

Lead Scoring Case Study

TRAPTI AGARWAL

DS C29 Cohort

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

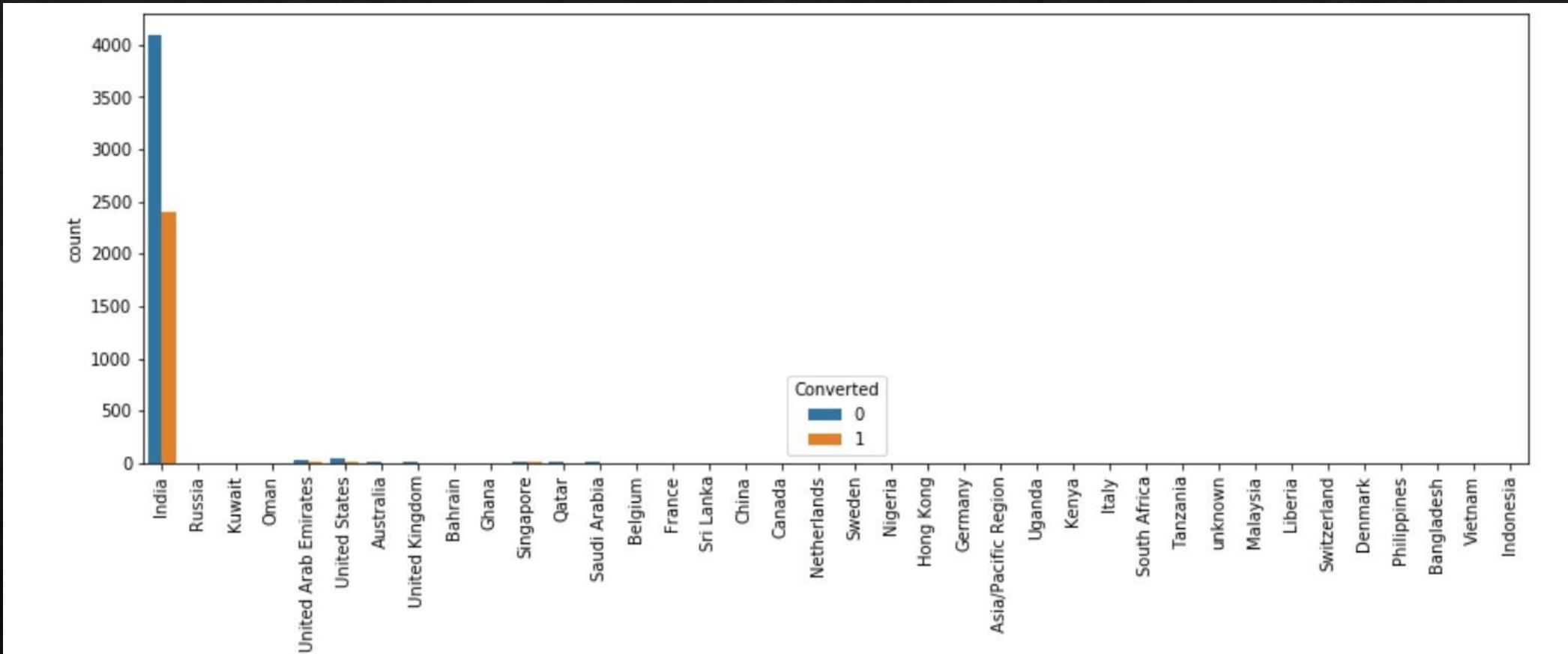
Business Goal

- ❖ X Education want to develop a model to 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 higher lead score have a higher conversion rate and the customers with lower lead score have a lower conversion rate.
- ❖ The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goal of the Case Study

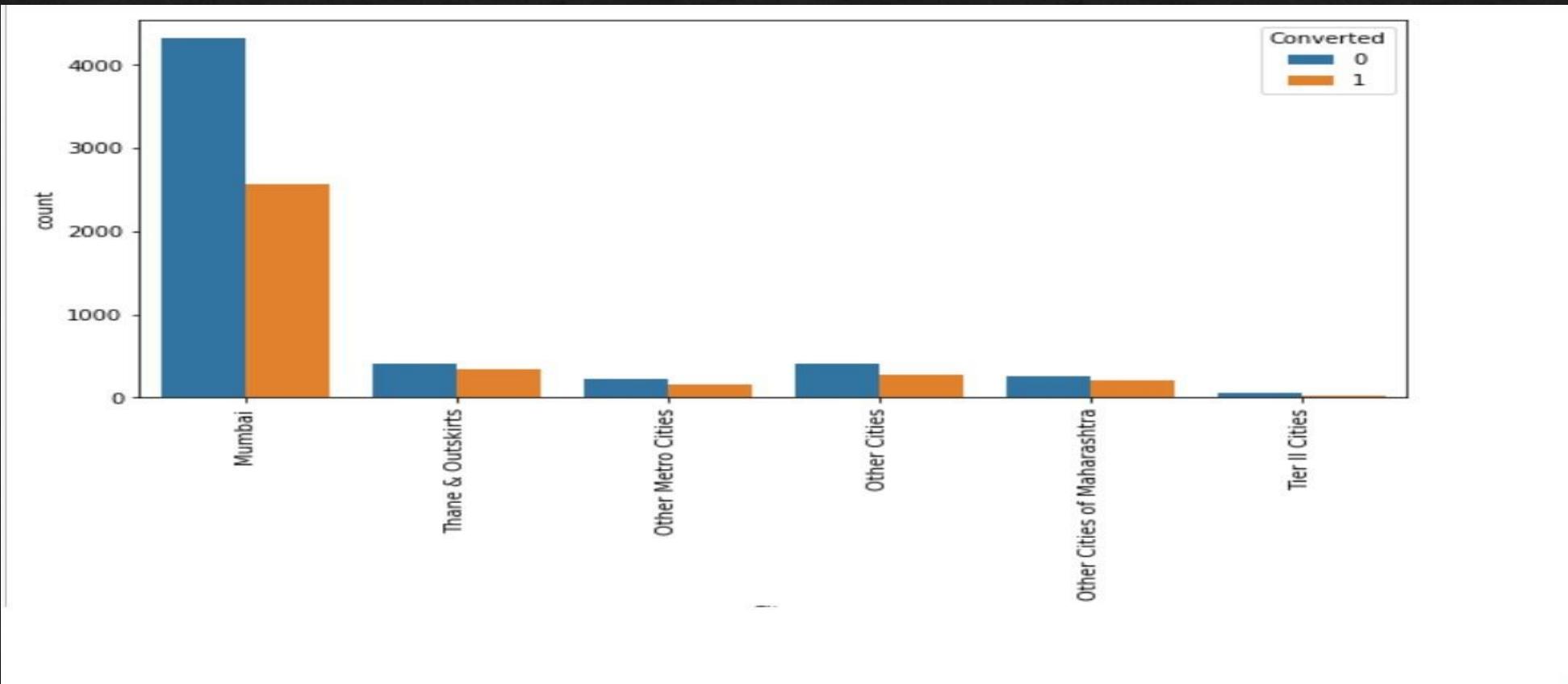
- ❖ To build a logistic regression model and assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- ❖ A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold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

Univariate Analysis- Country



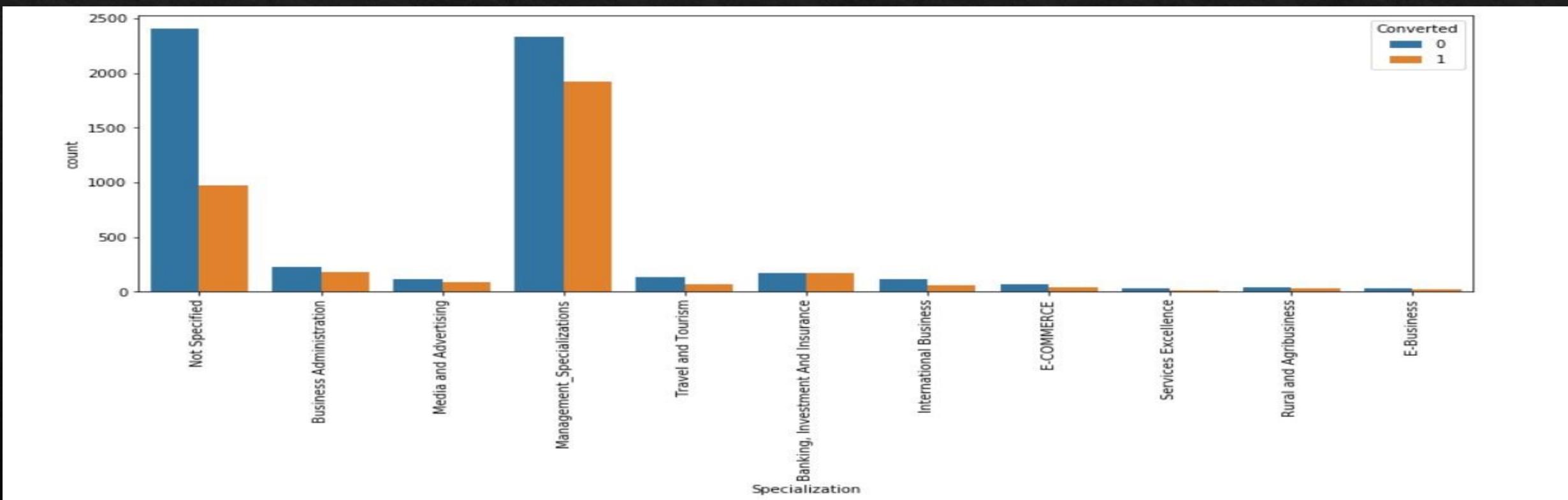
Nearly 90% of the leads are from India. The conversion rate is very low and we will address this issue in our analysis

Univariate Analysis- City



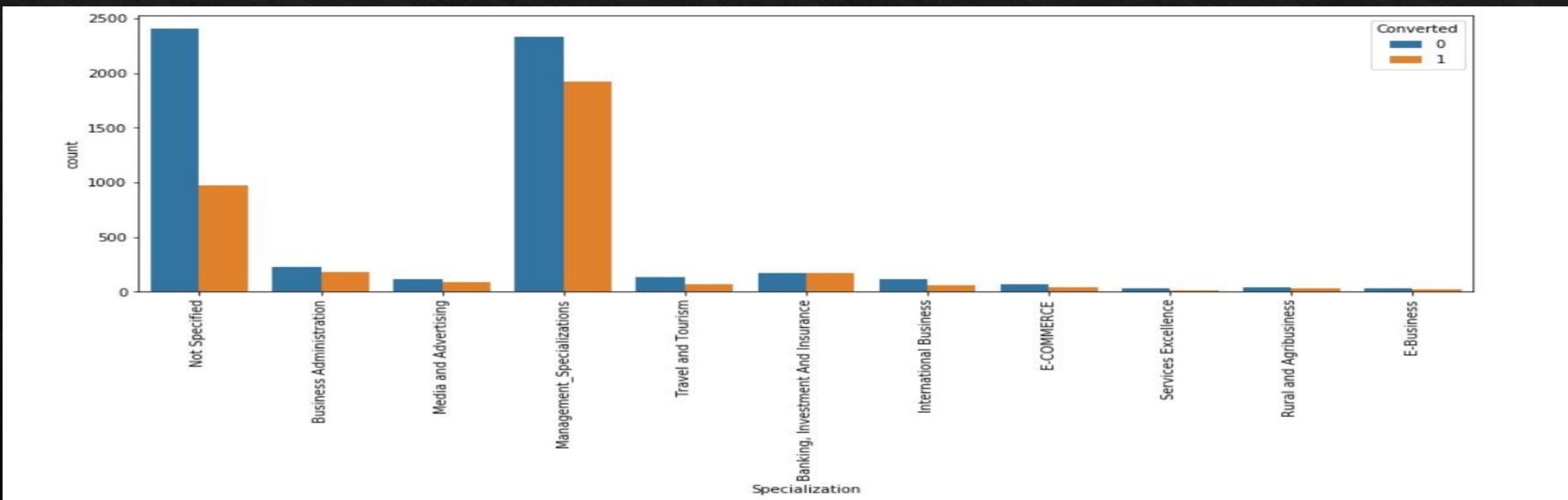
We can see that most of the leads are from Mumbai with ~ 30% conversion rate compared to other cities in India.

Univariate Analysis - Specialization



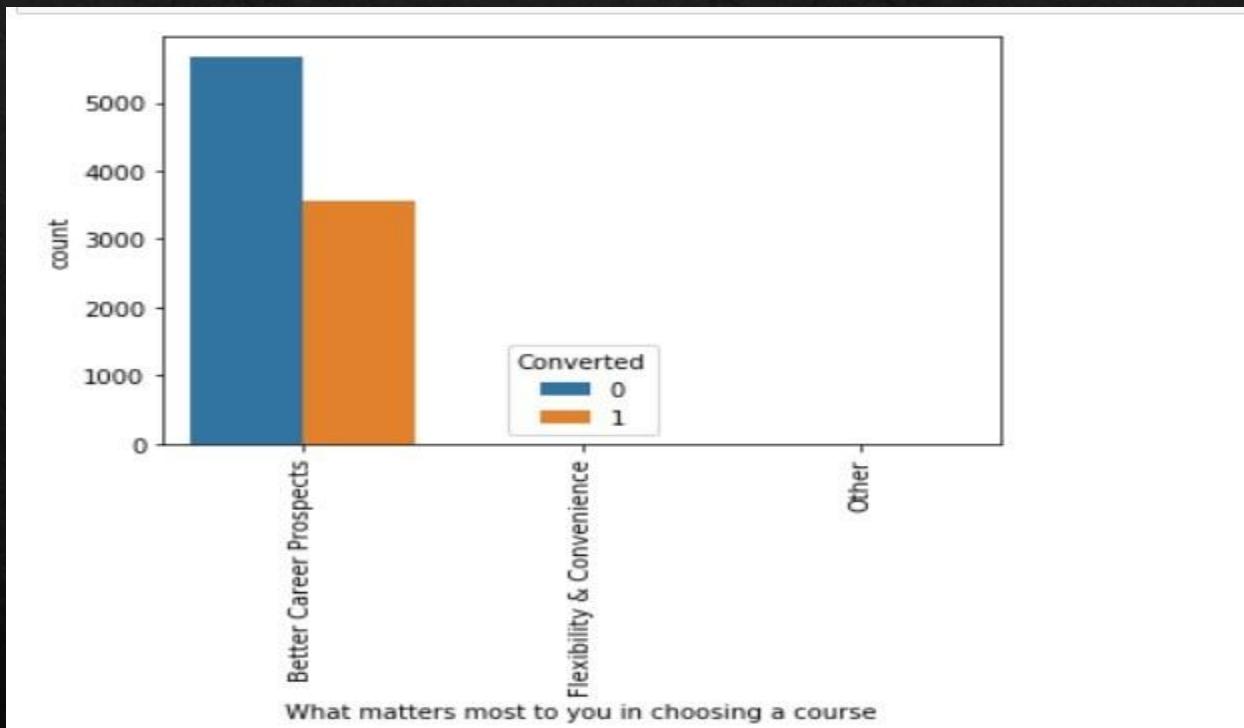
- We see that specialization with Management in them have higher number of leads as well as leads converted.
- Specialization is significant variable and should not be dropped from our analysis

Univariate Analysis - Occupation



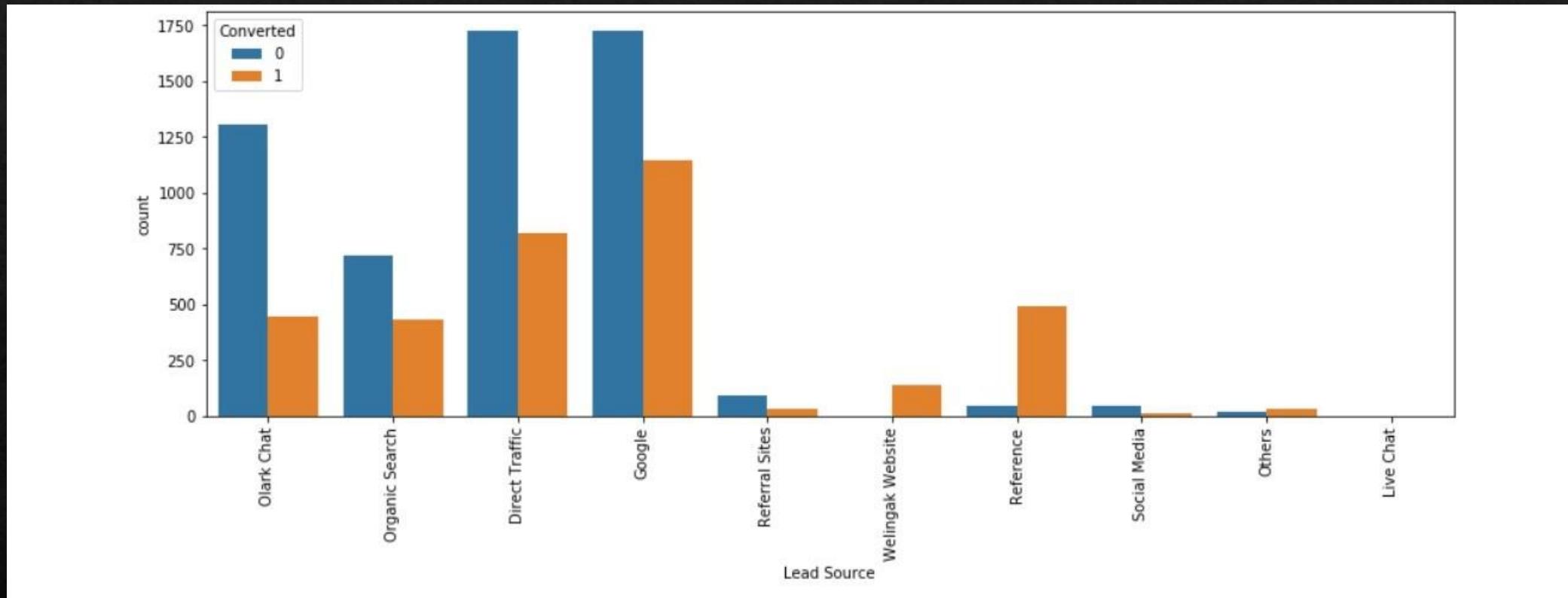
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.

Univariate Analysis – What matters most when choosing a course



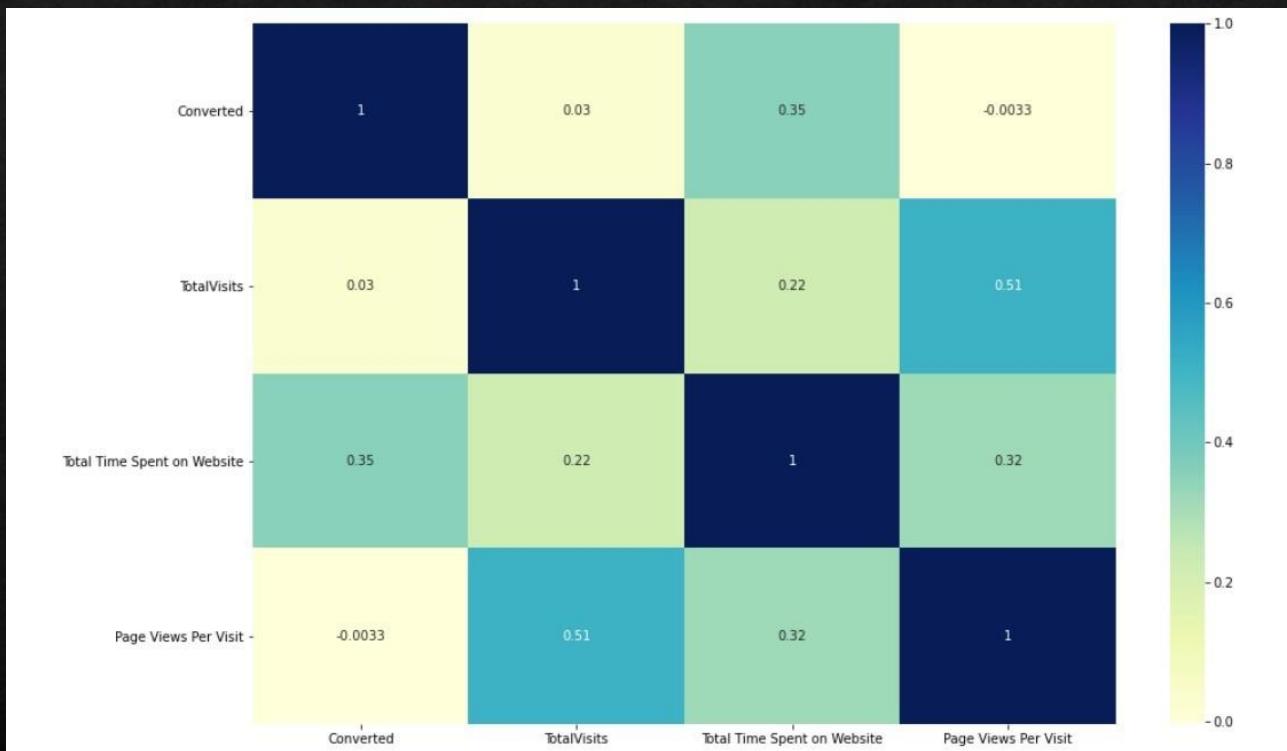
- From the given data we could see that most of the leads what to improve their career prospects and this is the main reason why they want to pursue an online higher education course
- The lead conversion rate also looks to be better for this parameter

Univariate Analysis – Lead Source



- From the above bar plot we can see that the lead conversion ratio for the leads obtained from Google, Direct Traffic, Organic Search, Olark Chat, Reference are better compared to all other sources
- It is to be noted that the conversion rate for referrals is far better than any other sources as leads usually would like to take feedback from previous learners before opting for an online course
- The sales team should hence focus more on these sources and may ignore concentrating efforts on the remaining sources

Numerical Attributes Analysis



- We can see that there is a strong correlation between the number of leads who visited the website and spent time on the website
- The sales team should focus more on these leads who have a better conversion rate as they are spending time to go through the website and courses offered

Data Conversion

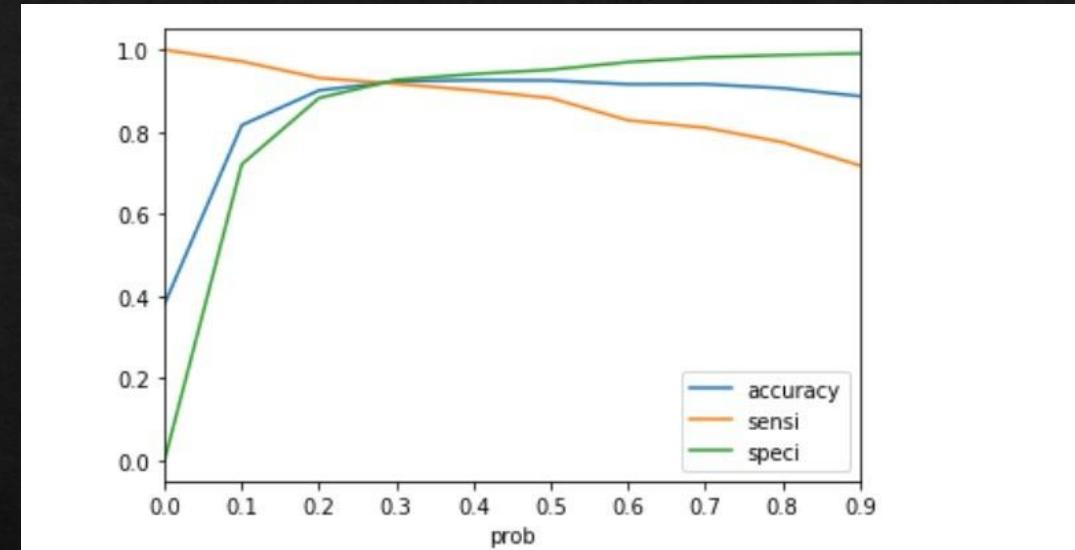
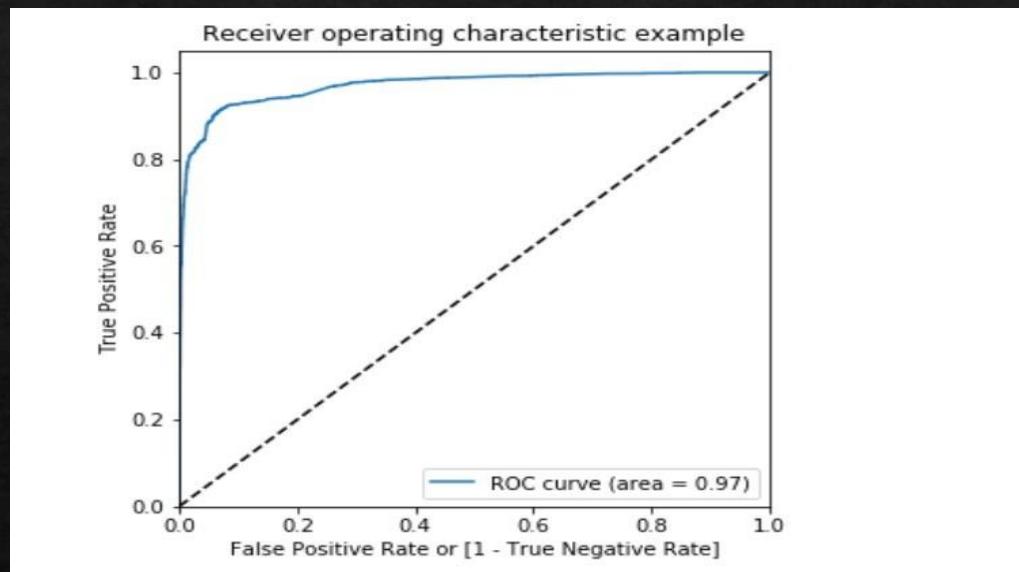
- Numerical Variables are normalized
- Categorical Variables are imputed
- Outlier Treatment
- Dummy Variables are created for object type variables

Model Building

Steps Performed

- Splitting the data into Train and Test Sets
- The initial step for logistic regression is to perform a train-test- split and we have chosen 70:30 ratio respectively as our train and test sets
- Using Recursive Feature Elimination for Feature Selection
- Running RFE with 15 variables as the output
- Model building & removing the variables whose p-values are greater than 0.05 and Variable Inflation Factor (VIF) greater than 5
- Predicting on the test data set
- The overall accuracy of the train set is 92.5%

ROC Curve



- We plotted the ROC curve to find the optimal cut off point
- Optimal cut off probability is where there is a trade off between sensitivity and specificity and they are balanced
- From the above graph we can see that the optimal cut off is 0.3

Predictions on the test set

After running the model on the Test Data below are the values that were obtained

Accuracy: 92.78%

Sensitivity: 91.98%

Specificity: 93.26%

Recall : 91.98%

Conclusion and Recommendation

It was found that the variables that contribute to a positive lead conversion ratio are as below.

1. Lead Origin - Lead Add Form
2. Total Time spent on Website
3. Last Activity
4. Referrals
5. Occupation – Working Professionals

Basis our analysis we recommend that the sales and marketing concentrate their efforts more on these variables to meet a lead conversion rate of 80% . Keeping these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind to subscribe to their courses