CASE STUDY: LEADS SCORING

BY:

SHRUTI DHANGE

PRAFULLA MEHROTRA

AMIT BARMAN

PROBLEM STATEMENT:

Lead Conversion Rate of X Education is very poor which is approximately 30%.

BUSINESS OBJECTIVE:

1. X Education wishes to identify the most potential leads, also known as ‘Hot Leads’ so that the Lead Conversion Rate should go up as the sales team will be focusing more on communicating with the potential leads only.

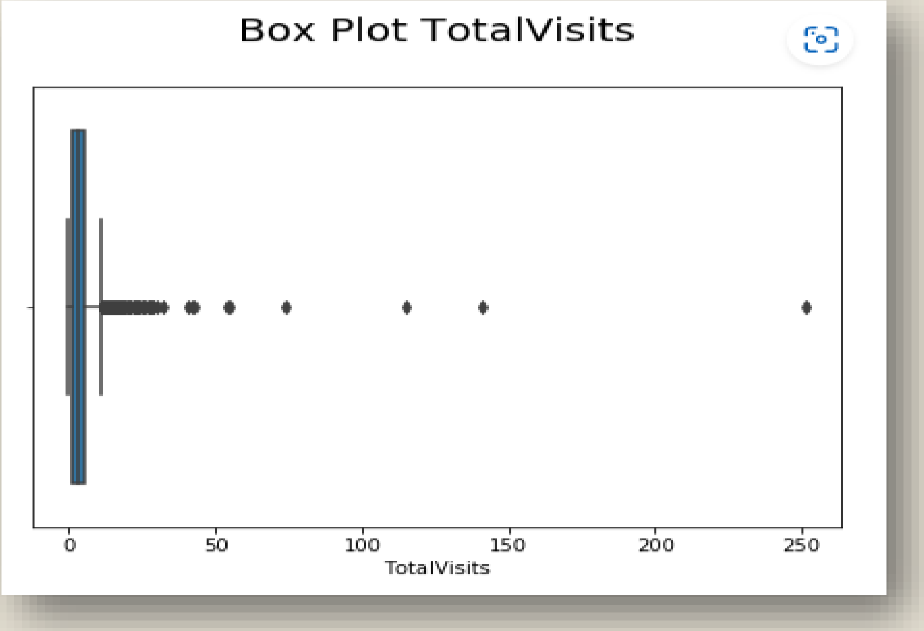
2. Building a model wherein a lead score is assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

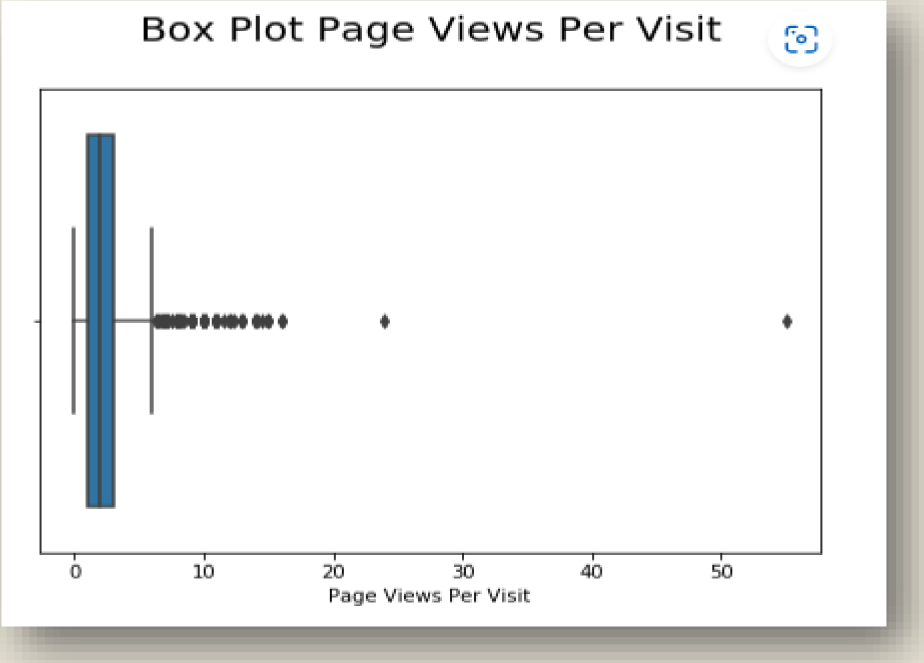
Approach of the analysis

I. We started by reading and understanding the data.

II. Next, we checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached on the right side.

III. We did Data Cleaning and removed columns with high missing value percentage >40%.



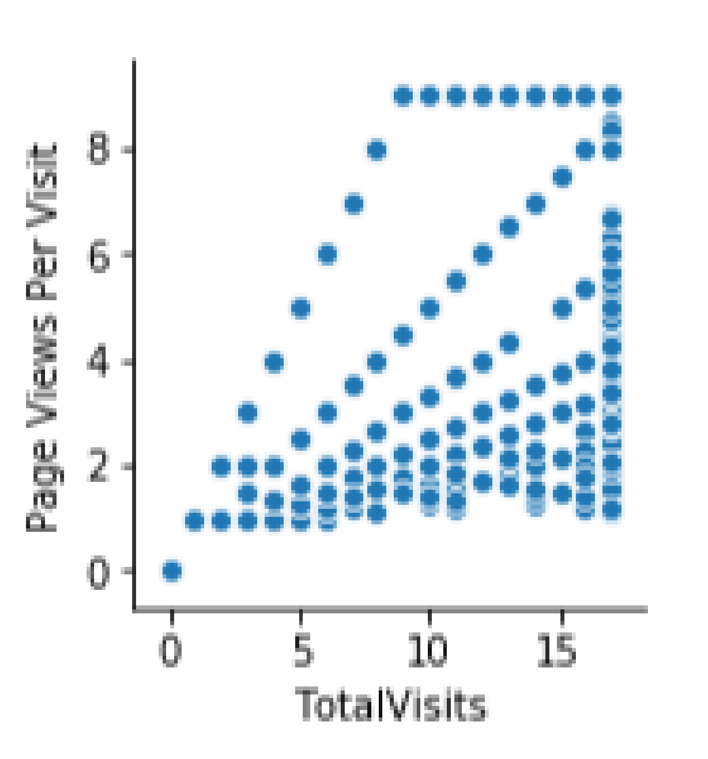


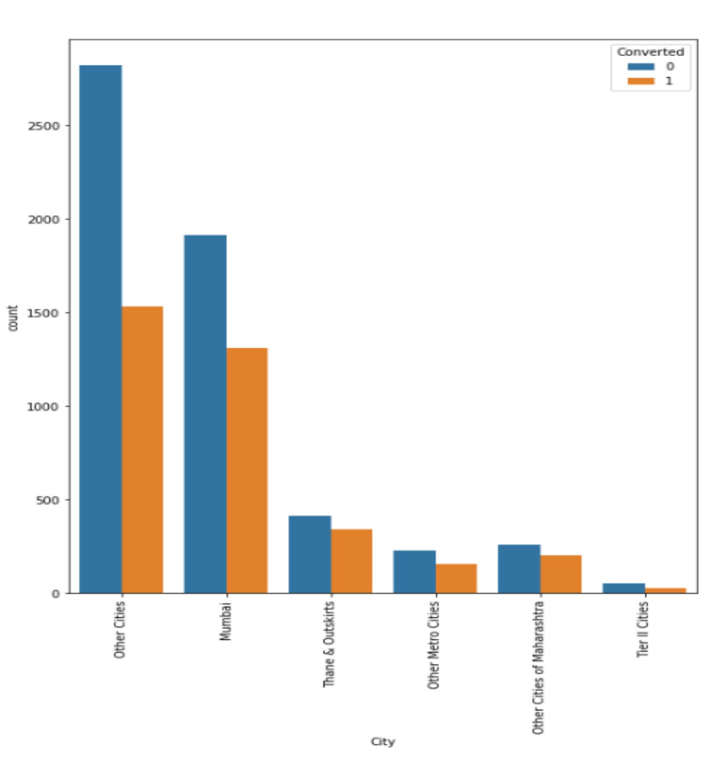
Data Visualization

We did data visualization for both categorical and continuous columns and noticed that:

I. For continuous columns variable TotalVisit has some correlation with variable Page Views Per Visit.

II. For categorical variables City Mumbai has good impact on potential lead.





Correlation

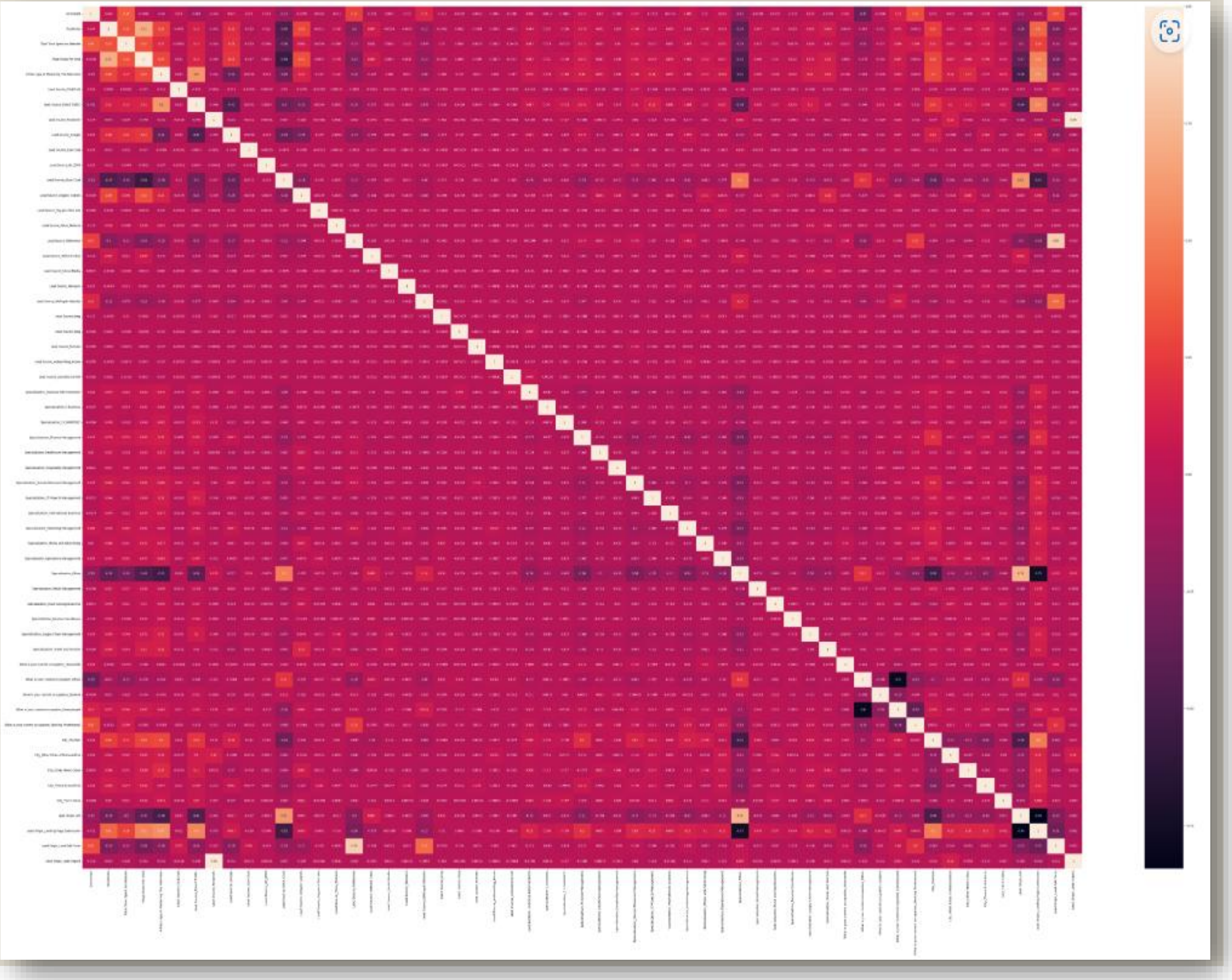
After fixing the outliers and dummy creation we proceed with our next step of analysis which is data preparation.

I. We split the dataset into train and test set and do standardization on the features.

II. Standardization is required in order to keep all the variables in same scale which will help us in computation in more efficient way.

III. Checked the correlation of the dataset. Attached heatmap is showing the correlation of all features present in the dataset.

IV. There were some high correlations in the heatmap which we dropped further.



Building a Model RFE

I. We build a model with all the features included and found there were many insignificant variables present in our model.

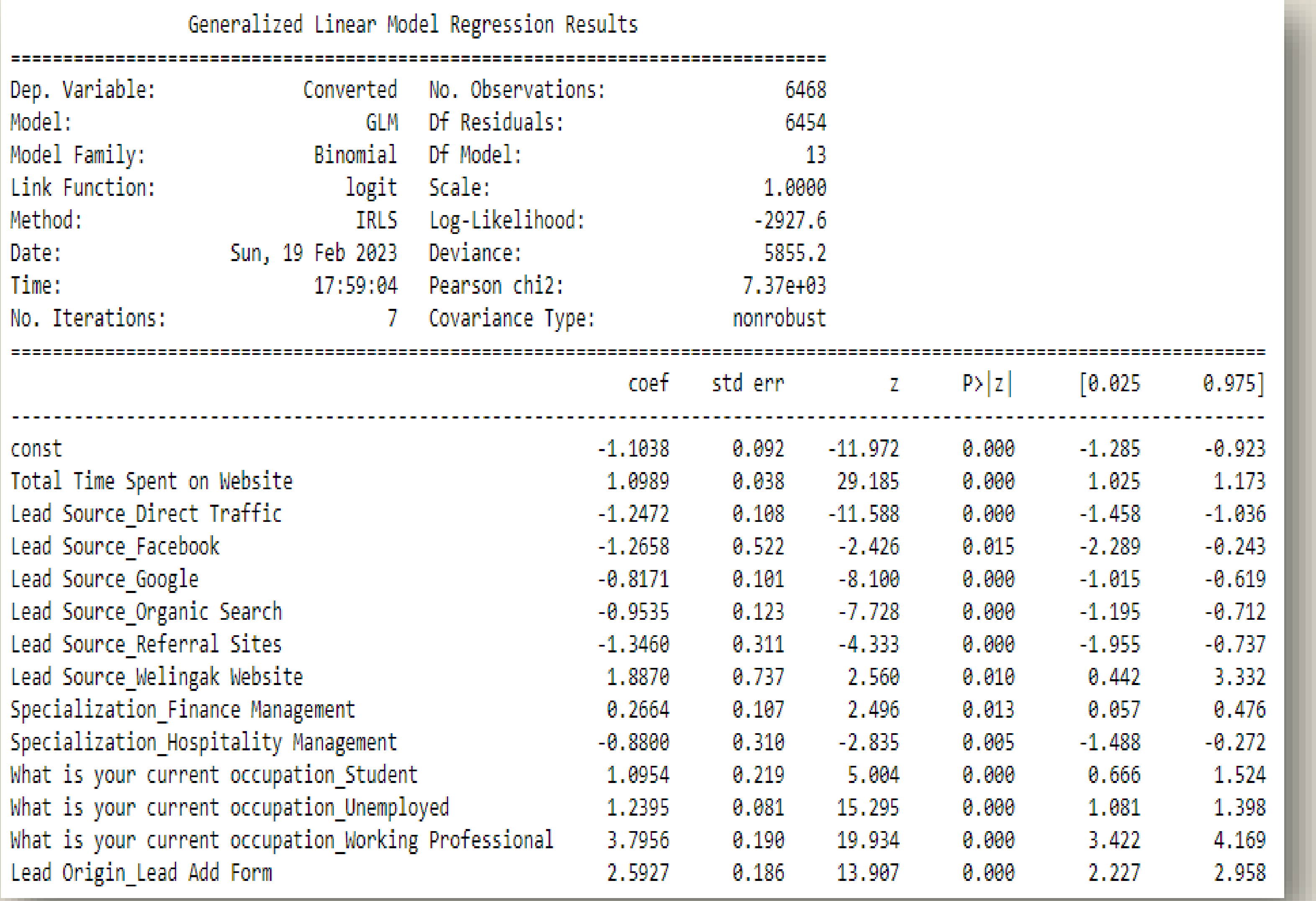
II. We need to drop them, but we can’t do it one by one as it is time consuming and not an efficient way to do so.

III. Hence, we started with RFE method to deduct those insignificant variables. We choose with RFE count 25.

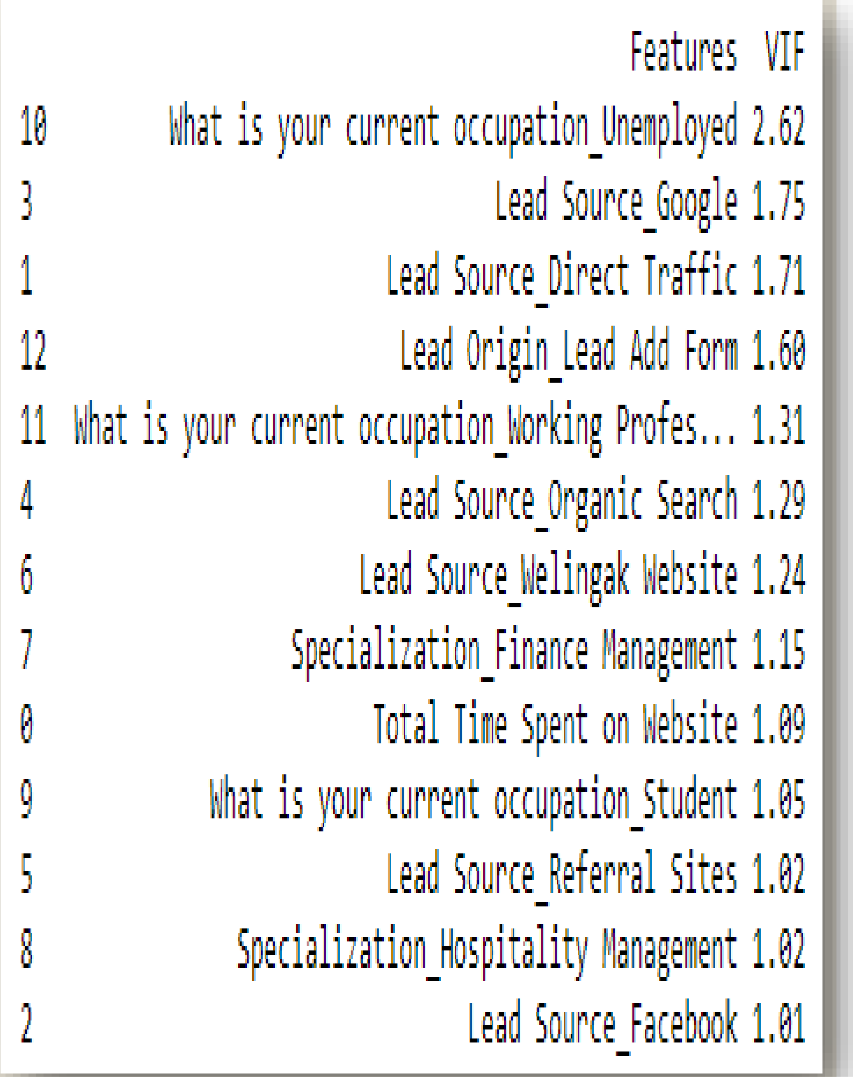
IV. We started creating our model with RFE count 25 and went dropping variables one by one until we reach the point where the model is having all significant variables and low VIF values.

V. Now we evaluated our model by first predicting it. We created new dataset with original converted values and the prediction values.

Final model visualization with VIF



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6454
Model Family:                Binomial   Df Model:                           13
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2927.6
Date:                Sun, 19 Feb 2023   Deviance:                       5855.2
Time:                        17:59:04   Pearson chi2:                 7.37e+03
No. Iterations:                     7   Covariance Type:             nonrobust
========================================================================================================================
                                                           coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                                   -1.1038      0.092    -11.972      0.000      -1.285      -0.923
Total Time Spent on Website                              1.0989      0.038     29.185      0.000       1.025       1.173
Lead Source_Direct Traffic                              -1.2472      0.108    -11.588      0.000      -1.458      -1.036
Lead Source_Facebook                                    -1.2658      0.522     -2.426      0.015      -2.289      -0.243
Lead Source_Google                                      -0.8171      0.101     -8.100      0.000      -1.015      -0.619
Lead Source_Organic Search                              -0.9535      0.123     -7.728      0.000      -1.195      -0.712
Lead Source_Referral Sites                              -1.3460      0.311     -4.333      0.000      -1.955      -0.737
Lead Source_Welingak Website                             1.8870      0.737      2.560      0.010       0.442       3.332
Specialization_Finance Management                        0.2664      0.107      2.496      0.013       0.057       0.476
Specialization_Hospitality Management                   -0.8800      0.310     -2.835      0.005      -1.488      -0.272
What is your current occupation_Student                  1.0954      0.219      5.004      0.000       0.666       1.524
What is your current occupation_Unemployed               1.2395      0.081     15.295      0.000       1.081       1.398
What is your current occupation_Working Professional     3.7956      0.190     19.934      0.000       3.422       4.169
Lead Origin_Lead Add Form                                2.5927      0.186     13.907      0.000       2.227       2.958
```

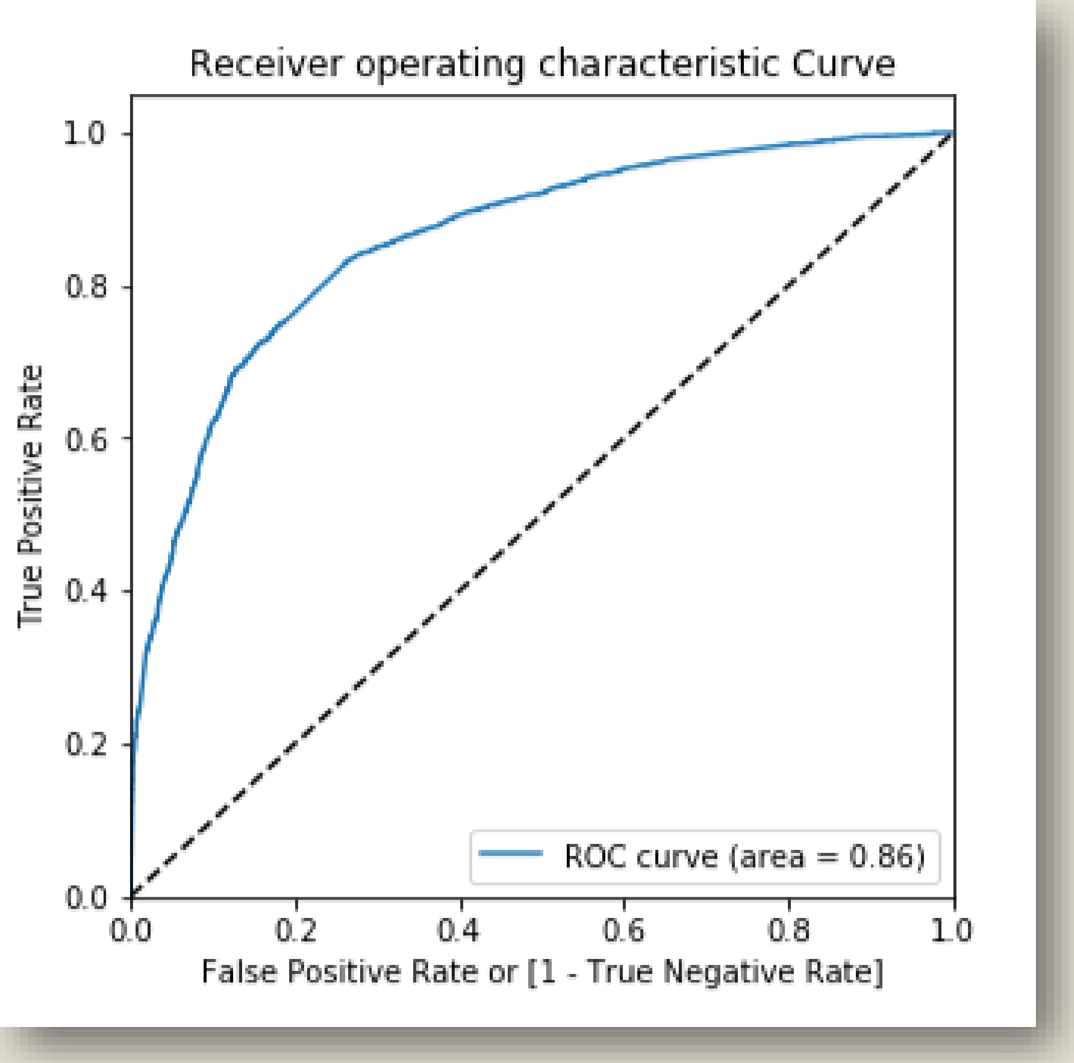


```
                                           Features   VIF
10       What is your current occupation_Unemployed  2.62
3                                Lead Source_Google  1.75
1                        Lead Source_Direct Traffic  1.71
12                        Lead Origin_Lead Add Form  1.60
11  What is your current occupation_Working Profes...  1.31
4                       Lead Source_Organic Search  1.29
6                      Lead Source_Welingak Website  1.24
7                 Specialization_Finance Management  1.15
0                       Total Time Spent on Website  1.09
9           What is your current occupation_Student  1.05
5                        Lead Source_Referral Sites  1.02
8             Specialization_Hospitality Management  1.02
2                              Lead Source_Facebook  1.01
```

Evaluating the model

I. After building the final model making prediction on it(on trainset), we created ROC curve to find the model stability with auc score(area under the curve) As we can see from the graph plotted on the right side, the area score is 0.86 which is a great score.

II. And our graph is leaned towards the left side of the border which means we have good accuracy.

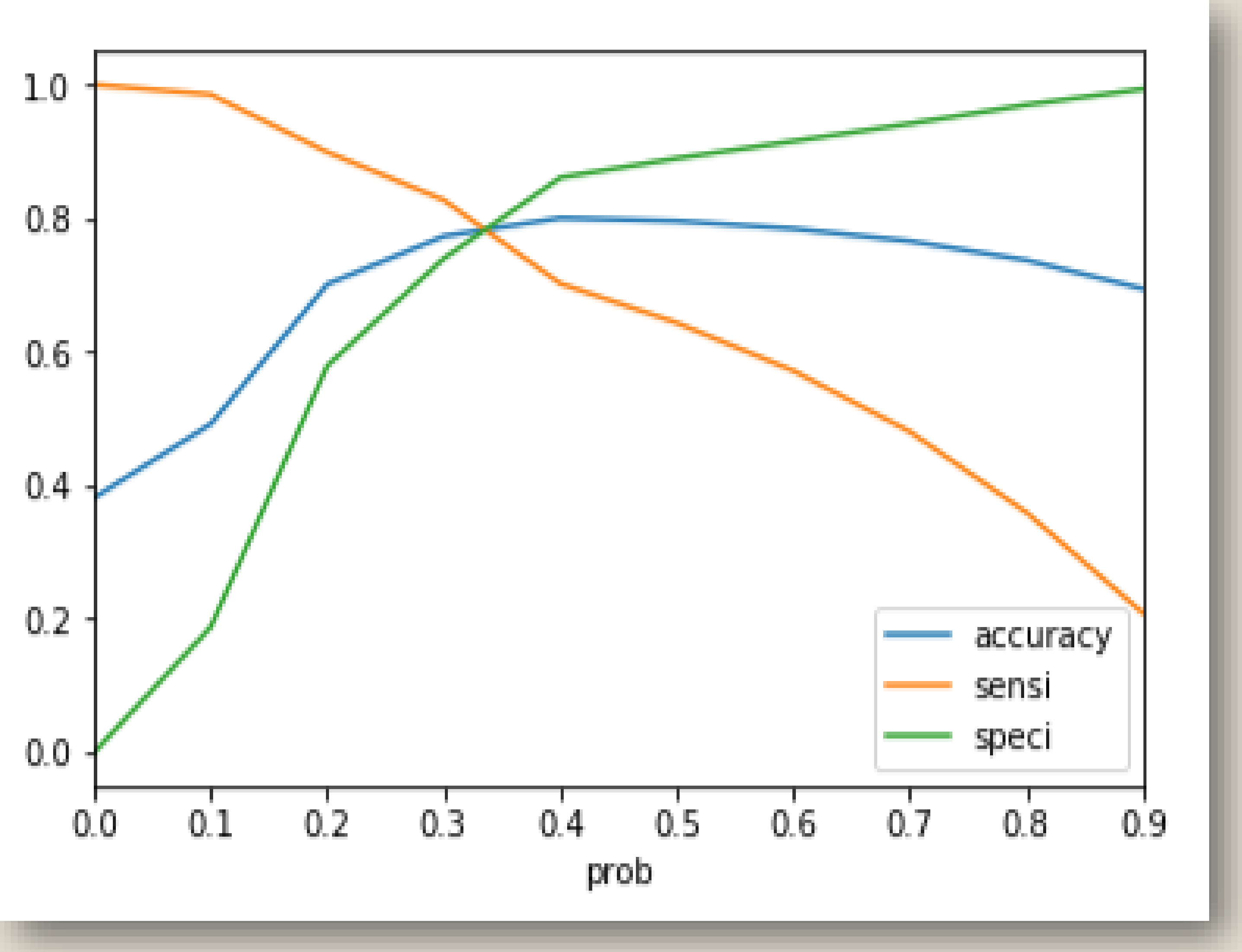


Finding the optimal cutoff point

I. Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cutoff.

II. We found that on 0.3 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected.

III. To verify our answer we plotted this in a graph line plot which is on the right side and we stand corrected that the meeting point is close to 0.3 and hence we choose 0.3 as our optimal probability cutoff.



Precision and Recall

I. We used this cutoff point to create a new column in our final dataset for predicting the outcomes.

II. After this we did another type of evaluation which is by checking Precision and Recall.

III. As we all know, Precision and Recall plays very important role in build our model more business oriented and it also tells how our model behaves.

IV. Hence, we evaluated the precision and recall for this model and found the score as 0.66 for precision and 0.82 for recall.

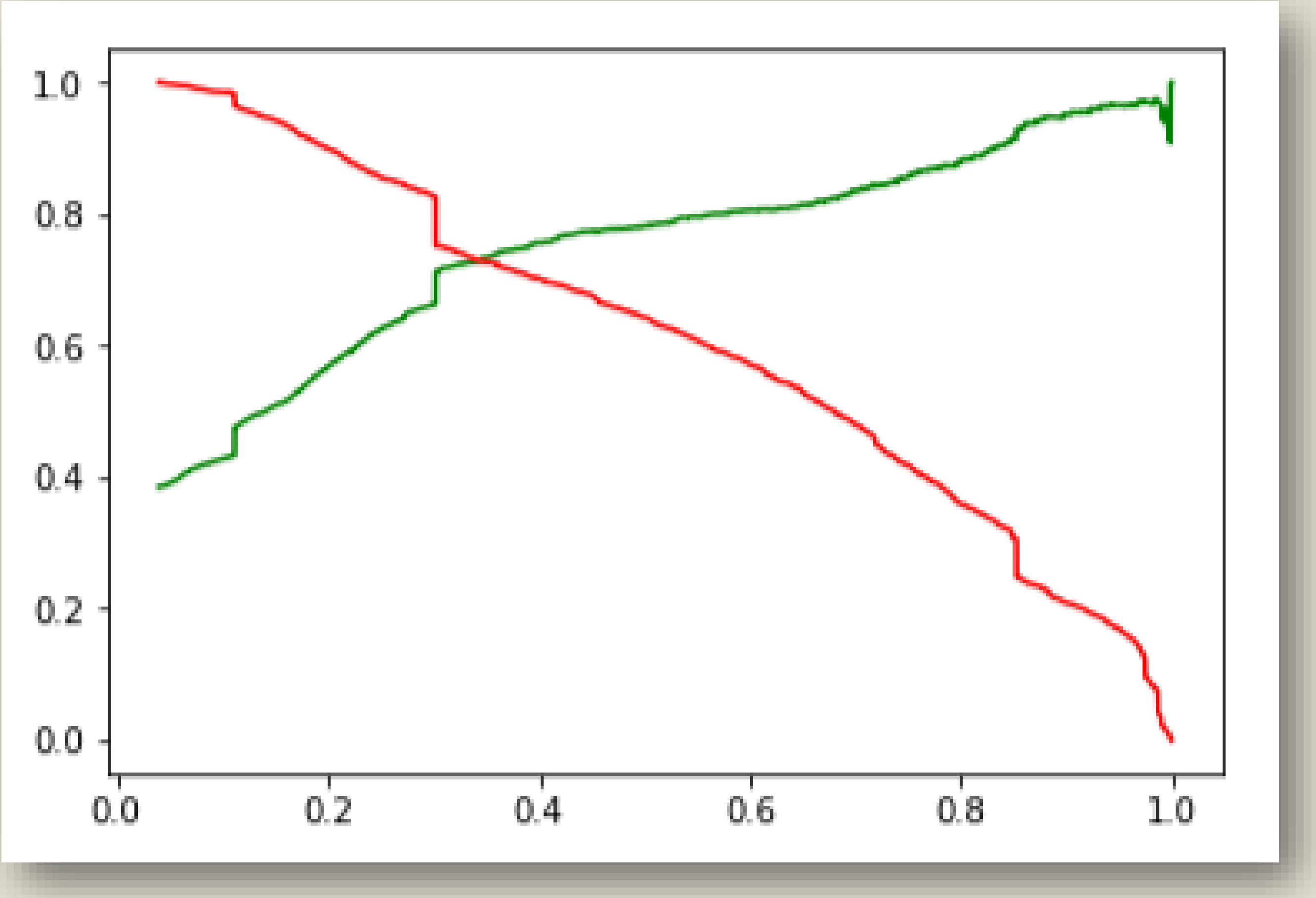
V. Now, recall our business objective the recall percentage I will consider more valuable because it is okay if our precision is little low which means less hot lead customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision.

VI. We get more relevant results as many as hot lead customers from our model .

Precision and Recall

I. We created a Precision and Recall graph to ensure optimum point to take it as a cut-off probability.

II. We found that Precision and Recall meeting point is approximately at 0.3.



Prediction on test set

I. Before predicting on test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.

II. After doing the above step, we started predicting the test set and the new predictions values were saved in new dataframe.

III. After this we did model evaluation i.e. finding the accuracy, precision and recall.

IV. The accuracy score we found was 0.77, precision 0.66 and recall 0.83 approximately.

V. This shows that our test prediction is having accuracy , precision and recall score in an acceptable range.

VI. This also shows that our model is stable with good accuracy and recall/sensitivity.

VII. Lead score is created on test dataset to identify hot leads high the lead score higher the chance of converted, low the lead score lower the chance of getting converted.

Conclusion

Valuable Insights:

I. The Accuracy, Precision and Recall/Sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps. Means the recall is having high score value than precision which is acceptable for business needs.

II. In business terms, this model has an ability to adjust with the company's requirements in coming future.

III. This concludes that the model is in stable state. Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted and X Education Company needs to focus on following key aspects to improve the overall conversion rate:

-Increase user engagement on Welingak website.

-Approach to Working Professionals who wants to gain extra knowledge, since this helps in higher conversion.

-Focus on lead form on the website.

-Add quality contents on website so that user will spend more time exploring information about courses.

THANK YOU!