LEAD SCORING CASE STUDY

Upgrad Assignment

Problem Statement

An education company named X Education sells online courses to industry professionals. although X Education gets a lot of leads, its lead conversion rate is very poor.

X Education has appointed you to build a model to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Business Objective

- 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

What have we covered in this case study? / Strategy

1) Reading, cleaning and understanding the data

2) Preparing the data for modeling

3) Splitting the data and Feature Scaling

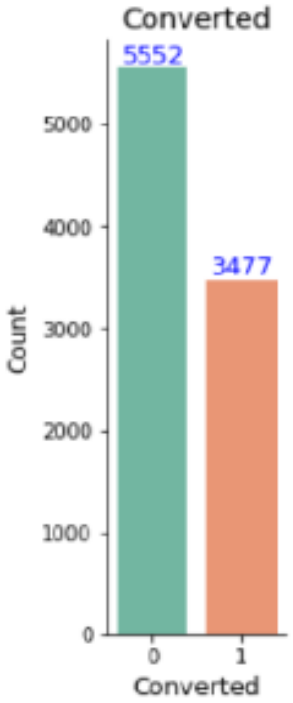
4) Model Building

5) Assigning a Lead Score

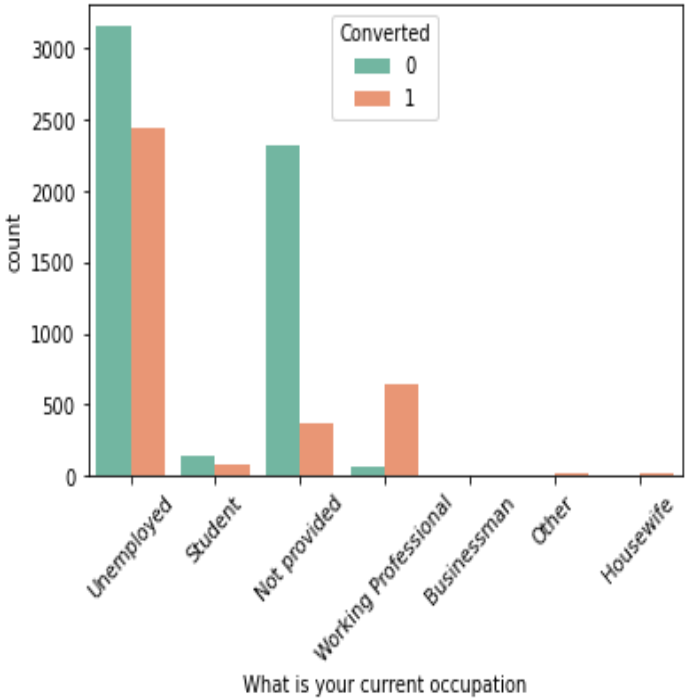
6) Running metrics such as Accuracy, Sensitivity Specificity, Precision and Recall on Train and Test Sets

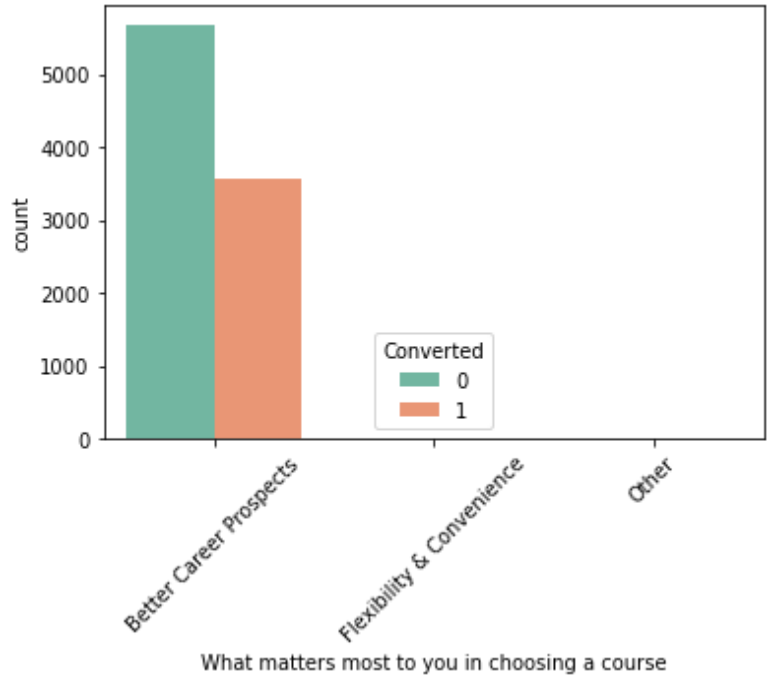
Exploratory Data Analysis

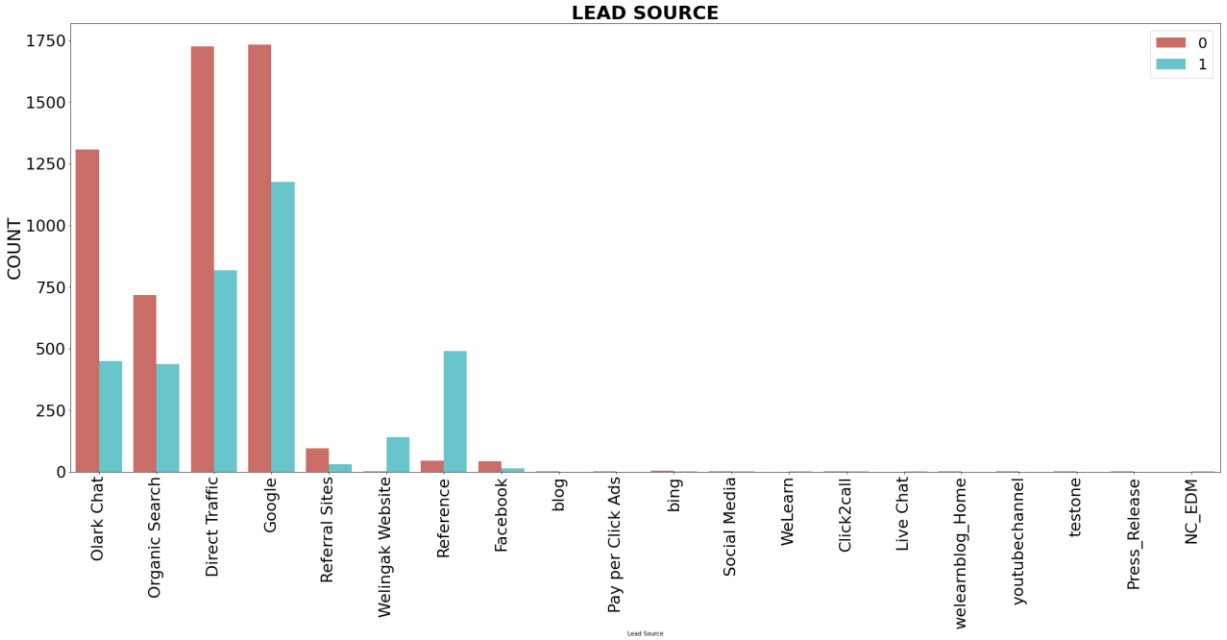
- We saw a 38% conversion rate

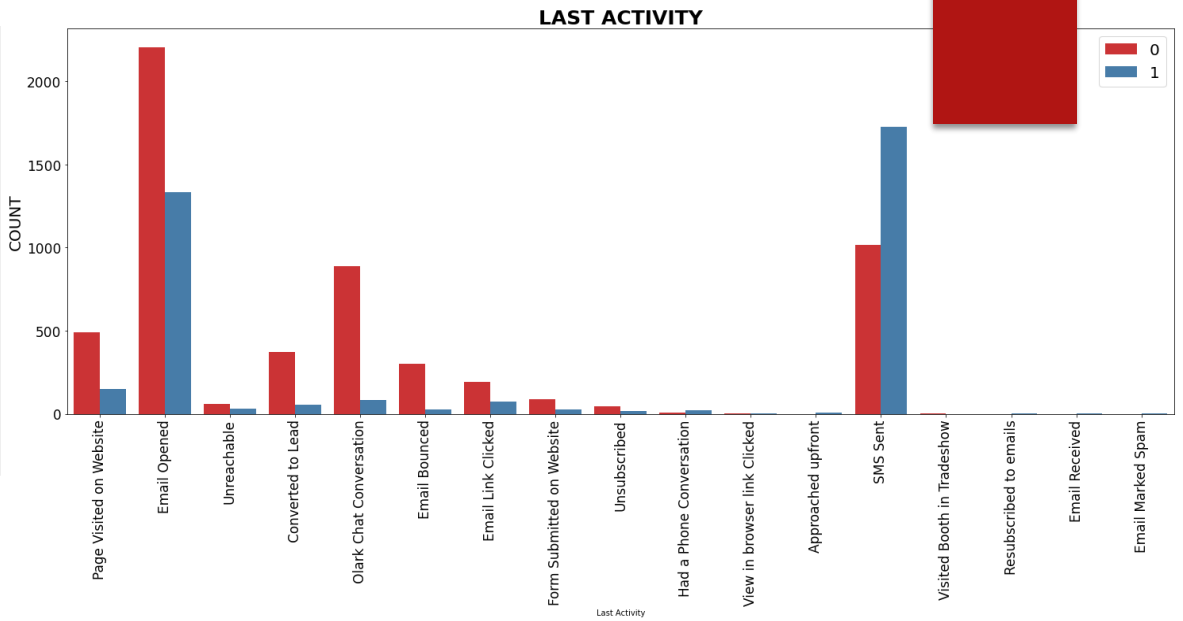


Current Occupation vs What matters in choosing a career
(Target variable – Converted)





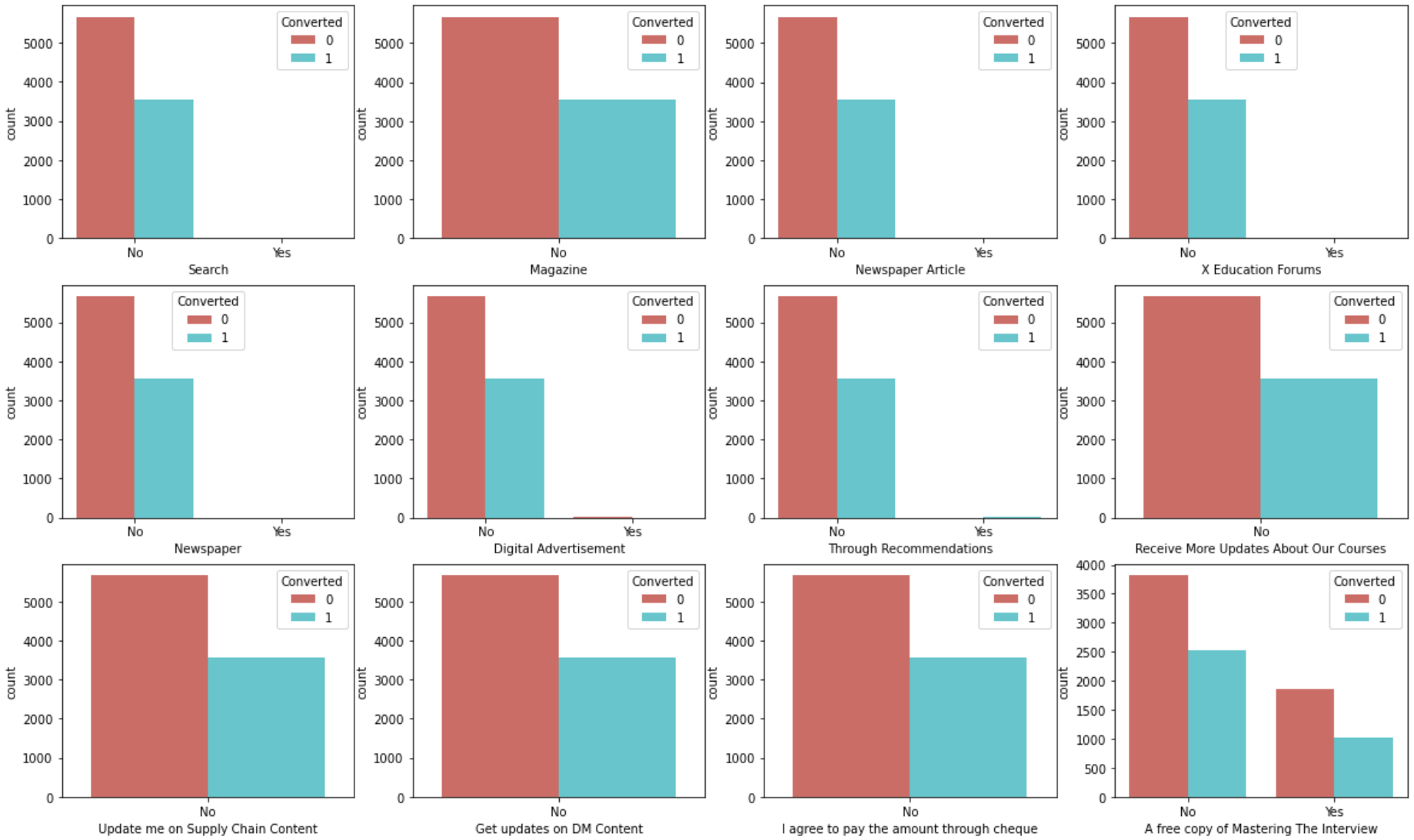




Last Activity vs Lead Source

Last Activity: Email Opened and SMS sent are notable labels
Lead Source: Clubbed all low frequency labels to 'Others'. Lot of lead generating labels here

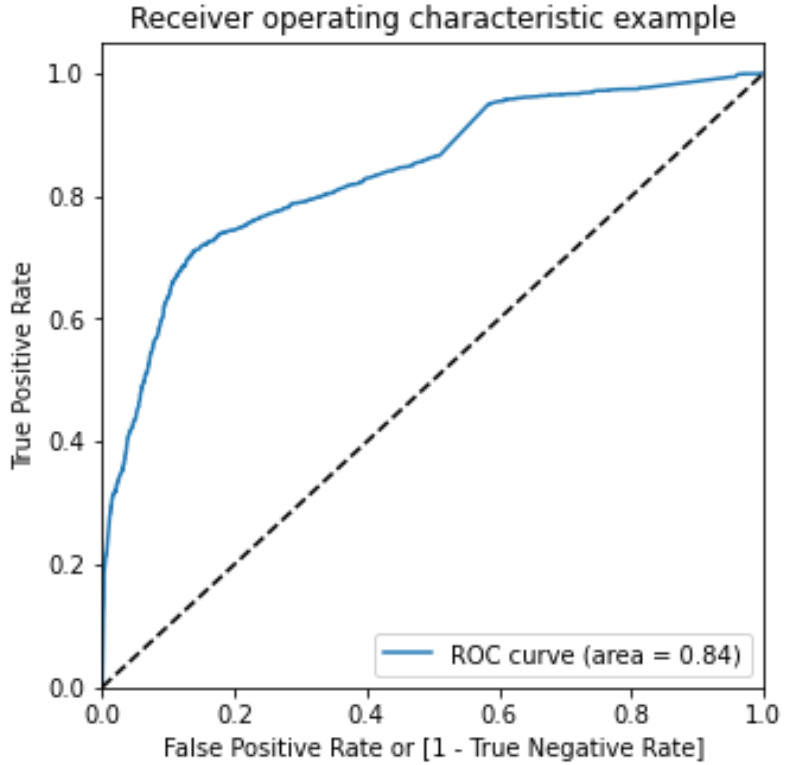
A lot of variables (especially categorical) are skewed and thus dropped.



Variables impacting the conversion rate

- Do Not Email
- Total Time Spent on Website
- Lead Source_Facebook
- Lead Source_Others
- Lead Source_Reference
- Lead Source_Welingak Website
- What is your current occupation_Housewife
- What is your current occupation_Student
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional

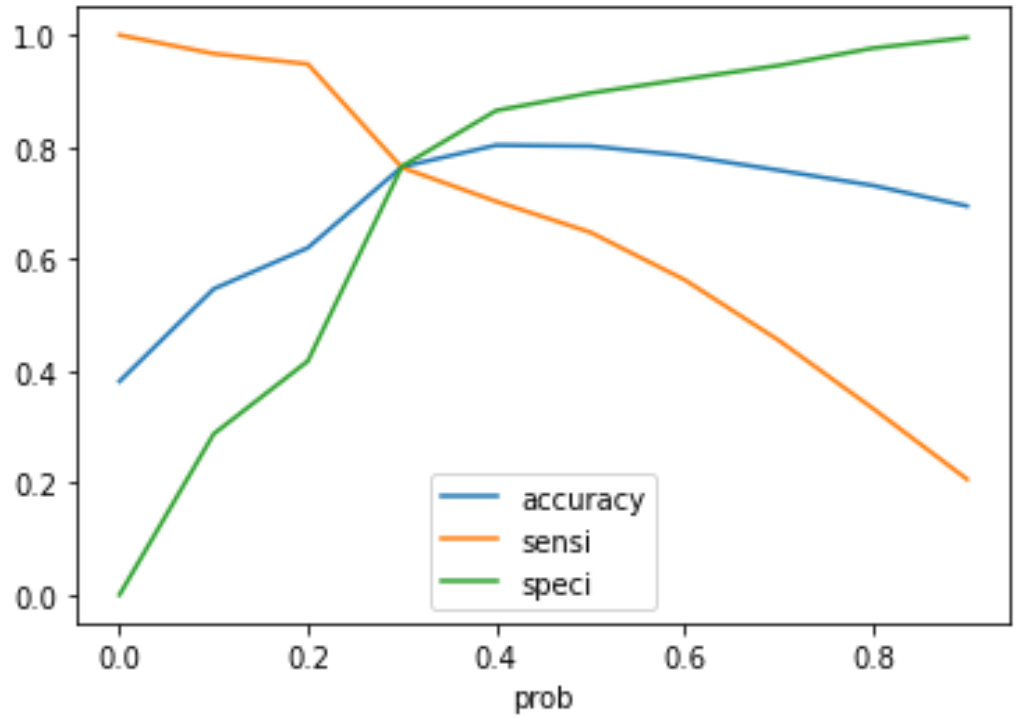
Model Evaluation on Train Data Set



Accuracy - 80.14 %

Sensitivity - 64.76 %

Specificity - 89.63 %



ROC curve close to 1(0.84) , concluding it a good model

Model Evaluation on Test Set

- Accuracy : 69.11%
- Sensitivity :81.64%
- Specificity : 60.94%
- Precision : 57.71%
- Recall :81.64%
- **Final Prediction: 83.01%**

Conclusion and Recommendations

Accuracy, Sensitivity and Specificity values of test set are approximately closer to the respective values calculated using trained set.

Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%. Meeting the CEO's requirement!

Would recommend to focus on the below features which contribute towards the probability of a lead getting converted are :

- What is your current occupation_Working Professional
- Total Time Spent on Website
- Lead Source_Reference
- Lead Source_Welingak Website