LEAD SCORING CASE STUDY

SUBMISSION

(3rd March 2019)

Group Name:

1. Varun Hasteer
2. Prajwal Rao
3. Mavallapalli Saraschandra
4. Anuj Garg

Business Objective

- 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.

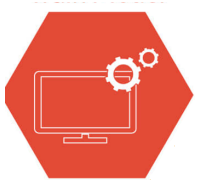
- The company requires the Lead Conversion Rate of around 80%

- Leads with higher scores may have more chance to take course in X Education, If leads having lower score may have less chance to take course in X Education

Methodology

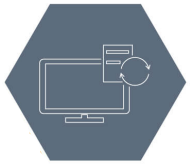
Data Cleansing and Preparation

Exploratory Data Analysis



Model Building

Model Evaluation



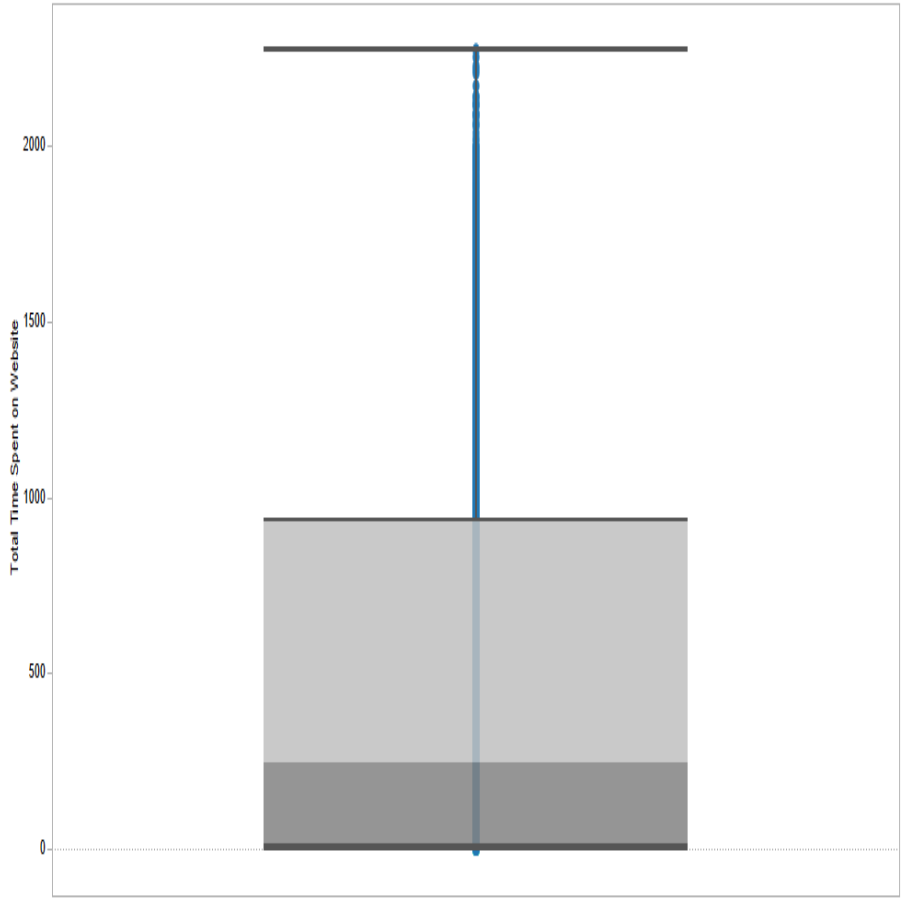
Model Testing

EDA

Lead Analysis

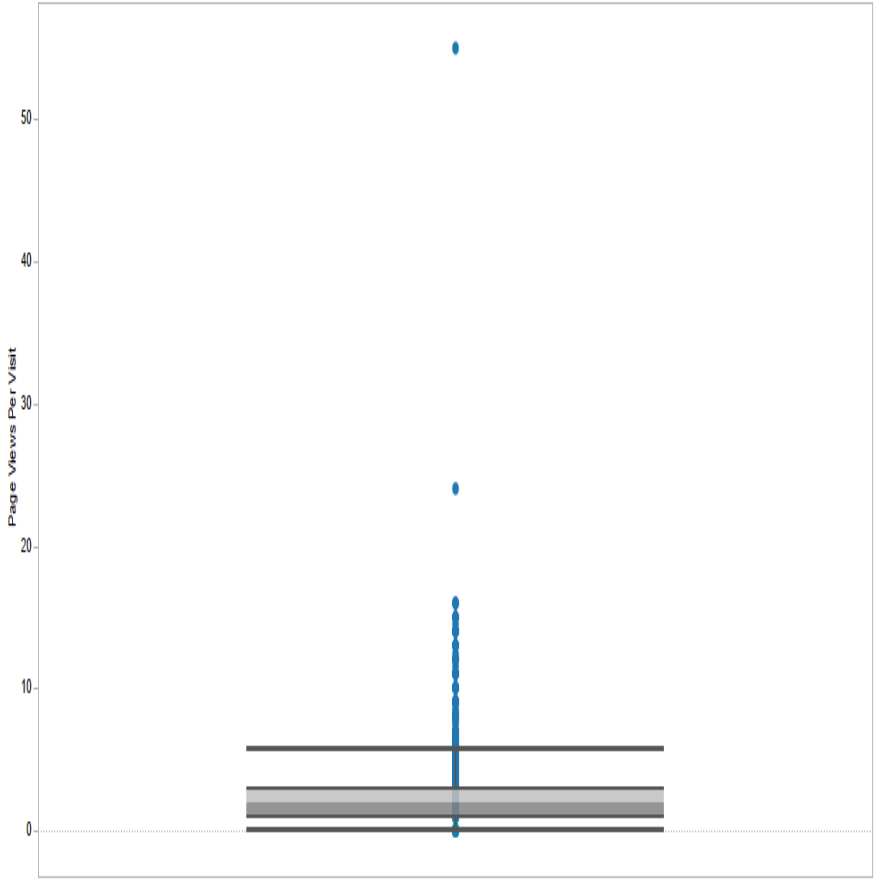
Leads-Converted?	Number of Leads	Avg. Page Views Per Visit	Avg. Total Time Spent on Website	Avg. Total Visits
No	5,679	2.37	330	3.33
Yes	3,561	2.35	739	3.63

Outlier Analysis on Total Time Spent on Website



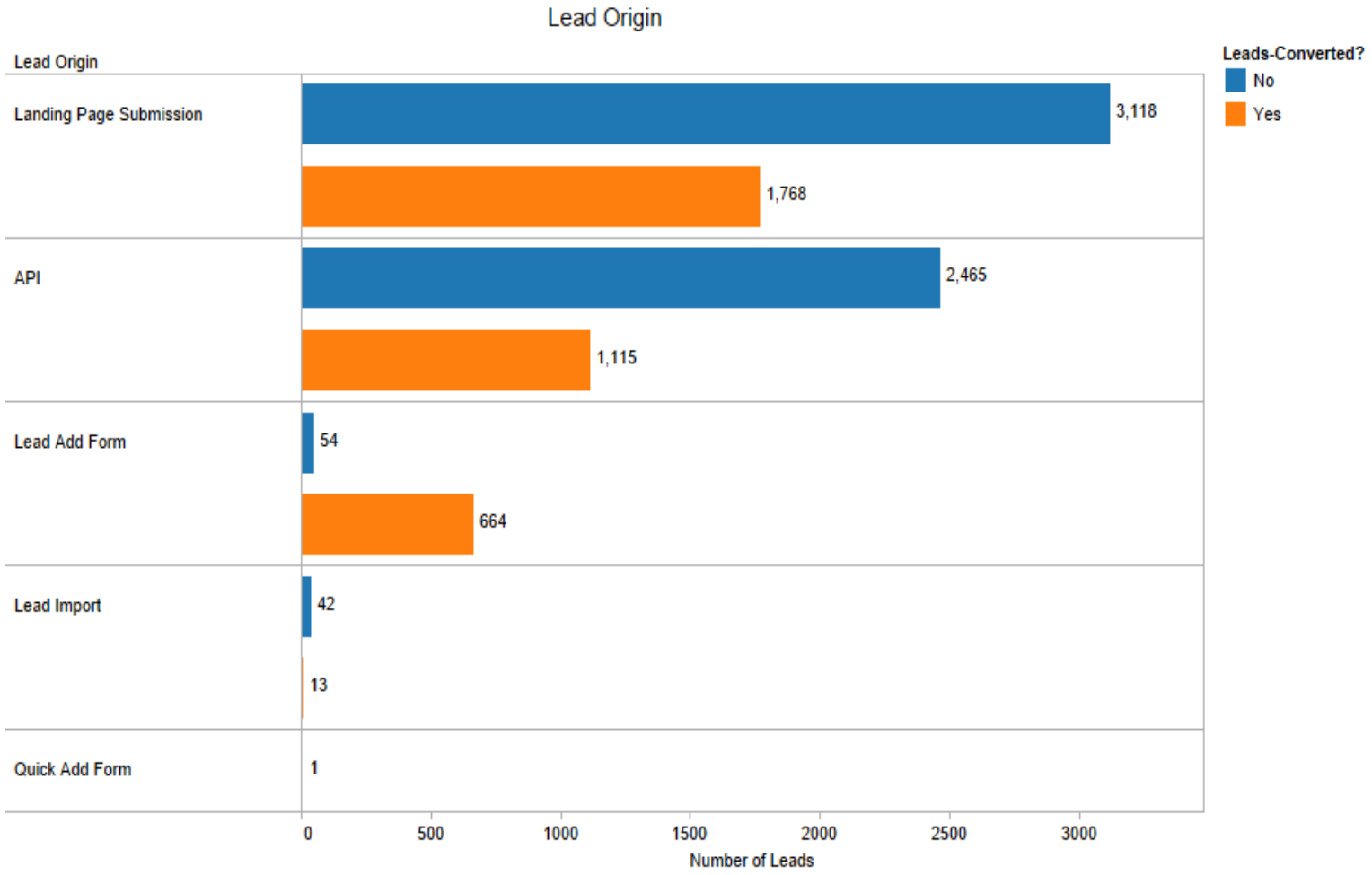
Sum of Total Time Spent on Website. Details are shown for Lead Number.

Outliers Analysis on Page Views Per Visit



Sum of Page Views Per Visit. Details are shown for Lead Number. The view is filtered on sum of Page Views Per Visit, which keeps non-Null values only.

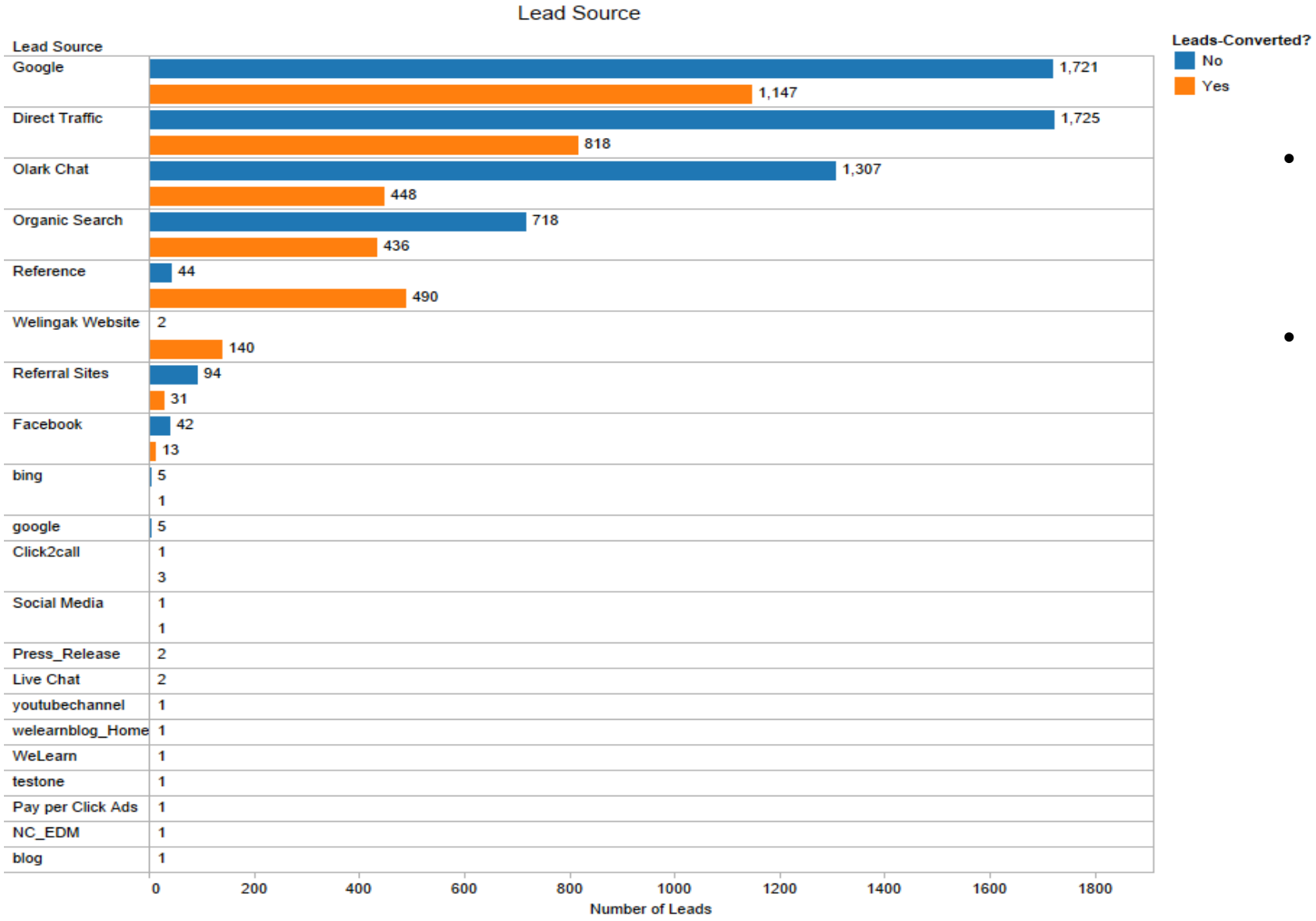
EDA …. cont.



Sum of Number of Leads for each Leads-Converted? broken down by Lead Origin. Color shows details about Leads-Converted?. The marks are labeled by sum of Number of Leads.

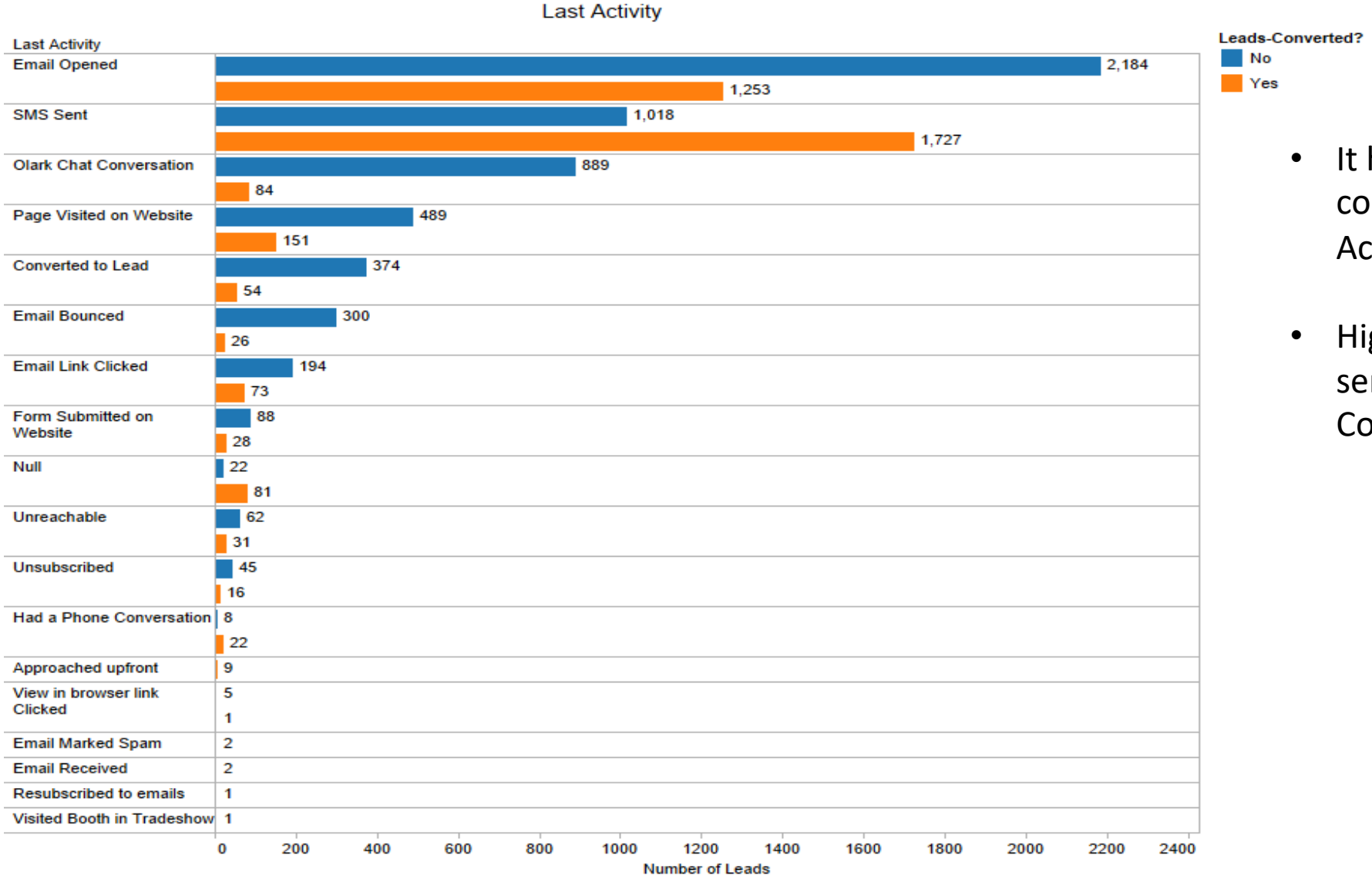
- It has been observed that more Conversion happened in “Landing Page submission” type of Lead Origin
- Higher conversion rate for the leads originating from “Lead Add Form”

EDA …. cont.



- It has been observed that more conversion happened in “Google” as type of Lead Source
- Higher conversion rate for the leads from the sources ”Reference” and “Welingak Website”

EDA …. cont.



Sum of Number of Leads for each Leads-Converted? broken down by Last Activity. Color shows details about Leads-Converted?. The marks are labeled by sum of Number of Leads.

- It has been observed that more conversion happened with Last Activity as “Email Opened”
- Higher conversion rate in the ”SMS sent” and “Had a Phone Conversation” categories

Model Building

- 7:3 data has been splitted to do train and test while building the prediction model
- Feature Elimination has been done in RFE(Automated Feature Elimination), VIF(Manual Feature Elimination) to build better model
- Total 17 features are selected

Features
Last Notable Activity_Modified
Last Activity_Olark Chat Conversation
Last Activity_Email Bounced
Do Not Email
Lead Source_Olark Chat
Lead Origin_Lead Add Form
Last Notable Activity_Olark Chat Conversation
Last Activity_Converted to Lead
Lead Source_Welingak Website
Total Time Spent on Website
Last Activity_Page Visited on Website
Last Notable Activity_Email Opened
Last Activity_Form Submitted on Website
Lead Source_Referral Sites
Last Notable Activity_Email Link Clicked
Lead Origin_Lead Import
Last Activity_Had a Phone Conversation

Model Evaluation- Step 1

- 79.08 % Model Accuracy Has been observed
- Confusion Matrix

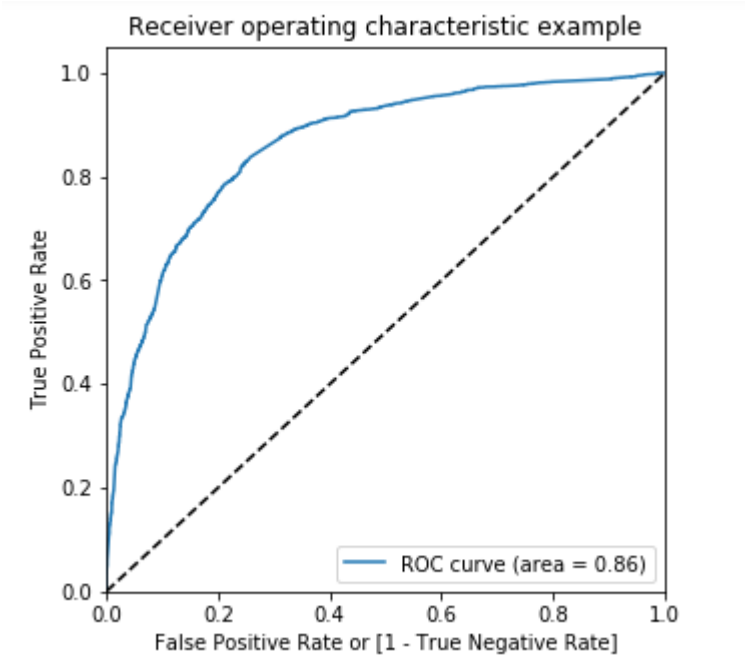
Actual/ Predicted	Converted	Not Converted
Converted	3012	873
Not Converted	429	1910

- Metrics in Model Evaluation

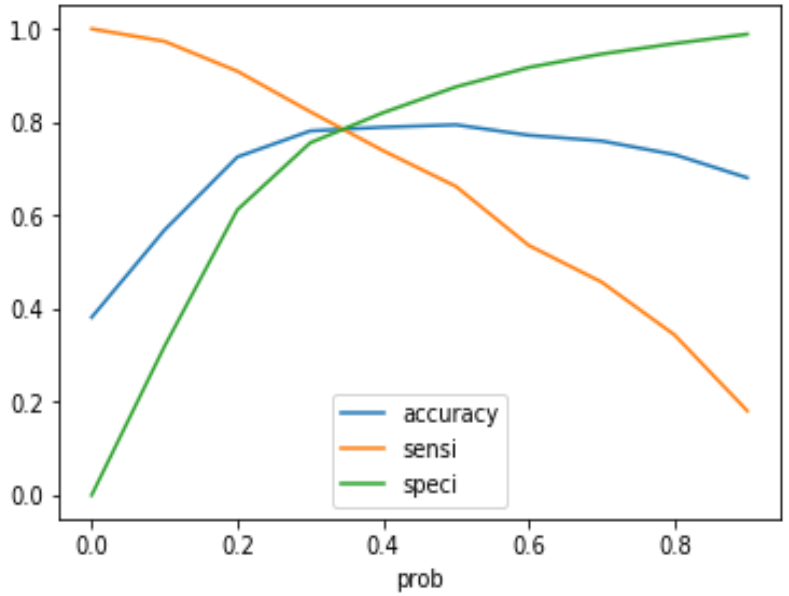
Metric	Percentage
Sensitivity	81.6%
Specificity	77.5%
False Positive Rate	22.4%
Positive Predictive Rate	68.63%
Negative Predictive Rate	87.53%

Metric	Percentage
Precision	87%
Recall	34%

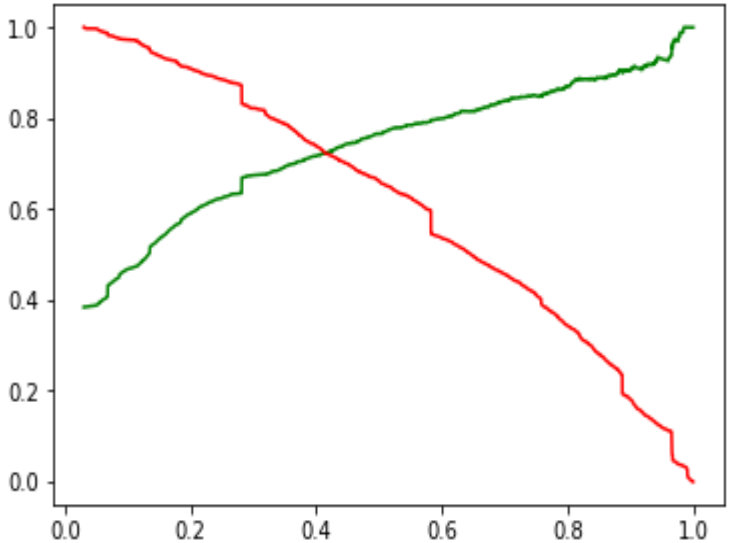
Model Evaluation- Step 2



86% area is under curve denotes better Model



0.3 probability is the optimum cut-off rate



Precision & Recall Trade-off found at 0.4 Probability

Prediction results on Test dataset

- 78 % Model Accuracy Has been observed
- Confusion Matrix

Actual/ Predicted	Converted	Not Converted
Converted	1266	411
Not Converted	193	902

- Metrics in Model Evaluation

Metric	Percentage
Sensitivity	82%
Specificity	75%
False Positive Rate	25%
Positive Predictive Rate	69%
Negative Predictive Rate	87%

Metric	Percentage
Precision	69%
Recall	82%

Conclusion

- Using our model the company can achieve **Lead Conversion Rate** : **82 %**
- We have also scored each Lead on scale 0 to 100. Marketing director can assign importance of the Lead based on score.

Top 3 Variables contributing more Probability

- Lead Origin
- Lead Source
- Last Activity

Top 3 Categorical Variables contributing more Probability

- Lead Origin-Lead Add Form
- Lead Source-Welingak website
- Last Activity-Had a Phone

Strategic Prediction Approach for Good Lead Scoring:

- Decreasing Cut-Off Values
- Focus on Lead Origin and Lead Source
- Follow-up Methods for High Conversion rates