Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. When people fill up a form providing their email address or phone number, they are classified to be a lead. Although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Overall Approach:

1. Data reading, understanding and cleaning.
2. Data analysis(visualisation)
    - Univariate analysis
    - Bivariate analysis
3. Data preparation
4. Model building
5. Model evaluation

Data reading, understanding and cleaning: In first step we import data then we drop unwanted columns. Like Columns contain missing values more than 40%. Or columns with minimum variation in data levels. Then we impute missing values of categorical data by mode and of numerical data by median. If there will any outlier then we cap it to 99 percentile or to 1 percentile.

Data analysis: In second step we did univariate analysis on both numerical and categorical data and bivariate analysis on numerical and categorical data. For numerical univariate analysis we used box plot. For categorical univariate analysis we used count plot. Pair plot and bar plot were used for bivariate numerical analysis, bivariate categorical analysis respectively.

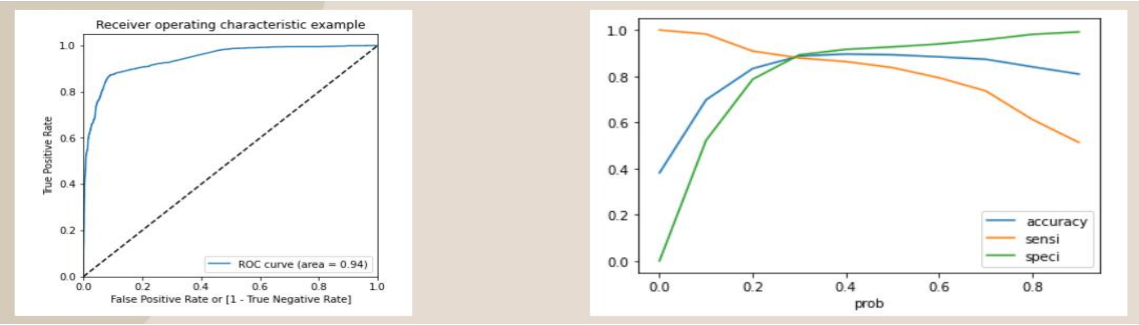
Data preparation: In this step we create dummy features for categorical data. By using dummy features, we can create binary data so model can understand it. After that we split data into train and test data. First, we build model on train data and then then evaluate it on test data. Data splitting ratio was 7:3. Then we rescale the data so data can look like symmetric in nature.

Model building: We build model by using logistic regression. We used RFE method for automatic feature selection. RFE method select top 15 features. After that we checked P-value and VIF of the model. Then we dropped columns with P-value greater than 0.05 and VIF greater than 5. And build final model. We select probability on the basis of optimal threshold that was 0.3.

Model evaluation: On the basis of optimal threshold probability, we build model. And we checked Recall on both, train and test data. Formula for checking recall is True Positive/(True Positive + False Negative).

Report:

ROC and threshold curve:



Confusion matrix:

Train data=

	FN	FP
TN	3572	430
TP	298	2168

Test data=

	FN	FP
TN	1483	194
TP	114	981

Recall:

Train data = 0.88

Test data = 0.89