LEAD SCORE CASE STUDY

SUBMITTED BY:

SUSHIL SANTOSHRAO MULI

YASHI NAIK

MADHUREMA NAG

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, 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.

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

GOALS OF THE CASE STUDY

- There are quite a few goals for this case study:

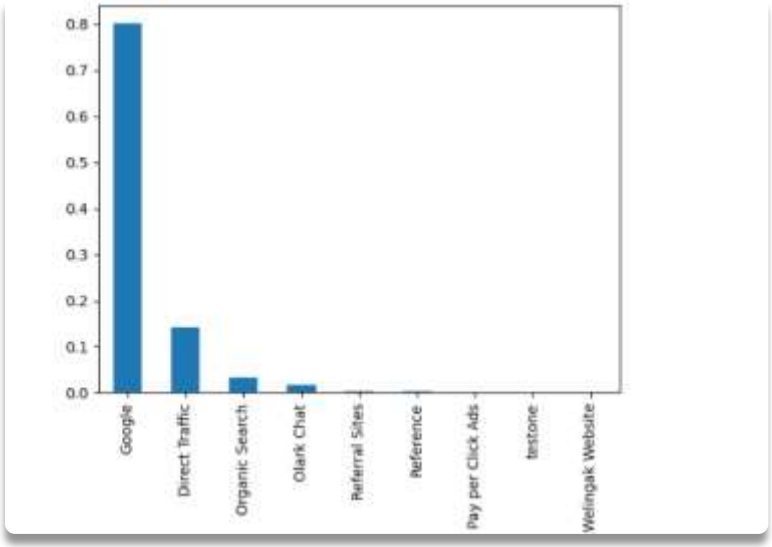
1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

STRATEGY

- Import data
- Reading and understanding data
- Data cleaning
- EDA
- Test train split
- Dummy variable creation and feature scaling
- Building a logistic regression model
- Model evaluation-specificity & sensitivity or precision recall
- Predictions on the test set

Exploratory data analysis

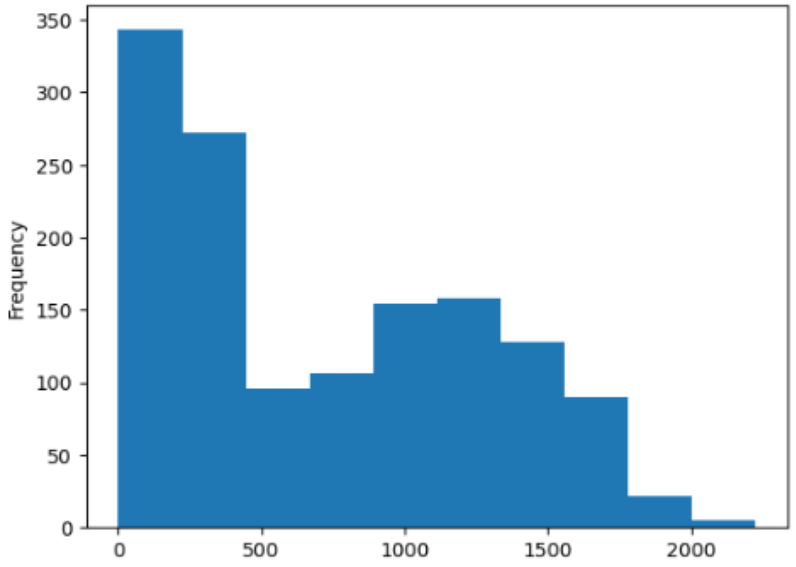
UNIVARIATE ANALYSIS



Lead source (count plot)

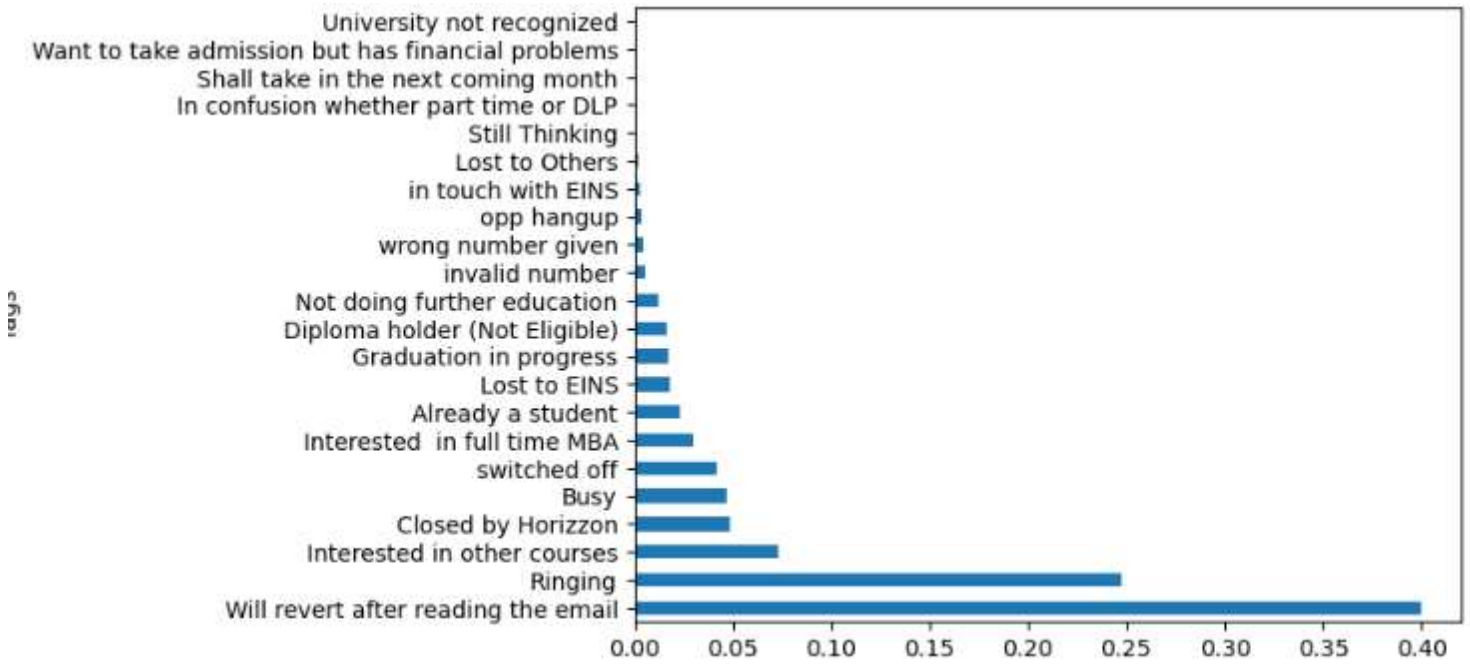
- # most of the lead source was found in google only

UNIVARIATE ANALYSIS(cont)



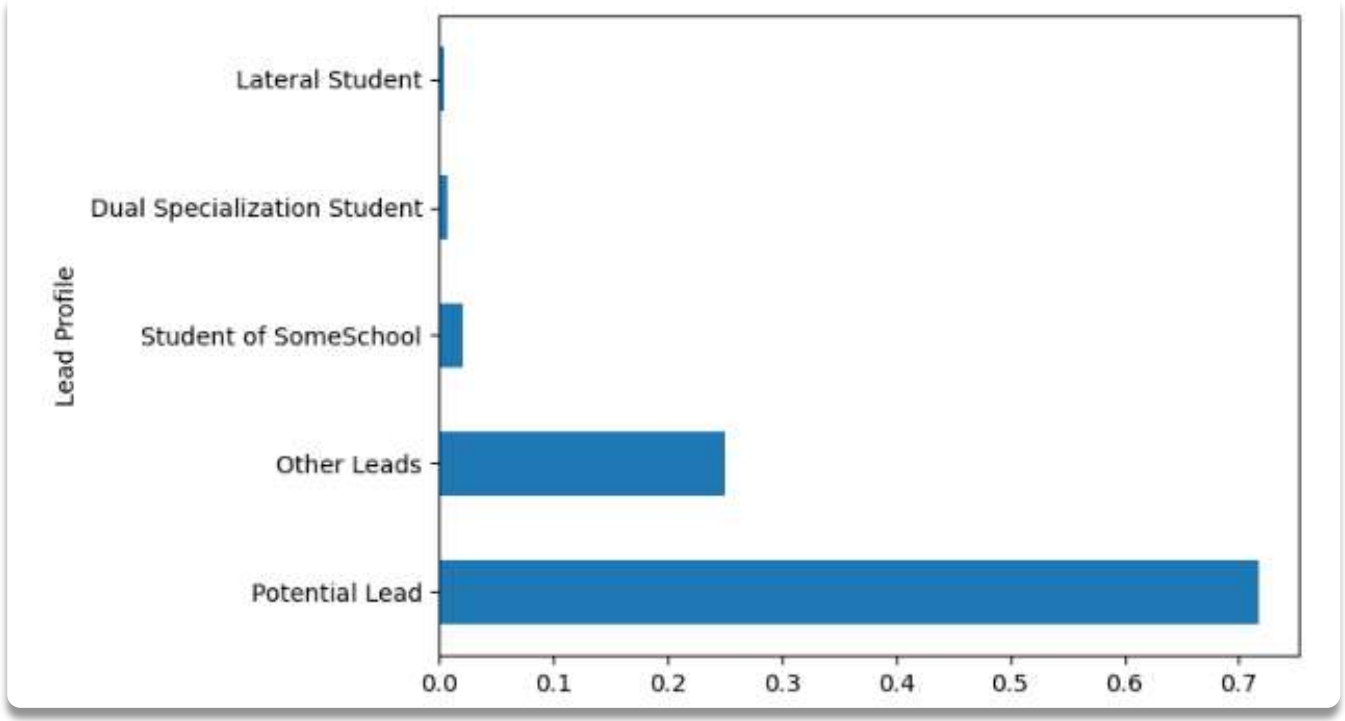
- Total time spent on website (count plot)
- # Max Number of people who spent time on website fall between 0 to 500

UNIVARIATE ANALYSIS(cont)



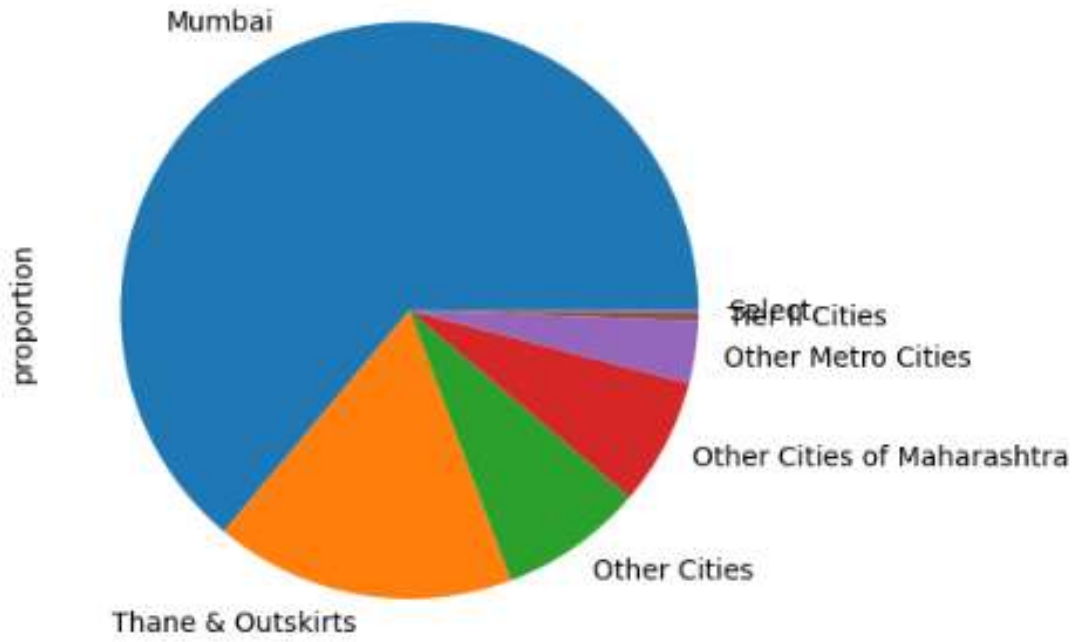
- Tags (count plot)
- almost 5% of 1400(approx) people are about to join this programe

UNIVARIATE ANALYSIS(cont)



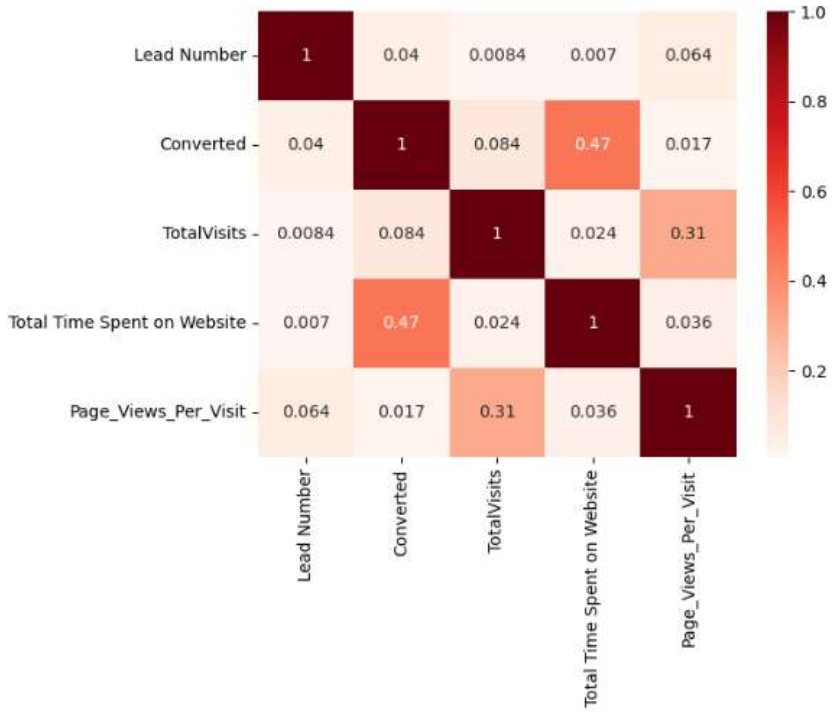
- Lead profile (count plot)
- almost 27% people are willing to join this programme

UNIVARIATE ANALYSIS(cont)



- City (count plot)
- #Most of the customers are from Mumbai.

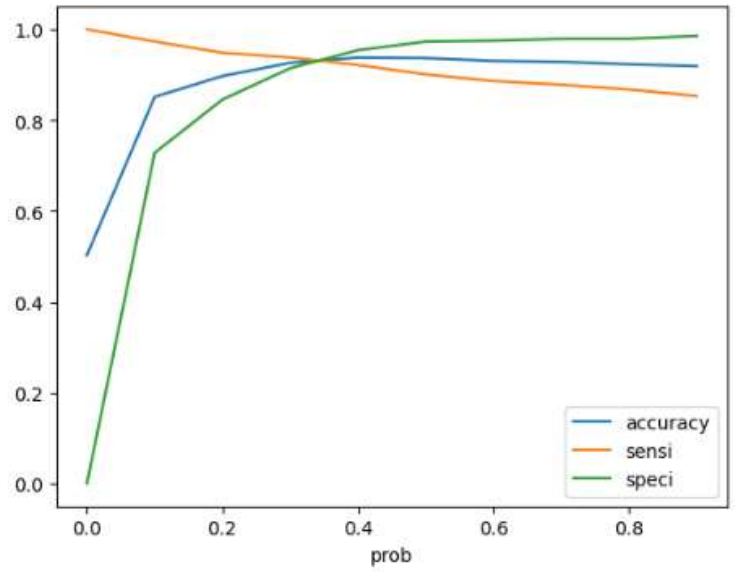
Bi-variate analysis



- Total time spent on website has high positive correlation with converted.
- Pages views per visit has high correlation with total visits

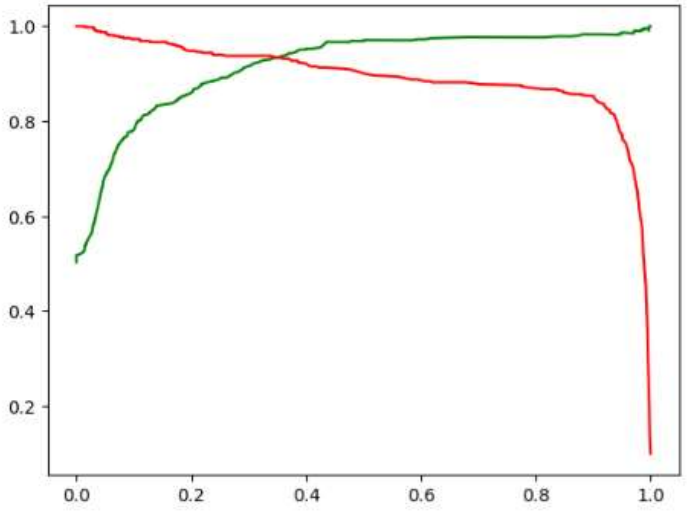
MODEL EVALUATION-TRAIN SET

From the curve above, 0.35 is the optimum point to take it as a cutoff probability



precision_recall_curve vs .Converted_Prob

The graph depicts optimal cutoff at 0.37 based on precision & recall



MODEL EVALUATION
Sensivity & specificity on Test Dataset

- Accuracy-0.91
- Sensivity-0.88
- Specificity-0.94

RESULT

- Accuracy, sensitivity and specificity values of test set are close to training set
- Accuracy, sensitivity and specificity values of training set are –

Accuracy-0.93

Sensivity-0.94

Specitivity-0.93

- Accuracy, sensitivity and specificity values of test set are-

Accuracy-0.91

Sensivity-0.88

Specitivity-0.94

- Hence, overall this model proves to be accurate