A Data-Driven Approach to Predict the Success of Bank Telemarketing

Curated By:
Pyimoe Than
&
Zain Mirza

San Francisco State University

Table of Contents

1. Executive Summary

This document outlines our approach to help a Portuguese bank overcome its challenges in effectively advertising its ‘bank term deposit’ service. Through the use of machine learning techniques, we are confident in our ability to predict which individuals are most likely to subscribe to this service based on various factors in a dataset. Our project involves identifying the most effective ML model for predicting responses and ensuring that our phone call campaigns are personalized and targeted toward the proper audience.

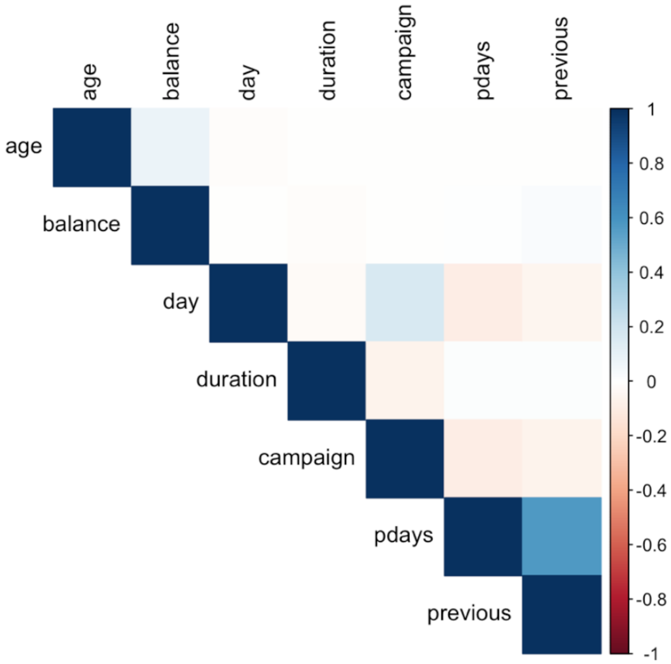
2. Introduction

The dataset, acquired from the University of California, Irvine Repository, from this project contains information on direct marketing campaigns conducted by a Portuguese banking institution. The ML techniques aim to enhance the bank to understand the predicting factors that potentially determine a classified “yes” or “no” response to a ‘term deposit’ subscription.

3. Data Description

The data utilized in this project is sourced from the UCI Repository, specifically from the Bank Marketing Data Set. This dataset comprises 45,211 observations including 17 columns.

3.1 Correlation Matrix Graph Between Non-numeric Predictor Variables



Correlation coefficients between many pairs of predictor variables in the data appear to be less than 0.5. This suggests a weak linear relationship between the variables.

Variable overview:

1. ***age***: (numeric)
2. ***job***: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur',' housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3. ***marital***: marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4. ***education*** (categorical:)
5. ***default***: has credit in default? (categorical: 'no', 'yes', 'unknown')
6. ***Balance***: (numeric)
7. ***housing***: has a housing loan? (categorical: 'no', 'yes', 'unknown')
8. ***loan***: has a personal loan? (categorical: 'no', 'yes', 'unknown')
9. ***contact***: contact communication type (categorical: 'cellular', 'telephone')
10. ***month***: last contact month of the year (categorical: 'Jan', 'Feb', 'mar', ..., 'Nov', 'Dec')
11. ***day_of_week***: last contact day of the week (categorical: 'mon', 'Tue', 'wed', 'Thu', 'Fri')
12. ***duration***: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call, y is known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
13. ***campaign***: number of contacts performed during this campaign and for this client (numeric, includes the last contact)
14. ***pdays***: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means the client was not previously contacted)
15. ***previous***: number of contacts performed before this campaign and for this client (numeric)
16. ***poutcome***: outcome of the previous marketing campaign (categorical: 'failure',' nonexistent',' success')

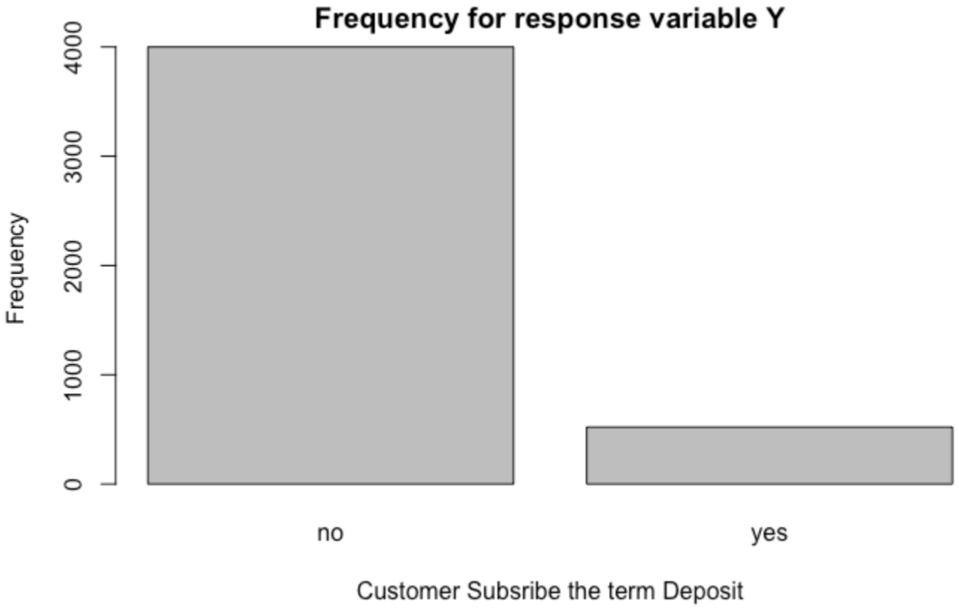
Output variable (desired target)

17. ***y***: has the client subscribed to a term deposit? (binary: ‘yes’, ‘no’)

3.2 Data Cleaning

During the data cleaning process, we conducted a series of integrity checks to ensure the dataset's accuracy, completeness, and consistency. We were pleased to discover that there were no missing values or duplicated data in the dataset. However, we did notice an unbalanced response variable, with a higher count of "No" responses than "Yes" responses. This class imbalance has the potential to introduce bias in our prediction model, which is something we must address.

Figure 1: Frequency of Responses



```
Percentage of 'yes' subscription of term deposit: 11.524 %
Percentage of 'no' subscription of term deposit: 88.476 %
```

The figure above depicts the total number of ‘yes’ responses and ‘no’ responses in the dataset. From the graph above we can see there are far more ‘no’ responses to the bank service than there are ‘yes’.

An oversampling method was applied to address this issue to balance the data. The result of this technique exhibits a balanced distribution in our dataset and is now ready for further analysis using machine learning techniques. Below, in Figure 2, you can visualize the frequency of responses after including the oversampling method.

Figure 2: Frequency of Responses (After Oversampling Method)

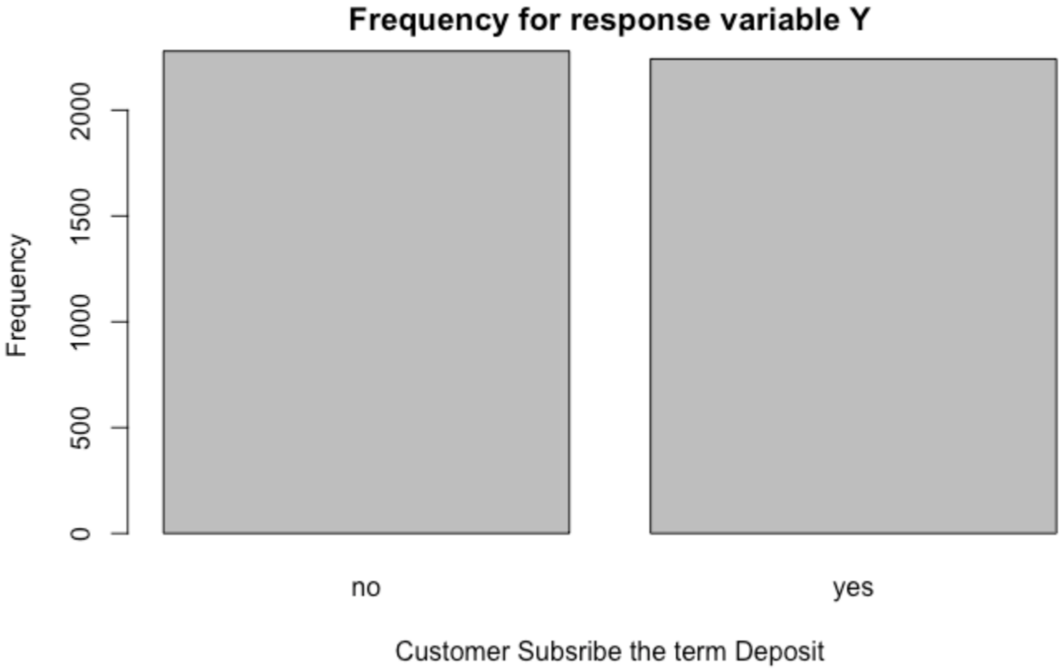
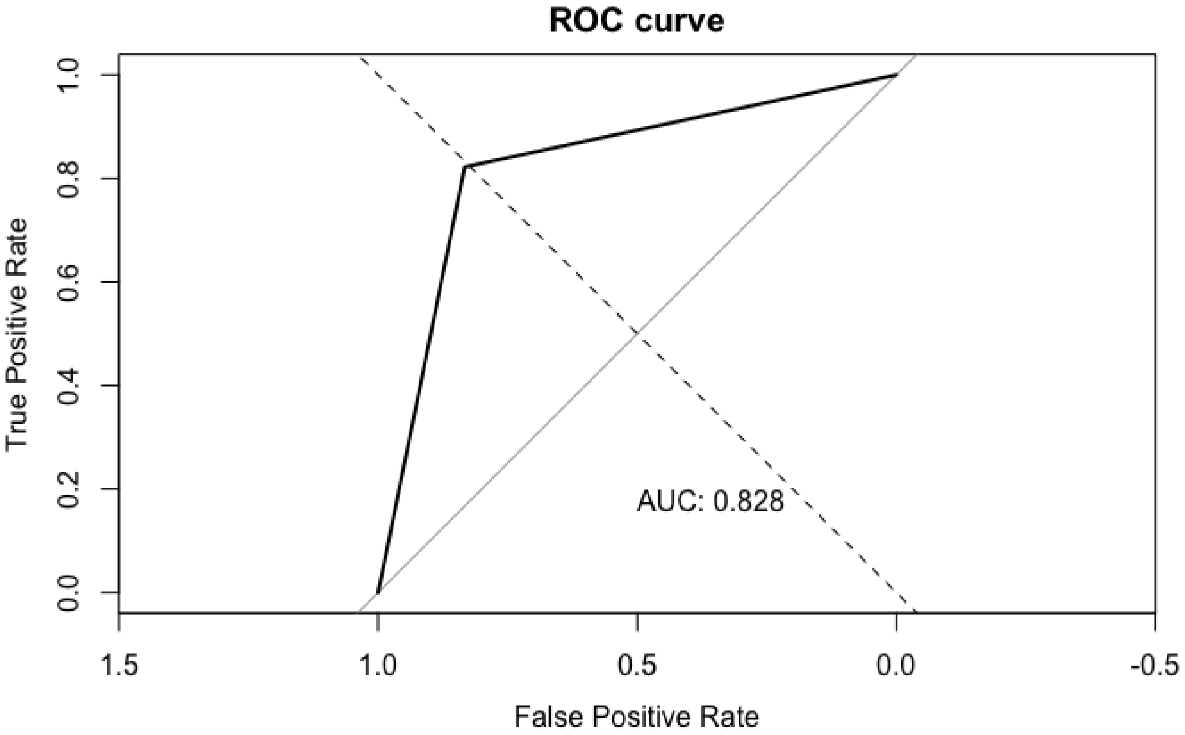


Figure 2 above displays a balanced number of ‘yes’ and ‘no’ responses, after incorporating the oversampling method.

4. Fit the logistic Regression with all predictors

Following the data cleaning process, we divided the data into 80% training and 20% testing data prior to fitting. The Logistic Regression model was trained using both k-fold cross-validation on the training data. This comprehensive validation approach provides robust estimates of the model's performance. We generated Confusion Matrix which show the accuracy, AUC, sensitivity, and specificity

Accuracy	Sensitivity	Specificity	AUC	misClassification Rate
0.8276	0.8257	0.829	0.8283	0.17



- Accuracy score of 0.8276 means the model correctly predicted the outcome for **82.76 %** of the bank customers.
- AUC score of 0.8283 means the model performs better than the random guessing since it is greater than 0.5
- Sensitivity means the percentage of positive outcomes the model is able to detect
- Specificity means the percentage of negative outcomes the model is able to detect.
- misClassificationRate of 0.17 means the model incorrectly predicted the outcome for **17%** of the bank customers.

We also found a significant number of predictor variables from the null model that can be used to predict the response for a 'term deposit' bank service. However, this approach carries the risk of over-fitting since all features are used to fit the model. To mitigate this risk, we will use feature selection to identify the most important predictors responsible for our response outcome. By selecting only the most relevant predictors, we can boost the accuracy and interpretability of our predictive models. In this analysis, we will ensure that our selection of predictors is optimal to produce the best results. We will use the forward and backward elimination process to determine the proper feature selection.

5. Feature Selection

5.1 Forward Selection

The Forward Selection method starts with an empty feature set and iteratively adds one feature at a time based on its performance and contribution to the model's predictive power. From the features selection using the forward method, we found these variables to be most significant:

job	education	loan	day	duration	pdays	poutcome	marital	housing	contact	month	campaign	previous

The table above displays all the significant predictor variables found using the forward selection process.

5.2 Backward Selection

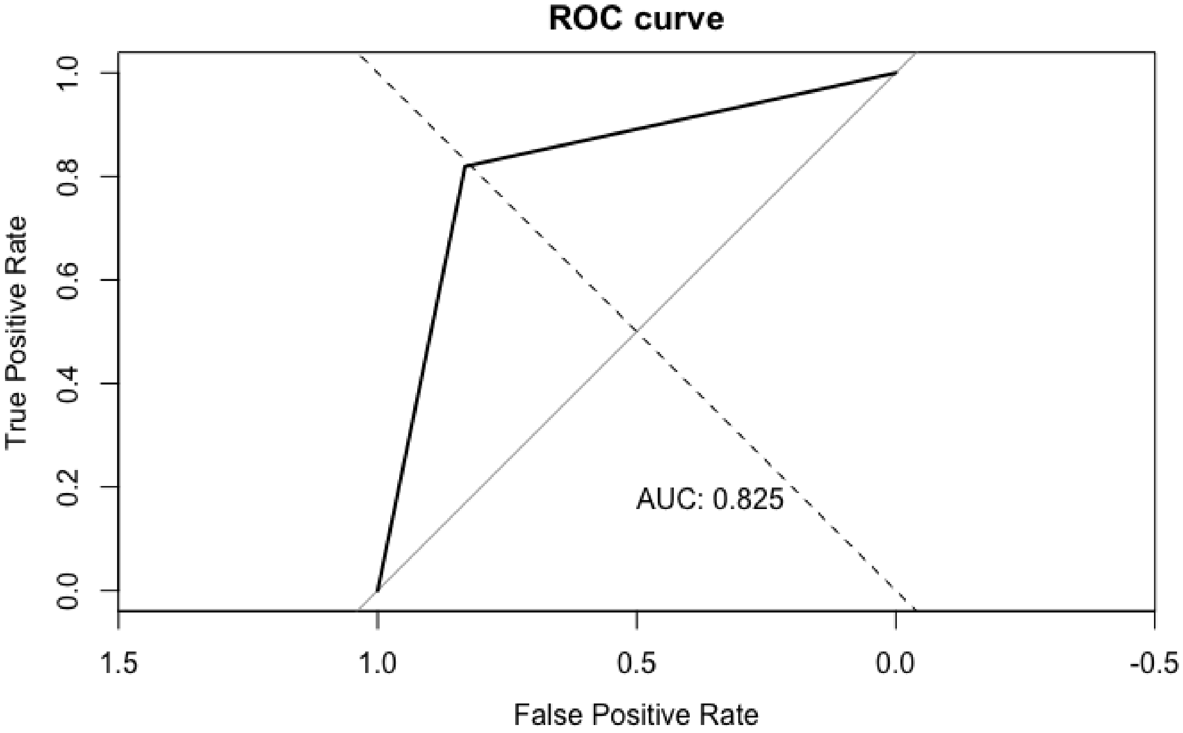
The Backward Selection method, in contrast to the Forward Selection, commences with a complete set of features and gradually eliminates one feature at a time based on its performance and impact on the model's predictive ability. Our previous findings have demonstrated that the Backward Elimination approach produces the same outcome as the Forward Selection. Both Forward and Backward methods have identified the same 13 predictor variables from the dataset's 16.

6. Model Selection

6.1 Logistic Regression with Logit Link

After the feature selection process, to ensure a reliable evaluation, we divided the data into 80% training and 20% testing data prior to fitting. The Logistic Regression model was trained using both k-fold cross-validation and Leave-One-Out Cross-validation (LOOCV) on the training data. This comprehensive validation approach provides robust estimates of the model's performance. We generated Confusion Matrix which show the accuracy, AUC, sensitivity, and specificity

Accuracy	Sensitivity	Specificity	AUC	misClassification Rate
0.8254	0.8276	0.823	0.825	0.17

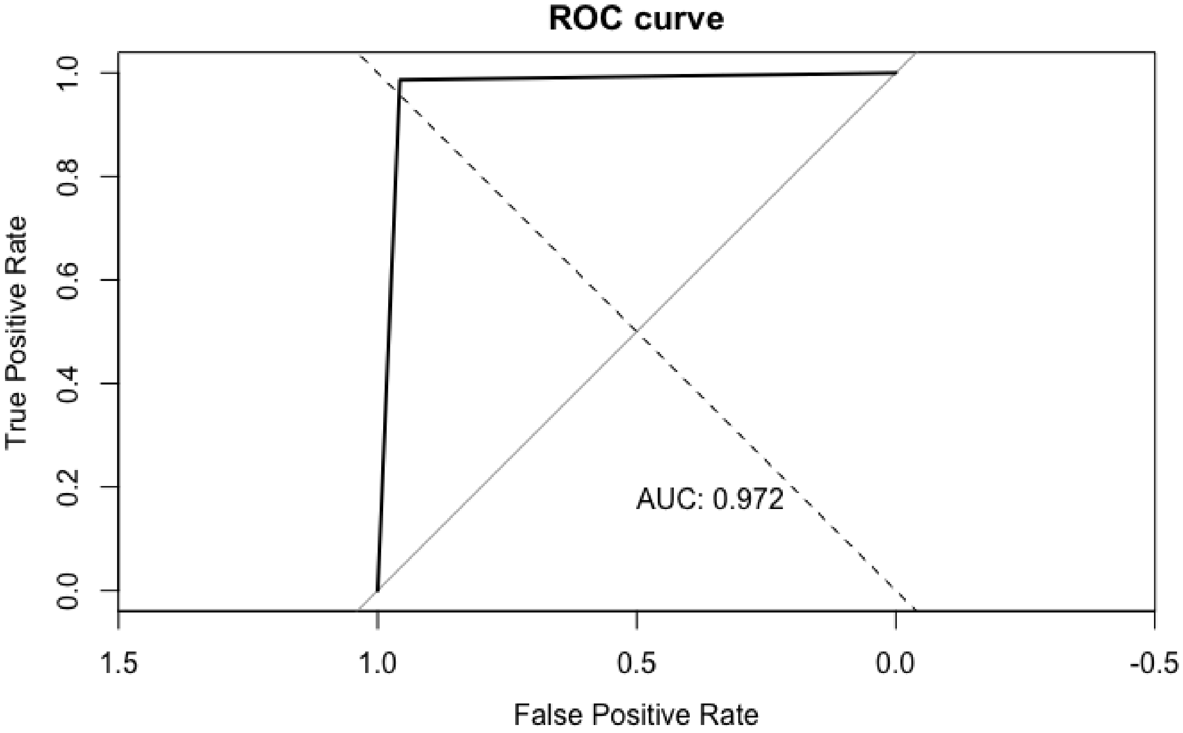


- Accuracy score of 0.8254 means the model correctly predicted the outcome for **82.54 %** of the bank customers.
- AUC score of 0.825 means the model performs better than the random guessing since it is greater than 0.5
- Sensitivity means the percentage of positive outcomes the model is able to detect.
- Specificity means the percentage of negative outcomes the model is able to detect.
- misClassificationRate of 0.17 means the model incorrectly predicted the outcome for **17%** of the bank customers.

6.2 Random Forest

After fitting the logistic regression, I then fit the model using the random Forest machine learning model on the training data. We generated a Confusion Matrix which shows the accuracy, AUC, sensitivity, and specificity. We further create a ROC curve. It gives us the AUC score.

Accuracy	Sensitivity	Specificity	AUC	misClassificationRate
0.9713	0.9567	0.9865	0.972	0.028



-
- Accuracy score of 0.9713 means the model correctly predicted the outcome for **97.13 %** of the bank customers.
- AUC score of 0.972 means the model performs better than the random guessing since it is greater than 0.5
- Sensitivity means the percentage of positive outcomes the model is able to detect.
- Specificity means the percentage of negative outcomes the model is able to detect.
- misClassificationRate of 0.028 means the model incorrectly predicted the outcome for **2.8%** of the bank customers.

So far, the Random Forest machine learning model gives the lowest misClassification Rate. Some disadvantages of using it is

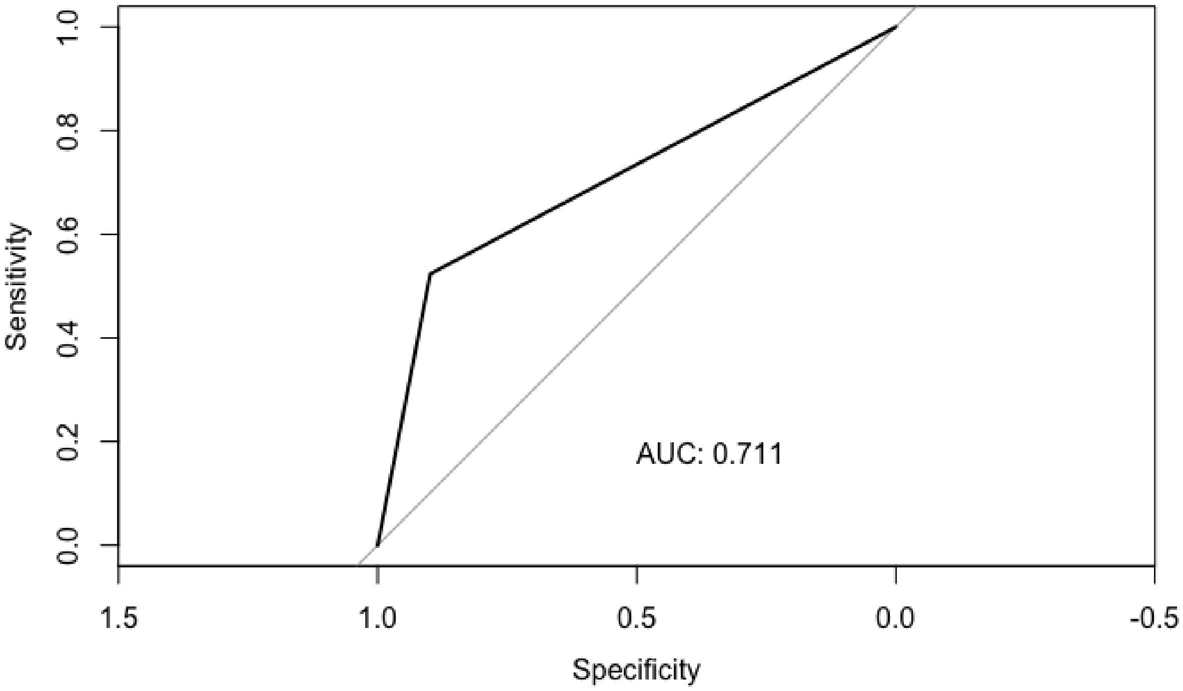
1. Computationally expensive
2. Require significant amount of memory
3. Hard to interpret the features

6.3 Naive Bayes

Since Random Forest is memory intensive, I will try to fit the model using a fast, efficient, and simple method which is Naive Bayes.

Accuracy	Sensitivity	Specificity	AUC	misClassificationRate

0.7149	0.898	0.523	0.711	0.285



- Accuracy score of 0.7149 means the model correctly predicted the outcome for **71.49 %** of the bank customers.
- AUC score of 0.711 means the model performs better than the random guessing since it is greater than 0.5
- Sensitivity means the percentage of positive outcomes the model is able to detect.
- Specificity means the percentage of negative outcomes the model is able to detect.
- misClassificationRate of 0.285 means the model incorrectly predicted the outcome for **28.5 %** of the bank customers.

Although Naive Bayes is a fast, efficient and simple method, it gives a high misclassification rate compared to Random Forest and Logistic Regression. Here are the some disadvantages of using Naive Bayes

1. Strong independence assumption
2. Inability to capture complex relationships

6.4 Proposed Solution

When selecting a machine learning(ML) algorithm, it is important to carefully consider their advantages and disadvantages. Additionally, the available budget should be taken into account.

The decision between Random Forest and Naive Bayes should be based on the project requirements and available resources. If computational efficiency is a priority, Naive Bayes is recommended. However, if accuracy and handling complex relationships are crucial, and computational resources are sufficient, Random Forest may be the better option.

Appendix

7.1 Data Source - UCI Repository: https://archive.ics.uci.edu/ml/datasets/bank+marketing.

7.2 Code - GitHub :

Project Code