

# Bank Marketing Analysis

---

## Data Science project

**By:**

| Name                         | Section | ID |
|------------------------------|---------|----|
| Bodour Mohamed Ahmed Alsisy  | 1       | 28 |
| Aya Abdelhameed Mohamed Gad  | 1       | 24 |
| Amaal Nader Sabri Abdullah   | 1       | 21 |
| Alaa Naser Ezat Abu-elsaud   | 1       | 19 |
| Aya Naser Ezat Abu-elsaud    | 1       | 27 |
| Ebtesam Saber Abdelsatar Ali | 1       | 2  |
| Hekmet Reda Samir            | 2       | 34 |

**Under supervision of**

**Eng. Farah Mohamed**

## Table of Contents

1. Problem Domain
  - 1.1 Objective
  - 1.2 Dataset Overview
2. Project Workflow Summary
  - 2.1 Data Acquisition
  - 2.2 Data Preprocessing
  - 2.3 Exploratory Data Analysis (EDA)
  - 2.4 Feature Engineering
  - 2.5 Model Building and Evaluation
  - 2.6 Clustering Analysis (Bonus)
3. Source Code
  - 3.1 Repository Link
4. Visualization Snapshots
  - 4.1 Age Distribution
  - 4.2 Subscription Trend by Month
  - 4.3 Education Level Distribution
  - 4.4 Job Category Distribution
  - 4.5 Housing Loan Distribution
  - 4.6 Age vs Loan Status
  - 4.7 Correlation Matrix
  - 4.8 Confusion Matrix
  - 4.9 KMeans Clustering Results
5. Conclusion
  - 5.1 Model Performance Summary
  - 5.2 Feature Importance
  - 5.3 Clustering Insights

## Section 1: Problem Domain

The **Bank Marketing** dataset, sourced from the UCI Machine Learning Repository , pertains to direct marketing campaigns conducted by a Portuguese banking institution. These campaigns primarily involved phone calls to clients, aiming to persuade them to subscribe to term deposits.

**Objective:** Develop a predictive model to determine whether a client will subscribe to a term deposit based on their demographic information, past interactions, and campaign-related attributes.

### Dataset Overview:

- **Instances:** 41,188
- **Features:** 20 (comprising both numerical and categorical variables)
- **Target Variable:** y (binary: 'yes' or 'no')

## Section 2: Project Workflow Summary

The project followed a structured data science pipeline encompassing the following stages:

### 1. Data Acquisition:

- Utilized the ucimlrepo Python package to fetch the dataset.

### 2. Data Preprocessing:

- Handled missing values using the most frequent strategy for categorical columns.
- Removed the poutcome column due to its high percentage of missing values.
- Addressed duplicate entries and ensured data cleanliness.

### **3. Exploratory Data Analysis (EDA):**

- Conducted univariate and bivariate analyses using histograms, boxplots, and count plots.
- Identified and treated outliers using the Interquartile Range (IQR) method.
- Performed correlation analysis to understand relationships between variables.

### **4. Feature Engineering:**

- Created new features such as no\_previous\_contact and total\_contacts.
- Applied transformations to variables like pdays and previous to handle skewness.
- Encoded categorical variables using Label Encoding.
- Standardized numerical features using StandardScaler.

### **5. Model Building and Evaluation:**

- Implemented three classification algorithms: Logistic Regression, Random Forest, and Support Vector Machine (SVM).
- Performed hyperparameter tuning using GridSearchCV.
- Evaluated models using metrics such as Accuracy, Precision, Recall, F1 Score, and ROC AUC.
- Assessed the impact of preprocessing techniques like Scaling, Normalization, and Principal Component Analysis (PCA) on model performance.

### **6. Clustering Analysis (Bonus):**

- Applied KMeans clustering to identify potential customer segments.
- Evaluated clustering performance using the Adjusted Rand Index (ARI).
- Visualized clusters using PCA-reduced components.

### Section 3: Source Code

The complete source code for this project is available in the following GitHub repository:

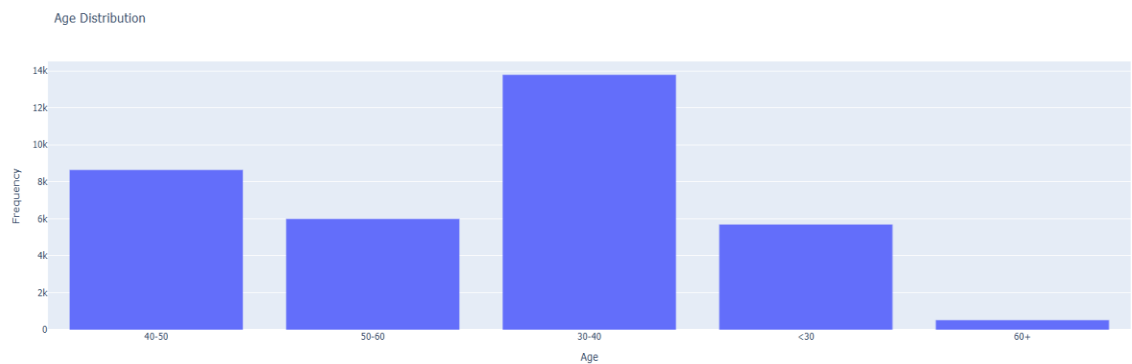
<https://github.com/Budur4/Data-science-project.git>

### Section 4: Visualization Snapshots

Below are key visualizations that provide insights into the data and model performance:

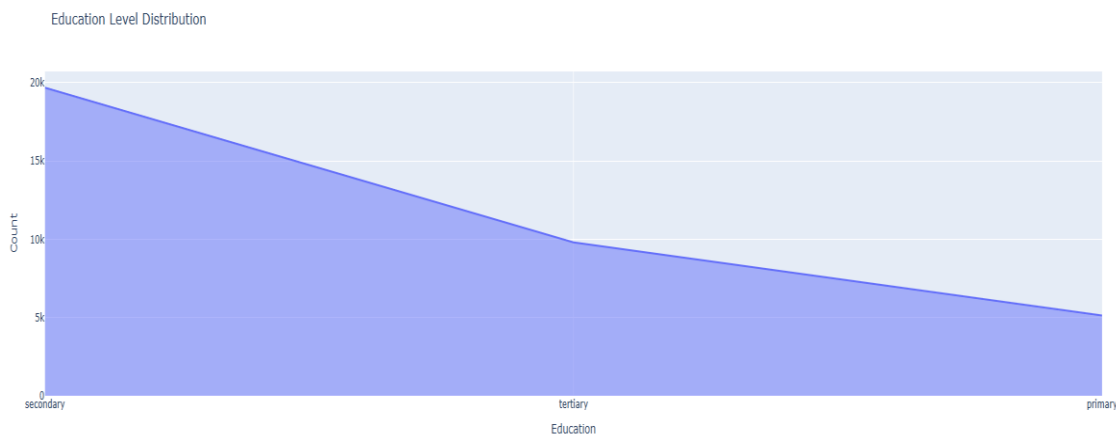
#### 1. Age Distribution:

- Histogram depicting the distribution of clients' ages



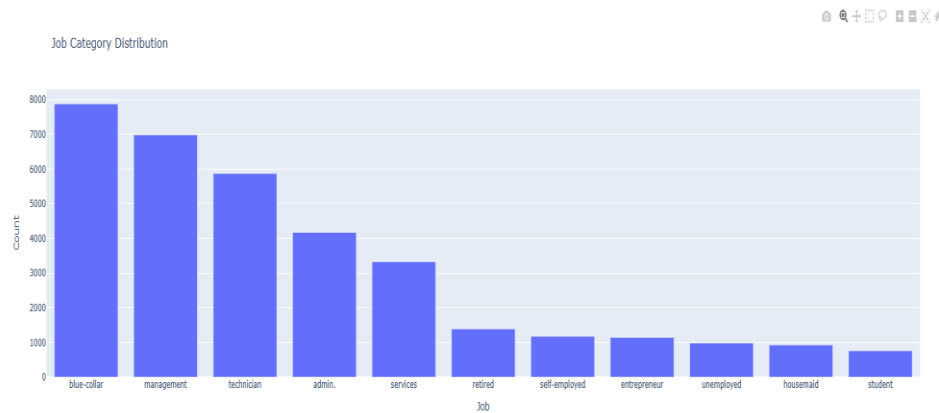
#### 2. Education Level Distribution:

- Area plot illustrating the distribution of clients' education levels.



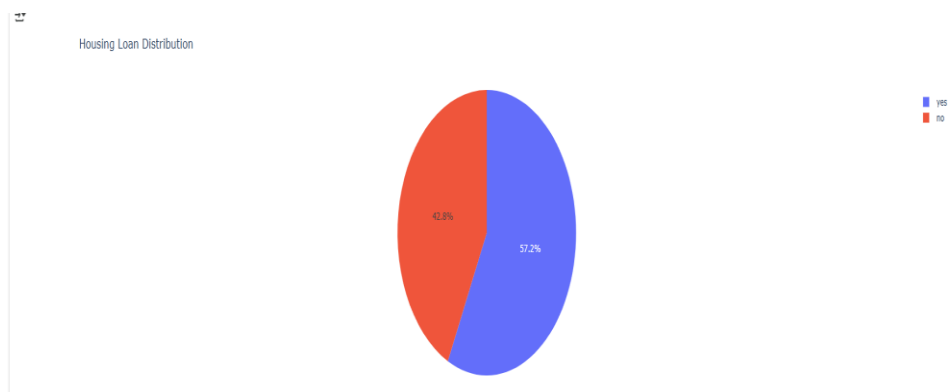
### 3. Job Category Distribution:

- Bar chart represents the count of clients across various job categories.



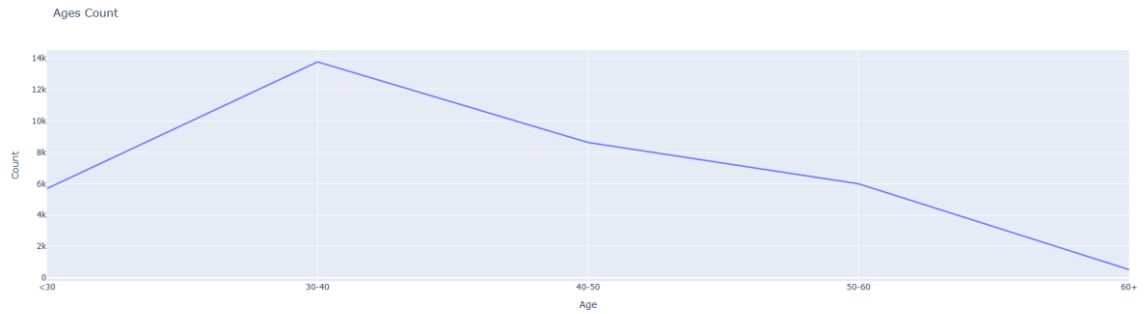
### 4. Housing Loan Distribution:

- Pie chart indicating the proportion of clients with housing loans.



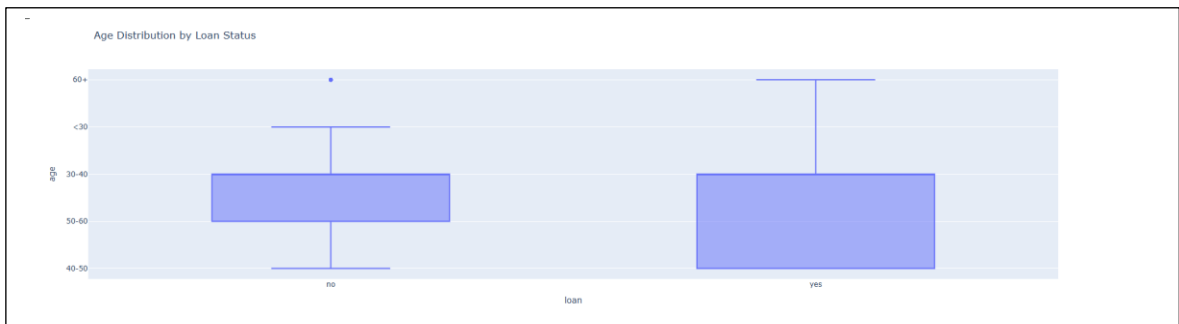
## 5. Line Plot — Contacts Over Months

- Visualizes how many contacts were made in each month.



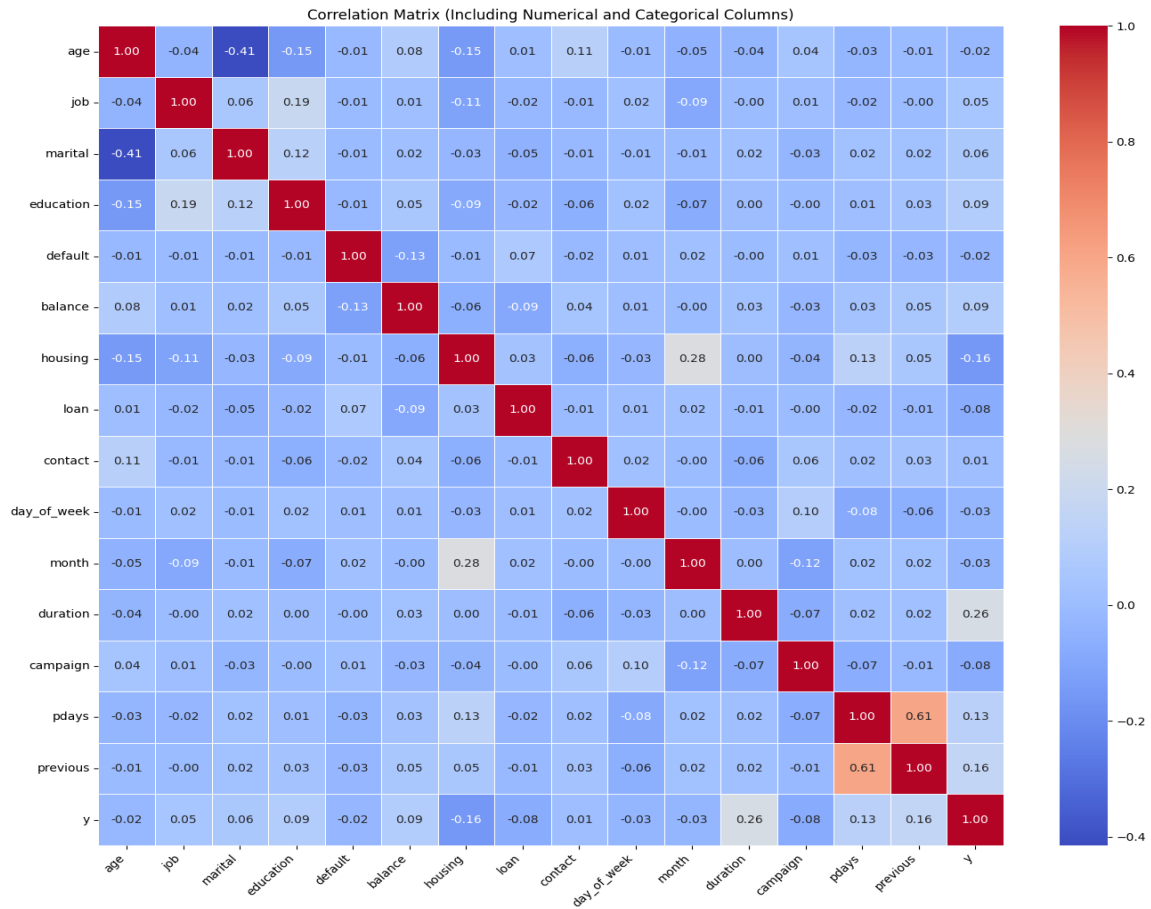
## 6. Age Distribution by Loan Status:

- o Box plot comparing age distributions based on loan status.



## 7. Correlation Matrix:

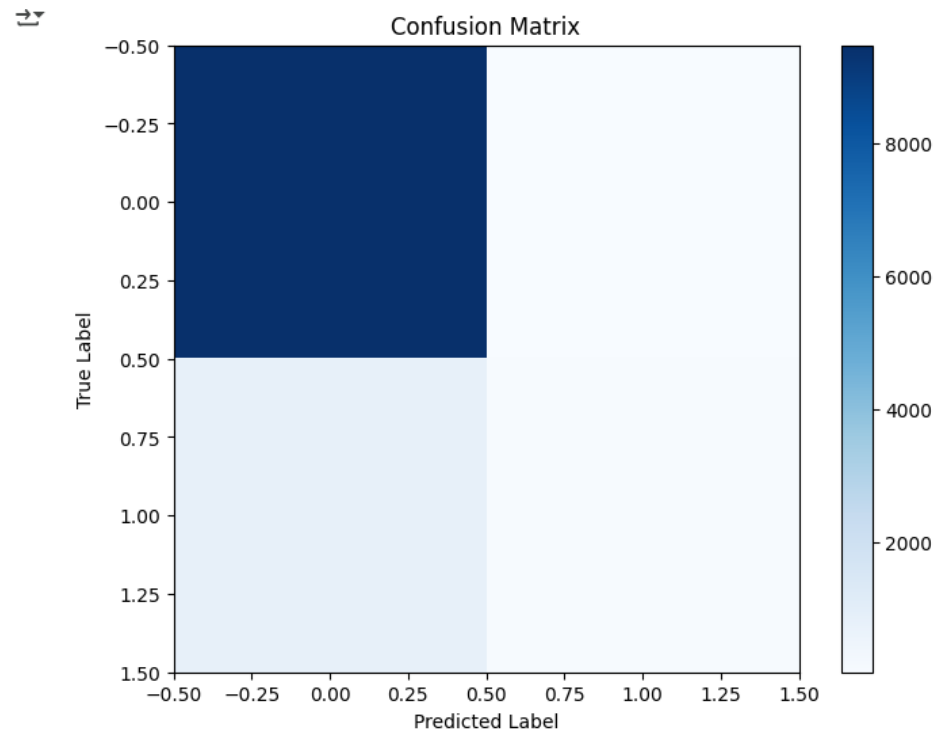
- Heatmap displaying correlations between numerical and encoded categorical variables.





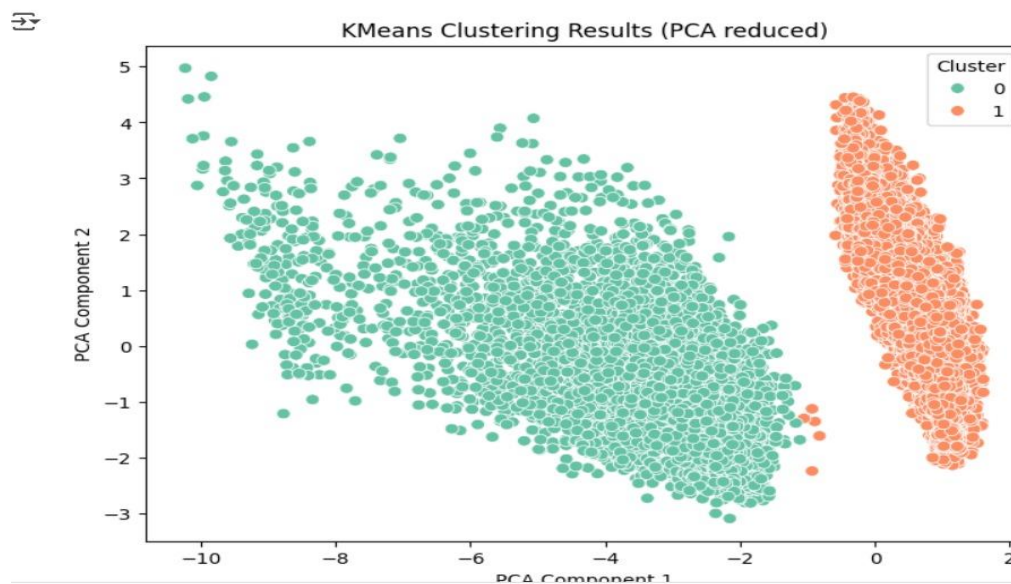
## 8. Confusion Matrix:

- Matrix illustrating the performance of the best-performing model (Random Forest) on the test set.



## 9. KMeans Clustering Results:

- Scatter plot visualizing clusters identified through KMeans after PCA reduction.



- **Model Performance:**

- The Random Forest classifier achieved the highest accuracy among the tested models.
- Scaling had minimal impact on model performance, while normalization and PCA slightly reduced accuracy, indicating model robustness.

- **Feature Importance:**

- Features such as duration, pdays, and previous were significant predictors of subscription likelihood.

- **Clustering Insights:**

- Unsupervised clustering revealed distinct customer segments, which can inform targeted marketing strategies.

Overall, the project demonstrates the efficacy of machine learning techniques in predicting customer behavior, providing valuable insights for marketing campaign optimization.