LEAD SCORING CASE STUDY

Vignesh Satheesh
Asha Nair

OBJECTIVE

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein a lead score is assigned 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.
- A target lead conversion rate to be around 80%.

CASE STRATEGY

- Data Understanding
- Data Cleaning
- EDA
- Data Preparation
- Modelling

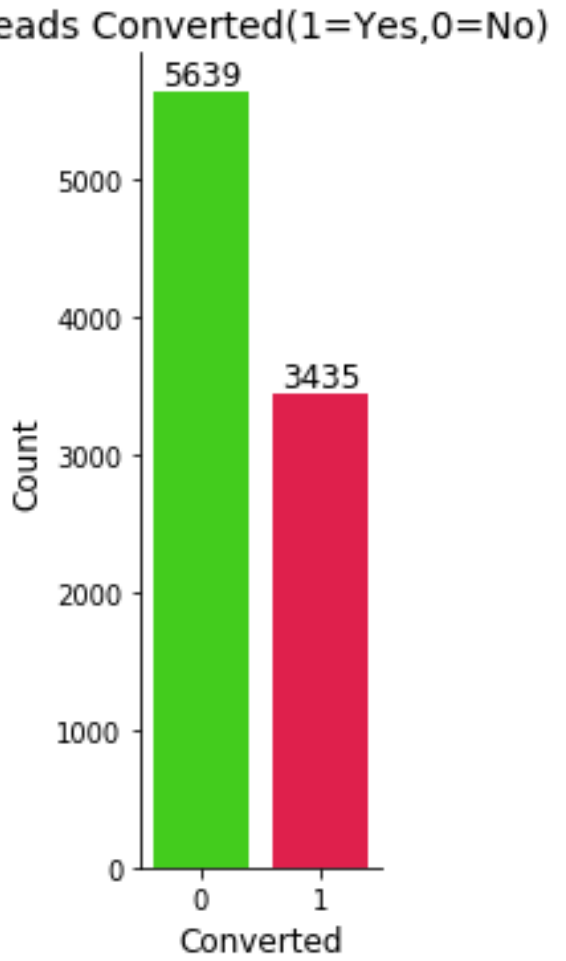
DATA UNDERSTANDING & CLEANING

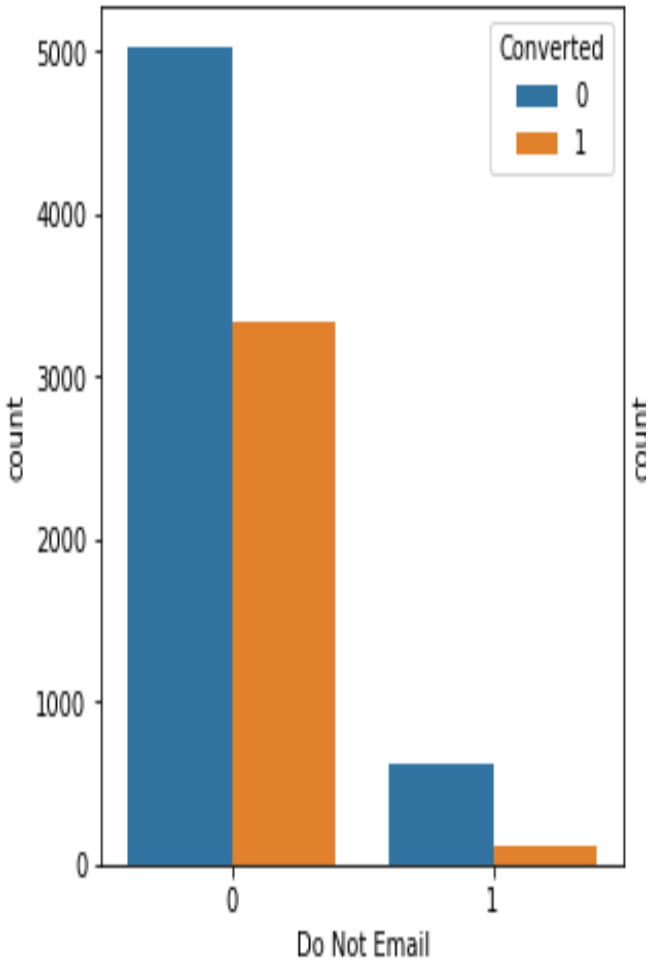
- Over 9000 columns and 37 columns were found in the data
- It was found that a number of questions were not answered by the customers and the sales teams.
  - A number of rows with ‘Select’ values were found.
  - These values were replaced with null values
- The data was found to have more than 25% null values.
- Columns with more than 70% null values were dropped to retain as much data as possible to arrive at a good model.
- The data with high variation were dropped from the original dataframe.
- Remaining data with more than 2% null values were imputed with mode or median values
- Some null values were replaced with appropriate names.
- The null values lesser than 2% were dropped as they were insignificant.
- Approximately, 98% data was retained after cleaning

DATA PREPARATION

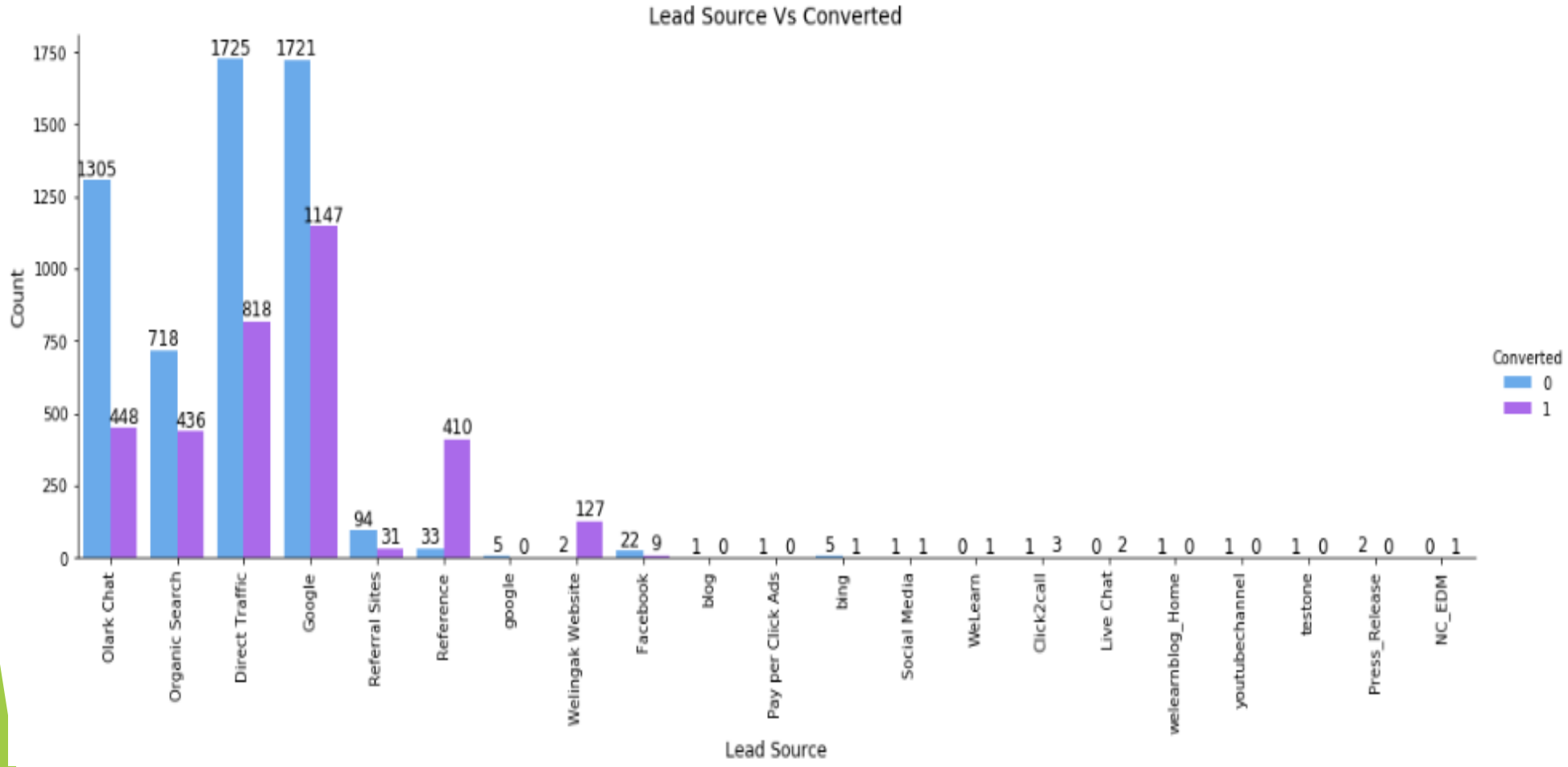
- The data was prepared for modelling:
  - The columns with Yes and No values were converted to binary values
  - Dummy variables were created for categorical variables.
- The data was split into two
  - Train data-70% of original data
  - Test data: 30% of original data.
- The X and y variables were created :
  - X – Dataframe without Prospect ID and Converted.
  - y – ‘Converted’ , the target variable.
- The standard scaler was used to standardise the numerical variables.

EXPLORATORY DATA ANALYSIS.





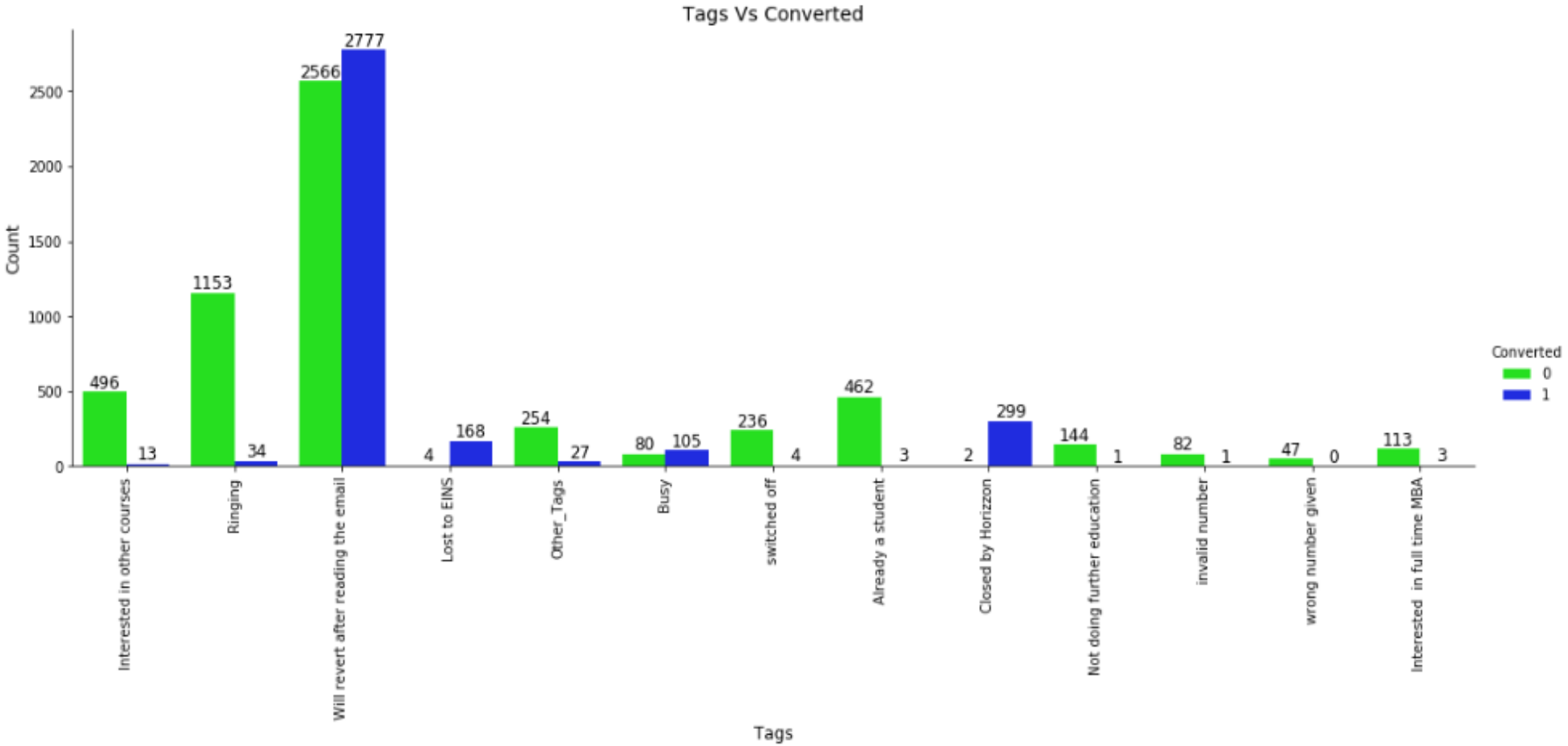
- Initial Conversion rates : 38%
- Aim : Improve conversion rates to around 80 %
- Most of the leads were from people who choose Do Not Email feature as No



Inference

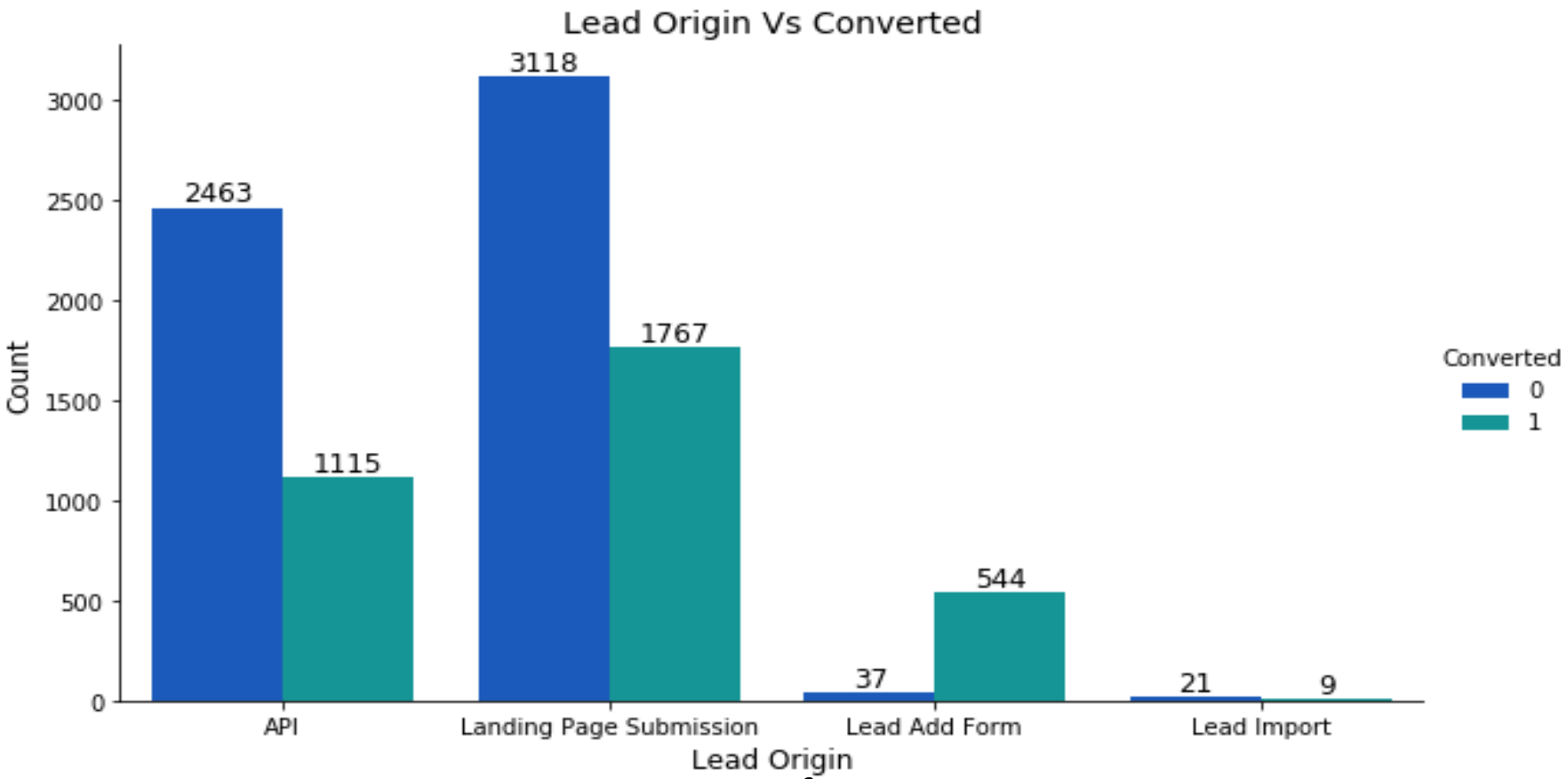
1. Google and Direct traffic generates maximum number of leads.
2. Conversion Rate of reference leads and leads through welingak website is high.

To improve overall lead conversion rate, focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website



Inference

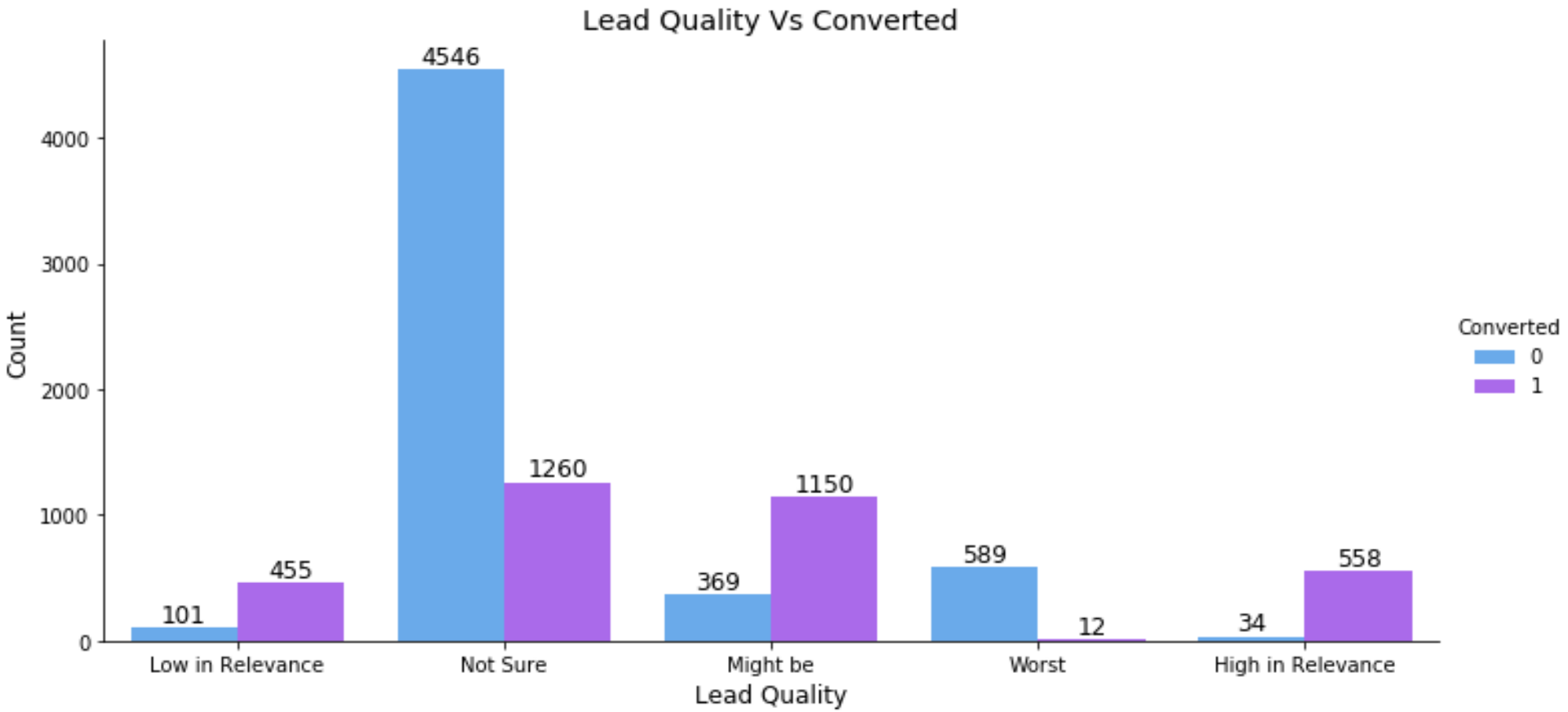
1. Conversion Rate of Will revert after reading the email, Lost to EINS and Closed by Horizon are the highest
2. Some values were replaced into Others Tags



Inference

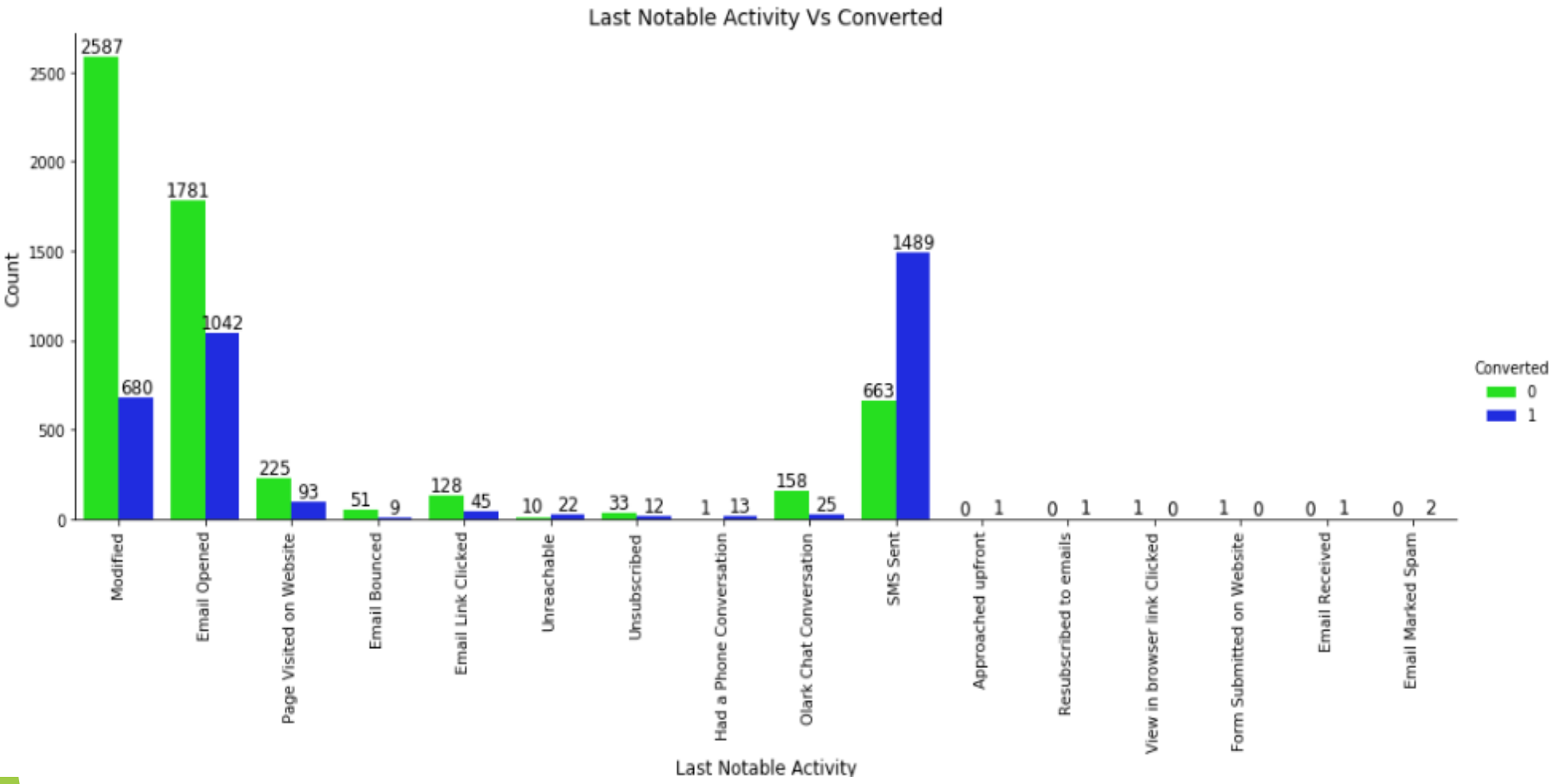
1. API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
2. Lead Add Form has more than 90% conversion rate but count of lead are not very high.
3. Lead Import are very less in count.

To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



Inference:

1. The conversion of Might be, Low in relevance and High in relevance are the highest
2. Although the Not Sure value has some conversions, it is low as compared to others



Inference

1. Most of the lead have their Email opened as their last activity.
2. Conversion rate for leads with last activity as SMS Sent is almost 60%.

FINAL MODEL

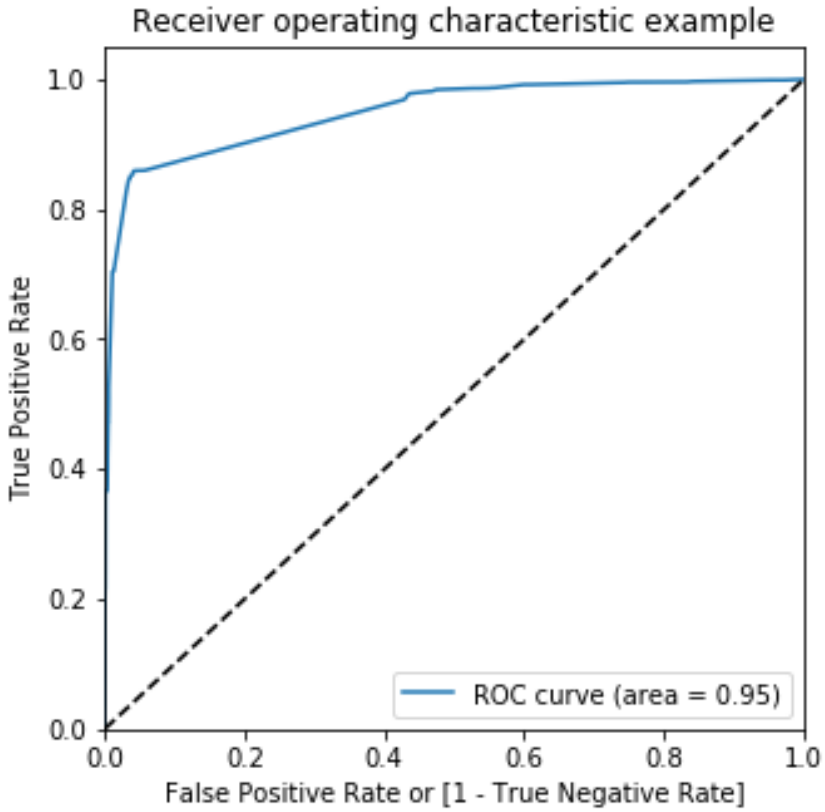
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0888	0.216	-9.654	0.000	-2.513	-1.665
Do Not Email	-1.3012	0.212	-6.134	0.000	-1.717	-0.885
Lead Origin_Lead Add Form	1.0894	0.363	3.001	0.003	0.378	1.801
Lead Source_Welingak Website	3.4138	0.818	4.173	0.000	1.810	5.017
What is your current occupation_Working Professional	1.3403	0.291	4.602	0.000	0.769	1.911
Tags_Busy	3.8040	0.330	11.532	0.000	3.157	4.450
Tags_Closed by Horizzon	7.9562	0.763	10.433	0.000	6.461	9.451
Tags_Lost to EINS	9.1785	0.754	12.177	0.000	7.701	10.656
Tags_Ringing	-1.6947	0.337	-5.036	0.000	-2.354	-1.035
Tags_Will revert after reading the email	3.9665	0.229	17.311	0.000	3.517	4.416
Tags_switched off	-2.2882	0.587	-3.900	0.000	-3.438	-1.138
Lead Quality_Not Sure	-3.3406	0.128	-26.026	0.000	-3.592	-3.089
Lead Quality_Worst	-3.7624	0.850	-4.426	0.000	-5.428	-2.096
Last Notable Activity_SMS Sent	2.7406	0.120	22.847	0.000	2.506	2.976

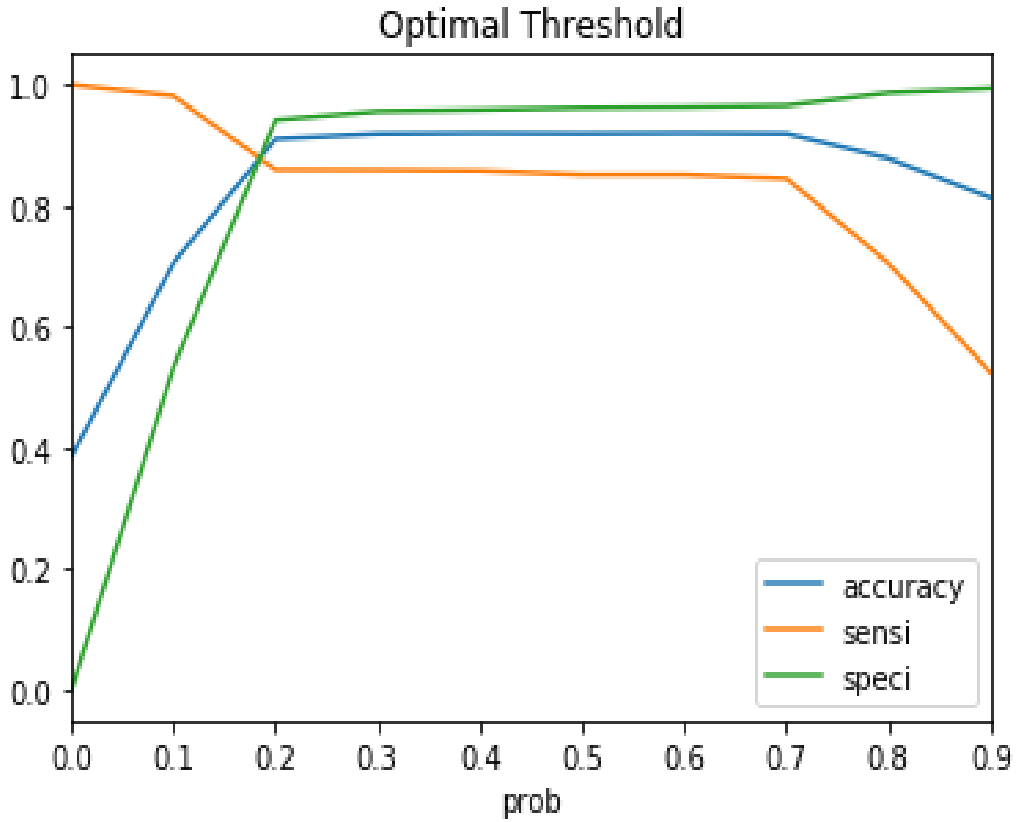
Final Regression Model with 12 variables

	Features	VIF
8	Tags_Will revert after reading the email	2.81
10	Lead Quality_Not Sure	2.76
1	Lead Origin_Lead Add Form	1.62
7	Tags_Ringing	1.54
12	Last Notable Activity_SMS Sent	1.52
2	Lead Source_Welingak Website	1.36
3	What is your current occupation_Working Profes...	1.26
5	Tags_Closed by Horizzon	1.15
4	Tags_Busy	1.11
0	Do Not Email	1.10
9	Tags_switched off	1.10
6	Tags_Lost to EINS	1.05
11	Lead Quality_Worst	1.03

VIF table with 12 features

TRAIN DATA: ROC & THRESHOLD

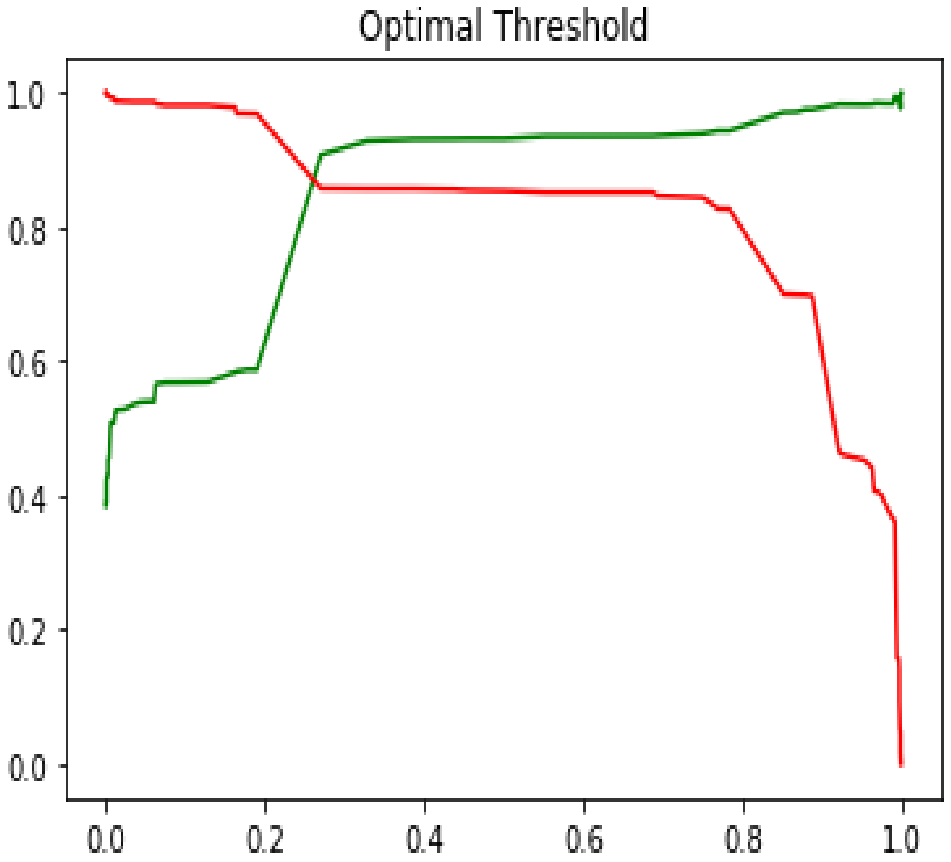




1. The ROC(Receiver Operating Characteristic Curve) shows the tradeoff between the True Positive Rate and False Positive Rate, in our case pretty high, which is good for us
2. The Threshold that we choose was 1.8, according to the plot that we got

3. Later we used it to determine the probability

TEST DATA: THRESHOLD



- For Precision-Recall tradeoff, we took the optimal threshold as 0.25
- Used this cut-off in a new column called as final predicted where a probability greater than 0.25 would be converted as our hot lead

RESULTS

TRAIN DATA SET	
Overall accuracy after building model	0.92
Accuracy after VIF	0.92
Sensitivity	0.85
Specificity	0.96
False Positive Rate	0.038
Positive Predictive Value	0.93
Recall	0.91
Probability Threshold/Optimal Cut off	1.8

TRAIN DATA SET AFTER CUT OFF-1.8	
Overall accuracy after building model	0.92
Accuracy after VIF	0.92
Sensitivity	0.97
Specificity	0.57
False Positive Rate	0.42
Positive Predictive Value	0.58
Negatively Predicted Value	0.97

TEST DATA SET	
Overall Accuracy	0.91
Sensitivity	0.84
Specificity	0.95
Precision	0.9
Recall	0.84
F1 Score	0.87

CONCLUSION

- **Top three variables:**
  - Tags_Lost to EINS
  - Tags_Closed by Horizzon
  - Tags_Will revert after reading the email

Top 10 hot leads.

	Prospect ID	Converted	Converted_prob	final_predicted	Lead_Score
1361	2118	1	0.999049	1	99
2452	3817	1	0.997423	1	99
265	7517	1	0.990228	1	99
2168	6040	1	0.990228	1	99
549	3285	1	0.997423	1	99
2455	5685	1	0.990228	1	99
1502	6026	1	0.999049	1	99
1174	5804	1	0.999132	1	99
1176	7805	1	0.990228	1	99
2161	2664	1	0.996691	1	99