Lead Scoring Case Study

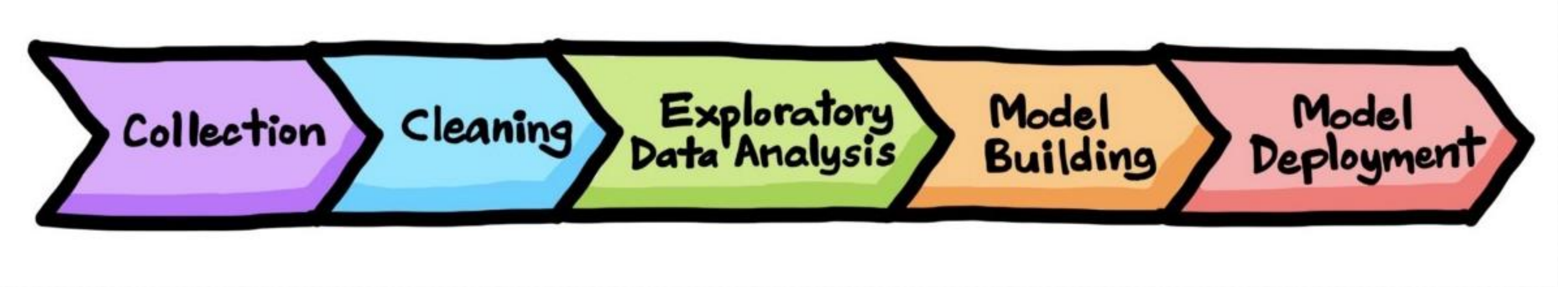
Nithyashree PV

Problem Statement

- X Education sells online courses for professionals.
- People who land on the website and fill up a form are considered leads.
- Their typical conversion rate is around 30%.
- We have been appointed by X education to identify the most promising leads by building a model.
- The CEO has given a ballpark lead conversion rate of around 80%.

Approach

- This problem can we solved by using a Logistic Regression Classfication model to identify the hot leads from the less promising leads.
- The following steps are involved and the steps form Data Cleaning to Model Building have been followed for our problem.

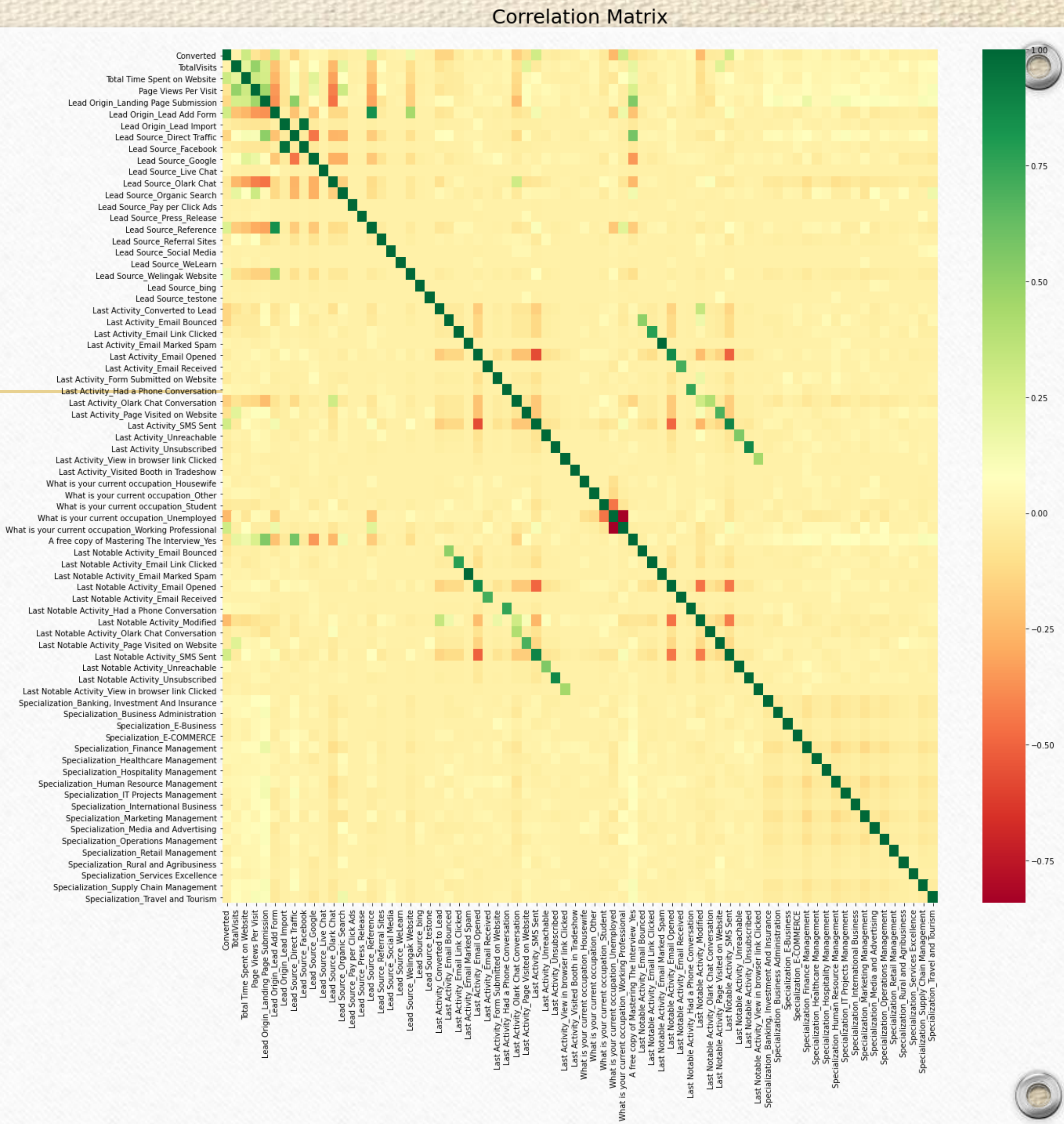


Approach

- First we have dropped columns with over 3000 missing values, and columns with high imbalanced or data that isn't very informative for our business purposes.
- We have created dummy variables for categorical columns, Split our data set into training and test data. Then, scaled the numerical variables for leveled results.

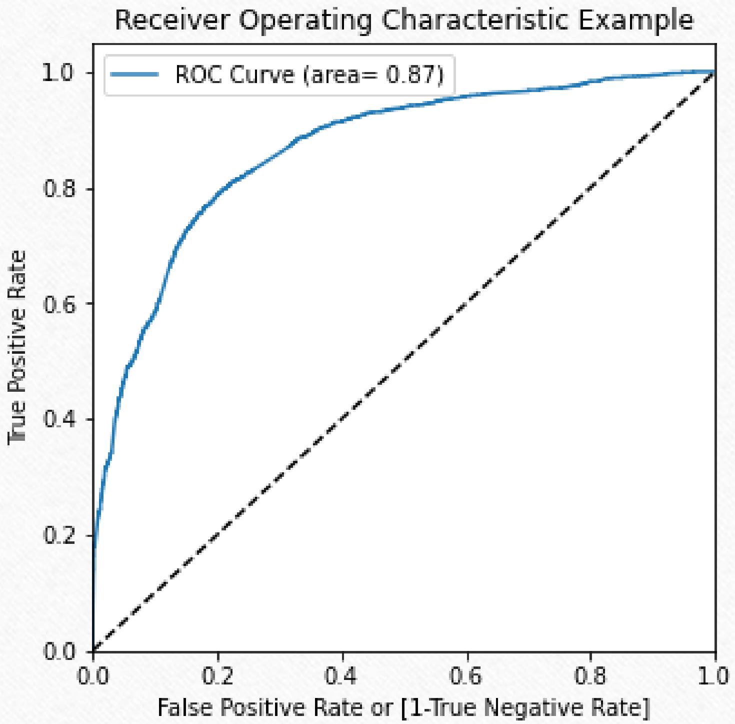
Correlation Matrix

- A correlation matrix of all variables after dummification to idenfity multicollinearity among different columns within the data.
- We reduced these columns through RFE for top 15 features and built our model through elimination of less significant variables.

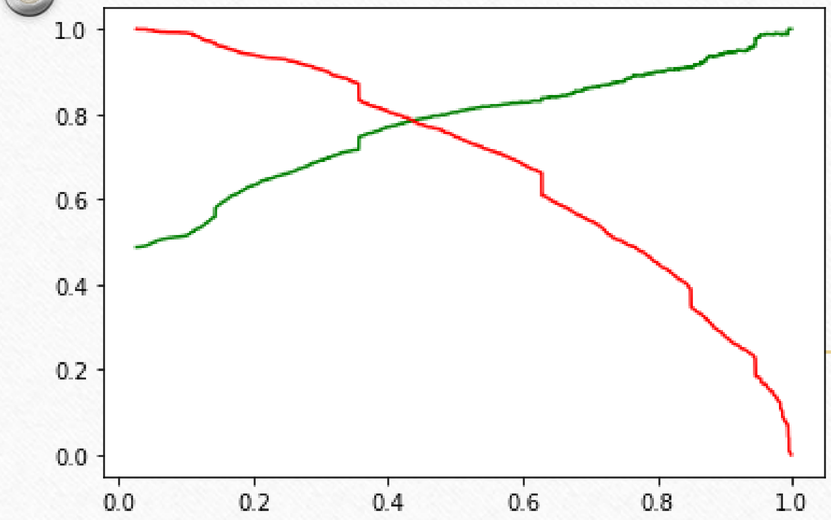


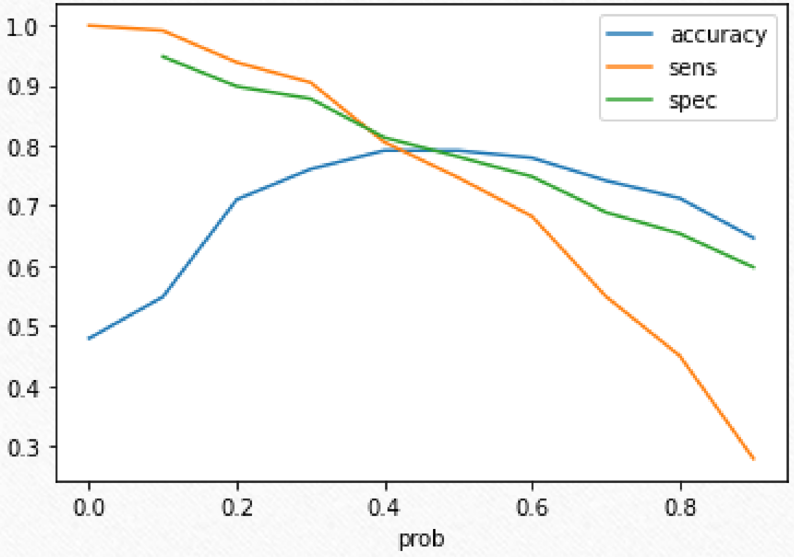
ROC Curve

- We stopped at a model with 13 variables with p value<0.05 and vif<5.
- We calculated the Predicted Conversion and Conversion Probability to plot the ROC curve.
- The area under the ROC curve is 0.87 which means are model is good.



Precision Recall Cutoff



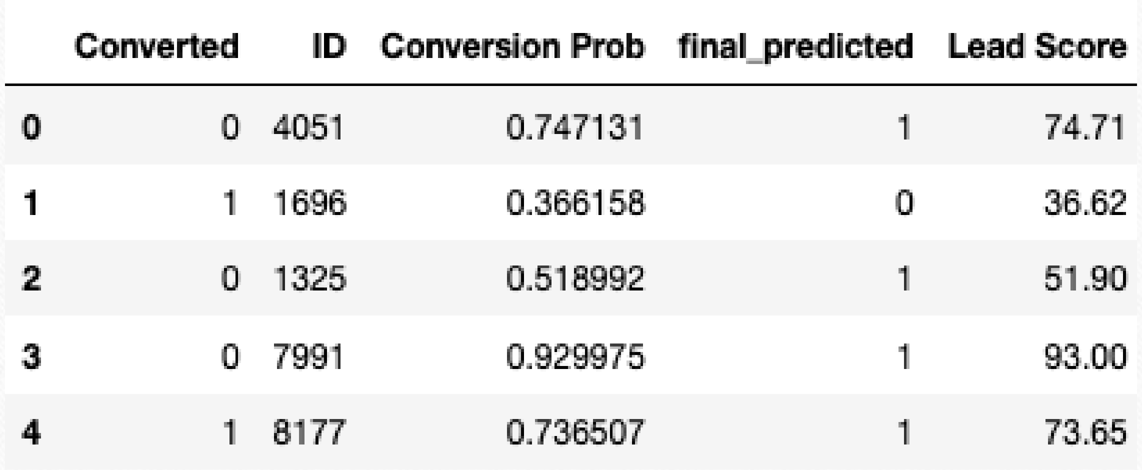


- Graph 1: Precision Recall Tradeoff where we can see the optimal cutoff point is 0.4-0.5.

- Graph 2: The plot of accuracy, sensitivity and specificity where they converge between 0.4-0.5 further confirming the optimal cutoff point required for our model's performance.

Train vs Test Data

	Training Data	Test Data
Precision	0.79	0.78
Recall	0.78	0.77
Accuracy	0.79	0.78



	Converted	ID	Conversion Prob	final_predicted	Lead Score
0	0	4051	0.747131	1	74.71
1	1	1696	0.366158	0	36.62
2	0	1325	0.518992	1	51.90
3	0	7991	0.929975	1	93.00
4	1	8177	0.736507	1	73.65

We can see form the above metrics the model performs well on both Training and Test data as the values are quite similar. We assign a lead score for each potential ID and leads with higher than 0.45 are promising leads.

Business Recommendations

- X education should prioritise leads based on:
  1. Total time spent on the website, and total vists: more time spent on the website as well as more visits by the same persons indicates higher lead score.
  2. Lead Source: Leads from Welingak Chat or Olarkchat are more likely to be converted.
  3. Phone Conversation/Add form: Leads who have called or filled the form have more chance for conversion.
  4. Occupation: Working Professional are the most likley to lead to conversion to hot leads among all occupations.
  5. The business should still prioritise higher lead scores when the sales team is in its busier season while still prioritising lesser promising leads when time allows as phone conversation can increase potential conversion of leads.