upGrad

Lead Scoring Model

Group Case Study

By Suchitra, Sunil and Sukhwinder (DS C62)

Problem Statement & Objective

Company: X Education, An Education Company sells Online courses to Industry Professionals. They Market their courses through various websites which sends them leads of potential customers.

Situation: The current conversion Rate of this leads is around 30%

Objective: The company wishes to identify the most potential leads, also known as Hot Leads for a more focused marketing and conversion strategy. Thus, aiming for a target lead conversion rate of 80% or higher

Data: 9,000 data points of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

Goal: Build a logistic regression model to assign a lead score to all leads. Customers with higher lead scores will have a higher chance of conversion.

Approach for Problem Solving

1. Import Data
   - *Read The Data file in Python*
   - *Check if the Data gets read completely and correctly*
2. Data Cleanup
   - *Understand the Data Types*
   - *Review Missing values*
   - *Check Outliers*
3. Exploratory Analysis
   - *Perform Univariate Analysis*
   - *Data Visualization*
   - *Identify Key Impact Variables*
4. Build and Test Model
   - *Select Key variables*
   - *Create Dummy variables for Categorical Data*
   - *Create Train Test Data*
   - *Build Model*
   - *Test Model*
5. Conclusion
   - *Identify Top Impact variable*
   - *Summary Report*

Import Data

- **Reading Data: 9240 Rows and 37 Columns**

```
In [2]: # Import file
        ls= pd.read_csv('Leads.csv')
```

- **Review Data Heads to know if the columns were read correctly. Also review Data Tail to know if the Data was read till the last row**

```
In [5]: # Glance on the data to know if the data and data headers were read corrected (A high level view)
        # Review Head

        ls.head(10)
```

Out[5]:

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How did you hear about X Education	W c occu
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select	Unem

- **Check the Datatypes**

```
In [7]: # check for null and datatype
        ls.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
```

Data Clean up

Review Columns with Missing Data

- *17 Columns have missing value*
- *Drop Columns with more than 25% Data missing*

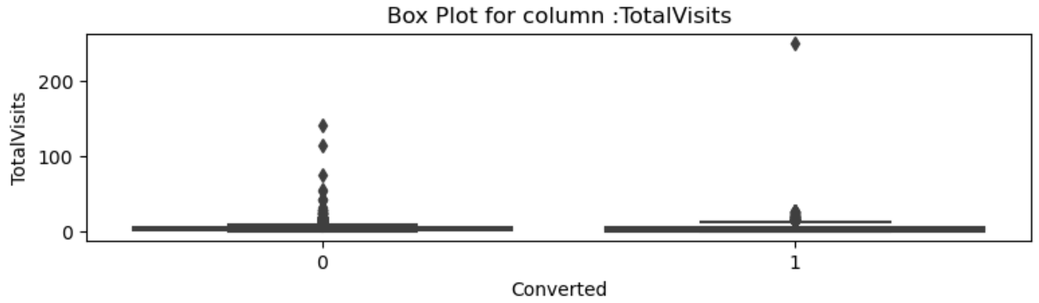
```
The Lead Scoring Dataframe has 37 columns.
There are 17 columns that have missing values.

Lead Quality                                        51.590909
Asymmetrique Profile Score                          45.649351
Asymmetrique Activity Score                         45.649351
Asymmetrique Profile Index                          45.649351
Asymmetrique Activity Index                         45.649351
Tags                                                36.287879
Lead Profile                                        29.318182
What matters most to you in choosing a course       29.318182
What is your current occupation                     29.112554
Country                                             26.634199
How did you hear about X Education                  23.885281
Specialization                                      15.562771
City                                                15.367965
TotalVisits                                          1.482684
Page Views Per Visit                                 1.482684
Last Activity                                        1.114719
```

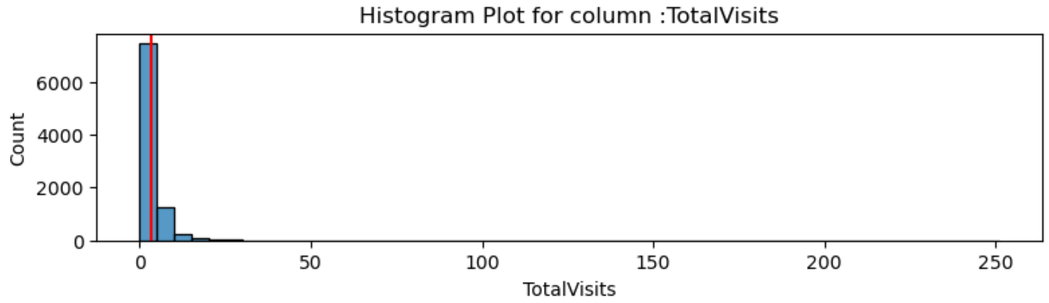
- *Discovered Data field which should also be treated as missing data*

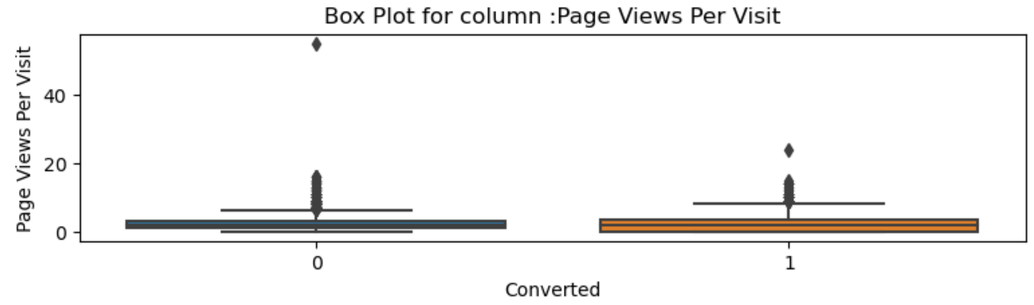
```
 Summary view of column: How did you hear about X Education

 How did you hear about X Education
 Select                   53.846154
 Online Search             8.904562
 Word Of Mouth             3.824113
 Student of SomeSchool     3.416354
 Other                     2.049813
 Multiple Sources          1.675116
 Advertisements            0.771435
 Social Media              0.727353
 Email                     0.286533
 SMS                       0.253471
 Name: count, dtype: float64

  Summary view of column: Specialization

 Specialization
 Select                               20.398942
 Finance Management                   10.568658
 Human Resource Management             9.224157
 Marketing Management                  9.069870
 Operations Management                 5.499229
```

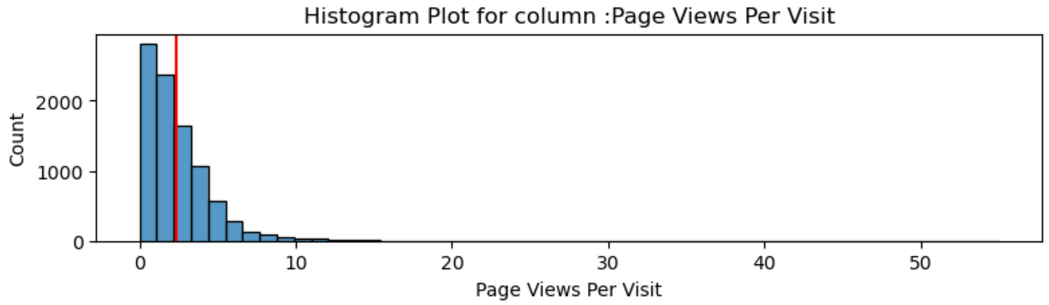
Data Clean up

Review Columns for Outliers







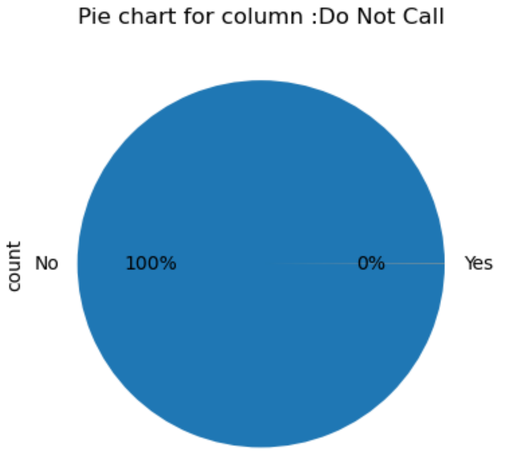


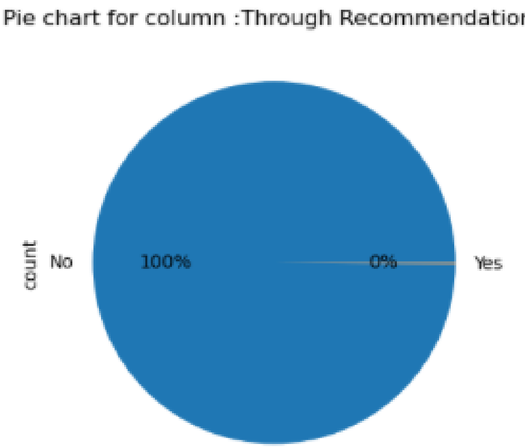
Data Clean up

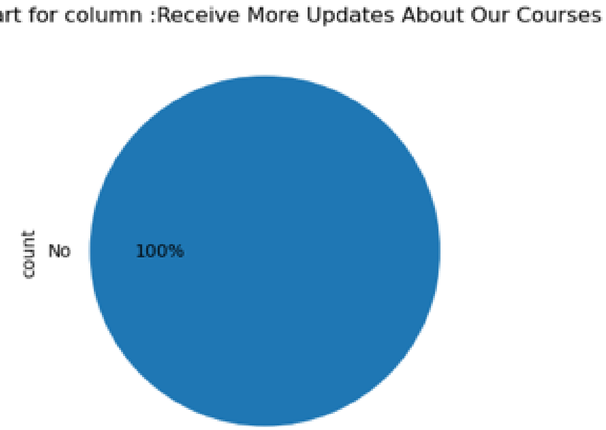
Short-listing of Columns

Below Columns have only one type of data value and can dropped

- Do Not Call,
- Search,
- Magazine,
- Newspaper Article
- X Education Forums,
- Newspaper,
- Digital Advertisement,
- Through Recommendations,
- Receive More Updates About Our Courses,
- Update me on Supply Chain Content,
- Get updates on DM Content,
- I agree to pay the amount through cheque,





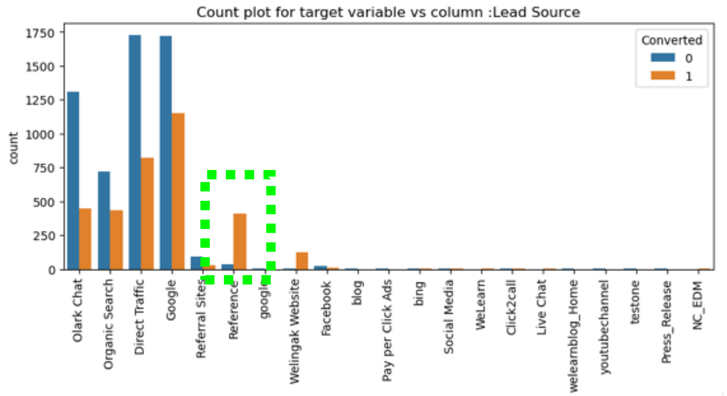


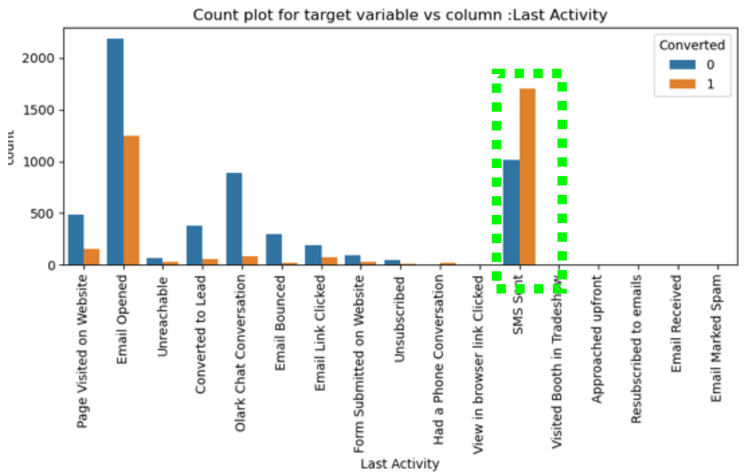
Exploratory Data Analysis

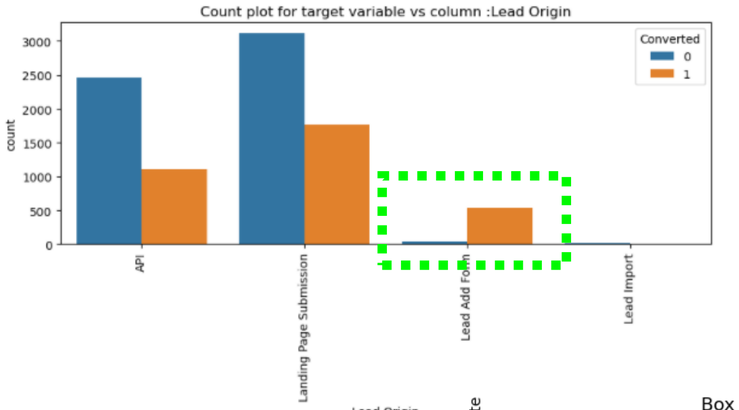
Impactful variables on Conversion

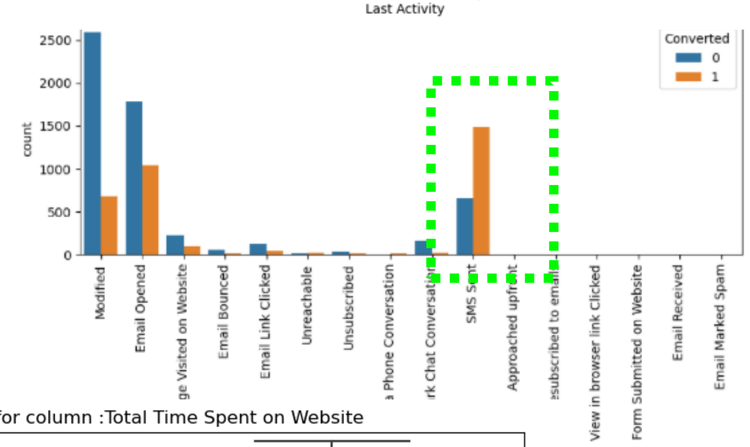
Key Impactful Variables:

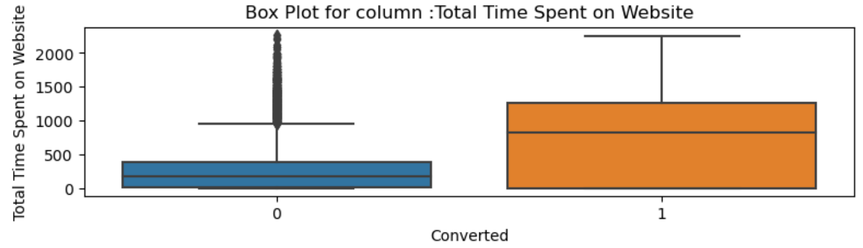
1. Lead Origin: Lead Add Form has a significant conversion rate.
2. Lead Source: Conversion Rate is highest for Lead Welingak Website and Reference.
3. Last Activity: SMS sent has highest conversion rate followed by Email Opened
4. Last Notable Activity: Top 3 Last Notable Activity are - Modified, Email Opened and SMS Sent with SMS Sent having the highest Conversion Rate
5. Total time spent on the website







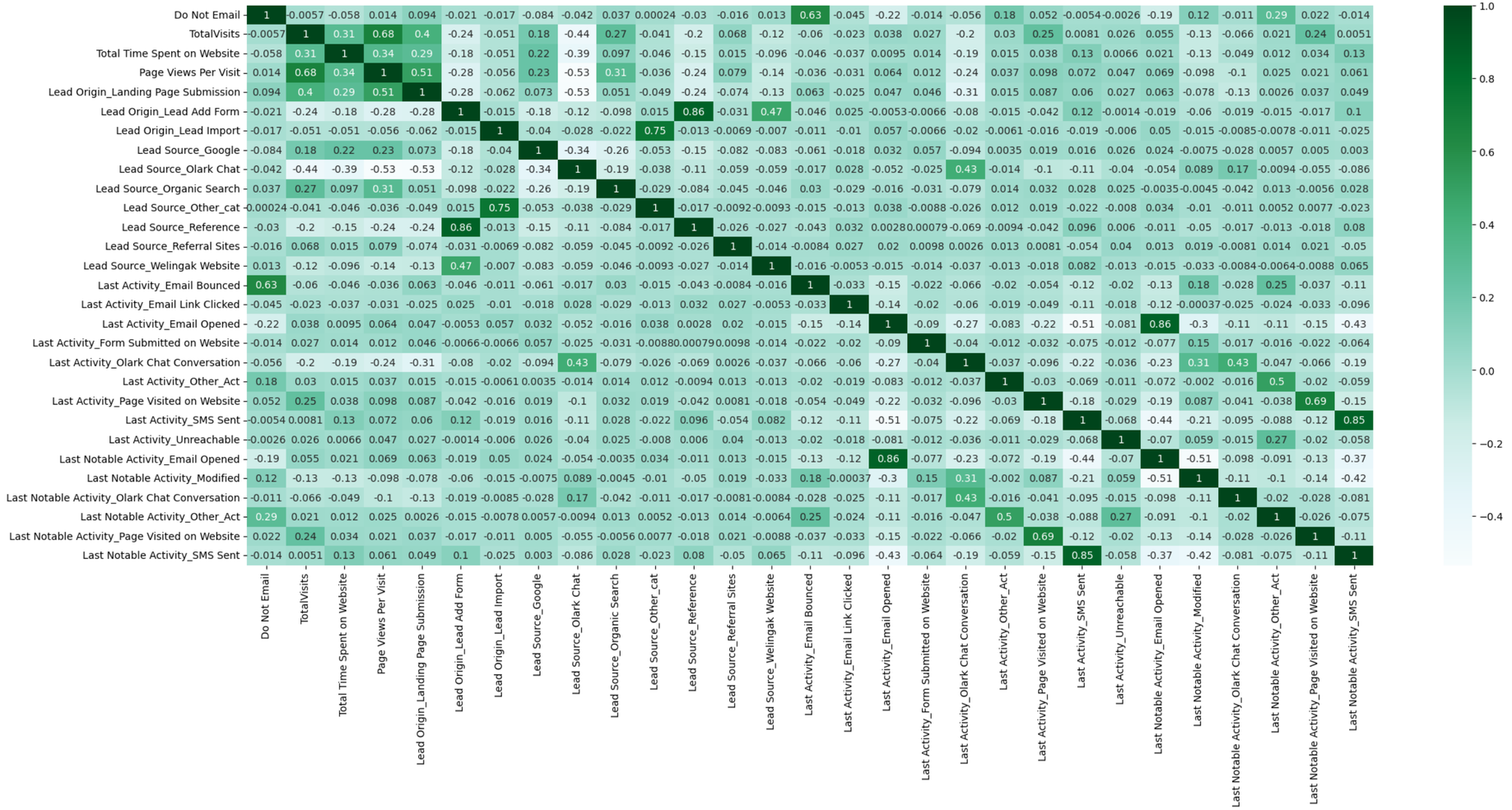




Exploratory Data Analysis

Scaling and Correlation of Variables

- Standardisation of variables using StandardScaler()
- Created Dummy Variables using pd.get_dummies()
- Checked Correlation and found a few darker shades highlighting strong relationship



Build the Model

Approach for building the Model

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif
- value is greater than 5
- Predictions on test data set
- Overall accuracy 80%

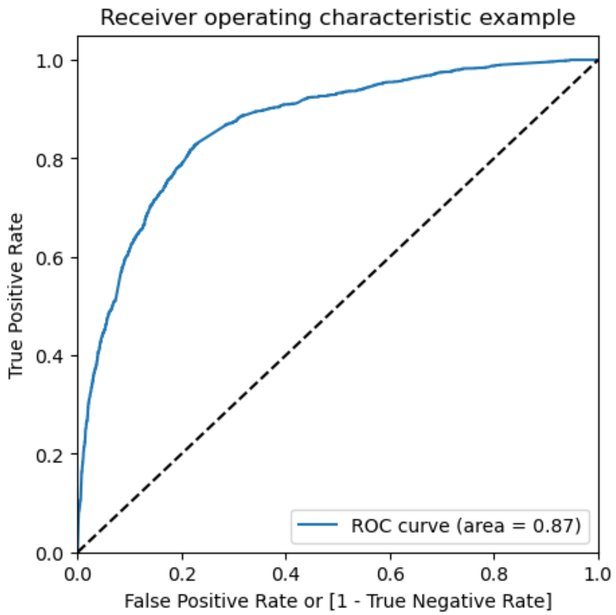
P-Value is less than 5%

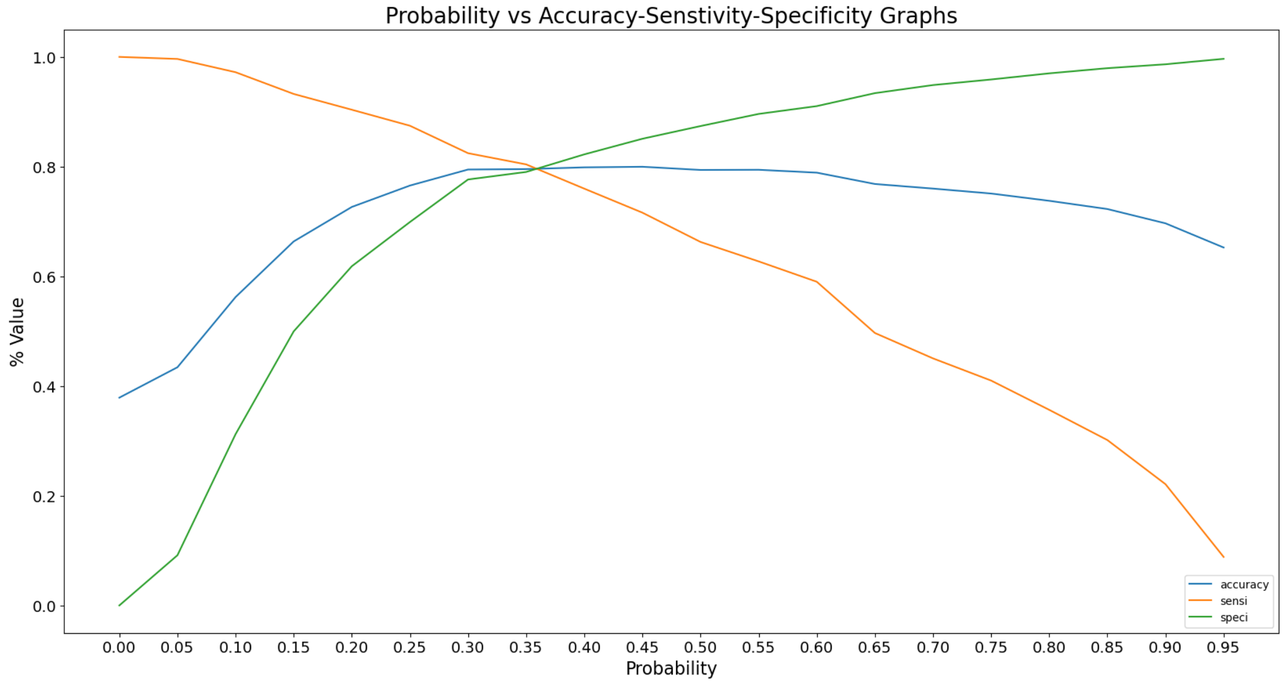
	coef	std err	z	P>|z|
const	-1.3742	0.106	-12.985	0.000
Do Not Email	-1.6519	0.176	-9.362	0.000
Total Time Spent on Website	1.1376	0.039	29.241	0.000
Lead Origin_Lead Add Form	4.7573	0.235	20.213	0.000
Lead Source_Olark Chat	1.1293	0.100	11.316	0.000
Last Activity_Email Opened	0.3005	0.105	2.867	0.004
Last Activity_Olark Chat Conversation	-1.1752	0.174	-6.759	0.000
Last Activity_Other_Act	1.2745	0.350	3.641	0.000
Last Notable Activity_Modified	-0.2844	0.099	-2.864	0.004
Last Notable Activity_Other_Act	1.0615	0.321	3.311	0.001
Last Notable Activity_SMS Sent	1.7732	0.122	14.486	0.000

VIF is less than 5

	Features	VIF
3	Lead Source_Olark Chat	1.78
5	Last Activity_Olark Chat Conversation	1.59
8	Last Notable Activity_Other_Act	1.48
7	Last Notable Activity_Modified	1.44
6	Last Activity_Other_Act	1.36
1	Total Time Spent on Website	1.28
0	Do Not Email	1.24
9	Last Notable Activity_SMS Sent	1.18
2	Lead Origin_Lead Add Form	1.17
4	Last Activity_Email Opened	1.17

ROC Curve and the Metrics





Confusion Matrix

Accuracy vs Sensitivity vs Others

```
In [90]: # Creating final confusion matrix on training predictions with Probability Cutoff 0.37
         confusion_train = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.final_predicted )
         confusion_train

Out[90]: array([[3157,  776],
                [ 514, 1884]])
```

```
In [116]: confusion_test = metrics.confusion_matrix(y_test_pred_final.Converted, y_test_pred_final.final_predicted )
          confusion_test

Out[116]: array([[1336,  350],
                 [ 220,  808]])
```

Training Data Scores calculated at probability cutoff threshold = 0.37

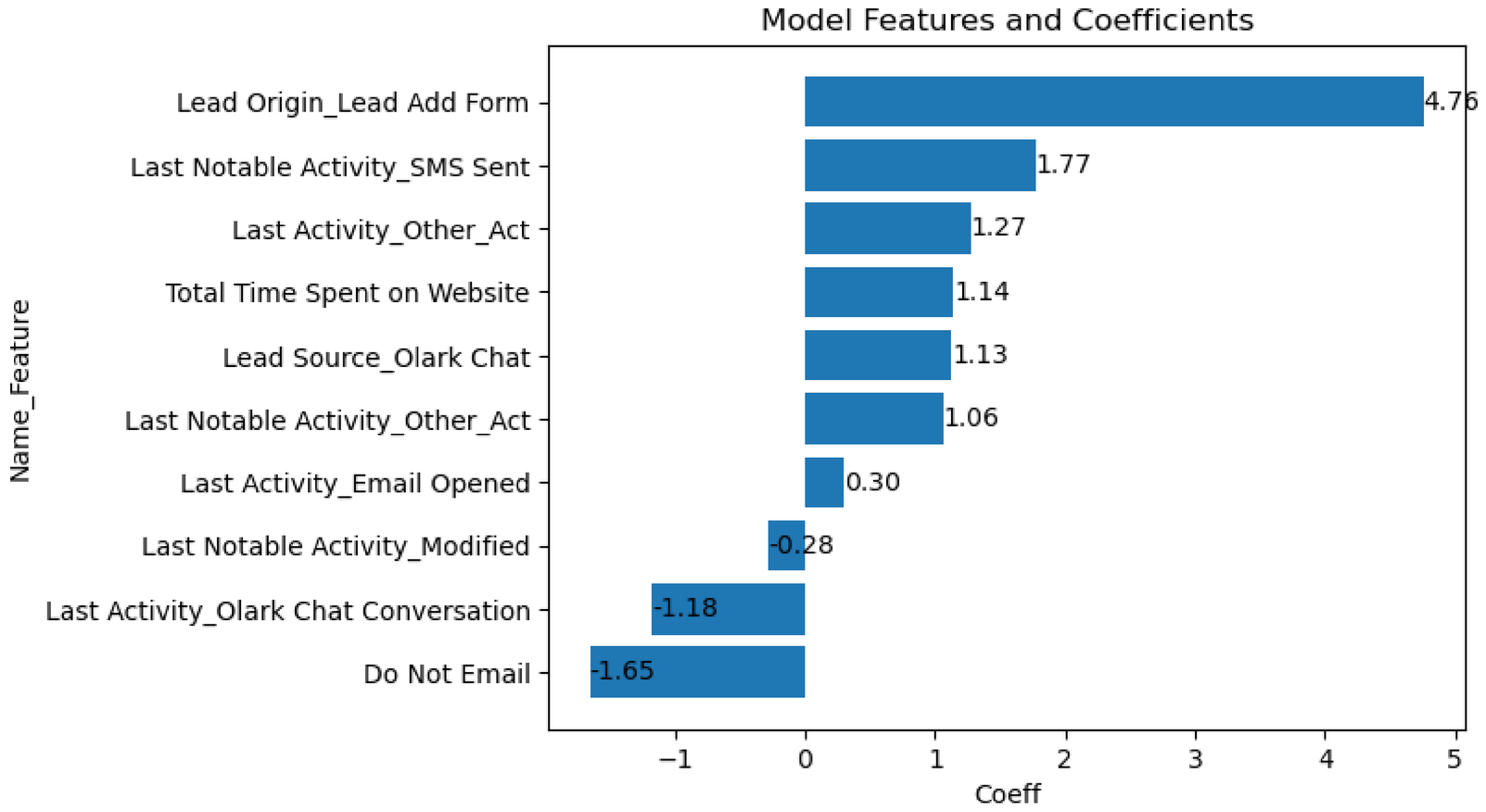
- Accuracy = 79.6%
- Sensitivity = 78.6%
- Specificity = 80.3%
- Precision = 70.8%
- Recall = 76.6%

Test Data Scores calculated at probability cutoff threshold = 0.37

- Accuracy = 79%%
- Sensitivity = 78.6%
- Specificity = 79.2%
- Precision = 69.7%
- Recall = 78.6%

The Final Features

Conclusion



Top 5 features responsible for good conversion rate are:

- Lead Origin_Lead Add Form
- Last Notable Activity_SMS Sent
- Last Activity Other Act
- Total Time Spent on Website
- Lead Source_Olark Chat