LEAD SCORING CASE STUDY

By – Devanshi Sinha & Priyanka Solanke

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- X Education gets a lot of leads, however 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'.
- After successfully identifying 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 OBJECTIVE

- To help the company select the most promising leads
- 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 where a higher score would mean that the lead is 'hot lead'.
- The model should be able to adjust to if the company's requirement changes in the future.

ANALYSIS APPROACH AND METHODOLOGY

Reading and Understanding the data

- *Shape*
- *Info*
- *Datatypes*
- *Statistical summary*

Data Cleaning

- *Checking null/missing values*
- *Handling null values*
- *Removing redundant columns if any*

Analyzing the Data

- *Univariate Analysis of all the variables*
- *Bivariate Analysis of all the variables*
- *Correlation between different variables*

Data Preparation

- *Treating outliers*
- *Creating dummy variables for categorical variables*
- *Splitting data into train and test dataset*
- *Scaling the continuous variables*

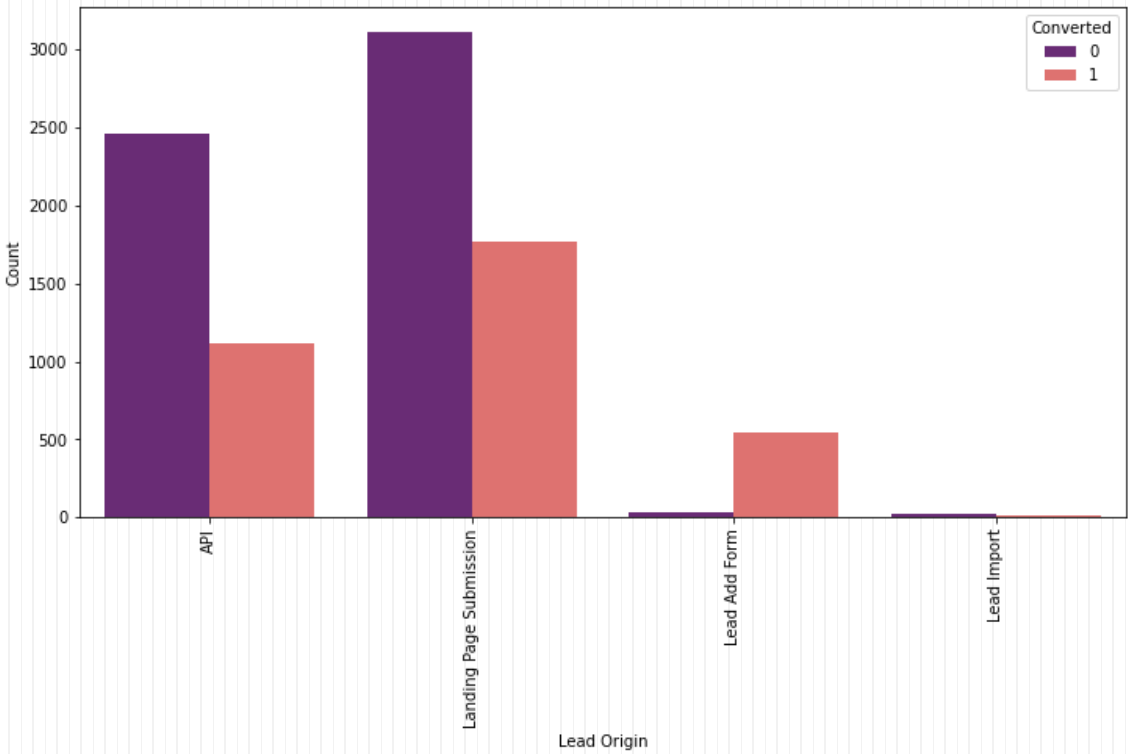
Model Building

- *Using automated feature selection approach: RFE*
- *Building Logistic Regression Model*
- *Using manual approach of feature selection using p-value & vif*

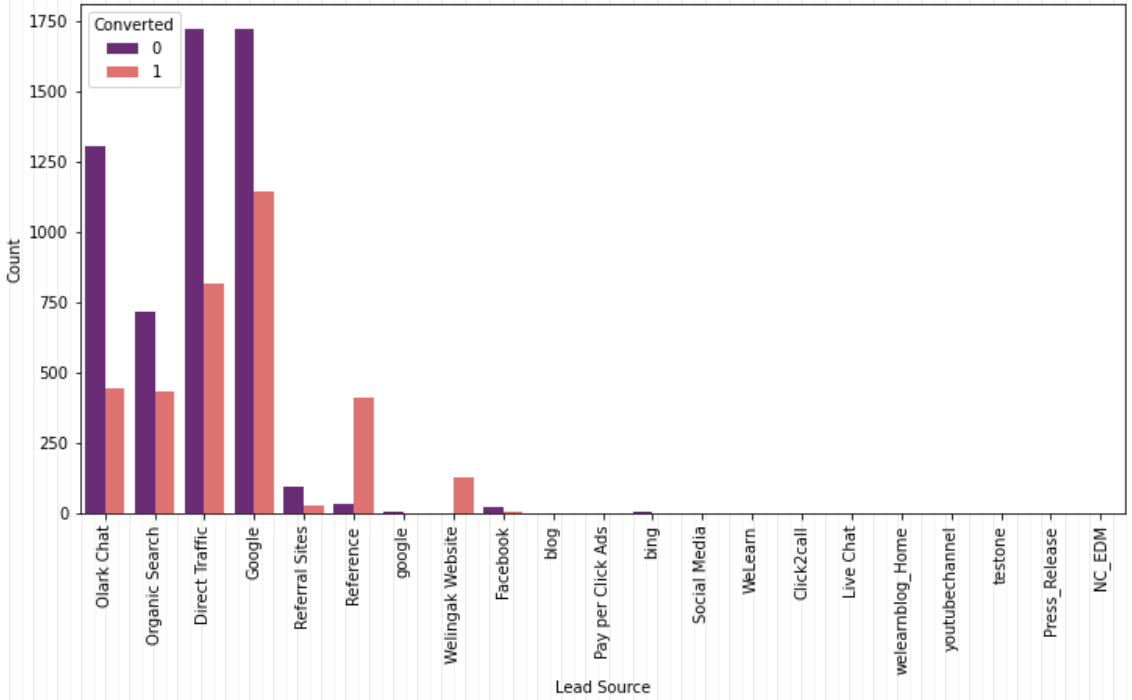
Model Evaluation & Deployment

- Evaluating the model using ROC curve, accuracy, sensitivity, recall, etc.
- Assigning lead score to find most promising leads
- Validating model using test data set making predictions

UNIVARIATE ANALYSIS: CATEGORICAL VARIABLES

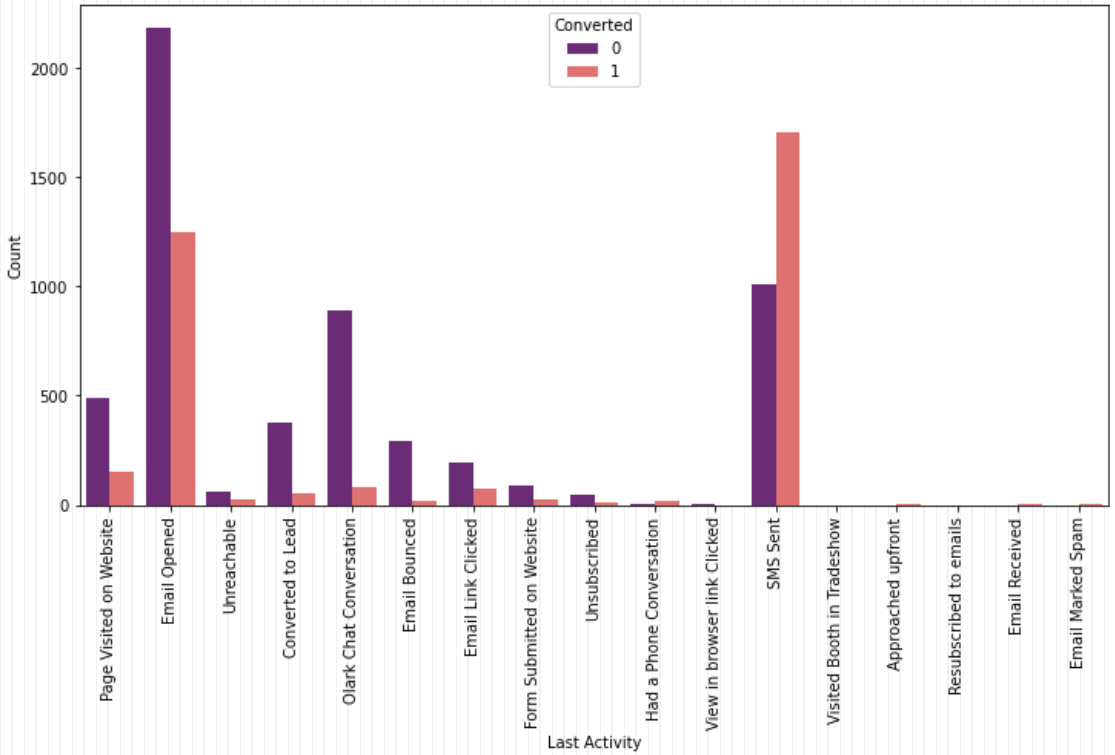


- The lead count from lead origins 'API' and 'Landing Page Submission' are quite significant.
- Lead origin from ‘Lead Add Form' has a high conversion rate (about 90%) however the count of leads is very low.

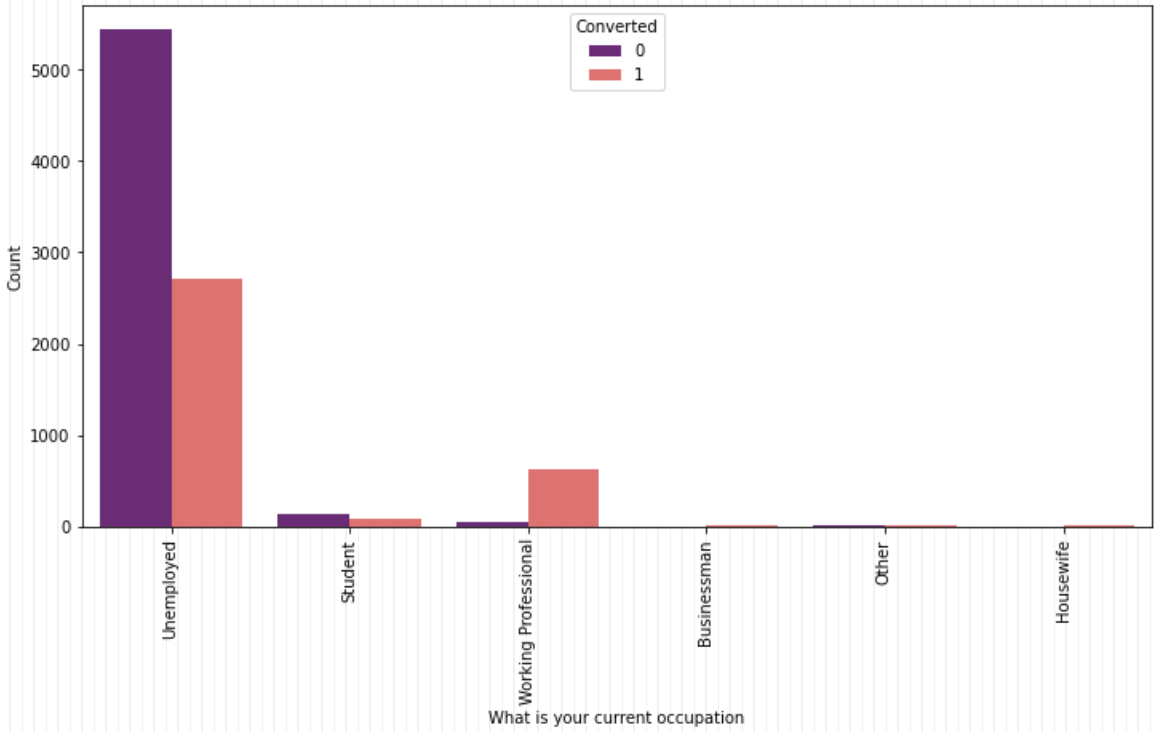


- The source of leads coming from 'Google', ‘Direct traffic' and 'Olark Chat' are more however their conversion rate is low.
- Leads from 'Reference' and 'welingak website' have a very high conversion rate but the count of leads is very low.

UNIVARIATE ANALYSIS: CATEGORICAL VARIABLES

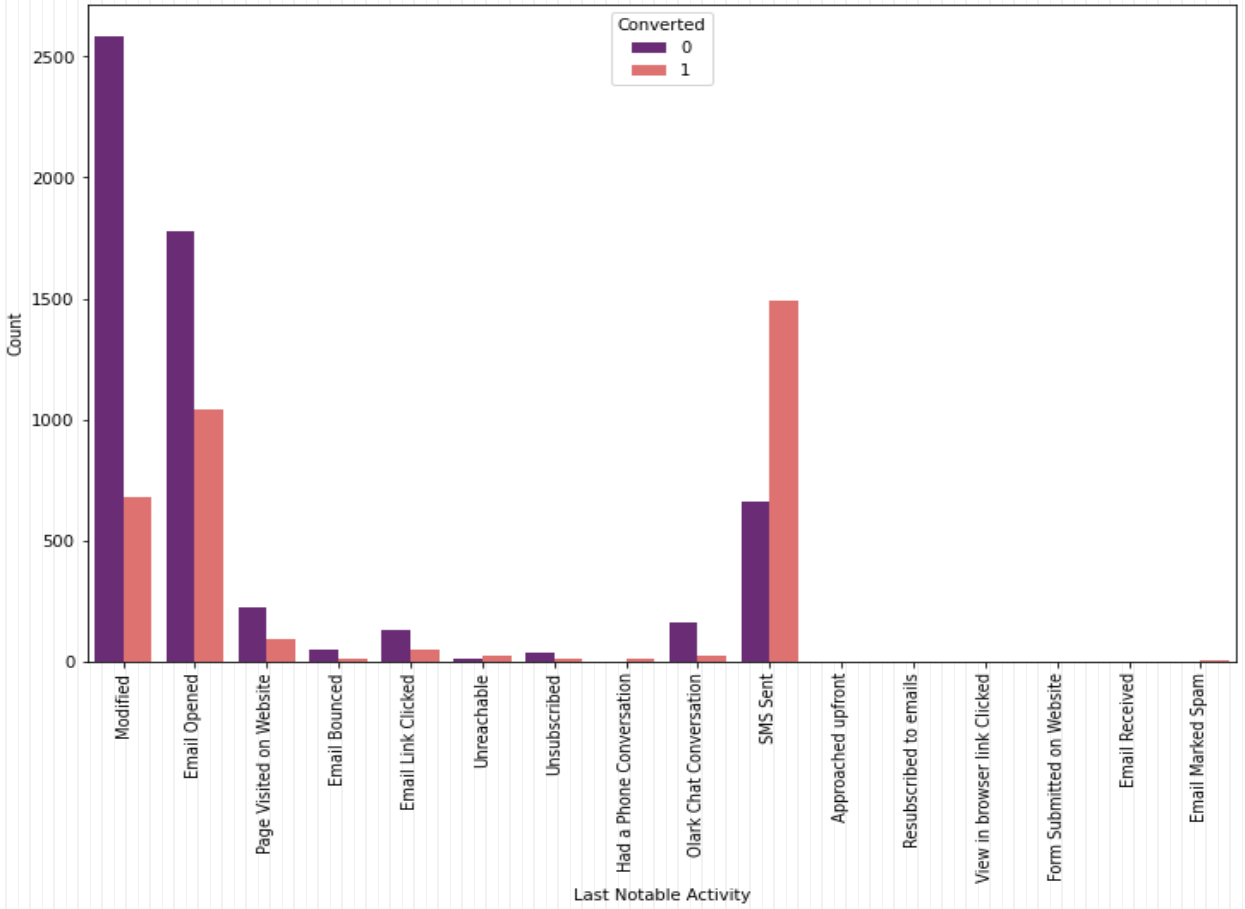


- Most leads have last activity as 'Email Opened' and 'SMS Sent'.
- The 'SMS sent' category has a pretty high conversion rate.

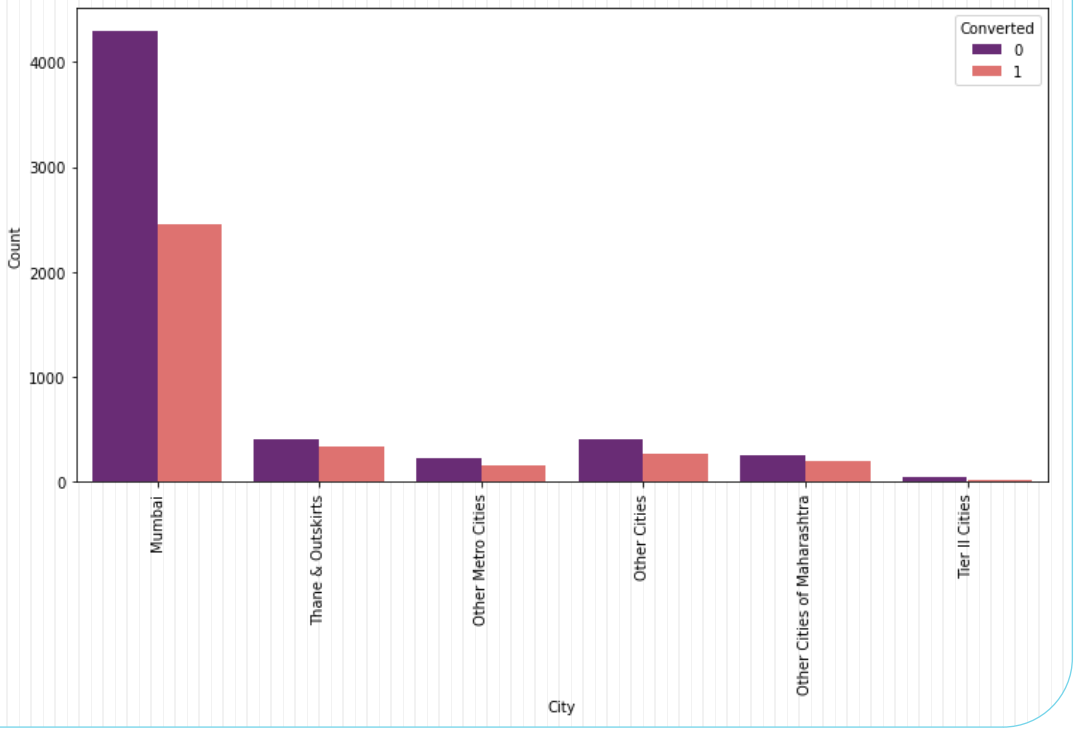


- Most of the leads are 'unemployed' but with a low conversion rate.
- Working professionals have a pretty good conversion rate but the lead count is quite low.

UNIVARIATE ANALYSIS: CATEGORICAL VARIABLES

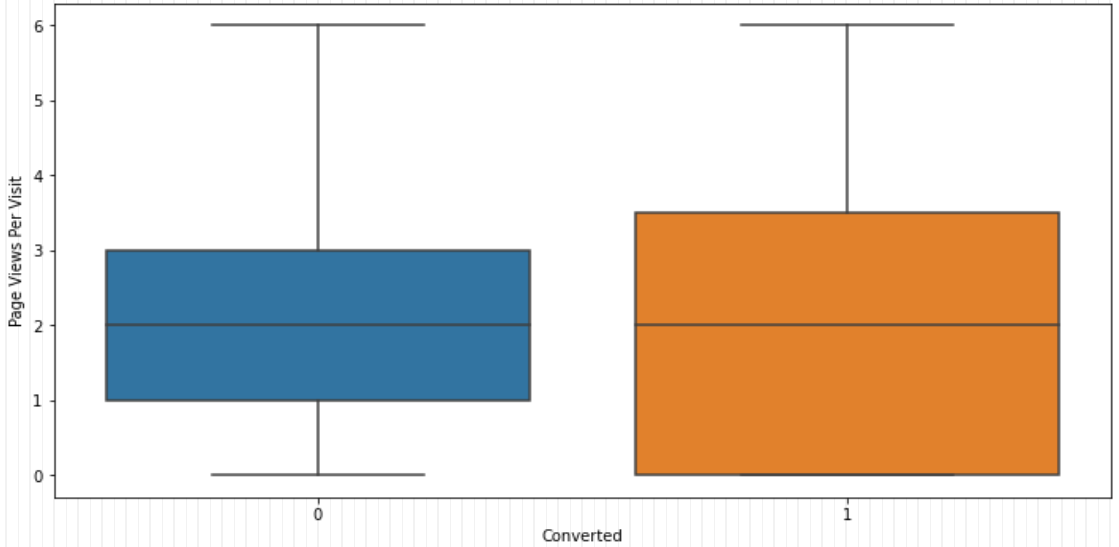


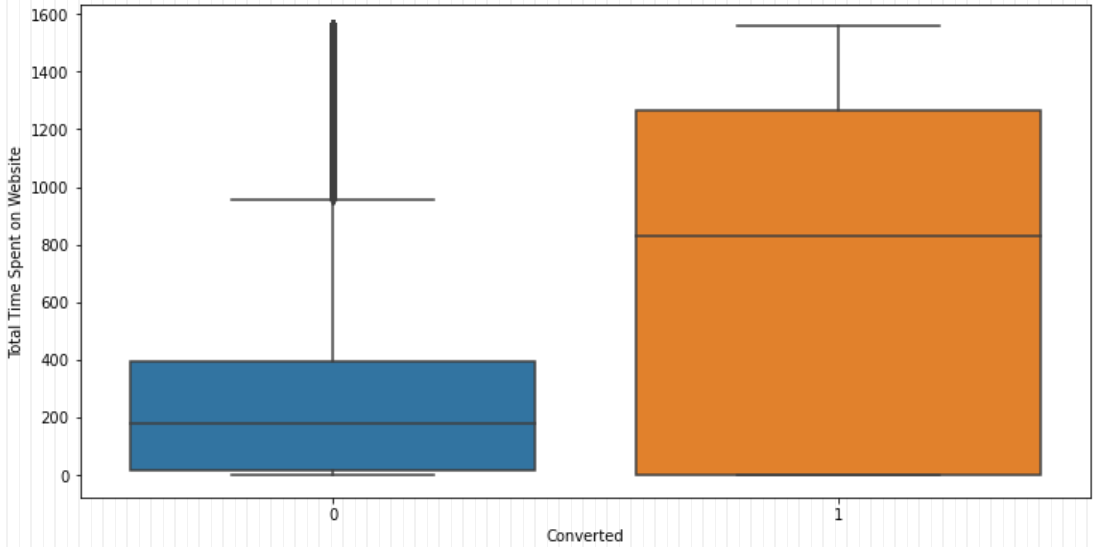
- Students with last notable activity as 'SMS Sent' have a high conversion rate.
- There are many leads with last notable activity from 'Modified' and 'Email Opened' but the conversion rate is low.



- Most of the leads are from 'Mumbai'.
- Leads from 'Mumbai', 'Thane & Outskirts', and from 'other cities of Maharashtra' have a pretty good conversion rate.

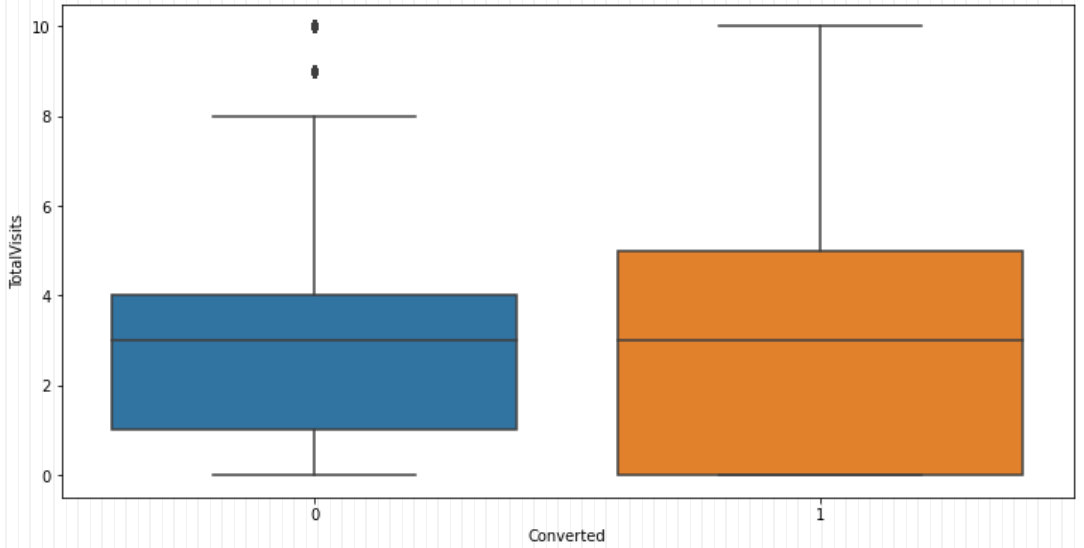
UNIVARIATE ANALYSIS: CONTINUOUS VARIABLES



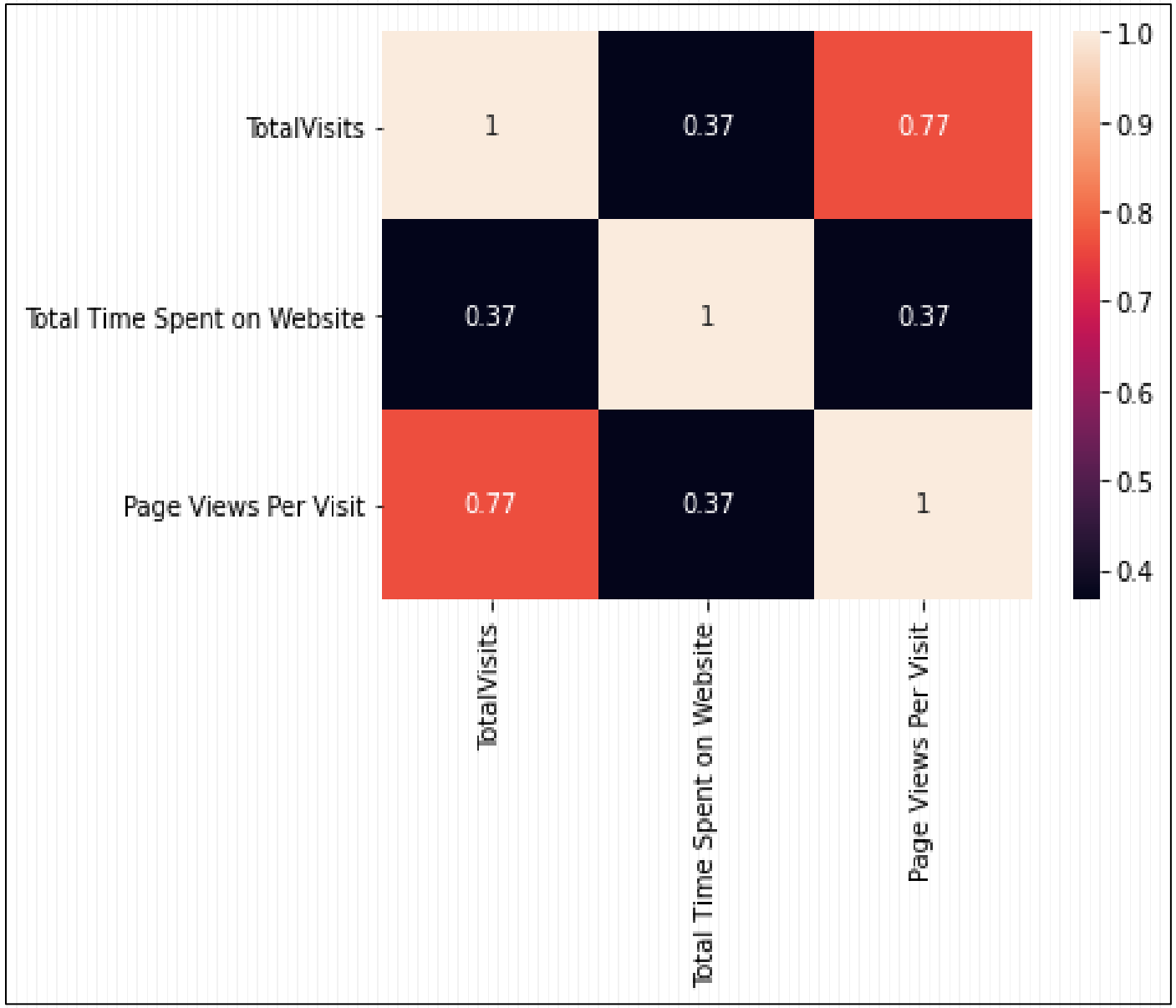


After capping the outliers in 'TotalVisits' and 'Page Views Per Visit' columns to their 95th percentile:

- The median value of 'TotalVisits' and 'Page Views Per Visit' is similar for both converted and non converted leads.
- The converted leads spent more time on the website, as seen by the significant rise in the median and IQR.

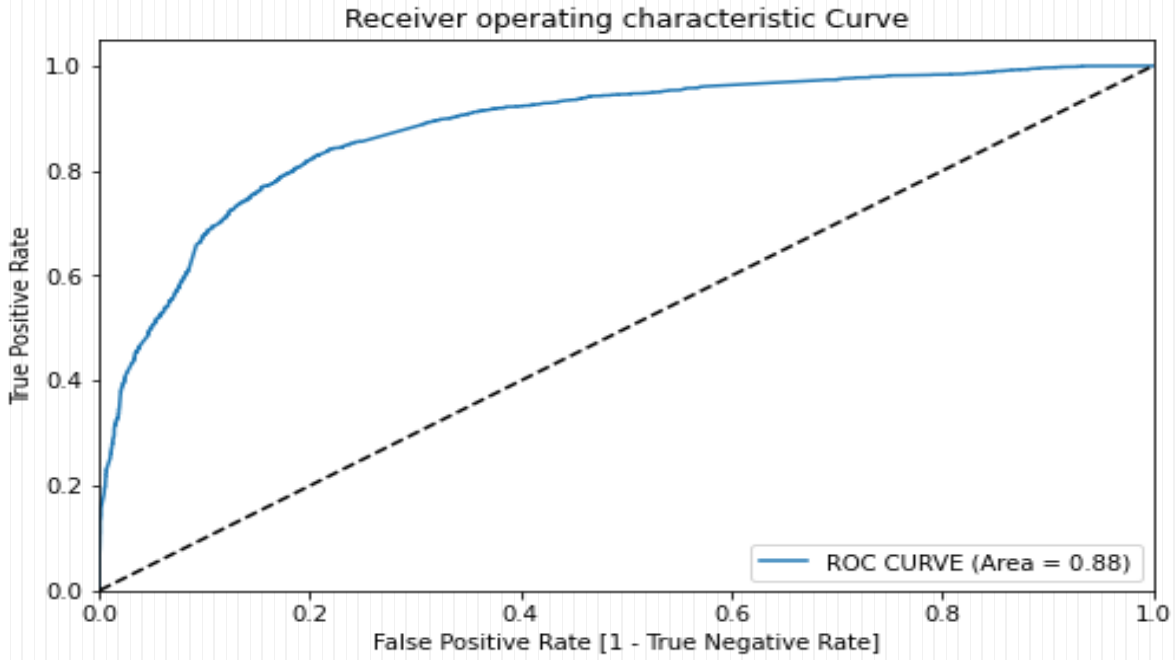


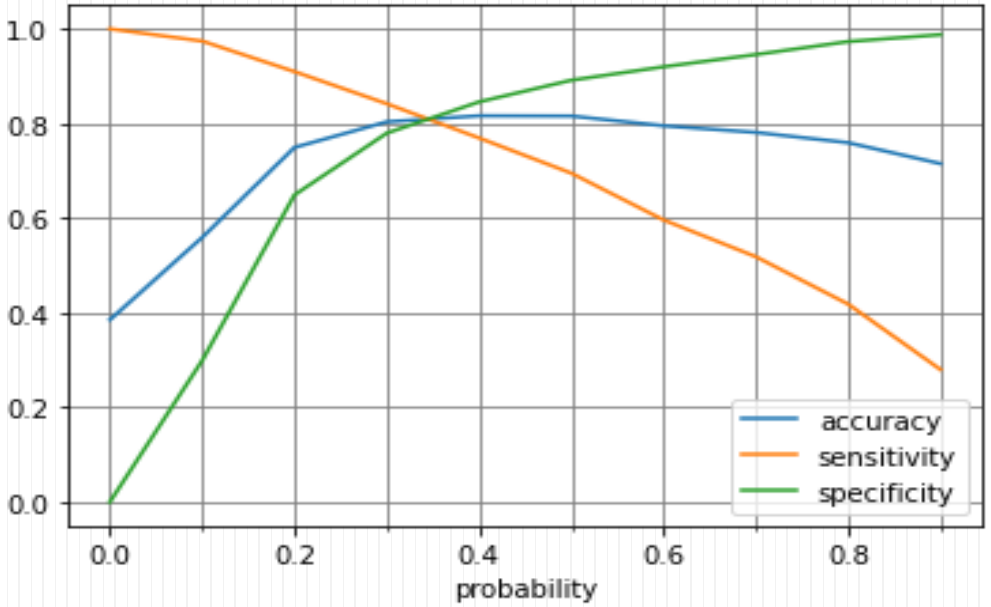
BIVARIATE ANALYSIS



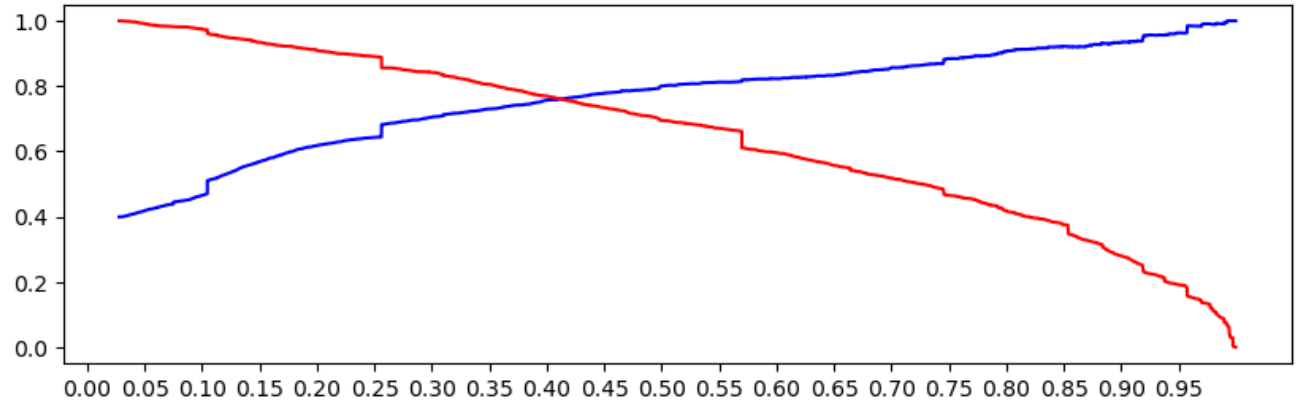
From the given heat map, we can observe that variable "Total visits" is highly correlated with "Page Views Per Visit" with correlation coefficient of 0.77.

MODEL EVALUATION





Precision and recall trade off



Logistic Regression Model Parameters:

- By determining the area under the curve (AUC) of ROC curve, the goodness of the model is defined. The larger the AUC, the better the model. AUC for our model is 0.88.
- Based on accuracy, sensitivity and specificity our optimal threshold value is 0.34
- Based on precision and recall curve trade off we found optimal cut off point as 0.41.

MODEL EVALUATION

EVALUATION METRIC	TRAIN DATASET	TEST DATASET
Accuracy	80.83	80.38
Sensitivity	81.07	79.47
Specificity	80.69	80.91
Precision	80.03	70.36
Recall	69.50	79.47

F1 SCORE
74.39

- Since CEO in particular, has given a ballpark of the target lead conversion rate to be around 80%. Our Logistic Regression Model performs admirably and has a high level of sensitivity, which is exactly what we need.
- The results in the table show that the model is not over-trained and is performing well on the test data set.

CONCLUSION

The X Education company needs to focus on the following factors to improve conversion rate of leads:

- The total time spent on the website impacts the conversion rate.
- Leads from lead origin 'Lead Add Form' have a high conversion rate.
- Leads both sourced from 'Olark Chat' also have decent conversion rate.
- Leads with lead source 'Welingak website' have a high conversion rate.
- Leads with last notable activity as 'SMS sent' have a high conversion rate.
- Leads who are working professionals have a pretty high conversion rate.
- Keep in mind leads with other last notable activities as: modified, page visited on website.