LEAD SCORING CASE STUDY

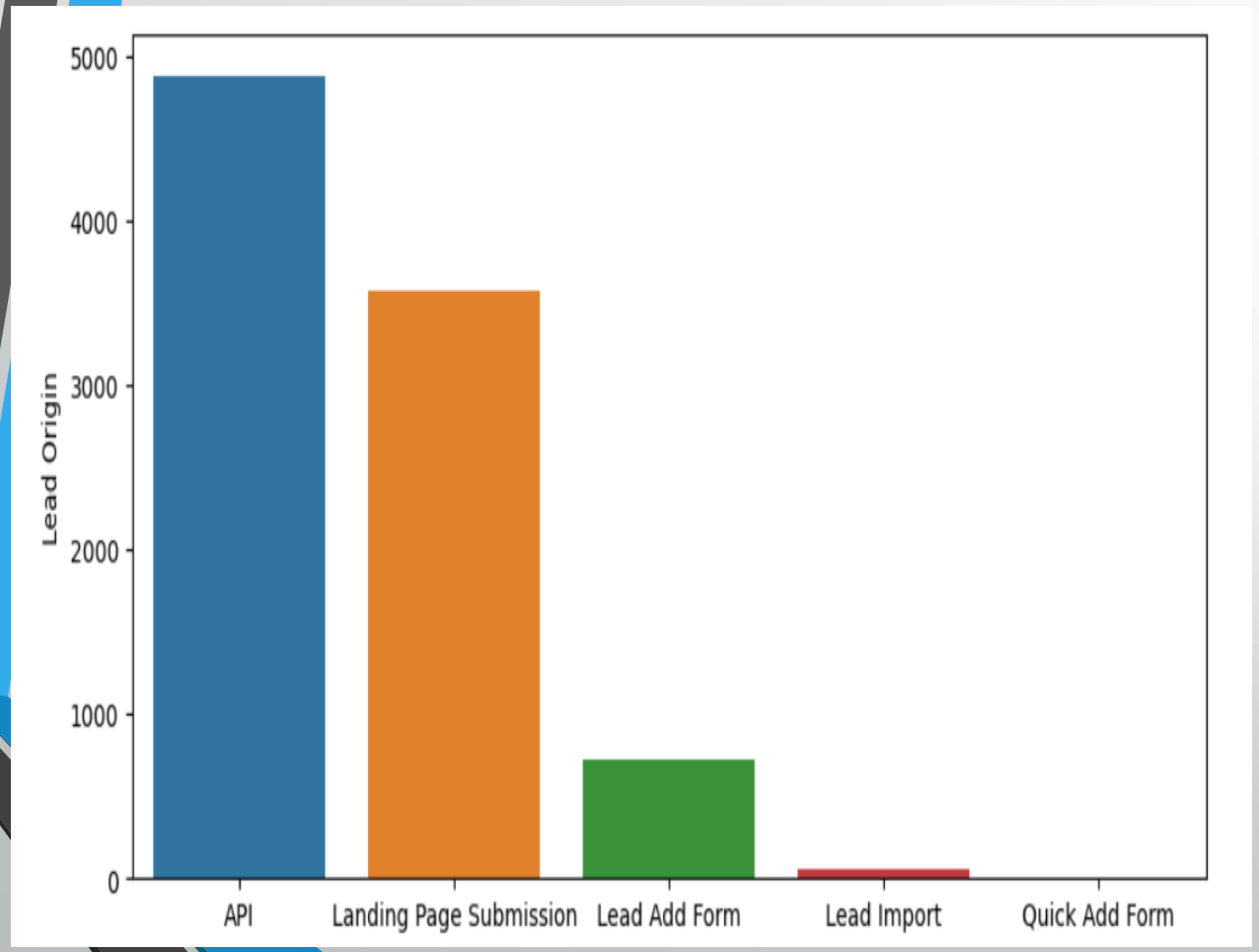
Problem Statement:

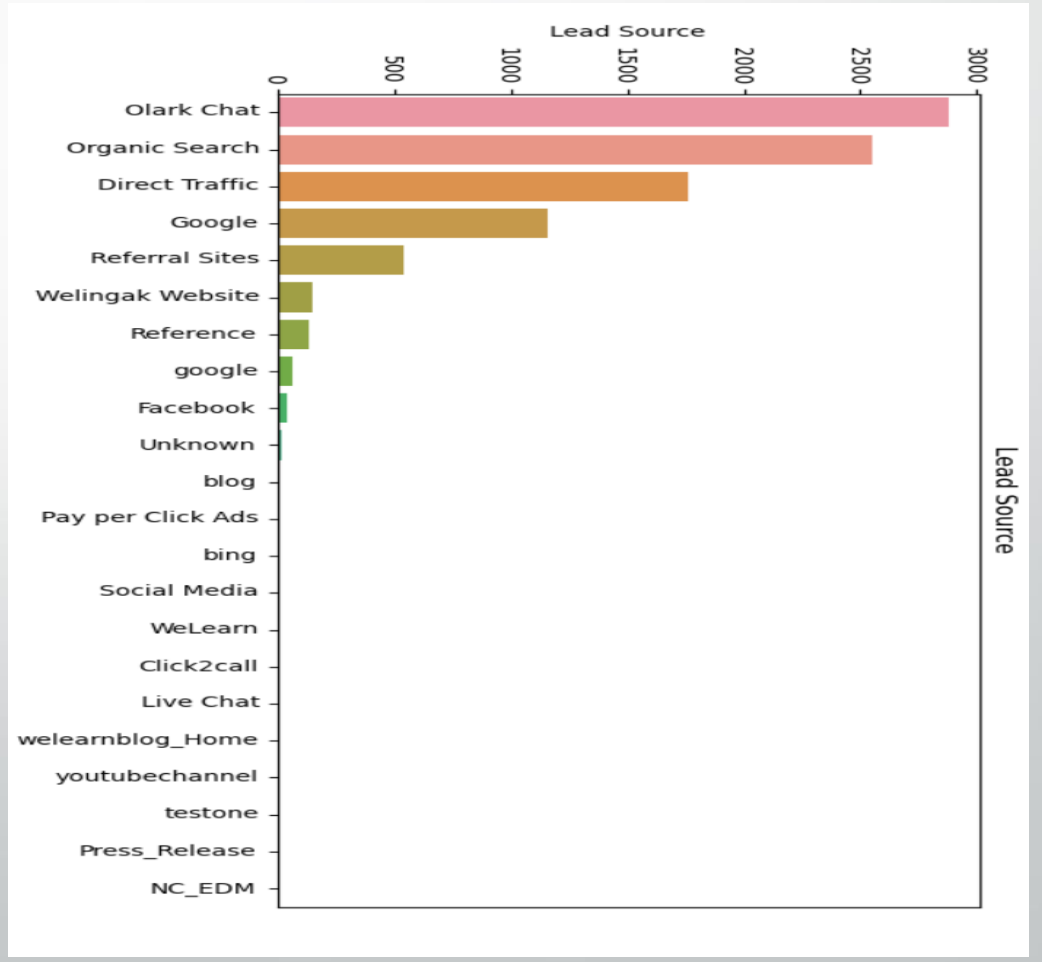
- X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, 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.
- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom.
- X Education has appointed you 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Steps for Analysis

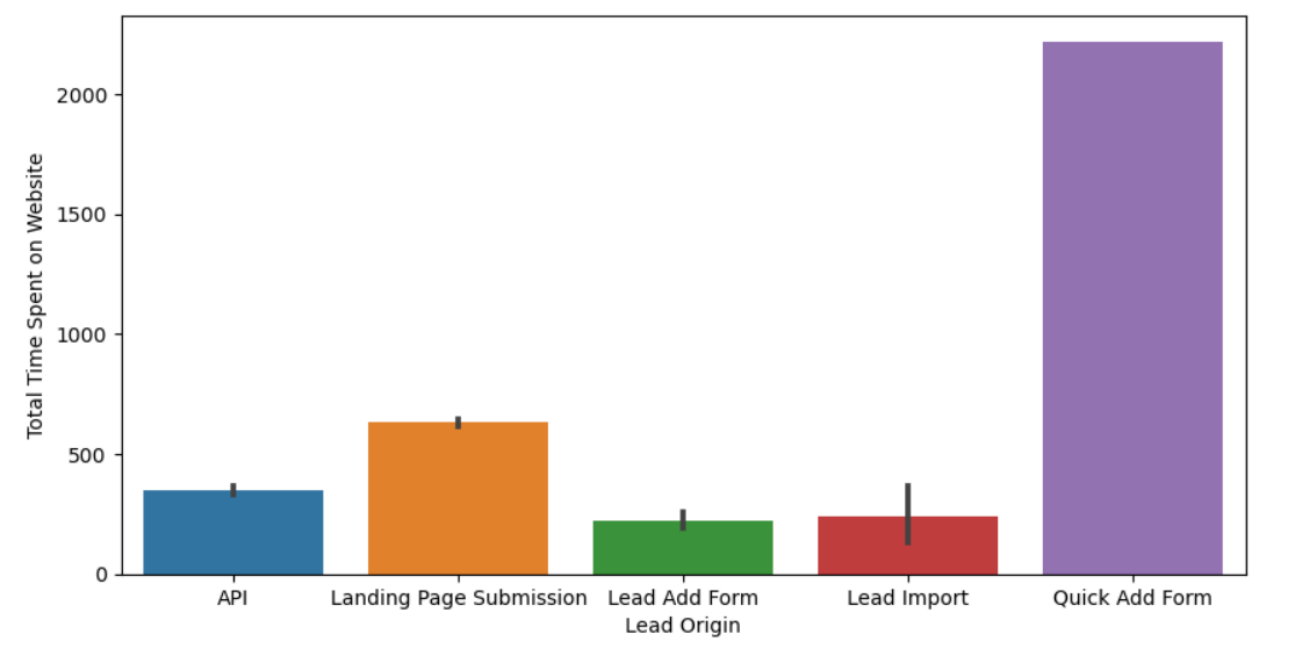
1. Reading and understanding the data
2. Cleaning the data
    a) Replacing the values that do not make any sense
    b) Dropping columns with more than 40% null values
    c) Dropping columns that are not necessary in data analysis
3. Performing Visualization on the clean data
4. Getting Dummy Variables
5. Training the model
6. Building the model
7. Predicting Probabilities
8. Model evaluation

Most Popular Lead Origin and Source

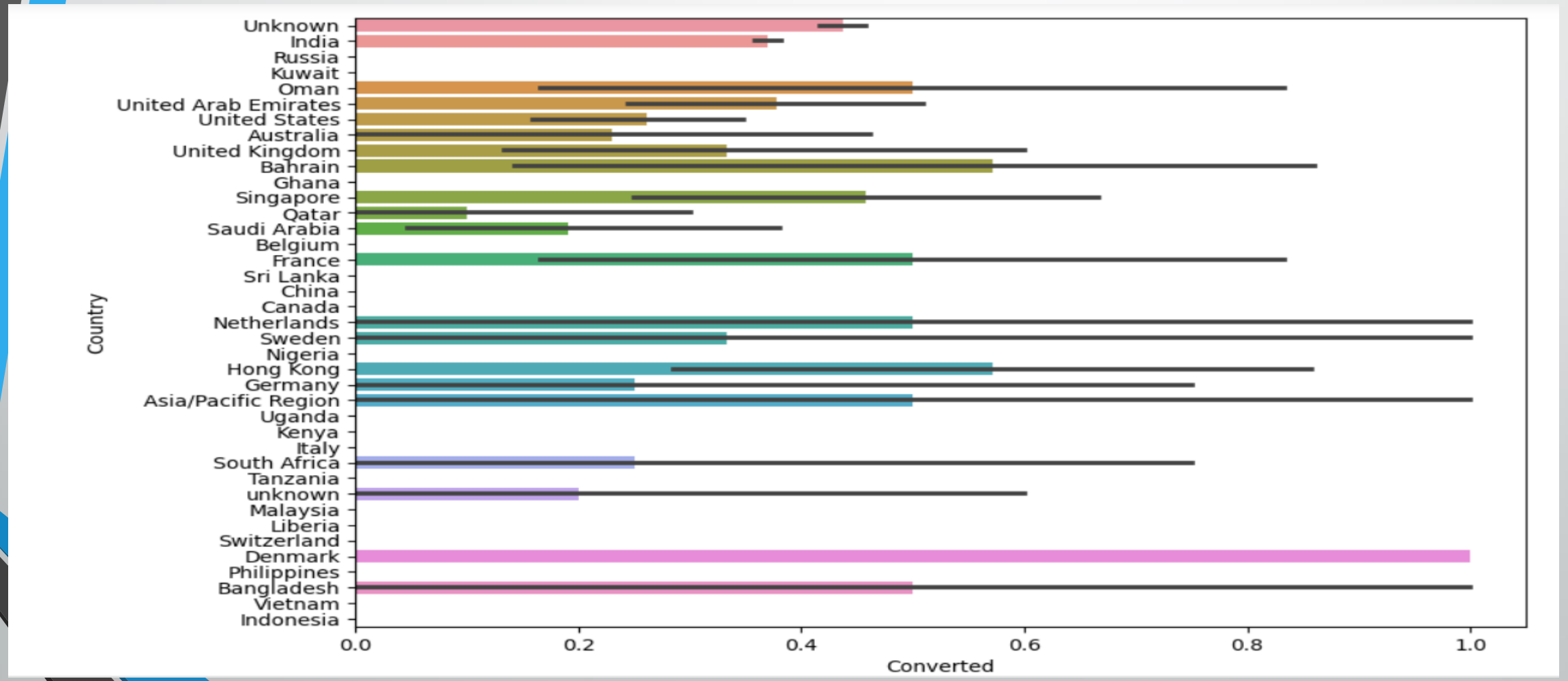




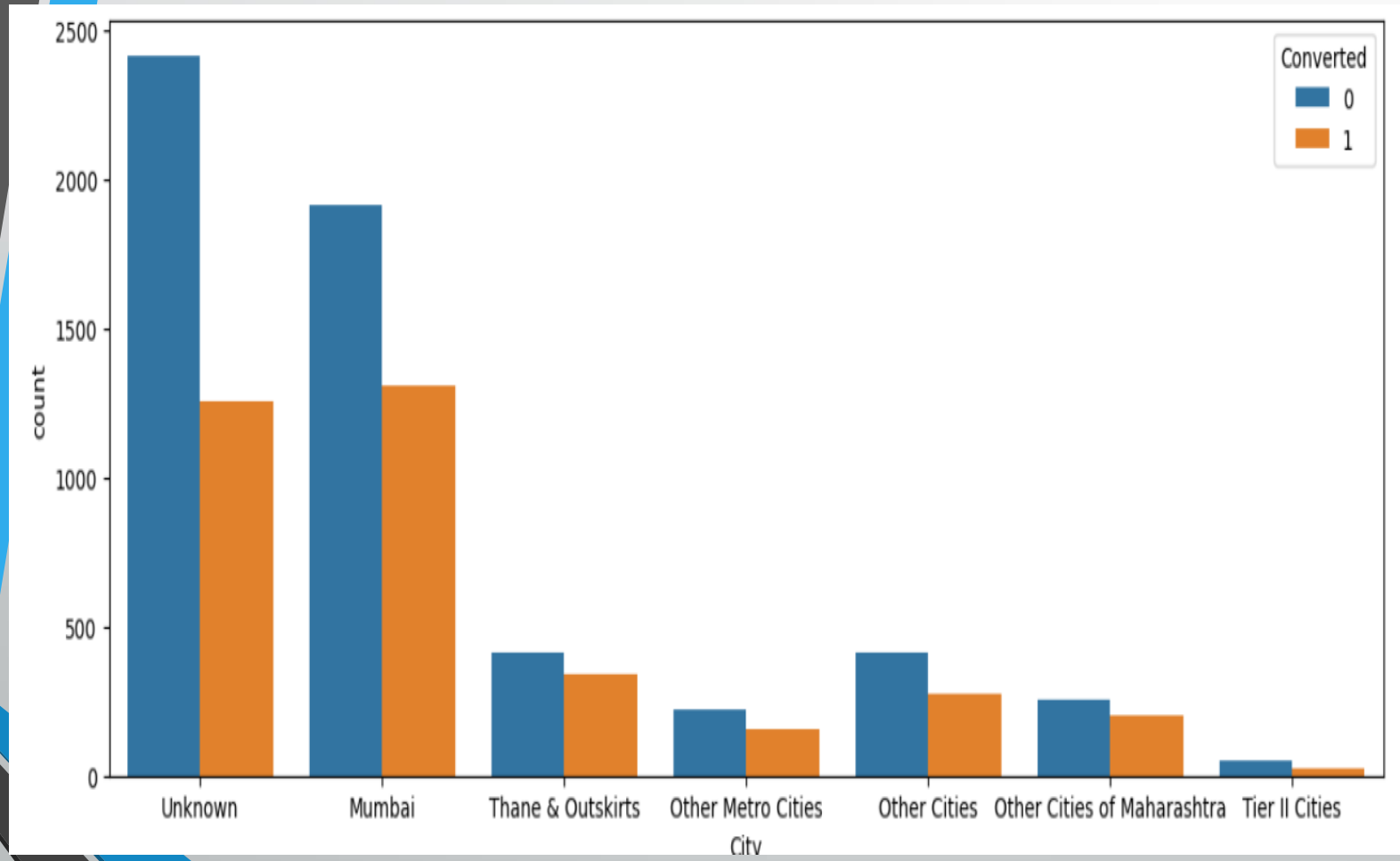
Total Time Spent on Different Origins

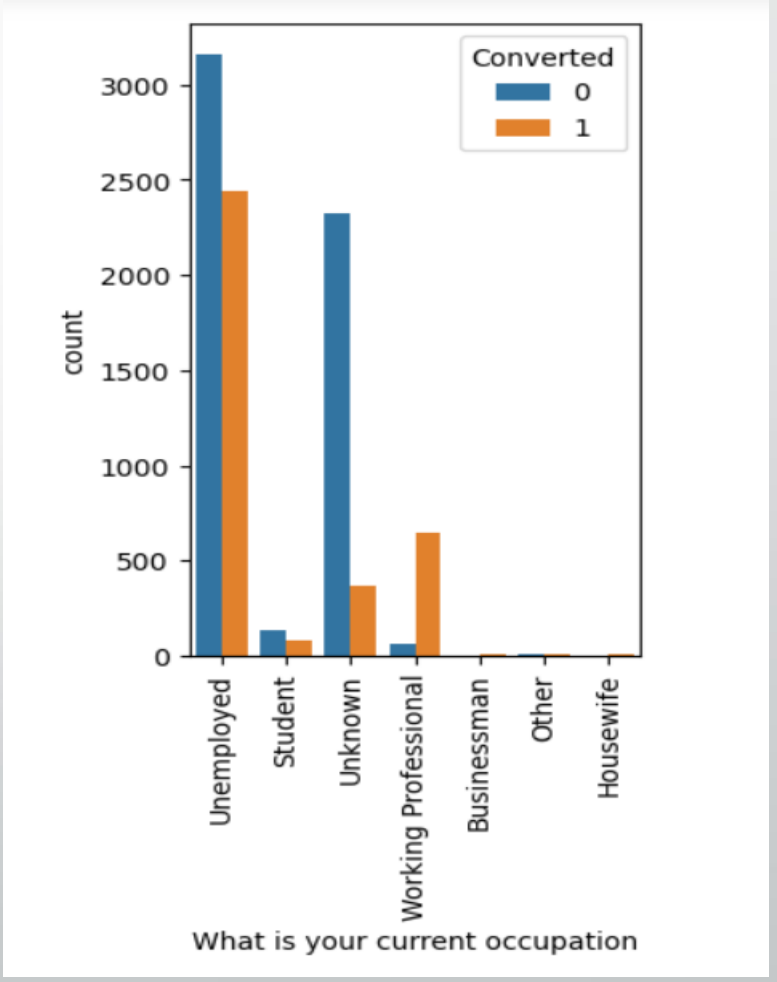


Country-wise Converted Data

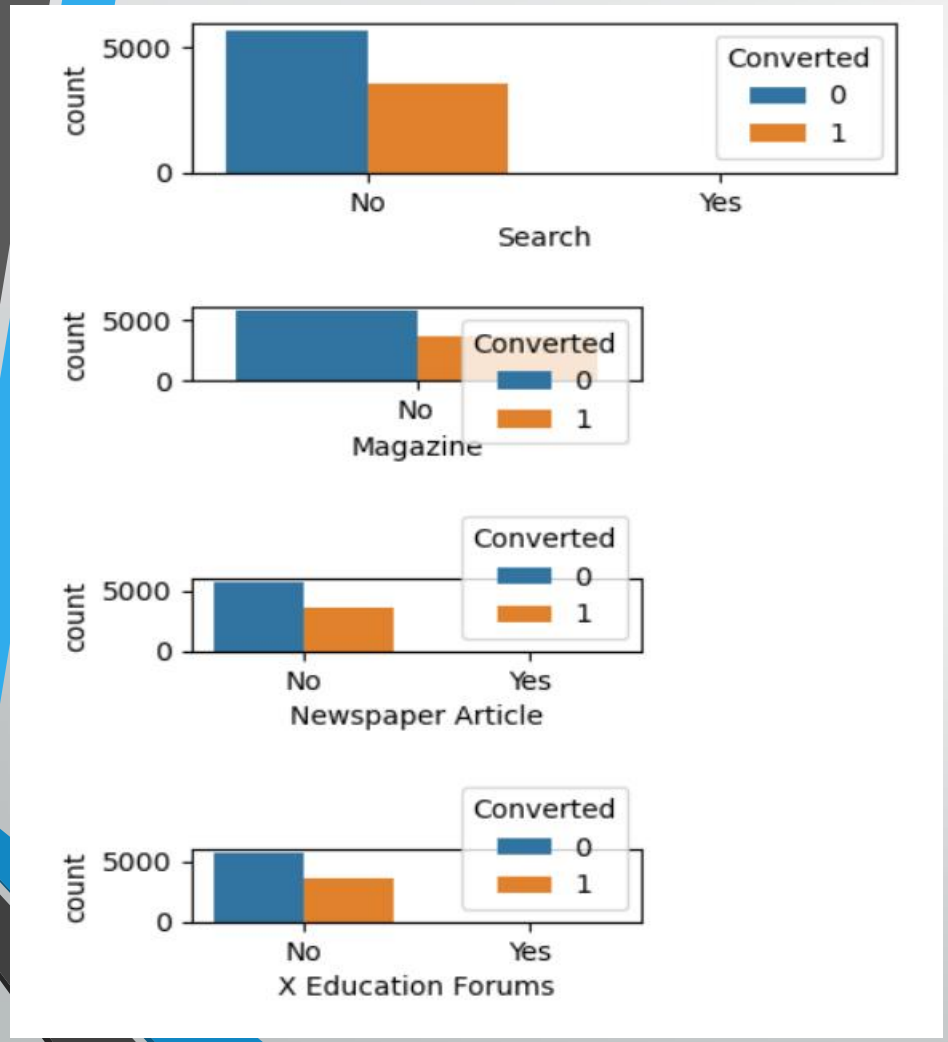


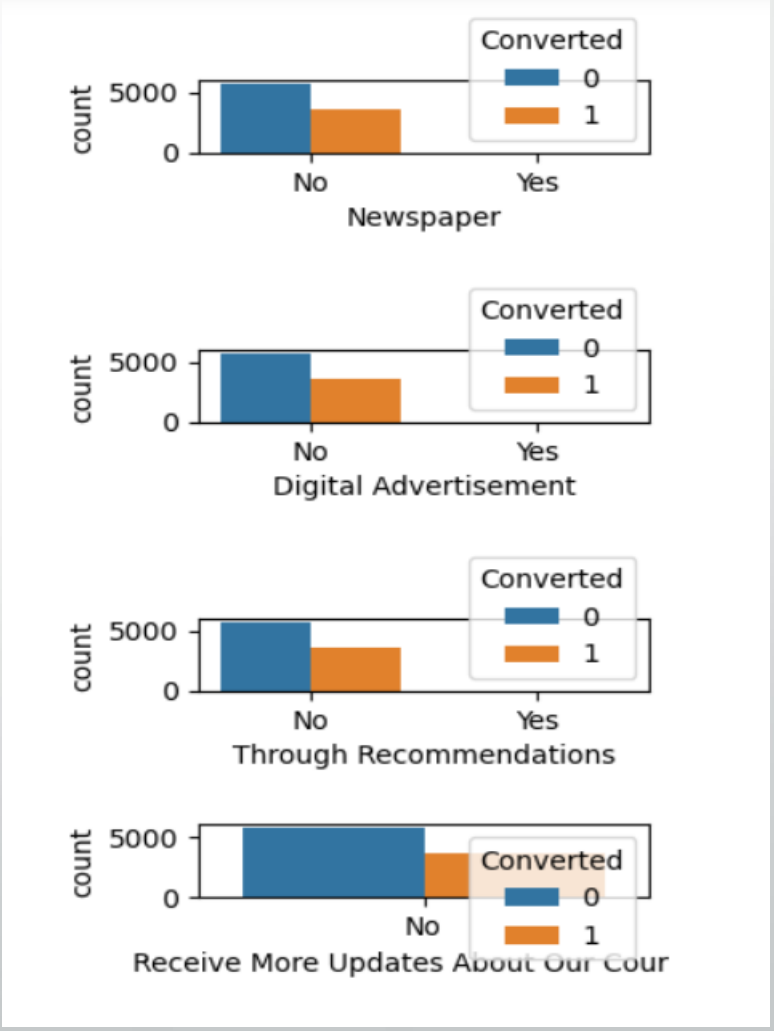
Converted Data as per City and Occupation





Converted Data according to Other Factors





Model Building

1. Splitting the Data into Training and Testing Sets. The first basic step for regression is performing a train-test split, we have chosen **70:30** ratio.
2. Use **RFE** for Feature Selection. Running RFE with **15** variables as output.
3. Building Model by removing the variable whose **p- value is greater than 0.05** and **VIF value is greater than 5.**
4. Predictions on test data set.
5. Overall **82% accuracy, sensitivity** of around **70%** and **specificity** of around **88%**

Conclusion

According to the logistics(p-values and VIF),the most important columns in the data set are:

1. What matters most to you in choosing a course.
2. Country
3. Total Time Spent on Website
4. Lead Origin:
    a. Lead Add Form
5. Last Activity
    a. SMS Sent
    b. Olark Chat Conversation
6. TotalVisits
7. What is your current occupation
8. Last Notable Activity