**BANK TELEMARKETING ANALYSIS**

*Final project paper*

*By*

**KARTIK VENKATESHWARAN**

*Table of contents*

*Introduction*

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 Novo Banco which is the largest Portuguese bank by assets. 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.

*Our dataset variables*

Below is a description of our dataset variables.

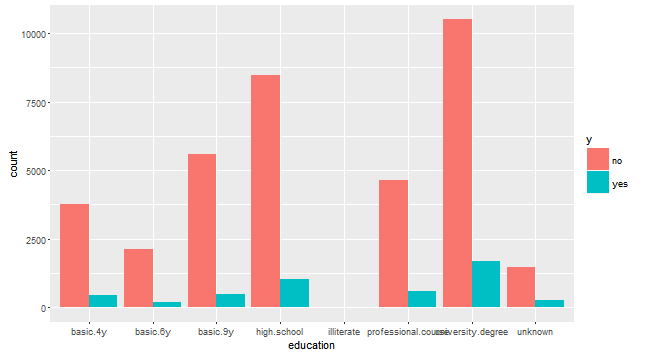
|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Data type** |
| Age | Age of the respondent | Numeric |
| Job | Job of the respondent | Categorical |
| Marital | Marital status of respondent | Categorical |
| Education | Educational status of respondent | Categorical |
| Default | 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*

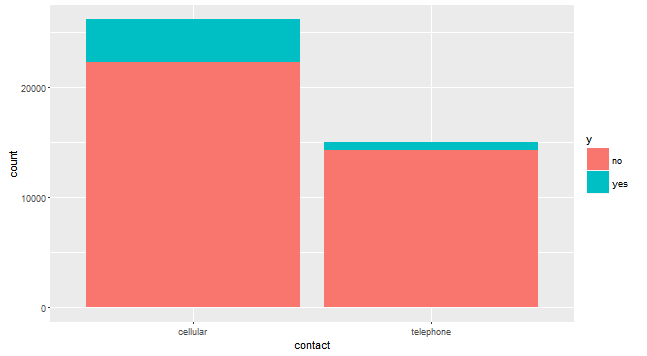
In addition to the variables above, we added new variables “year” and “gold price”. We felt that adding these new variables, especially gold price, which is usually seen as an economic indicator, would add value to our analysis and help us build a better model. We also tested our dataset for missing 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 also took a look at the distribution of some critical variables. Looking at the marriage variable we find that people with a university degree or a high school degree are more likely to make a deposit than someone with a basic 6 year education. 10% of respondents with a high school diploma and 13% of respondents with a university degree responded positively which might indicate a high degree of collinearity between these variables and the outcome variable.

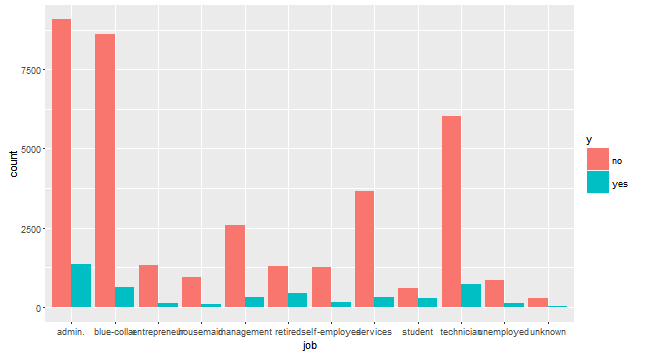


Looking at the contact variable we found that people were more likely to make a deposit if they were contacted by cellular (~15% said yes) than by a fixed telephone line (5%).

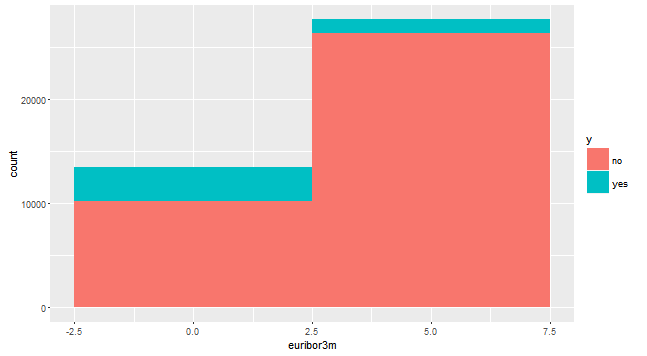
Following plot shows the variance:



Now let’s look at the job variable. We would assume that people with a steadier job would be more inclined to make a deposit than someone with an unsteady job. What we find in our analysis of the job variable is that out of the people who made a deposit after the campaign, about 30% identified as “admins”. If we include the population, people with a job profile of admins made up 25% of the group. Among this 25%, 13% of users made a deposit after the campaign which is greater than our population estimate of 11%. This means we can expect significant success from our campaign if we market to people with job profile of admins.



We also take a look at the euribor3m variable viz. the European interbank interest rate and assess how it affects customers’ decision to make a deposit. We see a clear trend where the number of people willing to make a deposit is greater when the interest rate is between -2.5% to 2.5 percent. As the interest rate goes beyond 2.5%, more people are hesitant to make a deposit with the bank.



*Building the model*

Since the variable we will be predicting is a binary variable (yes/no), we will treat this as a classification problem and build classifier models. We use three popular algorithms used to build classifiers – logistic regression, random forests and gradient boosted machines. Since we have a large dataset with more than 40000 records, we will use the h2o package which is widely used for big data related processing tasks. We have divided our dataset into testing and training sets in an 80:20 ratio.

For our first model, we used logistic regression as our algorithm as this is the go-to method when predicting a categorical variable. We used h2o’s inbuilt glm function with all our variables as we felt due to the small number of variables and also due to the inherent bias in the dataset (almost 90% of results are for one side of the dependent variable), using all variables would yield a more accurate prediction than using just some of the variables.

The significant results from our logistic regression model are as follows:

*MSE: 0.078041*

*R^2: 0.219311*

*AUC: 0.796865*

The factor we took into consideration as most accurately reflecting model accuracy was the area under curve (AUC). We have an AUC of 0.79 which we felt was good enough accuracy for our model. The top 3 factors that came out as impacting our model most are listed in below table:

*Positive impact*: euribor3m, month.mar, year

*Negative impact*: emp.var.rate, month.may, poutcome.failure

We decided to create 2 more models for our analysis – Random Forest and Gradient Boosted machines. Our RF model returned an AUC of 0.77, with age, the euribor rate and job the top 3 factors influencing our model. The gradient boosted model returned the best AUC of 0.82, with the top 3 factors influencing the model being number of employees, month and number of days the client was last contacted by a previous campaign.