McMaster University

STATS/CSE 780

Final Project Report 1

Seyed Mohammad Mehdi Hassani Najafabadi(Student ID: 400489126)

17 April, 2023

Contents

Introduction

The problem we discussed in this project is Identifying the best strategies to improve the effectiveness of future marketing campaigns by analyzing the patterns from the last marketing campaign performed by the bank.The data used in this project pertains to direct marketing campaigns conducted by a Portuguese banking institution through phone calls. The dataset was sourced from UCI Machine Learning Repository [1], and is publicly available for research. A Portuguese retail bank was addressed, with 17 social-economic variables collected from 4521 costumers (observation) between 2008 to 2013, thus including the effects of the 2008 financial crisis [2]. This dataset has been widely used for data analysis in terms of improve next marketing campaigns and leverage the financial performance of marketing campaign Kaggle bank marketing analysis projects, such as [3] and [4]. In this project the problem statement can be broken down into two main tasks:Predicting whether a client will subscribe to a term deposit product (variable 'y') using the provided dataset and different classification Techniques, and finding meaningful patterns to improve the credibility for telemarketing campaign.

Methods

In order to address the problems posed in the problem definition, four different machine learning classification algorithms have been used, all of these methods have unique strengthener to examine the best method: 1- Decision tree(DT): DT has the advantage of fitting models that tend to be easily understood by humans, while also providing good predictions in classifications tasks.In addition, DT is capable of handling mixed data types which we have in this problem.The rationale for selecting the tuning parameters for DT is : 1- max_depth: Controls the maximum depth of the tree, 2- min_samples_split: Determines the minimum number of samples required to split an internal node, 3- min_samples_leaf: Specifies the minimum number of samples required to be at a leaf node.

2- Random Forest(RF): RF is an ensemble learning method that constructs multiple decision trees during training and outputs the class with the majority vote from individual trees. It offers improved accuracy and reduced overfitting compared to a single decision tree. RF can handle mixed data types, making it suitable for this problem, and provides an added advantage of feature importance estimation.The rationale for selecting the tuning parameters for RF is :1-n_estimators: Refers to

the number of trees in the forest, 2- max_features: Specifies the maximum number of features to consider when looking for the best split,3-max_depth, min_samples_split, and min_samples_leaf: These parameters have the same purpose as in the decision tree and help to control the complexity of individual trees in the forest.

3- Gradient Boosting classifier(GBC): GBC is another ensemble learning technique that combines multiple weak learners, usually decision trees, to build a strong predictive model. It builds trees sequentially, with each tree attempting to correct the errors of its predecessor. GB is known for its high predictive accuracy and can handle mixed data types. However, it might require more computational resources compared to other methods.The rationale for selecting the tuning parameters for GBC is :1- n_estimators: Determines the number of boosting stages or weak learners, 2-learning_rate: Controls the contribution of each weak learner in the ensemble. 3- max_depth, min_samples_split, and min_samples_leaf: As in the decision tree and random forest, these parameters help control the complexity of the base learners.

4- Support Vector machine(SVM): SVM is a powerful machine learning algorithm that is widely used in classification tasks. It finds the optimal hyperplane that separates the data points of different classes with the maximum margin. SVM can effectively handle high-dimensional data and is less prone to overfitting. Though it may require more computational resources and preprocessing steps, such as feature scaling, it can deliver good classification performance on diverse problems.The rationale for selecting the tuning parameters for SVM is :1- C: This regularization parameter controls the trade-off between maximizing the margin and minimizing the classification error.2-kernel: Specifies the kernel function to use in the algorithm. 3- gamma: This parameter is only applicable for certain kernel functions, such as the radial basis function (RBF) kernel. To effectively compare the four classification algorithms, we use the following evaluation metrics: ROC curve, accuracy, precision, recall, and F1 score. The ROC curve assesses the trade-off between sensitivity and specificity, making it useful for imbalanced datasets. Accuracy measures the proportion of correct predictions, while precision and recall evaluate the costs of false positives and false negatives, respectively. The F1 score balances precision and recall, providing a comprehensive measure of classifier performance. Analyzing these metrics helps determine the most suitable classifier for the given problem based on specific requirements and performance aspects.

Results

The number of observation is 4521 and there are 17 variables which are described follows:

Variable names

Variable names in this data sets are: 1- age (numeric),2-job(categorical), 3- marital status(categorical),4-education(categorical),5- has credit in default? (binary: "yes","no"),6- average yearly balance, in euros (numeric),7- has housing loan? (binary: "yes","no"),8- has personal loan? (binary: "yes","no"),9- contact communication type(categorical),10- last contact day of the month (numeric), 11- month: last contact month of year(categorical),12- last contact duration(numeric),13-number of contacts performed during this campaign(numeric),14- number of days that passed by after the client(numeric), 15- number of contacts performed before this campaign and for this client (numeric),16- outcome of the previous marketing campaign(categorical),17- **Target Variable**: has the client subscribed a term deposit? (binary:"yes","no")

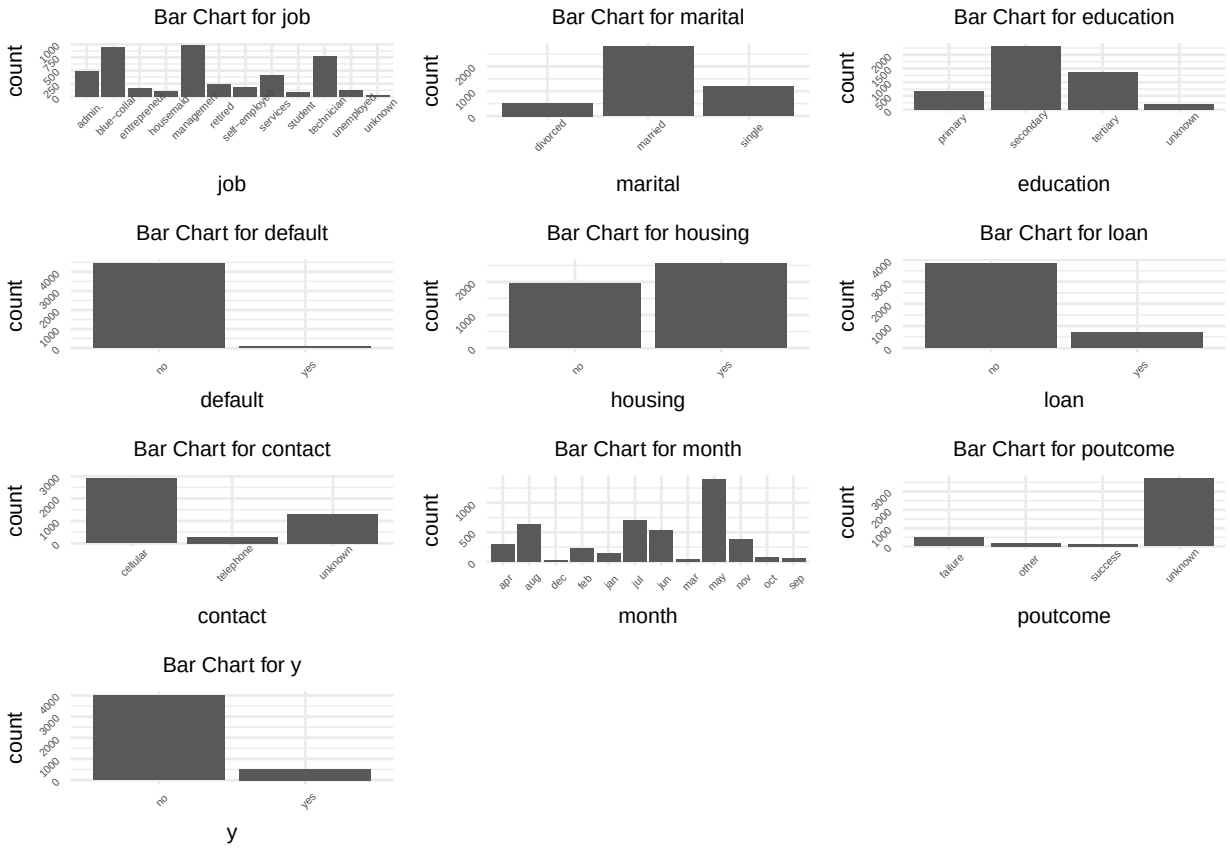
Exoplanetary data analysis

Missing Vlaue

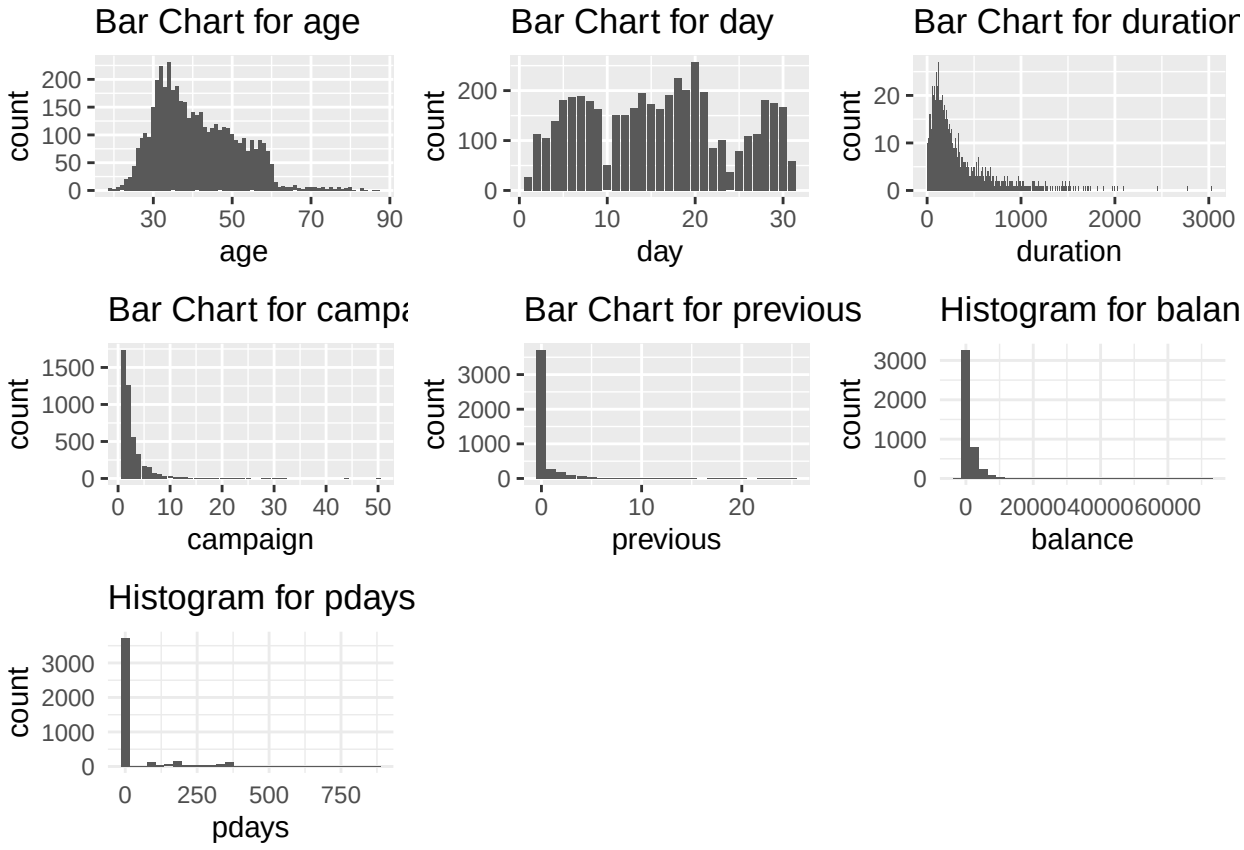
After importing the dataset, we have to look at the total number of rows in the dataset and analyze the number of missing values.So As it is shown , there are no missing values.

Summeries and Data type

In the dataset we have both categorical and numerical columns. Let's look at the values of categorical columns first.

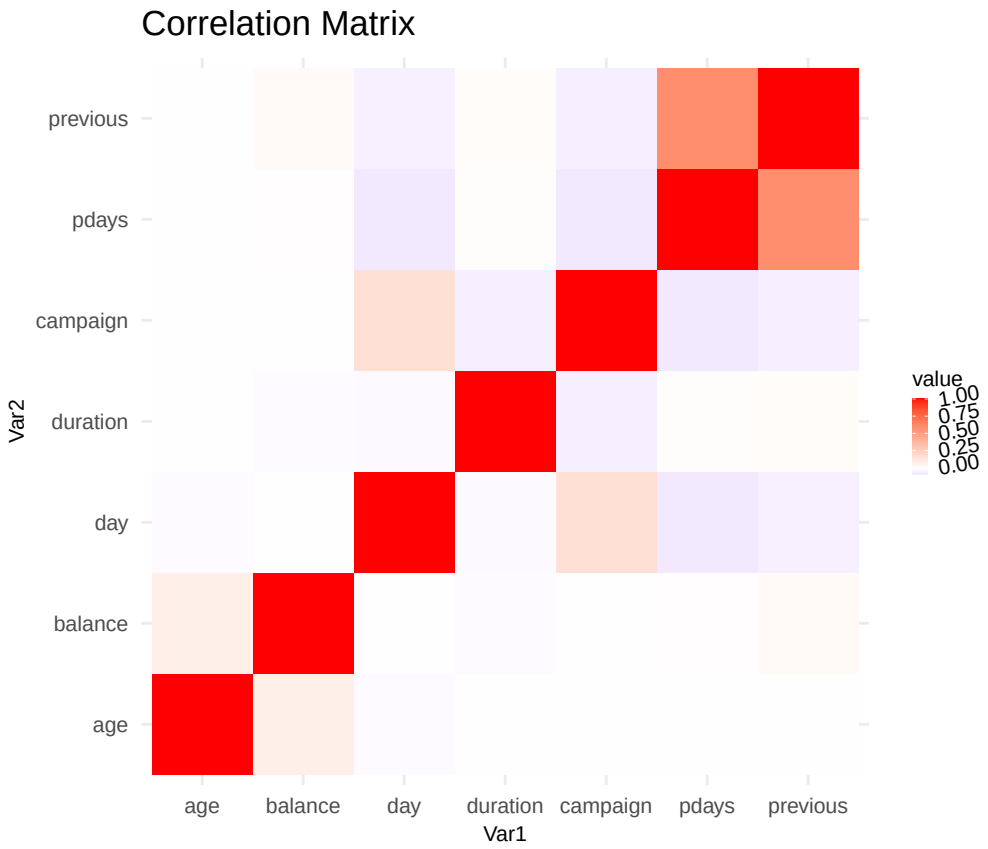


Now let's look at the numerical columns' values. The most convenient way to look at the numerical values is plotting histograms.



Correlation Analysis

Since some of the variables in the dataset are categorical, the correlation analysis has been performed only on the numerical variables.Upon examining the heatmap, we found that the color patterns revealed little to no correlation between the features in our dataset. This means that the numerical variables under study do not exhibit strong linear relationships with each other. Consequently, this lack of correlation suggests that the variables are relatively independent, which can be an advantageous attribute when employing certain machine learning algorithms or statistical techniques that assume minimal multicollinearity between predictors. In the context of our study, this finding implies that further investigation may be necessary to uncover more complex, non-linear relationships, or to explore the potential influence of the categorical variables on the outcome of interest.

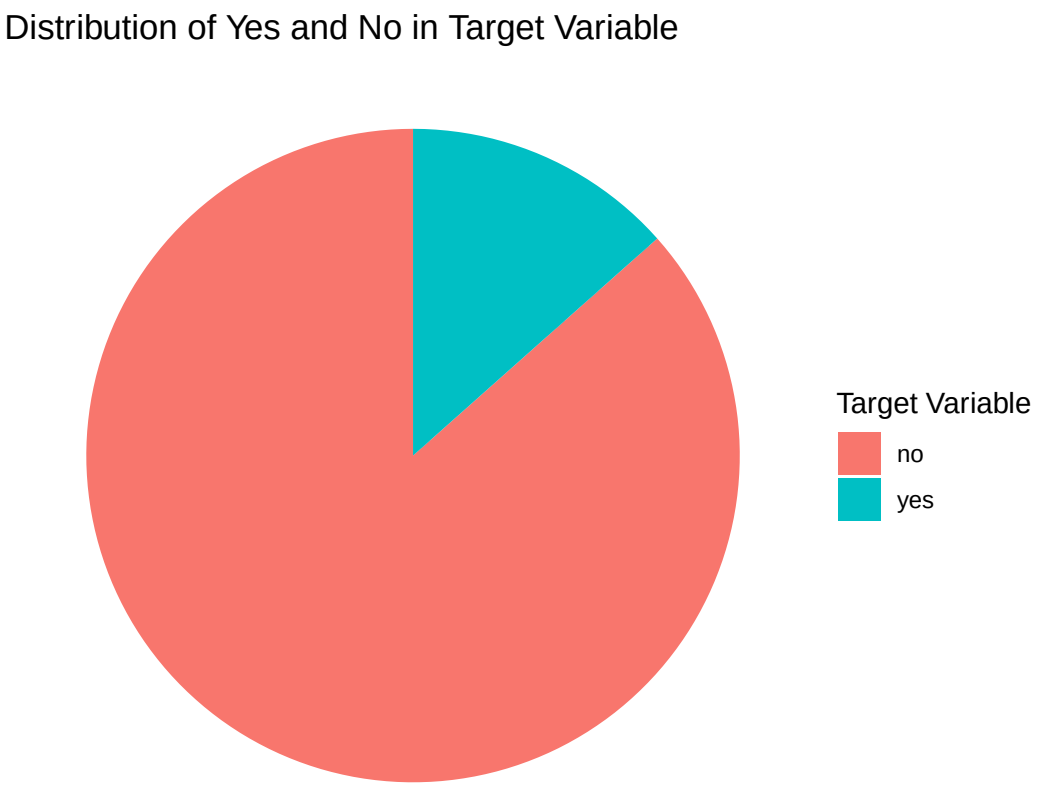


Outliers analysis

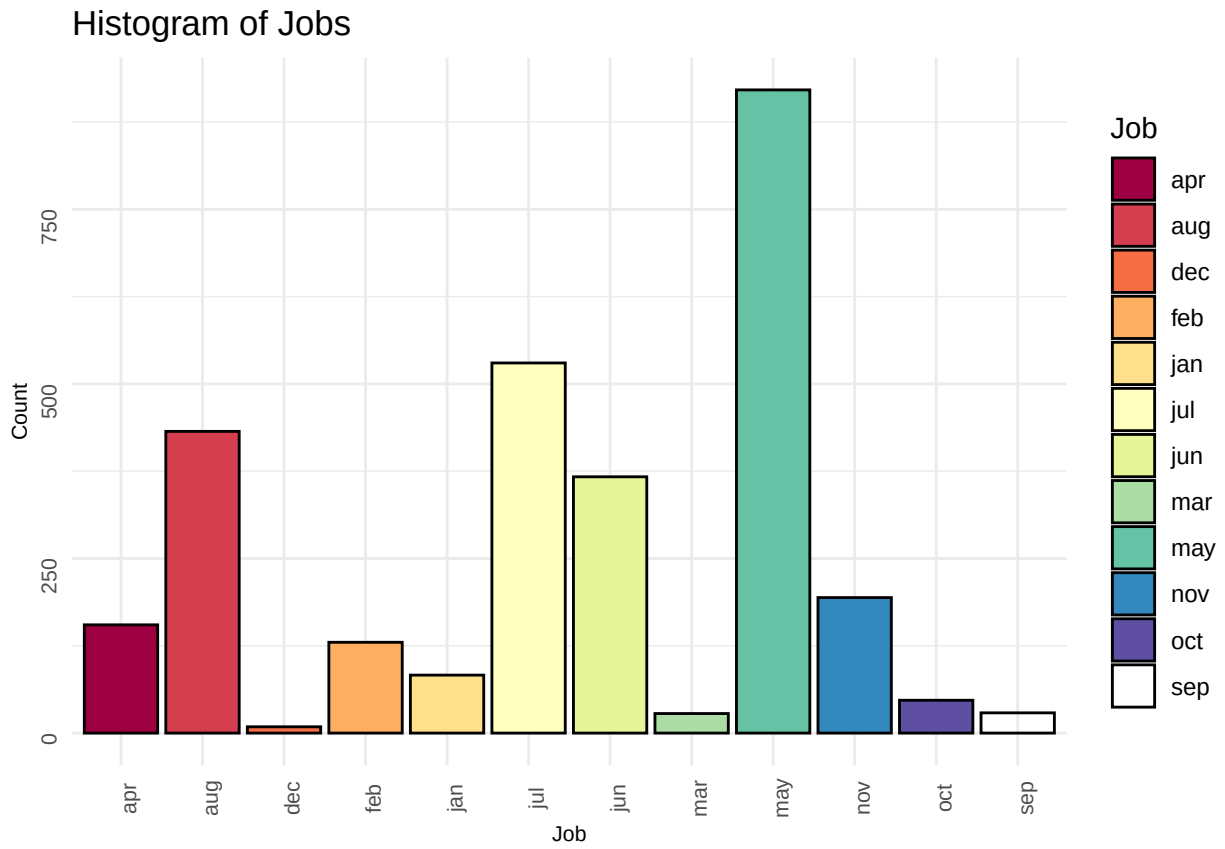
According to the figure(), there are some outliers which can be handled by inter quarterly Rang technique(IQR) techniques:

Analysis of the Response Column

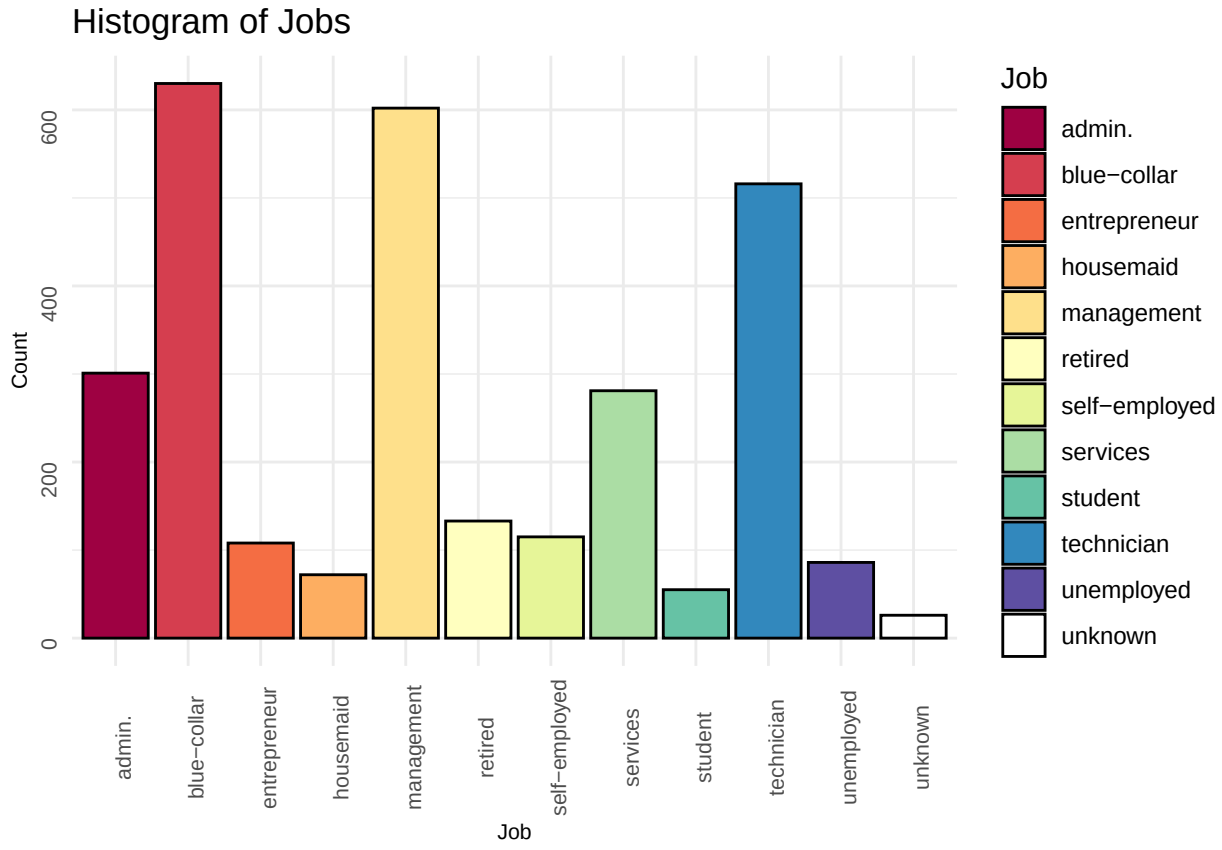
It is very important to look at the response column, which holds the information, which we are going to predict. In our case we should look at 'deposit' column and compare its values to other columns. First of all we should look at the portion of 'yes' and 'no' values in the response column 'deposit'.



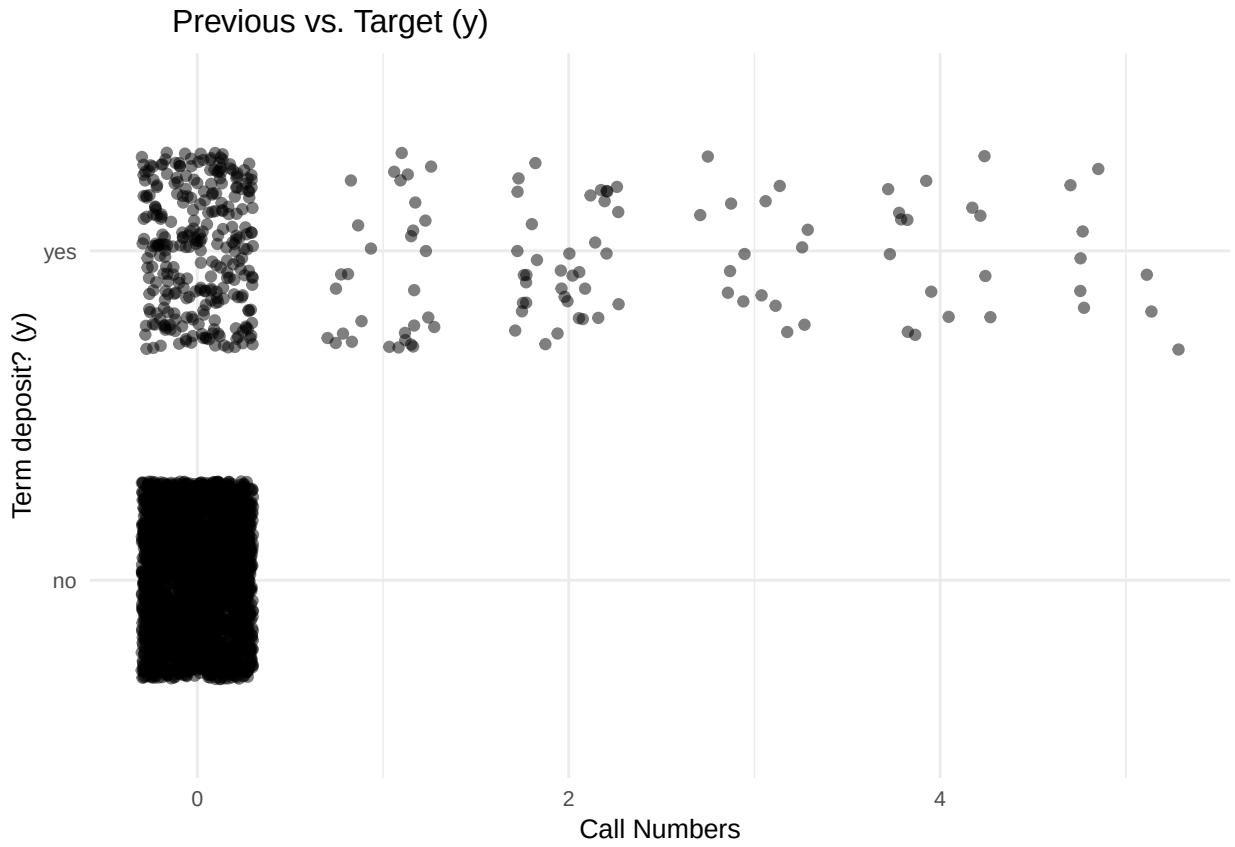
Our findings show that the months of May, June, July, and August, which constitute the summer season, have a higher number of phone call campaigns. This could be referred to as the "Summer Camp" period for marketing efforts.



it's crucial to focus on specific target groups for these campaigns. Our data suggests that targeting individuals in management positions, blue-collar workers, and technicians may yield better results in terms of term deposit subscriptions. By concentrating our efforts on these groups, we can enhance the efficiency and success of our direct marketing campaigns.



Based on the following analysis, we recommend not calling clients too many times during this period. Overwhelming clients with phone calls may lead to reduced effectiveness of the marketing campaignand a negative perception of the bank.



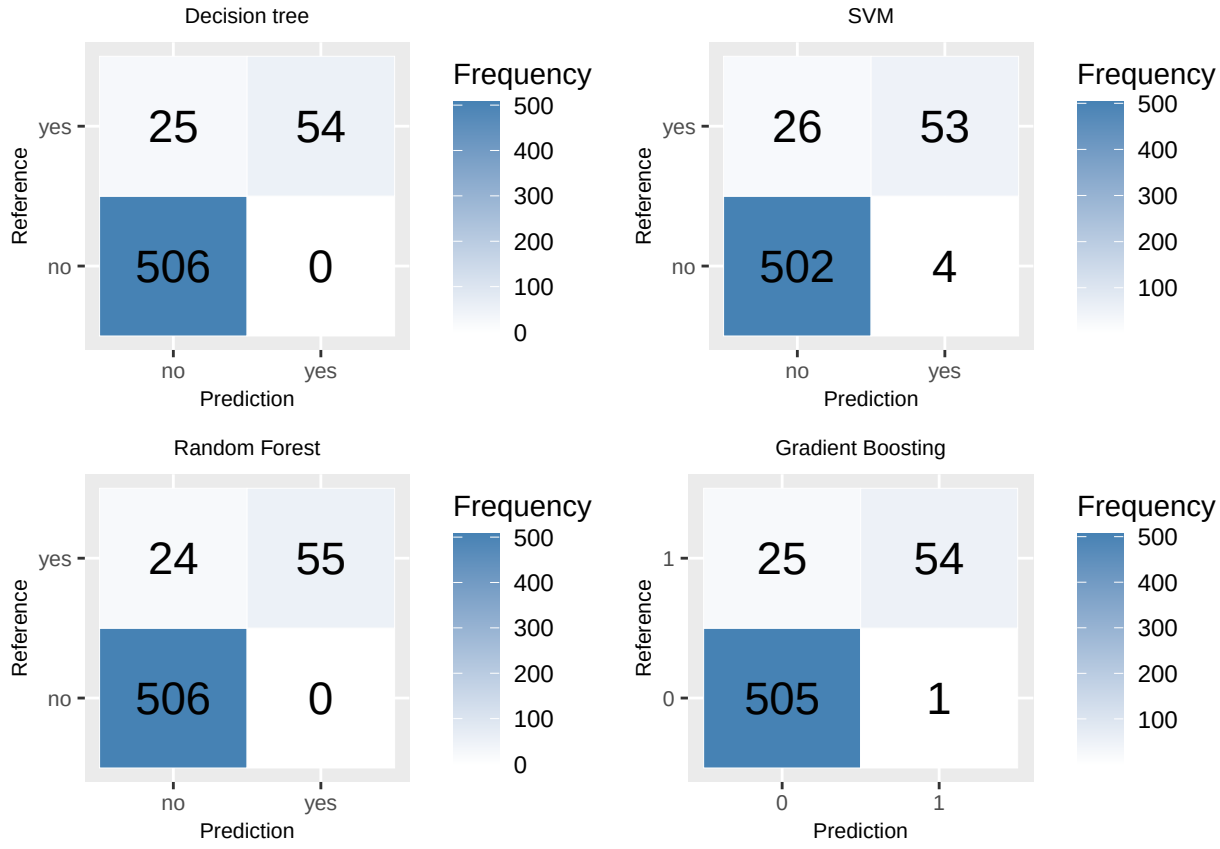
Interpretation of results applying two methods

We considered four different machine learning models: Decision Trees, Random Forest, Support Vector Machines (SVM), and Gradient Boosting. Each model has its unique strengths and applications. To validate our models, we employed K-folds cross-validation, which involves splitting the data into train and test sets multiple times, ensuring a more accurate and reliable evaluation of the models' performance.

Split Data set to train and test by Kfolds and Decision trees

To validate our models, we employed K-folds cross-validation, which involves splitting the data into train and test sets multiple times, ensuring a more accurate and reliable evaluation of the models'performance.

Training



Rank

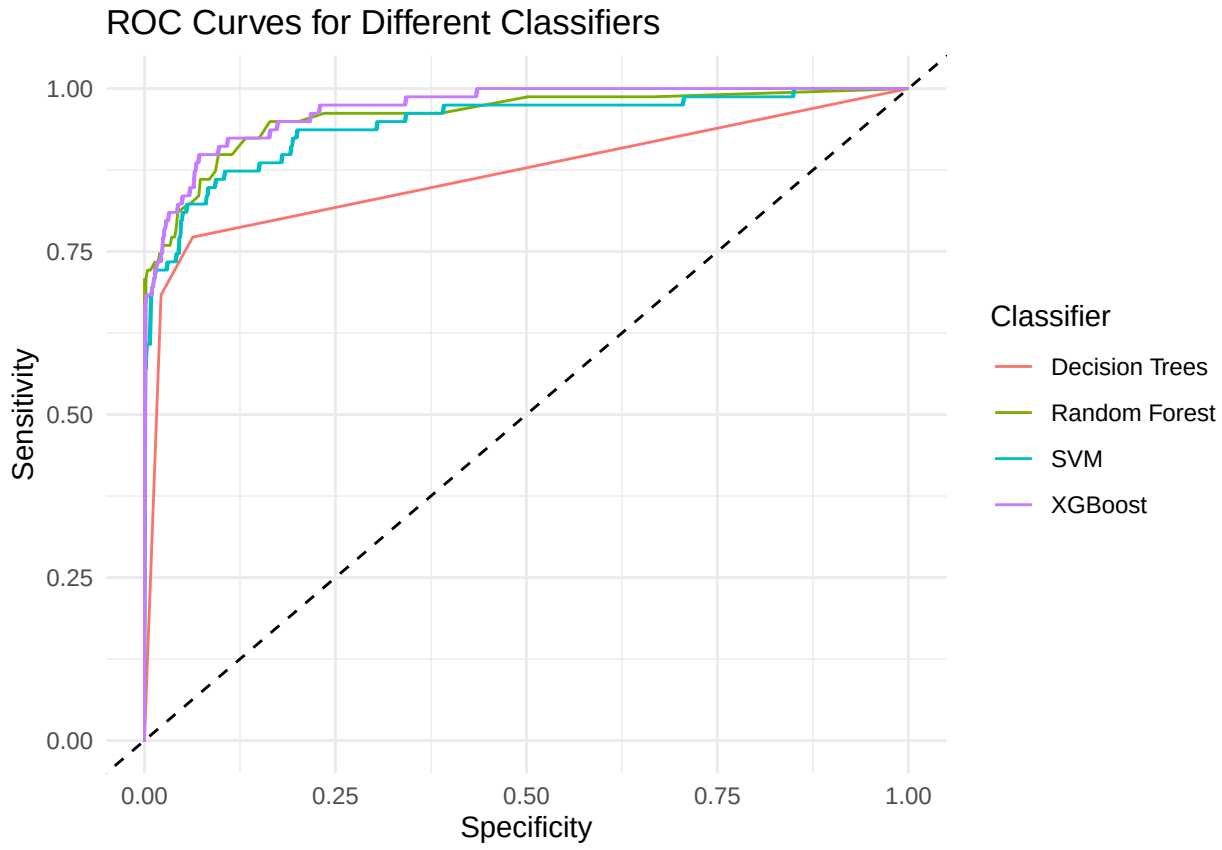
This section calculates the evaluation metrics (accuracy, F1 score, precision, and recall) for four classifiers (Decision Trees, SVM, Random Forest, and XGBoost). It then ranks the classifiers based on each of the evaluation metrics and computes their average rank. The classifiers are sorted by their average rank to compare their performance.The reason for using ranking instead of relying on a single metric is that it allows for a more comprehensive comparison of the classifiers. By taking into account multiple evaluation metrics, the ranking method provides a better overview of the overall performance of each classifier. It helps to identify the classifiers that perform consistently well across all the metrics, rather than just excelling in one particular aspect.

Table 1: Classifier Evaluation Metrics and Ranking

	classifier	accuracy	f1_score	precision	recall
2	SVM	0.9487179	0.9709865	0.9920949	0.9507576
4	XGBoost	0.9555556	0.9749035	0.9980237	0.9528302
1	Decision Trees	0.9572650	0.9758920	1.0000000	0.9529190
3	Random Forest	0.9589744	0.9768340	1.0000000	0.9547170

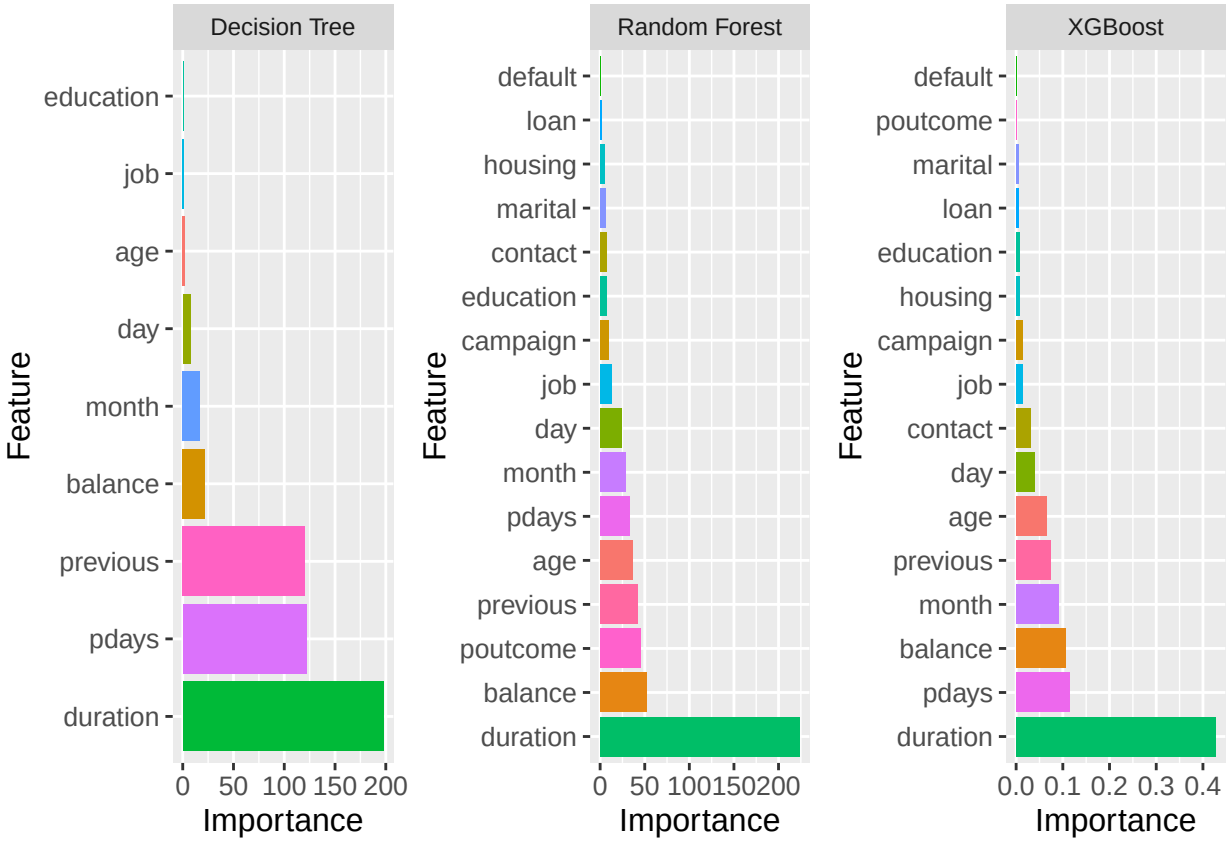
ROC evaluation

In this analysis, we compared the performance of four different classifiers using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values. Upon inspecting the ROC curves, it was observed that the Decision Tree classifier had a curve that was close to the minimum ROC, indicating comparatively lower performance than the other classifiers. A higher ROC curve represents better classifier performance, while a curve closer to the minimum ROC suggests that the classifier's performance is closer to that of a random classifier.



Feature importance

Feature importance analysis is conducted to identify the most influential variables in a predictive model, providing insights into the underlying relationships between features and the target variable. By understanding these relationships, we can make more informed decisions about feature selection and model improvement, ultimately enhancing the model's predictive performance.



The analysis revealed a 34% difference between the impact of contact duration and the number of contacts. Some features such as having housing or personal loans, and the client's education levelwere found to be less significant in determining term deposit subscriptions. made.

Conclusion

1- Key factors influencing the outcome of phone call campaigns include the number of days since the client was last contacted, the number of contacts performed, and the client's account balance.
2- Targeting individuals with longer call durations may prove advantageous, as they exhibit an average subscription rate of 78%.

3- There is a noteworthy relationship between balance and loans: individuals with low or no balance are more likely to have a housing loan compared to those with average and high balance categories. This can impact the probability of opening a term deposit account.
4- Our analysis indicates that marketing campaigns are particularly active during the months of May, June, and July.

Methods Comparison

1- The Decision Tree model outperformed other models in our study, making it the most effective option among the evaluated methods. 2- All four methods achieved overall accuracies greater than 95%, demonstrating strong predictive capabilities.
3- Crucial features impacting term deposit subscription include: 1- Last contact duration, 2- Number of days since the client was last contacted, 3- Number of contacts performed, and 4- Client's account balance.
4- A significant 34% difference was observed between the impact of contact duration and the number of contacts made.
5- Interestingly, factors such as housing or personal loans and the client's education level were found to be less significant in determining term deposit subscriptions. These findings provide valuable insights to help inform future marketing campaigns, enabling banks to optimize their strategies for increased success rates in selling term deposit products.

Analytical challenges

1- Since we don't have enough positive outcomes in your dataset for a binary classification problem, it can be challenging to train and evaluate a machine learning model. One technique that can be used in such cases is using decision trees, which can help mitigate the impact of imbalanced classes and provide more reliable estimates of model performance.

2- Interpretability of the results is a challenge because the decision tree model outperformed other models, but it may be difficult to understand why this was the case. Decision trees are known for their interpretability, but as the model becomes more complex, it can be challenging to understand the decision-making process. This is especially true for large datasets with numerous features, where decision trees can quickly become very complex. Thus, it may be difficult to explain why

certain features were given higher importance than others, which could limit the ability to make actionable insights.

3- Reproducibility of the results is another challenge because the analysis may not be easily reproducible due to factors such as changes in the dataset, the use of different machine learning models, or differences in software or hardware. This can make it difficult to verify or replicate the findings, which could limit the confidence in the conclusions drawn from the analysis. Reproducibility is an important consideration in data analysis, as it allows others to verify the results and build on them, potentially leading to new insights and improved methods.

Additionally, computational cost is also a challenge that can arise during data analysis. Certain machine learning algorithms can be computationally expensive, especially when working with large datasets with many features. This can limit the scalability of the analysis, making it difficult or impossible to work with larger datasets or to run the analysis in a reasonable amount of time. Therefore, it is important to carefully consider the computational cost of different methods when selecting an appropriate approach for a given dataset.

References 1. Bank Marketing Data Set,UCI Machine Learning Repository,https://archive.ics.uci.edu/ml/datasets/bank+marketing,2008

2. S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014 DOI:https://doi.org/10.1016/j.dss.2014.03.001 https://doi.org/10.1016/j.dss.2014.03.001

3. Bank Marketing Dataset,Kaggle,https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset?select=bank.csv,2017

4. Bank Marketing Analysis,Kaggle,https://www.kaggle.com/code/aleksandradeis/bank-marketing-analysis,2019