ACQUISITION ANALYTICS CASE STUDY

Suman Sarkar

ABSTRACT

Bank wanted to predict the probability of a response from each prospect and target the ones most likely to respond to the next telemarketing campaign.

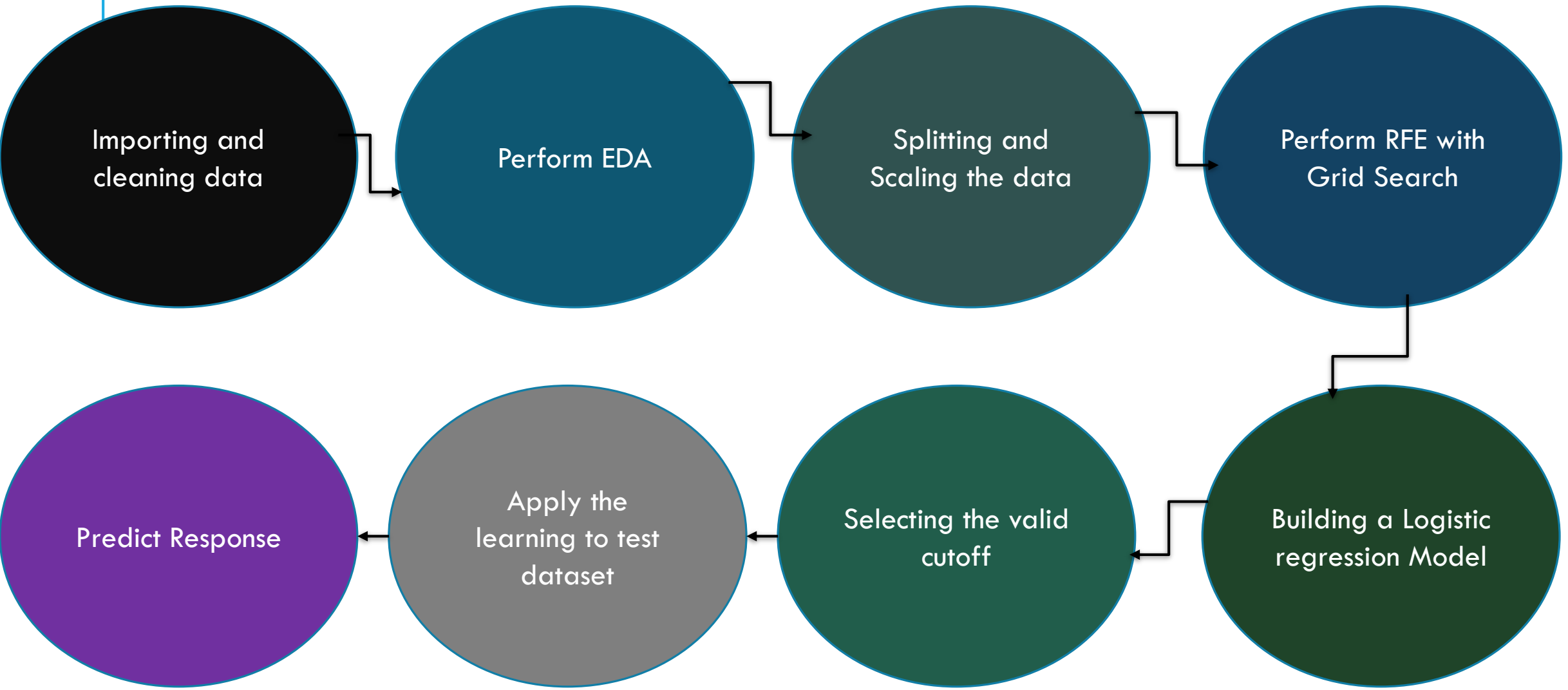
Goals:

Identify relevant predictor variables for a response using EDA.

Build predictive models and choose the best one.

Sort the prospects in order of decreasing probability of response (predicted by the best model) and target the top X% (or top Y deciles), where X would be determined by your business objective (e.g., maximising the overall response rate/number of responders at a fixed marketing cost).

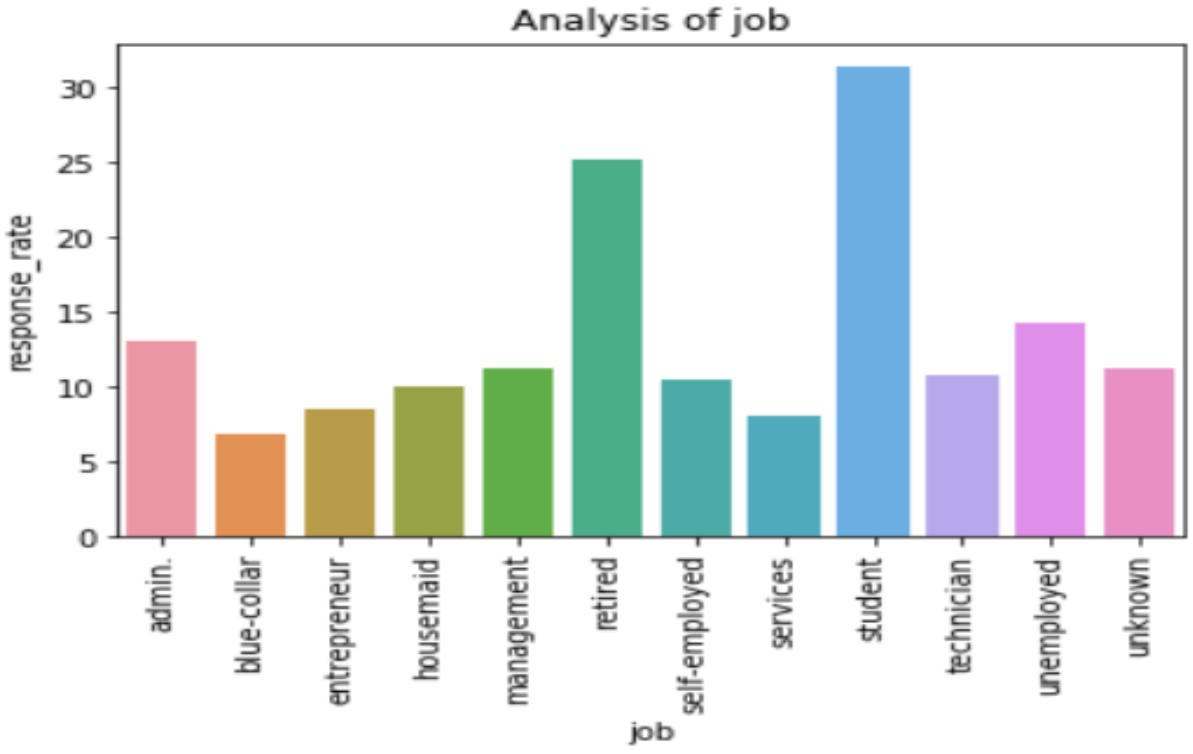
PROBLEM SOLVING METHODOLOGY

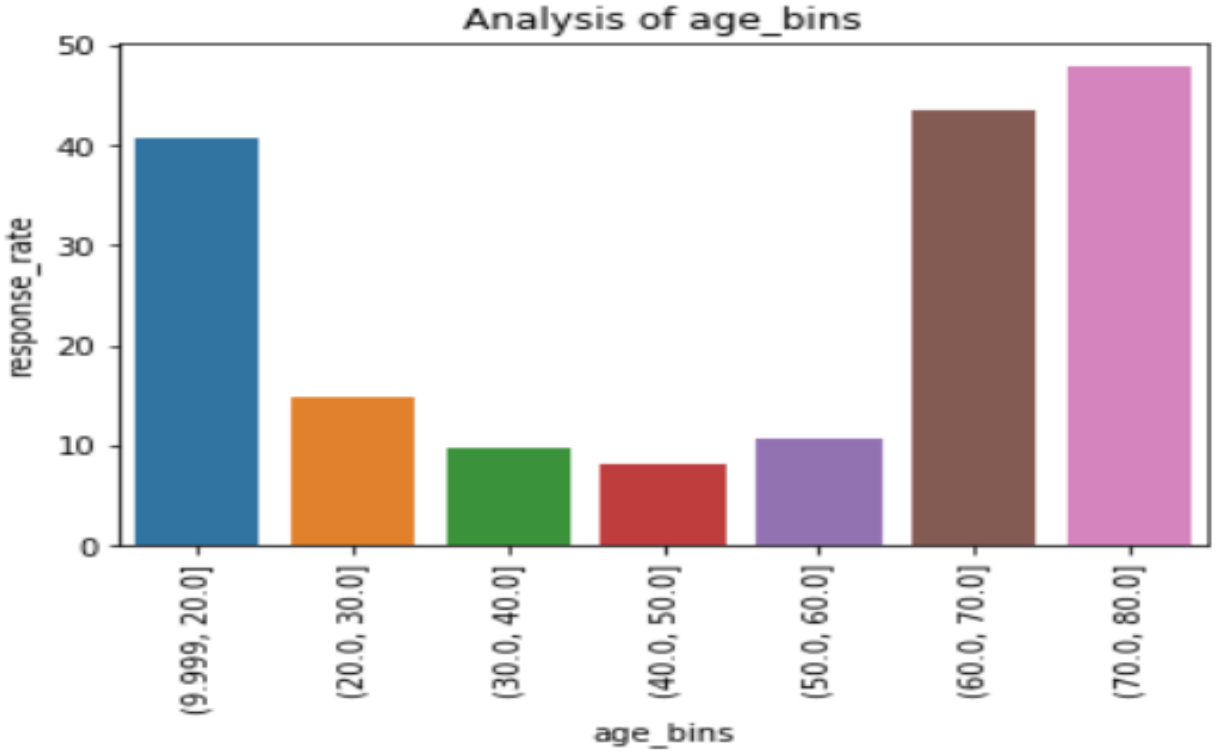


EDA & INSIGHT

We have selected each column and tried to find the relationship with the response variable.

For Example from the below graph we can identify that young and aged persons are more interested in the campaign and also we can cross verify that from the job graph that retired and students are more interested in this campaign. For students we can conclude that most probably there father are going to invest.





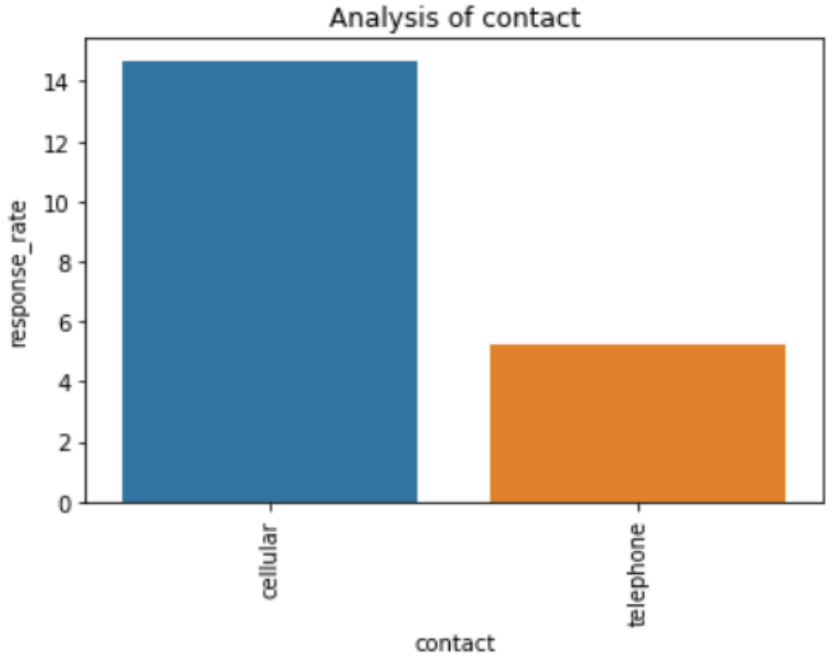
EDA & INSIGHT

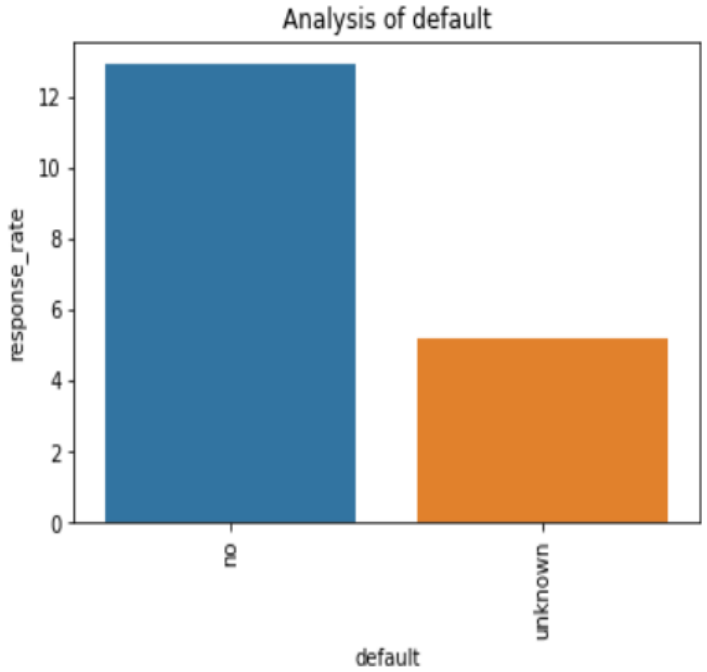
Few more insights :

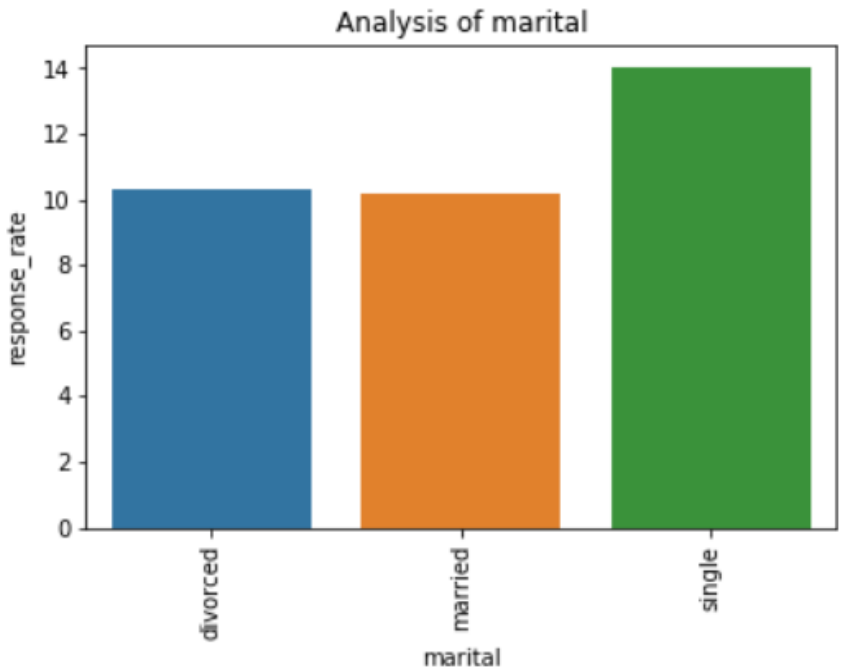
1 . Cellular people are more interested in this bank campaign.

2 . Non defaulted customers are the ideal candidates.

3 . Single people can be targeted for the next campaign.







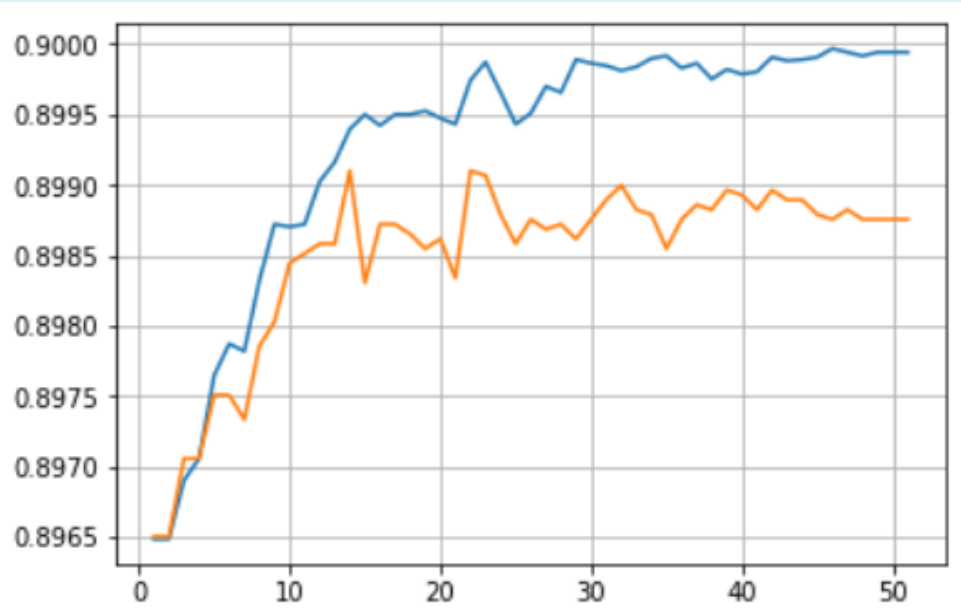
DATA ENCODING & FEATURE SELECTION

For logistic regression we are having a lot of categorical columns so we are going to convert them into numerical using one hot encoding

	job	marital	education	default	housing	loan	contact	month	day_of_week	pdays	previous	poutcome
0	housemaid	married	Primary_Education	no	no	no	telephone	may	mon	First_time_contacted	Never contacted	nonexistent
1	services	married	Secondary_Education	unknown	no	no	telephone	may	mon	First_time_contacted	Never contacted	nonexistent
2	services	married	Secondary_Education	no	yes	no	telephone	may	mon	First_time_contacted	Never contacted	nonexistent
3	admin.	married	Primary_Education	no	no	no	telephone	may	mon	First_time_contacted	Never contacted	nonexistent
4	services	married	Secondary_Education	no	no	yes	telephone	may	mon	First_time_contacted	Never contacted	nonexistent

Also we are going to perform grid search on top of RFE to get the optimum number of variables

```
{'n_features_to_select': 14}
```

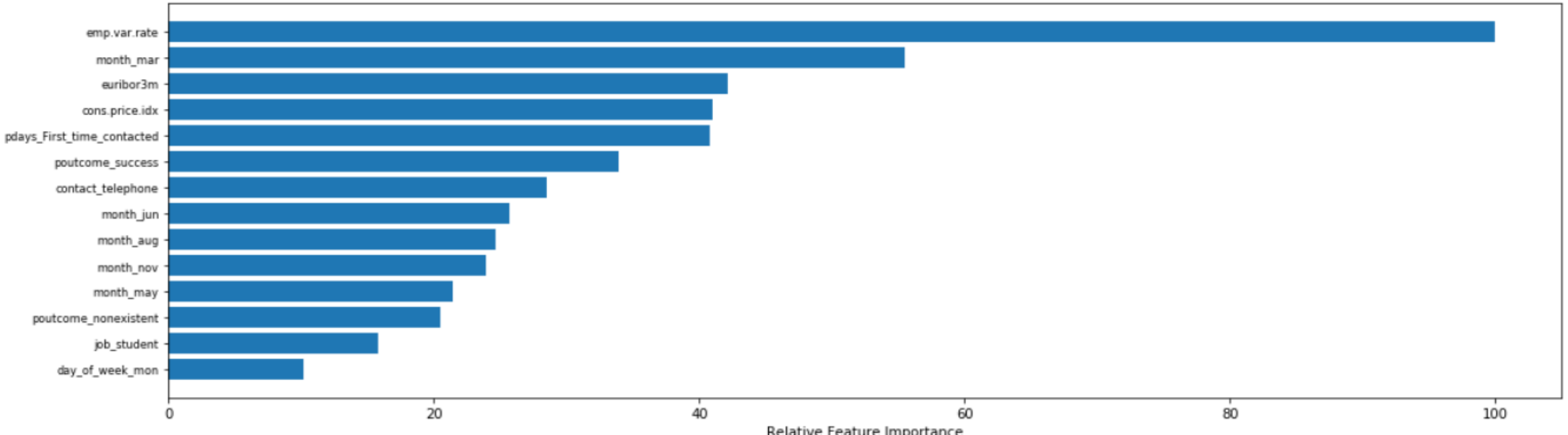


LOGISTIC MODEL & FEATURE IMPORTANCE

Fitting Logistic Regression model on the data set with class weight balance to handle class imbalance

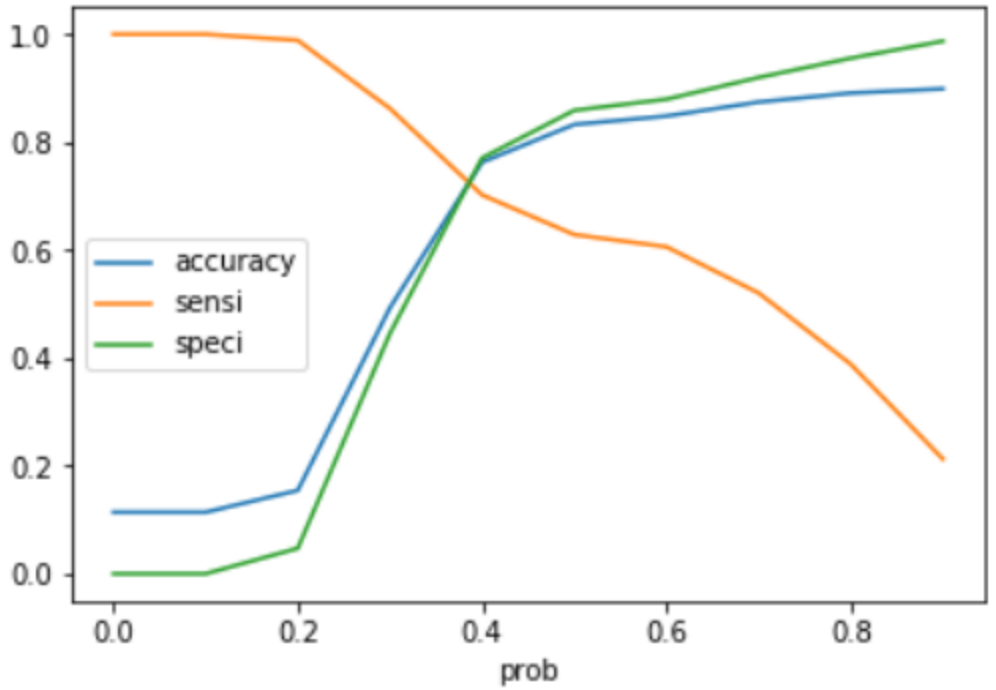
```
]: l_model = LogisticRegression(class_weight='balanced')
   X_train = X_train[rfe_selected_columns]
   l_model.fit(X_train,y_train)
   train_prob = l_model.predict_proba(X_train)[:,1]
```

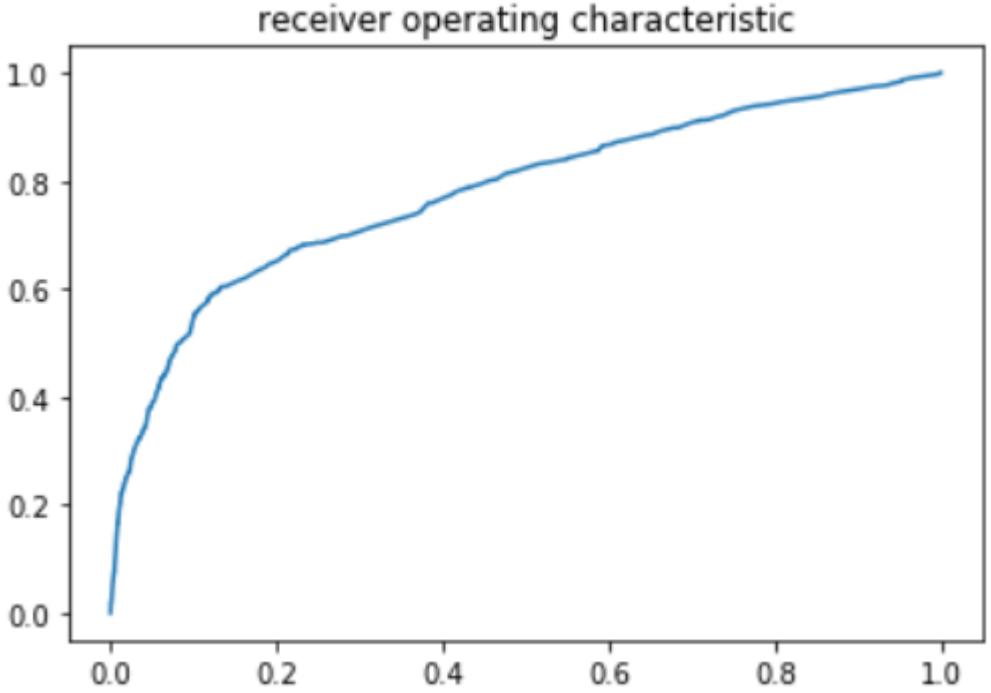
After fitting we can see the feature importance.



OPTIMAL CUT OFF & AUC

We can see that the optimal cutoff is 0.4 and the other graph is showing AUC graph





VALIDATION METRICS

The most important part is to check all the validation metrics which can tell you the performance of your Logistic model.

Metrics

- Accuracy
- Confusion Metrix
- Classification report
- Log Loss

```
log_loss(test_df_new.actual,test_df_new.pred_class)
```

```
8.340677359767966
```

```
accuracy_score(test_df_new.actual,test_df_new.pred_class)
```

```
0.7585174395079712
```

```
print(classification_report(test_df_new.actual,test_df_new.pred_class))
```

```
              precision    recall  f1-score   support

           0       0.95      0.77      0.85     11004
           1       0.27      0.68      0.38      1353

    accuracy                           0.76     12357
   macro avg       0.61      0.72      0.62     12357
weighted avg       0.88      0.76      0.80     12357
```

```
confusion_matrix(test_df_new.actual,test_df_new.pred_class)
```

```
array([[8451, 2553],
       [ 431,  922]], dtype=int64)
```

```
roc_auc_score(test_df_new.actual,test_df_new.pred_class)
```

```
0.7247210447964497
```

```
roc_auc_score(test_df_new.actual,test_df_new.pred_class)
```

```
0.7247210447964497
```

IMPORTANT COLUMN

Below Creating a data frame with the variables prospect ID, actual response, predicted response, predicted probability of response, duration of the call in seconds and cost of the call

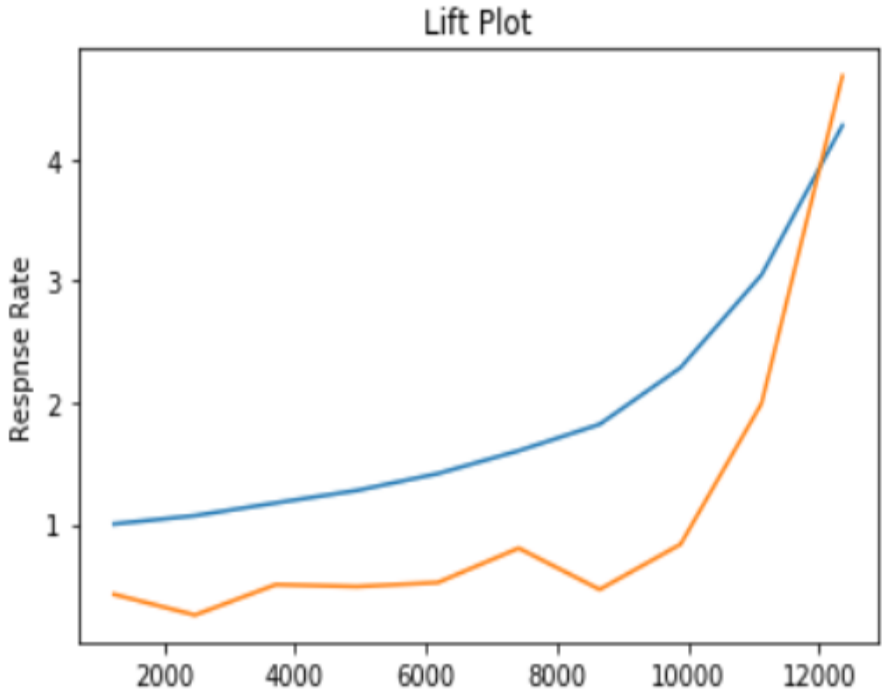
	predicted_probability	actual_response	predicted_response	prospect_id	duration	cost_of_the_call
0	0.743247	1	1	37771	712	11.866667
1	0.211843	0	0	2371	147	2.450000
2	0.172266	0	0	2097	832	13.866667
3	0.358745	0	0	24834	214	3.566667
4	0.211740	0	0	1720	646	10.766667

DECILE,GAIN & LIFT

Below table is showing the corresponding metric for all the decile like gain and Lift, Also we can verify our model using lift chart where we can see the performance after and before applying model.

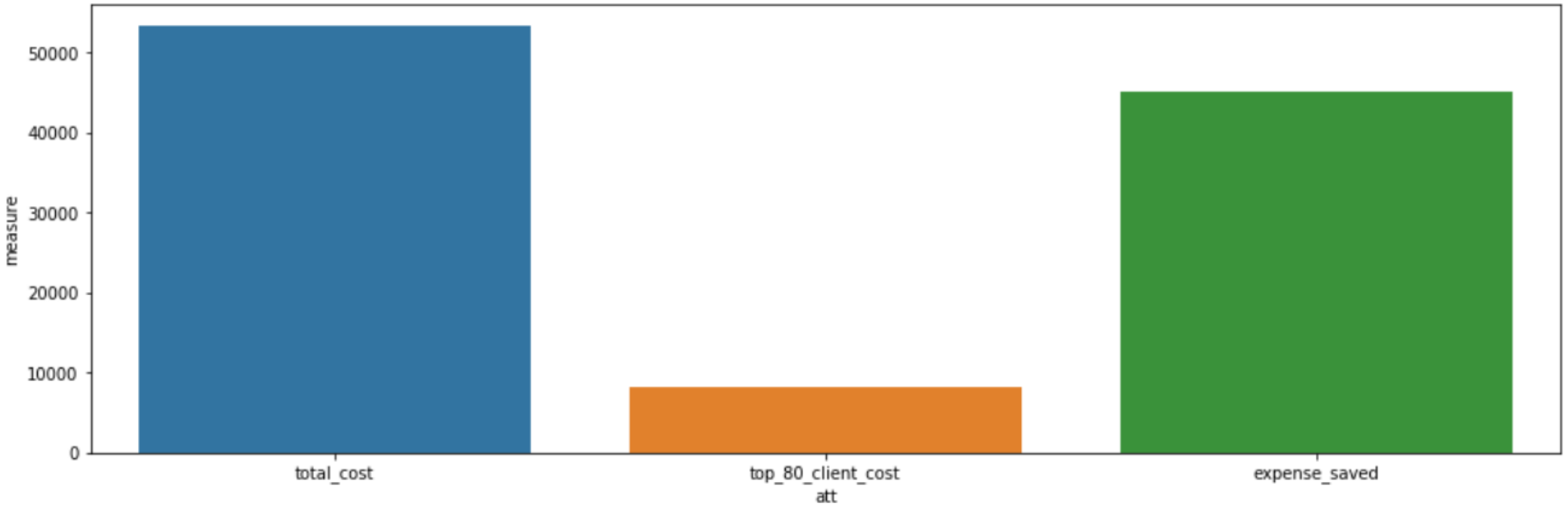
:

	decile	total	actual_response	cumresp	gain	cumlift	base_line
9	1	1235	579	579	42.793792	4.279379	4.685603
8	2	1190	246	825	60.975610	3.048780	1.990774
7	3	1235	103	928	68.588322	2.286277	0.833536
6	4	1109	57	985	72.801183	1.820030	0.461277
5	5	1348	99	1084	80.118256	1.602365	0.801165
4	6	1264	64	1148	84.848485	1.414141	0.517925
3	7	1122	60	1208	89.283075	1.275472	0.485555
2	8	1291	62	1270	93.865484	1.173319	0.501740
1	9	1027	31	1301	96.156689	1.068408	0.250870
0	10	1536	52	1353	100.000000	1.000000	0.420814



COST OF ACQUISION

Below graph is showing if we can target only 50% of customer we can save a lot of money on the telemarketing campaign



THANK YOU