Capstone Project2 – Milestone report

1. Define the problem

Banks and financial institutions have huge data and information about their customers. This data can be used to create and keep clear relationship and connection with the customers in order to target them individually for definite products or banking offers. Usually,the selected customers are contacted directly through:personal contact, telephone cellular, mail, and email or any other contacts to advertise the new product/service or give an offer, this kind of marketing is called direct marketing. In fact,direct marketing is the main strategy of many of the banks, financial institutions and insurance companies for interacting with their customers.

But, the challenge is to improve the effectiveness of these direct marketing campaigns. They are often ‘hit and miss’. However, if they are effective, they have tremendous impact on the topline.

1. Identify your client

This problem is faced by banks, financial institutions and insurance companies during their targeted product marketing efforts.

I run a Financial services company I do engage in direct product sales and marketing. I face this problem everyday and I would like to solve this problem.

1. Describe your data set, and how you cleaned/wrangled it

I would like to use ‘Bank Marketing Data Set ’ from UCI data reporsitory.

It has 41188 records. I plan to do a train-test split.

<https://archive.ics.uci.edu/ml/datasets/bank+marketing>

There are three datasets

1. Bank\_full.csv (45211 records)
2. Bank\_additional\_full.csv (41188 records)
3. Bank.csv ( 10% of records randomly chosen from bank\_full)

Data Wrangling:

<https://github.com/dipanjan123/Capstone_project2/blob/master/Data%20Wrangling%20and%20EDA.ipynb>

1. I first checked the datasets in pandas using info, shape, describe etc.
2. Here are the attributes:

Input variables:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue-collar','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')

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')

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - 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)

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')

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)

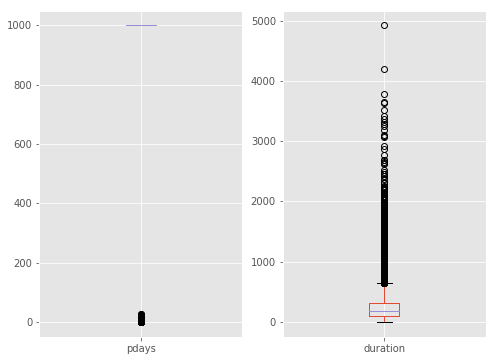
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

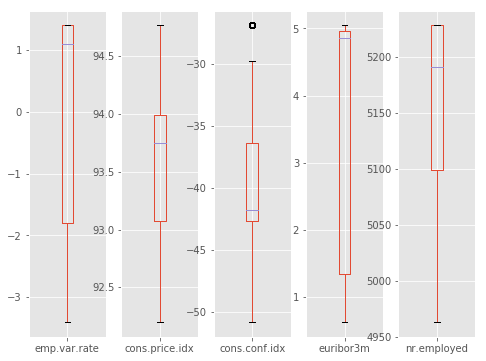
Output variable (desired target):

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

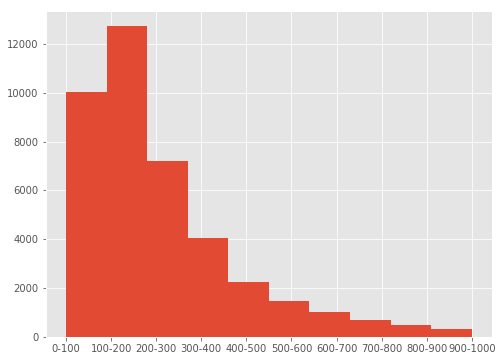
1. I performed boxplot of variables ‘days past previous campaign’ and ‘duration of campaign

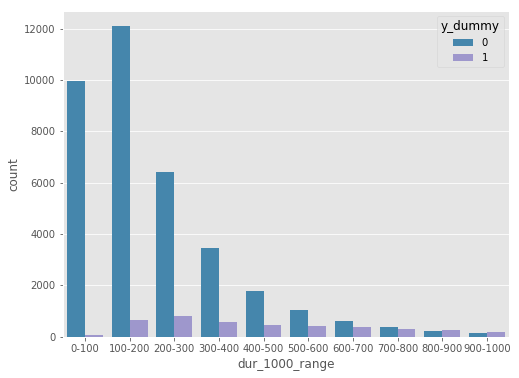


1. I performed boxplot of five macroecenomic variables (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed)



1. I created six new variables: This step makes several continuous variables range bound and also makes the target variable binary(0 and 1)
   1. euribor3m\_range : Creates a range of euribor3m values
   2. cons\_price\_range: Creates a range of consumer price index values
   3. Conf\_index\_range\_neg: Creates a range of confidence index values
   4. duration\_range : Creates a range of campaign duration values
   5. age\_binned : Creates a range of age values
   6. y\_dummy : Makes the target variable 1 and 0.
2. I made a composite group of marital status and age
   1. The average positive response rate is 0.112 for the entire dataset
   2. There are seven group whose response rate was better than average :
      1. Divorced - Senior
      2. Married - Senior
      3. Single - Young
      4. Single - Lower middle
      5. Unknown - Young
      6. Unknown - Middle
      7. Unknown – Senior
3. I also created a composite group for housing, loan and default and observed the response rates.
4. Duration is a very important parameter
   1. I observed that most number of called have been of duration 20 minutes or less
   2. I created a dataframe which contains records of 1000 seconds duration or lower.
   3. I plotted histogram
   4. I observed response rate in different duration ranges
   5. If the duration exceeds 500 seconds chance of success increases drastically.

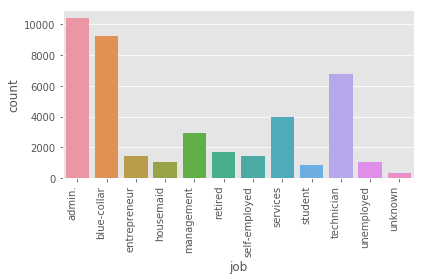


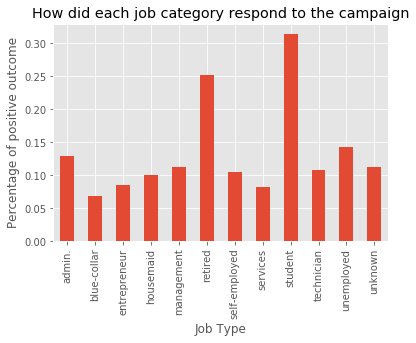


Data Visualization:

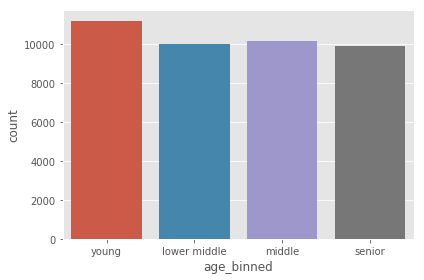
<https://github.com/dipanjan123/Capstone_project2/blob/master/Data%20Visualisation%20Capstone2.ipynb>

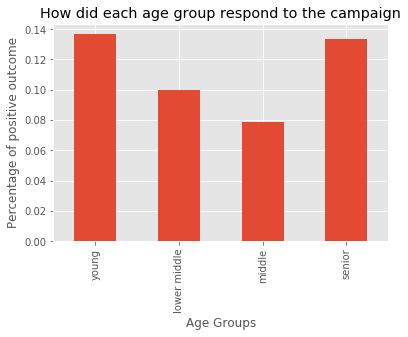
1. I checked the different job categories and how they responded to the campaign. Students, Retired, Unemployed and Admin people responded better than other job categories.



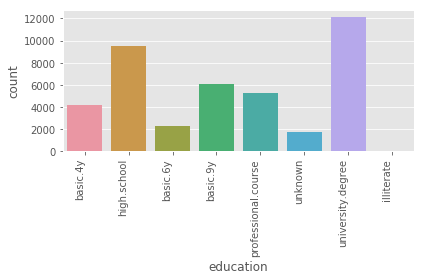


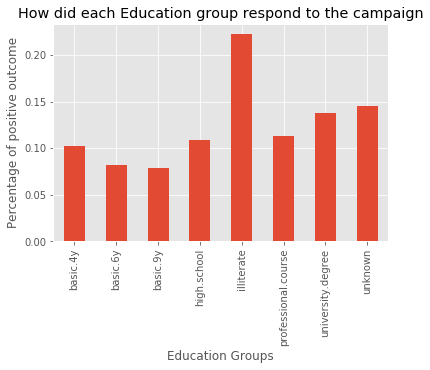
1. I checked what the age groups and how they responded. Young and senior age groups responded better.



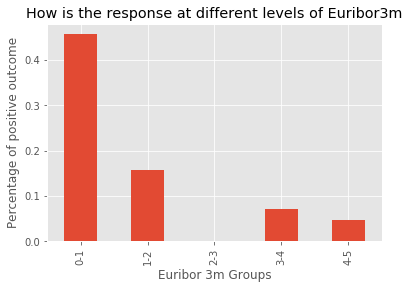


3. I checked the educational groups and how they responded. Illeterate and University degree hoders responded better.

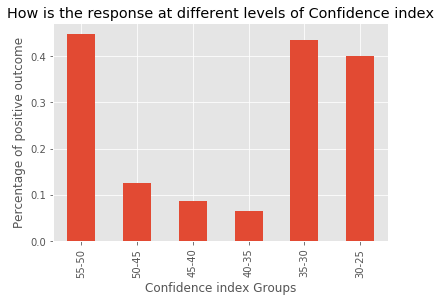




1. I checked the the reponses at different euribor3m levels. Response is better at lower Euribor 3m levels.



1. I checked reponses at different confidence index levels. Response is better at high and low confidence index levels.

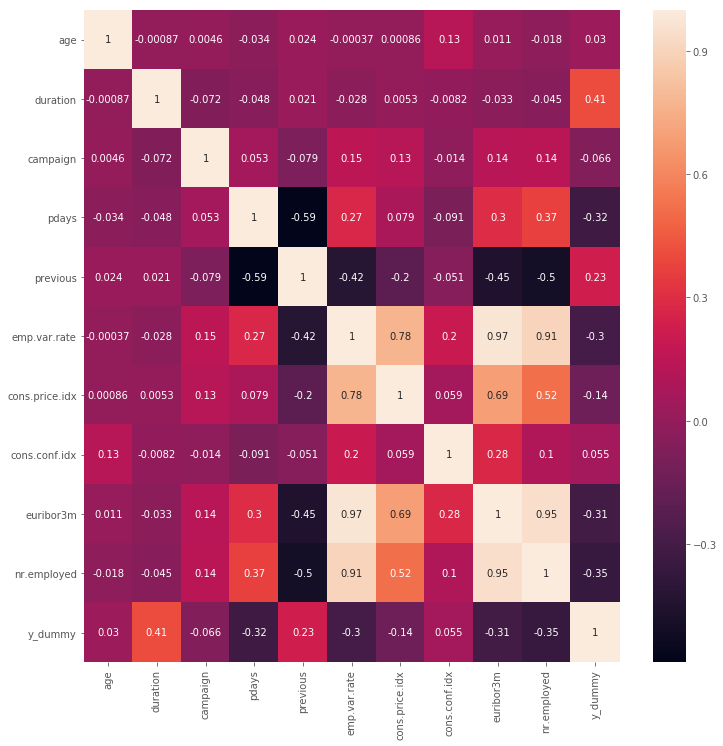


1. I checked responses at different months and ‘day of the week’. No pattern found.

Inferential statistics

<https://github.com/dipanjan123/Capstone_project2/blob/master/Inferential%20statistics%20capstone%20project2.ipynb>

1. I checked the correlation between parameters and tried to figure out which input variable impact the target variable. Here are the findings:
   1. Duration impacts response very strongly positively
   2. Previous contact impacts response positively
   3. Number of days past from last contact impacts negatively
   4. Macro parameter nr\_employed(number of employees) impacts negatively
   5. Macro parameter euribor 3m impacts negatively



1. Next I checked the categorical variables
   1. I conducted chi-square test for each categorical variable.
      1. Education:

Chi2 value for education, p-value, Expected counts (193.10590454149565, 3.3051890144025054e-38, 7, array([[ 3.70555618e+03, 2.03379664e+03, 5.36400554e+03, 8.44309556e+03, 1.59722249e+01, 4.65235418e+03, 1.07972240e+04, 1.53599563e+03], [ 4.70443819e+02, 2.58203360e+02, 6.80994464e+02, 1.07190444e+03, 2.02777508e+00, 5.90645819e+02, 1.37077595e+03, 1.95004370e+02]]))

With high Chi-square value(193) and low p-value, I concluded that education has impact on target variable

* + 1. Job

Chi2 value for job, p-value, Expected counts (961.24244032895535, 4.1897632875638613e-199, 11, array([[ 9247.91822861, 8211.49830048, 1291.97552685, 940.58657862, 2594.59920365, 1526.23482568, 1260.91842284, 3521.87559483, 776.42760027, 5983.37292415, 899.76867049, 292.82412353], [ 1174.08177139, 1042.50169952, 164.02447315, 119.41342138, 329.40079635, 193.76517432, 160.08157716, 447.12440517, 98.57239973, 759.62707585, 114.23132951, 37.17587647]]))

With high Chi-square value and low p-value, I conclude that job has impact on target variable.

* + 1. Day of week

Chi2 value for day of week, p-value, Expected counts (26.144939075871971, 2.9584820052785324e-05, 4, array([[ 6945.25580266, 7554.8623871 , 7651.58308245, 7178.62775566, 7217.67097213], [ 881.74419734, 959.1376129 , 971.41691755, 911.37224434, 916.32902787]]))

With low Chi-square value and low p-value, I conclude that ‘day of week’ has relatively no impact on target variable.

* + 1. Month

Chi2 value for month, p-value, Expected counts (3101.1493514116773, 0.0, 9, array([[ 2335.49422162, 5482.02253083, 161.49694086, 6365.8189764 , 4718.905118 , 484.49082257, 12217.86471788, 3639.00524425, 637.11430514, 505.78712246], [ 296.50577838, 695.97746917, 20.50305914, 808.1810236 , 599.094882 , 61.50917743, 1551.13528212, 461.99475575, 80.88569486, 64.21287754]]))

Examining the chi-square and p-value, I conclude that month has relatively no impact on target variable.

* + 1. Marital status

Chi2 value for marital, p-value, Expected counts (122.65515182252989, 2.0680146484422109e-26, 3, array([[ 4.09243896e+03, 2.21197568e+04, 1.02648165e+04, 7.09876663e+01], [ 5.19561037e+02, 2.80824318e+03, 1.30318345e+03, 9.01233369e+00]]))

With high Chi-square value and low p-value, I conclude that marital status has impact on target variable.