**BANK TELEMARKETING ANALYSIS**

*Final project paper*

*By*

**KARTIK VENKATESHWARAN**

*Abstract*

Major international and national financial institutions create marketing campaigns to gauge the popularity of a newly launched product or service. This kind of campaign allows institutions to collect relevant data on a collection of random users and helps them assess the success/ failure of the product or service. This paper will attempt to analyze data from a Portuguese bank which gathered data around a telemarketing campaign to see which factors influence a customer to subscribe to a deposit with the bank. This will help the bank market more effectively to a particular group of users and be able to potentially drive more revenue and gain customers.

The client in our problem is Banco de Portugal which is the Portuguese central bank. Since the dataset includes European financial market indicators, we decided to pick this client. We could use our classifier to any bank’s marketing campaign results data with a little feature engineering to eliminate the European financial market indicators if required.

In this paper, we will detail the problem statement and look at the results of our exploratory data analysis and examine relationships between various variables. After looking at these relationships, we will create three models to predict what factors most influence a customer’s decision to make a deposit. We will then make recommendations to the bank on how to tailor their next campaign based on the results of our analysis.

*Problem statement*

Today major institutions conduct surveys and polls of their audience as a means of understanding what drives consumer behavior. These polls often help institutions determine whether a potential customer finds a particular product appealing, and if yes, what are the factors influencing the customer’s decision to buy and use the product. In our problem, the bank in question has collected data around a sample of their customers including demographic, financial and data around the last campaign interaction if any. The challenge is to predict customer behavior and to analyze if any of these factors influence their willingness to invest money with the bank via a deposit.

*Our dataset variables*

Below is a description of our dataset variables.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Data type** |
| Age | Age of the respondent | Numeric |
| Job | Job of the respondent | Categorical |
| Marital | Marital status of respondent | Categorical |
| Education | Educational status of respondent | Categorical |
| Default | Credit default status | Categorical |
| Housing | Indicates if the respondent has a housing loan or not | Categorical |
| Loan | Indicates if respondent has an active loan | Categorical |
| Contact | Type of communication. Values include cellular, telephone | Categorical |
| Month | Contact month of year | Categorical |
| Year | Contact year | Categorical |
| Day of week | last contact day of the week | Categorical |
| Duration | Last contact duration in seconds | Numeric |
| Campaign | Number of contacts performed during this campaign and for this client | Numeric |
| pdays | number of days that passed by after the client was last contacted from a previous campaign | Numeric |
| previous | number of contacts performed before this campaign and for this client | Numeric |
| poutcome | outcome of the previous marketing campaign | Categorical |
| emp.var.rate | employment variation rate | Numeric |
| cons.price.idx | consumer price index - monthly indicator | Numeric |
| gold.price | Price of gold – monthly indicator | Numeric |
| cons.conf.idx | consumer confidence index - monthly indicator | Numeric |
| euribor3m | euribor 3 month rate - daily indicator | Numeric |
| nr.employed | number of employees - quarterly indicator | Numeric |
| y (target variable) | has the client subscribed a term deposit? | Binary (Y or N) |

*Data Wrangling*

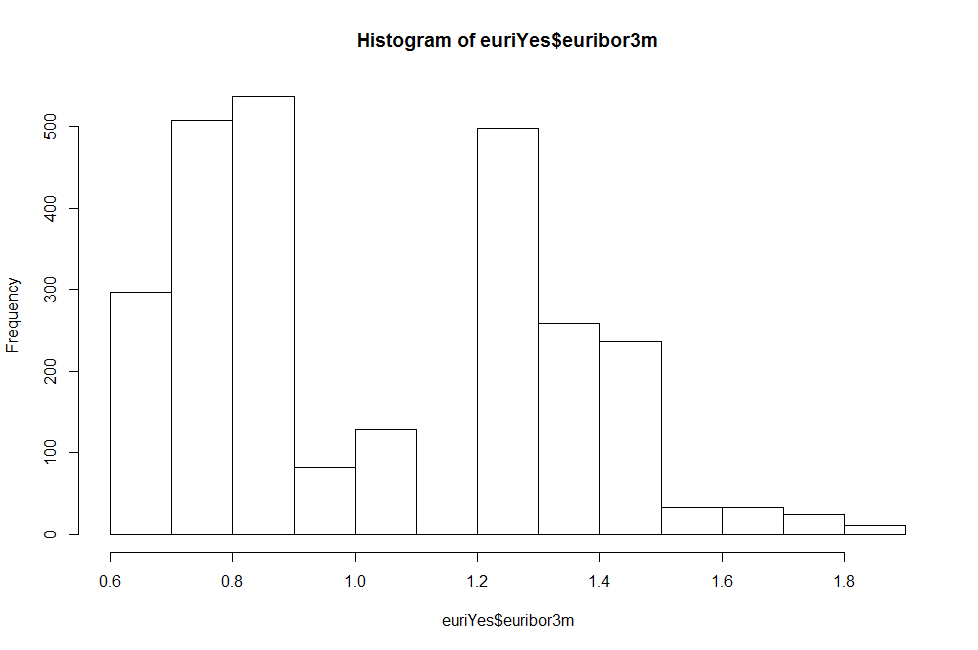
In addition to the variables provided in the problem set, we added new variables “year” and “gold price”. We felt that adding these new variables, especially gold price, which is usually seen as an economic indicator, would add value to our analysis and help us build a better model. We also tested our dataset for missing values and found none. We took out the variable “duration” from our dataset as it was indicated in the problem statement that this variable is highly correlated to our dependent variable. We decided to split our data into a training and testing set in an 80:20 ratio.

*Exploratory Data Analysis*

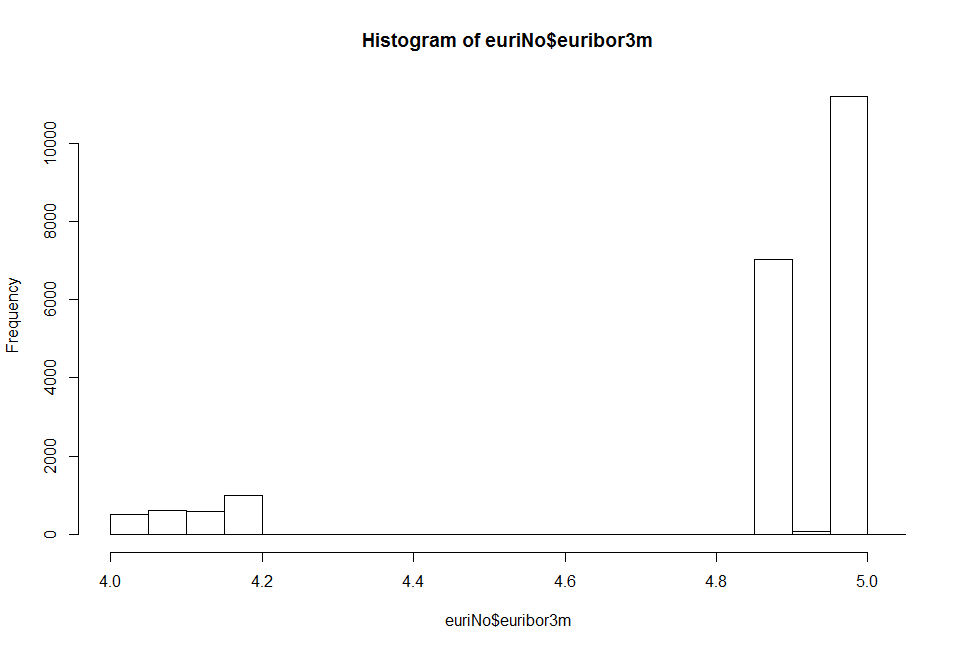
Looking at our data, we find that data is skewed towards one class in our dependent variable. Almost 89% of values say “no” while only about 11% have a value of yes. This is a fairly common scenario seen in data related to marketing campaigns and fraud detection initiatives. We will be accounting for this imbalance while building and analyzing our model and will be judging our model performance against this benchmark. Since this is a marketing campaign initiative where the goal is to reach out to as many people as possible, we will go with a model which forecasts more people with a yes than the data we have.

One observation that immediately stands out is when we compare customer behavior with the EURIBOR interest rate. The Euro Interbank Offered Rate is a daily reference rate, published by the European Money Markets Institute, based on the averaged interest rates at which Eurozone banks offer to lend unsecured funds to other banks in the euro wholesale money market (or interbank market). It is observed that customers are more likely to make a deposit with the bank when the Euribor rate is low as compared to when the Euribor rate is high, which can be used as an indicator of consumer financial behavior.

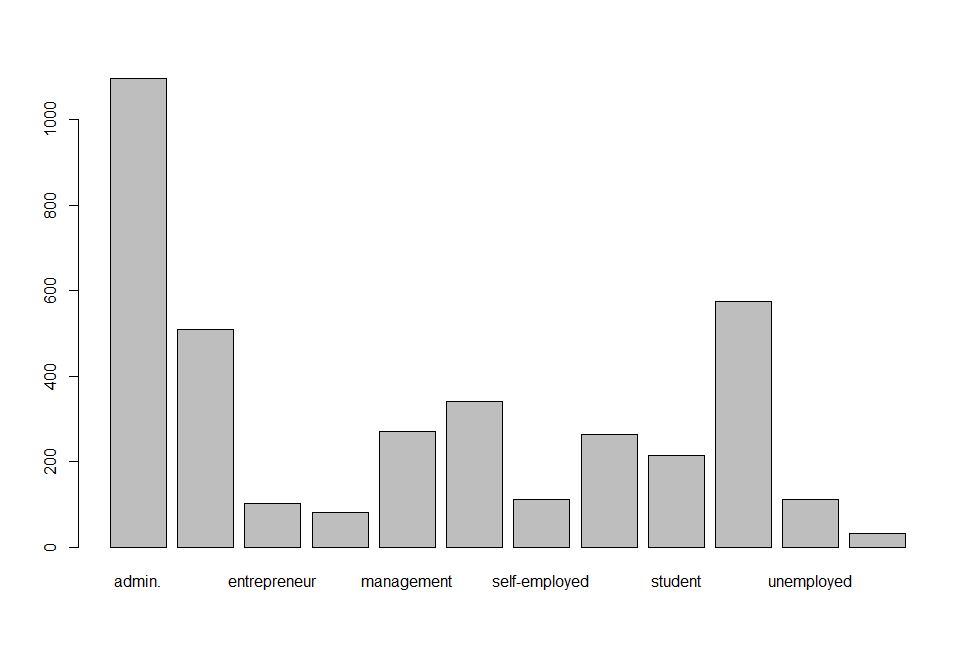
Graph below shows euribor rate for all customers who responded positively. As we can see a large number of respondents provided the desired response when the euribor rate was below 2. Hence we can infer from the distribution that customers would tend to make a deposit when the financial indicator euribor rate is below 2.



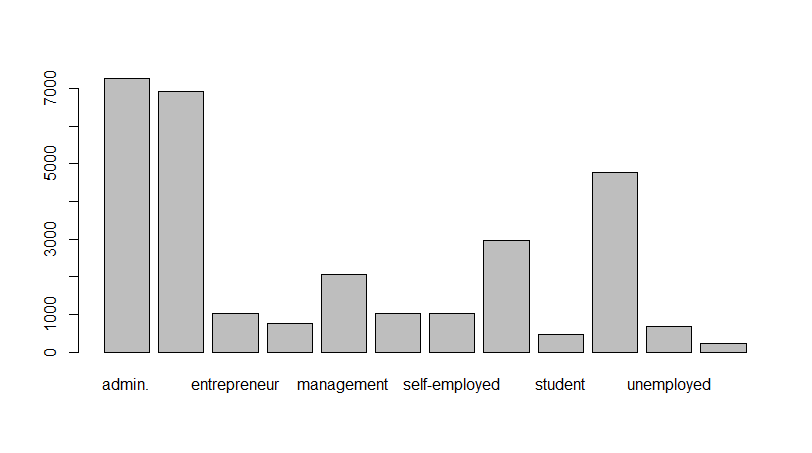
In graph below, we look at the euribor rate distribution for people who said no. As we can see a significant proportion of our population said no when the euribor rate was above 4. Looking at this we can come to the conclusion that when the euribor rate is above 4 people are generally hesistant to make a deposit with the bank.



Next let us look at the job variable to see if there is any relationship between a customer’s occupation and their choice of making a deposit or not. First we examine the job variable for people who said yes. We notice that the distribution shows admin, technicians and blue collar workers are the top 3 job categories for people who subscribed to a deposit.

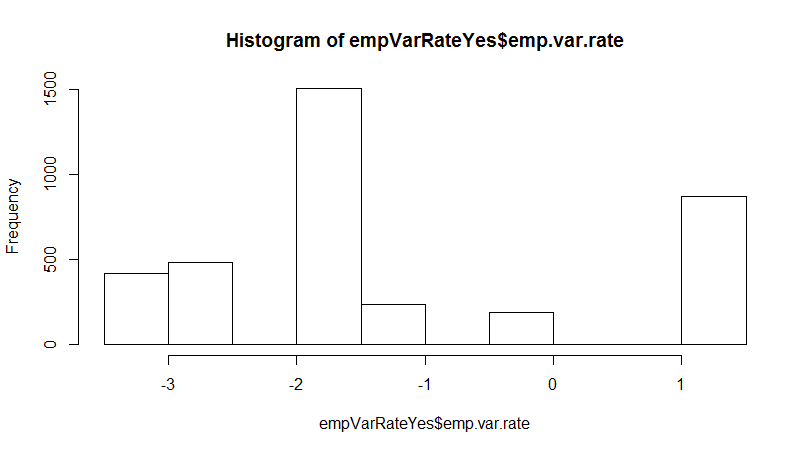


Next we will look at the job description distribution for people who said “no”. We notice that the distribution is fairly the same for all categories except for the “retired” and “student” job categories. The number of people with these job categories shows a decline in the “no” group. We will examine these variables in more detail



When we look at the distribution of the dependent variable for people in the “retired” and “student” groups we find that about 25% of retirees and 30% of students are more likely to subscribe to a deposit. When compared to our population proportion of 11% these numbers are high and could be considered a significant factor. Therefore we can assume that the job categories of “retired” and “student” have a significant, but not high, influence on the dependent variable.

We also look at our dependent variable against the employment variation indicator. This is an indicator to denote if the unemployment rate is rising or falling. Our data indicates that when the unemployment rate is falling, people are more likely to invest in a deposit than when the unemployment rate is rising. This follows fairly conventional financial wisdom which shows that customers are careful with their investments during times of financial stress.



We looked at all other variables and concluded that the other independent variables have little to no influence on our outcome as the distributions are fairly even for both categories of our dependent variable.

*Building the model*

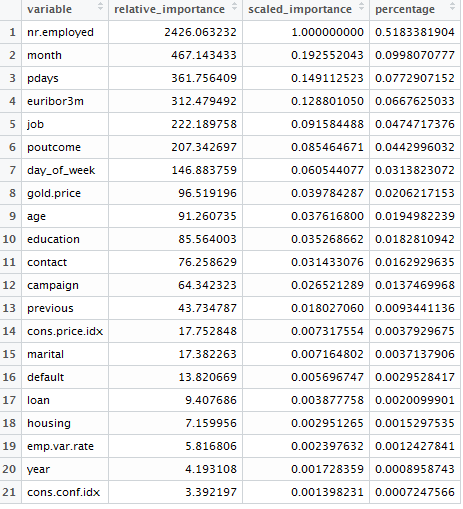
Since our dependent variable is a categorical variable, we will stick to our approach initially specified where we will use the generic logistic regression model to predict our outcome. We will also use Random Forest and Decision Tree algorithms as a test to see how our logit model performs against other benchmarks. We will be using the h2o package to build our models and test them.

We start out by building a logistic regression model. In this model we compare the outcome dependent variable to all our independent variables. We will use h2o’s glm function to build our model using the training data. After we have our model, we used our model with our training dataset to evaluate the dependent variable.

We observed that our logistic model returned an area-under-curve (AUC) of 0.79 which is significantly high enough to be considered a good indicator of model accuracy. When we used this model to predict the outcome for our test data set, this model returned predictions with the value “yes” for about 14% of our variables, which is slightly higher than our calculated proportion of “yes” responses for the whole data set. We also looked at which variables were highly influencing our dependent variable and found the employment variation rate, euribor rate and the month value of “march” to be highly correlated with our outcome variable.

Next we used the Random Forest algorithm to build another model using the same training dataset. We observed a high AUC of 0.77 here, which indicated the model’s accuracy. When we used this model to predict our dependent variable, we got a 12.5 % of our dependent variable as a yes which, while higher than the 11% for our consolidated data set, is short of the almost 14% proportion that our logistic model result gives us.

We also used the Gradient Boosting algorithm to evaluate our classifier. We found an AUC of 0.82, thus finding that this was the strongest classifier out of the 3 models. On evaluation of the results of this model/ classifier, we got 9.8% of our target variable to have the “yes” value. Below is a table showing the variable importance as it was ranked by the Gradeint Boosting algorithm.



Considering that this is a marketing problem and the ideal result of this initiative being to reach out to more customers, we will consider the logistic model to be the most appropriate one to be used by the bank as it will help reach out to more people and the prediction was made with a significantly high degree of accuracy.

*Recommendation based on our analysis*

Based on our EDA and results with the logistic regression model, we would recommend that the bank monitor financial indicators like the employment variance indicator and euribor quarterly rate and ramp up its marketing efforts when both these indicators show a declining pattern. Additionally we would ask the bank to campaign more amongst senior citizens and students as these demographic groups seem to have a high degree of probability of opening a deposit with the bank. Our model also says that the bank can expect more success in their marketing efforts in the month of March, so while we would recommend trying to make more calls than usual that month, we would make this recommendation with a degree of caution as currently there is not enough data as compared to the population for us to make a strong case.

*Recommendations for future research*

For future research and refinement of our findings, we recommend that the bank add more variables around economic indicators such as price of Brent crude, rate of variation in private consumption, the EUR/USD exchange rate etc. to evaluate if our classifier performs any better with these indicators. We would also propose using more computational and demanding algorithms like Support Vector Machines to evaluate and see if we can generate a better classifier. We would also advise the bank to continue to collect more data around this use case as having more data will enable us to generate a better model and predict if there are any additional factors that influence customers’ decision to open a deposit with them.