Bank Marketing Effectiveness Prediction

By

Manmohan Madhu

**Abstract**

This project aims to predict the outcome of a telemarketing call campaign by training a machine learning model.

**Problem Statement**

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

**Introduction**

Each row of this dataset contained information regarding a customer such as his/her age, loans taken, education, marital status etc. After preprocessing and EDA, a model was trained so that in future, given a few customer information a prediction can be made if a customer subscribes to a term deposit or not.

**Challenges Faced**

* The dataset did not contain null values, however, column pdays contained a majority of entries with -1.
* The dataset had outliers for columns campaign and previous.
* Yes and no in the dataset had to be replaced with boolean values.
* One hot encoding had to be performed for categorical variables.

**The Approach Used to Solve the Problem**

After handling the null values and taking care of outliers, the data was split into a training set and a test set. This was followed by training two models, namely, XGBoost and Random Forest Classifier. The models were used to predict the values of the dependent variable on the test set and the results were evaluated using metrics such as accuracy and roc\_auc score.

**Libraries used for analysis**

1. Pandas : To load the data into a dataframe object and analyze.
2. Matplotlib : To help visualize the data.
3. Seaborn : For added functionality to matplotlib.
4. Numpy : To use the numpy functions in analysis.

From sklearn

1. Test train split
2. Grid search CV
3. XGBoost
4. Random Forest Classifier
5. Accuracy score
6. ROC\_AUC score

**Dataset**

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Data Description:

Input variables:

Bank Client data:

* age
* job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
* marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
* education (categorical: ‘primary’, ‘secondary’, ‘tertiary’)
* default: has credit in default? (categorical: 'no','yes','unknown')
* housing: has a housing loan? (categorical: 'no','yes','unknown')
* loan: has a personal loan? (categorical: 'no','yes','unknown')

Related with the last contact of the current campaign:

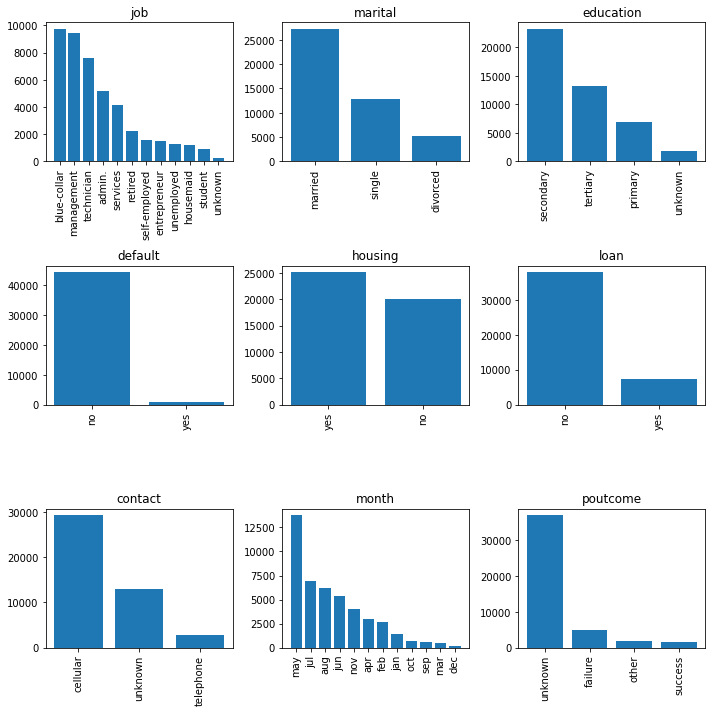
* contact: contact communication type (categorical: 'cellular','telephone')
* month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
* day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
* duration: last contact duration, in seconds (numeric).
* Other attributes:
* campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
* pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
* previous: number of contacts performed before this campaign and for this client (numeric)
* poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
* Output variable (desired target):
* y - has the client subscribed to a term deposit? (binary: 'yes','no'.

**Dataset preparation before analysis**

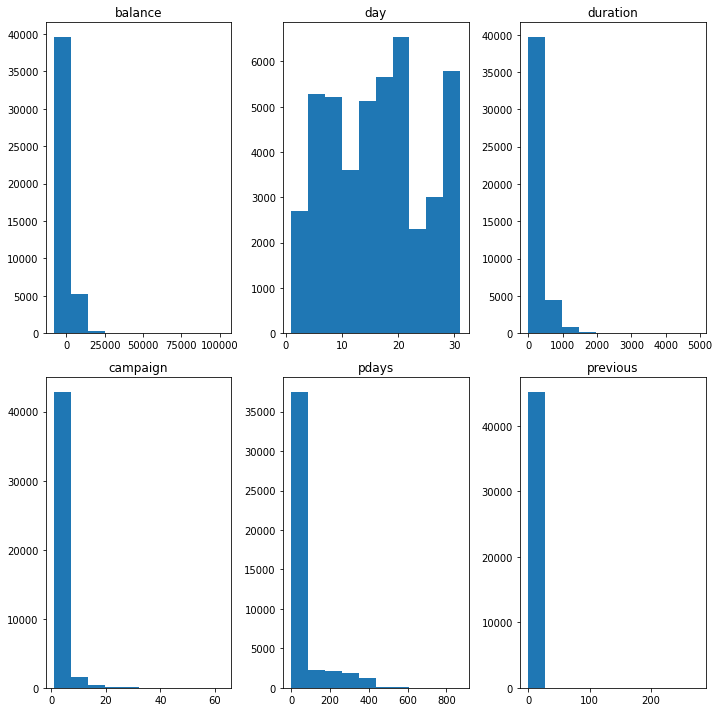
The dataset did not contain null values.

* Column pdays contained a majority of entries with -1. This can’t be the case because the number of days cannot be negative. Therefore, this column was dropped.
* The dataset had outliers for columns campaign and previous. A threshold was set at 34 after interpreting the distribution plots. Entries above this threshold was replaced with the mean.
* Yes and no in the dataset had to be replaced with boolean values.
* One hot encoding had to be performed for categorical variables.

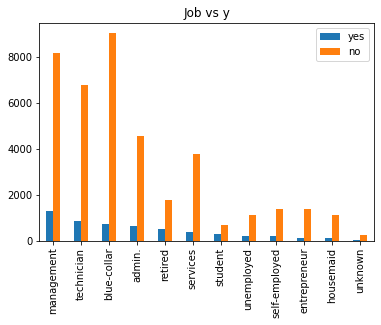
**Plots**

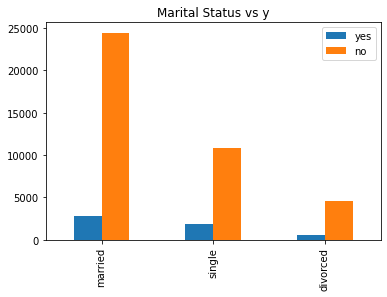
Univariate analysis of categorical variables

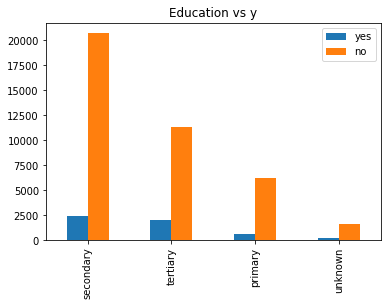
Univariate analysis of numerical variables

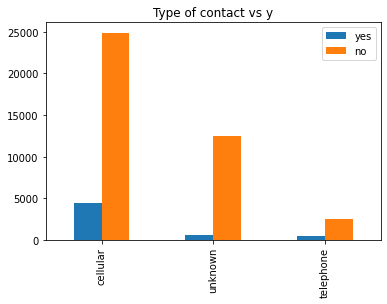


Bivariate analysis









**Analysis**

* Customers working blue collar jobs are less likely to subscribe whereas management and technicians are more likely.
* Single customers are more likely to subscribe for a term deposit whereas married customers are less likely.
* Customers with tertiary education are more likely to subscribe whereas customers with secondary education are less likely to subscribe.

**Feature Engineering**

* Replacing yes or no with boolean values for columns 'default', 'housing', 'loan' and 'y'.
* Replacing with mean if the value is more than a specified threshold of 34 for columns ‘campaign’ and ‘previous’.
* One hot encoding of categorical columns such as 'job', 'marital', 'education', 'contact', 'month' and 'poutcome'.

**Training of the model**

* The entire dataset is split into two sets, train and test, by a 70:30 ratio.
* Two modeling libraries are used to train the dataset, XGBoost and Random Forest Classifier.
* Hyperparameter tuning: Grid search CV is used to optimize the roc\_auc score for the Random Forest Classifier.

**Evaluation of the models**

* Since for this business problem a false positive and false negative does not have a major impact, we can focus on accuracy as the evaluation metric to consider.
* In addition to this we also consider the ROC\_AUC score to test our models’ classification capabilities.
* XGBoost : Accuracy 90.6% ROC\_AUC 71.66%.
* Random Forest Classifier: Accuracy 89.53% ROC\_AUC 90.88%.

**Conclusion**

With an accuracy of 89.5 percent and an roc\_auc score of 90.88 percent we can confidently use this model to predict whether the customer will subscribe or not even before the call is made. This will save time and resources as we have to only call the customers who are more likely to subscribe thereby reducing the workload of the marketing team.

—-------------End—--------------------