Bank Marketing Data Analysis Project Report

Chak Pu Patrick Tong

Kexin Zhu

Oluwatosin Kajotoni

Vitaly Sukhinin

Contents

[1. Introduction 1](#_Toc163159802)

[1.1 Objective of the project 1](#_Toc163159803)

[2. Methodology 2](#_Toc163159804)

[2.1 Data Exploration and Cleaning 2](#_Toc163159805)

[2.2 Data Organization 2](#_Toc163159806)

[2.3 Visualization 2](#_Toc163159807)

[2.4 Outlier Handling and Data Normalization 3](#_Toc163159808)

[2.5 Categorical Data Encoding 3](#_Toc163159809)

[2.6 Learning Methods Used 4](#_Toc163159810)

[2.7 Machine Learning Implementation 4](#_Toc163159811)

[3. Result and Discussion 6](#_Toc163159812)

[3.1 Insight on Data 6](#_Toc163159813)

[3.2 Insight on Model Prediction 13](#_Toc163159814)

[3.3 Conclusion 15](#_Toc163159815)

# 1. Introduction

The essence of this project lies in its commitment to harnessing the power of predictive analytics in the realm of financial marketing. By analyzing the data of a Portuguese bank's marketing campaigns, we aim to forecast client engagement with term deposit subscriptions—a key indicator of the campaign's success.

## 1.1 Objective of the project

The primary objective of this project was to analyze direct marketing campaigns, via phone calls, by a Portuguese banking institution. Specifically, we aimed to develop a predictive model to forecast whether a client would subscribe to a term deposit. This initiative could potentially enhance the efficiency of future marketing strategies by enabling the institution to target individuals more likely to respond positively.   
  
This report encapsulates the rigorous processes of data cleaning, preprocessing, machine learning modeling, and evaluation, underscoring the strategic decisions made to ensure data integrity, model robustness, and insightful visualizations. Each step was executed to address specific challenges within the dataset and ultimately, provide actionable recommendations for targeted marketing strategies.

# 2. Methodology

The dataset of this project included a mix of numerical and categorical attributes, such as age, job type, marital status, education, contact details and other circumstances, culminating in the binary target attribute indicating whether a client subscribed to a term deposit.  
  
In this project, Data cleaning was a crucial step that involved encoding categorical variables into a format suitable for machine learning models and handling missing values marked as "unknown."  
These steps addressed missing values, outliers, and the encoding of categorical variables, as detailed below.

## 2.1 Data Exploration and Cleaning

Conversion of Unknown Values: This initial step involved the transformation of 'unknown' entries across the dataset to np.NaN, enabling us to employ uniform data handling techniques.

Assessment of Missing Data: An exhaustive review quantified missing values, utilizing .isnull() and .sum() to inform strategic decisions for data imputation or omission.

Data Reduction: Approximately 7% of data records exhibiting missing values were removed, with careful consideration to maintain the balance of the dataset, particularly the representation of our target variable.

## 2.2 Data Organization

The dataset's features were systematically classified into integer, nominal, and ordinal groups to facilitate tailored preprocessing and ensure their appropriateness for machine learning applications.

## 2.3 Visualization

Visualization plays an indispensable role in both the exploratory phase and the presentation of results in data analysis. In this project, visualization served multiple purposes:

Data Understanding: Initial visualizations such as histograms, bar charts, and scatter plots were employed to explore the distributions and relationships between different variables. These visuals were pivotal in identifying patterns, trends, and anomalies within the dataset.

Insight Communication: The visualization of the model's results was crucial in elucidating complex findings. For example, ROC curves were used to compare the performance of different classifiers, while confusion matrices offered insights into the true positives and false negatives yielded by the models.

Feature Importance: Tree-based models like the Random Forest Classifier provided an intrinsic method of determining the importance of each feature. Visualizing these importances helped in understanding which factors were most predictive of a client's likelihood to subscribe to a term deposit, enabling a more focused approach to future campaigns.

Through strategic visualization, the project transformed raw data into a compelling narrative, providing stakeholders with intuitive access to complex data insights.

## 2.4 Outlier Handling and Data Normalization

Outliner treatment: A outliner treatment is created to capping all the outliner using IQR method with upper bound and lower bound method to optimize the data.

Standardization: Numerical features were standardized to have zero mean and unit variance, ensuring that no single feature would dominate the model's learning process due to its scale.

## 2.5 Categorical Data Encoding

The categorical data underwent a two-fold encoding process:

**Categorical Data Encoding**

Encoding Strategy: Categorical variables were encoded to numeric formats, with ordinal data encoded using Ordinal Encoding and nominal data through One-Hot Encoding. This distinction preserved the inherent order in ordinal data while treating nominal data as separate binary features.

Implementation: The encoding process converted categorical columns to a format compatible with machine learning algorithms, using scikit-learn's OneHotEncoder and OrinalEncoder for efficient transformation.

**Concatenation of Processed Features**

After separately processing numerical and categorical features, the data was recombined into a single dataframe. This consolidated dataset, now devoid of missing values, outliers, and with all categorical variables suitably encoded, was prepared for machine learning modeling.

## 2.6 Learning Methods Used

For the case, due to the nature of the task is related to binary classification. The selection of learning algorithm of the following was used:

Logistic Regression

Logistic Regression is a regression fitting on the logistic function which will predict the possibility of a data to be in 1 or 0, creating a threshold and predict the output.

Random Forest Classification

Random Forest Classification mean performing the Decision Tree Classifier on with different subset of data on the same predictor.

## 2.7 Machine Learning Implementation

The application of machine learning was the cornerstone of this project. The following steps were taken to ensure the implementation was as robust and effective as possible:

Preprocessing for Machine Learning: Before the Machine learning, Data Cleaning process has been performed to ensure the effect of specific data has reduced to minimal. After that both models will be fit with the train data and to observe the result on rest data.

Ensemble Techniques: The use of ensemble methods such as the Random Forest Classifier provided a more accurate and stable model by combining the predictions of several decision trees and reducing variance.

Performance Metrics: A variety of metrics were used to evaluate the models. Accuracy was measured to see the overall level of correct classifications, while the ROC curve and the area under it (AUC) were used to assess the trade-offs between true positive rate and false positive rate at various threshold settings.

Model Comparison and Selection: The performance of different models was compared robustly, considering both statistical metrics and business objectives. The model that not only provided the best statistical results but also aligned with the pragmatic goals of the campaign was selected.

By rigorously applying machine learning techniques, the project ensured that the models developed were both statistically sound and relevant to the practical needs of the marketing campaign

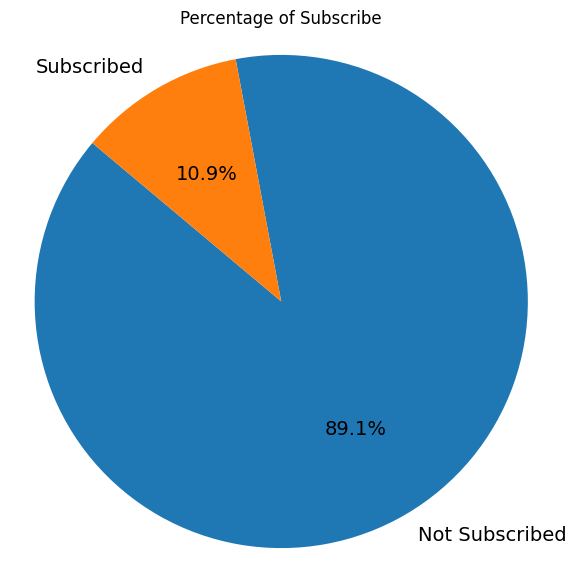
# 3. Result and Discussion

## 3.1 Insight on Data

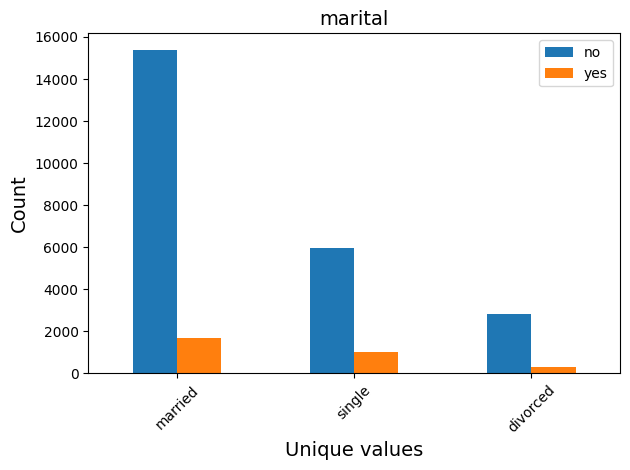
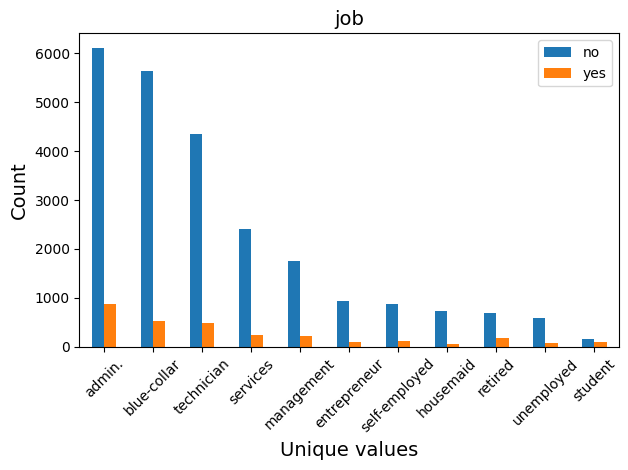
The variables in the dataset can be divided into two types: categorical and numerical.

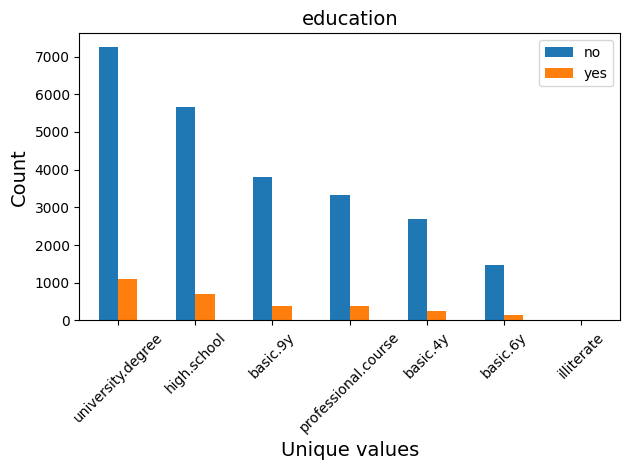
**Categorical variables analysis:**

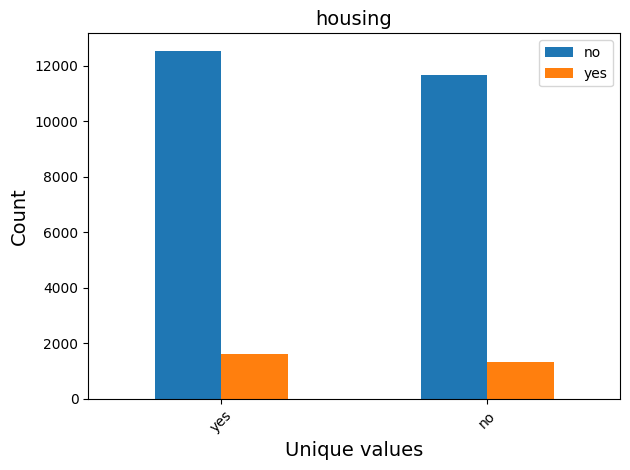
The pie chart demonstrates that 89% of the observations are “Not subscribed” as compared to just 11 % data as “Subscribed”. Clearly this dataset is an imbalanced dataset.

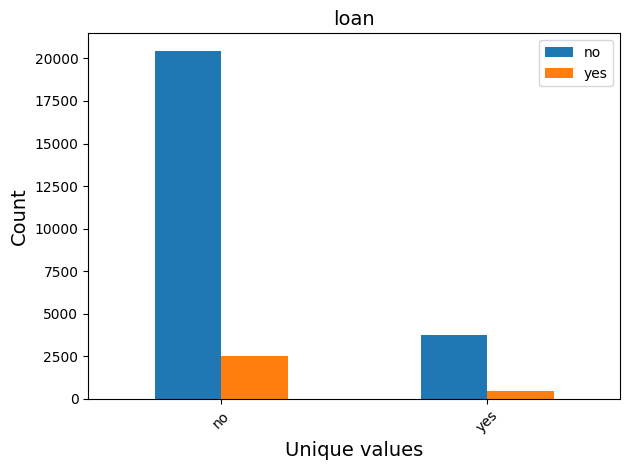


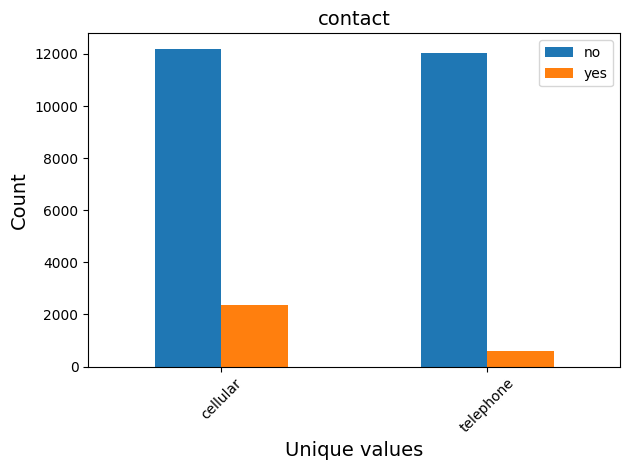
After plotting the histograms of each variable, the graphs are shown below:

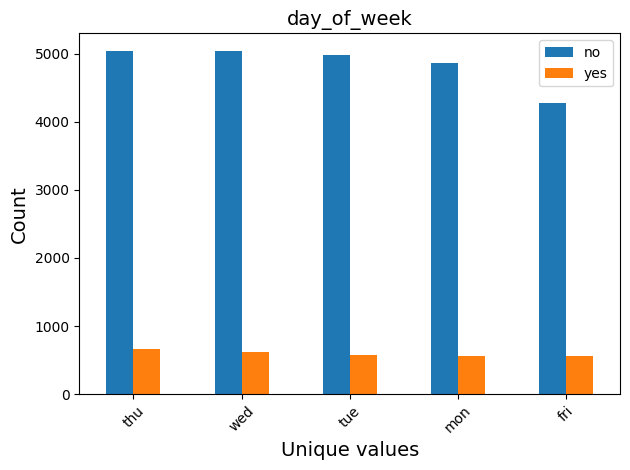


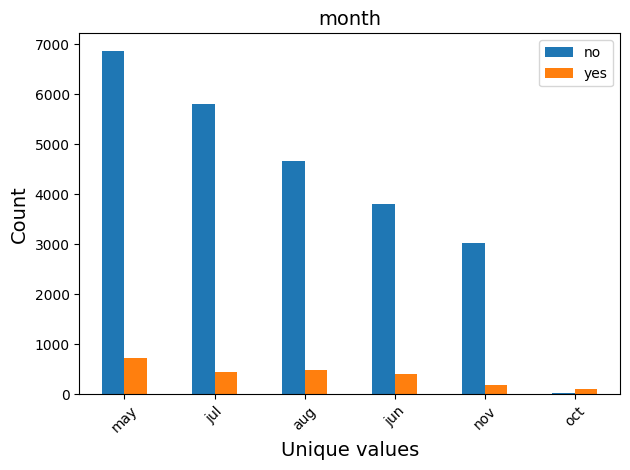


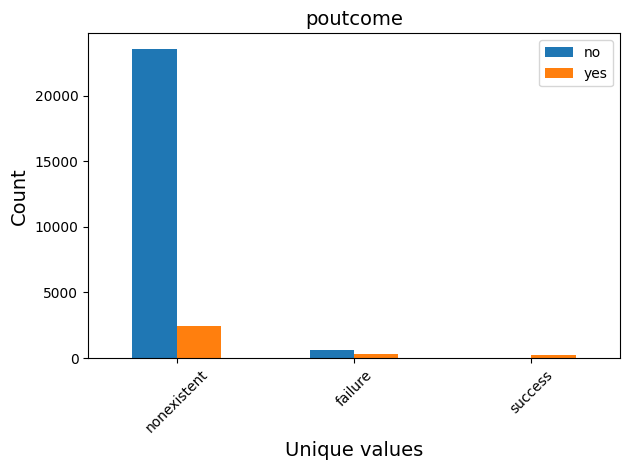












By analyzing these histograms, we can gain a deeper understanding of how different categorical features relate to the likelihood of subscription.

Many features are seems not related with subscription. Including “housing”, “loan”, “day\_of\_week”.

Some important features are related:

**Job**: retired and students are more likely to subscribe.

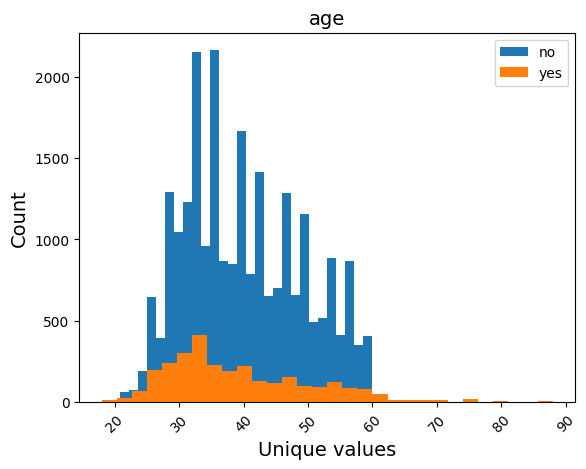
**Marital**: people who are not in a marriage are more likely to subscribe.

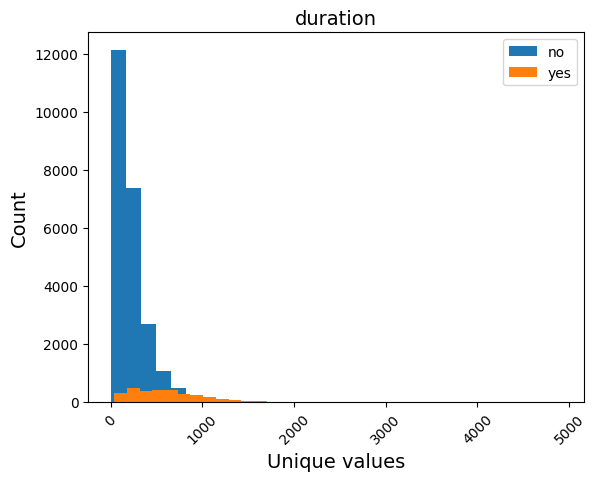
**Contact**: contact by cellular are more likely to subscribe.

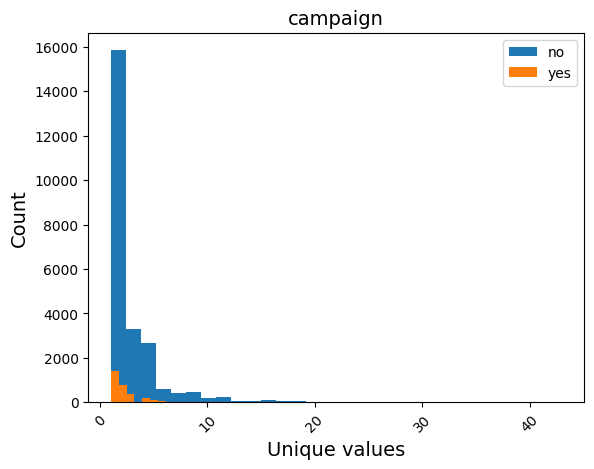
**Month**: Seasonal effect can been seen in the data.

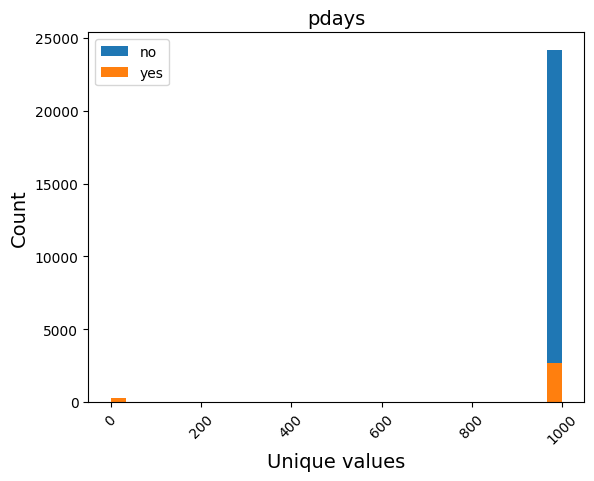
**Numerical variables analysis:**

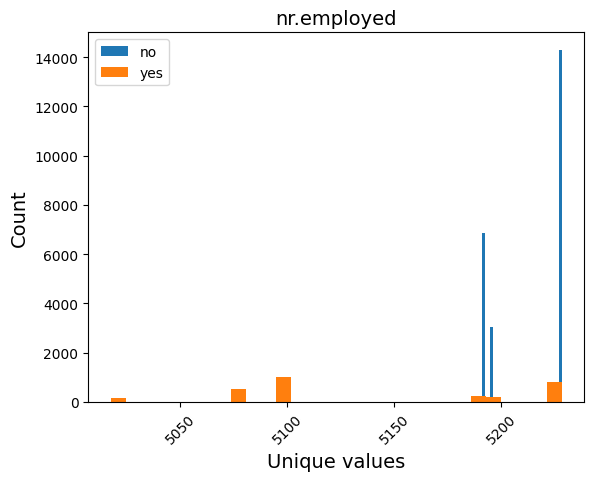
After plotting the histograms of each variable, the graphs are shown below:









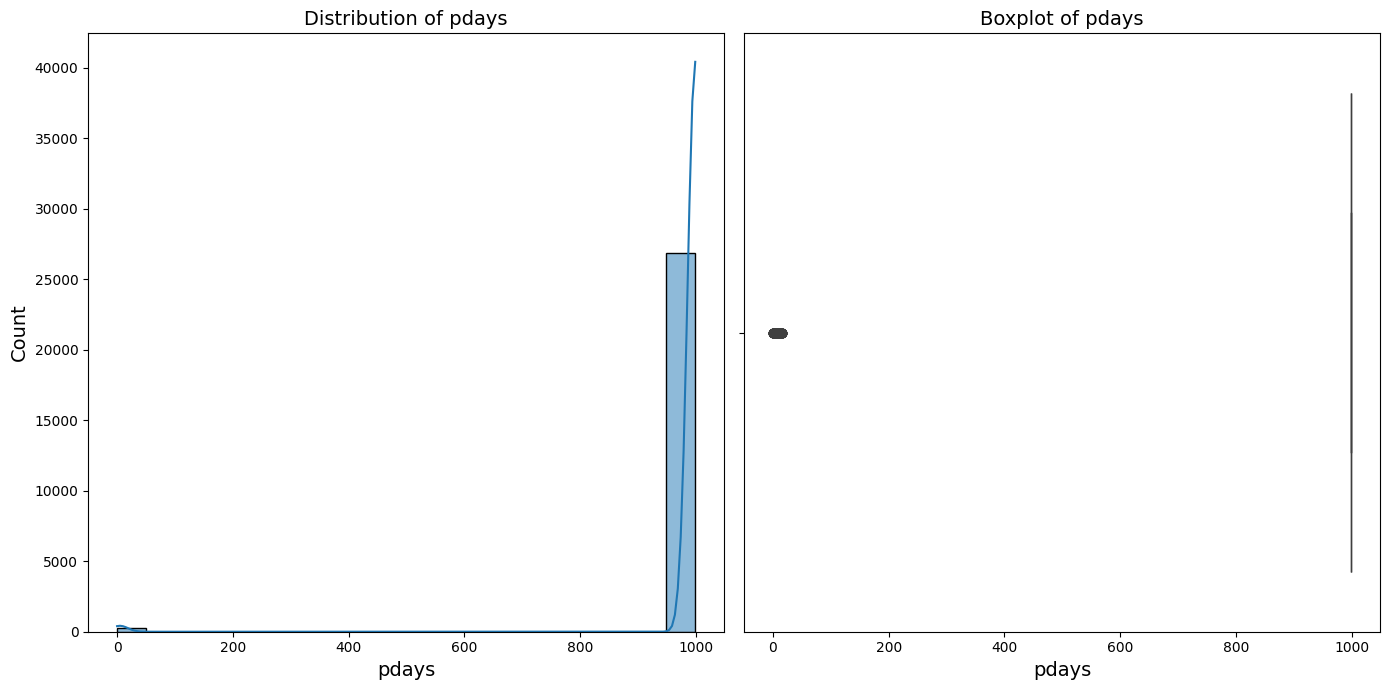


People under 25 and over 60 are more likely to subscribe, this also matches the conclusion from “job”.

Contact duration seems important. The longer the call last the more likely to subscribe.

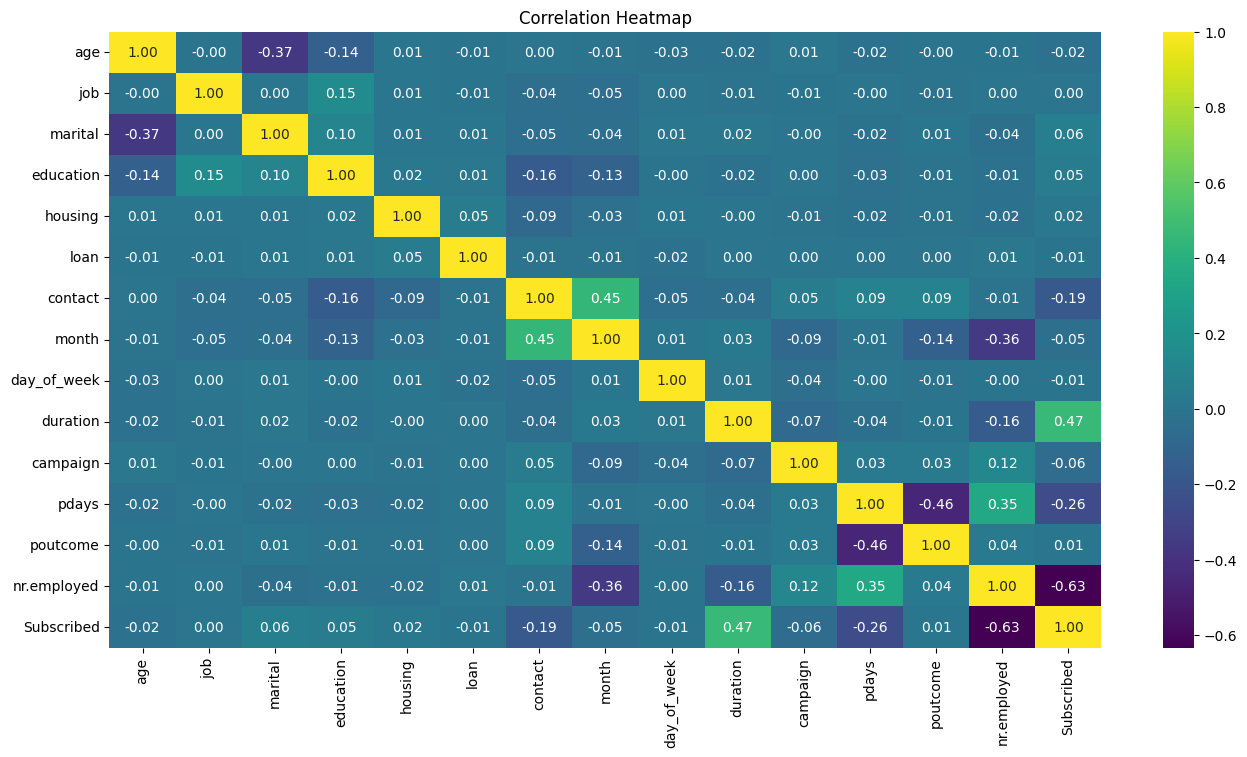
People who subscribed were exposed to fewer campaigns than those who didn't.

Since the very majority is not previously contacted, cannot get a conclusion from “pday”.



**Correlation analysis:**

To generate the heatmap, we encoded all categorical variables. The graph is shown below:



From the graph we can sort the variables that infect the subscription.

From the heatmap, we can see that “subscribed” (our target variable) has good correlation with “nr.employed”, “duration”, “pdays” and “contact”. We expect to see these independent variables as significant while building the models.

## 3.2 Insight on Model Prediction

**Model prediction outcome:**

After performing both the regression of the confusion matric for both model is shown below:

A diagram of a logistic regression

Description automatically generatedA blue squares with white text

Description automatically generated

3.2.1 Confusion Matrix for both model

For both model, both model has a similar output on the test set while the Logistic Regression has a high count on the FP index while the Random Forest has a higher count of FN.

The Accuracy score for both model in test set is very similar both 95.5% and 96.6% score for Logistic Regression and Random Forest Classification respectively.

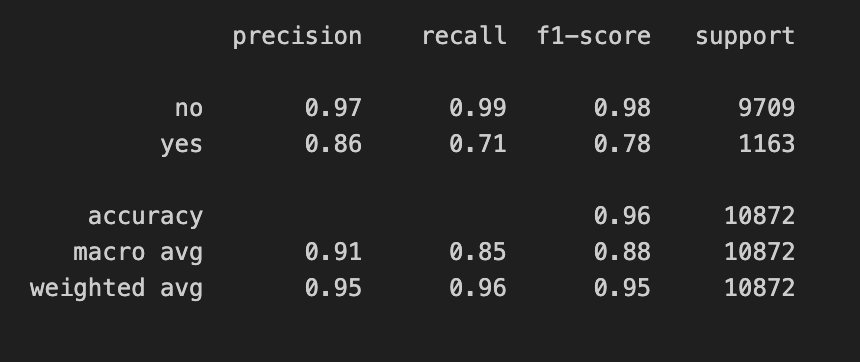
As for the Precision Score for the Logistic Regression is lower with 74.98% while the Recall score is 84.78% and with a F1 score of 79.58%.

A screenshot of a computer screen

Description automatically generated

3.2.2 Classic Report of the Logistic Regression Model

While the Precision Score for the Random Forest is 85.27% and the Recall score is 71.19% with the F1 score is 77.70%.



3.2.3 Classic Report of the Random Forest Model

The Recall indicate the percentage score that the correct data over to the all Actual true data while Precision score is the correct label over all the yes label. And F1 score is the overall weight balance of both score. So overall the Logistic Regression is perform better than Random Forest Model in term to precision and recall score.

The major reason for the lower precision and recall compare to the high accuracy result with Random Forest may due to overfitting with a very high precision score on the train set.

Below is the ROC and AUC curve for both of the Regression Model:

A graph with a line and a point

Description automatically generated

A graph with a line

Description automatically generated

With the ROC-AUC of the Logistic Regression is higher compared to Random Forest Classification model. I also support that the Logistic Regression are able to have a ability to create a correct classification.

## 3.3 Conclusion

In conclusion, this project perform a predict financial marketing which focus on direct marketing campaigns by a Portuguese bank. Through rigorous data analysis, preprocessing, and machine learning modeling, insightful patterns emerged regarding client engagement with term deposit subscriptions.

The data exploration revealed a number of insights into the dataset, highlighting the imbalance in subscription outcomes, significant features influencing subscription likelihood, and correlations between variables such as contact duration and subscription rates. In order to perform a robust prediction of the marketing campaign effectiveness Two machine learning models, Logistic Regression and Random Forest Classification, were employed and thoroughly evaluated.

While both models showcased high accuracy rates, Logistic Regression outperformed Random Forest in terms of precision and recall scores. Despite Random Forest's high accuracy, its lower precision and recall suggest potential overfitting issues. The ROC-AUC analysis further reinforced the superiority of Logistic Regression in correctly classifying outcomes.

Overall, the project underscores the importance of robust data preprocessing, careful model selection, and thorough evaluation metrics in harnessing the power of predictive analytics for targeted marketing strategies. Also it provide an insight of the effectiveness of the campaign and provide a parameter for stakeholder to compare with other campaign on the market.