Lead Scoring Case Study

Problem statement

An education company named X Education sells online courses to industry professionals. Many professionals who are interested in the courses visit their website and browse for courses, and such professionals are classified as leads. Employees from the sales team reach out to such professionals to convince them to buy courses, but the percentage of conversion is very low (around 30%).

Problem statement is to increase the conversion rate by building 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. The target lead conversion rate should be around 80%.

Also, problem statement is to quickly increase the number of potential leads from model when sales teams increases (Interns join) as sales team can target more leads during that time.

Analysis Approach

Step 1: Importing Data

Step 2: Inspecting the Data frame (Null checks, check data set size, categorical columns levels, and ranges of numerical columns values)

Step 3: Data Preparation and looking at correlations (Imputing null values, finding misspelled data, outlier(s) treatment, and looking at numeric columns correlations)

Step 4: Test-Train Split (Splitting the dataset into test, and train data set)

Step 5: Feature Scaling (Applying feature scaling on test set)

Step 6: Logistic Regression Model (Creating model with all features)

Step 7: Feature selection using RFE (Identifying list of features necessary for model)

Step 8: Plotting the ROC curve and Finding optimal cutoff point (Finding the cutoff value of the lead score to identify a converted user)

Step 9: Making predictions on the test set (Applying the model to test set, and analyzing the results)

Results

Model was able to successfully identify users which can convert based on some parameters. Model can be adjusted also to give more leads when sales team is more in numbers.

Following parameters have been shortlisted to identify a conversion. Some of the following parameters are levels of original dataset categorical columns

1. Lead Source_Welingak Website
2. Tags_Busy
3. Tags_Closed by Horizzon
4. Tags_switched off
5. Tags_Lost to EINS
6. Asymmetrique Activity Index_03.Low
7. Lead Origin_Lead Add Form
8. Lead Quality_Worst
9. Lead Quality_Not Sure
10. Tags_Ringing
11. Tags_Will revert after reading the email
12. Last Notable Activity_SMS Sent

Results

Recursive Feature Elimination (RFE), and Variance Inflation Factor (VIF) parameters were used to identify parameters which impacted the conversion factor the most, and model was built using such parameters.

Conversion Matrix values:

True positives - Leads who got converted, and our model correctly predicted it

(Test set = 874, Train set = 2121)

True negatives - Leads who did not get converted, and our model also correctly predicted it

(Test set = 1615, Train set = 3664)

False positives - Model predicted lead will be converted, but it did not convert

(Test set = 94, Train set = 250)

False negatives - Model predicted lead will not be converted, but it did got converted

(Test set = 141 , Train set = 318)

Train set output parameters

Accuracy – 92%

Sensitivity / True positive rate – 87%

(Model predicted that user will convert, and 87% of such users converted)

Specificity – 94%

(Model predicted that user will not convert, and 94% of such users did not convert)

Test set output parameters

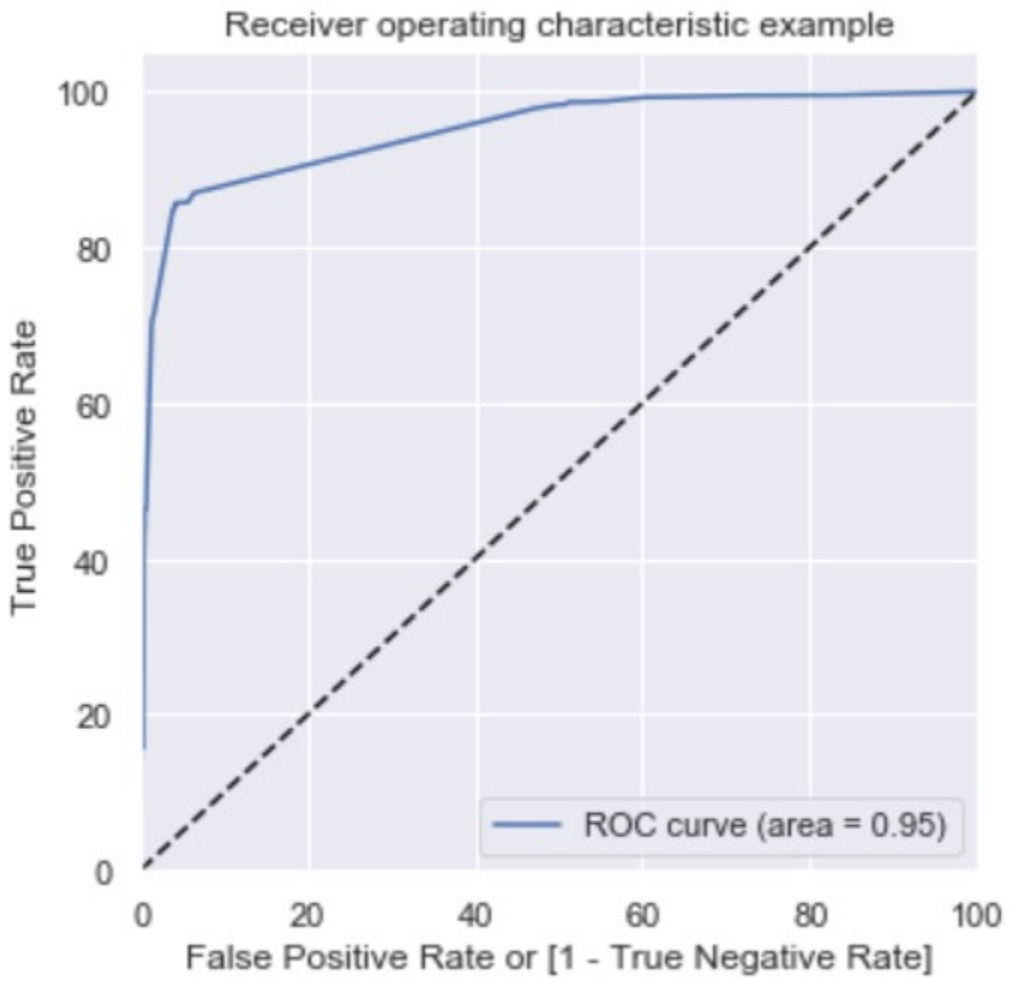
Accuracy – 91%

Sensitivity / True positive rate – 86%

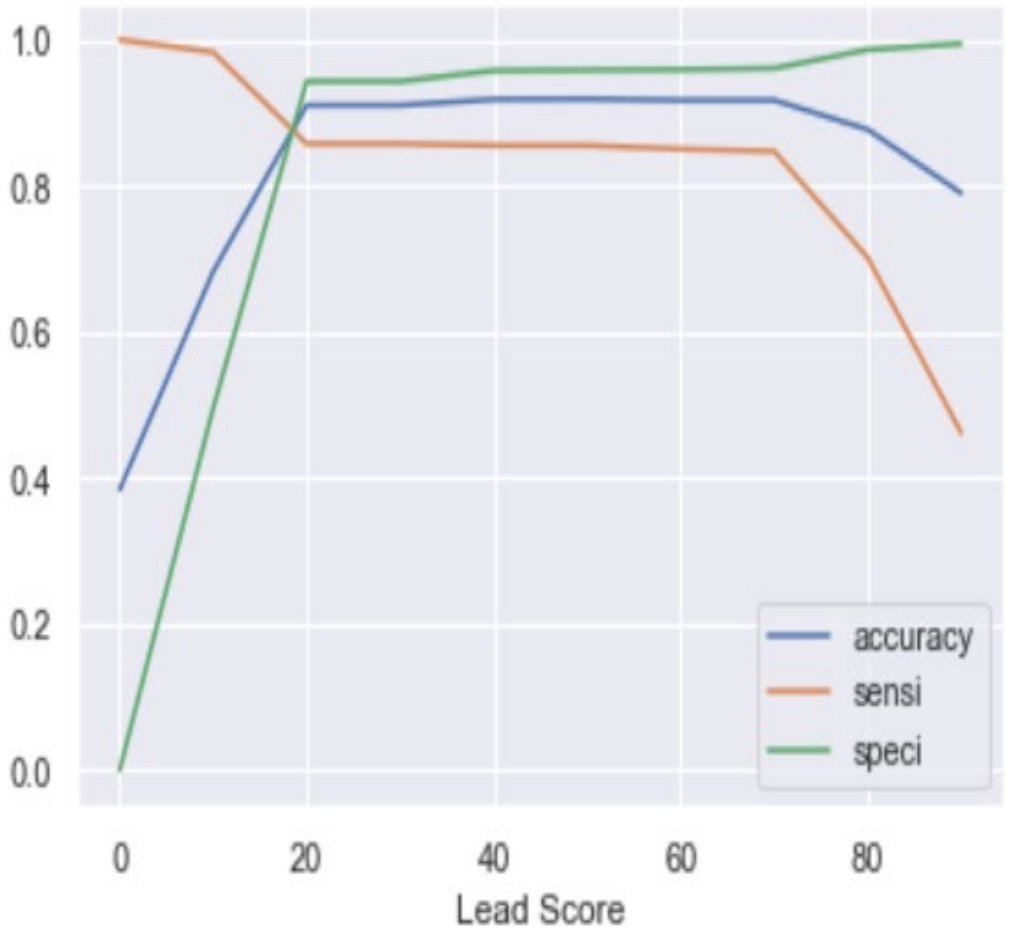
Specificity – 94%

Since, the train/test set Sensitivity/Specificity rates are approximately same, we can say that model is not overfitted, and is delivering good results.

ROC Curve



Specificity/Sensitivity/Accuracy Trade off Graph



Leads based on company's requirement changes

Sensitivity/Specificity/Accuracy trade off graph holds the key of this requirement.

If interns can handle more work, then we can lower the threshold to mark the user as potential lead. In this process, we will decrease specificity, and increase sensitivity but it will fulfill our requirement.

If less work is needed once quarter targets have been achieved, then we can increase the threshold to only provide us high quality hot leads. Using this process, we can increase the revenue of the company without spending much effort on the process.