (.79) is misleading due to the underlying class imbalance. The low specificity tells us it fails to flag customers at risk of churning, which is the goal of this research. The findings from Rish (2001) align with our results, indicating again that Naive Bayes struggles when feature distributions are not fully independent and not functionally related (Rish, n.d., 46). From a business perspective, this could lead to missed opportunities for retention strategies to prevent churn. Improvements could include rebalancing the data, applying further feature selection or transformation, or incorporating hybrid models that address the assumption of feature independence before considering utilizing a Naive Bayes model for deployment.

Discussion:

Based on the results from our model evaluations, the k-Nearest Neighbors (kNN) and CART models performed most effectively in predicting customer churn, each achieving an accuracy above 85% and well-balanced AUC scores. These findings suggest simpler, interpretable models such as decision trees and instance-based learning capture complex customer behavior patterns more effectively than linear approaches (Linear Regression, Logistic Regression, and Naive Bayes). While Logistic Regression and Naive Bayes struggled with sensitivity and specificity, respectively, the kNN model, in particular, demonstrated high sensitivity (97.40%), indicating strong potential in flagging likely churners, an essential need for bank retention strategies.

From a practical standpoint, the results emphasize the critical role of variables such as tenure, account balance, and customer activity in predicting churn. These findings confirm industry intuition: customers with short tenure or low engagement are most likely to leave.

Models that prioritize these predictors can help banks intervene earlier, for example, by offering onboarding incentives or personalized service outreach to at-risk customers.

Theoretically, the study supports the importance of combining behavioral and demographic data to improve model performance. Unlike models limited to static traits (e.g., age or geography), models incorporating dynamic indicators like activity level and loyalty scores offer better insights. This aligns with previous research by Verbeke et al. (2012) and Khandani et al. (2010), which supports the inclusion of engagement metrics in churn prediction.

Researchers can benefit from our results by focusing future work on hybrid models that balance interpretability with performance, such as random forests or other methods. Banks should adopt models with high sensitivity to reduce customer loss, even if specificity is modest. Ultimately, our analysis shows that predictive modeling, when combined with thoughtful feature engineering, can meaningfully guide customer retention strategies and improve decision-making in competitive banking environments.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1986). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Halder, R. K., Roy, A., Rahman, M. M., & Khatun, M. (2024). *Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications*. *Journal of Big Data, 11*(1). https://doi.org/10.1186/s40537-024-00973-y

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). *Consumer credit-risk models via machine-learning algorithms*. Journal of Banking & Finance, 34(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Reichheld, F. F., & Sasser, W. E., Jr. (1990). Zero defections: Quality comes to services. *Harvard Business Review, 68*(5), 105–111. https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Rish, I. (n.d.). An empirical study of the naive Bayes classifier. *T.J. Watson Research Center*, 41-47. https://faculty.cc.gatech.edu/~isbell/reading/papers/Rish.pdf

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2012). *Building comprehensible customer churn prediction models with advanced rule induction techniques*. European Journal of Operational Research, 218(1), 211–229.

Image 1

Deutsche Bank exterior main office.

Reprinted from *Deutsche Bank exterior main office*, by A. Go, n.d., Adobe Stock (https://stock.adobe.com/images/deutsche-bank-exterior-main-office/563221035). Copyright [n.d.] by Adobe Stock. Used under license.

Appendix A

Appendix A includes graphs illustrating model performance and key feature impacts across multiple algorithms: Logistic Regression, K-Nearest Neighbors, CART, Linear Regression, and Naive Bayes. These visualizations highlight differences in predictive accuracy, sensitivity, and feature influence among the models used.

Figure A1.

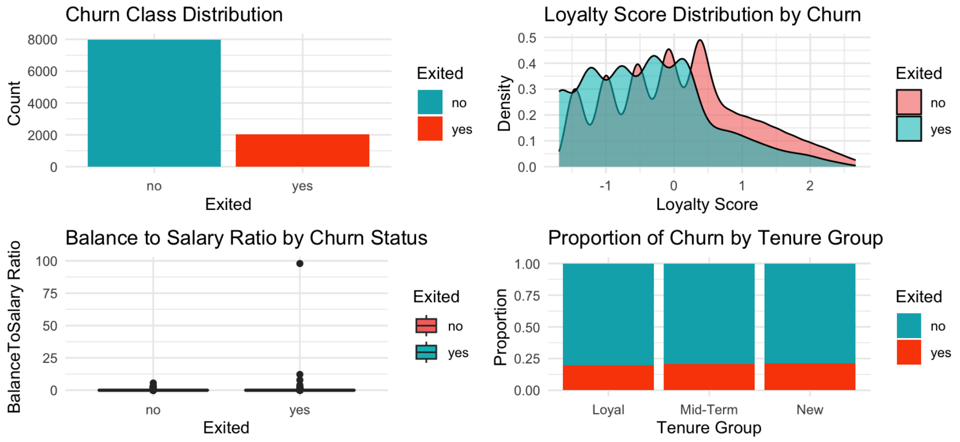


Figure A2.

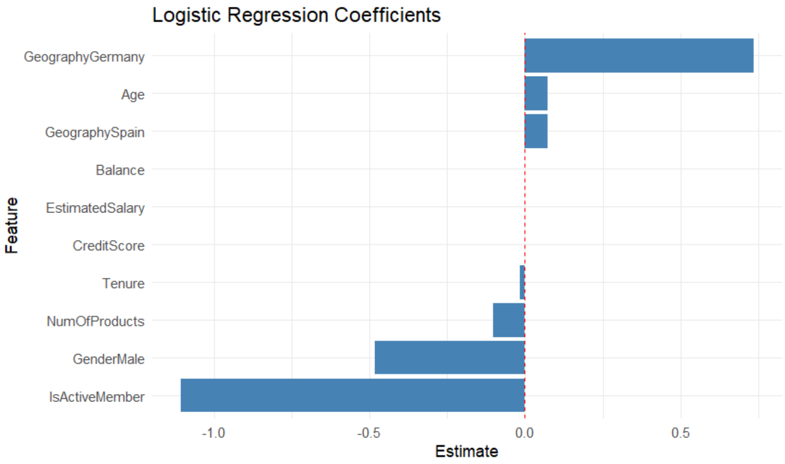


Figure A3.

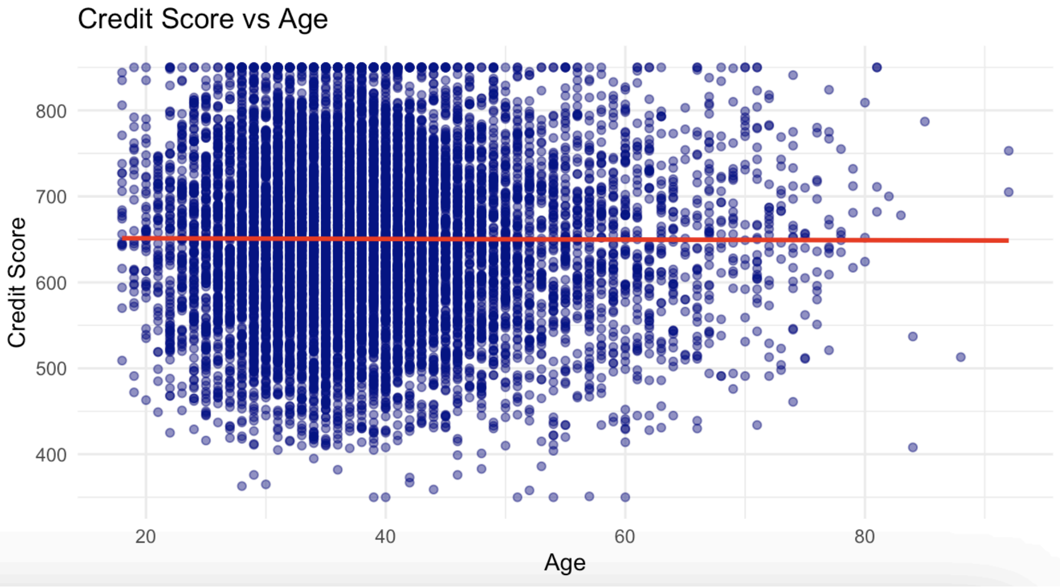


Figure A4.

CART Training performance (Confusion Matrix and Statistics)

	Reference	
Prediction	no	yes
no	5340	791
yes	234	636

And corresponding statistics:

```
Accuracy : 0.854

                 95% CI : (0.845, 0.862)

    No Information Rate : 0.796

    P-Value [Acc > NIR] : <0.0000000000000002


                  Kappa : 0.472
```

```
 Mcnemar's Test P-Value : <0.0000000000000002


            Sensitivity : 0.958

            Specificity : 0.446

         Pos Pred Value : 0.871

         Neg Pred Value : 0.731

             Prevalence : 0.796

         Detection Rate : 0.763

   Detection Prevalence : 0.876

      Balanced Accuracy : 0.702


       'Positive' Class : no
```

Figure A5.

CART holdout performance (Confusion Matrix and Statistics)

	Reference	
Prediction	no	yes
no	2284	337
yes	104	274

And corresponding statistics:

```
Accuracy : 0.853

                 95% CI : (0.84, 0.865)

    No Information Rate : 0.796
    P-Value [Acc > NIR] : 0.000000000000000698

                  Kappa : 0.472

 Mcnemar's Test P-Value : < 0.0000000000000002

            Sensitivity : 0.956
            Specificity : 0.448
         Pos Pred Value : 0.871
         Neg Pred Value : 0.725
             Prevalence : 0.796
         Detection Rate : 0.762
   Detection Prevalence : 0.874
      Balanced Accuracy : 0.702

       'Positive' Class : no
```

Figure A6.

k	ROC	Sens	Spec
1	0.6723572	0.8737010	0.4710135
2	0.7279151	0.8754947	0.4689353
3	0.7605977	0.9300380	0.4576480
4	0.7812109	0.9280635	0.4464198
5	0.7935991	0.9454663	0.4408598
6	0.8027954	0.9483382	0.4394169
7	0.8046425	0.9592810	0.4163006

```
 8  0.8053554  0.9585632  0.4127844

 9  0.8102391  0.9603553  0.4057963

10  0.8138336  0.9608930  0.4057815

11  0.8156765  0.9641226  0.3952970

12  0.8178659  0.9668140  0.3889885

13  0.8194194  0.9680685  0.3868955

14  0.8213707  0.9686065  0.3819856

15  0.8228667  0.9696827  0.3770954

16  0.8245484  0.9707596  0.3743081
```

Figure A7.

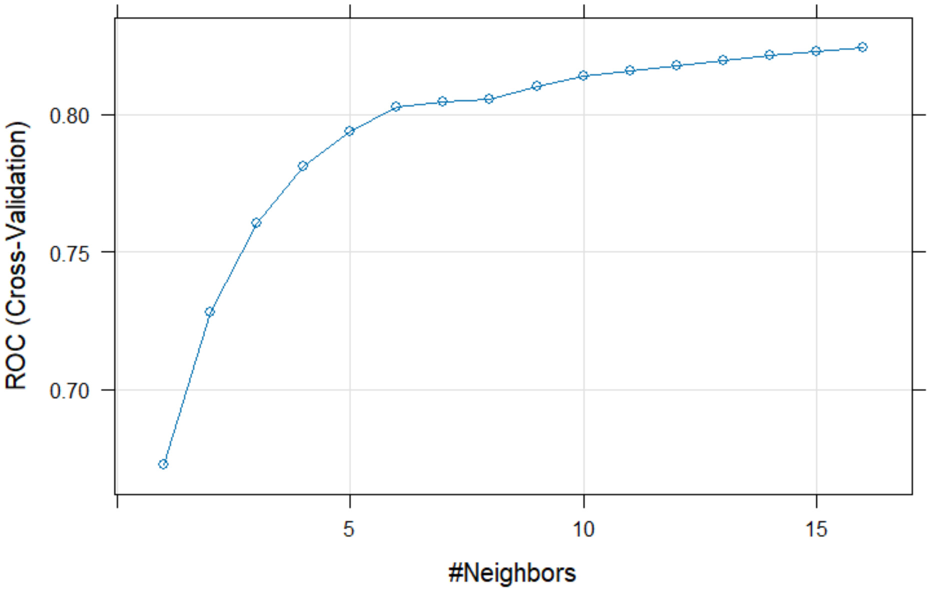


Figure A8.

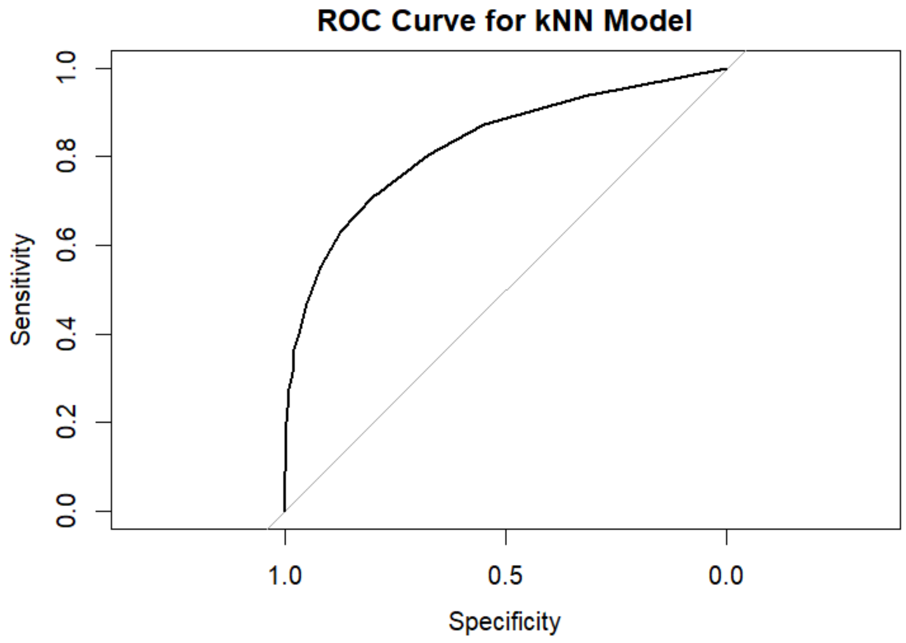
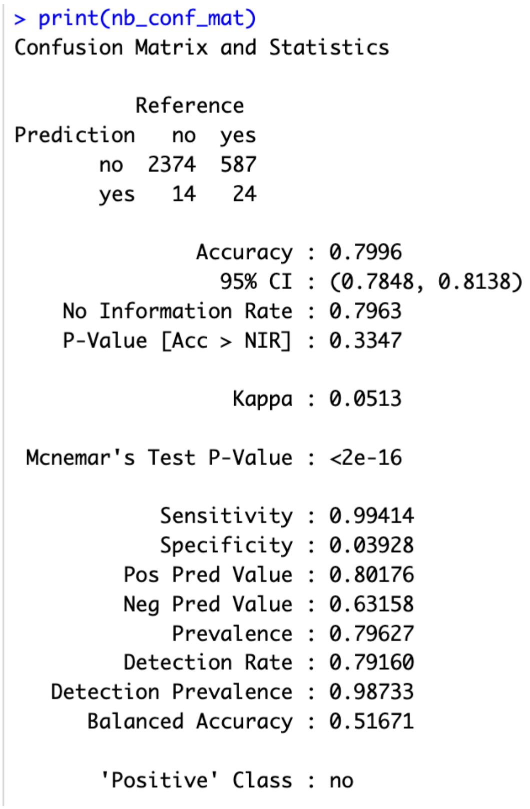
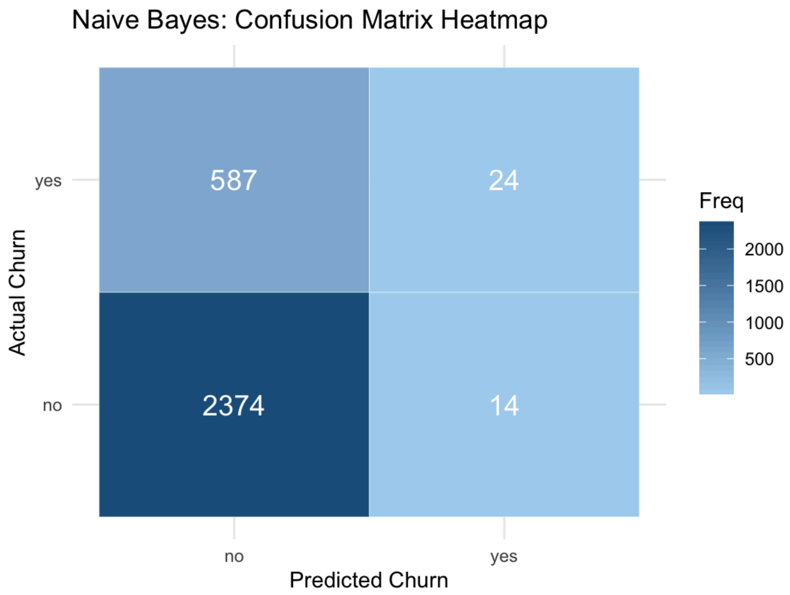


Figure A9.



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

          Reference
Prediction   no  yes
       no  2374  587
       yes   14   24

               Accuracy : 0.7996
                 95% CI : (0.7848, 0.8138)
    No Information Rate : 0.7963
    P-Value [Acc > NIR] : 0.3347

                  Kappa : 0.0513

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99414
            Specificity : 0.03928
         Pos Pred Value : 0.80176
         Neg Pred Value : 0.63158
             Prevalence : 0.79627
         Detection Rate : 0.79160
   Detection Prevalence : 0.98733
      Balanced Accuracy : 0.51671

       'Positive' Class : no
```

Figure A10.


Naive Bayes: Confusion Matrix Heatmap
yes
no
Actual Churn
587
24
2374
14
no
yes
Predicted Churn
Freq
2000
1500
1000
500