Lead Score Case Study

Submitted By:

Thilokesh

Abhishek Tandon

Problem Statement

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, 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'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

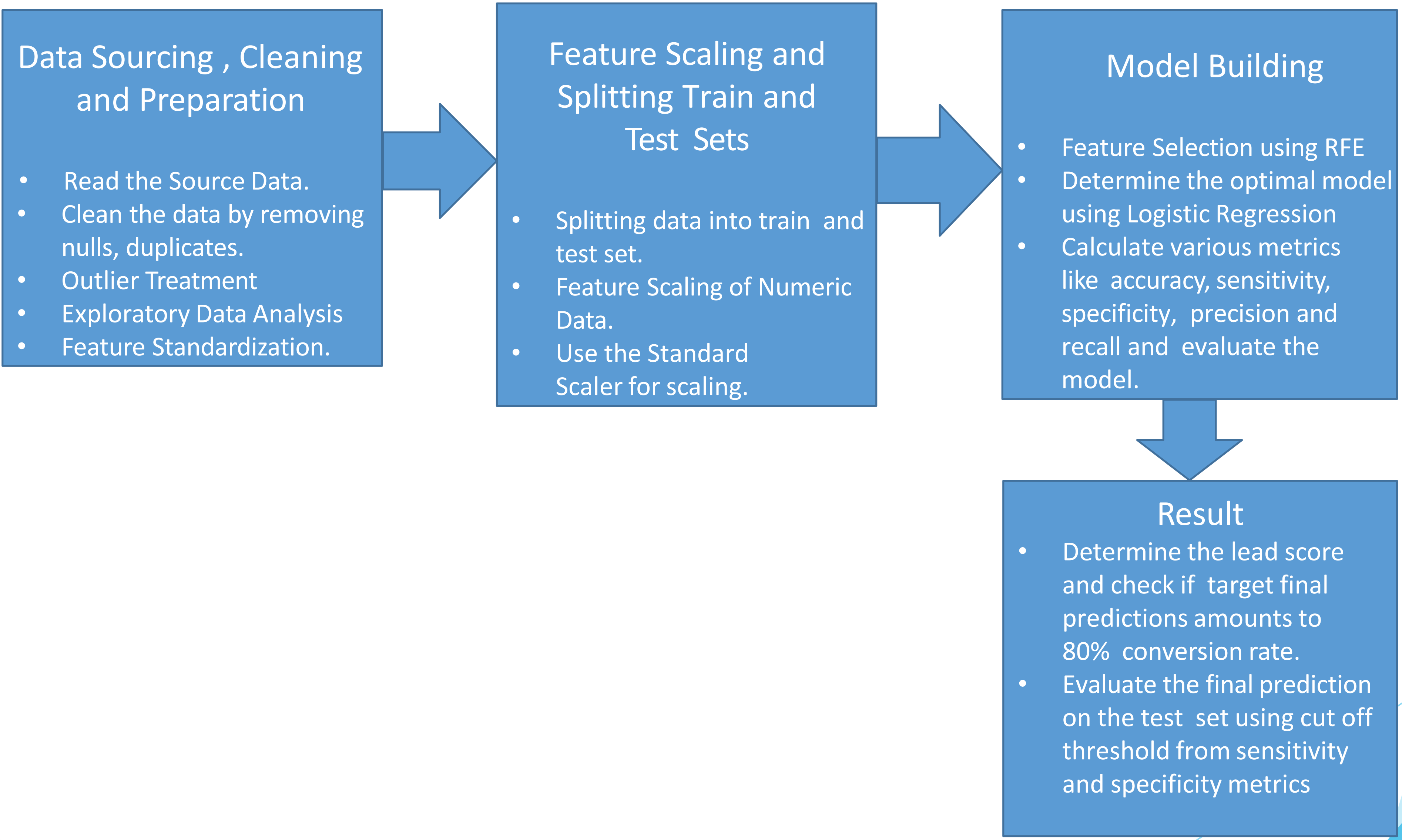
Business Goals

- X education wants to know the most promising leads.
- 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.
- Deployment of the model for the future use.

Solution Strategy

- Data cleaning and data manipulation.
  - Check and handle NA values and missing values.
  - Drop columns, if it has more than 40% nulls as it will not be useful for analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.
- EDA (Exploratory Data Analysis)
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: Logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

Solution Methodology



Data Sourcing , Cleaning and Preparation

- Read the Source Data.
- Clean the data by removing nulls, duplicates.
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

Feature Scaling and Splitting Train and Test Sets

- Splitting data into train and test set.
- Feature Scaling of Numeric Data.
- Use the Standard Scaler for scaling.

Model Building

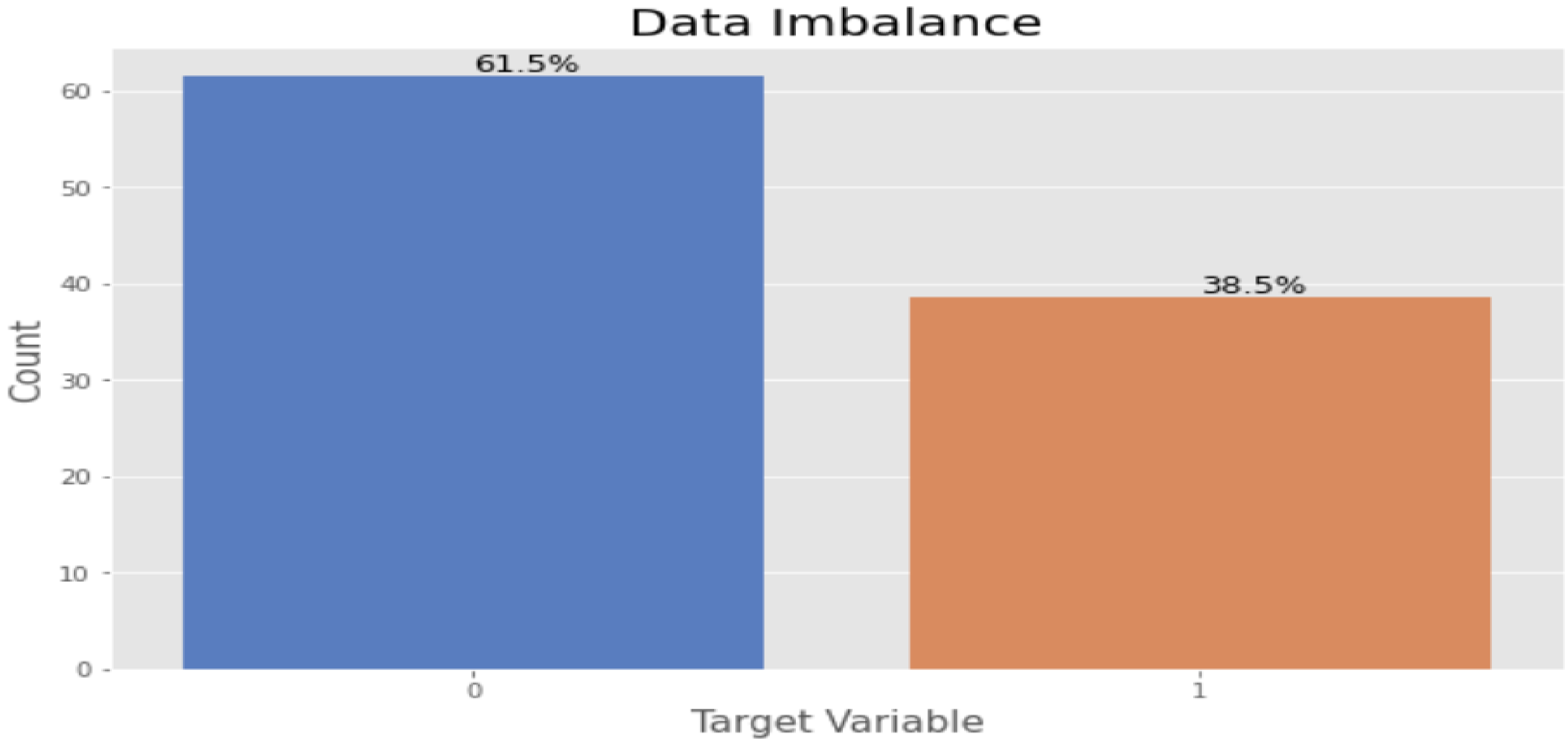
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

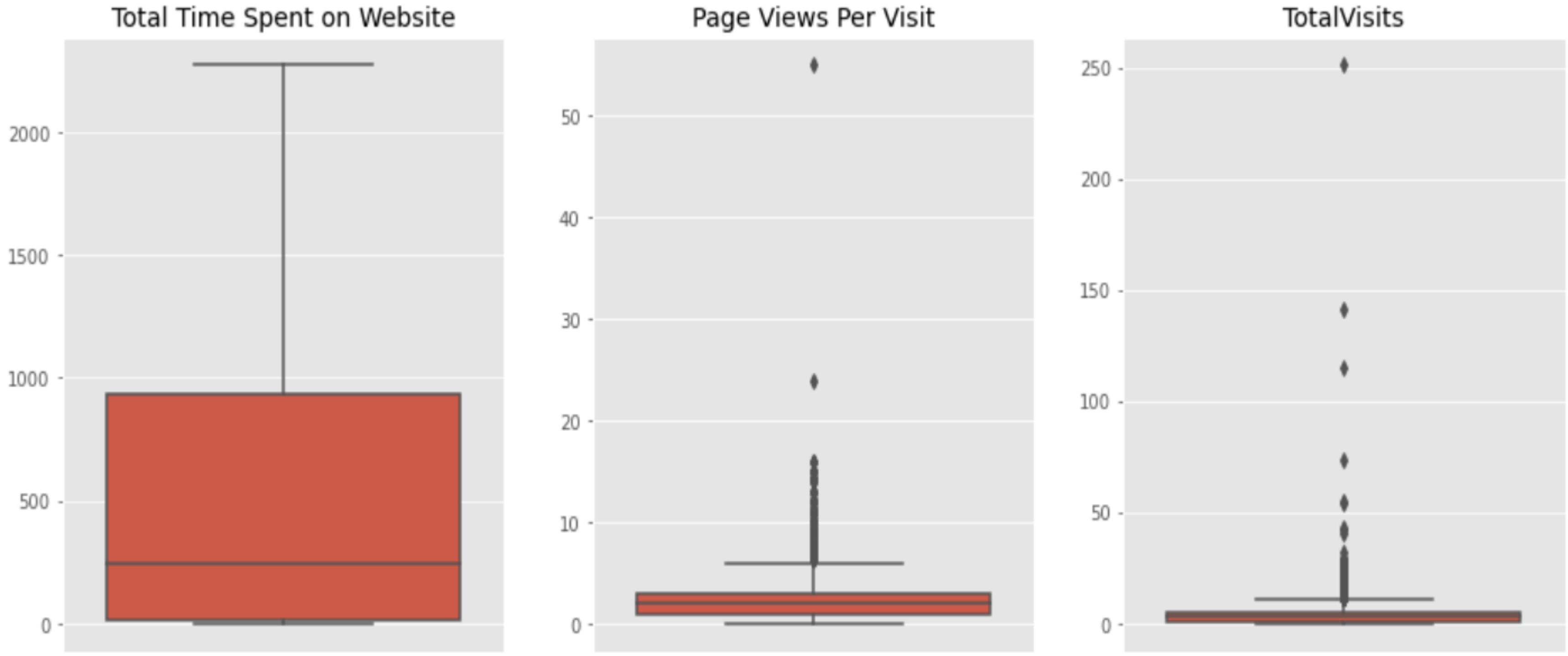
Exploratory Data Analysis

Current Lead Conversion Rate



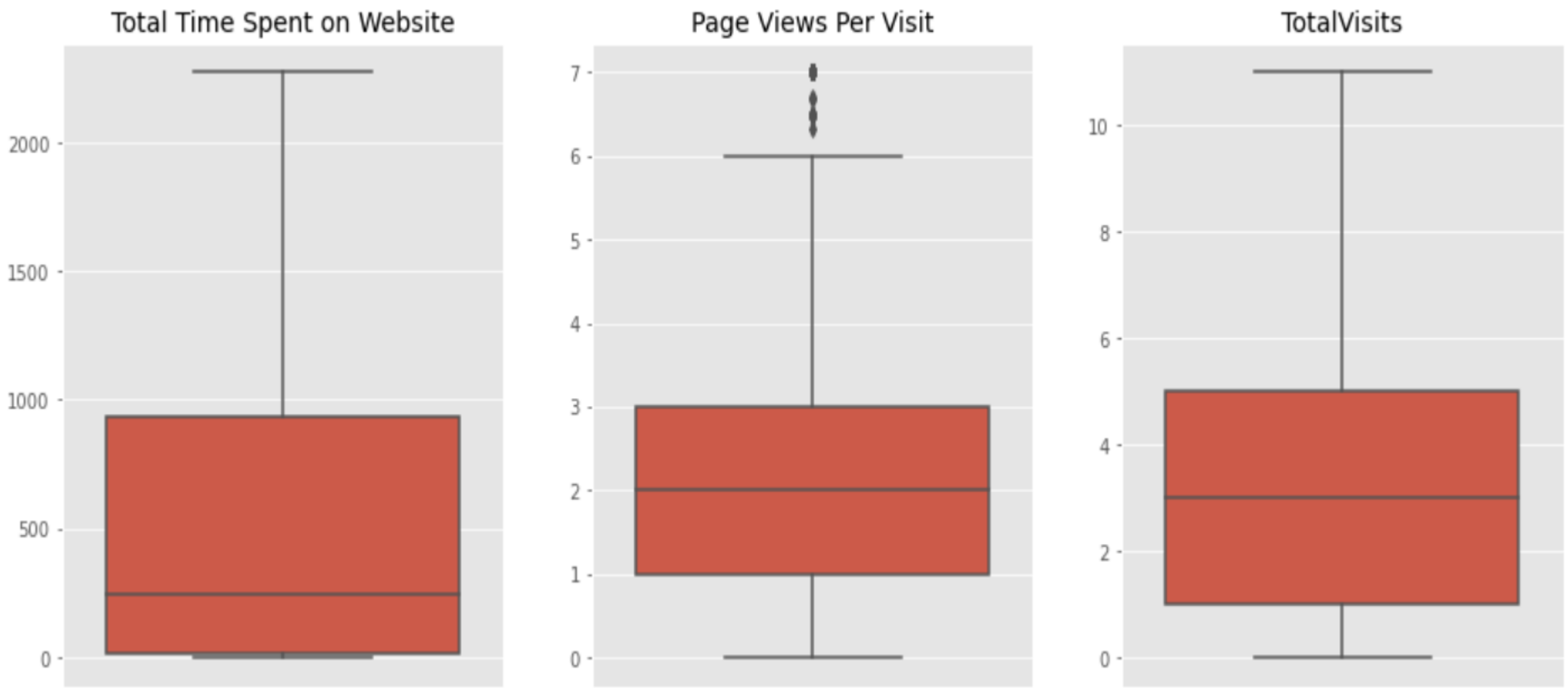
- We have around 39% Conversion rate in Total in the given Data set.
- We can see that the current conversion rate is quite less than the expected conversion rate.

Box Plot of Numerical Variables



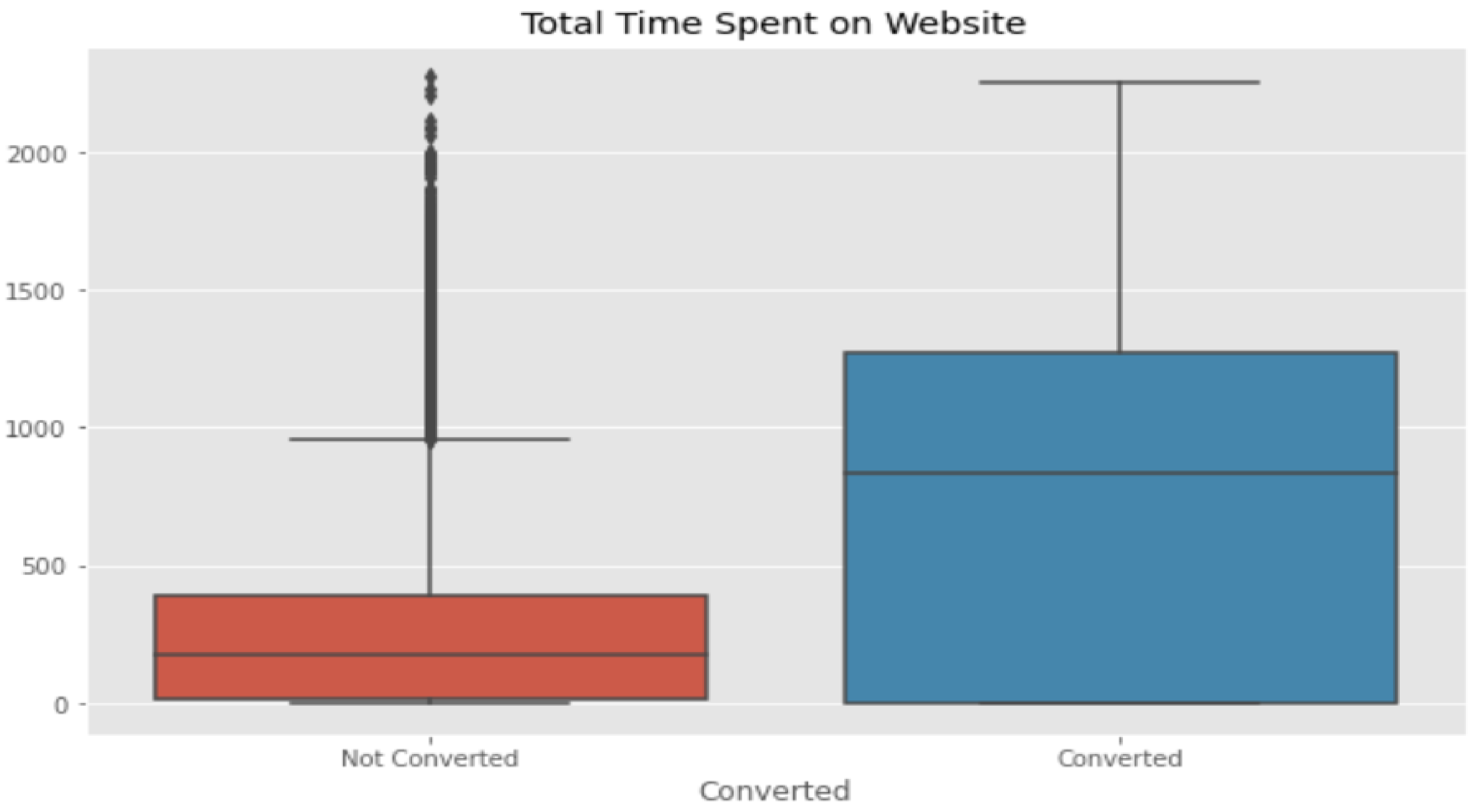
- We can clearly see that there are outliers in the Page Views Per Visit and TotalVisits.

Box Plot of Numeric Variables After Outlier Handling



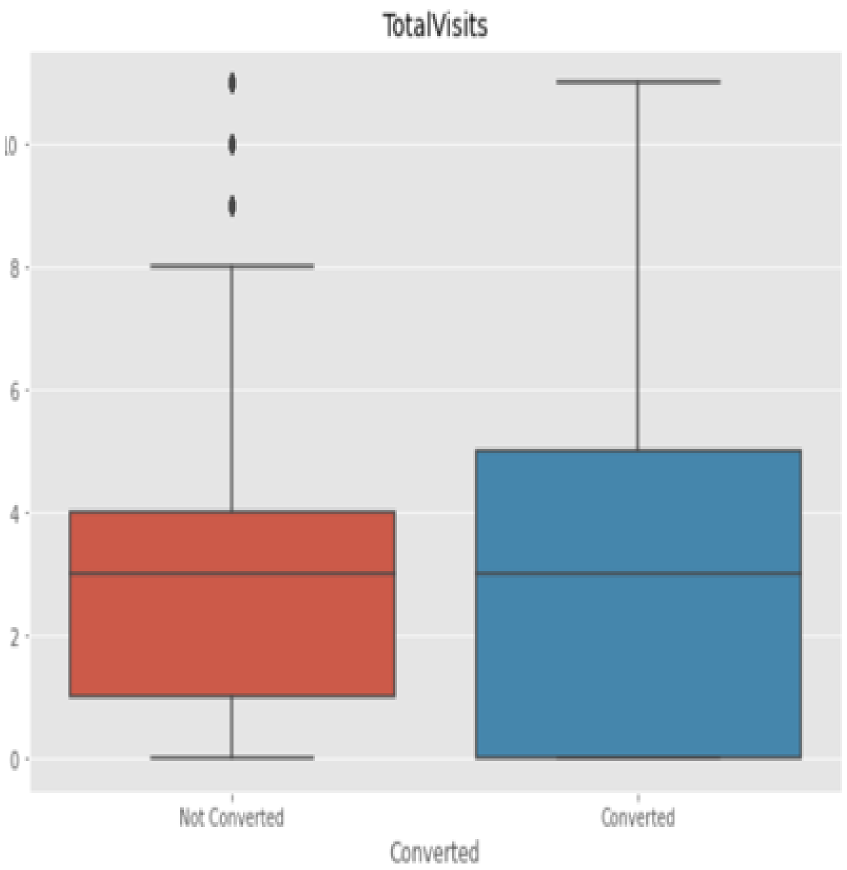
- Even after capping the variables there still seems to be some outliers in the Page Views Per Visit column but these outliers seem to be in an acceptable range.

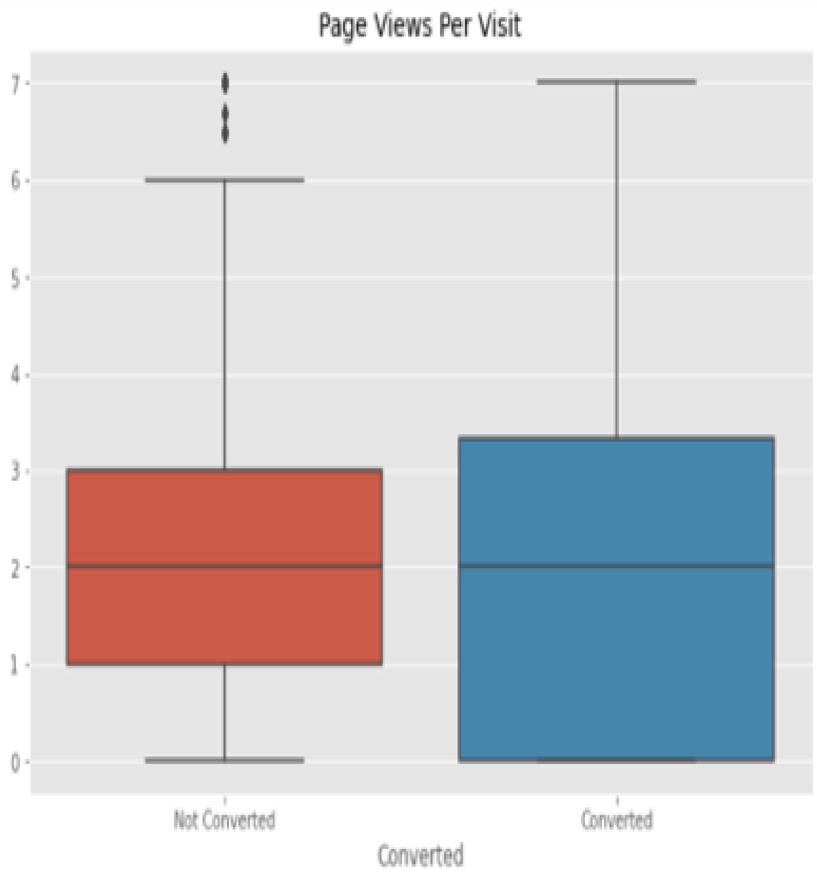
Lead Conversion Plot Total Time Spent on Website



- We can clearly see converted leads have spent more time on the website

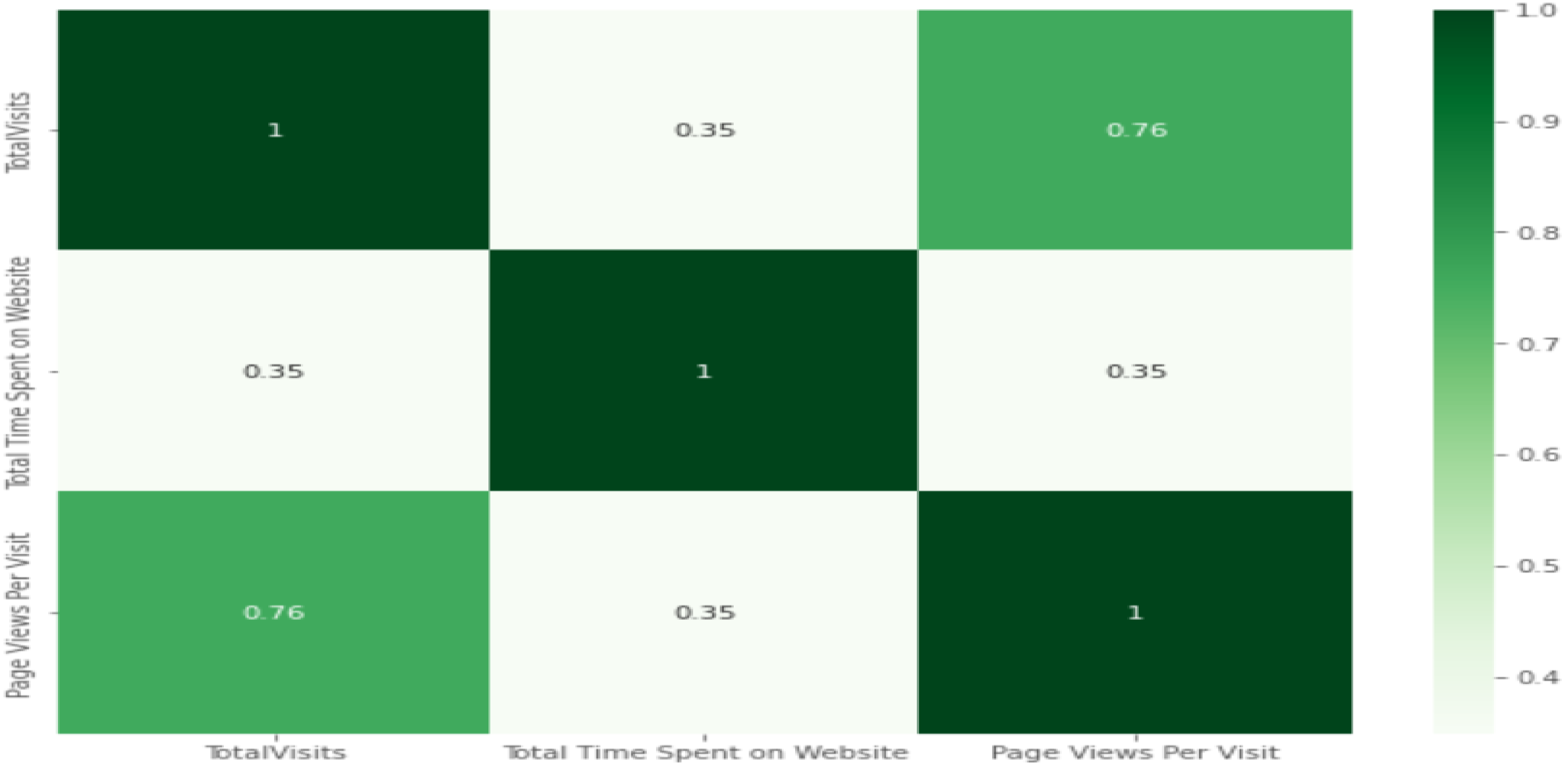
Lead Conversion Plot : Total Visits & Page Views Per Visit





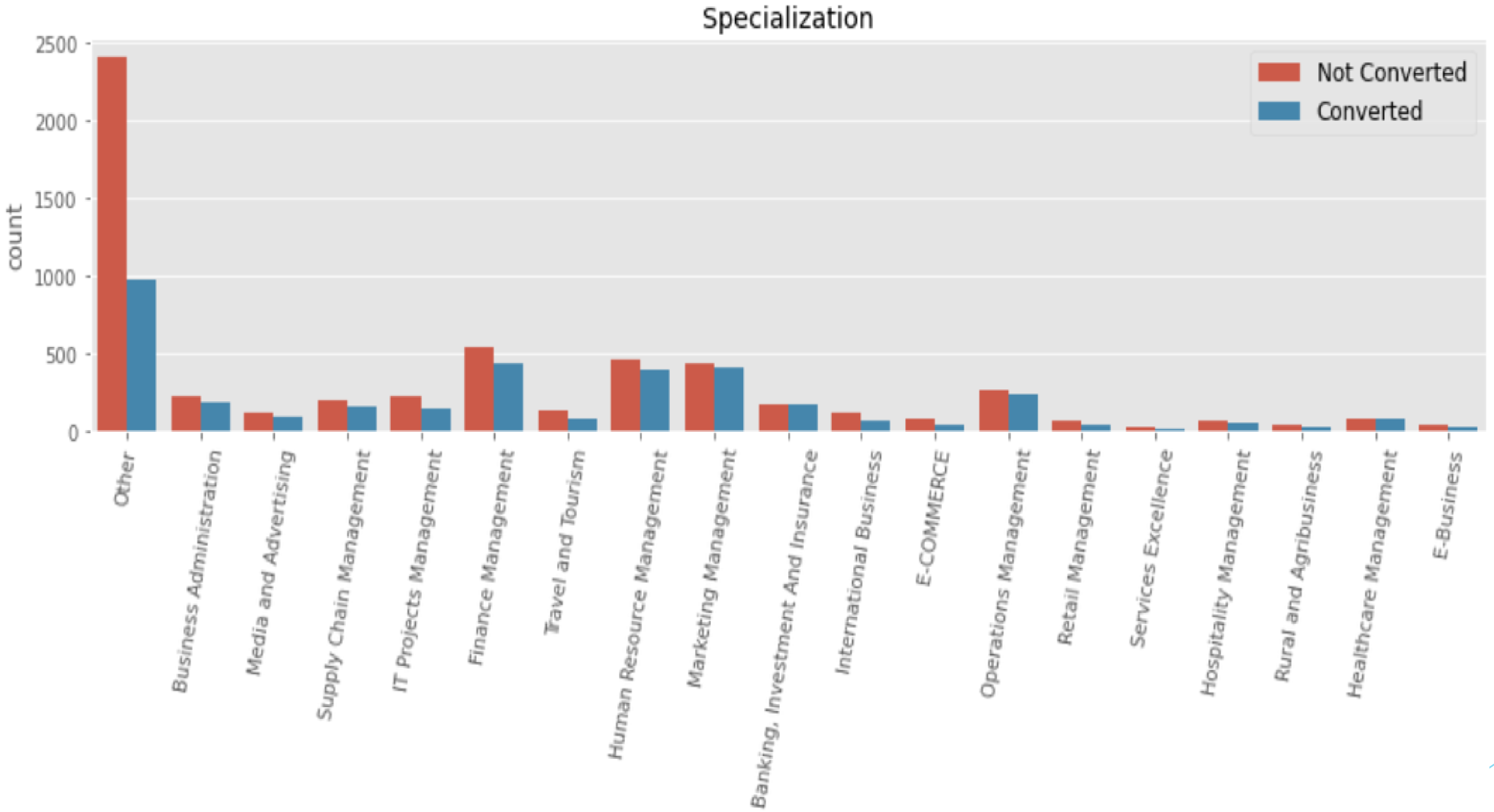
- On average(Median) we can clearly see that there is not much difference in the total visits in both the categories for TotalVisits and Page Views Per Visit.
- We can also observe some outliers in the not converted category when we take total visits in to consideration

Correlation Heatmap of Numeric Variables



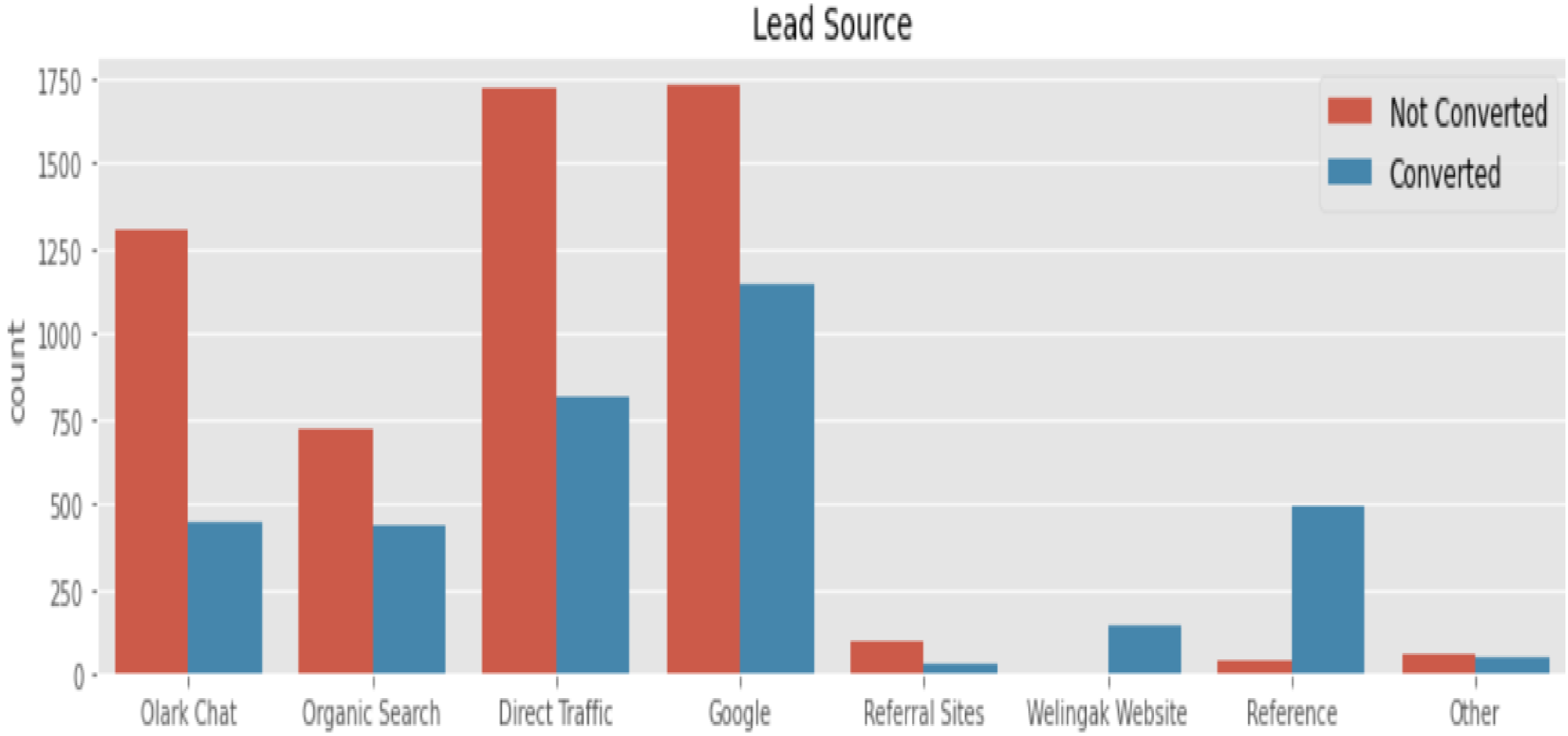
- We can observe there is high correlation between Page Views Per Visit and Total Visits

Conversion Plot of Specialization



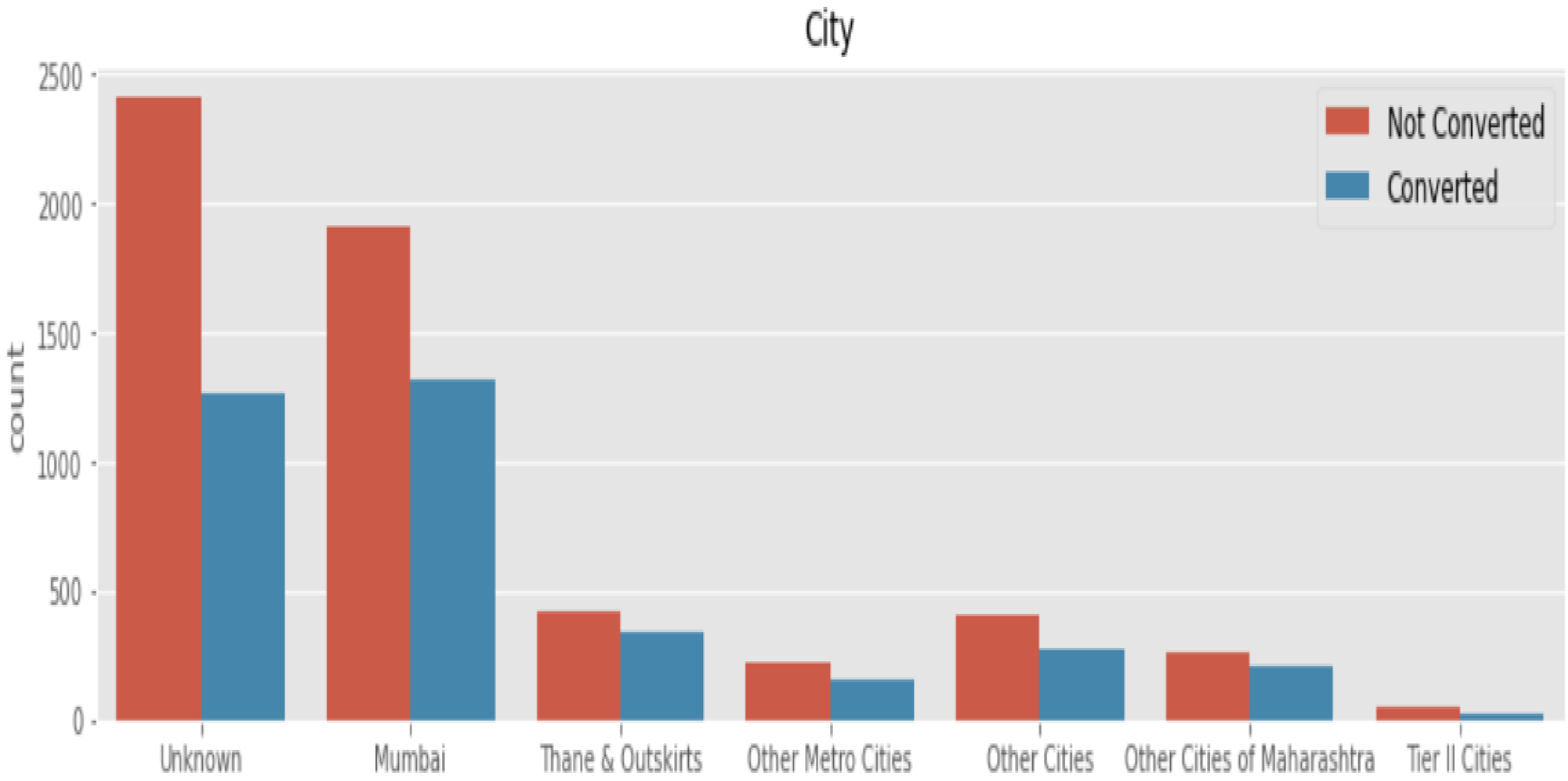
- We can see from the count plot of 'Specialization' that if user has a specialization in any field then the chance of conversion id high.

Conversion Plot of Lead Source



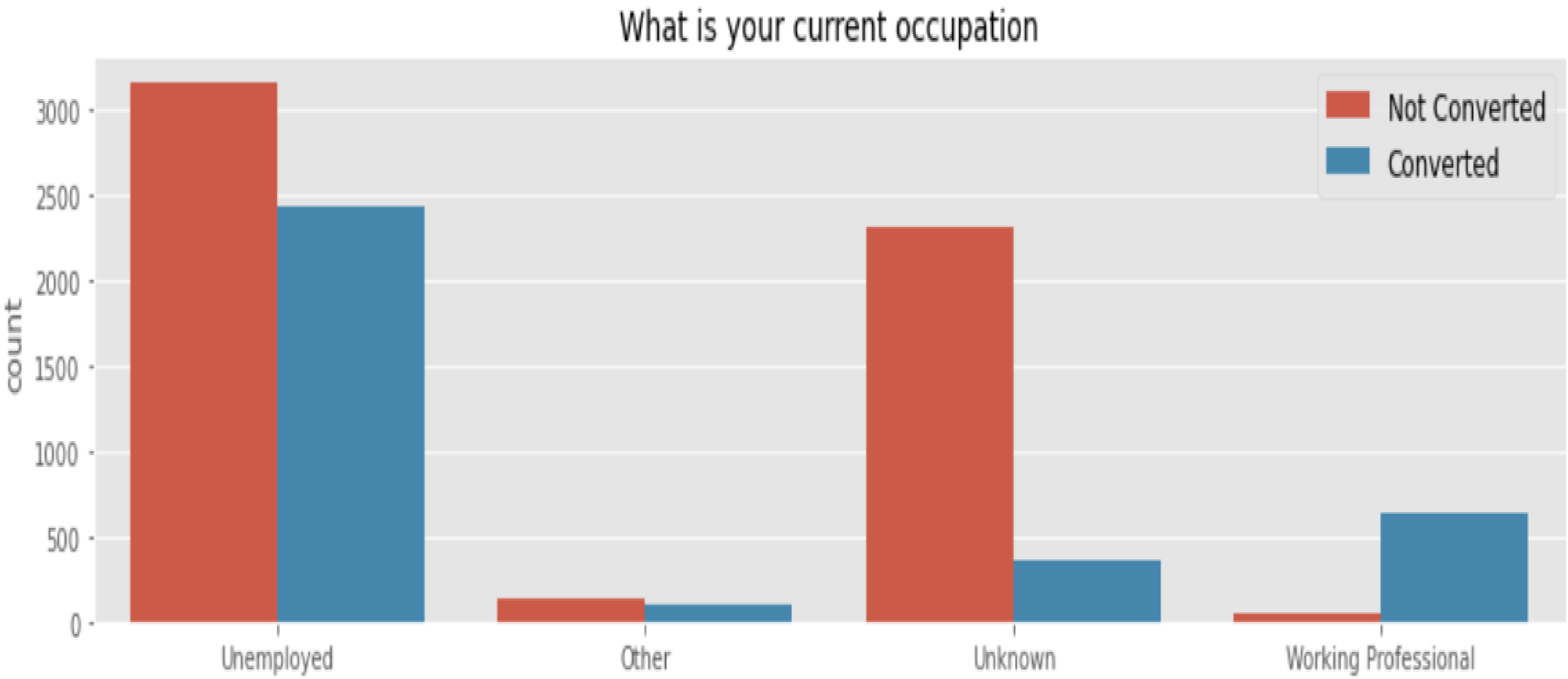
- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.

Conversion Plot of City



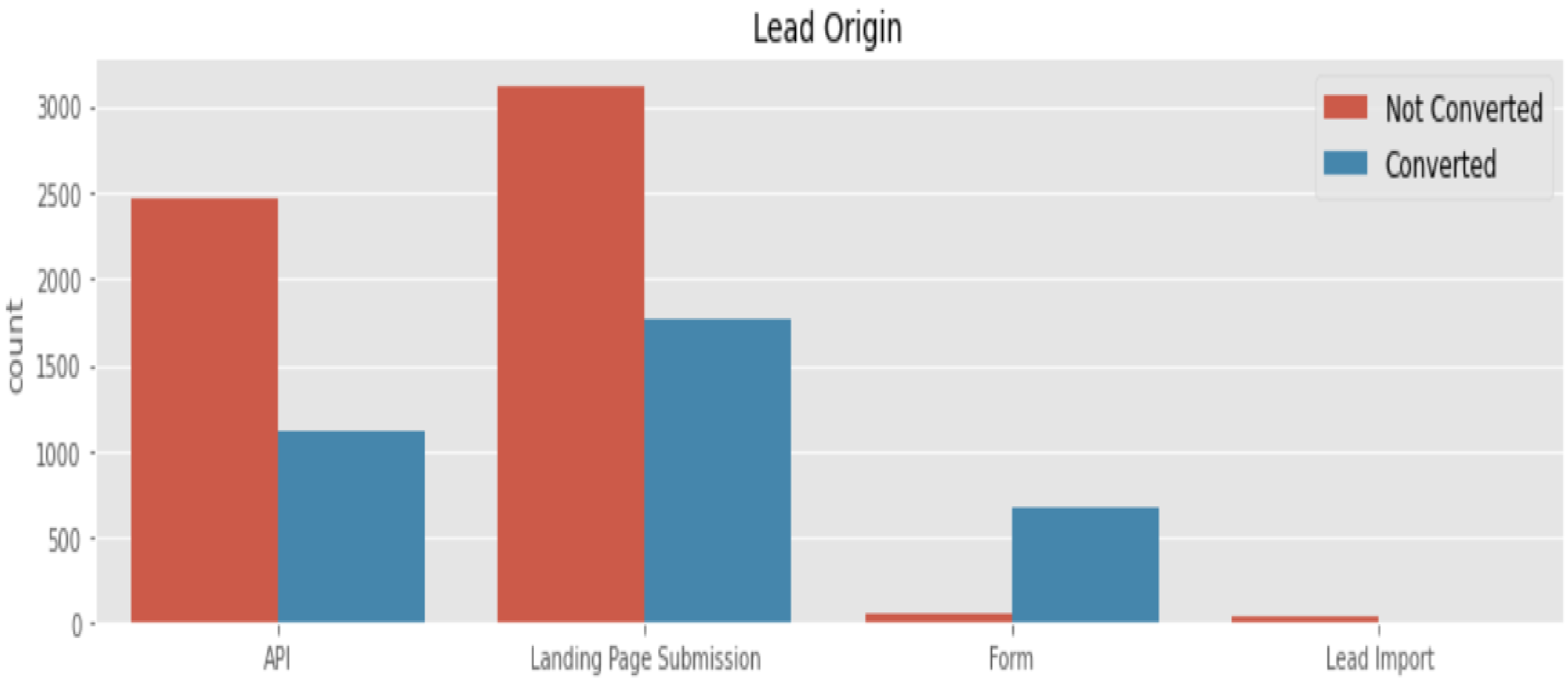
- Most leads are from Mumbai with around 50% conversion rate.
- From the count plot of the city column, we can see that there are many users not giving their city name and therefore if these missing values were imputed then would skew the data towards the city of Mumbai.

Conversion Plot of Occupation



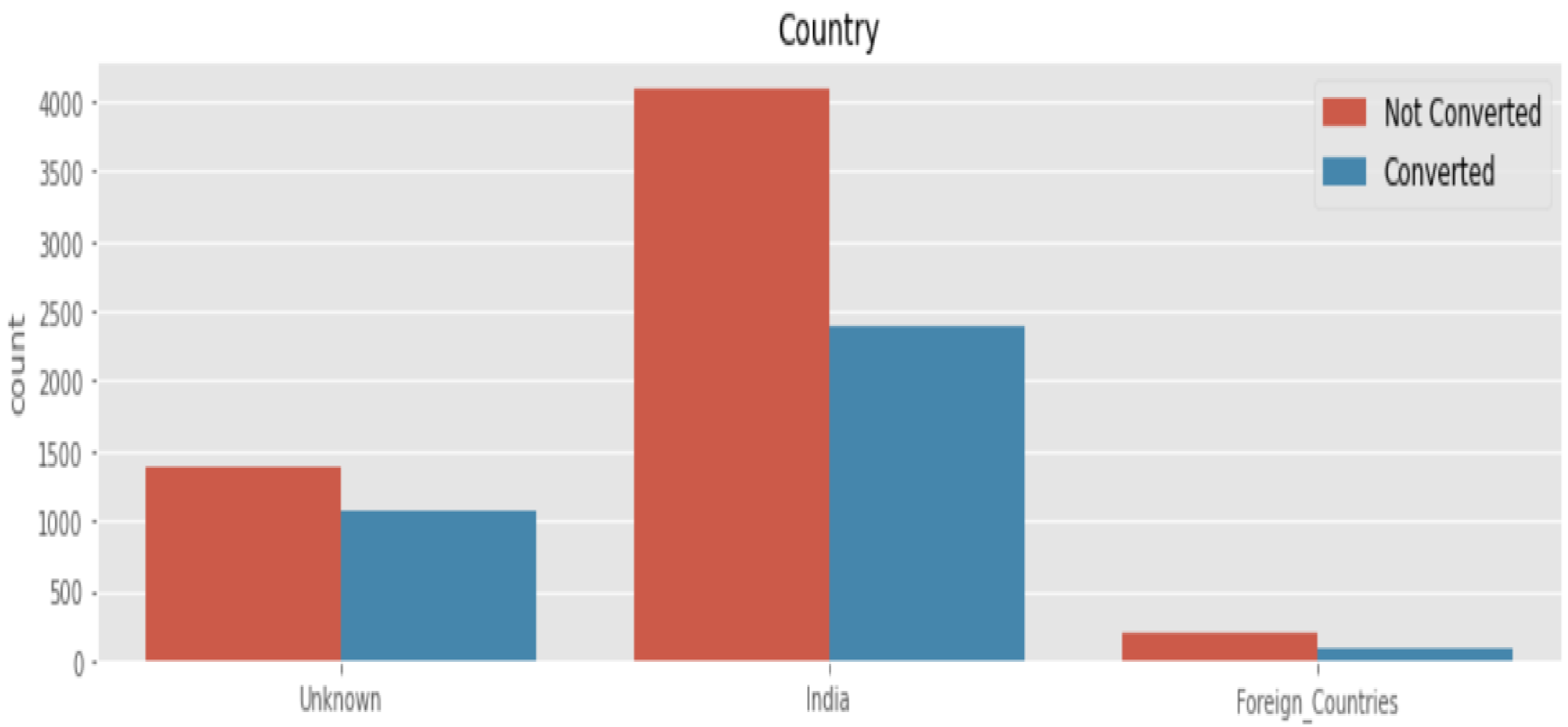
- Working Professionals going for the course have high chances of joining it.
- When the count plot of the user occupation is plotted we can observe that most of our users are unemployed, this also proves the fact that our users are looking for a better career prospect.

Conversion Plot of Lead Origin



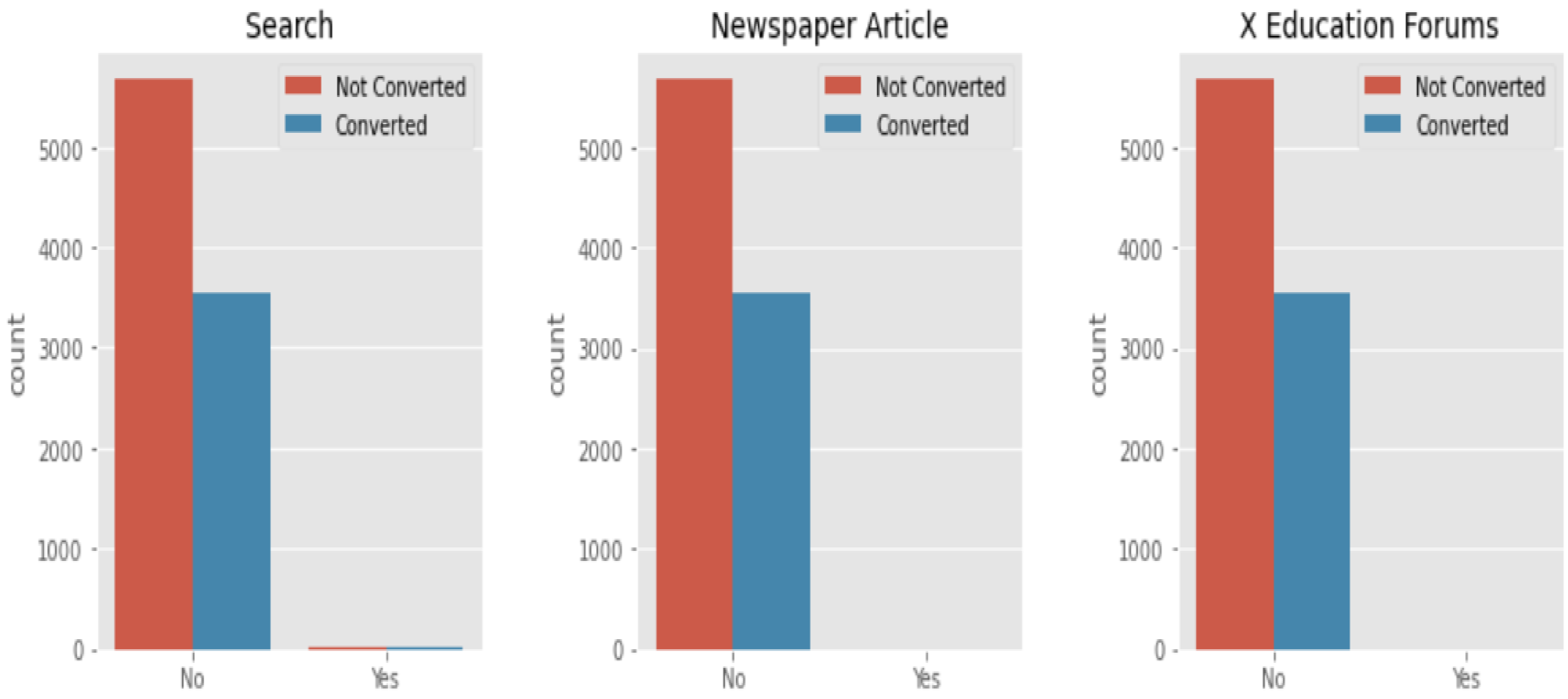
- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.

Conversion Plot of Country



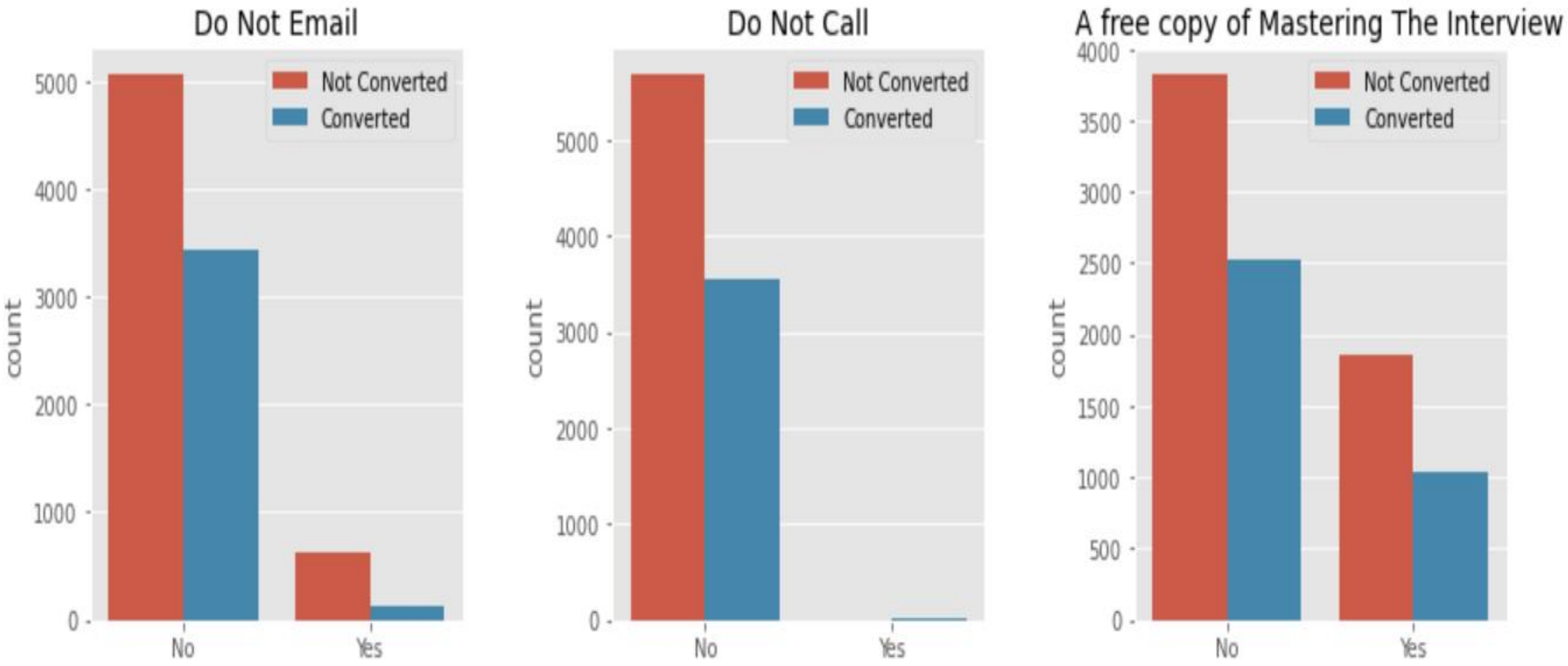
- We can observe that most of our user base is from India, we can also observe that there are many users that have not provided any information on the country.
- Since, the values in it are skewed we would drop this column

Conversion Plot for Binary Variables



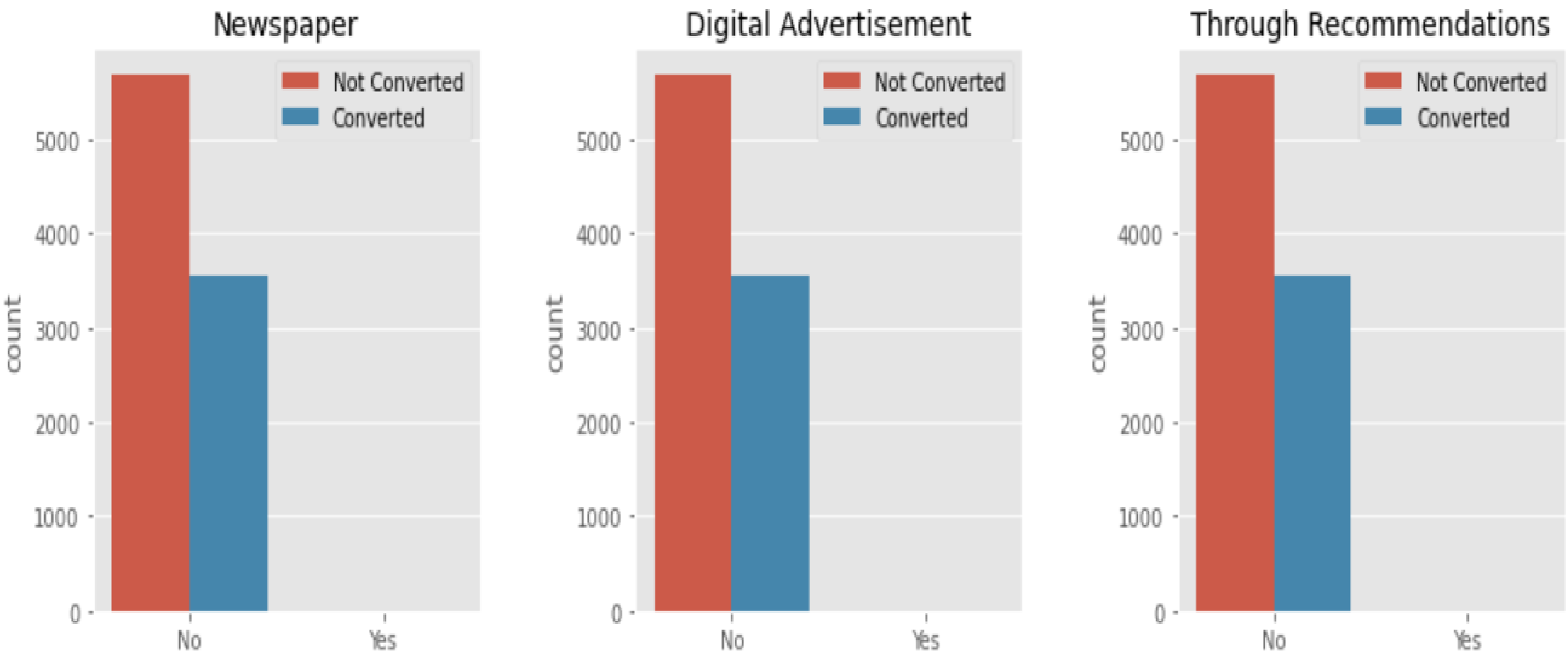
- Search : Most entries are 'No'. No Inference can be drawn with this parameter.
- Newspaper Article: Most entries are 'No'. No Inference can be drawn with this parameter.
- X Education Forums: Most entries are 'No'. No Inference can be drawn with this parameter.

Conversion Plot for Binary Variables



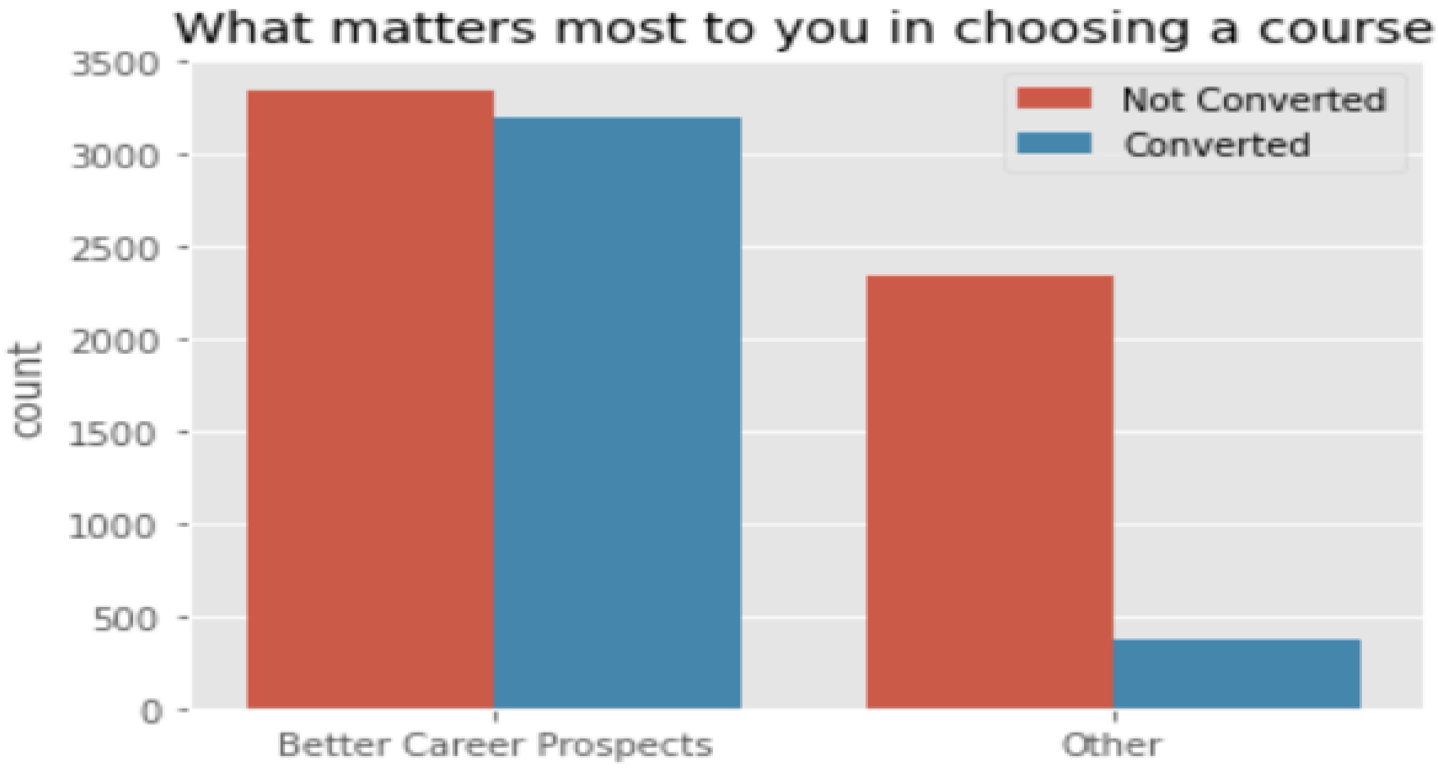
- Do Not Email : Most entries are 'No'. No Inference can be drawn with this parameter.
- Do Not Call : Most entries are 'No'. No Inference can be drawn with this parameter.
- A free copy of Mastering The Interview: Most entries are 'No'. No Inference can be drawn with this parameter

Conversion Plot for Binary Variables



- Newspaper : Most entries are 'No'. No Inference can be drawn with this parameter.
- Digital Advertisement : Most entries are 'No'. No Inference can be drawn with this parameter.
- Through Recommendations : Most entries are 'No'. No Inference can be drawn with this parameter.

Conversion Plot of Reasons for Choosing a course

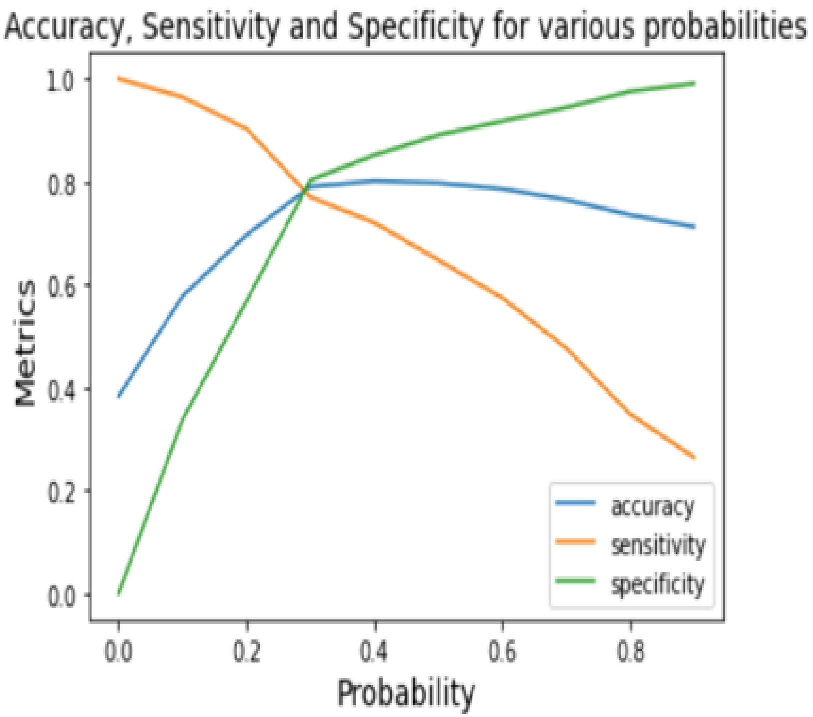


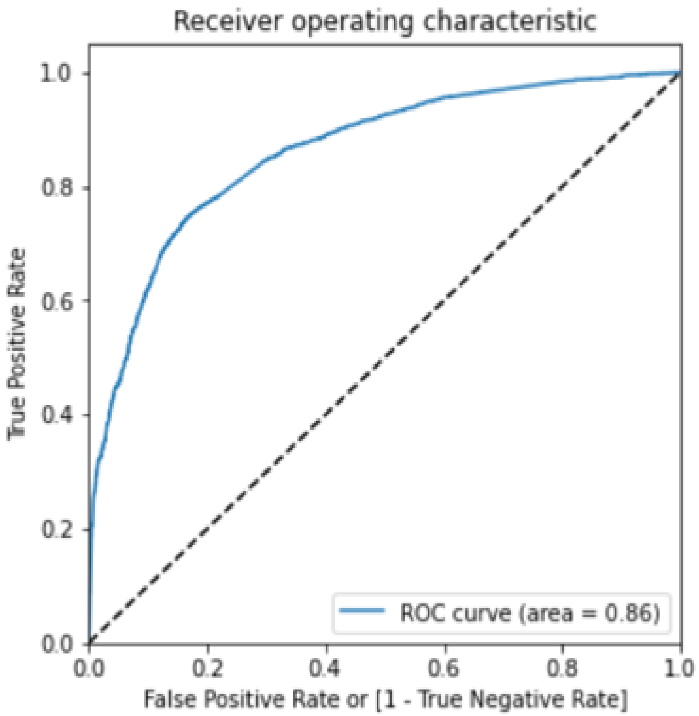
- We can see that this is highly skewed column so we can remove this column.

Model Building

- 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.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output.
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
- Predictions on test data set.
- Calculate the Accuracy , Sensitivity and Specificity of the Train and Test Data Set.
- Model Evaluation.

Model Evaluation (ROC Curve)

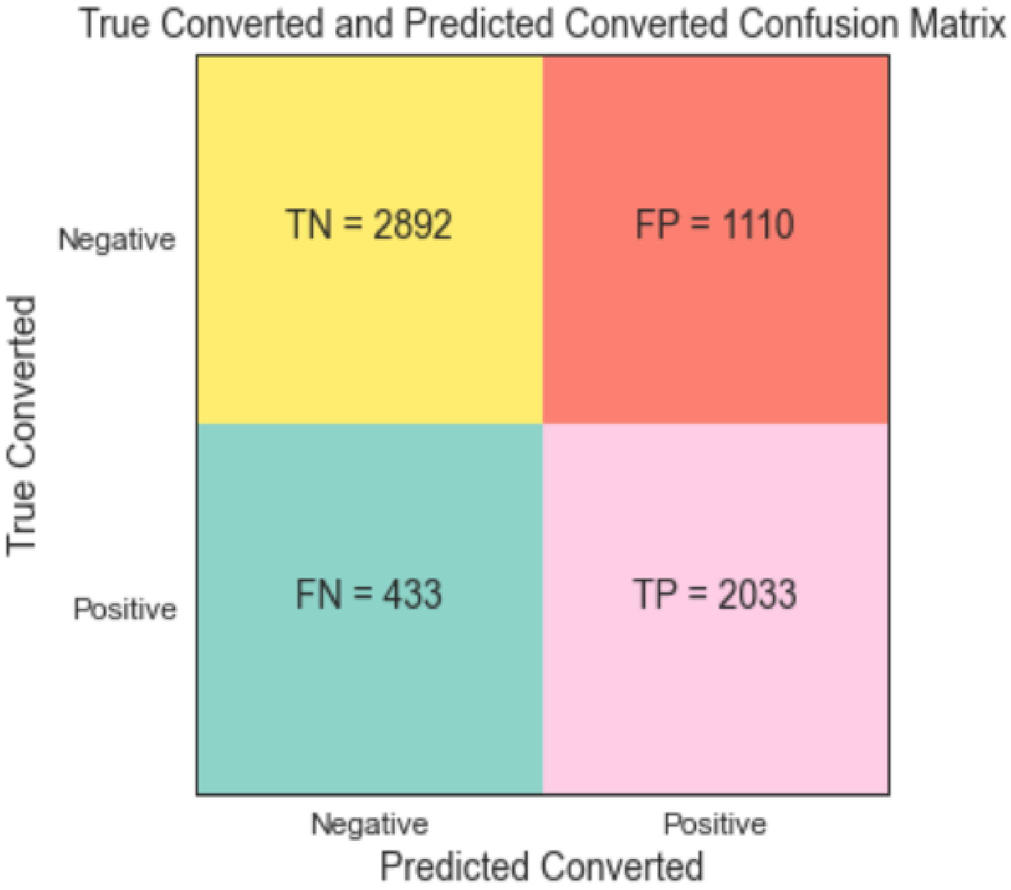




Finding Optimal Cut off Point.

- Optimal cut off probability is that Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.29.

Model Evaluation On Train Data

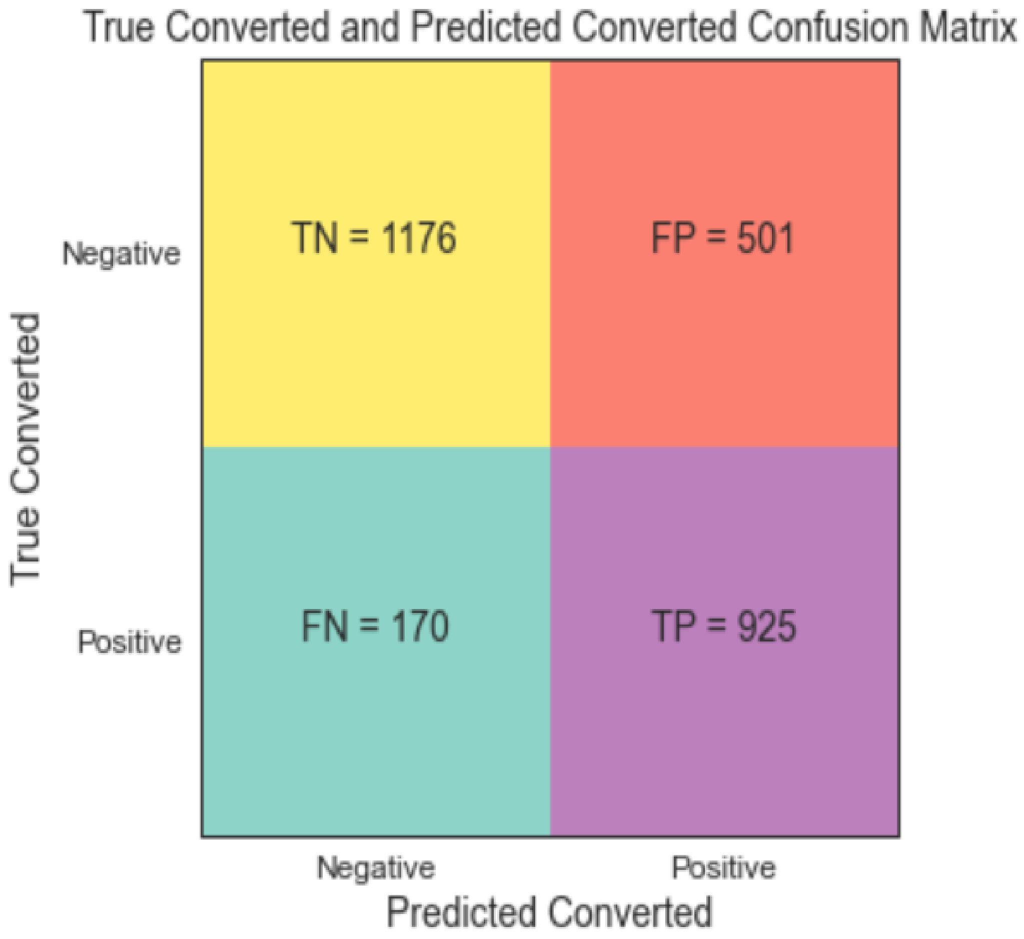


True Converted and Predicted Converted Confusion Matrix

True Converted \ Predicted Converted	Negative	Positive
Negative	TN = 2892	FP = 1110
Positive	FN = 433	TP = 2033

- Accuracy - 76%
- Sensitivity - 82 %
- Specificity - 72 %
- Recall – 82 %

Model Evaluation On Test Data



True Converted and Predicted Converted Confusion Matrix

True Converted \ Predicted Converted	Negative	Positive
Negative	TN = 1176	FP = 501
Positive	FN = 170	TP = 925

- Accuracy - 76%
- Sensitivity - 84 %
- Specificity - 71 %
- Recall – 84 %

Conclusion & Recommendation

- Increase user engagement on Welingak website since this helps in higher conversion.

- Focus on Working Professional which has high conversion certainty.

- Give incentives to improve the amount Referrals this would give a boost to the conversion rate.

- Make the user spend more time on the website which would also increase the chance of conversion, one recommended way to increase the time spent on the website is to give free content or free intro to the course.

- Improve the Olark Chat service, improving the response time, connecting to the right person etc..

- Make sure that the users visiting the website give information as to their specialization, current occupation, etc by making the user register or fill a form.

- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.