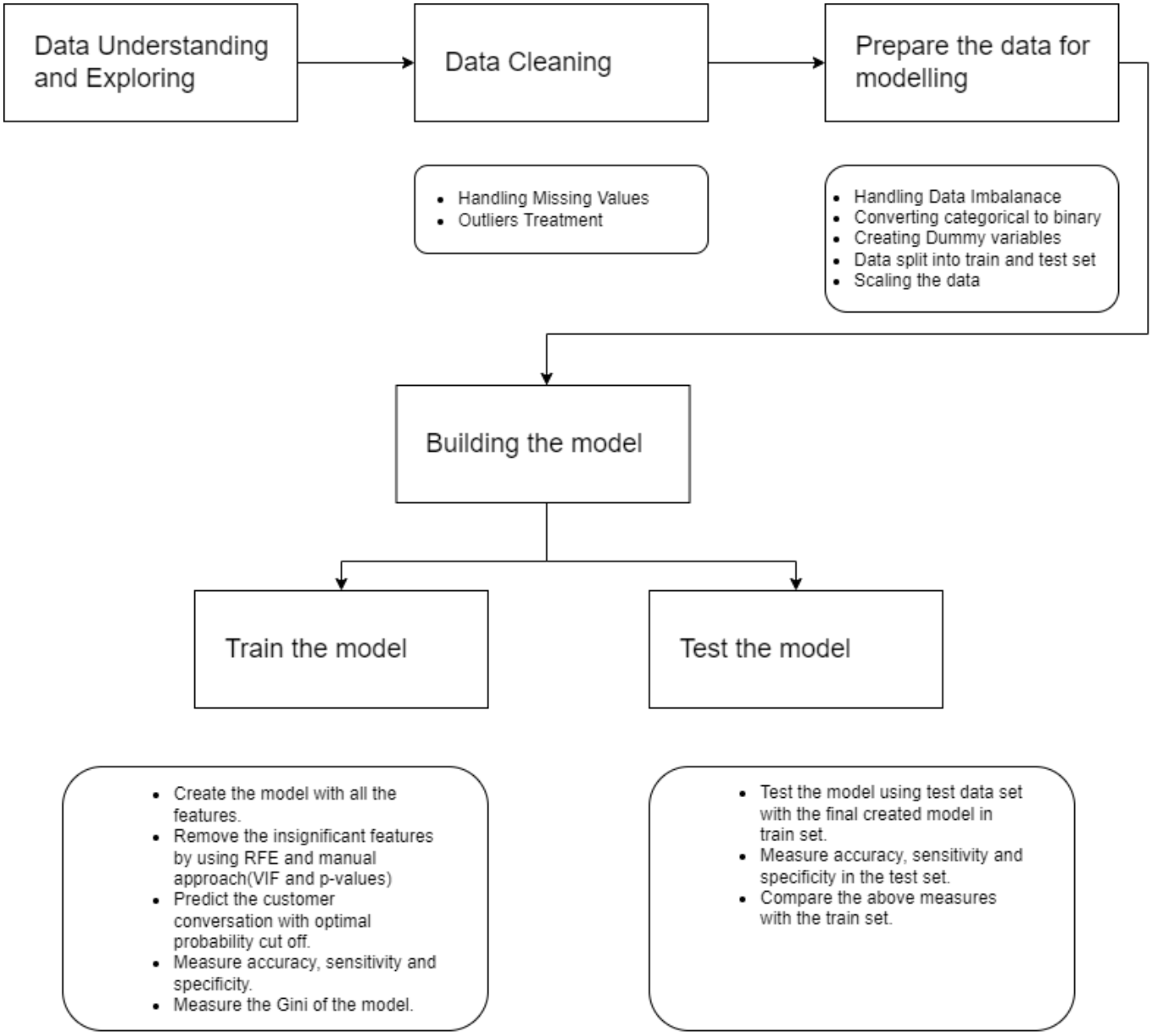
Summary Report

Problem statement:

Identify the set of leads of X Education so that the lead conversion rate should go up and the sales team of the company focus more on communication with the potential leads rather than making calls to every customer.

Analysis approach:

Flow diagram of step by step approach.



Model outcome:

Optimum probability cut off: 0.35

Lead score:

The lead score is calculated based on the probability of customer being converted. According to the final model, if the lead score is more than 35, then the customer is likely to be converted. Higher the lead score, higher the chance the lead/customer being converted.

Average Lead Score of the converted leads = 68

Average Lead Score for the not converted leads = 15

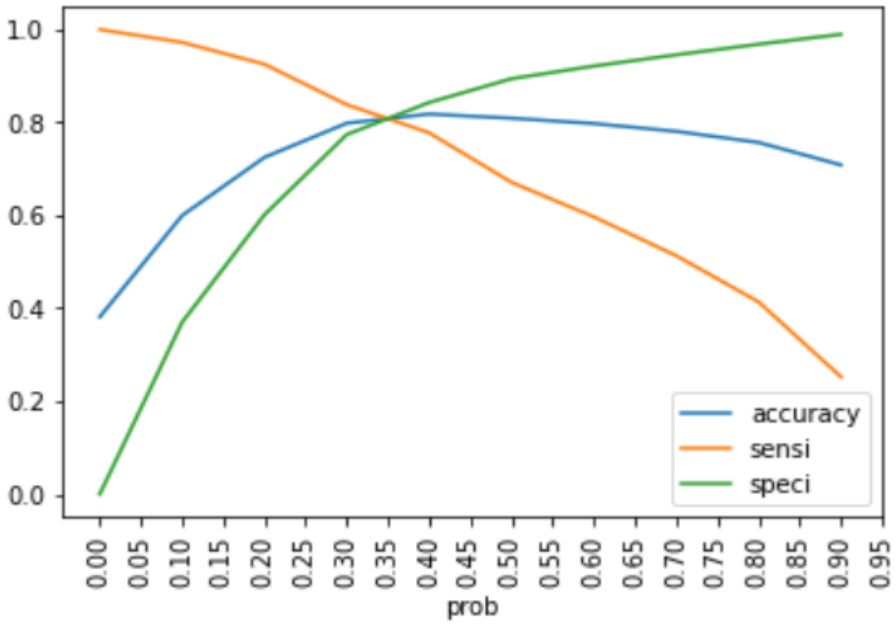
Features of the final model:

Below are the significant features of the final model arranged in the descending order of the impact on the basis of their coefficients.

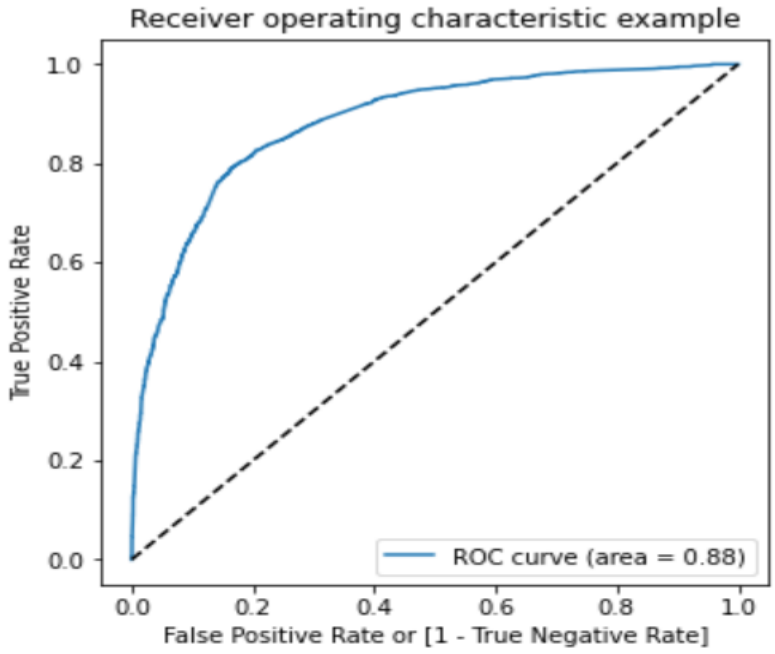
Features	Coefficient	Impacts
Lead Origin_Lead Add Form	4.4017	Positive
Current Occupation_Working Professional	2.6908	Positive
Last Activity_SMS Sent	1.9097	Positive
Lead Source_Olark Chat	1.2101	Positive
Time Spent	1.0574	Positive
Last Activity_Email Opened	0.8166	Positive
Last Activity_Others	0.6463	Positive
TotalVisits	0.1944	Positive
Free Copy	-0.3186	Negative
Last Activity_Olark Chat Conversation	-0.6859	Negative
Lead Origin_Landing Page Submission	-1.0199	Negative
const	-1.0565	Negative
Specialization_Not Specified	-1.138	Negative
Lead Source_Reference	-1.1764	Negative
Last Activity_Email Bounced	-1.1921	Negative

Important measures of the model:

Measures	Train set	Test set
Accuracy	81.19%	80.08%
Sensitivity	67.11%	67.11%
Specificity	89.40%	89.40%



Gini of the model: 0.88



Conclusion:

Since the difference between the test and train data parameters is 0.5 to 1% , we can conclude that our model doesn't over-fit the training model.
The model has good accuracy, sensitivity and specificity. Overall, the model performs well in the test set, what it had learnt from the train set.

Business recommendation for higher conversion rate:

Highly likely to be converted leads:

1. Lead score more than 68.
2. Working Professionals, Total time spent on website(>12 hrs), Total number of visits, lead source.
3. Lead source Google, Direct Traffic, Organic Search, Welingak Website and Reference.
4. Last activity of the customers is any of SMS sent, Email Opened.

Very less likely to be converted leads:

1. Customers opted for 'Do not email' option.
2. Lead score less than 15.
3. Total time spent on website less than 5 hrs.
4. Last activity of the customers is any of 'Olark chat conversation', 'page visited on the website', 'Email bounced'.

Learnings gathered:

1. Data preparation for modelling

a) It is important to treat missing values and also get rid of the outliers present in the data.

b) If there is huge data imbalance in the features, then it is better to either drop that particular feature or remove the imbalance by merging the imbalanced values to other values.

c) All the features should be in the same scale.

2. Model building

a) There shouldn't be any multicollinearity between the variables.

b) Find the optimal probability cut off to get a balance between Sensitivity and Specificity with good Accuracy.

c) The model should perform well in the test set in terms of Sensitivity, Specificity and Accuracy.