Lead Scoring Case Study

SUBMITTED BY:

Preeti
Praveen
Preethi

Problem Statement

- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- The typical lead conversion rate at X education is around 30%. Now, this means 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 in order to let the conversation rate go up.

Business Objective

- 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.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches

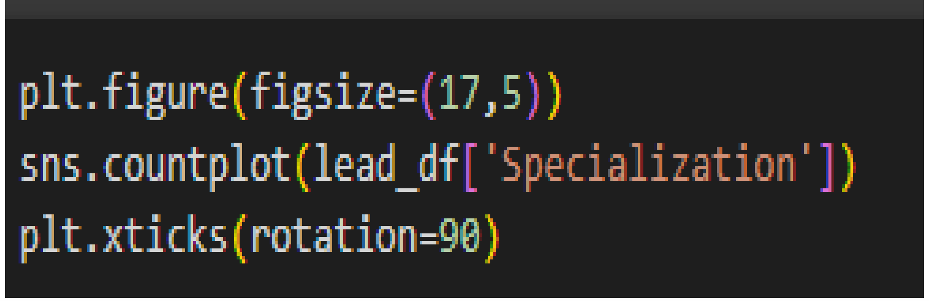
Problem Approach

- Reading & understanding the data
- Data cleaning ❖ EDA
- Feature scaling
- Splitting the data into test & train dataset
- Prepare the data for modelling
- Model building
- Model evaluation-specificity & Sensitivity or precision recall
- Making predictions on the test
- Assigning lead score
- Feature Importance Determination

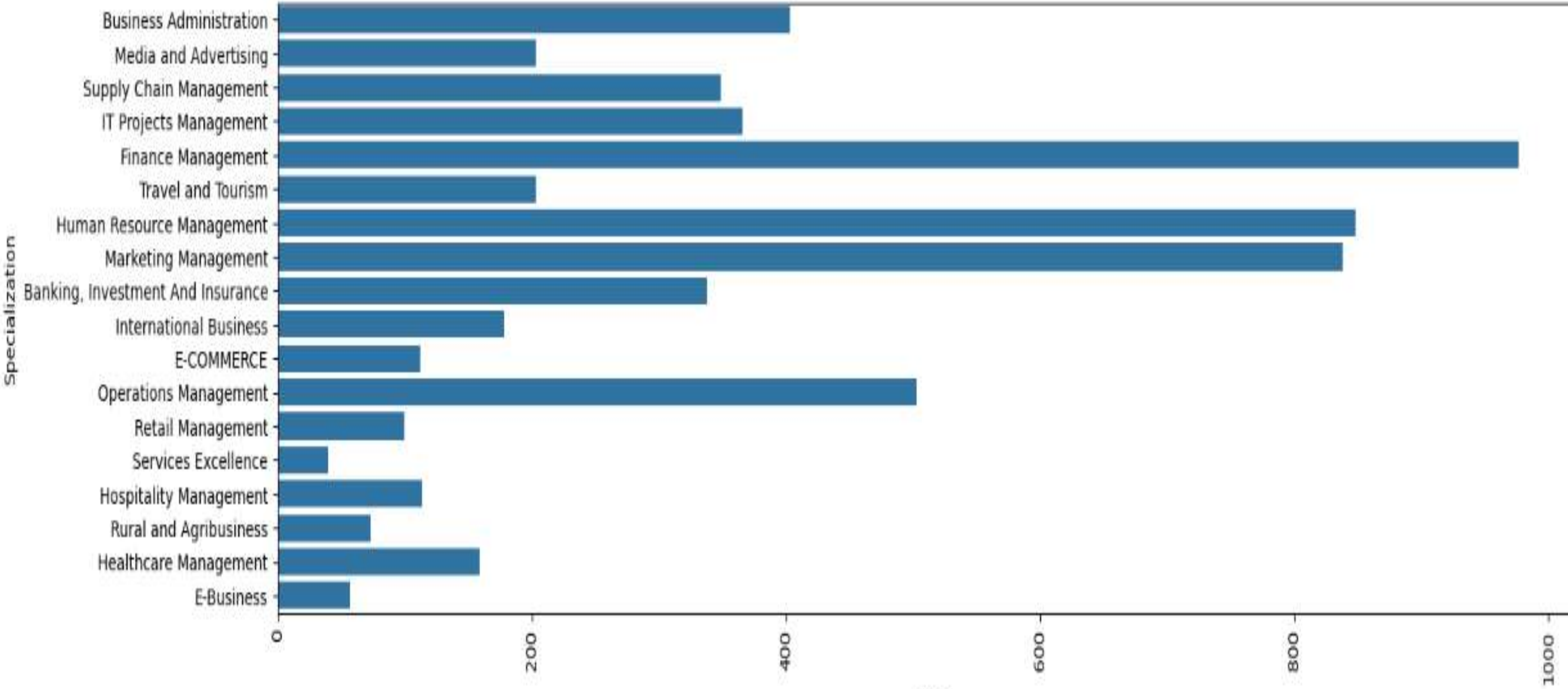
Data Cleaning

- There are some columns with high percentage of missing values are dropped straight away.
- Further, we can not eliminate all columns which might be useful and were having a great impact on our model but are having a strong possibility of high number of null values so we can treat those columns by imputing with mean and median by observing the type of the variables i.e. (continuous or categorical).
- The outliers which were found during the analysis has also been removed. After those process up to 98% data has been retained and on this cleaned data the analysis was performed

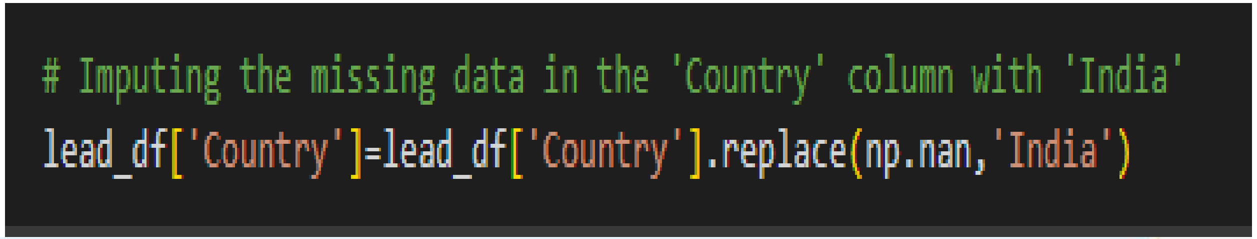
Column: 'Specialization' has 37% missing values



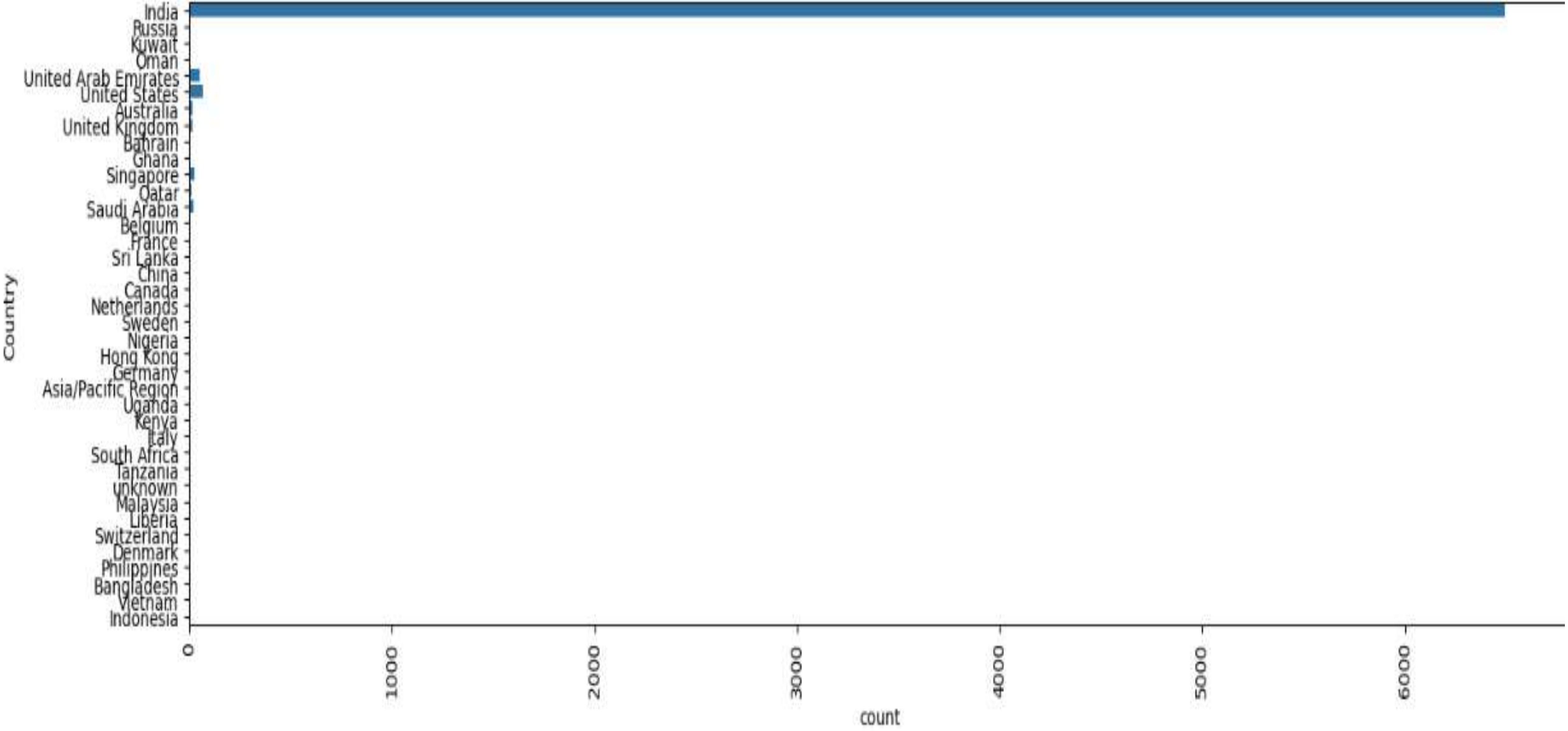
```
plt.figure(figsize=(17,5))
sns.countplot(lead_df['Specialization'])
plt.xticks(rotation=90)
```



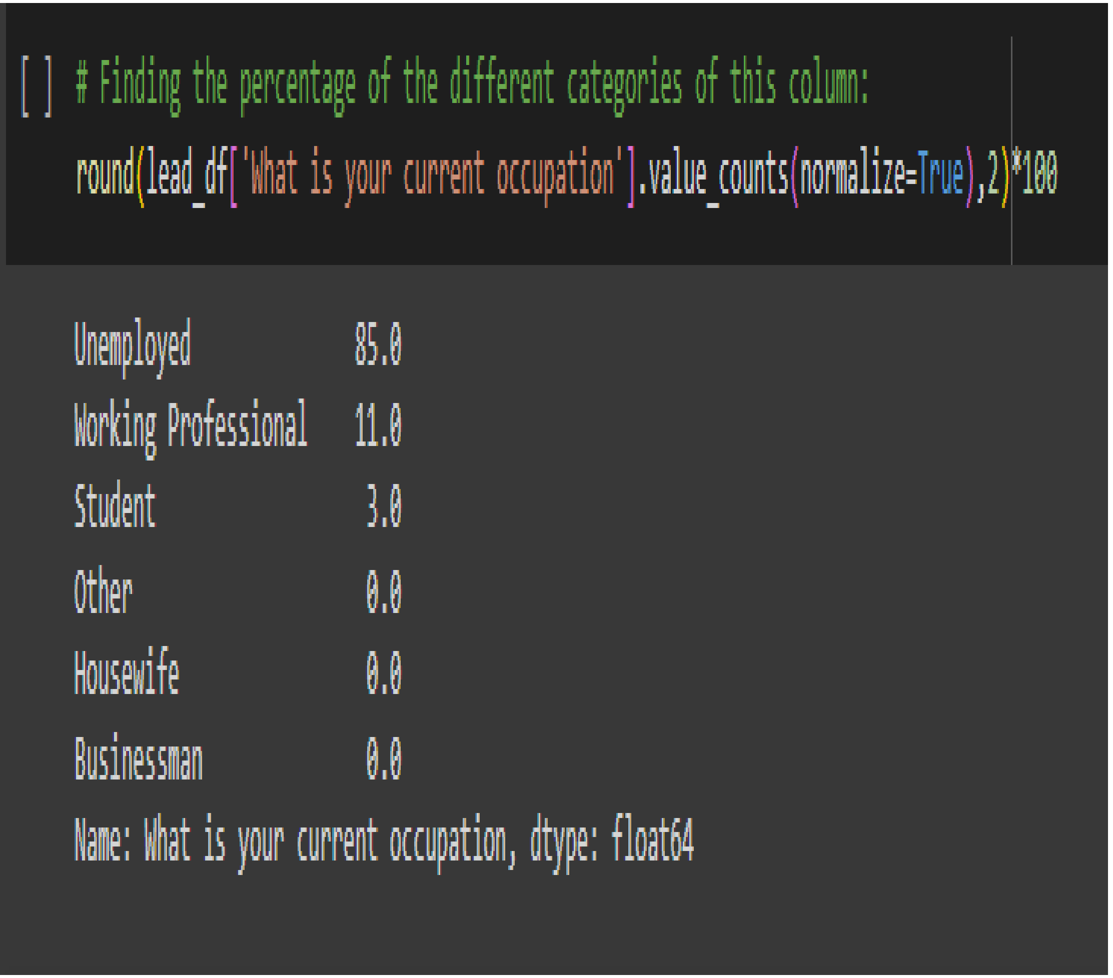
Column: 'Country' has 27% missing values



```
# Imputing the missing data in the 'Country' column with 'India'
lead_df['Country']=lead_df['Country'].replace(np.nan,'India')
```

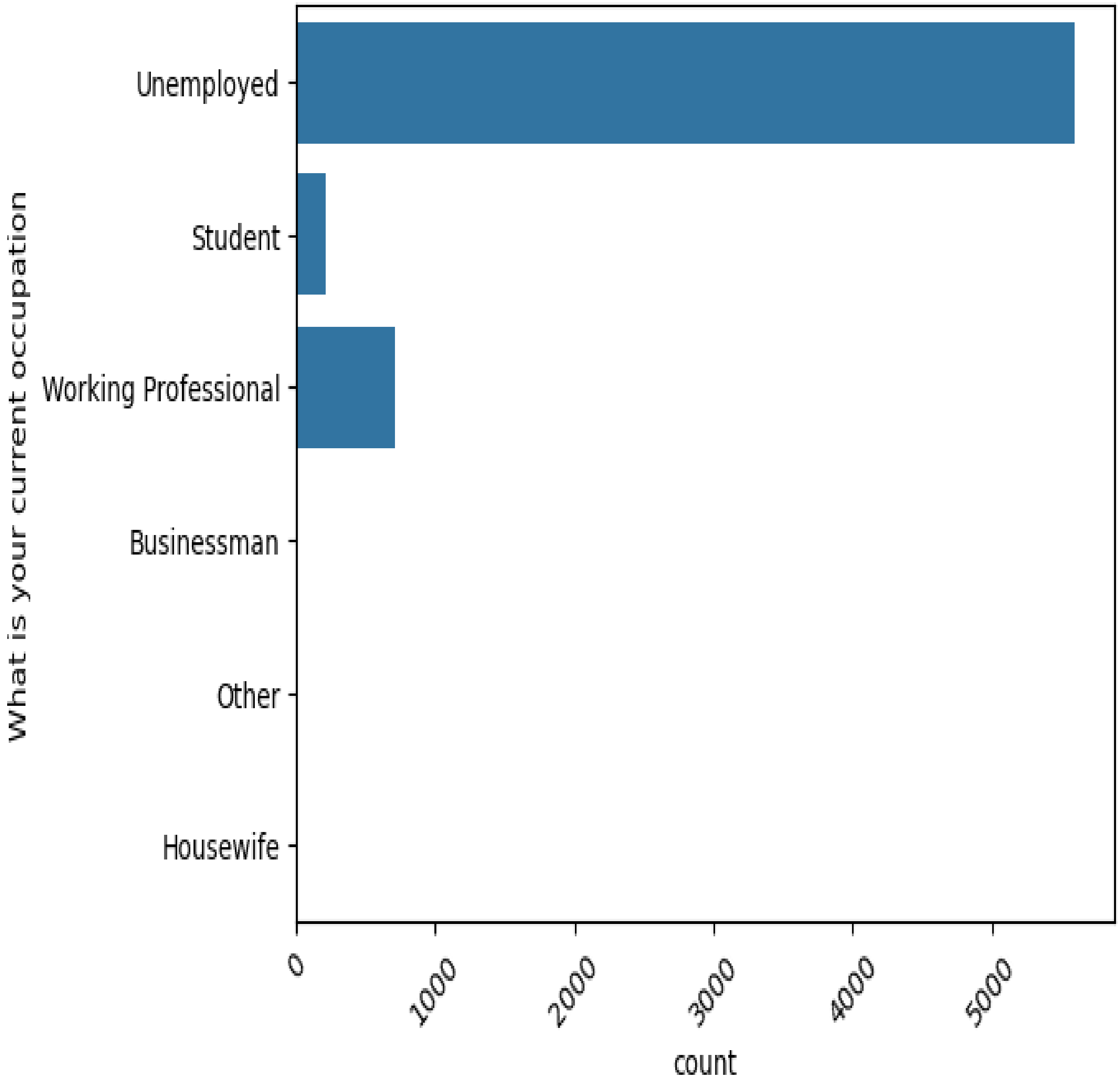


the most values are 'Unemployed' , we can impute missing values in this column with this value.



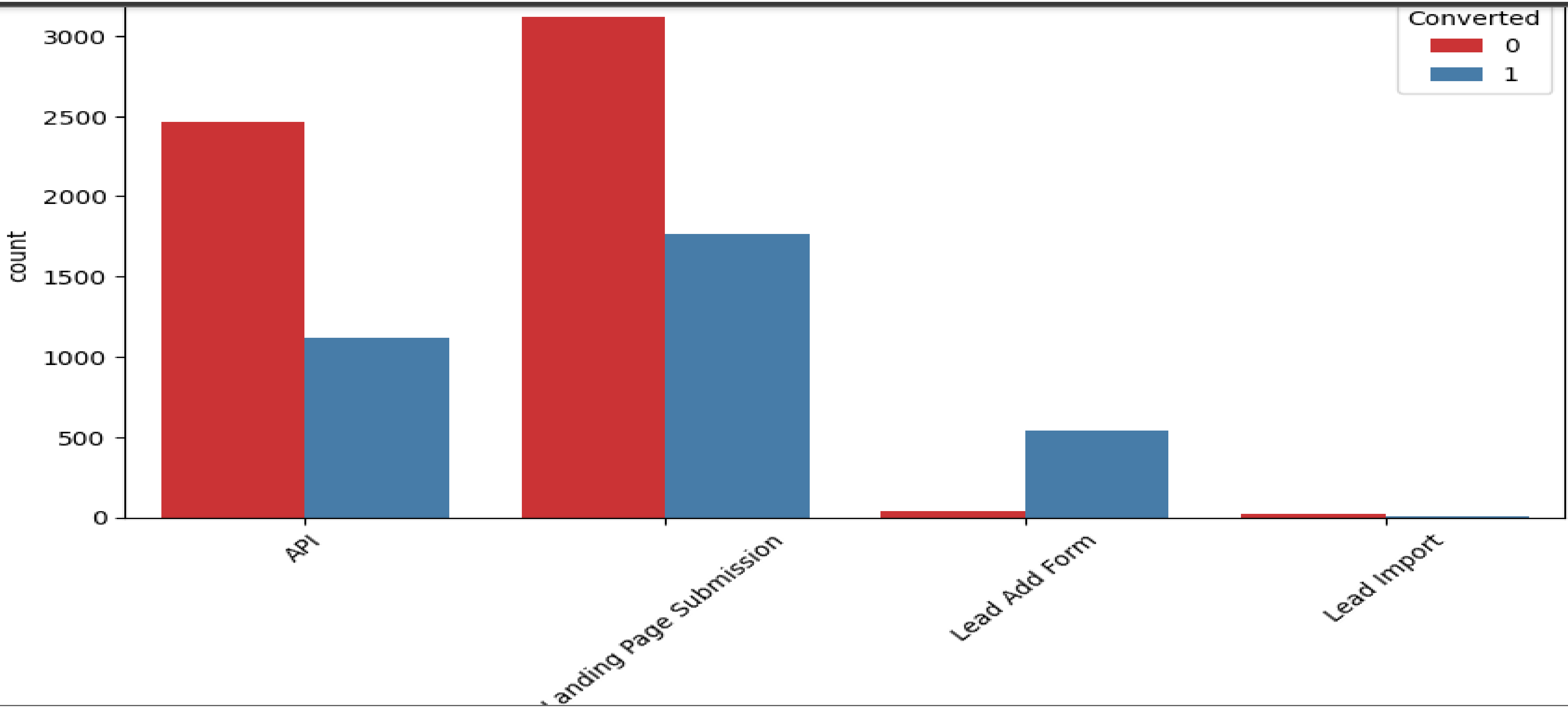
```
[ ] # Finding the percentage of the different categories of this column:
    round(lead_df['What is your current occupation'].value_counts(normalize=True),2)*100
```

```
Unemployed              85.0
Working Professional    11.0
Student                  3.0
Other                    0.0
Housewife                0.0
Businessman              0.0
Name: What is your current occupation, dtype: float64
```

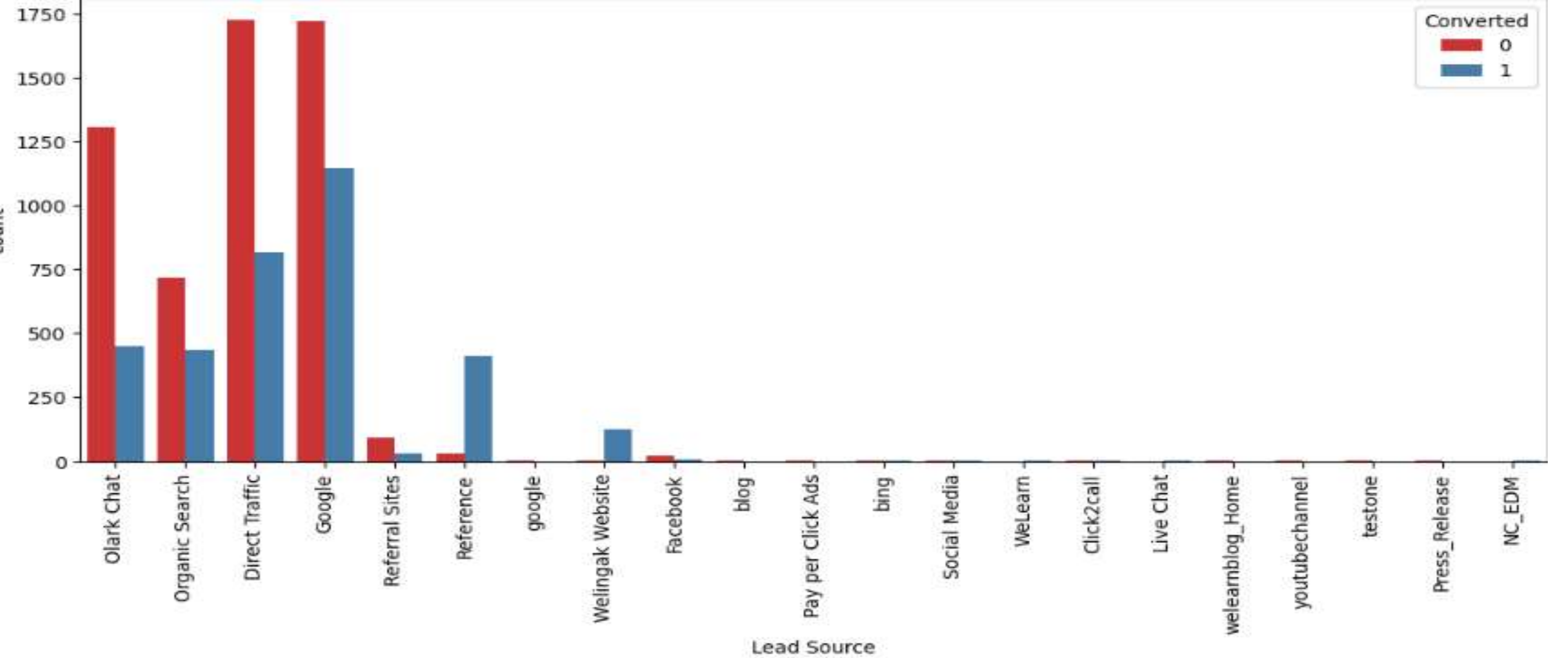


EDA

EDA was performed on the cleaned data by plotting different types of plots and analyzing both the variables which is continuous and categorical. Univariate analysis was done against the target variable for better understanding. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable and it also describes each variable on its own.



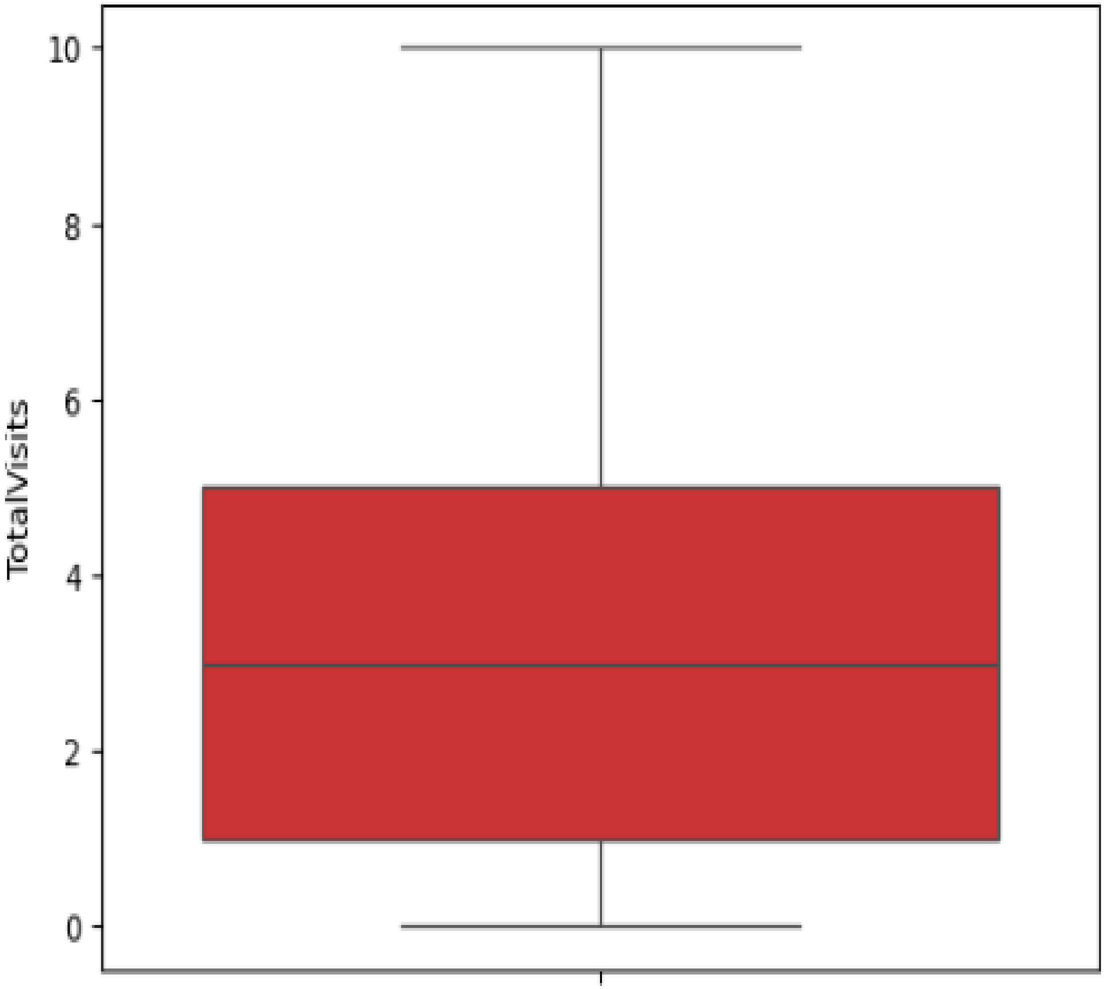
- PI and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.

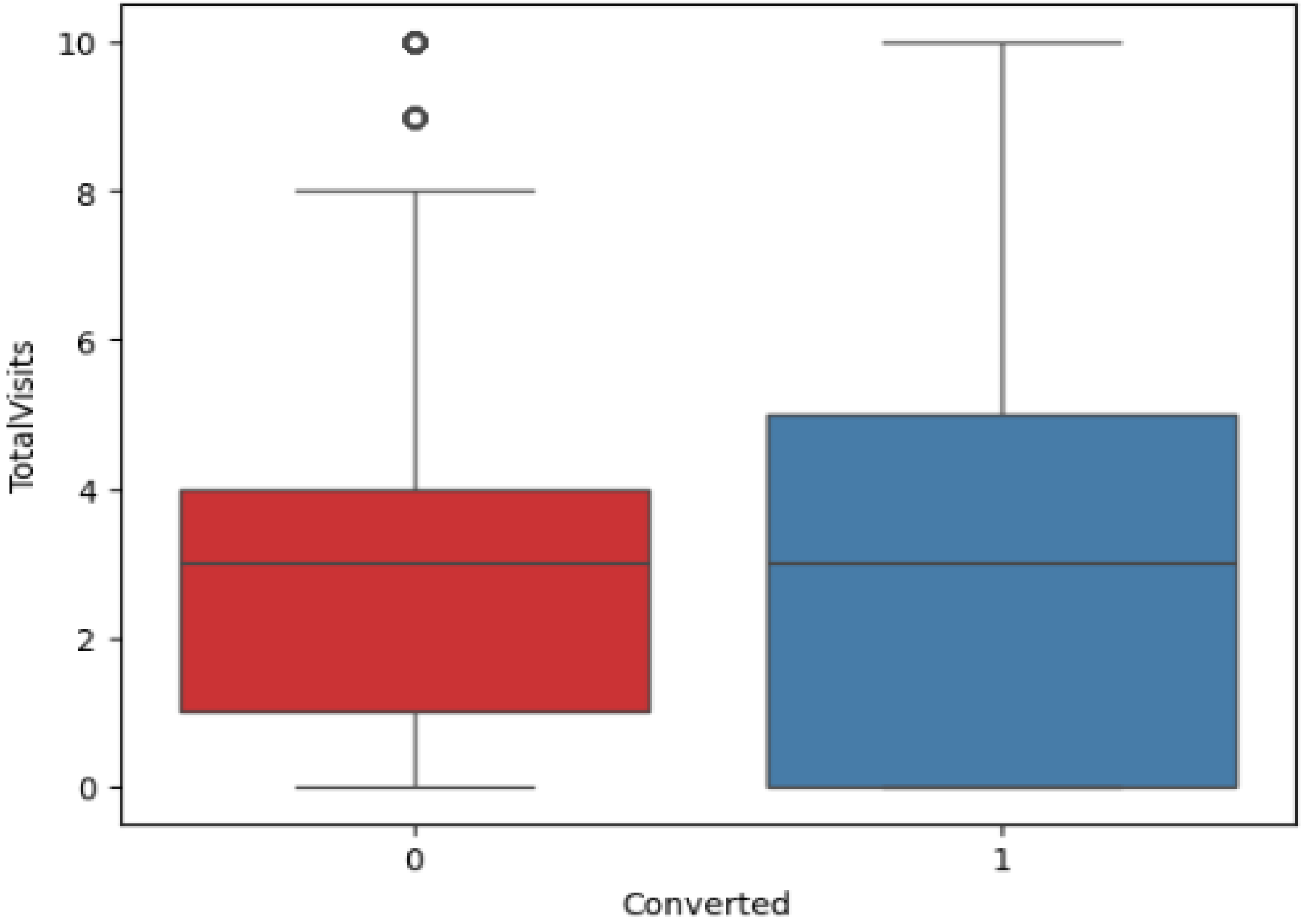


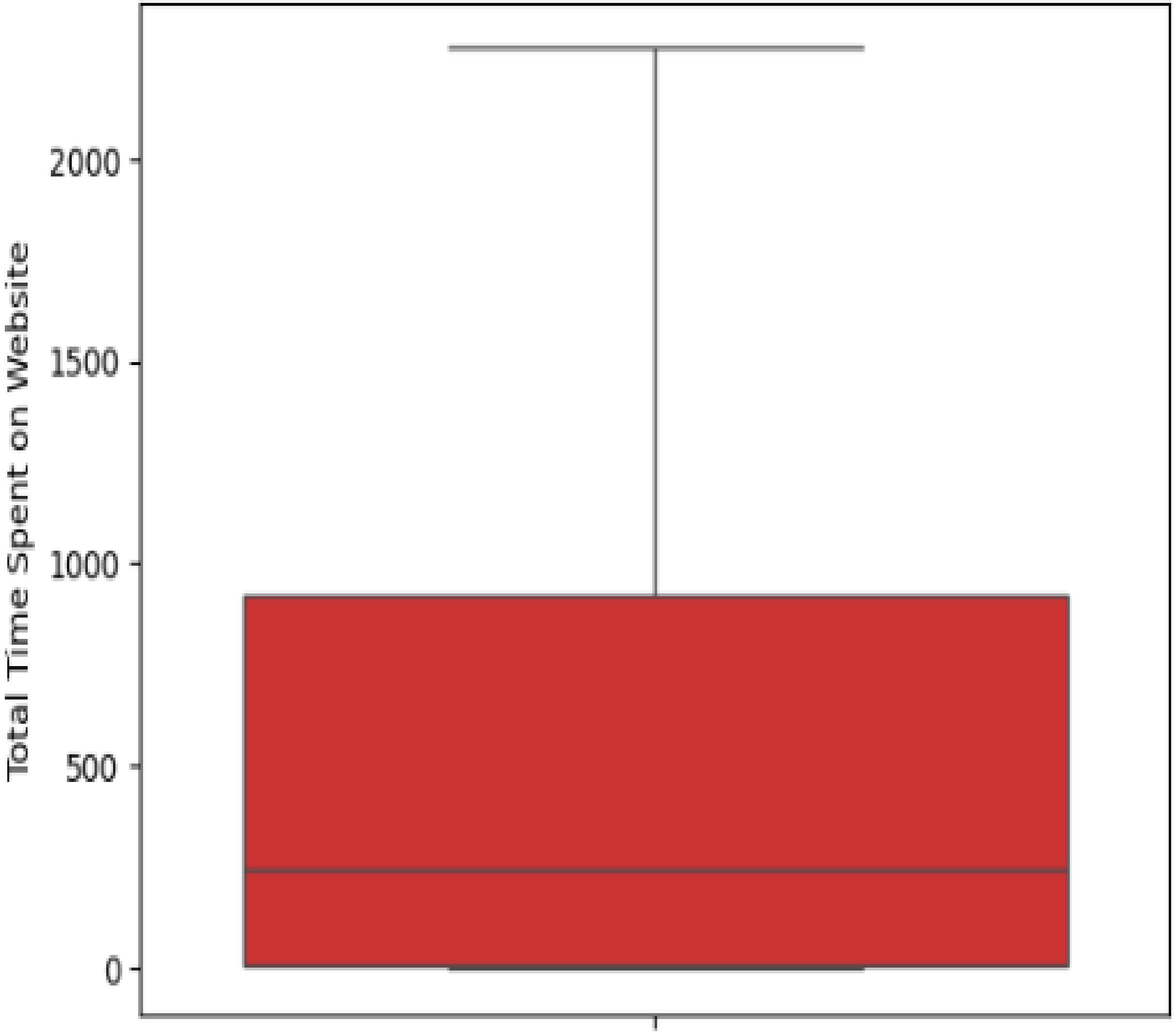
- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

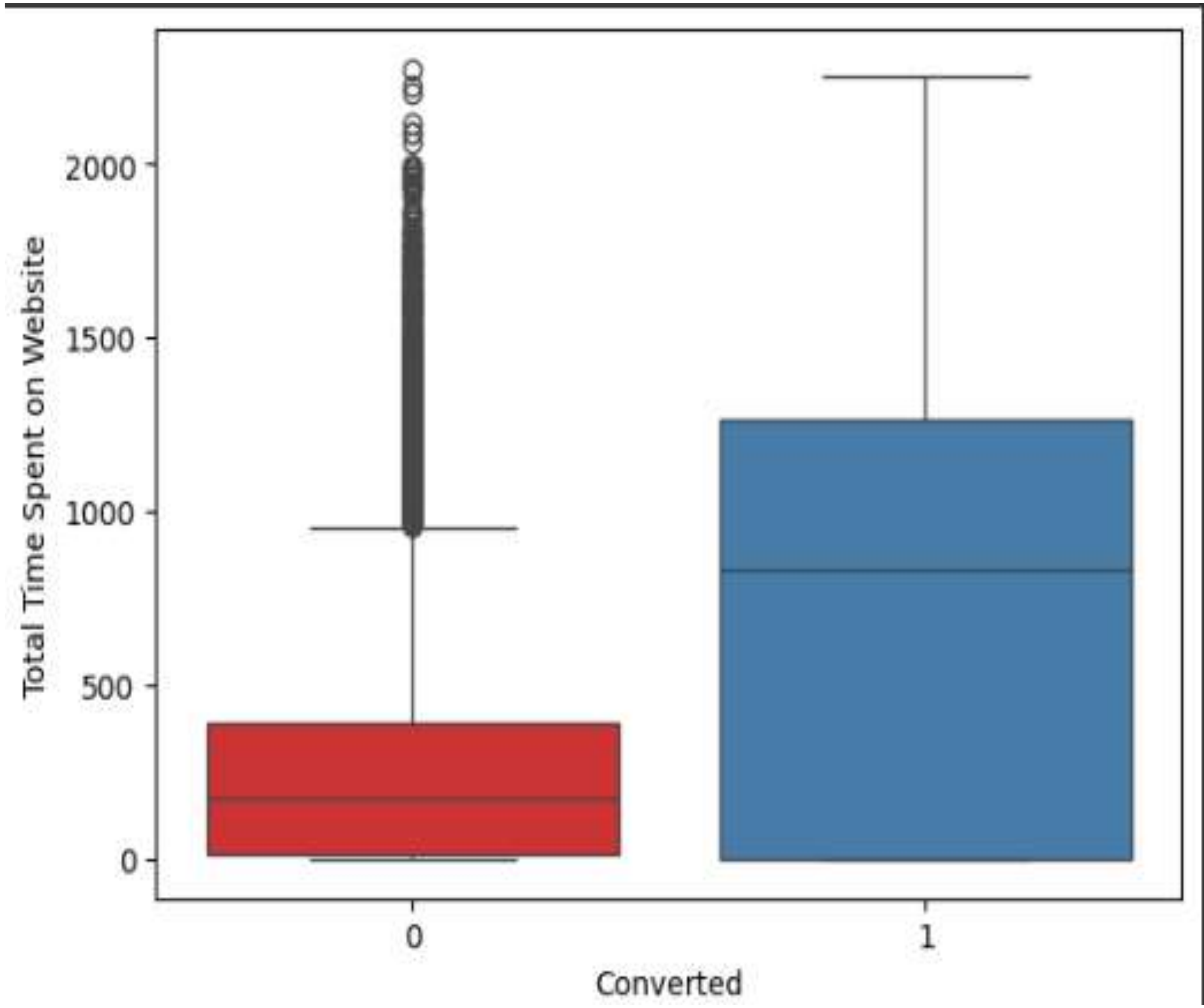
Numerical Analysis

- Median for converted and not converted leads are the same. Nothing can be concluded on the basis of Total Visits.



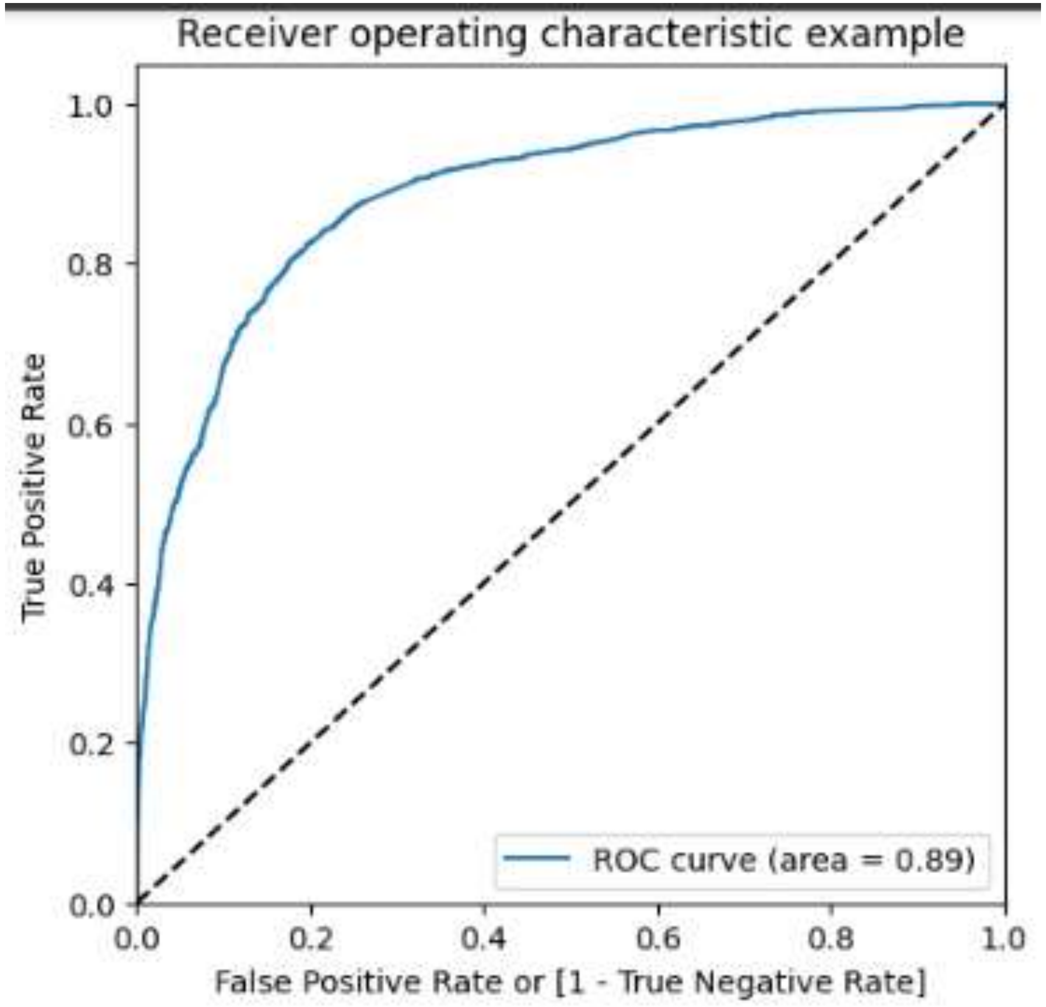


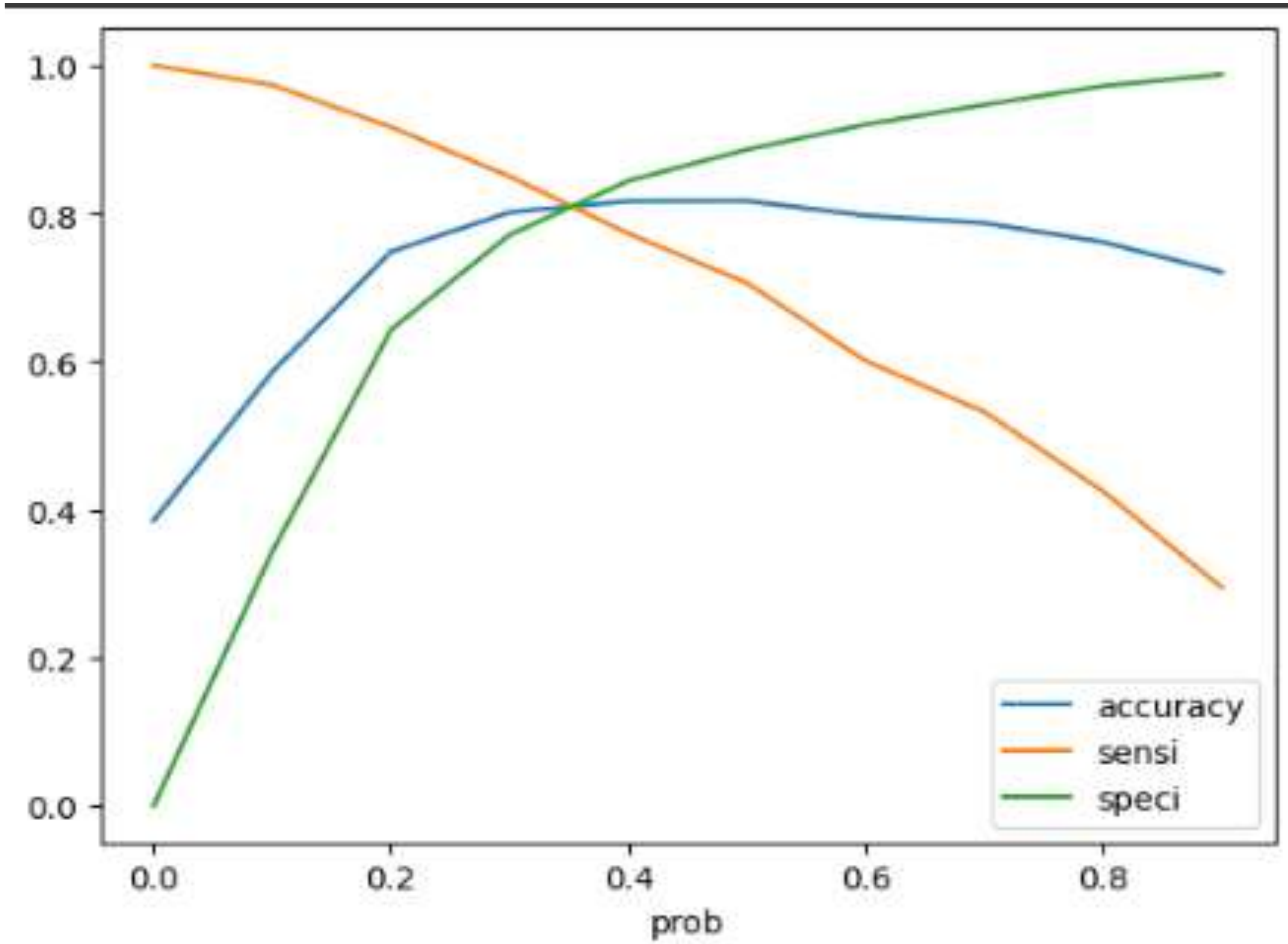




- Leads spending more time on the website are more likely to be converted.

Model Evaluation: Plotting the ROC Curve





Observations:

After running the model on the Test Data , we obtain:

- Accuracy : 80.4 %
- Sensitivity : 80.4 %
- Specificity : 80.5 %

Results : Comparing the values obtained for Train & Test:

Train Data:

- Accuracy : 81.0 %
- Sensitivity : 81.7 %
- Specificity : 80.6 %

Test Data:

- Accuracy : 80.4 %
- Sensitivity : 80.4 %
- Specificity : 80.5 %

Conclusion:

- People sending higher than average are promising lead, so targeting them and approaching them can be helpful in conversion.
- Reference and offers for referring a lead can be a good source for higher conversion.
- Leads who spent more time on website, more likely to convert
- We see that the conversion rate is 30-35% (close to average) for API and Landing page submission. But very low for Lead Add form and Lead import. Therefore we can intervene that we need to focus more on the leads originated from API and Landing page submission
- The model shows high close to 81% accuracy.
- The model shows 76% sensitivity and 83% specificity.
- The model finds correct promising leads and leads that have less chances of getting converted.

Thank you