1. **INTRODUCTION**

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking Institution. The classification goal is to predict if the client will subscribe a term deposit (variable y). A Term deposit is a deposit that a bank or a financial institution offers with a fixed rate (often better than just opening deposit account) in which your money will be returned at a specific maturity time. For more information with regards to Term Deposits please click on this link from Investopedia: <https://www.investopedia.com/terms/t/termdeposit.asp>

**Data Description**

This is the classic marketing bank dataset uploaded originally in the UCI Machine Learning Repository. The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyse to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

**Feature**

1. age | int64 | age in years
2. job | object | type of job (categorical: ['admin.' 'technician' 'services' 'management' 'retired' 'blue-collar' 'unemployed' 'entrepreneur' 'housemaid' 'unknown' 'self-employed' 'student'])
3. marital | object | marital status (categorical: ['married' 'single' 'divorced'])
4. education | Object | education background (categorical: ['secondary' 'tertiary' 'primary' 'unknown'])
5. default | Object | has credit in default? (categorical: ['no' 'yes'])
6. balance | int64 | Balance of the individual
7. housing | object | has housing loan? (categorical: ['yes' 'no'])
8. loan | object | has personal loan? (categorical: ['no' 'yes'])
9. contact | object | contact communication type (categorical: ['unknown' 'cellular' 'telephone'])
10. day | int64 | last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11. month | object | last contact month of year (categorical: ['may' 'jun' 'jul' 'aug' 'oct' 'nov' 'dec' 'jan' 'feb' 'mar' 'apr' 'sep'])
12. duration | int64 | last contact duration, in seconds (numeric)
13. campaign | int64 | number of contacts performed during this campaign and for this client
14. pdays | int64 | number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
15. previous | int64 | number of contacts performed before this campaign and for this client
16. poutcome | object | outcome of the previous marketing campaign (categorical: ['unknown' 'other' 'failure' 'success'])

**Label**

1. deposit | object | has the client subscribed a term deposit? (binary: 'yes','no')

**DATA CLEANING**

The First step after importing the CSV file is to find the missing values and perform some data manipulation in other to make our data clean and ready to use for the analysis. We performed some basic data cleaning and manipulation by cleaning the Quotation marks in the data while importing and converting the data types to the appropriate data types, we then perform our days mapping to set the dates to the correct format. These preprocessing steps are common when preparing a dataset for analysis or machine learning, ensuring that data types are appropriate for analysis and handling any inconsistencies in the data.

**DATA VISUALISATION**

After the cleaning of the data, we start performing our Visualisation and doing our summary and descriptive statistics to get a real feel of the data, we created boxplots, bar plots and line chart to check the relationships between the variables that are going to assist us in predicting the term deposit.

A graph of ageing

Description automatically generated

*Fig 1: Histogram of Age variable*

Another step that we must do is to visualize our categorical variables, this shows the spread of our categorical variables across each other and as well the relationship between the categorical variables and the label variable.

A close-up of a graph

Description automatically generated

*Fig2: Distribution of some categorical variables*

Furthermore, we would carryout a Correlation analysis in order to try and see which of our variables have a strong correlation with each other, this would give us some clearer insight as to which data we need to drop due to the multi-collinearity effect.

A screenshot of a graph

Description automatically generated

*Fig3: Correlation Matrix*

This correlation plot will help you understand the relationships between different numerical variables in your dataset. Positive values indicate positive correlation, negative values indicate negative correlation, and values close to zero suggest a weak correlation. The strength and direction of the correlation are represented by the color intensity in the heatmap.

The step is called the Exploratory data Analysis, after we have done the EDA we then begin with the Feature Engineering.

**FEATURE ENGINEERING**

This process involves deriving a subset of our whole population which would be used for our analysis purposes, from our data we a going to get a 20% subset of this from this subset we would use to train our model by making the train and test labels. That is going to be the first step to go through before running any of the models which we are going to be using to get our findings.

* **HANDLING IMBALANCED DATA**

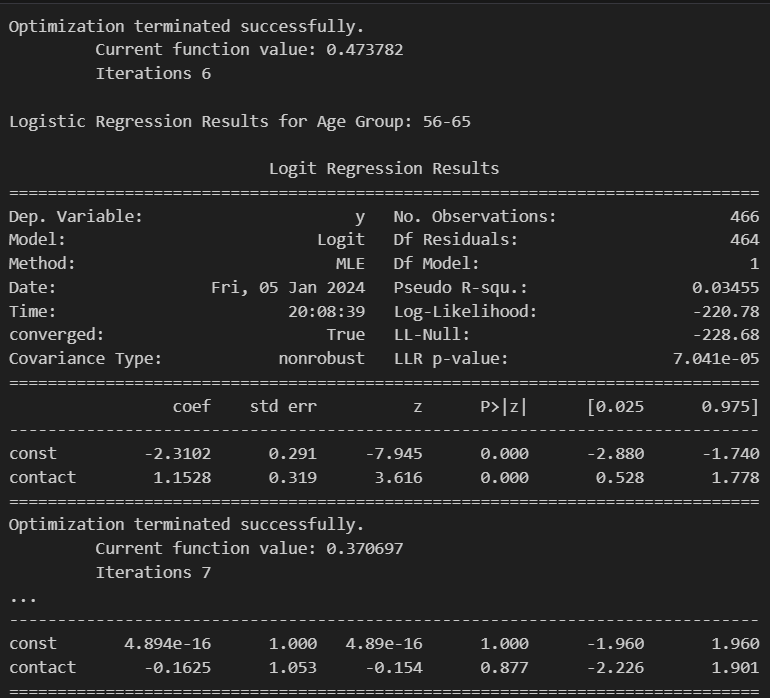
***Undersampling the majority class:*** When dealing with an imbalanced dataset, where one class (e.g., 'yes' or 'no') significantly outnumbers the other, it can affect the performance of machine learning models, especially those that are sensitive to class distribution. Henceforth, we do the Under sampling of the majority class to reduce the number of instances of the majority class by randomly removing samples.

* **Discretize Continuous Variables:** Another action performed was to ensure that the continuous variables were discretized.

**TRAINING THE MODEL**

This part of the data engineering process involves splitting the data into the training dataset and the test dataset which would be split into 80 to 20 percent. When that is done we then scale our variables so that we can then build a Logistic regression Model which would assist us in finding answers to the question we need to.

**RESULTS AND FINDINGS**

****

*FIG4: LOGISTIC REGRESSION MODEL*

*A screenshot of a computer screen

Description automatically generated*

*A screenshot of a computer

Description automatically generated  
FIG 5: Odds Ratio coefficients*

**QUESTION 1:**

**Do the job backgrounds and education levels affect customers’ responding rate significantly?**

From the above result we can conclude that there is Significant Impact of Individuals in the "Retired" job category have significantly higher odds of responding. It is also note-worthy that "Blue-collar" and "Management" job categories have a significant impact on reducing the odds of response.

We also can deduce the effect of Education Levels on responding rate, we would find out that "Basic.4y" education level has a significant impact on reducing the odds of response.

"Professional Course," "University Degree," and "Unknown" education levels have a significant impact on increasing the odds of response.

**QUESTION 2:**

**Is there a relationship between customers' response rates and the contact communication type?**

From the information garnered from the Odds ratio, Individuals contacted through "Cellular" communication have approximately 6.98% higher odds of responding compared to "Telephone" communication.

**QUESTION 3:**

**Is the contact communication type -responses relationship identified in the previous question the same for all age groups?**

To answer this question, we divided our ages into different age groups, and we ran a logistic regression model to assess the interaction effect between "Age" and "Contact Communication Type." This involves looking at the coefficients for both variables and their interaction term. If the interaction term is significant, it suggests that the relationship between contact communication type and response varies with age.

Age Group: 56-65

* **Constant (Intercept):** -2.3102
* **Contact Coefficient:** 1.1528
* **Interpretation:** In this age group, the log-odds of the target variable (y) decreases by -2.3102 when the "contact" variable is zero The log-odds increase by 1.1528 for a one-unit increase in the "contact" variable.

Age Group: 26-35

* **Constant (Intercept):** -2.7401
* **Contact Coefficient:** 1.0490
* **Interpretation:** In this age group, the log-odds of the target variable (y) decreases by -2.7401 when the "contact" variable is zero The log-odds increase by 1.0490 for a one-unit increase in the "contact" variable.

Age Group: 36-45

* **Constant (Intercept):** -2.8849
* **Contact Coefficient:** 0.9382
* **Interpretation:** In this age group, the log-odds of the target variable (y) decreases by -2.8849 when the "contact" variable is zero The log-odds increase by 0.9382 for a one-unit increase in the "contact" variable.
* .

Age Group: 0-25

* **Constant (Intercept):** -1.7272
* **Contact Coefficient:** 0.9540
* **Interpretation:** **:** In this age group, the log-odds of the target variable (y) decreases by -1.7272 when the "contact" variable is zero The log-odds increase by 0.9540 for a one-unit increase in the "contact" variable.

Age Group: 46-55

* **Constant (Intercept):** -2.5649
* **Contact Coefficient:** 0.7468
* **Interpretation:** In this age group, the log-odds of the target variable (y) decreases by -2.5649 when the "contact" variable is zero The log-odds increase by 0.7468 for a one-unit increase in the "contact" variable.

Age Group: 66-75

* **Constant (Intercept):** -1.6094
* **Contact Coefficient:** 1.7864
* **Interpretation:** In this age group, the log-odds of the target variable (y) decreases by -1.6094when the "contact" variable is zero The log-odds increase by 1.7864 for a one-unit increase in the "contact" variable.

Age Group: 76-100

* **Constant (Intercept):** ~4.894e-16 (close to zero)
* **Contact Coefficient:** -0.1625
* **Interpretation:** Due to the small magnitude of the constant, the log-odds for this age group is close to zero. The contact coefficient is -0.1625, indicating a small decrease in the log-odds for a one-unit increase in the "contact" variable.

**KEY INSIGHTS AND CONLUSION**

There are key insights which we could observe from the data and that includes the feature importance of the variables.

A graph of a number of blue bars

Description automatically generated with medium confidence

The most important features are on the left, and importance decreases as you move to the right. The x-axis shows the names of the features**.** The heights of the bars are proportional to the importance of the corresponding features relative to each other. It is good to know that Feature importances represent the relative importance of each feature in predicting the target variable and Features with higher bars have higher importance according to the XGBoost model.