Lead Score Case Study

Group members

1. Amoolya K
2. Amit Dubey
3. Amitanshu

Problem Statement

X Education sells online courses to industry professionals. 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.

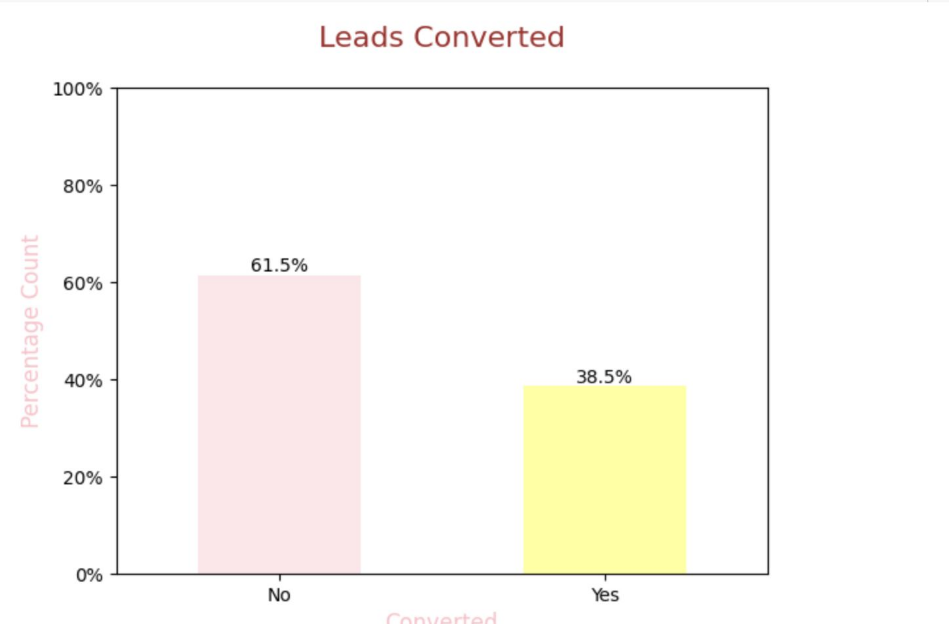
Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Data Cleaning

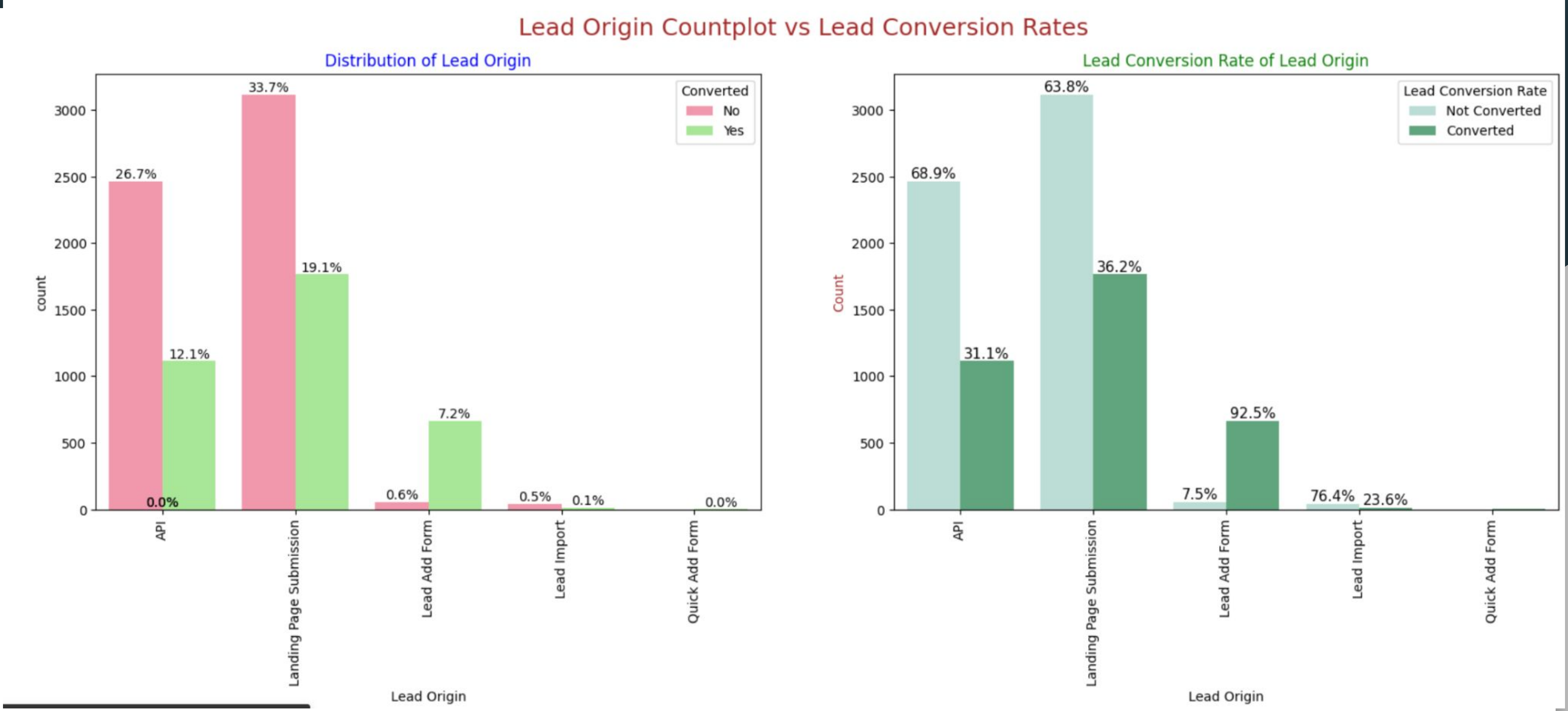
- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modelling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

EDA



Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)

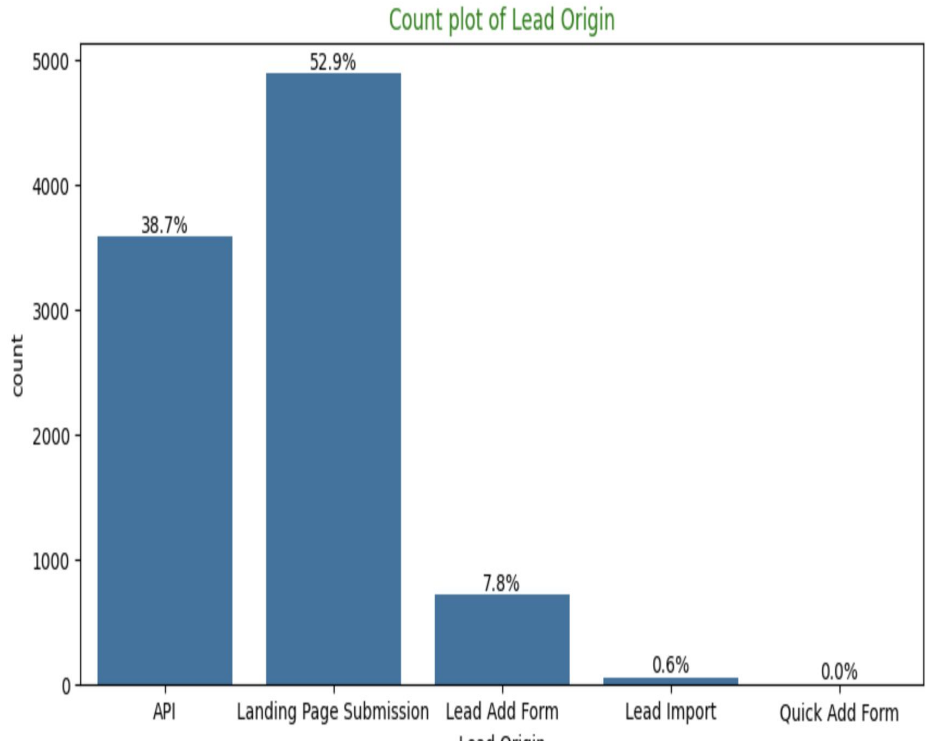
- While 61.5% of the people didn't convert to leads. (Majority)

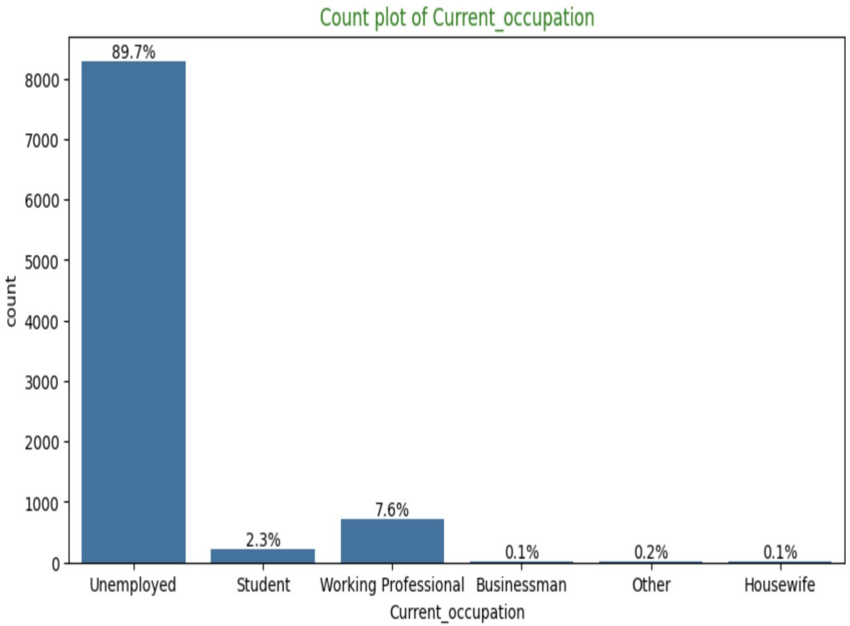


Lead Source: 58% Lead source is from Google & Direct Traffic combined.

Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities.

EDA


Count plot of Lead Origin

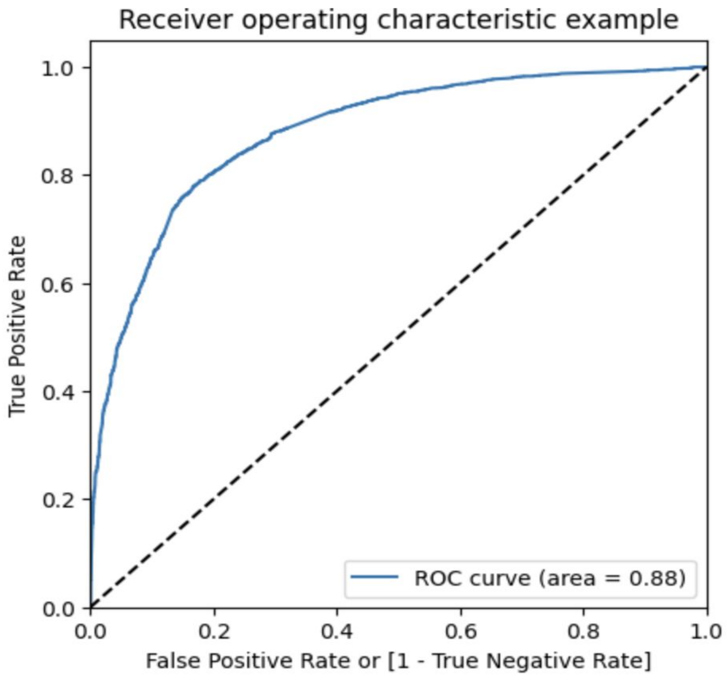

Count plot of Current_occupation

Data Model

ROC Curve – Train Data Set

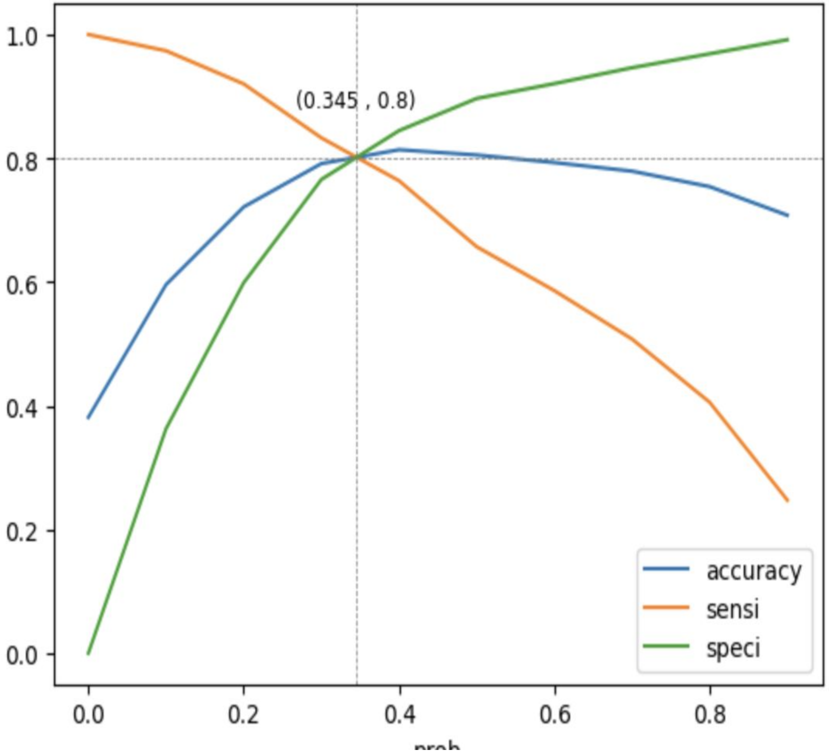
● Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.

● The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



It was decided to go ahead with 0.345 as cutoff after checking evaluation metrics coming from both plots

Confusion Matrix & Evaluation Metrics with 0.345 as cutoff



Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spent on the Website.
- Total number of visits.
- When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website

- When the last activity was:
  a. SMS
  b. Olark chat conversation

- When the lead origin is Lead add format.
- When their current occupation is as a working professional.
- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.