How to improve effectiveness of bank marketing campaigns

Marketing campaigns:
Marketing campaign is usually a promotion to reach certain sales goals within a period of time. The word 'Goal' is very important. The more specific the goal, the better is the chance of the campaign being successful. The specificity also helps the design of the campaign.
Modes and Channels:
The campaigns can be delivered through several channels like, email, sms, phone call, social media, word of mouth, in-store signage etc.
Target Audience:
It is very important to know whom the campaign is targetted to. Customer segmentation is a step that is usually carried out before a campaign. The better the product-market fit, the more effective the campaign will be.
The cons of the marketing campaigns:
The increasing number of marketing campaigns over time has reduced their effects on general public. Firstly, due to competition, positive response rates to mass campaigns are tpically low. Secondly, 'intrusion of privacy' is a drawback of direct marketing. In order to save time and cost, it is imporatnt to to filter the target audience and still maintain a certain success rate.

This Project:

Our objective is to build a classifier to predict whether or not a client will subscribe a term deposit. If the classifier has high accuracy, the banks can arrange a better management of available resources by focusing on the potential customers "picked" by the classifier, which will improve their efficiency a lot. Besides, we plan to find out which factors are influential to customers' decision, so that a more efficient and precise campaign strategy can be designed to help to reduce the costs and improve the profits.

Data Source:

Our data were collected from a Portuguese marketing campaign related with bank deposit subscription for 45211 clients and 20 features, and the response is whether the client has subscribed a term deposit. Our data set is downloaded from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing .

The marketing campaigns were based on phone calls. Sometimes more than one contact to the same client was required.

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)

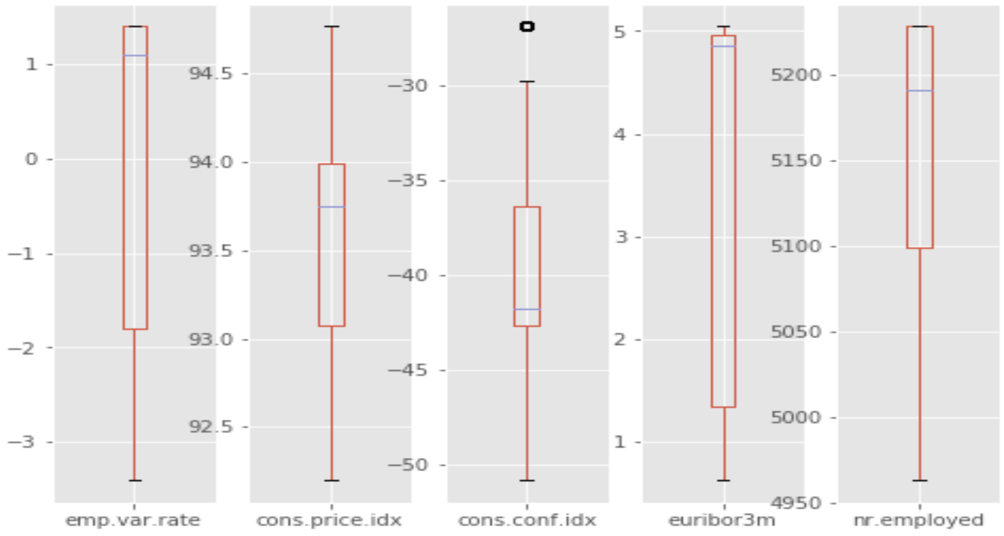
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal 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')

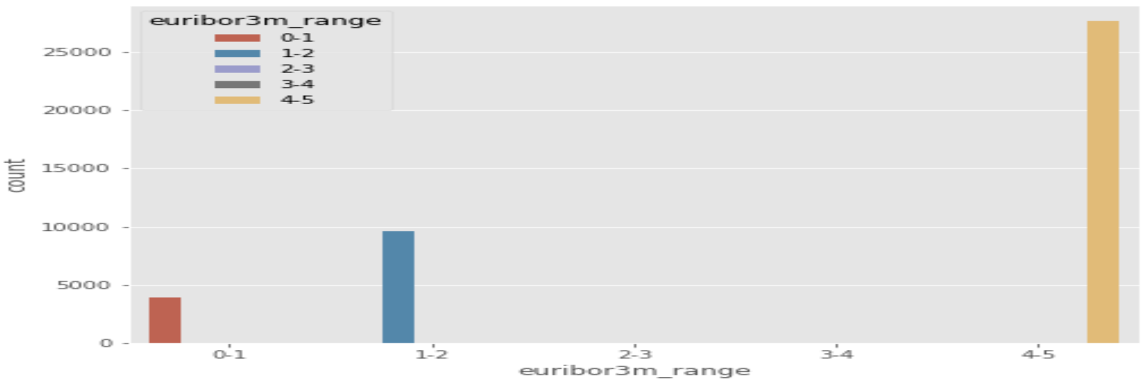
<u>Initial impression about data</u>
Let's discuss about initial impression that can be created using the dataset
Total 45211 records
7 numeric attributes : age, balance, day, duration, campaign, pdays, previous
10 Factors:
6 multi-valued categorical attributes : job, marital, education, contact, month, poutcome
3 yes/no binary attributes: default, housing, loan
1 target attribute y
No missing values: Preprocessing should be easier

Data wrangling and EDA:

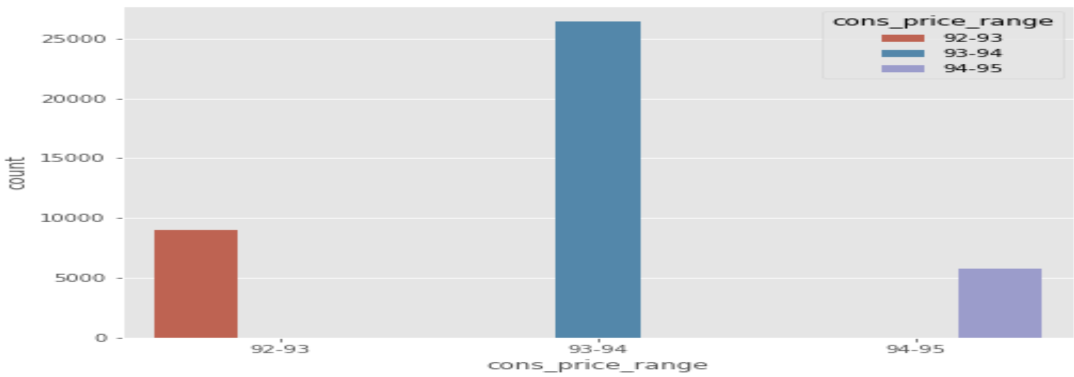
We started with boxplots of the macroeconomic parameters –



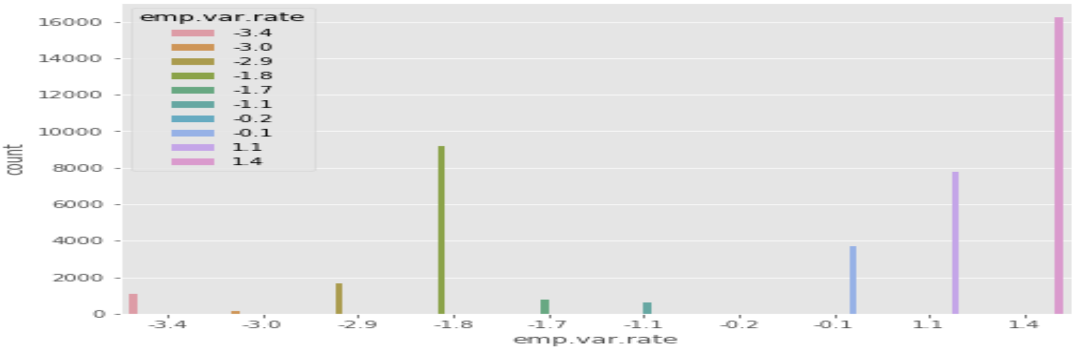
We plotted Euribor -3m rate - quarterly indicator for interbank lending rate



We plotted Consumer price index - CPI



And we plotted Employment variation rate



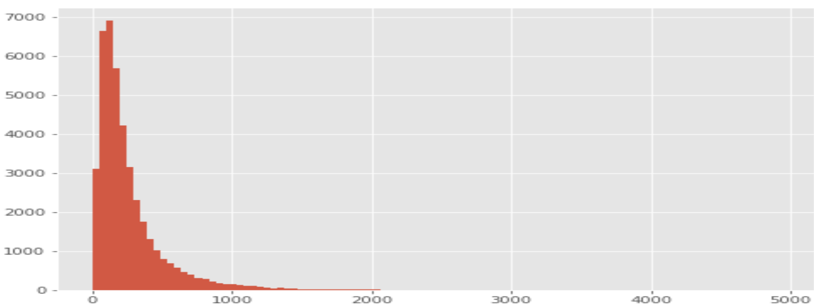
We created a composite group containing marital status and age ranges. The overall average positive response rate for the entire dataset is 11.26%. But, seven groups had better than average response –

1. Divorced - Senior
2. Married - Senior
3. Single - Young
4. Single - Lower middle
5. Unknown - Young
6. Unknown - Middle
7. Unknown - Senior

We created another composite group containing three parameters – 'Has housing loan?', 'Has personal loan?', 'has credit in default?'. The groups having better than average response are –

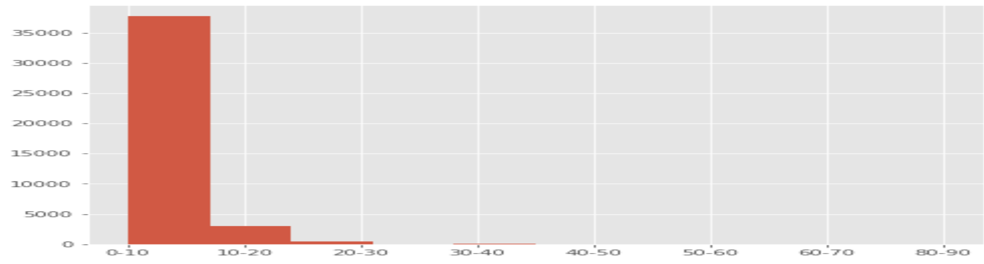
1. Housing loan – yes, Personal loan –no, Default – no
2. Housing loan – yes, Personal loan –yes, Default – no
3. Housing loan – no, Personal loan –no, Default – no (Very much expected)
4. Housing loan – no, Personal loan –yes, Default – no

<u>Duration:</u>
Duration is an important, but, tricky input variable in this dataset. Histogram shows that duration of most of the phone calls have been less than 1000 seconds.

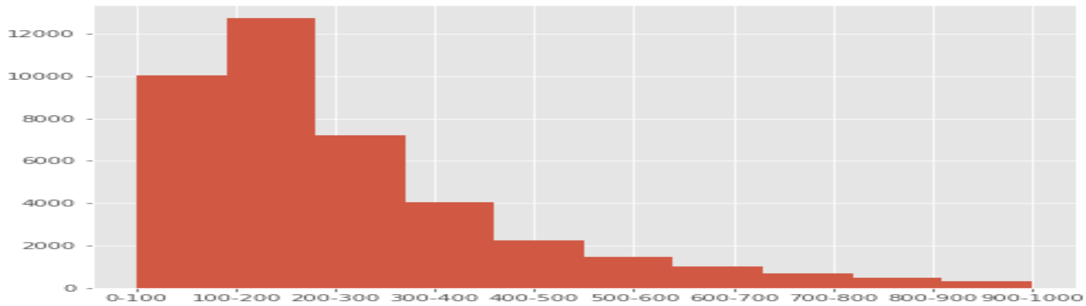



```
max duration 4918
min duration 0
mean duration 258.2850101971448
```

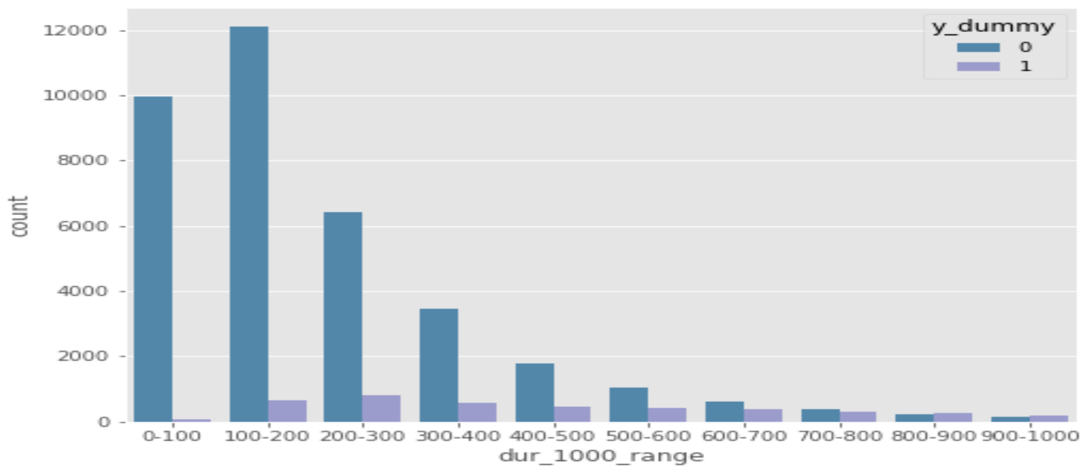
Interesting more than 90% of the calls are in the 0-10 minutes range-



Further investigation reveals that it is highly likely that the call will last between 100 and 200 seconds.

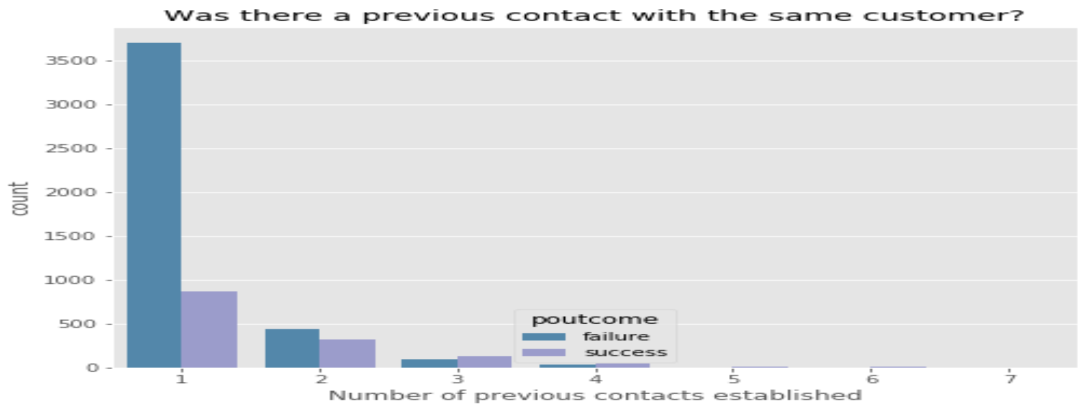


However, if the call lasts more than 700 seconds the likelihood that there will be a positive outcome is almost 100%



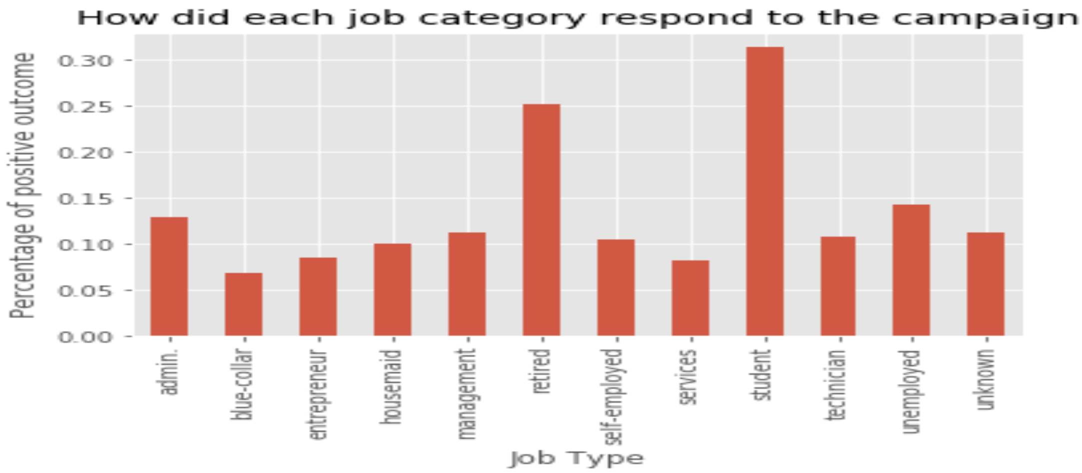
<u>Previous Contacts:</u>

If more than one previous contact were established with the customer, chance of a positive outcome is very high.



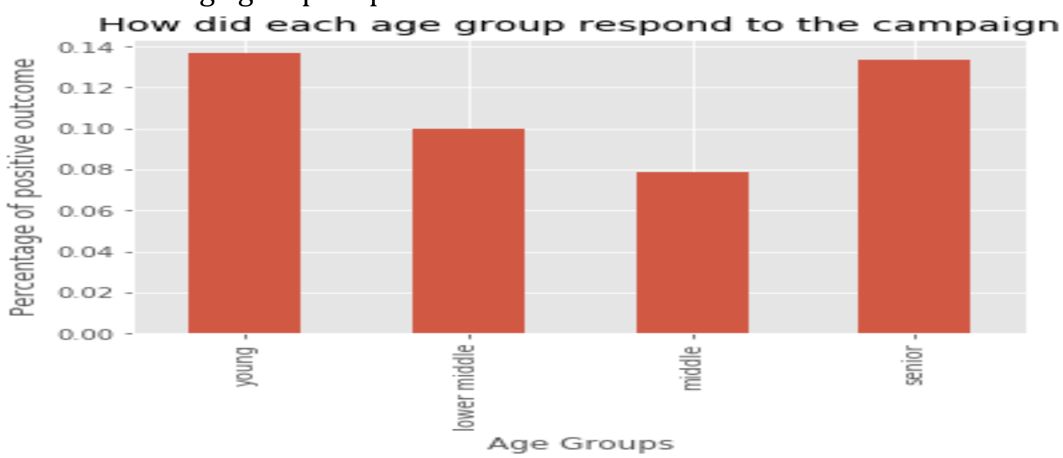
Data Visulaization:

How did each job category respond to the campaign:

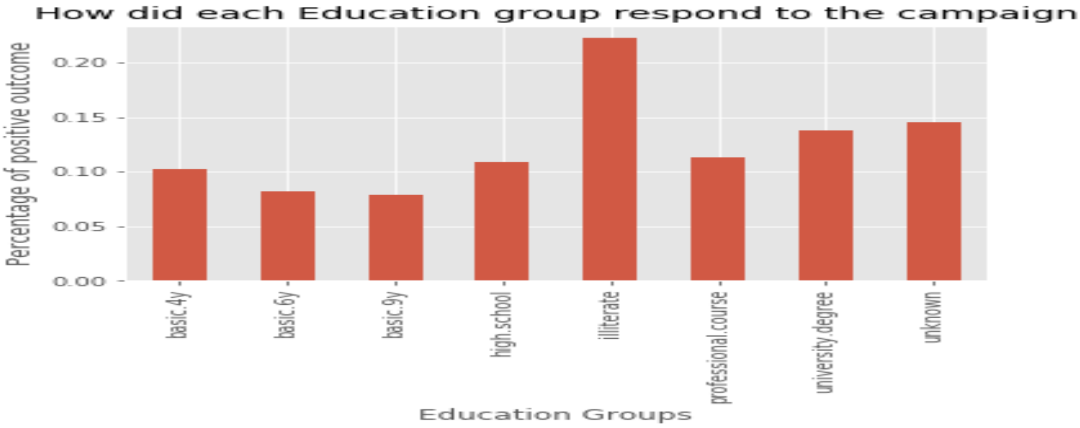


Students and retired people responded more positively than other groups, followed by unemployed and admin. But, the proportion of students and unemployed in the dataset is very small. So, we can infer that retired and admin job categories respond better.

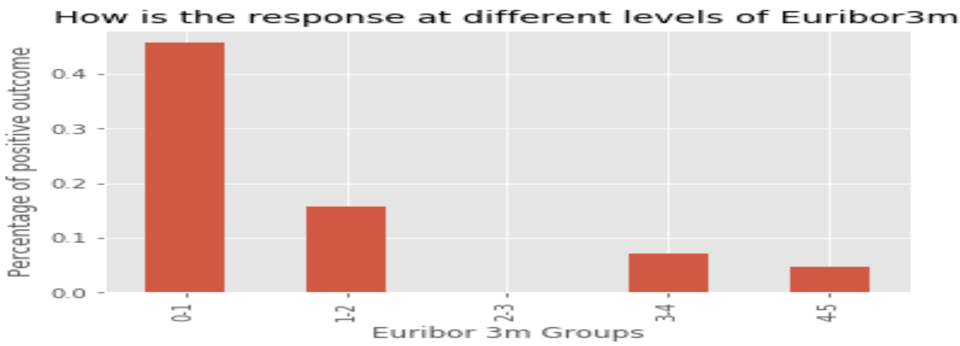
How did each age group respond?



Youg people and seniors responded better. So, one pattern is gradually emerging out of age and job category analysis. Young/Student and Senior/Retired respond better.

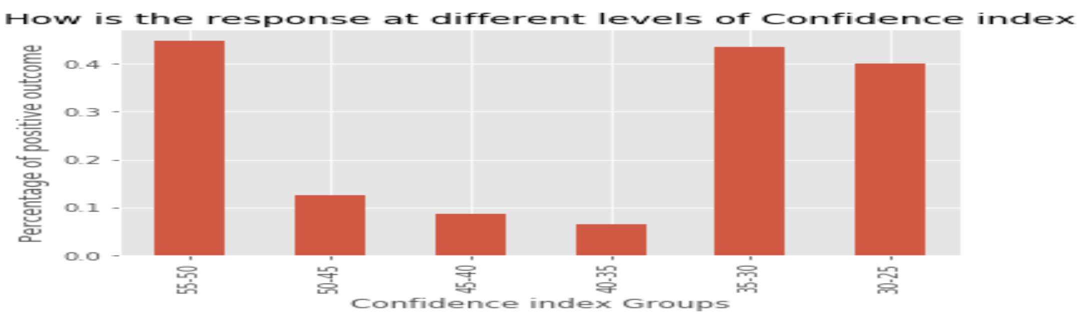


University degree, Professional course and High School – are the three educational groups who responded better. The proportion of illeterates is very low in the dataset.

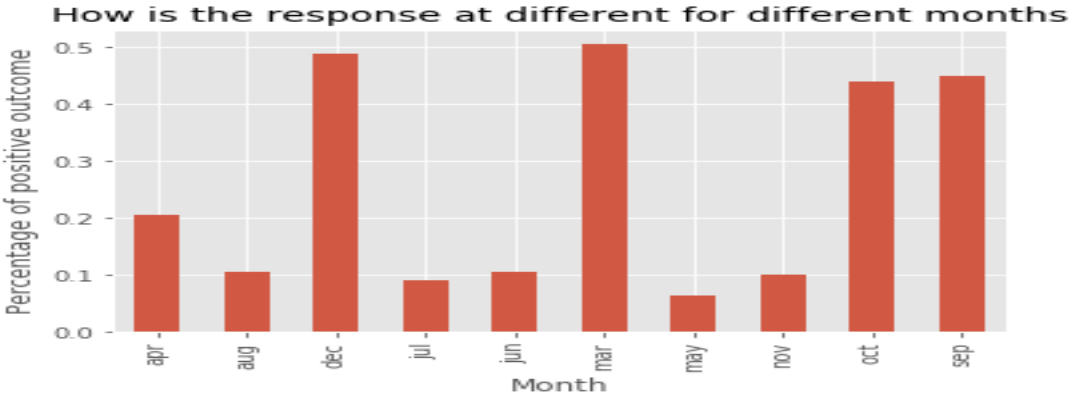


Response is the best when Euribor-3m is between 0 and 1.

We could not find any specific pattern for consumer price index.



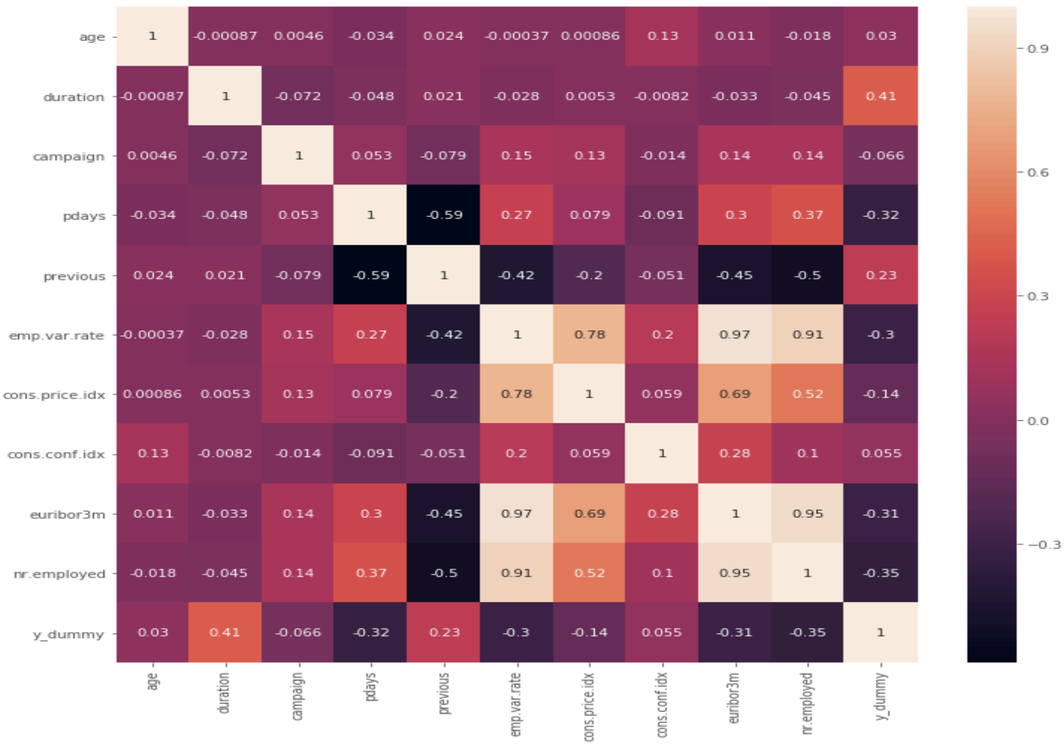
If confidence index is low (25 to 35) or high (above 50), the response rate is better. Response rate is low in the middle range(35 to 50) of confidence index.



While there is no specific pattern for the month, March, December, Sepetember and October show better response rates. May be quarter ends have some impact.

Inferential Statistics:

Correlation matrix



Positive correlation:
Duration of the call has highest positive correlation followed by number of previous contacts.
Negative correlation:
Nr_employed, pdays(number of days that passed by after the client was last contacted from a previous campaign), euribor-3M, emp-var-rate.

Next I checked the categorical variables

a. I conducted chi-square test for each categorical variable.

i. Education:

```
Chi2 value for education, p-value, Expected counts (193.10590454
149565, 3.3051890144025054e-38, 7, array([[  3.70555618e+03,   2
.03379664e+03,   5.36400554e+03,          8.44309556e+03,   1.5
9722249e+01,   4.65235418e+03,          1.07972240e+04,   1.535
99563e+03],        [  4.70443819e+02,   2.58203360e+02,   6.8099
4464e+02,          1.07190444e+03,   2.02777508e+00,   5.906458
19e+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

ii. Job

```
Chi2 value for job, p-value, Expected counts (961.24244032895535
, 4.1897632875638613e-199, 11, array([[ 9247.91822861,  8211.498
30048,  1291.97552685,   940.58657862,         2594.59920365,
1526.23482568,  1260.91842284,  3521.87559483,           776.427
60027,  5983.37292415,   899.76867049,   292.82412353],        [
 1174.08177139,  1042.50169952,   164.02447315,   119.41342138,
         329.40079635,   193.76517432,   160.08157716,   447.12
440517,           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.

iii. Day of week

```
Chi2 value for day of week, p-value, Expected counts (26.1449390
75871971, 2.9584820052785324e-05, 4, array([[ 6945.25580266,  75
54.8623871 ,  7651.58308245,  7178.62775566,          7217.67097
213],        [  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.

iv. Month

```
Chi2 value for month, p-value, Expected counts (3101.14935141167
73, 0.0, 9, array([[  2335.49422162,   5482.02253083,    161.496
94086,   6365.8189764 ,          4718.905118  ,    484.49082257
,  12217.86471788,   3639.00524425,          637.11430514,
505.78712246],        [   296.50577838,    695.97746917,     20.
50305914,    808.1810236 ,          599.094882  ,     61.50917
743,   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.

v. Marital status

```
Chi2 value for marital, p-value, Expected counts (122.6551518225
2989, 2.0680146484422109e-26, 3, array([[  4.09243896e+03,   2.2
1197568e+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.

The final part of this exercise is predictive modelling. I used several machine learning techniques as well as deep learning to build a classifier to predict whether a customer will subscribe to a term deposit or not. Several steps were executed to reach that goal:

<u>Step1:</u>
Preprocessing: Label Encoder was used for preprocessing of the input variables – job, education, housing, loan, default, poucome, marital.

<u>Step2:</u>
Train-Test-Split:
We split the data into training and test data with 30% of data preserved for testing the models with test data.

<u>Step3:</u>
Apply Logistic Regression
The results are:

```
Training score:  0.908570635774 Testing score:  0.91130533301

[[10705   263]
[  833   556]]

               precision    recall  f1-score   support
0              0.93          0.98      0.95     10968
1              0.68          0.40      0.50      1389
avg / total   0.90           0.91      0.90     12357
```

This is a model having reasonable amount of accuracy and good F1-score. However, there is class imbalance in the dataset with only 11.26% people positively responded with term deposit subscription.

<u>Step4:</u>
We applied class imbalance treatment with class_weight = 'balanced', and applied logistic regression.

The results are

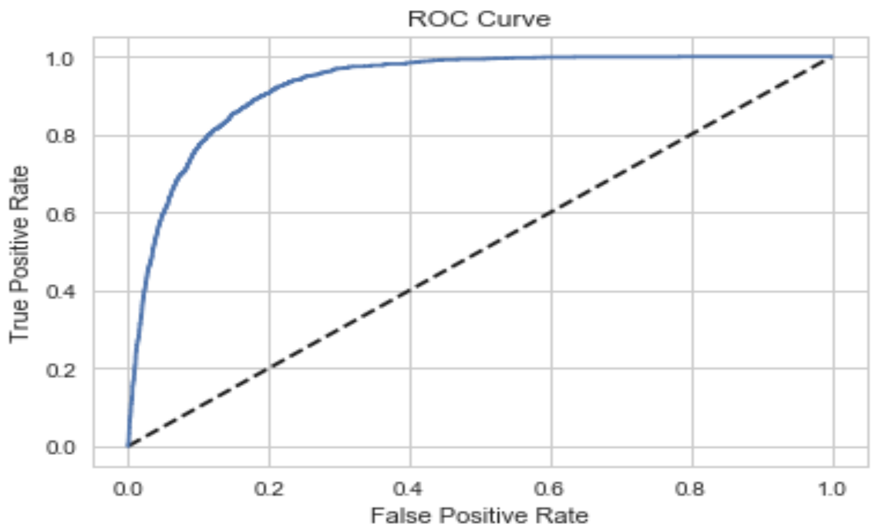
```
Training score:  0.851687419791 Testing score:  0.850206360767
[[9303 1665]
[ 186 1203]]
               precision    recall  f1-score   support            0
       0.98        0.85      0.91      10968
1              0.42       0.87      0.57      1389
avg / total     0.92       0.85      0.87     12357
```

Precision improved; but f1-score slightly declined.

Step 5:
We tried to improve the model with hyper parameter tuning with C value ranging 0.001 to 1000. We used GridSearchCV of Python and obtained an optimum value of C. We got best C value of 1000 and best accuracy score of 0.908.

Step 6:
In this step, we instantiate the best tuned model and find out whether the model has improved or not, by comparing the relevant outcome parameters. We see significant improvement in all aspects. Accuracy score is 0.9115. The ROC curve has improved a lot.

```
[[10702   266]
[  828   561]]
               precision    recall  f1-score   support
0              0.93          0.98      0.95     10968
1              0.68          0.40      0.51      1389
avg / total    0.90          0.91      0.90     12357
```




```
Area under the curve(AUC) is 0.928672073849
```

Step 7:
We started to check whether other types classification models can give better results than Logistic Regression. The first model we tried is Support Vector machine. We found couple of shortcomings of this model –

a. The model ran slower than Logistic Regression classifier
b. The accuracy score was only 0.89

Step 8:
We applied the Random Forest Classifier. This model ran real fast.
a. Accuracy score without class imbalance treatment is 0.9078
b. Accuracy score with class imbalance treatment is 0.9079
c. We also obtained the feature importance matrix from the Random Forest Classifier –

Input Parameter	Feature Importance
duration	0.415486
euribor3m	0.115036
emp.var.rate	0.080496
nr.employed	0.070722
age	0.067059
pdays	0.035494
campaign	0.032359
education	0.031553
job	0.030863
cons.price.idx	0.028275
cons.conf.idx	0.025789
marital	0.017451
default	0.012865
housing	0.012775
loan	0.009214
previous	0.007743
poutcome	0.00682

Step 9:
At this point we decided to explore all possible classification models to see if there are other models that can give better results. We looked into Linear Discriminant Analysis(LDA), K- Nearest-Neighbor(KNN), Decision Tree Classifier (CART)and Gaussian Naïve Bayes(NB). Following are the results:

```
Accuracy scores in different models
LDA: 0.906282
K-NN: 0.902778
CART: 0.887552
NB: 0.844022
```

Step 10:
Now it is time to apply Artificial neural network. We implemented a very basic Keras Tensor flow Neural Network model with one input layer and one output layer. The input layer has 10 neurons and output layer has 1 neuron. We are operating with 17 dimensions.

```
Epoch 1/10 28831/28831 [==============================] - 1s 27us/step
 - loss: 0.3970 - acc: 0.8872
Epoch 2/10 28831/28831 [==============================] - 0s 14us/step
 - loss: 0.3520 - acc: 0.8872
Epoch 3/10 28831/28831 [==============================] - 0s 13us/step
 - loss: 0.3502 - acc: 0.8872
Epoch 4/10 28831/28831 [==============================] - 0s 12us/step
 - loss: 0.3304 - acc: 0.8872
Epoch 5/10 28831/28831 [==============================] - 0s 13us/step
 - loss: 0.3173 - acc: 0.8872
```

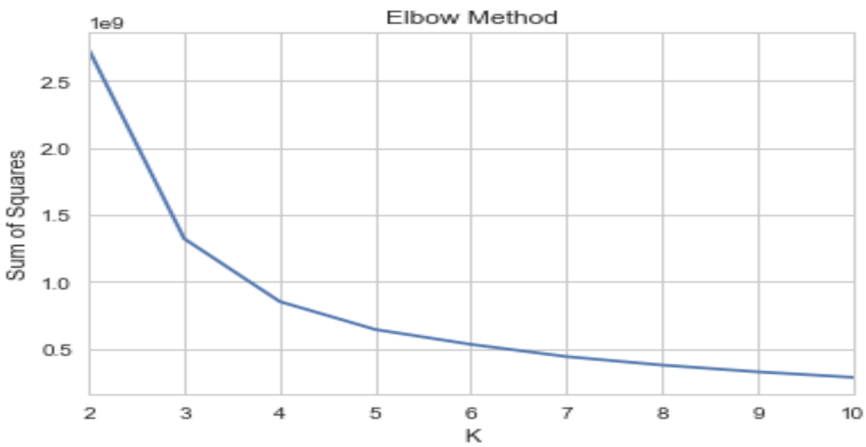
```
Epoch 6/10 28831/28831 [==============================] - 0s 13us/step
 - loss: 0.3158 - acc: 0.8872
Epoch 7/10 28831/28831 [==============================] - 0s 12us/step
 - loss: 0.3116 - acc: 0.8872
Epoch 8/10 28831/28831 [==============================] - 0s 13us/step
 - loss: 0.3086 - acc: 0.8872
Epoch 9/10 28831/28831 [==============================] - 0s 14us/step
 - loss: 0.3087 - acc: 0.8872
Epoch 10/10 28831/28831 [==============================] - 0s 13us/ste
p - loss: 0.3057 - acc: 0.8872
```

The Accuracy is: 88.76%

Step 11:

Clustering and segmentation of customers is almost an essential precursor of any marketing campaign effort. We used K-means clustering which is an unsupervised learning algorithm. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

We first used elbow method to find optimum number of clusters:

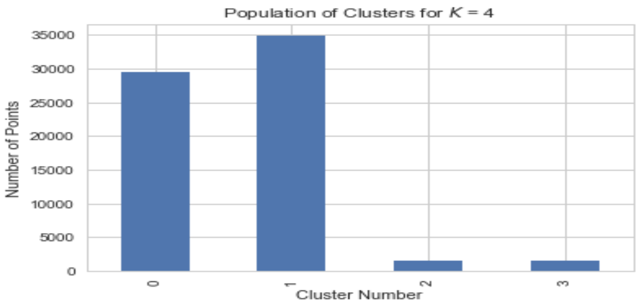


We could use either 3 or 4 clusters.

a. 3 Clusters:
```
0     4820
1    34859
2     1509
```

b. 4 clusters:
```
0    29406
1     8698
2     1582
3     1502
```

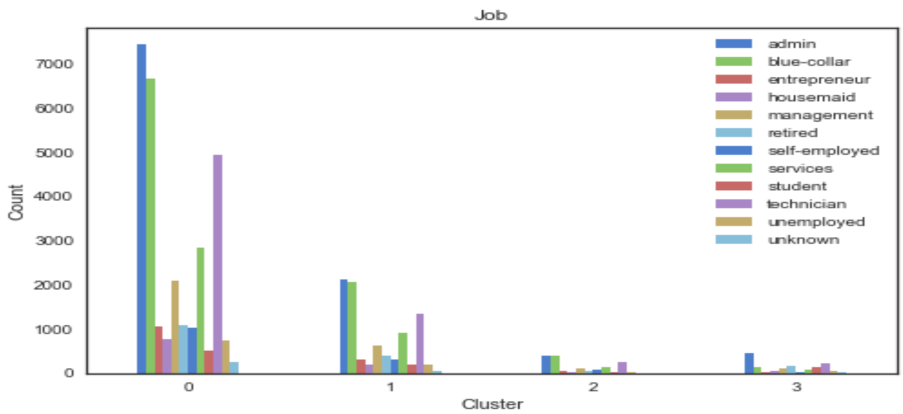
c. We decided t go with 4 clusters



Step 12:

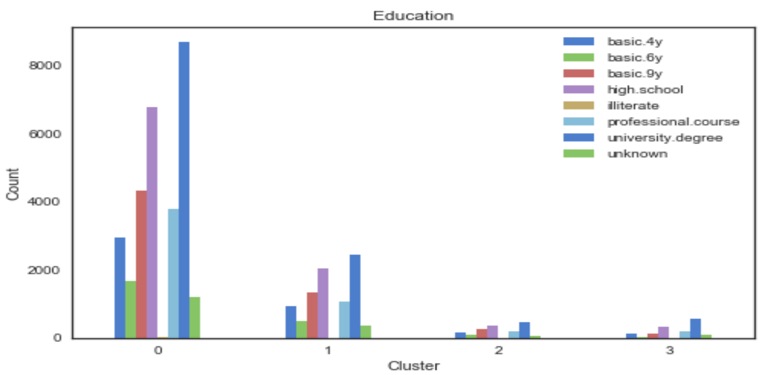
We started to check the characteristics of the clusters. How are they similar and how they are dissimilar. We looked at the input parameters like: job, marital, and duration.

a. How clusters are different in terms of job category



Cluster 1 and Cluster2 have relatively higher proportion of blue color job category

b. How clusters are different in terms of Education



No Siginificant difference observed.

c. How clusters are different in terms of Duration

cluster	count	mean	std	min	25%	50%	75%	max
0	29406.0	142.756342	75.783721	0.0	83.00	135.0	199.00	309.0
1	8698.0	473.838009	135.230907	304.0	360.00	436.0	563.00	822.0
2	1582.0	1172.082807	398.335190	823.0	917.25	1061.5	1285.75	4918.0
3	1502.0	309.368842	213.623888	1.0	164.00	252.0	386.75	1260.0

The 'mean' column clearly reveals that average call duration of cluster 2 is much higher that cluster 0. Cluster 1 and Cluster 3 have moderate amounts of duration.

So, one of the major identifying criteria of a cluster is the duration, which actually makes sense. The customers of custer2 take longer to get convinced. The chance of the customer agreeing to subscribe to the term deposit is also higher for cluster2.

Cluster	Term Deposit success rate
0	0.035
1	0.2009
2	0.5702
3	0.6378

We can see that success rate of cluster2 is way higher than cluster0 and cluster1. Cluster 3 actually, has the highest success rate.

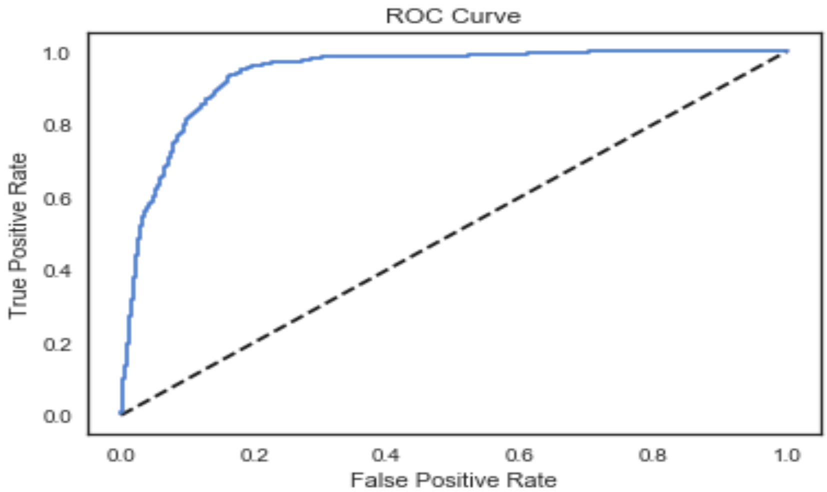
Step 13:

We believe it would be good to apply logistic regression classification to each of the clusters. Cluster 0 is a highly unbalanced so far as target variable is concerned. Hence, we applied logistic regression with class imbalance treatment. The results are fantastic.

```
LogisticRegression(C=1.0, class_weight='balanced', dual=False,
fit_intercept=True, intercept_scaling=1, max_iter=100,          multi_cl
ass='ovr', n_jobs=1, penalty='l2', random_state=None,          solver='l
iblinear', tol=0.0001, verbose=0, warm_start=False)
Training score:  0.931208705791
Testing score:  0.931874858309
[[8018  484]
[ 117  203]]

              precision    recall  f1-score   support
```

0	0.99	0.94	0.96	8502
1	0.30	0.63	0.40	320
avg / total	0.96	0.93	0.94	8822




```
Area under the curve:  0.938198732651
```

Conclusion:

While marketing campaign is a complex area, we have simplified it to a great extent. Through EDA and data visualization, We have figured out how to give high level guidelines for the campaign to make it more effective. We can choose the right kind of demographic to improve the success rate of the campaign. For example, success rate for students and retired people are higher; and success rate goes down as too many days pass from the last campaign. Whereas, success rate is higher if there were more number of previous contacts established with the customer. If the customer doesn't have any housing loan, personal loan and any previous credit default, the chance of success is higher.

Here are the guidelines for executing bank marketing campaign in future – We should get the complete dataset containing customer information and apply our clustering algorithm. For each cluster, we should apply our Logistic Regression classification classification algorithm and do train test split. For each cluster we should get the probability of success for each customer. We should plan decide the script and duration of each call based on those success probabilities.