**Summary report:**

***PROBLEM STATEMENT:***

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. 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'.

Data Set:

1. *'Leads.csv'* leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
2. 'Leads Data Dictionary.xlsx' is data dictionary which describes each attribute in *Leads.csv* dataset.

***Analysis Approach:***

- Firstly Importing the Data (Leads.csv) file.
- The next step is to understand the data set. (Understanding the shape of data set, not null columns etc.)
- After Understanding the data, the next step is to handle the data as shown below.
    - Converting the values "Select" to nan values
    - Dropped columns with high missing values . Dropped columns "'How did you hear about X Education' ,'Lead Profile', 'Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score' since these have >30% missing values and we couldn't find anything suitable to handle the missing values
    - Handled Missing Values
    - Replaced nan with "Not Sure" in 'Lead Quality' column
    - Replaced nan with "Mumbai" in 'City' column
    - Replaced nan with "Others" in 'Specification' column
    - Replaced nan with "Other issues" in 'Tags' column
    - Replaced nan with "Better Career Prospects" in 'What matters most to you in choosing a course' column
    - Replaced nan with "Unemployed" in 'What is your current occupation' column
    - Replaced nan with "India" in 'Country' column

- Replaced nan with median in ‘Page Views Per Visit’ column
- Replaced nan with “Email Opened” in ‘Last Activity’
- Replaced nan with “Google” in ‘Lead Source’
- Dropped columns 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content‘ and 'I agree to pay the amount through cheque‘ since these columns has all values as No and it does not help in our analysis

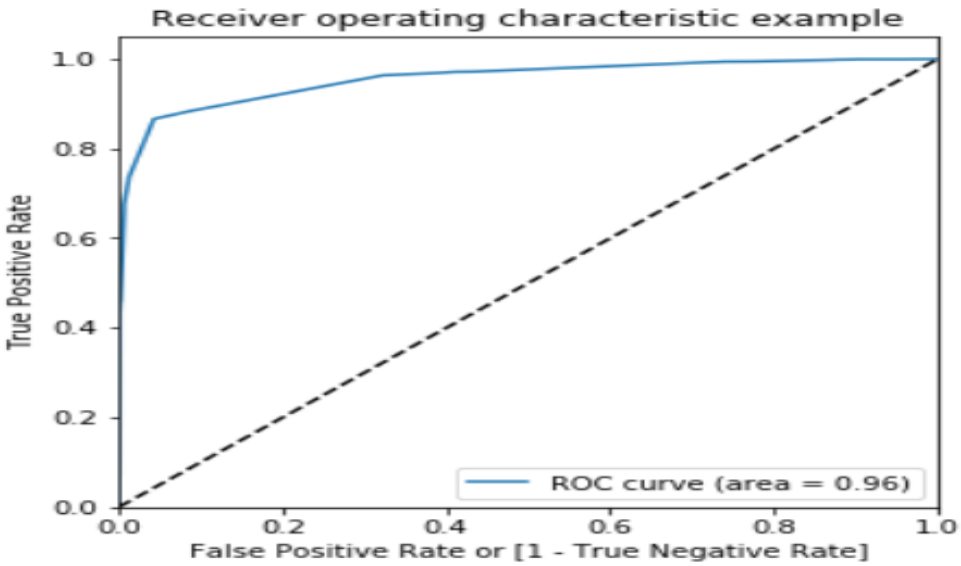
- Scaling the data set
  - Generated Dummy Variables for all categorical columns
  - Scaled data using StandardScaler for Numerical variables
- Modelling
  - Split the data into Test and Train
  - Apply Logistic Regression model
  - Feature Selection using RFE
    - Identified top 15 variables which are of significant importance
    - We will drop variable one by one so that P value is close to zero and model again with left out features.
    - Once the P value is close to zero(<0.05), we will check variable’s VIF which is should be less than 5, otherwise drop the variables and model again with left out Features.
  - Arriving to final model with 11 variables.

| Variable | Coef. |
|---|---|
| Tags_Closed by Horizzon | 8.22 |
| Tags_Lost to EINS | 7.26 |
| Tags_Will revert after reading the email | 4.97 |
| Lead Source_Welingak Website | 4.87 |
| Last Activity_SMS Sent | 2.17 |
| Last Notable Activity_Olark Chat Conversation | -1.14 |
| Last Notable Activity_Modified | -1.81 |
| Lead Quality_Worst | -2.94 |
| Tags_Ringing | -3.23 |
| Tags_invalid number | -3.49 |
| Tags_switched off | -3.67 |

  - Arriving to final model with 11 variables and apply the model on Test data set

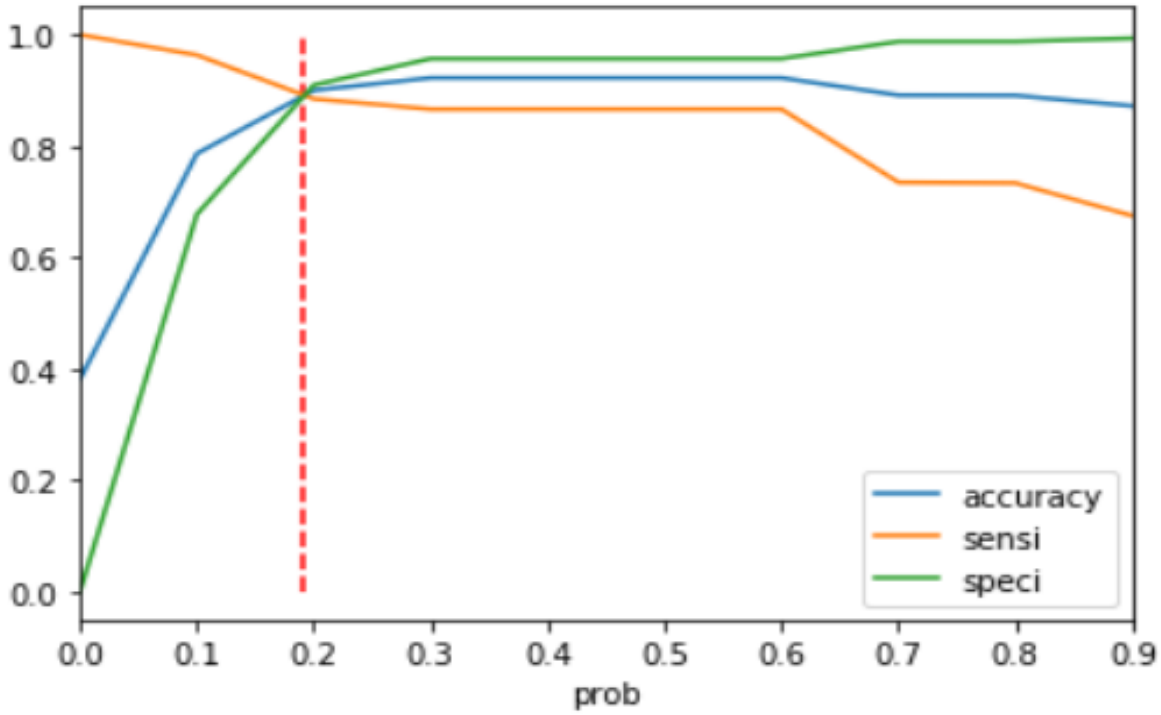
***Results:***

- **ROC Curve:**




  - ROC curve is more towards the upper-left corner of the graph, it means that the model is very good
  - Area under the ROC curve is 0.96 which detonates it's a good model

- **Finding optimal Cutoff:**




  - Based on Accuracy, Sensitivity and Specificity, we have considered the cut off probability as 0.19. it's a point where all three metrics meet

- Accuracy @ 90%. Accuracy is expected to measure how well the test predicts both categories
- Sensitivity @ 88%. Sensitivity indicates, how well the test predicts one category( True Positives or True Leads)
- Specificity @ 91%. Specificity measures how well the test predicts the other category(True Negatives)
- Precision @ 93% and Recall @ 87%
- Based on the business needs we need to change the cut off. If business goal is to predict the true leads then we should change the prob cut off close to 0.

***Conclusion:***

- Identified 11 variables which are of high significance in determining a lead customer
- Sensitivity @ 88%. Sensitivity indicates, how well the test predicts one category( True Positives)
- Specificity @ 95%. Specificity measures how well the test predicts the other category(True Negatives)
- ***Focus Areas***:
    - Tags : Ringing, Invalid Number and Switched off are having –ve coefficient. If we could restrict invalid phone number entry in forms through mobile number verification will help and also capturing alternate number will help
    - Advertise more in Welingak Website