LEAD SCORING CASE STUDY

BY: JYOTSNA SHUKLA

SYEDA SHEEBA NYER

Problem Statement:

- X Education sells online courses to industry professionals. 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%.

Business Goal:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Methodology

- **Data Cleaning and Data Manipulation.**

1. Check and handle duplicate data.
2. Check and handle N/A and missing values.
3. Drop columns, if it contains large number of missing values and not useful for further analysis.
4. Imputation of values, if needed.
5. Check for outliers and outlier treatment.

- **EDA**

1. Univariate data analysis: value count, distribution of variables etc.
2. Bivariate data analysis: correlation between variables.

- **Feature scaling and creation of dummy variables.**
- **Classification techniques: Logistic regression used for the model and prediction.**
- **Validation of the model**
- **Model Presentation**
- **Conclusions and recommendations**

Data Manipulation

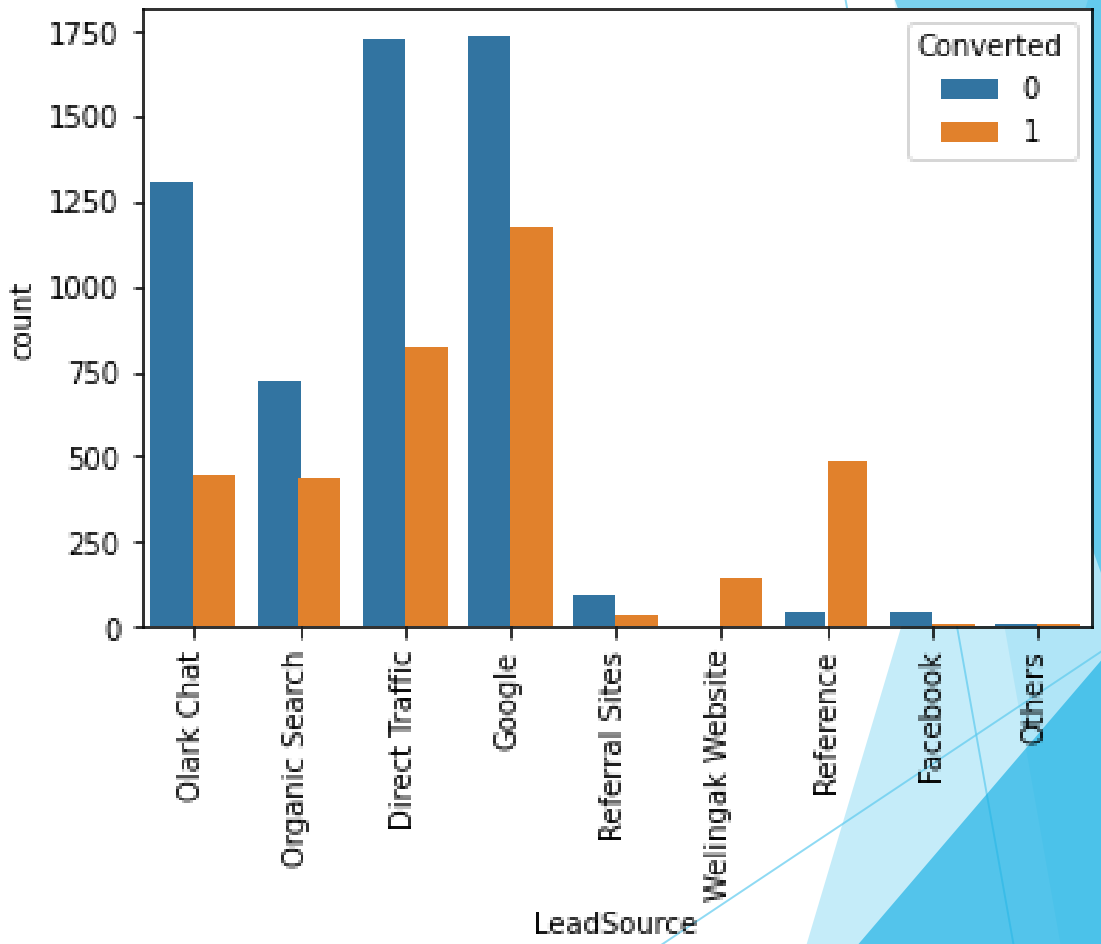
- At first, there were total 9240 rows and 37 columns.
- Removing the Lead Number and Prospect ID columns which are not needed for the analysis.
- Dropping the columns having more than 40% of missing values.
- Single value features like “Update me on supply chain content”, “Magazine”, “Receive more updates about our courses”, “I agree to pay the amount through cheque”, “Get updates on DM Content” have been dropped.
- After checking for value counts for some of the variables of object type, we found that there is no much variance, so we dropped the features like, “Do Not Call”, “Search”, “Newspaper” etc

EDA

- **Inferences:**

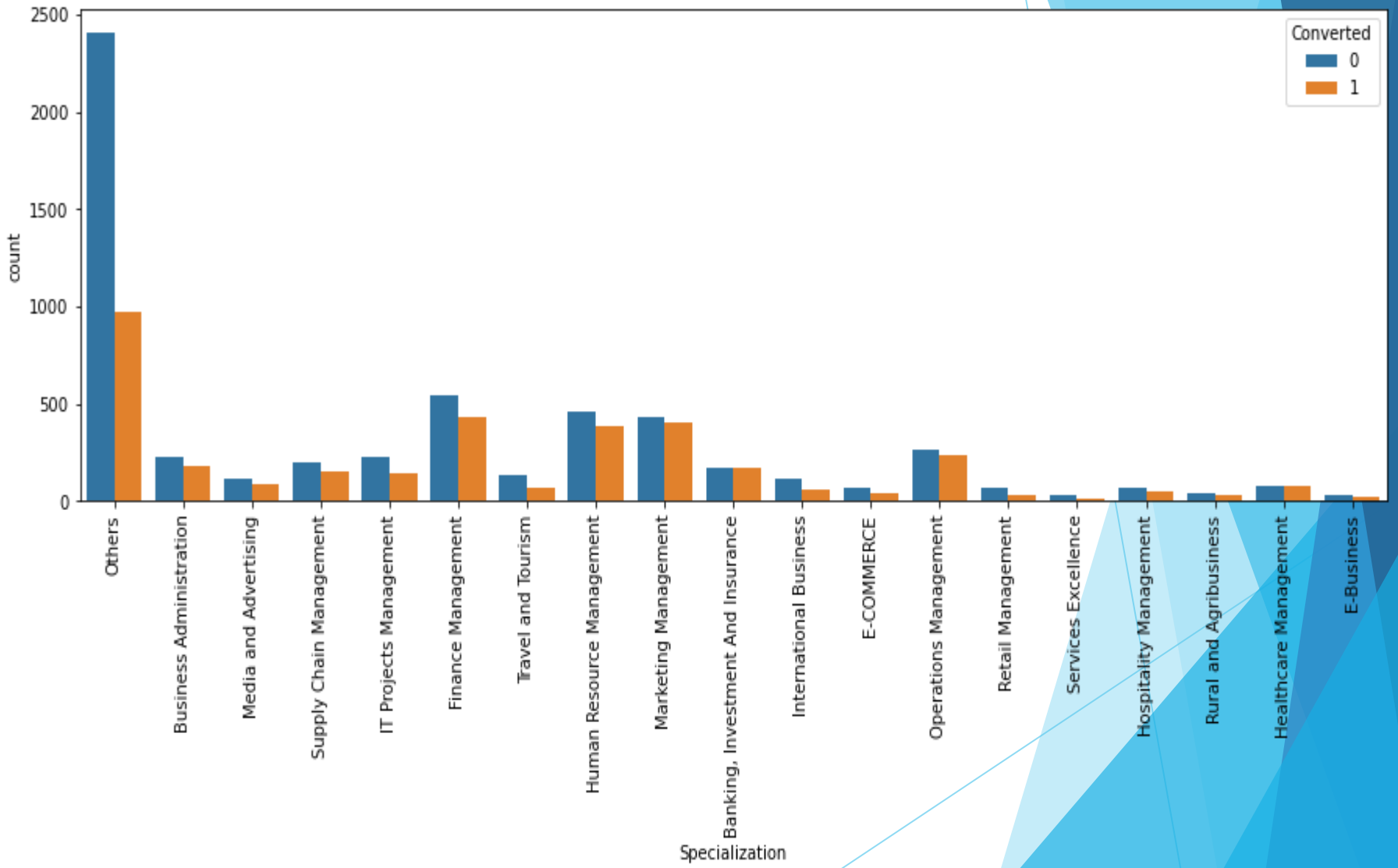
1. Google and Direct traffic generates maximum number of leads.
2. Conversion Rate of reference leads and leads through welingak website is high.

- **To improve overall lead conversion rate, our 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.**



Specialization

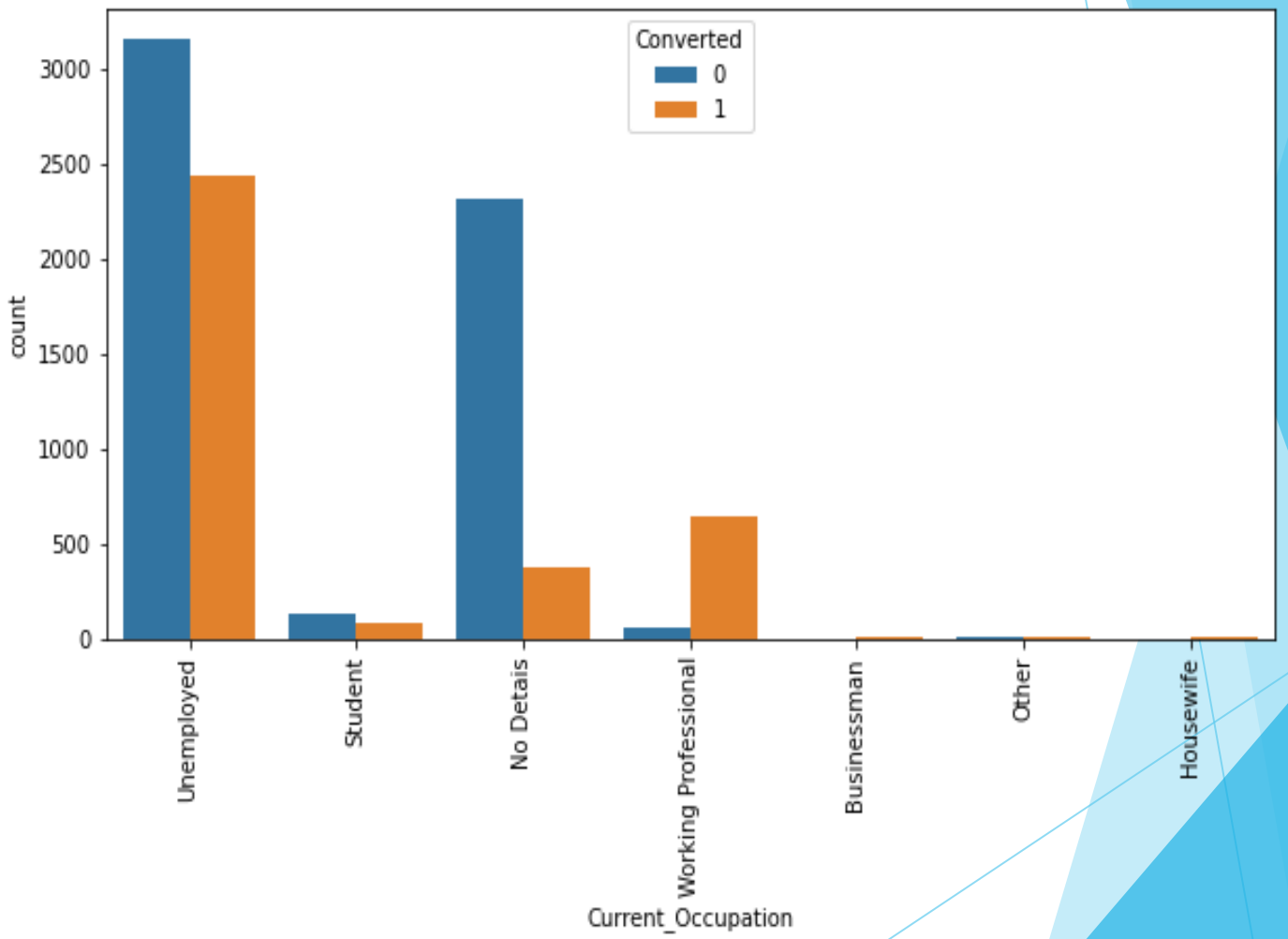
- It maybe the case that lead has not entered any specialization if his/her option is not available on the list, he/she may not have any specialization or is a student. Hence we have made a category "Others" for missing values.
- After “Others” category, the lead conversion can be seen highly in “Finance Management” category.



Current Occupation

Inferences:

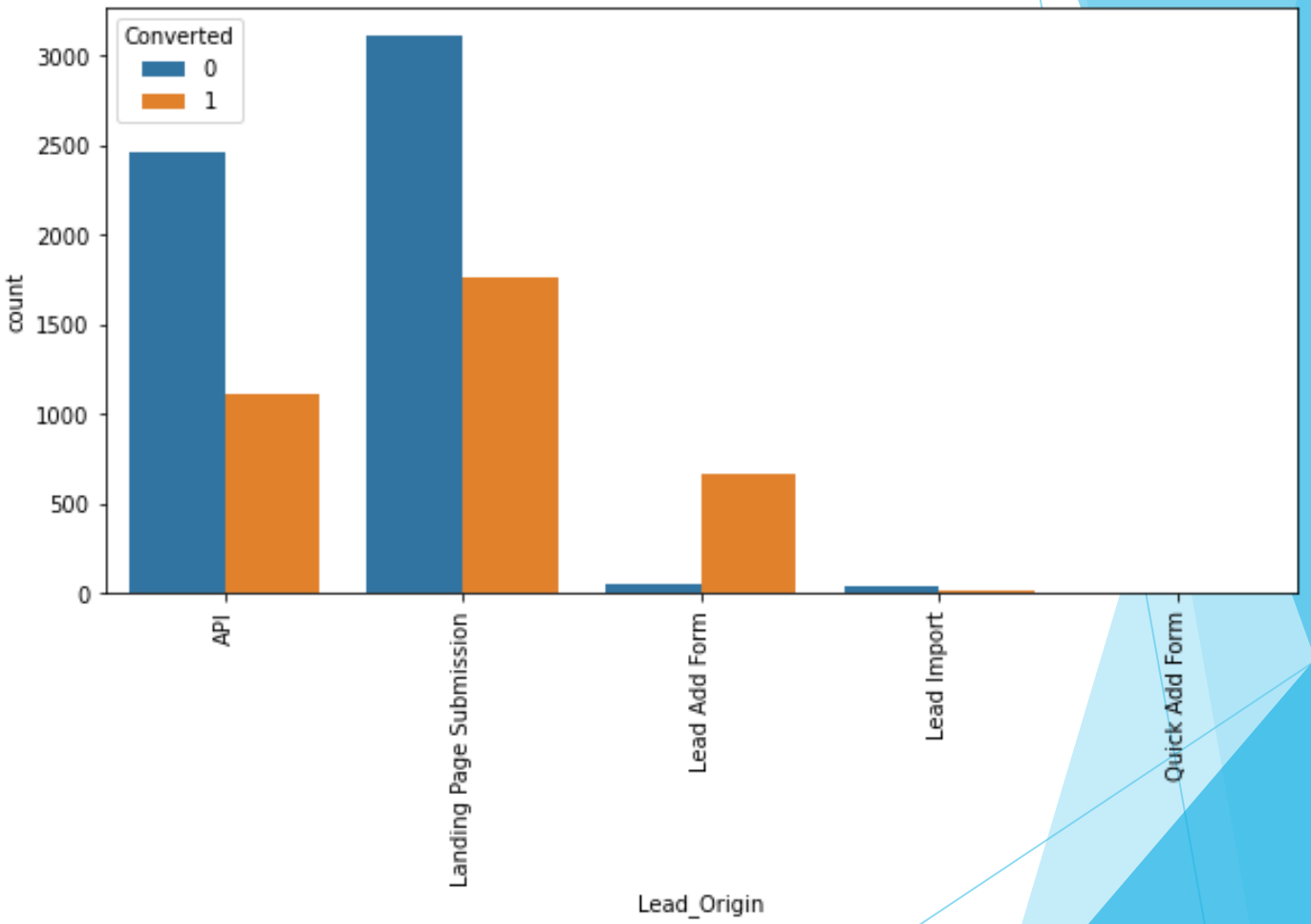
1. Working Professionals going for the course have high chances of joining it.
2. Unemployed leads are the most in numbers and have high conversion rate.



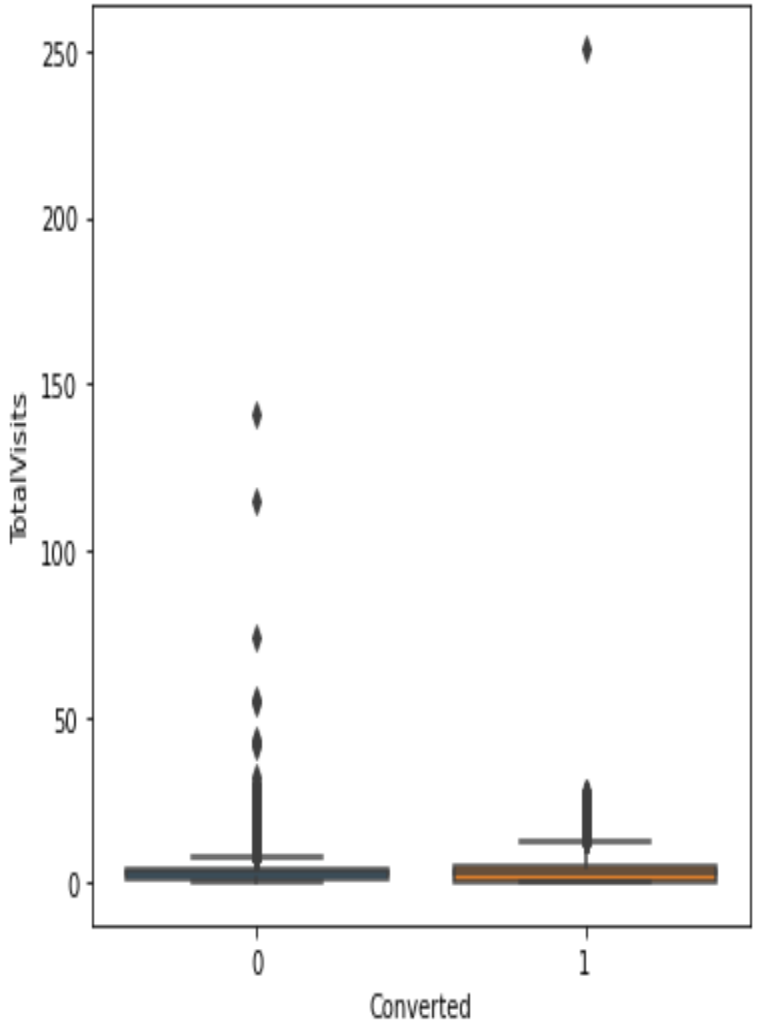
Lead Origin

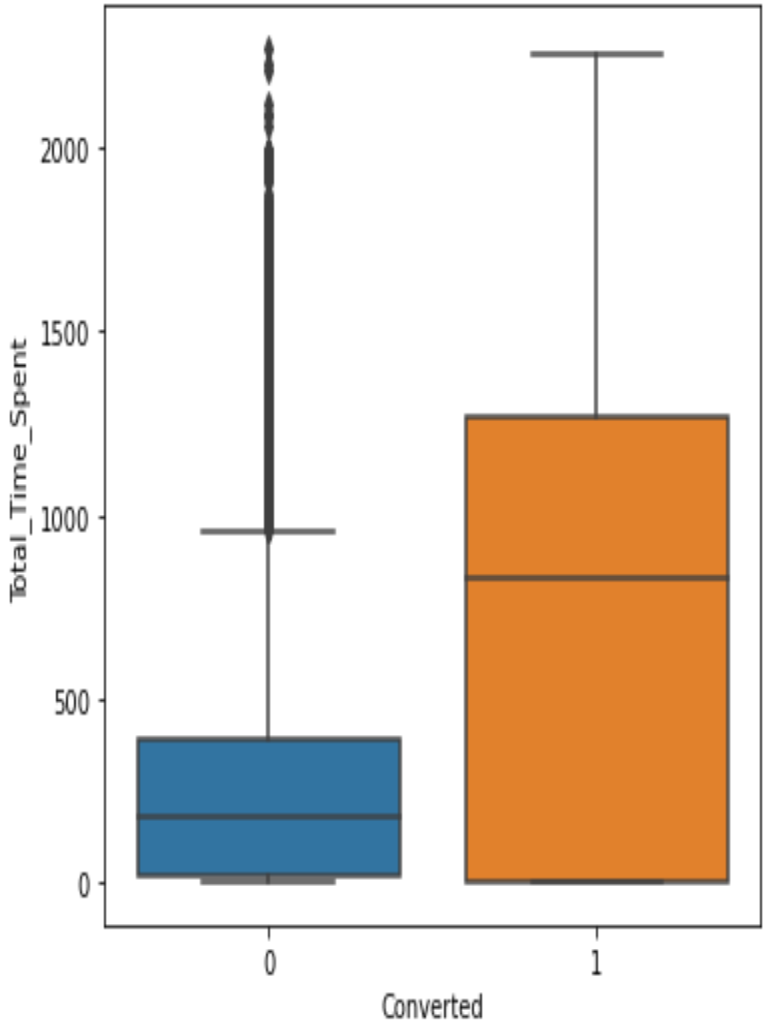
Inferences:

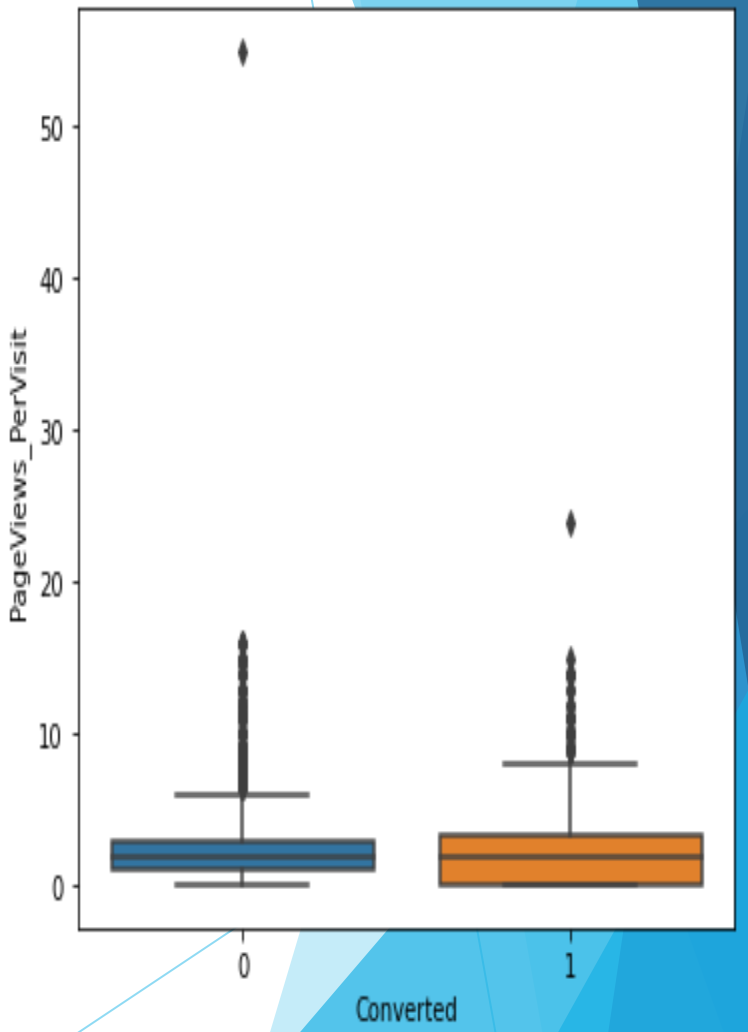
- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



Checking for Outliers



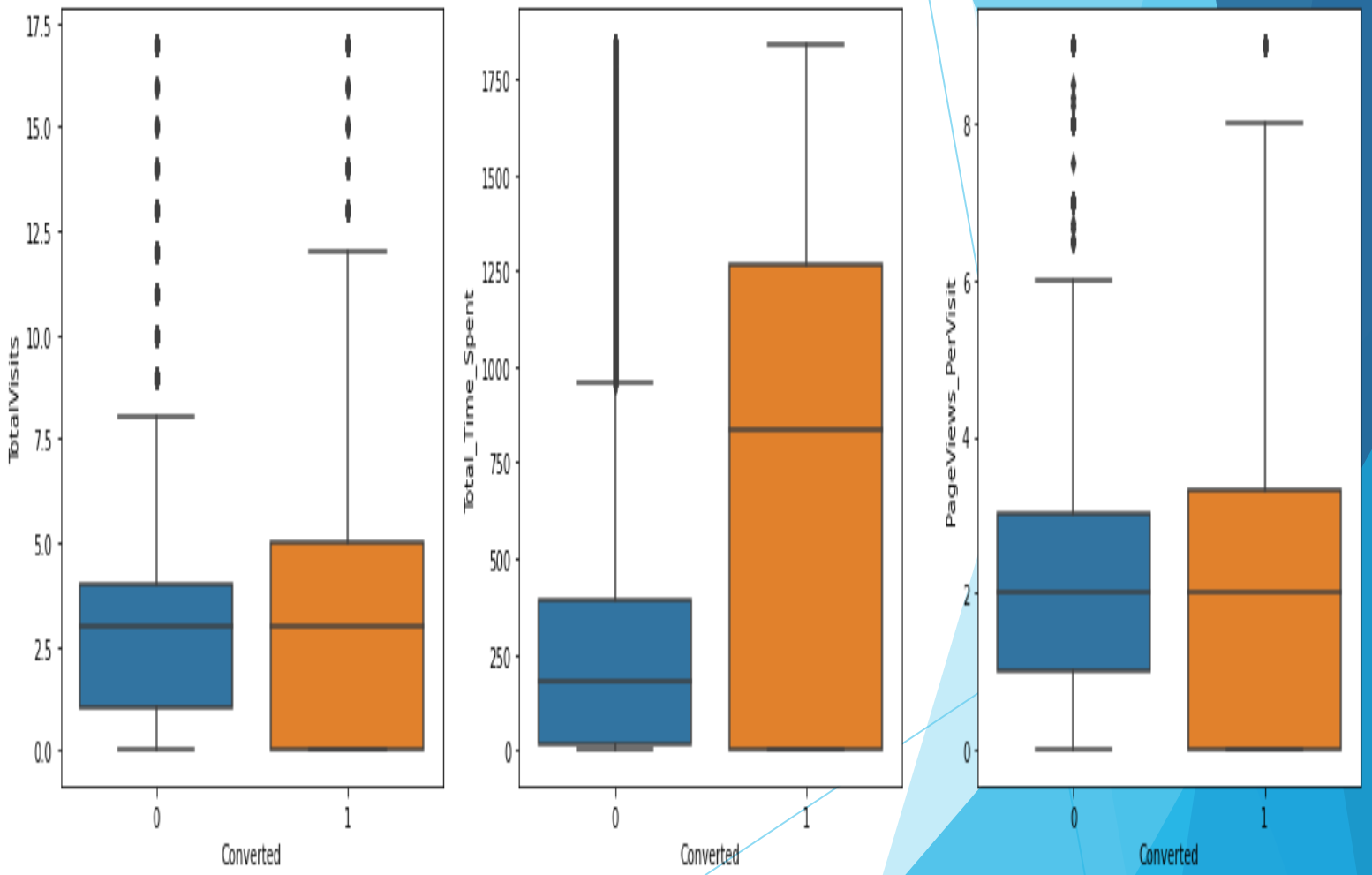




Outlier Treatment

Inferences: After treating outliers, we can see that,

1. 'Total Visits' - Median for converted and not converted leads are the same. Nothing can be concluded on the basis of Total Visits.
2. 'Total Time Spent' - Leads spending more time on the website are more likely to be converted. So, the Website should be made more engaging to make leads spend more time.
3. 'Page Views Per Visit' - Median for converted and unconverted leads is the same. Nothing can be said specifically for lead conversion from Page Views Per Visit.



Data Conversion

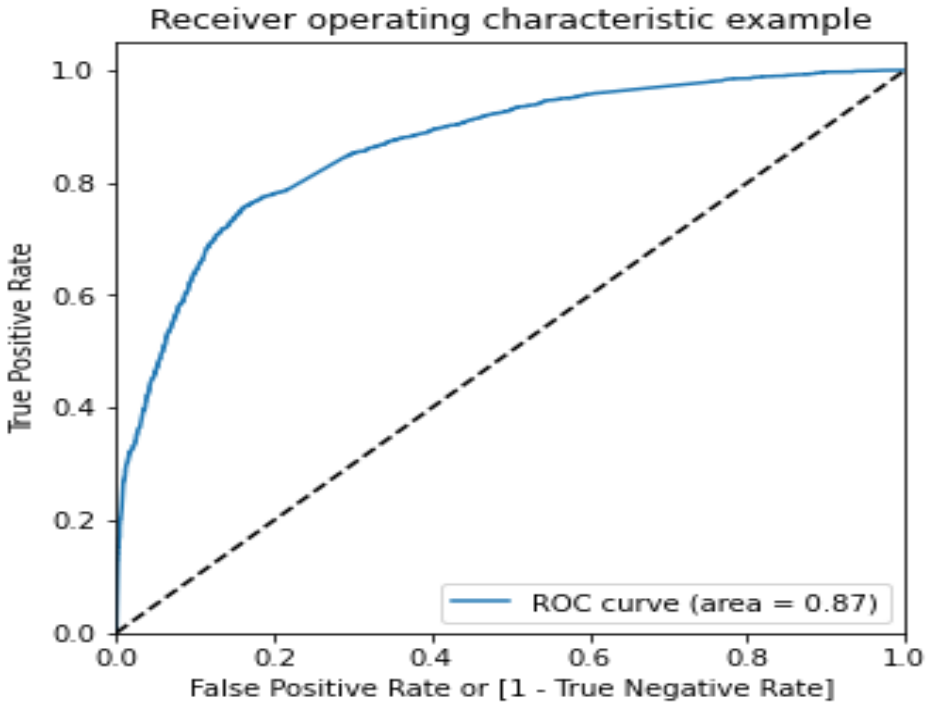
- Numerical Columns are Normalised
- Dummy variables are created for categorical type variables.
- After Dummy variables creation we got,

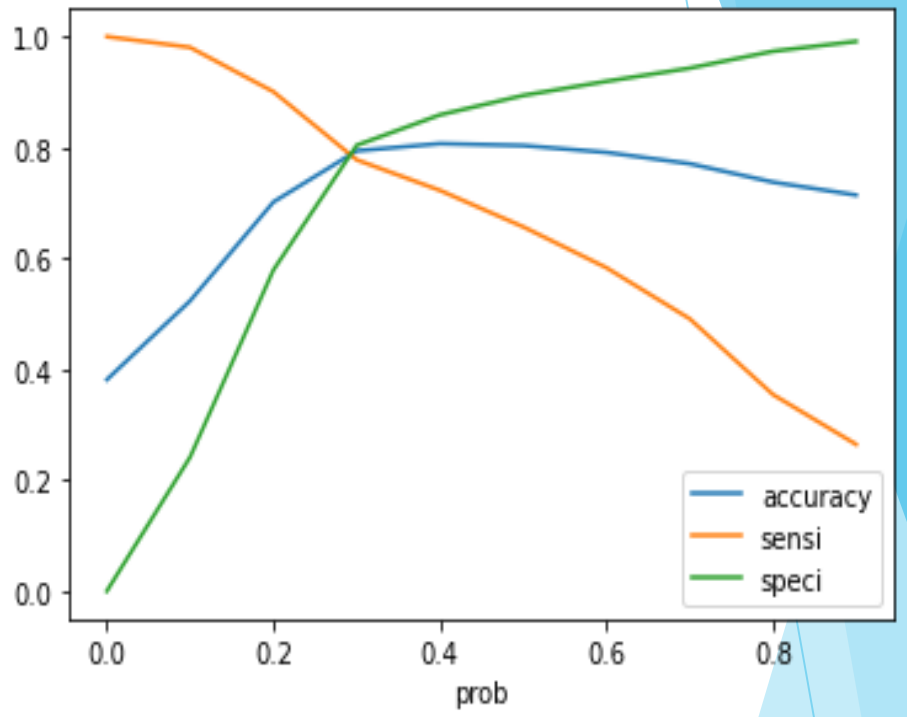
1. Total no. of rows = 9240
2. Total no. of columns = 48

Model Building

- At first, Splitting of data into Training and Test sets is done.
- Then, the very basic step for regression is performing a train-test split by choosing a ratio, we have chosen 70:30 ratio.
- Used RFE with 20 variables as output.
- Building Model by dropping the variables whose p-value is greater than 0.05 and VIF greater than 5.
- Predictions on test dataset.
- We got overall Accuracy around 80%

ROC Curve





- Finding optimal cutoff point
- Optimal cutoff probability is that probability where we get balanced sensitivity and specificity.
- From the second curve, we can see that 0.3 is the optimum point to take it as a cutoff probability.

Conclusion/Recommendation

- The important features responsible for good conversion rate or the one's which contribute more towards the probability of a lead getting converted are:

1. Total Time Spent: Recommendation is to add more features/knowledge on site that visitors spent more time having correct/engaging information.
2. Lead_Origin_Lead Add Form: The form filling should be reachable with one click on every page so that more visitors fill the form. Also, the current form could be shortened so as to have more complete fill rate.
3. LeadSource_Welingak Website: The commercials should be looked at with Welingak so that we get more leads from this site.

- As Overall Accuracy of model is 80 percent, Sensitivity is 66 percent and Specificity is 89 percent, model seems to be good fit predict the lead probability to conversion. The Model, based on accuracy, achieves the business goals as it meets the target of 80% conversion rate.
- As we get more data, we can train the model again and could get improved model and corresponding prediction of leads conversion.
- Also, we can add more features by collecting the data point for each lead so as to improve the accuracy.