Lead Scoring Case Study

Shilpa Kulkarni
Shashwat Shirvastava
Shivam Agrawal

Problem Statement

- X Education wanted to improve its lead conversion rate which is currently at 30%
- Company would like to identify Hot Leads to make the process more efficient
- Company would like to build a logistic regression model which would assign a lead score to each of the leads
- Leads having higher conversion chance should be assigned higher lead score
- Lead conversion target post this exercise is 80%

Analysis Approach

- Import the provided leads dataset from the past
- Inspect the dataframe and the data dictionary and perform EDA
  - Check and correct the attribute data types
  - Impute the null values. Use visualization as required
  - Handle the 'Select' level in the categorical variables
  - Handle the outliers. Use visualization as required
  - Convert the discreet categorical variables into dummy variables
- Proceed to model building
  - Create train-test split and apply feature scaling
  - Check for correlations and drop highly correlated variables
  - Run the logistics regression model on the train set and check for p-values
  - Use RFE to select top 15 variables and perform the required model building iterations till the p-value and VIF is within acceptable limits

Analysis Approach

- Making predictions
  - Get the predicted probability values of the target variable on the train set
  - Use a threshold of 0.5 and accordingly compute the predicted values of the target variable('Converted' in this case)
  - Derive the Confusion matrix and calculate the different metrics
  - Find the optimal cutoff probability for balanced sensitivity and specificity and derive different metrics
  - Make the predictions on the test set and derive the different metrics
- Calculate the lead score value in the range of 0 to 100
- Calculate the Top 3 features which contribute most towards the probability of a lead getting converted
- Calculate the Bottom 3 features that need improvement to convert a lead

Visualization And Results

Null Values

- As a first step replaced 'select' values with 'nan'
- Deleted the columns having null % greater than 70%
- For the remaining, imputed nan with either mode values or created a new category called Others

Before

How did you hear about X Education	78.46
Lead Profile	74.19
Lead Quality	51.59
Asymmetrique Activity Index	45.65
Asymmetrique Profile Score	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
City	39.71
Specialization	36.58
Tags	36.29
What matters most to you in choosing a course	29.32
What is your current occupation	29.11
Country	26.63
Page Views Per Visit	1.48
TotalVisits	1.48
Last Activity	1.11
Lead Source	0.39

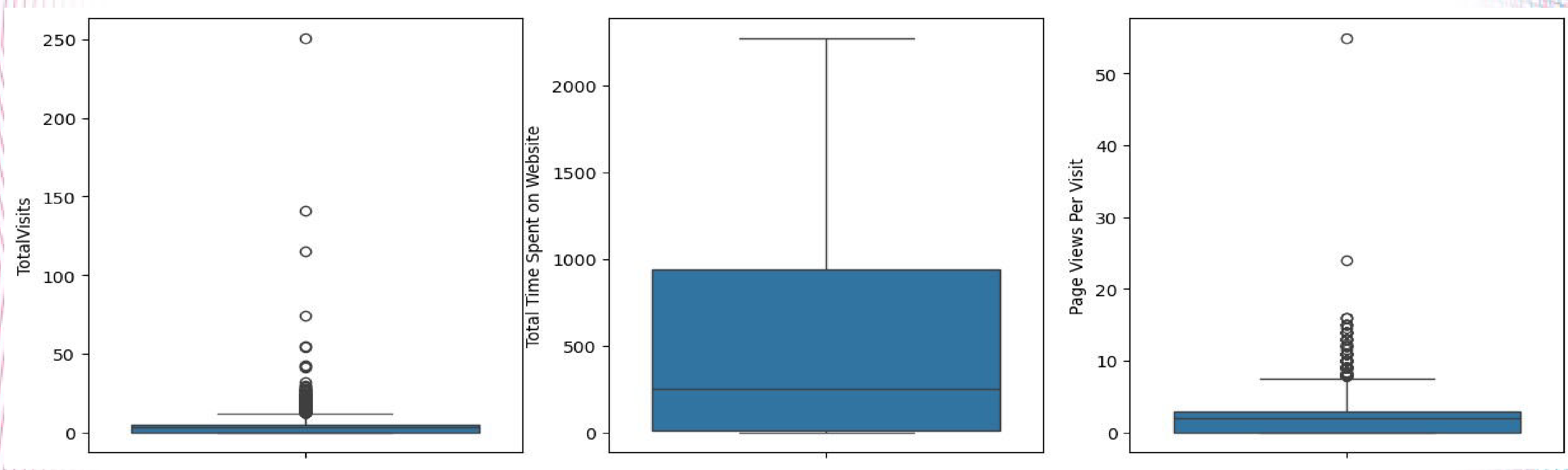
After

```
<class 'pandas.core.frame.DataFrame'>
Index: 8528 entries, 0 to 9239
Data columns (total 28 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Lead Number                                 8528 non-null   int64
 1   Lead Origin                                 8528 non-null   object
 2   Lead Source                                 8528 non-null   object
 3   Do Not Email                                8528 non-null   int64
 4   Do Not Call                                 8528 non-null   int64
 5   Converted                                   8528 non-null   int64
 6   TotalVisits                                 8528 non-null   float64
 7   Total Time Spent on Website                 8528 non-null   int64
 8   Page Views Per Visit                        8528 non-null   float64
 9   Last Activity                               8528 non-null   object
 10  Specialization                              8528 non-null   object
 11  What is your current occupation             8528 non-null   object
 12  Search                                      8528 non-null   int64
 13  Magazine                                    8528 non-null   int64
 14  Newspaper Article                           8528 non-null   int64
 15  X Education Forums                          8528 non-null   int64
 16  Newspaper                                   8528 non-null   int64
 17  Digital Advertisement                       8528 non-null   int64
 18  Through Recommendations                     8528 non-null   int64
 19  Receive More Updates About Our Courses      8528 non-null   int64
 20  Tags                                        8528 non-null   object
 21  Lead Quality                                8528 non-null   object
 22  Update me on Supply Chain Content           8528 non-null   int64
 23  Get updates on DM Content                   8528 non-null   int64
 24  City                                        8528 non-null   object
 25  I agree to pay the amount through cheque    8528 non-null   int64
 26  A free copy of Mastering The Interview      8528 non-null   int64
 27  Last Notable Activity                       8528 non-null   object
```

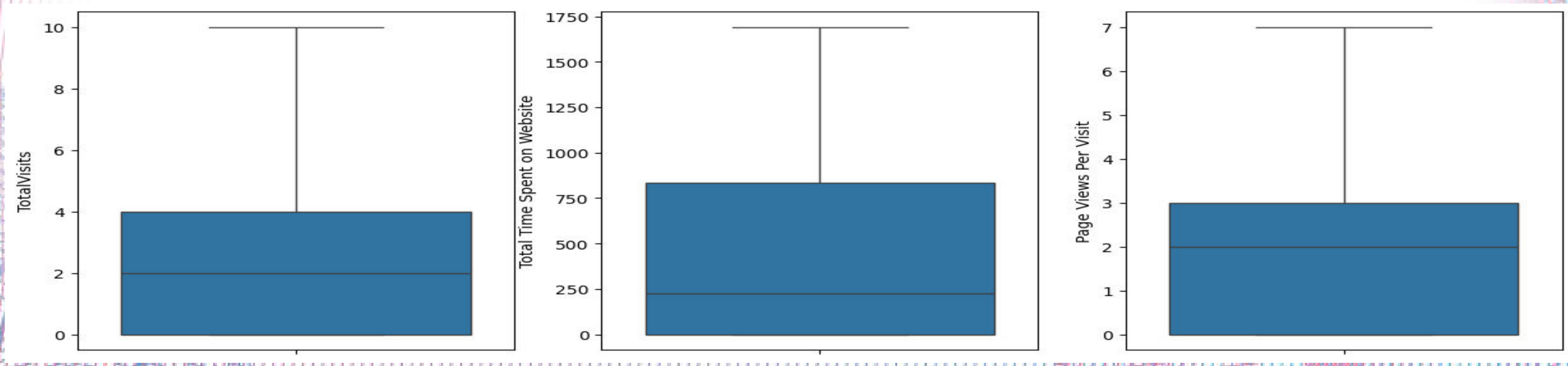
Outliers

Handled the outliers for the continuous numerical variables by using 1.5 IQR method - ['TotalVisits','Total Time Spent on Website','Page Views Per Visit']

Before



After



Dummy Variables

Create dummy variables for columns with dtype as object -

Index(['Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags', 'Lead Quality', 'City', 'Last Notable Activity'],dtype='object')

```
<class 'pandas.core.frame.DataFrame'>
Index: 8528 entries, 0 to 9239
Data columns (total 81 columns):
 #   Column                                   Non-Null Count  Dtype
---  ------                                   --------------  -----
 0   Lead Number                              8528 non-null   int64
 1   Do Not Email                             8528 non-null   int64
 2   Do Not Call                              8528 non-null   int64
 3   Converted                                8528 non-null   int64
 4   TotalVisits                              8528 non-null   float64
 5   Total Time Spent on Website              8528 non-null   int64
 6   Page Views Per Visit                     8528 non-null   float64
 7   Lead Origin_Landing Page Submission      8528 non-null   int32
 8   Lead Origin_Lead Add Form                8528 non-null   int32
 9   Lead Origin_Lead Import                  8528 non-null   int32
 10  Lead Source_Facebook                     8528 non-null   int32
 11  Lead Source_Google                       8528 non-null   int32
 12  Lead Source_Olark Chat                   8528 non-null   int32
 13  Lead Source_Organic Search               8528 non-null   int32
 14  Lead Source_Other_Source                 8528 non-null   int32
 15  Lead Source_Reference                    8528 non-null   int32
 16  Lead Source_Referral Sites               8528 non-null   int32
 17  Lead Source_Welingak Website             8528 non-null   int32
 18  Last Activity_Email Bounced              8528 non-null   int32
 19  Last Activity_Email Link Clicked         8528 non-null   int32
 20  Last Activity_Email Opened               8528 non-null   int32
 21  Last Activity_Form Submitted on Website  8528 non-null   int32
 22  Last Activity_Olark Chat Conversation    8528 non-null   int32
 23  Last Activity_Other_Activity             8528 non-null   int32
 24  Last Activity_Page Visited on Website    8528 non-null   int32
 25  Last Activity_SMS Sent                   8528 non-null   int32
 26  Last Activity_Unreachable                8528 non-null   int32
 27  Last Activity_Unsubscribed               8528 non-null   int32
 28  Specialization_Business Administration   8528 non-null   int32
 29  Specialization_E-Business                8528 non-null   int32
 30  Specialization_E-COMMERCE                8528 non-null   int32
 31  Specialization_Finance Management        8528 non-null   int32
 32  Specialization_Healthcare Management     8528 non-null   int32
 33  Specialization_Hospitality Management    8528 non-null   int32
 34  Specialization_Human Resource Management 8528 non-null   int32
 35  Specialization_IT Projects Management    8528 non-null   int32
 36  Specialization_International Business   8528 non-null   int32
 37  Specialization_Marketing Management      8528 non-null   int32
 38  Specialization_Media and Advertising     8528 non-null   int32
 39  Specialization_Operations Management     8528 non-null   int32
 40  Specialization_Others                    8528 non-null   int32
 41  Specialization_Retail Management         8528 non-null   int32
 42  Specialization_Rural and Agribusiness    8528 non-null   int32
 43  Specialization_Services Excellence       8528 non-null   int32
 44  Specialization_Supply Chain Management   8528 non-null   int32
 45  Specialization_Travel and Tourism        8528 non-null   int32
 46  Occupation_Housewife                     8528 non-null   int32
 47  Occupation_Other                         8528 non-null   int32
 48  Occupation_Student                       8528 non-null   int32
 49  Occupation_Unemployed                    8528 non-null   int32
 50  Occupation_Working Professional          8528 non-null   int32
 51  Tags_Busy                                8528 non-null   int32
 52  Tags_Closed by Horizzon                  8528 non-null   int32
 53  Tags_Graduation in progress              8528 non-null   int32
 54  Tags_Interested  in full time MBA        8528 non-null   int32
 55  Tags_Interested in other courses         8528 non-null   int32
 56  Tags_Lost to EINS                        8528 non-null   int32
 57  Tags_Not doing further education         8528 non-null   int32
 58  Tags_Other_Tags                          8528 non-null   int32
```

Dummy Variables

```
 59  Tags_Ringing                                8528 non-null   int32
 60  Tags_Will revert after reading the email    8528 non-null   int32
 61  Tags_invalid number                         8528 non-null   int32
 62  Lead Quality_Low in Relevance               8528 non-null   int32
 63  Lead Quality_Might be                       8528 non-null   int32
 64  Lead Quality_Not Sure                       8528 non-null   int32
 65  Lead Quality_Worst                          8528 non-null   int32
 66  City_Other Cities                           8528 non-null   int32
 67  City_Other Cities of Maharashtra            8528 non-null   int32
 68  City_Other Metro Cities                     8528 non-null   int32
 69  City_Thane & Outskirts                      8528 non-null   int32
 70  City_Tier II Cities                         8528 non-null   int32
 71  Notable_Email Link Clicked                  8528 non-null   int32
 72  Notable_Email Opened                        8528 non-null   int32
 73  Notable_Had a Phone Conversation            8528 non-null   int32
 74  Notable_Modified                            8528 non-null   int32
 75  Notable_Olark Chat Conversation             8528 non-null   int32
 76  Notable_Other_Last Notable Activity         8528 non-null   int32
 77  Notable_Page Visited on Website             8528 non-null   int32
 78  Notable_SMS Sent                            8528 non-null   int32
 79  Notable_Unreachable                         8528 non-null   int32
 80  Notable_Unsubscribed                        8528 non-null   int32
dtypes: float64(2), int32(74), int64(5)
```

Steps Before Model Creation

- Train-Test Split

```
# Splitting the data into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
```

- Feature Scaling

```
scaler = StandardScaler()

X_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']] = scaler.fit_transform(X_train[['TotalVisits','Total Time Spent on Website'
```

- Checking the Conversion Rate

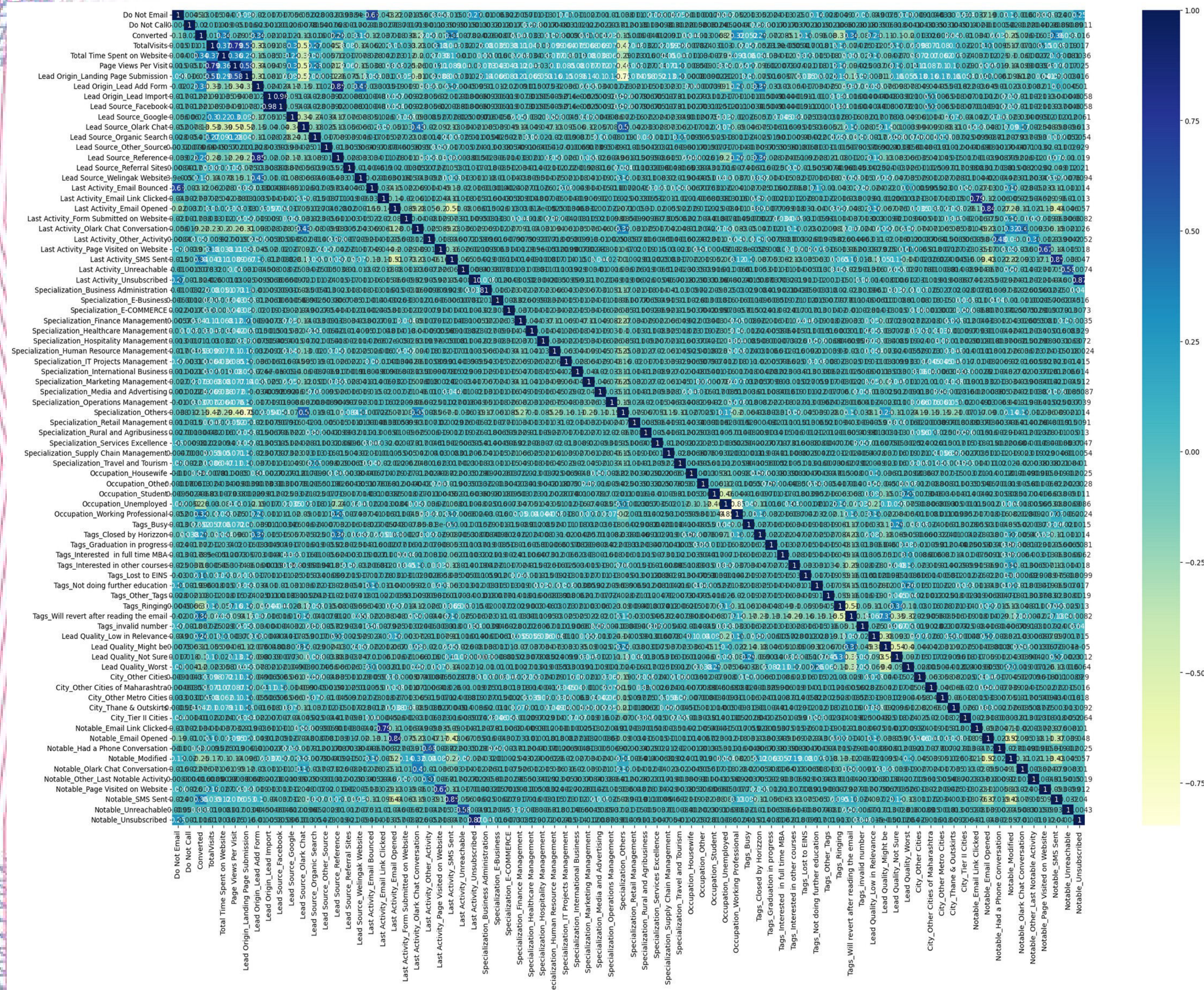
```
### Checking the Conversion Rate
convert = (sum(lead_df['Converted'])/len(lead_df['Converted'].index))*100
convert
```

```
37.453095684803
```

We have around 38% conversion rate. This is neither exactly 'balanced' (which a 50-50 ratio would be called) nor heavily imbalanced. So we'll not have to do any special treatment for this dataset.

Correlations

Heatmap



Do Not Email	0.63
TotalVisits	0.79
Page Views Per Visit	0.79
Lead Origin_Landing Page Submission	0.58
Lead Origin_Lead Add Form	0.85
Lead Origin_Lead Import	0.98
Lead Source_Facebook	0.98
Lead Source_Olark Chat	0.50
Lead Source_Reference	0.85
Last Activity_Email Bounced	0.63
Last Activity_Email Link Clicked	0.79
Last Activity_Email Opened	0.84
Last Activity_Page Visited on Website	0.67
Last Activity_SMS Sent	0.85
Last Activity_Unreachable	0.58
Last Activity_Unsubscribed	0.87
Specialization_Others	0.50
Notable_Email Link Clicked	0.79
Notable_Email Opened	0.84
Notable_Page Visited on Website	0.67
Notable_SMS Sent	0.85
Notable_Unreachable	0.58
Notable_Unsubscribed	0.87

The highly correlated variables are displayed above. We will proceed ahead with model building using RFE and based upon p-values decide to drop the highly related variables

Model Building

- Model 1 identified several variables which aren't really significant
- Hence we used the RFE method to identify the 15 most relevant features

['Lead Origin_Lead Add Form', 'Lead Source_Welingak Website','Occupation_Working Professional', 'Tags_Busy','Tags_Closed by Horizzon', 'Tags_Interested in full time MBA','Tags_Lost to EINS', 'Tags_Not doing further education', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Tags_invalid number', 'Lead Quality_Might be', 'Lead Quality_Not Sure', 'Lead Quality_Worst','Notable_SMS Sent']

- Model 2 built using the above features still had insignificant variables. We dropped the insignificant variables one at a time and validated the p-values and VIF after each model building
- At the end of Model 6 we got acceptable VIF and p-values

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6490	0.270	-9.827	0.000	-3.177	-2.121
Lead Origin_Lead Add Form	1.7007	0.297	5.718	0.000	1.118	2.284
Lead Source_Welingak Website	2.4144	0.785	3.077	0.002	0.877	3.952
Occupation_Working Professional	2.6095	0.232	11.235	0.000	2.154	3.065
Tags_Busy	3.5436	0.299	11.864	0.000	2.958	4.129
Tags_Closed by Horizzon	9.8689	1.045	9.441	0.000	7.820	11.918
Tags_Lost to EINS	9.6282	0.637	15.105	0.000	8.379	10.877
Tags_Will revert after reading the email	5.3223	0.243	21.875	0.000	4.845	5.799
Lead Quality_Might be	-3.9916	0.227	-17.613	0.000	-4.436	-3.547
Lead Quality_Not Sure	-2.0455	0.291	-7.030	0.000	-2.616	-1.475
Lead Quality_Worst	-3.1952	0.725	-4.405	0.000	-4.617	-1.774
Notable_SMS Sent	2.8425	0.114	24.946	0.000	2.619	3.066

	Features	VIF
6	Tags_Will revert after reading the email	2.82
7	Lead Quality_Might be	2.70
0	Lead Origin_Lead Add Form	1.65
10	Notable_SMS Sent	1.45
1	Lead Source_Welingak Website	1.30
4	Tags_Closed by Horizzon	1.23
2	Occupation_Working Professional	1.21
8	Lead Quality_Not Sure	1.19
3	Tags_Busy	1.12
5	Tags_Lost to EINS	1.05
9	Lead Quality_Worst	1.00

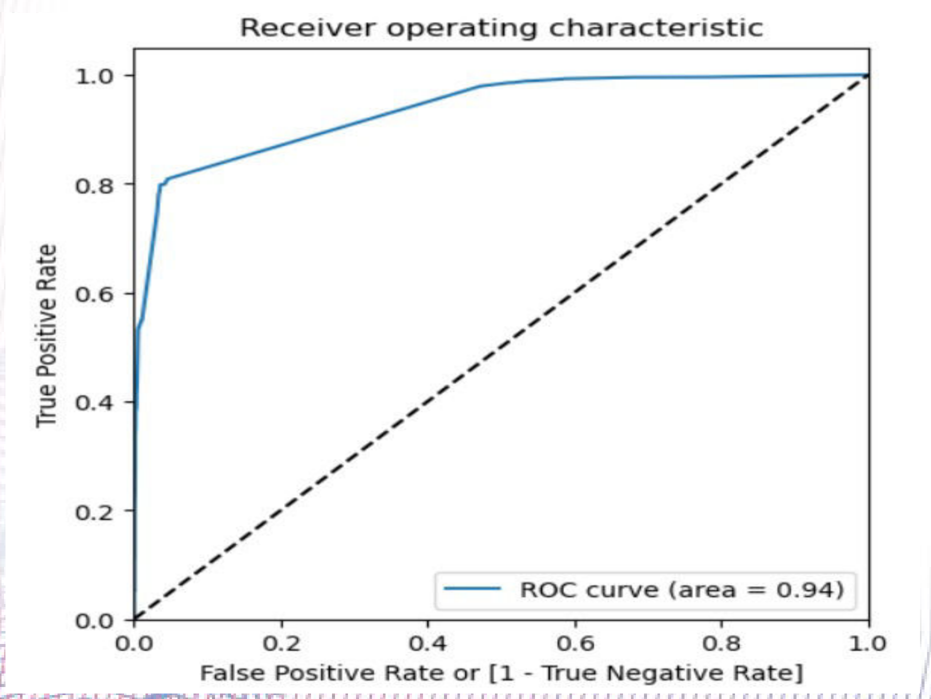
Making Predictions – Train Set

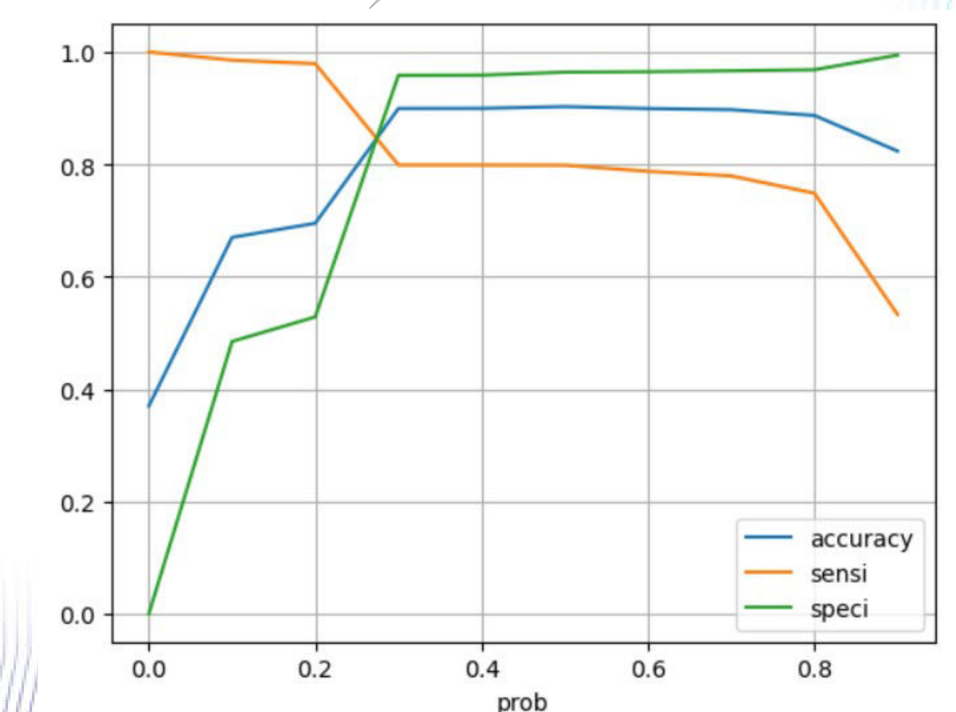
- Predictions on the train set with Probability threshold as 0.5 gave the following metrics
- Accuracy - 90%
- Sensitivity – 80%
- Specificity - 96%

```
Metrics for cutoff: 0.5
---------------------------------------------
confusion metrics
[[3626  135]
 [ 445 1763]]
Accuracy 0.9028312950242922
Sensitivity  0.7984601449275363
Specificity  0.9641052911459718
False positive rate  0.035894708854028186
Positive predictive value 0.928872497365648
Negative predictive value 0.8906902480962908
```

Optimal Probability 0.3

- Using ROC curve to find trade off between sensitivity and specificity





Making Predictions – Train Set

- Predictions on the train set with Probability threshold as 0.3 gave the following metrics

```
Metrics at threshold 0.3
confusion metrics
[[3605  156]
 [ 444 1764]]
Accuracy  0.8994806500251299
Sensitivity  0.7989130434782609
Specificity  0.9585216697686786
False positive rate  0.041478330231321456
Positive predictive value 0.91875
Negative predictive value  0.890343294640652
```

- Accuracy - 89%
- Sensitivity – 80%
- Specificity - 95%

- Calculating Precision And Recall – The Precision And Recall Scores are as below

```
precision_score(y_train_pred_final.Converted, y_train_pred_final.predicted)
```

```
0.928872497365648
```

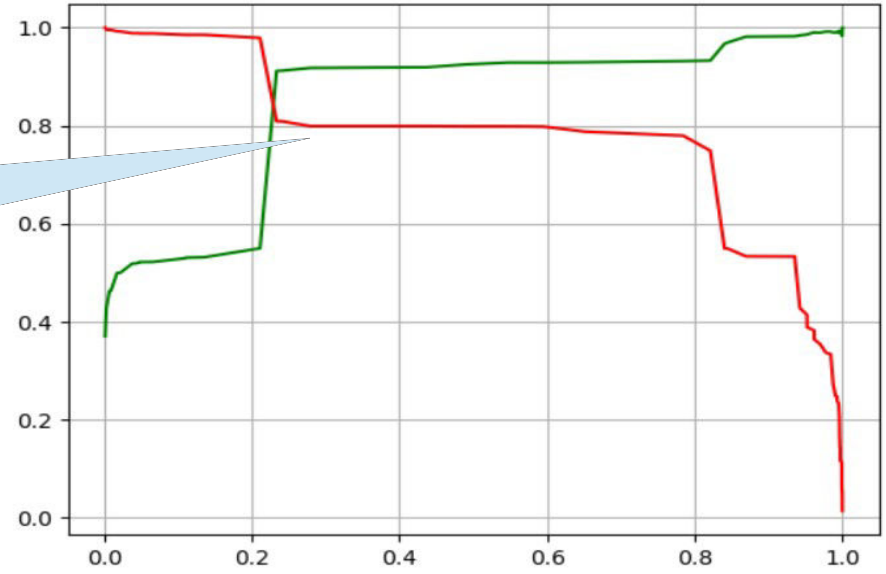
```
recall_score(y_train_pred_final.Converted, y_train_pred_final.predicted)
```

```
0.7984601449275363
```

Making Predictions – Train Set

- Precision and Recall trade off probability

Optimal Probability 0.23




```
Metrics at threshold 0.23
confusion metrics
[[3588  173]
 [ 422 1786]]
Accuracy  0.9003183112749205
Sensitivity  0.8088768115942029
Specificity  0.9540015953203935
False positive rate  0.04599840467960649
Positive predictive value 0.9116896375701888
Negative predictive value  0.8947630922693267
```

- Metrics at above threshold

- 0.23 is the optimal value. We will use this to make predictions on the test set

Making Predictions – Test Set

- Metrics on test set at 0.23 threshold gave the below metrics

```
Metrics at threshold 0.23
confusion metrics
[[1502   71]
 [ 178  808]]
Accuracy  0.902696365767878
Sensitivity  0.8194726166328601
Specificity  0.9548633184996821
False positive rate  0.045136681500317864
Positive predictive value 0.919226393629124
Negative predictive value  0.8940476190476191
```

- Conclusion – From the above metrics and after comparing it with the train set metrics we can conclude that the model is a good fit and it is not over trained

Lead Score Calculation

- We created a consolidated single dataframe for test and train dataset
- The lead score was calculated by using the formula

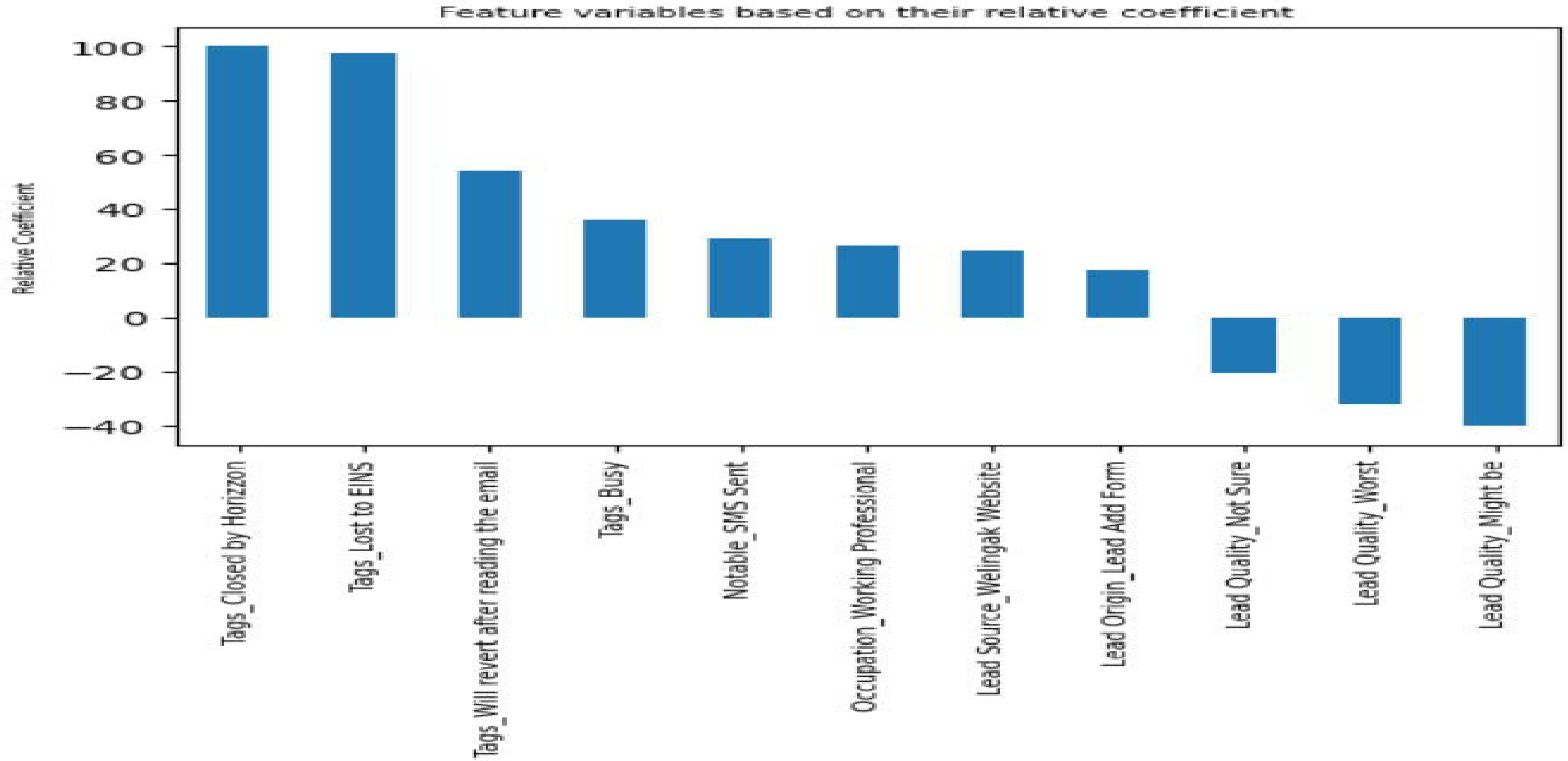
Converted Probability * 100

	Converted	Converted_Prob	final_predicted	Lead_Score
Lead Number				
579533	1	0.21	0	21
579538	1	0.82	1	82
579545	0	0.14	0	14
579546	0	0.02	0	2
579615	1	0.21	0	21

Deriving Top and Bottom 3 features

- We created relative coefficient values for all the features from the final model , except interept
- The Top and Bottom features are depicted below

```
Lead Origin_Lead Add Form                     17.23
Lead Source_Welingak Website                  24.46
Occupation_Working Professional               26.44
Tags_Busy                                     35.91
Tags_Closed by Horizzon                      100.00
Tags_Lost to EINS                             97.56
Tags_Will revert after reading the email      53.93
Lead Quality_Might be                        -40.45
Lead Quality_Not Sure                        -20.73
Lead Quality_Worst                           -32.38
Notable_SMS Sent                              28.80
dtype: float64
```



Conclusion And Recommendations

- Following are the top 3 features that contribute most towards the probability of a lead getting converted
  - Tags_Closed by Horizzon
  - Tags_Lost to EINS
  - Tags_Will revert after reading the email
- Following are the bottom 3 features that need improvement in order to convert a lead
  - Lead Quality_Might be
  - Lead Quality_Worst
  - Lead Quality_Not Sure
- This model will help to identify the hot leads i.e. a lead which is most likely to convert based on the computed lead score for each of the leads in the dataset. The sales team can target leads with a higher lead score so as to -
  - Increase revenue
  - Informed prioritization of sales cycle
  - Increase market effectiveness
  - Reduce opportunity loss
- In addition the model will also enable the sales team to become more aggressive when they have more manpower to chase the hot leads by tweaking the conversion probabilities
- Also when the targets are met and the sales team wants to focus more on other work and less on unnecessary phone calls they can choose higher threshold for conversion probability and target only those leads which have a very very high conversion probability

THANK YOU