# Predicting Bank's Term Deposit Subscription

Brown University: DATA 1030 Project Final Report

Author: Guansu (Frances) Niu

GitHub Link: https://github.com/francesniu/Data-1030-Bank-Marketing-Project.git

Date: 12/03/2019

## 1. Introduction:

### 1.1 Project Goal

The goal is to predict if clients will subscribe to a term deposit from a bank. The problem is classification, and the target variable is 'has the client subscribed to a term deposit' (binary). According to the dataset, 41,188 clients have already made the decision. By analyzing the correlation between those clients' data and their decision outcomes, the bank will be able to forecast the future client adaptability. Based on the results, the bank can improve the term deposit so that it will be more widely accepted by its clients.

### 1.2 Dataset description

The data of direct marketing campaigns is provided by a Portuguese banking institution. The dataset consists of 20 attributes and 1 target variable belonging to 41,188 clients.

### 1.3 Feature description:

Input variables:

**Bank client data attributes:**
1 - age (numeric)
2 - job: type of job (categorical: 'admin.','bluecollar', 'entrepreneur', 'housemaid', 'management', 'retired','self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown').
3 - marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed).
4 - education (categorical:'basic.4y','basic.6y','basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown').
5 - default: has credit in default? (categorical: 'no', 'yes',' unknown').
6 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown').
7 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown').

**Current campaign attributes:**
8 - contact: contact communication type (categorical: 'cellular', 'telephone').
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec').
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri').
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end

of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact).

**Previous campaign attributes:**
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means clients were not previously contacted).
14 - previous: number of contacts performed before this campaign and for this client (numeric).
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success').

**Social and economic context attributes:**
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric).
17 - cons.price.idx: consumer price index - monthly indicator (numeric).
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric).
19 - euribor3m: euribor 3-month rate - daily indicator (numeric).
20 - nr.employed: number of employees at the bank - quarterly indicator (numeric).

Target variable:

21 - y - has the client subscribed to a term deposit? (binary: 'yes', 'no').

## 1.4 Public projects using the same dataset

### *"A data-driven approach to predict the success of bank telemarketing"*

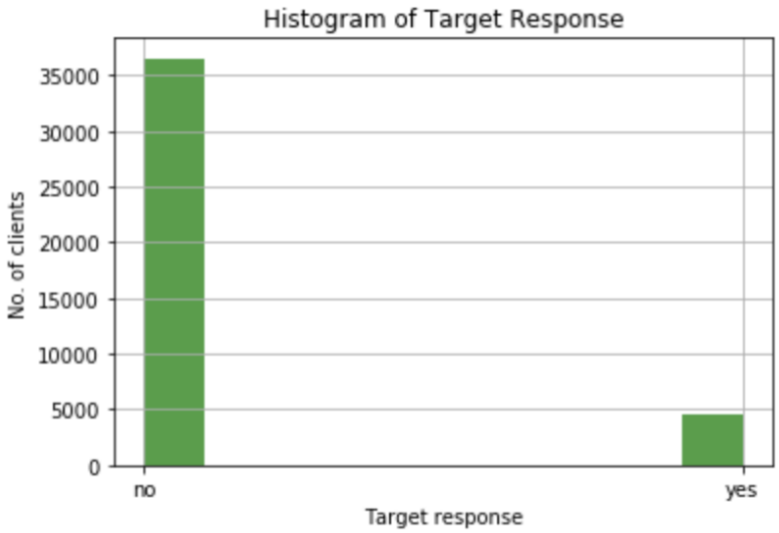
This project proposes a data mining methodology to predict the success of telemarketing calls for selling bank long-term deposits. By integrating the dataset, four models including logistic regression, decision trees, neural network and support vector machine were tested. By comparing the two metrices, area of the receiver operating characteristic curve and area of the LIFT cumulative curve, that neural network presents the best result, allowing to reach 79% of the subscribers by selecting the half better classified clients.

### *"Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology"*

The project describes a data mining approach to bank direct marketing campaigns using the dataset. It performed three iterations of the CRISP-DM and the best model, materialized by a Support Vector Machine, achieved high predictive performance. It also confirmed that open-source technology in the DM field is able to provide high quality models for real applications meanwhile allow a cost reduction of DM projects.

## 2. Exploratory Data Analysis:

2.1 Balance of Target variable:

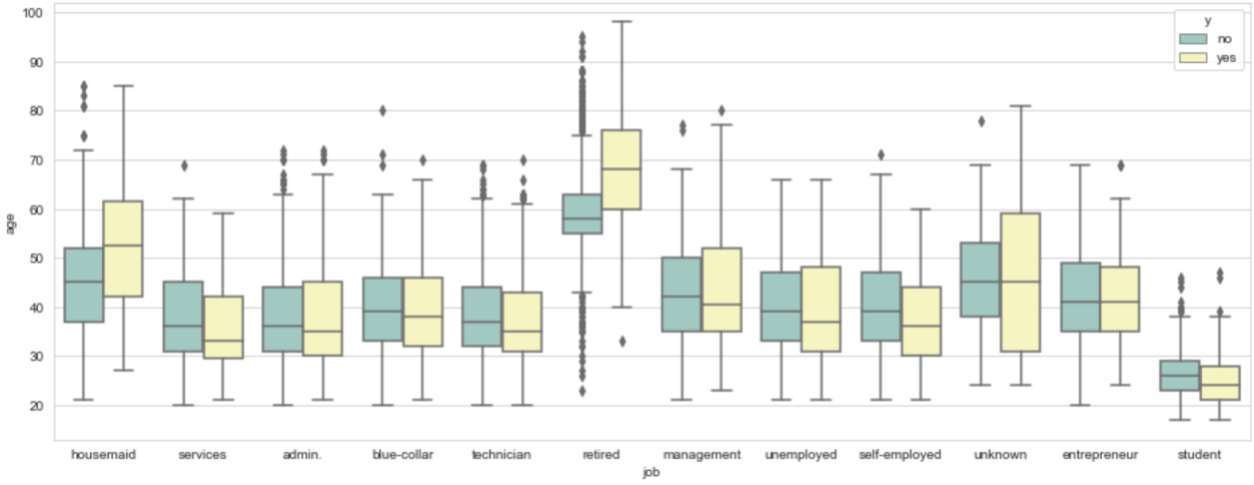



The target variable of the dataset refers to 'has the client subscribed to a term deposit?' which gives a binary result. About 88.7% of the clients have not subscribed to the deposit, and 11.3% have subscribed. Therefore, the dataset is imbalanced.

2.2 Examples of Feature Visualization:

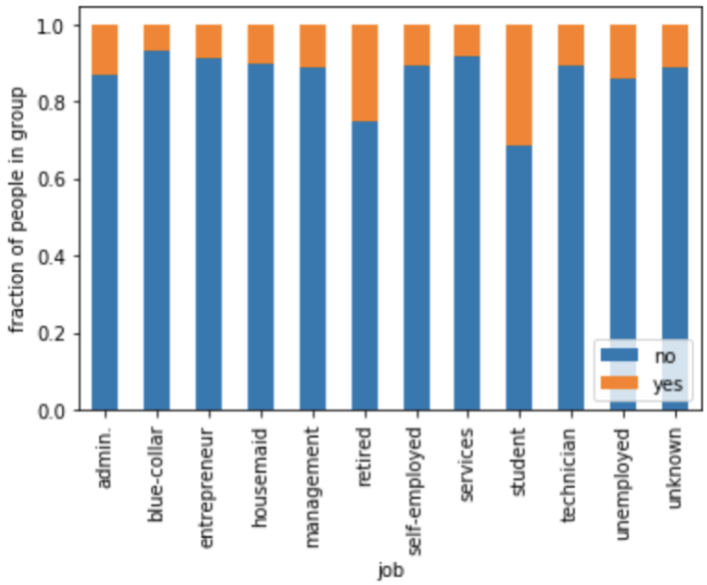
Example 1: job

Distribution of Age by Jobs under Subscription Condition




- The retired group has the highest median age, and the student group has the smallest median age.

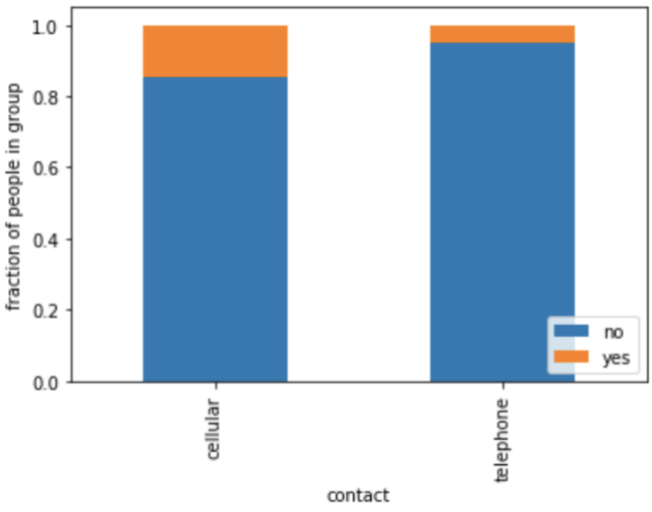
Distribution of Subscription Condition among Jobs




- Among all categories, the student group has the highest subscription percentage, and it is also the youngest group. The retired group has the second highest subscription percentage, and it is the oldest group. This might be because most of them do not have jobs that allow them to earn income directly, and they are looking for other opportunities to invest.
- The blue-collar group has the lowest subscription percentage probably because most of them do not work frequently with the bank's financial products, so the term deposit might not seem trustworthy to them. The entrepreneur group has the second lowest subscription percentage probably because most of them have their own financial plans, so it could be hard to persuade them to subscribe to the product.

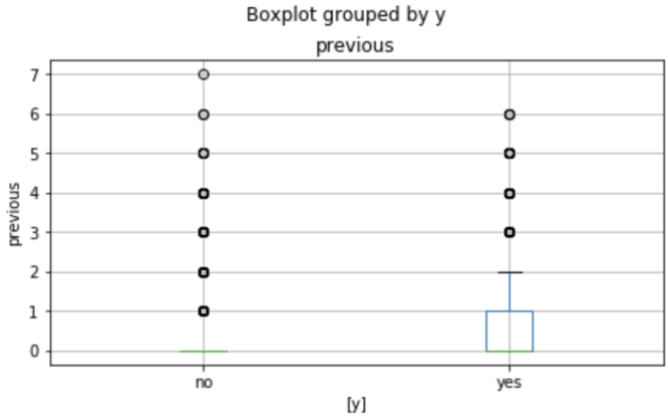
Example 2: contact

Distribution of Subscription Condition between Communication Methods



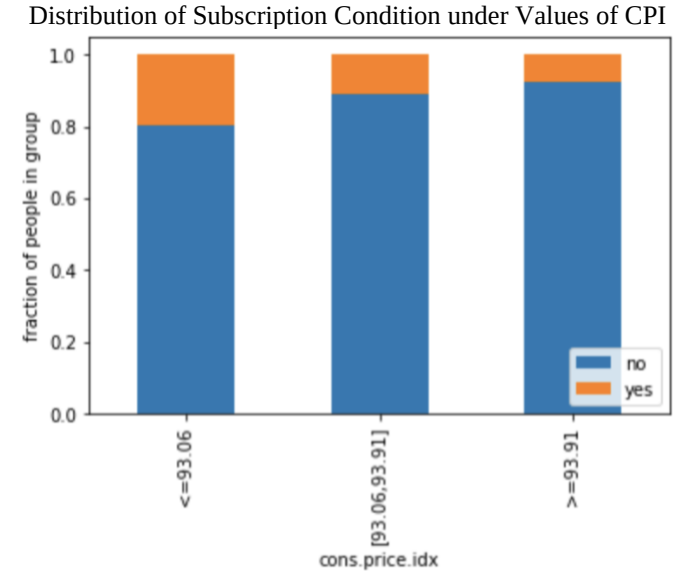
- Higher subscription rate appears in the cellular contact which implies that using cell phones to do marketing is more effective. This is because nowadays the adoption of cell phones is much higher than telephones and people pick up cellular phones calls more easily.

Example 3: previous (number of contacts performed before this campaign and for this client)



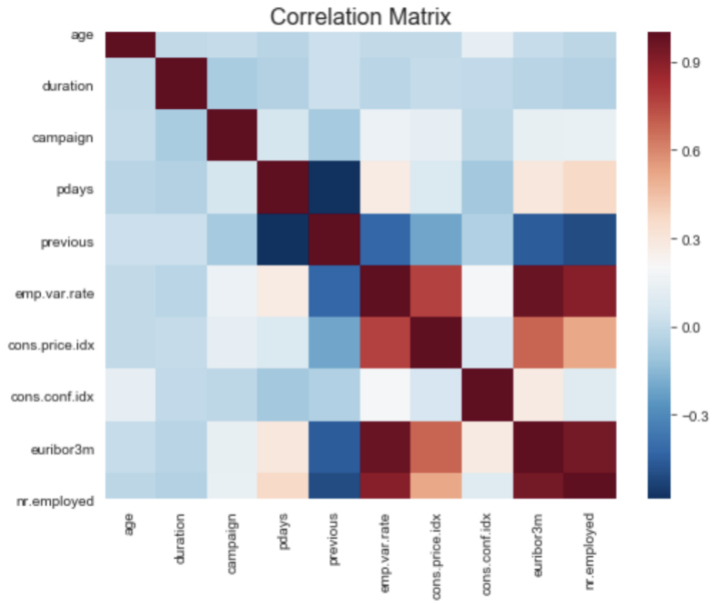

- The average number of contacts under ‘no’ is about 0.13, while under ‘yes’ is 0.49. It is obvious that the bank needs to keep contact with clients frequently and build friendly relationship with them in order to make successful marketing.

Example 4: consumer price index




- Consumer price index is directly proportional to inflation. During high inflation periods, consumers tend to simply turn to saving through options, and maybe they will buy gold or Treasury Inflation Protected Securities. Therefore, during the period of high inflation, clients have a lower possibility to subscribe.

Example 5: Correlation matrix of numerical features




- Employment variation rate shows strong positive correlation with consumer price index, Euribor 3-month rate and number of employees.
- Consumer price index shows relatively strong positive correlation with Euribor 3-month rate and number of employees.
- Euribor 3-month rate shows a very strong positive correlation with number of employees.

## 3. Methods:

3.1 Preprocess on Missing Values:

Two continuous features (‘pdays’, ‘duration’) have missing values. For ‘pdays’ the MCAR test gives a p-value of 1. Since the p-value is greater than 0.05, the null hypothesis should be retained, and so those rows should be removed. However, due to more than 79% of the entire rows are included, the column 'pdays' is dropped from the model instead of the rows. Since ‘duration’ has a dominant impact on the target variable as suggested by the description of the dataset, this column is also dropped from the model.

Six categorical features ('job', 'marital', 'education', 'default', 'housing', 'loan') have missing values. Since they only take very small percentages, they are rolled into the largest category under each feature.

3.2 Preprocess based on EDA:

The EDA process shows some rows that are low percentage of the features. ‘yes’ in ‘default’ feature is only about 0.0073%, and ‘illiterate’ in ‘education’ is only 0.044%. Since those rows are not representative enough, they are dropped from the model input.

3.3 Preprocess using Ordinal, Onehot, Label Encoders and Standardscalar:

After preprocessing missing values, 18 features are classified into categorical features and continuous features and require further preprocess.

| Methods | Features | Comments |
|---|---|---|
| **Onehot encoder** | 'job', 'marital', 'default', 'housing', 'loan','contact','poutcome' | Unordered categorical features. |
| **Ordinal encoder** | 'education','month','day_of_week' | Ordered categorical features. |
| **Standardscalar** | 'age','campaign','previous','emp.var.rate', 'cons.price.idx','cons.conf.idx', 'euribor3m','nr.employed' | Unlimited numerical features |
| **Label encoder** | 'y' | Binary target variable |

After preprocessing, the input data of the model contains 36 columns and 41,167 rows.

3.4 Machine Learning Pipeline:

This ML pipeline is specifically designed for reducing average time of running the large dataset on a working laptop. If the dataset has thousands of rows, these steps can be combined and the pipeline will be less redundant.

- Read the preprocessed data
- Calculate baseline of F1 score
- Split the data into train and test with a test size of 20%
- Tune parameters based on best F1 score
- Cross validate with n_splits = 10 on F1 score using the tuned parameters
- Calculate uncertainties using control variable method:
    - Calculate the uncertainty of F1 score caused by splitting through running 6 simulations with different random states of splitting using the tuned parameters
    - Calculate the uncertainty caused by non-deterministic ML models through running 6 simulations with different random states of classifiers using the tuned parameters
- Calculate confusion matrix and accuracy score based on the first-time simulation

The project uses F1 score to rank the model because the raw dataset is highly imbalanced, and the False Negatives and False Positives are more crucial. F1 score gives the harmonic mean of Precision and Recall and emphasizes more on the incorrectly classified cases. The accuracy score in the pipeline is only used to give a general perspective on the models. Since it is not an important metric in this case, there is no further tuning and CV based on it.

3.5 Model Description:

| Classification Models | Tuned Parameters | F1 Score | Uncertainties due to Splitting | Uncertainties due to non-deterministic model | Accuracy Score |
|---|---|---|---|---|---|
| **Random forest** | max_feature = 15 ∈ [10, 20], max_depth = 9 ∈ [5, 10] | 0.367 | 0.00450 | 0.00186 | 0.900 |
| **XGboost** | max_depth = 11 ∈ [7, 14] | 0.379 | 0.00760 | 0.0 | 0.887 |
| **Gradient boost** | max_depth = 11 ∈ [10, 14] | 0.391 | 0.00692 | 0.000856 | 0.842 |

The baseline of F1 score is 0.202 using the calculation suggested by: https://stats.stackexchange.com/questions/390200/what-is-the-baseline-of-the-f1-score-for-a-binary-classifier.

The baseline of accuracy score is 0.887.

## 4. Results:

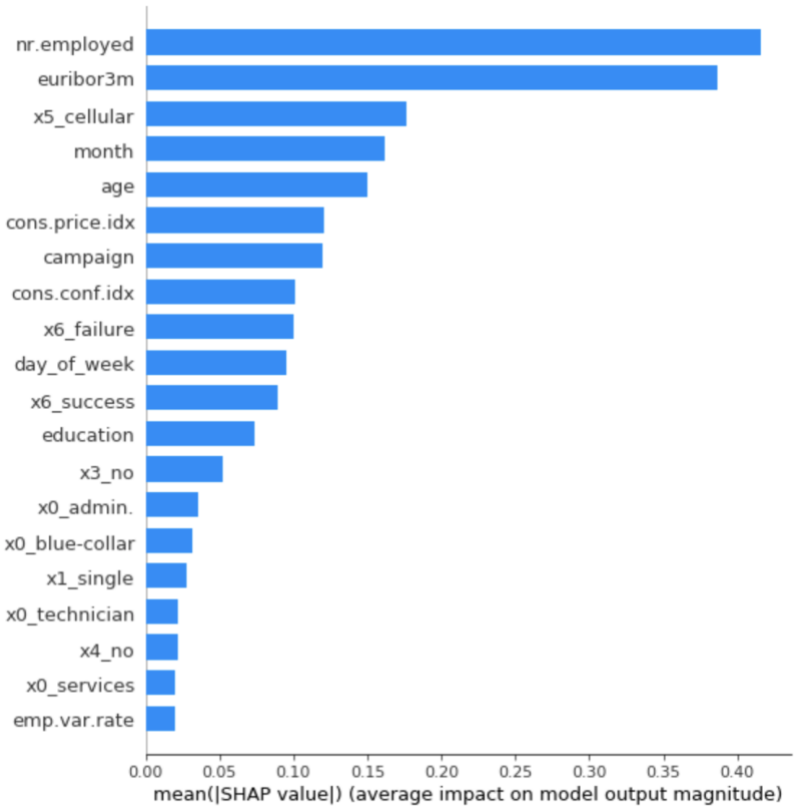
4.1 Model Comparison:

All the F1 scores are above the baseline, and only the accuracy score of gradient boost is below the baseline. Based on the F1 score, gradient boost performs the best and random forest performs the worst. The accuracy score flips the result. In general, the three models give very similar results by adding uncertainties to the F1 scores.

4.2 Feature Importance (based on XGboost):

The following figure shows SHAP explanation force plots for a the 10th person in the dataset.


SHAP values explain the predicted subscription probabilities of this person. The baseline-the-average predicted probability- is -2.301. The reason of getting the negative values is that the '0' responses of the dataset are 9 times the '1' responses. This person has a lower subscription possibility mostly because of the external factors. For example, in May (month = 2), people have just paid taxes a month or two ago, and a lot of them might not have enough savings to buy deposit. Also, as illustrated by the global feature importance graph below, the big economic environment can affect the subscription heavily.




Number of employees at the bank is the most important feature because the bank can delegate more people to work on this campaign. Those employees can also specialize in their jobs, making their work more efficient, therefore successful. Euribor3m reflects the bank to bank interest rate, which is a representation of the banks' confidence in the market. Therefore, the more confident the banks are in the market, the more confident the consumers, making them more susceptible to the campaign.

## 5. Outlook:

According to the raw data, the ratio of the binary response of the target variable is around 89:11. It would be better to collect similar numbers of responses. This can allow the prediction model adequately to describe what variables make a difference between the responses. Here, the dataset has too many 'no' responses, which would push the model to weigh the features that cause 'no' more. Also, tuning the models' parameters in wider ranges and running more simulations on them can improve the accuracy.

Total words: 1997.

## Reference:

Data Resource: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Publications:

Moro, S., Cortez, P. and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, pp.22-31.

Moro, S., Laureano, R.M., & Cortez, P. (2011). Using data mining for bank direct marketing: an application of the CRISP-DM methodology.