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
Nada Kamal-eldin Ahmed	3	94

Under supervision of

Eng. Farah Mohamed

Table of Contents

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

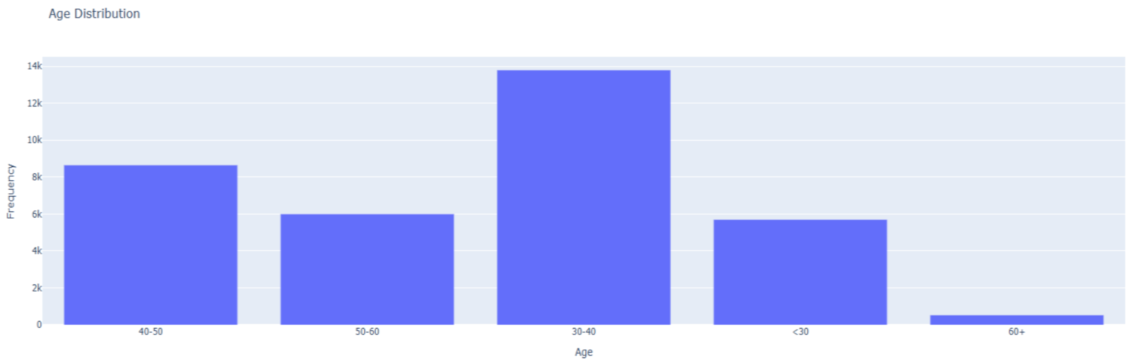
The complete source code for this project is available in the following GitHub repository:

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

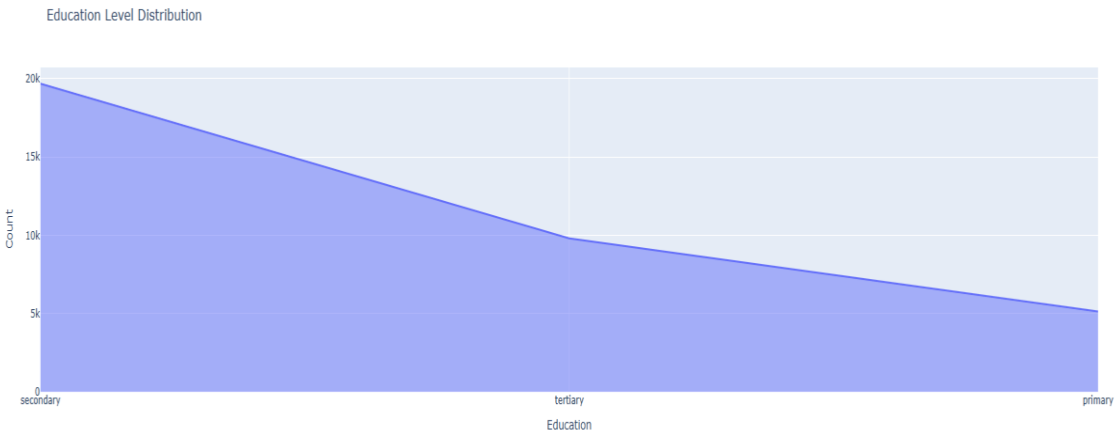
Section 4: Visualization Snapshots

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

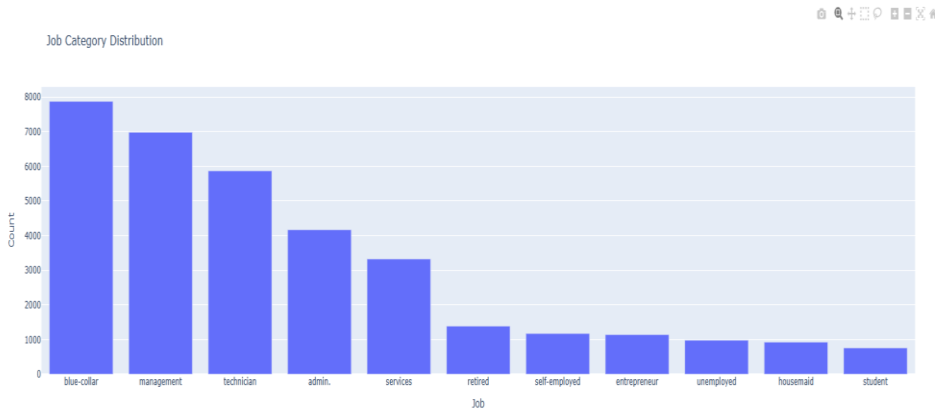
1. **Age Distribution**:
   - Histogram depicting the distribution of clients' ages



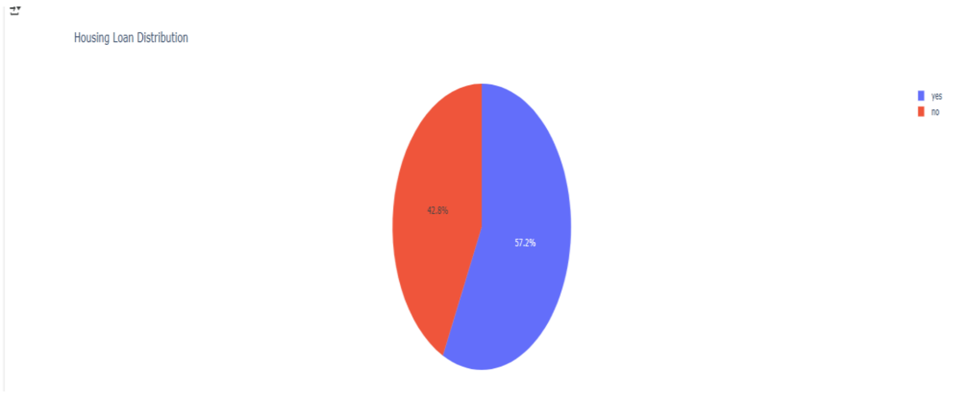
2. **Education Level Distribution**:
   - Area plot illustrating the distribution of clients' education levels.



3. **Job Category Distribution**:
   - Bar chart represents the count of clients across various job categories.

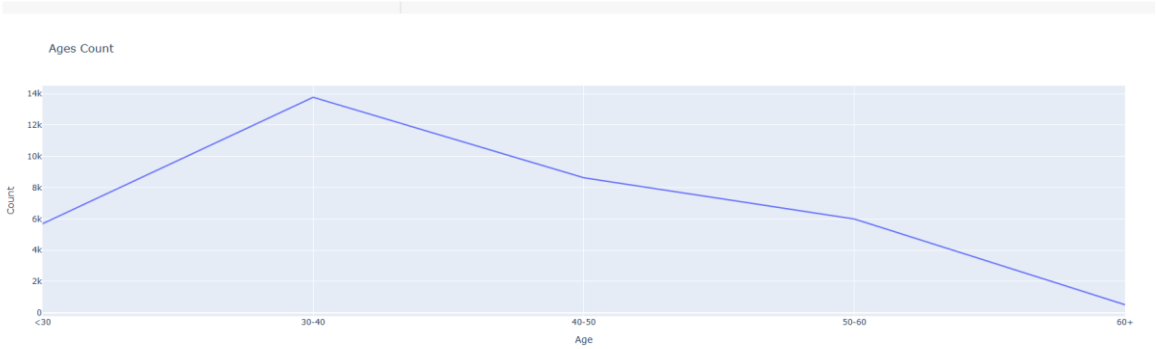


4. **Housing Loan Distribution**:
   - Pie chart indicating the proportion of clients with housing loans.



5.Line Plot — Contacts Over Months

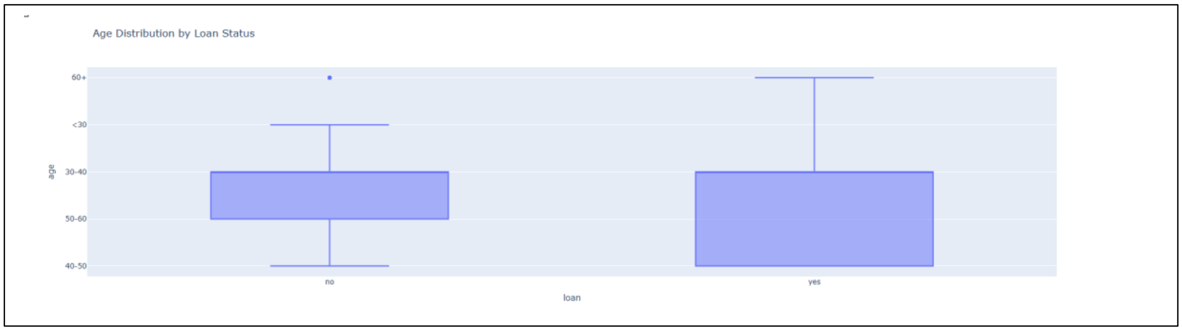
- Visualizes how many contacts were made in each month.



-

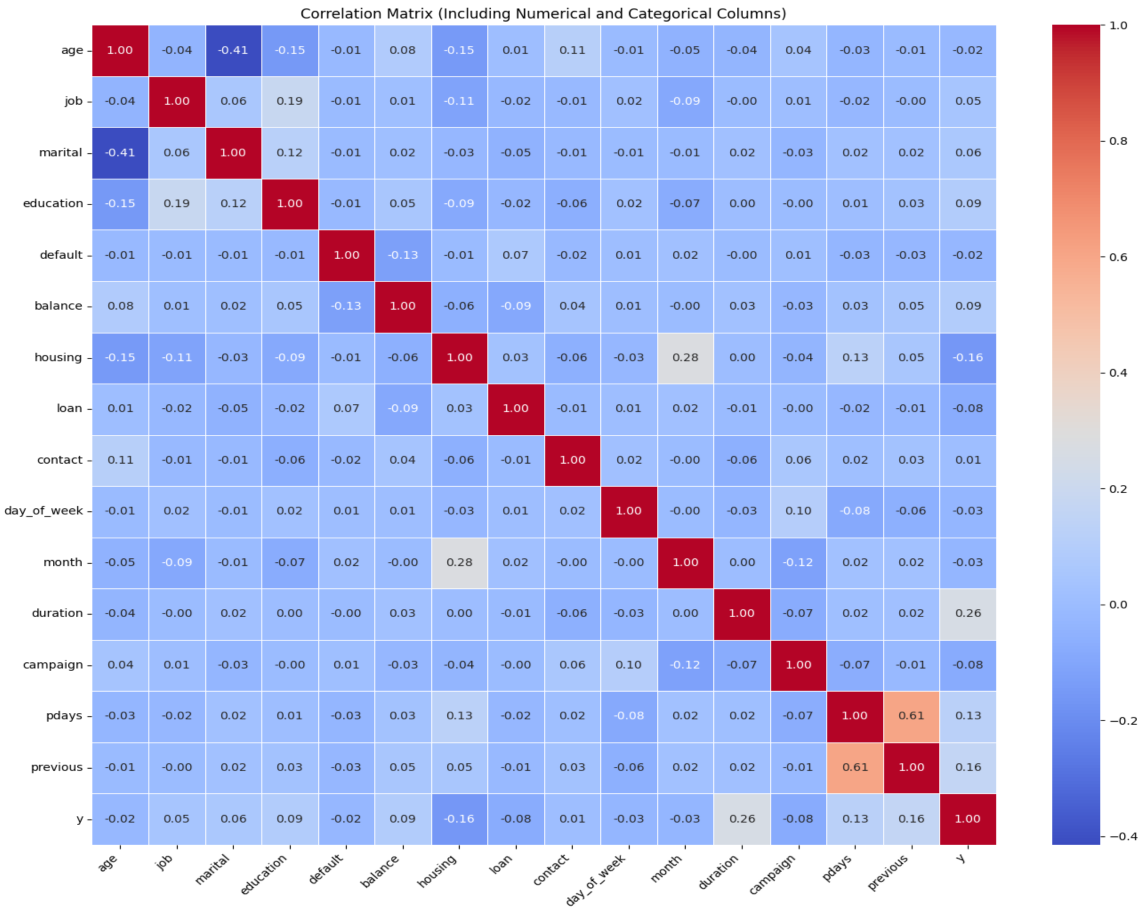
6.Age Distribution by Loan Status:

- Box plot comparing age distributions based on loan status.



7.**Correlation Matrix**:

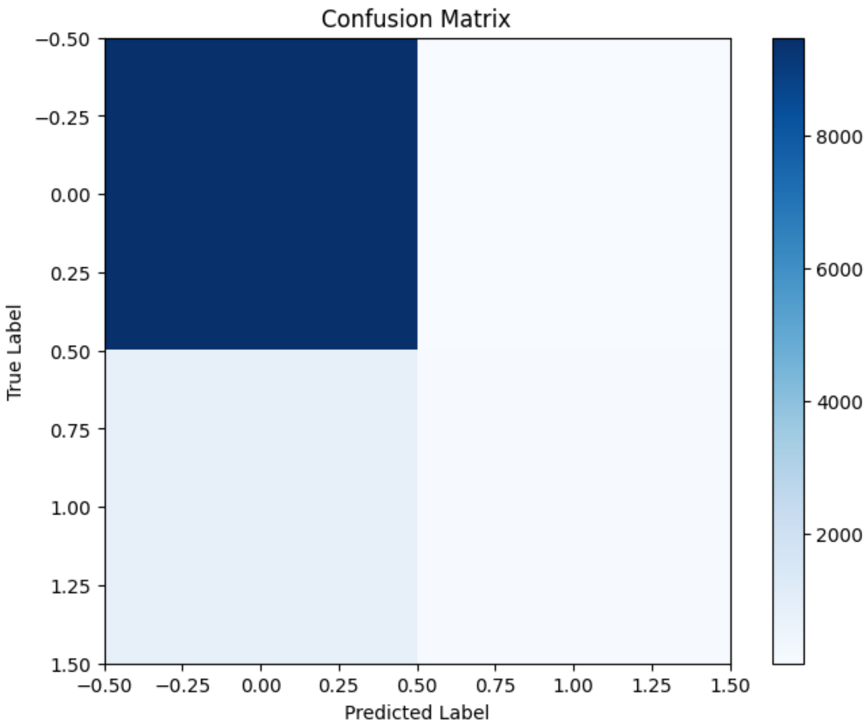
- Heatmap displaying correlations between numerical and encoded categorical variables.



	age	job	marital	education	default	balance	housing	loan	contact	day_of_week	month	duration	campaign	pdays	previous	y
age	1.00	-0.04	-0.41	-0.15	-0.01	0.08	-0.15	0.01	0.11	-0.01	-0.05	-0.04	0.04	-0.03	-0.01	-0.02
job	-0.04	1.00	0.06	0.19	-0.01	0.01	-0.11	-0.02	-0.01	0.02	-0.09	-0.00	0.01	-0.02	-0.00	0.05
marital	-0.41	0.06	1.00	0.12	-0.01	0.02	-0.03	-0.05	-0.01	-0.01	-0.01	0.02	-0.03	0.02	0.02	0.06
education	-0.15	0.19	0.12	1.00	-0.01	0.05	-0.09	-0.02	-0.06	0.02	-0.07	0.00	-0.00	0.01	0.03	0.09
default	-0.01	-0.01	-0.01	-0.01	1.00	-0.13	-0.01	0.07	-0.02	0.01	0.02	-0.00	0.01	-0.03	-0.03	-0.02
balance	0.08	0.01	0.02	0.05	-0.13	1.00	-0.06	-0.09	0.04	0.01	-0.00	0.03	-0.03	0.03	0.05	0.09
housing	-0.15	-0.11	-0.03	-0.09	-0.01	-0.06	1.00	0.03	-0.06	-0.03	0.28	0.00	-0.04	0.13	0.05	-0.16
loan	0.01	-0.02	-0.05	-0.02	0.07	-0.09	0.03	1.00	-0.01	0.01	0.02	-0.01	-0.00	-0.02	-0.01	-0.08
contact	0.11	-0.01	-0.01	-0.06	-0.02	0.04	-0.06	-0.01	1.00	0.02	-0.00	-0.06	0.06	0.02	0.03	0.01
day_of_week	-0.01	0.02	-0.01	0.02	0.01	0.01	-0.03	0.01	0.02	1.00	-0.00	-0.03	0.10	-0.08	-0.06	-0.03
month	-0.05	-0.09	-0.01	-0.07	0.02	-0.00	0.28	0.02	-0.00	-0.00	1.00	0.00	-0.12	0.02	0.02	-0.03
duration	-0.04	-0.00	0.02	0.00	-0.00	0.03	0.00	-0.01	-0.06	-0.03	0.00	1.00	-0.07	0.02	0.02	0.26
campaign	0.04	0.01	-0.03	-0.00	0.01	-0.03	-0.04	-0.00	0.06	0.10	-0.12	-0.07	1.00	-0.07	-0.01	-0.08
pdays	-0.03	-0.02	0.02	0.01	-0.03	0.03	0.13	-0.02	0.02	-0.08	0.02	0.02	-0.07	1.00	0.61	0.13
previous	-0.01	-0.00	0.02	0.03	-0.03	0.05	0.05	-0.01	0.03	-0.06	0.02	0.02	-0.01	0.61	1.00	0.16
y	-0.02	0.05	0.06	0.09	-0.02	0.09	-0.16	-0.08	0.01	-0.03	-0.03	0.26	-0.08	0.13	0.16	1.00

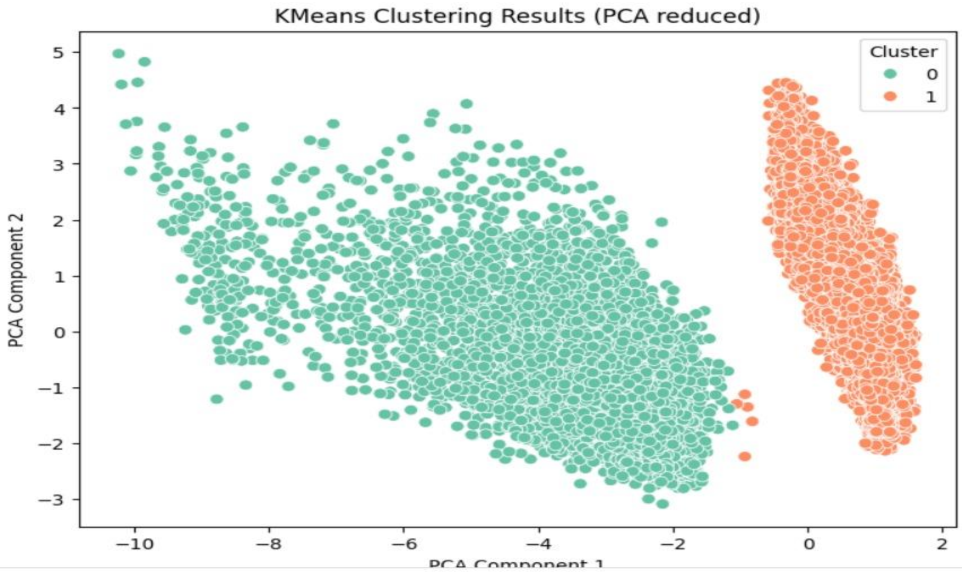
8.Confusion Matrix:

- 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.



The project successfully developed a predictive model for term deposit subscription using the Bank Marketing dataset.

- **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.