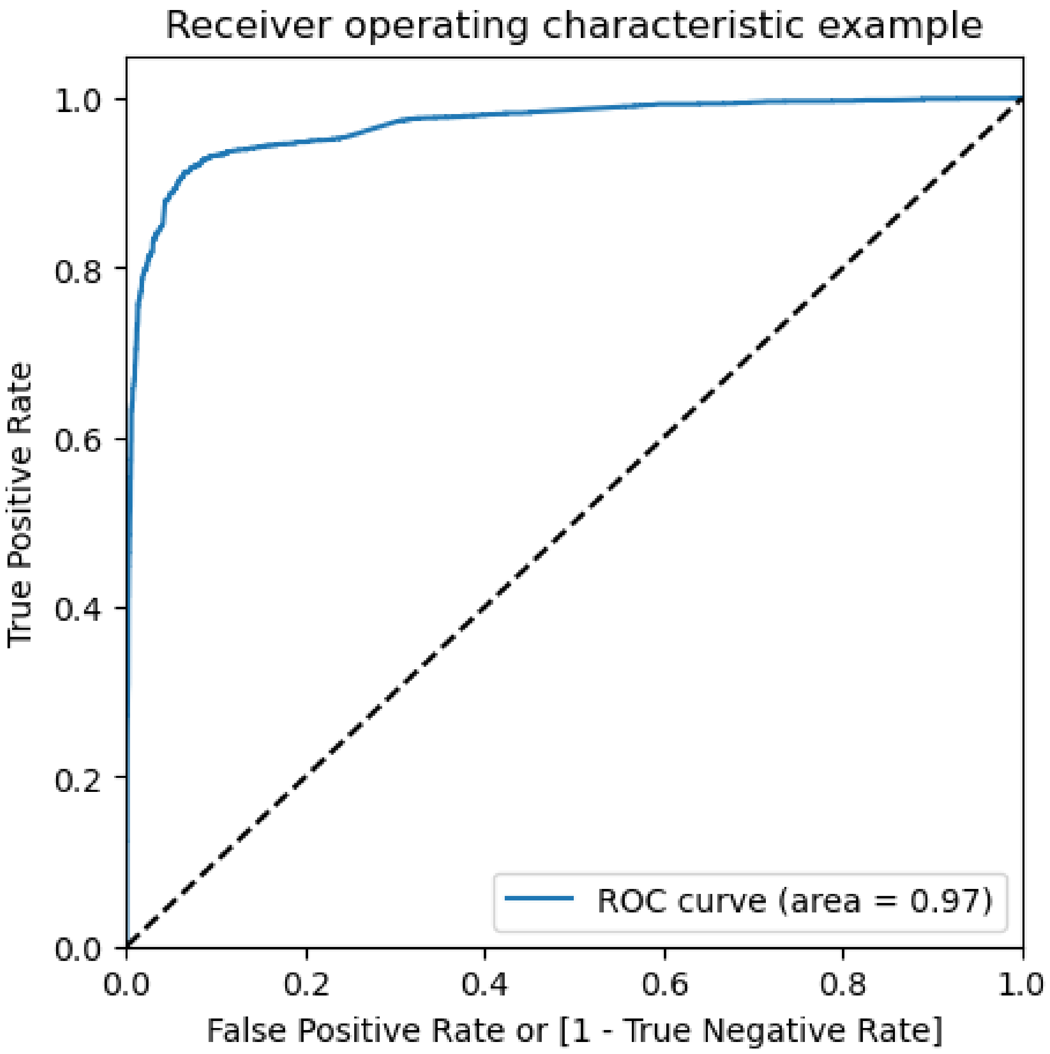
SUMMARY REPORT

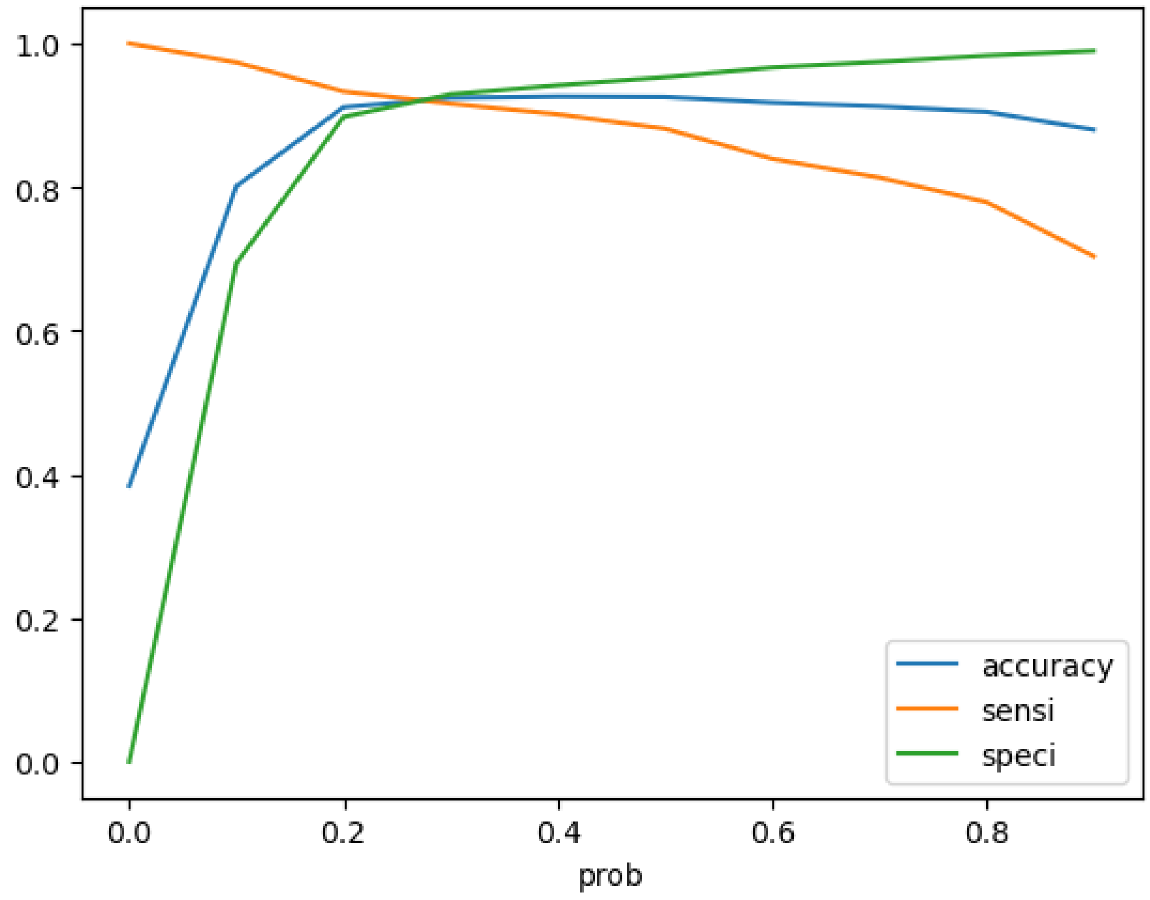
The dataset consists of 9240 leads with 37 variables in which ‘Converted’ column is the target i.e, dependent variable, whereas others are independent variables. The aim of the project is to build a Logistic Regression model to find out the ‘Hot Leads’ and increase the conversion rate from 30% to 80%.

- ***Libraries***: numpy, pandas, matplotlib, seaborn, sklearn, statsmodel.
- ***Null values***: Null values along with 'Select' (assumed to have been result of not selecting the value by the customer) are taken as Null.
  Dropped variables where null value is greater than 45% and are not extremely significant for the business i.e, ‘How did you hear about X Education’, ‘Lead Profile’, ‘Lead Quality’, ‘Asymmetrique Profile Score’, ‘Asymmetrique Activity Score’, ‘Asymmetrique Activity Index’, ‘Asymmetrique Profile Index’.
  Checked each column one by one to find the alternate values to fill null values such as median in ‘Page Views Per Visit’, ‘TotalVisits’ variables.
- ***Repetitive Values***: Also drop columns having high imbalance of data, because it would not contribute to unbiased target prediction, these are ‘What matters most to you in choosing a course’, ‘What is your current occupation’, City’, ‘Country’, ‘Lead Source’, ‘Receive More Updates About Our Courses’, ‘I agree to pay the amount through cheque’, ‘Get updates on DM Content’, ‘Update me on Supply Chain Content’, ‘Prospect ID’, ‘Newspaper Article’, ‘Through Recommendations’, ‘Digital Advertisement’, ‘Newspaper’, ‘X Education Forums’, ‘Magazine’, ‘Search’, ‘Do Not Call’, ‘Do Not Email’.
- ***Similar Variables***: The repeated variables are also removed. These are 'Prospect ID' and 'Last Activity'.
- ***Low Frequency Categorical Values***: Some categorical variables have multiple values with low frequency which are clubbed together into ‘other’ category so as to reduce the number of variables to work with. These variables are ‘Tags’, ‘Specialization’.
- ***Boolean Variables***: Convert boolean variable to binary variable, where ‘Yes’ represents 1 and ‘No’ represents 0. After data cleaning only one such variable ‘A free copy of Mastering the Interview’.
- ***Categorical Variable***: The variables with multiple categories converted to dummy variables and dropped the original one after concatenating them with original dataframe. These are ‘Lead Origin’, ‘Last Notable Activity’, ‘Specialization’ and ‘Tags’.
- ***Outliers***: There are outliers identified in the numerical variables namely ‘TotalVisits’ and ‘Page Views Per Visit’ and are managed by changing the values above 99 percentile by the value of 99th percentile.
- ***Correlation***: To see the correlation between different variables, build correlation matrix and since the matrix is too big, we can unstack the variables having greater than 0.5 and less than -0.5 correlation pairs, then drop the variable having higher

correlation accordingly. Dropped ‘TotalVisits’, ‘A free copy of Mastering The Interview’, ‘Last Notable Activity_Email Opened’ due to multicollinearity.

- ***Tain-Test Data***: Split data into train and test data (80:20), then scaled the data using RobustScaler to standardized the values of variables ‘Total Time Spent on Website’, ‘Page Views Per Visit’.
- ***RFE method***: Used RFE method to identify the important 15 variables affecting the target variable 'Converted', with 92% model accuracy, out of the total 64 variables present.
- ***P-value and VFI***: The factors with high p values are ‘Tags_number not provided’, ‘Tags_Lateral student’ and ‘Tags_invalid number’; they are removed from the model one by one until p value lower than 0.05 is achieved. The remaining variables have p value lower than 0.05 and VIFs of all are below 5 as desired.
- ***ROC curve***: It shows the trade-off between Sensitivity and Specificity as desired as can be seen below was used to find the cut-off value.





With the cut off as 0.3, we got Precision around 88% , Specificity around 92% and Recall around 91% on Training Data. Whereas got Precision around 89% , Specificity around 93% and Recall around 92% on Testing Data.