# Report: Analysis of A Bank Marketing Campaign

Table of Contents

[Report: Analysis of A Bank Marketing Campaign 1](#_Toc201566910)

[**Dataset Overview** 1](#_Toc201566911)

[**Data Cleaning & Preprocessing** 2](#_Toc201566912)

[**Identifying patterns and relationship** 2](#_Toc201566913)

[**Modeling** 3](#_Toc201566914)

[**Check for Accuracy** 3](#_Toc201566917)

[**Model Saving** 4](#_Toc201566918)

[**Model Deployment & Testing** 4](#_Toc201566919)

[**Insights** 5](#_Toc201566920)

[**Recommendations** 6](#_Toc201566930)

[How to Handle Class Imbalance 6](#_Toc201566931)

[Resampling 6](#_Toc201566932)

### **Dataset Overview**

The dataset contains information on direct marketing campaigns (phone calls) by a bank. The goal is to predict whether a client will subscribe to a term deposit product.

#### Key Features:

* **Client Data**: age, job, marital, education, default, housing, loan
* **Campaign Data**: contact, month, day\_of\_week, duration, campaign, pdays, previous, poutcome
* **Macroeconomic Indicators**: emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed
* **Target Variable**: y (whether the client subscribed to a term deposit)

### **Data Cleaning & Preprocessing**

#### Handling csv data file

The csv file was loaded into a dataframe using the delimiter(;)

#### *Data Observation*

The head function was used on the data frame to view a few data rows and have a fair idea of values of the dataset

#### *Data Columns and Values*

The columns of the dataset were also evaluated for unique values, the distribution and their data types( categorical/numerical) to have a deeper understanding of the dataset ,

#### Handling outliers and missing Data

Screened dataframe to find outliers and missing data.

Using…df[df.isnull().any(axis=1)] it was realized that the dataset has no missing values however unique values generated from columns indicated that some clients did not disclose certain info and hence it was flagged as unknown. Nonetheless, the spread of the data exposed the total number of such to be small and insignificant.

### **Identifying patterns and relationship**

The relationships between features and the target feature was explored through visualizations and filtering.

Distribution of target feature was also observed using a barplot. It showed 12% yes and 88% no’s

#### Correlation

* A Pearson correlation was performed on numerical Features such as age, duration, campaign, pdays, emp.var.rate, euribor3m, etc.
* A **Point-Biserial** correlation was performed on categorical values such as job, marital, education, contact, month, poutcome, etc., were converted using pd.get\_dummies().
* One-hot encoding was applied with alignment to ensure model consistency.

### **Modeling**

#### Handling target variable

The target value was hot encoded in 0s and 1s.

The dataset was split into training set(x-train) and test(x-test) set randomly in an 80-20% fashion.

#### Model Training

Model Used**:** Logistic Regression and Random Forest

Reason**:** Both models were used and their accuracy measured to see the most efficient model for this modelling

#### Logistic Regression

* StandardScaler was used to scale all numerical features.
* Trained using:
  + X\_train\_scaled
  + y\_train
* Tested on:
  + X\_test\_scaled
  + y\_test
* Achieved good performance, though affected by class imbalance.

#### Feature importance

Importance of features were ranked using the Random Forest model

### **Check for Accuracy**

Accuracy score and classification report was run on the model. The model proved to be efficient with accuracy scores of :

Logistic Regression:

Accuracy: 0.9124987860541905

precision recall f1-score support

0 0.93 0.97 0.95 9144

1 0.67 0.43 0.52 1153

accuracy 0.91 10297

macro avg 0.80 0.70 0.74 10297

weighted avg 0.90 0.91 0.90 10297

Random Forest:

Accuracy: 0.914829562008352

precision recall f1-score support

0 0.94 0.97 0.95 9144

1 0.66 0.49 0.56 1153

accuracy 0.91 10297

macro avg 0.80 0.73 0.76 10297

weighted avg 0.91 0.91 0.91 10297

### **Model Saving**

* Trained model and scaler were saved using pickle:
  + log\_model.pkl
  + scaler.pkl

#### 2nd Model Testing

The banking-full.csv dataset that contained 10% of the original dataset which was randomly generated was used as a testing set and the following were the accuracy scores.

Logistic Regression:

Accuracy: 0.9150279193979121

precision recall f1-score support

0 0.93 0.97 0.95 3668

1 0.68 0.43 0.52 451

accuracy 0.92 4119

macro avg 0.80 0.70 0.74 4119

weighted avg 0.90 0.92 0.91 4119

### **Model Deployment & Testing**

* A pipeline was created to:
  + Accept new client data
  + Encode categorical features
  + Scale numeric features
  + Align columns with training data
  + Predict probability of subscription
* Model tested using new unseen data points in memory.

### **Insights**

**1.** Customer Profile Insights

Certain job types are more likely to subscribe

* Clients in roles like **student**, **retired**, or **admin.** had higher conversion rates.
* These groups may be more responsive to long-term savings.

#### Education level matters

* Clients with **tertiary education** subscribed at higher rates.
* Marketing efforts can be tailored based on educational background.

#### Marital status plays a role

* **Single** clients may be more open to fixed deposit products than married or divorced clients.

#### 2. Campaign Performance Insights

Number of contacts affects success

* The **first contact** had the highest success rate.
* Repeated contacts (high campaign value) saw **diminishing returns** — potentially annoying customers.

Call duration is highly predictive

* Longer conversations correlated positively with conversion.
* But it’s not useful for real-time prediction (you don't know duration in advance).

**3.** Macroeconomic & Timing Insights

Timing matters (Month of contact)

* Subscriptions peaked during **March**, **October**, and **December**.
* Suggests seasonal behaviors or marketing effectiveness during certain months.

Economic indicators influence response

* Higher euribor3m and emp.var.rate values correlated **negatively** with subscriptions.
* Clients are **less likely to subscribe** during economically uncertain or tight periods.

#### 4. Model & Prediction Insights

Imbalanced classes affect accuracy

* Majority of clients (88%) did **not** subscribe.
* Accuracy alone is misleading — hence **recall, precision, and F1** are more meaningful.

Most important features:

* duration, pdays, euribor3m, job, education

### 

### **Recommendations**

| **Strategy Area** | **Recommendation** |
| --- | --- |
| **Targeting** | Focus on students, retirees, and well-educated clients |
| **Campaign Efficiency** | Reduce repeated contacts; aim for impactful first call |
| **Economic Context** | Intensify campaigns during favorable economic periods |
| **Call Center Tactics** | Train agents to engage clients early in the call |
| **Data Collection** | Reduce "unknown" responses — they reduce prediction power |

#### Sampling

**How to Handle Class Imbalance**

**Resampling**

* **Oversample minority** class (e.g., SMOTE)
* **Undersample majority** class