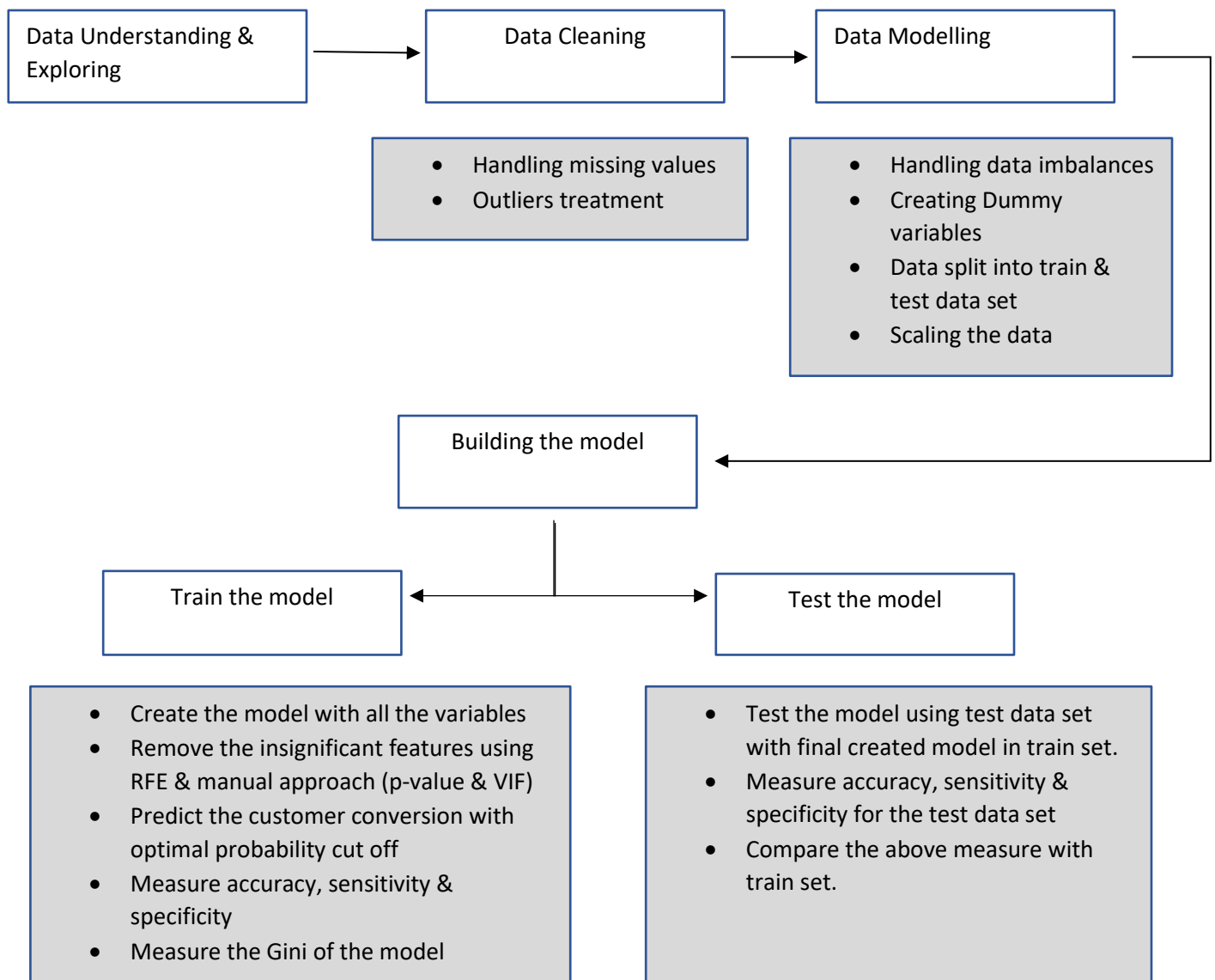# Summary Report

Problem Statement :-

Identify the sets of leads of X Education company so that the lead conversion rate should go up so the sales team of the company could focus more on communicating with the potential leads (Hot leads) rather than approaching each and every customer which in turn would help them in paying more attention to the customers that are more likely to purchase their product/course.

Analysis Approach :-

Data Understanding & Exploring	→	Data Cleaning	→	Data Modelling

Data Cleaning:
- Handling missing values
- Outliers treatment

Data Modelling:
- Handling data imbalances
- Creating Dummy variables
- Data split into train & test data set
- Scaling the data

Building the model

Train the model

Test the model

Train the model:
- Create the model with all the variables
- Remove the insignificant features using RFE & manual approach (p-value & VIF)
- Predict the customer conversion with optimal probability cut off
- Measure accuracy, sensitivity & specificity
- Measure the Gini of the model

Test the model:
- Test the model using test data set with final created model in train set.
- Measure accuracy, sensitivity & specificity for the test data set
- Compare the above measure with train set.

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 final model, if the lead score is <u>more than 35</u> then the customer is likely to be converted. Higher the lead score higher is the chance that the lead/customer being converted.

Average lead score of the converted leads = 71

Average lead score for the not converted leads = 13
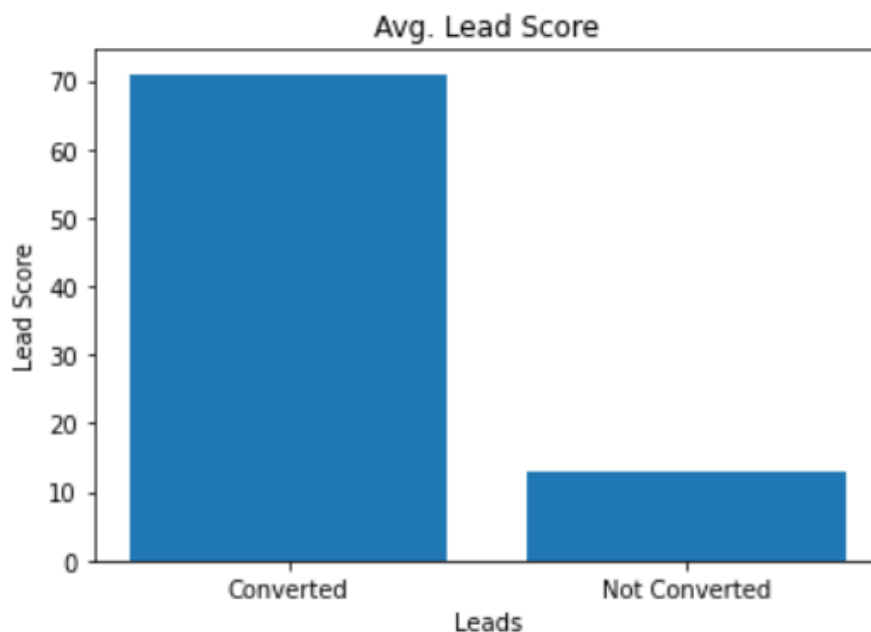

Features of the Final model: -

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

| Features	Coefficients	Impacts
Total Tome Spent on Website	4.5548	Positive
Lead Origin_Lead Add Form	4.1770	Positive
What is your current occupation_Working Professional	3.6230	Positive
Last Notable Activity_Had a Phone Conversation	3.3499	Positive
What is your current occupation_Others	2.8814	Positive
Last Notable Activity_Unreachable	1.6213	Positive

Lead Source_Olark Chat	1.4727	Positive
Last Notable Activity_SMS Sent	1.3846	Positive
What is your current occupation_Student	1.3101	Positive
TotalVisits	1.1152	Positive
What is your current occupation_Unemployed	1.1035	Positive
Last Activity_Converted to Lead	-1.1308	Negative
Last Activity_Olark Chat Conversation	-1.4674	Negative
Last Activity_Email Bounced	-2.0182	Negative

Important Measures of the Model: -

Measures	Train Set	Test Set
Accuracy	0.80	0.80
Sensitivity	0.82	0.81
Specificity	0.80	0.80
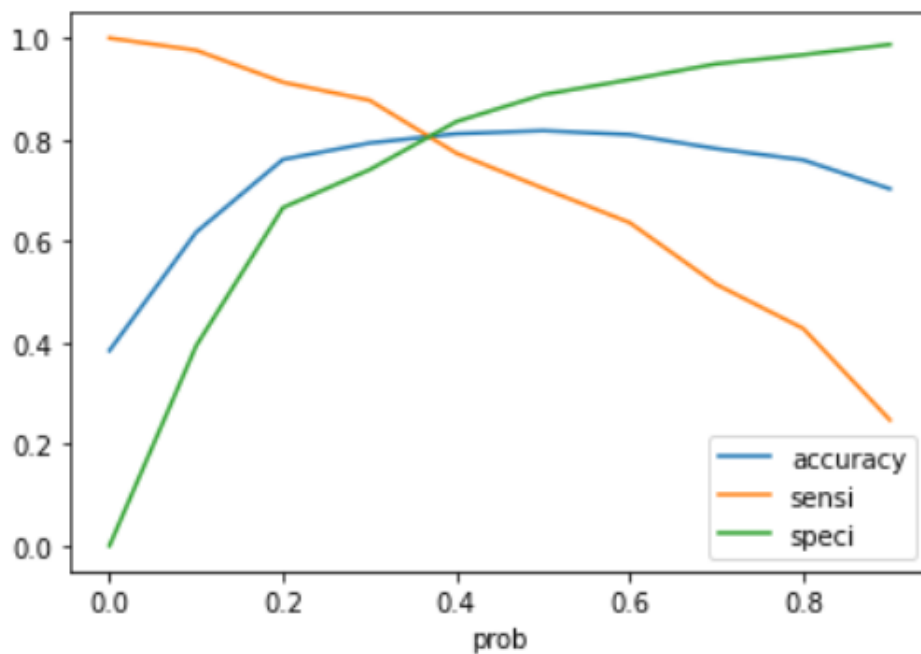
Gini of the model: - 0.89
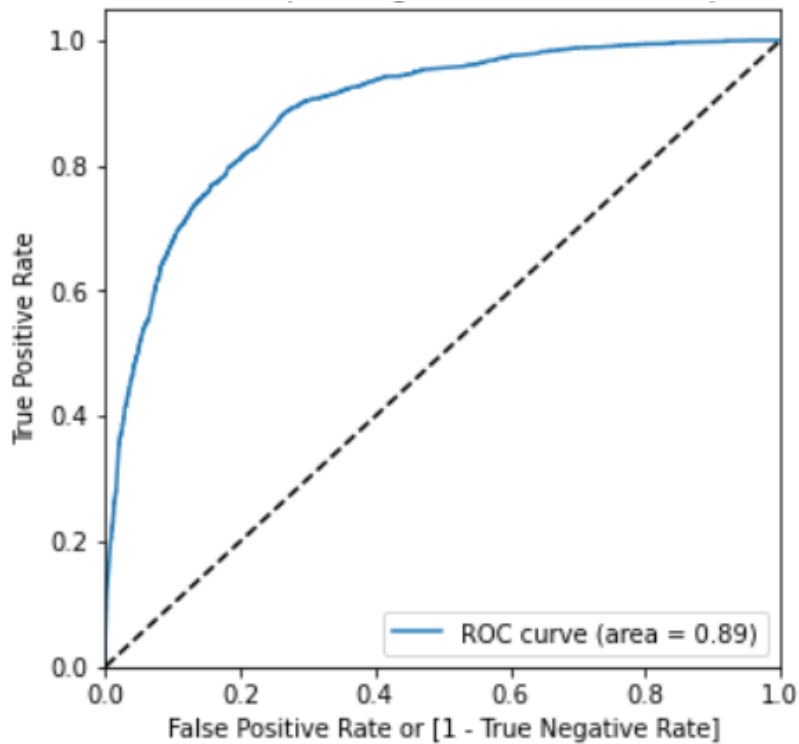
Conclusion: -

The model has good accuracy, specificity & sensitivity. Overall, the model performs well on the test data set from what it had learned from train data set.

Business recommendation for higher conversion rate:

Highly likely to be converted leads:

1. Lead score more than 71
2. Current occupation – Working professional, Student & Others
3. Lead source Olark Chat
4. Last notable Activity – SMS sent, Had a phone conversation

Less likely to be converted leads:

1. Lead score less than 13
2. Last Activity – Email bounced
3. Last Activity – Converted to lead, Olark chat conversation

Learnings gathered: -

1. **Data Preparation/Cleaning:**
    i.   It is important to treat missing values & also treat outliers as they may skew our data.
    ii. If there is huge data imbalance, then it is better to either drop that feature or merge the data/values with low percentage compared to others.
    iii. All the features should be adjusted to same scale.
2. Model Building: -
    i.   There shouldn't be any multicollinearity among the variables.
    ii. Find the optimal probability cut off to get a balance between Sensitivity & Specificity with good accuracy.