LEAD SCORE CASE STUDY

Sreedevi Ramachandran

Reefat Shaikh
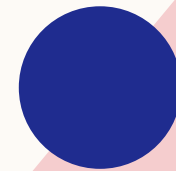
Masthan Raja

Date: 18-Jun-2023

HIGHLIGHTS

Problem Statement

Steps followed for Logistic Regression analysis

Data Visualizations

Model Summary

Recommendations

PROBLEM STATEMENT

➤ X Education gets a lot of leads, i.e. professionals who are interested in their courses who land on their website and browse for courses.

➤ However, its lead conversion rate is very poor.

➤ X Education needs to identify the most promising leads, i.e. the leads that are most likely to convert into paying customers.

➤ To make their lead conversion process more efficient, the company wishes to identify the most potential leads, known as 'Hot Leads'.

➤ The ballpark of the target lead conversion rate should be around 80%.

Goals and Objectives of the Case Study:

➤ 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.

➤ The model should be able to adjust to the company's requirement changes in the future.

Dataset provided:

1. 'Leads.csv' - leads dataset from the past with around 9000 data points.

2. 'Leads Data Dictionary.xlsx' – data dictionary to learn more about the dataset.

STEPS FOLLOWED FOR LOGISTIC REGRESSION ANALYSIS

➢ Step 1: Importing Data

➢ Step 2: Inspecting the Data frame

➢ Step 3: Data Preparation: To do an initial data pre-processing

   ➢ Handling missing values & imputing them

   ➢ Handling columns with imbalanced data

   ➢ Handling outliers

➢ Step 4: Data Visualizations

➢ Step 5: Data Preparation – II

   ➢ Converting binary variables (Yes/No) to 0/1

   ➢ For categorical variables with multiple levels - create dummy features

   ➢ Test-Train Split

   ➢ Feature Scaling

   ➢ Looking at Correlations

➢ Step 6: Model Building

   ➢ Feature Selection Using RFE

   ➢ Assessing the model with Stats Models

   ➢ Evaluate model accuracy, confusion matrix & other metrics

   ➢ Plotting the ROC Curve

   ➢ Finding Optimal Cutoff Point

➢ Step 7: Making predictions on the test set - Assign Lead Score & Hot Leads
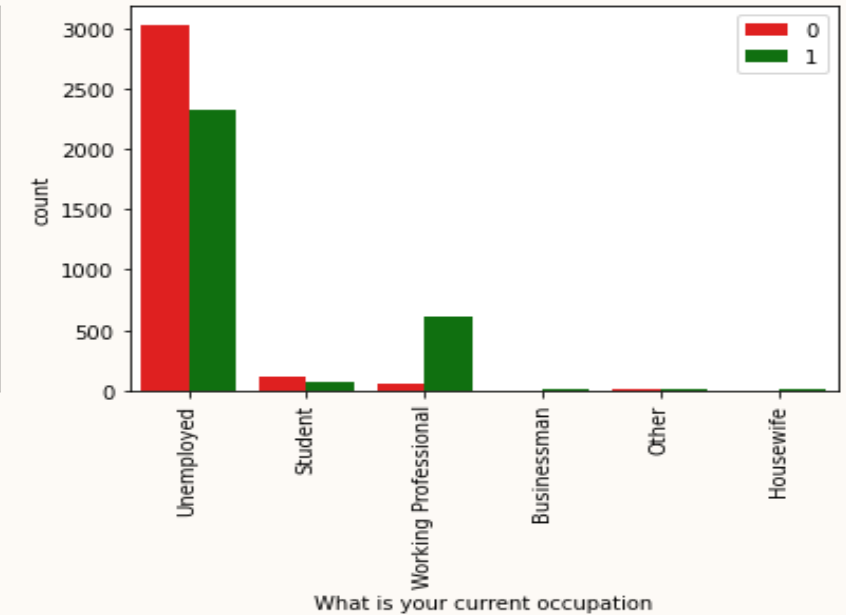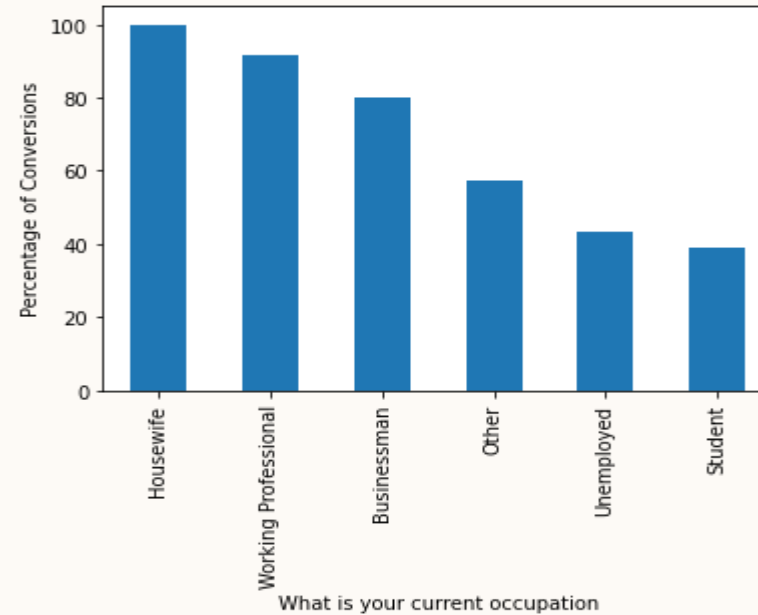
DATA VISUALIZATIONS

IMBALANCE ANALYSIS OF TARGET FEATURE - CONVERTED
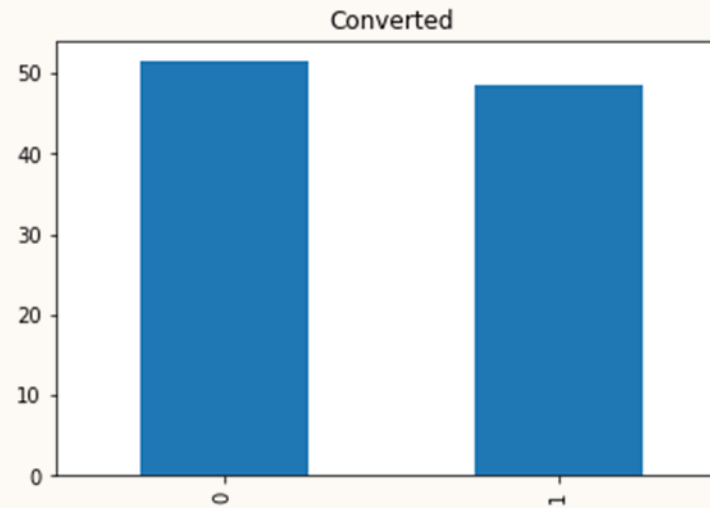
- ➤ We have a balanced dataset.
- ➤ As per the dataset, 48% leads have converted and 51% not converted which looks okay to start further EDA and analysis.
- ➤ The current Conversion rate is 49%.

FEATURE ANALYSIS – WHAT IS YOUR CURRENT OCCUPATION

- ➢ There is good lead conversion among Working Professionals (WP) ~91% of the total WP have converted. So, WP seems like a good crowd for targeting the campaigns.
- ➢ Of the Unemployed Professionals (UP), 43% have converted. Bu, There are large number of UP's who have not converted – which needs to be looked at.
- ➢ Of the 10 Housewives identified as leads, all have been converted. So, there seems like a good potential but need to dig in on why we have less Housewives as leads.
- ➢ There is also a small student population 34% of whom have converted from potential leads.

# FEATURE ANALYSIS – LAST ACTIVITY & LAST NOTABLE ACTIVITY
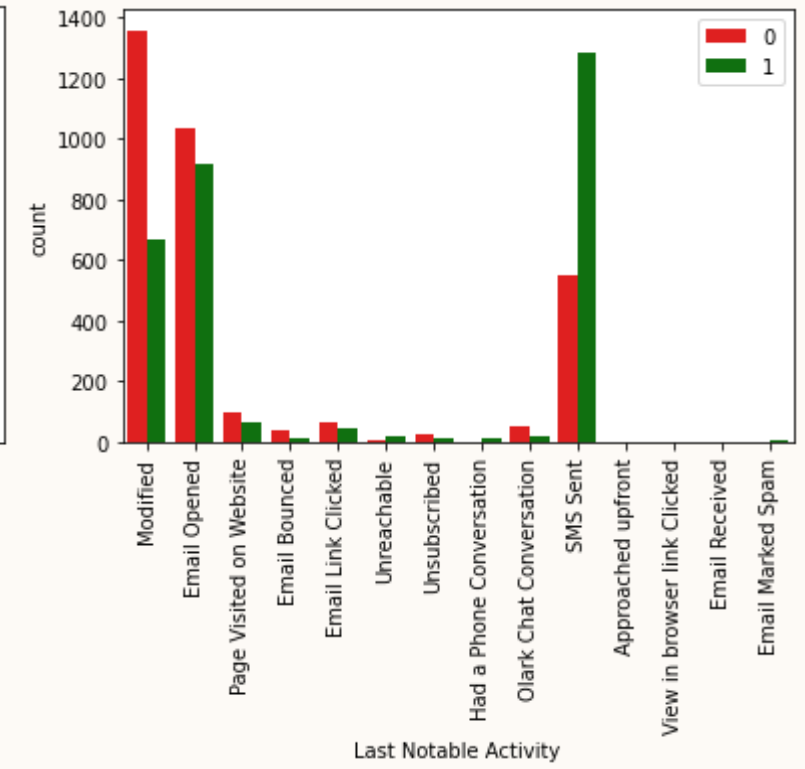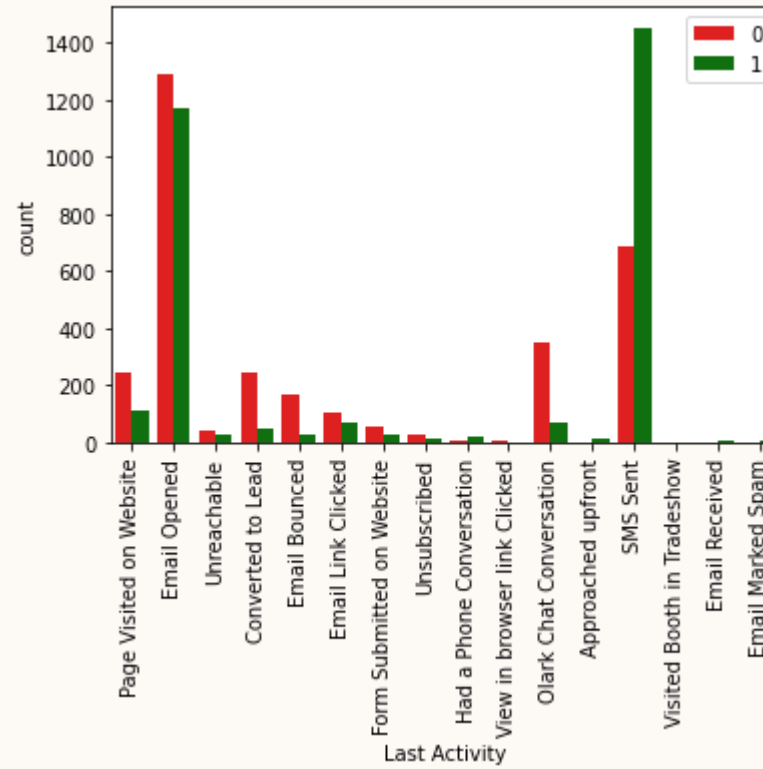
- The customers who have SMS Sent as Last Activity/Last Notable Activity have high conversion rate.
- For the feature Email Opened, both conversion & non-conversion counts are close to each other. Need to see how to be able to target the non-conversions here.
- The features - Modified and Olark Chat Conversion have lesser conversions.
- These 2 features - Last Activity/Last Notable Activity are almost the same - so one can be removed from further analysis.

FEATURE ANALYSIS – LEAD ORIGIN & LEAD SOURCE
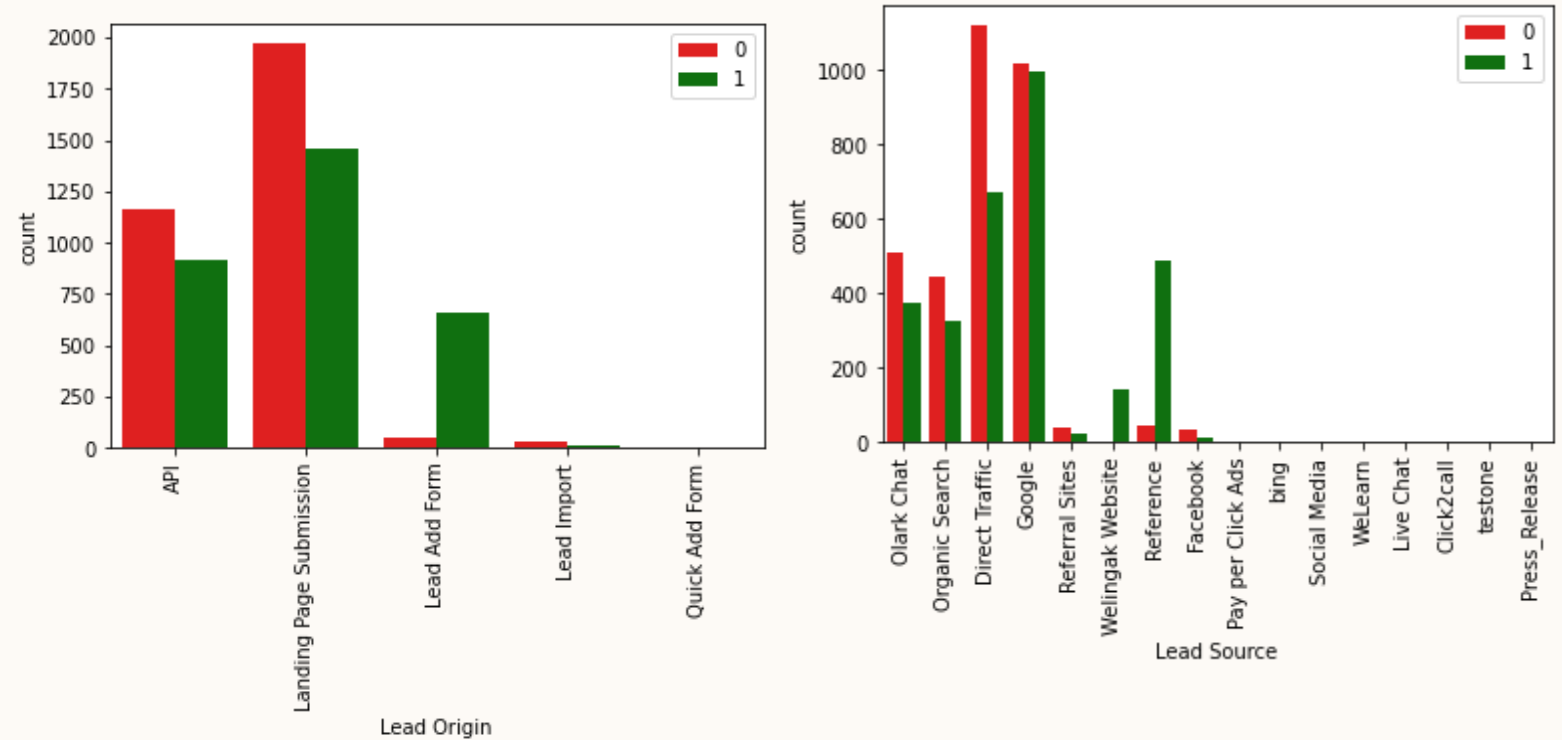
- ➢ Lead Source - Reference & Welingak Website have the best conversion rate.
- ➢ Direct Traffic, Google, Olark Chat & Organic Search seem to be the origin for majority of the leads where in each has 50% conversion rate.
- ➢ The majority of the leads are seen to be originating from the Lead Origin - API, Landing Page Submission & Lead Add Form. Lead Add Form has high conversion rate.
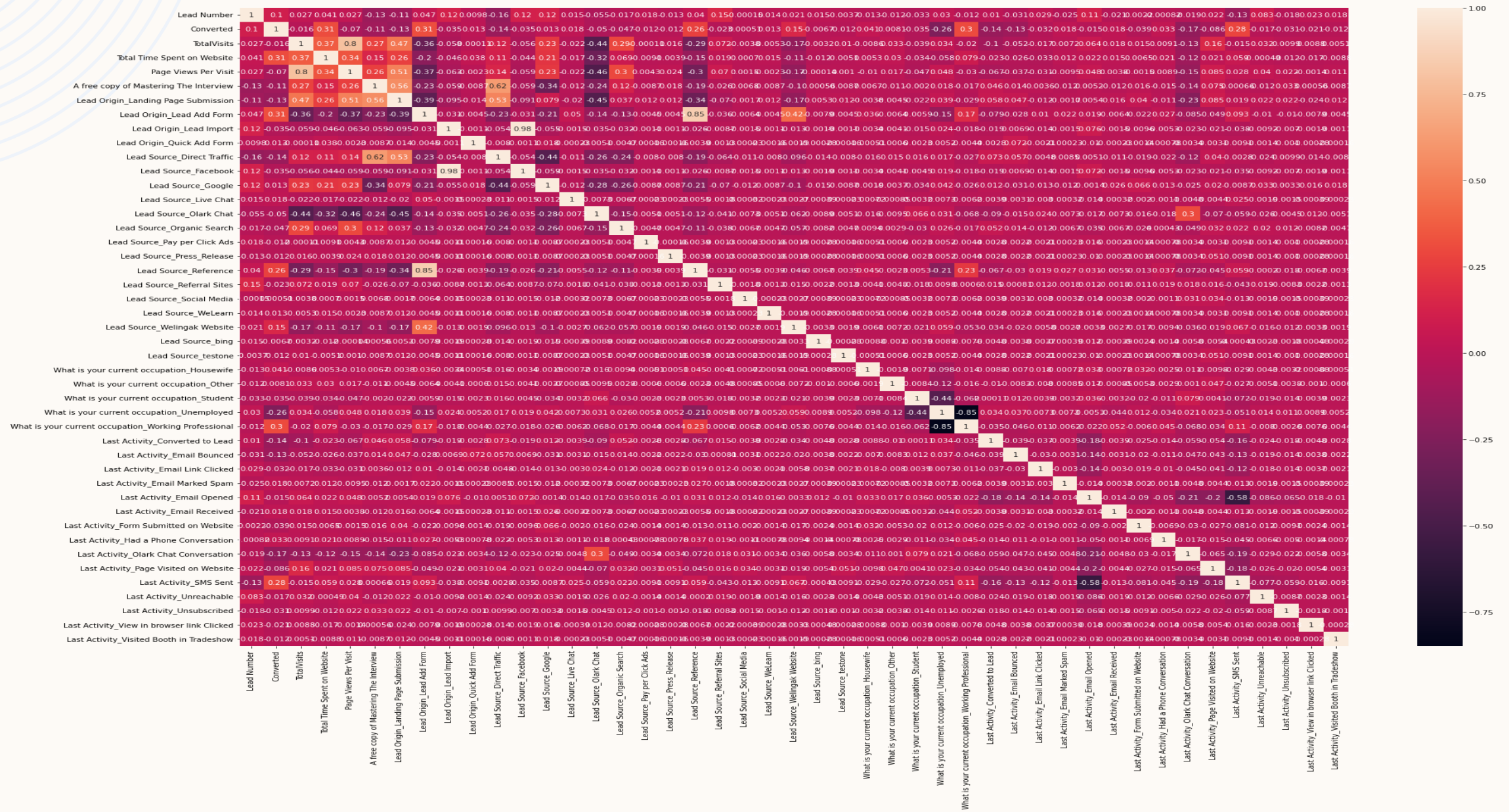
FEATURE ANALYSIS – A FREE COPY OF MASTERING THE INTERVIEW
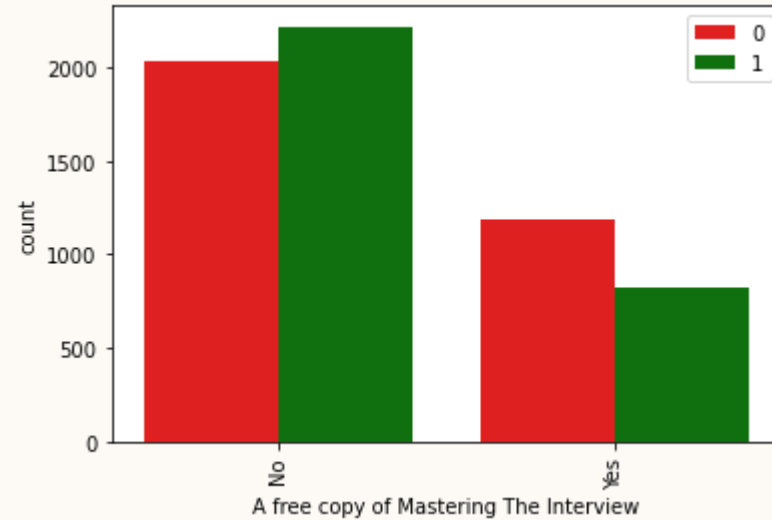
➢ 'A free copy of Mastering The Interview' does not seem to be a driving factor for conversions.

➢ Even those customers who have said No for freebie, have converted from Leads.

HEAT MAP ANALYSIS

HEAT MAP INSIGHTS

➢There is strong negative correlation between:

      What is your current occupation_Unemployed

      What is your current occupation_Working Professional

➢There is positive correlation between the below features:

      TotalVisits & Page Views Per Visit

      Lead Origin_Lead Add Form & Lead Source_Reference

      Lead Origin_Lead Import & Lead Source_Facebook

MODEL SUMMARY

➢ A Logistic Regression Model has been built and trained by the team to achieve 82% lead conversion rate and it achieves the same accuracy for test data as well.

➢ A lead score between 0 and 100 is assigned to each of the leads which can be used by the company to target potential leads.

➢ The most potential leads, also known as 'Hot Leads' can be predicted by the model created.

➢ As per the final model, the following features have Positive Coefficients:

- Total Time Spent on Website
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Last Activity_Email Opened
- Last Activity_SMS Sent
- Last Activity_Unreachable

➢ As per the final model, the following features have Negative Coefficients:

- Lead Source_Direct Traffic
- Last Activity_Converted to Lead
- Last Activity_Email Bounced
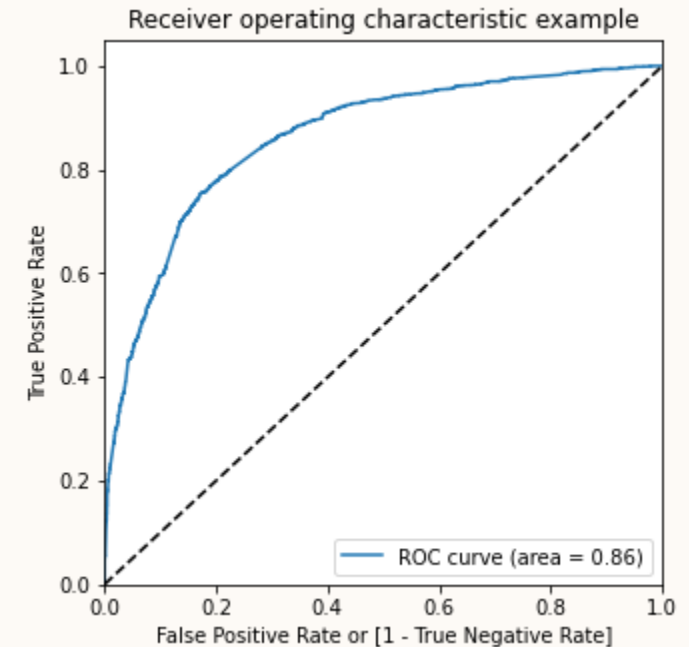- Last Activity_Olark Chat Conversation

MODEL METRICS

➢ Model Evaluation Metrics on train data:
   ▪ Accuracy - 77.92
   ▪ Sensitivity - 84.23
   ▪ Specificity - 71.83
   ▪ Precision - 74.25
   ▪ Recall - 84.23

➢ Model Evaluation Metrics on test data:
   ▪ Accuracy - 77.69
   ▪ Sensitivity - 82.25
   ▪ Specificity - 73.58
   ▪ Precision - 73.72
   ▪ Recall - 82.25

Receiver operating characteristic example

➢ Receiver Operating Characteristic curve:
   ▪ The ROC curve looks and the area under the curve is 0.86.
   ▪ An AUC score closer to 1 means that the model has the ability to separate the two classes and the curve would come closer to the top left corner of the graph.

RECOMMENDATIONS

➢ The company should target the leads who spent "more time on the Website" since they are more likely to get converted.

➢ The company should make calls to the leads coming from the lead sources – "Welingak Websites, Reference & Olark Chat" as these are more likely to get converted.

➢ The company should reach out to the leads who are "Working Professionals" since they have more chances of conversions.

➢ The company should make calls to the leads who have last activity as "Email Opened, SMS Sent & Unreachable".

➢ The company should NOT make any calls to the leads coming from "Direct Traffic" since it has very less chances of conversion.

➢ The company should NOT make calls to the leads who have last activity as "Converted to Lead, Email Bounced & Olark Chat Conversation".

THANK YOU