Analyzing the Bank Marketing Data by using different Classification Algorithms

Dipesh Chand

Østfold University College

Halden, Norway

[dipesh.chand@hiof.no](mailto:dipesh.chand@hiof.no)

**ABSTRACT**

This project applied different classification techniques to build the model to predict whether the customer will subscribe bank long-term deposit. A Portuguese retail bank collected data from 2008 through 2013. We will analyze the small set of data related to the bank client based on telephone communication.

The Portuguese Bank had an issue of revenue declined, so they conducted a survey and campaign to identify existing clients that have higher chance to subscribe for term deposit and focus marketing effort of such customers. A customer-based analysis of banking services allows for understanding of the possible effects of the concentration on a wide variety of banking resources into a small group of national enterprises[.](#page6) This kind of study projects could be helpful to determine the likelihood of procurement of financial services.

**1 INTRODUCTION**

Businesses today often launch marketing campaigns to boost the sale of their products and services. While digital marketing becomes popular and have many advantages over traditional marketing, traditional marketing methods are still providing physical customer experience difficult to be offered by digital marketing and could not be completely replaced by digital marketing 1 . One of the drawbacks of traditional marketing is that it is typically more expensive than digital marketing since it involves one or multiple types of activities such as phone calls, customer visits, or physical prints etc. These activities often require significant efforts and investment from businesses. Therefore, for traditional marketing, it is important to target marketing activities towards desirable customers who are more likely to buy products and services than others. It will not be cost-effective if marketing campaign targets are simply randomly chosen without going through thorough review and selection.

Machine learning can provide a data-driven approach to help marketing campaign more targeted to desirable customers. In this project, using bank telemarketing as an example of traditional marketing, a machine learning model was developed and demonstrated its effectiveness in maximizing business return while minimizing marketing effort.

**2 PROBLEM STATEMENT AND DATASET DESCRIPTION**

The most effective strategy to progress a business marketing at a least conceivable overhead is constantly seen as the fundamental issue by the supervisor. The foremost tremendous part of the challenge is to recognize the promising and potential clients of the displaying thing with limited data. Realizing that specific information which chooses the promising and potential client would empower executive to put more assets on positive portion towards the items and cut down the budget spent on non-promising client, so that to dispose of bottlenecks and make a progressively productive advancing way.

The dataset is related to direct marketing campaigns run by the Portuguese bank and contains information on various features of interest for approximately 41,188 customers. The dataset has been taken from the UCI machine learning repository <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>. The features of interest can be broken down as follows:

**age** - Age of the client- (numeric)

**job** - Client’s occupation - (categorical) (admin, bluecollar, entrepreneur,housemaid, management, retired, selfemployed, services, student, technician, unemployed, unknown)

**marital** - Client’s marital status - (categorical) (divorced, married, single, unknown, note: divorced means divorced or widowed)

**education** - Client’s education level - (categorical) (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)

**default** - Indicates if the client has credit in default – (categorical) (no, yes, unknown)

**housing** - Does the client as a housing loan? - (categorical) (no, yes, unknown)

**loan** - Does the client as a personal loan? - (categorical) (no, yes, unknown’)

**contact** - Type of communication contact - (categorical) (cellular, telephone)

**month** - Month of last contact with client - (categorical) (January- December)

**day of week** - Day of last contact with client - (categorical)

(Monday - Friday)

**duration** - Duration of last contact with client, in seconds - (numeric) For benchmark purposes only, and not reliable for predictive

modeling

**campaign** - Number of client contacts during this campaign -

(numeric) (includes last contact)

**pdays** - Number of days from last contacted from a previous campaign - (numeric) (999 means client was not previously contacted)

**previous** - Number of client contacts performed before this campaign - (numeric)

**poutcome** - Previous marketing campaign outcome - (categorical) (failure, nonexistent , success)

**emp.var.rate** - Quarterly employment variation rate - (numeric)

**cons.price.idx** - Monthly consumer price index - (numeric)

**cons.conf.idx** - Monthly consumer confidence index - (numeric)

**euribor3m** - Daily euribor 3 month rate - (numeric)

**nr.employed** - Quarterly number of employees - (numeric)

**Output variable** (desired target) - Term Deposit – subscription verified (binary: ‘yes’or ‘no’)

The initial impression that can be created using the dataset are as:

- Total 41188 records

- 10 numeric attributes : age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed

- 10 Factors:

- 10 multi-valued categorical attributes : job, marital, education, default, housing, loan, contact, month, day\_of\_week, poutcome

- 1 target attribute y

- No missing values: Preprocessing should be easier

**3 RELATED WORK**

Moro, Cortez, and Rita [1] has utilized semi-automatic modeling procedure. In their study they have picked information from July 2012 and utilized 22 highlights of data. They have compared about 4 distinctive sort of data mining model. They are "logistic regression, decision trees, neural network (NN) and support vector machine". There are two measurements utilized in the investigation known as AUC (area of the receiver operating characteristic curve) and ALIFT (area of LIFT cumulative curve). The testing model was utilized in advancement stage utilizing the most recent information of July 2012 and another model called a rolling window scheme. The research has shows the result of AUC as 0.8 and ALIFT as 0.7. This has permitted 79% of subscribers as half possible customers. There two extraction strategies are utilized which are sensitive analysis and DT. They were applied to NN. This uncovered several key traits which was tenable and important to telemarketing managers for the campaign.

Vajiramedhin and Suebsing[2] suggested that the performance of the predictive model with the number of smaller features can be improved. Their experiment on Direct Bank Marketing dataset can enhance the predictive model performance both of the TP rate and the ROC rate while it employs the smaller storage space, reduces the computation time and gains the higher predictive performance.

Another study Elsalamony and Elsayad [3], evaluate and compare the classification performance of the two different techniques models Multilayer perceptron neural network (MLPNN) and C5.0 on the bank direct marketing dataset to classify for bank deposit subscription. In their study they used statistical measures; Classification accuracy, sensitivity and specificity and found that C5.0 has slightly better performance than MLPNN. And also Importance analysis has shown that attribute "Duration" in both models has achieved the most important attribute.

**4 METHODOLOGY**

Since the dataset contain both numerical and categorical data, we could use various data mining techniques. This project particularly utilizes the Data Classification Analysis technique to examine a dataset related to direct marketing campaign of a Portuguese banking institution. The objective of this classification technique is to predict if the client will subscribe to a Term Deposit and to improve the performance of model and increase the classification accuracy using various approaches. In order to obtain more accurate and precise model to predict desired output, we will performed several classification techniques and model such as Decision Tree using C5.0, Random Forest Model and eXtreme Gradient Boosting (XGBoost). We will perform correlation analysis to see if there is any relation- ship between predicted attribute (client subscribe term deposit) and other explanatory attributes. The next method, classification model (decision tree), will be helpful to study the customer pattern and accuracy of the applied model. After we perform all of the above techniques, we would be able to understand the data and suggest the best fit model for prediction of “customer term deposit” more accurately and precisely. We perform the following steps in our study:

1) Data Understanding and Exploring: Cleaning, and Visualization

2) Feature Engineering

3) Data Modeling:

4) Model Evaluation

**4.1 Data Understanding and Exploring:**

We used R language with RStudio IDE in this project for analysis. The first step is to load the dataset into a dataframe for easy manipulation and exploration. The initial findings of this dataset are like:

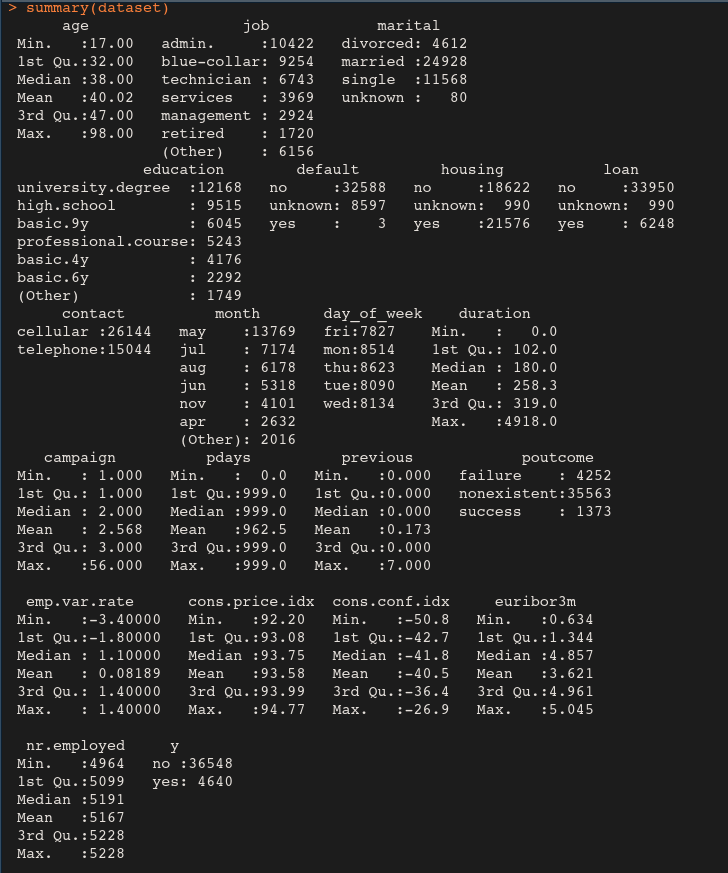
- Most of the clients have never been contacted since contact is unknown for 28.79%.

- 86.34% of the time outcome of previous marketing campaign is unknown.

- Duration seems to have lot more variation, it may be a good predictor

- Data is very imbalanced, only 11.26% yes in outcome.

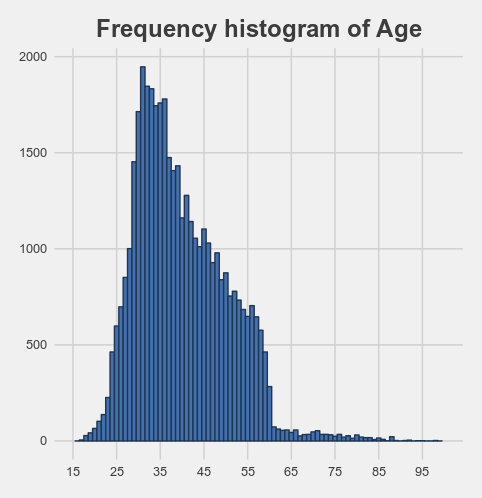
These insights were the very basic details found by just looking at the summary of the datasets, The Fig. 2 displays the summary and clear picture of the dataset.

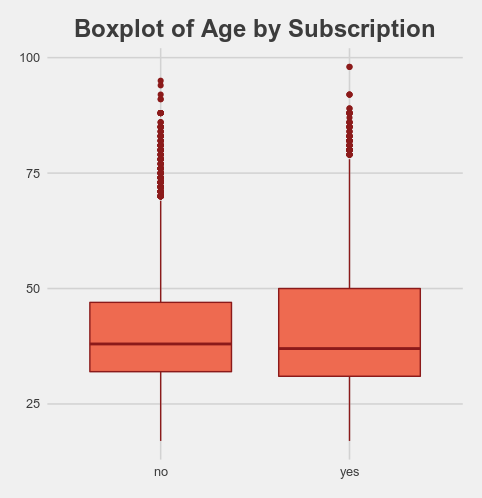
*Figure 1: Summary of the dataset.*

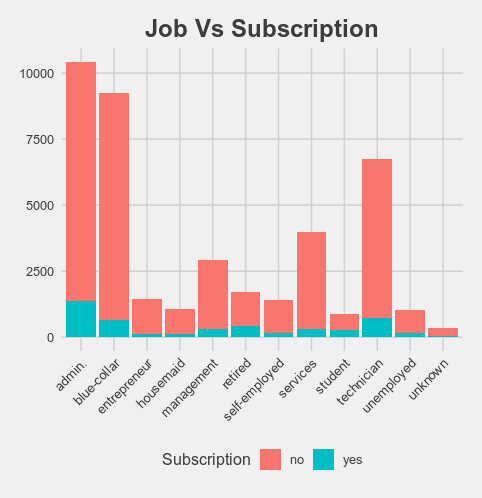
**Data Cleaning:** Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Although it is clearly mentioned that there are no missing values in the dataset but we didn’t want to take risk so we performed missing value check and found that the dataset is free from missing value. So we don’t have to perform any task to manage the missing value.

**Data Visualization:** The number of plots were created to visualize the data so that it would be easy to understand the information provided by the dataset. and more over we also plot combination of variable to the outcome to demonstrate the relation between them. By examining the histogram of numerical features, although most of features are not distributed normally, there are few features with extremely skewed distribution except ‘pdays’. However, UCI Repository intentionally used ‘999’ for ‘pdays’ to arbitrarily represent clients not previously contacted [2] and is not really outliers. Thus, no data transformation was performed.

This dataset is heavily biased for class label (‘y’) 88.7% of samples (36548) are labeled as “no”and only 11.3% (4640) are labeled as “yes”.

*Figure 2: Histogram showing frequency of Age*

*Figure 3: Box plot of Age by subscription*

*Figure 4: Plot demonstrating the distribution of Yes/No based on Job attribute*

**4.2 Feature Engineering :**

**Feature Binning:** The Age variable is widely scattered, shown in Fig. 2 and doing classification on this type of data is very challenging, so binning of data is the best option. We divided the data into four category as

0-19→ "Teens",

20-35→ "Young Adults",

36-60→ "Adults",

61-100→ "Senior Citizens"

By binning it may improve accuracy of the predictive models by reducing the noise or non-linearity.

**Feature Selection:** Checking the predictor variables that are highly correlated with each other. Two metrics are used - Correlation factor and VIF.

At first we created the correlation Matrix of the dataset and then find attributes that are highly correlated, here we set cutoff as 0.75. We got euribor3m and emp.var.rate attributes highly correlated. Then we applied variance inflation factor (VIF). VIF quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

Again we found euribor3m variable have greater value. So analyzing Correlation factor and VIF it is decided to remove the 'euribor3m' variable before performing any further activities.

**4.3 Data Pre-processing**

The dataset is divided into training data and test data with the intention of using the training data to find the parameters of the particular model being used (fitting the model on the training data) and then applying this to the test data to determine the model’s performance and to draw conclusions about its predictive capability. This can be done with a sample.split function call by specifying split ratio. In this project the dataset is split into training set and testing set with 80% for training and 30% for testing.

Since we have imbalanced data, as discussed already in the earlier section, we need to deal with splited dataset. Synthetic Data generation method is used to balance the training set. Specifically, SMOTE Technique is used and make set balance up to 55% No and 45% Yes.

In order to model the data, we performed three classification techniques, 1) Decision Tree Model with rpart 2) Decision Tree Model with C5.0 3) Random Forest Model 4) eXtreme Gradient Boosting (XGBoost)

**4.4 Data Modeling**

**4.5 Model Evaluation**

**5 RESULT**

**6 CONCLUSION**

**7 FUTURE PLAN**

**REFERENCES**