Acquisition Analytics

Vidhu Jain

PGDDS – DS C14

Problem Statement

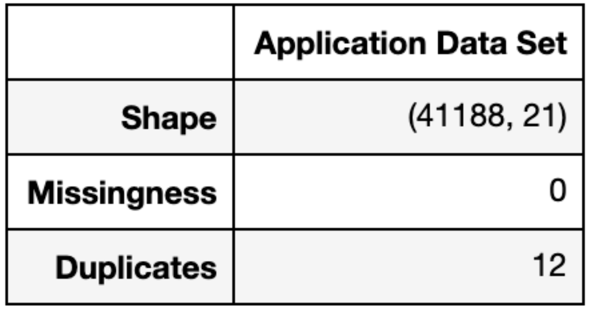
We have been provided the dataset related to a marketing campaign done by a **Portuguese Bank.** The campaigns were telephonic, i.e., sales agents made phone calls to sell a **term deposit** product.

The goal is to build a **response model** which will be used by marketing teams to create an acquisition strategy according to the budget constraints

Our main objective is to:

1. Identify relevant predictor variables for a response using EDA
2. Build predictive models and choose the best one.
3. Find the number of top X% prospects you should target to meet the business objective
4. Report the average call duration for targeting the top X% prospects to the CMO
5. Determine the cost of acquisition

Data Set

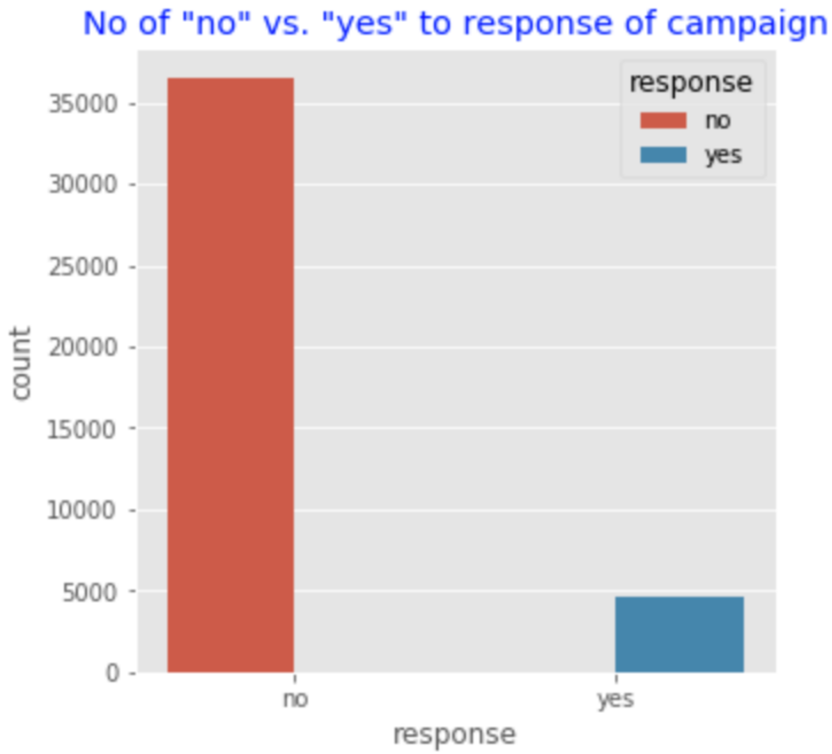


	Application Data Set
Shape	(41188, 21)
Missingness	0
Duplicates	12

There are no null values in the data set. However there are 12 duplicate rows. We will drop these duplicate rows to make our final dataset

Exploratory Data Analysis

Number of ‘no’ vs. ‘yes’ response given in data set

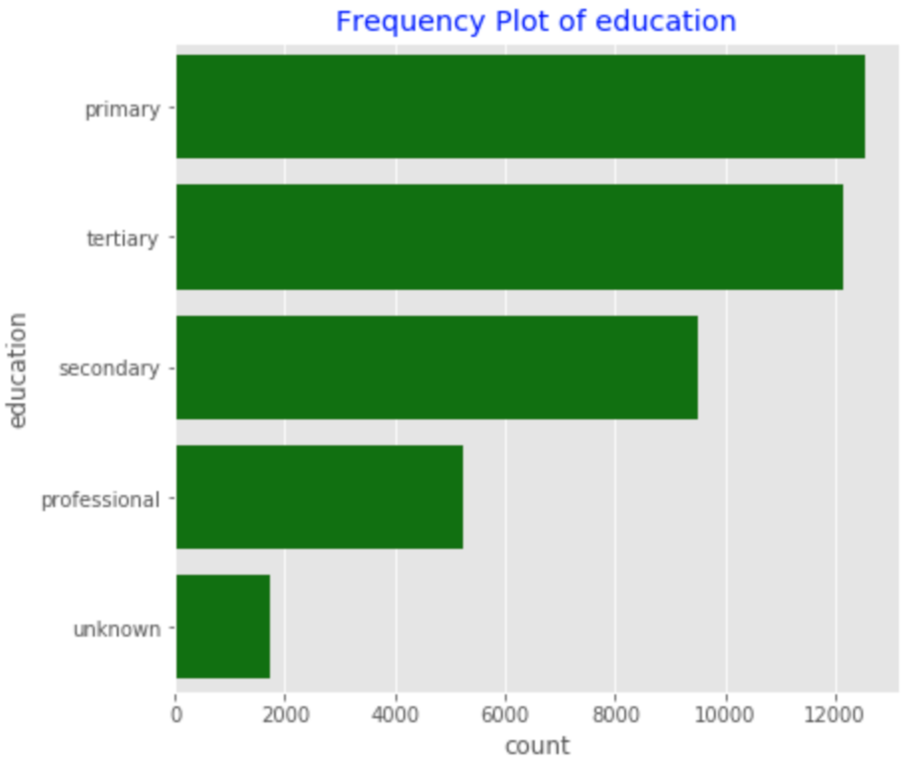


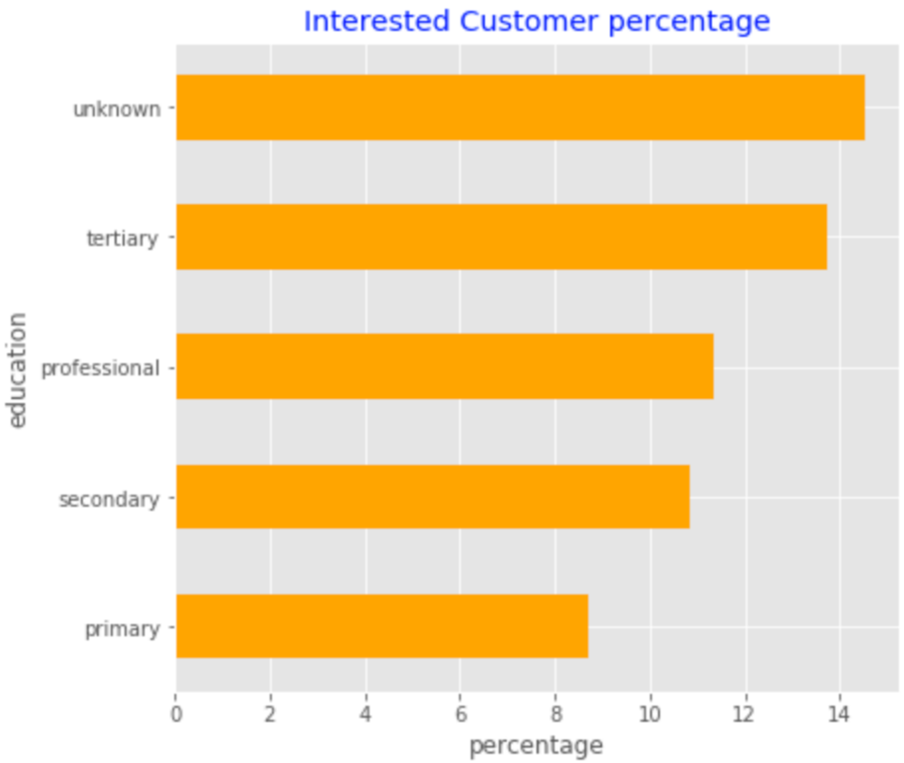
Observations

- The Target class is highly imbalanced with
- 88.73% of observations as "0" - labeled as “no” i.e. no response to campaign
- 11.27% of observations as "1" - labeled as “yes” i.e. yes response to campaign

Exploratory Data Analysis

Education



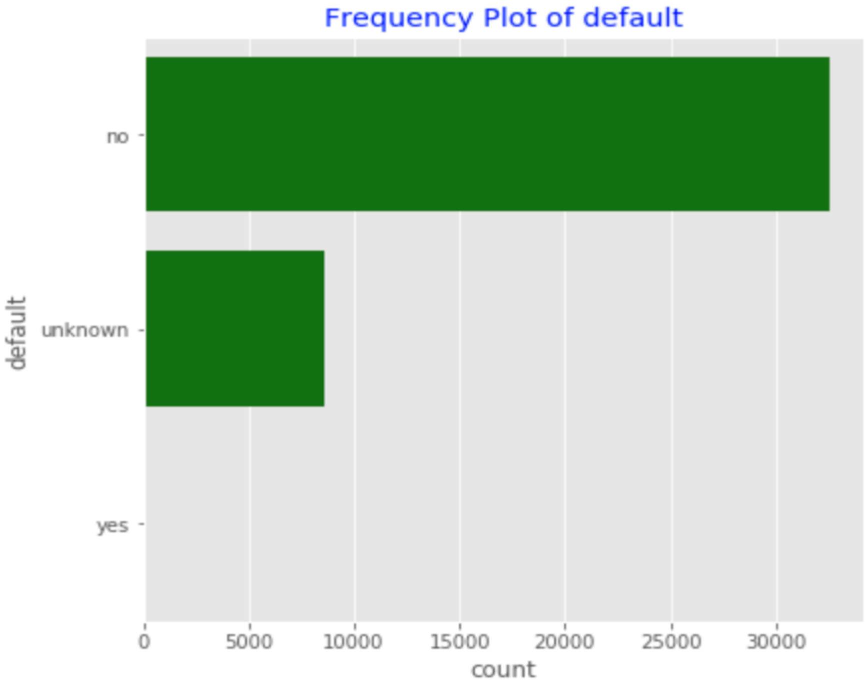


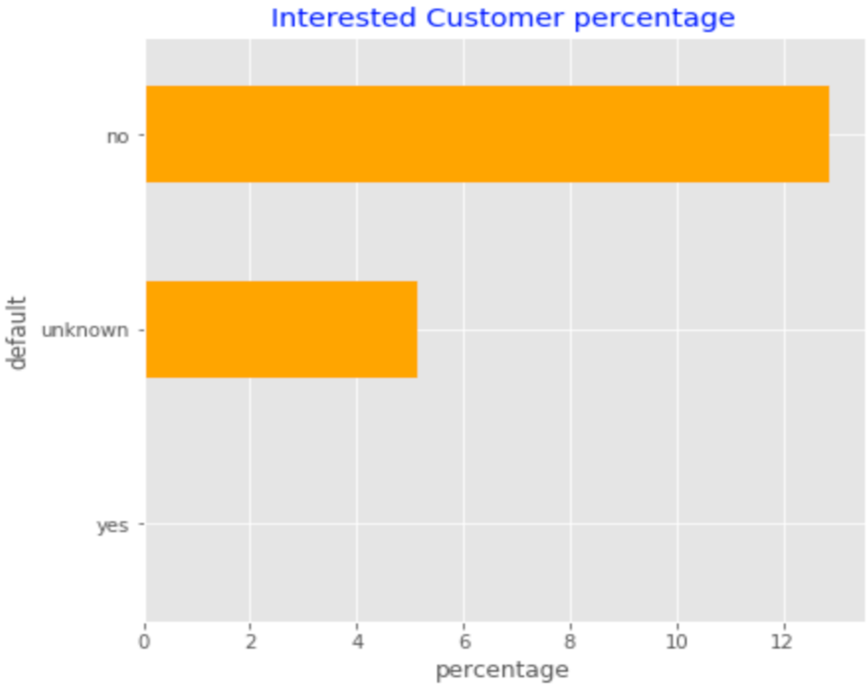
Observations

- Primary education forms the largest part of group of people contacted for campaign while the response rate is lowest for people with Primary education.
- There is group of very less people whose education is not known(1729 count). Out of those 14.52% responded to the campaign.
- Response rate is highest for people with Tertiary education. Perhaps they trust the bank more. This part of group has better relationship with the bank

Exploratory Data Analysis

Default



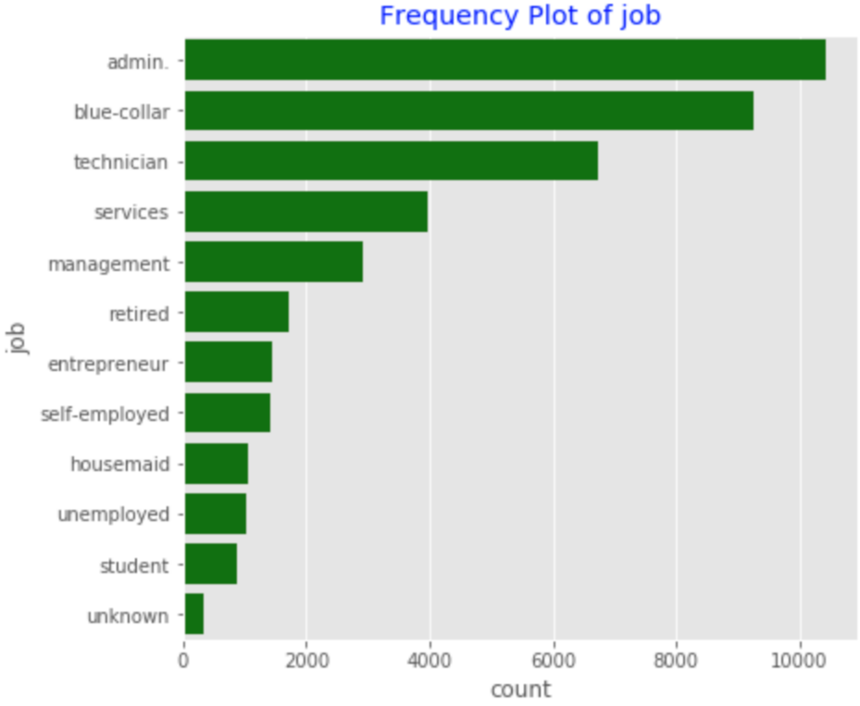


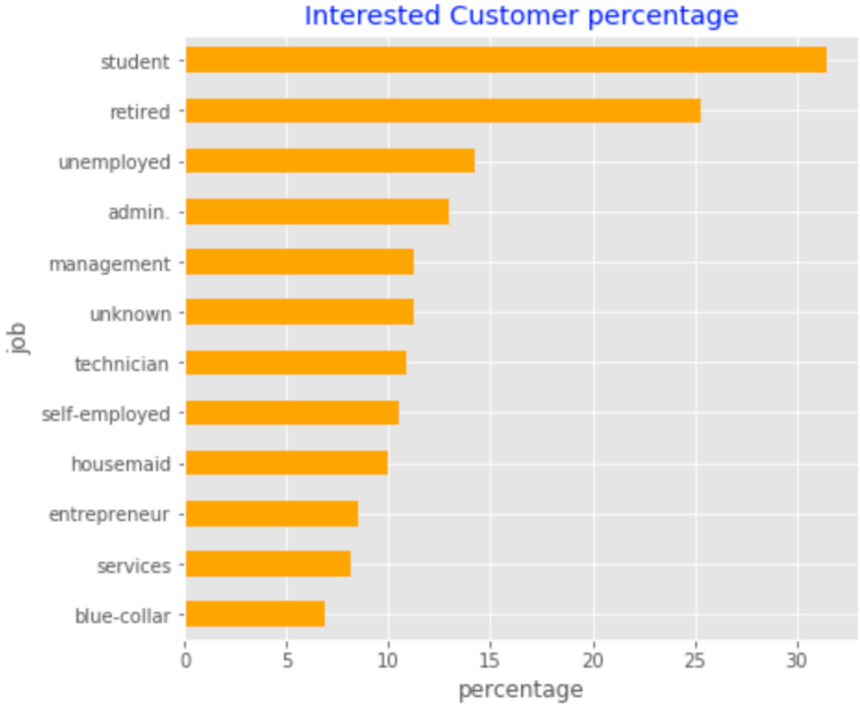
Observations

- There are only 3 people who have defaulted in credit. And none of these responded to the campaign
- People with no previous default have been contacted more during the campaign(32566)
- Also people with no previous default have responded fairly well(12.88%) in comparison to people who previously defaulted or whose default status is not known

Exploratory Data Analysis

Job



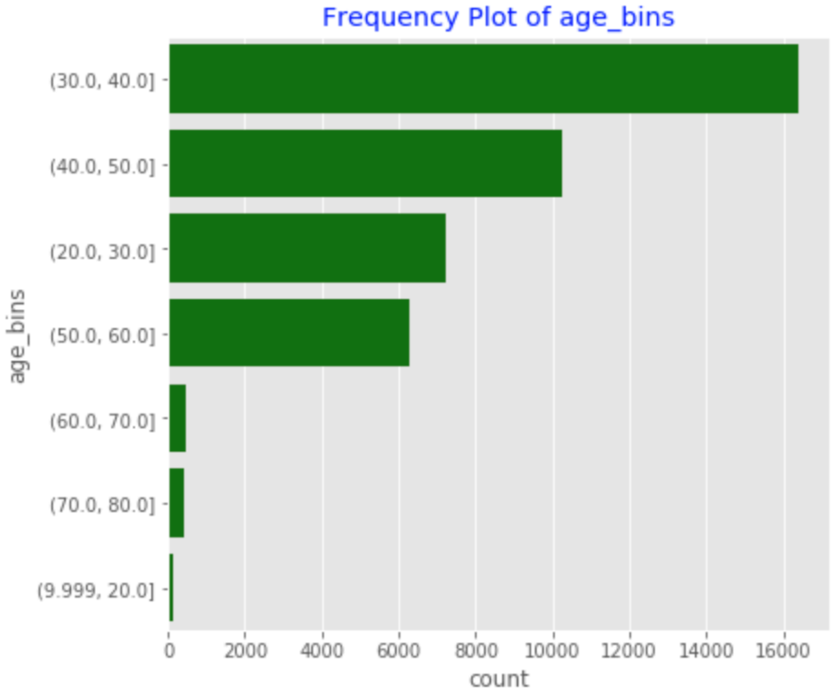


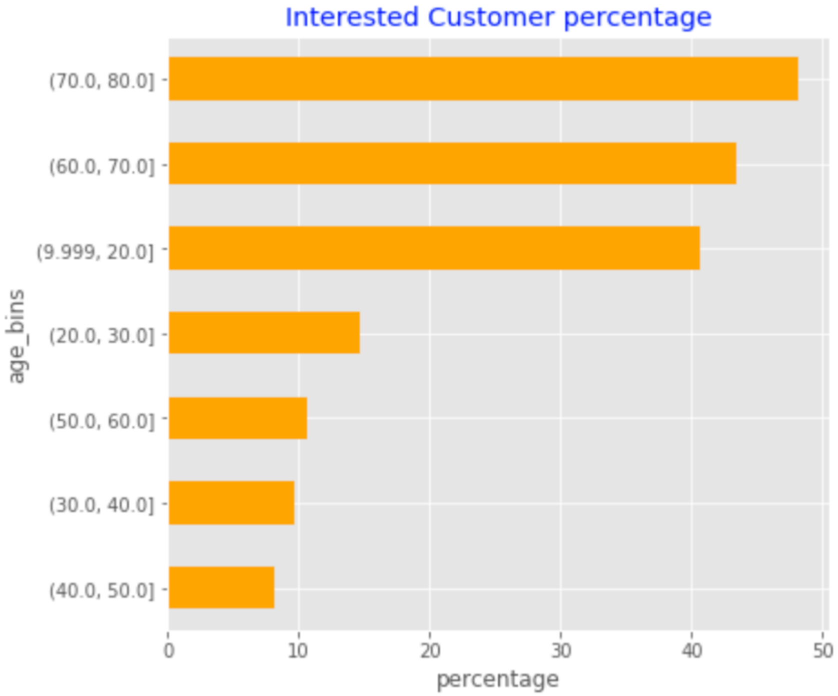
Observations

- People with in profession of administration(10416) and blue-collar profile(9252) were contacted more during the campaign
- Students(31.43%) and retired people(25.29%) responded positive to the campaign while the blue-collar(6.9%) profiled people responded least to the campaign. This may be due to the reason that students and retired people seek more money for their education and retirement respectively

Exploratory Data Analysis

Age Bins



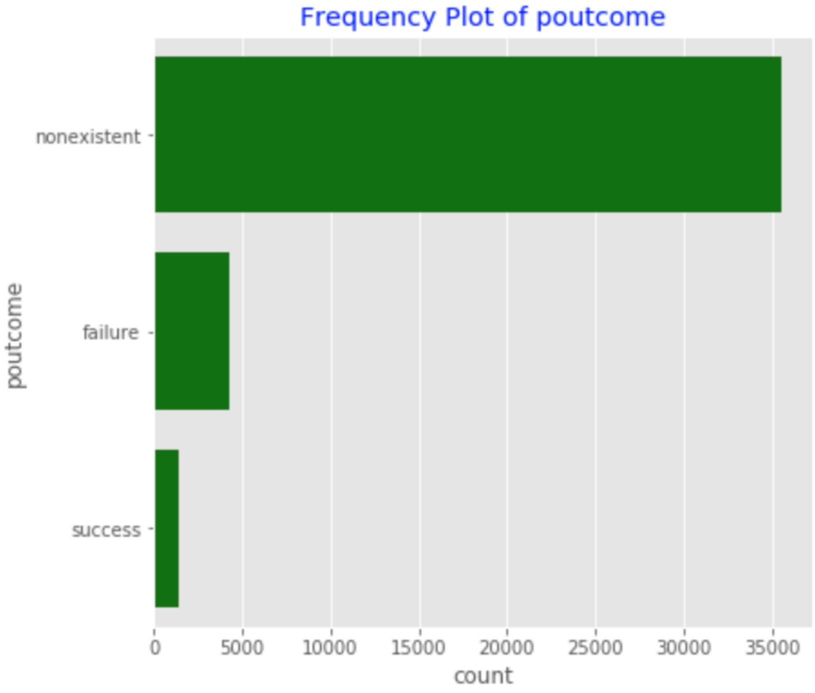


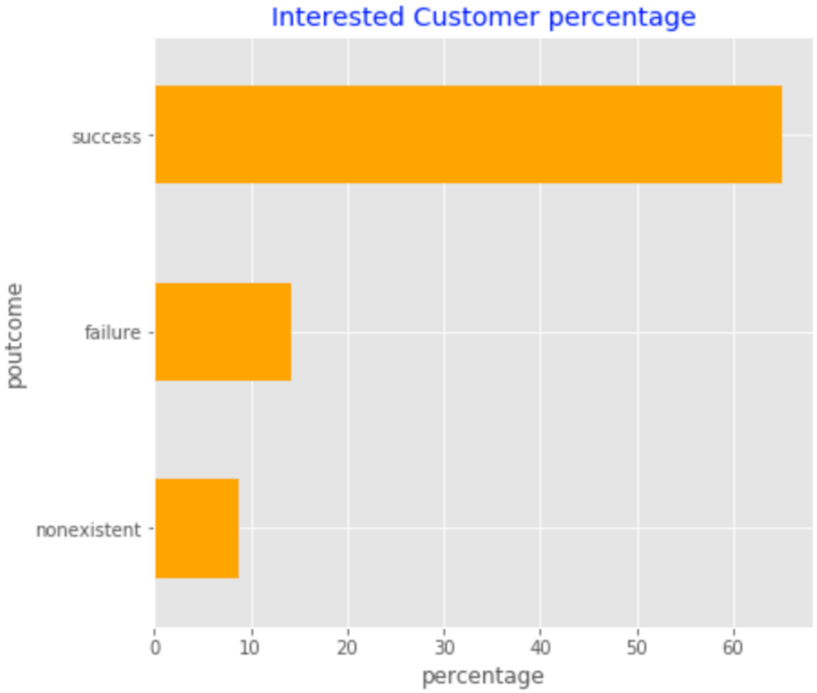
Observations

- People in the age group of 30 to 50 were contacted (16375) more during the campaign followed by people in the age group of 20 to 30 (7239) and 50 to 60 (6268)
- Senior citizens with age above 60 and people in the age group of 10 to 20 responded more to the campaign. This also aligns well with our earlier observation that student responded most to the campaign

Exploratory Data Analysis

Previous Outcome



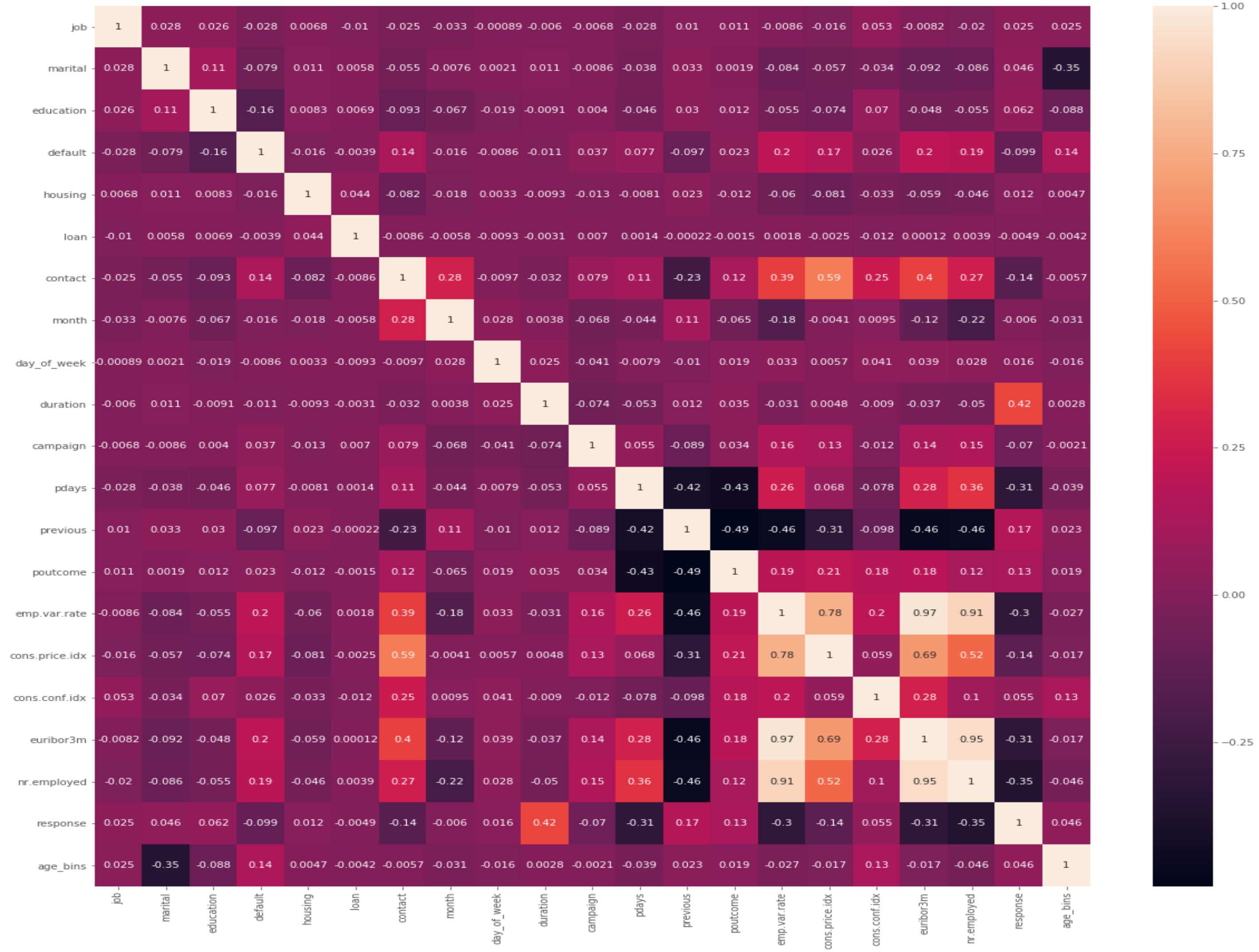


Observations

- Most of the people who form part of the campaign had non-existent outcome of the previous marketing campaign (35539)
- While the people with success as previous outcome(65.11%) responded more to the campaign

Exploratory Data Analysis

Heatmap of all variables to check correlation



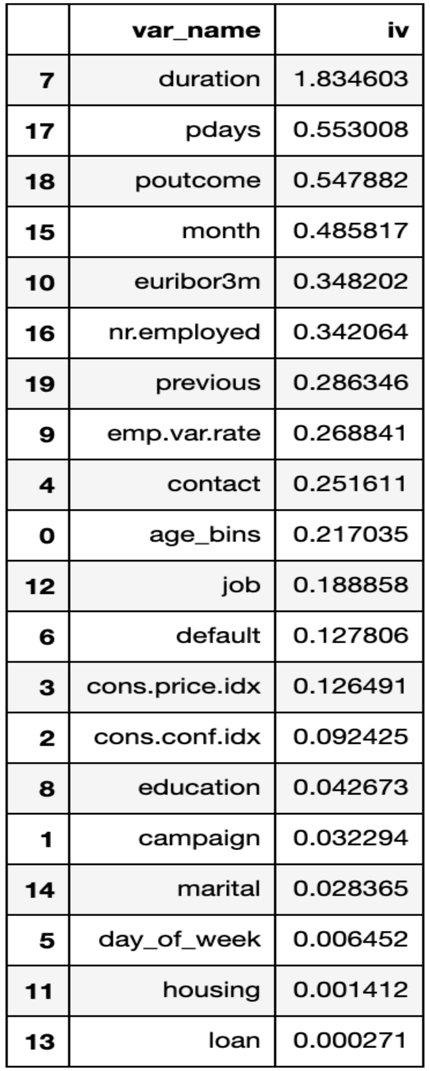
	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	response	age_bins
job	1	0.028	0.026	-0.028	0.0068	-0.01	-0.025	-0.033	-0.00089	-0.006	-0.0068	-0.028	0.01	0.011	-0.0086	-0.016	0.053	-0.0082	-0.02	0.025	0.025
marital	0.028	1	0.11	-0.079	0.011	0.0058	-0.055	-0.0076	0.0021	0.011	-0.0086	-0.038	0.033	0.0019	-0.084	-0.057	-0.034	-0.092	-0.086	0.046	-0.35
education	0.026	0.11	1	-0.16	0.0083	0.0069	-0.093	-0.067	-0.019	-0.0091	0.004	-0.046	0.03	0.012	-0.055	-0.074	0.07	-0.048	-0.055	0.062	-0.088
default	-0.028	-0.079	-0.16	1	-0.016	-0.0039	0.14	-0.016	-0.0086	-0.011	0.037	0.077	-0.097	0.023	0.2	0.17	0.026	0.2	0.19	-0.099	0.14
housing	0.0068	0.011	0.0083	-0.016	1	0.044	-0.082	-0.018	0.0033	-0.0093	-0.013	-0.0081	0.023	-0.012	-0.06	-0.081	-0.033	-0.059	-0.046	0.012	0.0047
loan	-0.01	0.0058	0.0069	-0.0039	0.044	1	-0.0086	-0.0058	-0.0093	-0.0031	0.007	0.0014	-0.00022	-0.0015	0.0018	-0.0025	-0.012	0.00012	0.0039	-0.0049	-0.0042
contact	-0.025	-0.055	-0.093	0.14	-0.082	-0.0086	1	0.28	-0.0097	-0.032	0.079	0.11	-0.23	0.12	0.39	0.59	0.25	0.4	0.27	-0.14	-0.0057
month	-0.033	-0.0076	-0.067	-0.016	-0.018	-0.0058	0.28	1	0.028	0.0038	-0.068	-0.044	0.11	-0.065	-0.18	-0.0041	0.0095	-0.12	-0.22	-0.006	-0.031
day_of_week	-0.00089	0.0021	-0.019	-0.0086	0.0033	-0.0093	-0.0097	0.028	1	0.025	-0.041	-0.0079	-0.01	0.019	0.033	0.0057	0.041	0.039	0.028	0.016	-0.016
duration	-0.006	0.011	-0.0091	-0.011	-0.0093	-0.0031	-0.032	0.0038	0.025	1	-0.074	-0.053	0.012	0.035	-0.031	0.0048	-0.009	-0.037	-0.05	0.42	0.0028
campaign	-0.0068	-0.0086	0.004	0.037	-0.013	0.007	0.079	-0.068	-0.041	-0.074	1	0.055	-0.089	0.034	0.16	0.13	-0.012	0.14	0.15	-0.07	-0.0021
pdays	-0.028	-0.038	-0.046	0.077	-0.0081	0.0014	0.11	-0.044	-0.0079	-0.053	0.055	1	-0.42	-0.43	0.26	0.068	-0.078	0.28	0.36	-0.31	-0.039
previous	0.01	0.033	0.03	-0.097	0.023	-0.00022	-0.23	0.11	-0.01	0.012	-0.089	-0.42	1	-0.49	-0.46	-0.31	-0.098	-0.46	-0.46	0.17	0.023
poutcome	0.011	0.0019	0.012	0.023	-0.012	-0.0015	0.12	-0.065	0.019	0.035	0.034	-0.43	-0.49	1	0.19	0.21	0.18	0.18	0.12	0.13	0.019
emp.var.rate	-0.0086	-0.084	-0.055	0.2	-0.06	0.0018	0.39	-0.18	0.033	-0.031	0.16	0.26	-0.46	0.19	1	0.78	0.2	0.97	0.91	-0.3	-0.027
cons.price.idx	-0.016	-0.057	-0.074	0.17	-0.081	-0.0025	0.59	-0.0041	0.0057	0.0048	0.13	0.068	-0.31	0.21	0.78	1	0.059	0.69	0.52	-0.14	-0.017
cons.conf.idx	0.053	-0.034	0.07	0.026	-0.033	-0.012	0.25	0.0095	0.041	-0.009	-0.012	-0.078	-0.098	0.18	0.2	0.059	1	0.28	0.1	0.055	0.13
euribor3m	-0.0082	-0.092	-0.048	0.2	-0.059	0.00012	0.4	-0.12	0.039	-0.037	0.14	0.28	-0.46	0.18	0.97	0.69	0.28	1	0.95	-0.31	-0.017
nr.employed	-0.02	-0.086	-0.055	0.19	-0.046	0.0039	0.27	-0.22	0.028	-0.05	0.15	0.36	-0.46	0.12	0.91	0.52	0.1	0.95	1	-0.35	-0.046
response	0.025	0.046	0.062	-0.099	0.012	-0.0049	-0.14	-0.006	0.016	0.42	-0.07	-0.31	0.17	0.13	-0.3	-0.14	0.055	-0.31	-0.35	1	0.046
age_bins	0.025	-0.35	-0.088	0.14	0.0047	-0.0042	-0.0057	-0.031	-0.016	0.0028	-0.0021	-0.039	0.023	0.019	-0.027	-0.017	0.13	-0.017	-0.046	0.046	1

Observations

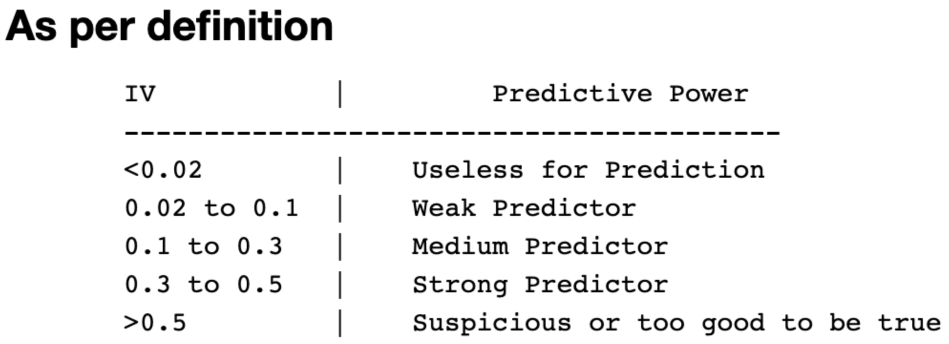
- No variable seems to have a strong correlation with target variable response
- There seems to be a fair positive relationship between duration and response. Also there is a negative correlation between response and pdays variable. Also there seems to be negative and small relationship between response variables and emp.var.rate and cons.price.idx
- There seems to be positive relationship between con.price.idx and contact
- Also there seems to be very high positive relationship between
    - euribor3m and emp.var.rate
    - nr.employed and emp.var.rate
    - nr.employed and euribor3m
    - euribor3m and cons.price.idx
- We will have to further check the multicolinearity amongst the variable and variance inflation factor as well

Exploratory Data Analysis

Information Value of all variables



	var_name	iv
7	duration	1.834603
17	pdays	0.553008
18	poutcome	0.547882
15	month	0.485817
10	euribor3m	0.348202
16	nr.employed	0.342064
19	previous	0.286346
9	emp.var.rate	0.268841
4	contact	0.251611
0	age_bins	0.217035
12	job	0.188858
6	default	0.127806
3	cons.price.idx	0.126491
2	cons.conf.idx	0.092425
8	education	0.042673
1	campaign	0.032294
14	marital	0.028365
5	day_of_week	0.006452
11	housing	0.001412
13	loan	0.000271



As per definition

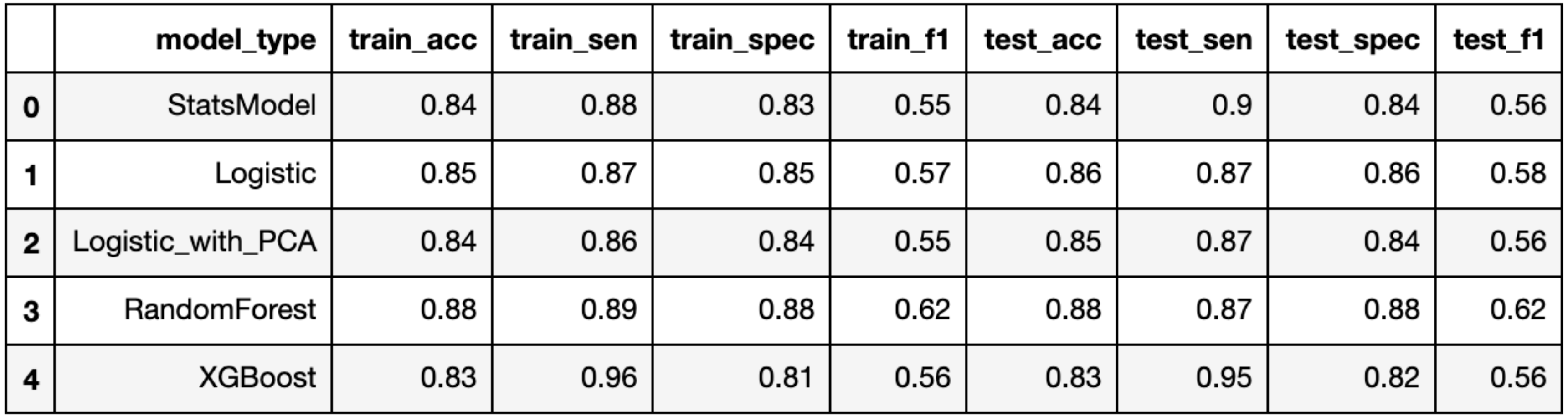
```
IV             |         Predictive Power
-----------------------------------------
<0.02          |    Useless for Prediction
0.02 to 0.1    |    Weak Predictor
0.1 to 0.3     |    Medium Predictor
0.3 to 0.5     |    Strong Predictor
>0.5           |    Suspicious or too good to be true
```

Observations

- Going by Information Value, feature **'duration'** seems to be too good to be true. Its **not a good predictor**
- poutcome, month, euribor3m, nr.employed seem to be **strong predictors**
- previous, emp.var.rate, contact, age_bins, job, default, cons.price.idx seem to be **medium predictors**
- We will build various models and check for model evaluation metrics and feature importance

Model Building

Evaluation metrics of various models



	model_type	train_acc	train_sen	train_spec	train_f1	test_acc	test_sen	test_spec	test_f1
0	StatsModel	0.84	0.88	0.83	0.55	0.84	0.9	0.84	0.56
1	Logistic	0.85	0.87	0.85	0.57	0.86	0.87	0.86	0.58
2	Logistic_with_PCA	0.84	0.86	0.84	0.55	0.85	0.87	0.84	0.56
3	RandomForest	0.88	0.89	0.88	0.62	0.88	0.87	0.88	0.62
4	XGBoost	0.83	0.96	0.81	0.56	0.83	0.95	0.82	0.56

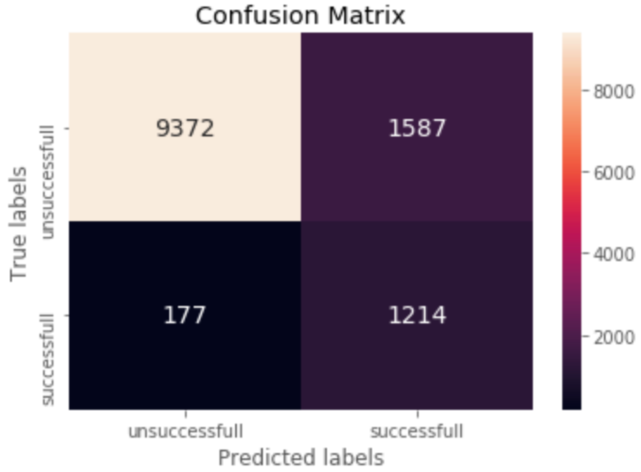
Observations

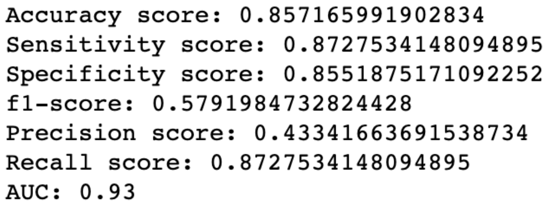
- Logistic Regression gave the best result. Train and Test scores are almost stable and Sensitivity and Test Sensitivity is better.
- We will use Logistic Regression to finally build our model

Model Evaluation

Evaluation metrics of Final Model

Logistic Regression **with** duration



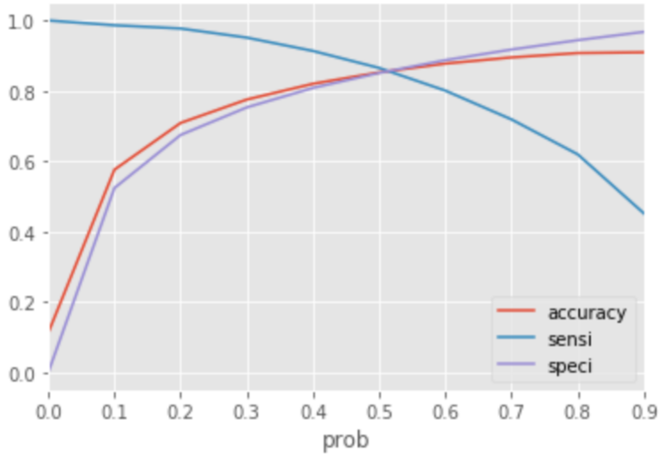


```
Accuracy score: 0.857165991902834
Sensitivity score: 0.8727534148094895
Specificity score: 0.8551875171092252
f1-score: 0.5791984732824428
Precision score: 0.43341663691538734
Recall score: 0.8727534148094895
AUC: 0.93
```

Observations

- Test Accuracy of the model is 85% with Sensitivity of 87%
- Area under the curve is 0.93

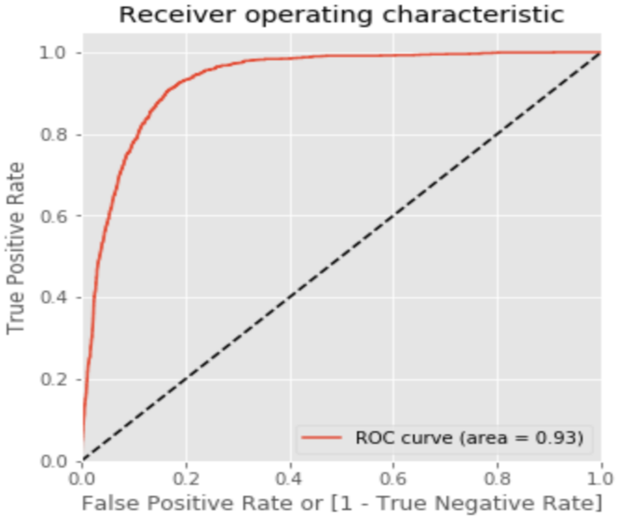
Cut-Off Probability



Observations

- 0.5 is the optimum point to take it as cut-off probability

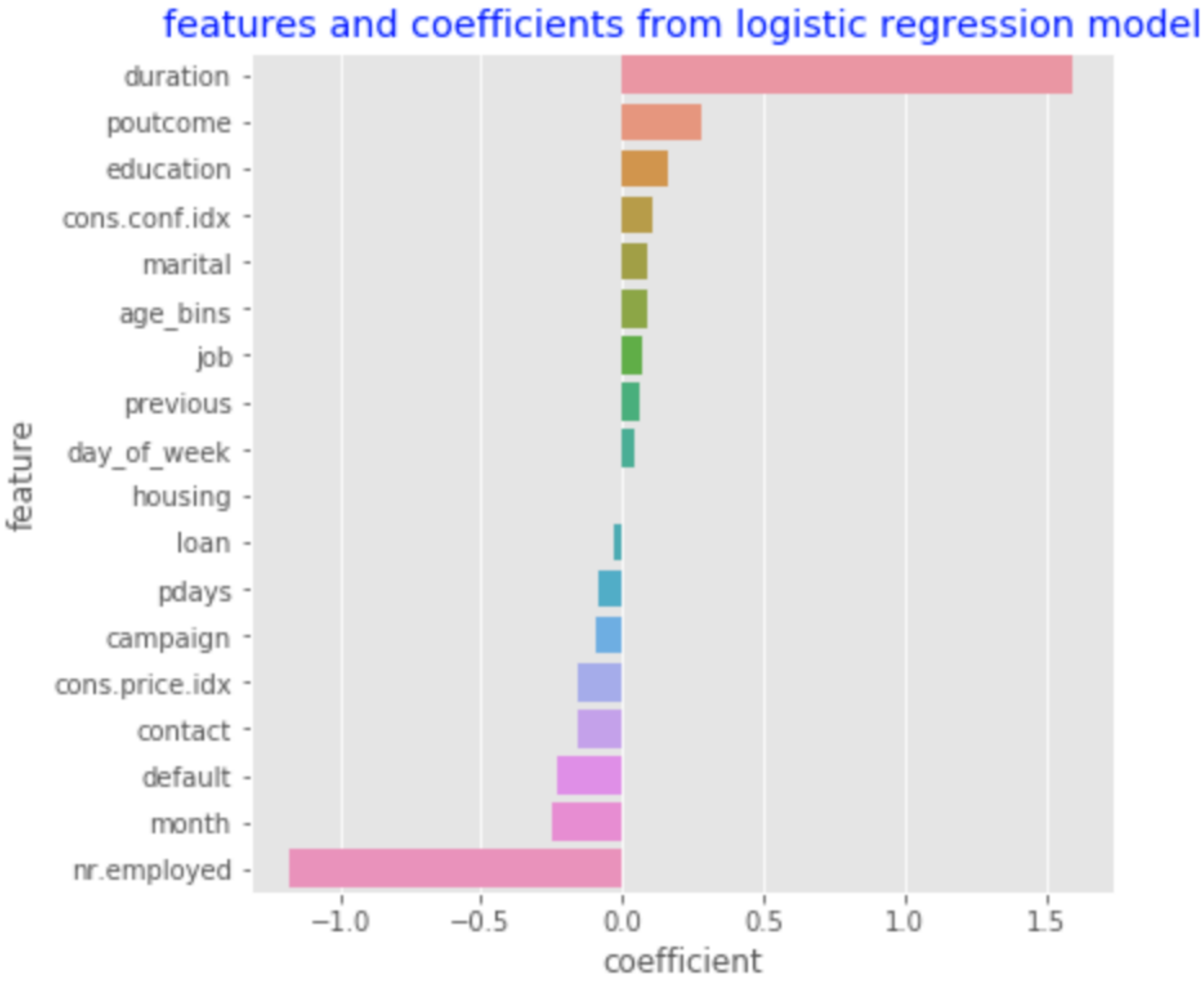
ROC-AUC Curve



Observations

- Area under the curve is 0.93

Feature Importance



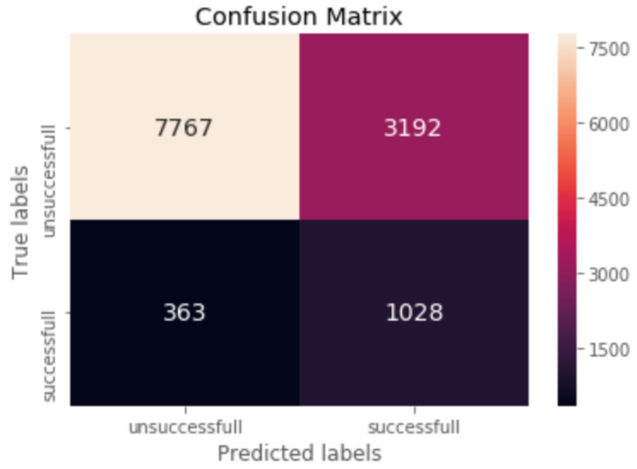
Observations

- duration forms the very important predictor with positive coefficient. This means more the duration of call, more likely the person going to respond to the campaign
- nr.employed also is an important predictor
- since duration of call is not known beforehand so we will drop duration feature and will rebuild the model

Model Evaluation

Evaluation metrics of Final Model

Logistic Regression **without** duration

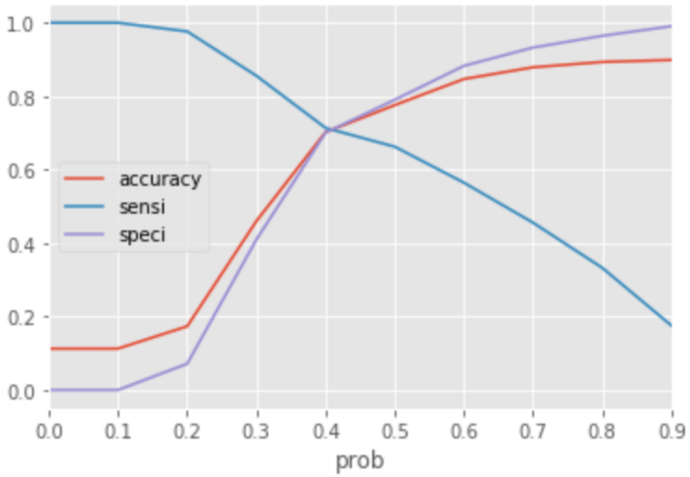


```
Accuracy score: 0.7121457489878542
Sensitivity score: 0.7390366642703091
Specificity score: 0.7087325485901999
f1-score: 0.3664230974870789
Precision score: 0.24360189573459715
Recall score: 0.7390366642703091
AUC: 0.79
```

Observations

- Test Accuracy of the model is 71% with Sensitivity of 73%
- Area under the curve is 0.79

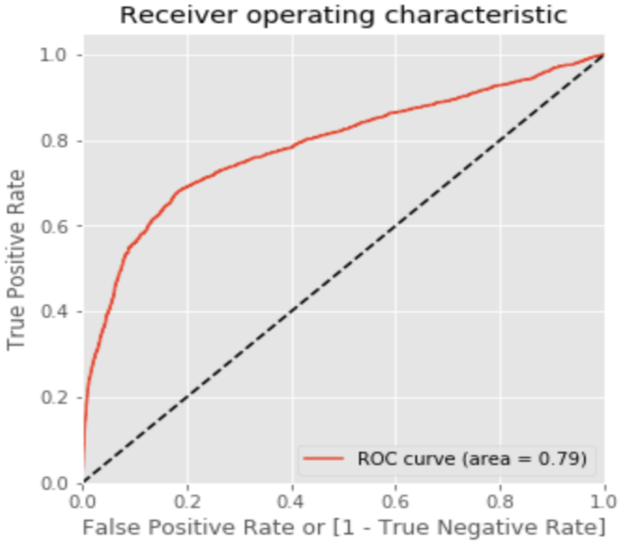
Cut-Off Probability



Observations

- 0.4 is the optimum point to take it as cut-off probability

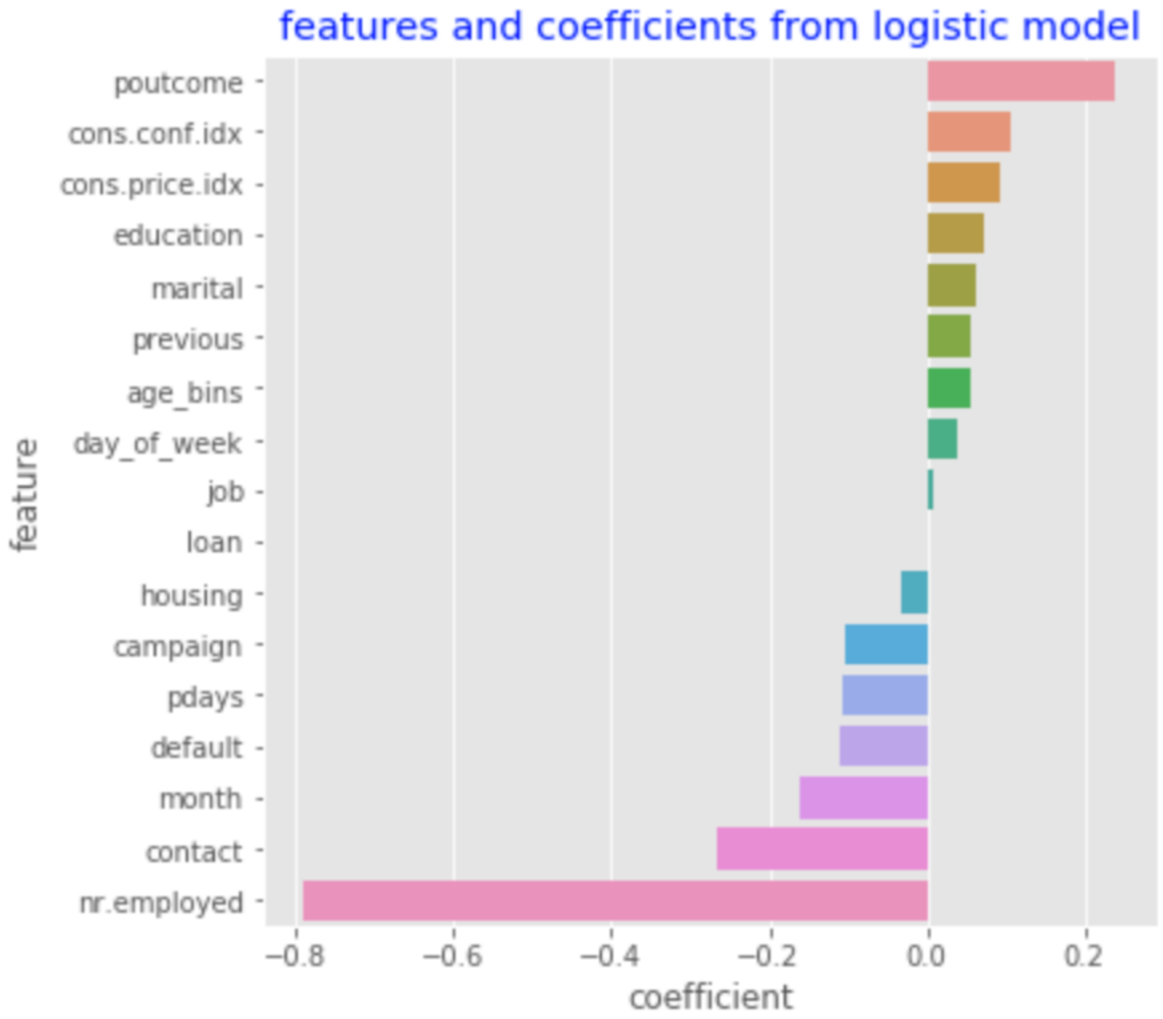
ROC-AUC Curve



Observations

- Area under the curve is 0.79

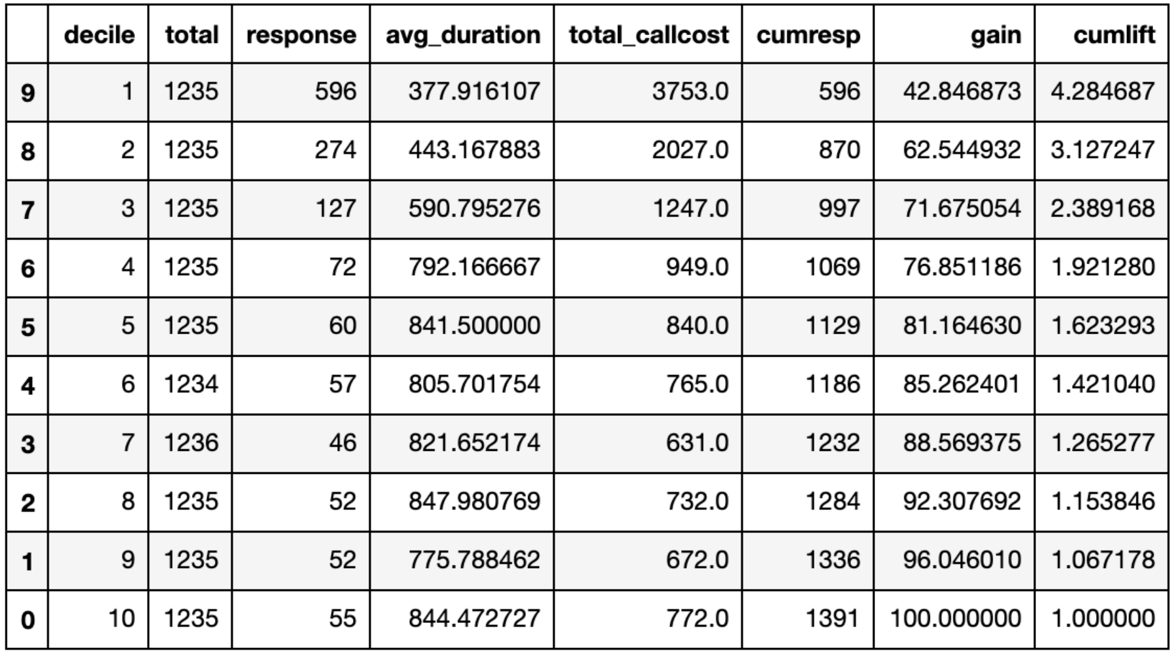
Feature Importance



Observations

- nr.employed forms the very important predictor with negative coefficient.
- contact is also is an important predictor with negative coefficient. This means if the person is contacted less then there are more chance of the person responding to the campaign

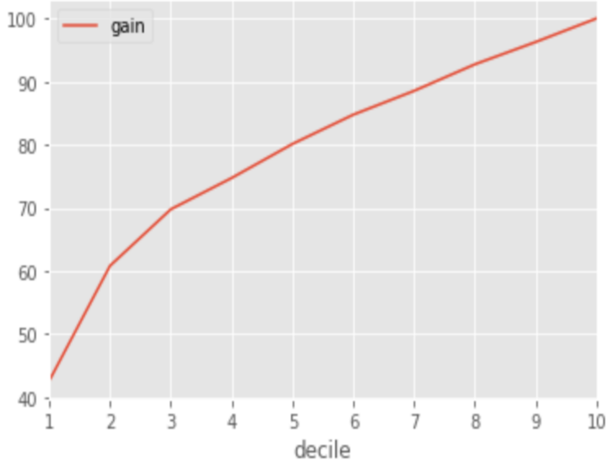
Lift and Gain Charts

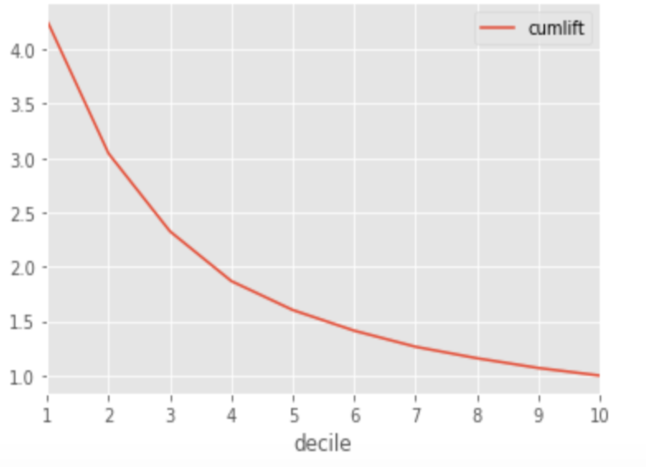


	decile	total	response	avg_duration	total_callcost	cumresp	gain	cumlift
9	1	1235	596	377.916107	3753.0	596	42.846873	4.284687
8	2	1235	274	443.167883	2027.0	870	62.544932	3.127247
7	3	1235	127	590.795276	1247.0	997	71.675054	2.389168
6	4	1235	72	792.166667	949.0	1069	76.851186	1.921280
5	5	1235	60	841.500000	840.0	1129	81.164630	1.623293
4	6	1234	57	805.701754	765.0	1186	85.262401	1.421040
3	7	1236	46	821.652174	631.0	1232	88.569375	1.265277
2	8	1235	52	847.980769	732.0	1284	92.307692	1.153846
1	9	1235	52	775.788462	672.0	1336	96.046010	1.067178
0	10	1235	55	844.472727	772.0	1391	100.000000	1.000000

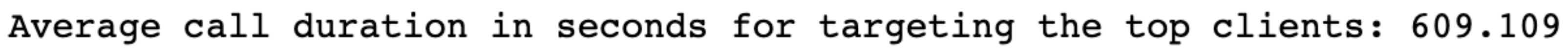
Observations

- One can attain more than 80% of total acquisitions by targeting top 50% of the total client base.
- Thus top 50% of the total client base can be targeted to achieve 80% of the response rate thereby reducing the campaign cost which otherwise would have been more if all client base have been contacted



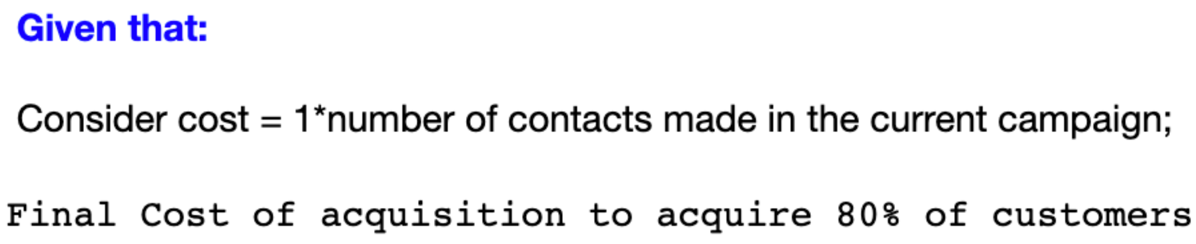


Average call duration for targeting top 50% of prospects to acquire 80% of customers



```
Average call duration in seconds for targeting the top clients: 609.109
```

Final Cost of acquisition for targeting top 50% of prospects to acquire 80% of customers



Given that:

Consider cost = 1*number of contacts made in the current campaign;

```
Final Cost of acquisition to acquire 80% of customers: 8816.0
```

Conclusion and Recommendations

- Given the duration call not known beforehand, excluding this feature, Logistic Regression gave the best results with accuracy of model 71% and Sensitivity of model around 71%
- nr.employed, contact and poutcome form the important predictors
- With this model, we can achieve 80% of acquisition by targeting top 50% of the total client base.
- The total cost of acquisition to acquire 80% of customers is 8816 units
- The average call duration for targeting top 50% clients is 609.109 seconds