Effectiveness of Marketing Campaigns

Dipanjan Bandy

6th July 2018

Topics

- Marketing Campaigns
- Project Dataset
- Initial impression of data
- Data Story, Data Visualization
- Inferential Statistics
- Predictive modeling
  - Supervised learning
  - Clustering
  - Deep learning
- Conclusion

Marketing Campaigns

- A promotion to reach certain sales goals within a specific time period
- More targeted, the better
- Need to have good idea of customer segments
- Product – market fit should be good
- Organizational resources need to be used efficiently
- Predictive models can improve effectiveness of the campaigns
- Can be delivered through several channels – email, sms, phone call, social media, word of mouth, in-store signage

Project Dataset

- 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.

Project Dataset

- 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')

Project Dataset – Initial impression

- Initial impression about data
- 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

Project Dataset – Those who subscribed to the term deposit

Exploration of data

- 54% of those who subscribed were married
- 90% of those who subscribed didn't have any credit default
- 83% of those who subscribed don't have a personal loan
- 33% of those who subscribed are young and 17% are seniors.
- Most of those were never contacted before
- The call duration for 64% of them were 0-10 minutes whereas for 30% of them it was 10 -20 minutes.

Data Story: Composite Groups

The overall average positive response rate for the entire dataset is 11.26%.

- We created a composite group containing marital status and age ranges. Seven groups had better than average response –
- Divorced - Senior
- Married - Senior
- Single - Young
- Single - Lower middle
- Unknown - Young
- Unknown - Middle
- Unknown - Senior

Data Story: Composite Groups

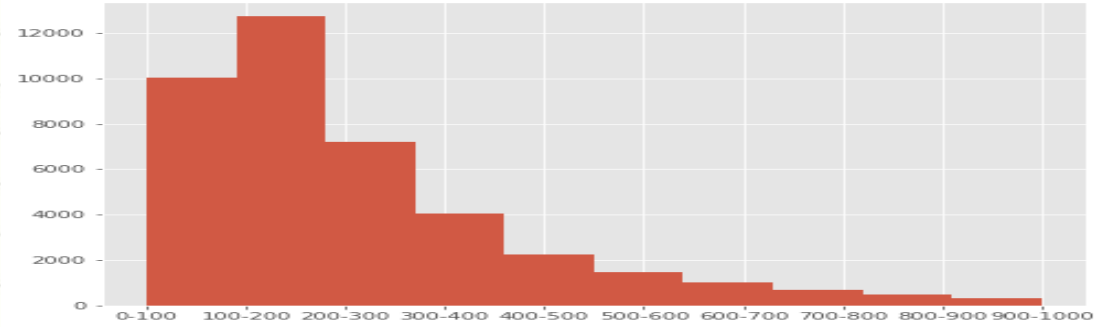
The overall average positive response rate for the entire dataset is 11.26%.

- 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 –

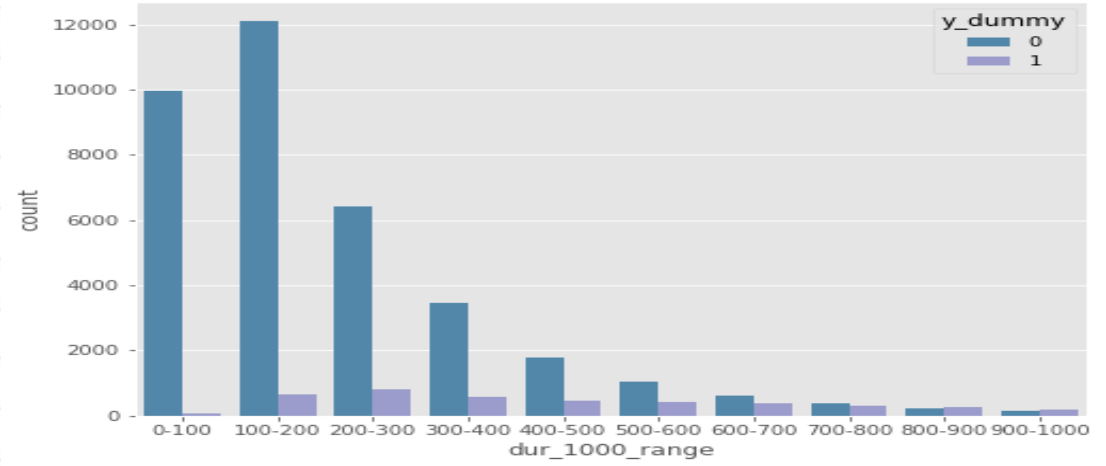
- Housing loan – yes, Personal loan –no, Default – no
- Housing loan – yes, Personal loan –yes, Default – no
- Housing loan – no, Personal loan –no, Default – no (Very much expected)
- Housing loan – no, Personal loan –yes, Default – no

Data Story: Call Duration

It is highly likely that the call will last between 100 and 200 seconds.

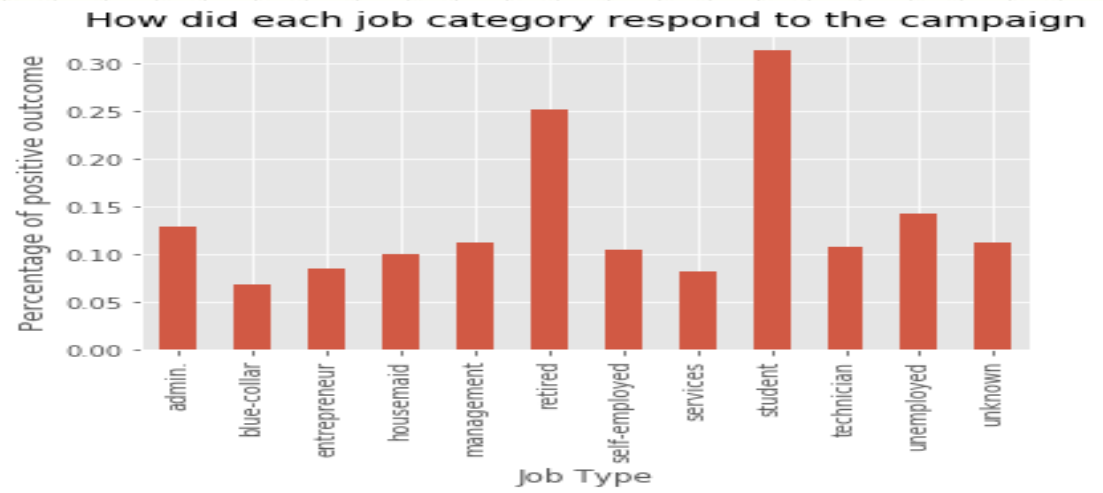


But, if the call lasts more than 700 seconds chance of a positive outcome is almost 100%

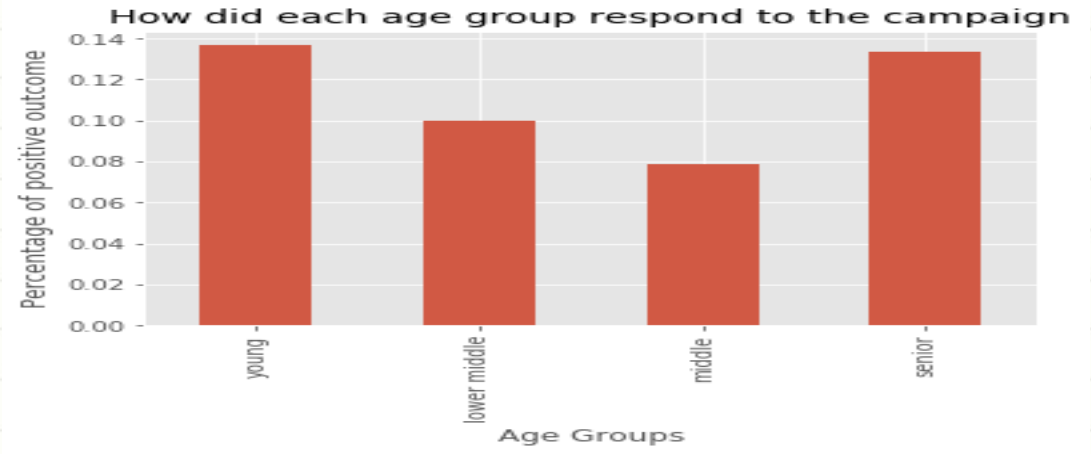


Data Story: Job Category and age groups

Students and retired people responded more positively than other job categories

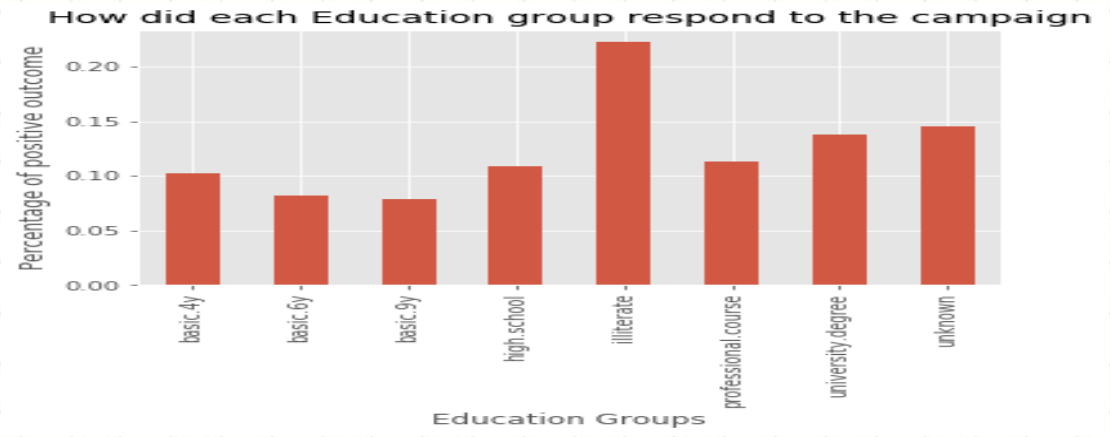


Young people and seniors responded better.

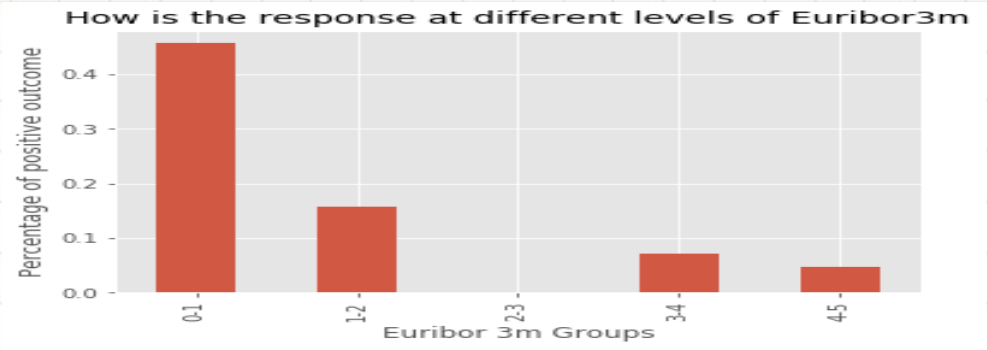


Data Story: Education, Euribor3m

University degree, Professional course and High School – are the three educational groups who responded better.

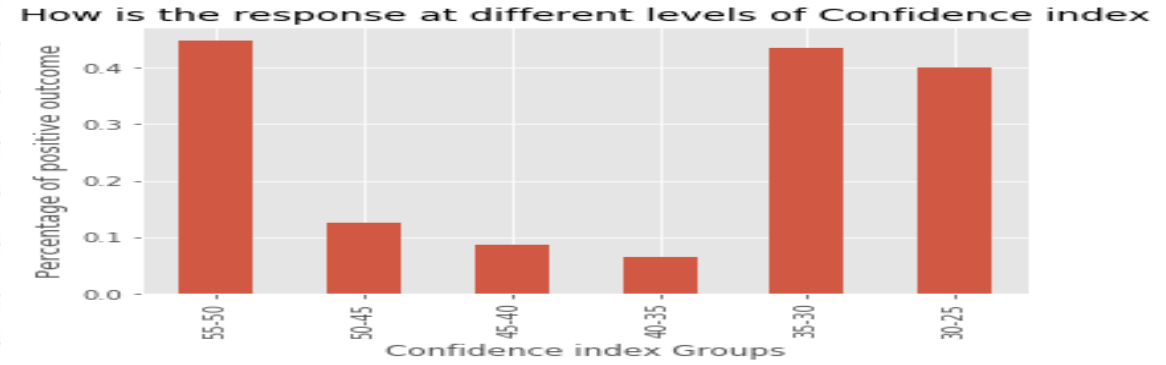


Response is the best when Euribor-3m is between 0 and 1.

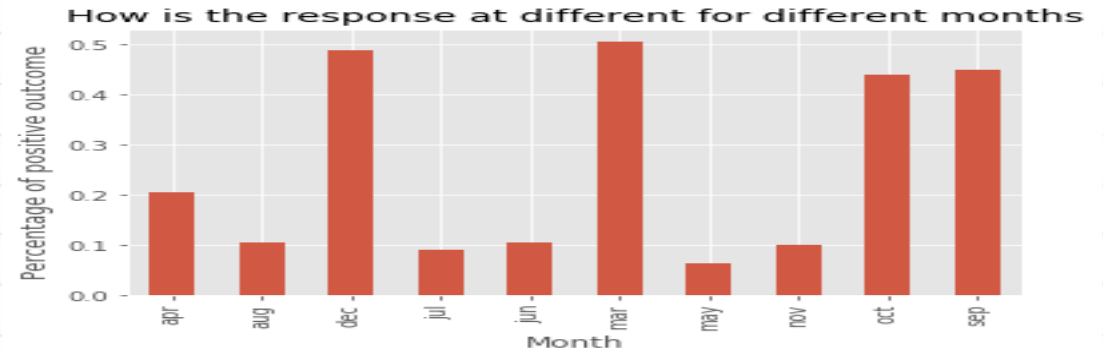


Data Story: Confidence index, Month

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.

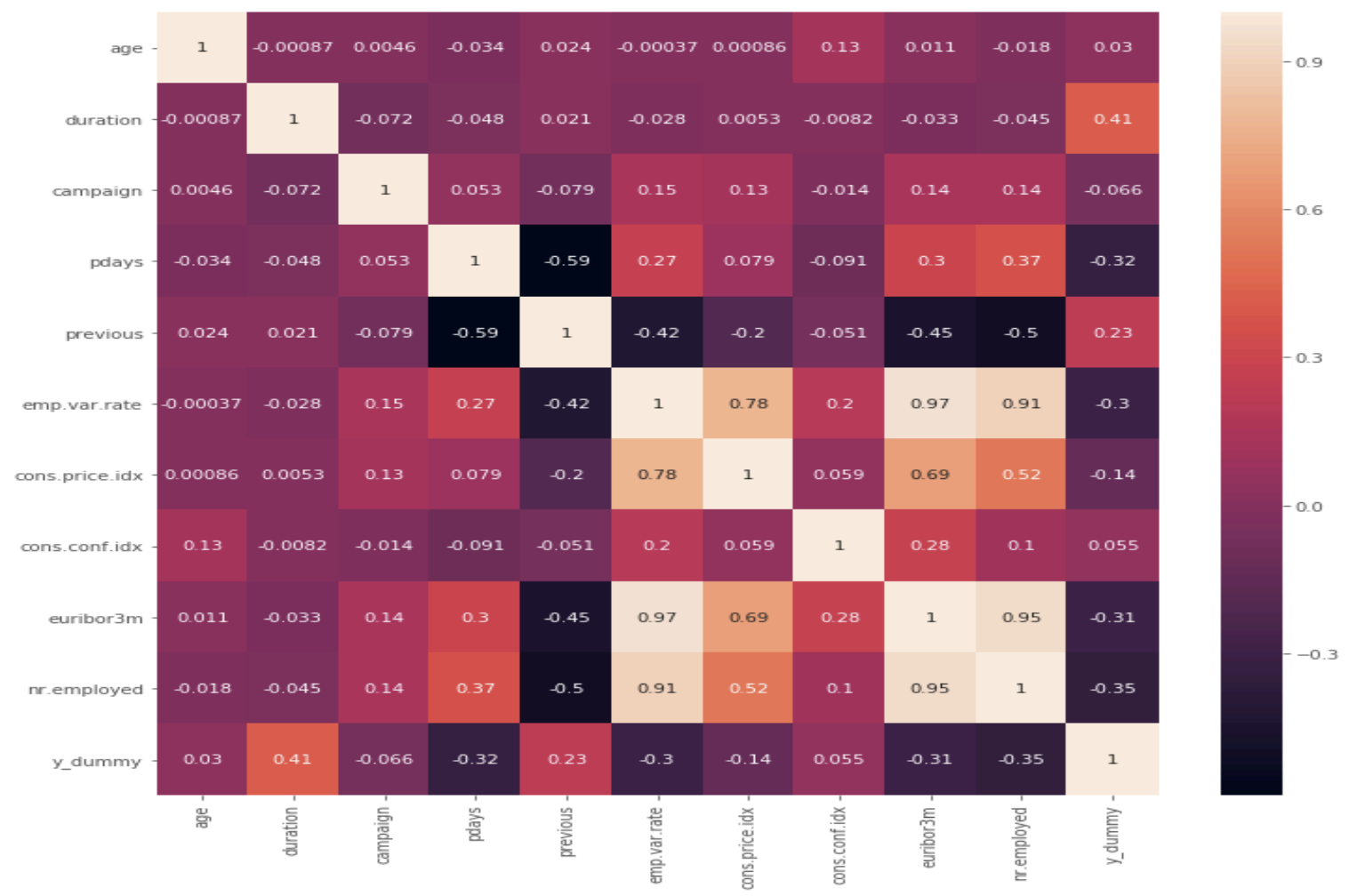


While there is no specific pattern for the month, March, December, Sepetember and October show better response rates. May be quarter ends have some impact.



Inferential Statistics: Correlations

How are the numerical input variables related?



Inferential Statistics: Correlations

How are the numerical input variables related to the target variable (Subscribed or not)?

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.

Predictive Modeling

Label Encoding, Train-test-split, Logistic Regression

- Preprocessing: Label Encoder was used for preprocessing of the input variables – job, education, housing, loan, default, poucome, marital.
- Train-Test-Split: We split the data into training and test data with 30% of data preserved for testing the models with test data.
- Apply Logistic Regression

  The results are:

  Training score: 0.908570635774

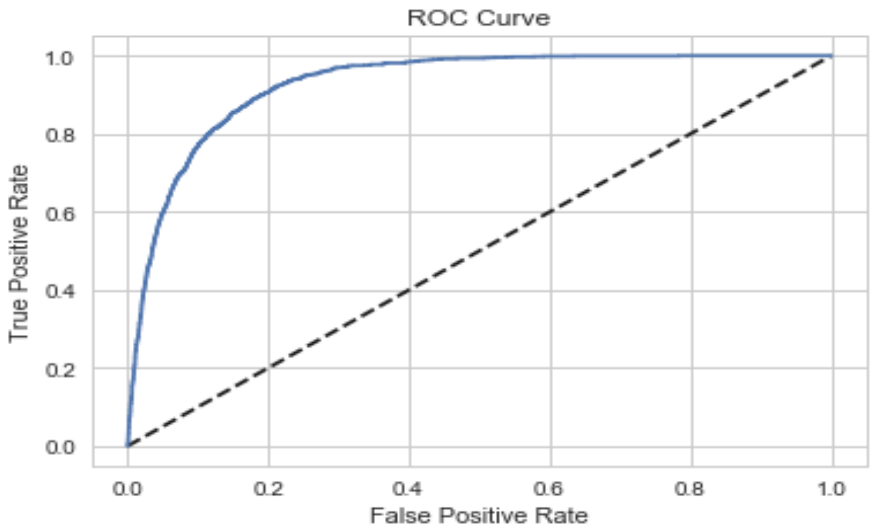
  Testing score: 0.91130533301

  F1-Score : 0.90

Predictive Modeling : Best tuned model

Applied GridSearchCV of sklearn

We instantiated 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.



Area under the curve(AUC) is 0.928672073849

F1-Score : 0.90

Predictive Modeling :Other classification models

Random Forest Classifier

- Accuracy score without class imbalance treatment is 0.9078
- Accuracy score with class imbalance treatment is 0.9079
- We also obtained the feature importance matrix from the

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

Predictive Modeling
Other Classification models

- Support Vector machine(SVM): Accuracy score is 0.89
- Linear Discriminant Analysis (LDA) Accuracy score is 0.91
- K-Nearest Neighbor(KNN) Accuracy score is 0.90
- Decision Tree Classifier(CART) Accuracy score is 0.89
- Gaussian Naïve Bayes(NB) Accuracy score is 0.84
- All the above models, other than SVM ran fast

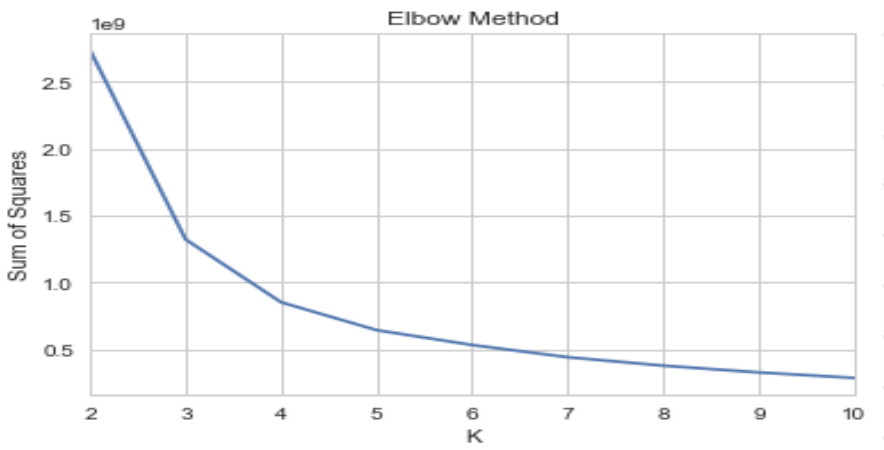
Predictive Modeling
Neural Network

Artificial neural network was applied. 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.

- The Accuracy is: 88.76%

Unsupervised learning
Segmentation and clustering

Clustering and segmentation of customers is almost an essential precursor of any marketing campaign effort. Elbow method was used to find out the optimum number of clusters.



- We decidedto use 4 clusters
- Applied K-means clsutering
- Call duration was found to be one differentiating factor among clusters

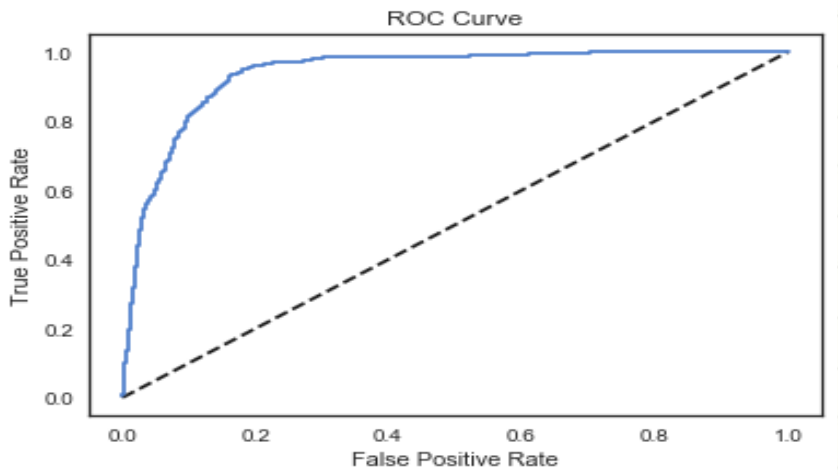
Apply Logistic Regression on clusters

Logistic regression was applied to the largest one – Claster 0

- Training score: 0.931208705791
- Testing score: 0.931874858309
- Classification matrix

[[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(AUC): 0.94

Bank Marketing Campaign Conclusion

- We simplied the the complex marketing campaign area with a methodical approach
- Data analysis and inferential statistics helps, for example
  - Success rate improves for seniors and students
  - Success rate goes down if too many days pass from last campaign
  - Success rate is higher if there were more previous contacts or call duration exceeds 700 seconds
- A combination of Random Forest and Logistic regression is the best apprach with Random Forest classifier providing a confirmation on significant input parmeters
- Clustering the entire dataset and subsequently applying Logistic Regression to each clsuter provided the best result.

HAPPY CAMPAIGNING !!!!