LEADS SCORING CASE STUDY

Sushma Subburayan
Lekha Priyadarshini

LEADS SCORING CASE STUDY

Problem Statement:

- X Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires wants a model 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
- The required target lead conversion rate is around 80%.

Data :

Leads data data set containing the features of the leads

GOALS OF THE ANALYSIS:

- To assign calculate the lead score assign it 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
- To find the leads score cut off to achieve 80% conversion rate.

BACKGROUND

- X Education, an education company, sells online courses to industry professionals who are interested in the courses, lands on their website and browses for courses.
- The company markets its courses on several websites and search engines, so people landing on the website, might browse the courses or fill up a form for the course or watch some videos.
- People filling up a form providing their email address or phone number are classified to be a lead. The company also gets leads through past referrals and recommendations.
- The sales and marketing team then try to convert these leads by making calls, writing emails, etc.

- The following are the features
  - Name of the country
  - Child Mortality - Death of children under 5 years of age per 1000 live births exports
  - Exports of goods and services. Given as %age of the Total GDP
  - Health - Total health spending as %age of Total GDP
  - imports - Imports of goods and services. Given as %age of the Total GDP
  - Income - Net income per person
  - Inflation - The measurement of the annual growth rate of the Total GDP
  - Life Expectancy - The average number of years a new born child would live if the current mortality patterns are to remain the same
  - Total Fertility - The number of children that would be born to each woman if the current age-fertility rates remain the same.
  - GDPP - The GDP per capita. Calculated as the Total GDP divided by the total population.

OBJECTIVE

Lead Conversion Process - Demonstrated as a funnel

DATA UNDERSTANDING

- There are 9240 records in the given lead scoring dataset.
- There are 37 variables in the given lead scoring dataset.
- A lot of fields had “Select” as the value which is treated as null in the data preparation.
- Removed the columns where the percentage of null value is greater than 30% and also the columns which didn’t had any significant variance in the values. Variables that had more than 90% of the records as the same value.
- There are 9074 record and 68 columns after data cleaning and the data preparation.

SOLUTION APPROACH

1 Data Understand

Read and understood the given leads data set

Data Cleaning

Cleaning the data by removing null values ouliers etc

3 Data Preparation

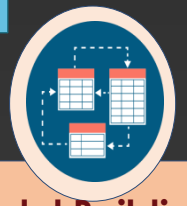
Assigning 1 and 0 to the values for the features with only two levels

Creating dummy variable for features with more than two levels

Splitting the dataset in to train and test dataset in the ratio of 70:30

Data Visulaization

Present the data in a visual format for easy understanding.



Model Building

Using RFE for Logistic regression

Using Generalized Linear Model Regression

Getting a better model on train data after multiplate modification

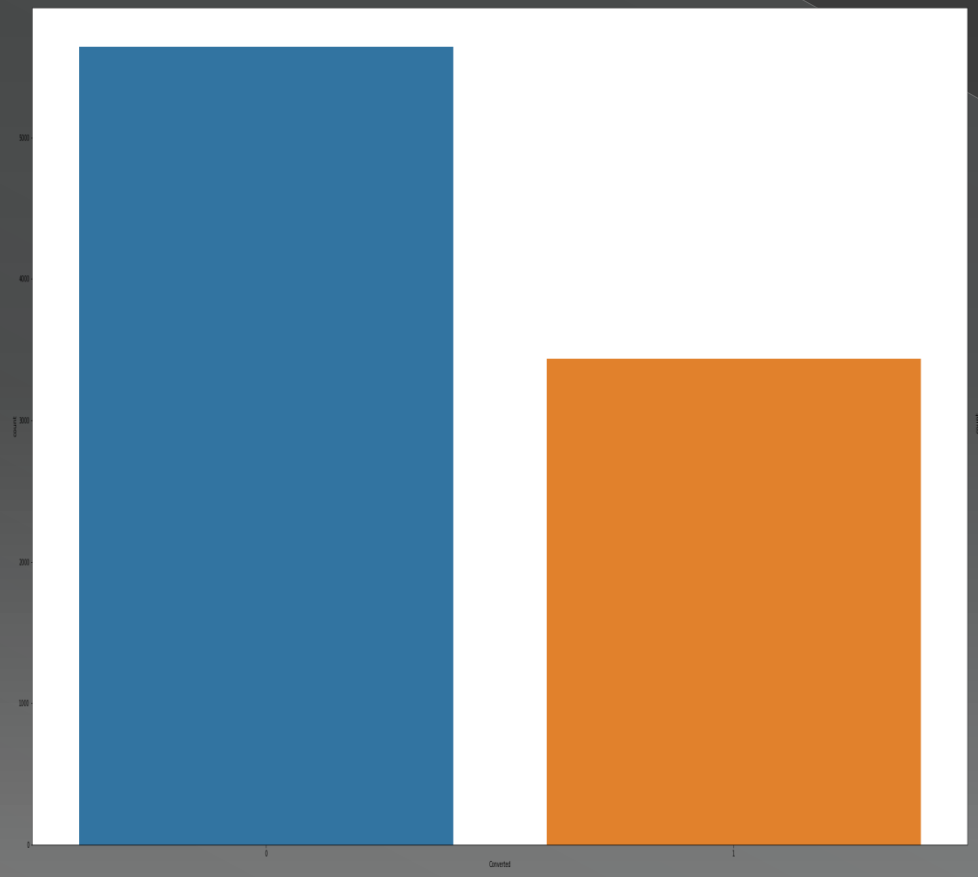
Testing the model with test data set.

Model Evaluation

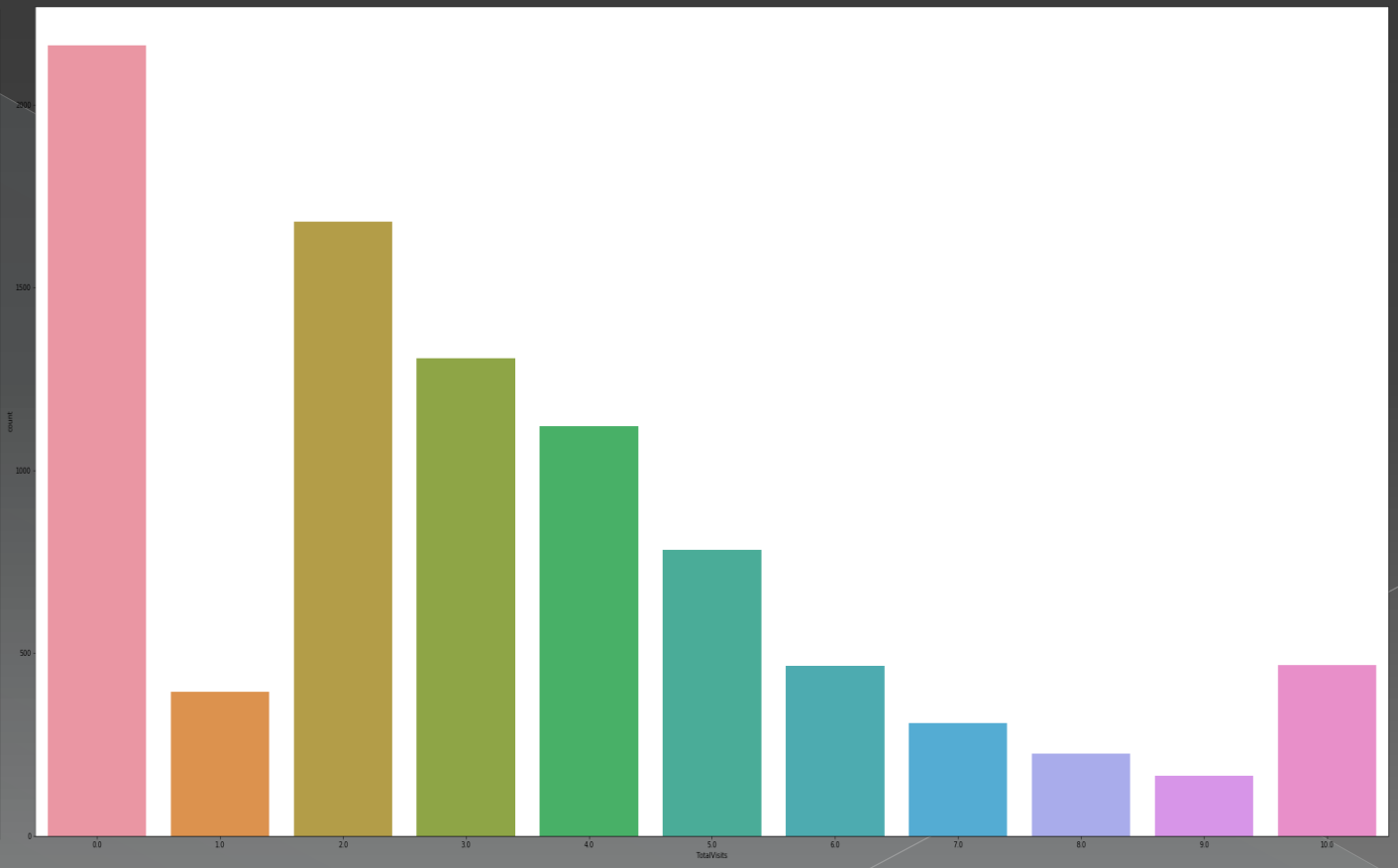
Making Predictions

EDA ANALYSIS

- From the below converted bar graph we can see that out of the total people applied almost 38% of people got converted.
- From the total visit bar graph we can infer that max people who have applied did not visit the site which is a concern followed by 2 visit per person

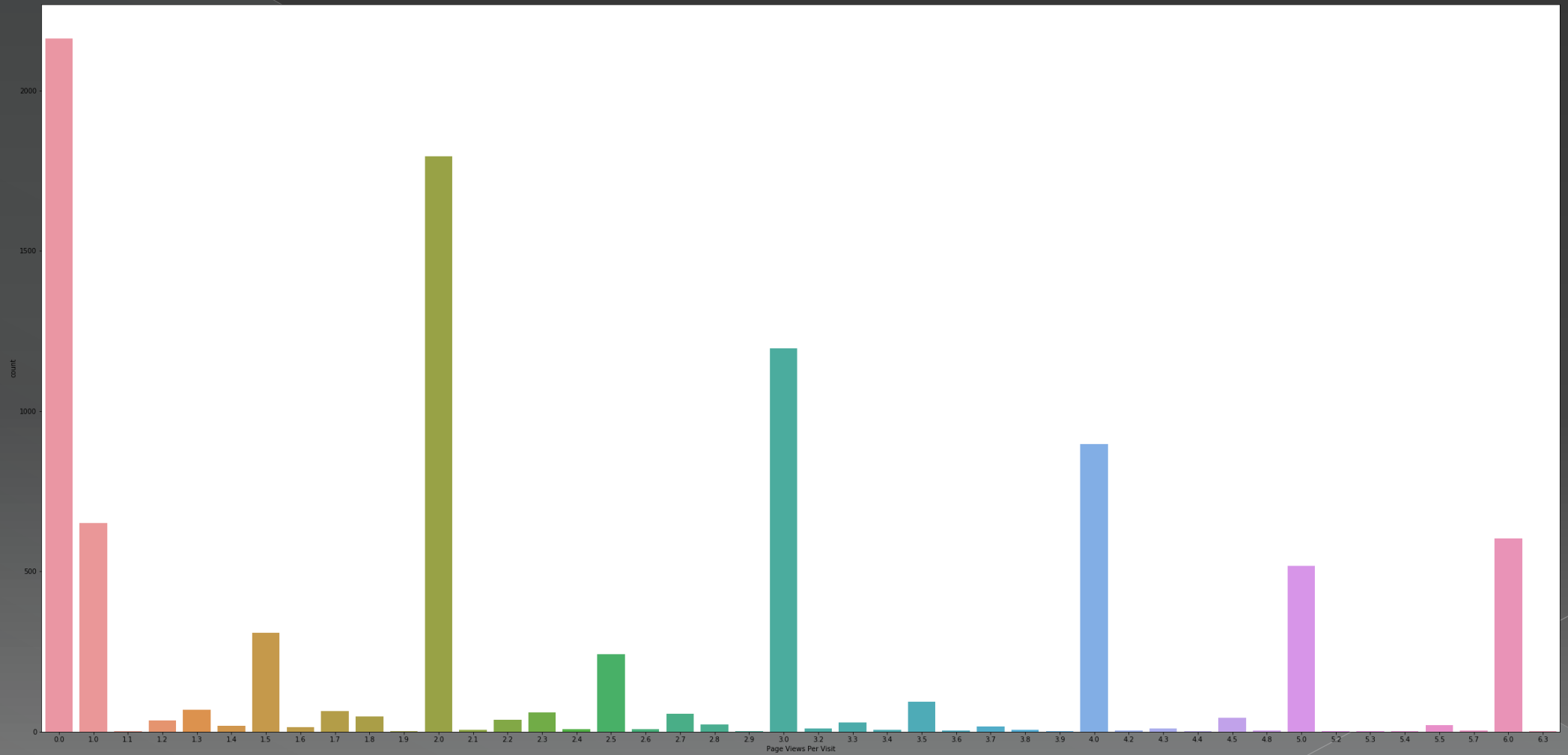


Convert



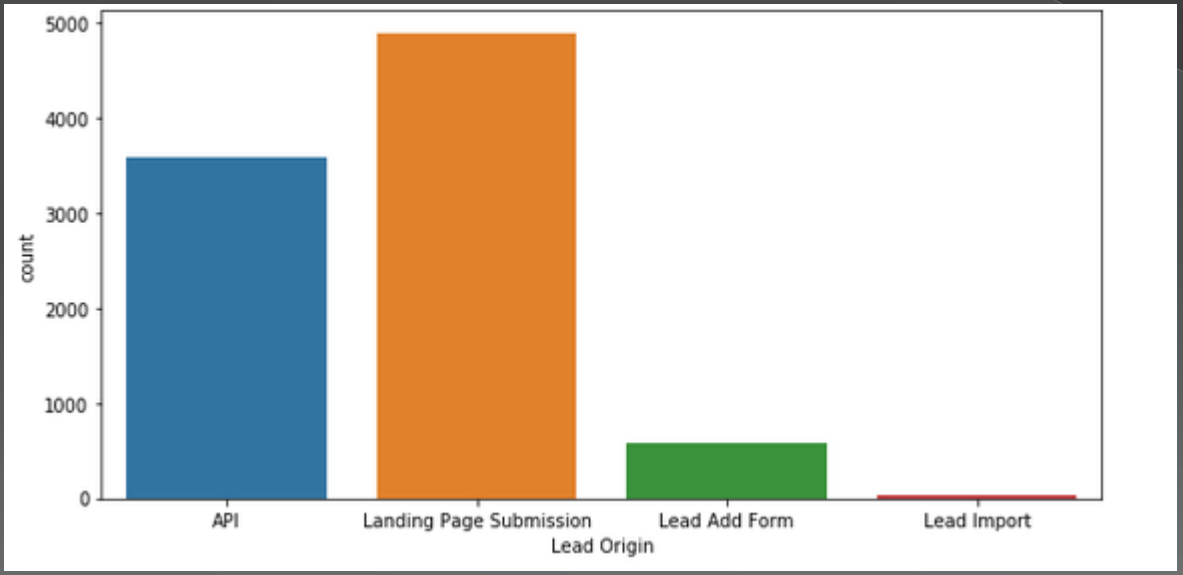
Total Visits

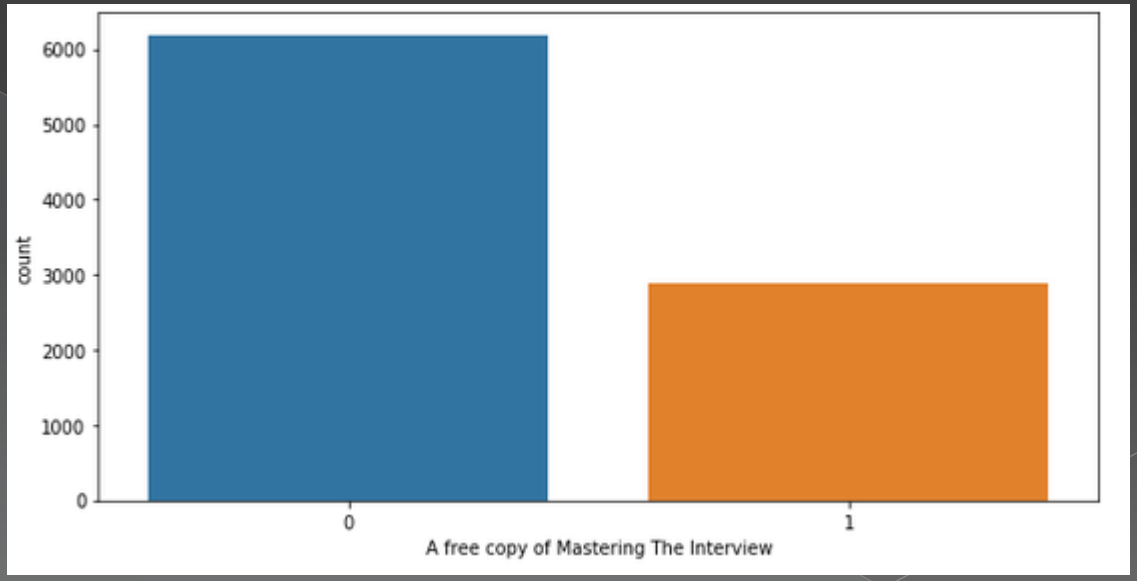
- From the page views per visit bar graph we can infer that max people who have applied did not visit the site which is a concern followed by 2 pages per visit per person



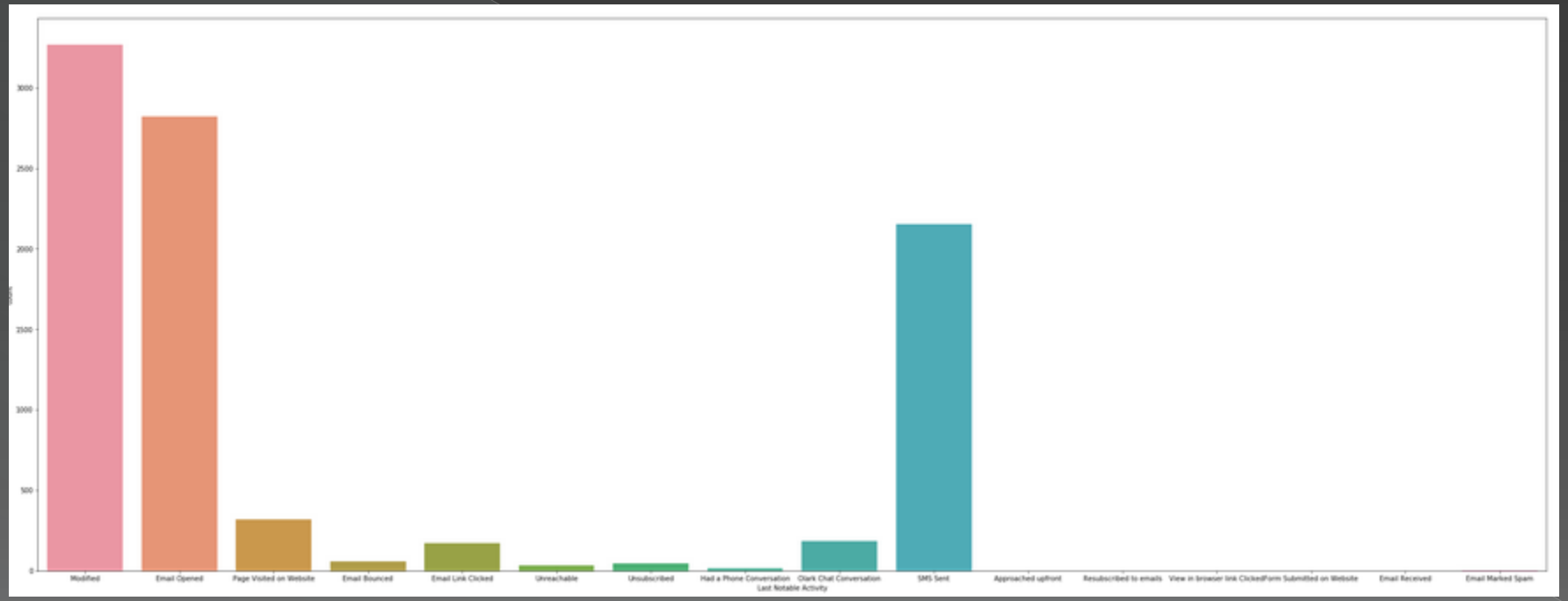
Page Views per Visit

- From the lead origin bar graph we can infer that maximum of the lead are obtained from landing page submission
- From the A free copy of Mastering the interview we can infer that maximum of the people do not want a copy of that.

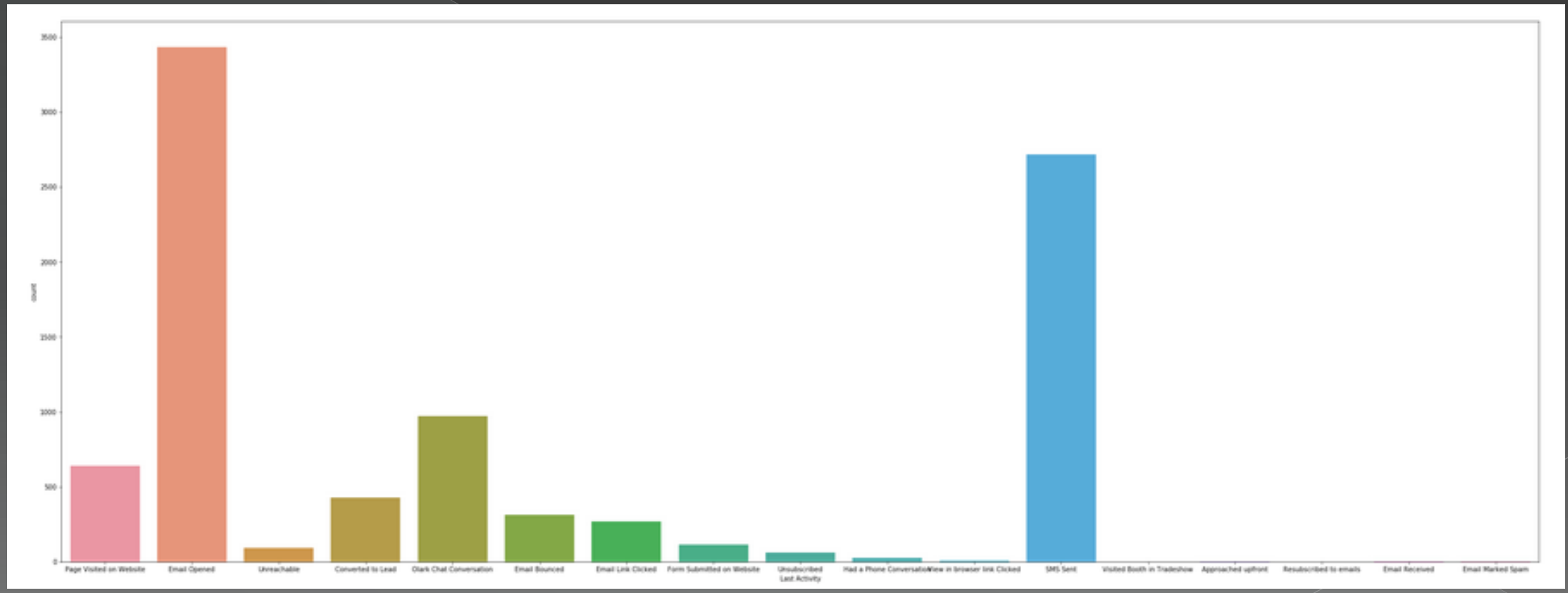




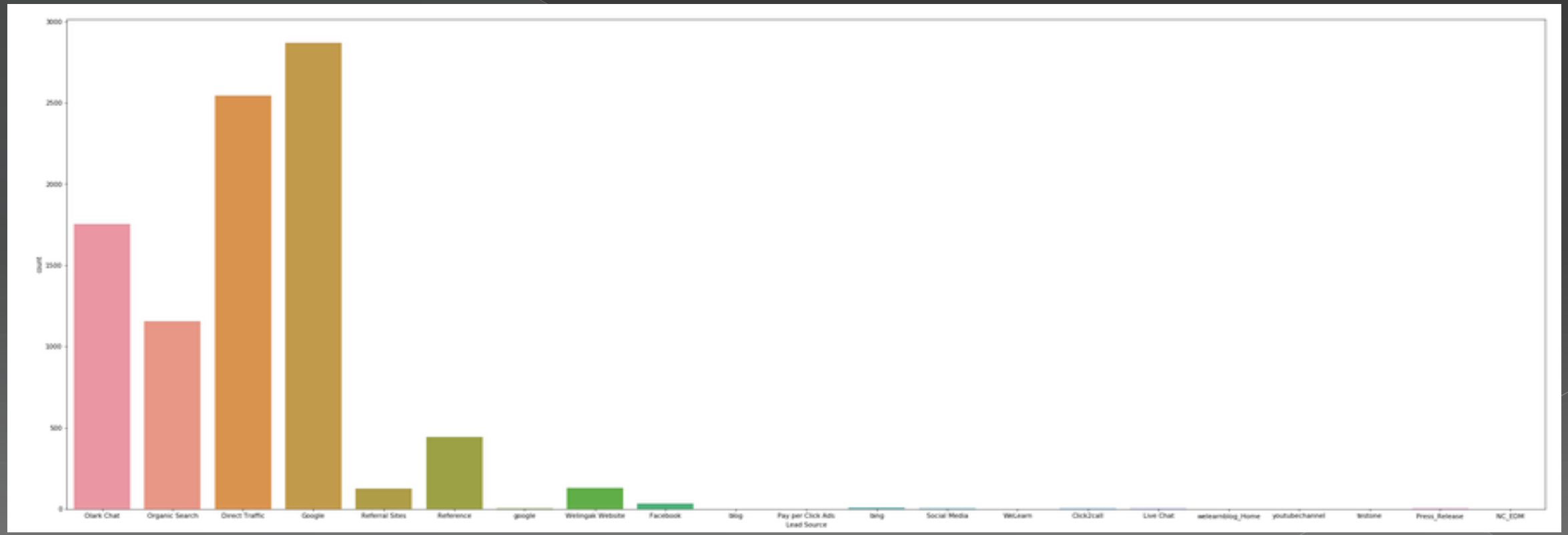
- From the Last Notable Activity bar graph we can infer that max people who have applied have done some modification or have opened their email , its good that people are following their application



- From the Last Activity bar graph we can infer that max people who have applied have opened their email followed by SMS sent



- From the Lead Source bar graph we can infer that max lead are got from Google followed by Direct Traffic



ANALYSIS

- Used Logistic Regression Model , GLM model with combination of feature selection technique called RFE.
- Selected 15 features using RFE and built around 3 models to get the best model, all the features in the model have p values less than 0.5 and the vif of the columns are less than 2. So it is very stable model
- The AUC score (Area under the Curve) is 86%.
- Assigned the score to each of the leads ranging from 0 to 100 based on the conversion probability
- To arrive at 80% conversion rate, considered leads with score greater than equal to 73

CONCLUSIONS

- The top three features effecting the conversion are
  - Lead Source
  - Through Recommendations
  - Last Activity
- The lead scores are based on the conversion probability. Higher the score greater is the conversion probability
- By increasing the value of score cut off the conversion rate increases
- By decreasing the value of score cut off the conversion rate decreases

THANK YOU