LEAD SCORING CASE STUDY

PREPARED BY – AKSHAY NALAWADE, YESHWANTH KESANI, RAHUL DANISH

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.

THE COMPANY MARKETS ITS COURSES ON SEVERAL WEBSITES AND SEARCH ENGINES LIKE GOOGLE. ONCE THESE PEOPLE LAND ON THE WEBSITE, THEY MIGHT BROWSE THE COURSES OR FILL UP A FORM FOR THE COURSE OR WATCH SOME VIDEOS. WHEN THESE PEOPLE FILL UP A FORM PROVIDING THEIR EMAIL ADDRESS OR PHONE NUMBER, THEY ARE CLASSIFIED TO BE A LEAD. MOREOVER, THE COMPANY ALSO GETS LEADS THROUGH PAST REFERRALS. ONCE THESE LEADS ARE ACQUIRED, EMPLOYEES FROM THE SALES TEAM START MAKING CALLS, WRITING EMAILS, ETC. THROUGH THIS PROCESS, SOME OF THE LEADS GET CONVERTED WHILE MOST DO NOT. THE TYPICAL LEAD CONVERSION RATE AT X EDUCATION IS AROUND 30%.

NOW, ALTHOUGH 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 Goal

- 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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Approach/Strategy

- ▶ Importing the data.
- ▶ Understanding and cleaning the data for analysis.
- ▶ Performing EDA to figure out most helpful attributes.
- ▶ Feature Scaling.
- ▶ Splitting the data into train and test.
- ▶ Data Modelling and building a logistic regression model.
- ▶ Assigning a lead score for each leads.
- ▶ Test the model on train and test set.
- ▶ Measuring the accuracy of the model for evaluation.

Exploratory data analysis:
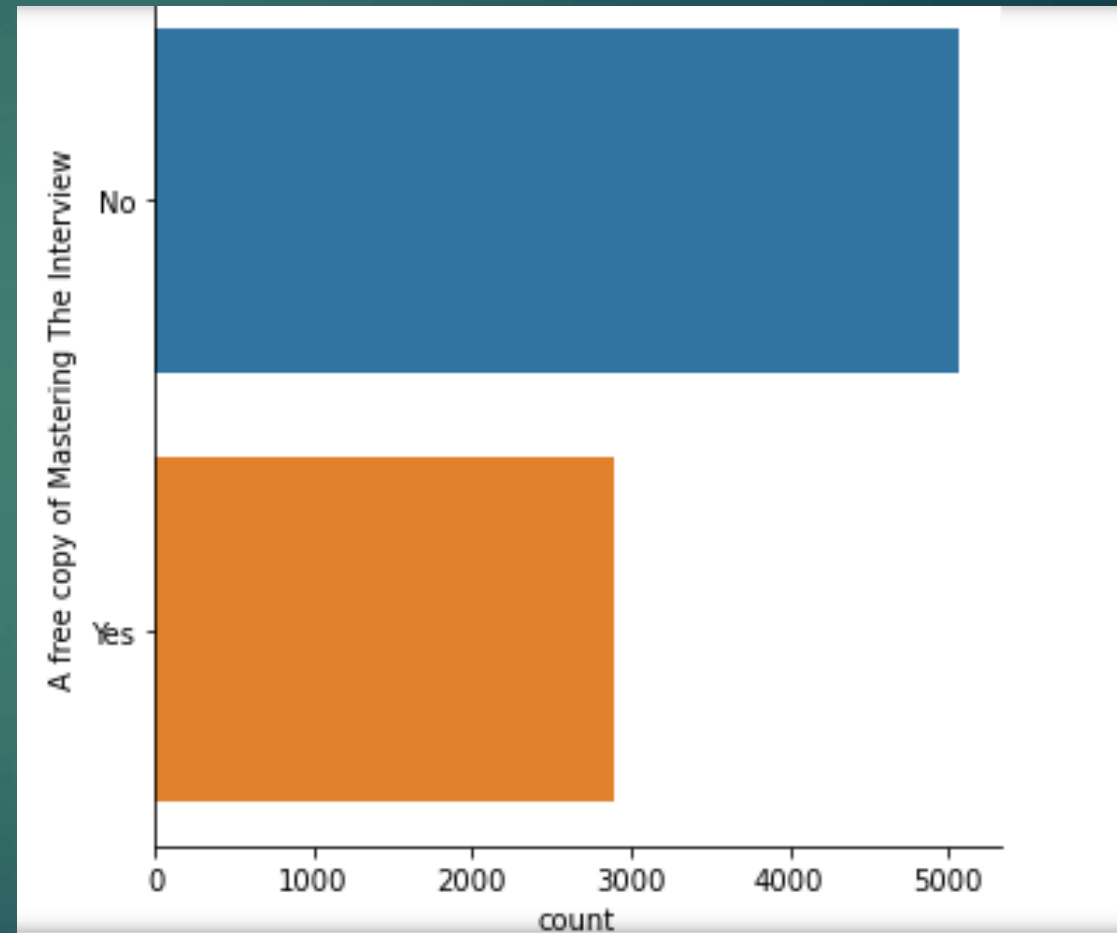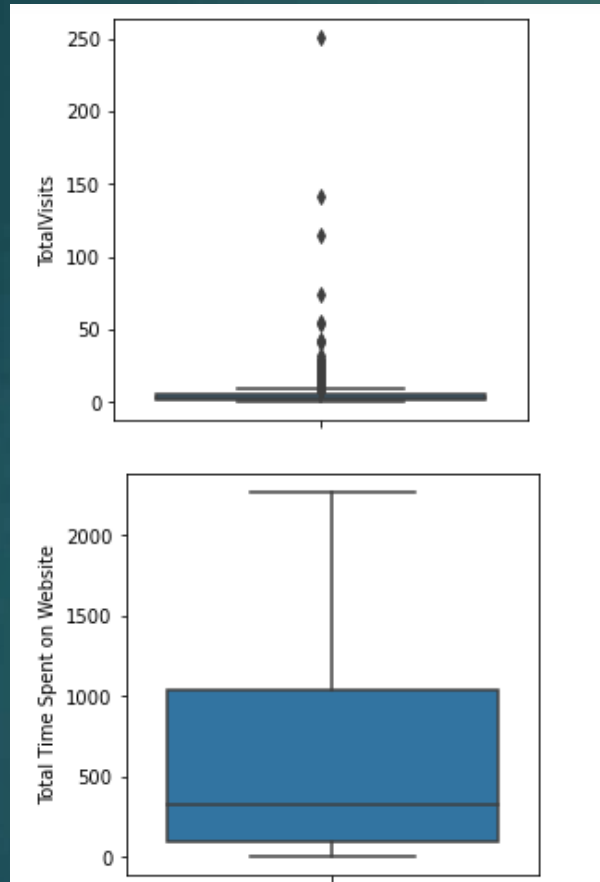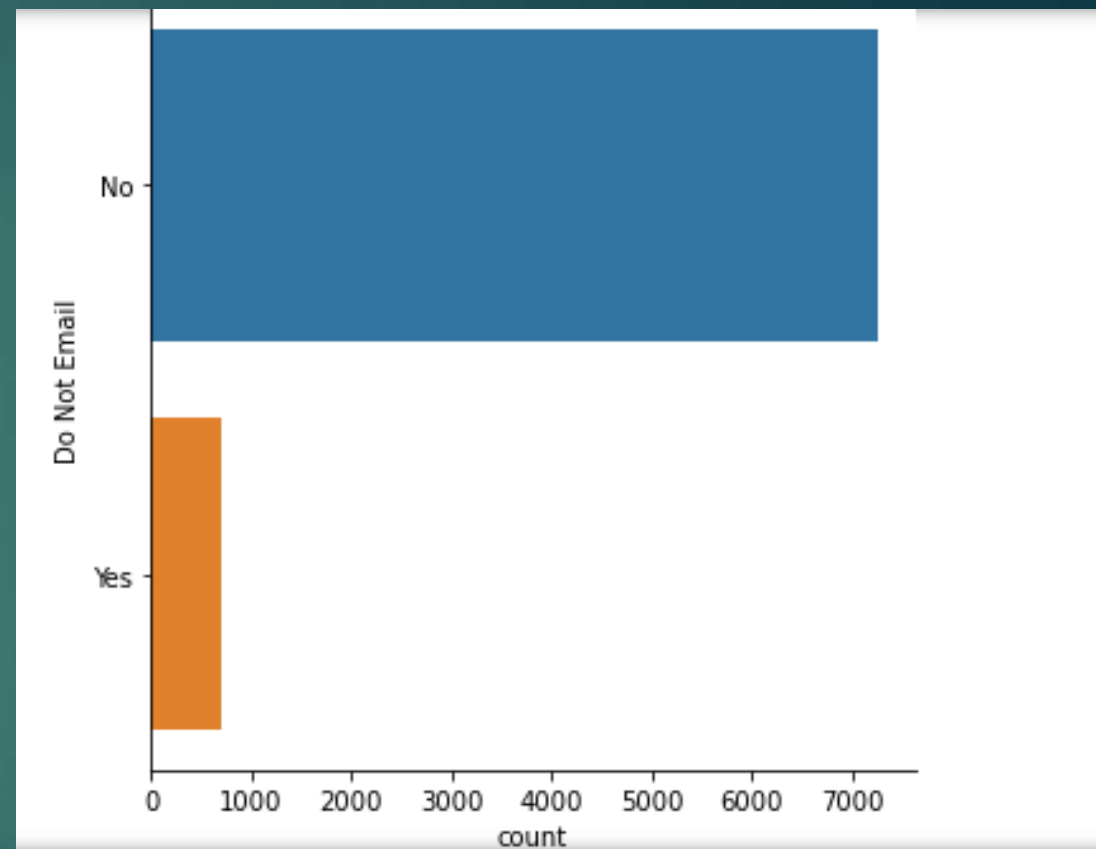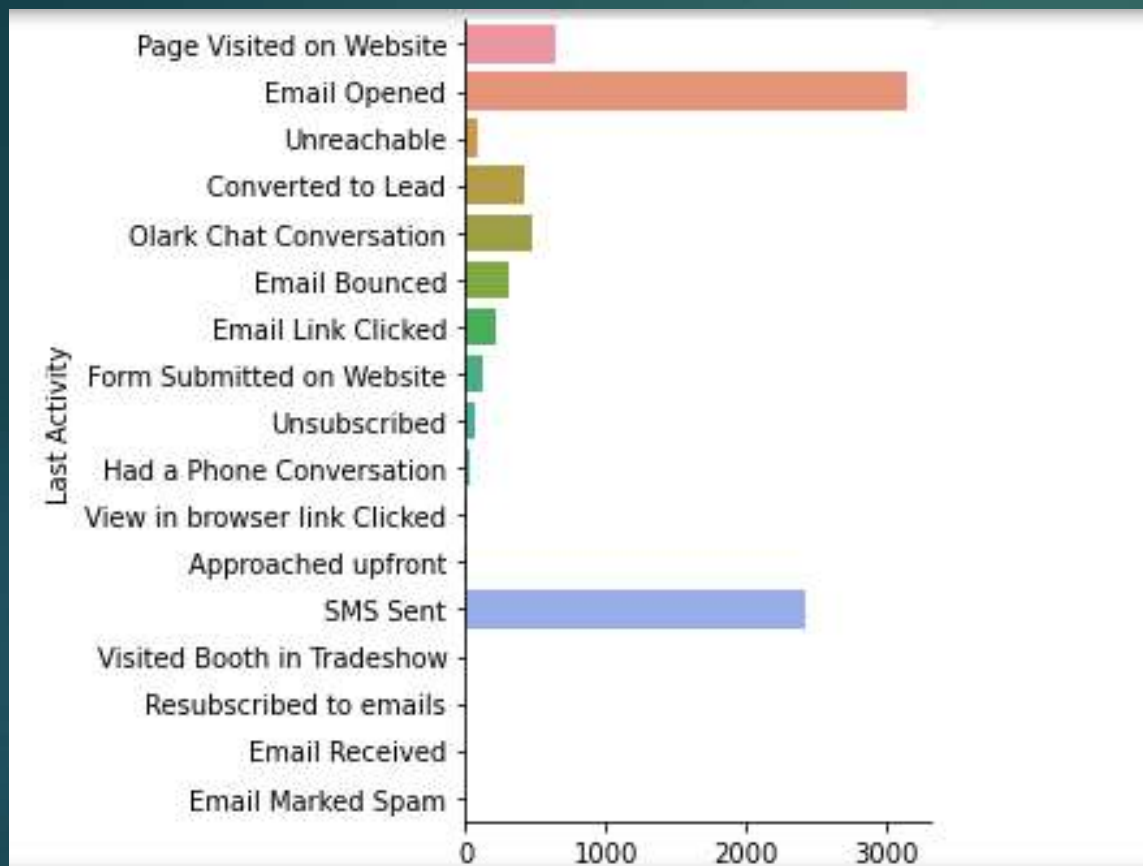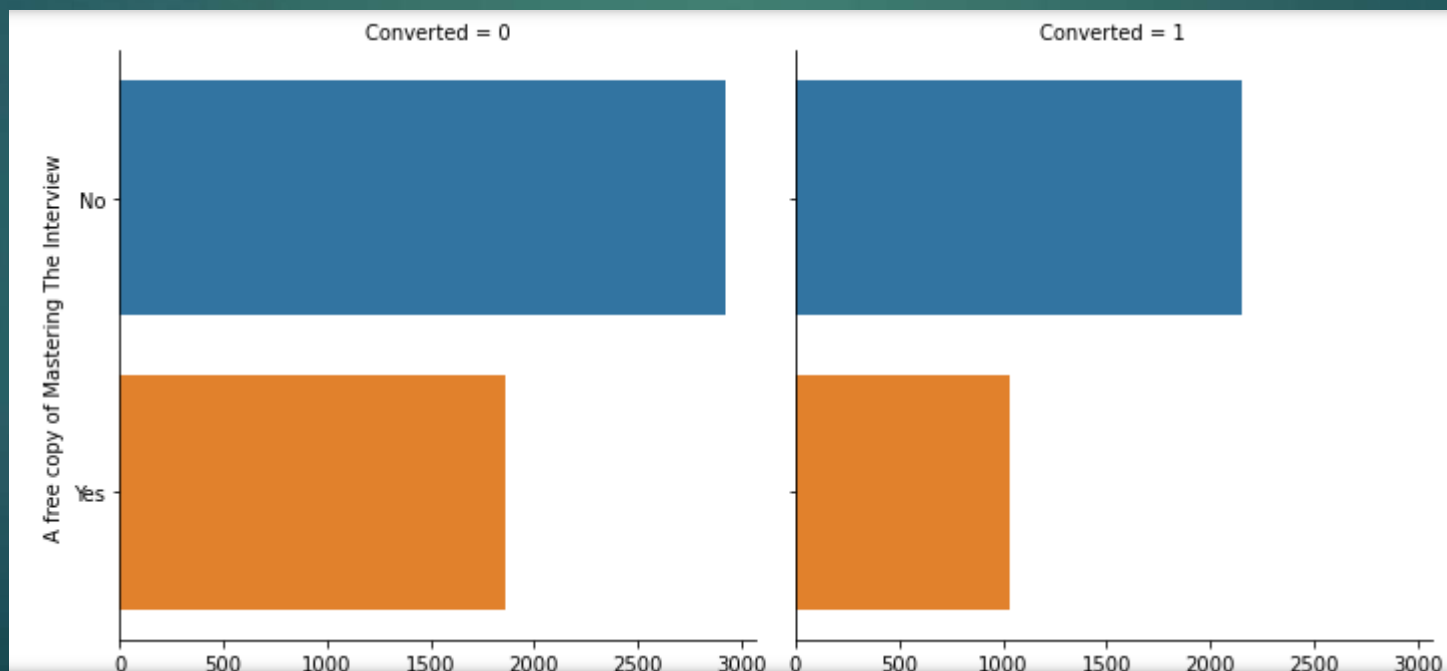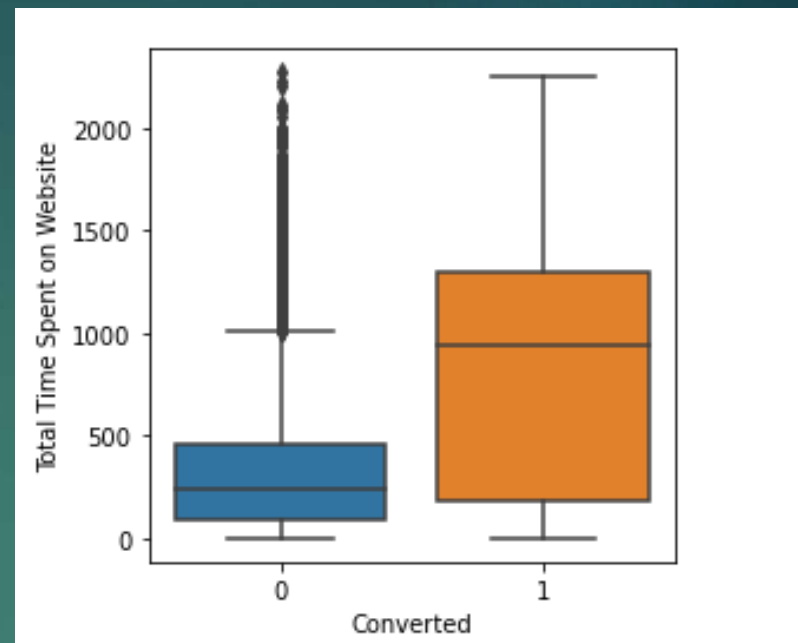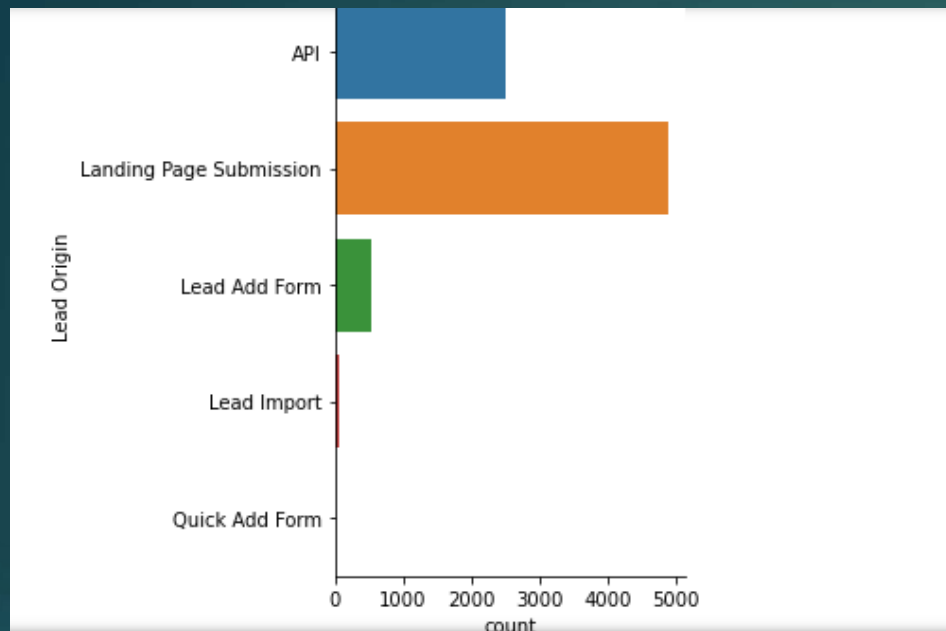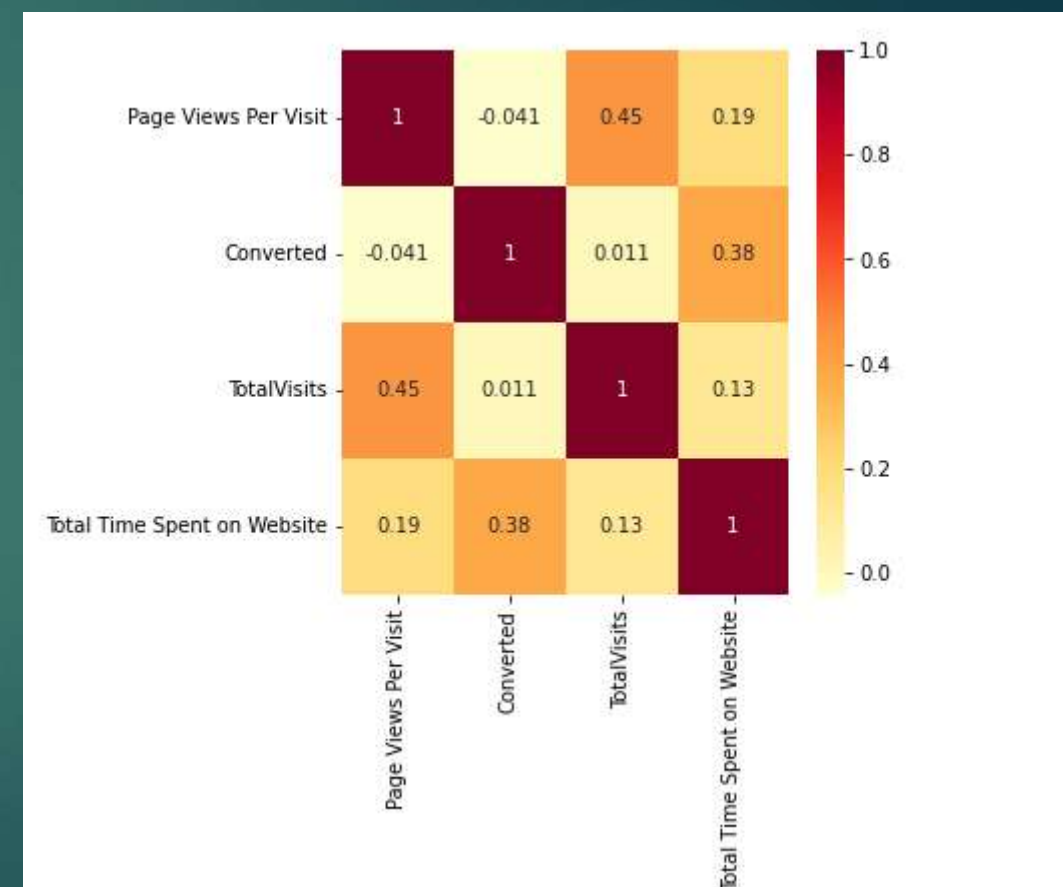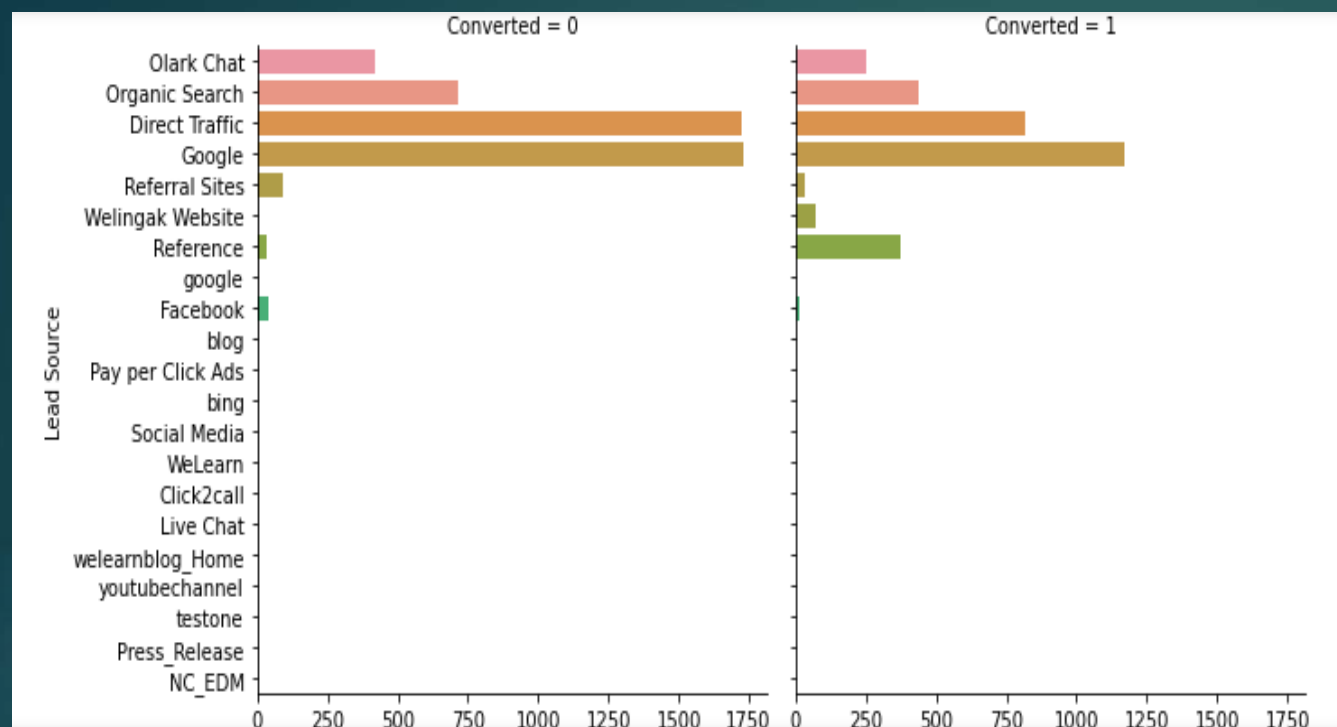
Model Building

- ▶ Splitting the data into train and test data.
- ▶ Scaling the variables in train set.
- ▶ Build the first model.
- ▶ Using RFE to eliminate less relevant variables.
- ▶ Check VIF value for existing columns.
- ▶ Predict using train set.
- ▶ Evaluate accuracy and other metrics.
- ▶ Precision and recall analysis on test predictions.

Model Evaluation (TRAIN)
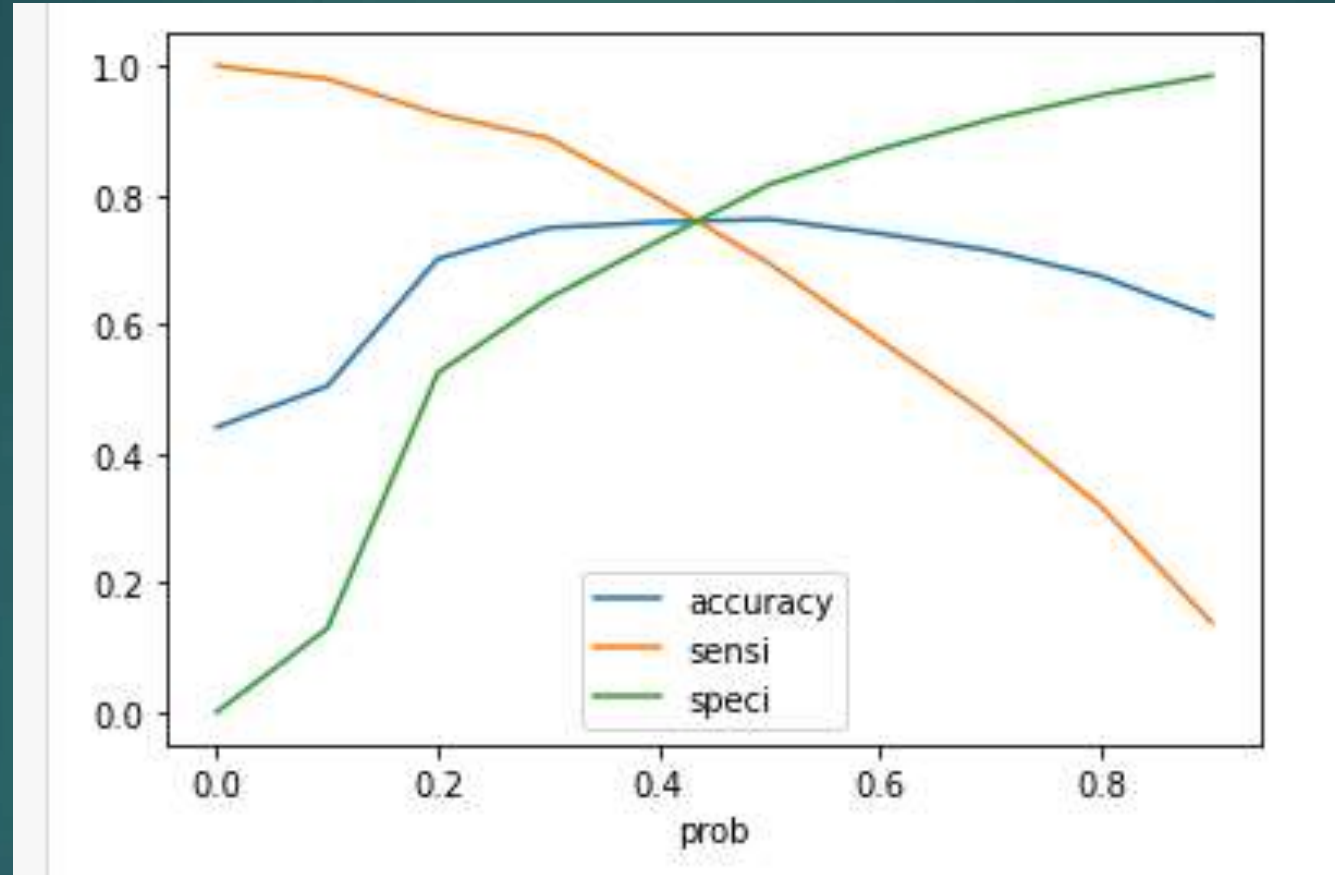
With the current cut off as 0.42 we have accuracy of 76 %, sensitivity at 84% and specificity of around 68%

Model Evaluation (TEST)
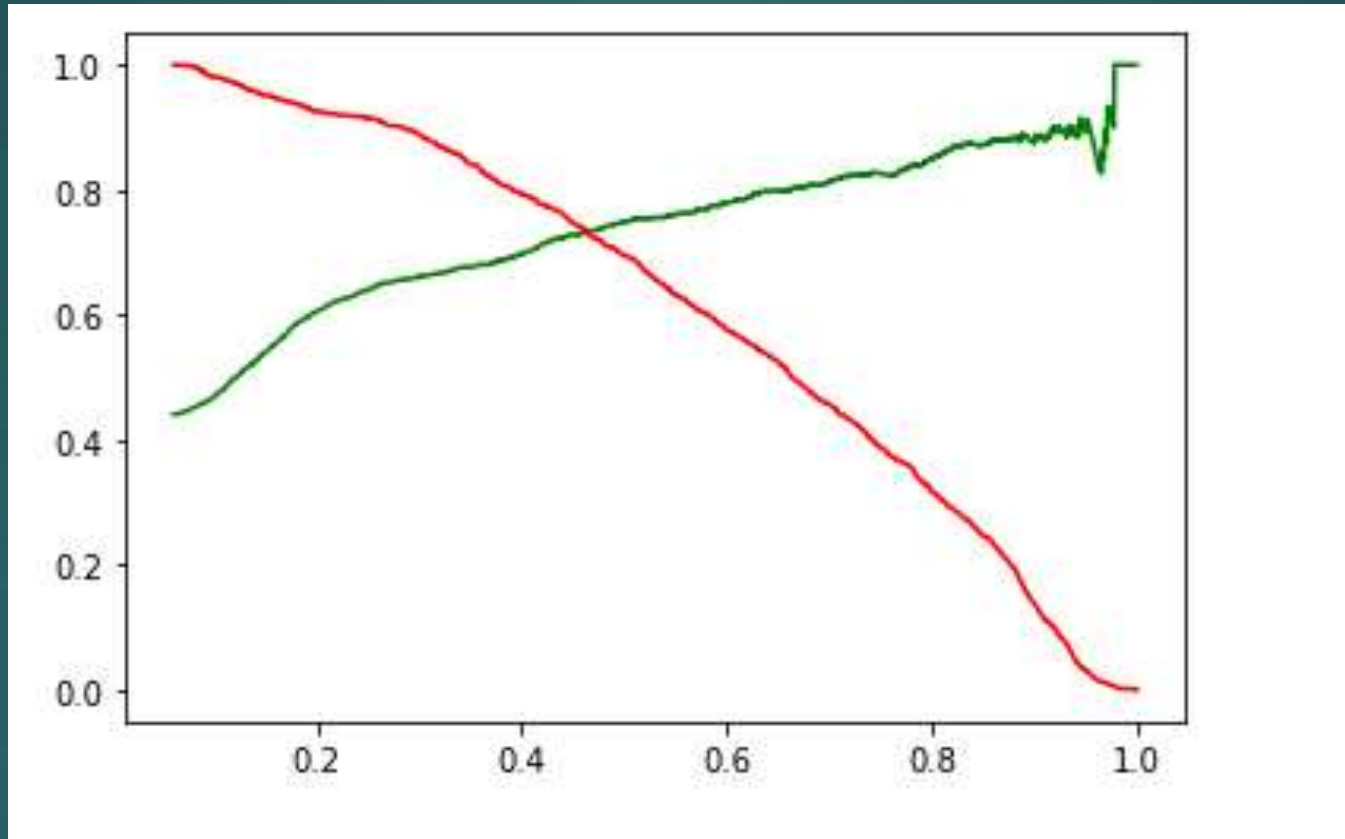
With the current cut off as 0.45 we have Precision around 70% and Recall around 78% and accuracy of 76 %.

Conclusion:

➢ *People spending higher than average time are promising leads hence targeting them would be helpful in conversions.*

➢ *Landing page submissions can help out find more leads.*

➢ *An alert message can have high lead conversion rate.*

➢ *The logistic regression model shows accuracy up-to 75 to 78 %*

➢ *The Threshold has been selected from accuracy, sensitivity, precision and recall curves.*

➢ *The model shows 80% sensitivity and 70% specificity*

➢ *The model finds correct promising leads and tends to have high chances of getting converted.*

➢ *Overall this model proves to be accurate.*