LEAD SCORING CASE STUDY

MANALI SHAH
SHIVANI MURDE
SONAL DAHALE

PROBLEM STATEMENT

- X Education is an online education company that provides courses to industry professionals. The company employs marketing channels, such as Google search engines and websites, to attract leads to their platform. Once on the platform, leads can browse through the available courses or fill out a form to learn more information. These leads can also come through referrals.

- The sales team at X Education works to convert the acquired leads into paying customers. However, despite many leads, the company's conversion rate is only around 30%. In order to increase the conversion rate, X Education needs a model to prioritize leads for the sales team to focus on. The CEO has set a target
- X Education is an online education company that provides courses to industry professionals. The company employs marketing channels, such as Google search engines and websites, to attract leads to their platform. Once on the platform, leads can browse through the available courses or fill out a form to learn more information. These leads can also come through referrals.

- The sales team at X Education works to convert the acquired leads into paying customers. However, despite many leads, the company's conversion rate is only around 30%. In order to increase the conversion rate, X Education needs a model to prioritize leads for the sales team to focus on. The CEO has set a target conversion rate of 80%.

OBJECTIVE

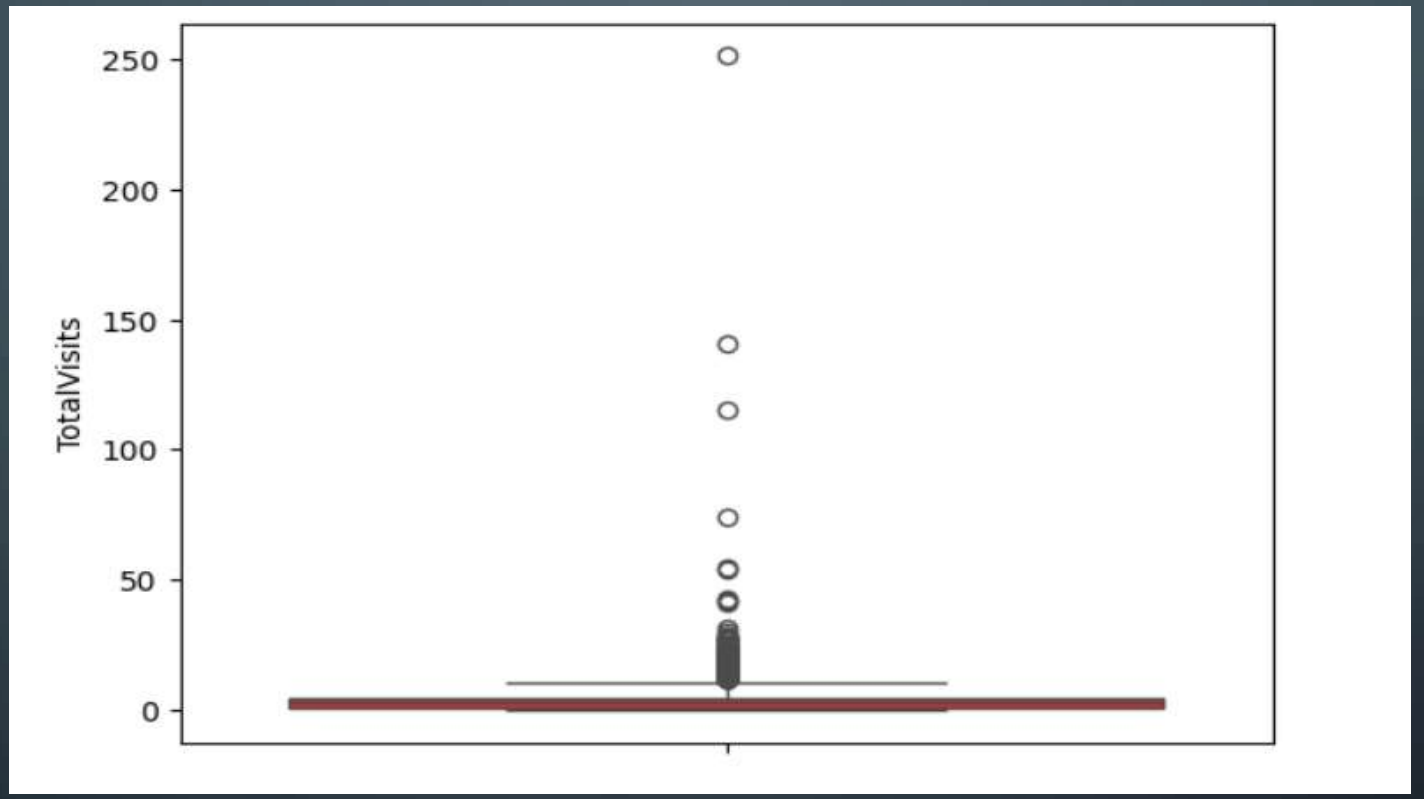
- To achieve the CEO's goal of an 80% conversion rate, X Education needs a lead scoring model. The model will assign a score between 0 and 100 to each lead based on their likelihood of conversion. Leads with higher scores will be considered more promising and will be prioritized by the sales team. The model will take into account data such as browsing behavior, form fill-ups, video views, and past referrals to calculate the score.

- The model will be built using logistic regression, and will provide a way for the sales team to focus their efforts on leads that are most likely to convert. By prioritizing the leads with higher scores, the company hopes to increase the conversion rate and achieve the CEO's target.

APPROACH

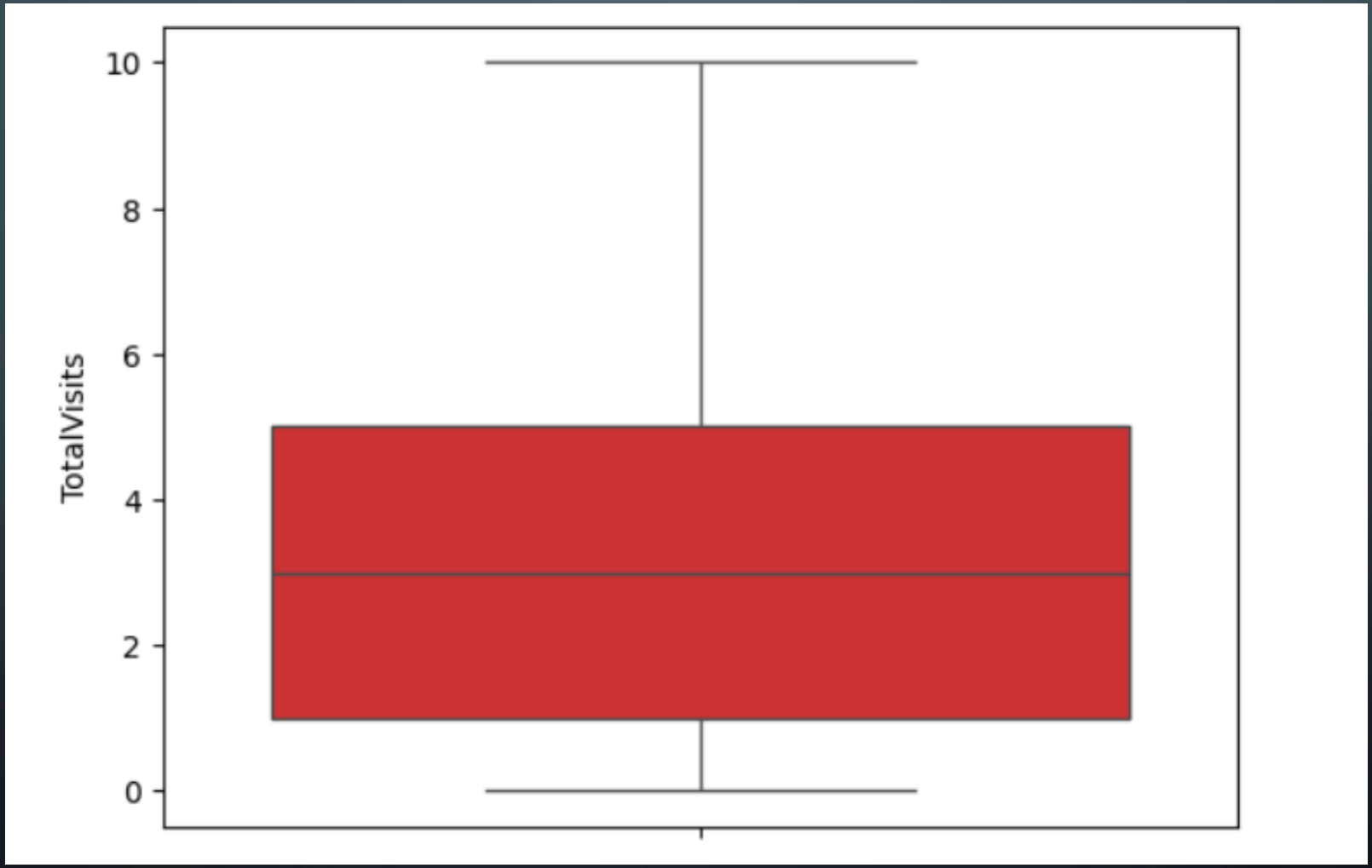
- Logistic Regression
- Data Understanding, cleaning and preparation
- Exploratory Data Analysis
- Data Preparation
- Features Selection using RFE
- Modal Building
- Modal Evaluation

OUTLIERS

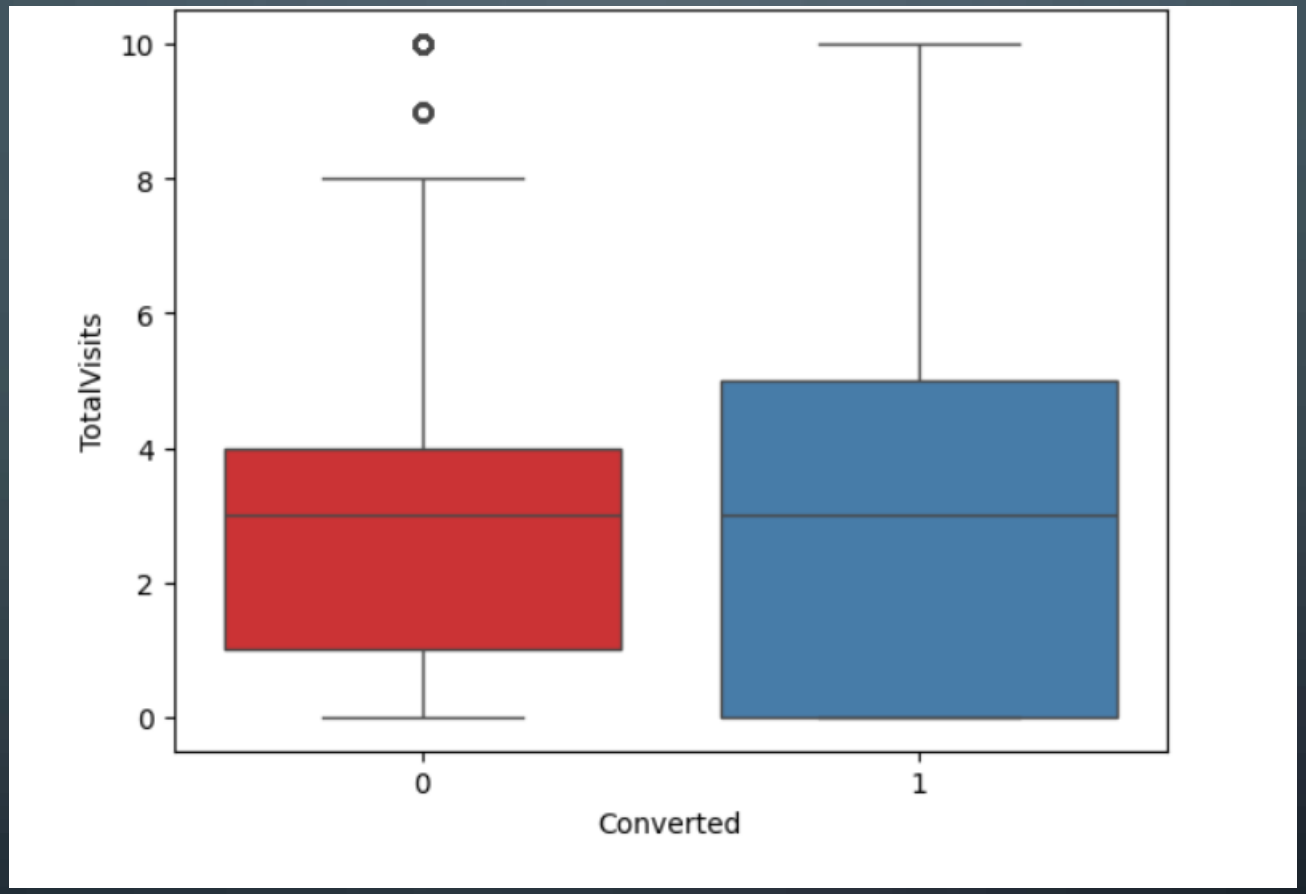


As we can see there are a number of outliers in the data. We will cap the outliers to 95% value for analysis.

OUTLIERS

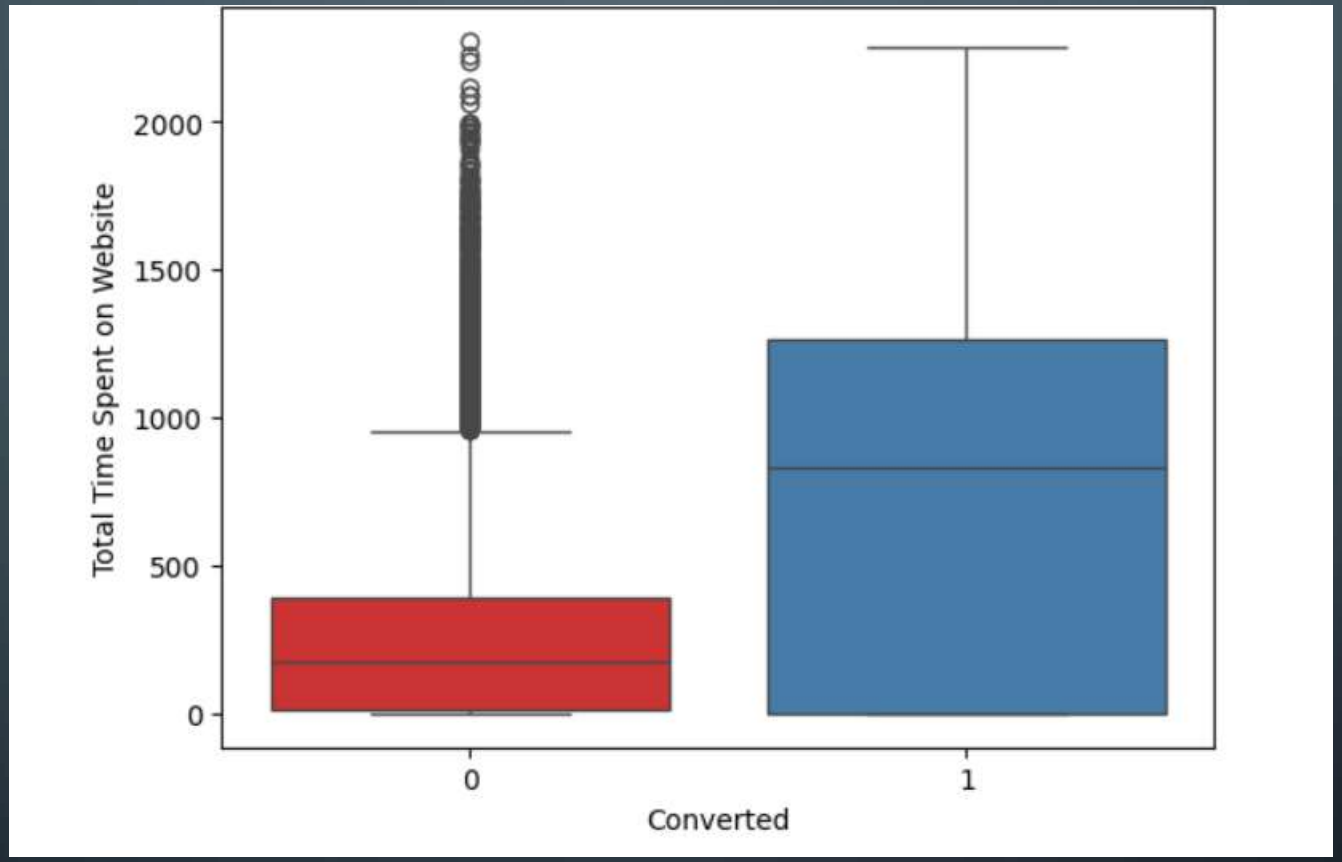


OUTLIERS



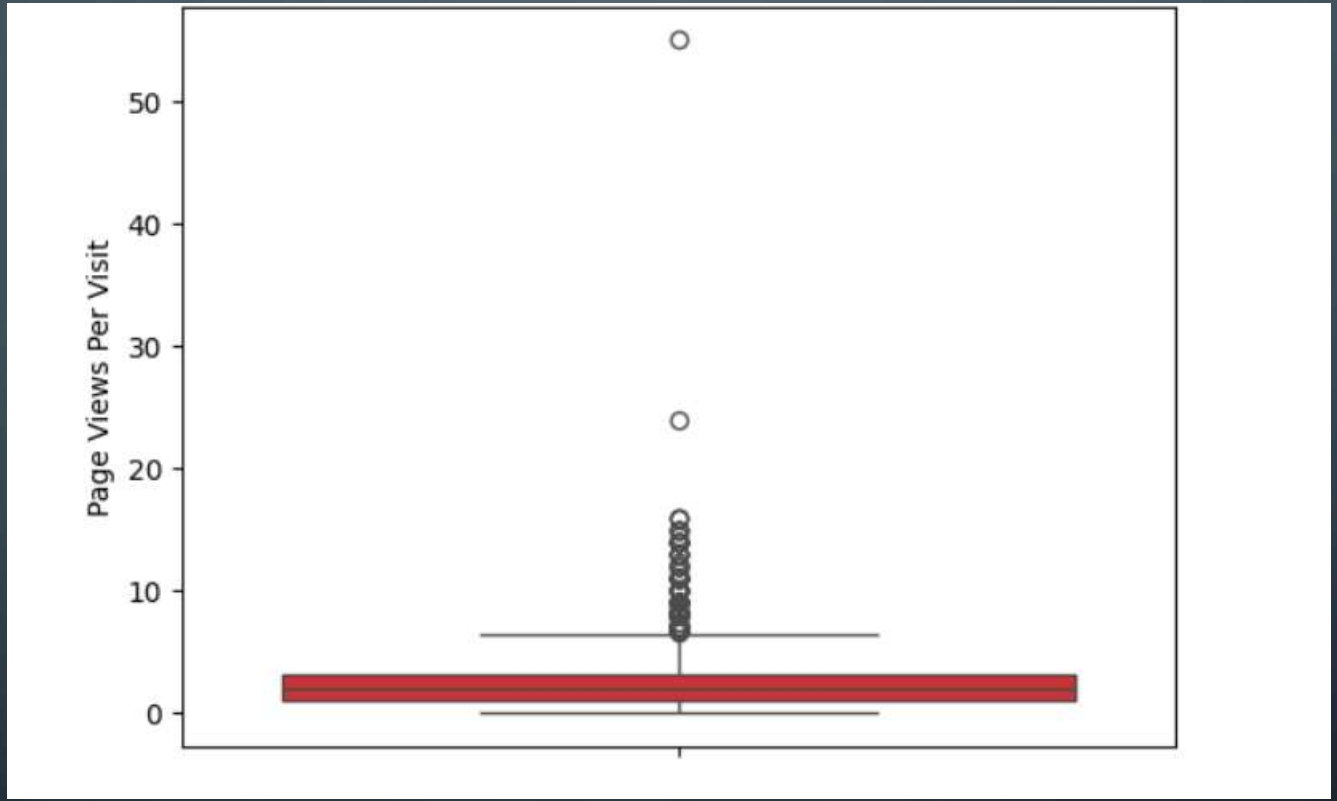
Median for converted and not converted leads are the same.
Nothing can be concluded on the basis of Total Visits.

OUTLIERS



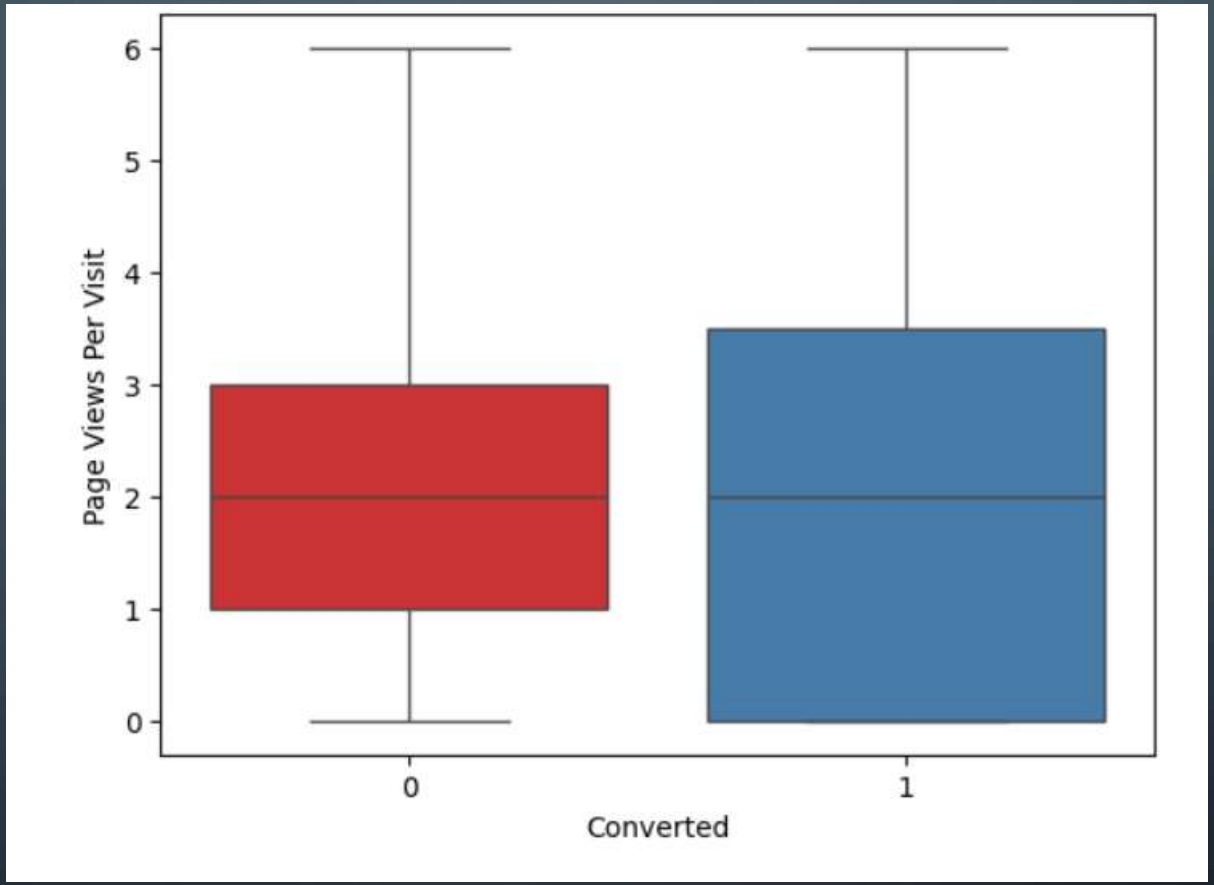
Leads spending more time on the weblise are more likely to be converted
Website should be made more engaging to make leads spend more time

OUTLIERS



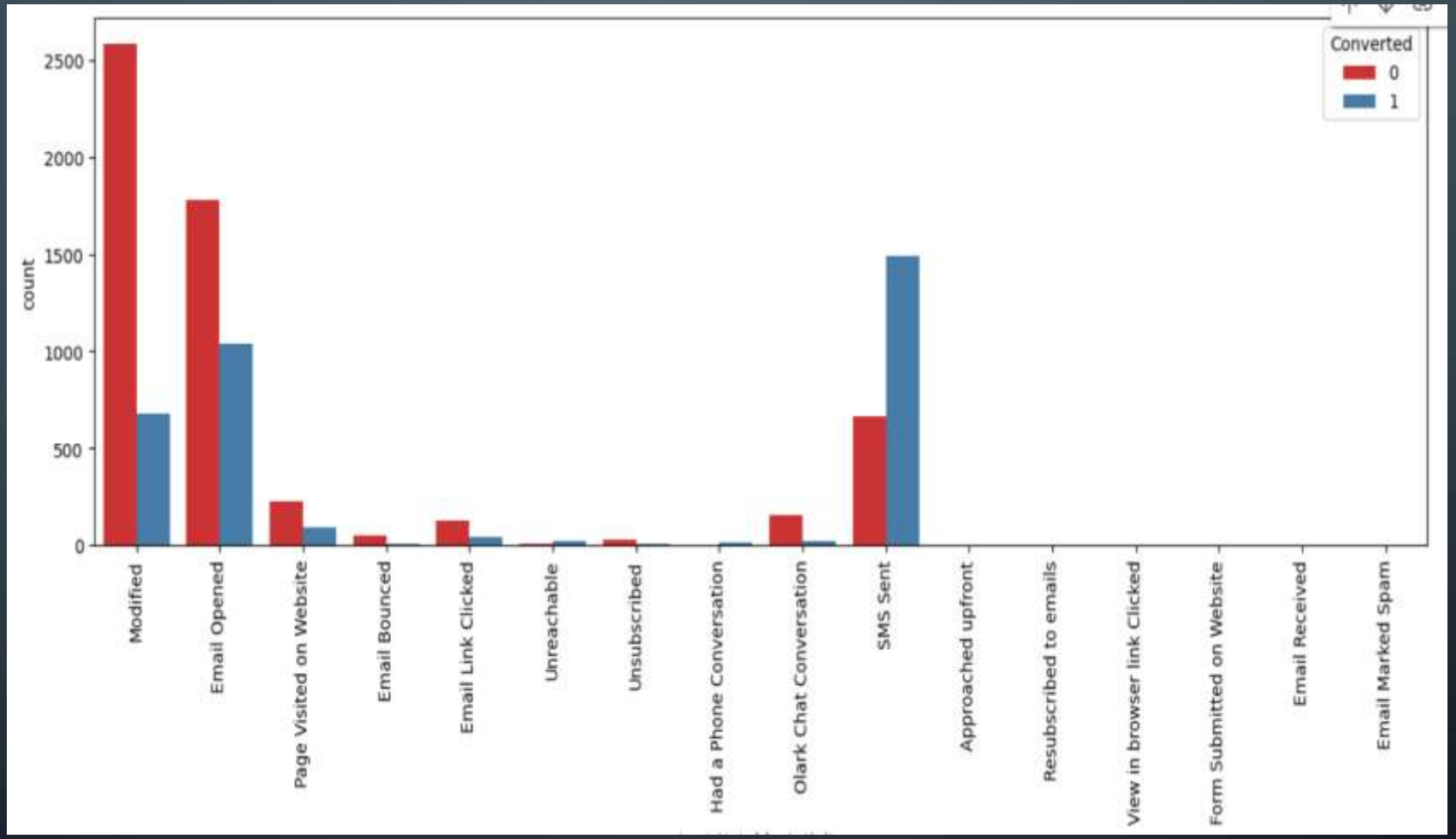
As we can see there are a number of outliers in the data. We will cap the outliers to 95% value for analysis.

OUTLIERS



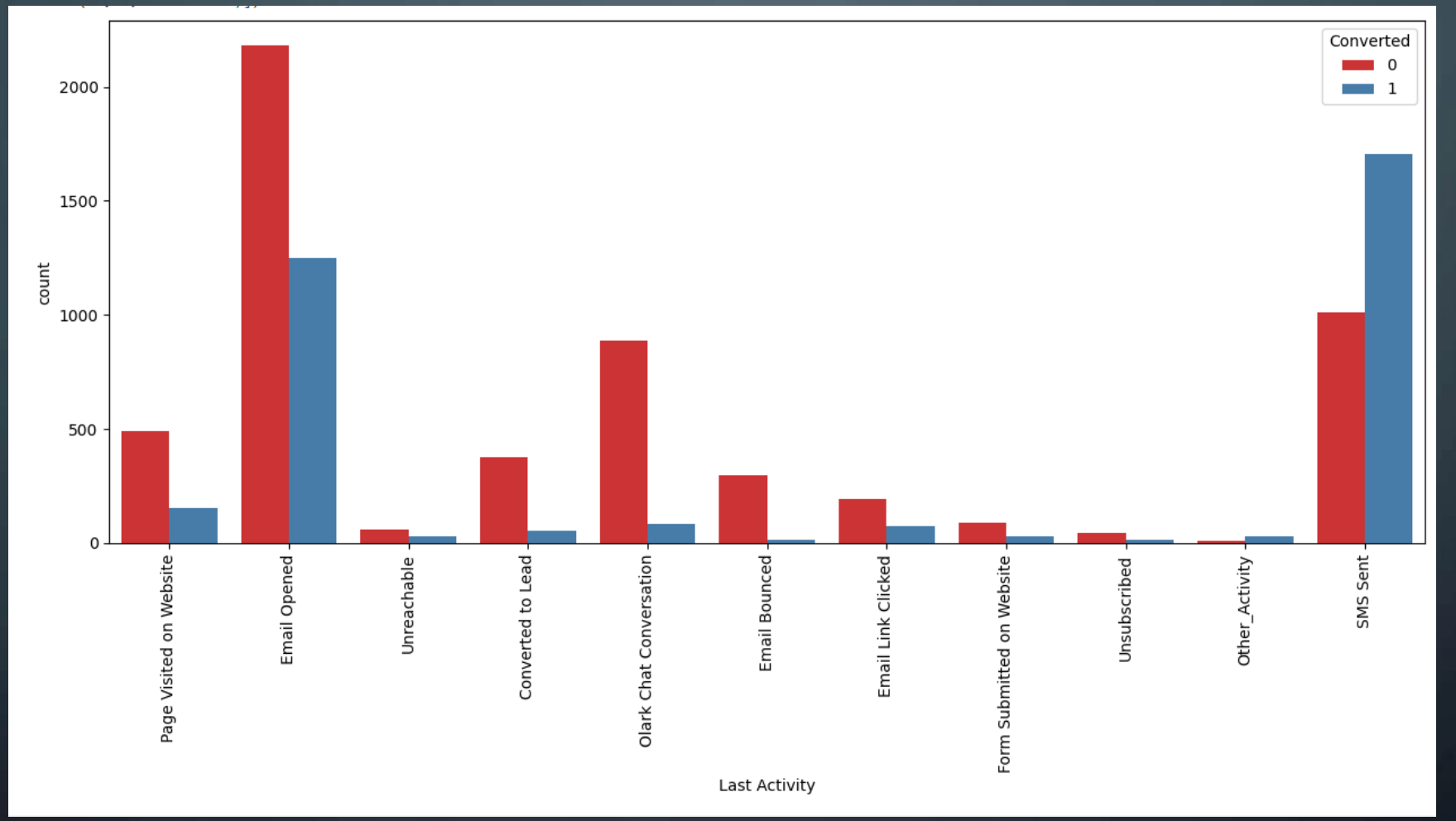
Median for converted and unconverted leads is the same.
Nothing can be said specifically for lead conversion from Page Views Per Visit

VISUALIZATION



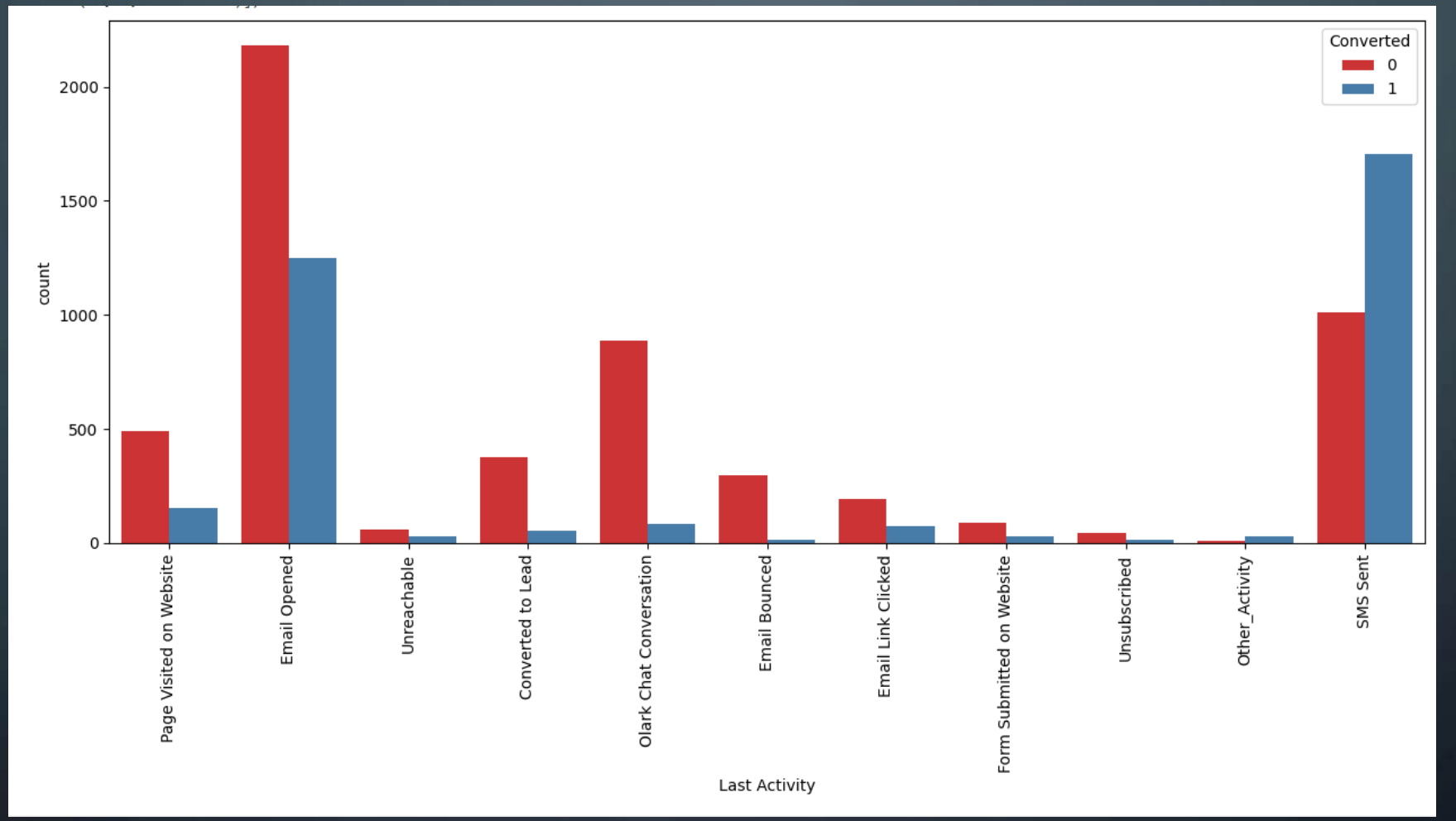
Based on the univariate analysis we have seen that many columns are not adding any information to the model, hence we can drop them for further analysis

VISUALIZATION



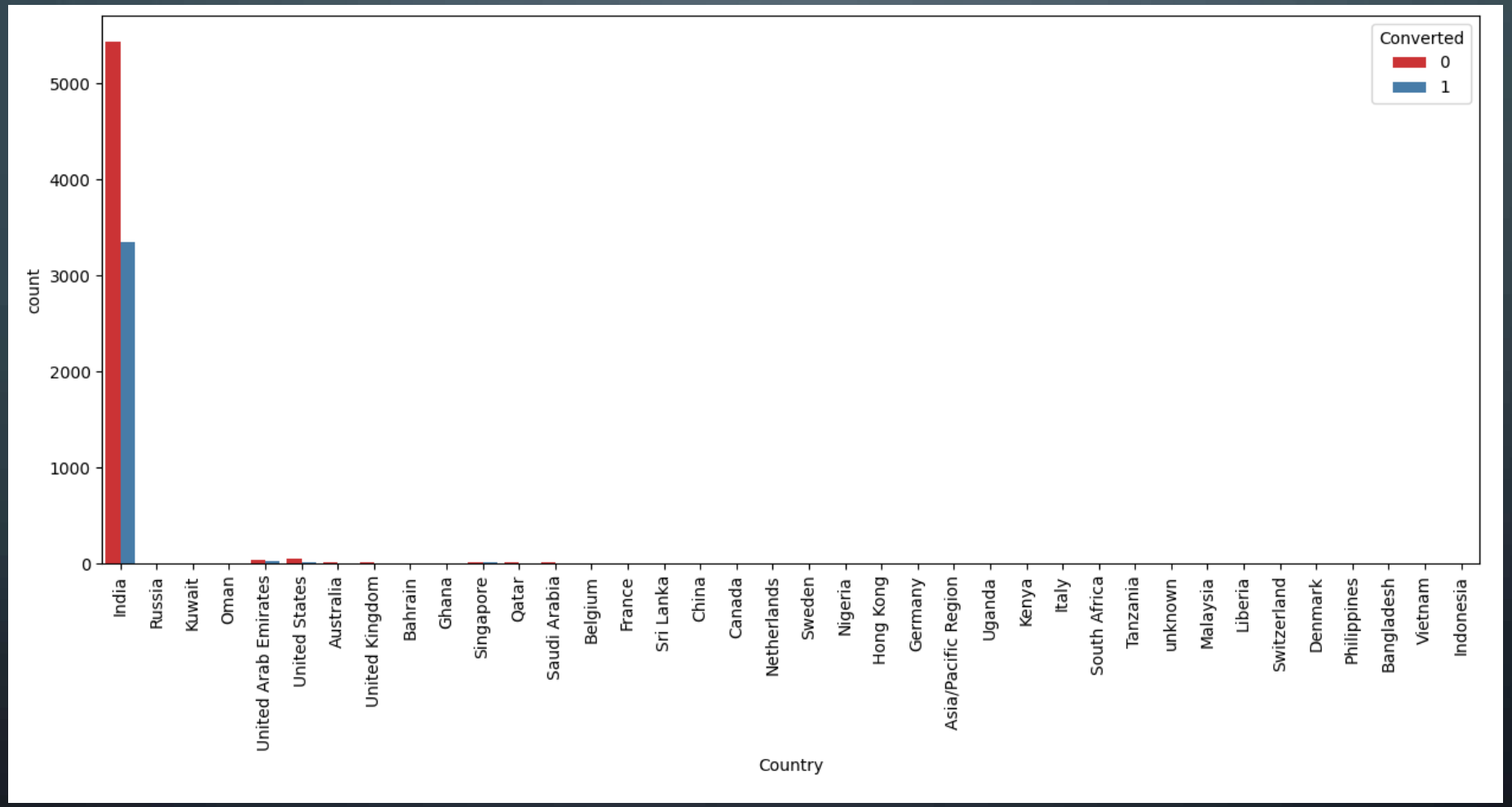
Most of the lead have their Email opened as their last activity.
Conversion rate for leads with last activity as SMS Sent is almost 60%.

VISUALIZATION



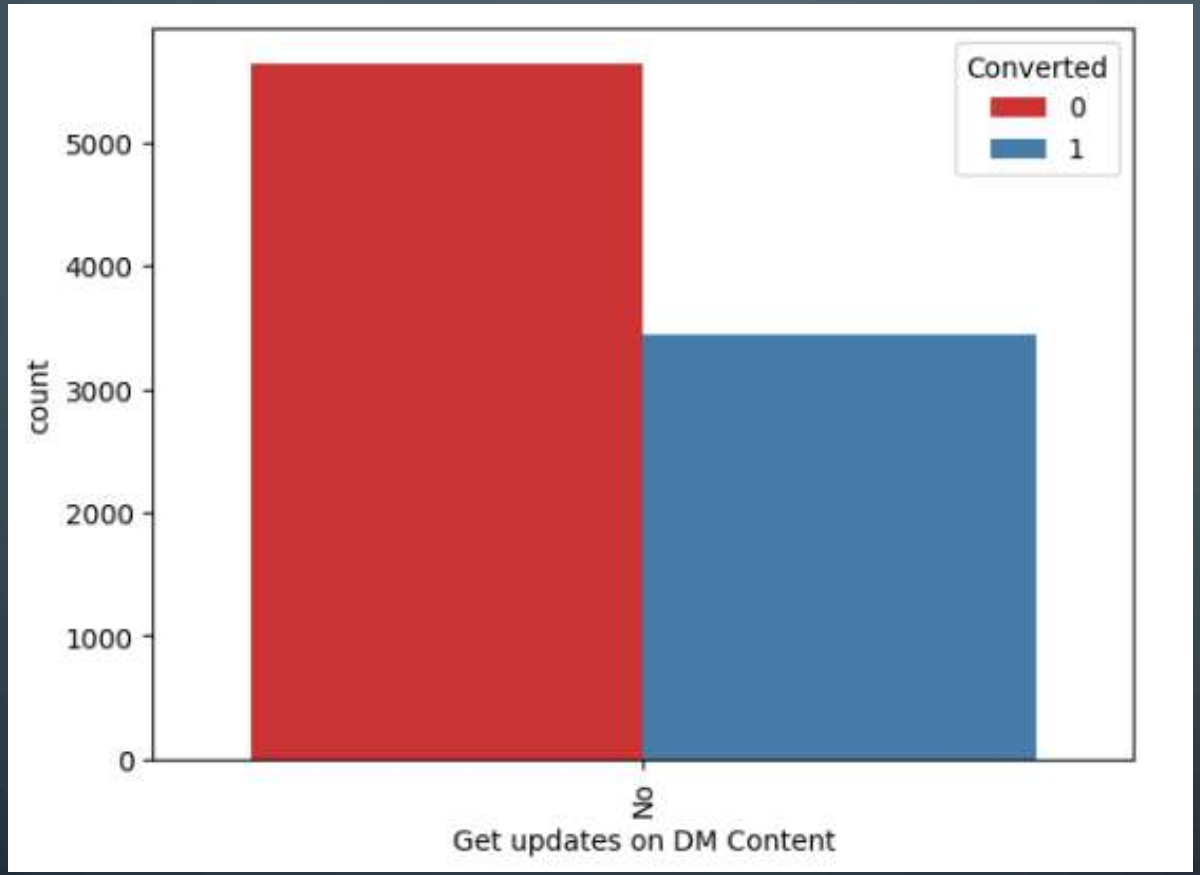
Most of the lead have their Email opened as their last activity.
Conversion rate for leads with last activity as SMS Sent is almost 60%.

VISUALIZATION



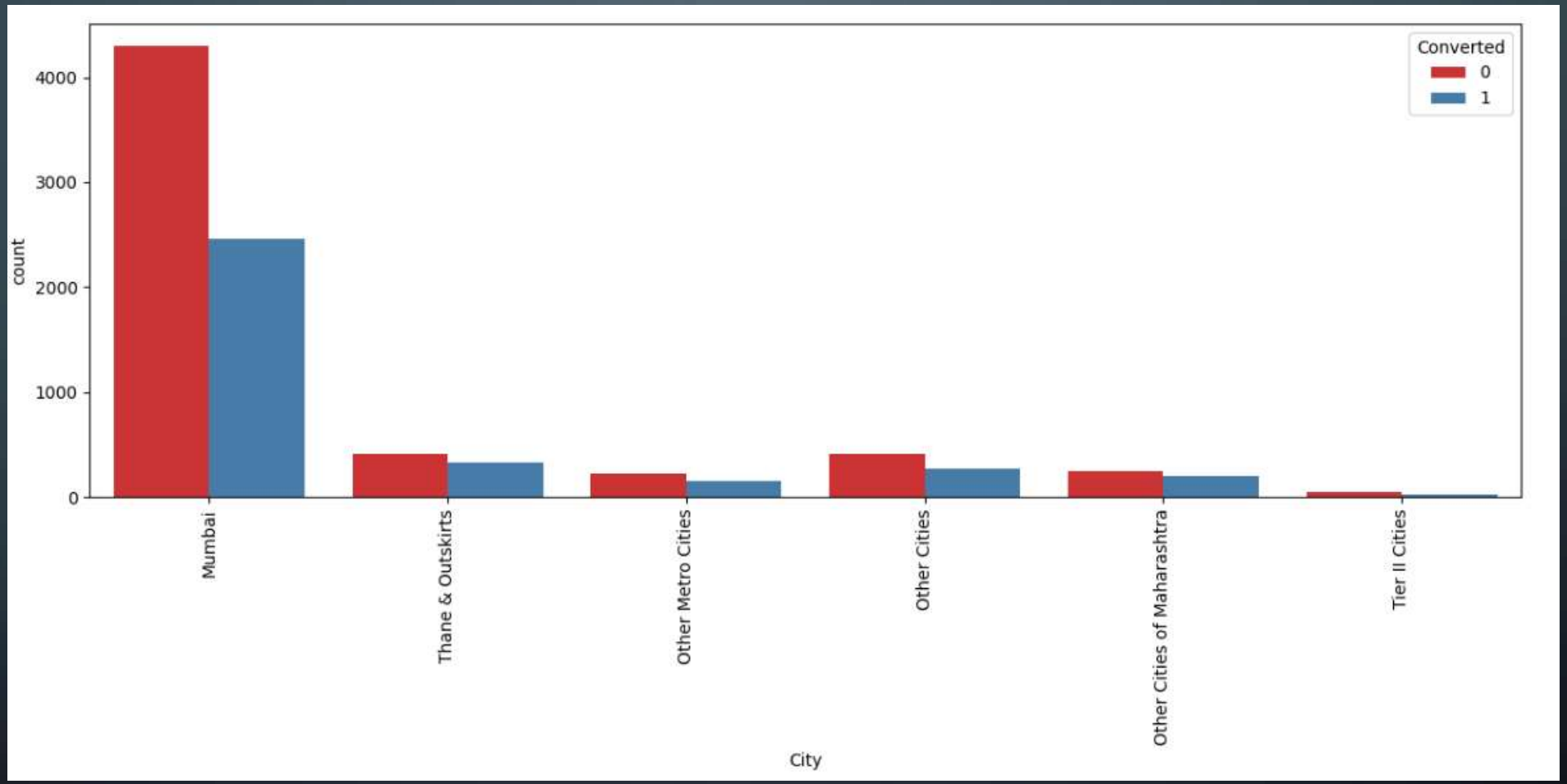
Most values are 'India' no such inference can be drawn

VISUALIZATION



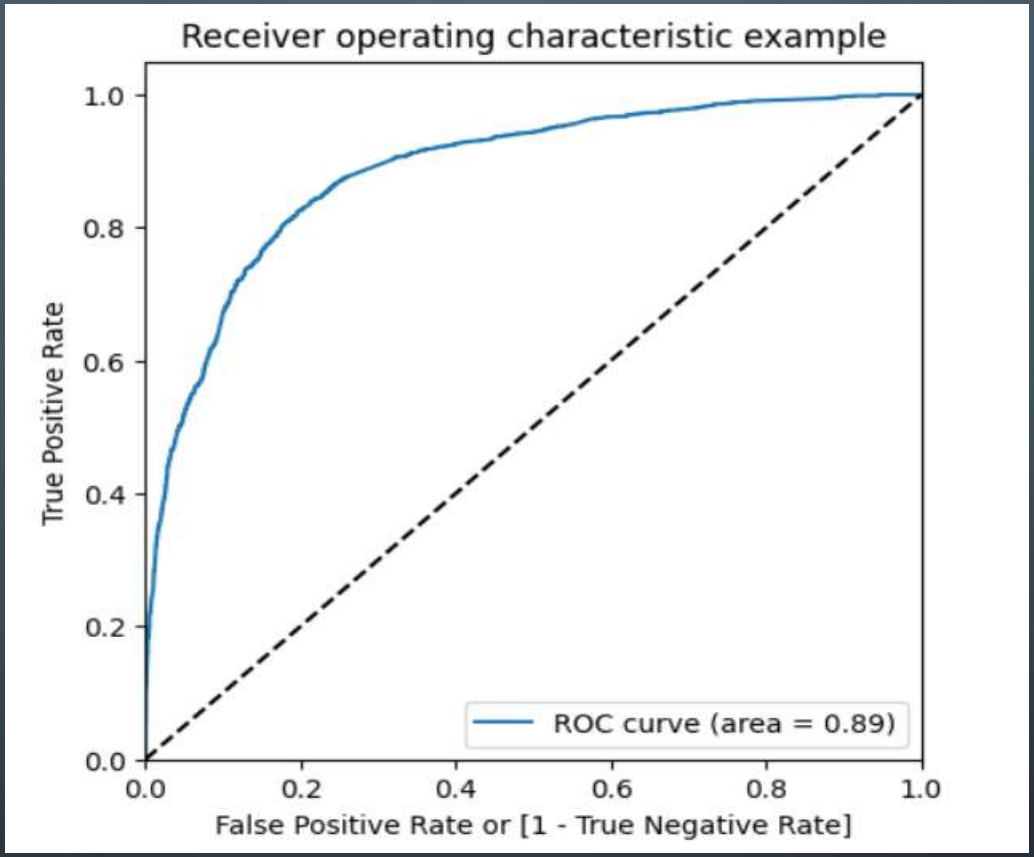
Most entries are 'No'. No Inference can be drawn with this parameter.

VISUALIZATION



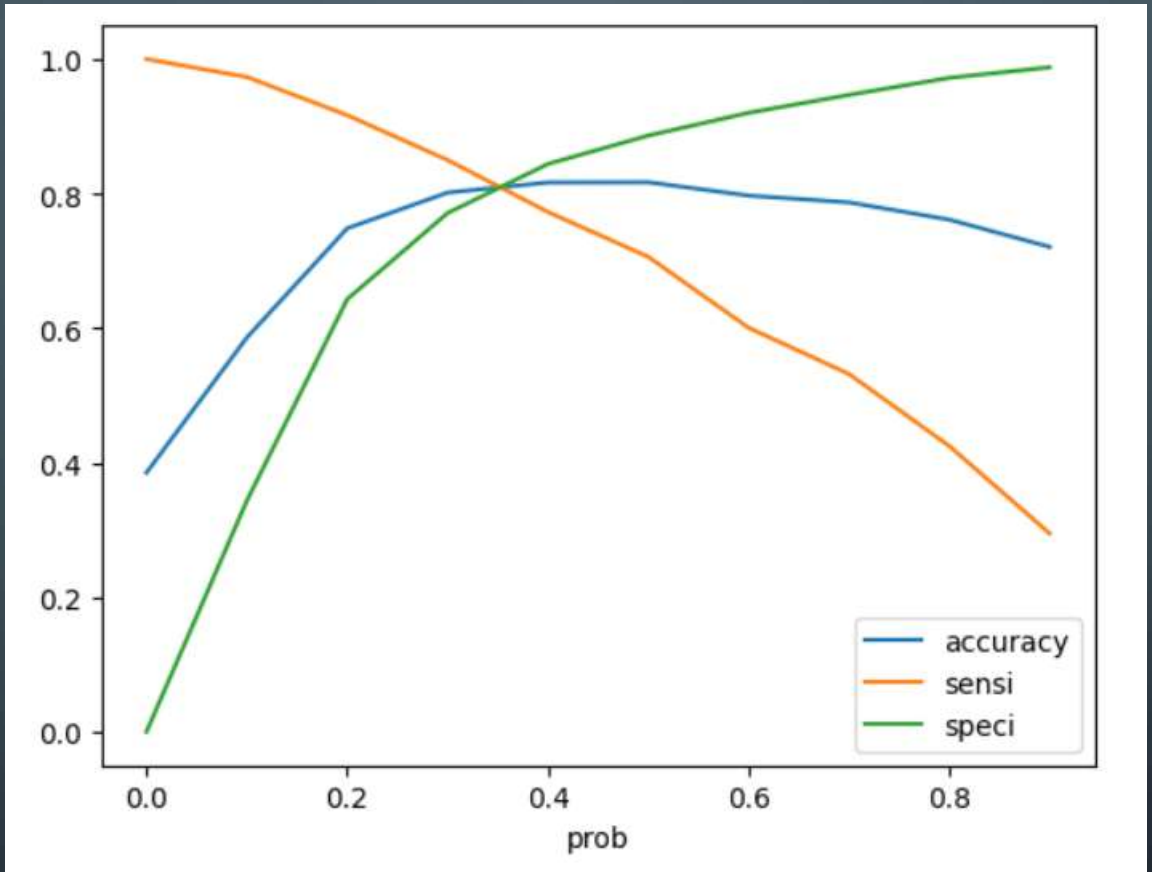
Most leads are from mumbai with around 50% conversion rate.

MAKING PREDICTION ON TRAIN SET



Since we have higher (0.89) area under the ROC curve , therefore our model is a good one.
Finding Optimal Cutoff Point
Above we had chosen an arbitrary cut-off value of 0.5. We need to determine the best cut-off value and the below section deals with that. Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

MAKING PREDICTION ON TRAIN SET



From the curve above, 0.34 is the optimum point to take it as a cutoff probability

OBSERVATIONS

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

- Accuracy : 80.4 %

- Sensitivity : 80.4 %

- Specificity : 80.5 %

RESULTS :

1) Comparing the values obtained for Train & Test:- Thus we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% . The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.

	Train Data:	Test Data:
Accuracy :	81.0%	80.4%
Sensitivity :	81.7%	80.4%
Specificity :	80.6%	80.5%

2) Finding out the leads which should be contacted:- The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 85. They can be termed as 'Hot Leads'.

THANK YOU