Project Report

Bank Customer Churn Analysis

Apoorv Vats

vats.ap@northeastern.edu

Submission Date: 04-21-2023

Table of Contents

Problem Setting

Problem Definition

Data Sources

Data Description

Exploring the Dataset and Cleaning

Dimension reduction and variable selection

Data Imbalance

Data Mining Models

Candidate models for the Project

Model Evaluation

Conclusion

Learnings from the Project

Future Scope

References

Problem Setting:

Knowing which clients are most likely to leave or stop using the bank service is known as customer churn prediction. This prognosis is significant for a lot of businesses. This is because it is frequently more expensive to acquire new clients than to keep old ones. When a customer is at risk of leaving, we need to know precisely what marketing efforts to make with them to increase the possibility that they will stay.

Problem Definition:

The aim of the project is to estimate the churn rate of a bank. A dataset that contains some customers' details who are withdrawing their accounts from the bank due to some loss andother issues, with the help of this data we try to analyze and maintain accuracy. It is important for a bank to analyze it and take the required measures.

Data Sources:

The data is taken from Kaggle.com.

https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset

Data Description:

Collect data on customer demographics, transaction history, and feedback. This dataset has 10,000 rows and 12 columns of data. It includes numerical predictors like credit score, age, tenure, balance, estimated salary, and categorical predictors like gender, country, age, active member, which will be used to predict the target variable churn (0 or 1). Cleaning and pre-processing the data to remove any inconsistencies or missing values is important to ensure that the data is ready for analysis. Create new features or variables that can be used to predict customer churn. These can include variablessuch as

the number of transactions per month, the length of time a customer has been with the bank, or the number of products a customer has with the bank.

Exploring the Dataset and Cleaning:

- There are numerical and categorical values in the dataset. Mean, median, and mode of the data are calculated for the numerical data. There are 10000 rows and 12 columns inthe dataset.
- The information about the distributions of the variables in the dataset, such as count, mean, standard deviation, etc.

Figure 1: Data description

	credit_score	age	tenure	balance
count	10000.000000	10000.000000	10000.000000	10000.000000
mean	650.528800	38.921800	5.012800	76485.889288
std	96.653299	10.487806	2.892174	62397.405202
min	350.000000	18.000000	0.000000	0.000000
25%	584.000000	32.000000	3.000000	0.000000
50%	652.000000	37.000000	5.000000	97198.540000
75%	718.000000	44.000000	7.000000	127644.240000
max	850.000000	92.000000	10.000000	250898.090000

	products_number	credit_card	active_member	estimated_salary
count	10000.000000	10000.00000	10000.000000	10000.000000
mean	1.530200	0.70550	0.515100	100090.239881
std	0.581654	0.45584	0.499797	57510.492818
min	1.000000	0.00000	0.000000	11.580000
25%	1.000000	0.00000	0.000000	51002.110000
50%	1.000000	1.00000	1.000000	100193.915000
75%	2.000000	1.00000	1.000000	149388.247500
max	4.000000	1.00000	1.000000	199992.480000

	churn
count	10000.000000
mean	0.203700
std	0.402769
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

- Box plots are used to display the distribution of a set of numerical data. A

box plot consists of median, box, whiskers, outliers, and notches.Here, we can conclude that there are no outliers in customer_id, tenure, balance, credit_card, active_member, estimated_salary. The data points that fall outside the whiskers are considered as outliers. Here, we have outliers in credit_score, age, and products_number. These can be potential outliers and to be investigated further.

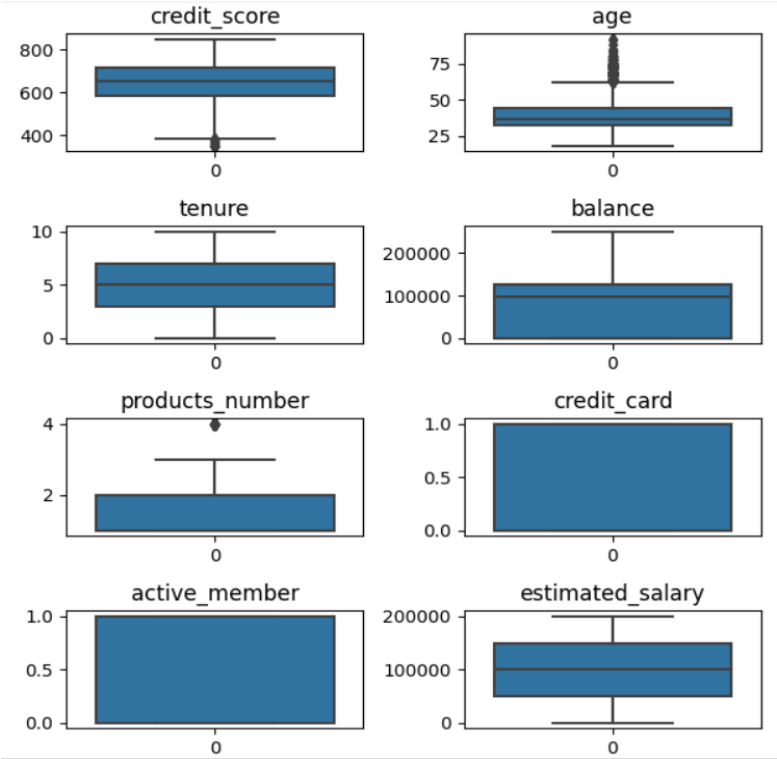
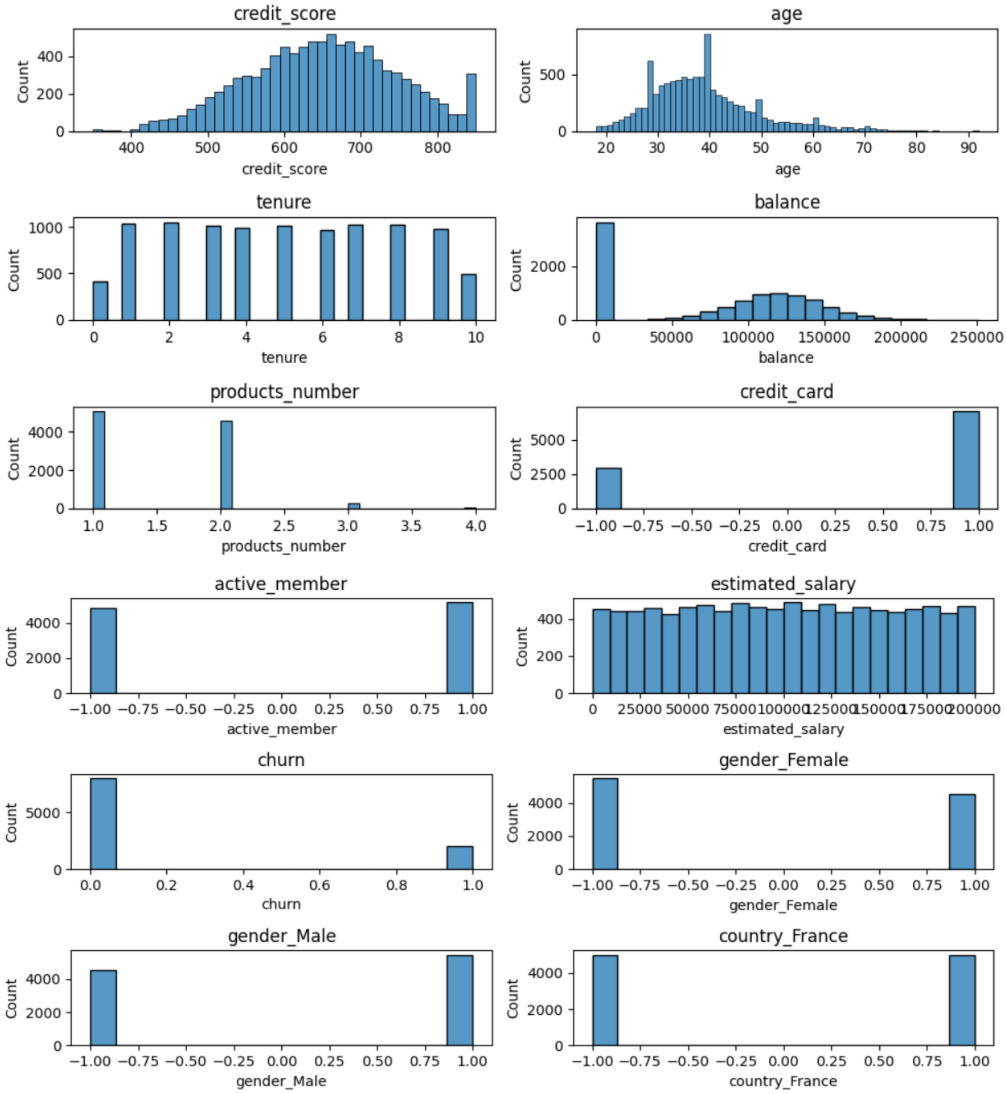


Figure 2: Box plot

- Histogram plot to check data distribution: A histogram plot is used to visualize the distribution of a numerical variable. In the above plot, data distribution of each attribute is used to understand the shape of the distribution. The x- axis represents the range of values for the variable, and the y axis represents the frequency of the observations. Here, the data distribution of customer idis

uniform. Similarly, we analyze the data distribution and remove the outliers existingin the data.



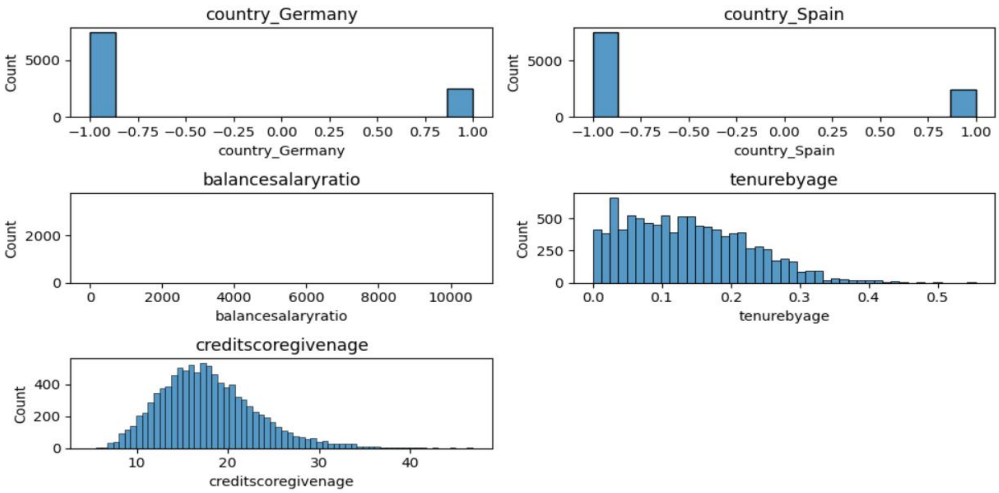


Figure 3: Histogram

- Correlation with Response variables: Correlation matrix: Correlation measures the degree of linear relationship between two variables, where a correlation coefficient value of -1 indicates a perfect negative correlation, a value of 0 indicates no correlation, and a value of +1 indicates a perfect positive correlation. Here,the value of credit_score, tenure, products_number, active_member is negative and indicates negative correlation.

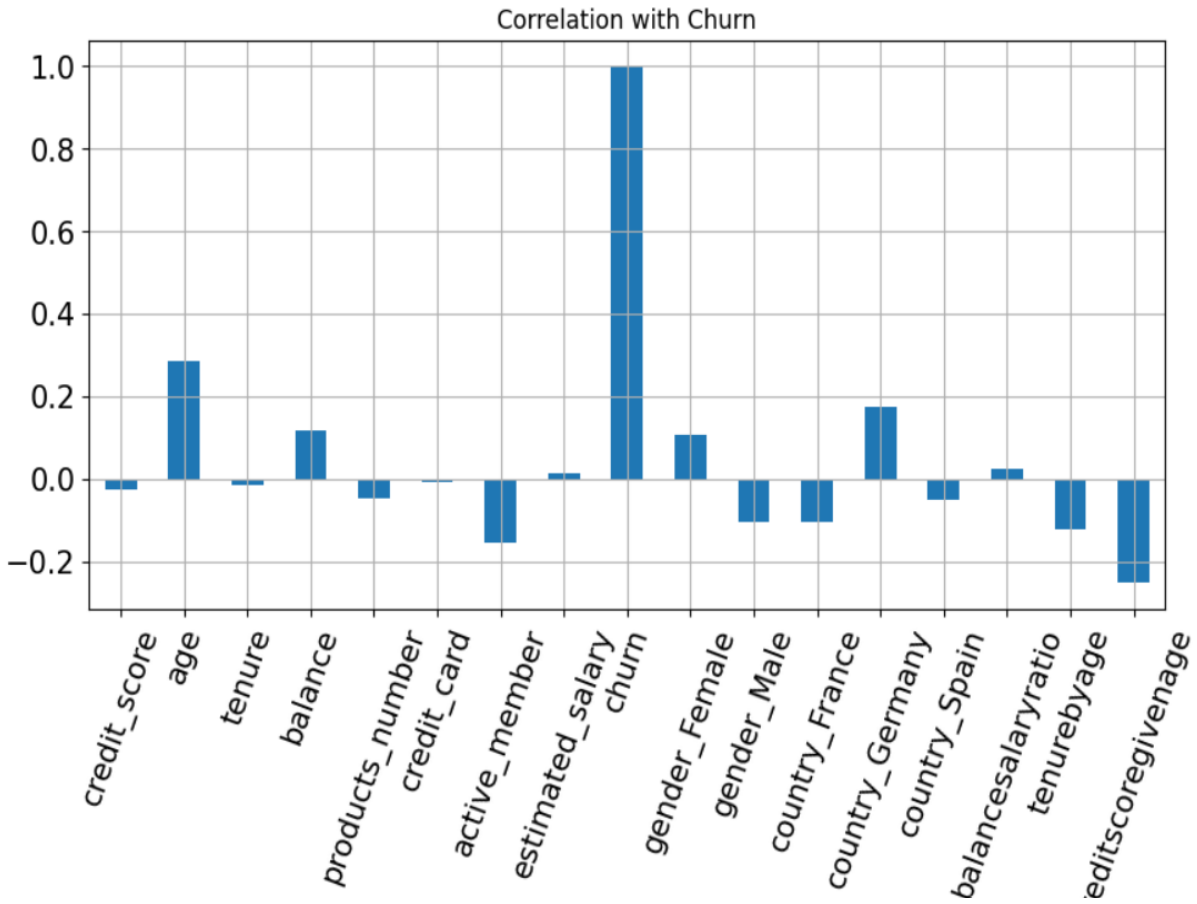


Figure 4: Correlation with response variables

- Heat maps are used to visualize and analyze data in a way that highlights patterns, trends, and correlations. We can analyze which columns are highly correlated by analyzing the plot.

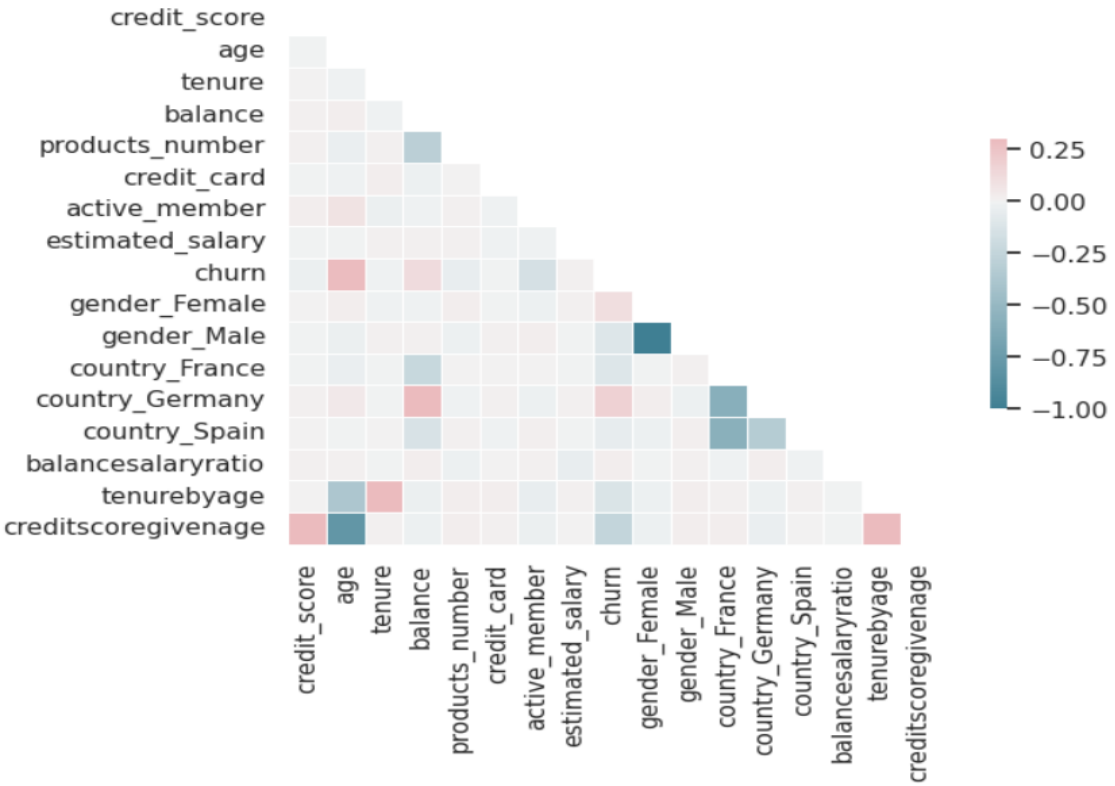


Figure 5: Heat Map

- Majority of the data is from customers in France.

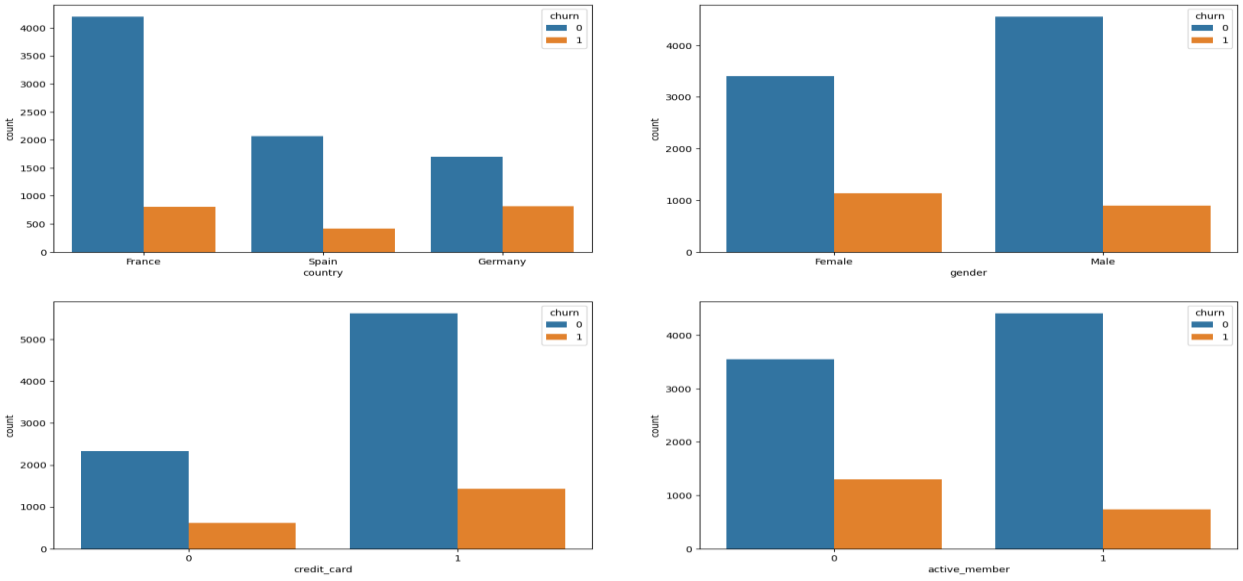


Figure 6. Count Plot

- To check whether the data is balanced:

  The bank has observed that more than 20% of their customers have churned, and the baseline model for predicting churn could be to assume that approx. 20% of the customers will churn. However, since identifying and retaining these 20% of customers is important to the bank, it is crucial to choose a model that can accurately predict this group with high precision, even if it may result in a lower accuracy.

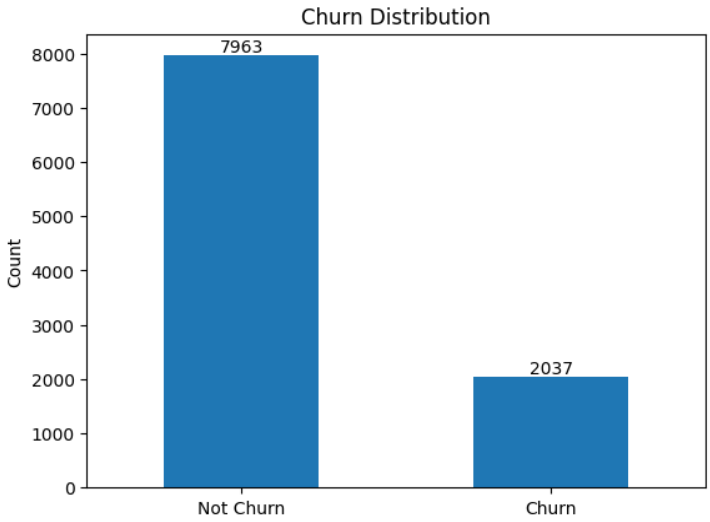


Figure 7. Data Imbalance

Dimension reduction and variable selection:

The column customer_id has no predictive significance, so it is removed from the dataframe.

The numerical and categorical variables of the data were separated. One Hot Encoding was performed on the categorical variables.

The numerical data was taken and MinMaxScaler was performed to standardize the data and furthermore to Principal Component Analysis was performed to reduce the dimensions, but MinMaxScaler did not affect the performance of the models and PCA significantly reduce the performance therefore these steps were removed from the project.

Data Imbalance:

The data is imbalanced roughly in 80-20 ratio.

Random under sampling technique is used in data preprocessing to balance the class distribution of a dataset.

Data Mining Models:

The following models have been used for the binary classification:

1. Logistic Regression
2. K-Nearest Neighbor
3. Naïve Bayes
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests

The following models have been compared based on the following metrics:

- **Accuracy**: It is calculated as the percentage of correctly classified instances over the total number of instances and is useful when the classes are balanced, and the costs of false positives and false negatives are equal.
- **Precision**: It measures the proportion of true positives among the predicted positives.
- **Recall**: It measures the proportion of true positives among the actual positives.
- **F1-Score**: It is the harmonic mean of precision and recall and provides a balanced evaluation of the model's performance.
- **AUC ROC** (Area Under the Receiver Operating Characteristic Curve) is a commonly used metric to evaluate the performance of a binary classification model. The ROC curve is a graphical representation of the model's true positive rate (sensitivity) against its false positive rate (1-specificity) at various classification thresholds. AUC ROC is the area under this curve and ranges from 0 to 1.

Candidate models for the Project:

Logistic Regression: A statistical algorithm used to predict the probability of an event occurring, given a set of input variables. In binary classification, logistic regression outputs a probability score that a given input belongs to a certain class (0 or 1), and the decision boundary is usually set at 0.5.

Advantages:

- Easy to implement and interpret.
- Can handle both linearly separable and non-linearly separable data with the rightchoice of features.
- Output probabilities that can be used to adjust the decision threshold.

Disadvantages:

- Assumes a linear relationship between the input features and the output.
- Can be sensitive to outliers and missing data.
- May not perform well when there are many irrelevant features.

K-Nearest Neighbor: A non-parametric algorithm that classifies data points based on the class labels of their k nearest neighbors in the feature space.

Advantages:

- Computationally efficient
- Requires relatively little training data.
- Popular for text classification tasks

Disadvantages:

- Assumes independence between features, which may not be true.
- Can be sensitive to irrelevant features or noisy data.
- May not perform well if the prior probabilities are significantly different from

reality.

Naïve Bayes: A probabilistic algorithm that is widely used for binary classification tasks. It works by calculating the probability of a data point belonging to each class based on the features, and then choosing the class with the highest probability.

Advantages:

- Computationally efficient.
- Requires relatively little training data.
- Popular for text classification tasks.

Disadvantages:

- Computationally expensive at test time, especially for large datasets.
- Can be sensitive to irrelevant features or noisy data.
- May require careful choice of k value to balance between overfitting and underfitting.

Decision Trees: A tree-based algorithm that partitions the input space into regions based on a series of if-else conditions. In binary classification, the tree is built to separatethe two classes based on the input features, and the leaves of the tree represent the predicted class.

Advantages:

- Easy to interpret and visualize.
- Can handle both categorical and numerical data.
- Can capture non-linear relationships between the input features and the output.

Disadvantages:

- Can be sensitive to small changes in the data. Prone to overfitting if the tree is too deep.
- May not perform well when the input features are highly correlated.

Support Vector Machines (SVMs): A linear algorithm that finds the best hyperplane that separates the input data into two classes. SVMs can handle both linearly separableand non-linearly separable data by using different kernel functions. In binary classification, SVMs aim to find the decision boundary that maximizes the margin between the two classes.

Advantages:

- Can handle both linearly separable and non-linearly separable data with the rightchoice of kernel function.
- Robust to outliers and noise.
- Can handle high-dimensional data well.

Disadvantages:

- Can be computationally expensive to train, especially with large datasets.
- Can be sensitive to the choice of hyperparameters.
- May not perform well when there are many irrelevant features.

Random Forests: A collection of decision trees that are trained on different subsets of the input data and features, and the final prediction is made by aggregating the outputsof the individual trees. In binary classification, random forests can improve the accuracyand robustness of the predictions.

Advantages:

- Can handle both categorical and numerical data.
- Can capture non-linear relationships between the input features and the output.Robust to overfitting due to the use of multiple trees.

Disadvantages:

- Can be computationally expensive to train and test.
- Can be difficult to interpret the results due to the use of multiple trees.
- May not perform well when the input features are highly correlated.

Model Evaluation:

Logistic Regression: After tuning the parameters using GridSearch CV, the model has an accuracy of 0.7140, which means that it correctly predicts the class label for about 71% of the instances. The precision is 0.399188, which means that when it predicts the positive class, it is correct about 40% of the time. The recall is 0.697400, which means that it correctly identifies about 70% of the positive instances. The F1 score is 0.507745, which is a measure of the harmonic mean of precision and recall, indicating that the model has a moderate overall performance.

K-Nearest Neighbor: After tuning the parameters using GridSearch CV, the model has an accuracy of 0.6665. The precision is quite low at 0.240426, indicating that when it predicts the positive class, it is correct only about 24% of the time. The recall is also low at 0.267139, which means that it correctly identifies only about 27% of the positive instances. The F1 score is 0.253080, which is the lowest among all the models, indicating poor overall performance.

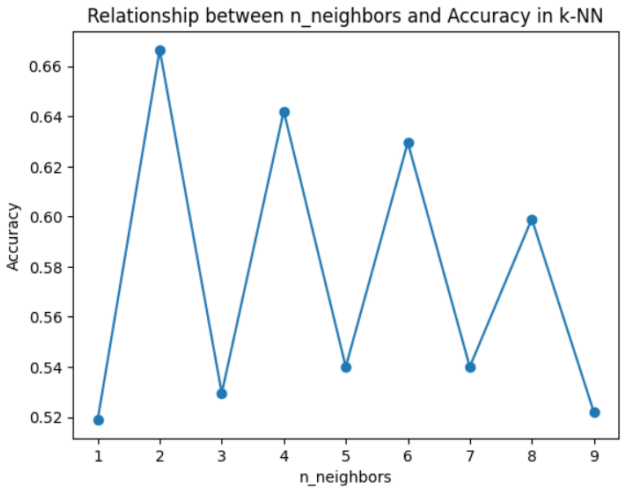


Figure 8: Accuracy vs n_neighbors for finding ideal k value

Naive Bayes: After tuning the parameters using GridSearch CV, the model has an accuracy of 0.7245, which is better than the logistic regression and K-Nearest Neighbor models. The precision is 0.409605, indicating that it correctly predicts the positive class about 41% of the time. The recall is 0.685579, indicating that it correctly identifies about 69% of the positive instances. The F1 score is 0.512821, indicating better overall performance.

Support Vector Machine: The model has an accuracy of 0.7200, which is slightly lower than the Naive Bayes model. The precision is 0.406036, indicating that it correctly predicts the positive class about 41% of the time. The recall is 0.699764, indicating that it correctly identifies about 70% of the positive instances. The F1 score is 0.513889, indicating moderate overall performance.

Decision Tree: After tuning the parameters using GridSearch CV, the model has an accuracy of 0.7880. The precision is 0.499218, indicating that it correctly predicts a positive class about 50% of the time. The recall is 0.754137, indicating that it correctly

identifies about 75% of the positive instances, which is like the logistic regression model. The F1 score is 0.600753, indicating moderate overall performance.

Random Forest: After tuning the parameters using GridSearch CV, the model has the highest accuracy of 0.7970, indicating that it correctly predicts the class label for about 80% of the instances. The precision is 0.513557, indicating that it correctly predicts the positive class about 50% of the time. The recall is 0.761229, indicating that it correctly identifies about 76% of the positive instances. The F1 score is 0.613333, which is the highest among all the models, indicating the best overall performance.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.7140	0.399188	0.697400	0.507745
1	K-Nearest Neighbour	0.6665	0.240426	0.267139	0.253080
2	Naive Bayes	0.7245	0.409605	0.685579	0.512821
3	Support Vector Machine	0.7200	0.406036	0.699764	0.513889
4	Decision Tree	0.7880	0.499218	0.754137	0.600753
6	Random Forest	0.7970	0.513557	0.761229	0.613333

Figure 9. Performance of Models

Receiver Operating Characteristic Curve:

The ROC curve is a graphical representation that depicts the performance of a binary classification model as the classification threshold varies. It is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold values.

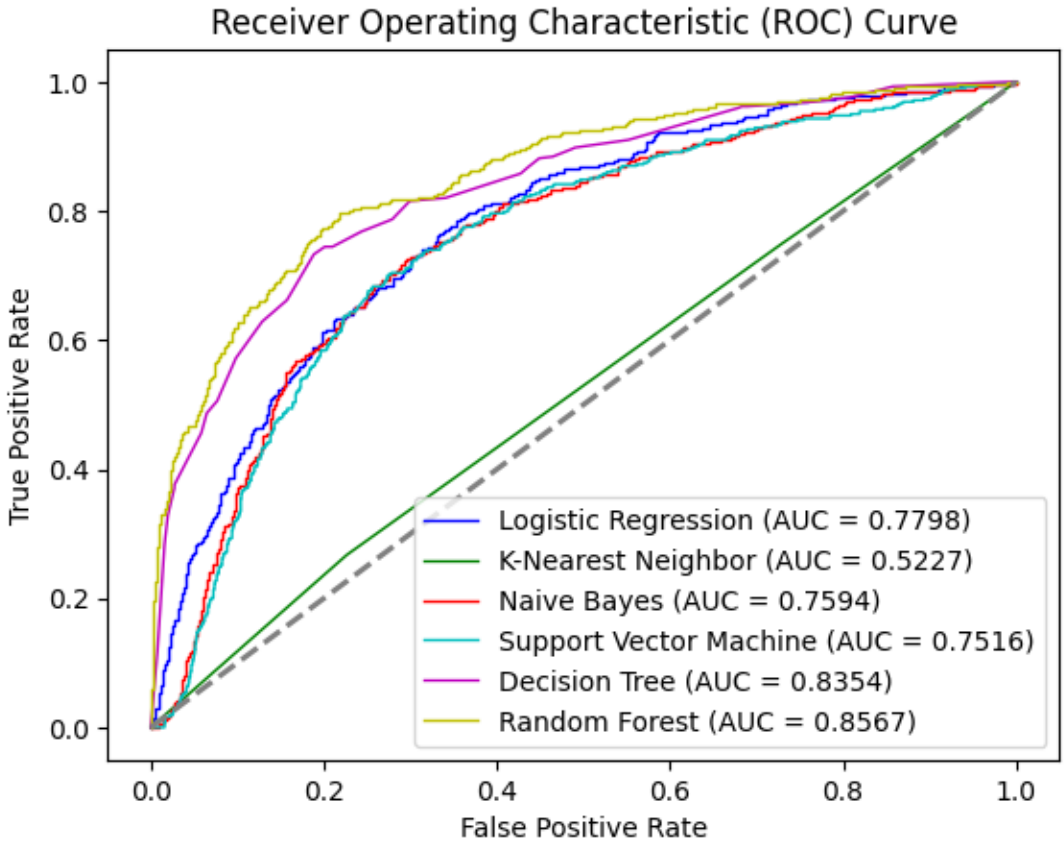


Figure 10: ROC Curve

From the above Figure 10, AUC of Random Forest is greater than the others. So, we conclude that Random Forest performs best among these models.

Classification Matrix:

The classification matrix is a tabular representation of the actual versus predicted classes. It is used to assess the effectiveness of a classification model. Each row represents instances in the actual class and each column represents instances in the predicted class. Performance metrics such as accuracy, precision, recall, and F1-score are then calculated from the classification matrix to evaluate the overall performance of the model.

Figure 11: Classification Matrix

Conclusion:

For our problem, Random Forest Classifier performs the best in all metrics.

It achieves an accuracy of 0.7970, precision 0.5135, recall 0.7612, f1-score 0.6133, and AUC of 0.8567.

Learnings from the Project:

During our data mining project, we learned about several important techniques for preparing data for analysis, such as cleaning, transformation, and normalization. Additionally, we became familiar with a variety of data mining algorithms, including decision trees, clustering, association rule mining, and neural networks. We were able to apply these techniques to extract valuable insights from the data.

One of the most crucial skills we acquired was the ability to evaluate the performance of different algorithms and compare them to determine the most effective method for a particular data set. We also learned how to create clear and compelling visualizations to communicate my findings to stakeholders.

Throughout the project, we discovered how critical it is to ensure data quality, select the appropriate features, and interpret results accurately to derive meaningful insights. Finally, we gained an understanding of the many diverse applications for data mining across various industries, from finance and marketing to healthcare and social media. Overall, the project provided us with valuable experience in data analysis and machine learning, and we feel confident in our ability to apply these skills in future projects.

Future Scope:

In the future, there is a significant scope for bank churning analysis due to the increasing competition among financial institutions to attract and retain customers. As banks continue to offer new and innovative products, features, and services, consumers will look for ways to take advantage of these offers and maximize their financial gains.

One potential area of growth for bank churning analysis is the use of advanced analytics and machine learning techniques to identify the most profitable offers and optimize the churning process. With the help of predictive models and data-driven insights, consumers can make more informed decisions about when to open and close accounts, which banks to target, and which promotions to take advantage of.

References:

1. Scikit-learn: Machine Learning in Python. https://scikit-learn.org/stable/

2. MatPlotLib: Visualization with Python. https://matplotlib.org/

3. Seaborn: Statistical Data Visualization. https://seaborn.pydata.org/

4. Random Forest Classifier using Scikit-learn. https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/

5. Medium. (2019, August 28). Understanding AUC - ROC Curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

6. Medium. (2020, May 6). Classification Metrics — Confusion Matrix Explained. https://towardsdatascience.com/classification-metrics-confusion-matrix-explained-7c7abe4e9543

7. Khani, A. H., Rezaei, S., & Oroumchian, F. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. Journal of Financial Innovation, 2(1), 1-23. doi: 10.1186/s40854-016-0029-6.