# Index

- Problem Statement
- Analysis Approach
- Data analysis (Plots)
- Precision and Recall metrics for the test
- Results and conclusions

# Problem Statement

- An education company named X sells online courses to industry professionals. 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 and this is classified as “lead”. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. Hence the company is interested in knowing it’s most potential leads and further increasing the conversion rate for the benefit of the company.
- The company wishes to identify the most potential leads, also known as ‘Hot Leads’. In order to do that we are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- We are expected to 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.

# ANALYSIS APPROACH

- To solve the business problem that was gives to us, we were required to build a logistic regression model wherein we had to assign a lead score to each of the leads, in this case between 0-100 to each of the leads which can be used by the company to target potential leads.
- We build the logistic regression model as follows :-

  a) Data Cleaning

  b) Data Analysis

  c) Data Preparation

  d) Test-Train Split

  e) Rescaling

  f) Model Building (Logistic Regression model)

  g) Feature selection using RFE

  h) Rebuilding model

  i) Plotting ROC Curve

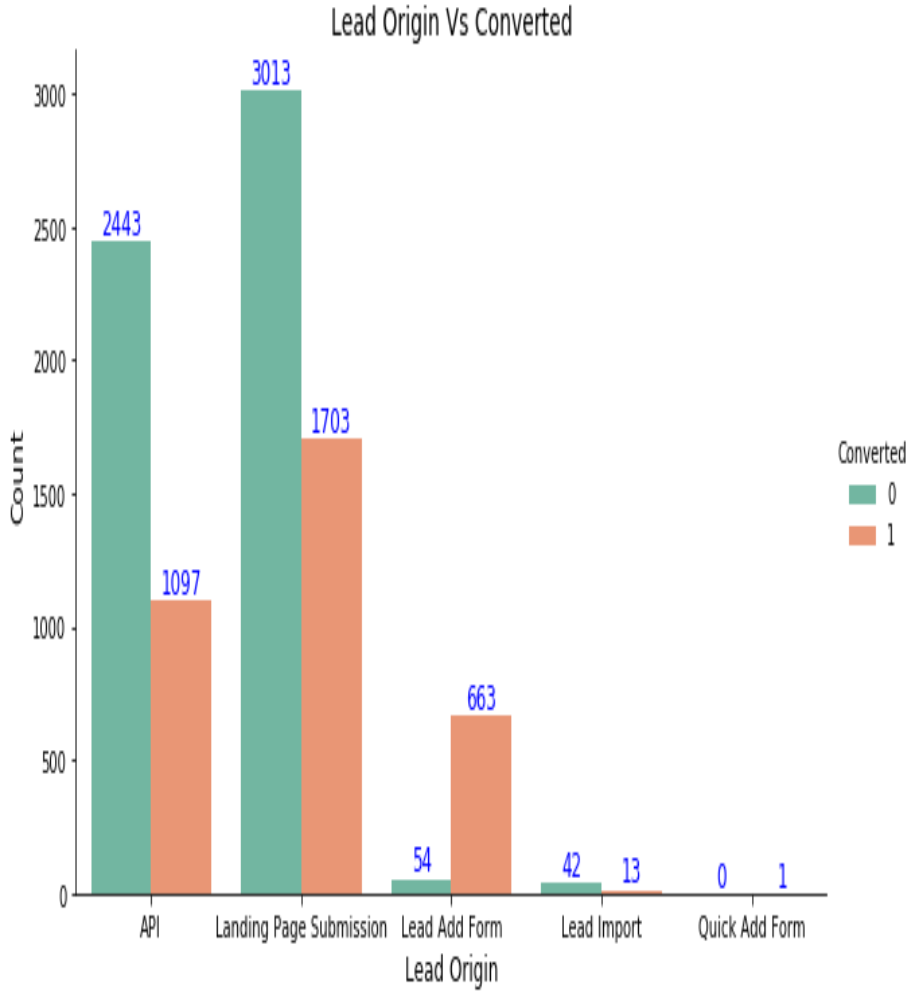
  j) Finding Optimal Cut off point

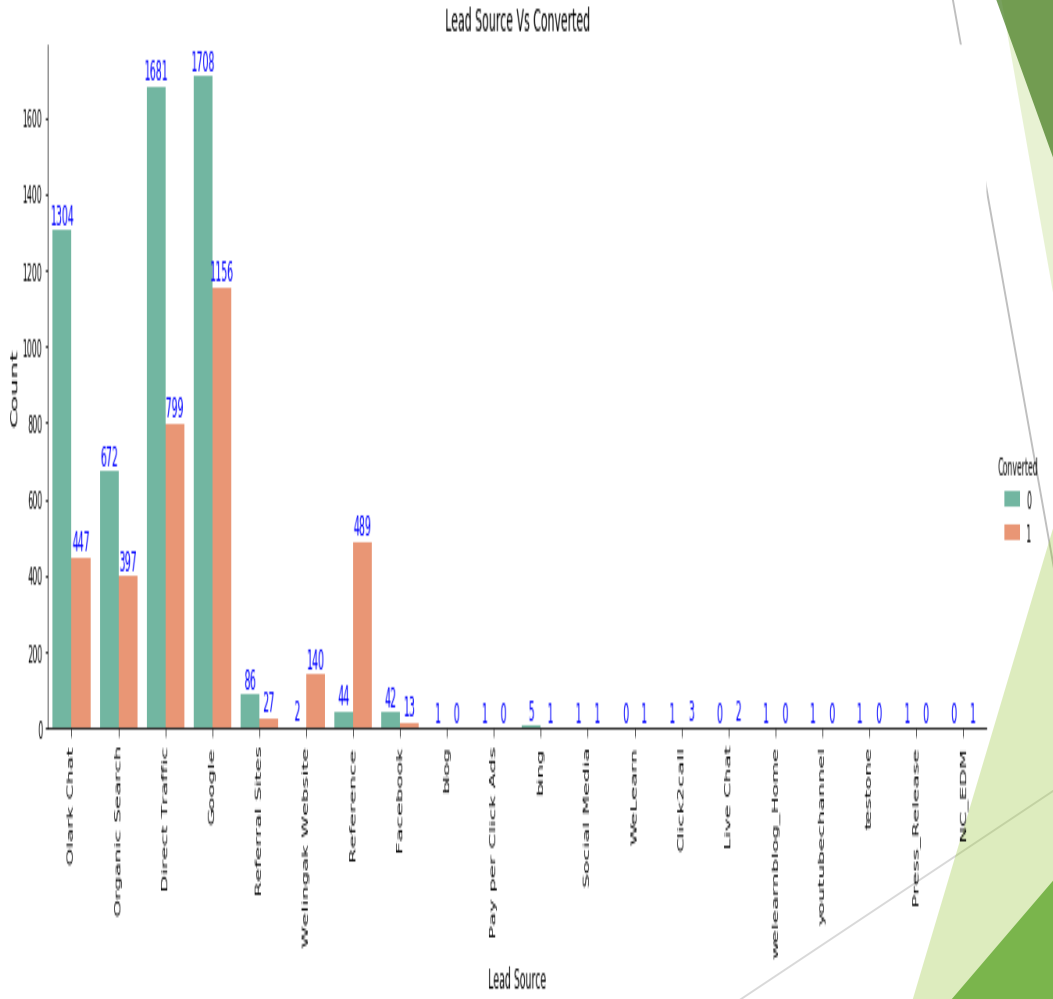
  k) Making prediction on the test set

# ANALYSIS APPROACH

- In order to build the logistic regression model we used the usual approach by first cleaning our data, checking the outliers and then doing the Data Analysis where we used plots to understand our data set better.
- We used columns such as lead origin, lead source, last activity digital ad etc to make plots in order to understand the conversion rate better.
- Further we did the Data Preparation by creating dummies for the columns and removing the columns that weren't helpful for the model building.
- We then divided the data into test and train set to further rescale.
- Finally we built the logistic regression model that was required.
- After feature selection using RFE, rebuilding the model and checking for confusion metrics & accuracy, VIFS and then plotting ROC curve we made the predictions for the company.

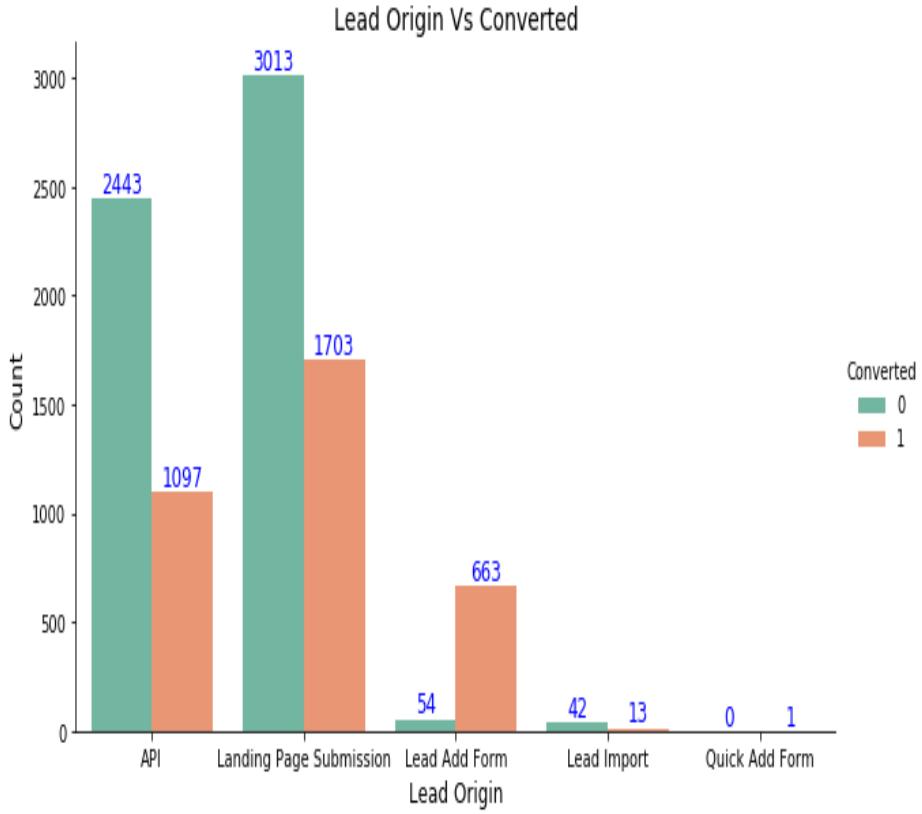
# Data Analysis (Plots)



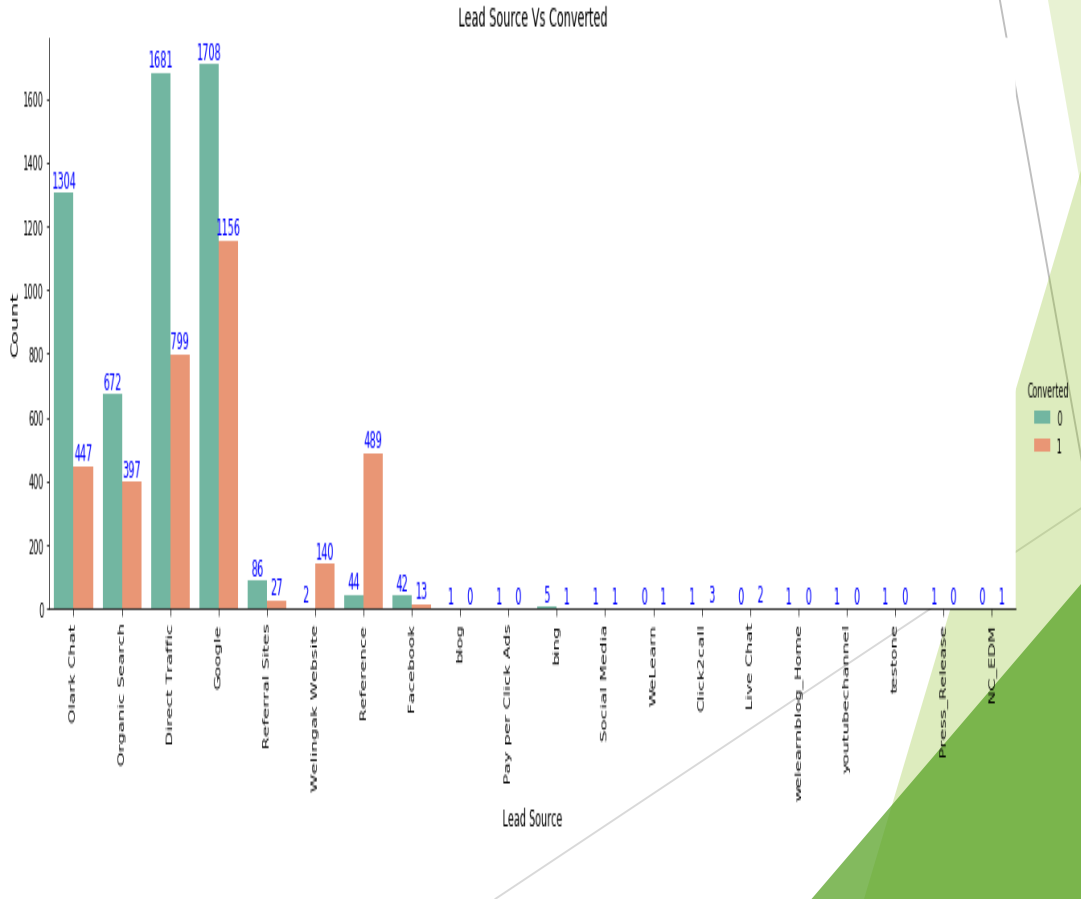



# Data Analysis (Plots)

From the above graph, it can be seen that the maximum conversion happened from Landing Page Submission

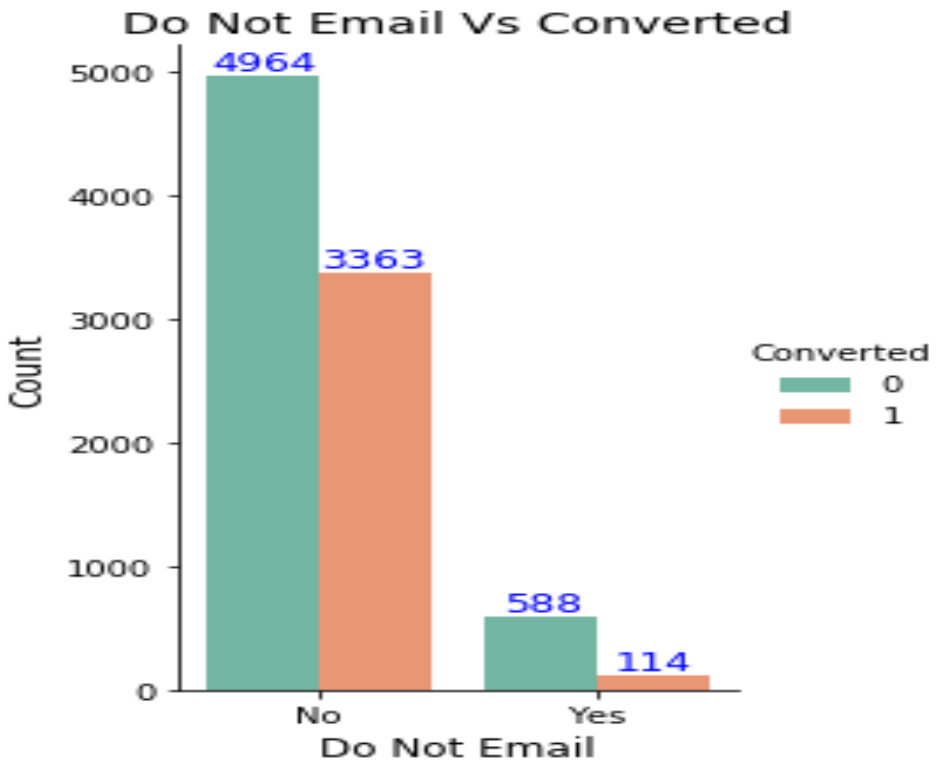



major conversion in the lead source is from google

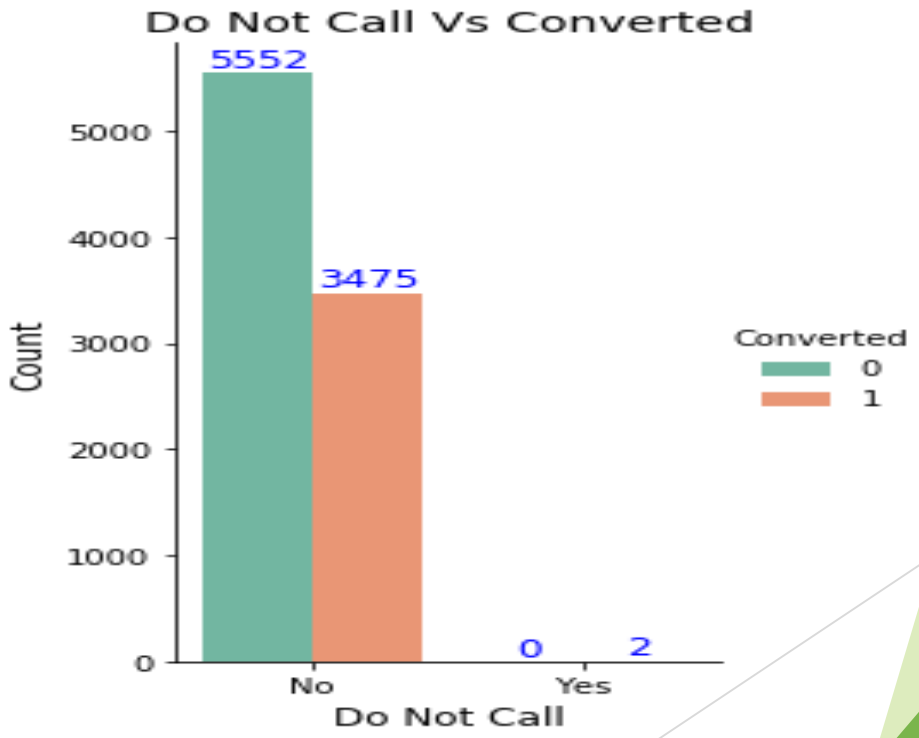


# Data analysis (Plots)

Major conversion has happened from the emails that have been sent

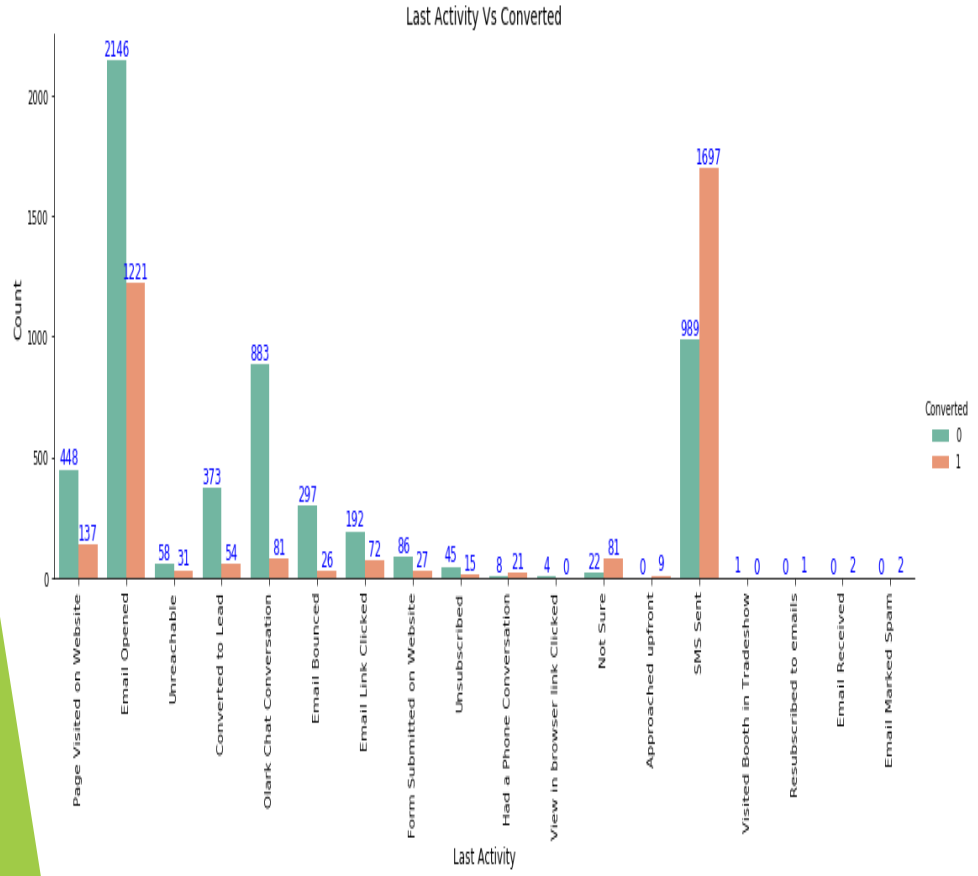



Major conversion has happened from the calls that have been made

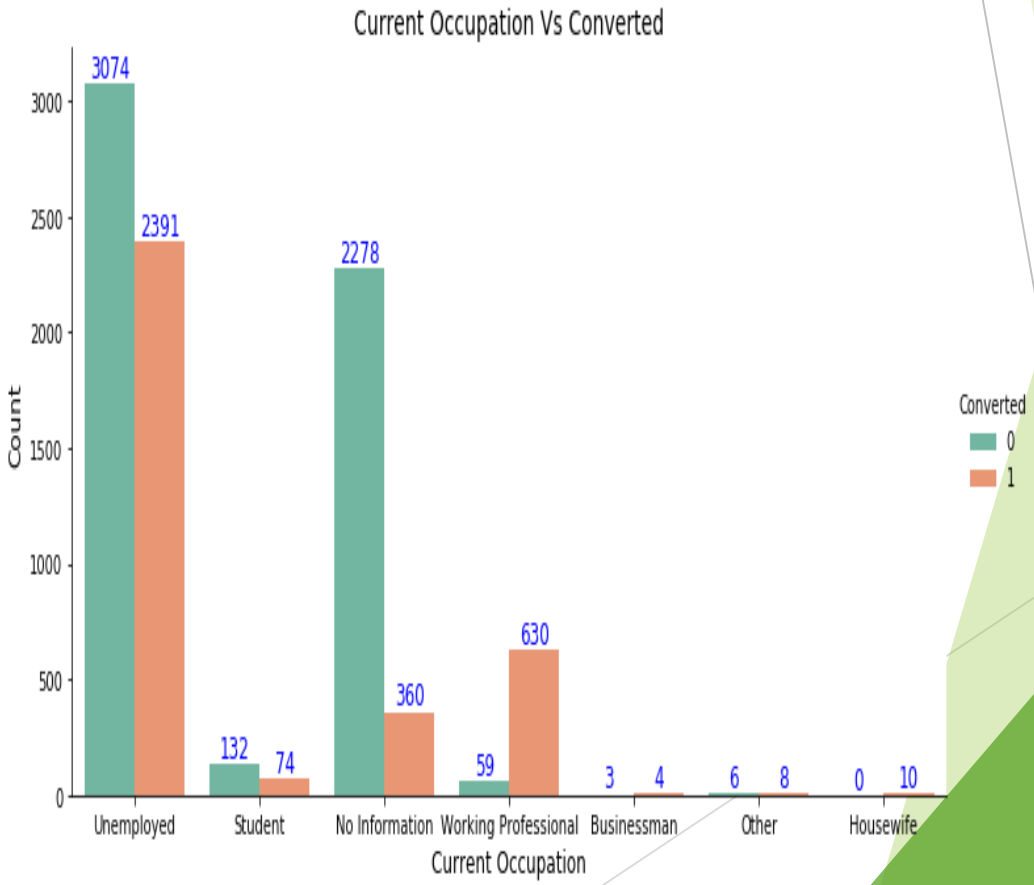


# Plots

Major conversion has made from the sms sent

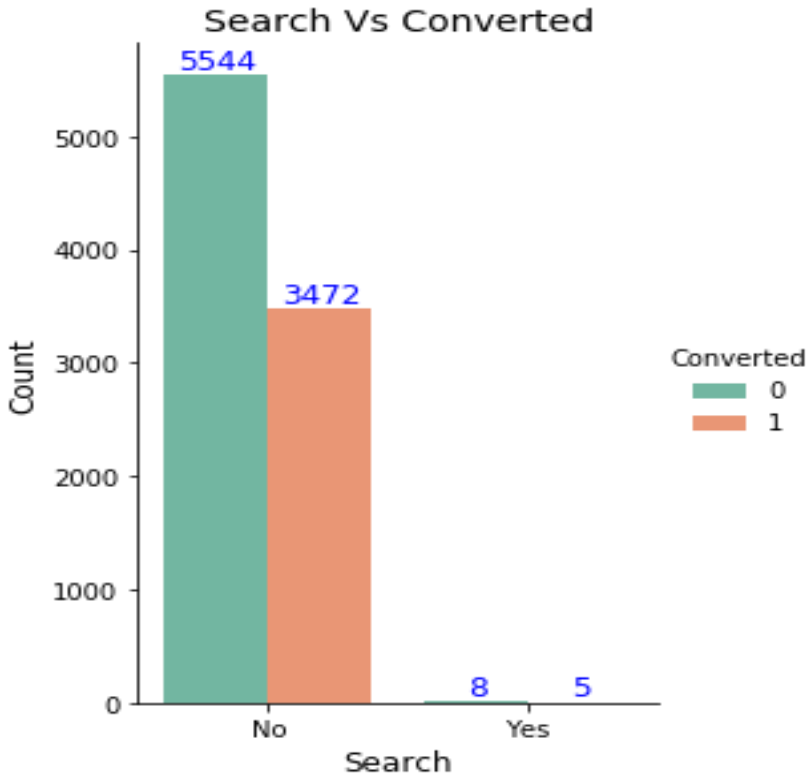



major conversion from unemployed people. All 10 housewives have been converted and 4 out of 7 businessmen converted

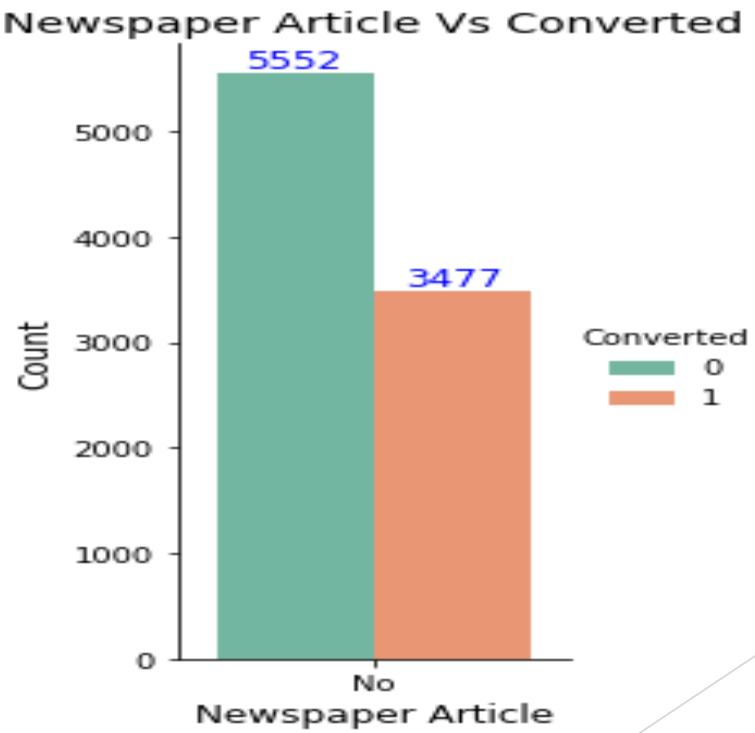


# DA (Plots)

conversions high for people who have not searched

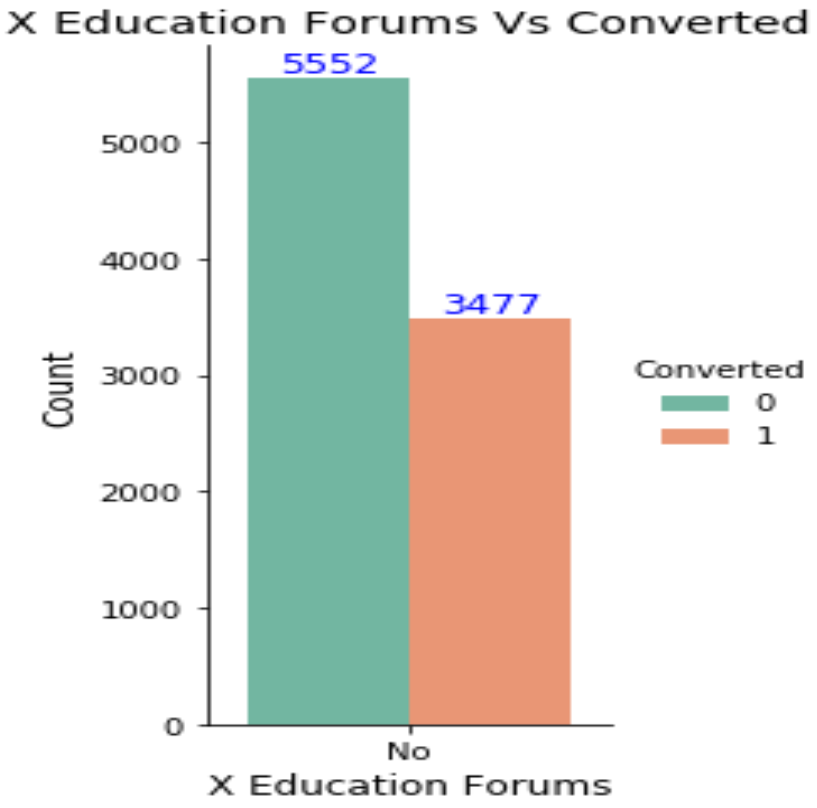



newspaper article has only 1 value i.e. No

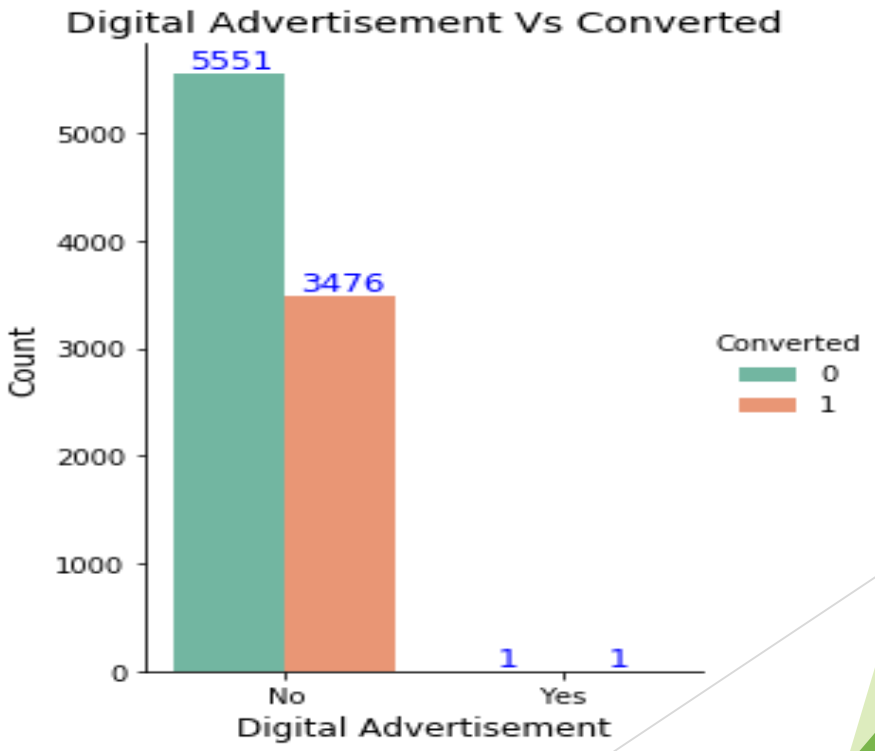


# DA (Plots)

X Education Forums has only value as No.

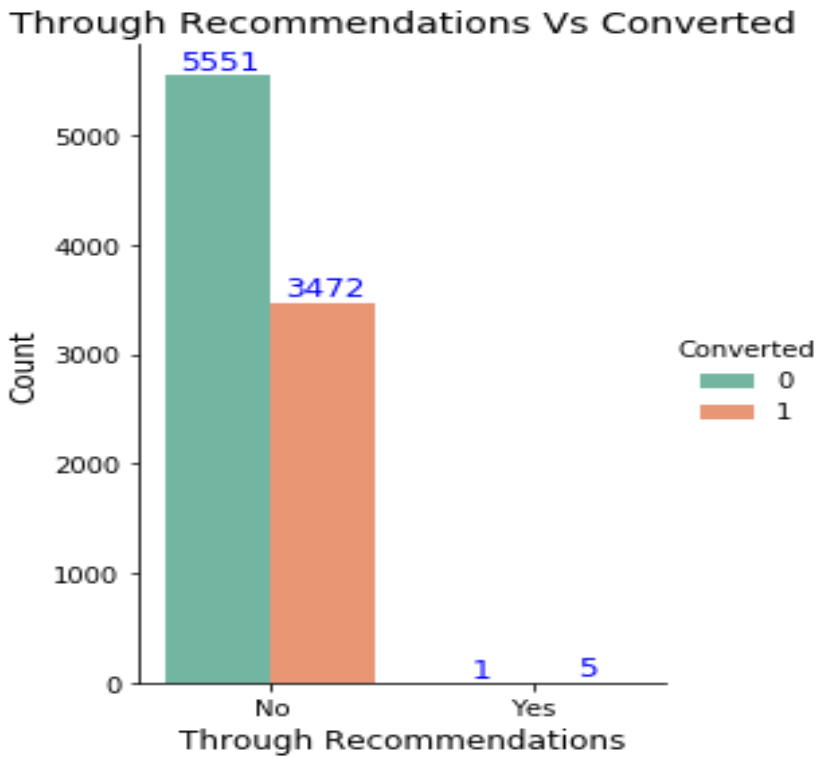



It can be noticed above that there were 2 leads that came from digital advertisement of which one lead got converted

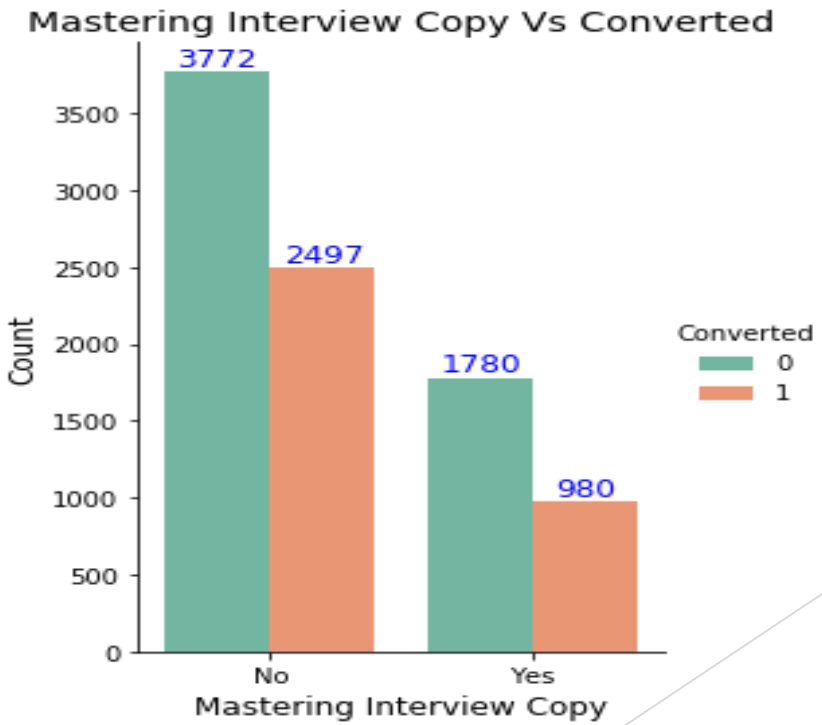


# DA (Plots)

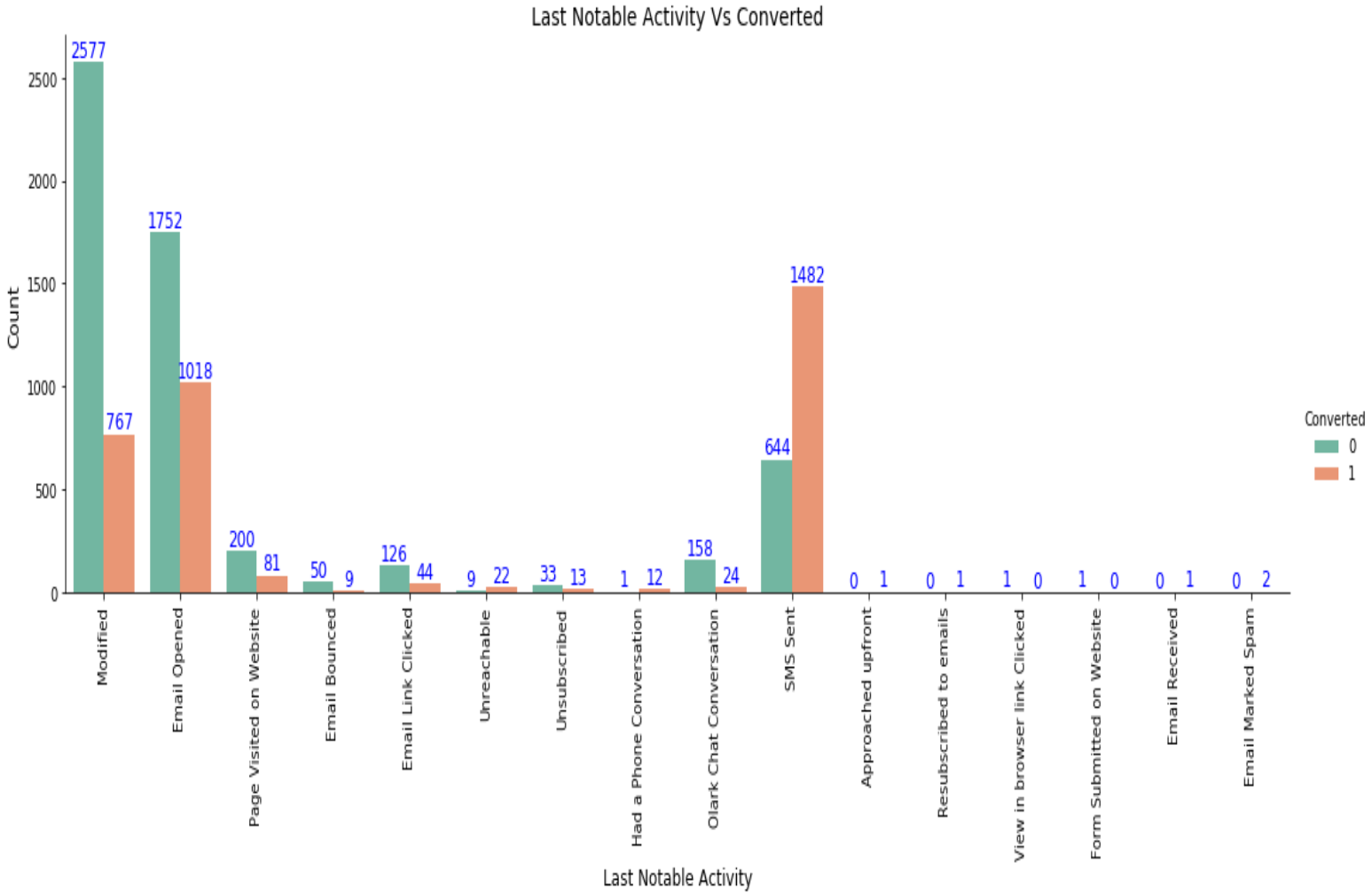
It can be seen that a total of 6 leads came through recommendations of which 5 leads got converted




Conversion rate is high on leads who do not want a free copy of Mastering Interviews

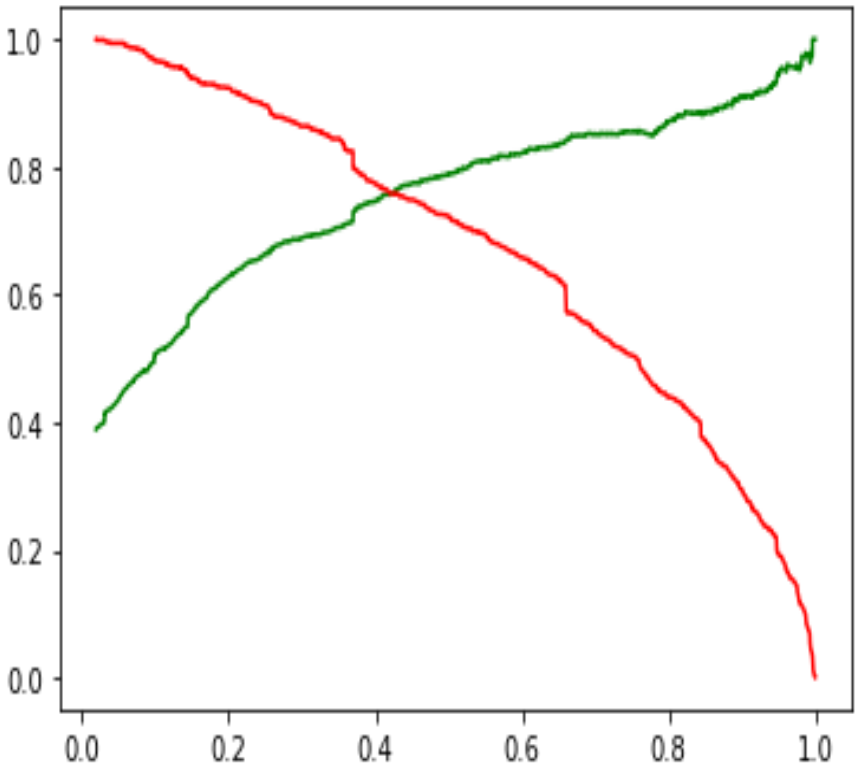


# DA (Plots)




- It can be noticed that the conversion rate is high for "SMS Sent"

# PRECISION AND RECALL METRICS FOR THE TEST




- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good

# RESULTS AND CONCLUSION

- The logistic model that we've build tells us that accuracy, sensitivity and specificity values of test set are around 81%, 79% and 82% respectively which are approximately closer to the respective values calculated using trained set.
- The lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%.
- Overall the model seems to be beneficial for the company as the target lead conversion rate is around 80%.