Assignment - Acquisition Analytics

Sachita Chauhan

22-July-2019

Tasks

- Business Objective: Achieve 80% of total responders at the minimum possible cost
- Predict the probability of response and target most likely respondents in the telemarketing campaign
  - Excluding “duration” from the model
- How many prospects should be called to meet the business objective?
  - Calculating the X in top X%
- Methodology

Methodology Overview

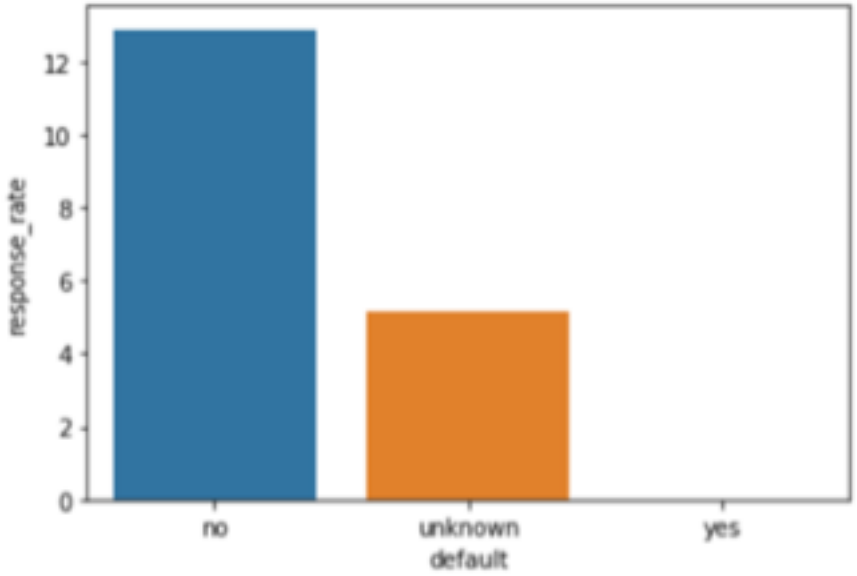
- EDA and Data Preparation – Different attributes and their relationship with “response” variable was studied to identify relevant predictors from the available data set classifications.
  - Client Data
  - Marketing Data (Last Contact Data)
  - Misc.
  - Social and Economic Data
- A unique ID for each prospect is created, to help in detailed analysis of test data
- Assumption for calculation of Call Rate: $1 USD/min

Methodology Overview

- Model Building – Logistic Regression Model, without using “duration” variable
  - Logistic Regression with all variables
  - Logistic Regression with RFE
  - Logistic Regression with PCA
- Identifying the top X% prospects to target to achieve Business Objective
- Creating a Lift Chart
- Identifying the Cost of Acquisition
- We should focus on “Sensitivity”, as our objective is to identify the true positive rate

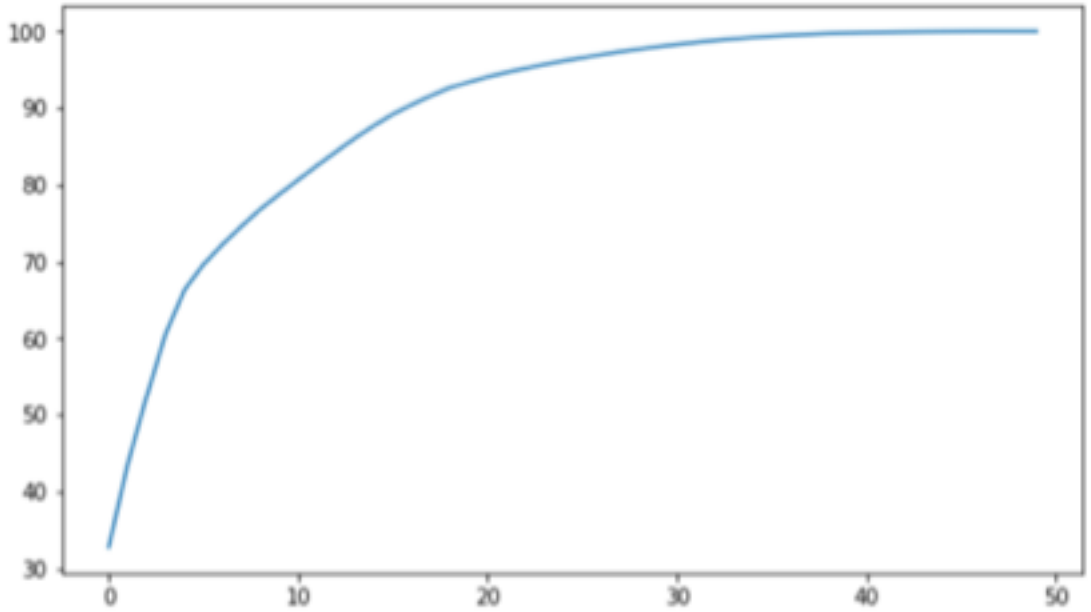
Model Building -Logistic Regression Model with all variables

- Model contain lots of insignificant values
- Tuning this model will eventually involve dropping multiple variables one by one.
- Additionally, this is also not considering class-imbalance.
- Since, the dataset has class-imbalance
  - 0: 88%
  - 1: 12%
- Therefore we proceed Logistic regression
  - With PCA
  - with GLM using RFE.



Model Building - Logistic Regression Model with PCA

- Scree Plot
  - Indicates that 16-18 components are sufficient to achieve more than 90% variance in dataset



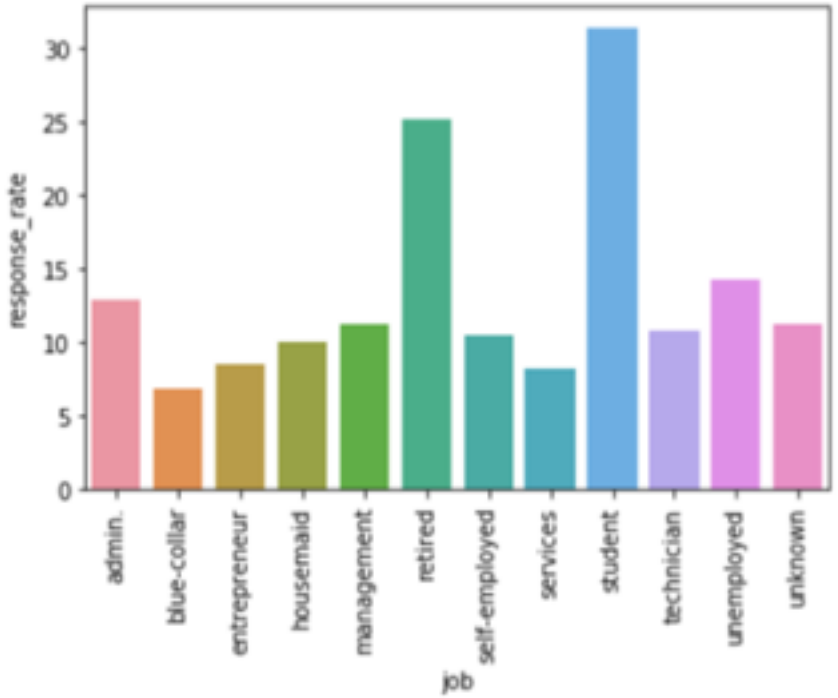
- Best hyperparameters:
  - 'logistic_C': 1
  - 'logisticpenalty': 'l2’
  - 'pca_n_components': 18
- Sensitivity : 0.61
- Specificity: 0.83
- Accuracy: 0.78

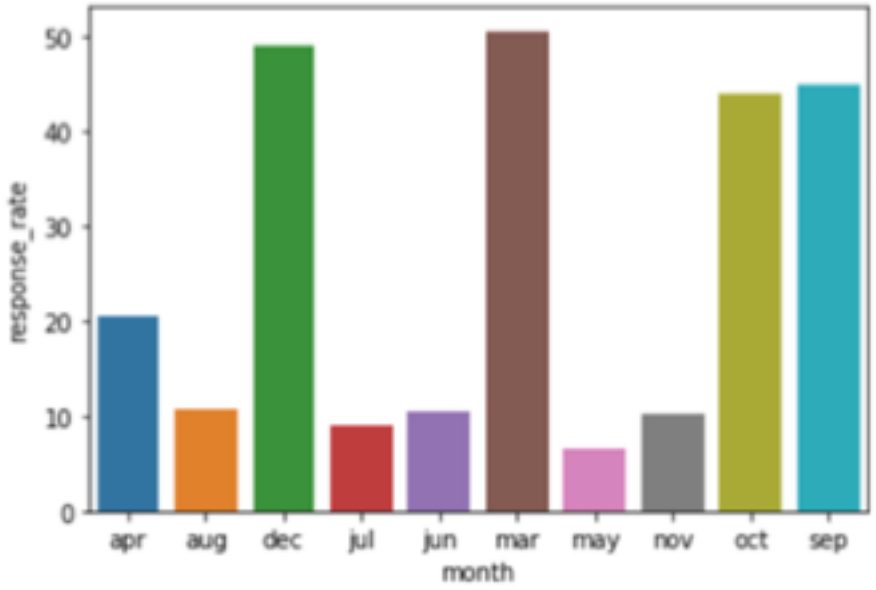
Model Building - Logistic Regression Model with RFE

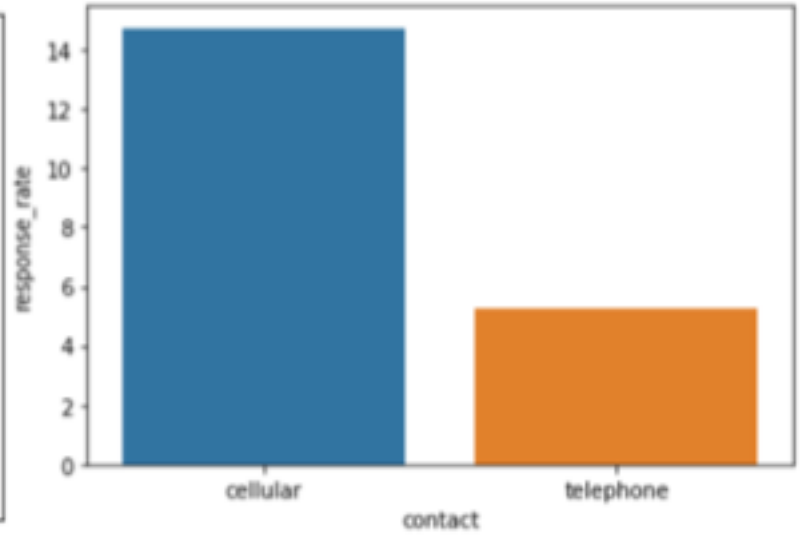
- Automated Approach: RFE (Recursive feature elimination) with number of features = 15.
- Drop insignificant variables using manual approach based on VIFs and p- values.
- The final tally of variables with their respective values
  - Significant p-values near to zero
  - VIFs < 4

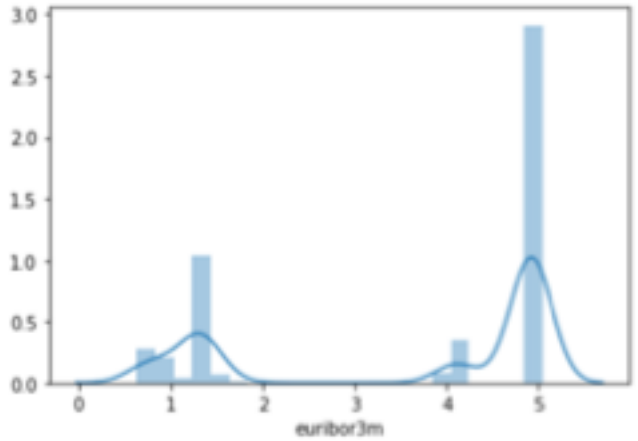
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5779	0.064	-40.418	0.000	-2.703	-2.453
job_retired	0.4339	0.081	5.370	0.000	0.276	0.592
job_student	0.4654	0.102	4.561	0.000	0.265	0.665
contact_telephone	-0.1845	0.059	-3.123	0.002	-0.300	-0.069
month_mar	0.7987	0.113	7.085	0.000	0.578	1.020
month_may	-0.9323	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

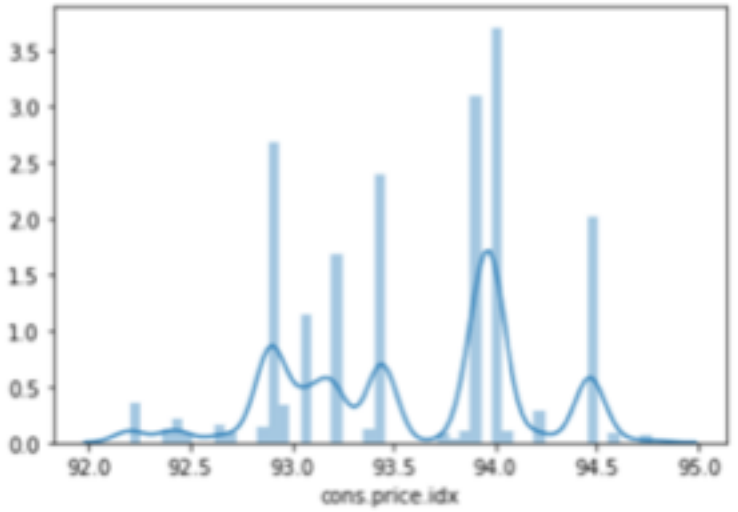
Relationship of predictor variable with response variable

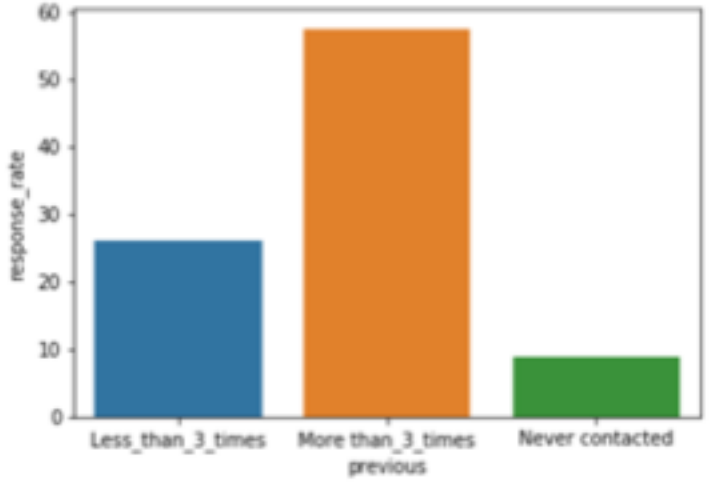






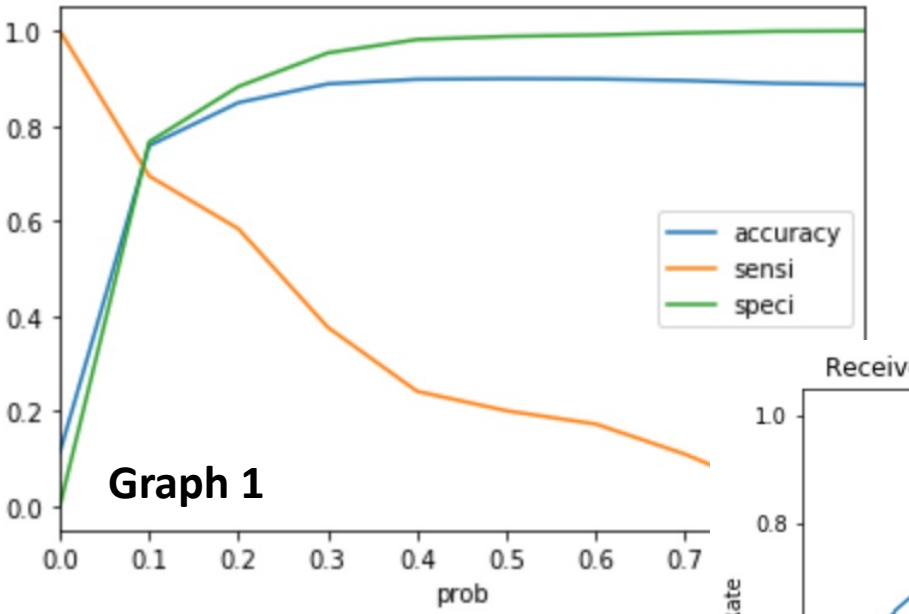


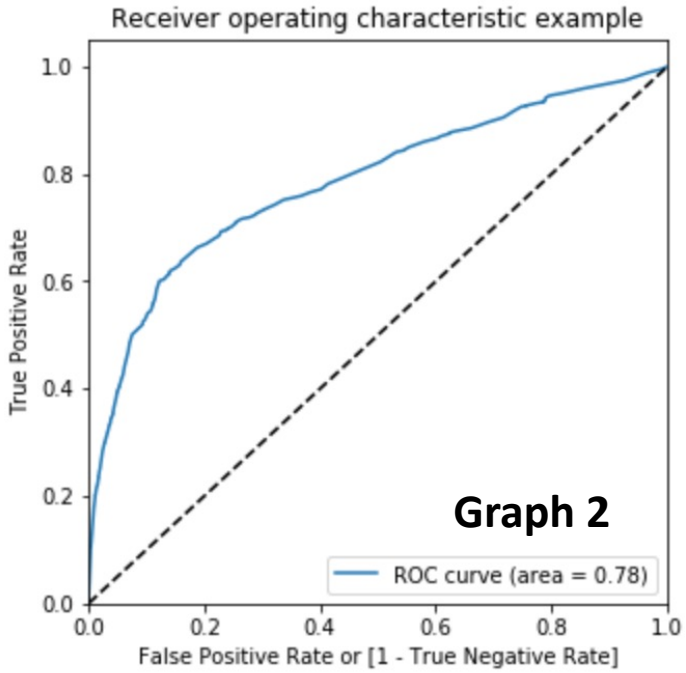




Model Building - Logistic Regression Model with RFE (Optimization)

- ROC Curve demonstrates tradeoff between sensitivity and specificity (Graph 2)
  - 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 (Graph 1)





Model Building - Logistic Regression Model with RFE (Evaluation)

Attribute	Train	Test
Accuracy	~76%	~75%
Sensitivity	~69%	~68%
Specificity	~76%	~76%
Precision	~27%	~26%
Recall	~69%	~68%

Confusion Matrix

Train Data

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

Test Data

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

Applying Model for Business Objective – **Test Data**

- To meet business objective, we need to achieve 80% response at a minimal cost
- From ”Table 1”, it is evident that 80% response can be achieved by targeting 60% (6th decile) of the total client base (12,357), which is ~7,414
- This can be used for cost optimization, and depending on cost/call we can determine the team-size of telemarketing campaign.
- Avg. call-duration per person for targeting top 80% prospect is ~255.36 seconds

Table 1

	decile	total	actual_response	cumresp	gain	cumlift
9	1	1235	582	582	42.112880	4.211288
8	2	1213	252	834	60.347323	3.017366
7	3	1239	114	948	68.596237	2.286541
6	4	1186	65	1013	73.299566	1.832489
5	5	1295	80	1093	79.088278	1.581766
4	6	1055	73	1166	84.370478	1.406175
3	7	1298	53	1219	88.205499	1.260079
2	8	1316	88	1307	94.573082	1.182164
1	9	878	31	1338	96.816208	1.075736
0	10	1642	44	1382	100.000000	1.000000

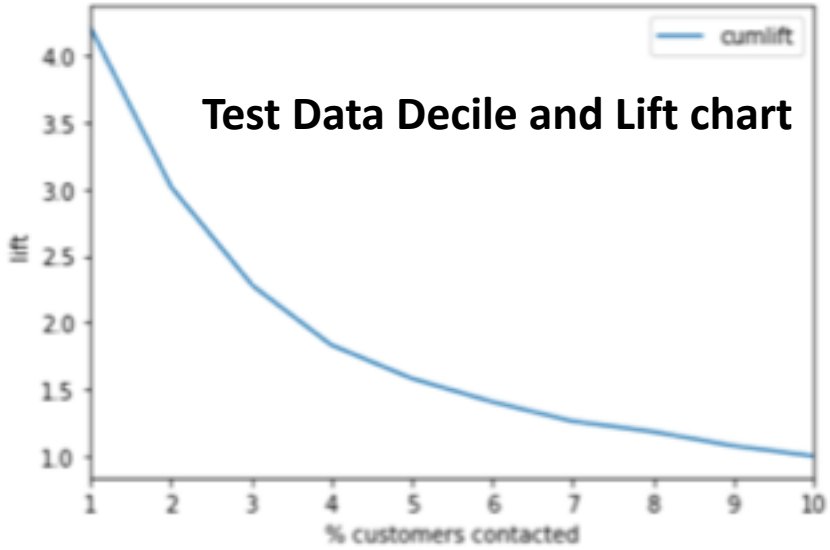
Applying Model for Business Objective – **Entire Data**

- To meet business objective, we need to achieve 80% response at a minimal cost
- From ”Table 2”, it is evident that 80% response can be achieved by targeting 50% (5th decile) 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.

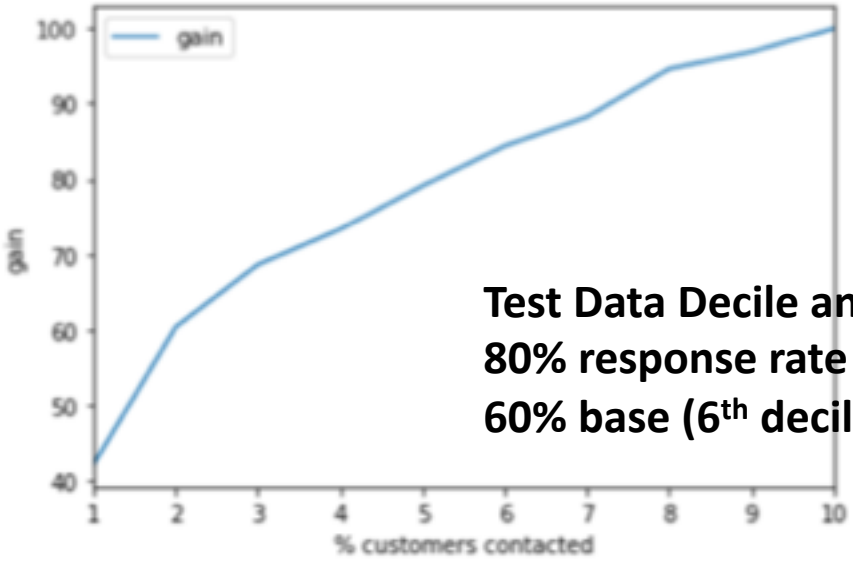
Table 2

	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

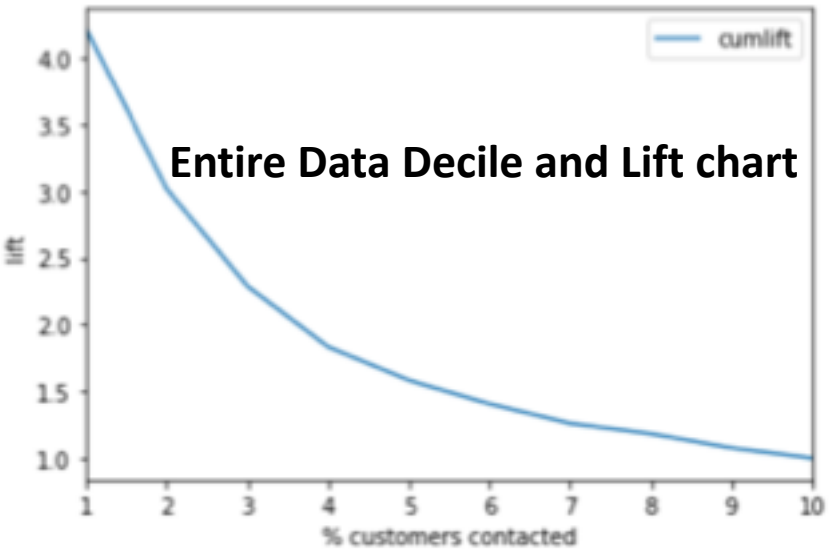
Model Performance – Lift Charts



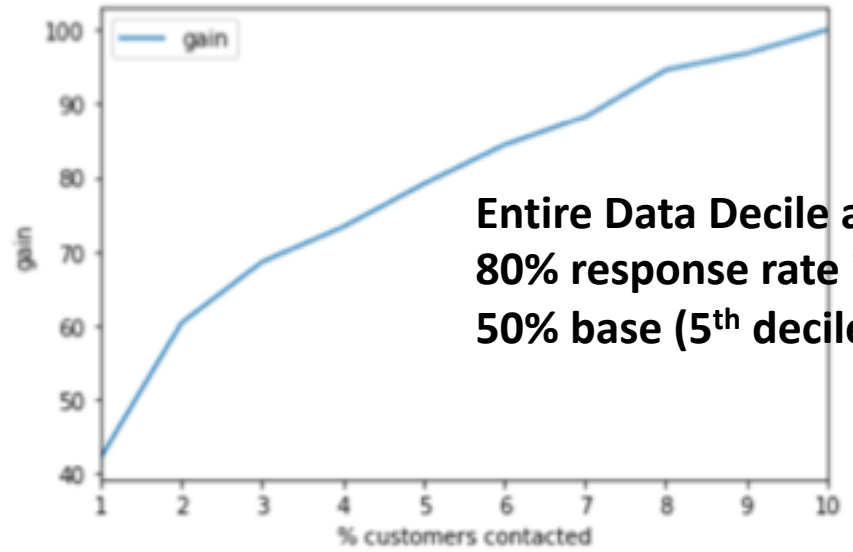
Test Data Decile and Lift chart



Test Data Decile and Gain chart
80% response rate is achieved at 60% base (6th decile)



Entire Data Decile and Lift chart



Entire Data Decile and Gain chart
80% response rate is achieved at 50% base (5th decile)

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 for both Test Data and Entire Data
- Test Data
  - **Cost** = 1* (60% of 12,357) = **7,414**
  - Since, 60% of base is required to be contacted to achieve 80%
- 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 our business objective of acquiring 80% of total responders at the minimum possible cost; we will need to target 50% of the total customer base for entire dataset. In case of test data, it is 60% of the test dataset.
- Significant variables identified by model:

job_retired	**month_mar**	**poutcome_success**
job_student	**month_may**	**cons.price.idx**
contact_telephone	**previous_Nevercontacted**	**euribor3m**

- Through our model we have improved 50% efficiency; as instead of calling the entire customer base, we can now achieve our objective by targeting just 50% of the entire customer base.