Summary Report

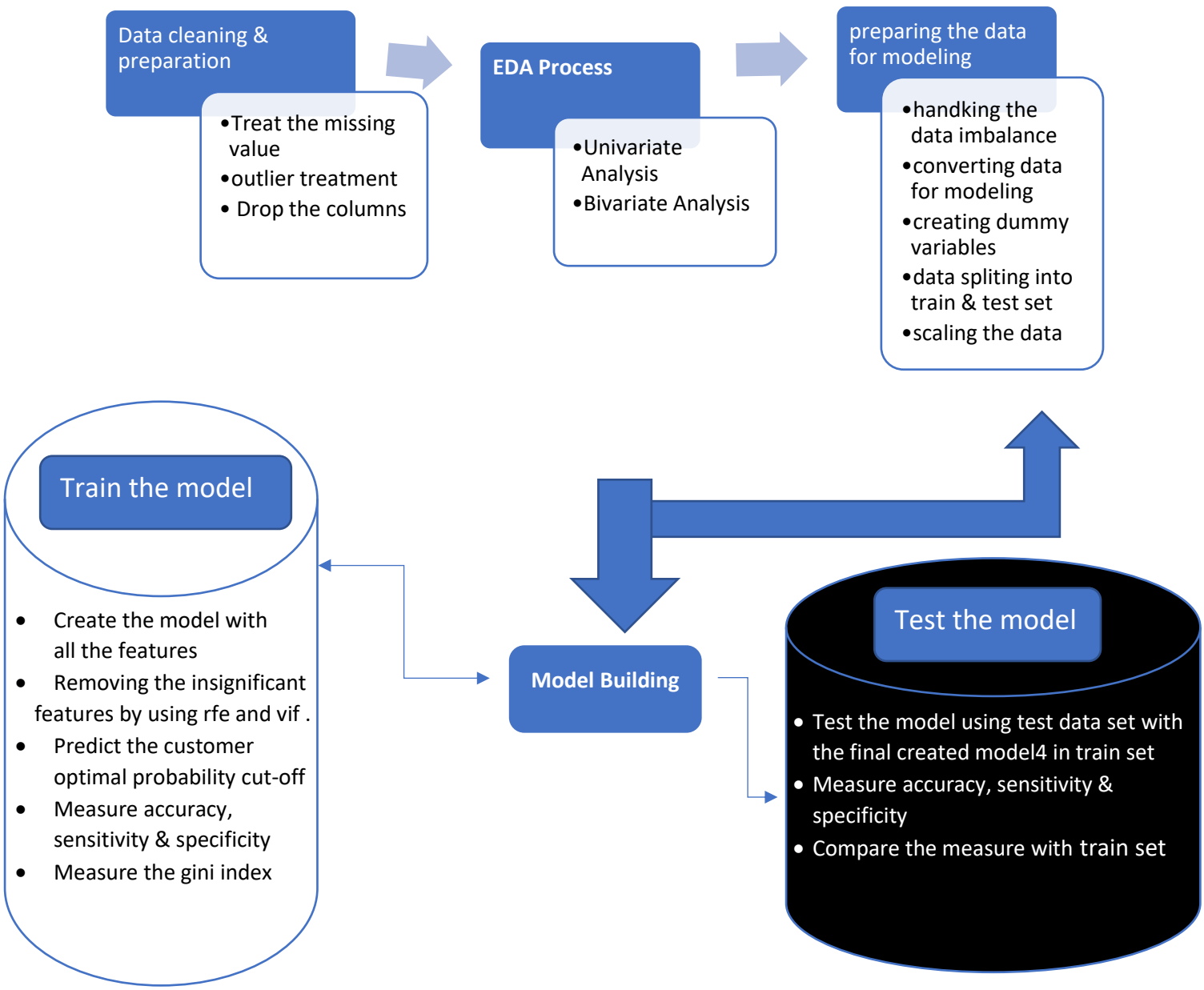
Problem statement:

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 chart of step by step approach.

Train the model



Cleaning data:

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not rovided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

EDA:

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values look good and no outliers were found.

Dummy Variables:

The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the Standard Scaler.

Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

Model Building:

Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

Sensitivity & Specificity :

This method was also used to recheck, cut off with Sensitivity 79.31% and Specificity 80.04% on the test data frame.

Model Output:

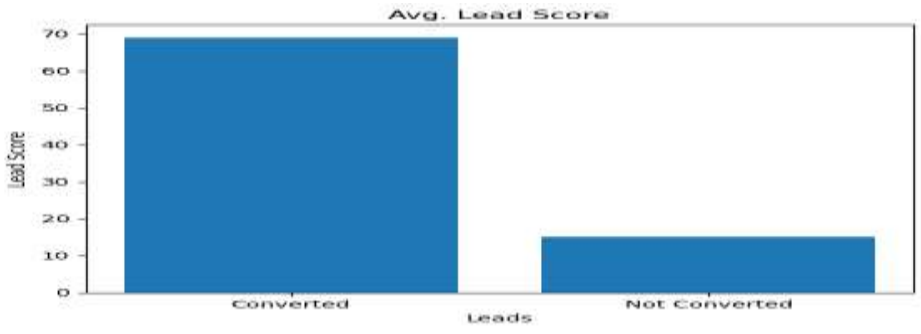
Optimum probability cut off: - 0.35

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.

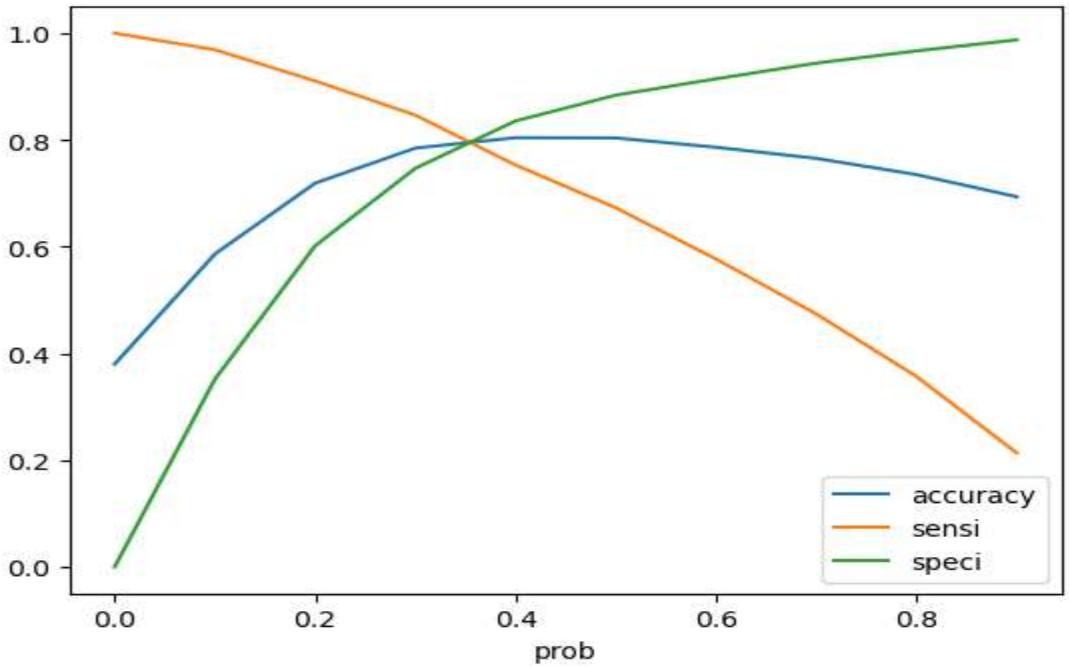
Average Lead Score of the converted leads = 69

Average Lead Score for the not converted leads = 20

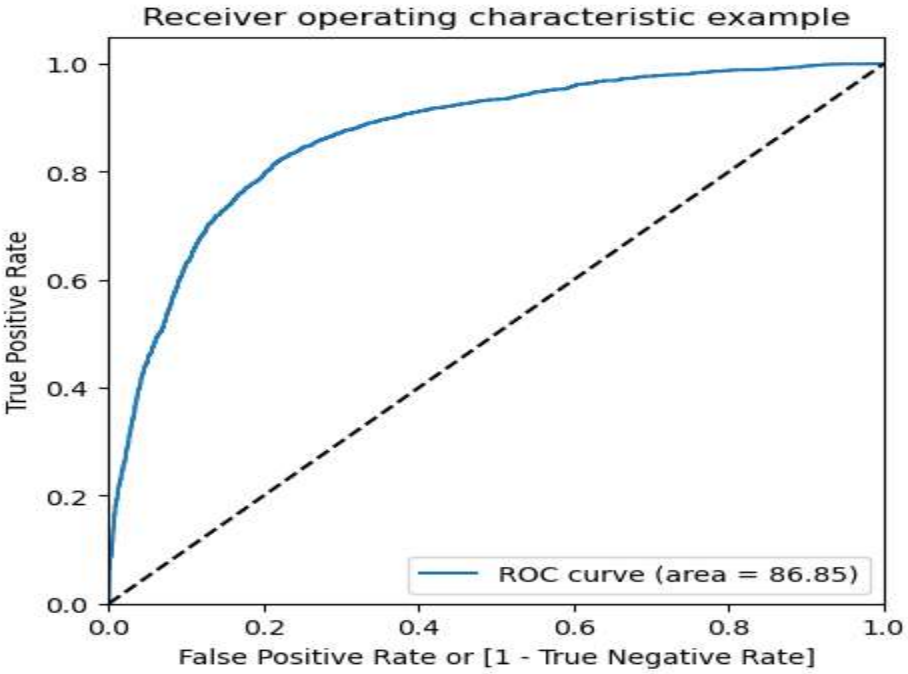


Important Measure of the model:

Measures	Train set	Test set
accuracy	79.77%	79.60%
specificity	80.04 %	78.83%
sensitivity	79.31%	80.90%



Gini index of the model: 86.85 %



Conclusion: - 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: -

- Lead score more than 68.
- Total time spent on website more than 12 hrs.
- Lead source Welingak Website and Reference.

Very less likely to be converted leads:

- Customers opted for ‘Do not email’ option.
- Lead score less than 15.
- Total time spent on website less than 5 hrs.
- Lead source Direct Traffic, Referral Sites, Organic Search and Google.
- Last activity of the customers is any of ‘Olark chat conversation’, ‘page visited on the website’, ‘Email bounced’, ‘Form submitted on website’, ‘Email link clicked’.

Learnings gathered: -

Data preparation for modelling:

- It is important to treat missing values and also get rid of the outliers present in the data.
- 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.
- All the features should be in the same scale.

Model building:

- There shouldn’t be any multicollinearity between the variables.
- Find the optimal probability cut off to get a balance between Sensitivity and Specificity with good Accuracy.
- The model should perform well in the test set in terms of Sensitivity, Specificity and Accuracy