Lead Scoring Case Study

Jitendra Havaldar

Shubham Rustagi

Suraj Gupta

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses, although X Education gets a lot of leads, its lead conversion rate is very poor which is only 30 %.

To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Objective

There are quite a few goals for this case study:

1. 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.
2. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Problem Solving Approach

Data Understanding and Preprocessing -

- Learn more about data such as shape of the data and statistical information about columns in the dataset
- Data cleaning:
  - Handle missing or null values - e.g drop columns with missing values >= 45%
  - Data Imputation – e.g. imputing missing country values with mode of the Country column in the dataset
  - Outlier Analysis and treatment – e.g. removing top and bottom 1% outliers

Solve problem -

- Feature Engineering – e.g. encode categorical features or make dummy features for multiple levels
- Exploratory Data Analysis - Univariate and Bivariate Analysis
- Train Test Split data
- Feature Scaling/Normalization – To speed up the Gradient Descent convergence process using Standardization technique
- Handling class imbalance
- Logistic regression Model Building

Model Evaluation -

- Evaluate metrics such as confusion matrix, precision, recall, ROC-AUC score
- Draw Conclusion and recommendations for model.

Data Preprocessing

Data Understanding and Preprocessing -

- Replaced Select with NaN
- Checked the % of missing values for all columns -
    - Columns with more than 45 % missing values were dropped
- Dropping unnecessary columns with only null values, single unique feature, rating columns.
- Imputed Values with highest count in particular columns
- Segregated all NA values into others as separate entity.
- Highly skewed columns were dropped.

Exploratory Data Analysis – Numerical Features

Following are the observations –

Total Time Spent on Website –

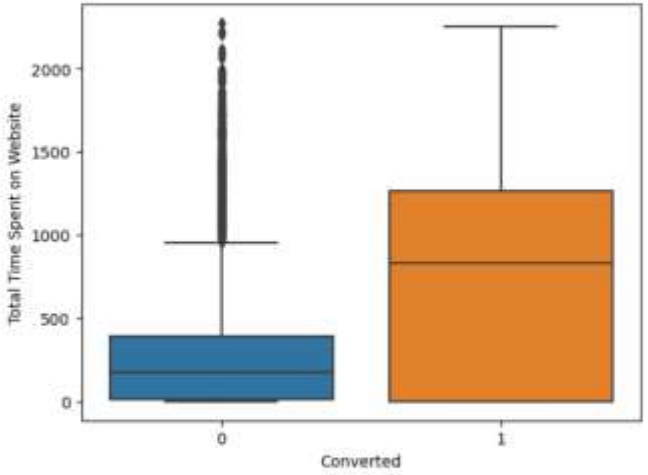
- Leads spending more time on the website are more likely to be converted
- Website should be made more enaging and immersive for the leads to spend more time

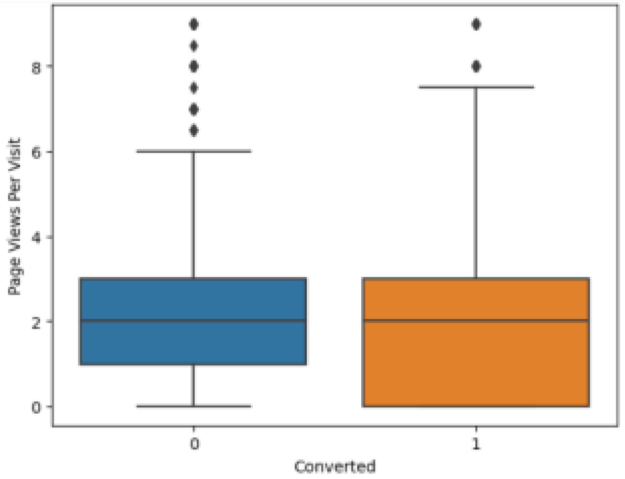
Page Views per visit –

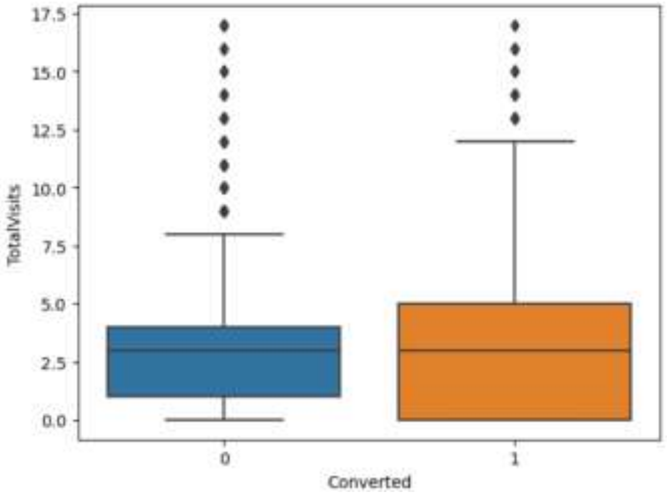
- No concrete finding from the boxplot for this feature

Total Visits –

- Median values for converted v/s non-converted are same
- Nothing conclusive based on the graph of this feature







EDA– Bi Variate Analysis

Following are the observations –

Tags –

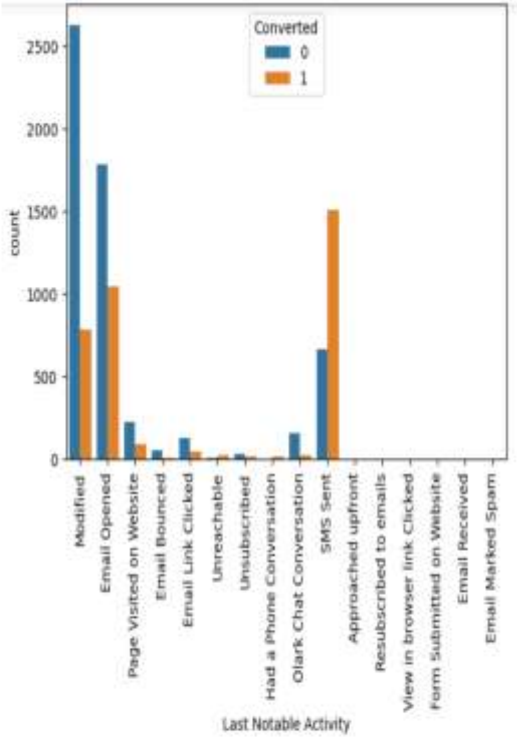
- Category “Will revert after reading the email” has a higher conversion percentage

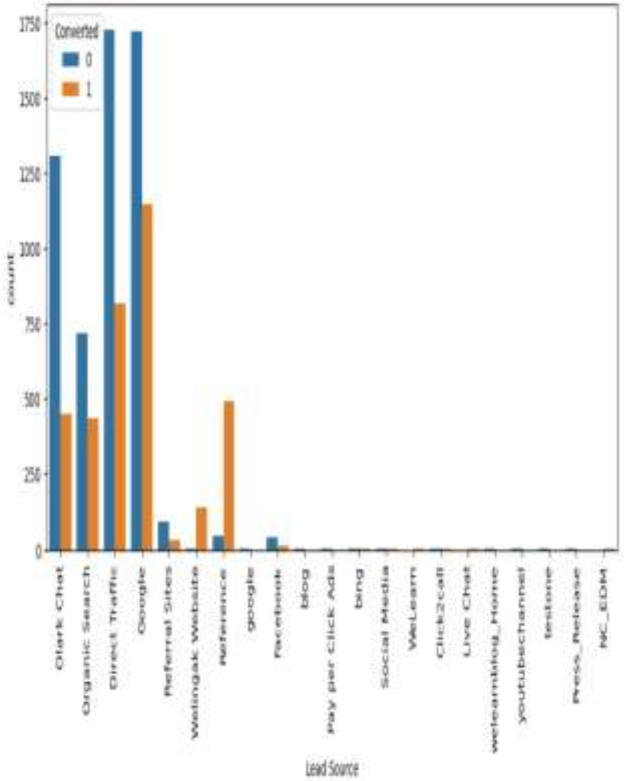
Lead Source –

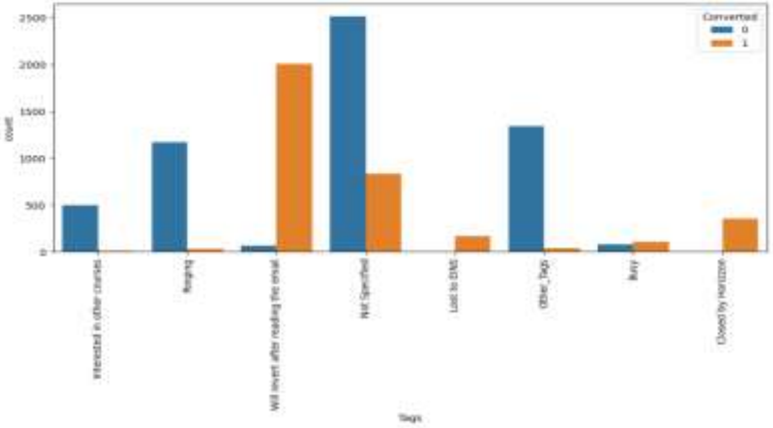
- Category “Reference” has a higher conversion percentage

Last Notable Activity –

- Category “Sms Sent” has a higher conversion percentage





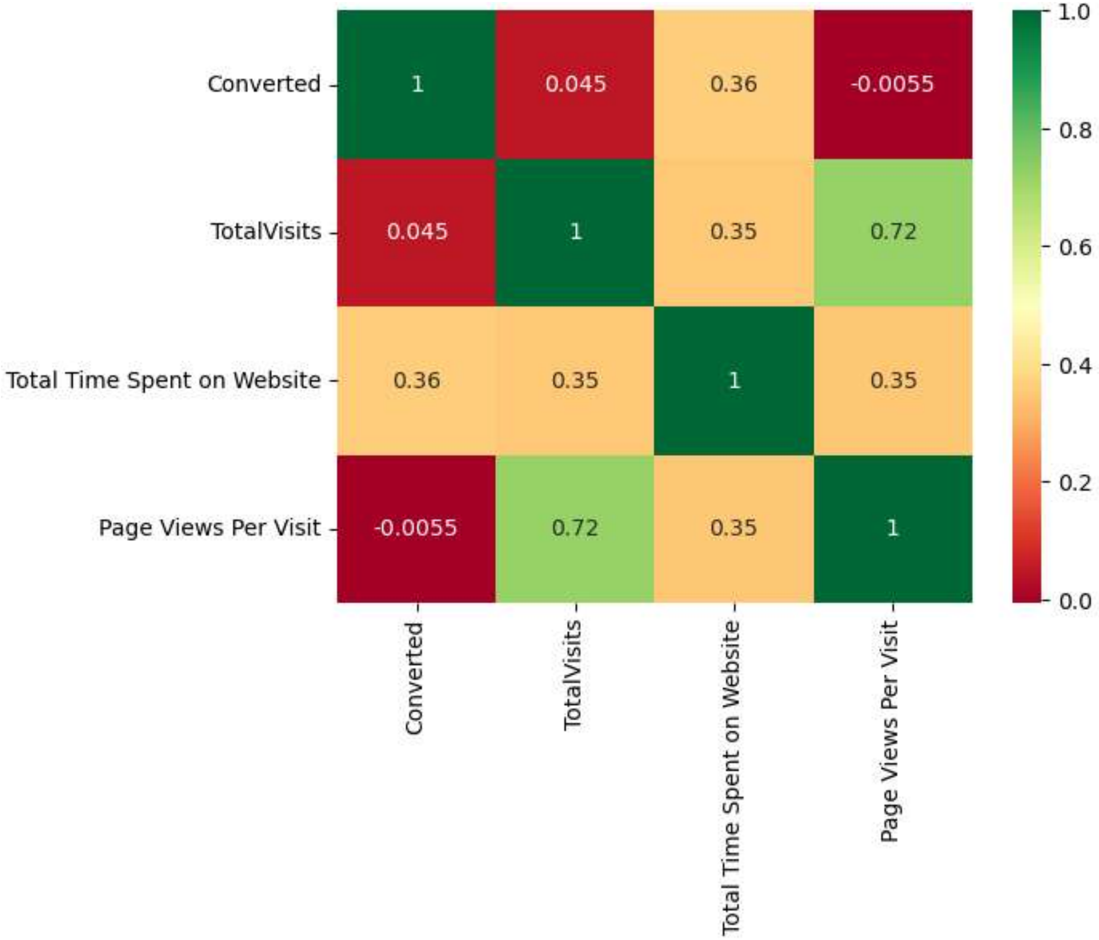


EDA– Multivariate Analysis

Following are the observations –

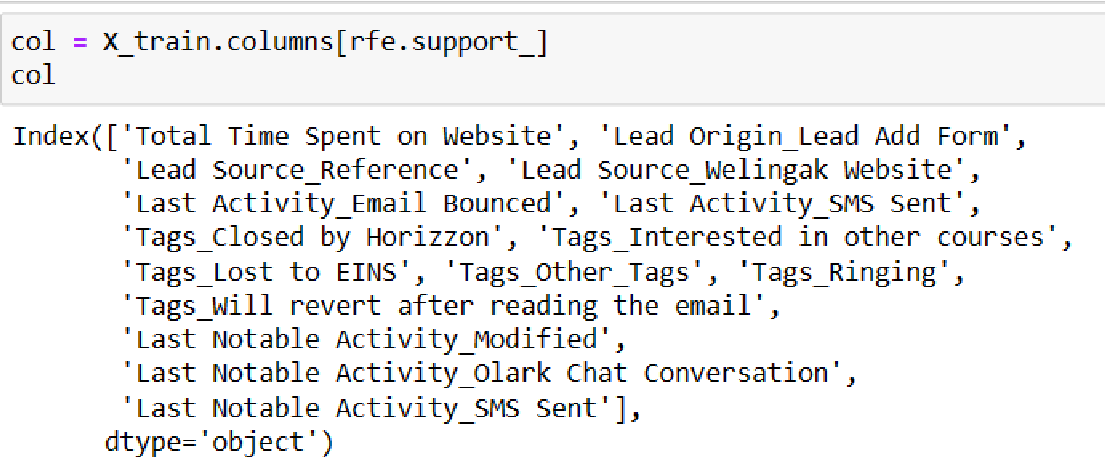
- Total Visits and Page Views per visit have a high correlation

- Target variable ‘Converted’ and ‘Total Time Spent on Website’ show some positive correlation



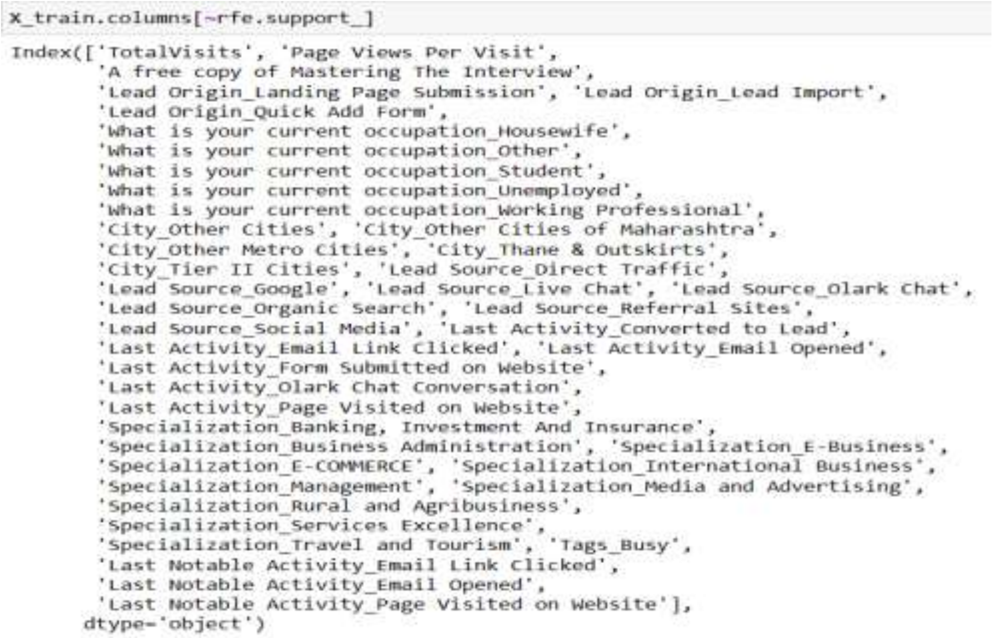
Model Building

- For Model building we need to scale and split data into train and test dataset.
- We will be using Logistic Regression for building the model.
- Initial variable selection will be done through RFE(recursive feature elimination) and further eliminating features with high p value and VIF value.
- Analyze various parameters for train dataset Specificity, Sensitivity, Accuracy, Precision and recall for train data.
- Plot the ROC Curve which shows trade off between sensitivity and specificity.



```
col = X_train.columns[rfe.support_]
col
```

```
Index(['Total Time Spent on Website', 'Lead Origin_Lead Add Form',
       'Lead Source_Reference', 'Lead Source_Welingak Website',
       'Last Activity_Email Bounced', 'Last Activity_SMS Sent',
       'Tags_Closed by Horizzon', 'Tags_Interested in other courses',
       'Tags_Lost to EINS', 'Tags_Other_Tags', 'Tags_Ringing',
       'Tags_Will revert after reading the email',
       'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation',
       'Last Notable Activity_SMS Sent'],
      dtype='object')
```



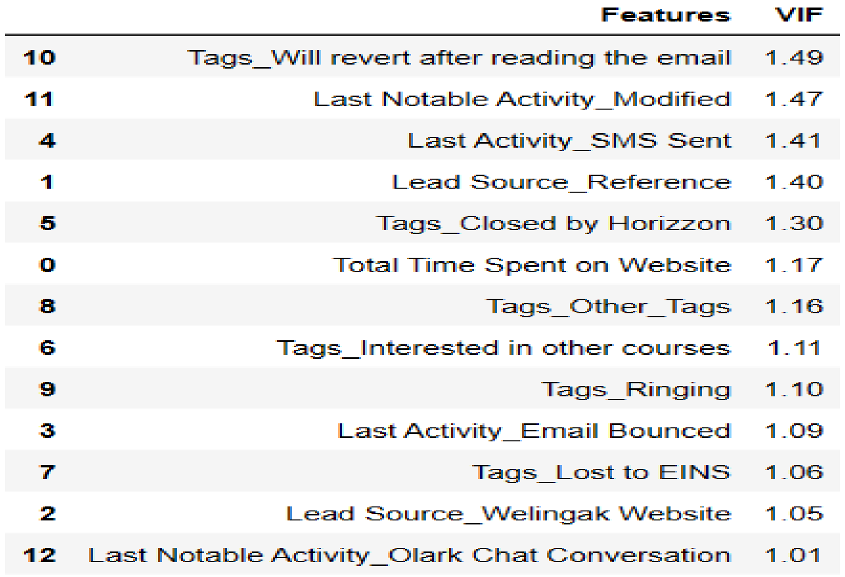
```
X_train.columns[~rfe.support_]
```

```
Index(['TotalVisits', 'Page Views Per Visit',
       'A free copy of Mastering The Interview',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Import',
       'Lead Origin_Quick Add Form',
       'What is your current occupation_Housewife',
       'What is your current occupation_Other',
       'What is your current occupation_Student',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'City_Other Cities', 'City_Other Cities of Maharashtra',
       'City_Other Metro Cities', 'City_Thane & Outskirts',
       'City_Tier II Cities', 'Lead Source_Direct Traffic',
       'Lead Source_Google', 'Lead Source_Live Chat', 'Lead Source_Olark Chat',
       'Lead Source_Organic Search', 'Lead Source_Referral Sites',
       'Lead Source_Social Media', 'Last Activity_Converted to Lead',
       'Last Activity_Email Link Clicked', 'Last Activity_Email Opened',
       'Last Activity_Form Submitted on Website',
       'Last Activity_Olark Chat Conversation',
       'Last Activity_Page Visited on Website',
       'Specialization_Banking, Investment And Insurance',
       'Specialization_Business Administration', 'Specialization_E-Business',
       'Specialization_E-COMMERCE', 'Specialization_International Business',
       'Specialization_Management', 'Specialization_Media and Advertising',
       'Specialization_Rural and Agribusiness',
       'Specialization_Services Excellence',
       'Specialization_Travel and Tourism', 'Tags_Busy',
       'Last Notable Activity_Email Link Clicked',
       'Last Notable Activity_Email Opened',
       'Last Notable Activity_Page Visited on Website'],
      dtype='object')
```

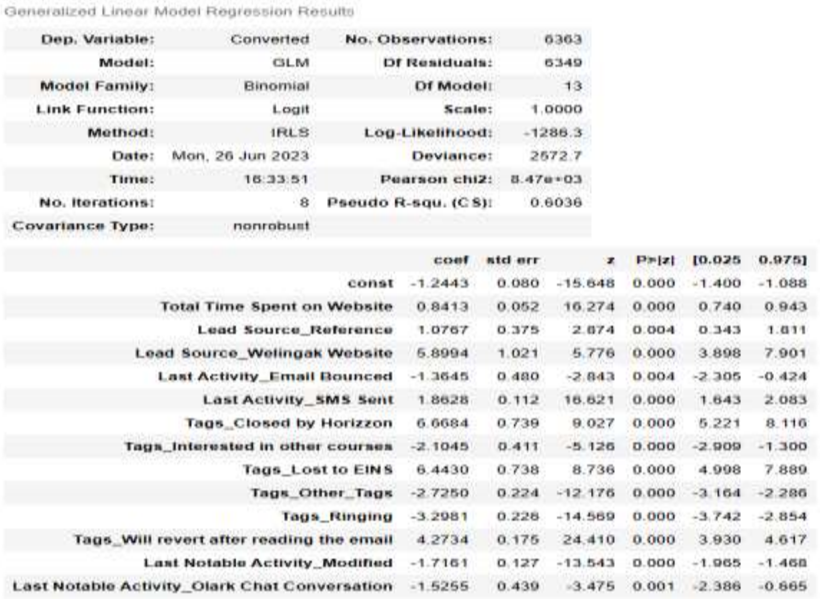
Logistic Regression Model

- Final Logistic Regression Model has been shown here after iteratively checking the p-value and VIF value

- The final model has p-values tending to zero and VIF values less than 2 indicating that this model can be used to make predictions on the test data



	Features	VIF
10	Tags_Will revert after reading the email	1.49
11	Last Notable Activity_Modified	1.47
4	Last Activity_SMS Sent	1.41
1	Lead Source_Reference	1.40
5	Tags_Closed by Horizzon	1.30
0	Total Time Spent on Website	1.17
8	Tags_Other_Tags	1.16
6	Tags_Interested in other courses	1.11
9	Tags_Ringing	1.10
3	Last Activity_Email Bounced	1.09
7	Tags_Lost to EINS	1.06
2	Lead Source_Welingak Website	1.05
12	Last Notable Activity_Olark Chat Conversation	1.01



Generalized Linear Model Regression Results

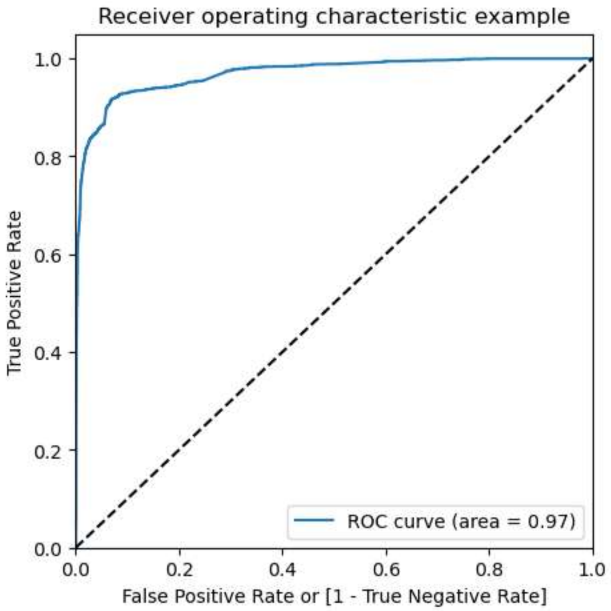
Dep. Variable:	Converted	No. Observations:	6363
Model:	GLM	Df Residuals:	6349
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1286.3
Date:	Mon, 26 Jun 2023	Deviance:	2572.7
Time:	16:33:51	Pearson chi2:	8.47e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.6036
Covariance Type:	nonrobust		

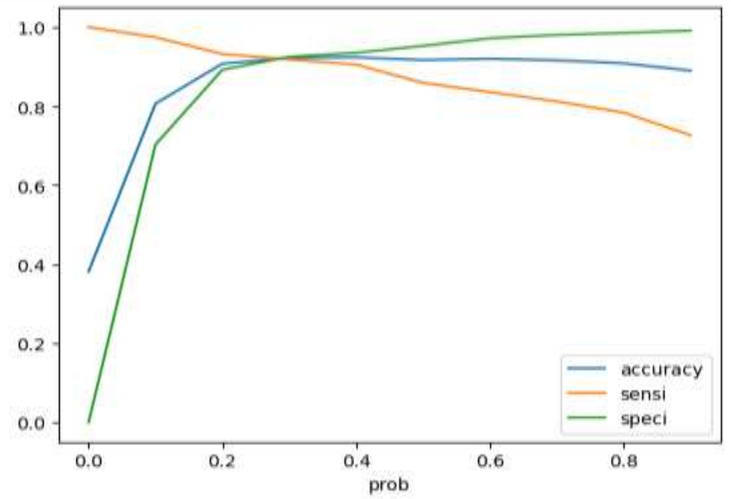
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2443	0.080	-15.648	0.000	-1.400	-1.088
Total Time Spent on Website	0.8413	0.052	16.274	0.000	0.740	0.943
Lead Source_Reference	1.0767	0.375	2.874	0.004	0.343	1.811
Lead Source_Welingak Website	5.8994	1.021	5.776	0.000	3.898	7.901
Last Activity_Email Bounced	-1.3645	0.480	-2.843	0.004	-2.305	-0.424
Last Activity_SMS Sent	1.8628	0.112	16.621	0.000	1.643	2.083
Tags_Closed by Horizzon	6.6684	0.739	9.027	0.000	5.221	8.116
Tags_Interested in other courses	-2.1045	0.411	-5.126	0.000	-2.909	-1.300
Tags_Lost to EINS	6.4430	0.738	8.736	0.000	4.998	7.889
Tags_Other_Tags	-2.7250	0.224	-12.176	0.000	-3.164	-2.286
Tags_Ringing	-3.2981	0.226	-14.569	0.000	-3.742	-2.854
Tags_Will revert after reading the email	4.2734	0.175	24.410	0.000	3.930	4.617
Last Notable Activity_Modified	-1.7161	0.127	-13.543	0.000	-1.965	-1.468
Last Notable Activity_Olark Chat Conversation	-1.5255	0.439	-3.475	0.001	-2.386	-0.665

Plotting ROC Curve

ROC curve -

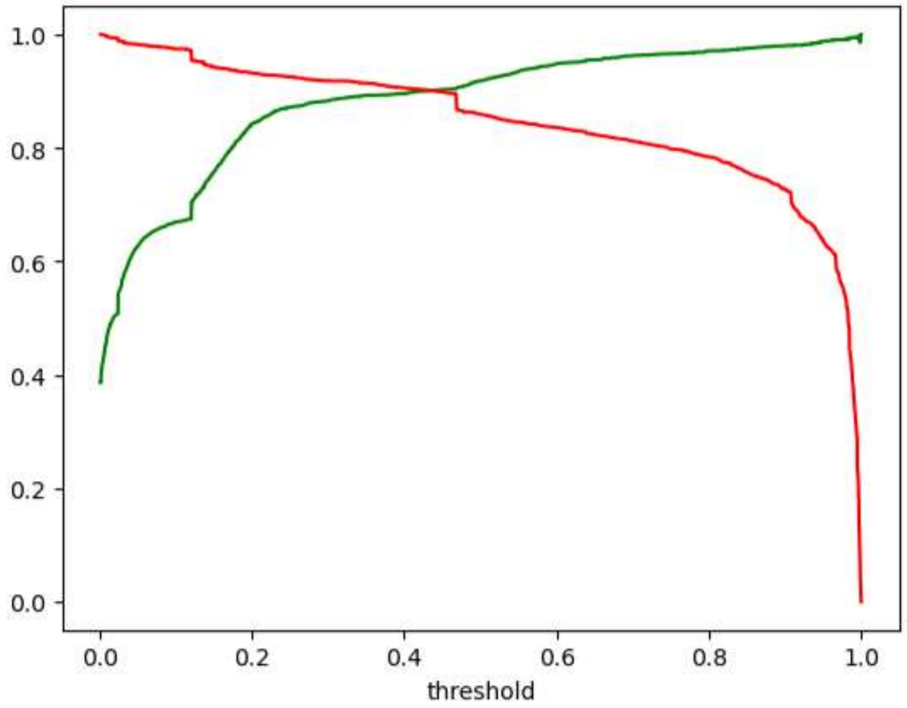
- It shows trade off between sensitivity and specificity.
- The closer the curve follows left hand border and then the top border of the ROC space, this proves better accuracy of the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.





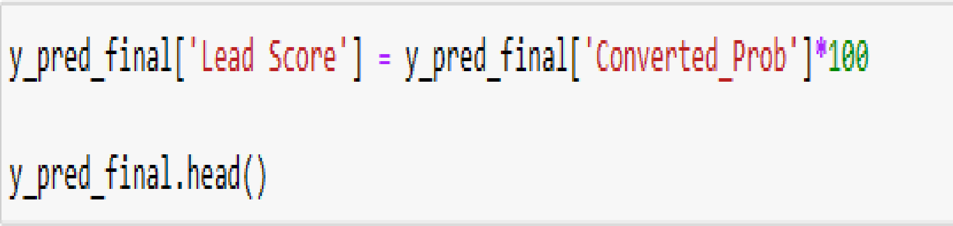
Precision-Recall Tradeoff

The threshold or cutoff is 0.4 as per the precision-recall tradeoff



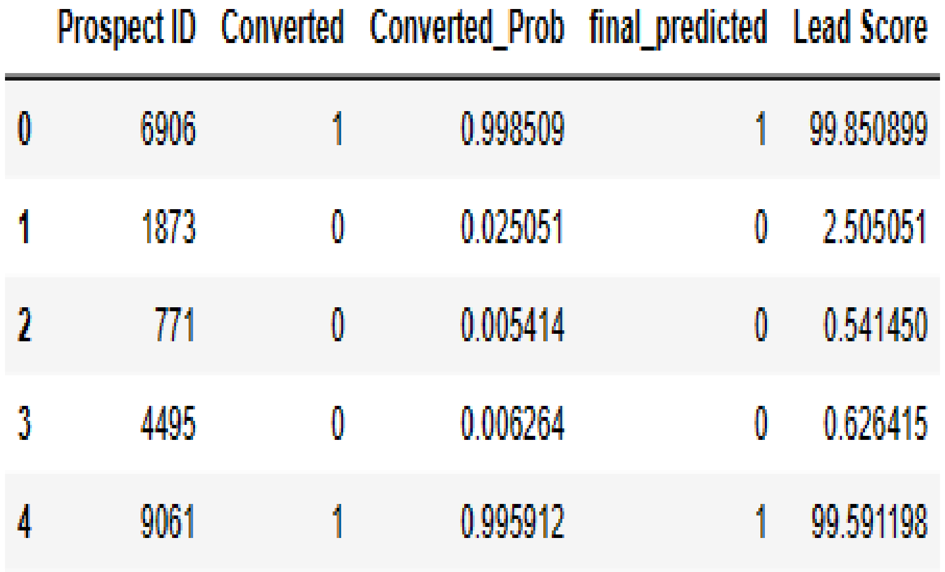
Lead Score calculation

The lead score is calculated as 100*Probability of lead conversion



```
y_pred_final['Lead Score'] = y_pred_final['Converted_Prob']*100

y_pred_final.head()
```



	Prospect ID	Converted	Converted_Prob	final_predicted	Lead Score
0	6906	1	0.998509	1	99.850899
1	1873	0	0.025051	0	2.505051
2	771	0	0.005414	0	0.541450
3	4495	0	0.006264	0	0.626415
4	9061	1	0.995912	1	99.591198

Recommendations

Top 3 features that contribute to lead conversion are –

- Total Time Spent on Website
- Leads that have come through source as reference
- Leads that have come through source as Welingak Website

Top 3 features that need to be improved are –

- Last Notable Activity_Olark Chat Conversation - Olark chat needs to be improved
- Last Notable Activity_Modified
- Tags_Will revert after reading the email - sales team need to follow up on the mails sent to leads