**Bank Marketing EDA Report**

**Author:** Arnav kalambe

**PRN:** 22070521097

**Course:** Data Science

**Title:** Bank Marketing

### **1. Introduction**

This report details the Exploratory Data Analysis (EDA) and initial machine learning model development for the Bank Marketing campaign dataset. The primary objective is to build a predictive model that can accurately identify which clients are likely to subscribe to a term deposit. This analysis provides a foundational understanding of the dataset's characteristics, highlights key insights, and documents the initial steps taken towards building a robust classification model.

**1.1 Dataset Details** ·

* Dataset Name: Bank marketing ·
* Source: [https://archive.ics.uci.edu/dataset/222/bank+marketing·](https://archive.ics.uci.edu/dataset/222/bank+marketing%C2%B7)
* GitHub Repository: <https://github.com/arnavkalambe27/-Bank_Marketing_DS_project>

### **2. Dataset Overview**

The dataset, bank-full.csv, contains a wide range of information about a bank's clients and the results of various marketing campaigns. Each row represents a client, and the columns contain attributes such as age, job, marital status, and previous campaign outcomes.

* **Dataset Source**: [Placeholder for Dataset Link]
* **Project GitHub Repository**: [Placeholder for GitHub Repo Link]
* **Number of Rows**: 45,211
* **Number of Columns**: 17
* **Target Variable**: y (binary: 'yes' or 'no', indicating subscription to a term deposit)
* **Feature Types**: The dataset includes both numerical and categorical features.

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

#### **3.1 Initial Data Inspection**

Initial inspection of the dataset revealed the presence of both numerical and categorical features. The categorical features were encoded into a numerical format using One-Hot Encoding to be compatible with machine learning algorithms.

#### **3.2 Descriptive Statistics for Numerical Features**

A summary of key statistics (mean, standard deviation, min, max, etc.) for each numerical feature was generated to understand the data's central tendency, spread, and potential outliers.

* **age**:
  + Minimum: 18 years
  + Maximum: 95 years
  + Mean: 40.9 years
* **balance**:
  + Minimum: -8019 euros
  + Maximum: 102127 euros
  + Mean: 1362.27 euros
* **duration**:
  + Minimum: 0 seconds
  + Maximum: 4918 seconds
  + Mean: 258.16 seconds
* ...and other numerical columns (day, campaign, pdays, previous).

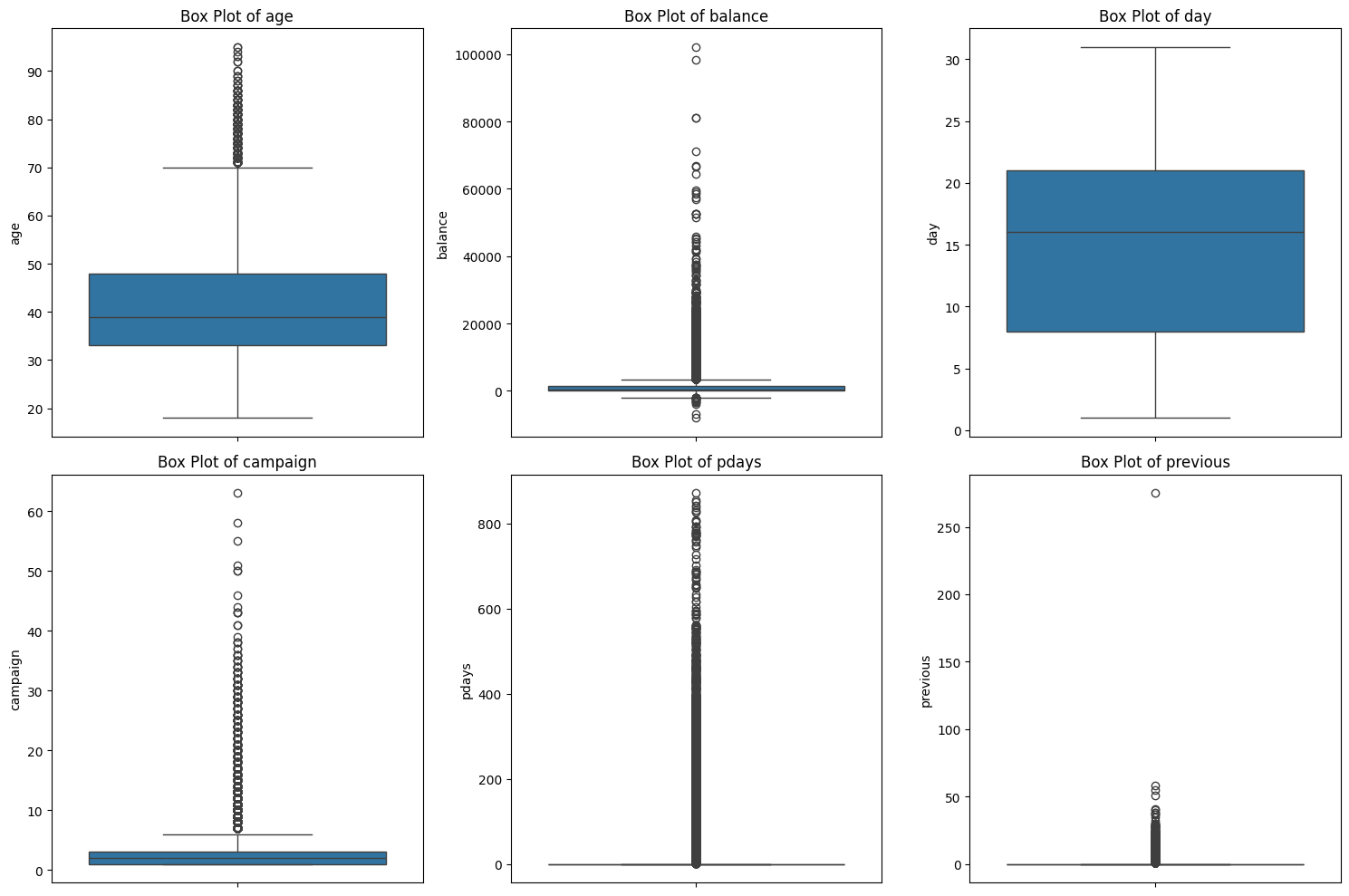
#### **3.3 Distribution of Numerical Features**

The distributions of the key numerical features provide insights into their characteristics. Most features show a positive skew, with a concentration of values at the lower end.

#### **3.4 Target Variable (y) Distribution**

The distribution of the target variable reveals a significant class imbalance, a crucial factor to address in the modeling phase. The majority of clients did not subscribe to a term deposit.

* **Value Counts**:
  + y = 0 (no subscription): 31,937 (Training), 7,985 (Testing)
  + y = 1 (yes subscription): 4,231 (Training), 1,058 (Testing)

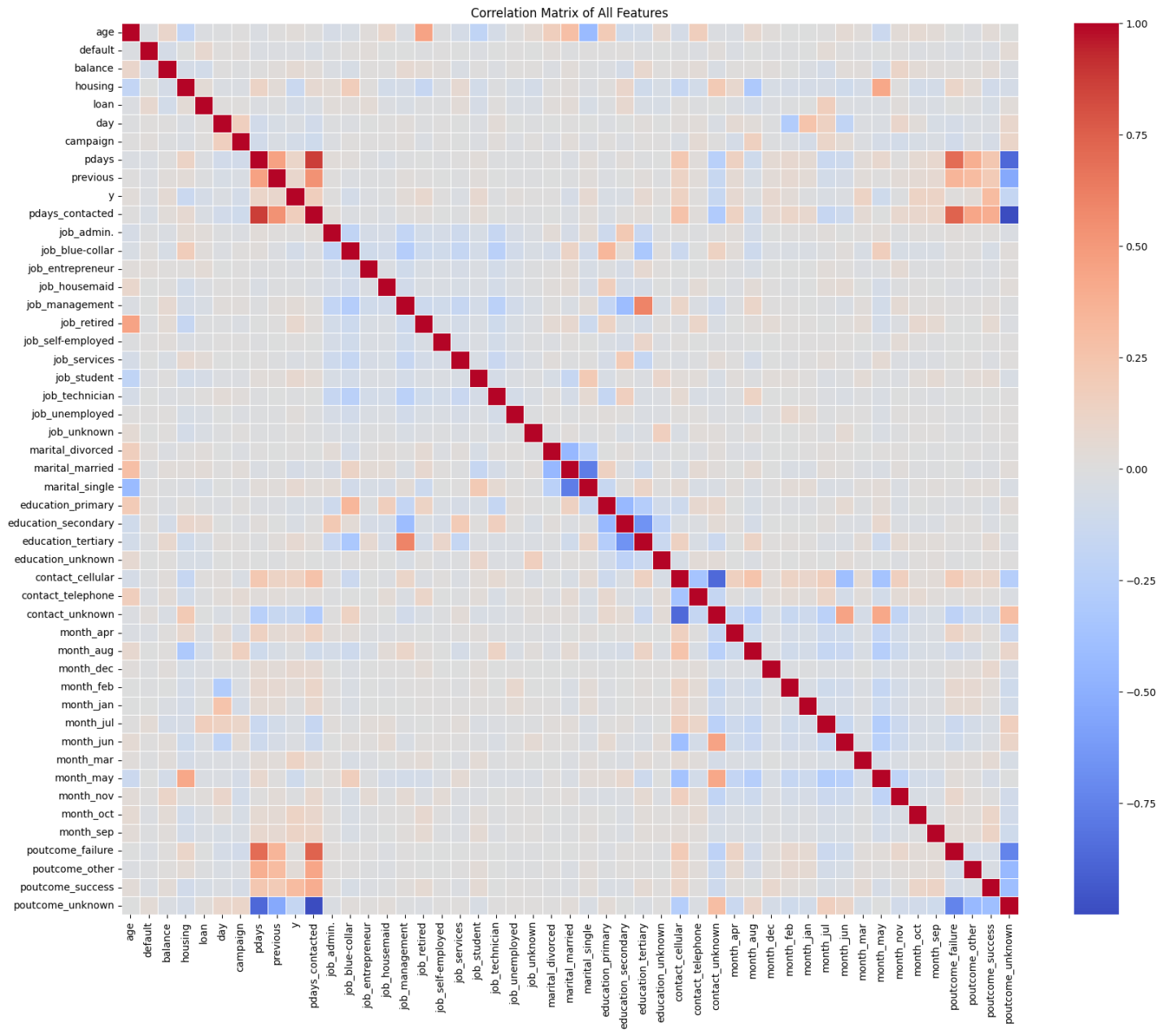
****

#### **3.5 Feature-Target Relationship: Age and Subscription**

Visualizing the age distribution segmented by the target variable y shows that while the overall population is concentrated in the 30-40 age group, the proportion of subscribers appears to be more evenly distributed.

#### **3.6 Categorical Feature Analysis**

An analysis of the categorical features (job, marital, education, etc.) showed the distribution of clients across different categories. Some categories have a disproportionately high or low subscription rate, which could be an important predictor.

****

#### **3.7 Feature Correlation Analysis**

A correlation matrix was generated to understand the linear relationships between all features. This helps in identifying highly correlated features that might lead to multicollinearity and also reveals potential relationships with the target variable.

### **4. Feature Engineering and Scaling**

To prepare the data for modeling, the following steps were taken:

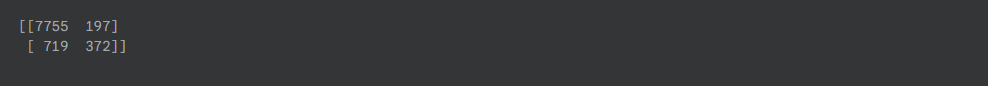
* **One-Hot Encoding**: All categorical variables were converted into numerical representations.
* **Feature Scaling**: Two scaling methods were applied to the numerical features to standardize their ranges.
  + **StandardScaler**: This approach centers the data to a mean of 0 with a standard deviation of 1. It is effective for models that assume a normal distribution.
  + **MinMaxScaler**: This method scales the data to a fixed range, typically [0, 1]. It is useful for algorithms that are sensitive to the magnitude of feature values, like Support Vector Machines.

### **5. Initial Model Training and Evaluation**

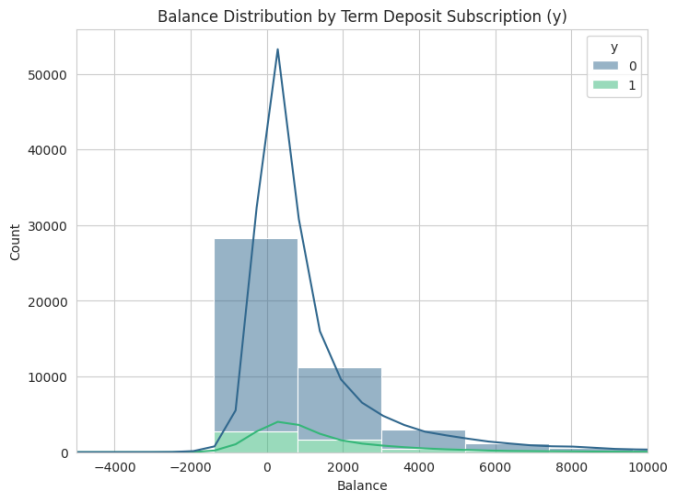
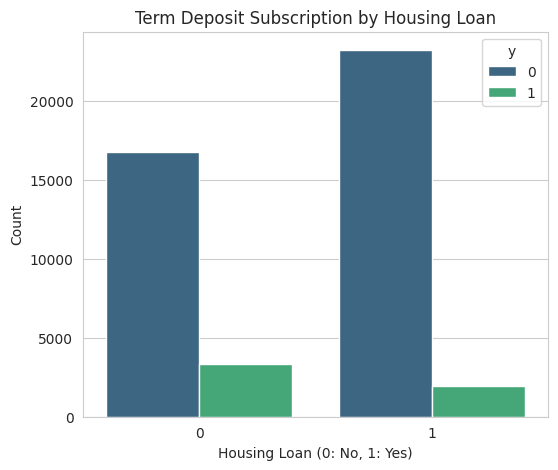
A Logistic Regression model was trained as a baseline using the data scaled with StandardScaler.

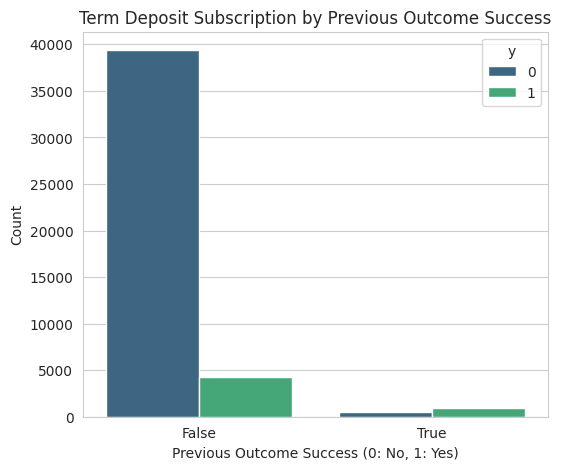
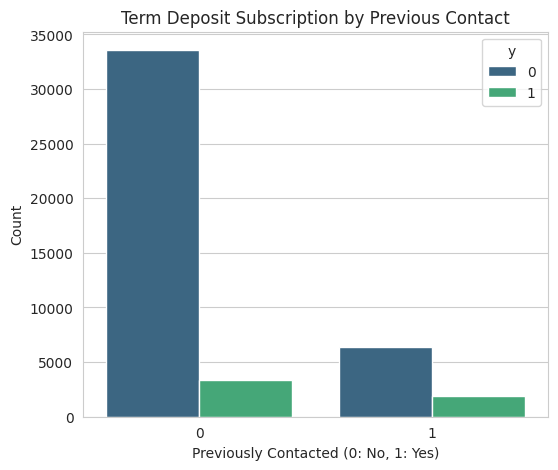
* **Model**: Logistic Regression
* **Scaling Method**: StandardScaler
* **Performance Metrics**:
  + **Accuracy**: 0.8987
  + **Precision (for Class '1')**: 0.6538
  + **Recall (for Class '1')**: 0.3410
  + **F1-Score (for Class '1')**: 0.4482

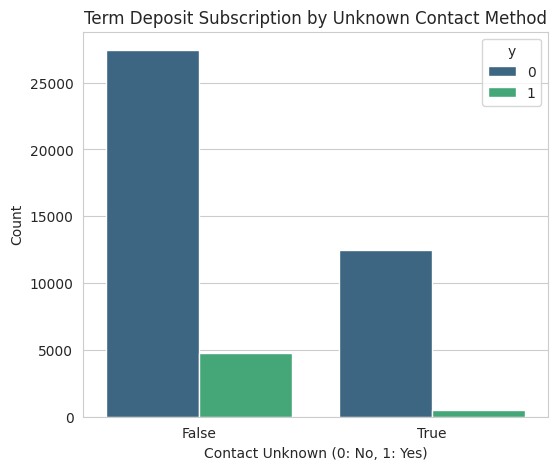
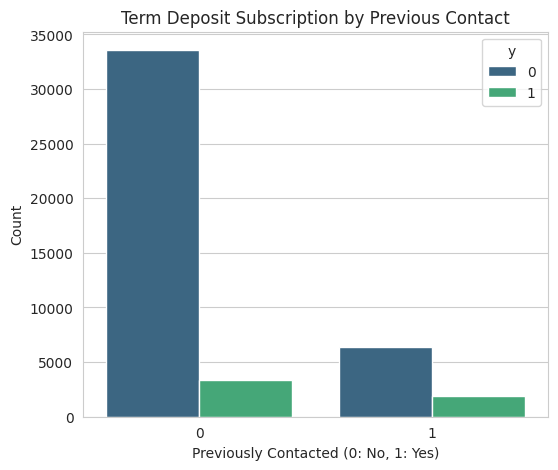
**Confusion Matrix**:



* **Interpretation**: The model correctly predicted 7755 non-subscribers and 372 subscribers. However, it failed to identify 719 actual subscribers (False Negatives), resulting in low recall for the minority class. This highlights the impact of the class imbalance.

****

****

****

### **6. Future Work**

To improve model performance, especially in identifying the minority class, the following steps are planned:

1. **Model Comparison**: Evaluate the Logistic Regression model using MinMaxScaler data to compare its performance.
2. **Advanced Models**: Implement and evaluate other classification algorithms such as Random Forest, Gradient Boosting Machines (XGBoost), and Support Vector Machines.
3. **Hyperparameter Tuning**: Optimize the parameters of the best-performing models using techniques like Grid Search or Random Search.
4. **Class Imbalance Handling**: Apply strategies such as SMOTE (Synthetic Minority Oversampling Technique) or adjusting class weights to improve the model's ability to predict the minority class.
5. **Feature Engineering**: Explore creating new features from existing data, for example, interaction terms or polynomial features.
6. **Cross-Validation**: Use k-fold cross-validation to ensure the model's performance is stable and generalizes well to unseen data.
7. 1.1 Dataset Details Dataset Name: Food Security Dataset Source: h ps://ndap.ni .gov.in/dataset/7115 GitHub Repository: h ps://github.com/Blaze-0903/Food-Security\_DSProject