Text

Description automatically generated

**Data Management and Big Data - ALY6110**

**Final Project Draft Report**

**Group 5:**

Himanshu Kumar

Zihan Ma

**Date:** 6st Dec’2023

**Professor:** Dr. Hema Seshadri

**Introduction**

***(Data Source:*** *https://www.kaggle.com/datasets/krantiswalke/bankfullcsv?select=bank-full.csv****)***

The dataset, sourced from Kaggle, encompasses 45,211 entries, each representing an individual's information. It contains 17 variables that offer a broad range of demographic and behavioral data related to banking clients and their engagement with marketing campaigns, particularly focusing on term deposit subscriptions. The dataset delves into personal demographics, financial status, and banking behaviors, covering factors such as age, occupation, marital status, education level, credit default history, account balance, and loan status. Additionally, it includes details about communication methods and frequency, as well as information about past marketing campaigns and their outcomes. The 'Target' variable signifies whether a client has subscribed to a term deposit, serving as a crucial performance metric. Essentially, the dataset paints a comprehensive picture of banking clients, their financial habits, and their responses to marketing efforts. Leveraging this data could potentially optimize future marketing strategies, enhance customer service, and improve the offerings of financial products.

**Data Variables:**

Among the 17 variables in the dataset, there is a categorization into 10 categorical variables and 7 numerical variables, distributed as follows

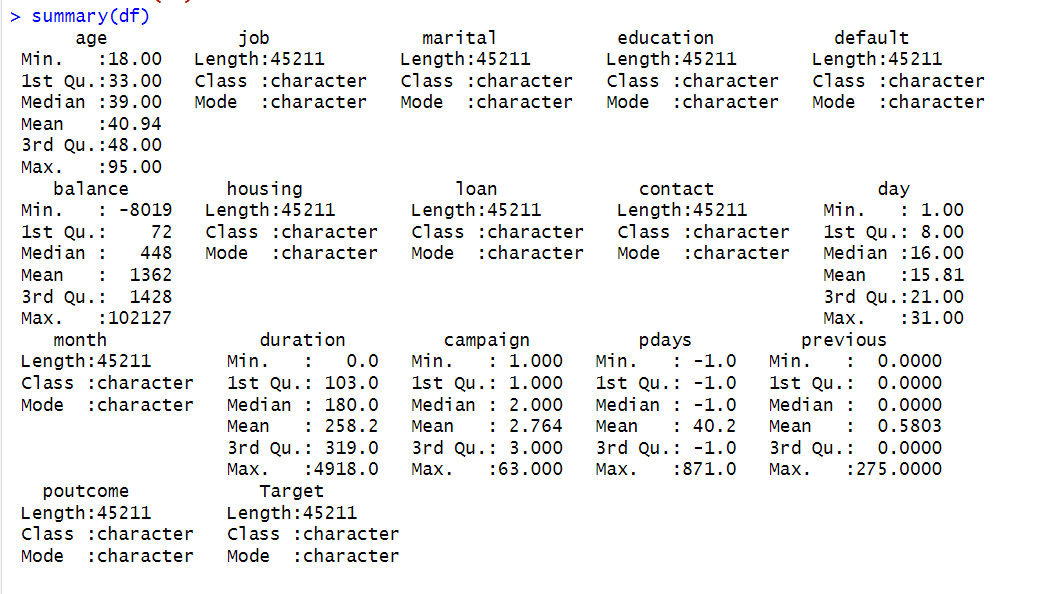
* **Categorical Variables**: Job, Education, Marital, Housing, Default, Loan, Month, contact, Poutcome, and Target.
* **Numerical Variables**: Age, Day, Balance, Duration, Pdays, Campaign and Previous.

**Objective and Framework:**

The primary aim of this project is to employ advanced data analytics techniques, including Logistic Regression, Decision Tree, and Random Forest models, to discern meaningful patterns within the extensive dataset obtained from Kaggle. With a focus on 45,211 individual entries, each representing a unique banking client, the goal is to unravel critical insights that can empower the bank in refining its marketing and promotional strategies. By leveraging demographic and behavioral data, spanning categories like age, occupation, marital status, education level, and financial indicators such as account balance and credit history, the objective is to identify distinct customer segments.

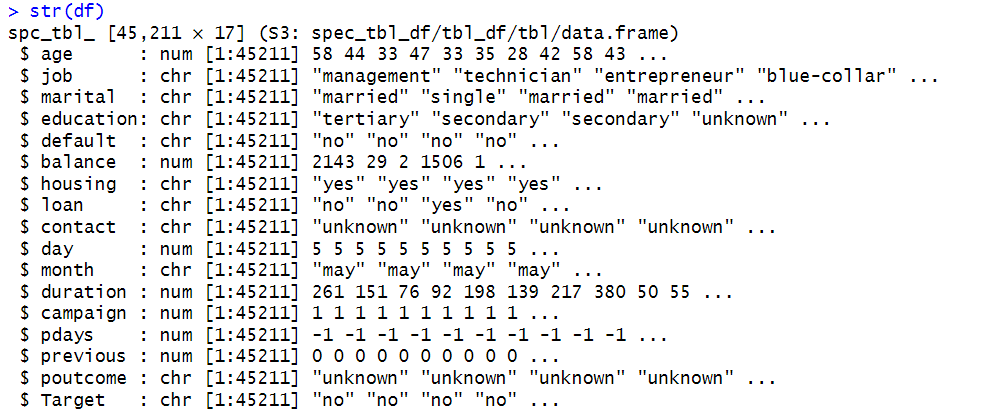
Through the application of machine learning models, we intend to develop predictive capabilities that enable the bank to anticipate which clients are more likely to subscribe to term deposits. This segmentation will aid in tailoring marketing campaigns to specific customer profiles, optimizing resource allocation, and enhancing the overall effectiveness of promotional efforts. Moreover, the project seeks to contribute to the bank's decision-making processes by providing actionable insights into customer behaviors, preferences, and responsiveness to marketing initiatives. Ultimately, the objective is to equip the bank with a data-driven strategy that not only increases the likelihood of successful term deposit subscriptions but also improves customer satisfaction and engagement with the bank's suite of financial products and services.

**Data Exploration:**

****

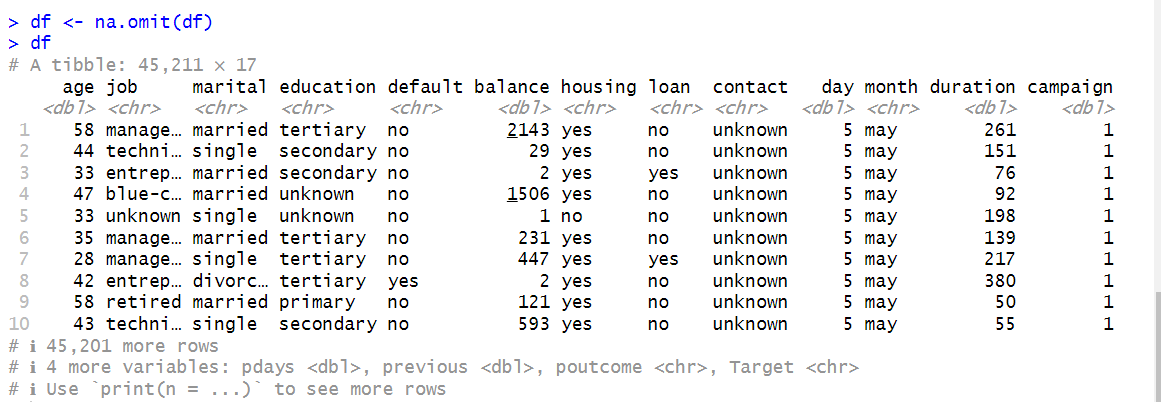
For the exploration of the dataset, the data was analyzed using R. The initial step involved loading the dataset head and gathering relevant information. Subsequently, descriptive statistics were performed to gain insights into the dataset. This involved examining measures such as mean, median, mode, and range to summarize the central tendency and spread of the data. Additionally, the standard deviation was calculated to understand the dispersion or variability within the dataset. These exploratory analyses were crucial in understanding the characteristics and distribution of the data, allowing for a thorough examination of the dataset. Overall, these steps laid the foundation for further analysis.

**Data Entries:**

****

The dataset, sourced from Kaggle, includes a substantial number of 45,211 entries, where each entry signifies a unique record. It encompasses 17 different variables, each offering various demographic and behavioral information about the individuals in question.

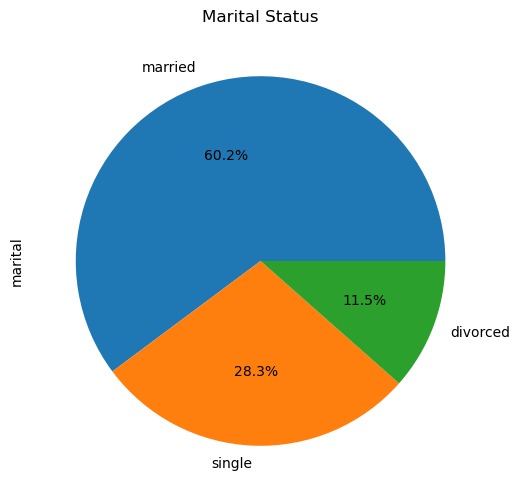
**Data Cleaning and Duplicate Values:**

****

The dataset we analyzed was very clean and well-organized. There were no missing values, which means we had all the necessary information for our analysis. This is important because missing data can cause problems and lead to inaccurate results. We also didn't find any duplicated entries, so each piece of information was unique and didn't repeat. This is good because it means we didn't have any unnecessary repetition in our dataset. Having clean and complete data will make our analysis more reliable and help us draw accurate conclusions from the dataset.

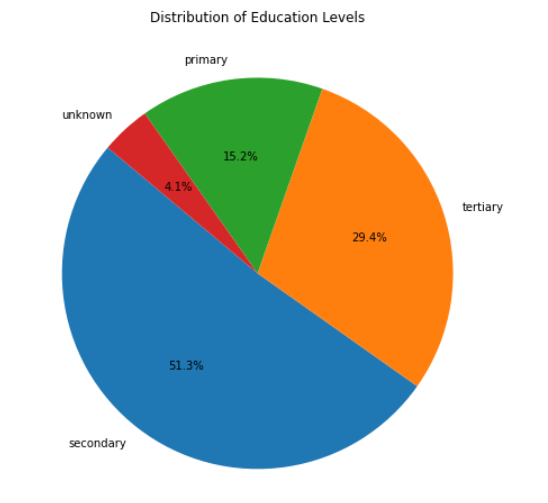
**Data Visualizations and EDA:**

1. **Pie Chart of Marital Status**



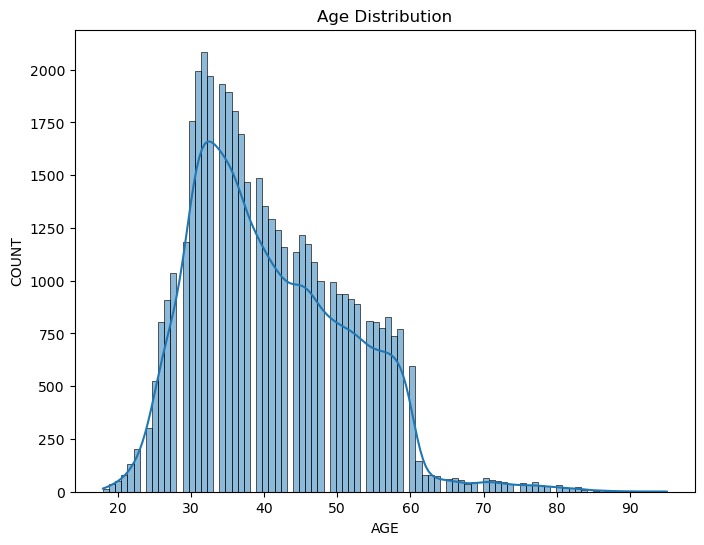
The above pie chart shows the marital status of all the customers who were contacted by the bank during their campaign. 60% of the customers are married while 28% are single. Also, 11% of the customers have their marital status as divorced. This breakup of marital status will help the bank to customize the services according to the needs of the customers such as type of loan and financial budget etc.

1. **Pie Chart of Education Level**



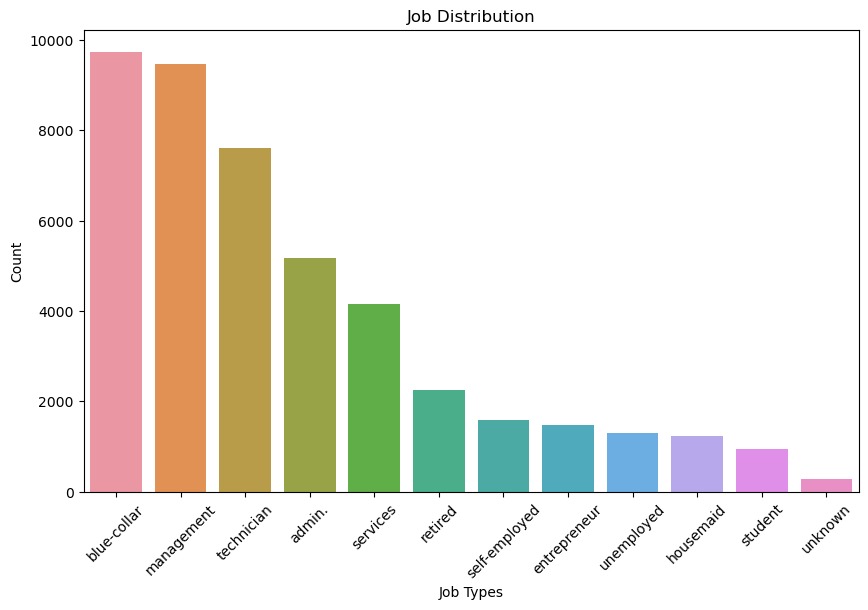
The above pie chart shows the education level of all the customers who were contacted by the bank during their campaign. 51% of the customers hold a secondary level of education while 29% of them have a tertiary level of education. Also, 15% of the customers have basic primary education while 4% of customers have their education level as unknown.

1. **Histogram of Count concerning age distribution**



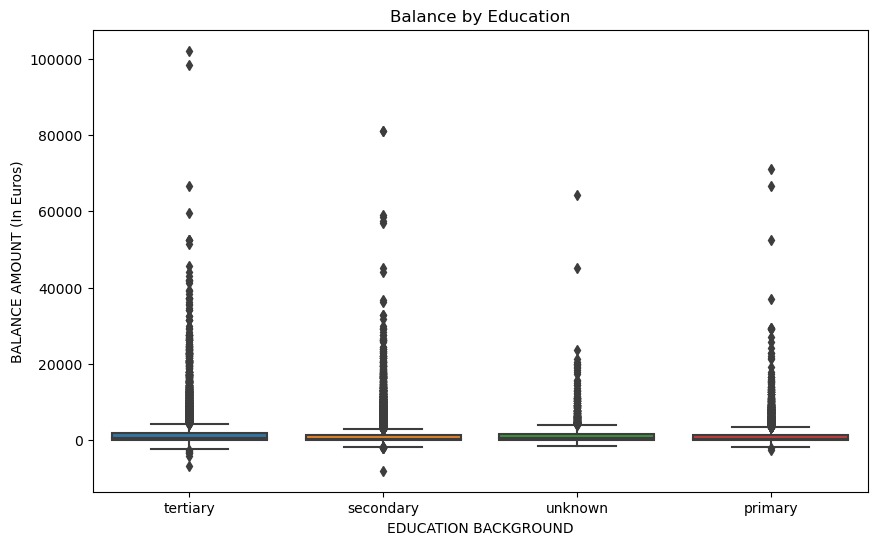
The above histogram shows how many clients are in different age groups. Most of the clients in the dataset are between 25 and 55 years old. This will help the bank to segment the target audience where there are the most clients and how to customize their services to enhance customer experience.

1. **Bar Plot of Count concerning Job Type**



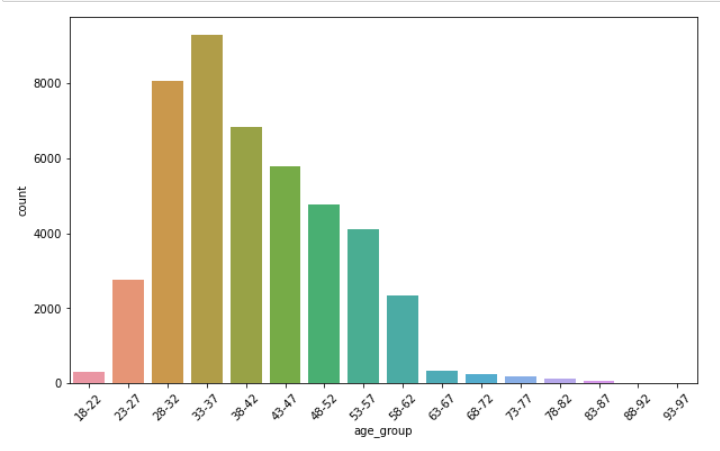
The above bar plot shows the breakup of occupation within 12 different categories. Most of the people do blue-collar jobs which count to around 9900 in number followed by customers practicing management in their jobs with almost the same number. The data has a good diversity of customers falling into different categories who have different financial backgrounds. The bank can use this information in shaping their campaigns based on the occupation followed by an individual.

1. **Box Plot of Balance Amount v/s Education Background**



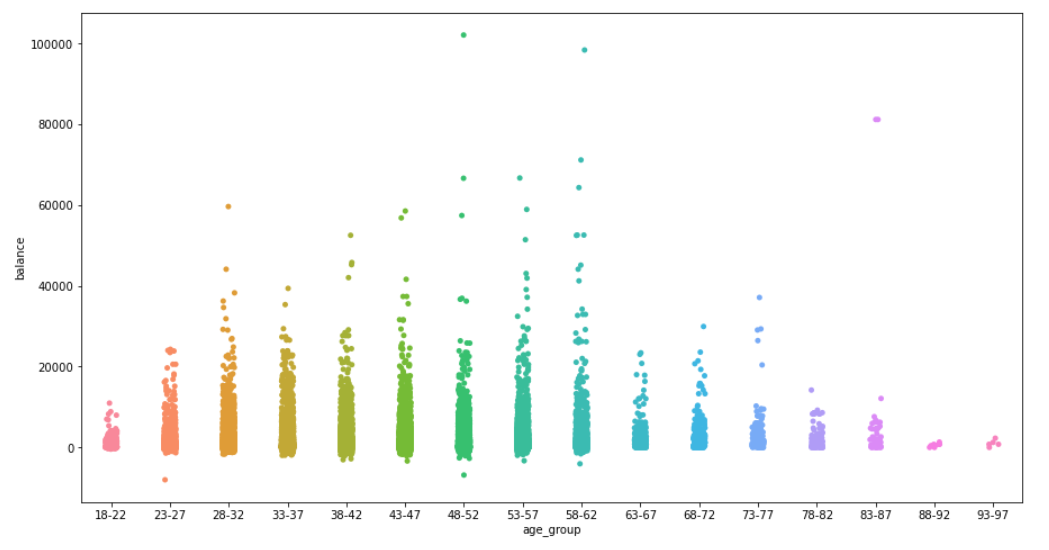
The boxplots depicting education categories (primary, secondary, tertiary, and unknown) against client balances reveal the presence of outliers in each group. These outliers represent individuals with exceptionally high or low balances based on their respective education levels. The boxplots effectively highlight the variations in balances within each education category, showcasing instances where certain clients have significantly different account balances compared to the majority within their educational group.

1. **Bar Chart of Customers (Age wise)**



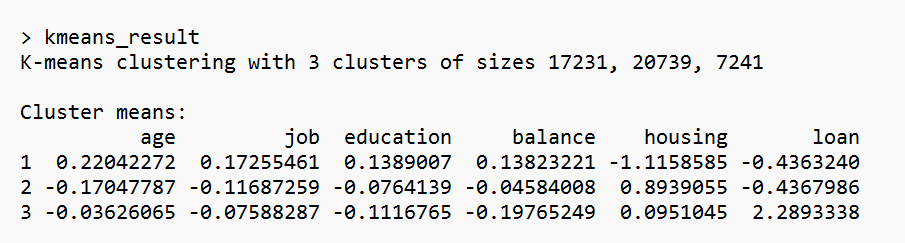
The above bar chart shows the age groups of the customers in the bank. The majority of the customers who were contacted during the marketing campaign belong to the age group of 33 to 37 years. This is a very important visualization from the bank’s perspective as they may segregate their customers based on the targeted campaign which may have a huge impact on that particular section.

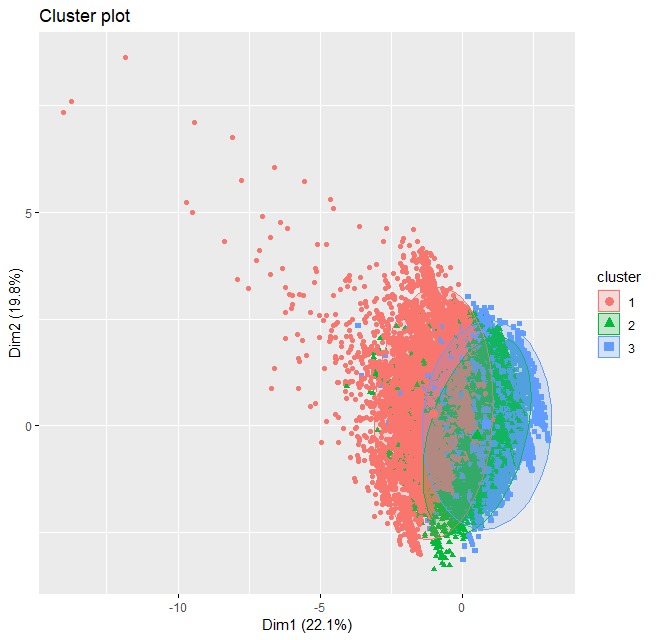
1. **Scatter Plot of Age Group and Balance**



Continuing the age group bar graph from the previous figure this scatter plot bar chart represents the age group of the customers on the x-axis and their bank balance on the y-axis. It can be seen that the bank balance is on the higher side by the age group of people belonging to 48 to 62 years. This means that bank can target their age group for their campaigns or schemes.

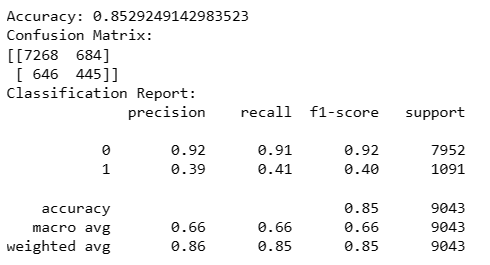
1. **Cluster Plot:**

****



So using the clustering method through R, we have developed three categories of customers based on variables like age, job, education, balance, housing, and loan. Each category represents the likelihood of a customer enrolling for the term deposit scheme that the bank is marketing to its clients. The customers who have the red dots are more prone to be enrolled in the bank scheme as they maintain a high level of balance and belong to a median age range. Customers with green dots depicting the borderline which can rest at the mid-level in the targeting funnel of the bank. The least likely customers are represented through blue dots, which the bank can omit from their list and can save their cost of marketing.

**Decision Tree Model:**

****

The Decision Tree model exhibited a commendable overall accuracy of approximately 85.29% on the test dataset, implying that it correctly predicted the term deposit subscription status for a substantial proportion of the clients. The confusion matrix provides further insight into the model's performance. In the context of binary classification (subscribed or not subscribed), the matrix reveals that the model made 7,268 correct predictions for clients who did not subscribe to a term deposit (True Negatives) and 445 correct predictions for clients who did subscribe (True Positives).

However, the model also made errors, misclassifying 684 instances where clients did not subscribe (False Positives) and 646 instances where clients did subscribe (False Negatives). The precision, recall, and F1-score metrics further elucidate the model's performance. Precision, representing the accuracy of positive predictions, is relatively high for clients not subscribing (92%), indicating a low rate of false positives. On the contrary, precision for clients subscribing is lower (39%), indicating a higher rate of false positives in this category.

Recall, representing the ability to capture all positive instances, is higher for clients not subscribing (91%) compared to clients subscribing (41%). The F1-score, a harmonic mean of precision and recall, reflects a balance between the two metrics. In summary, while the model demonstrates good overall accuracy, there is room for improvement in correctly identifying clients who are likely to subscribe to term deposits. Further model refinement, feature engineering, or the exploration of alternative algorithms may enhance predictive performance for the positive class.

**Random Forest Model:**

In this section, we explore the efficacy of a Random Forest model in addressing our core problem. The model construction involved tuning the number of trees (ntree), with a range from 100 to 500, to minimize the Out-of-Bag (OOB) error rate. The optimal ntree, determined to be 450, was chosen based on the stabilization of the OOB error rate, as illustrated in the accompanying plot. This tuning is crucial for balancing the model's complexity and generalization ability.

A graph with numbers and lines

Description automatically generated

The Random Forest model summary highlights an OOB error rate of 9.93%, with 450 trees and 3 variables tried at each split. The feature importance plot, integral to our analysis, identifies key variables driving the model's predictions. This insight is pivotal for understanding the underlying dynamics of our problem.

A screenshot of a computer

Description automatically generated

Evaluating the model's performance, we observe an accuracy of 90.1% and a Kappa statistic of 0.4011, as per the confusion matrix on the test data. These metrics underscore the model's proficiency in predicting outcomes, balancing sensitivity (97.35%) and specificity (34.95%).

Conclusively, the model's accuracy and insight into feature importance directly contribute to solving our problem. The high accuracy signifies reliable predictions, while the understanding of feature relevance aids in targeting key areas for intervention or further analysis. This synergy between accuracy and interpretability makes the Random Forest model a robust tool in our analytical arsenal.

**Logistic Regression Model:**

In this section, we delve into the Logistic Regression Model developed for our study. The process began with standardizing numerical variables to ensure uniform scaling. This standardization is vital for models like Logistic Regression, which are sensitive to variable scales.

A graph with red dots

Description automatically generated

Here, we examine two variations of Logistic Regression models in our analysis. The first model incorporates L1 regularization (Lasso), optimized through cross-validation to determine the best lambda value. This model achieved an accuracy of 89.86%, with a sensitivity of 97.69% and specificity of 30.29%, demonstrating a balanced prediction capability.

A screenshot of a computer

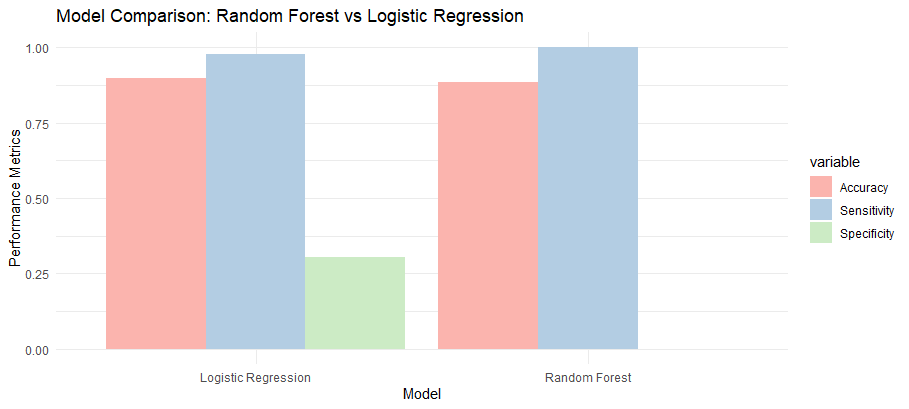
Description automatically generated

The second model involves further refinement through cross-validation based on the best lambda from Lasso regression. This approach slightly improved the model's performance, yielding an accuracy of 89.87%, with a sensitivity of 97.67% and specificity of 30.48%.

A screenshot of a computer program

Description automatically generated

Both models exhibit high accuracy and sensitivity, indicating their effectiveness in predictive analytics. The nuanced improvements in the second model underscore the value of meticulous cross-validation in enhancing model performance.



In our analysis comparing the Random Forest and Logistic Regression models, the performance metrics illustrate a nuanced picture. The Random Forest model exhibited a slightly higher accuracy (90.00%) compared to Logistic Regression (89.87%). However, Logistic Regression demonstrated marginally superior sensitivity (97.67%) over Random Forest (97.33%). In terms of specificity, Random Forest outperformed Logistic Regression with 34.19% against 30.48%. These findings suggest that while both models are closely matched in accuracy and sensitivity, Random Forest holds a slight edge in distinguishing negative cases, as indicated by its higher specificity.

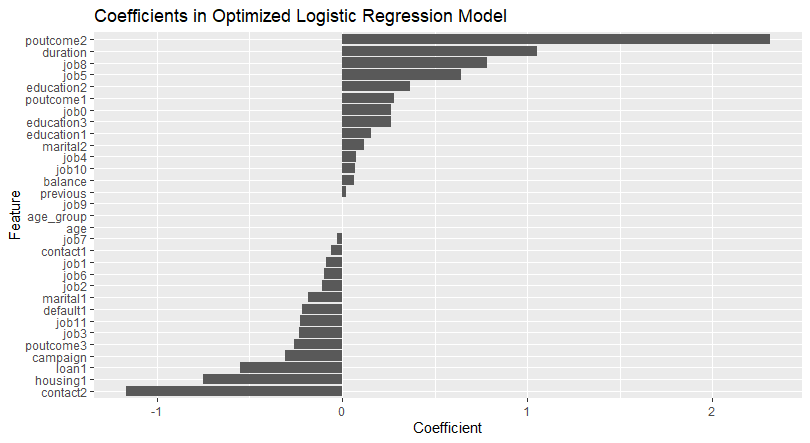
**Feature Importance:**

The Feature Importance plot for the Random Forest model and the Coefficients plot for the Optimized Logistic Regression model reveal different aspects of feature influence in each model.

A graph with lines and dots

Description automatically generated

Random Forest Feature Importance: This plot, based on Mean Decrease in Gini, shows that 'duration', 'balance', and 'age' are the most influential features. The Gini importance measures how each feature contributes to the homogeneity of nodes and leaves in the model. 'duration' has the highest score, indicating its strong impact on model decisions.



Logistic Regression Coefficients: The coefficients plot for Logistic Regression, which reflects the strength and direction of each feature's relationship with the target variable, shows 'poutcome2', 'duration', and 'job8' as the most significant. A high positive coefficient, like for 'duration', suggests a strong positive impact on the target variable, whereas a high negative coefficient, like for 'contact2', indicates a strong negative impact.

The differing results highlight the distinct ways each model interprets feature relevance. Random Forest assesses the features based on how well they improve the purity of the model splits, while Logistic Regression evaluates the direct influence of each feature on the predicted probability. This difference in interpretation is crucial for understanding how each model processes data and makes predictions.

**References:**

1. Scikit-learn. (2018). sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation. Scikit-Learn.org. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
2. Shin, T. (2021, February 26). Understanding Feature Importance and How to Implement it in Python. Medium. <https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285>
3. Portuguese Bank Marketing Data Set. (n.d.). Www.kaggle.com. <https://www.kaggle.com/datasets/yufengsui/portuguese-bank-marketing-data-set>

**‌**