Using Machine Learning Classification Techniques to Predict Bank Customer Churn

Introduction:

In the dynamic landscape of banking and financial services, maintaining a loyal customer base is paramount for sustained growth and competitiveness. However, the phenomenon of customer churn, where clients discontinue their relationship with a bank, presents a formidable challenge, often resulting in revenue loss and diminished market share. Addressing this challenge requires proactive measures and predictive analytics emerges as a powerful tool in this endeavor.

This project focuses on predicting bank customer churn using the R programming language, leveraging historical customer data to develop a robust predictive model. The primary objective is to identify customers who are likely to churn in the near future, enabling banks to take preemptive actions to retain them. By harnessing the power of data science and machine learning, we aim to uncover intricate patterns and correlations within the data that can provide valuable insights into the drivers of churn.

The significance of this project lies in its potential to revolutionize customer relationship management within the banking industry. By accurately predicting churn events, banks can tailor their retention strategies to address the specific needs and preferences of at-risk customers, thereby improving customer satisfaction and loyalty. Moreover, proactive churn management can lead to substantial cost savings compared to reactive approaches.

This project also underscores the practical application of data science techniques in solving real-world business challenges. Through a systematic approach encompassing data collection, preprocessing, feature engineering, model development, and evaluation, we aim to showcase the effectiveness of predictive analytics in enhancing decision-making processes within the banking sector.

Additionally, we will present the results obtained from our analysis, providing insights into the key factors influencing bank customer churn. By offering actionable recommendations based on our findings, we aim to assist banking stakeholders in devising effective strategies for customer retention in an increasingly competitive market environment.

Problem Statement:

Develop a predictive model using R to anticipate bank customer churn, leveraging historical data to identify at-risk individuals and enable proactive retention strategies, ultimately enhancing customer retention and reducing revenue loss.

Data Description:

The dataset provided contains customer information from a financial institution or a bank. It is a tabular dataset, where each row represents a customer, and each column represents a specific attribute or feature related to the customer. The dataset consists of the following columns:

1. `customer_id`: This column contains a unique identifier for each customer. It is an integer or a string data type.

2. `credit_score`: This column represents the credit score of the customer, which is a numerical value typically ranging from 300 to 850 or 500 to 900, depending on the credit scoring model used. It is an integer data type.

3. `country`: This column indicates the country where the customer resides. It is a categorical string data type.

4. `gender`: This column represents the gender of the customer, which is a binary categorical variable, either "Male" or "Female". It is a string data type.

5. `age`: This column contains the age of the customer, which is a numerical value representing the number of years. It is an integer data type.

6. `tenure`: This column indicates the number of years the customer has been with the institution. It is a numerical value representing the duration in years. It is an integer data type.

7. `balance`: This column represents the account balance of the customer. It is a numerical value representing the amount of money. It is a float or decimal data type.

8. `products_number`: This column indicates the number of products the customer has with the institution, such as savings accounts, loans, or investment products. It is an integer data type.

9. `credit_card`: This column is a binary categorical variable indicating whether the customer has a credit card with the institution or not. It is encoded as 0 or 1, or "Yes" or "No". It is a string or integer data type.

10. `active_member`: This column is a binary categorical variable indicating whether the

customer is an active member of the institution or not. It is encoded as 0 or 1, or "Yes" or "No". It is a string or integer data type.

11. `estimated_salary`: This column represents the estimated salary of the customer. It is a numerical value representing the amount of money. It is a float or decimal data type.

12. `churn`: This column is a binary categorical variable indicating whether the customer has left the institution or not. It is encoded as 0 or 1, or "Yes" or "No". It is a string or integer data type.

This dataset provides valuable insights into customer demographics, financial behavior, and potential risk factors.

Exploratory Data Analysis:

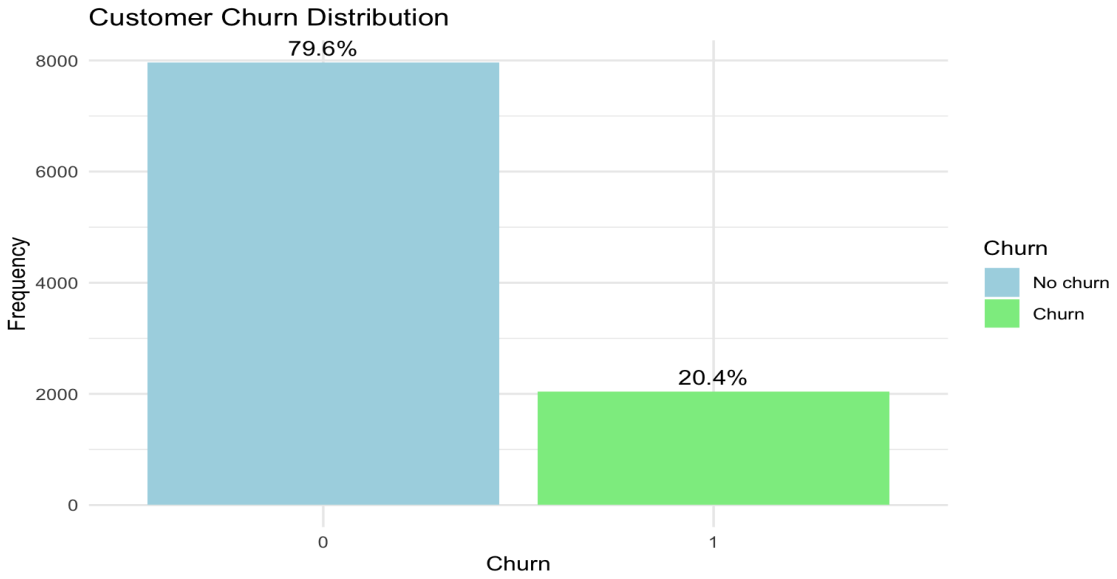

Figure 1: Customer Churn Distribution

Observations:-
There's a class imbalance with about 79% of the customers not churning, while around 20% did churn.

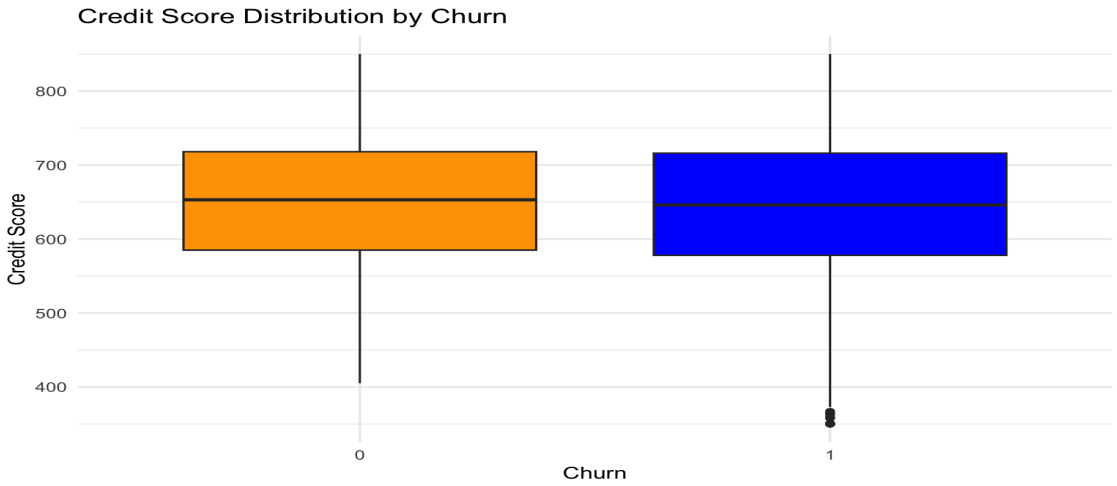

Figure 2: Credit Score Distribution by Churn

Observations:
The box plot suggests a possible correlation between lower credit scores and customer churn. The box (representing the middle 50% of data) for churned customers is positioned lower than the non-churned customers' box.

Figure 3: Balance vs. Estimated Salary

Observations:
The scatterplot shows an upward trend. As estimated salary increases, the balance also tends to increase. This suggests that people with higher estimated salaries tend to have larger balances in their accounts.

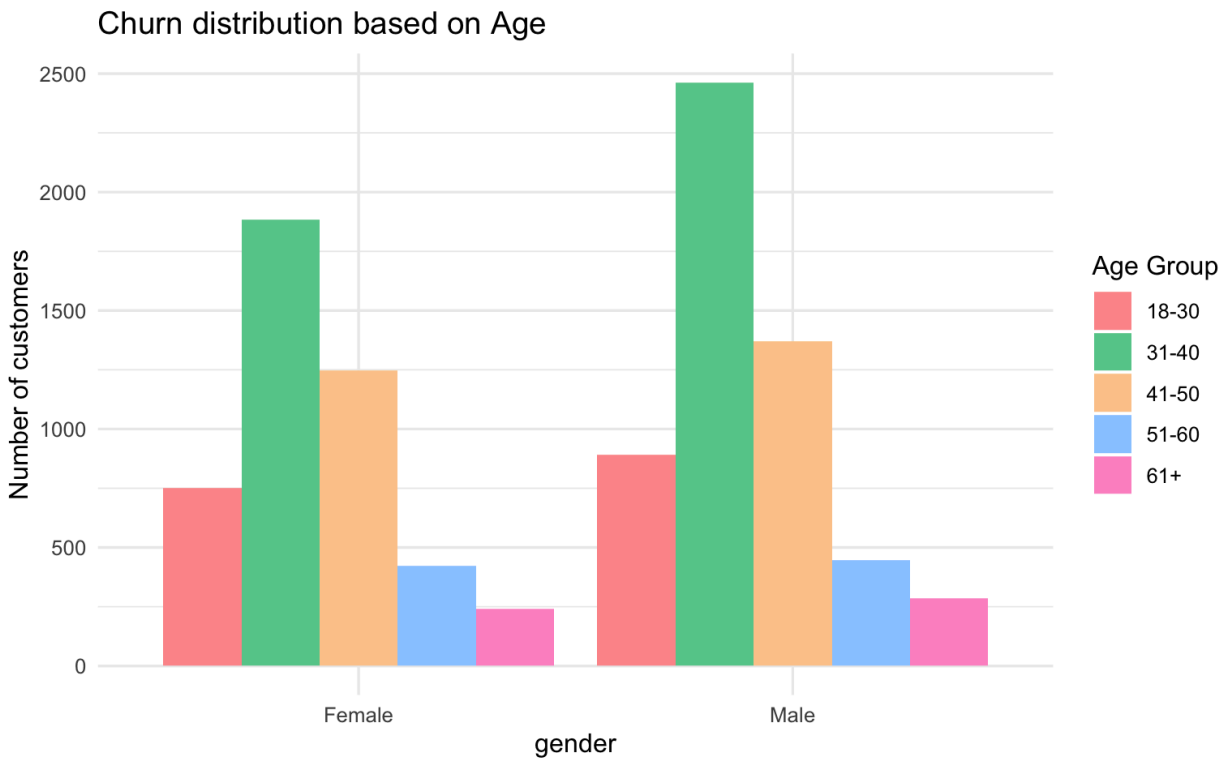

Figure 4: Distribution of Client Ages

Observations:
Majority of the customers are between the ages of 25 to 50. Most accounts are owned by people with less than 50 years. Especially 31-40 years Customers of ages 41-50 have the highest churn rate.

Methodology:

Data Cleaning:

First we made sure there were no missing values for the data set and once that was checked and no missing values were found then we started discussing variable selection for the models. For each of our models, we decided that it would be best to include every variable (categorical and numeric) in our models. This is because the dataset ended up having a smaller amount of variables to train with as seen in the Data Description. Since we are trying to predict bank customer churn, we felt that the best way to predict it is by using all of the variables except customer_id. We removed customer_id because it was

just a unique indicator for the customer but not anything meaningful in the dataset.. Since we are dealing with a binary variable in churn (0 and 1), we were able to use the numbers for LDA, QDA, and both Logistic Regression models, but did have to transform that column into factors for the Naive Bayes model. Finally, we used a random 80/20 train/test split for our data which we trained and tested our models with.

Methods Used:

1. Linear Discriminant Analysis (LDA)-In the realm of machine learning, linear discriminant analysis (LDA) is a statistical method used for both dimensionality reduction and classification. The goal of LDA is to identify the linear feature combinations that most effectively divide an item or event into two or more classes.LDA performs two main functions on the dataset that is Class separation and Dimensionality reduction. LDA is used for Feature Extraction,Classification,Pattern Recognition, Signal Processing and Medical Diagnostics.The underlying premise of LDA is that distinct classes produce data using Gaussian distributions, each class having a different mean but a same covariance matrix. The linear decision boundary can be found by LDA under this assumption. LDA might not work as well, though, if these presumptions are false (for example, the data is not regularly distributed).

2. Quadratic Discriminant Analysis (QDA)-Similar to Linear Discriminant Analysis(LDA), Quadratic Discriminant Analysis (QDA) is another statistical technique for classification; however, it differs significantly in the assumptions it makes about the data. QDA lets each class have its own covariance matrix, whereas LDA assumes that all classes have the same covariance matrix. Because of this, QDA is more adaptable when working with complicated datasets where the equal covariances assumption is broken.The two primary purposes of QDA are classification and discrimination based on class. Among the applications of QDA include machine learning, pattern recognition, and classification in complicated situations.QDA is a strong tool for classification in increasingly complicated settings since it can model each class with its own covariance; however, careful management is needed to prevent overfitting and guarantee robust results.

3. Logistic Regression (Normal)-Logistic Regression is a statistical method used for binary classification that can be extended to multiclass classification via techniques such as one-vs-rest (OvR) or multinomial logistic regression. It models the probabilities of the default classes of one or more independent variables, and is

particularly popular for its simplicity and efficiency in binary outcomes.A binary dependent variable is modeled using a logistic function in logistic regression. Applications for logistic regression can be found in the fields of medicine, marketing, social sciences, engineering, and credit scoring. Because of its versatility and ease of use, logistic regression is still a preferred option in many situations, especially when the problem involves binary classification and the underlying relationships are roughly linear.

4. Logistic Regression (Polynomial)-Logistic Polynomial Regression adds polynomial terms from the input features to the basic logistic regression model. In situations when there is a nonlinear relationship between the variables, this method can greatly enhance model performance by enabling the model to capture more intricate, nonlinear interactions between the independent and dependent variables.In polynomial logistic regression, a polynomial curve is fitted to the data as opposed to a linear line, as in simple logistic regression. Higher-degree terms of the predictor variables are incorporated into the logistic regression equation to accomplish this.Applications of polynomial logistic regression include marketing analytics, healthcare, economic forecasting, and environmental modeling.When linear assumptions are too restrictive, polynomial logistic regression is a useful extension; however, rigorous model design and validation are essential to prevent overfitting and guarantee the predictive capability of the model.

5. Naive Bayes- The method used by naive Bayes classifiers is to forecast which class has the highest posterior probability by first computing the posterior probability of each class using the Bayes theorem. In order to forecast which class has the highest posterior probability, naive Bayes classifiers first compute the posterior probability of each class using the Bayes theorem. Among the uses of Naive Bayes include spam filtering, text classification, medical diagnosis, and weather prediction.Naive Bayes is a fantastic place to start when it comes to classification tasks because of its simplicity and the strong assumptions it makes. It frequently performs unexpectedly well.

Model Building and Results Calculation:

To build our models, we used a variety of packages in R to get the model functions that

we could then train our models with. All of our models were trained on the train set and then compared with the test set to then get the Test MSE, Accuracy, Sensitivity and Specificity.

These were how our models were built:

1. Linear Discriminant Analysis (LDA)- Our linear discriminant analysis model was made by using the “lda” function from the “MASS” package. It compared the variable “churn” to every other variable using the train dataset. Afterwards, from the model, we were able to extract the predictions using the predict function and then comparing them to the test data to get the Accuracy. The Sensitivity and Specificity were calculated by using a predictions vs test value data frame.

2. Quadratic Discriminant Analysis (QDA)- Our quadratic discriminant analysis model was made by using the “qda” function also from the “MASS” package. It compared the variable “churn” to every other variable using the train dataset. Afterwards, from the model, we were able to extract the predictions using the predict function and then comparing them to the test data to get the Accuracy. The Sensitivity and Specificity were calculated by using a predictions vs test value data frame.

3. Logistic Regression (Normal)- Our normal logistic regression model was made by using the “glm” function from the “stats” package from base R. It compared the variable “churn” to every other variable using the train dataset. To make the model logistic, we specified the family as “binomial”. Afterwards, from the model, we were able to extract the predictions using the predict function and then comparing them to the test data to get the Accuracy. The Sensitivity and Specificity were calculated by using a predictions vs test value data frame.

4. Logistic Regression (Polynomial)- Our normal logistic regression model was made by using the “glm” function from the “stats” package from base R. To make the model logistic, we specified the family as “binomial”. It compared the variable “churn” to every other variable (besides age) normally using the train dataset. However, the age variable we decided could be interesting to look into from a polynomial standpoint. Thus, we decided to use cross validation to see which models had the lowest test MSE and then compared the models using ANOVA to get what we considered to be the best models from both standpoints. The polynomial model versus the test MSE graph is shown below.

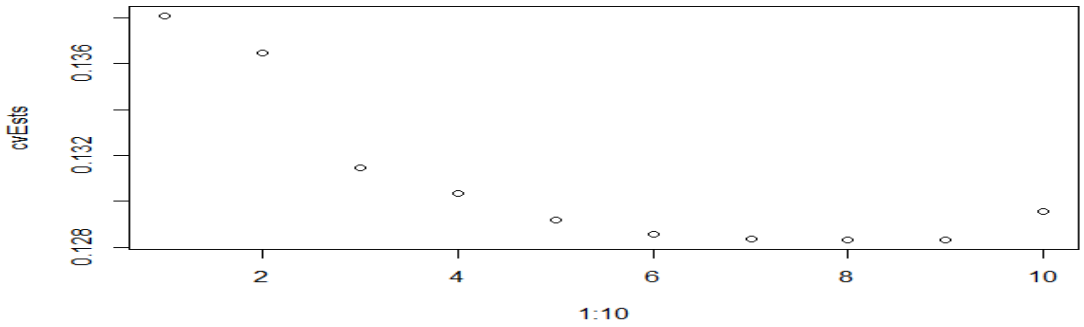

Figure 5: Cross validation estimates for a range of polynomial values for variable "age.

The model we chose to go with ended up choosing was polynomial 7 as it was chosen as one of the best models from both methods.
Afterwards, from the model, we were able to extract the predictions using the predict function and then comparing them to the test data to get the Accuracy. The Sensitivity and Specificity were calculated by using a predictions vs test value data frame.

5. Naive Bayes- Our Naive Bayes model was created by using the "naive_bayes" function from the "naivebayes" package. For this function to work, we had to convert the churn variable into a factor since it originally was numeric. Then it compared the variable "churn" to every other variable using the train dataset. Afterwards, from the model, we were able to extract the predictions using the predict function and then comparing them to the test data to get the Accuracy. The Sensitivity and Specificity were calculated by using a predictions vs test value confusion matrix.

Model Comparison Methods and Goal:

To compare model results we decided to use the Test Error and Accuracy of the model to determine the best model to use. The only model with an exception to this was the Logistic Regression with the Polynomial variable which used K-10 fold cross validation to choose the best model to compare test error with. For the modeling we used a random seed of 42 since we had random sampling and wanted to make sure our results were reproducible. Our goal of the modeling was to see which model had the best results at predicting customer churn in hopes to see which variables and types of customers that stayed or left. There were a variety of ways to build and choose these models but we wanted to limit our scope to classification as the goal of the project to provide a basic framework for future work on this data.

Results:

Model	Accuracy	Test Error	Sensitivity	Specificity
Linear Discriminant Analysis	81.42%	0.1858	23.59%	95.55%
Quadratic Discriminant Analysis	**84.54%**	**0.1546**	**40.26%**	**95.36%**
Logistic Regression (Normal)	81.87%	0.1813	21.54%	96.62%
Logistic Regression (Polynomial 7)	84.24%	0.1576	34.87%	96.30%
Naive Bayes	84.09%	0.1591	30.77%	97.12%

Table 1: Resulting metrics of the models

The results presented in the table showcase the performance metrics of different classification models applied to a particular dataset. Each model—Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression (with Normal and Polynomial 7 degrees), and Naive Bayes—has been evaluated based on several key metrics: accuracy, test error, sensitivity, and specificity.

Accuracy represents the overall correctness of the model's predictions on the test dataset. Quadratic Discriminant Analysis achieved the highest accuracy at 84.54%, closely followed by Logistic Regression (Polynomial 7) at 84.24%. However, accuracy alone does not provide a complete picture, as it doesn't account for the balance between correctly identifying positive cases (sensitivity) and negative cases (specificity).

Sensitivity measures the ability of the model to correctly identify positive cases (e.g., correctly predicting disease presence), while specificity measures the ability to correctly identify negative cases (e.g., correctly predicting disease absence). Models with high specificity (e.g., Logistic Regression with Normal) are better at avoiding false positives, while those with high sensitivity (e.g., Naive Bayes) are better at catching positive cases but might have a higher false positive rate.

The test error, which is simply (1 - accuracy), provides another angle on model performance, showing the proportion of incorrect predictions made by each model.

Lower test error values indicate better performance.

Overall, each model has its strengths and weaknesses based on these metrics. For instance, Quadratic Discriminant Analysis shows high accuracy but relatively lower sensitivity compared to Logistic Regression (Polynomial 7), which excels in sensitivity and specificity. Understanding these trade-offs is crucial in selecting the most appropriate model based on the specific objectives and constraints of the problem at hand.

Discussion:

The results we saw in the table happened because different types of models handle data in unique ways. Some models, like Quadratic Discriminant Analysis and Logistic Regression with a high-degree polynomial, are more flexible and can fit complex patterns in the data well.

However, this flexibility can sometimes lead to overfitting, where the model is too closely tuned to the training data and doesn't perform as well on new data.

Other models, like Linear Discriminant Analysis and Naive Bayes, are simpler and make certain assumptions about how the data is distributed. For example, Naive Bayes assumes that the features are independent of each other, which might not always hold true in real-world data.

For future research, we can improve these results by working on the following:-

1. Feature Improvement: We can try refining the features used by the models. This could involve adding new features or transforming existing ones to better match the assumptions of each model.

2. Model Adjustment: Tuning the settings of the models, like changing the degree of polynomial in Logistic Regression or adjusting regularization parameters, could help improve their performance.

3. Combining Models: Instead of relying on a single model, we can explore using multiple models together (ensemble methods) to get more reliable predictions by leveraging the strengths of each.

Conclusion:

In summary, choosing the right model depends on the goals of the analysis and the characteristics of the dataset. Continuous efforts to refine models and understand the data better will lead to improved predictive performance and deeper insights into the underlying patterns in the data.

References:

1. Khan, A. A., Jamwal, S., & Sepehri, M. M. (2010). Applying Data Mining to Customer Churn Prediction in an Internet Service Provider. International Journal of Computer Applications, 9(7), 8–14. https://doi.org/10.5120/1400-18894

2. Sharma, A., & Kumar Panigrahi, P. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. International Journal of Computer Applications, 27(11), 26–31. https://doi.org/10.5120/3344-4605

3. https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset?resource=download