Lead Scoring Case Study

Presented By

1. Surendra Bursu

2. Jayanth Srivastava

3.Sai Nikhil

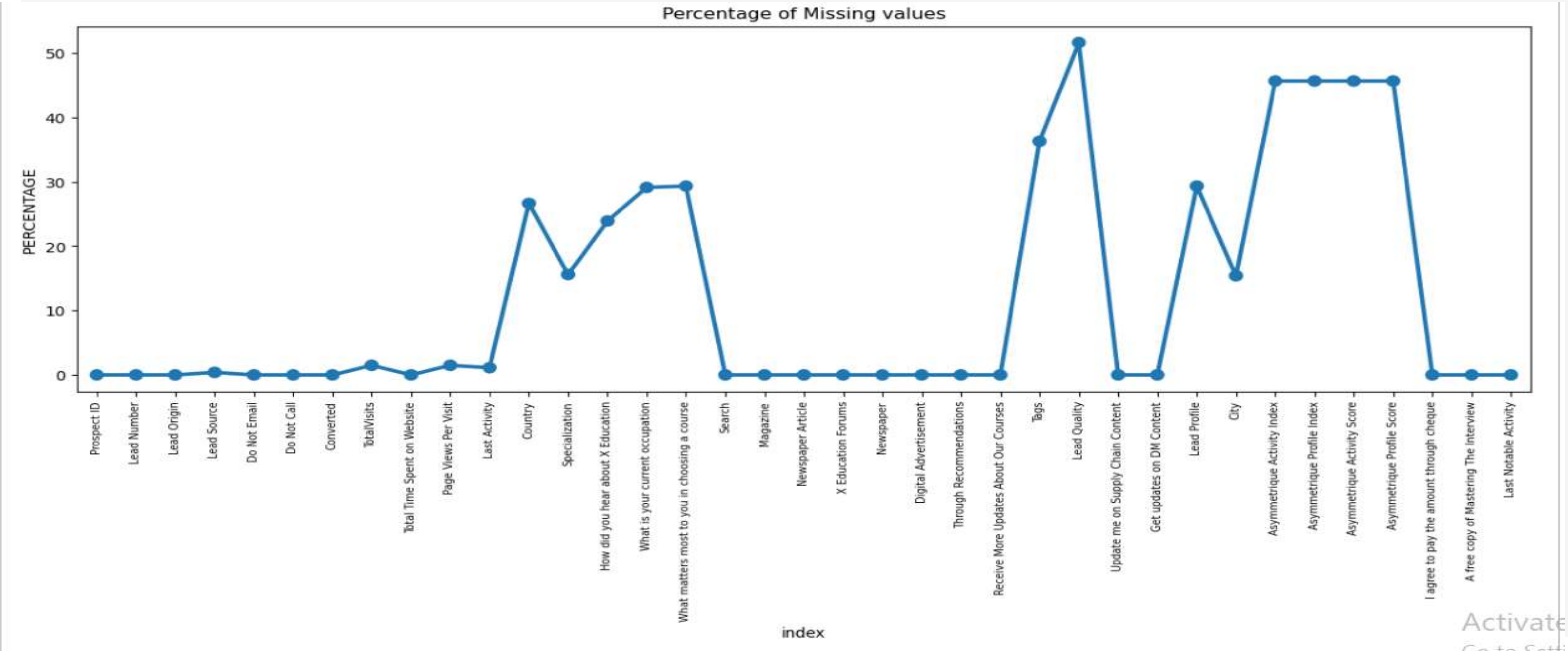
Problem Statement: X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Approach: We have build this model using Logistic regression along with RFE and VIF, to get top features and based on that we have provided recommendations to the company.

Below mentioned is the list of methodologies which we followed while building the model

1. EDA
2. Dummy Creation
3. Train_test Split
4. Model Building
5. Metrices score and Analysis

EDA: We checked for null values in the dataset, and found that there are many null values as well as 'select' values which needs to be addressed, we capped the null values to 40%, anything above 40% was dropped.



Missing Value Treatment: we treated missing values by imputing them with mode, also replaced 'Select' with other values as mentioned in problem statement

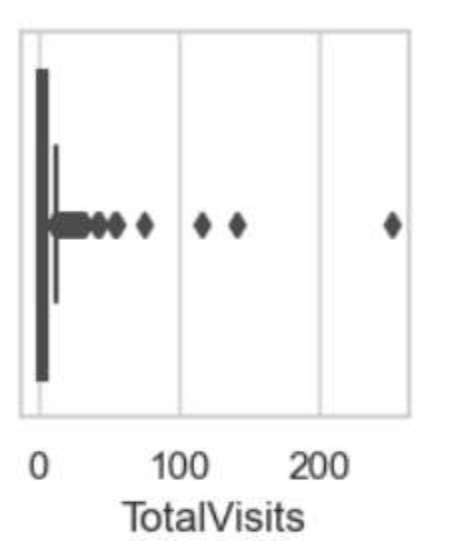
In [22]:
```
#checking value counts of Specialization column

leads_data['Specialization'].value_counts(dropna=False)
```

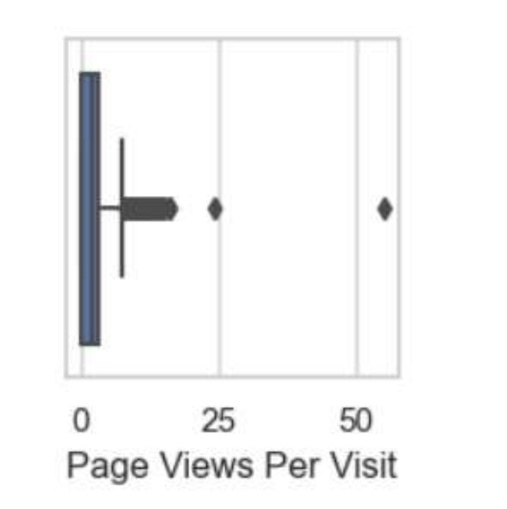
Out[22]:
```
Select                               1942
NaN                                  1438
Finance Management                    976
Human Resource Management             848
Marketing Management                  838
Operations Management                 503
Business Administration               403
IT Projects Management                366
Supply Chain Management               349
Banking, Investment And Insurance     338
Travel and Tourism                    203
Media and Advertising                 203
International Business                178
Healthcare Management                 159
Hospitality Management                114
E-COMMERCE                            112
Retail Management                     100
Rural and Agribusiness                 73
E-Business                             57
Services Excellence                    40
Name: Specialization, dtype: int64
```

In [23]:
```
# Lead may not have mentioned specialization because it was not in the list or maybe they are a students
# and don't have a specialization yet. So we will replace NaN values here with 'Not Specified'

leads_data['Specialization'] = leads_data['Specialization'].replace(np.nan, 'Specialization_Not Specified')
leads_data['Specialization'] = leads_data['Specialization'].replace('Select', 'Specialization_Not Specified')
```

Outlier Check: We did some univariate analysis and then outlier treatment these were some potential outliers we did capping of 99% .

```
In [83]: # Removing values beyond 99% for Total Visits

         nn_quartile_total_visits = leads_data['TotalVisits'].quantile(0.99)
         leads_data = leads_data[leads_data["TotalVisits"] < nn_quartile_total_visits]
         leads_data["TotalVisits"].describe(percentiles=[.25,.5,.75,.90,.95,.99])
```





Bivariate Analysis: We did some bivariate analysis, and these are the inferences

1)people living in Mumbai have slight good conversion ratio,

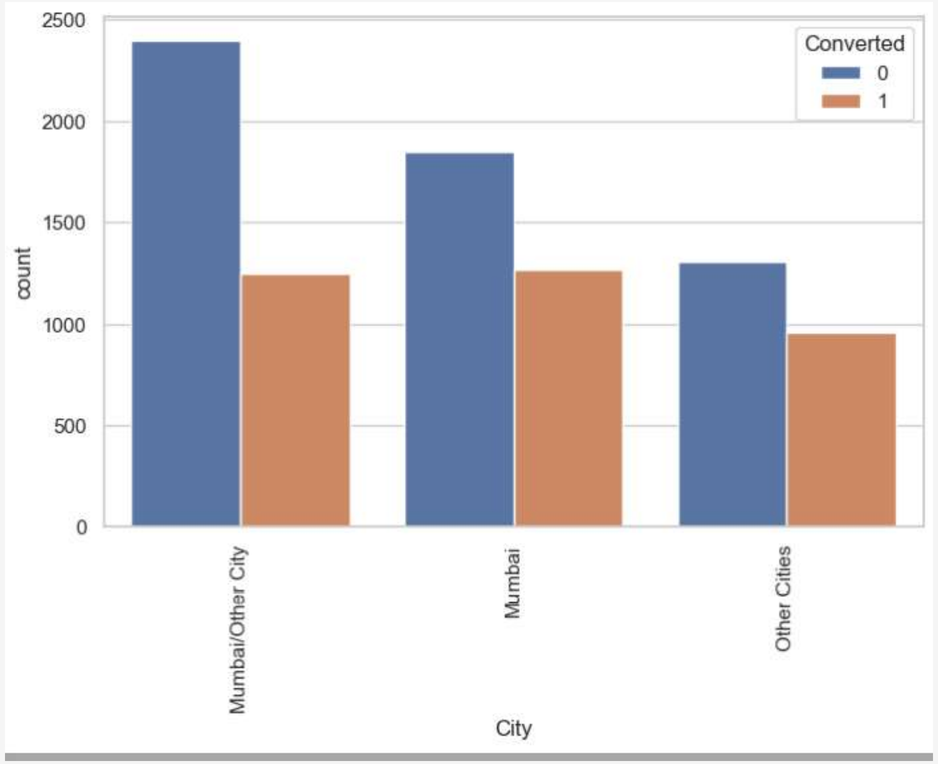
2) Management specialization's have good conversion ratio,

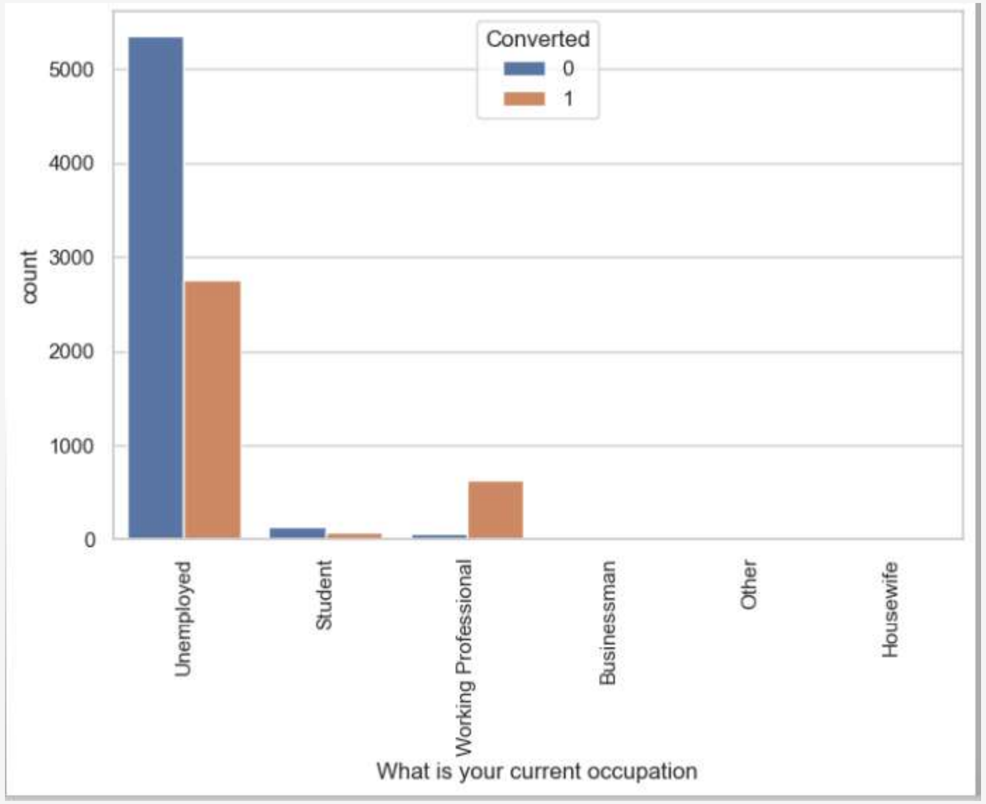
3)Unemployed people have good conversion ratio,

4)TAGS who will revert after reading email have better chance of getting converted into successful lead,

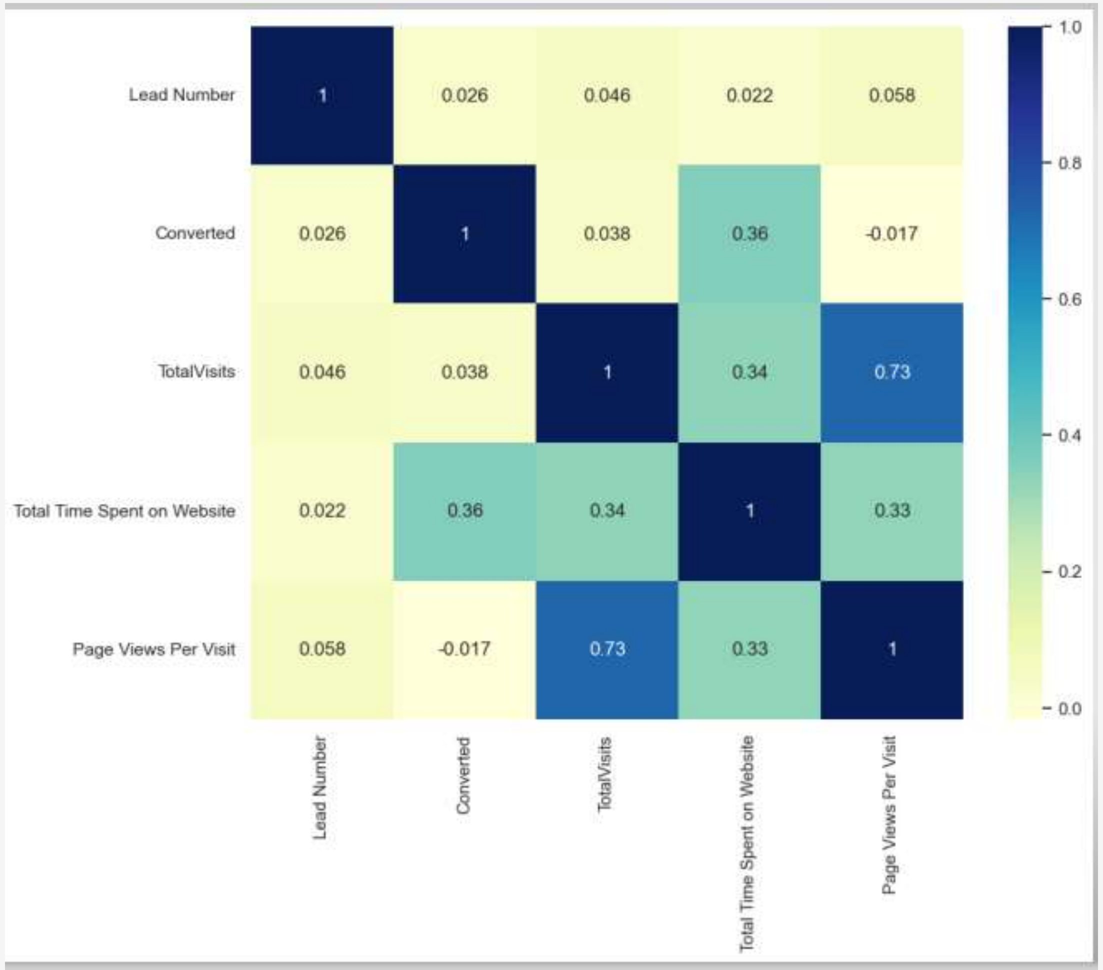
5)SMS sent have higher conversion ratio,

6) Those who said yes to receiving email have higher chance of getting converted


2500
2000
1500
1000
500
0
count
Converted
0
1
Mumbai/Other City
Mumbai
Other Cities
City


5000
4000
3000
2000
1000
0
count
Converted
0
1
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

Bivariate Analysis: Below is the correlation matrix, 'total visits' have high correlation with 'leads number'



Model Building: We build model using Logistic Regression, with help of Rfe and VIF we did 11 iterations and dropped columns with high pvalues and VIF with >5, we finally got the model on 11th iteration, Here is what the final model looks like

Generalized Linear Model Regression Results

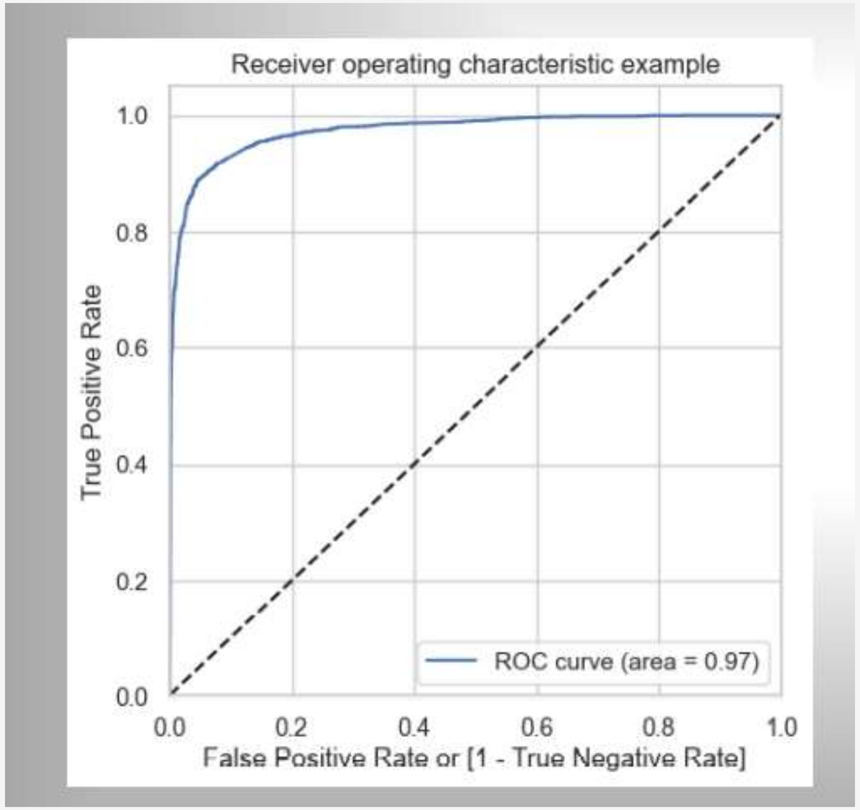
Dep. Variable:	Converted	No. Observations:	6320
Model:	GLM	Df Residuals:	6292
Model Family:	Binomial	Df Model:	27
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1210.8
Date:	Sun, 13 Nov 2022	Deviance:	2421.5
Time:	22:39:52	Pearson chi2:	8.72e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.6119
Covariance Type:	nonrobust		

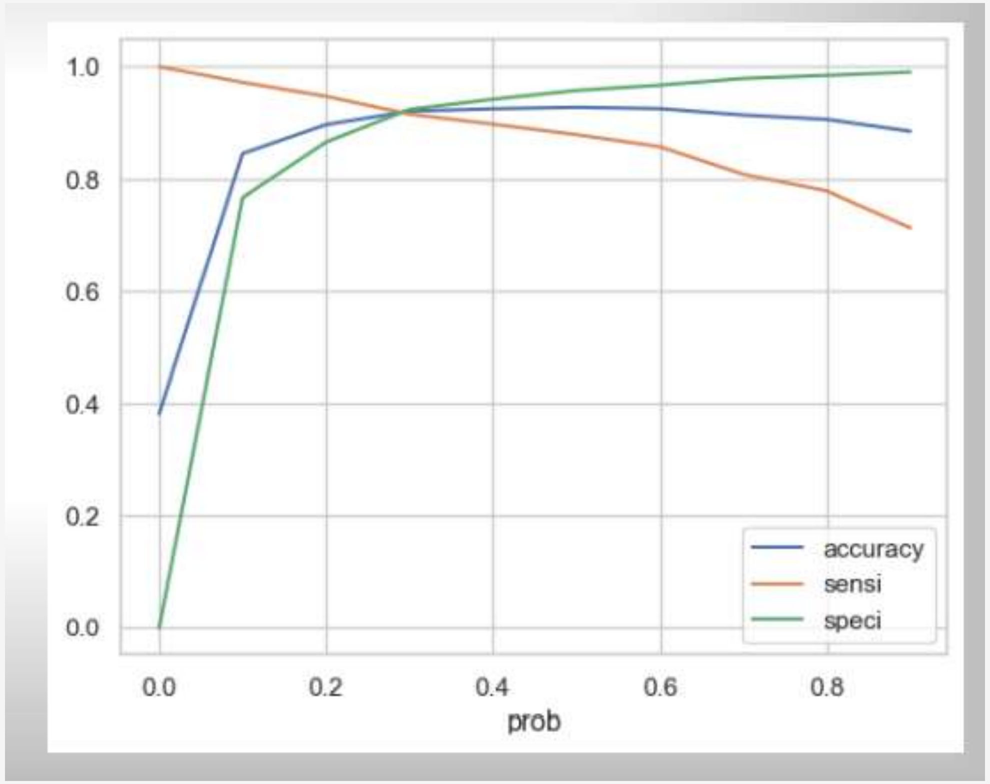
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7623	0.358	-2.130	0.033	-1.464	-0.061
Total Time Spent on Website	1.0238	0.062	16.475	0.000	0.902	1.146
Lead Origin_Landing Page Submission	-1.0234	0.226	-4.525	0.000	-1.467	-0.580
Lead Origin_Lead Add Form	2.2494	1.214	1.853	0.064	-0.130	4.628
What is your current occupation_Working Professional	0.6390	0.366	1.745	0.081	-0.079	1.357
Specialization_E-Business	-0.4411	0.667	-0.661	0.509	-1.749	0.867
Specialization_Specialization_Not Specified	-0.3933	0.218	-1.807	0.071	-0.820	0.033
Specialization_Travel and Tourism	-0.5026	0.436	-1.154	0.249	-1.356	0.351
Lead Source_Olark Chat	0.8853	0.171	5.186	0.000	0.551	1.220
Lead Source_Others	0.5275	0.889	0.594	0.553	-1.214	2.269
Lead Source_Reference	-1.4246	1.278	-1.115	0.265	-3.930	1.080
Lead Source_Welingak Website	3.0458	1.413	2.156	0.031	0.277	5.814
Last Activity_Email Bounced	-1.0168	0.494	-2.058	0.040	-1.985	-0.048
Last Activity_Email Opened	0.6047	0.199	3.042	0.002	0.215	0.994
Last Activity_Form Submitted on Website	0.8128	0.518	1.569	0.117	-0.202	1.828
Last Activity_Olark Chat Conversation	-0.4774	0.315	-1.517	0.129	-1.094	0.139
Last Activity_SMS Sent	1.3317	0.269	4.946	0.000	0.804	1.859
Last Notable Activity_Modified	-0.9571	0.203	-4.723	0.000	-1.354	-0.560
Last Notable Activity_Olark Chat Conversation	-0.6241	0.521	-1.198	0.231	-1.645	0.397
Last Notable Activity_Other_Notable_activity	1.1606	0.459	2.528	0.011	0.261	2.060
Last Notable Activity_SMS Sent	1.3058	0.299	4.374	0.000	0.721	1.891
Tags_Closed by Horizzon	6.3224	1.046	6.044	0.000	4.272	8.373
Tags_Interested in other courses	-3.0407	0.497	-6.119	0.000	-4.015	-2.067
Tags_Lost to EINS	5.0551	0.645	7.834	0.000	3.790	6.320
Tags_Not Specified	-0.6802	0.236	-2.886	0.004	-1.142	-0.218
Tags_Other_Tags	-3.2966	0.323	-10.203	0.000	-3.930	-2.663
Tags_Ringing	-3.8994	0.316	-12.340	0.000	-4.519	-3.280
Tags_Will revert after reading the email	3.6995	0.291	12.719	0.000	3.129	4.270

Features VIF

Metrices check and Analysis:

We did some analysis using roc curve and kept the threshold at 0.3, and using probability column we multiplied by 100 to get lead score





Metrices check and Analysis: We performed accuracy, recall, sensitivity, specificity, Here is a snapshot of result on test data set.

```
4    9182    0    0.025378    3    0

In [197]: # Let's check the overall accuracy.
          metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_Predicted)

Out[197]: 0.9276485788113695

In [198]: confusion2 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.final_Predicted )
          confusion2

Out[198]: array([[1532,  111],
                 [  85,  981]], dtype=int64)

In [199]: TP = confusion2[1,1] # true positive
          TN = confusion2[0,0] # true negatives
          FP = confusion2[0,1] # false positives
          FN = confusion2[1,0] # false negatives

In [200]: TP / float(TP+FN)
          #sensitivity is 91

Out[200]: 0.9202626641651032

In [201]: # Let us calculate specificity
          TN / float(TN+FP)

Out[201]: 0.9324406573341448

In [202]: precision_score(y_pred_final.Converted , y_pred_final.final_Predicted)

Out[202]: 0.8983516483516484

          precision score is 91.3

In [203]: recall_score(y_pred_final.Converted, y_pred_final.final_Predicted)

Out[203]: 0.9202626641651032

          recall score is 91.7
```

Inferences/Recommendations

Tags_Closed by Horizzon

Tags_Lost to EINS

Lead Source_Welingak Website

These are the top factors which can help in generating more successful leads, Also if there is a scenario where company wants lead conversion to be more aggressive then in that scenario , high sensitivity can be used. And if there is a scenario where company reaches a target before its quarter, for that we can use high specificity