Lead Score Case Study

Ravi Edla

Abstract Info

Problem statement:

An education company named X Education sells online courses to industry professionals. Though X Education lands a lot of people on their website and gets them to provide their email address or phone number, their lead conversion rate (paying customer) is very poor.

Objective:

To build a model such that a lead score is assigned to each of the leads in such a way that the customers with higher lead score have a higher conversion chance, and the customers with lower lead score have a lower conversion chance. And the target lead conversion rate is around 80%.

Data used for analysis:

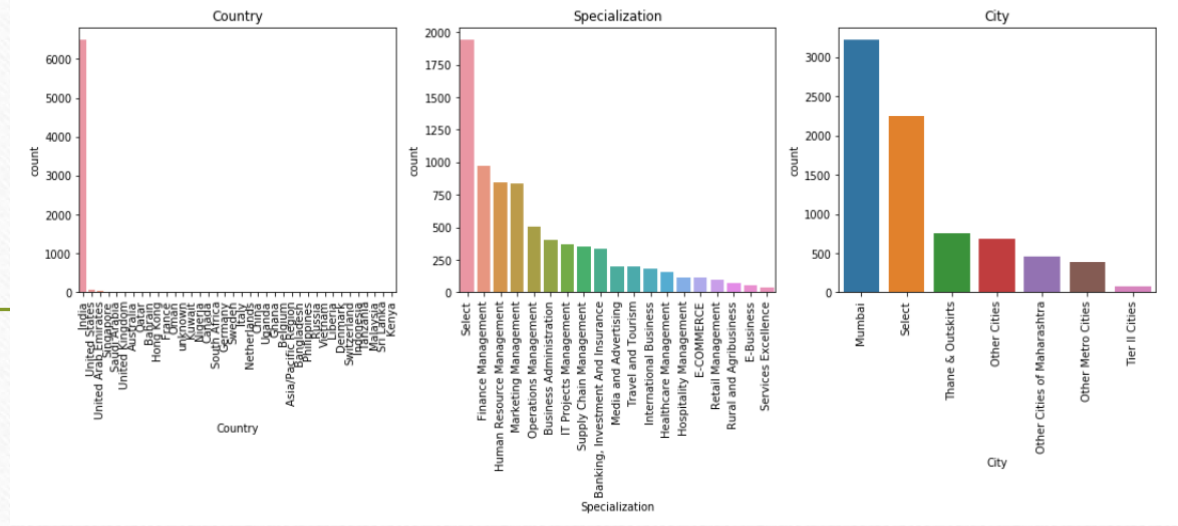
Leads dataset from the past with around 9240 data points.

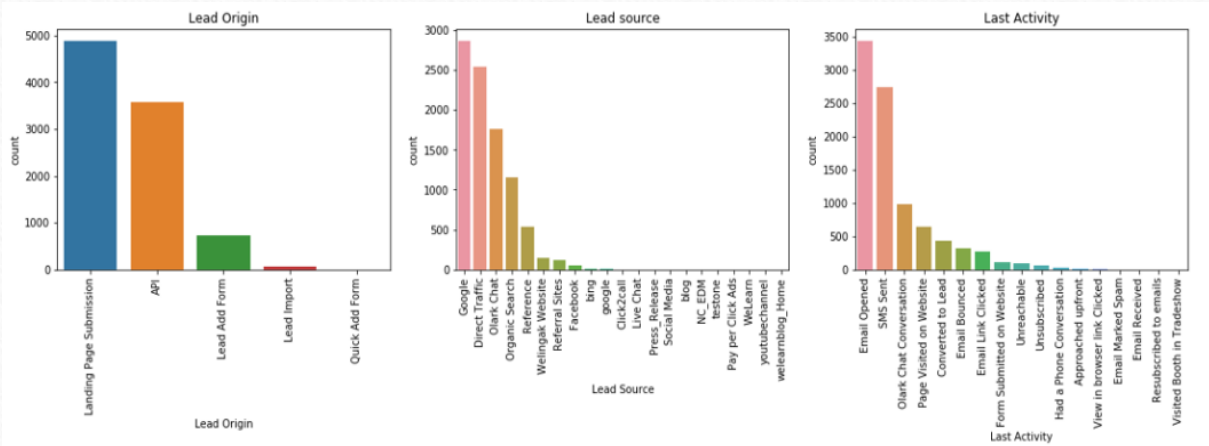
Analysis Methodology

- Data Collection, Importing and Cleaning
- Outlier Analysis and Removals
- Visualizing the Data – Univariate Analysis
- Data Preparation
- Train-Test Split
- Scaling the Data
- Model Building and Evaluation
- Making Predictions
- Conclusion

Univariate Analysis

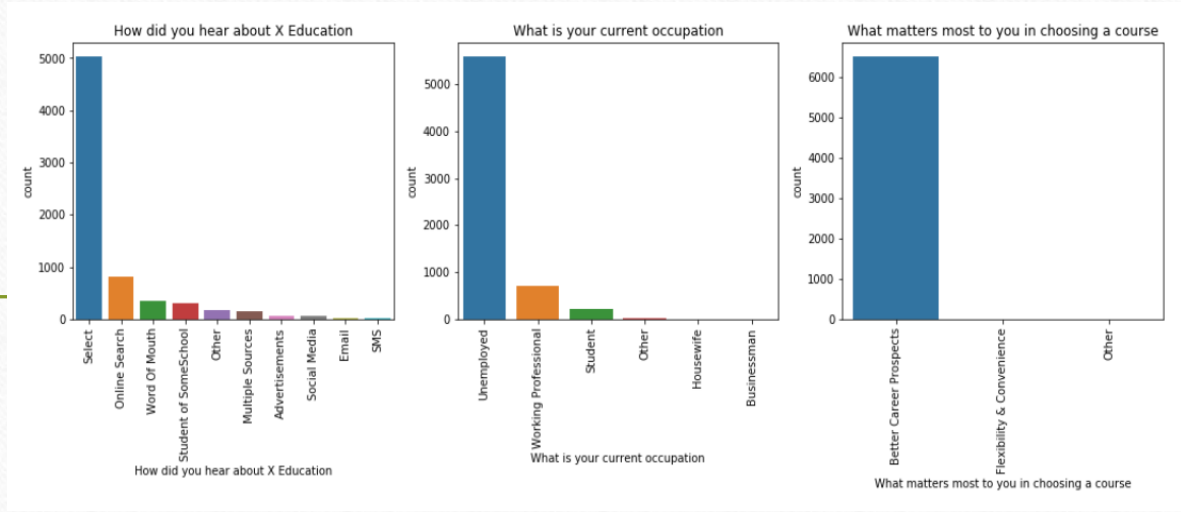
- Major lead origins from 'Landing Page Submission' and 'API'. Major source of leads are 'Google', 'Direct traffic', 'Olark Chat' with most traffic coming from 'India'.

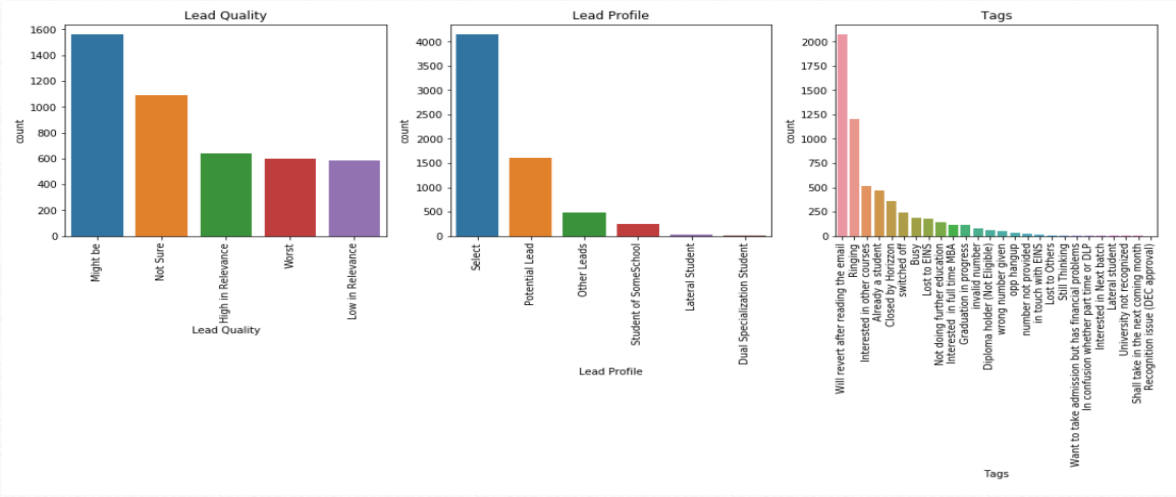




Univariate Analysis

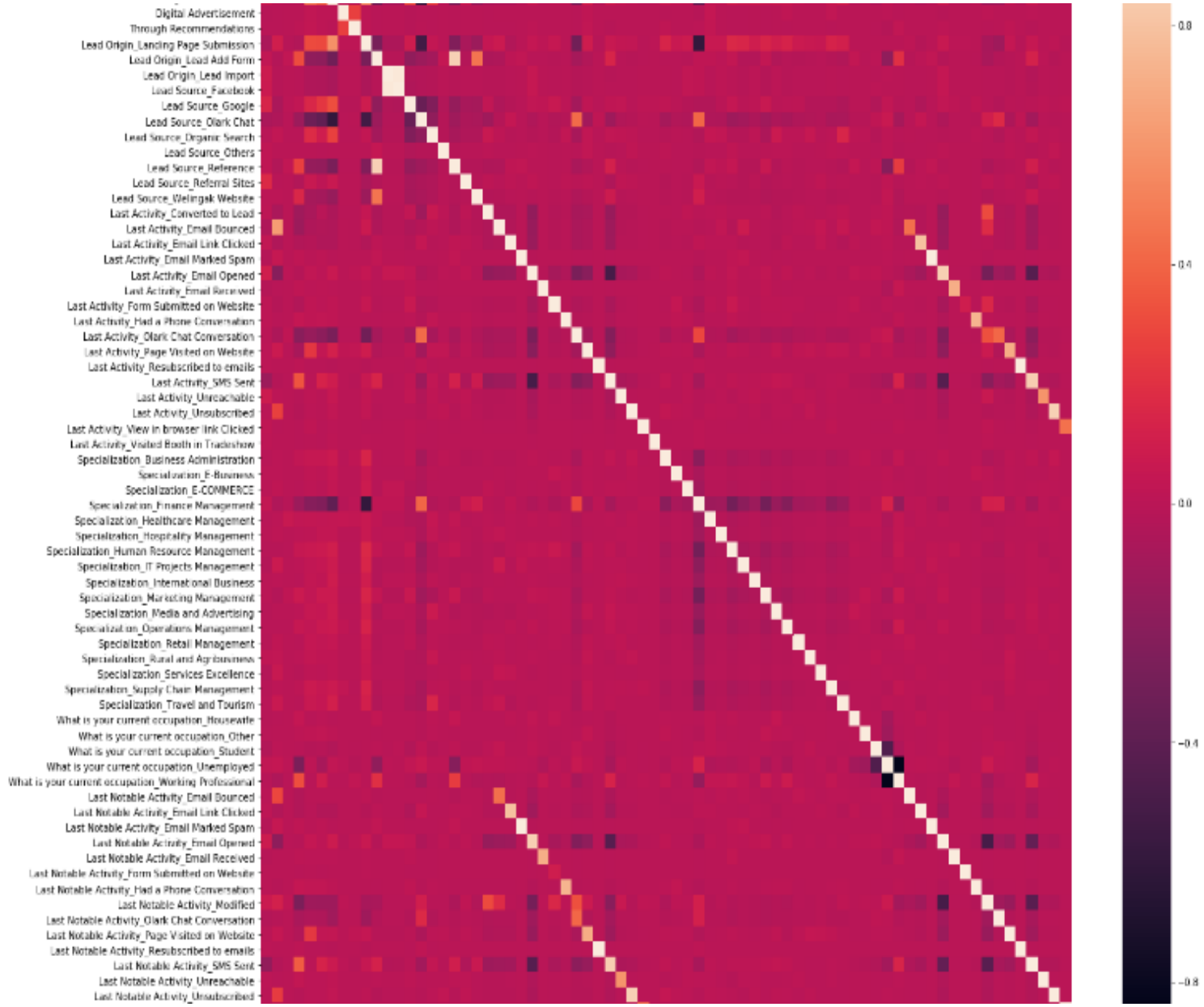
- We can see that there are few columns where 'Select' values are high in numbers. These data points are not adding any valuable information so we can treat them as NULL

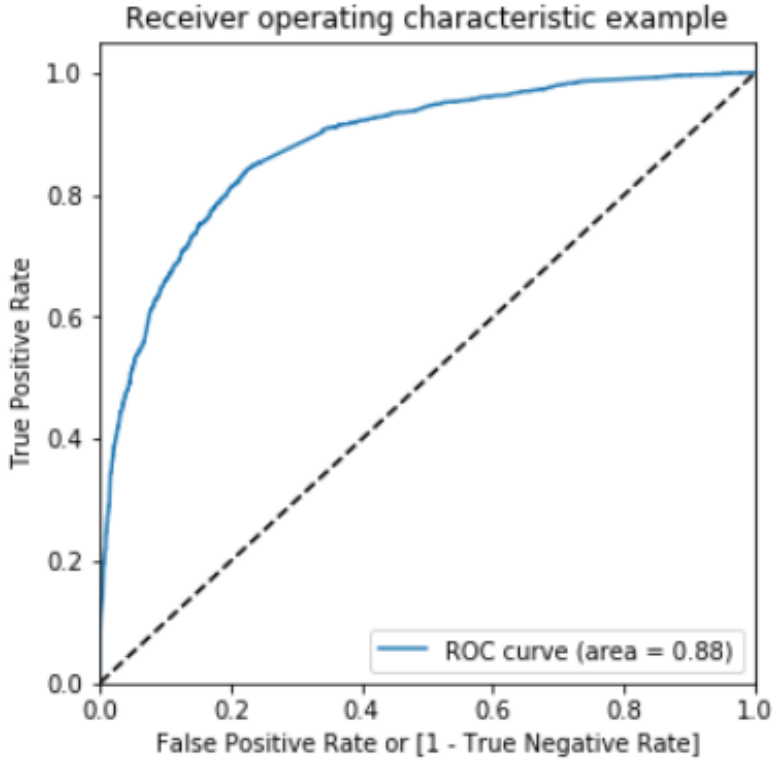




Data Correlation

- Created Dummy Variables for multi-level categorical column.



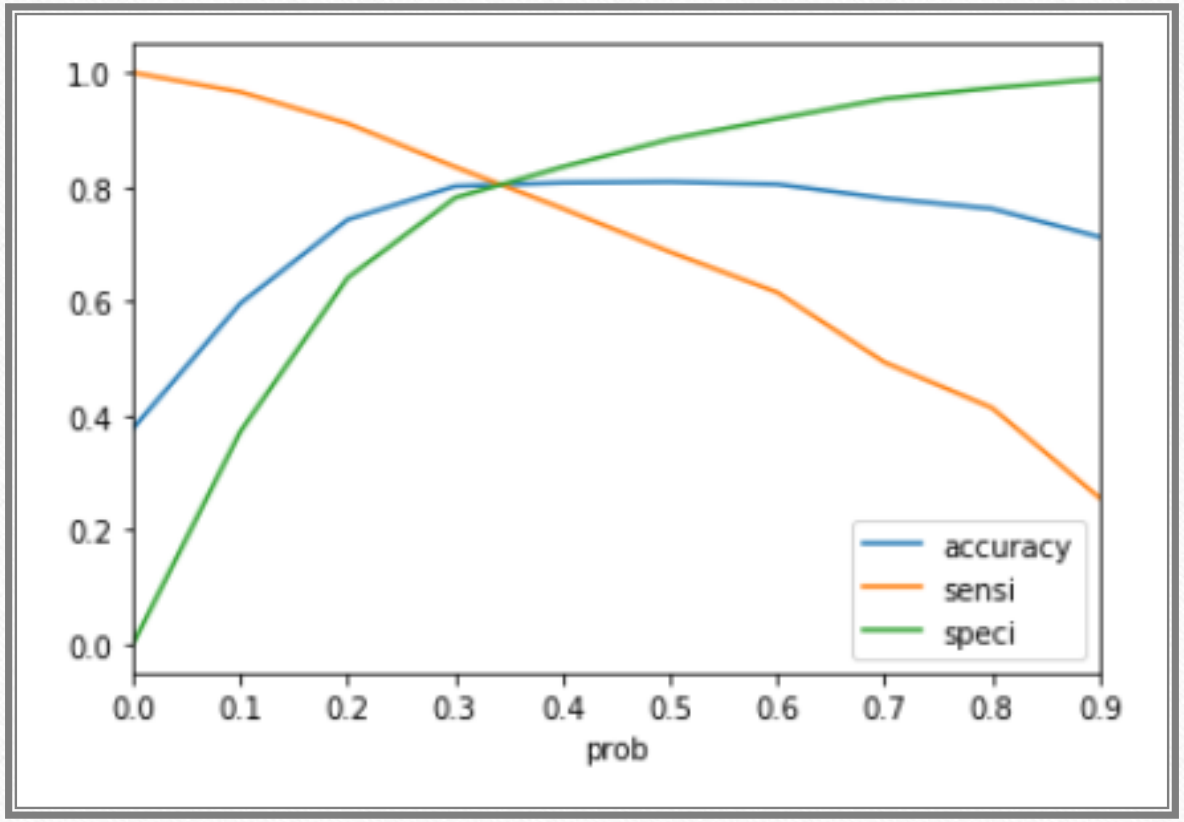


ROC Curve

- The ROC curve is showing the trade between sensitivity and specificity.
- An increase in sensitivity will be accompanied by the decrease in specificity.
- We can see, the curve is closer towards left hand border of ROC space. Thus it is showing the accuracy of our model.
- Area under our ROC curve is 0.88.

Optimal Cut-off point

- we can see that when the probability thresholds are very low, the sensitivity is very high and the specificity is very low.
- for longer probability thresholds, the sensitivity values are very low but the specificity values are very high.
- At around 0.3, the there metrics seem to be nearly equal with decent values.
- Hence we choose 0.3 as the optimal cut-off point.



Model Evaluation

Metrics	Train set	Test set	Final Model
Sensitivity	0.68	0.83	0.83
Specificity	0.88	0.78	0.78
False positive value	0.11	0.21	0.21
Positive predictive value	0.78	0.69	0.69
Negative predictive value	0.82	0.88	0.88

Predictions & Recommendations

- Priority should me on targeting working professionals as they are having higher rate of conversion.
- "Lead Source_Welingak Website", "Lead Origin_Lead Add Form" and "What is your current occupation_Working Professional" also has higher positive impact on conversion. Hence, more efforts should be kept to focus on individuals from these origin points.
- Olark chat is also pretty good lead source and more focus should be put towards this source for acquiring potential leads.
- Individual having phone conversations are having more inclination towards joining the course. Hence, more efforts can be made to reach out to potential leads for increasing the conversion rate.
- The more time the individual spends on the website, the more is the chance of conversion. Hence, more focus should be made in targeting the individuals with last notable activity of visiting the website.

Reference links

- Quick commands help:
  - www.hackerrank.com
  - www.geeksforgeeks.org
  - https:/pandas.pydata.org/pandas-docs
  - https:/stackoverflow.com/