Lead Scoring

Group Members:

Kuldeep

Dhanush

Manish

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.

- Although X Education gets a lot of leads, its lead conversion rate is very poor **around 30%.**
- As 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 total base.

Business Objective :

As 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. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education wants priorities 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.**

Problem Statement & Datasets :

The data given below contains the information about the Leads such as time spent on website, Visits and activities.:
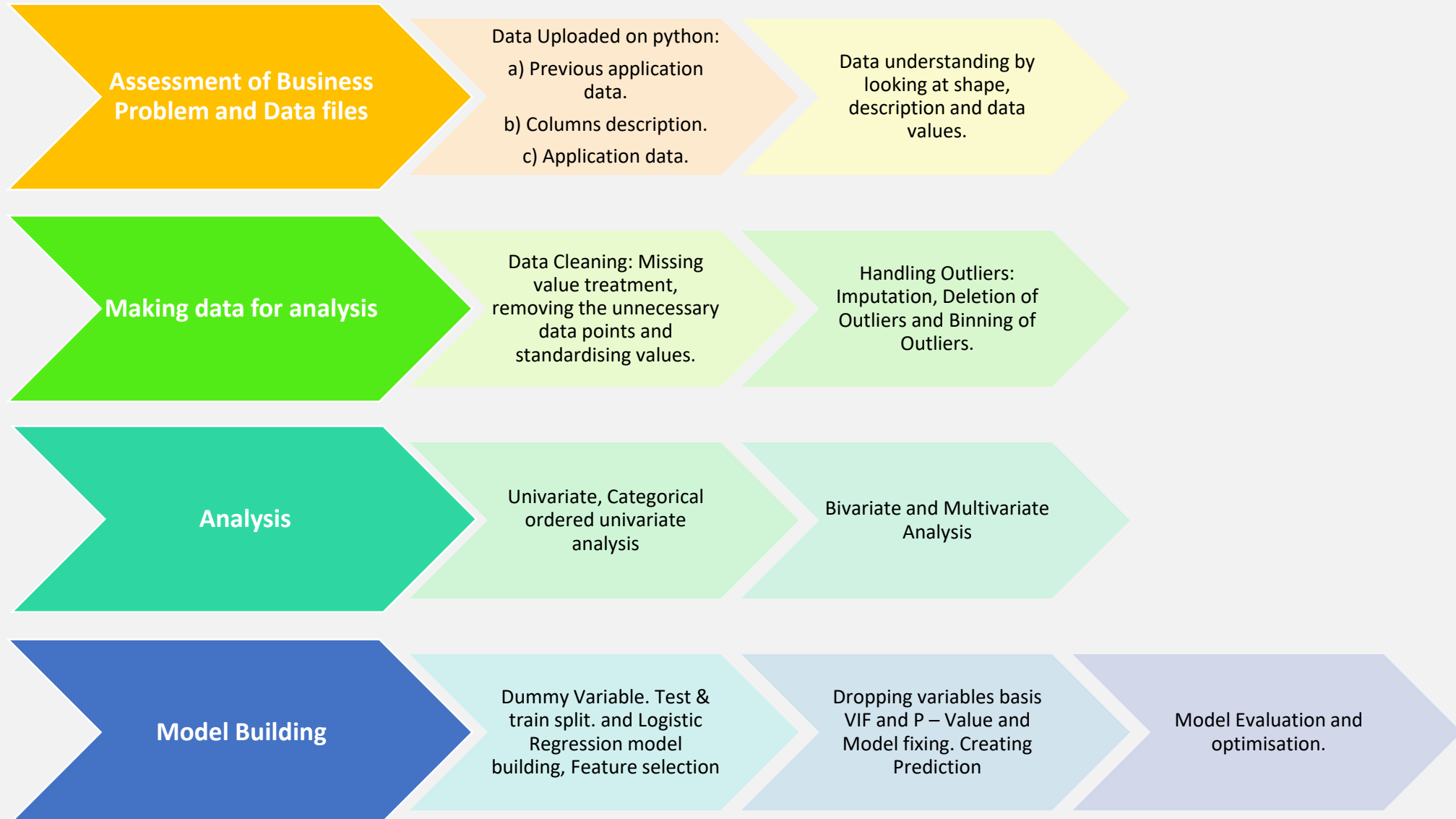
Datasets:

1. leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

2. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

3. *'Leads Data Dictionary'* is data dictionary which describes the meaning of the variables.
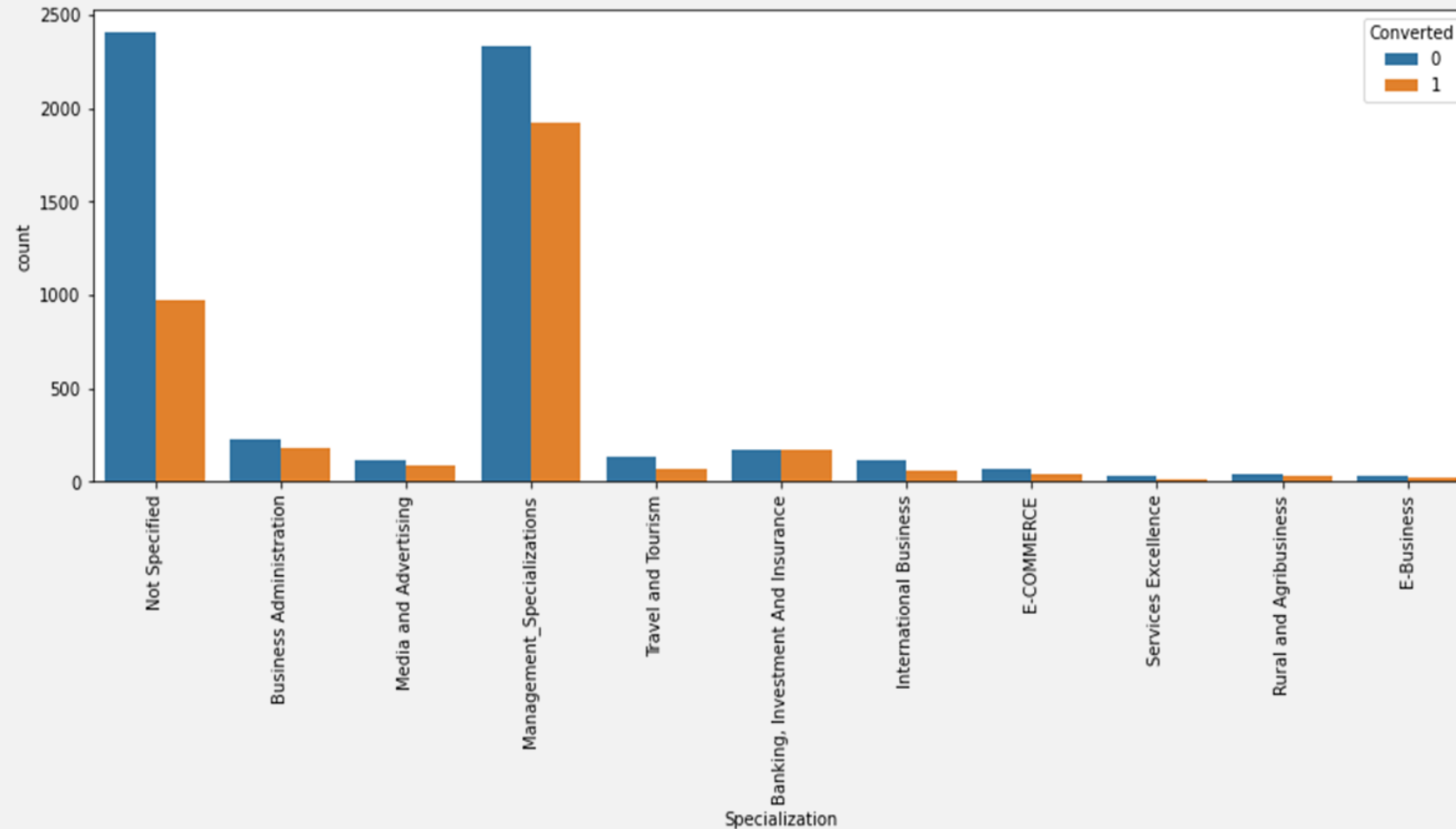
Assumption:

Looking at the application data , assumed that following variables are going to shape the result of overall analysis.

- Total Visits
- The total time spend on the Website.
- Lead Origin Lead Add Form
- Lead Source Direct Traffic
- Lead Source Google
- Lead Source Welingak Website
- Lead Source Organic Search
- Lead Source Referral Sites
- Do Not Email - Yes
- Last Activity Email Bounced
- Last Activity Olark Chat Conversation

# Approach: Following approach followed to answer the business problem

Assessment of Business Problem and Data files

Data Uploaded on python:

a) Previous application data.

b) Columns description.

c) Application data.

Data understanding by looking at shape, description and data values.

Making data for analysis

Data Cleaning: Missing value treatment, removing the unnecessary data points and standardising values.

Handling Outliers: Imputation, Deletion of Outliers and Binning of Outliers.

Analysis

Univariate, Categorical ordered univariate analysis

Bivariate and Multivariate Analysis

Model Building

Dummy Variable. Test & train split. and Logistic Regression model building, Feature selection

Dropping variables basis VIF and P – Value and Model fixing. Creating Prediction

Model Evaluation and optimisation.
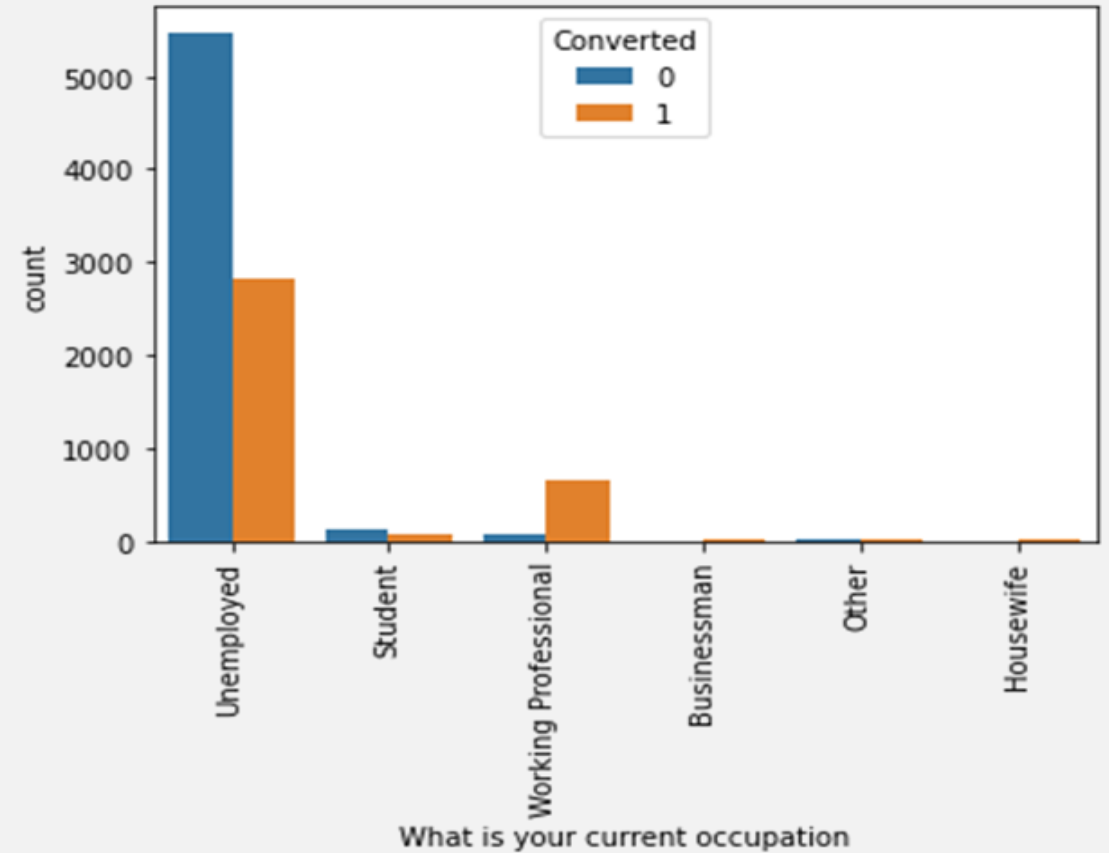
Analysis slide 1

- Maximum lead conversion is from "Management_Specializations". However, highest conversion rate is from "Banking, Investment and Insurance".
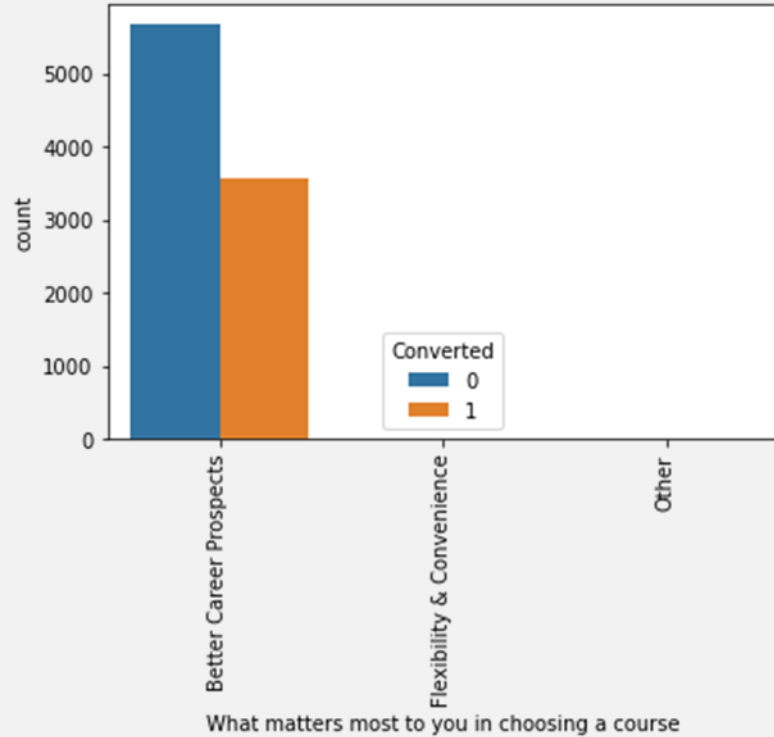
Analysis slide 2
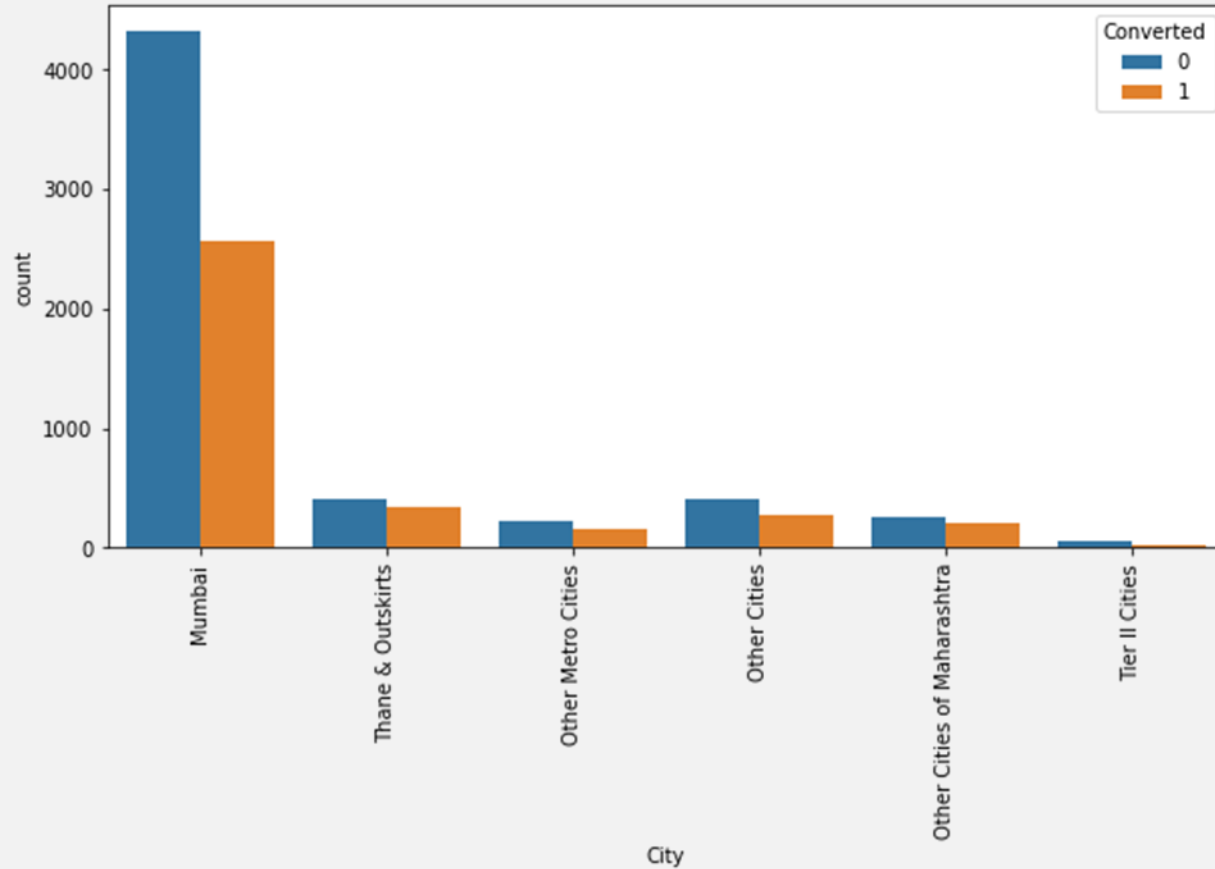
- Number wise Mumbai has highest number of leads and their conversion.
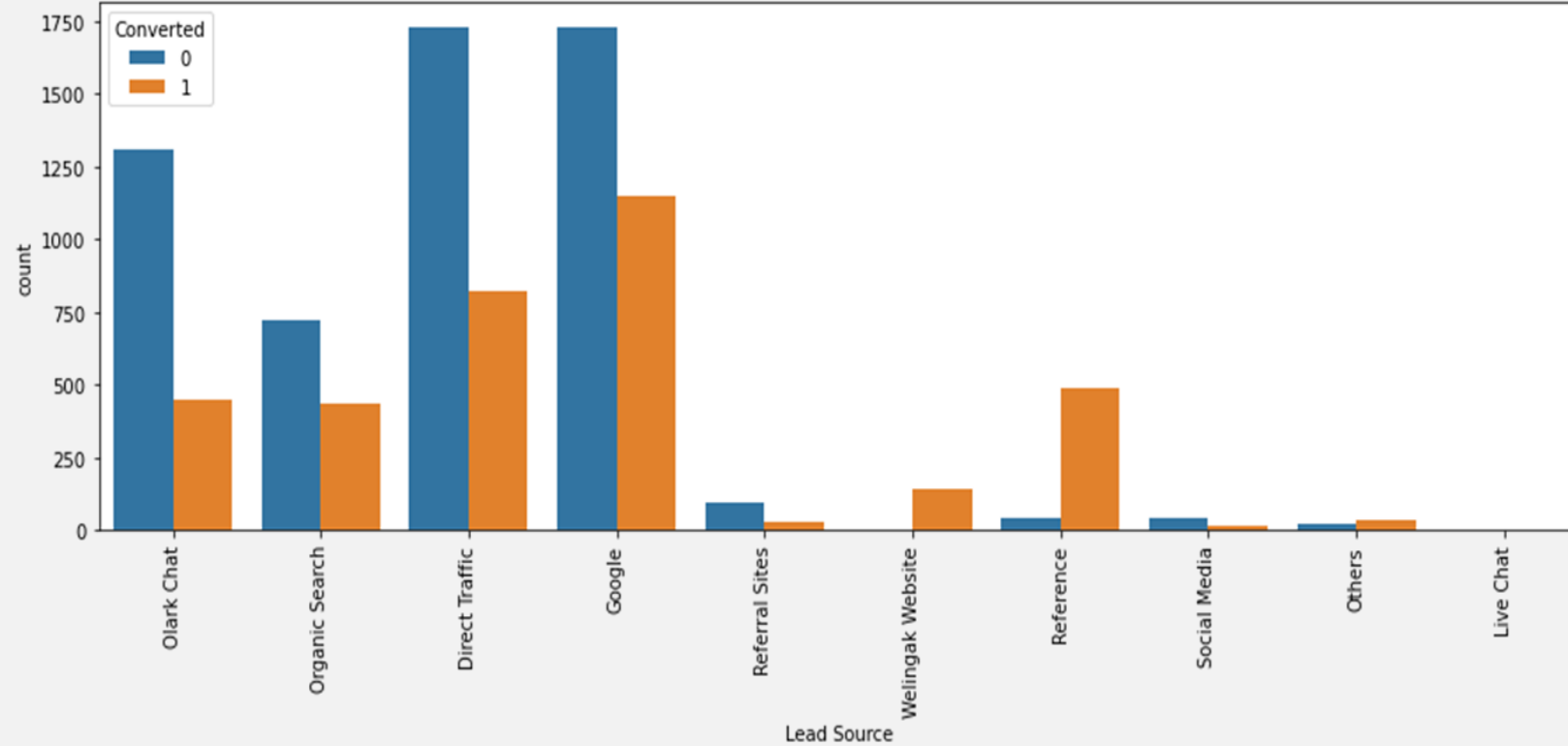
- Higher Lead conversion from working professionals.
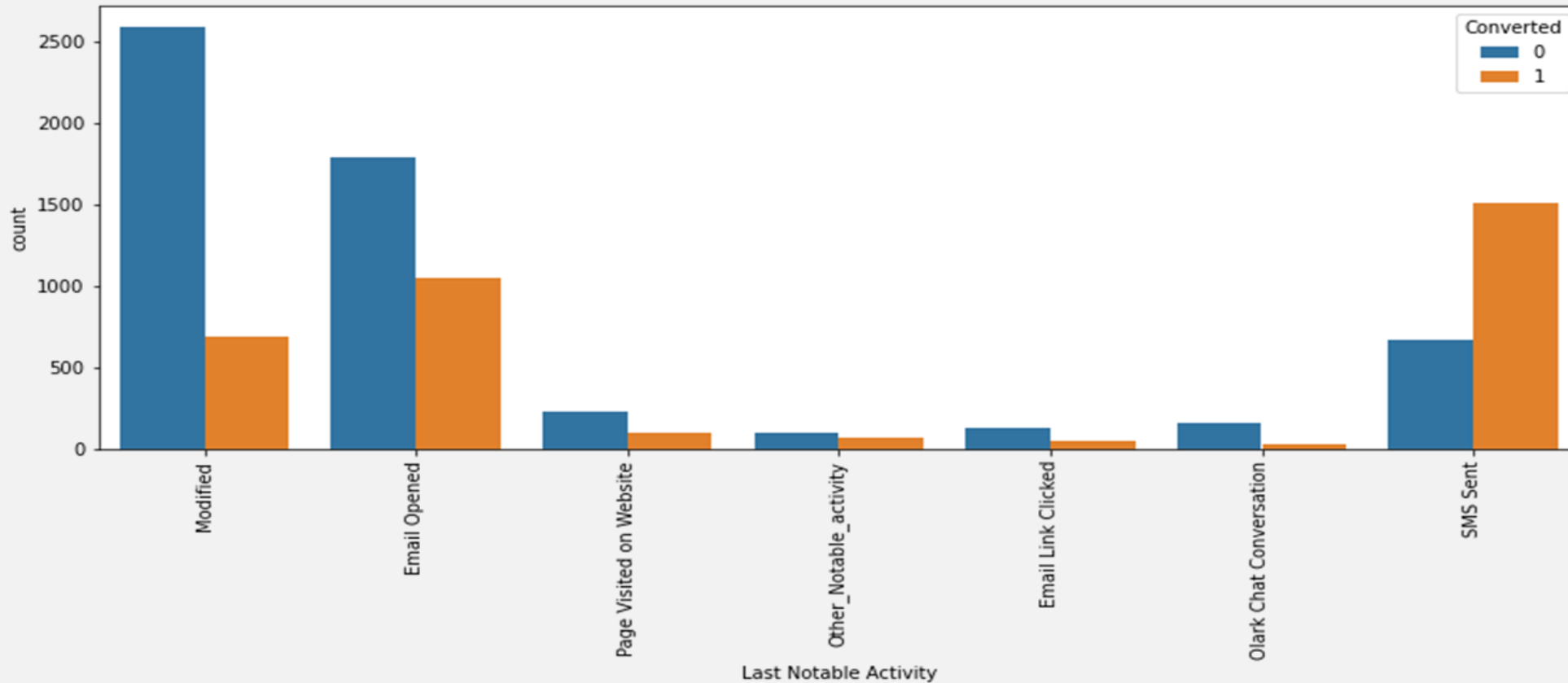
Analysis slide 3

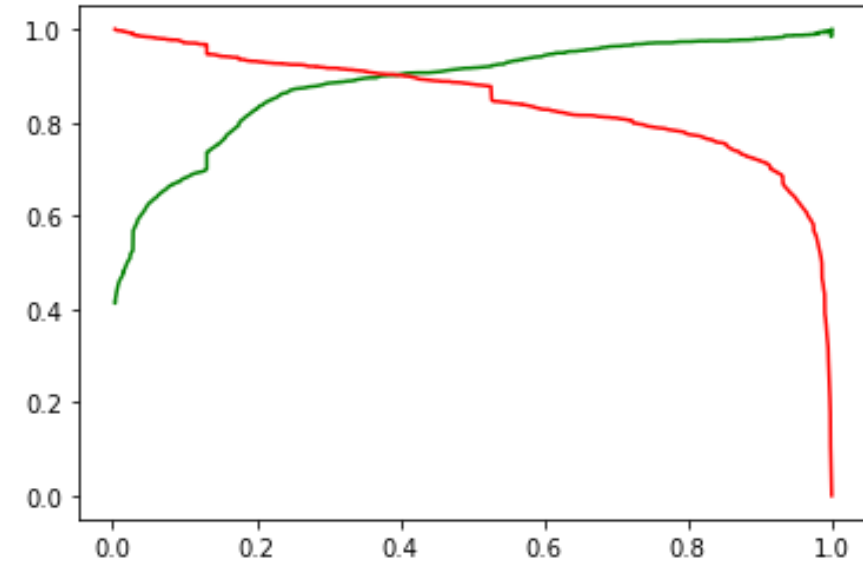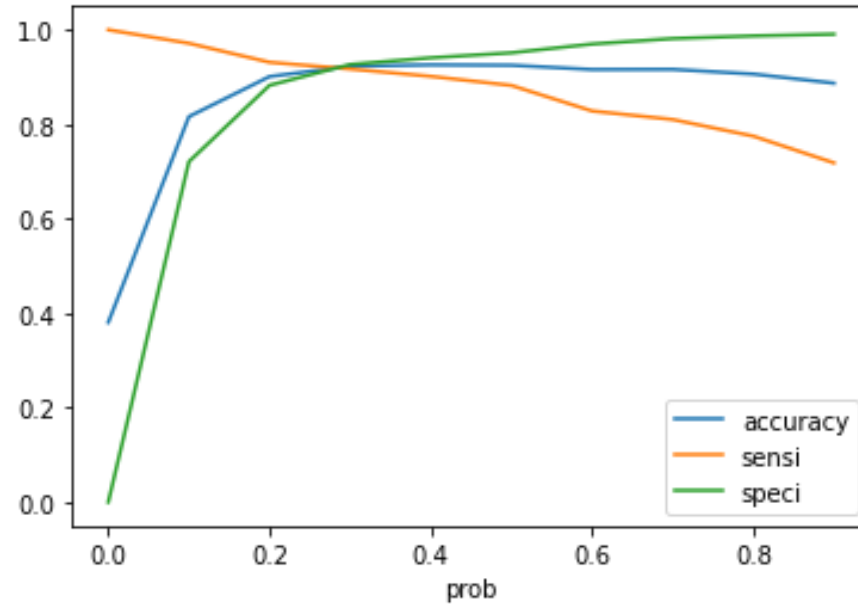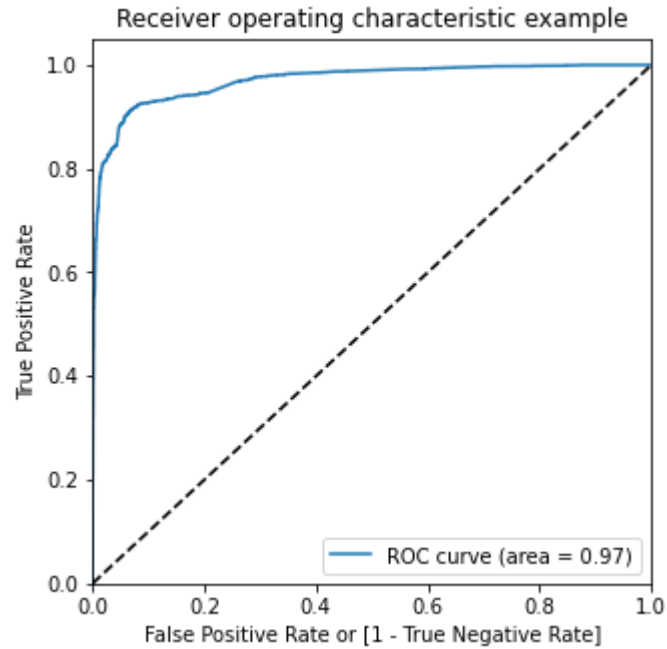- Clearly visible that reason for course enquiry is "Better Career Prospects".

- Maximum number of Lead generation and lead conversion is from "Google" and "Direct Traffic".

Analysis slide 4

- Last SMS sent and Email opened notable have higher number of Lead generation and conversion rate.

Analysis slide 5

Linear Regression Final Model Parameters

Area under ROC = 0.97

Intermediate cut-off = 0.3

Final cut-off = 0.4

Model performance analysis

- Overall accuracy on Test set: 92.7%

- Sensitivity of our logistic regression model: 91.9%

- Specificity of our logistic regression model: 93.2%

Conclusion

Top 3 variables that contribute towards lead conversion are as follow:

- Total time spent on website.

- Total Visit.

- Last Notable Activity_SMS Sent.

Top 3 variables that should be focused are as follow:

- Last Activity_SMS Sent (positively impacting)

- Last Activity_Olark Chat Conversation (negatively impacting)

- Lead Source_Olark Chat (negatively impacting)

Recommendation

X Education Company should focus on following area to improve the Lead conversion rate:

- Website should be made more engaging to make leads spend more time

- Increase on sending SMS notifications since this helps in higher conversion

- Get Total visits increased by advertising etc. since this helps in higher conversion

- Improve the Olark Chat service since this is affecting the conversion negatively.

Thank You.