LEAD SCORING CASE STUDY

BY:

Vishal More

Manaswini Pattanaik

Ujjwal Mishra

PROBLEM STATEMENT

To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers.

OBJECTIVES

- To build a Logistic Regression model that assigns lead scores to all leads such as higher conversion and lower conversion according to their lead scores.
- Identify the variables and understand their significance which leads to conversion.
- Summarize the data with its accuracy, sensitivity, specificity and precision.

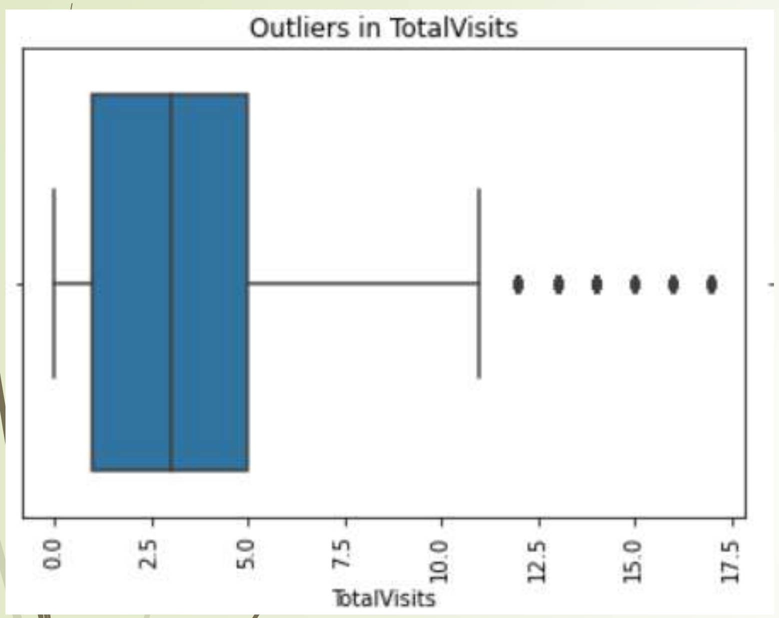
DATA EXPLORATION

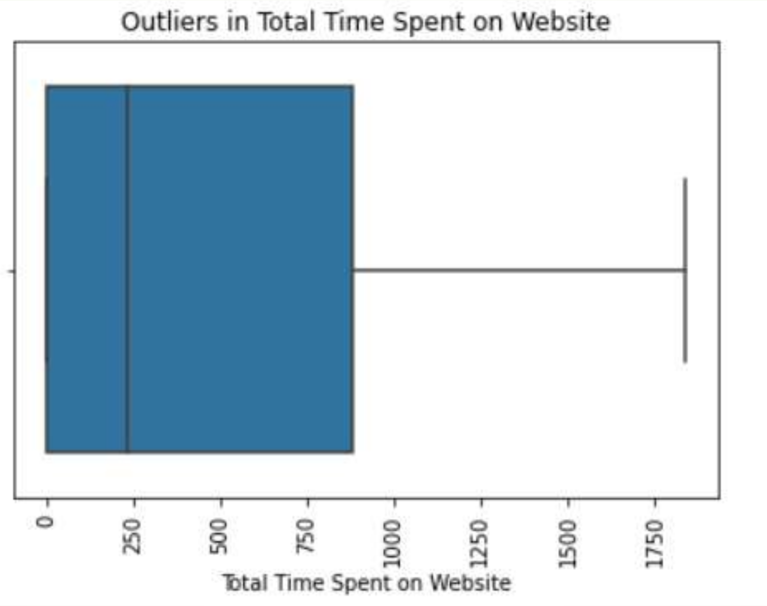
- Importing and Observing the data
- The dataset contains all the information about the leads generated through various sources.
- The dataset contains 9240 rows and 37 columns.
- The dataset has 7 numeric and 30 categorical columns.

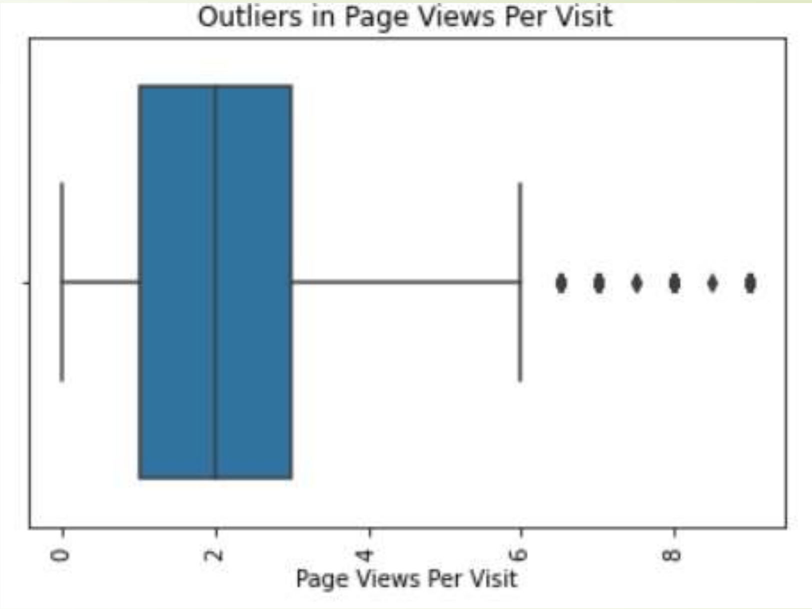
DATA CLEANING

- Checked for missing values and dropped unnecessary columns which have only one unique value.
- The missing values of quantitative columns have been imputed with median as the difference between mean and median is insignificant.
- Clubbing all the similar values into one category.
- The missing values of categorical columns have been imputed with ‘NA’

DATA VISUALIZATION



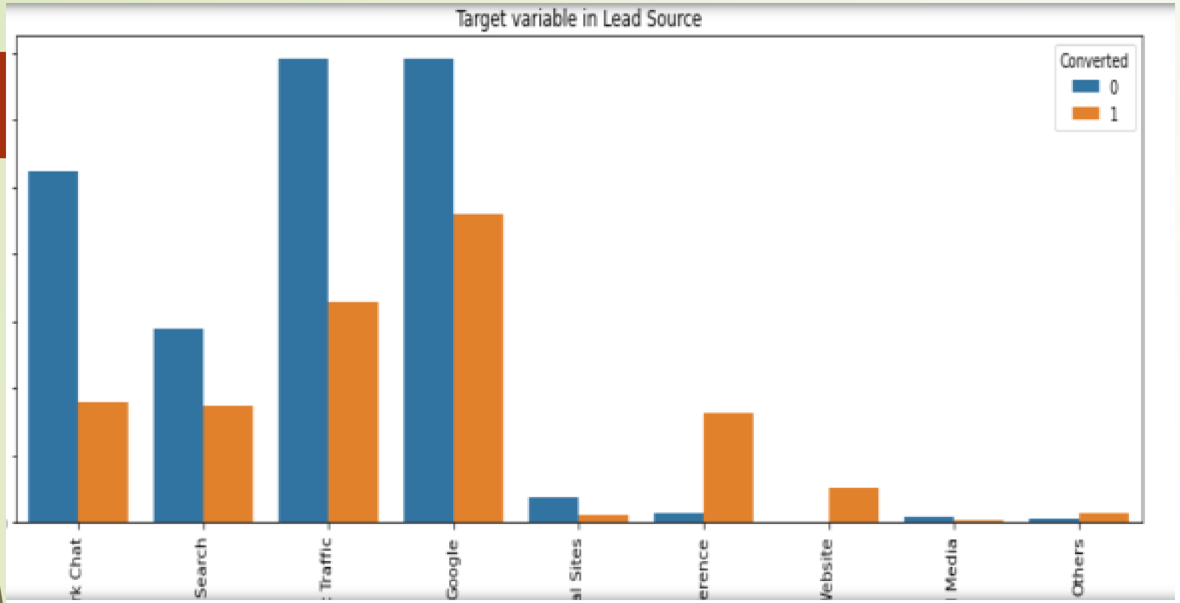




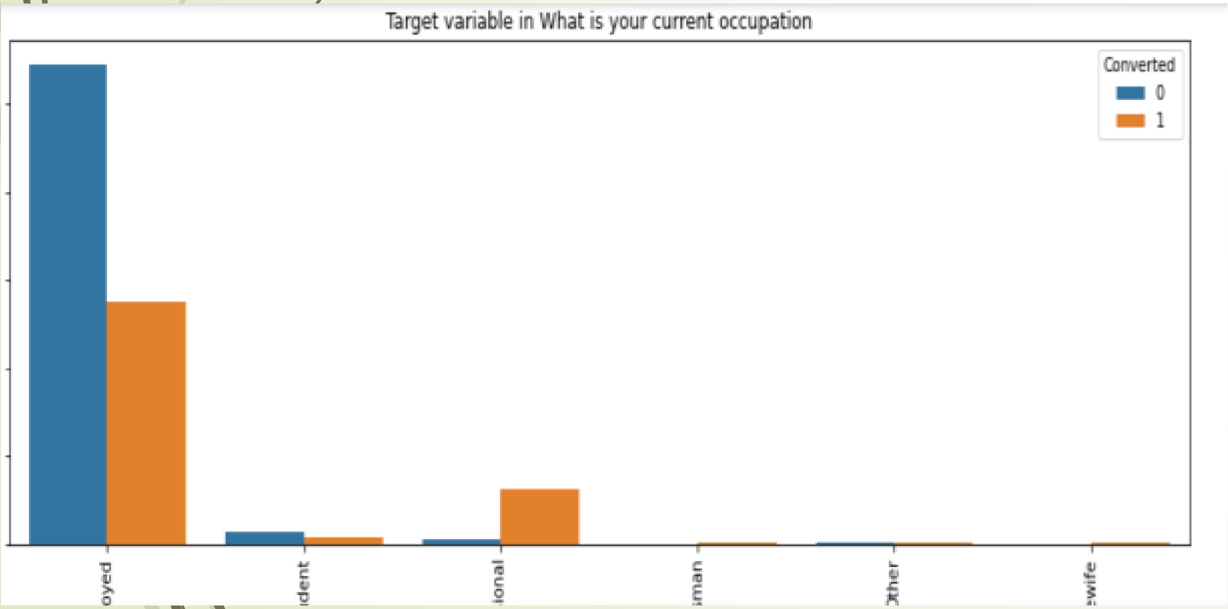
With the help of boxplot it is observed that TotalVisits', 'Total Time Spent on Website' and 'Page Views Per Visit' has outliers.

Inter Quantile Range (IQR) method has been used to treat outliers in the data.

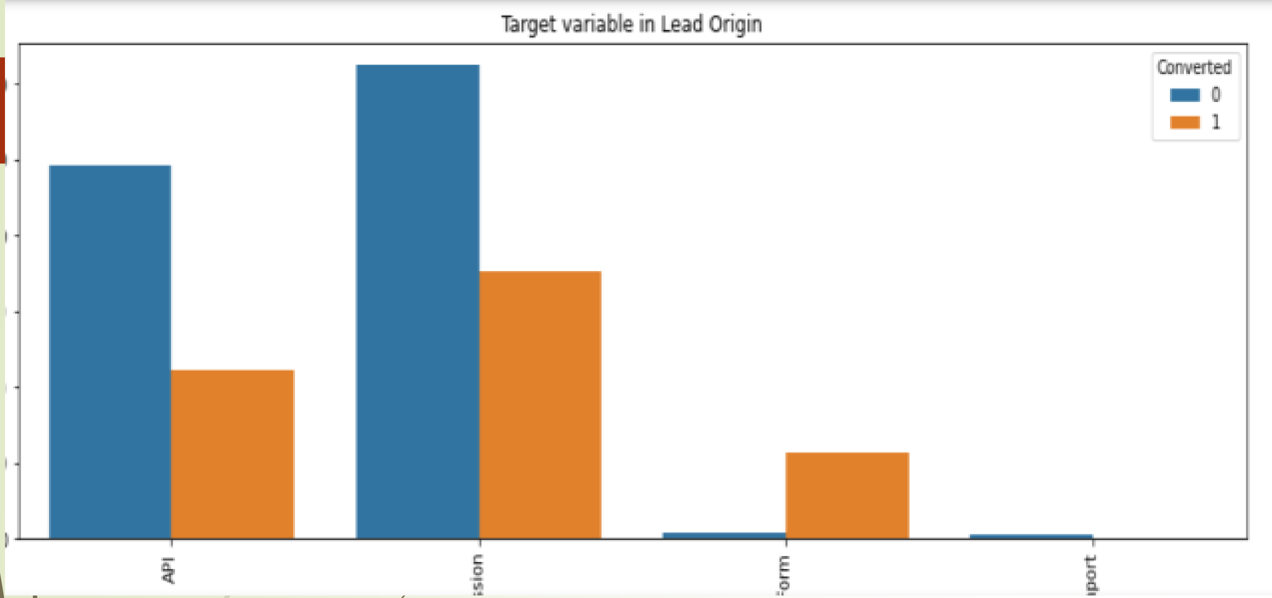
It is observed that people spending more time in website are likely to get converted.

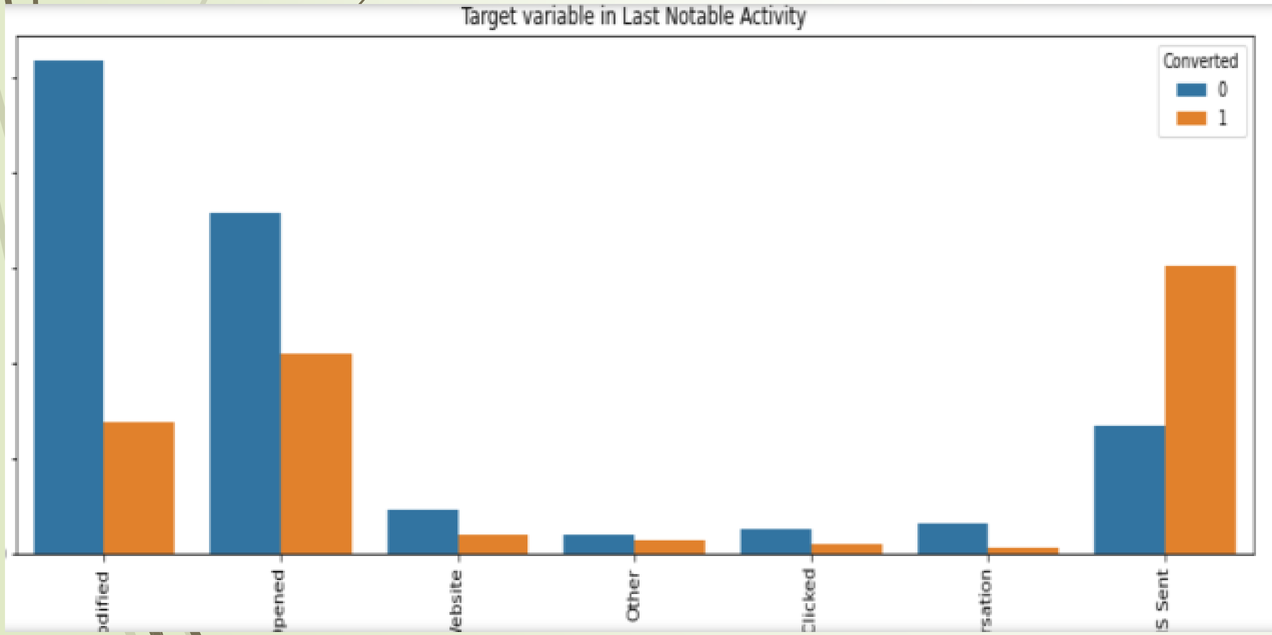


- From the graph it is observed that the conversion rates are high in ‘Reference’ and ‘Welingak Website’.
- Most leads are generated through ‘Direct Traffic’ and ‘Google’.



- Working Professionals are most likely to get converted.
- However most number of leads are generated for unemployed people category.





‘API’ and ‘Landing Page Submission’ generate the most leads but have less conversion rates,but ‘Lead Add Form’ generates less leads with high conversion rate.

- Highest conversion rate is for the last notable activity ‘SMS Sent’.
- ‘Email Opened’ also has high conversion rate.

DATA PREPARATION

- Converted binary variables like Yes/No into 1 and 0 format.
- Created dummy variables for the categorical variables.
- Dropped the variables which were converted into Dummies.
- Spliting the dataset into train and test.
- Scaling the dataset using the Standard Scaler function.

MODEL BUILDING

- Built the model using GLM for the detailed statistics.
- To get the selected features, used RFE technique.
- Making Predictions on the Train Set.
- Through train dataset the predicted values are:
- Accuracy- 92%
- Precision - 92%
- Recall - 86%

- Making predictions on the test dataset.
- Through test dataset the predicted values are :
- Accuracy - 92%.
- Precision - 90%
- Recall - 91%.

CONCLUSION

- Three variables which contribute most towards the probability of a lead conversion are :
- Lead Origin
- Total Time Spent on Website
- Last Activity

- Top three categories that contribute to decision are :
- Lead Origin ==> Lead Add Form
- Total Time Spent on Website ==> The total time spent by the customer on the website.
- Last Activity ==> SMS Sent