Acquisition Analysis

BY

JAGDISH MIRCHANDANI

3 – Nov 2019

Business Objective

- Achieve 80% of total responders at the minimum possible cost.
- Predict the probability of response and target most likely respondents in the campaign.
- Excluding the feature “duration” from the model.
- How many prospects should be called to meet the business objective.

Problem solving methodology

EDA and Data Preparation –

- Different attributes and their relationship with “response” variable were analyzed to identify relevant predictors .
- A unique ID was assigned to each prospect.
- Assumption for calculation of Call Rate: $1 USD/min
- Model Building
  - Logistic Regression Model, without using “duration” variable
  - Logistic Regression with PCA
  - Logistic Regression with RFE
- Identifying the top X% prospects to target to achieve Business Objective
- Lift Chart demo
- Identifying the Cost of Acquisition
- Focus should be on “Sensitivity” as the objective is to identify the true positive rate.

Problem solving methodology Contd..

Model Building -Logistic Regression Model with all variables

- Model contained lots of insignificant features.

- Dataset had class-imbalance
  - 0(No): 88%
  - 1(Yes) 12%

This was handled while creating the model.

Model building using PCA

- built model with 16-18 PCs, sensitivity achieved was 61%

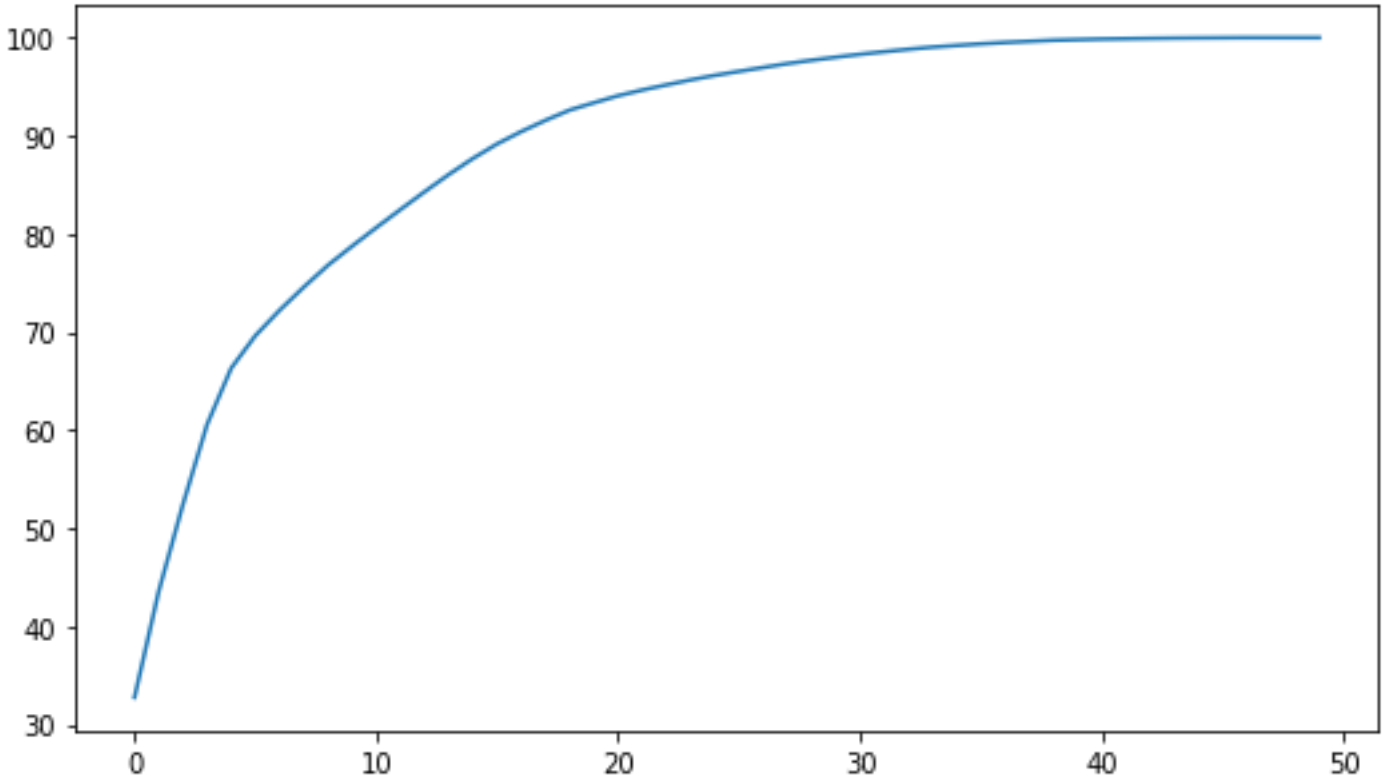
Model building using RFE

- built a model with GLM using RFE (started with 15 features and finally reduced to 9 features, variables were dropped one by one based on their p-values and VIF).
- 69% of sensitivity was achieved on the test data and 68% on the whole data set, since the results were decent, rest of the analysis was done on the results produced by this model.

Logistic Regression Model with PCA(Model Evaluation)

Scree Plot

16 Principal Components can explain 90% Variance in the dataset



Best hyper parameters:

- 'logistic_C': 1
- 'logisticpenalty': 'l2'
- 'pca_n_components': 18
- Sensitivity : 0.61
- Specificity: 0.83
- Accuracy: 0.78

Sensitivity achieved was 61%

Logistic Regression Model using RFE

- Automated Approach: RFE (Recursive feature elimination) with number of features = 15.
- Dropped insignificant variables one by one based on p-values and VIF.
- Finally the model was built using 9 features.

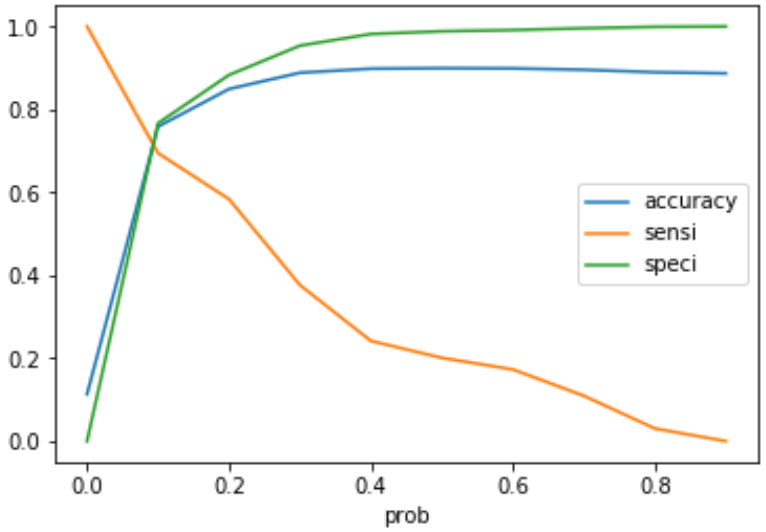
	Features	VIF
2	contact_telephone	3.12
7	cons.price.idx	2.61
8	euribor3m	2.37
5	previous_Never contacted	2.06
4	month_may	2.02
6	poutcome_success	1.13
0	job_retired	1.06
3	month_mar	1.06
1	job_student	1.05

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5779	0.064	-40.418	0.000	-2.703	-2.453
job_retired	0.4339	0.081	5.370	0.000	0.276	0.592
job_student	0.4654	0.102	4.561	0.000	0.265	0.665
contact_telephone	-0.1845	0.059	-3.123	0.002	-0.300	-0.069
month_mar	0.7987	0.113	7.085	0.000	0.578	1.020
month_may	-0.9323	0.051	-18.139	0.000	-1.033	-0.832
previous_Never contacted	0.4048	0.063	6.457	0.000	0.282	0.528
poutcome_success	1.9043	0.090	21.275	0.000	1.729	2.080
cons.price.idx	0.1799	0.024	7.383	0.000	0.132	0.228
euribor3m	-0.9710	0.027	-35.707	0.000	-1.024	-0.918

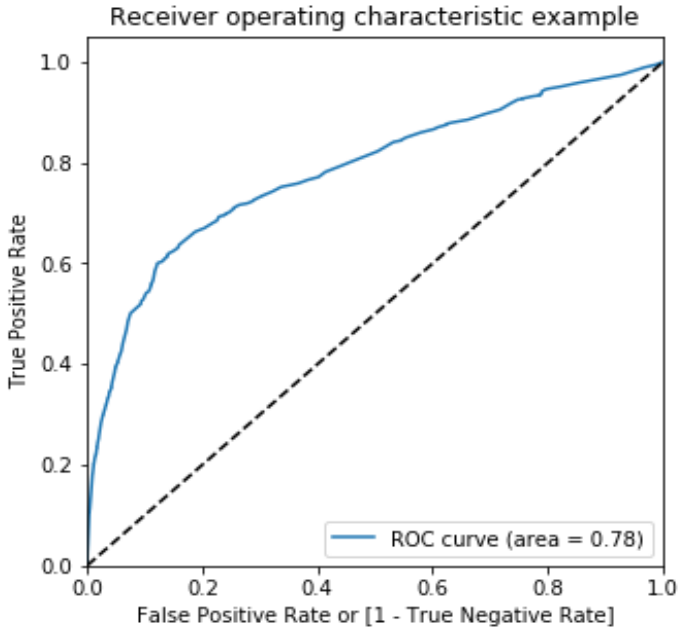
Model Building -Logistic Regression Model with RFE (Optimization)

- ROC Curve demonstrates tradeoff between sensitivity and specificity.
- Closer the curve follows the left-hand border and then the top border of ROC space, the more accurate the test.
- Cut Off Point is 0.1 where, accuracy, sensitivity and specificity coincide.

Trade off graph



ROC Curve



Model Evaluation -Logistic Regression Model with RFE

Metrics	Training Data Set	Test Data Set
Accuracy	76%	75%
Sensitivity	69%	68%
Specificity	76%	76%
Precision	27%	26%
Recall	69%	68%

Confusion Matrix

Training Data Set

Actual/Predicted	Not Converted	Converted
Not Converted	19603	5970
Converted	995	2263

Test Data Set

Actual/Predicted	Not Converted	Converted
Not Converted	8349	2626
Converted	445	937

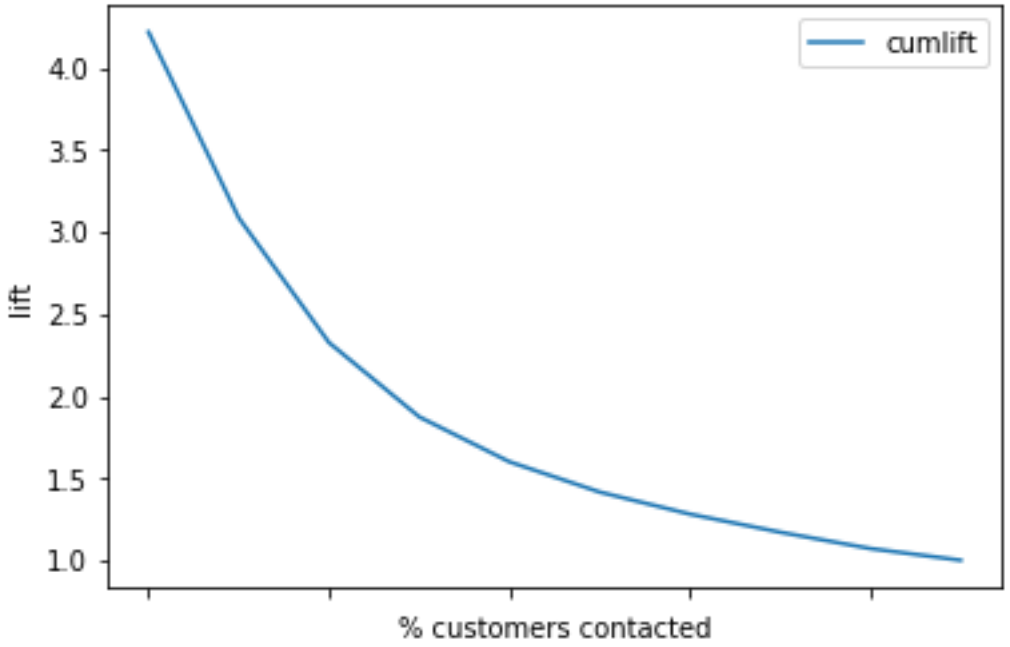
Apply model on the whole dataset

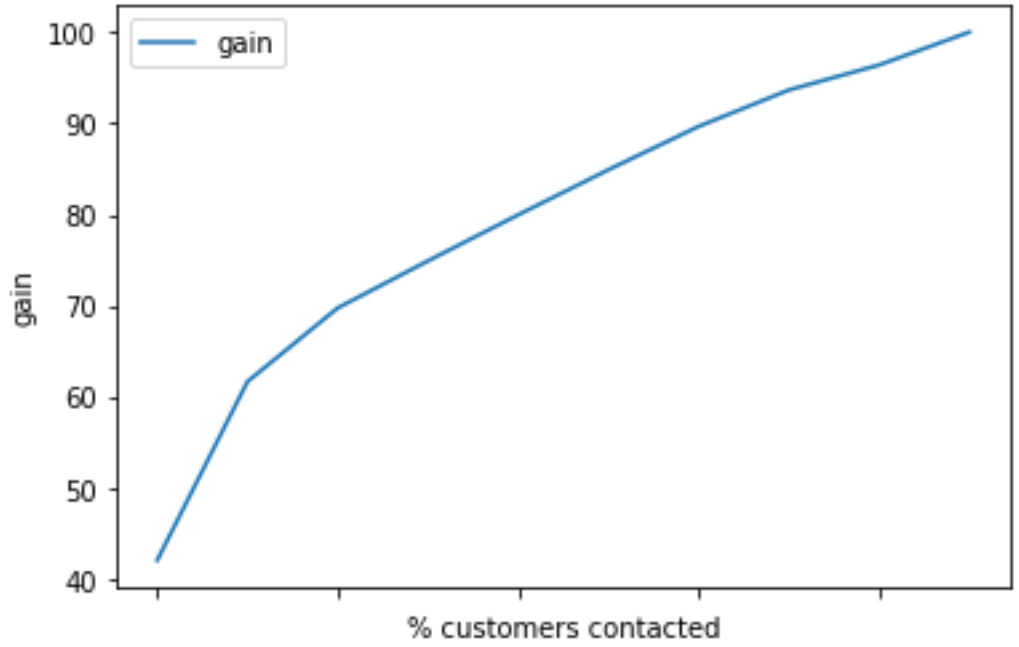
- To meet business objective, we need to achieve 80% response at a minimal cost
- From the table depicted here, it is evident that 80% response can be achieved by targeting 50% (5thdecile) of the total client base (41,188), which is 20,594
- Avg. call-duration per person for targeting top 80% prospect is 260.79 seconds.

	decile	total	actual_response	cumresp	gain	cumlift
9	1	4101	1958	1958	42.198276	4.219828
8	2	4060	907	2865	61.745690	3.087284
7	3	4182	374	3239	69.806034	2.326868
6	4	4085	239	3478	74.956897	1.873922
5	5	3898	234	3712	80.000000	1.600000
4	6	4295	229	3941	84.935345	1.415589
3	7	4014	220	4161	89.676724	1.281096
2	8	4205	185	4346	93.663793	1.170797
1	9	3518	127	4473	96.400862	1.071121
0	10	4830	167	4640	100.000000	1.000000

Lift Chart (Entire Dataset)

80% Response Rate is achieved at 5th Decile





The x-axis should show the number of prospects contacted; the y-axis should show the ratio of the response rate

Cost Of Acquisition

Cost of Acquisition for 80% response rate

- Cost to be considered = 1*number of contacts made in the current campaign

- We will calculate the value on the Entire Data
  Cost= 1* (50% of 41,188) = 20,594

- Since, 50% of base is required to be contacted to achieve 80%

Conclusion

- To achieve the business objective of acquiring 80% of total prospects at minimum possible cost, we will need to target 50% of the total customer base for entire dataset.

- Significant variables identified by the model:

	Features	VIF
2	contact_telephone	3.12
7	cons.price.idx	2.61
8	euribor3m	2.37
5	previous_Never contacted	2.06
4	month_may	2.02
6	poutcome_success	1.13
0	job_retired	1.06
3	month_mar	1.06
1	job_student	1.05

- Logistic model created above has improved 50% efficiency, as instead of contacting the whole set of customers, we can contact only 50% of the customers to achieve the business objective.