Lead Scoring Case Study

Submitted by : Prateek, Ihsan, Ankit

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’ . If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Objective

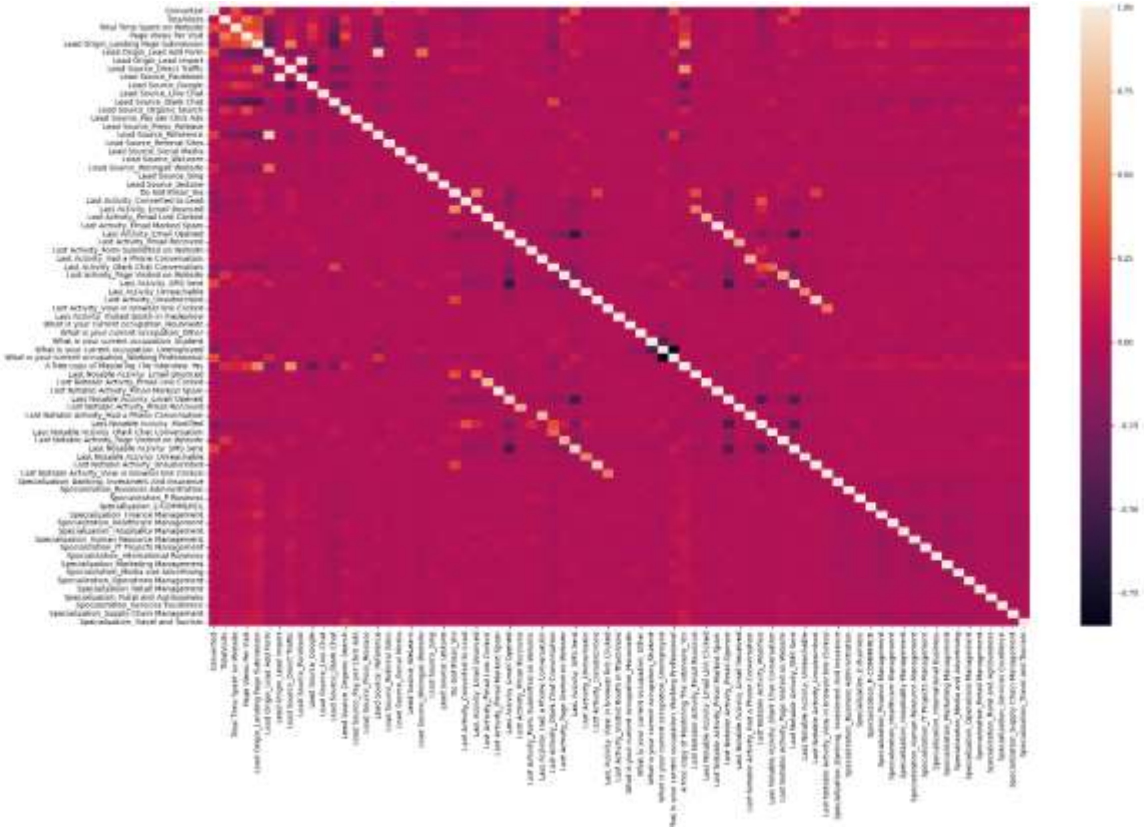
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- Target lead conversion rate has to be around 80%.

Data Cleaning & Preparation

- The leads dataset has 9240 rows and 37 columns
- Out of all the columns, we dropped the columns which have more than 3000 missing values.
- We then drop City, Country, Lead Profile and How did you hear about X Education, Prospect ID and Lead Number columns as they are not useful for our analysis.
- Around 12 columns have only one value which is "No" which will not help our analysis and hence is dropped.
- Null values are dropped from Specialization, Lead Source, Total Visits, What is your current occupation and What matters most to you in choosing a .course columns

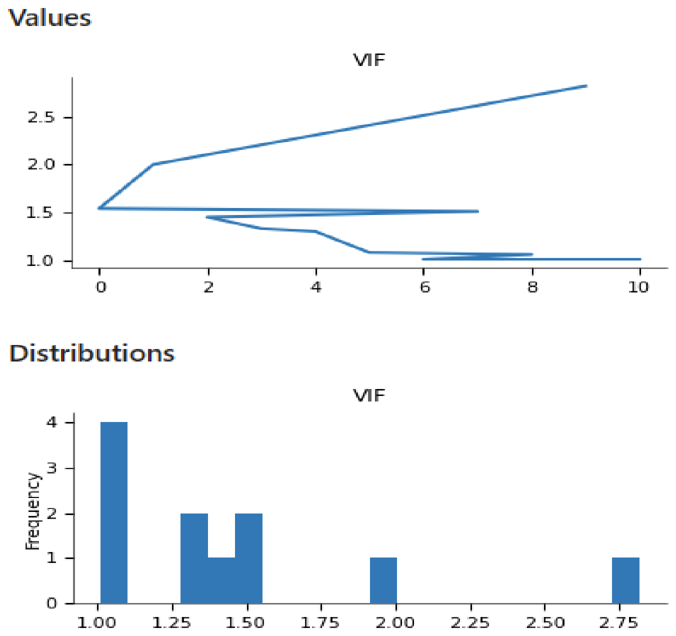
Data Preparation for Modelling

- Dummy variables are created for 'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity' , 'What is your current occupation',' A free copy of Mastering The Interview', 'Last Notable Activity' and 'Specialization' .
- Test Train split of 70:30 ratio
- Min Max scaling of numeric variables is then done
- Analyse the correlation heatmap of the variables



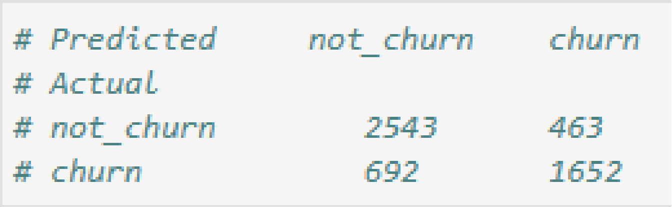
Model Building

- Recursive Feature Elimination (RFE) used to select 15 features
- Initial model built but the column “Last Notable Activity_Had a Phone Conversation” dropped due to p-value of 0.99
- “What is your current occupation_Housewife” column dropped because of p-value of 0.99 and “What is your current occupation_Working Professional” because of 0.212 p-value
- Final model has all VIFs and p-values in range



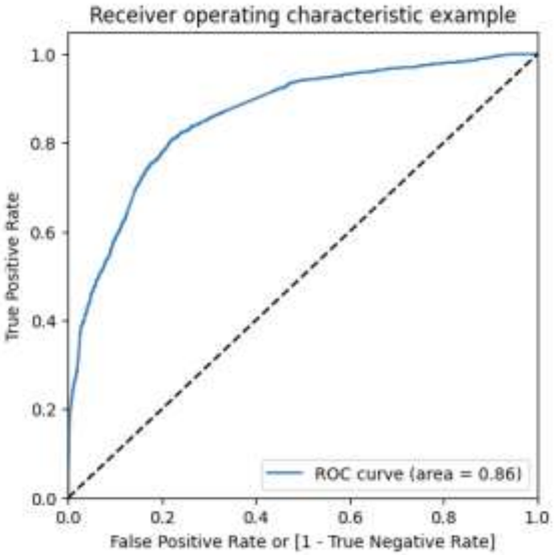
Model Evaluation

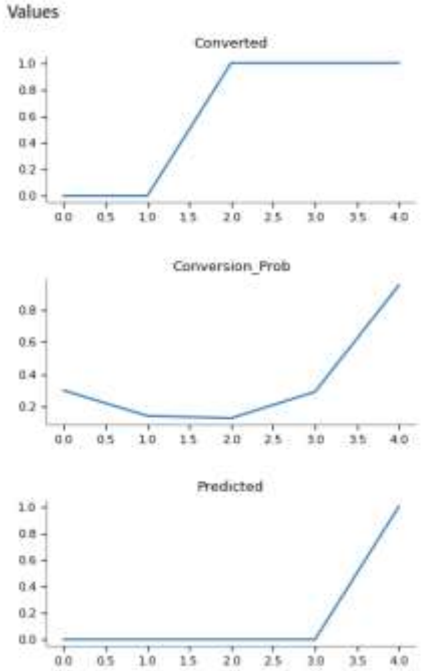
- Y train predictions pn train set done and new column with values of 1 and 0s based on cut off value of 0.5 created.
- Confusion matrix generated and sensitivity and specificity calculated to be around 73% and 83% respectively



```
# Predicted      not_churn     churn
# Actual
# not_churn         2543       463
# churn             692        1652
```

- ROC curve shows a good value of 0.86



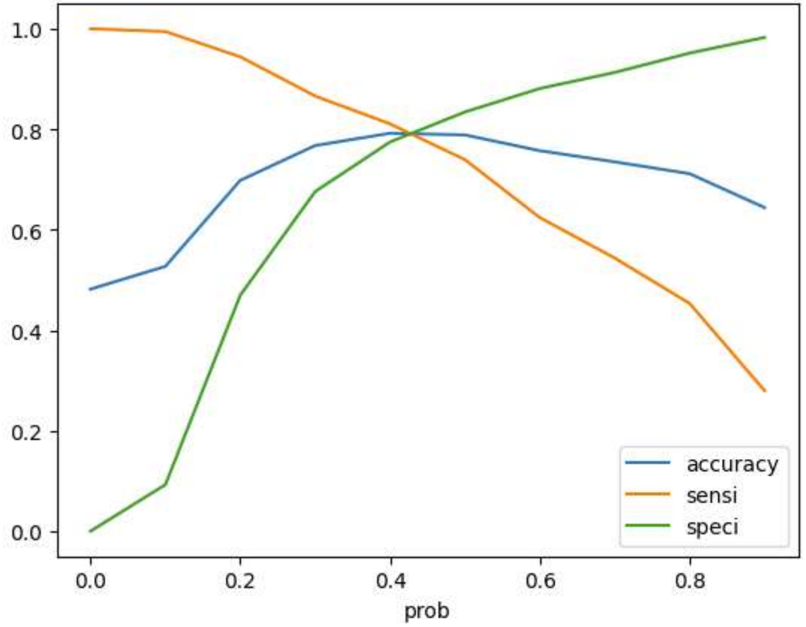


Model Evaluation

- Accuracy, specificity and sensitivity of model for different cutoffs is plotted and 0.42 is chosen as optimal cutoff
- Transformations which were done on training set done on test set and predictions done.
- New column created using 0.42 as cutoff for Converted or not.
- Overall accuracy, sensitivity and specificity found to be 78.4%, 77.9% and 78.9% respectively.
- Precision(TP / TP + FP) and recall(TP / TP + FN) found to be 80.5% and 73.9% respectively
- Confusion Matrix

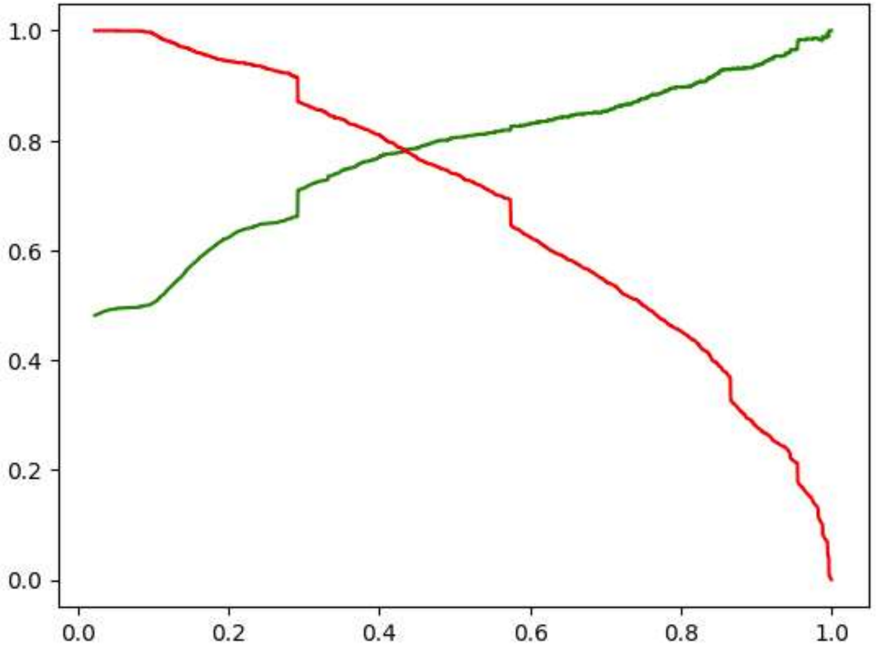
  [1852, 460]

  [479, 1670]



Model Evaluation

- Based on precision recall tradeoff cutoff point is changed to 0.44
- Predictions are done on the test set using 0.44 as cutoff to decided "Converted or not" .
- Final models accuracy, precision and recall on test set found to be 78.6%, 78.2% and 76.7% respectively



Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

- Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.

- Also the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 79% in test set.

- The top 3 variables that contribute for lead getting converted in the model are
  - ❑ Total time spent on website
  - ❑ Lead Add Form from Lead Origin
  - ❑ Had a Phone Conversation from Last Notable Activity.

- Hence overall this model seems to be good.