LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

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. They have process of form filling on their website after which the company that individual as a lead.
- 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, this means 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

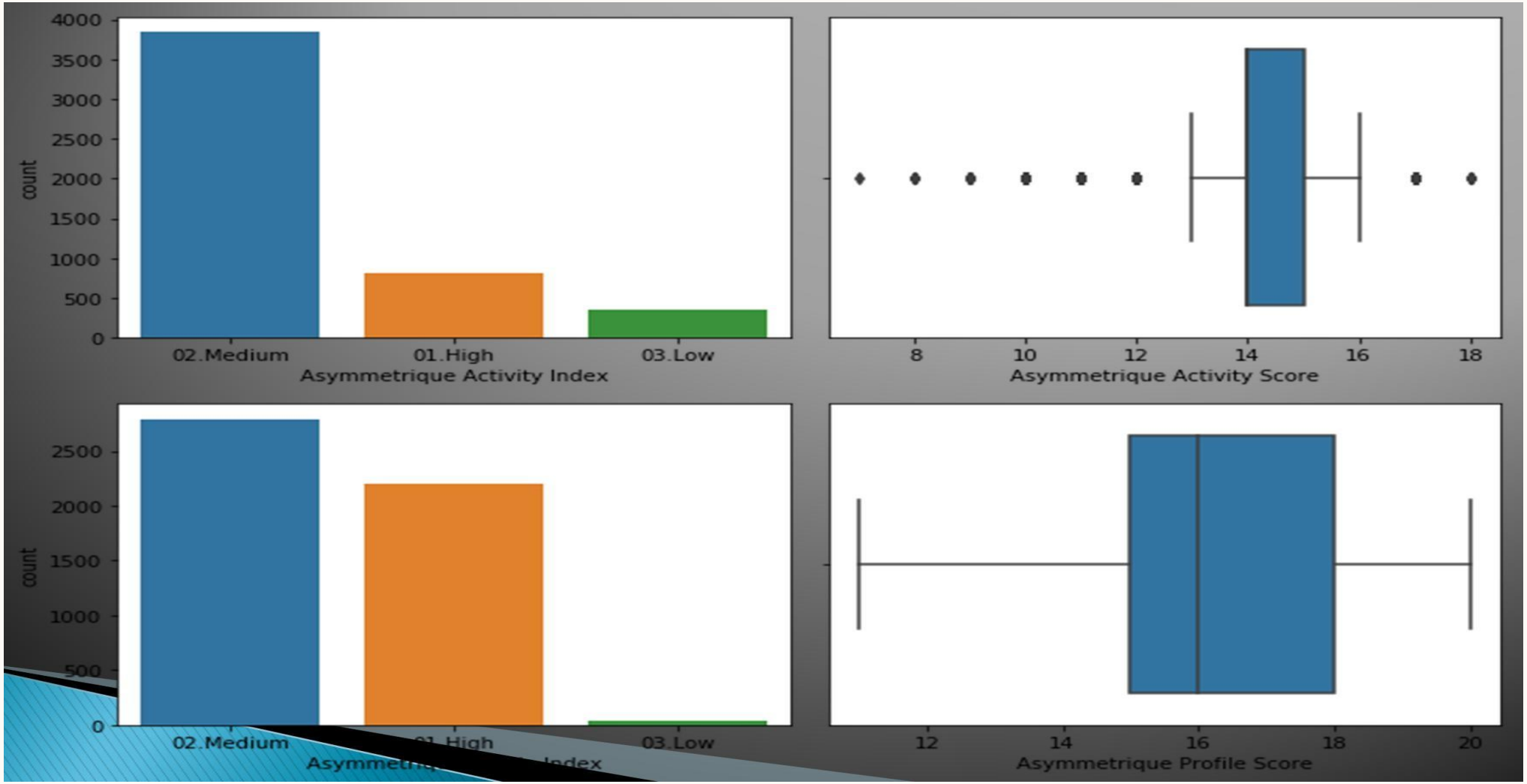
BUSINESS OBJECTIVE

- Lead X wants us to build a model to give every lead a lead score between 0 - 100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

PROBLEM APPROACH

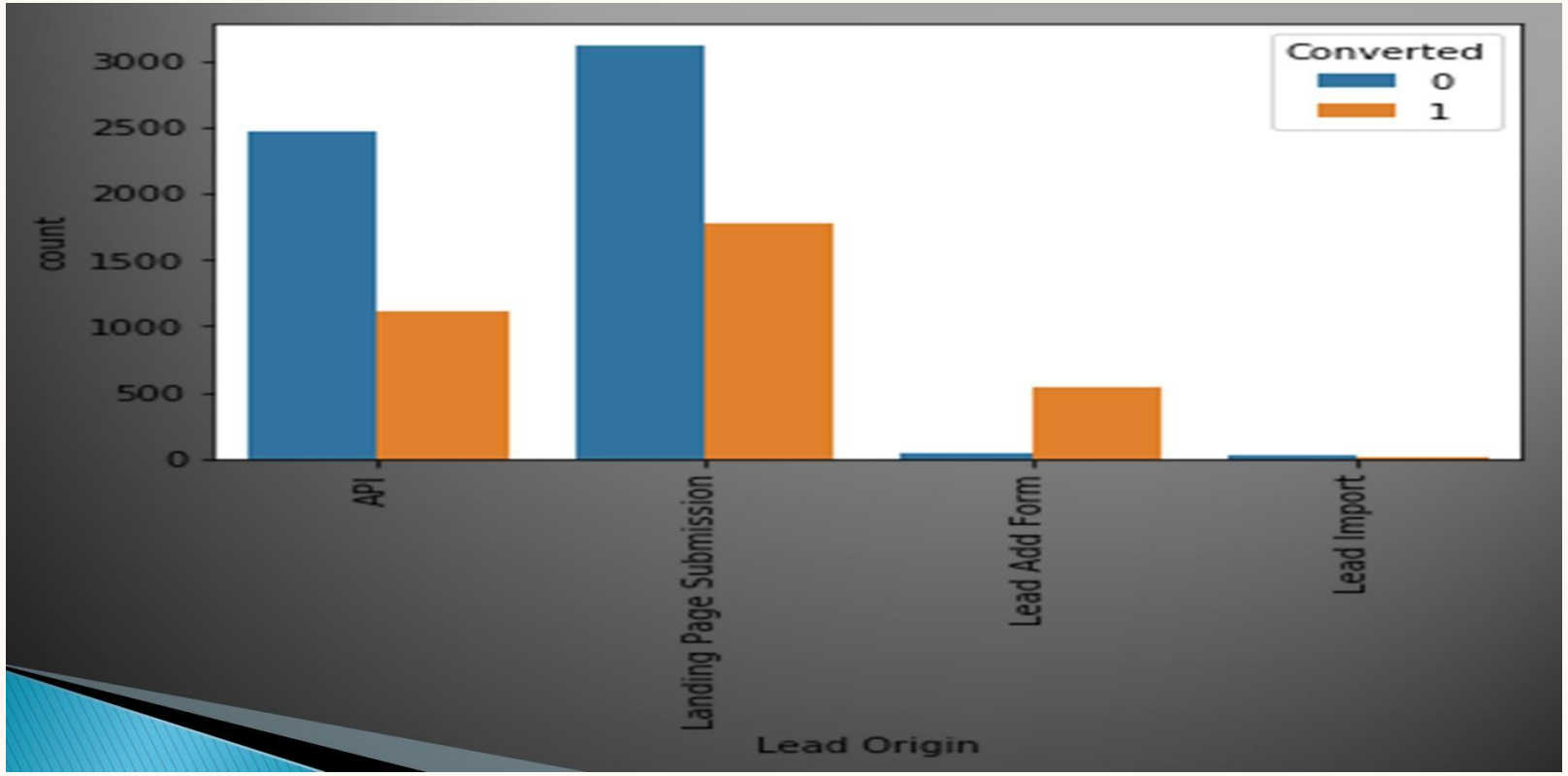
- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Assign a lead score for each leads
- Test the model on train set
- Evaluate the model by different measures and metrics
  Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation

- An index and score assigned to each customer based on their activity and their profile.
- There is too much variation in this parameters so its not reliable to impute any value in it.



LEAD ORIGIN VS CONVERTED

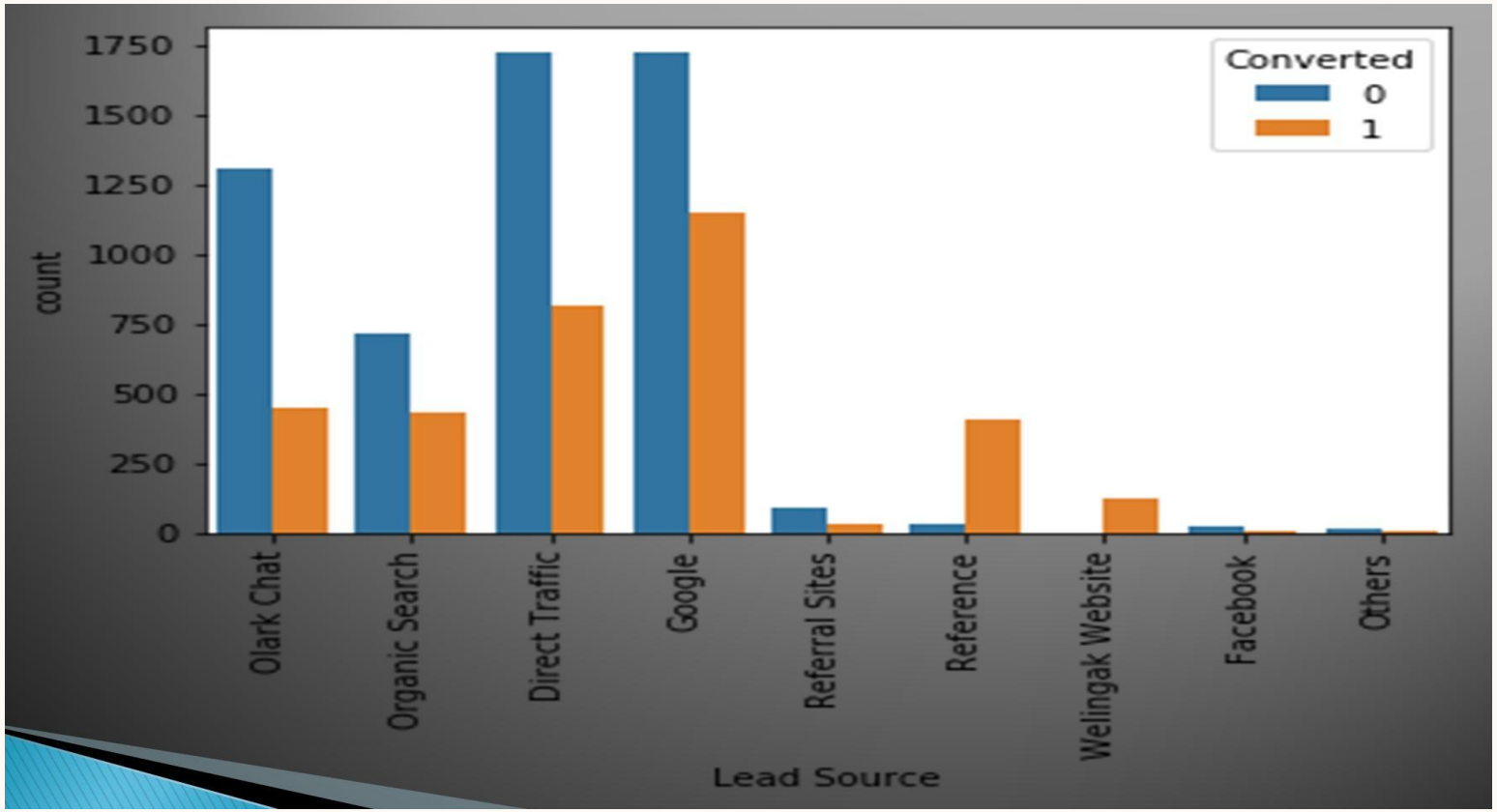
- API and landing page submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add form has more than 90% conversion rate but count of lead are not very high.
- Lead import are very less in count.



LEAD SOURCE VS CONVERTED

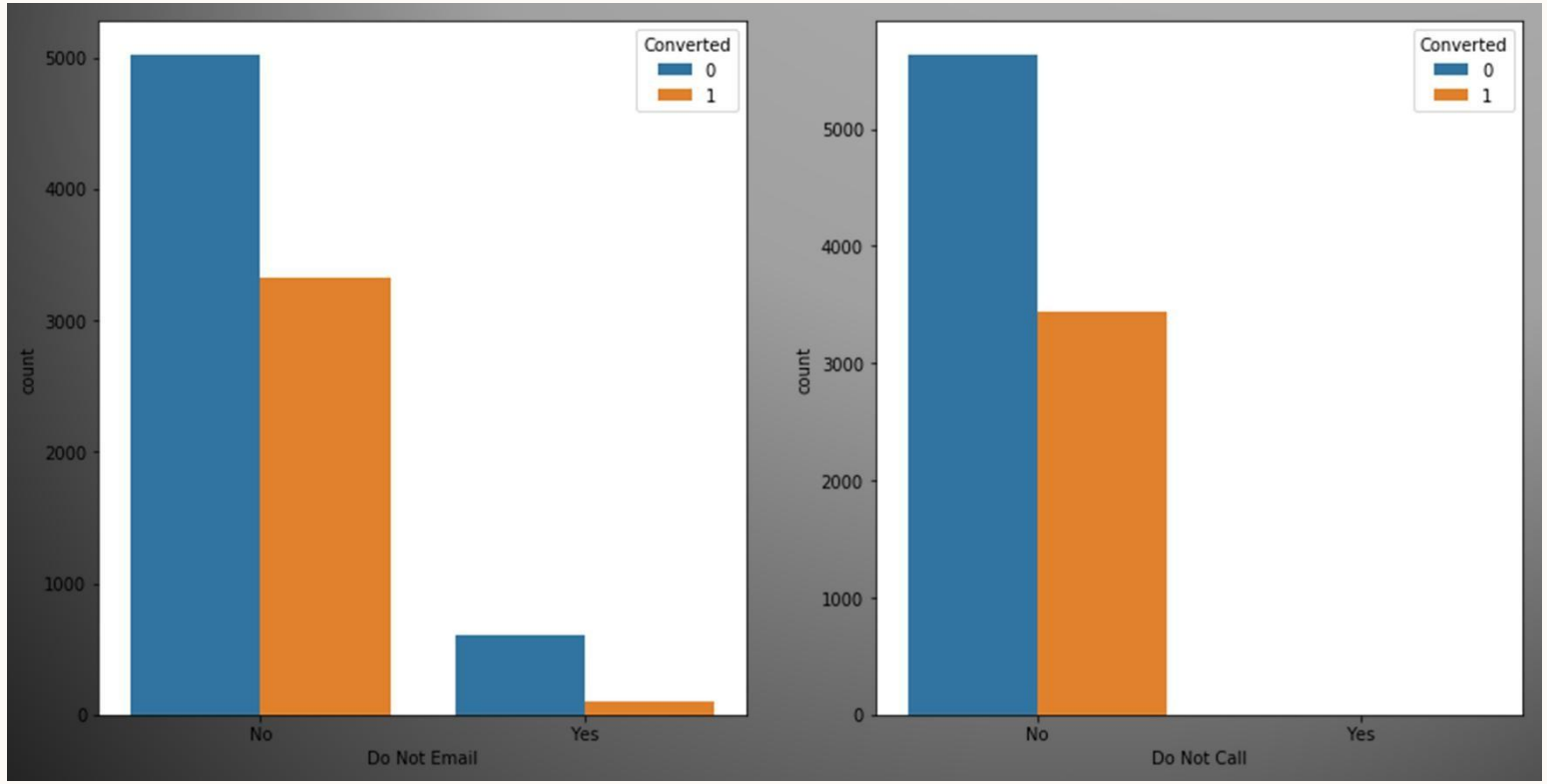
- Google and direct traffic generates maximum number of leads.
- Conversion rate of reference leads and leads through welingak website is high.

To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic and google leads and generate more leads from reference and welingak website.



DO NOT EMAIL & DO NOT CALL VS CONVERTED

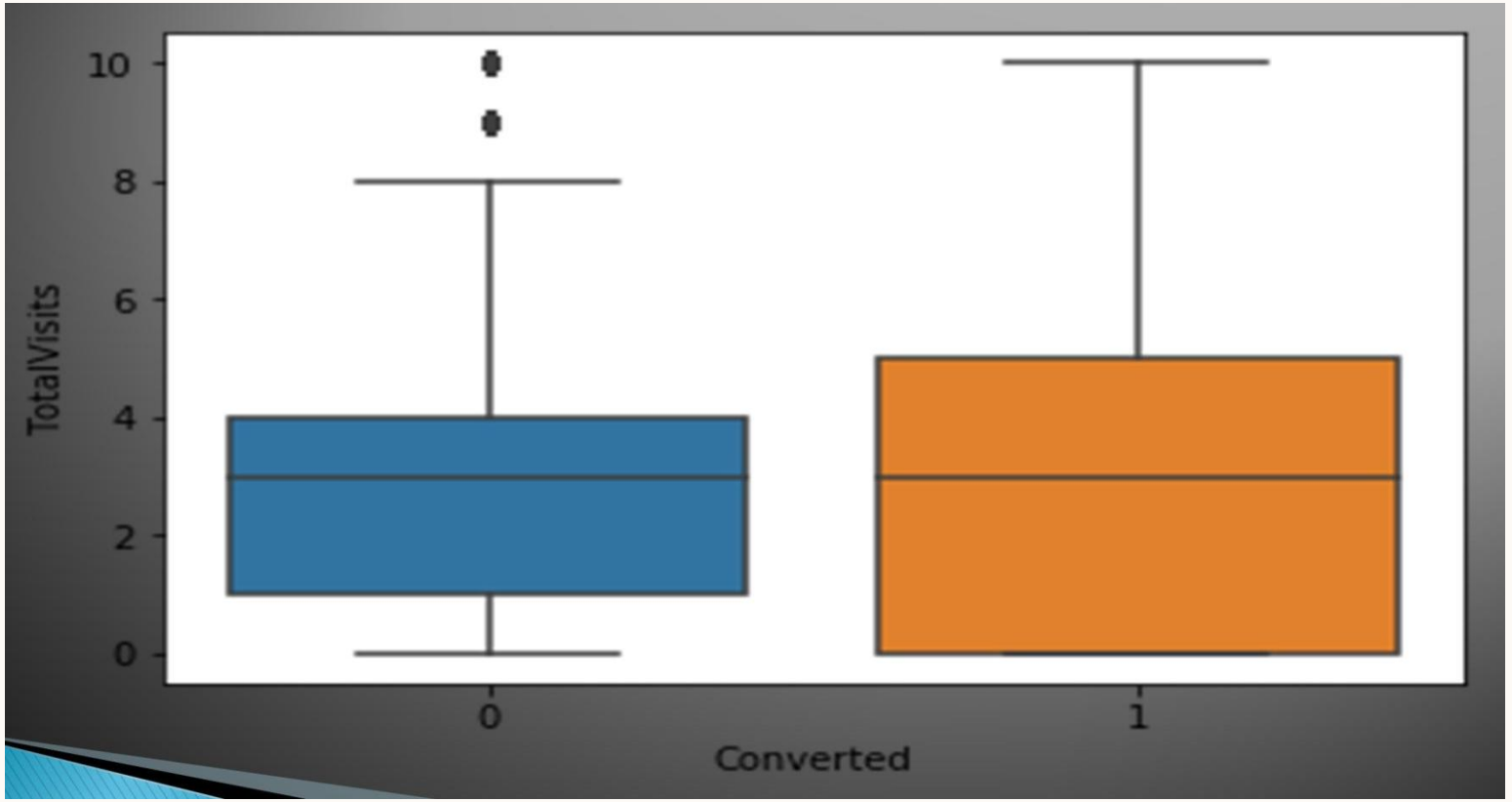
- Most leads prefer not to be informed through phone.
- Google searches has high conversion compared to other models, whilist references has had high conversion rate.



TOTAL VISITS VS CONVERTED

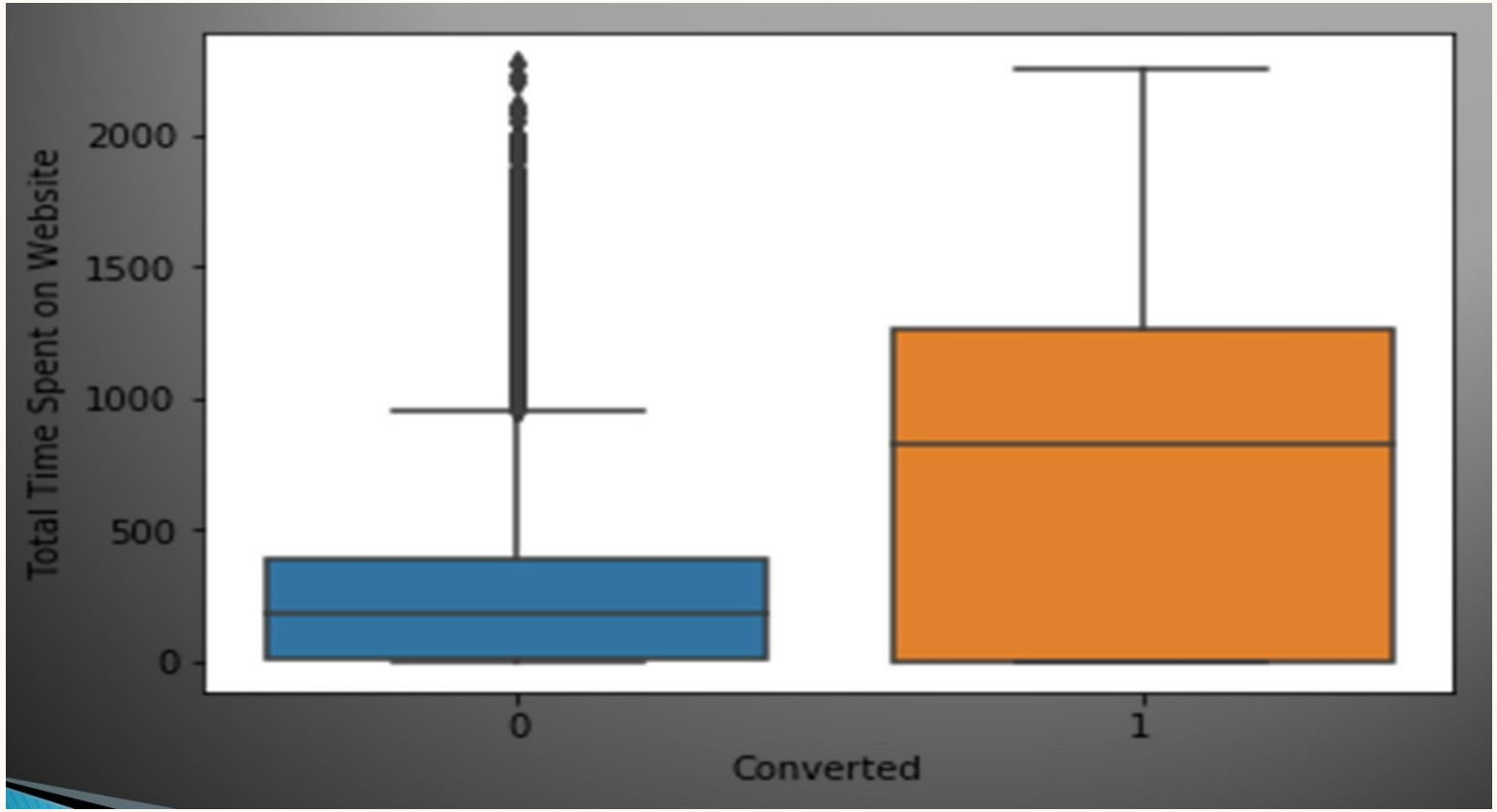
- Median for converted and not converted leads are the same.

Nothing conclusive can be said on the basis of total visits.



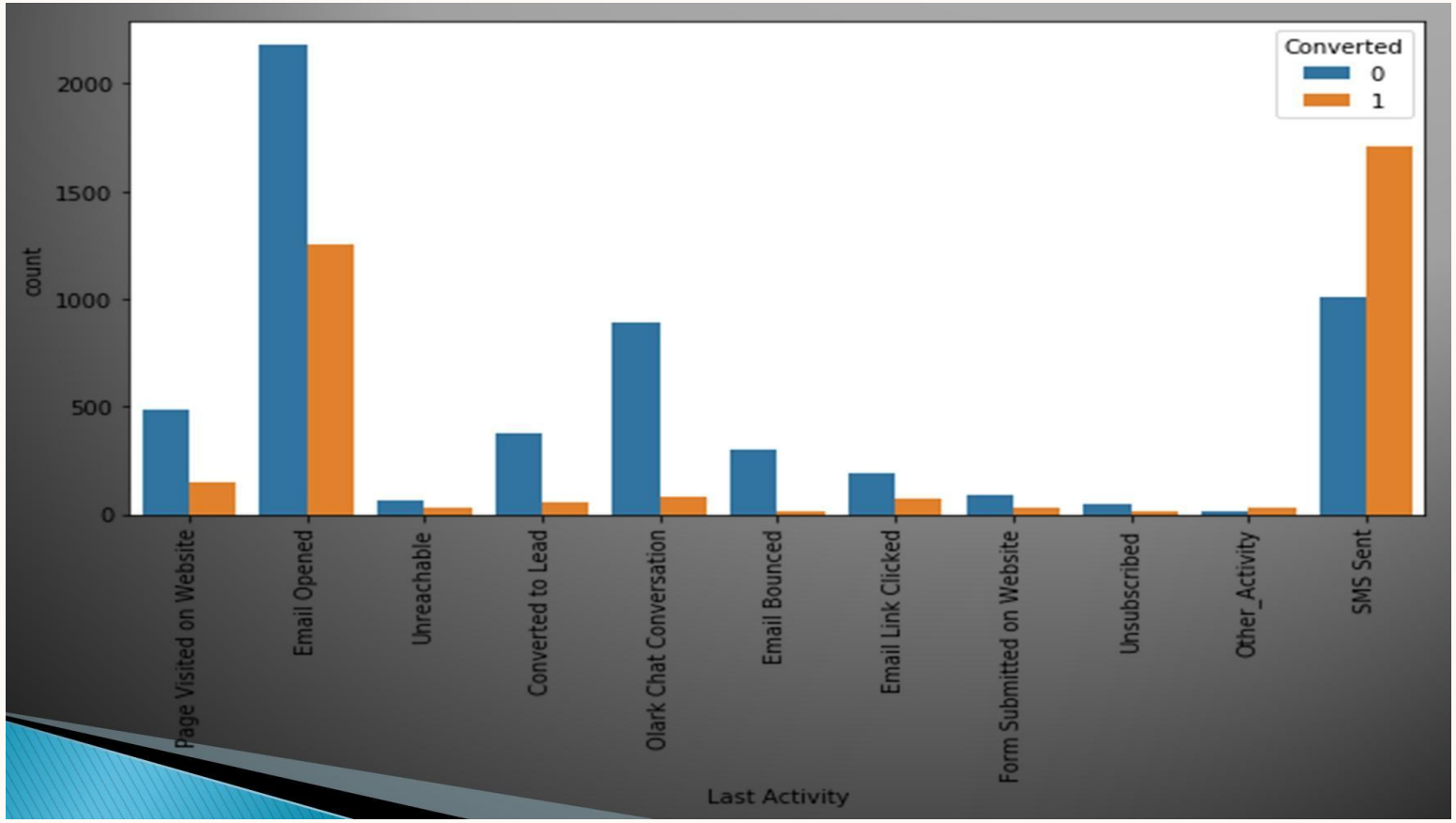
TOTAL TIME SPENT ON WEBSITE VS CONVERTED

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



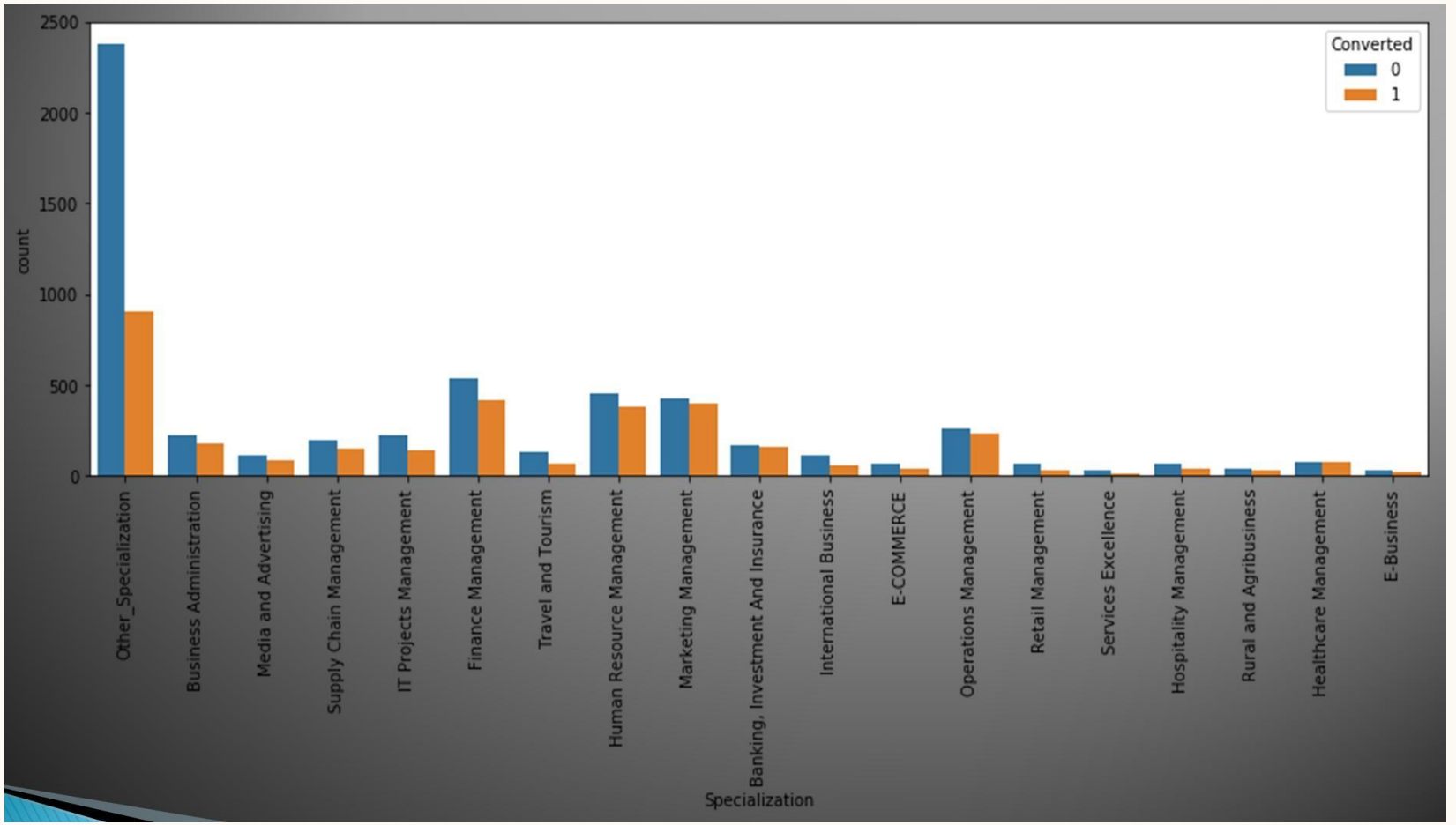
LAST ACTIVITY VS CONVERTED

- Most of the lead have their email opened has their last activity.
- Conversion rate for leads with last activity as SMS sent is almost 60%.



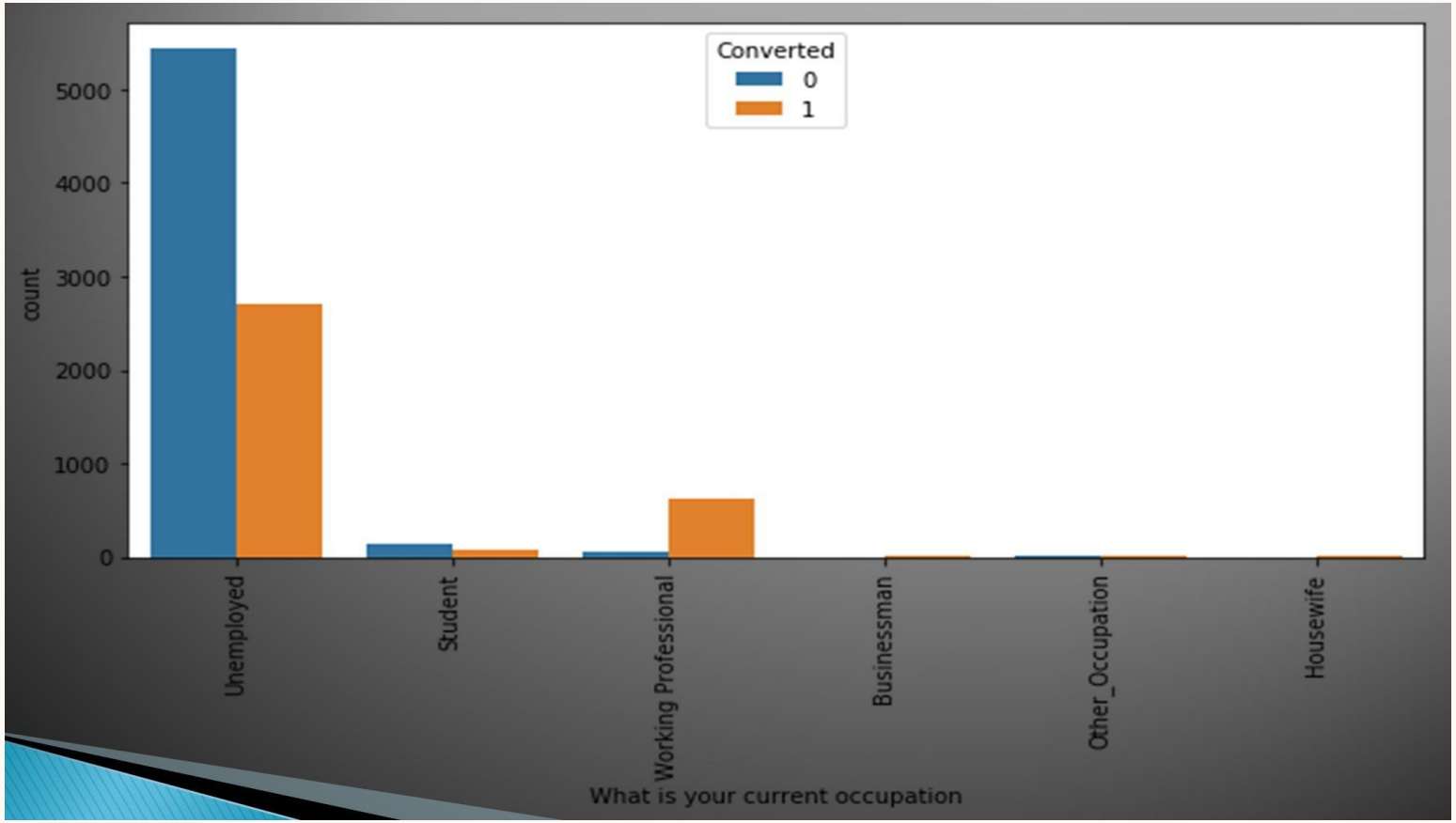
LAST ACTIVITY VS CONVERTED

Focus should be more on the specialization with high conversion rate.



WHAT IS YOUR CURRENT OCCUPATION VS CONVERTED

- Working professionals going for the course have high chances of joining it.
- Unemployed leads are the most in numbers but has around 30-35% conversion rate.



MODEL BUILDING

- Splitting into train and test
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on high p-value
- Check VIF value for all existing column
- Predict using train set
- Evaluate accuracy and other metrics
- Predict using test set
- Precision and recall analysis on text predictions

MODEL EVALUATION (TRAIN)

Accuracy, Sensitivity, Specificity

- 80.9% → Accuracy
- 77.6% → Sensitivity
- 82.9% → Specificity

Precision and recall

- 73.4% → Precision
- 77.6% → Recall

MODEL EVALUATION (TEST)

Accuracy, Sensitivity, Specificity

- 80.1% → Accuracy
- 75.5% → Sensitivity
- 83.1% → Specificity

Precision and recall

- 74.4% → Precision
- 75.5% → Recall

CONCLUSION

EDA

- People spending higher than average time are promising leads,so targeting them and approching them can be helpful in conversions
- SMS messages can have a high impact on lead conversion
- Landing page submission can help find out more leads
- Marketing management, human resources management has high coonversion rates.People from this specialization can be promising leads
- References and offers for referring a lead can be good source for higher conversions
- An alert messages or information has seen to have high lead conversion rate

Logistic Regression Model

- The model shows high close to 81% accuracy
- The threshold has been selected from Accuracy,Sensitivity,Specificity measures and precision recall curves
- The model shows 76% sensitivity and 83% specificity
- The model finds correct promising leads and leads that have less chances of getting converted
- Overall this model proves to be accurate.