Lead Scoring Case Study

Problem Statement

- The X Education company requires you to build a logistic regression model wherein we need to assign a lead score 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.

Roadmap

- Created train and test set by splitting the original cleaned data set after treating missing values.

- Selected 15 features using Recursive Feature Elimination (RFE) after creating dummy variables and scaling the data.

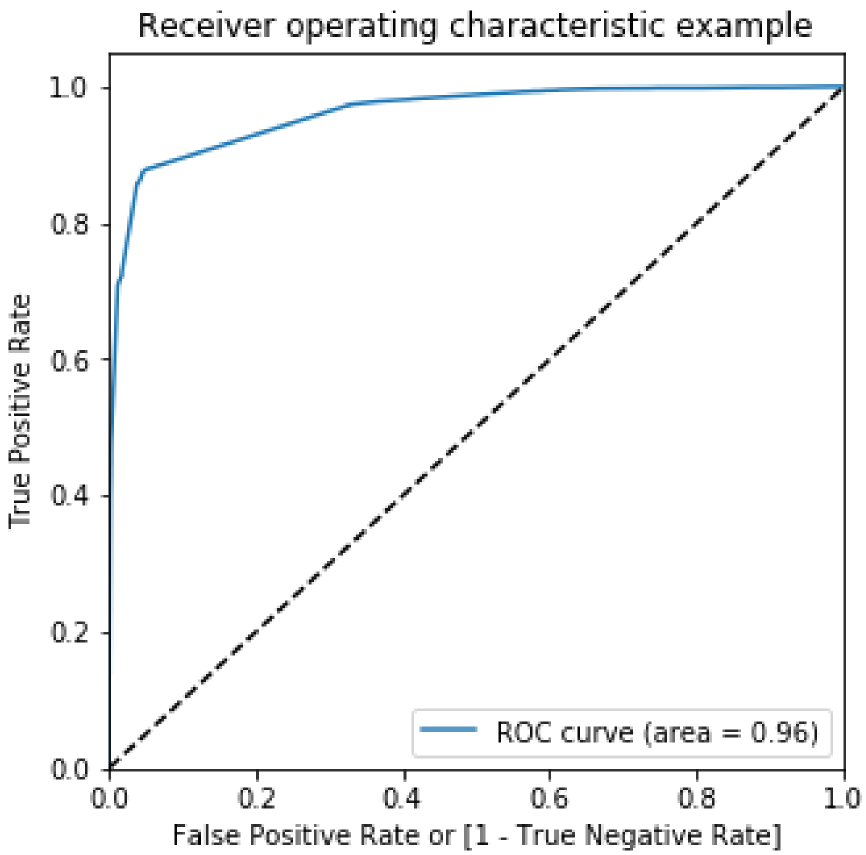
- Applied Logistic Regression algorithm to build a model and more than 92% accuracy and 87% sensitivity.

- Identified the optimal probability cutoff from the accuracy, sensitivity and specificity.

- Applied the model on the test data to identify the conversion probability. (accuracy 92%, sensitivity 87%)

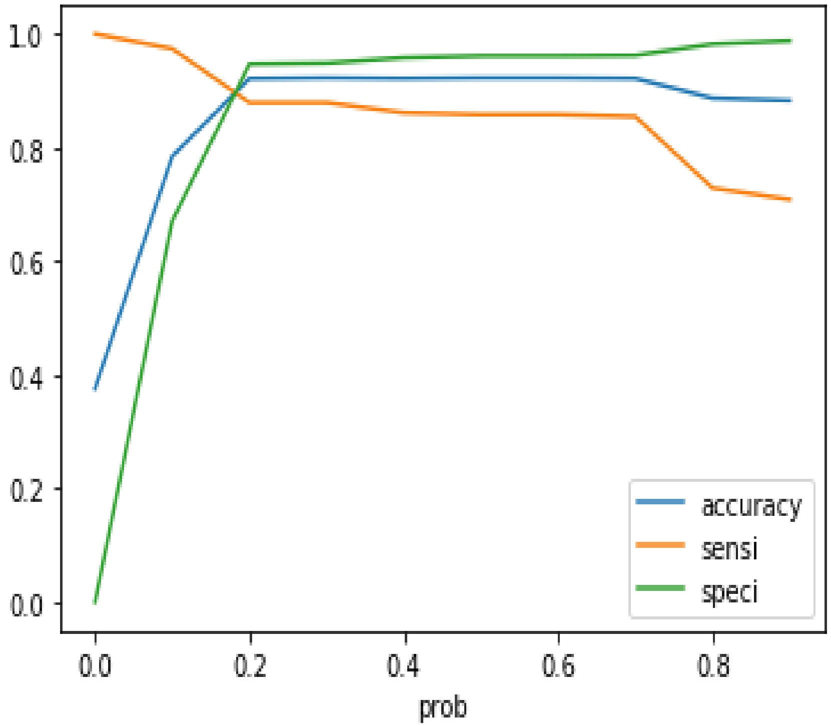
- Based on the calculated predicted probability, and optimal probability cutoff, all the leads are assigned with a lead score value (lead score = predicted probability x 100)

ROC Curve



- The ROC curve shows that 96% of the area is under the curve.
- The classification probability of lead conversion (1/0) is very high by the model.

Optimal probability cutoff



- Optimal probability cutoff is identified as 0.2 for better accuracy of the classification of lead conversion.
- With 0.2 cutoff the model has
  - Accuracy: 92%
  - Sensitivity: 87%
  - Specificity: 94%

Confusion matrix on Test data

Actual/Predicted	Not Converted	Converted
Not Converted	1498	82
Converted	121	842

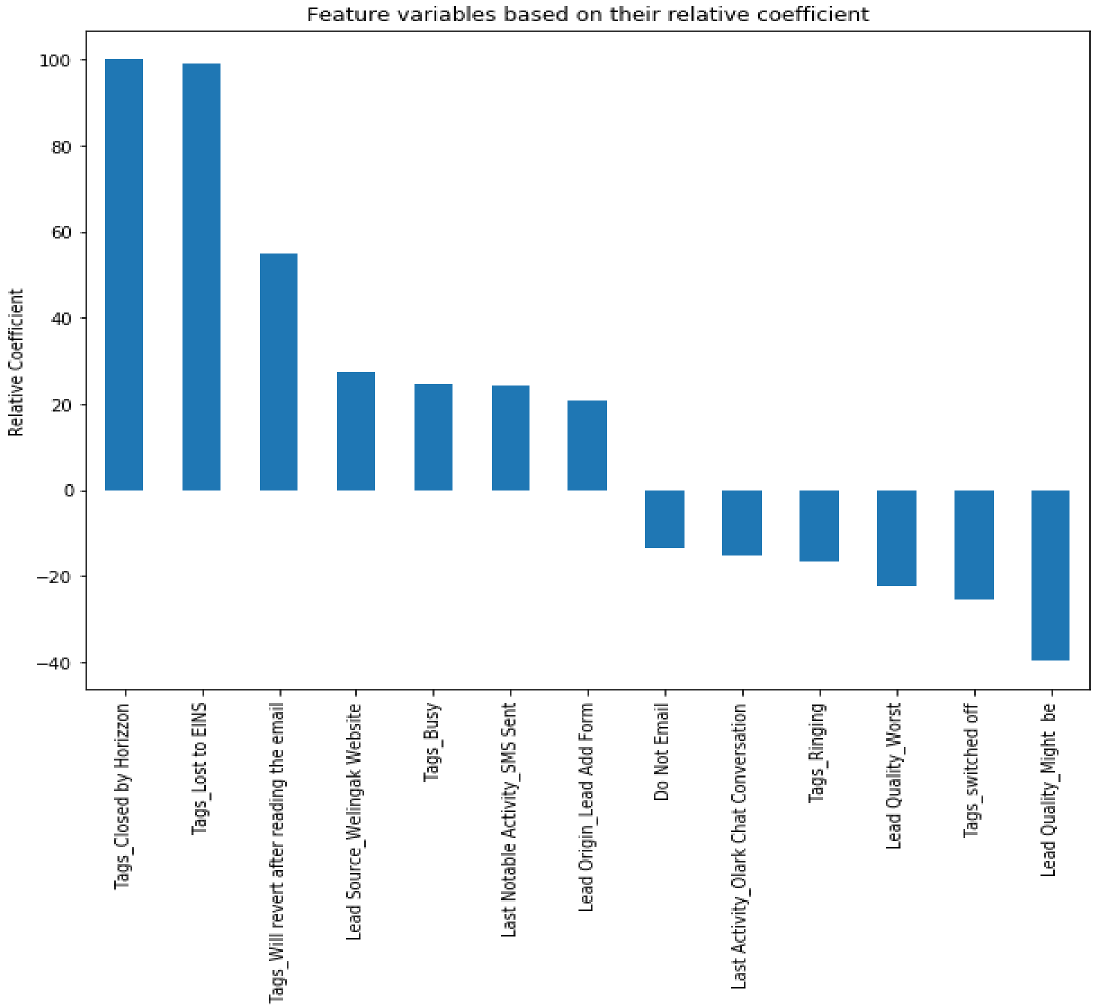
Accuracy: 92% | Sensitivity : 87% | Specificity : 94%

The model can predict if a lead can be converted or not with 92% accuracy on unseen data. This will help the company to predict the probability of ‘hot’ leads with 92% accuracy.

Also, the model can predict the probability of a lead which is actually converted over total converted lead with 87% chances.

The model’s prediction of a lead not getting converted is also very high (94% over unseen data). This means that X education company will save lot of time and resources by discarding low scoring leads.

Important Features



- The top 3 variables that contribute to convert a lead are:
  - Tags Closed by Horizon
  - Tags Lost to EINS
  - Tag We will revert after reading the email

- The top 3 variables that need improvement to convert a lead are:
  - Lead Quality_ Might Be
  - Tag switched off
  - Lead Quality Worst

Recommendation

- The leads which have high score can be treated as “hot” leads and sales team need to follow up as there is high possibility to convert those leads.
- Leads who have applied to ‘Do Not Email’ already does not need to be attended again.
- Based on the previous chat conversations if the lead is classified as ‘Might be’ or ‘Worst’ then those leads can be ignored.

Thank You