Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary:

In this analysis we were given a task to find out who will become a **hot lead** (chances of conversion more). We were provided with the data set with multiple features like how much time user spent on the website, source of the lead, how many times he visited, what is the occupation, whether they like to be contacted via mail or phone etc.

Data Cleaning

So, we started with data cleaning where we removed features which had more than 40% of null or missing values as they can deviate our analysis.

Converting missing values into null values and handling them in 2 ways:

- Replacing them with highest occurring value (frequency > 60%), if not then
- Categorizing them as ‘Not Available’

EDA Process

Then we did EDA process and found that all the management specialisations were showing high number of leads so all were grouped into one category.

Then all the categories with lower frequency were categorized to ‘Other’ category.

Facebook in Lead Source was merged with Social media.

Data Preparation for modelling

Then we created dummy variables for modelling process where all the categorical values are converted into 0 or 1 based on the occurrence.

Numerical values were scaled using standard scaler to transform the data to comparable scales.

Data was split into train and test datasets in ratio of 70:30.

Modelling Process

We used automated feature selection using RFE where 15 top features were selected then we went forward used manual feature elimination technique by checking features having p-value which should be less than 5% and VIF value less than 2%.

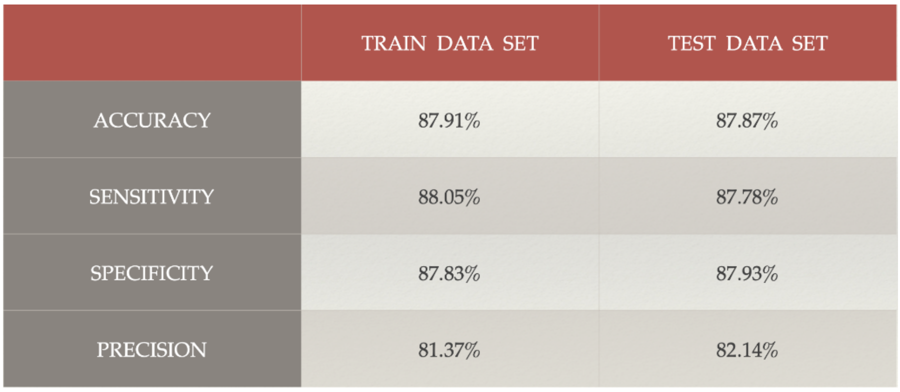
Then we found the optimal cut-off probability using the ROC curve and it came out to be 0.3.

Model Evaluation

Using these final values were predicted and a confusion matrix was created to calculate the accuracy, sensitivity, precision etc. The accuracy of around 87% and precision of around 81% was obtained on train dataset.

Prediction

Prediction was done on the test data set with the optimal cut off probability as 0.3 with accuracy around 87% and here is the comparison of the metrics on train and test dataset.



	TRAIN DATA SET	TEST DATA SET
ACCURACY	87.91%	87.87%
SENSITIVITY	88.05%	87.78%
SPECIFICITY	87.83%	87.93%
PRECISION	81.37%	82.14%