REPORT ON TERM DEPOSIT SUBSCRIPTION PREDICTION OF BANK

The data set (bank-additional-full.csv) consists of demographics data on 41,188 people.

The dataset has 20 input variables (mix of numeric and categorical variables) and 1 predictor variable (whether they responded "yes" or "no") to the marketing campaign. The target variable is y (column # 21) - whether the clients have responded 'yes' or 'no' to term deposit.

Snapshot of business case:

Assumptions

Cost of mailing to each person is $ 2

Revenue per person who responds positively to marketing campaign is $ 10

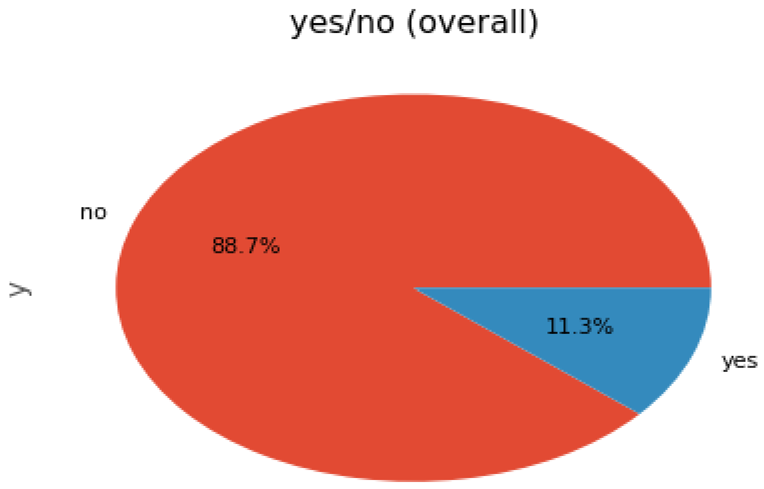
Average response rate of people on list is 11 %, marketing to every person on the test set incurs a loss of $ 7,000 with a ROI of negative 44 %.

The best machine learning model identifies the subset of people with high probability of conversion of around 36 % (3.6 X the average) by rank scoring the people on list (creating a targetted list) and provides a positive ROI of 79 %.

Average response rate (overall):

Overall distribution of campaign response in the original dataset:

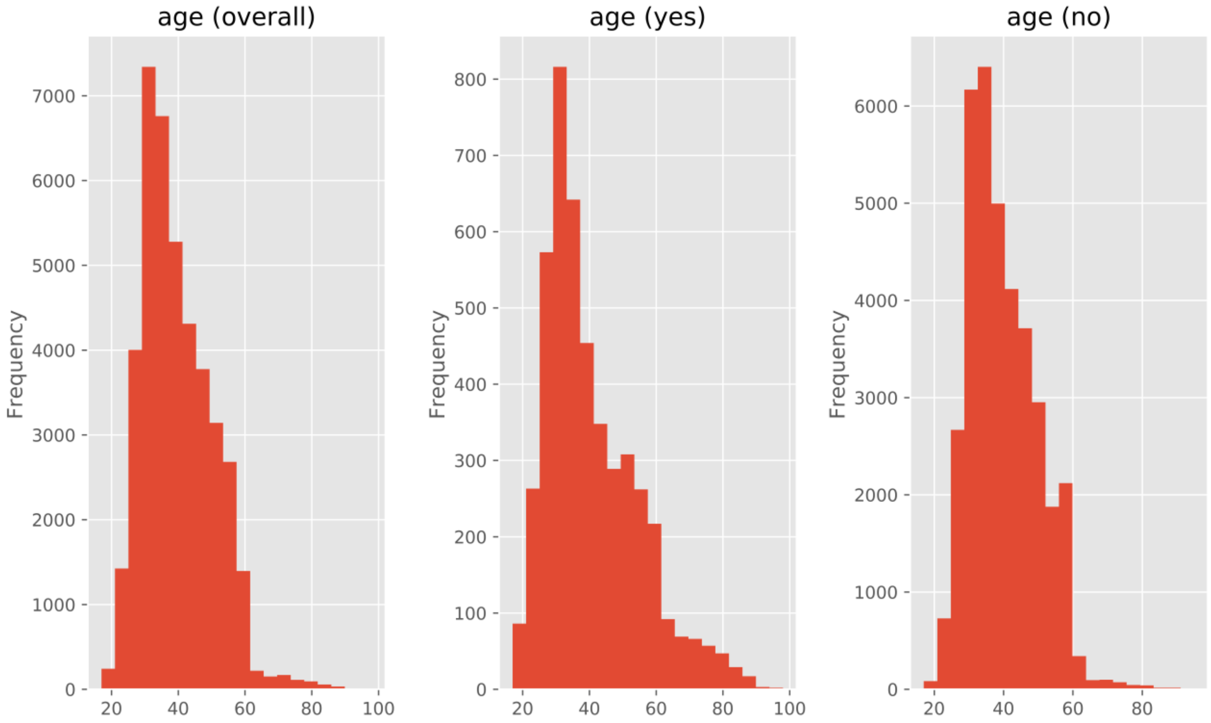
Overall response:



Data Exploration of variables in dataset:

Age (Numerical Variable):

age_distribution:



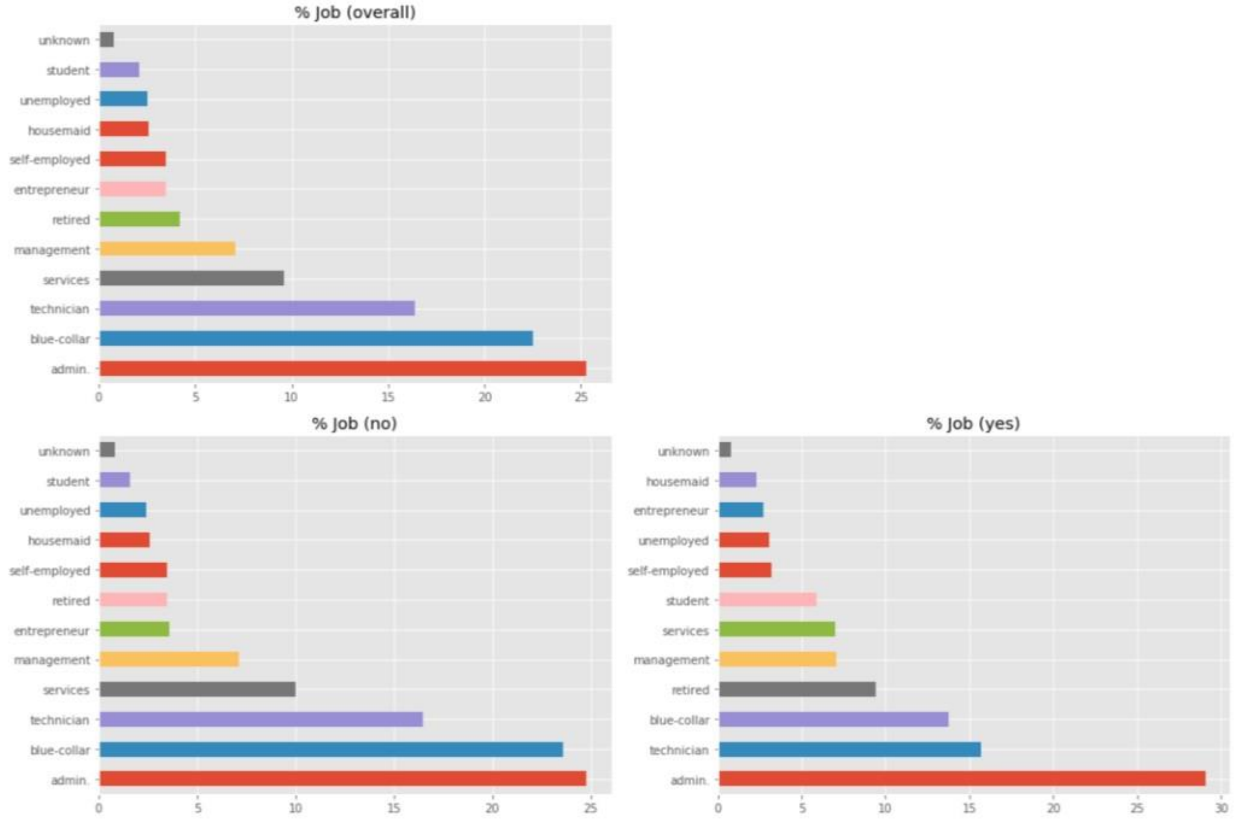
Statistical test on Age:

Average of mean age of people who responded "Yes" vs those who responded "No"

The 95% confidence interval of the difference of means is [0.591 , 1.413]. Hence, the difference of the mean of age is statistically significant at 95% confidence interval because the 95% confidence interval does not include 0.

Job (Categorical variable):

job 1



Statistical test on Job Type:

Chi square test of Job type between 2 groups of people who responed "Yes" vs those who responded "No"

p value of chi-square test: 4.232405679993378e-200

The chi square test shows that job is statistically significant at 95% confidence interval is whether people responded yes or no to the marketing campaign as the p value of the chi square is less than 5 %. From the chart shown above for the proportion of each job type in each groups, for people with the following job types had a higher % of responding "yes" than those who responded "no",

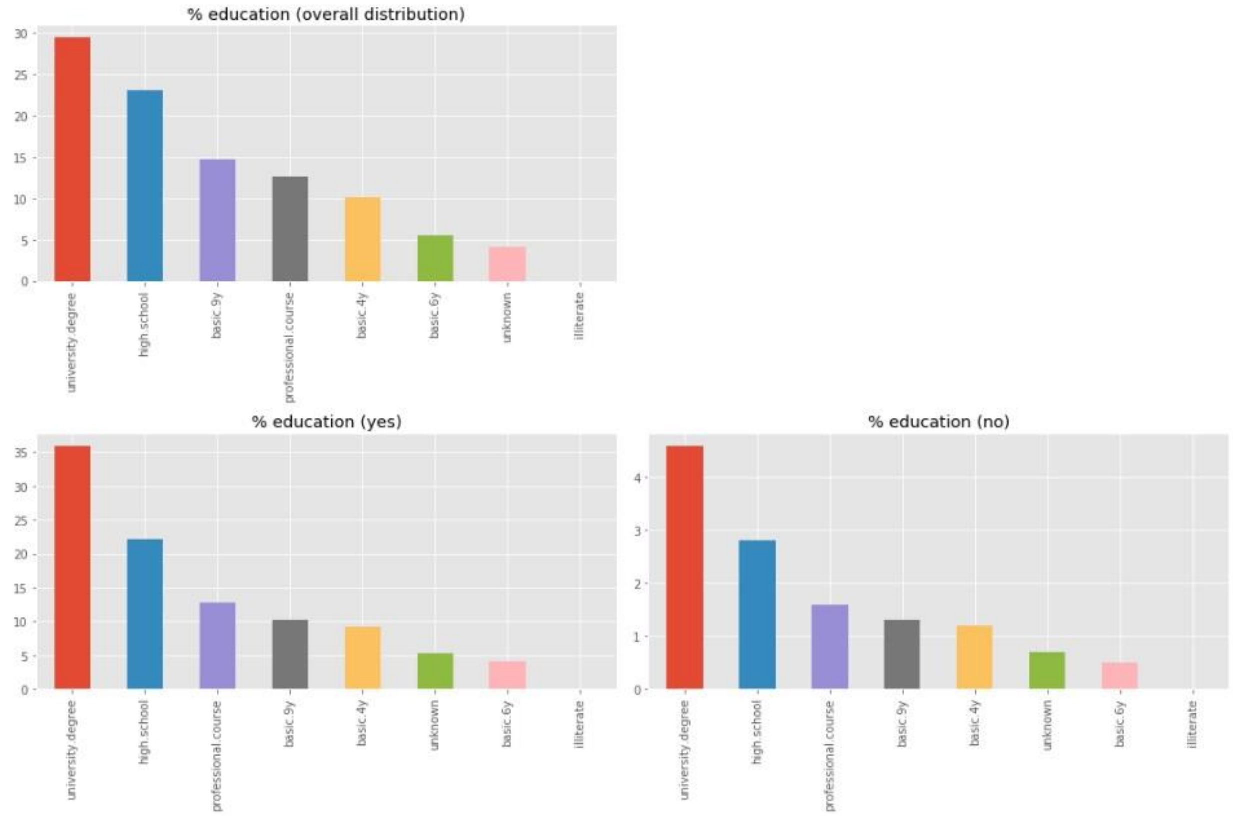
Admin jobs

Retired

Students

Education Level (Categorical variable):

education 1



Statistical test on Education Level:

Chi square test of Education level between 2 groups of people who responed "Yes" vs those who responded "No"

p value of chi-square test: 2.2494049169426562e-35

Analysis:

The chi square test reveals that education level is statistically significant at 95% confidence interval as p value is less than 5 %. From the chart above, those who responded "yes" had a much higher proportion of people with higher education (university degree, high school & professional course) than those who responded "no".

Analysis of other variables in dataset

The same process (visualization and statistical testing) has been repeated for all the other variables in the dataset.

Machine learning models applied on dataset:

Logistic regression model (with L1 regularization)

Naive bayes classifier model

Random forest classifier model

KNN classifier model

Gradient boosted tree model

Decision tree classifier model

SVM classifier model

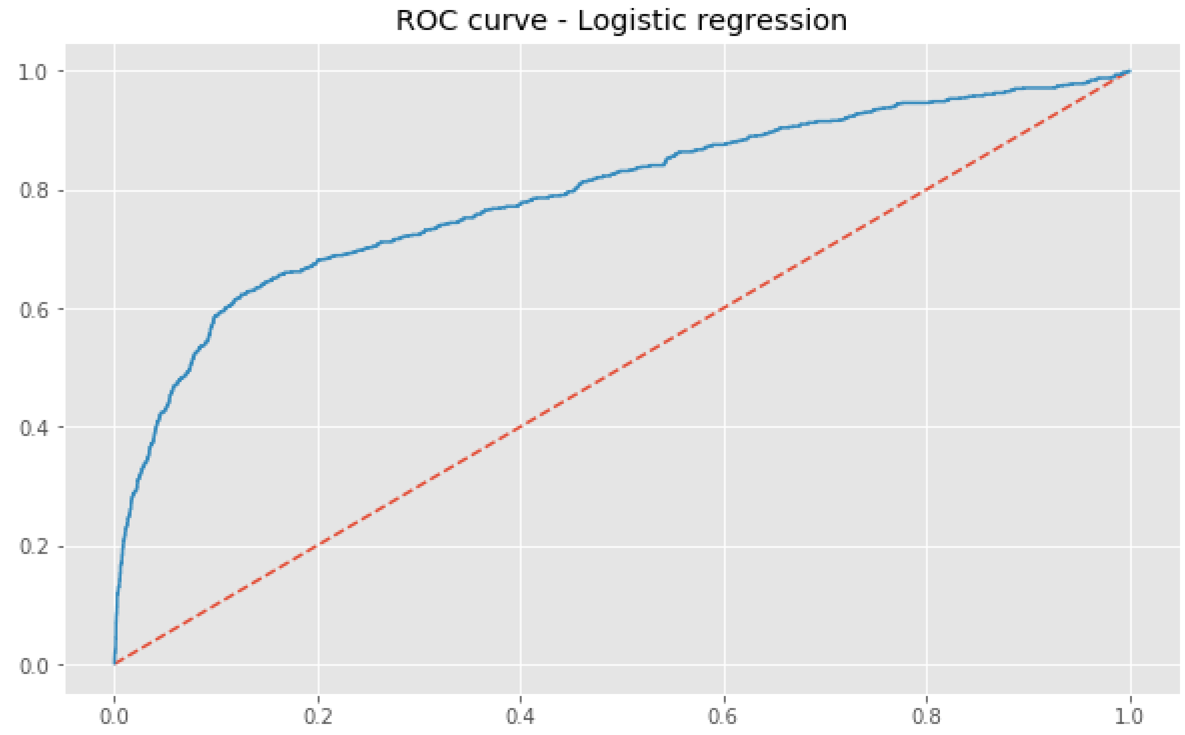
Principal component analysis:

PCA was applied to achieve variable reduction and save computation time for KNN, gradient boosted tree and SVM classification algorithm. Using PCA, the total varaibles were reduced to 34 principal components that explained 90 % of variation in original dataset of 426 variables.

Logistic Regression with L1 regularization:

Applying the L1 regularization reduced the input variables from 432 variables in original dataset to 234 variables with non-zero coefficient due to shrinkage but accuracy metric (test set AUC) was same for both the logistic regression models with and without regularization.

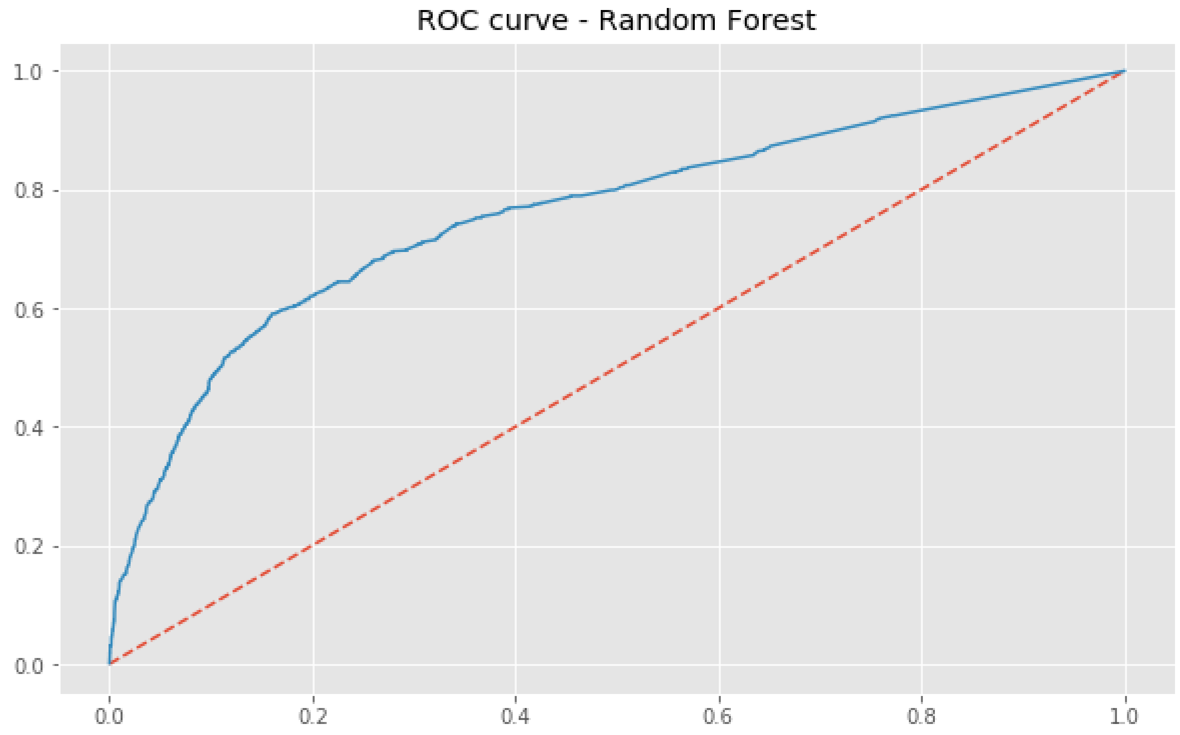
auc_logistic__l1_reg



AUC score for logistic regression with L1 regularization: 79 %

Random Forest:

auc_random_forest



AUC score for random forest: 78 %

Top 10 most importamt variables based on variable importance in random forest model:

age

poutcome_success

housing_yes

housing_no

previous

job_admin.

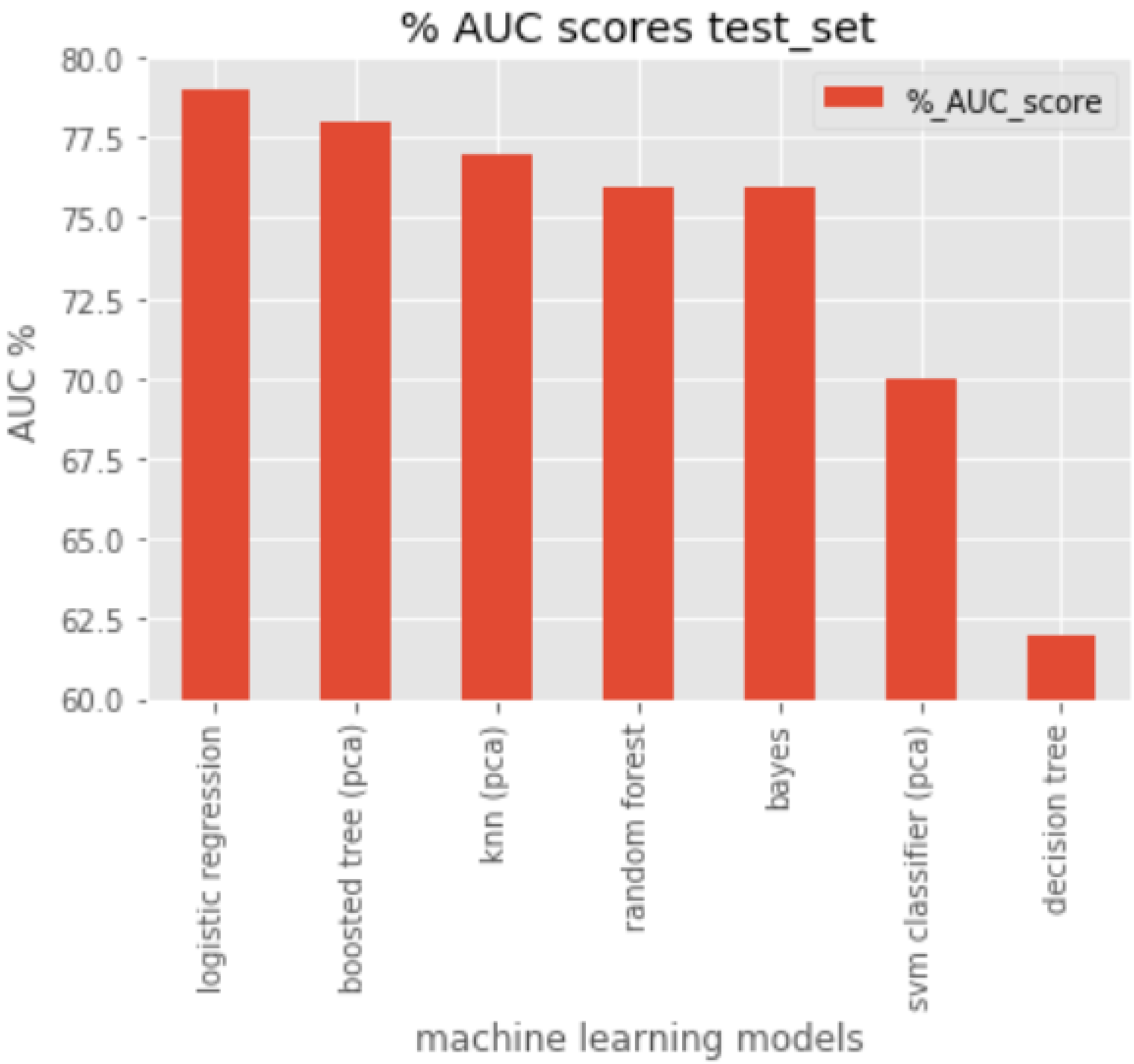
marital_married

education_university.degree

education_high.school

marital_single

Model Comparison:



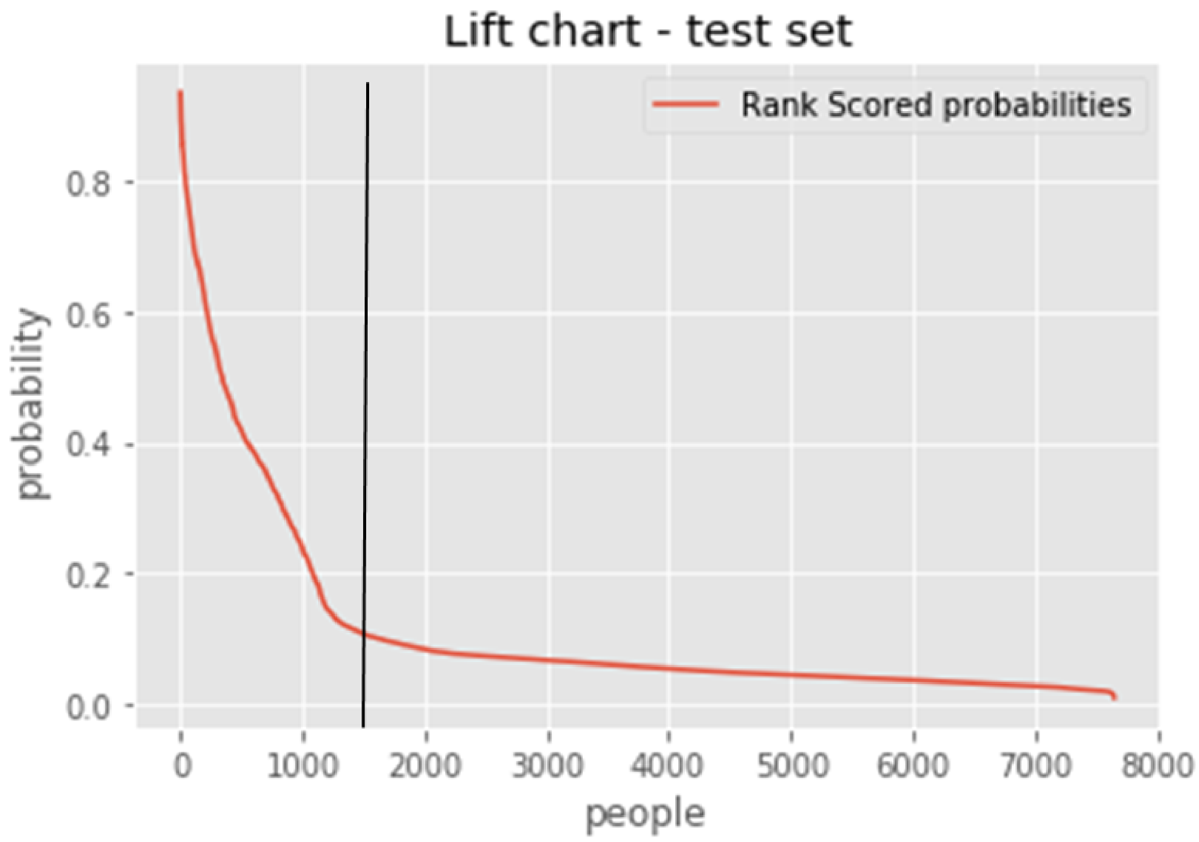
image

Model selection:

Linear logistic regression gives the best results, based on the highest AUC score on test data set (out of bag sample/ dataset not used to train the machine learning models).

Lift Chart

How well my model is performing compared to campaigning to everyone on the list using the best model?



Plot shows how much improvement the rank scored response is performing decile of customers targetted

image

Business Application

Case I: Campaign by mailing everyone on list

Baseline performance

Mailing to everyone in list without using model

Average response rate from customers in test set list is 10.93 %

Total customers in test set list = 7,649

Expected rev from customers for every positive response is $ 10

Cost of each mailing is $ 2

Mailing to everyone in list

Total size of test set: 7649

Average response rate of people in test set: 10.93 %

Total campaign profit: $ -6938

ROI: -45 %

Marketing to every person on the list results in a loss of $ 6,938 and a ROI of negative 45 %

Case II: Campaign based on model

Model performance

Mailing only to people in target list:

Average response rate in the top customer segment is 35.8 % (Applying elbow method on graph shown above)

Total customers in list = 1500 (Obtained from graph shown above corresponding to 36% response rate)

Expected revenue from customers for every positive response is $ 10

Cost of each mailing is $ 2

Marketing only to targeted list of people suggested by model:

Total campaign profit: $ 2374.0

ROI %: 79.0

Marketing to people on targetted list results in a profit of $ 2,374 and improves campaign ROI to 79 %.

Marketing to new list of people with high probability of conversion identified by the machine learning model, improves response rate to 36 % (best model), 3.6 X the overall average response rate of 11 % (baseline) and improves ROI from -45 % (baseline) to 79 % (best model).