LEAD SCORING CASE STUDY

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 us to build a model wherein we 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

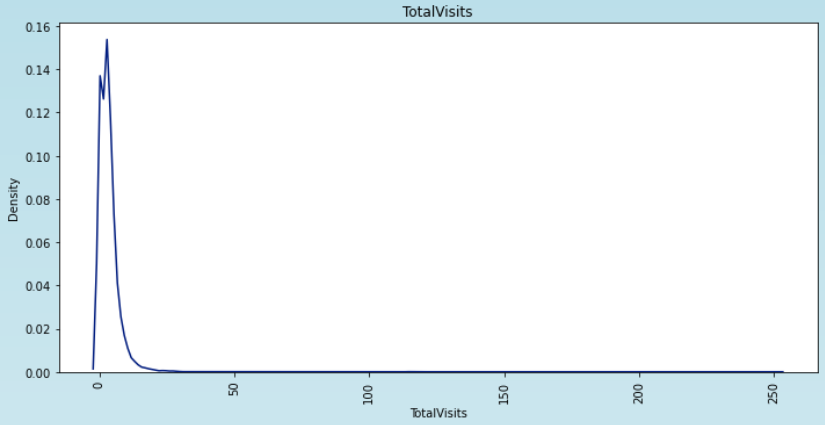
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

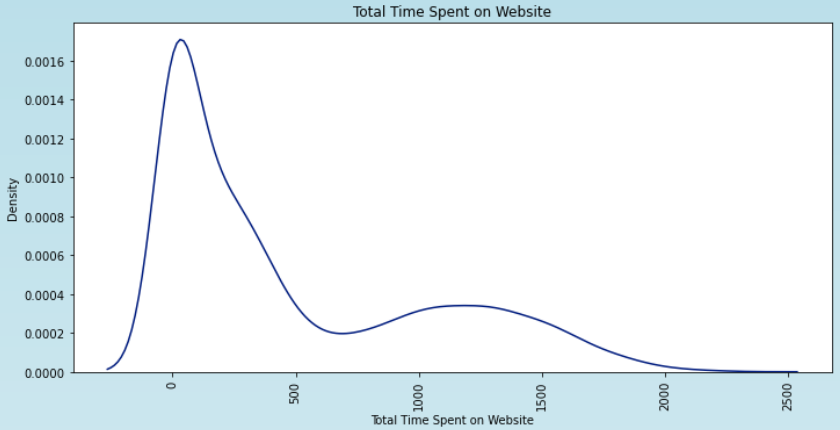
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

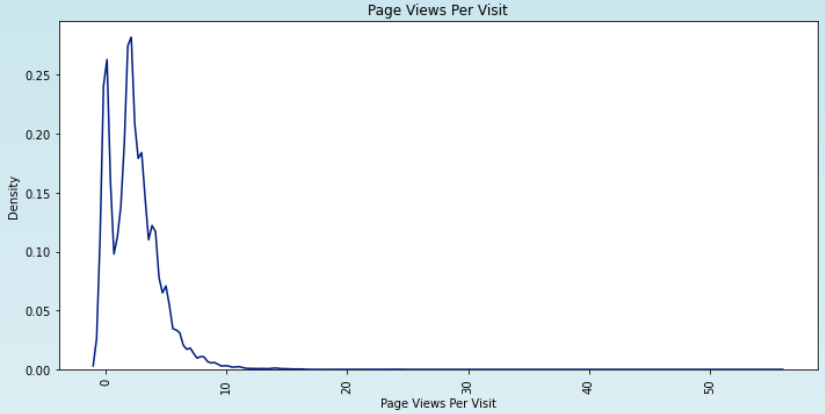
Strategy to solve the problem

- Importing and understanding the data.
- Cleaning and preparation of the data.
- Exploratory data analysis – Univariate and Bivariate
- Outlier Handling
- Creating dummy variables
- Train-Test Split
- Feature Scaling
- Feature Selection using RFE
- Model Building
- Creating Prediction
- Model Evaluation
- Optimising Cut Off (ROC Curve)
- Prediction on Test Set

EXPLORATORY DATA ANALYSIS



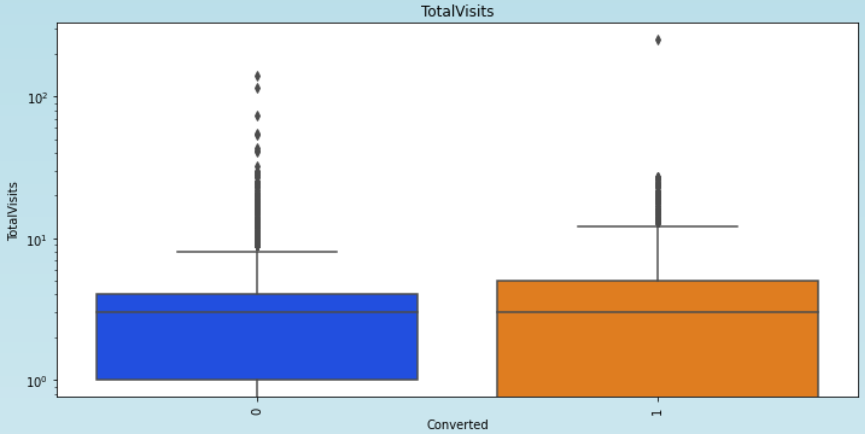


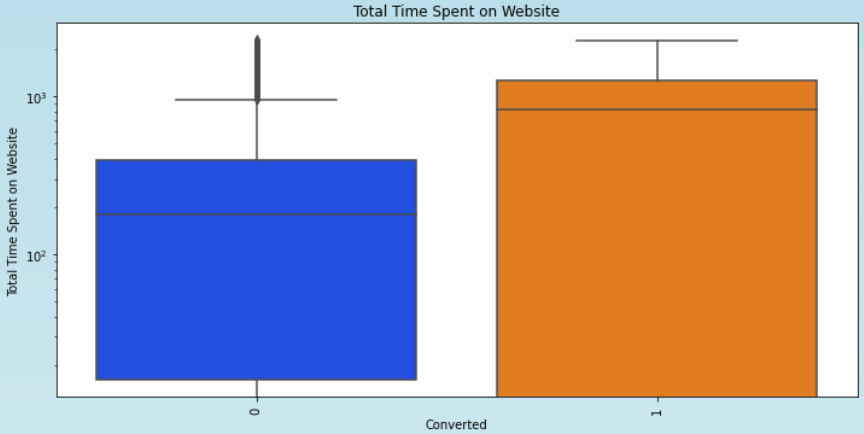


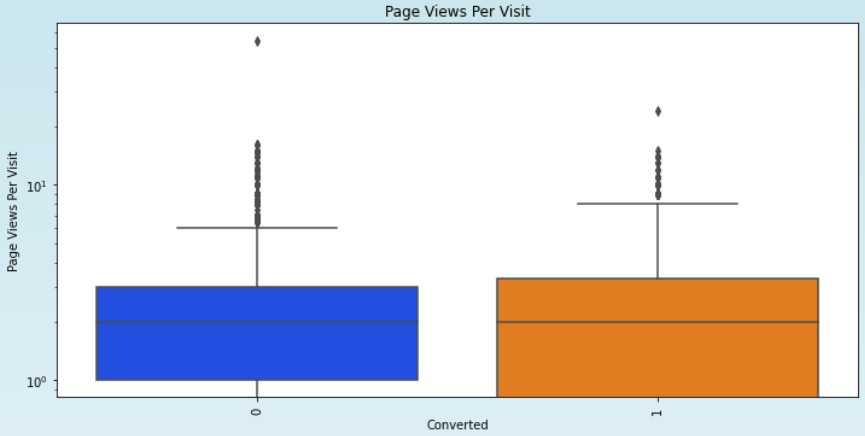
Uni-variate Analysis - Numerical values

- The max probability for TotalVisits is found to be around 15-20. It increases initially but decreases further.
- The max probability for PageViewsPerVisit is found to be around to be 3-5
- The probability of time spent is found to be high for time between 0-300 seconds and decreases further.

EXPLORATORY DATA ANALYSIS



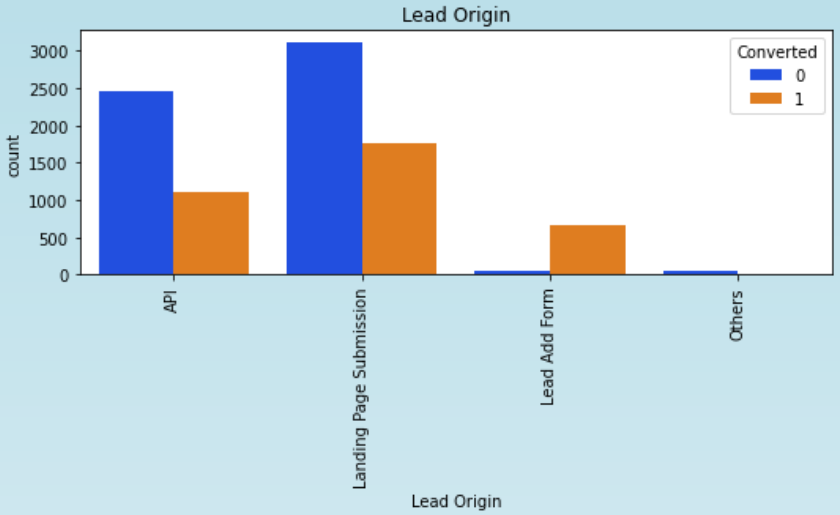


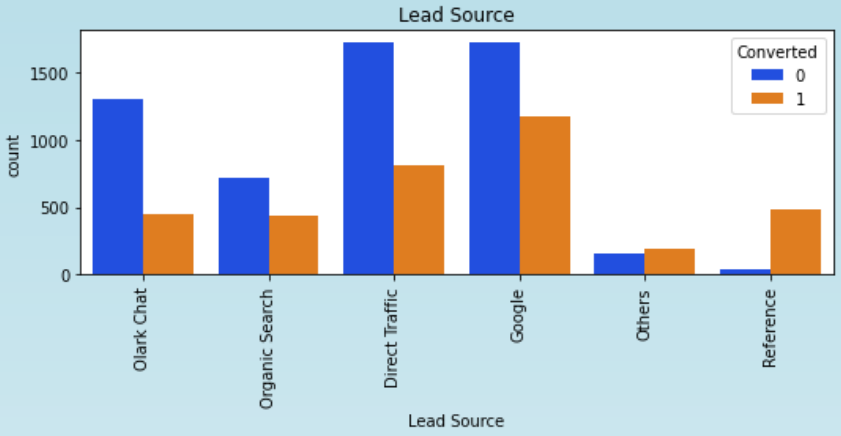


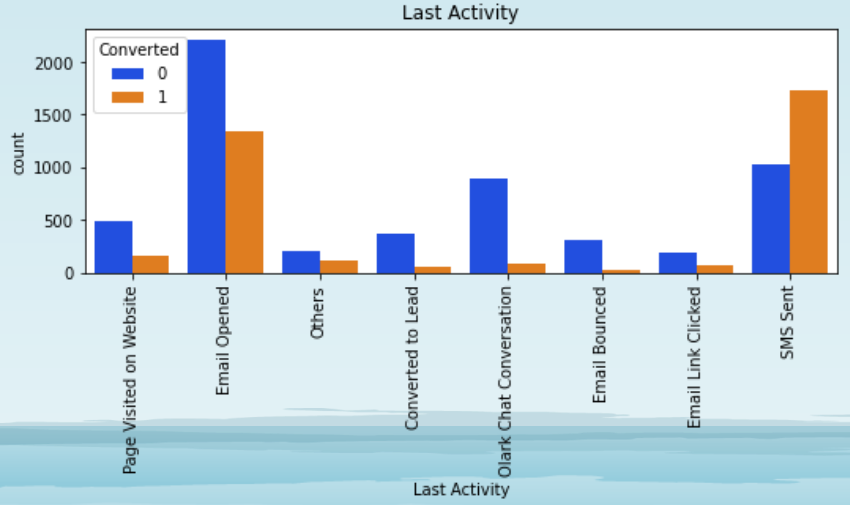
Bi-variate Analysis - Numerical values

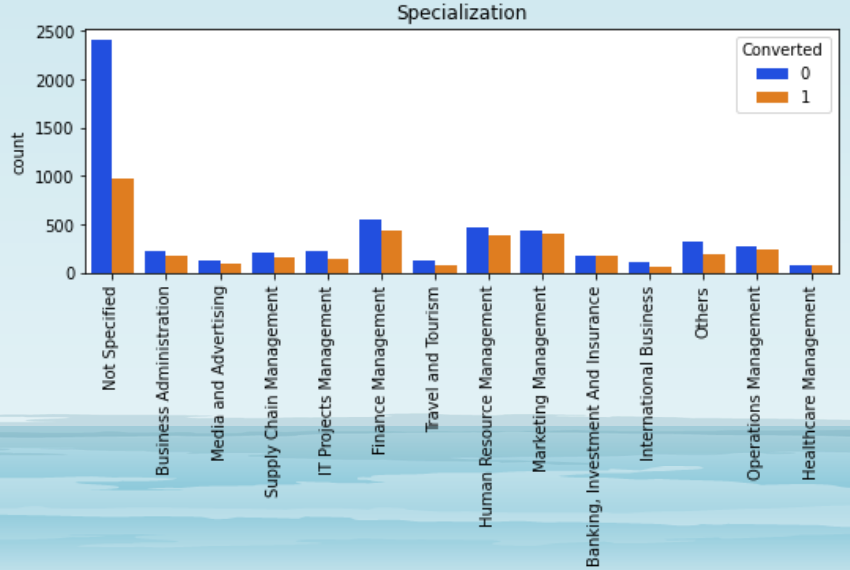
- The mean is found to be higher in case of Converted people rather than non-converted people.
- The average page views for both converted and non converted is found to be the same.
- The average total visits for both converted and non converted people is found to be the same.

EXPLORATORY DATA ANALYSIS

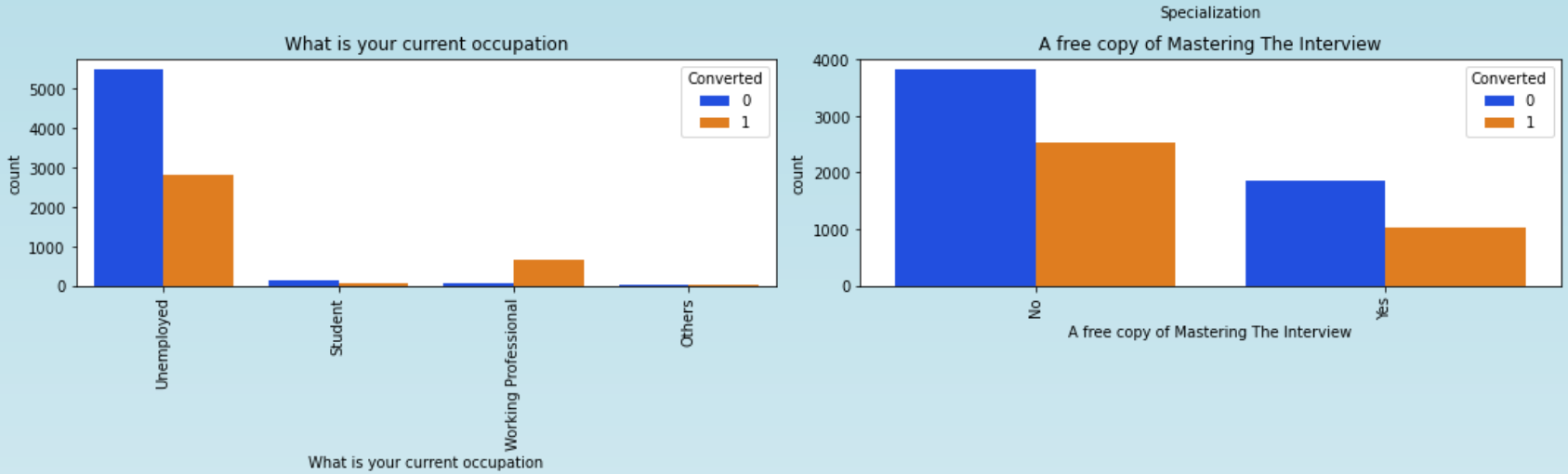








EXPLORATORY DATA ANALYSIS



Univariate analysis for Categorical data

- The percentage of Converted people is found to be greater for Landing Page Submission. We can also see that if Lead source is Add Form, the ratio of lead conversion is very high(almost not converted is very less).
- Google is found to be the important source for Lead Conversion
- We need to target people via Emails and SMS as it is found that the probability of response in case Converted leads is found to be higher.
- We cannot infer much about conversion rate from specialisation as people who do not select any specialisation can also be converted to a lead. But the ratio of non converted leads is higher than converted ones if they didn't choose specialisation.
- It is clearly visible from the graph that we need to target the Unemployed and Working Professional to get a higher conversion rate. The ratio of conversion rate is higher than not converted people for working professionals.
- People usually do not subscribe for a free copy of mastering the interview.

CONCLUSION

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

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

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

5. When the lead origin is Lead add format.

6. 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.