Lead Score Case Study

Group Members : Karan Gupta
Dhruva E Seshasayee
Vijayraj K Poojary

Problem Statement:

- X Education sells online courses to industry professionals, due to which they require help in selecting the most promising leads i.e. the leads that are most likely to convert into paying customers.
- Their current lead conversion rate is very poor. For example, if they acquire 100 leads in one day, only about 30 leads get converted
- To make this process more efficient, the company wants to identify the potential leads, also termed as 'Hot Leads'
- The company needs a model wherein a lead score is assigned to each of the leads, such that customers with a higher lead score have a higher conversion rate and those with a low lead score have a low conversion rate.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%

Our Solution Approach →

Data Cleaning & manipulation

- Check and handle the duplicate data
- Check and handle outliers
- Check and handle NA values and missing values
- Drop columns, if it has a large number of missing values
- Imputation of the values, wherever possible

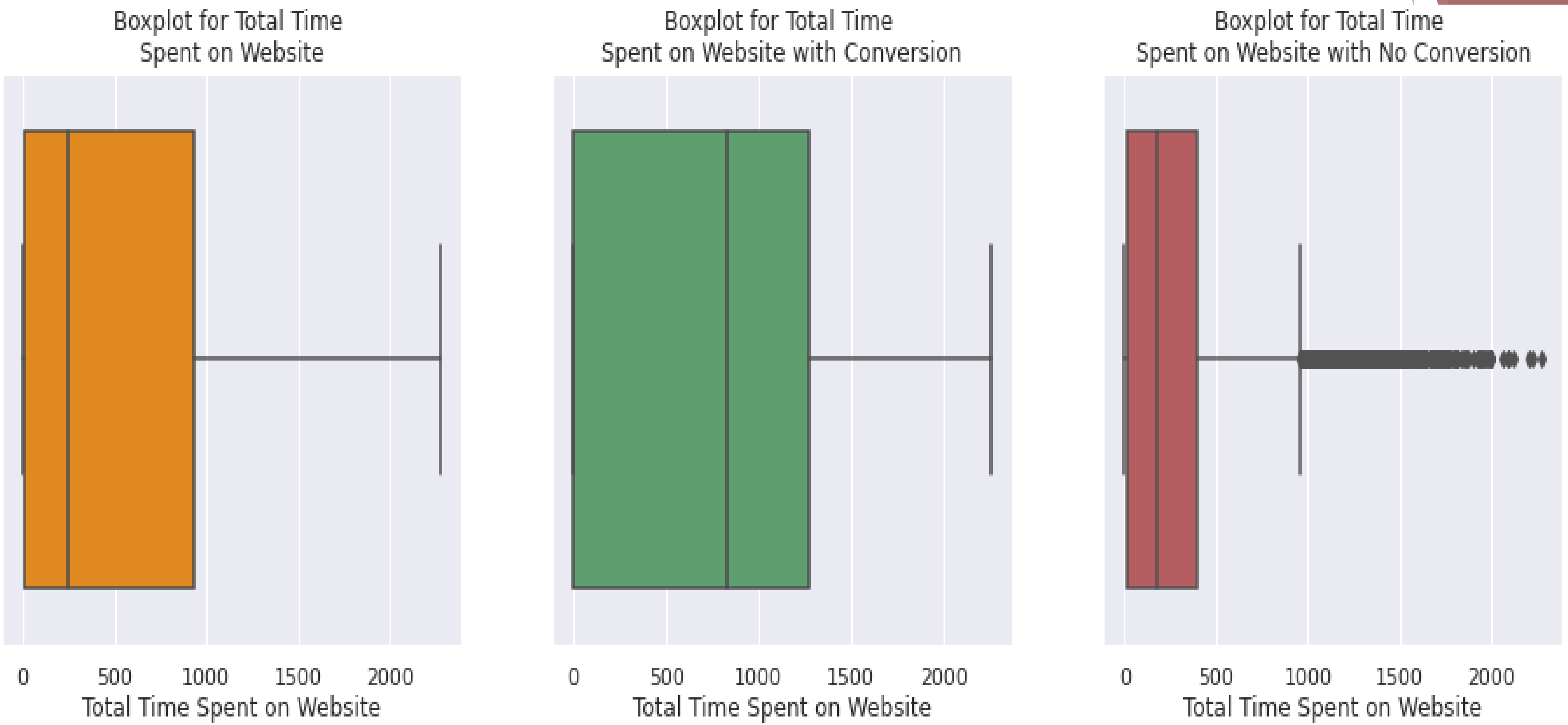
EDA

- Univariate data analysis – distribution of variable, value count, etc.
- Bivariate analysis – patterns between variables, correlation coefficients, etc.
- Classification technique – logistic regression used for model making, prediction
- Model presentation and validation
- Feature scaling, dummy variables and encoding of the data
- Conclusion

Data Manipulation→

- Total number of rows = 9240 and 32 columns for EDA analysis
- "Chain Content", "I agree to pay the amount through cheque", "Magazine", "Prospect ID", "Lead number" etc. have been dropped, as these aren't required for the analysis
- Some features that don't have enough variance have been dropped as well. These include →"Do not call", "What matters most to you in choosing course", "Newspaper article", "Search", "X Education Forums", "Digital Advertisement", etc.
- We also dropped columns that have more than 35% missing values
- Data imbalance was also checked, wherein 61% did not get converted but 39% did

EDA, Boxplot output upon analysis →

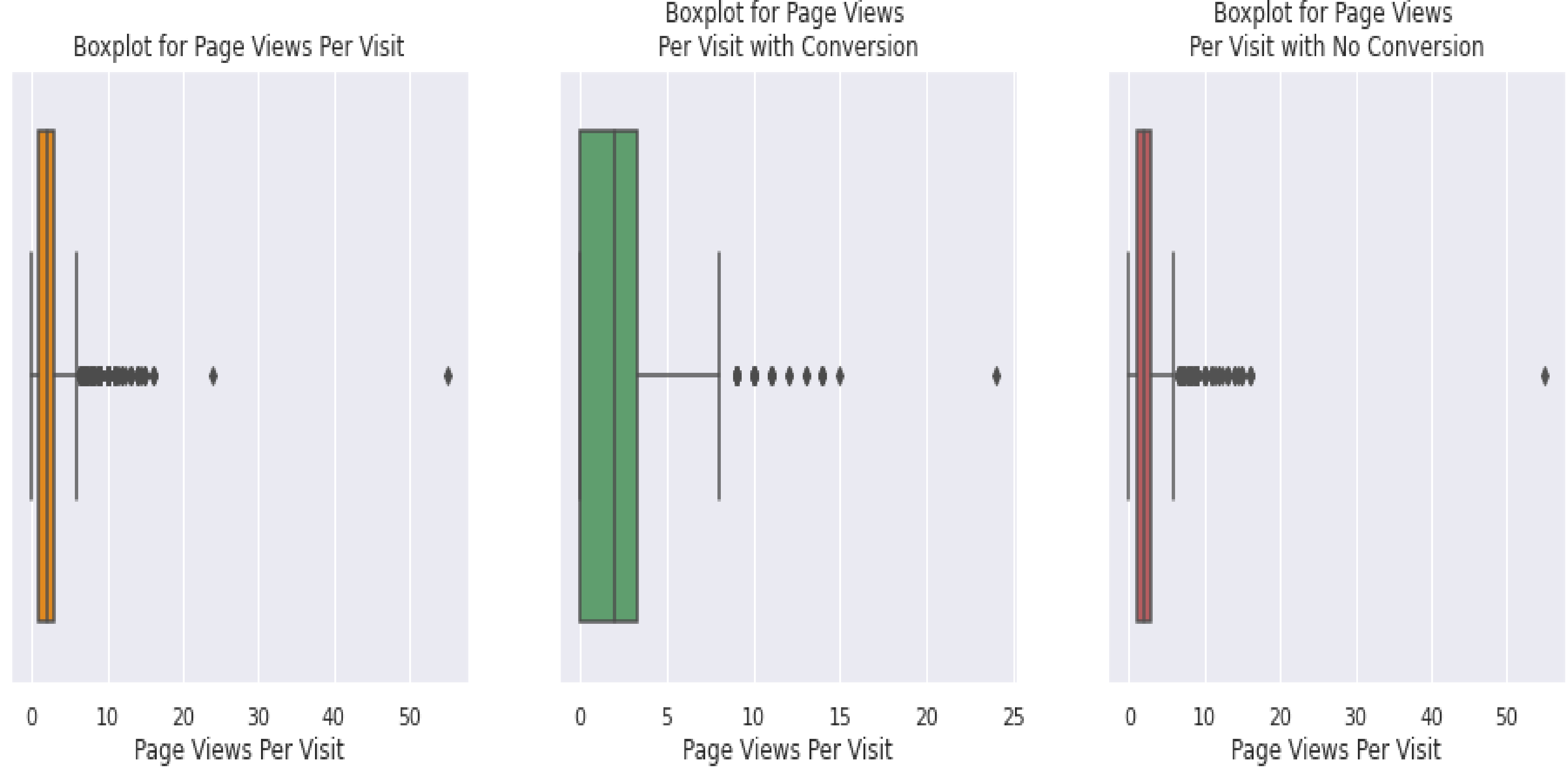


Boxplot for Total Time Spent on Website

Boxplot for Total Time Spent on Website with Conversion

Boxplot for Total Time Spent on Website with No Conversion

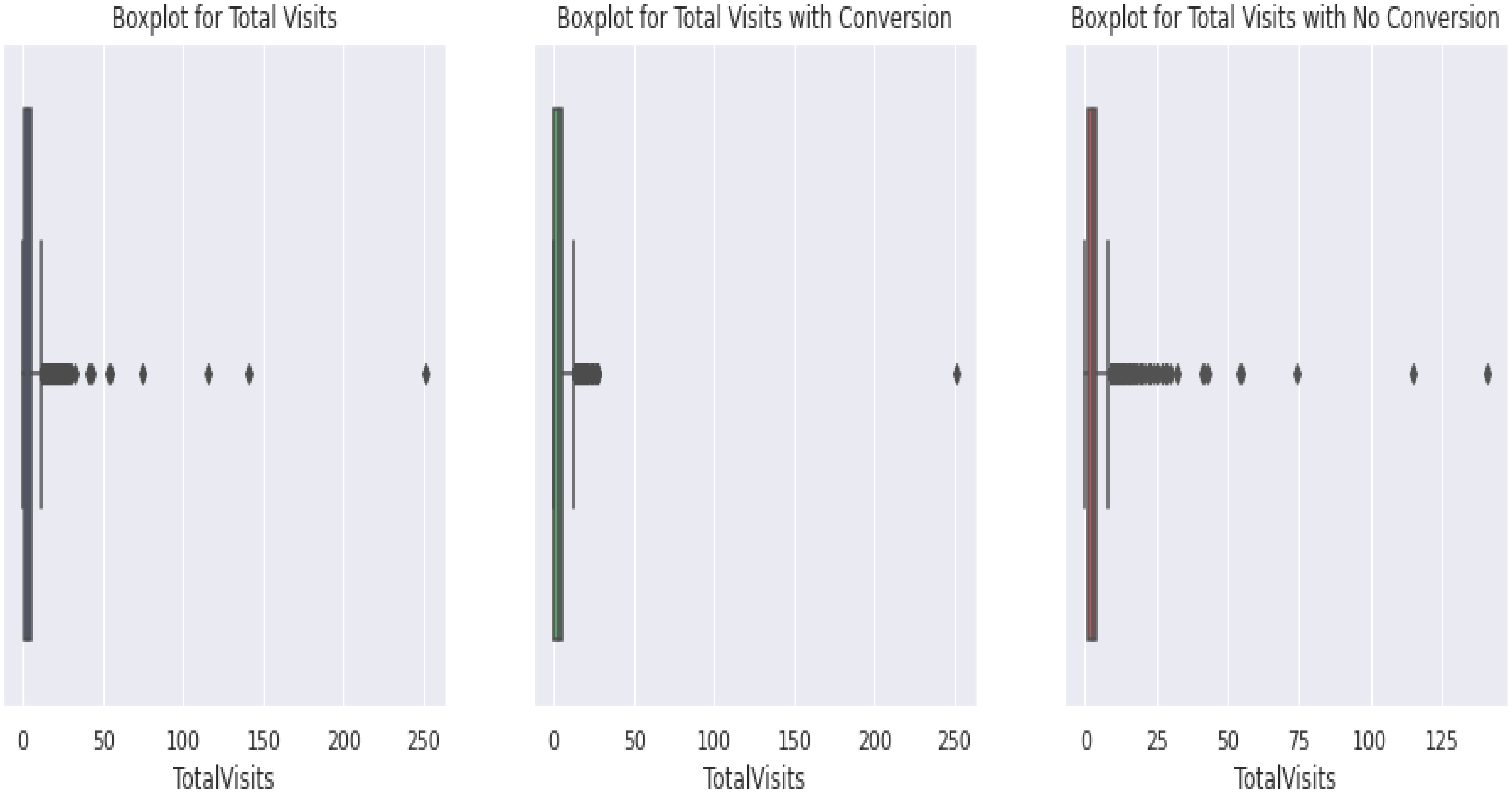
Here we can see some potential outliers for the class with "**No Conversion**" but when we look data at the aggregate level the effect of "**No Conversion**" seems to be neutralize because of "**Conversion** class seems to be far less skewed as compared to "**No Conversion**" class.


Boxplot for Page Views Per Visit

Boxplot for Page Views Per Visit with Conversion

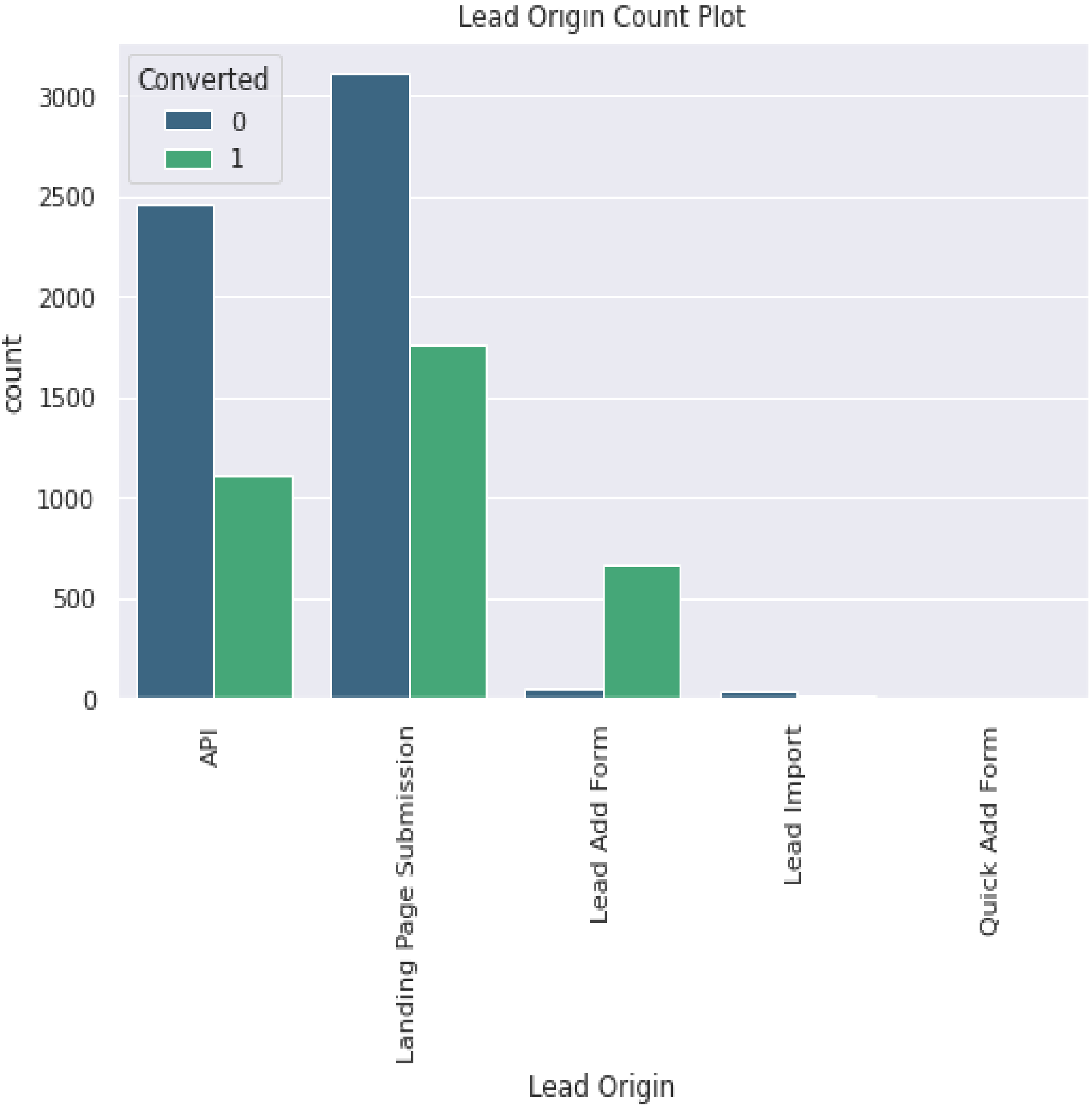
Boxplot for Page Views Per Visit with No Conversion

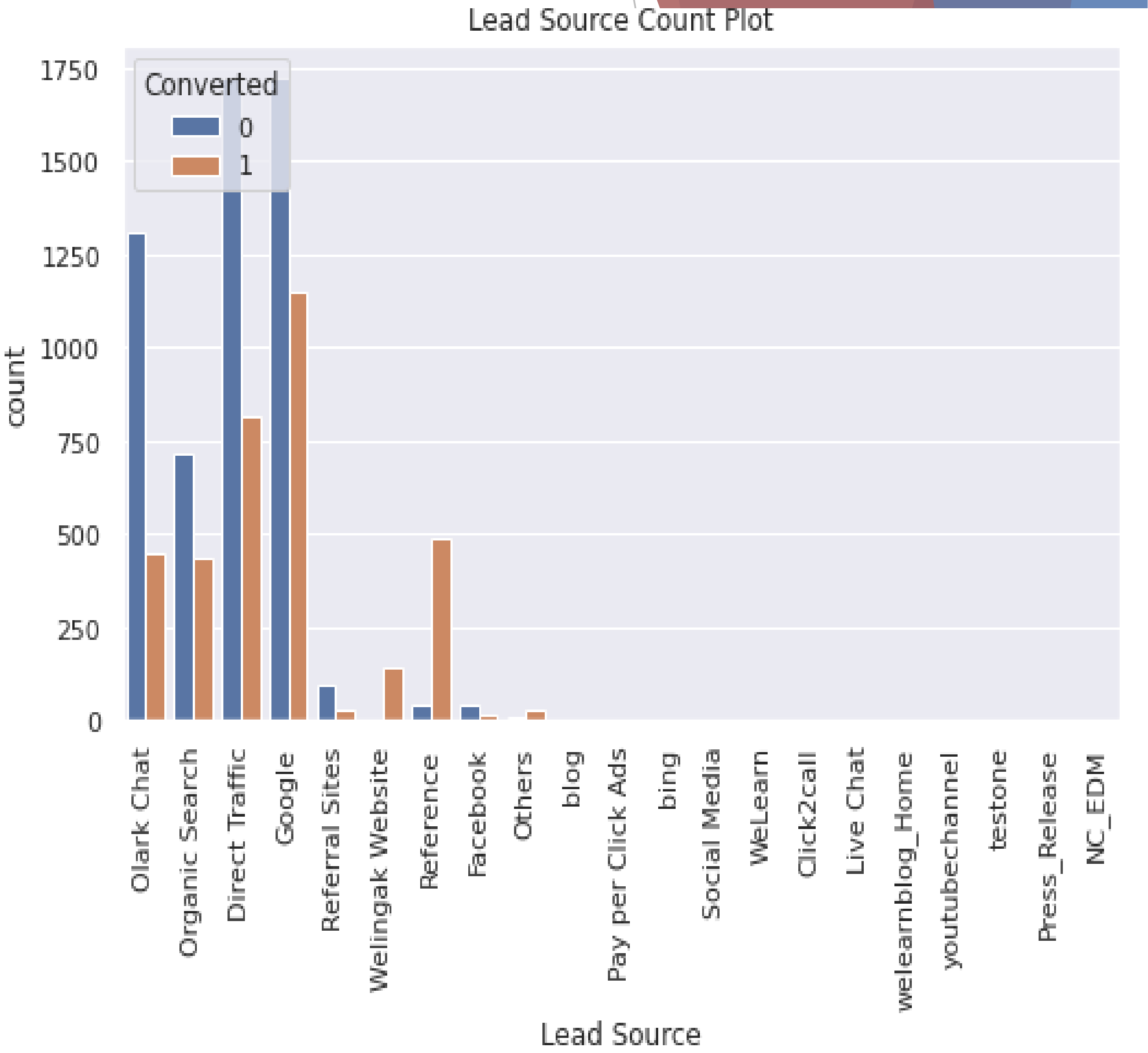
We can also see some extreme values here as well which indicated that we need to fix these outliers before procedding with our analysis.



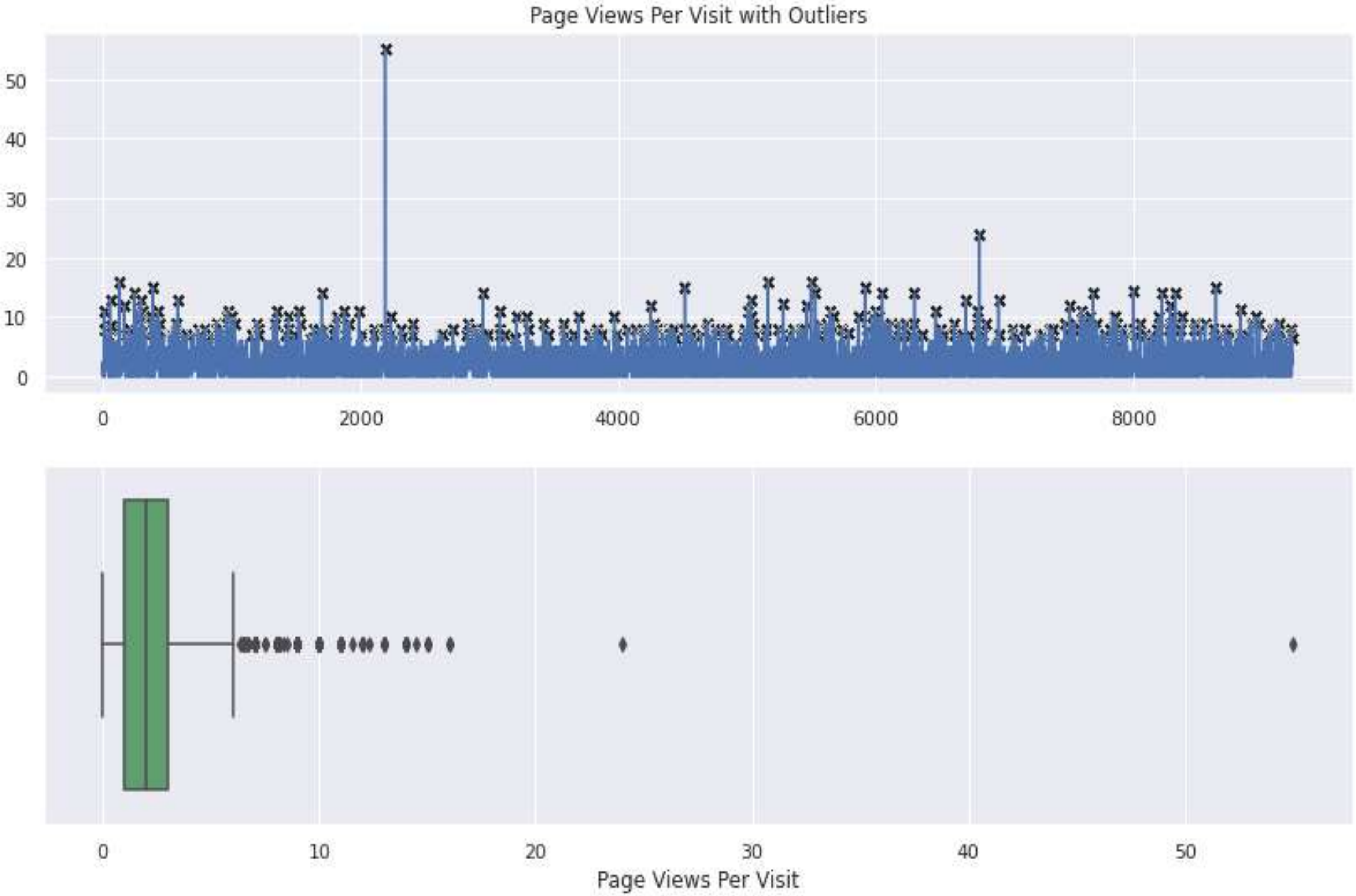
As we already see in our pairplot for the extreme values but via this boxplot, now we are clearly see some high extreme values in our data for the Total Visits.

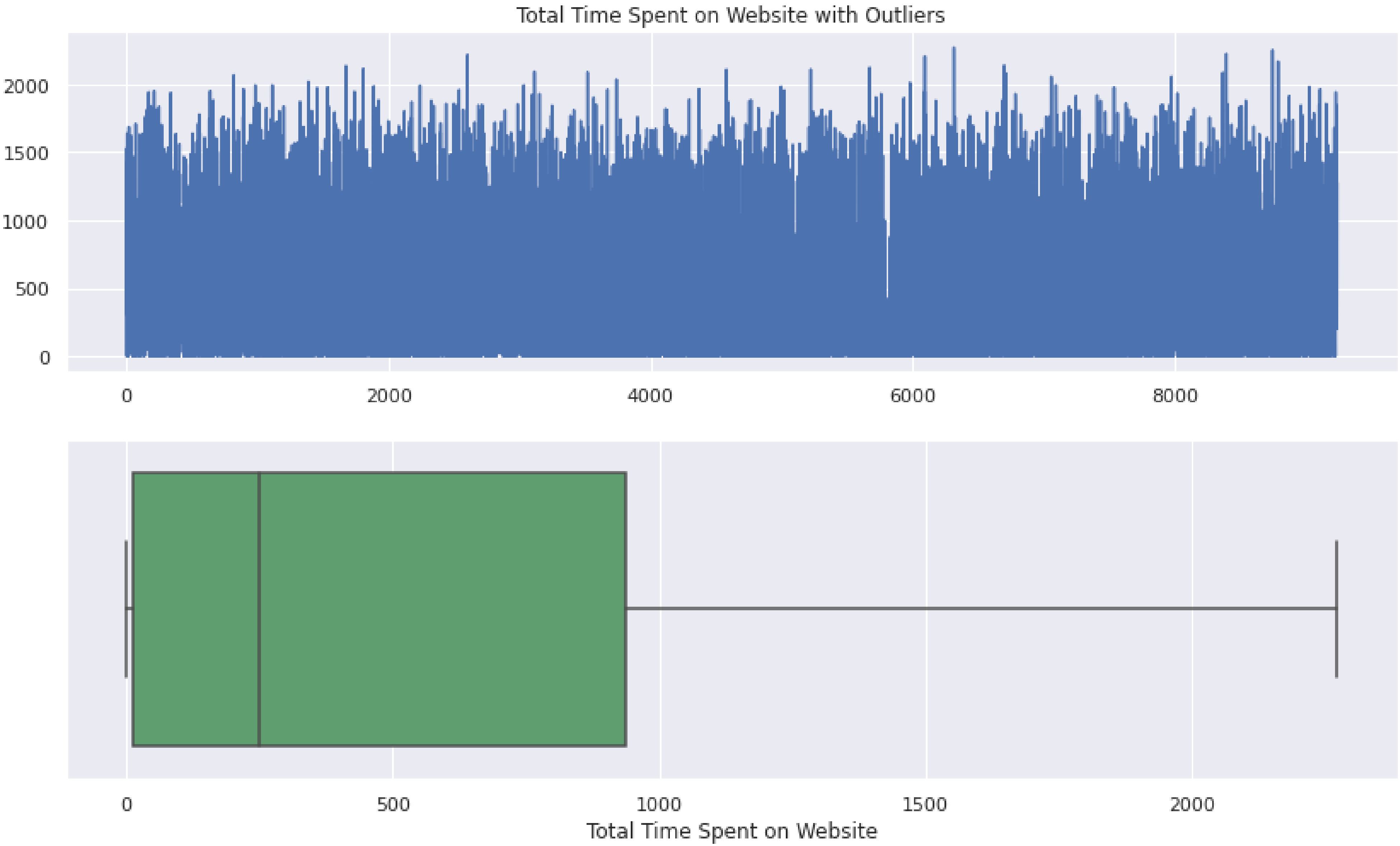
Categorical Variable Relation →





Outlier Analysis →

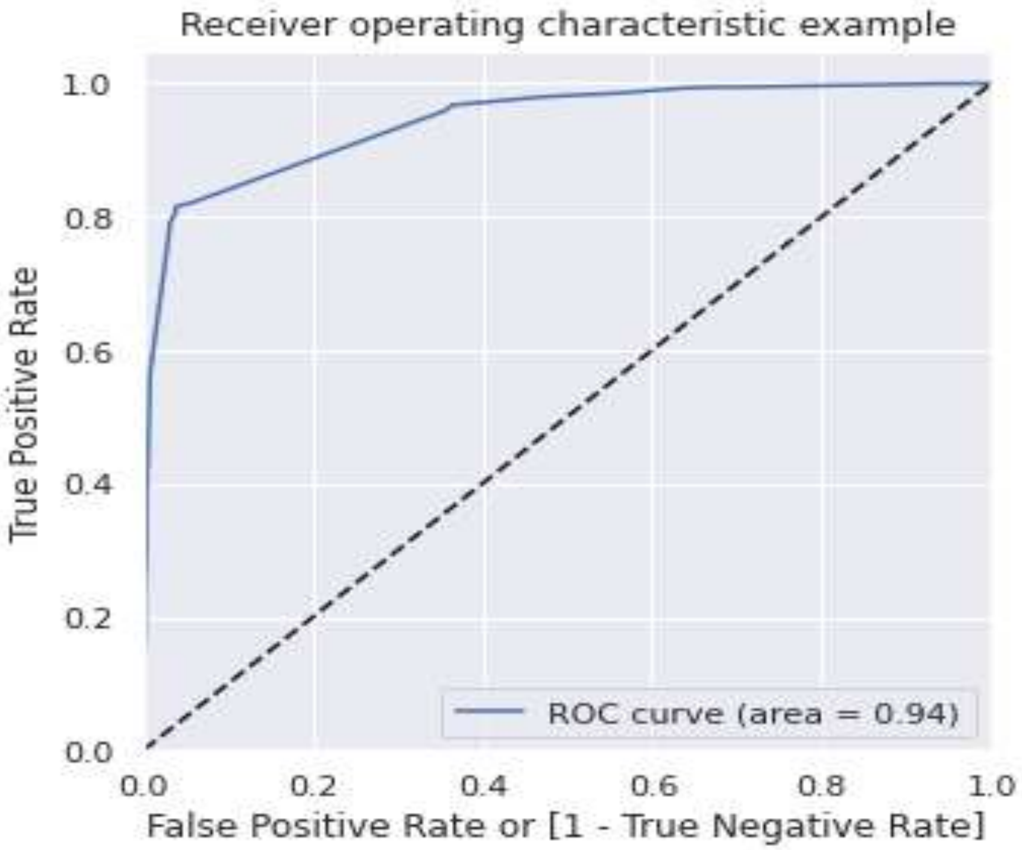



Total Time Spent on Website with Outliers
2000
1500
1000
500
0
0
2000
4000
6000
8000
0
500
1000
1500
2000
Total Time Spent on Website

Model Building →

- Splitting the data into train and test datasets
- The first-step for logistic regression is conducting a train-test split, which has been done in the ratio 70:30
- Using RFE for Feature Selection
- Running RFE with 15 variables as our output
- Model building where p value > 0.05 and VIF > 5
- Creating optimal lead score which is approximately 0.35
- Predictions on test dataset
- Overall accuracy attained is approximately 89%

ROC Curve & Gini of the model→



Gini of the model

We can see from the ROC curve, that the area of the curve is 0.92, which is the Gini of the model.

The curve is hugging the true positive rate axis.

Conclusion →

- The lead score calculated in the test dataset shows the conversion rate of 83% which was clearly more than the ballpark of the target rate which was around 80%. Thus, this clearly meets the expectation of the CEO.
- It was found that the variables that mattered the most were (in descending order)→
- Total time spent on the website
- Total number of visits (when lead source was – Google, direct traffic, organic search, Welingak website)
- When the last activity was →
- SMS
- Olark chat conversation
- When the lead origin is Lead add Format
- When they are working professionals
- Keeping the above points in mind, X Education can flourish as they have a good opportunity to convert almost all the potential buyers to full-time customers.