Lead Scoring Case Study

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

The company requires to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%. Here our goal is to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Model Building Steps

- Importing Dataset
- Inspecting the dataframe
- Data Preparation: Missing values and Outlier Handling
- Dummy variables creation
- Train-Test Split
- Feature Scaling
- Model Building
- Plotting ROC and finding optimal cut-off point
- Predictions on Test dataset
- Confusion matrix and accuracy
- Conclusion

Step1: Importing Dataset

- First imported required library of pandas and numpy.
- Imported the data provided ‘Leads.csv’ as ld_data
- The basis observation of data shows it contains some numerical and most of the categorical columns.

Step2: Inspecting the dataframe

- The primary shape of data is (9240,37)
- The most of columns contains high amount of missing value and numerical columns have outliers.
- Replaced ‘Select’ with Null as instructed.

Step3: Data Preparation

- Dropped the columns having missing value more than 40%
- For categorical columns , replaced missing value with mode value of the column if it is more than 50% in column., created a new value for combining low frequency values in a separated category
- For numerical columns, replaced missing value with their mean value.
- For outliers , removed top 1% values to remove it
- Dropped the columns which have only a single unique value

Step4: Dummy values creation

- First mapped binary variables (Yes/No) into 1/0
- For categorical variables with multiple levels created dummy variables and removed the repeated variables
- The final shape of data is now (9157,92)

Step5: Train-Test Split & Feature Selection

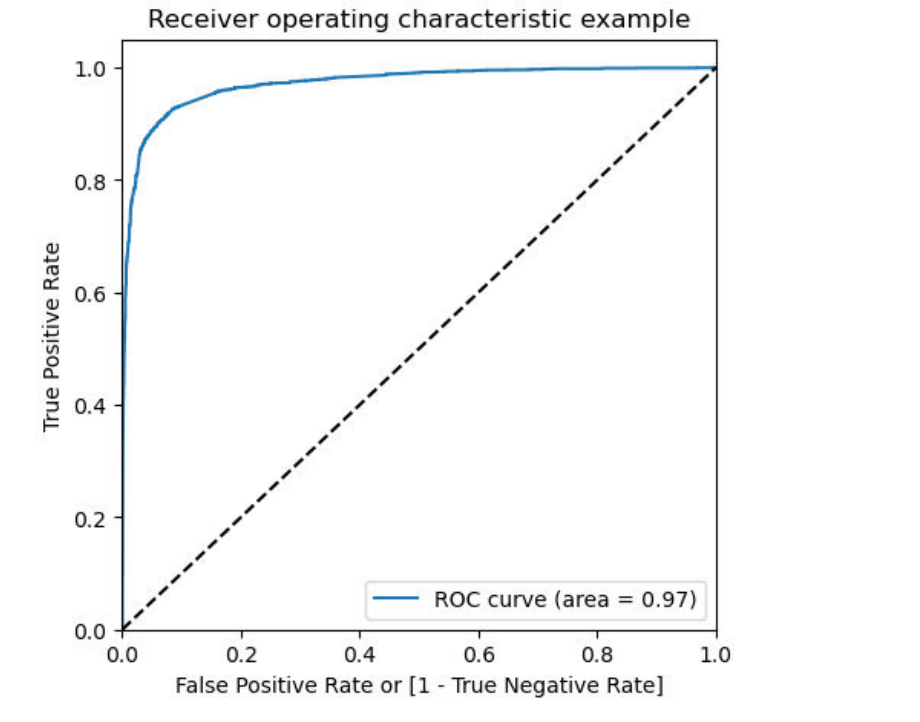
- Split the data into Train – Test dataset taking of 70%, 30% of data size.
- Scaled the numerical columns(other than having 1/0) using standard scaling method
- Used RFE method for feature selections, total 13 feature selected in primary basis

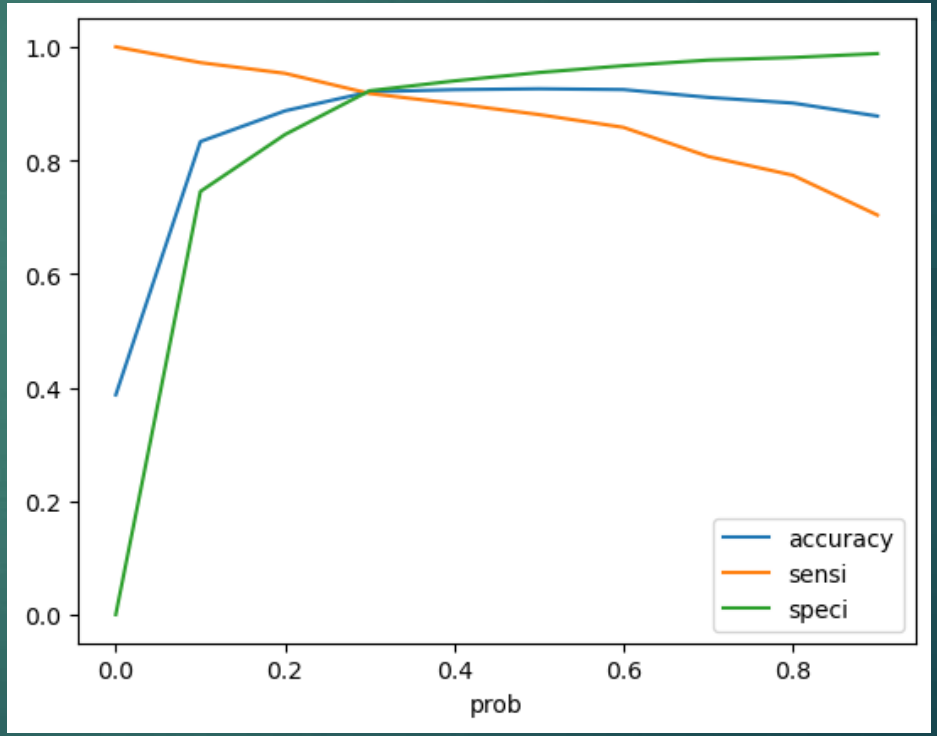
Step6: Model Building

- Built model using selected features using RFE method
- Checked P- value and VIF score which was under range for all the features so no need to change anything in the model further
- Created dataframe with actual converted flag and predicted probabilities
- Created predicted column with if converted_Prob > 0.4 else 0

	converted	converted_prob	leadId	predicted
0	0	0.001798	6490	0
1	1	0.419513	4026	1
2	1	0.690014	6453	1
3	1	0.889102	8949	1
4	0	0.018406	6467	0

Step7:Plotting ROC and finding optimal cut-off point





Step8:Predictions on Test dataset

- First scaled test data numerical values using standard scaler
- Applied the model on test dataset

	Converted	leadId	converted_Prob	final_predicted	Leadscore
0	0	7359	0.118463	0	12
1	0	2201	0.002173	0	0
2	1	2473	0.998153	1	100
3	0	8388	0.002164	0	0
4	0	2949	0.000285	0	0

Conclusion:

- Train Data

ACCURAY 92.07%

Sensitivity 91.77%

Specificity 92.26%

- Test Data

ACCURAY 92.24%

Sensitivity 92.24%

Specificity 93.25%