X Education Lead Scoring Case Study

By

Pravin Pawar

Ashish Sharma

Business Objective

Objective

X Education company sell online course to industry professionals. On any given day, many professionals who are interest interested in the courses land on their website and browser for courses.

Company get leads from different sources.

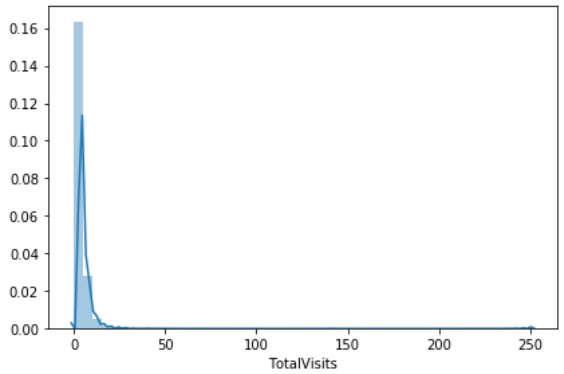
Once these leads acquired, employee from X education start making calls, writing emails, etc. Through this process some of the leads get converted while most do not. They typical conversion rate at X education is around 30%.

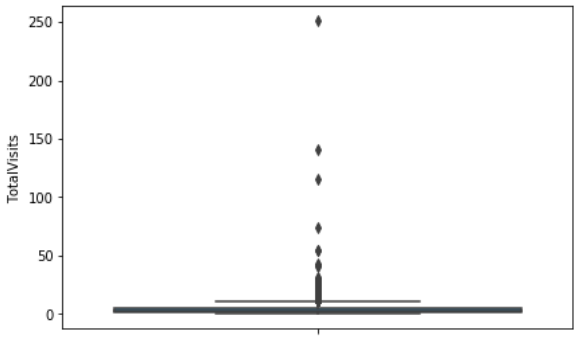
- Generate Lead Score for each leads.
- Predict Hot Lead (person with high chance of conversion) and Cold Lead (person with low chance of conversion)
- Target lead conversion is 80%, which means when sales team contact 100 Hot Leads predicted by model, at least 80 professional should convert/enroll for course.

Data Observation & Preparation

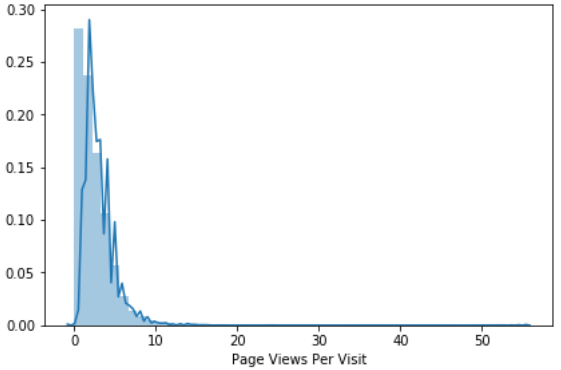
- Total 9240 observation and 37 features available in data set.
- There are no duplicate rows in data set
- Lot of Missing data in multiple columns, dropped such columns and rows using different criteria.
- Some of the features have outlier
  - Total Visits: 95% of data having less than 10 Total Visits
  - Page Views Per Visit: 95% of data have less than 6 page views per visit.
  - So will take 95% as upper cut off and remove all outlier above that.

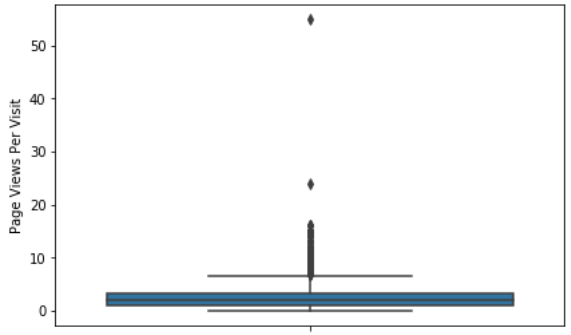
Total Visits





Page Views Per Visit

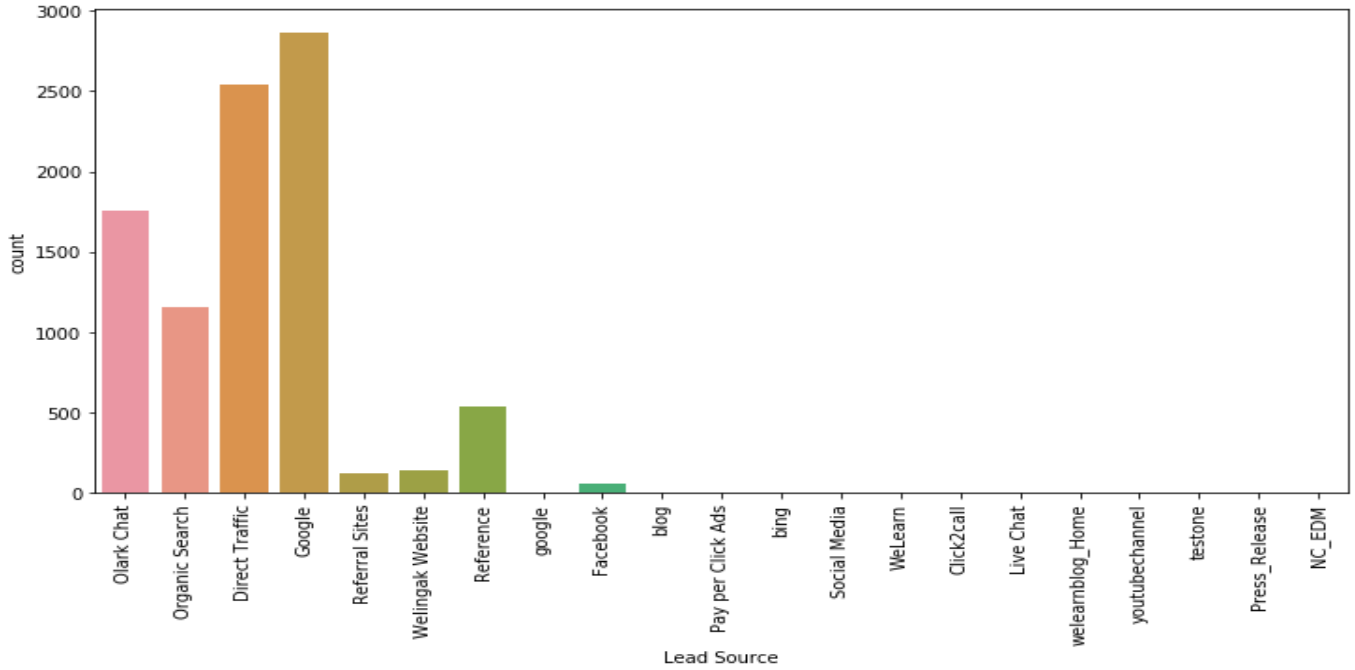


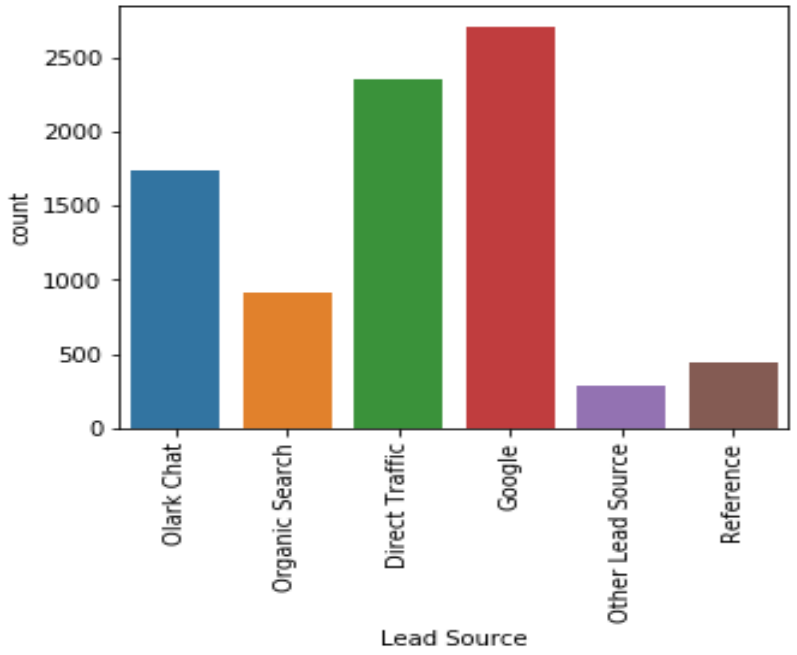


Analysis

Observation

- ❖ Some of feature variable have 10-20 different categories, creating so many dummy variable is not a good solution.
- ❖ Lead source feature variable have 21 categories

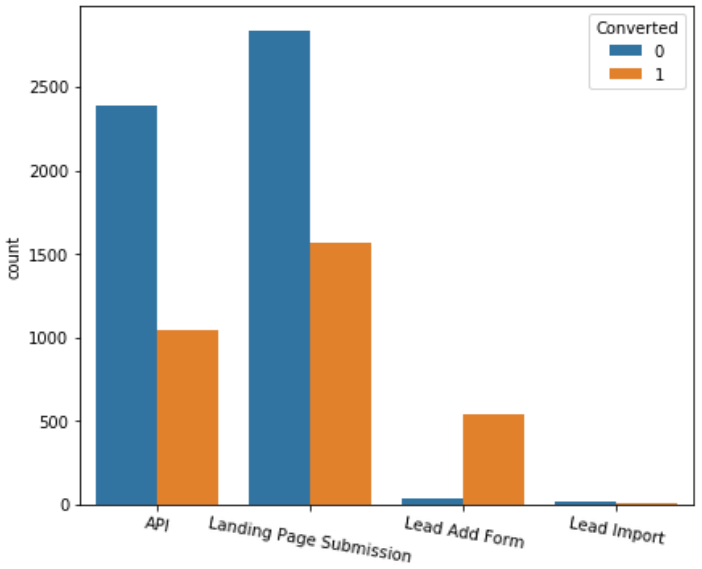




- ❖ Instead of analysing data on all categorical value, will merge less important categories.
- ❖ Same problem with other feature variable like Last Activity, and Last Notable Activity.

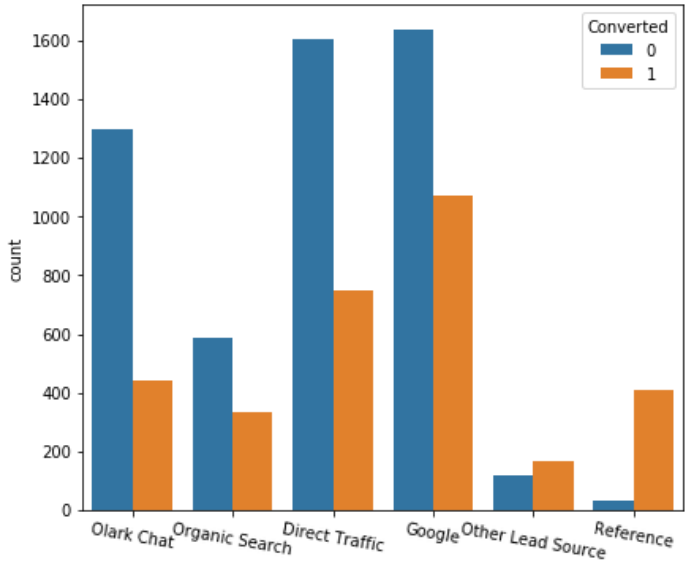
Analysis

Lead Origin



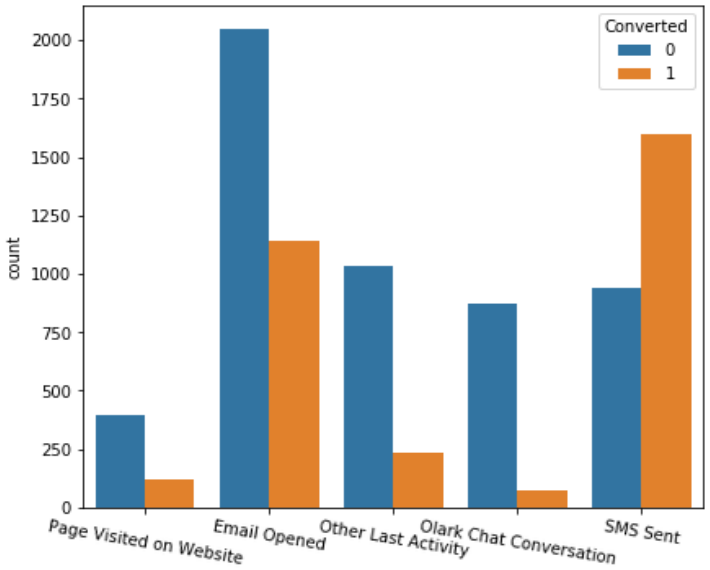
- More leads are generated through API & Landing Page Submission category.
- Out of that 40-50% leads are getting converted
- Where as for Lead Add Form category less lead generated, but almost 90% get converted.

Lead Source



- Lead frequency is more in Google , Direct Traffic and Olark Chat.
- Leads from Google have more chance of converting.

Last Activity

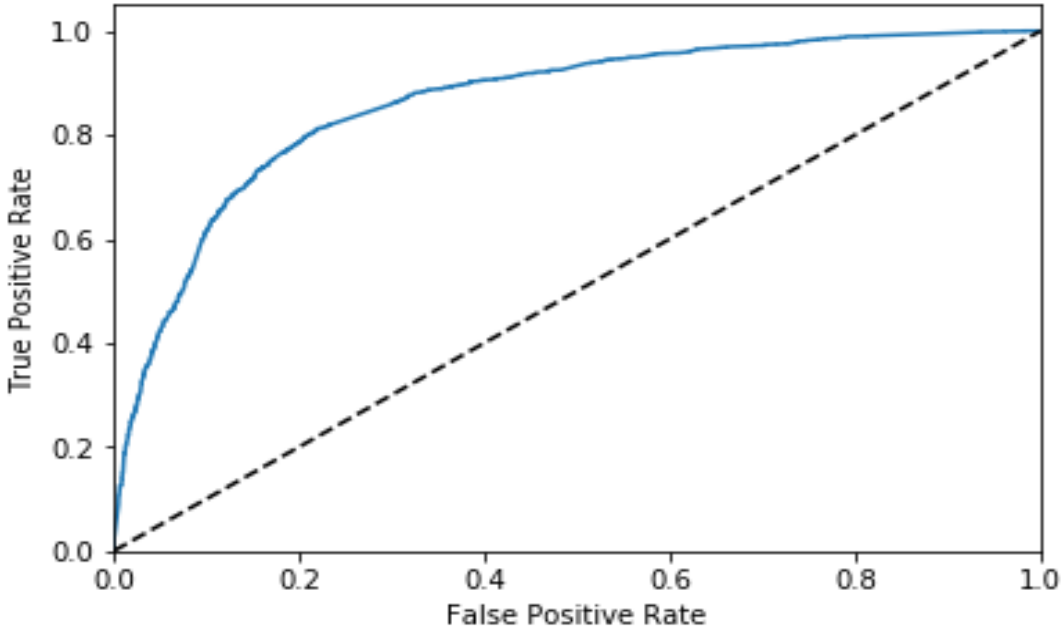


- Olark Chat conversation shows negative impact on lead conversion.
- Some trend shows in Lead Source -> Olark chat
- SMS sent category have approx. 70-75% conversion ration.

Analysis

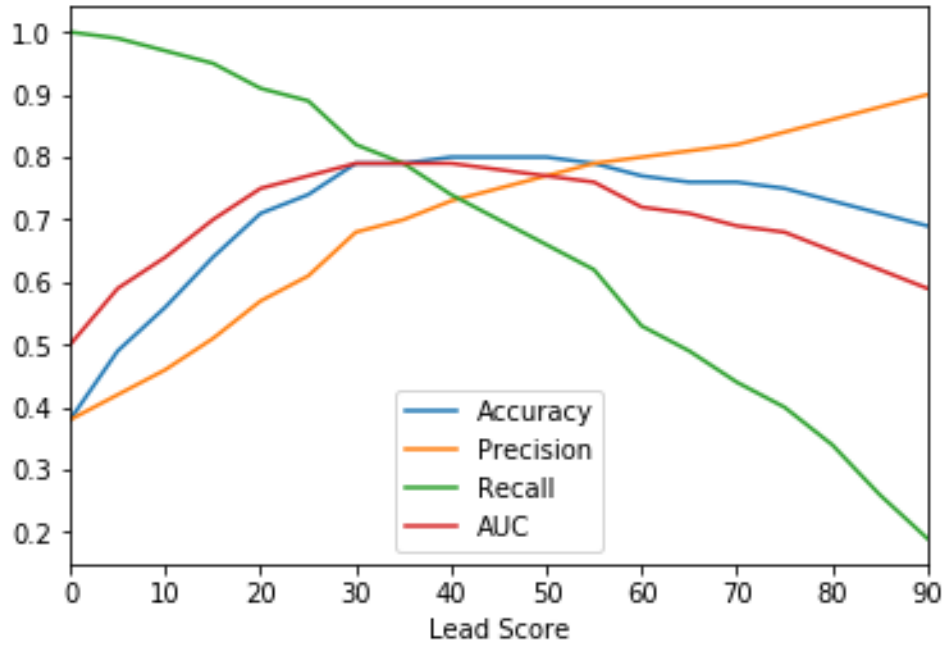
Logistic Regression:

- Built machine learning model using Logistic Regression algorithm
- Select 13 Feature for model building using RFE method.
- There are multiple metrics to evaluate Logistic Regression model.
- Based on business problem we are using Precision – Recall metrics as well as AUC curve.
- ROC curve is away from 45-degree diagonal line and it's closer to Left & Top border which is indicator of strong model.
- Using Precision - Recall metrics will choose cut off with High Precision Value, which should be more than 80.



Analysis

Logistic Regression: Lead Score Cut Off



Lead Score	Accuracy	Precision	Recall	AUC
60.0	0.77	0.80	0.53	0.72
65.0	0.76	0.81	0.49	0.71
70.0	0.76	0.82	0.44	0.69
75.0	0.75	0.84	0.40	0.68
80.0	0.73	0.86	0.34	0.65
85.0	0.71	0.88	0.26	0.62
90.0	0.69	0.90	0.19	0.59

- So we can take 60 as our lead score cut off value.
- Using this cut value mark Hot Lead (probability more than 60) Or Cold Lead (probability less than 60)

Conclusion

- Top 3 Features which signify lead in Hot Or Cold are: **Lead Origin, Last Activity, Lead Source.**
- Top 3 categories which signify lead in Hot Or Cold are**: Lead Add Form(from Lead Origin), Olark Chat Conversation (from Last Activity) and Olark Chat (from Lead Source).**
- **X Education company** should target leads / professionals from above mentioned categories.