# Lead Scoring Case Study

By Krunal Nagda

Problem Statement

- Identifying the most promising leads, that is, the leads that are most likely to convert into paying customers and assigning a lead score to each lead based on the chance of the lead to convert

- Target of lead conversion rate of around 80% has to be achieved

- This could be done by fitting a Logistic Regression model to the given data which would output the lead score for each lead by successfully meeting the target of 80%

Data Handling

- Variables with more than 35% missing values dropped
- Dropped variables having only 1 value throughout or centered around only one value, that is, highly skewed
- Grouped certain categories under the label Others for Lead Source and Last Activity which were in a small proportion of the data
- Dropped missing values of variables with small proportion of missing values, that is, Total Visits, Page Views Per Visit and Last Activity

Data Visualization
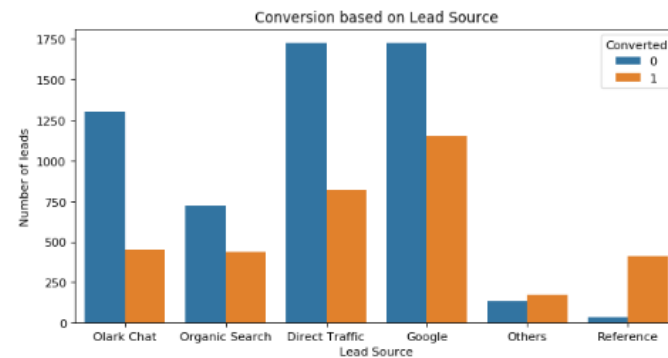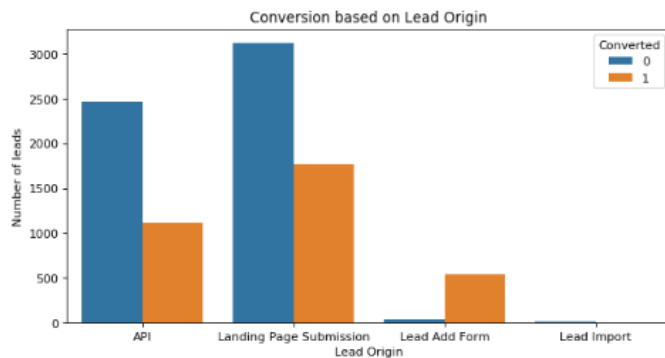
- Lead Add Form has the highest conversion rate as the Lead Origin
- Reference followed by Others has the highest conversion rate as the lead source

Data Visualization
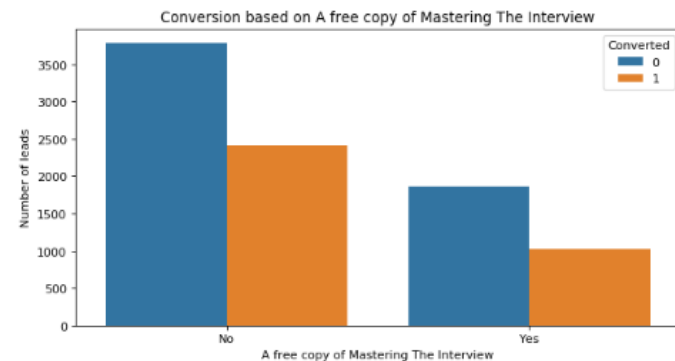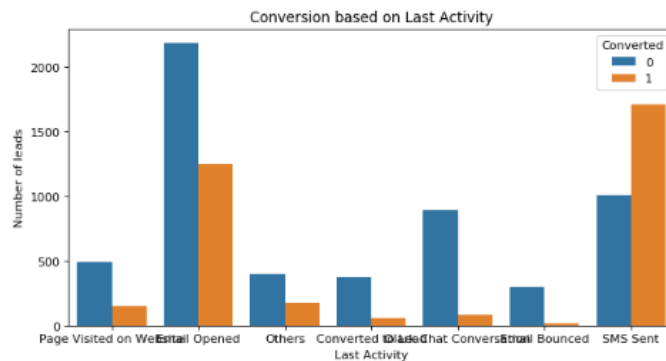
- When Last Activity of customer is SMS Sent, conversion rate is the highest
- Lead conversion rate is slightly higher when a free copy of Mastering The Interview is provided

Data Visualization
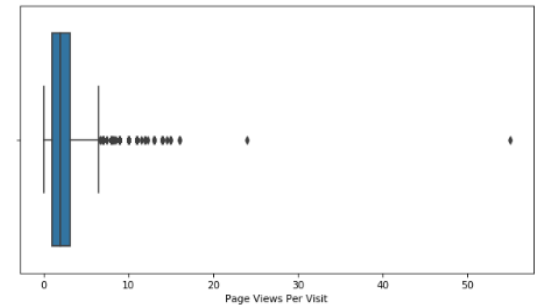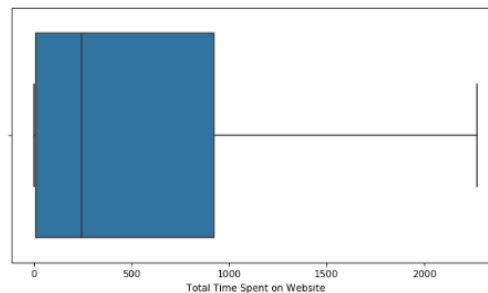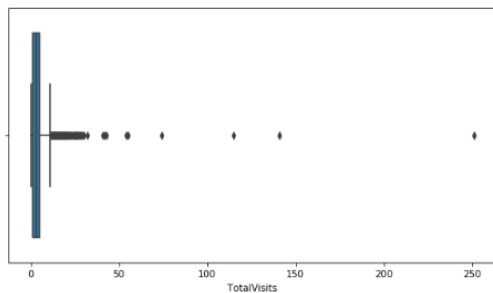
- Clearly, outliers exist in the columns Total Visits and Page Views Per Visit
- Soft capping done at 99th percentile so that the outliers do not affect our required observations

Logistic Regression Model

The logistic regression model is based on the following features, maintaining the 80% conversion rate target:

- Lead Origin:
  - Landing page submission
  - Lead add form
  - Lead import
- Lead Source:
  - Google
  - Olark chat
- Last Activity:
  - Email bounced
  - Email opened
  - Olark chat conversation
  - SMS sent
  - Others
- Total Visits
- Total Time Spent on Website
- Page Views Per Visit

Logistic Regression Model

- For the train data set, the model accuracy and sensitivity is 79%
- For the test data set, the model accuracy is 79% and model sensitivity is 78%
- Here, sensitivity is the overall lead conversion rate and hence it meets the target
- Lead scores are calculated by taking the product of the output of the model, that is probability of conversion, and 100
- 42% of leads have a lead score greater than 35, that is the threshold set to consider a lead to have a chance of being converted
- 33% of leads have a lead score of greater than 50
- 15% of leads have a lead score of greater than 80, that is having a very high chance of converting

Recommendations

- Main leads to target should originate from Lead Add Forms or Lead Imports, lead source should be Olark chat or Google, their last activity is SMS sent and they spend large amount of time on the website
- In a period when there is loads of manpower, the company should focus on the leads having a lead score greater than 35, which is 42% of the population, which could lead upto 90% of lead conversion rate
- In a period when the company wants to focus on new business and wants to contact only the leads who have a very chance of converting, the company should focus on leads with a lead score greater than 80, which is only 15% of the entire population, which could lead to almost a 100% lead conversion rate

Conclusion

- The sales team should focus mainly on these top 33% to get a conversion rate of around 80% instead of going all out and achieving a lower conversion rate
- By marketing more in the most trending Lead Origin (i.e. Google), or communicating more in the SMS mode amongst Unemployed people will help their company grow more
- The advertisement should be made in a tempting way in Google more so that people feel more likely to click on it and register on the landing page.
- The exciting deals can be communicated via SMS, so that they can go through it in their leisure time. By calls, mostly people feel reluctant to pay any attention and can go out of opportunity.

Thank You.