Lead Scoring Case Study

Presented by:

1. Ajay Sharma
2. Parshav Shivnani
3. Nags Sai Siva Kumar Koppsetty

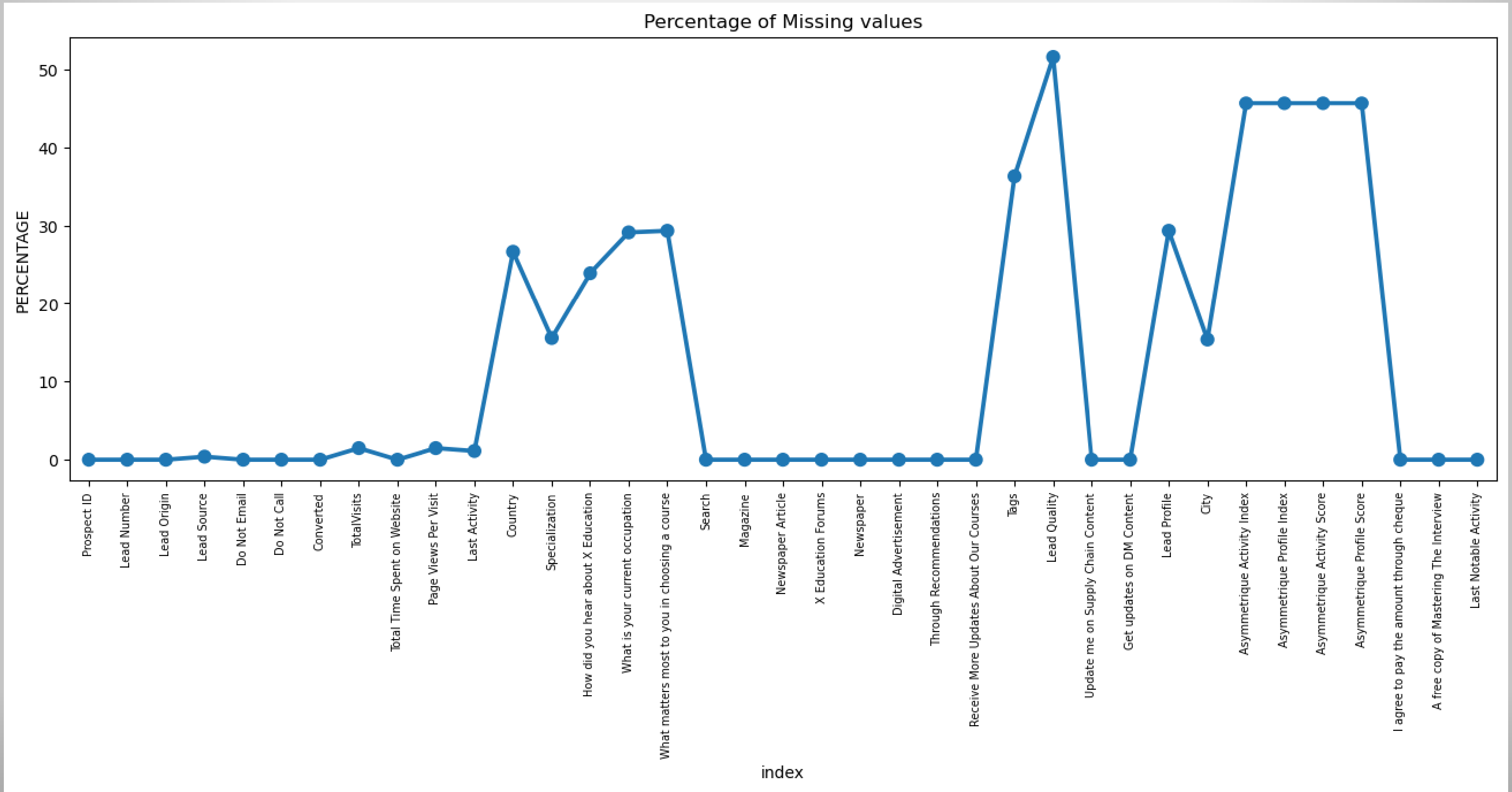
Problem Statement: X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher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

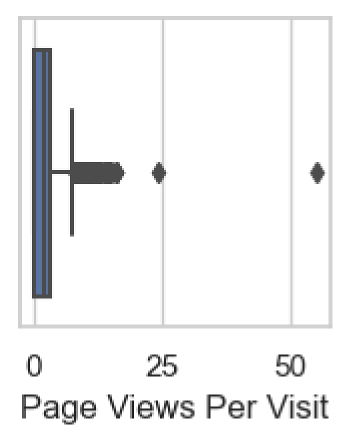
```
leads_data['Specialization'].value_counts(dropna=False)
```

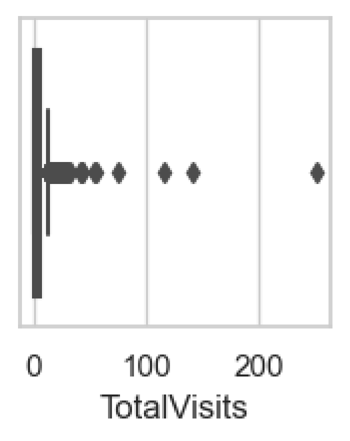
```
Out[22]: Select                               1942
         NaN                                  1438
         Finance Management                    976
         Human Resource Management             848
         Marketing Management                  838
         Operations Management                 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])

Out[83]: count     9141.000000
```





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

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

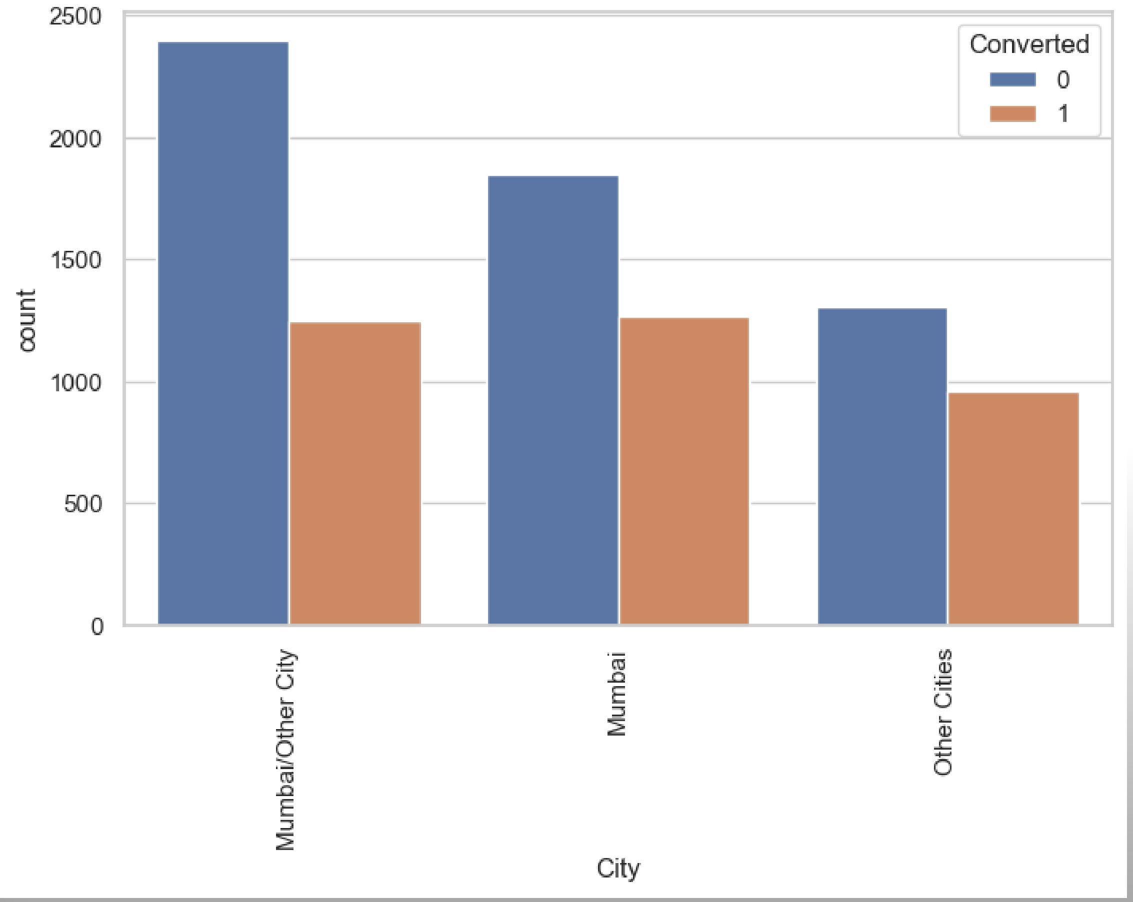
2) Management specialisations have good conversion ratio,

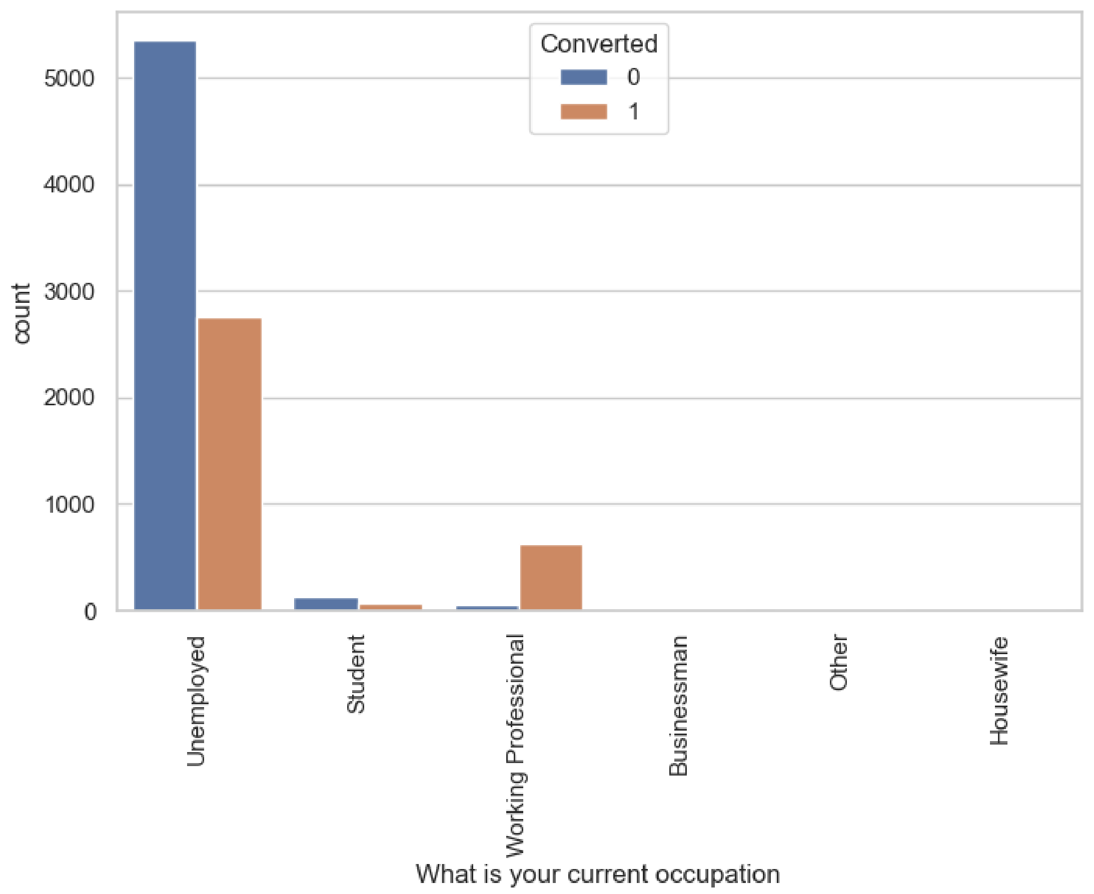
3Unemployed people have good conversion ratio,

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

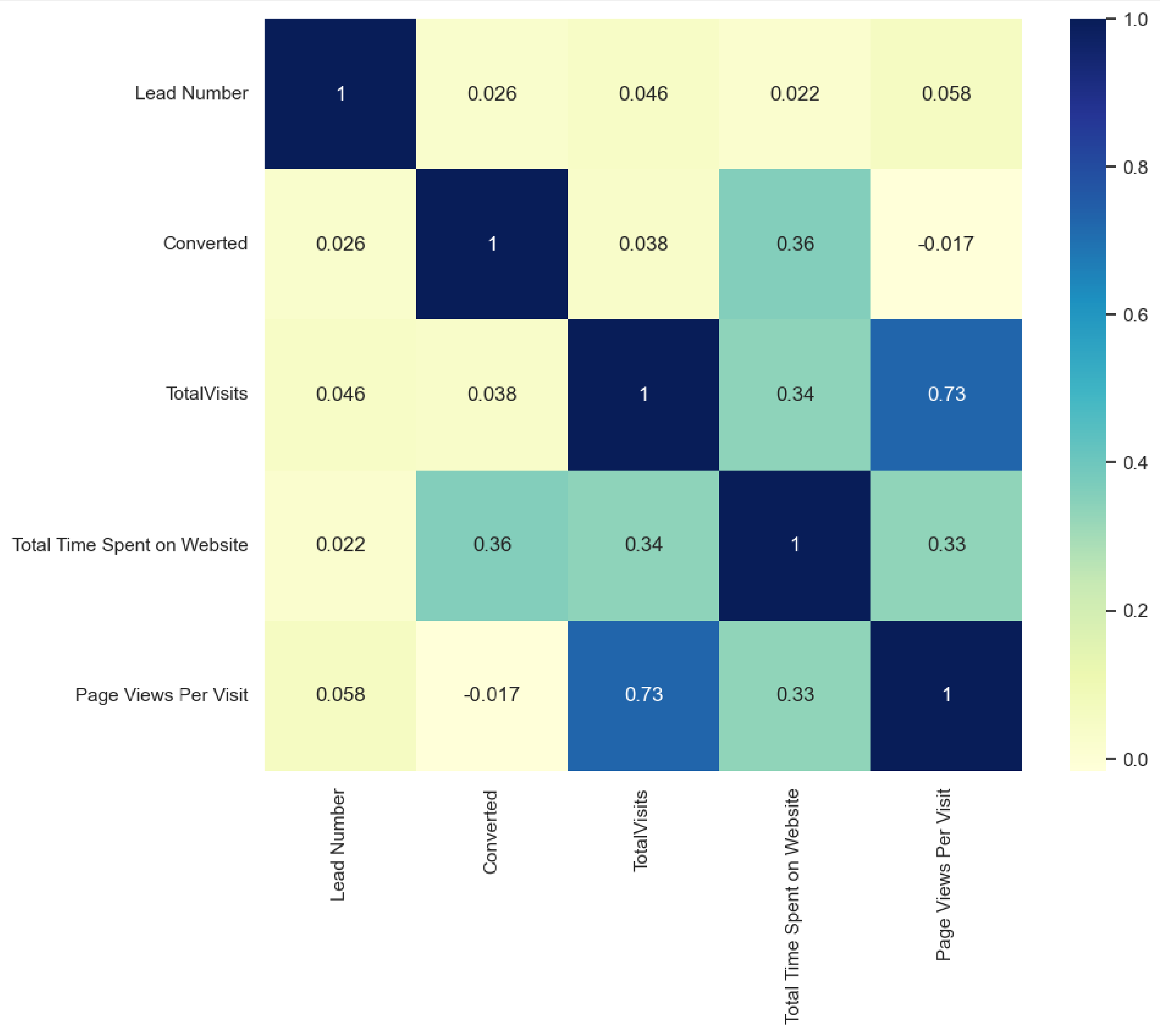
5)SmS sent have higher conversion ratio,

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





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



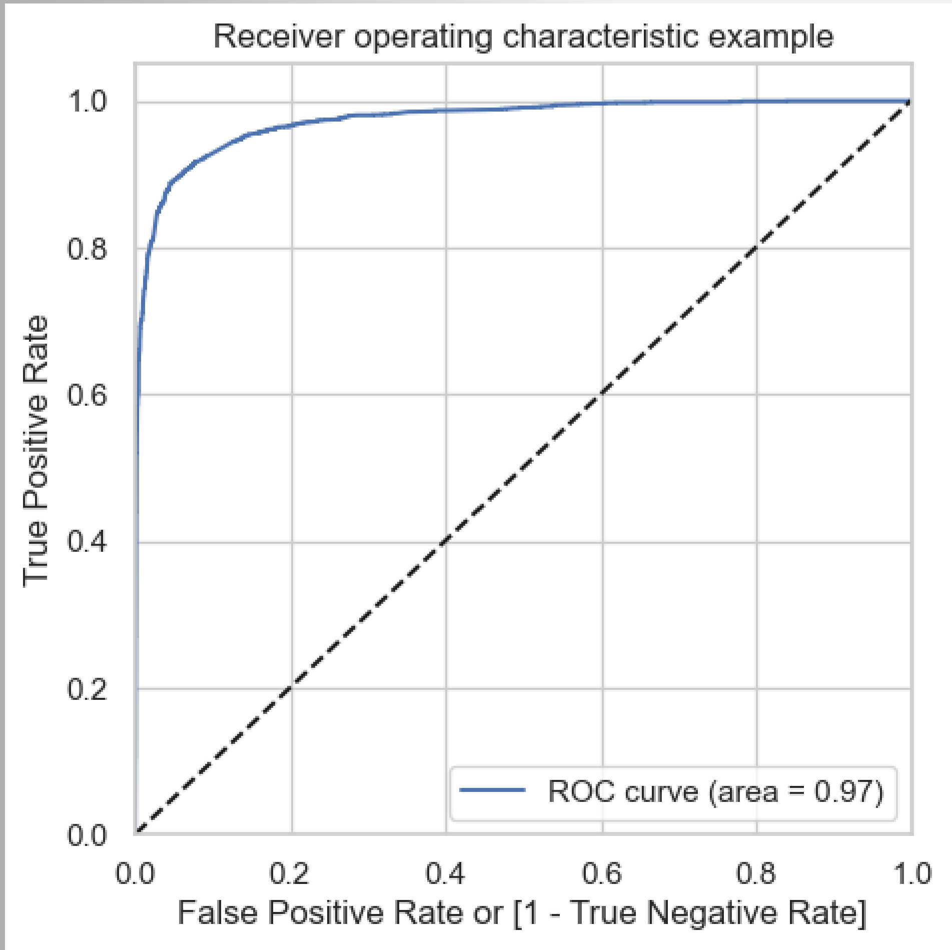
Model Building: We build model using Logistic Regression, with help of Rfe and VIF we did 11 iterations and dropped columns with high pvalues and VIF with >5, we finally got the model on 11th iteration, Here is what the final model looks like

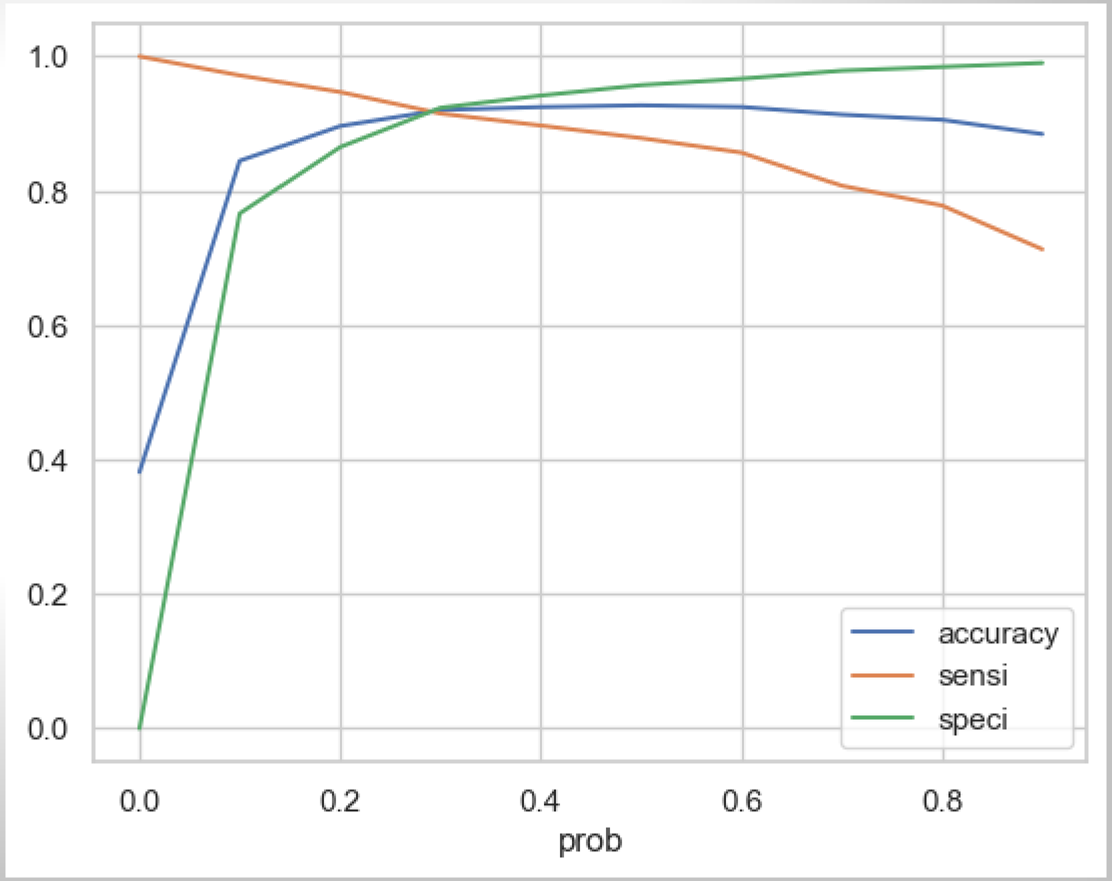
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6320
Model:	GLM	Df Residuals:	6302
Model Family:	Binomial	Df Model:	17
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1233.6
Date:	Sun, 13 Nov 2022	Deviance:	2467.1
Time:	22:39:55	Pearson chi2:	9.98e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.6091
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8949	0.257	-3.488	0.000	-1.398	-0.392
Total Time Spent on Website	1.0027	0.061	16.524	0.000	0.884	1.122
Lead Origin_Landing Page Submission	-1.1994	0.211	-5.688	0.000	-1.613	-0.786
Specialization_Specialization_Not Specified	-0.5360	0.210	-2.557	0.011	-0.947	-0.125
Lead Source_Olark Chat	0.7746	0.162	4.768	0.000	0.456	1.093
Lead Source_Others	1.7086	0.591	2.889	0.004	0.549	2.868
Lead Source_Welingak Website	5.2574	0.744	7.062	0.000	3.798	6.717
Last Activity_Email Opened	0.3228	0.164	1.965	0.049	0.001	0.645
Last Activity_Form Submitted on Website	1.2700	0.499	2.543	0.011	0.291	2.249
Last Activity_SMS Sent	2.2247	0.163	13.660	0.000	1.906	2.544
Last Notable Activity_Modified	-1.7084	0.142	-12.004	0.000	-1.987	-1.429
Last Notable Activity_Olark Chat Conversation	-1.4552	0.447	-3.257	0.001	-2.331	-0.580
Tags_Closed by Horizzon	7.7194	1.016	7.598	0.000	5.728	9.711
Tags_Interested in other courses	-2.3763	0.450	-5.285	0.000	-3.257	-1.495
Tags_Lost to EINS	5.9733	0.608	9.831	0.000	4.782	7.164

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.

```
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
```

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