Lead Score Case Study

Team Members:
Kulmeet Singh Dusaj
Nuthan Teja
AP Sharma

Problem Statement:

upGrad

- This Case Study is basically from the edtech industry. X Education, an online course provider for industry professionals, aims to improve its lead conversion process.
- Despite generating numerous leads, their conversion rate is low (30%). The company wants to identify potential high-converting leads ("Hot Leads") to increase the overall conversion rate to around 80%.
- The goal is to build a lead scoring model that assigns scores to leads based on their likelihood to convert. This involves analyzing historical data, preprocessing, feature engineering, model building, evaluation, optimization, and implementation into the existing sales process.
- The lead scoring model should help the sales team focus on leads with higher conversion potential, improving overall efficiency. Regular monitoring and iteration are emphasized for continuous improvement.

Overall Solution Steps:

upGrad

- DataFrame Analysis
- Data Cleaning and Manipulation
  - Checking for duplicates
  - Checking for missing values and dropping columns with high missing values
  - Imputation of missing values if necessary
  - Outlier check and handling
- Exploratory Data Analysis
  - Univariate Data Analysis: Distribution of a single variable using countplot(categorical variables) and histograms(numerical variables)
  - Bivariate Data Analysis: Correlation and pattern between 2 or more variables. Mostly used this to see trend of variables with the target variable.
- Feature Scaling and Dummy Variables and encoding of the data.
- Classification Technique :Logistic Regression Model for making predictions
- Validation of model
- Model Presentation
- Conclusion and Recommendation

Analysis Approach(Preliminary Checks and Manipulation)

upGrad

Data Manipulation:

- I have tried to start with 15 most important features which I selected using Recurrsive Feature Elimination. I have done a pre assignment research on different platforms like google and went through many edtech blogs to spot variables which mainly affects such lead conversion rates.
- I have tried keeping as many columns as I could for EDA, except the columns having to many null values(35% and above)
- I have also dropped columns which had only 1 category within it as it won't affect our analysis much. Example of such columns are Magazine, Receive More Updates About Our Courses.
- I have removed columns like Prospect ID and Lead Number as it won't affect our Analysis
- I have done data cleaning such as missing value treatment, outlier treatment etc.

Analysis Approach:

upGrad

STEPS IN ANALYSIS:

Understanding the Domain/variables:

- Before starting the Analysis it is very important to go through the data dictionary provided to understand the attributes and also I did some research on the variables that edtech and online education companies evaluates before considering a hot lead.
- I have spent a good amount of time just to gain some domain knowledge so that I can bring out the best insights from the data.

Import/Load the Data:

- I have then loaded the file into the pandas DataFrame for Analysis.

Analysis Approach:

upGrad

Check the Structure/Metadata of the data:

- It is very important to check on few things before starting with the analysis. We should be aware of the shape and size of the dataframe, we should also know the datatypes of the variables we have and if some datatype is needs a change we should go ahead and change the data types. In this assignment all the columns were already in correct format.
- I have used functions like shape , info() and describe() for metadata Analysis.

Missing Value Check and Imputation:

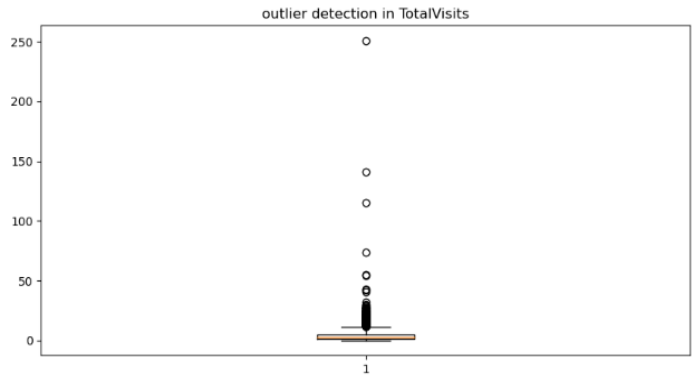
- There were lot of column values in the file that were missing. I have taken 35% as a threshold and after proper research the what impact a particular column could make to my analysis I have dropped columns.
- After dropping the columns I have checked the dataframe for the presence of null values using df.isnull().sum().

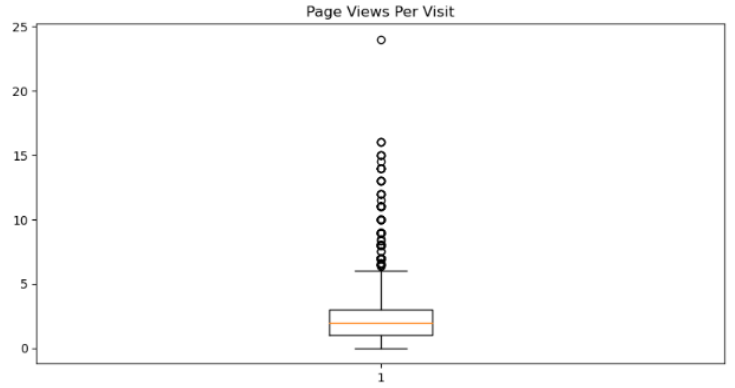
Analysis Approach (Exploratory Data Analysis) :

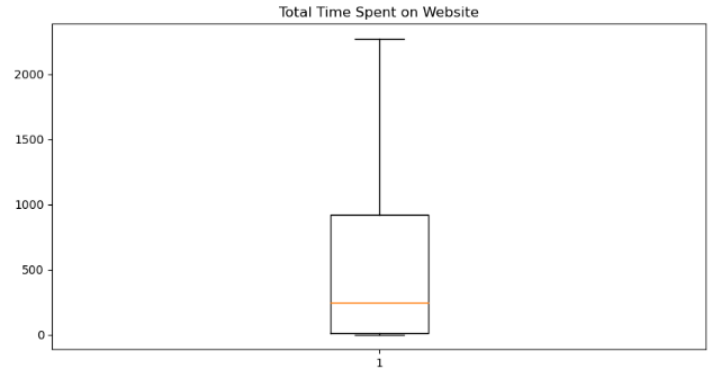
upGrad

OUTLIER TREATMENT

- I have analysed almost all the numerical columns for outliers
- I have used box plot in order to visualize the outliers.
- I found some outliers in columns like ‘TotalVisits’ and 'Page Views Per Visit’ which I dropped at appropriate values.





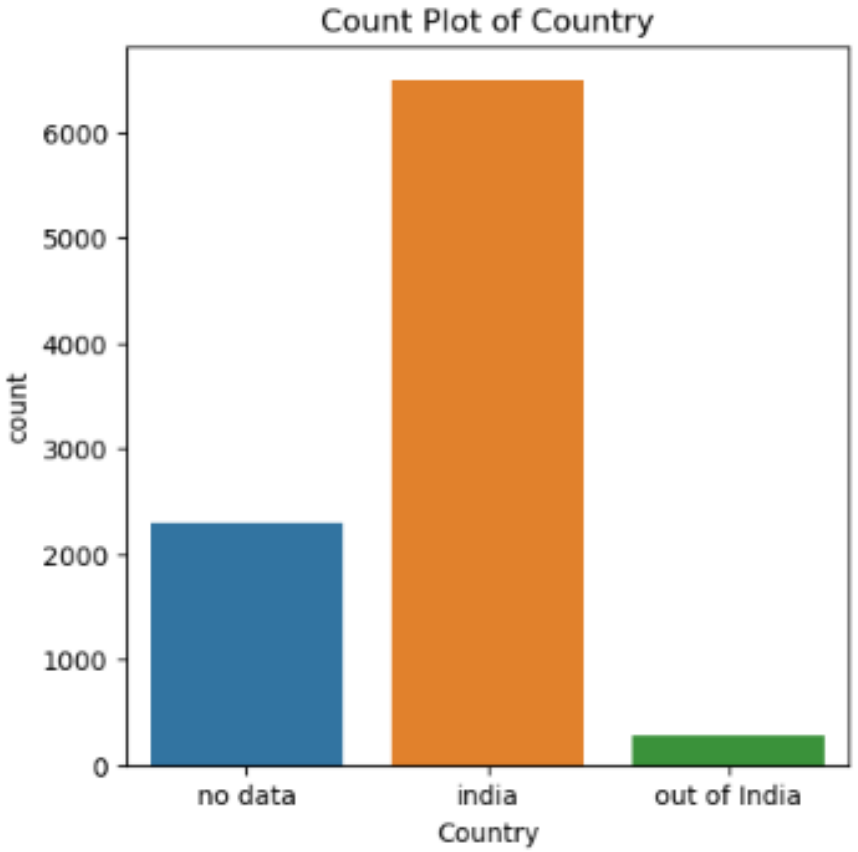


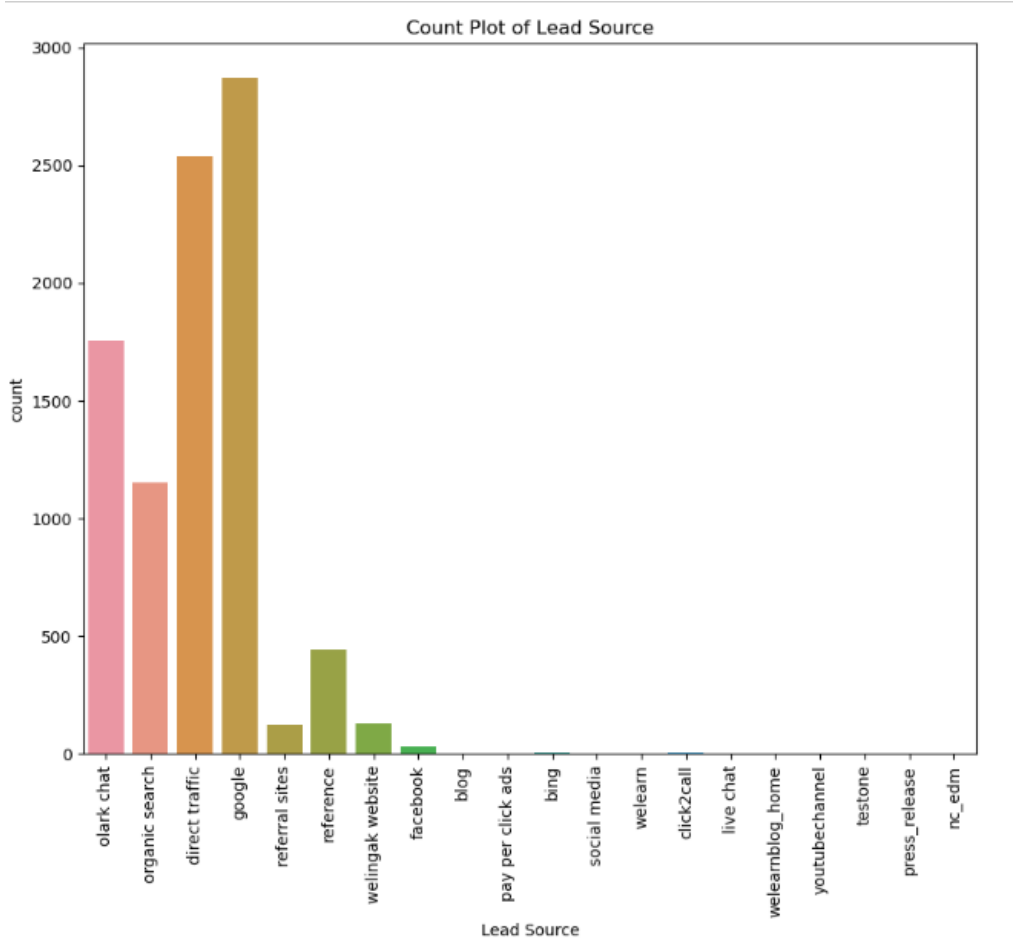
Analysis Approach:

upGrad

UNIVARIATE ANALYSIS:

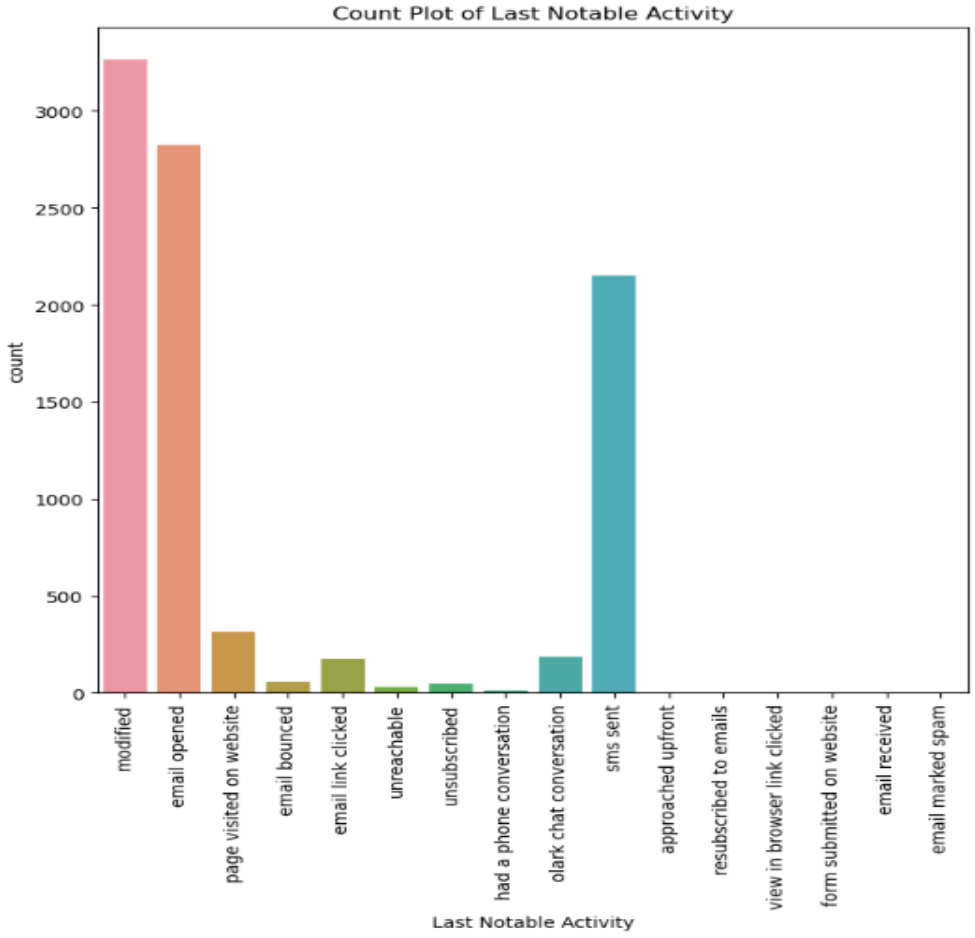
- For Univariate Analysis I have tried to plot count plots for categorical variables and Histogram plot for each major numerical variable. Some examples are listed below. Some of the important visuals are show below.

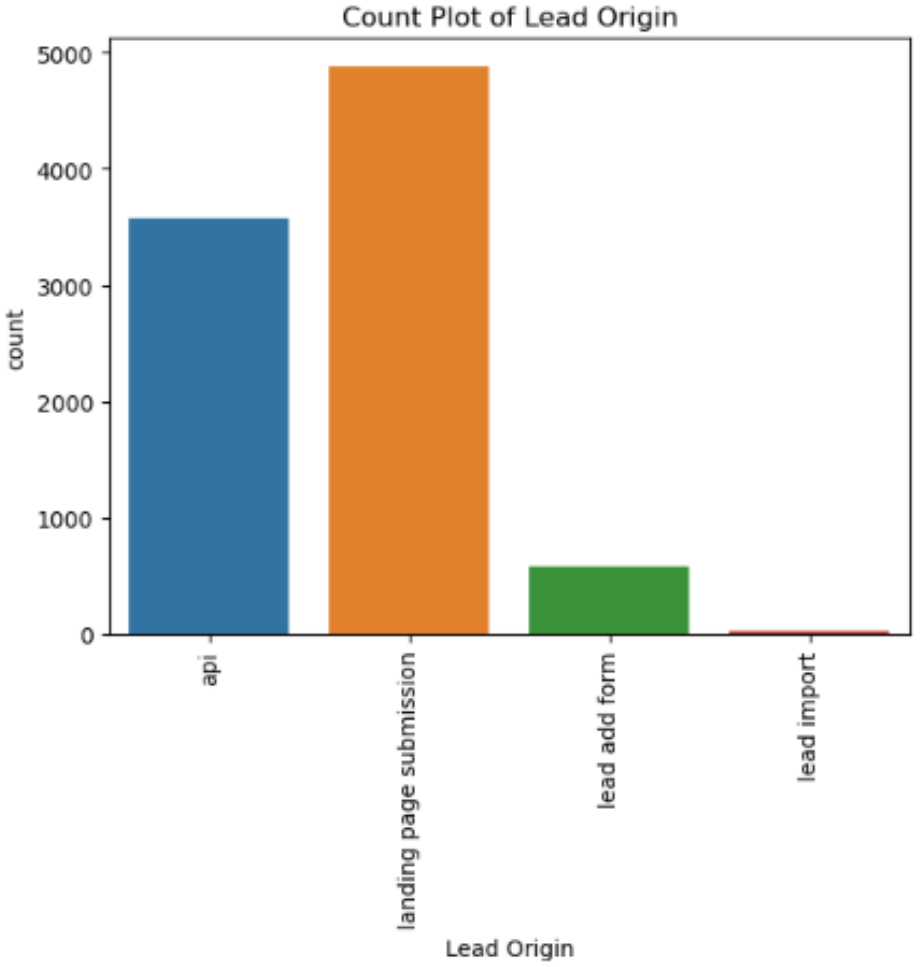




Analysis Approach:

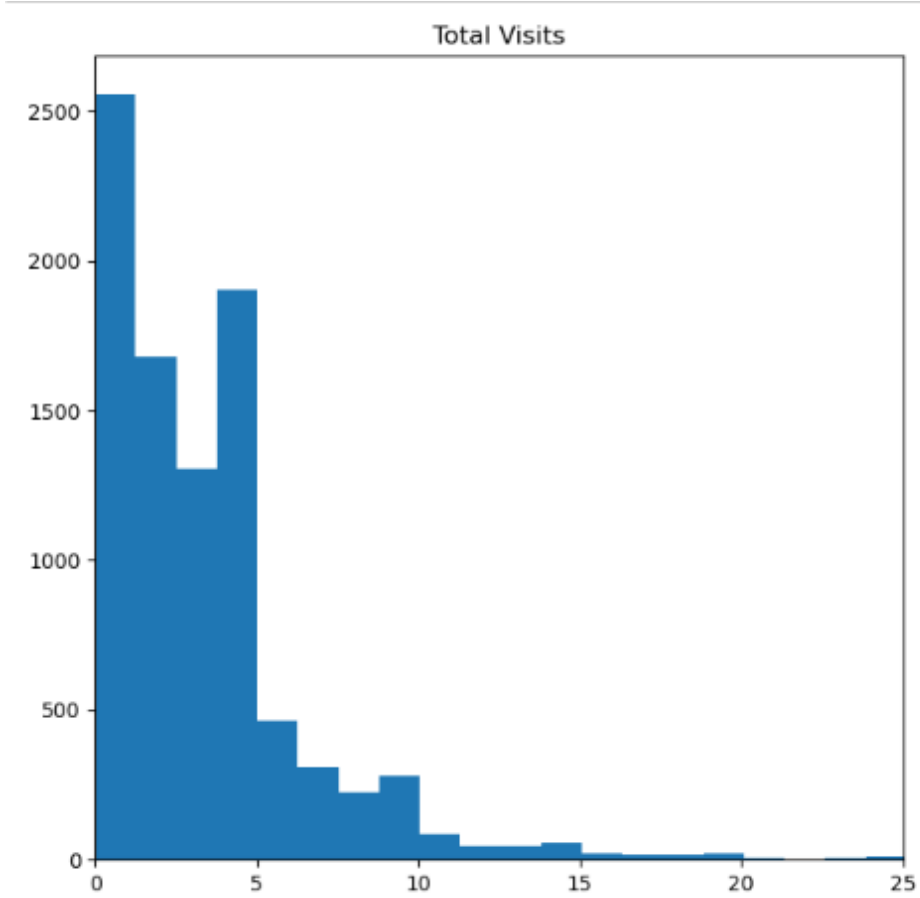
upGrad

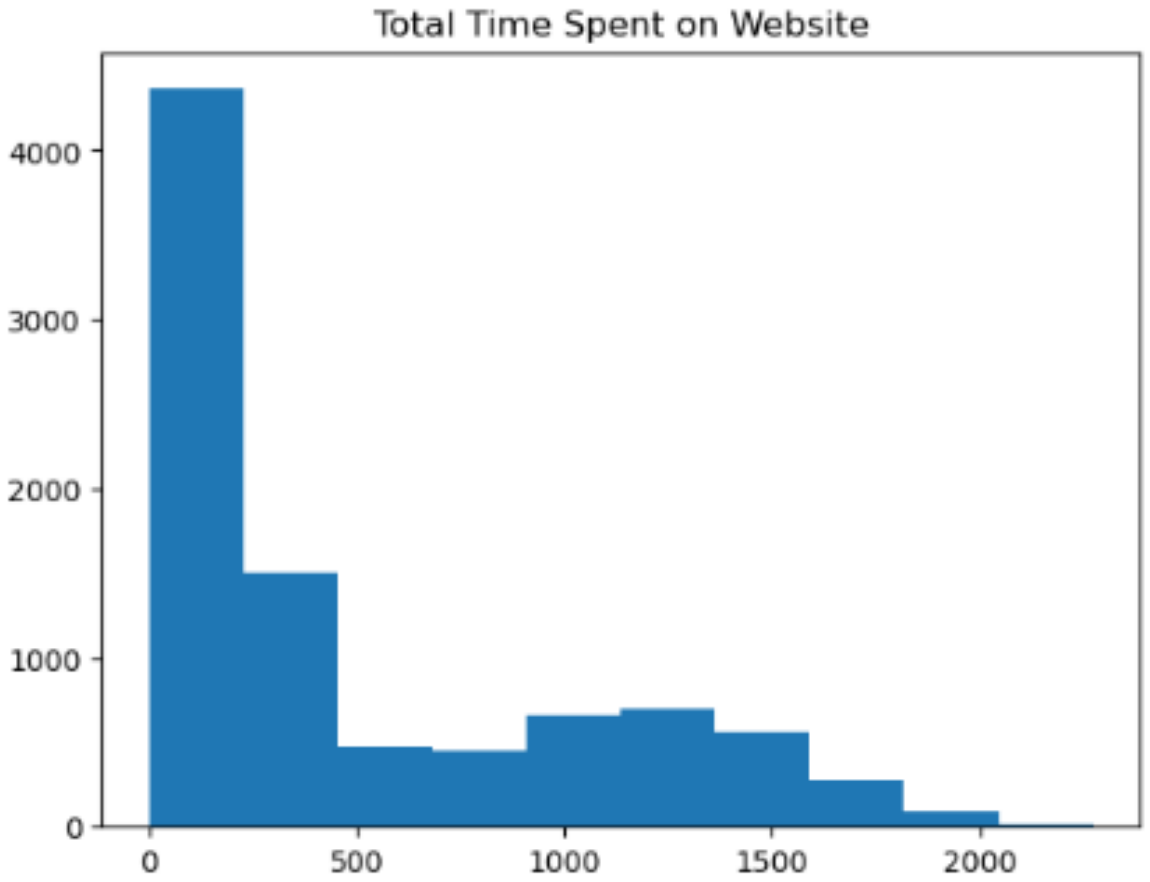




Analysis Approach:

upGrad



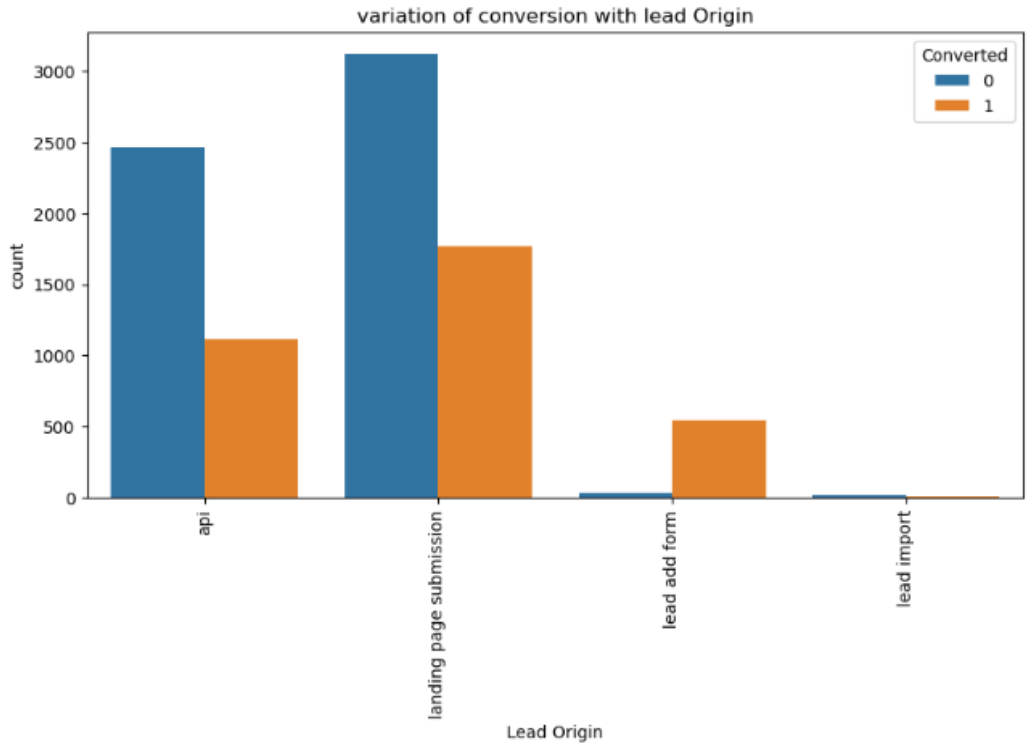


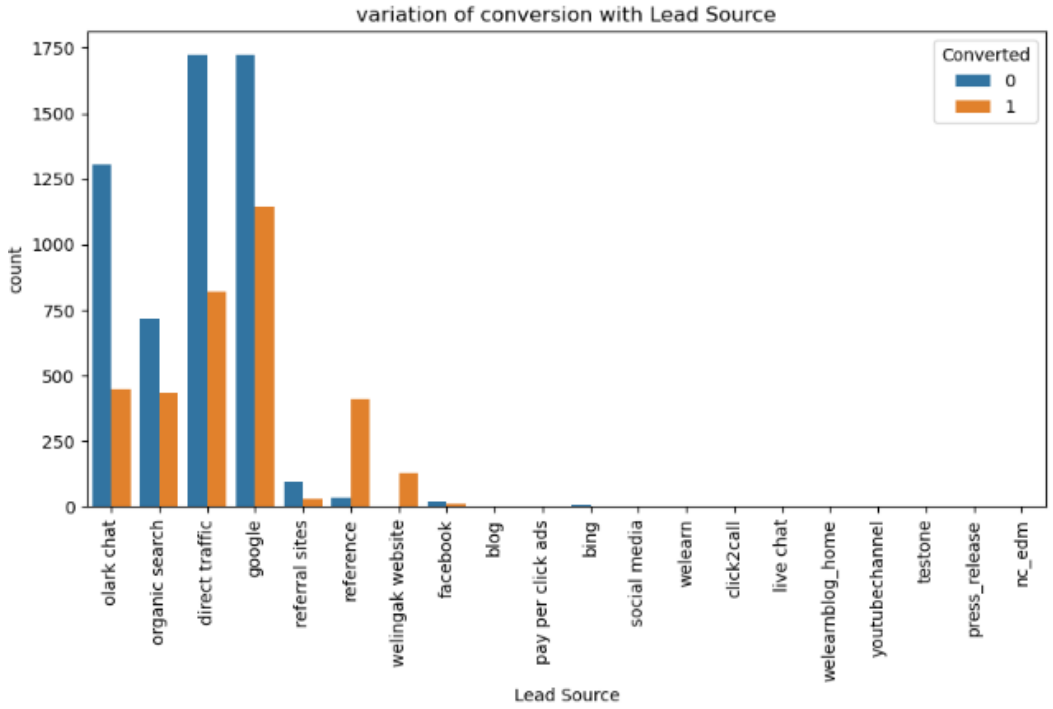
Analysis Approach:

upGrad

BIVARIATE ANALYSIS/MULTIVARIATE ANALYSIS

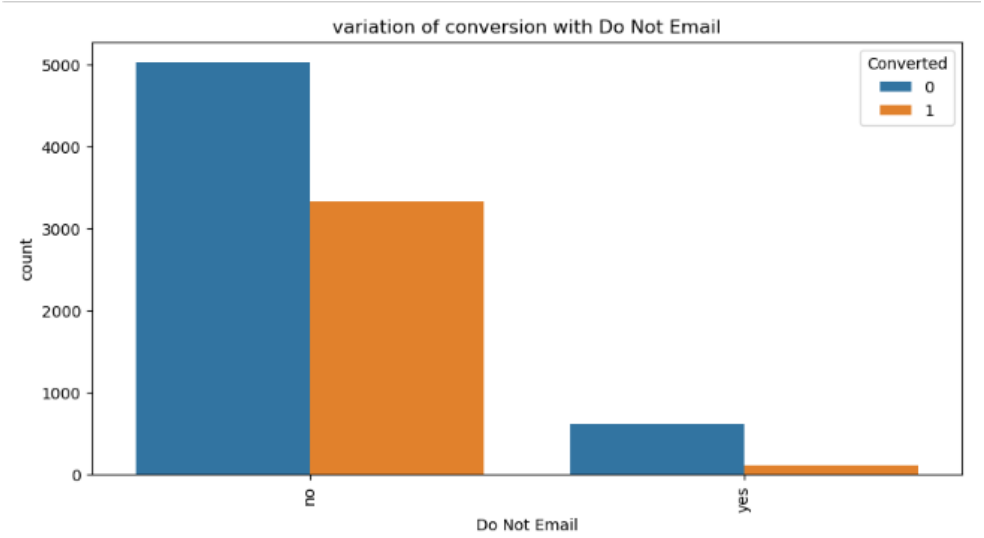
For bivariate and multivariate analysis I have used count plot to see relationship between different variables with target variable, to find correlation between different columns I have used correlation heatmaps.

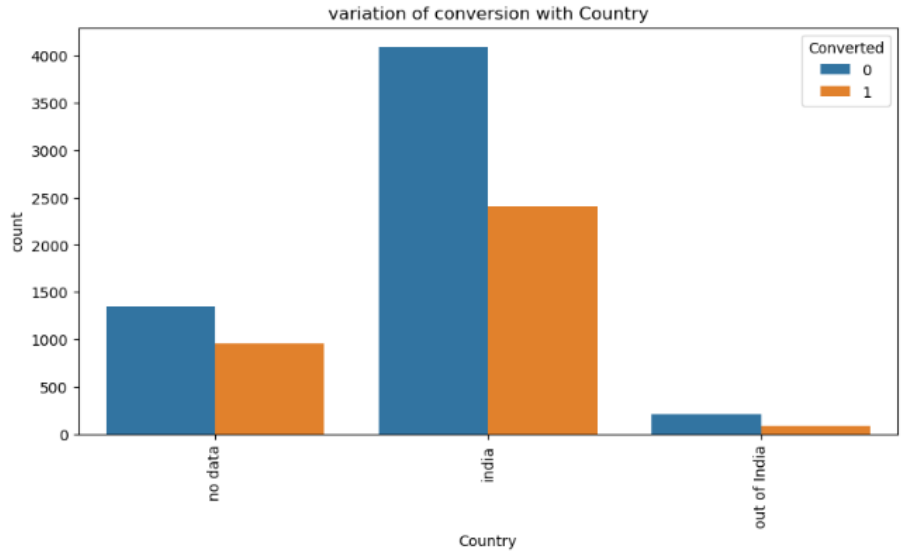


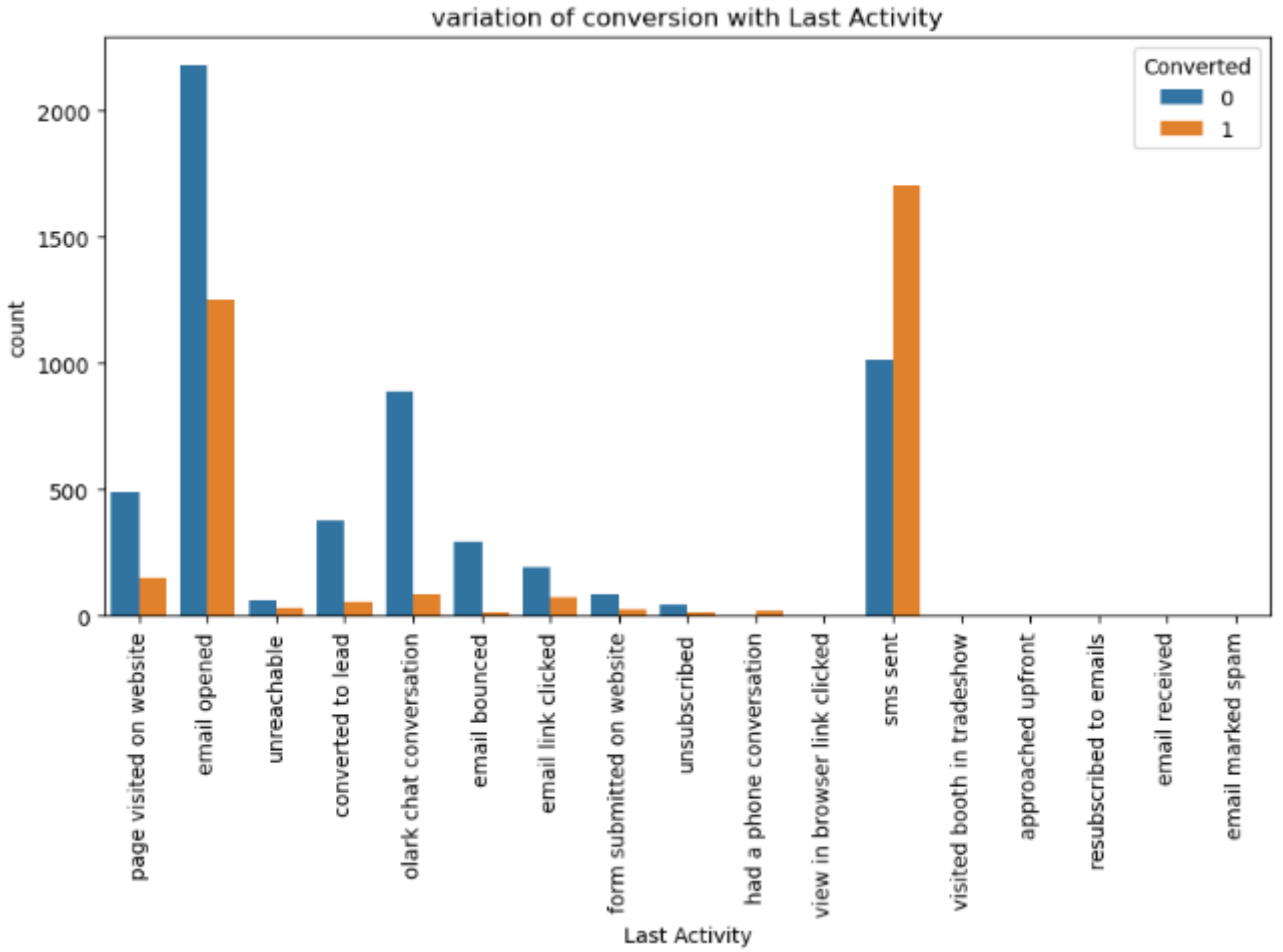


Analysis Approach:

upGrad

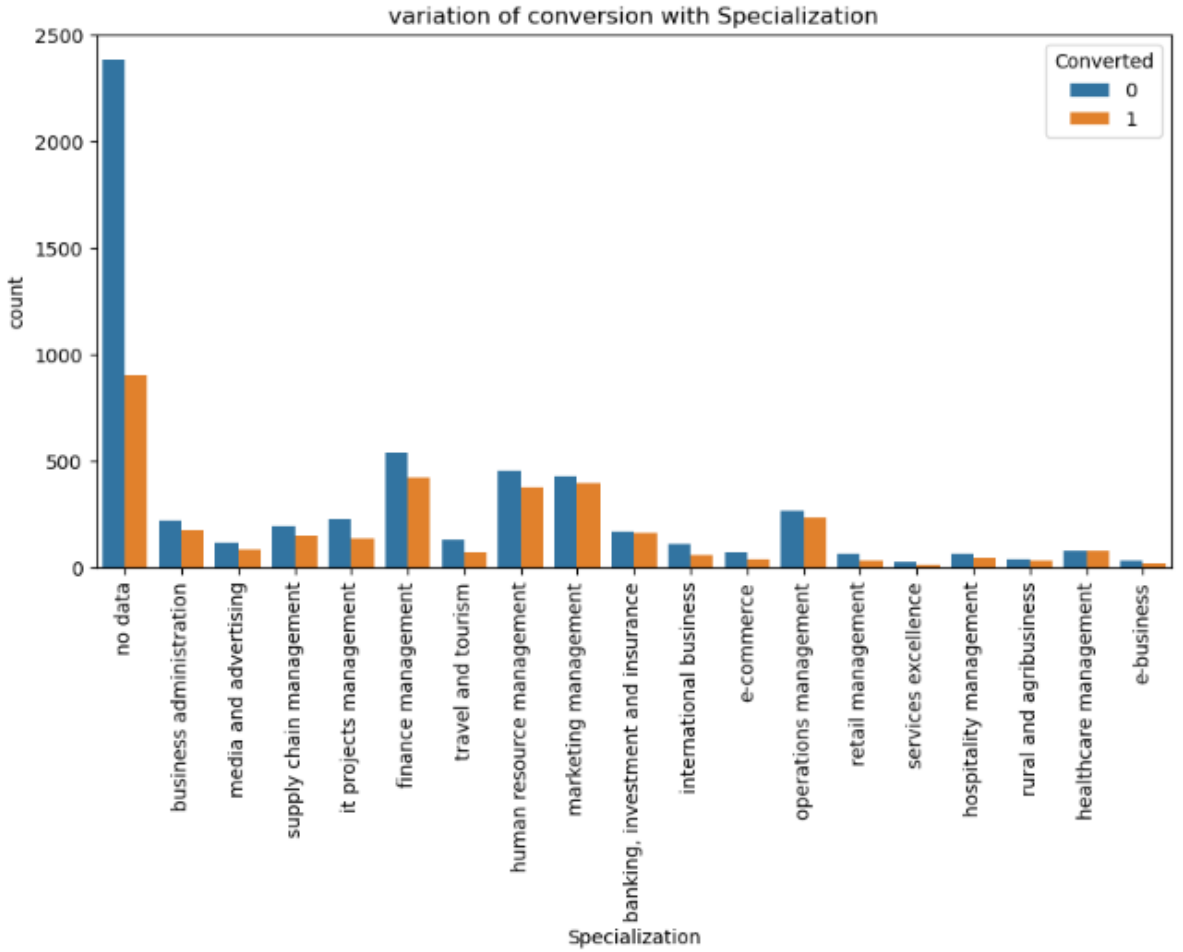


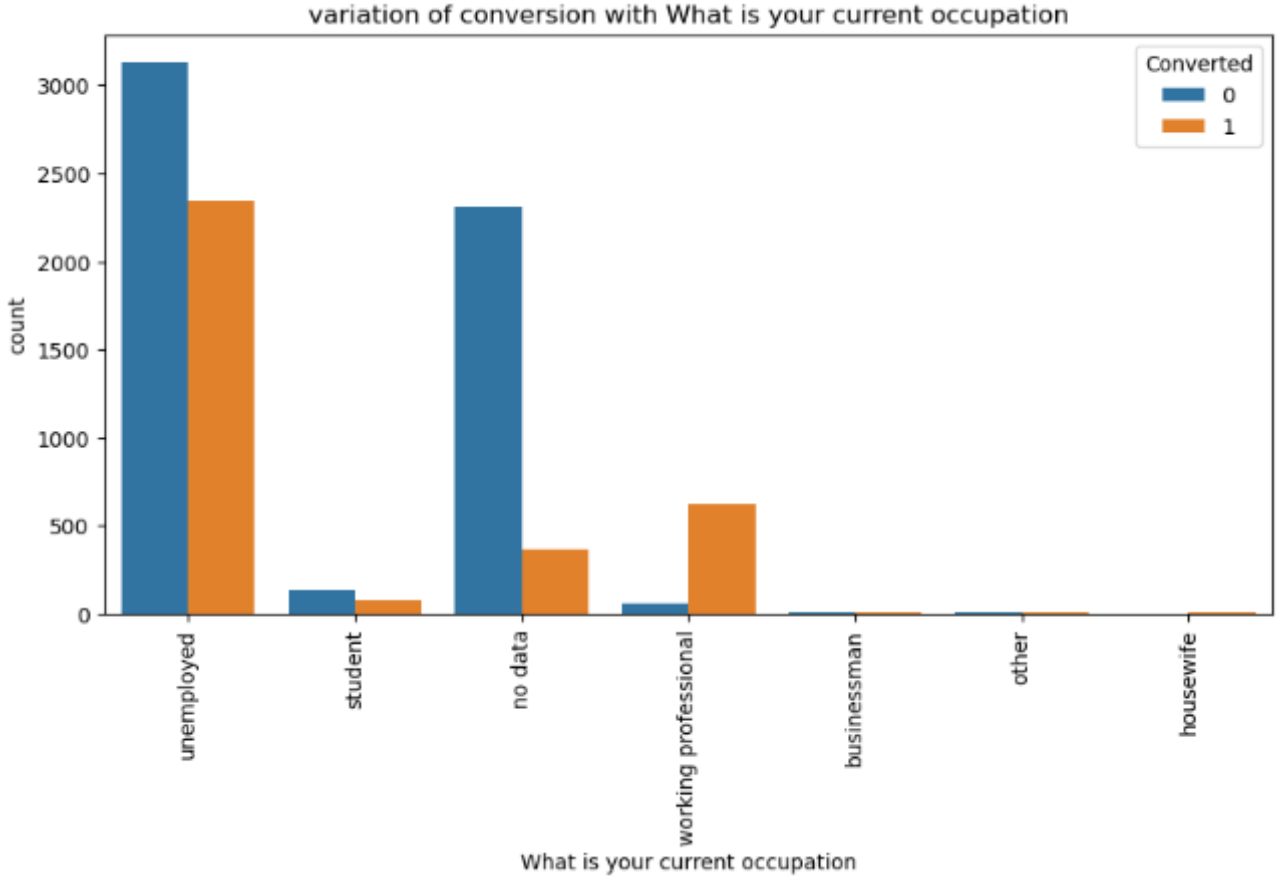




Analysis Approach:

upGrad

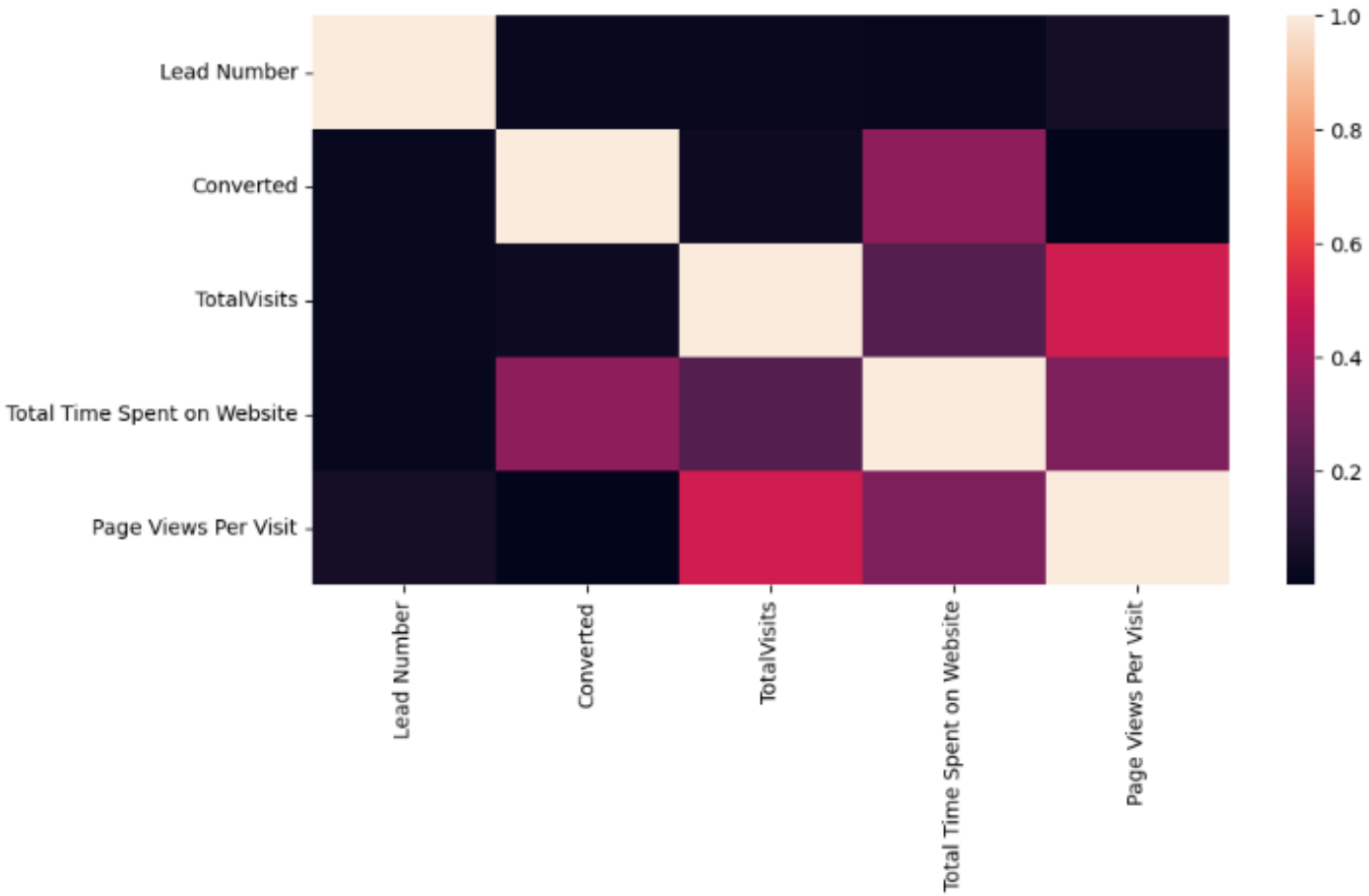




Analysis Approach:

upGrad

To find correlation among different columns I have used heatmaps.



Analysis Approach:

upGrad

Model Building:

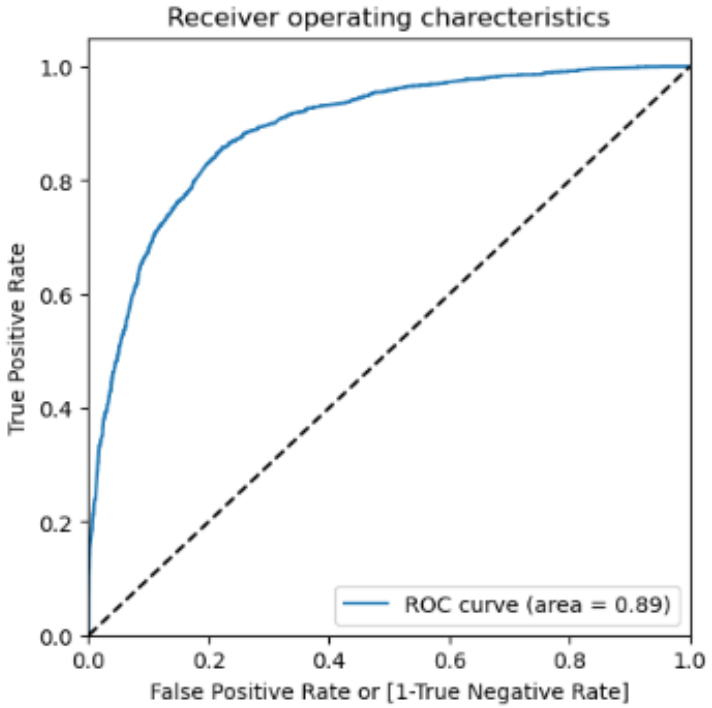
- We have split the data into training and test set. We have take 70% as train data and 30% as test data.
- Then we have used Recursive Feature Selection technique to bring out 15 most relevant features for model building.
- Then we started looking to the p values of the model built by these features.
- We dropped features one by one where p value was greater than 0.05 and variance inflation factor was going above 5.
- Then once the model was seem to be satisfactory we did predictions on test data with initial cutoff threshold of 0.5.
- With Initial cutoff of 0.5 we got an Accuracy of ~82%, sensitivity of ~66% and specificity of ~88%.

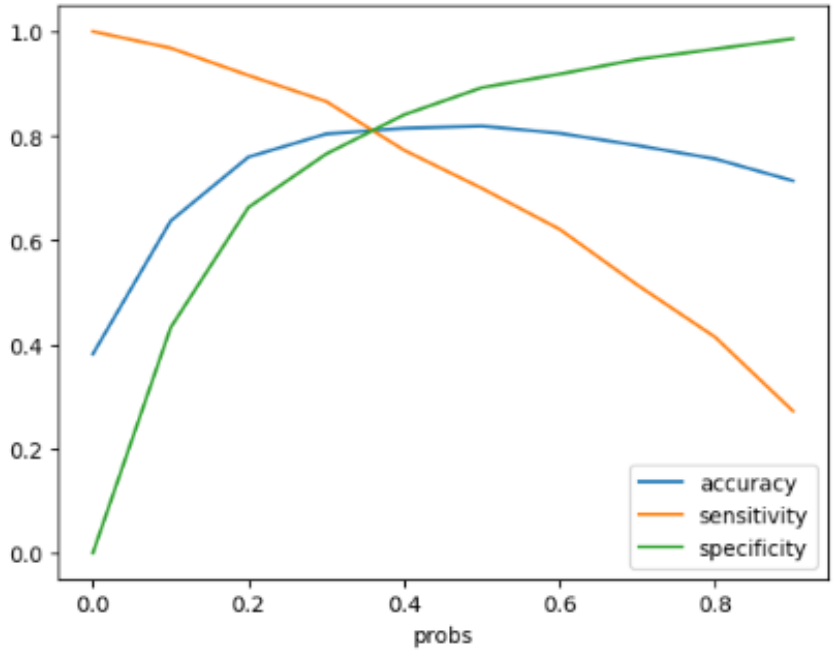
Analysis Approach:

upGrad

Receiver Operating Characteristics Curve:

- We use this to find optimal Cutoff points
- The optimal cutoff point lies where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cutoff would be 0.35
- With optimized cutoff of 0.35 we get accuracy of 79%, sensitivity of 80% and specificity of 78%



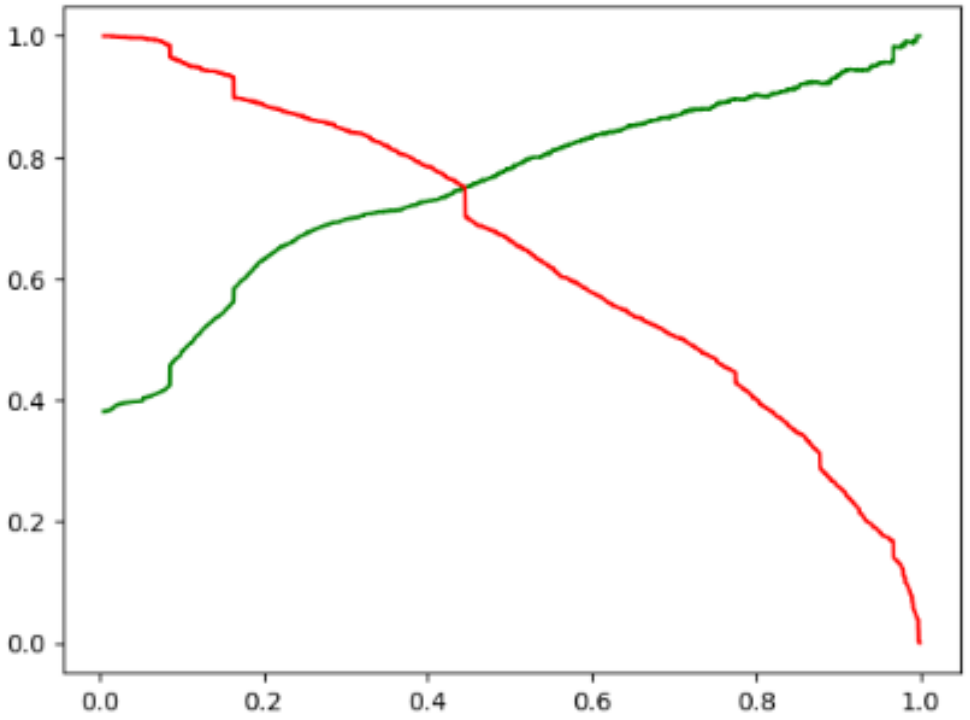


Analysis Approach:

upGrad

Precision Recall Tradeoff:

- We used the precision and recall tradeoff to bring a balance between precision and recall
- The optimal cutoff point from here we got as 0.43.
- The Accuracy here was 80% with precision as 72% and Recall as 74% which are good numbers



Conclusion

upGrad

Factors for Lead Conversion:

It was found that the variables that matters the most in building the regression model-

1. Total Time Spent on Website- This shows more the person is researching on the website and spending time on the website is a lead which has more chances to be converted.
2. TotalVisits- This shows more the number of times a customer visits the website more is the possibility of the conversion.
3. Page Viewed per visit alse shows how much a person is interested in joining the course. The more he researches the more hwe is interested and the more pages he will go through.
4. When lead source is from welingak website then there is a high chance of lead been converted to customer(learner).
5. When Last Notable Activity is email opened, olark chat conversation, page visited on website or sms sent then there is a high chance of lead been converted to customer(learner).
6. There is a high chances of conversion when Lead origin is lead add form
7. There is a high chances of conversion when the current occupation of the potential customer working professional.

Keeping these factors in mind X Education can make best of the profits by converting maximum potential leads to buy there courses.