LEAD SCORE CASE STUDY

PREPARED BY

SAI KUMAR MIRYALA

BEN BOBY

SATHIYASURIYA R

Problem statement:

- Ed-tech company named X Education sells online courses to industry professionals. Many interested candidates land on their websites to get more details.
- Once, interested person fills basic details. X Education, gets a data for leads, its lead conversion rate is very poor approximately 30%.
- To make good business out of it, the company wants to identify the most potential leads, also known as ‘Hot Leads’.
- If they are able to identify this set of leads, the lead conversion rate increases as the sales and marketing team will now be focusing more on communicating with the potential leads instead of making calls to everyone

Solution:

- The company wants to identify the most potential leads.
- to build a Model which identifies the hot leads.
- For future use user cases need to generate a model

Steps in Case study :

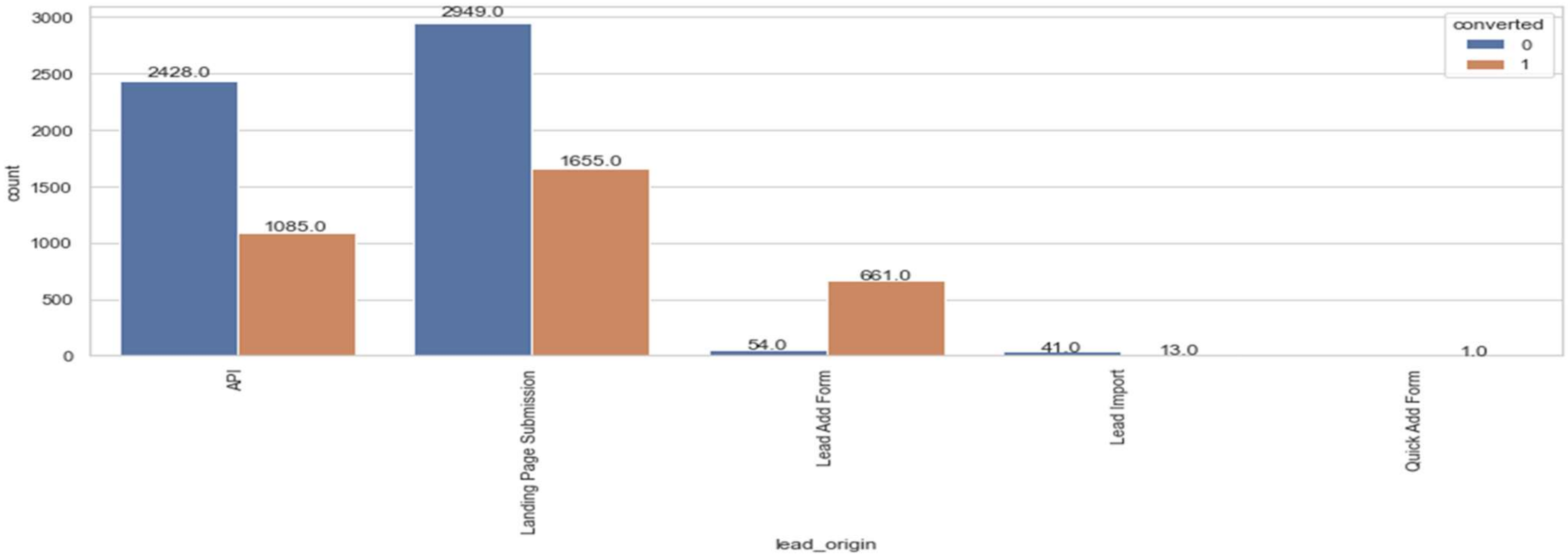
- Data Inspection.
- Data Cleaning.
- Exploratory Data Analysis.
- Univariate Analysis and Bivariate Analysis.
- Data Preparation
    1. Converting some binary variables
    2. Creating Dummy variables for the categorical features
    3. Splitting the data into train and test set
    4. Scaling the features
- Feature Selection Using RFE.
- Model Building
- Checking for P-values and VIF values.
- Making Prediction on the Train set
- Choosing an arbitrary cut-off probability point of 0.5 to find the predicted labels.
- Making the Confusion matrix
- Plot ROC Curve
- Find the optimal cut off point
- Evaluation of model
- Precision & Recall
- Making final predictions on test data
- Recomendations

Data Set info

- The Data set contains 9240 rows and 37 columns
- Prospect Id , Lead Number are unique Columns
- Input File contains lot of columns with Null values, there are specific columns where the value is ‘select’.
- Dropped the columns with missing values greater than 40%.(i.e. 'How did you hear about X Education','Lead Quality',’Lead Profile','Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score', 'Asymmetrique Profile Score’).

Exploratory Data analysis(EDA)

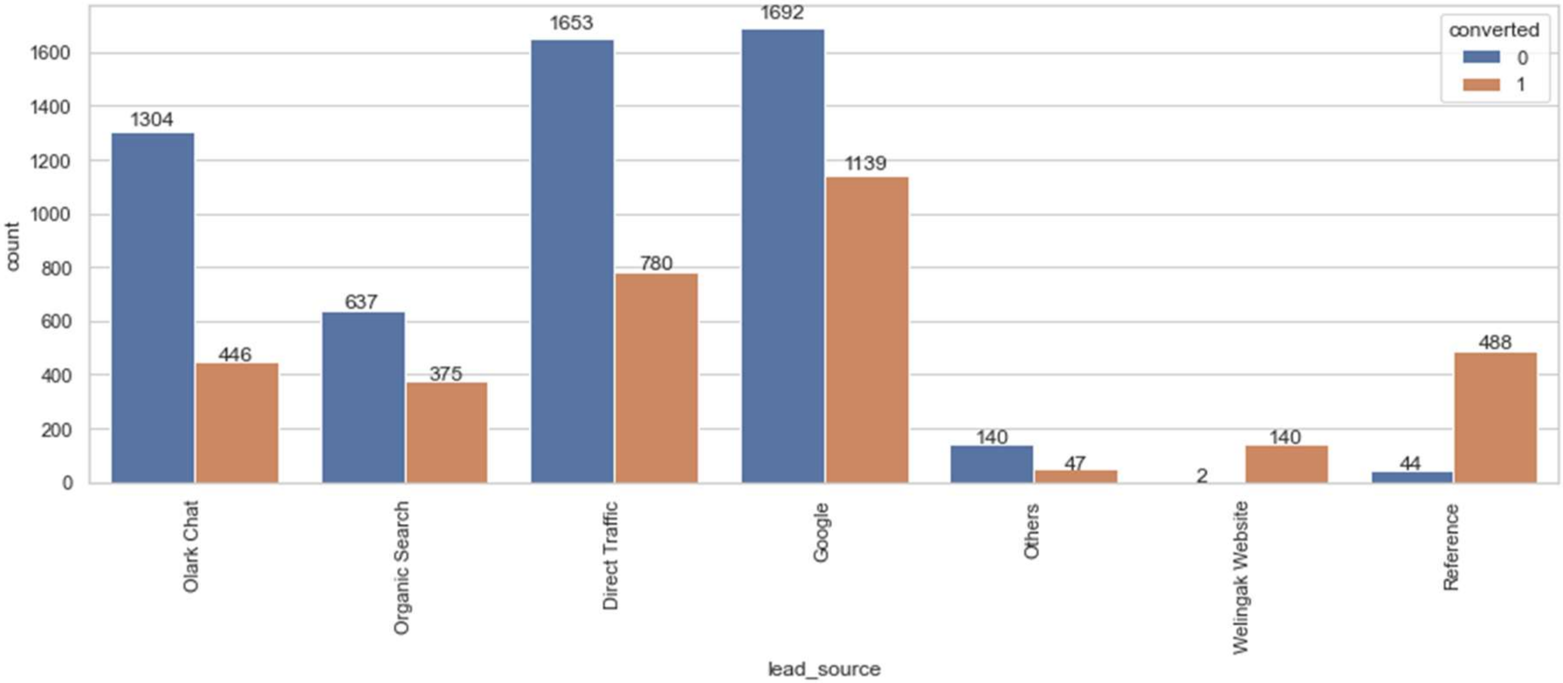
Lead Origin:
Customers identified via Lead Add Form tend to have a high rate of conversion.



Exploratory Data analysis(EDA)

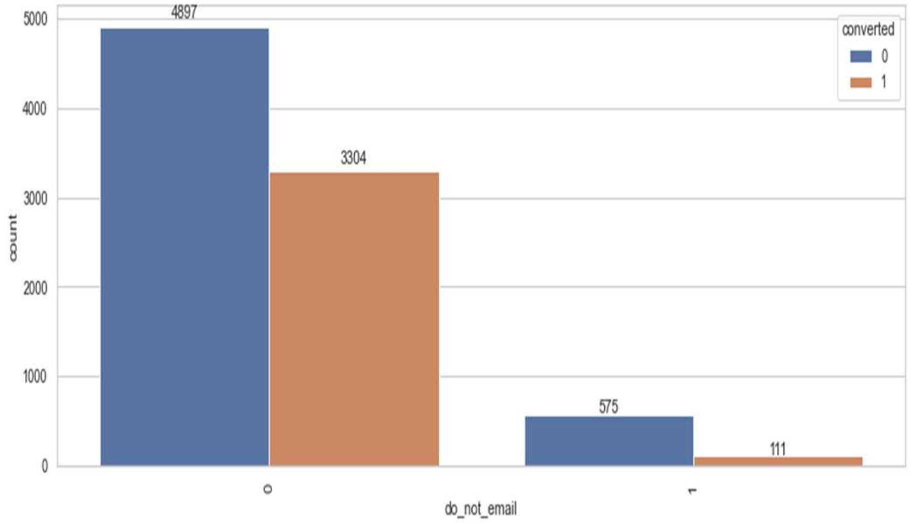
Lead source:

1. Source of Lead via Google and Direct Traffic have high negative conversion as compared to others.
2. Leads coming via Reference have the highest conversion rate.



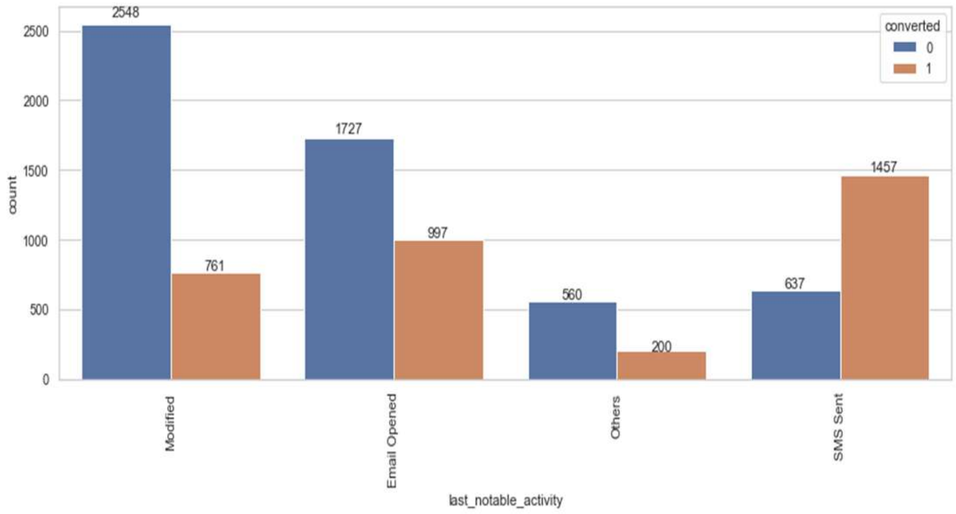
EDA (Contd..)

Email: Leads coming via email about the course shown high conversion rate.



Last Notable activity:

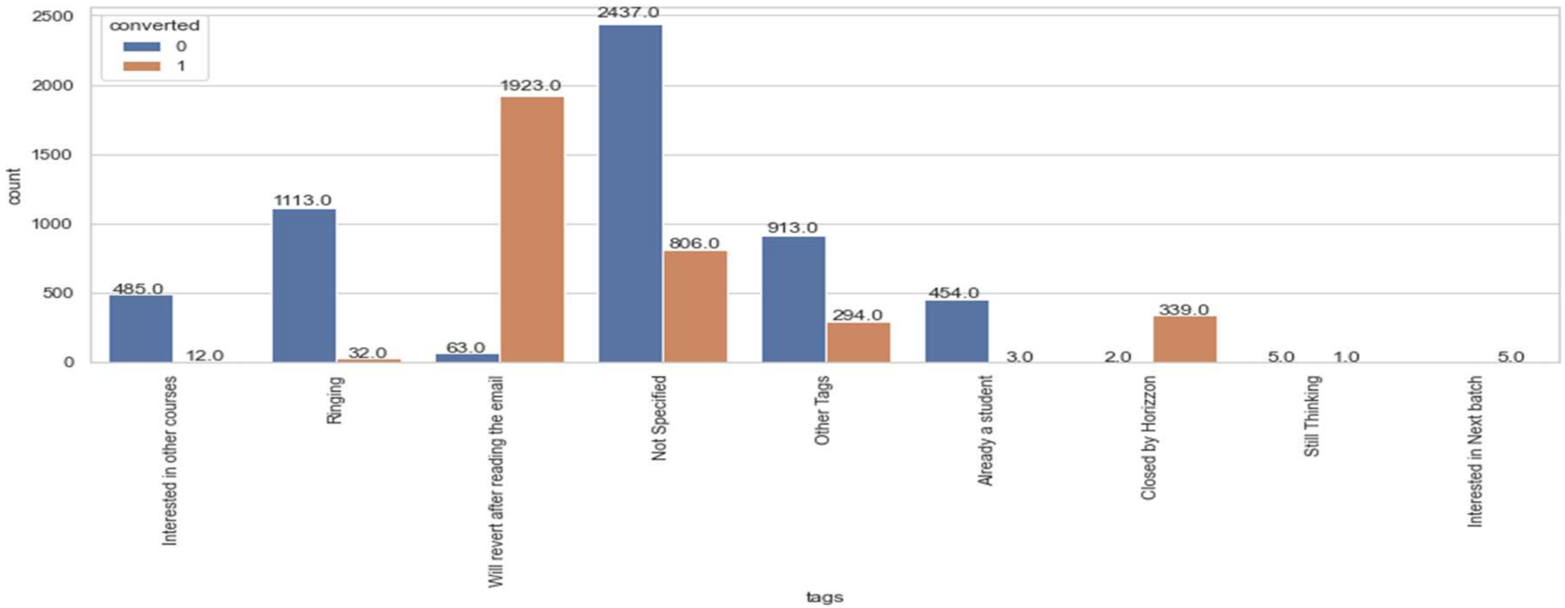
1. For individuals with Last Notable activity as SMS sent, 1457 people converted.
2. For individuals with Last Notable activity as Email opened , 997 people converted



EDA(contd…)

Tags:

1. email revert showed high rate of conversion..
2. Closed by Horizon also showed a high rate of conversion.

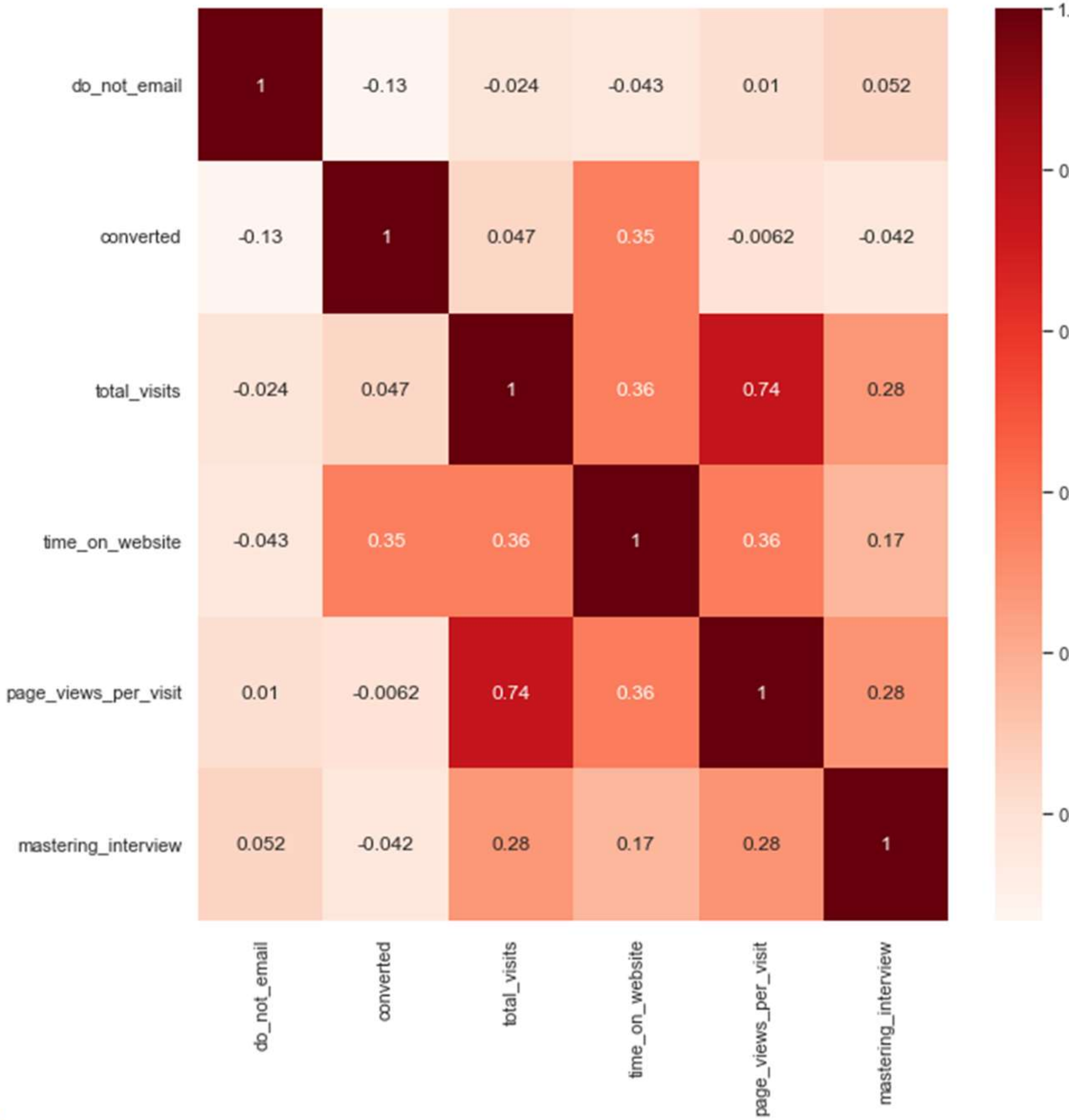


EDA(contd…)

Inferences from Heatmap

High Correlation is observed between:

- Page views per visit and total visits has high correlation of 0.74
- Time on website and converted has high correlation of 0.35



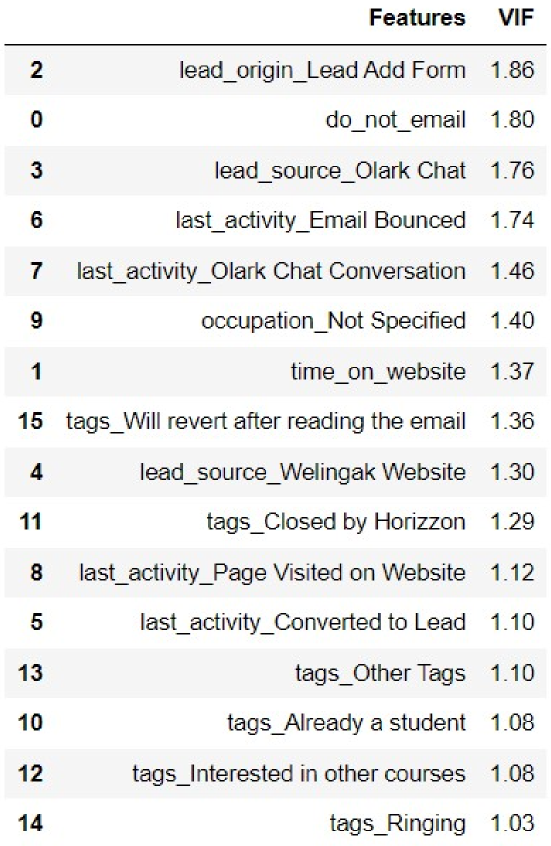
RFE Analysis and Future Selection

Following are the features identified in RFE.

- lead_origin_Lead Add Form
- do_not_email
- lead_source_Olark Chat
- last_activity_Email Bounced
- last_activity_Olark Chat Conversation
- occupation_Not Specified
- time_on_website
- 5tags_Will revert after reading the email
- lead_source_Welingak Website
- 1tags_Closed by Horizzon
- last_activity_Page Visited on Website
- last_activity_Converted to Lead
- tags_Other Tags
- tags_Already a student
- tags_Interested in other courses
- tags_Ringing

Final Model

- Logical Regression model built on the Initial 45 features received via RFE method. Analyzed p-value and RFE values and eliminated.
- on every step we have perform model evaluation by verifying factors like Accuracy, Precision, Sensitivity and precision.
- Logical Regression Final model has 16 features on which we did testing as well as prediction.



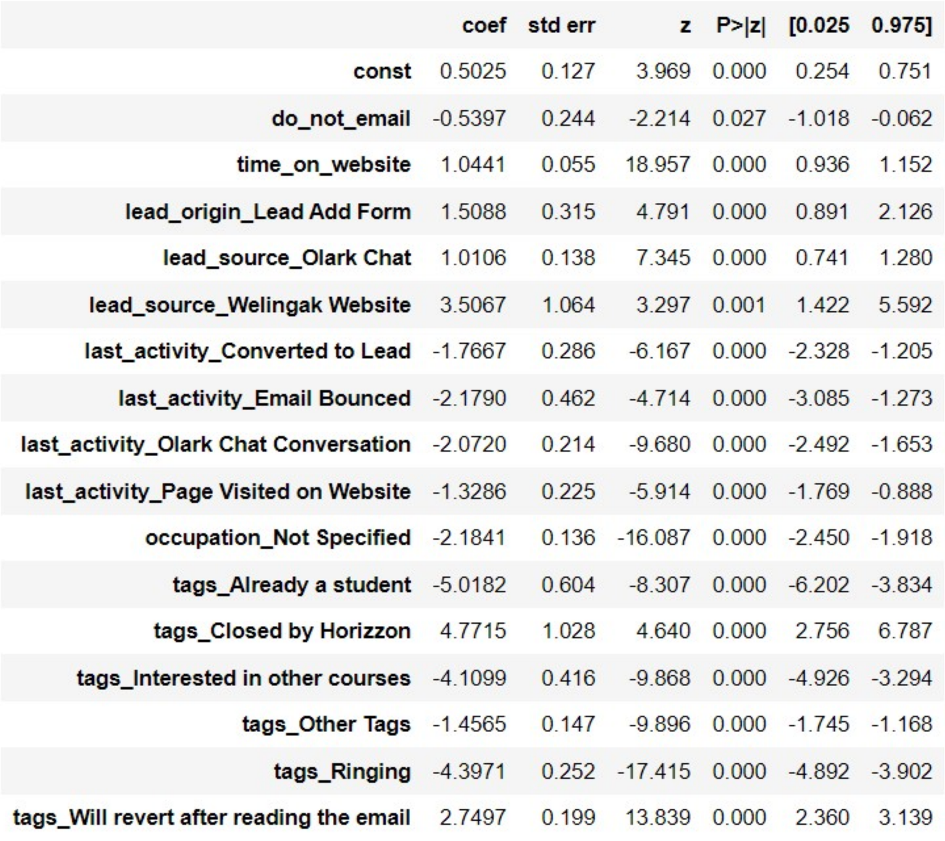
	Features	VIF
2	lead_origin_Lead Add Form	1.86
0	do_not_email	1.80
3	lead_source_Olark Chat	1.76
6	last_activity_Email Bounced	1.74
7	last_activity_Olark Chat Conversation	1.46
9	occupation_Not Specified	1.40
1	time_on_website	1.37
15	tags_Will revert after reading the email	1.36
4	lead_source_Welingak Website	1.30
11	tags_Closed by Horizzon	1.29
8	last_activity_Page Visited on Website	1.12
5	last_activity_Converted to Lead	1.10
13	tags_Other Tags	1.10
10	tags_Already a student	1.08
12	tags_Interested in other courses	1.08
14	tags_Ringing	1.03

Final Model – Regression Summary

- All p-values are zero which indicates all these columns are statistically significant.
- If we see coefficients following attributes are positively
- Positively impacting features:
- Total Time Spent on Website
- Lead_Origin_Lead Add Form
- Lead_Source_Olark Chat

Following negatively impacted features:

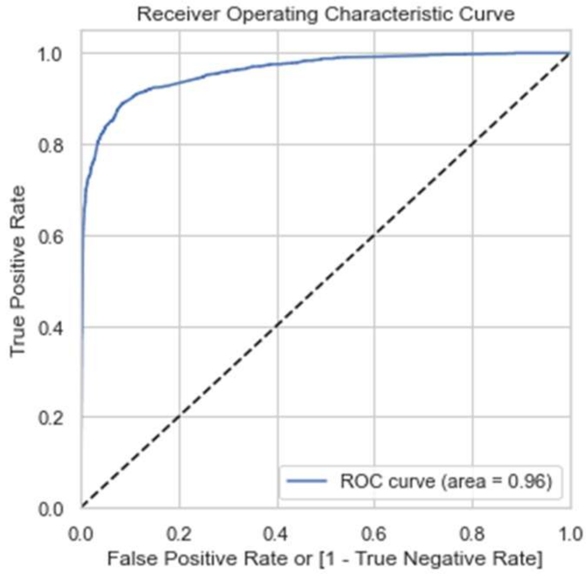
- occupation_Not Specified
- tags_Ringing
- tags_Already a student

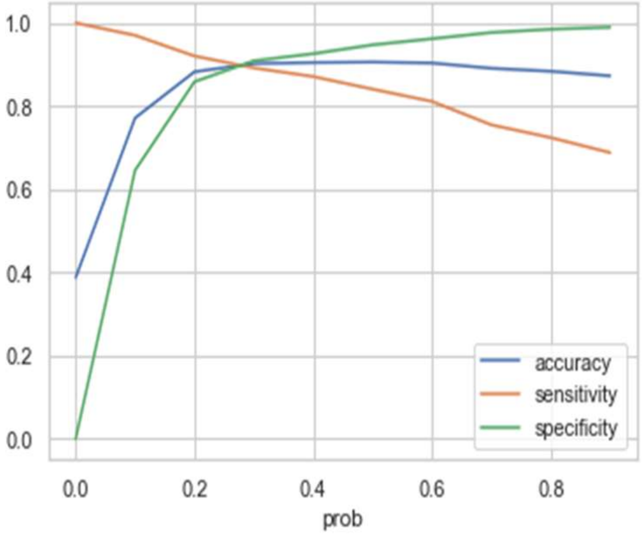


	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5025	0.127	3.969	0.000	0.254	0.751
do_not_email	-0.5397	0.244	-2.214	0.027	-1.018	-0.062
time_on_website	1.0441	0.055	18.957	0.000	0.936	1.152
lead_origin_Lead Add Form	1.5088	0.315	4.791	0.000	0.891	2.126
lead_source_Olark Chat	1.0106	0.138	7.345	0.000	0.741	1.280
lead_source_Welingak Website	3.5067	1.064	3.297	0.001	1.422	5.592
last_activity_Converted to Lead	-1.7667	0.286	-6.167	0.000	-2.328	-1.205
last_activity_Email Bounced	-2.1790	0.462	-4.714	0.000	-3.085	-1.273
last_activity_Olark Chat Conversation	-2.0720	0.214	-9.680	0.000	-2.492	-1.653
last_activity_Page Visited on Website	-1.3286	0.225	-5.914	0.000	-1.769	-0.888
occupation_Not Specified	-2.1841	0.136	-16.087	0.000	-2.450	-1.918
tags_Already a student	-5.0182	0.604	-8.307	0.000	-6.202	-3.834
tags_Closed by Horizzon	4.7715	1.028	4.640	0.000	2.756	6.787
tags_Interested in other courses	-4.1099	0.416	-9.868	0.000	-4.926	-3.294
tags_Other Tags	-1.4565	0.147	-9.896	0.000	-1.745	-1.168
tags_Ringing	-4.3971	0.252	-17.415	0.000	-4.892	-3.902
tags_Will revert after reading the email	2.7497	0.199	13.839	0.000	2.360	3.139

Final Model – Roc Curve

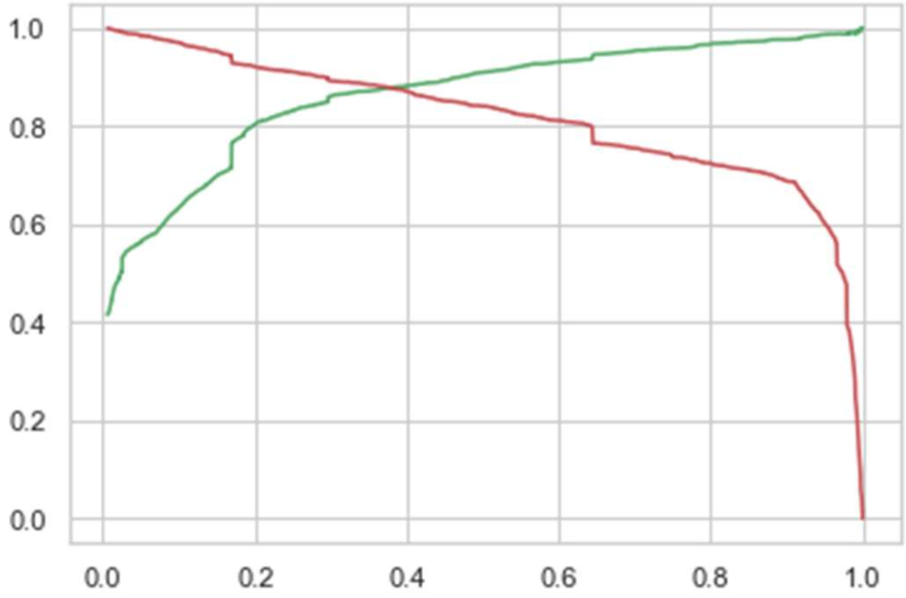
- Accuracy , sensitivity, specificity cut off is at 0.30
- Area under curve is 0.96, which seems to be indication of good model





Final Model – Precision & Recall Curve

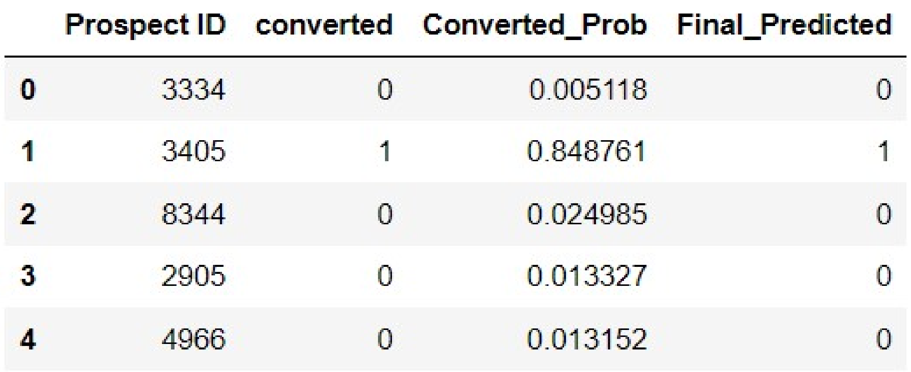
- Optimal cutoff for probability is near 0.38
- Precision ~ Recall Statistics remain close to 0.85



Final Model evaluation

Confusion metrics created at threshold of 0.35.

Overall Model Accuracy is 90%

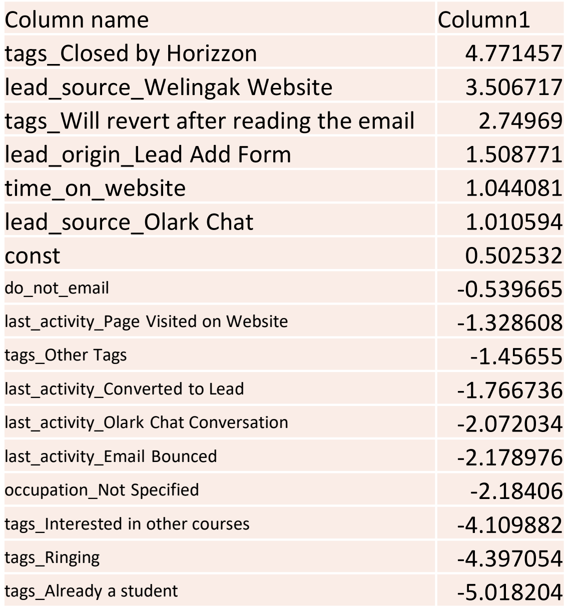


	Prospect ID	converted	Converted_Prob	Final_Predicted
0	3334	0	0.005118	0
1	3405	1	0.848761	1
2	8344	0	0.024985	0
3	2905	0	0.013327	0
4	4966	0	0.013152	0

Confusion metrics		
Actual/Predicted	Positive	Negative
Positive	1511	152
Negative	100	904

Evaluation metrics	Score
Accuracy	90
Precision	90
Recall	85
Sensitivity	90

Final model –coefficients.



Column name	Column1
tags_Closed by Horizzon	4.771457
lead_source_Welingak Website	3.506717
tags_Will revert after reading the email	2.74969
lead_origin_Lead Add Form	1.508771
time_on_website	1.044081
lead_source_Olark Chat	1.010594
const	0.502532
do_not_email	-0.539665
last_activity_Page Visited on Website	-1.328608
tags_Other Tags	-1.45655
last_activity_Converted to Lead	-1.766736
last_activity_Olark Chat Conversation	-2.072034
last_activity_Email Bounced	-2.178976
occupation_Not Specified	-2.18406
tags_Interested in other courses	-4.109882
tags_Ringing	-4.397054
tags_Already a student	-5.018204

Recommendations

- The company should make calls to the leads coming from "tags_Closed by Horizzon " as they are more likely to get converted.
- The company should make calls to the leads who spent "time_on_website" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose tags is "Already a student" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

Thank you