Lead Scoring Case Study

- Hardikkumar Babulal Panchal
- Ankita Patel

Problem Statement:

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- X Education gets a lot of leads through various resources however its lead conversion rate is very poor which is about 30% only

Goal:

- The CEO of the company gave the target lead conversation reaches to around 80%
- Hence, we need to identify leads which has higher chances to get converted and need to create the finest predictive logistic regression model for the same

Steps Followed – Before Analysis

1. Imported important libraries and Dataset
2. Read and understood the data - Through shape, describe, info

Clean and prepare the data with

1. Treatment on Null values
2. Dropped columns with missing values more than 45%
3. Imputed missing values with appropriate methods for Categorical Columns
4. Treated outliers with Median after Box plot analysis for Numerical variables

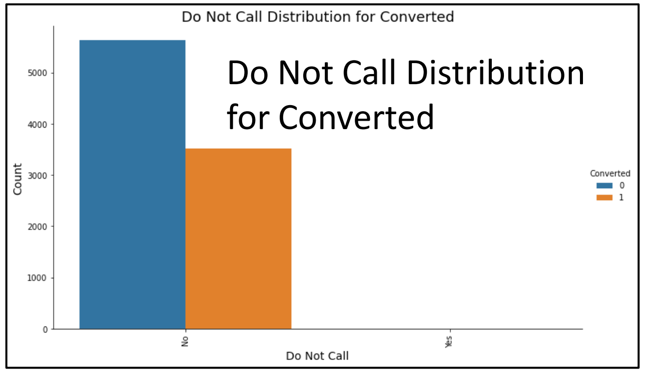
Steps followed - Analysis

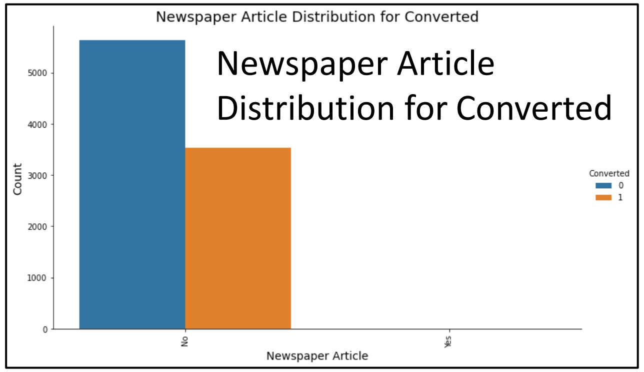
1. Performed EDA to fetch a greater understanding of Categorical and Numerical variables
2. Then started with Model Building where first created Dummy variables for categorical variables
3. Data split for Test and Train model
4. Used RFE to reduce the variables to 15 and started working on those selected variable to build model

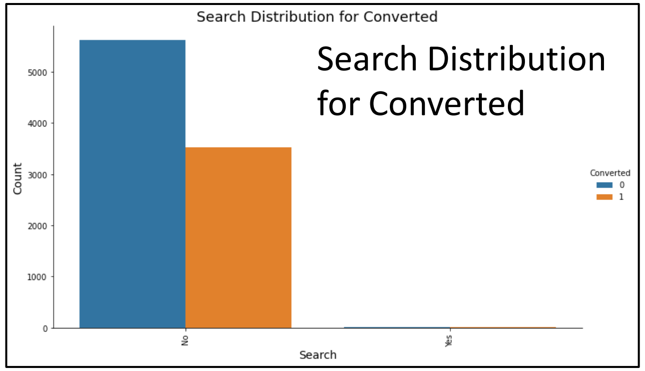
Model Building

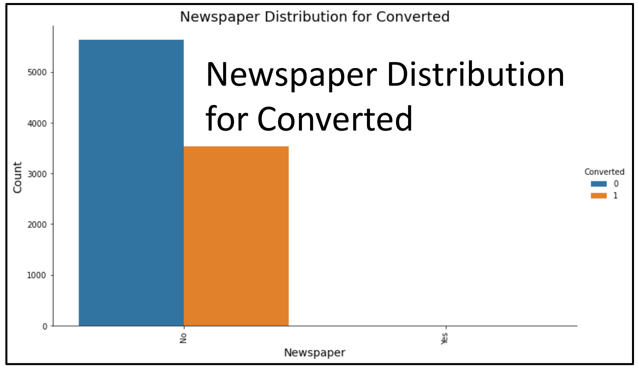
1. For Model building we used ‘sklearn’ and ‘statsmodels’ both modules
2. We divided data in 70% training set and 30% in test set
3. Created dummy variables for all categorical columns
4. We used MinMax Scaler for scaling of numerical columns
5. We started with automatic feature selection RFE with 15 features
6. And with those selected features we started building model with statsmodel
7. And based on p-value and VIFs checks we finalized one model which had 11 features with all having less than 0.05 p-value and VIFs less than 5
8. After this we evaluated model with both Accuracy, Sensitivity & Specificity which was ~91% on both training and test sets
9. With this trade-method optimum cut-off arrived is 0.3
10. In Precision-Recall trade-off also we achieved very good values of Precision, Recall and F1-Score and all values are ~90%, and optimum cut-off arrive 0.4.
11. ROC Curve value also we got 0.97 which indicates excellent predictive value.

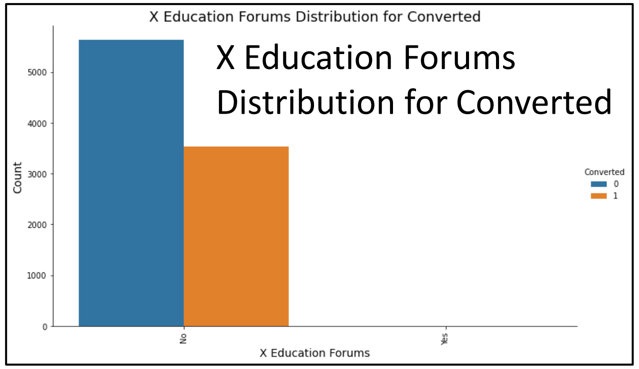
EDA – Categorical Variables

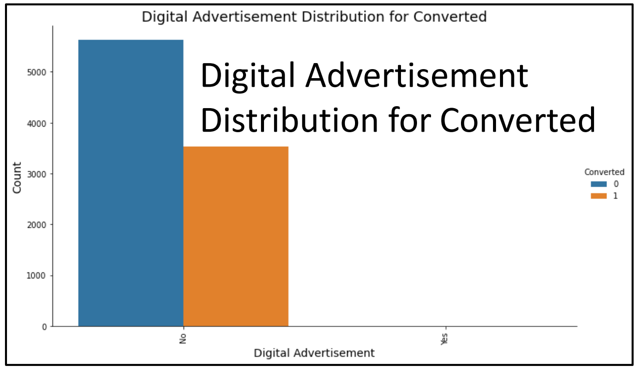

Do Not Call Distribution for Converted

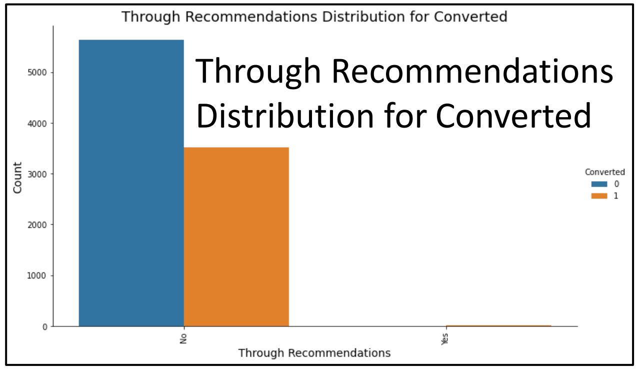

Newspaper Article Distribution for Converted


Search Distribution for Converted


Newspaper Distribution for Converted

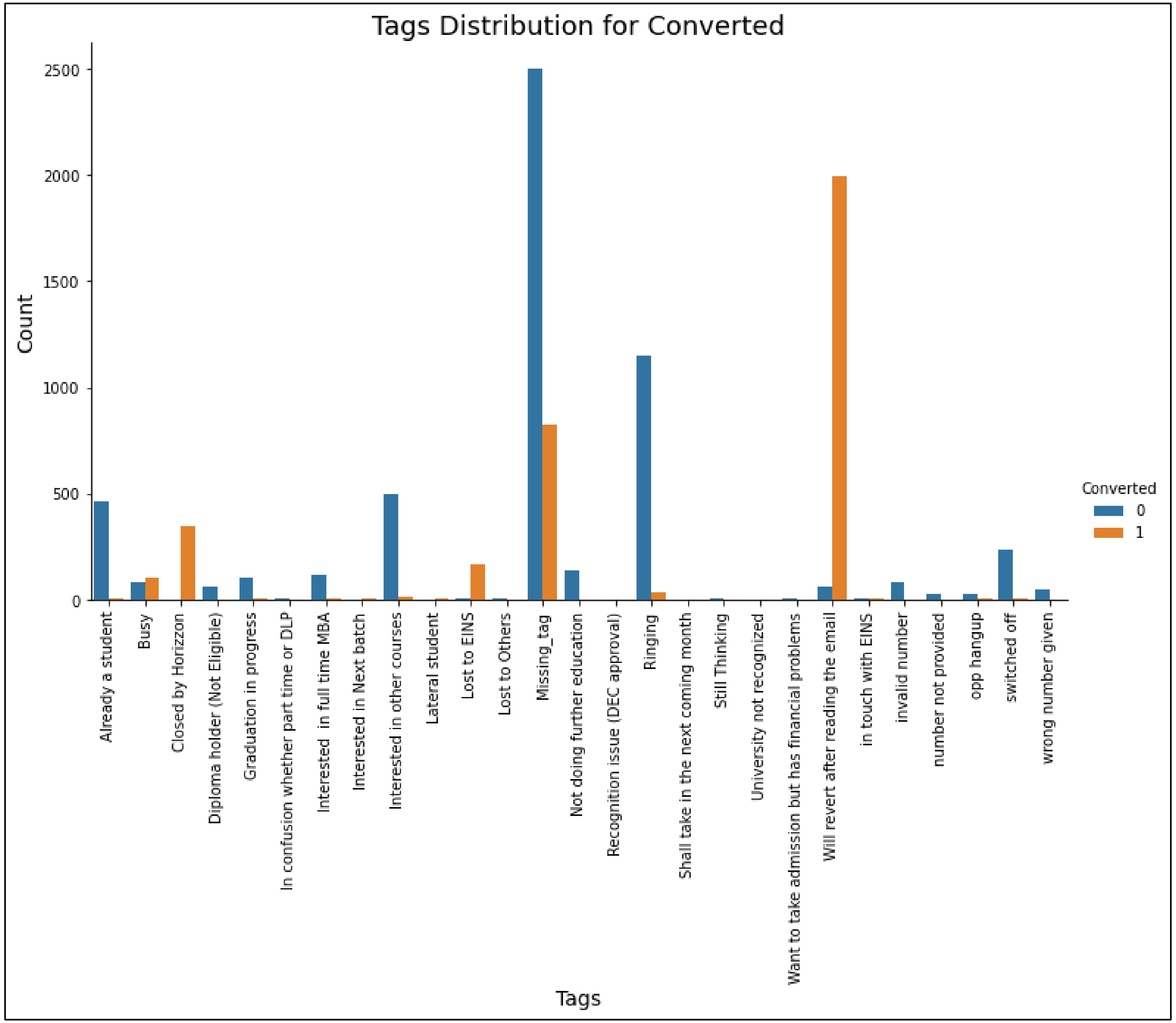

X Education Forums Distribution for Converted


Digital Advertisement Distribution for Converted


Through Recommendations Distribution for Converted

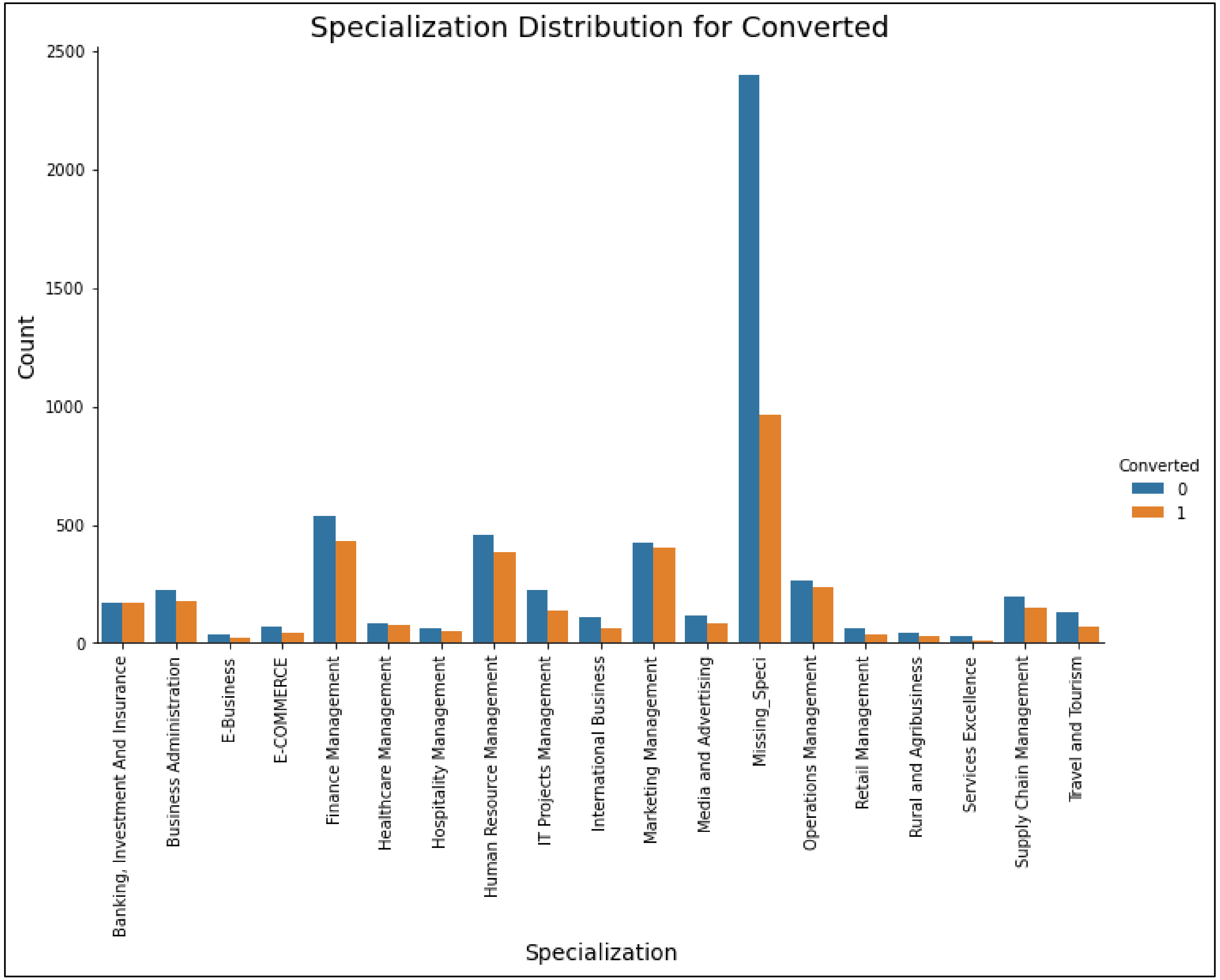
- From these plots, we can see that below columns have only 1 category which is contributing maximum values of columns.
- 'Do Not Call', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement’, 'Through Recommendations', 'Search’
- Hence, dropped all these columns as it won't be useful for analysis

EDA – Categorical Variables



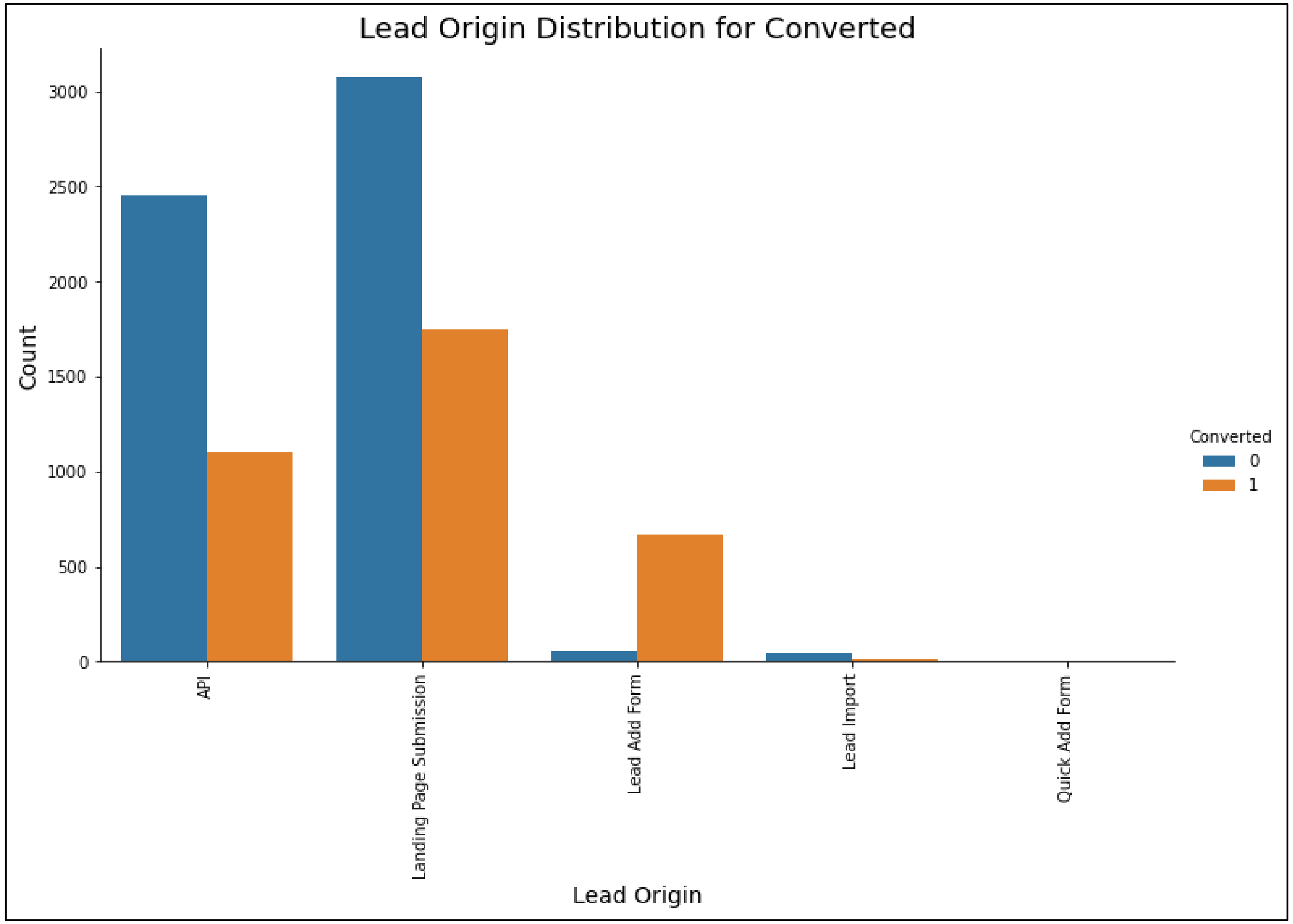
- Will Revert After Reading Email & Closed by Horizzon have more conversion numbers.
- And Ringing, Interested in other course have very few conversions.

EDA – Categorical Variables



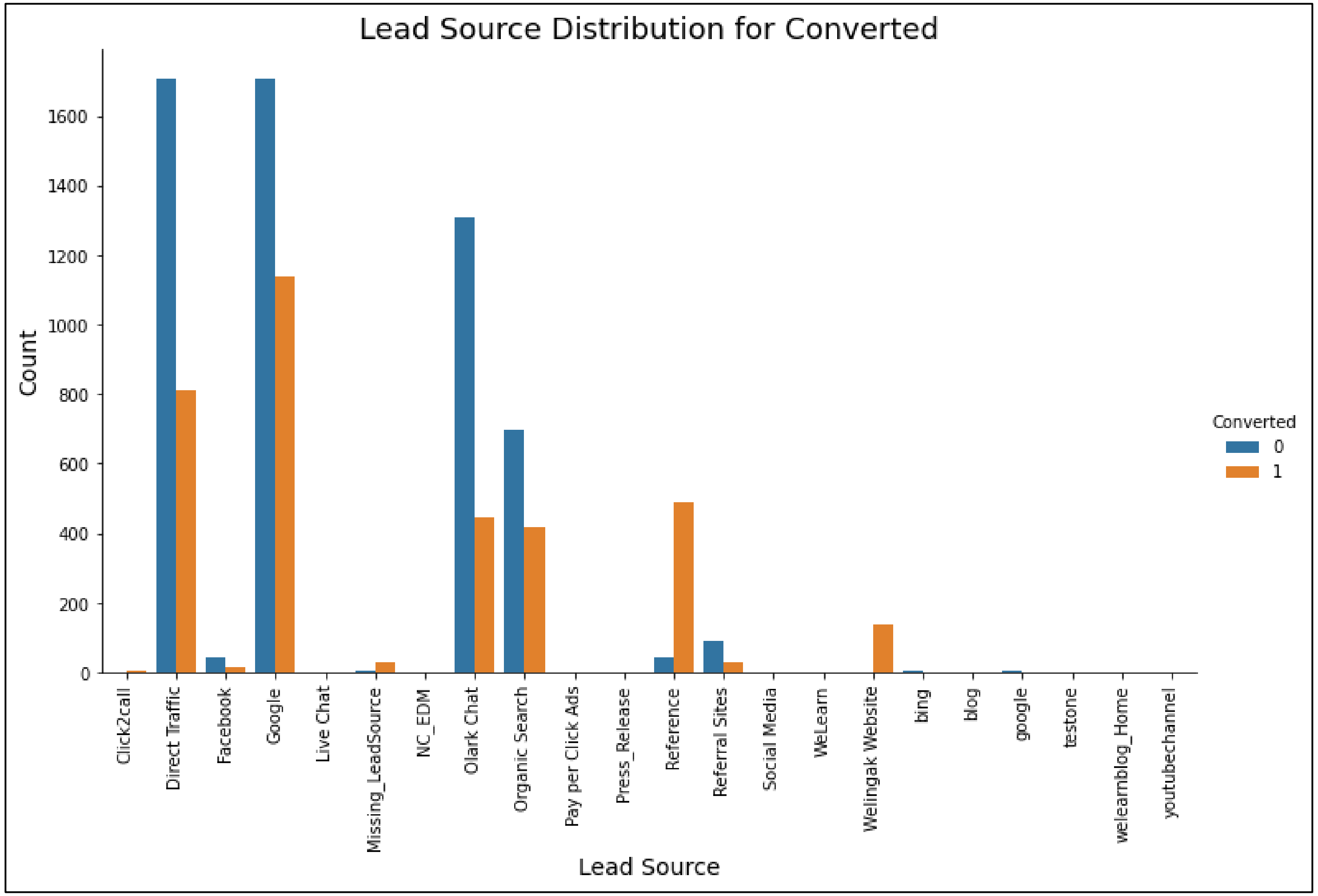
- We can see that There are enough numbers available for all categories.
- We can see that all categories have significant conversions present.

EDA – Categorical Variables



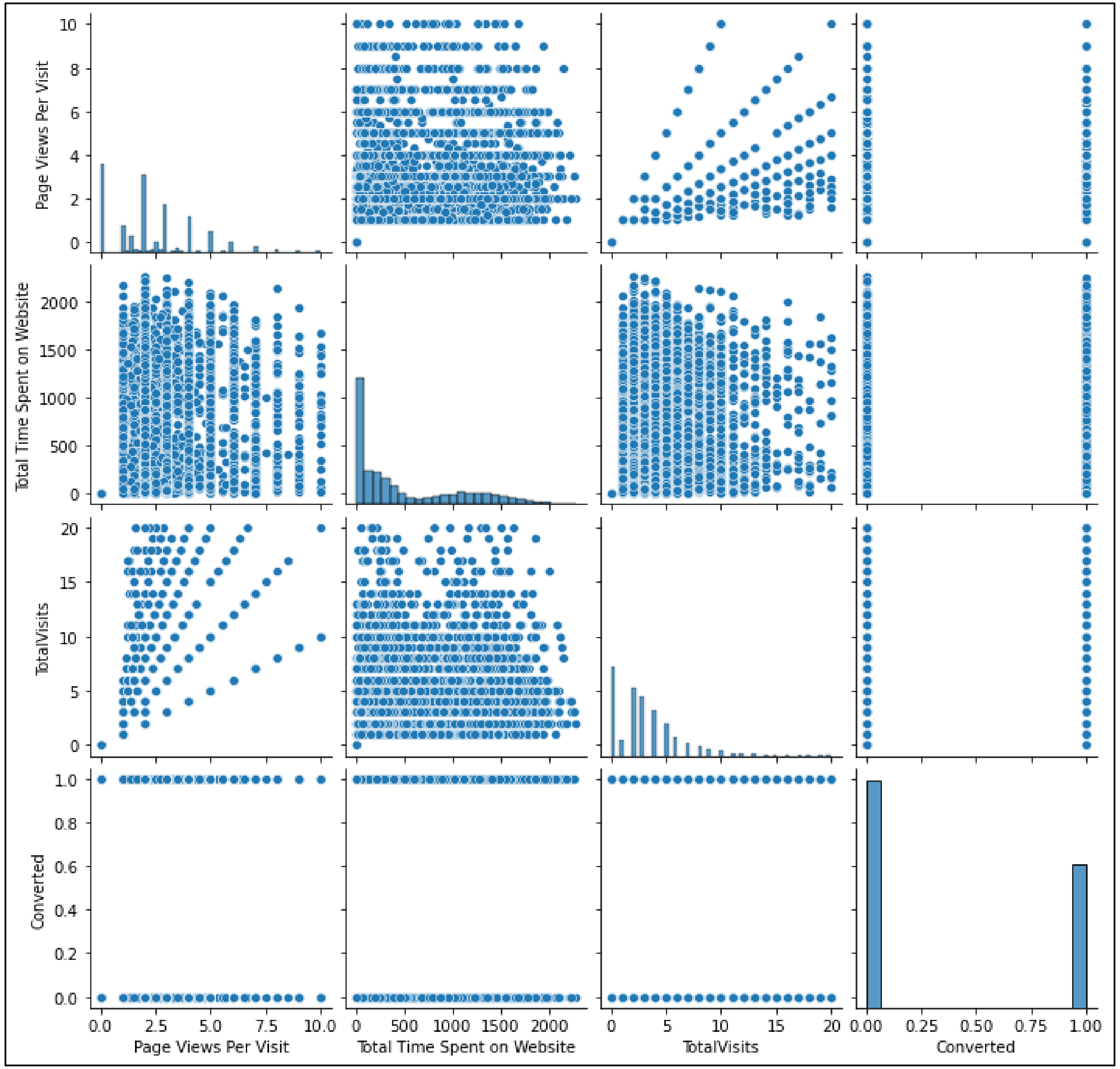
- Lead Origin from API and Landing page submission has high number of not conversion than conversion.
- Overall, they both have higher conversion leads from other Lead Origin categories.

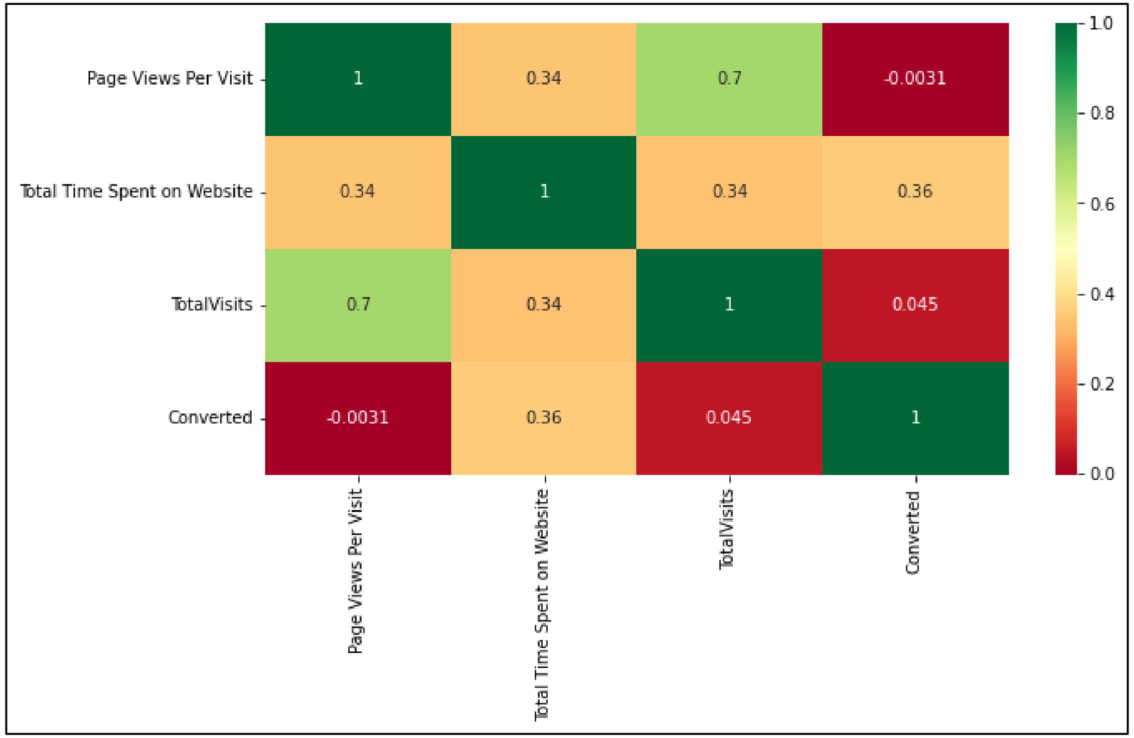
EDA – Categorical Variables



- There are good number of conversion for Google, Reference, Direct Traffic, Olark Chat & Organic Search.
- We can see that There are enough numbers available for all categories.

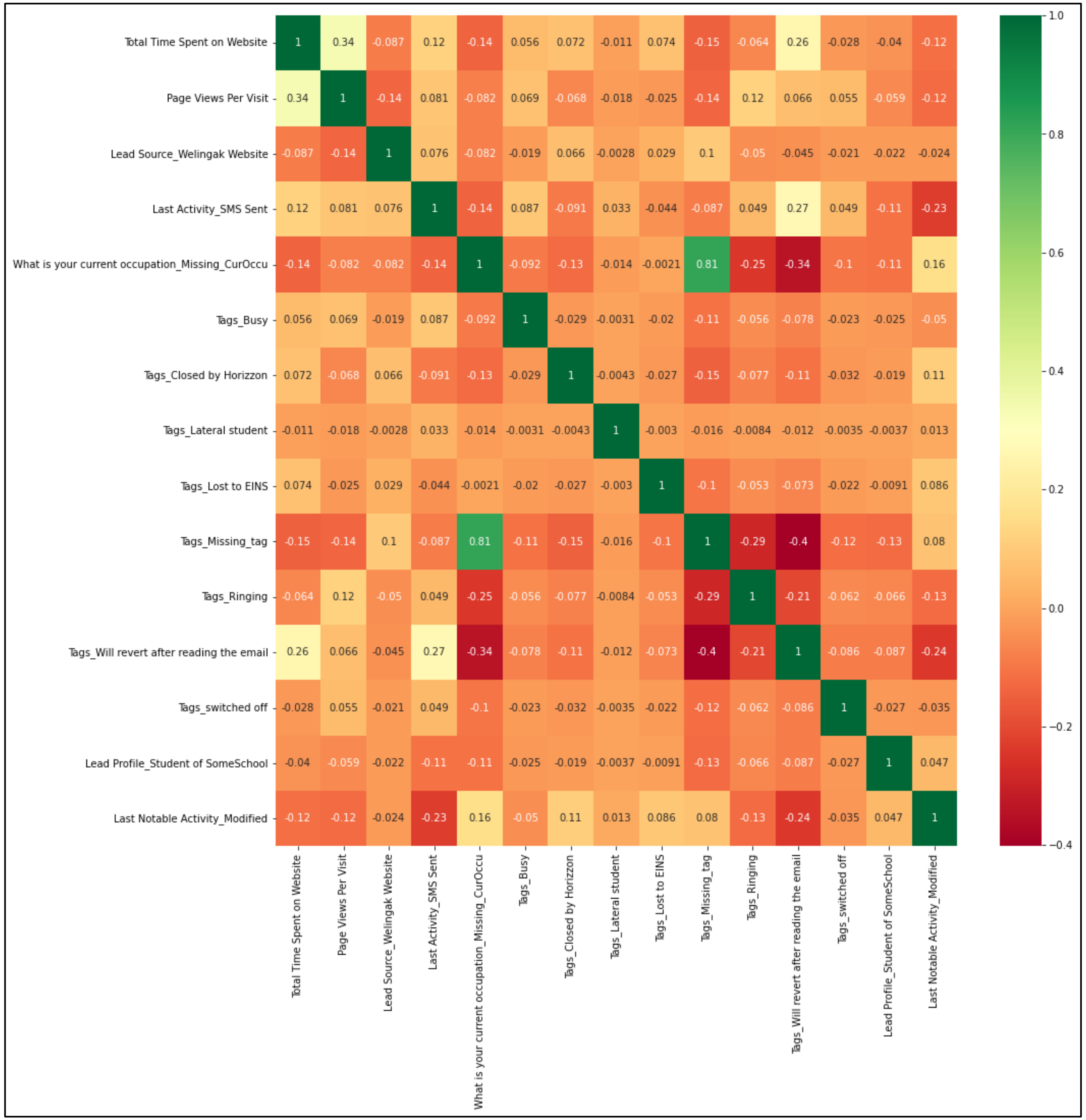
EDA – Numerical Variables





- From pair plots and Heat Map we don’t see any correlations between these Numerical Variables.

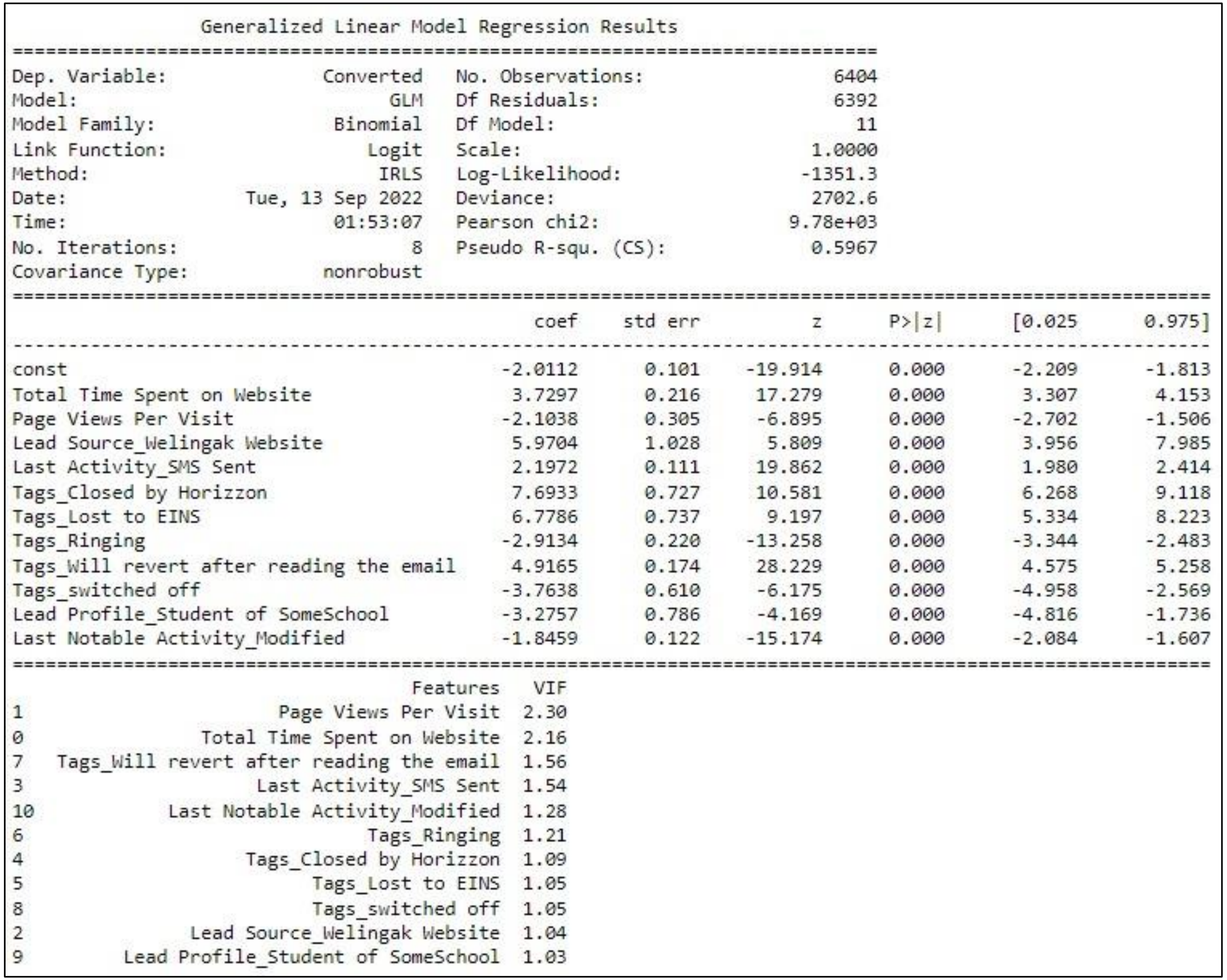
Correlation between selected feature by RFE



	Total Time Spent on Website	Page Views Per Visit	Lead Source_Welingak Website	Last Activity_SMS Sent	What is your current occupation_Missing_CurOccu	Tags_Busy	Tags_Closed by Horizzon	Tags_Lateral student	Tags_Lost to EINS	Tags_Missing_tag	Tags_Ringing	Tags_Will revert after reading the email	Tags_switched off	Lead Profile_Student of SomeSchool	Last Notable Activity_Modified
Total Time Spent on Website	1	0.34	-0.087	0.12	-0.14	0.056	0.072	-0.011	0.074	-0.15	-0.064	0.26	-0.028	-0.04	-0.12
Page Views Per Visit	0.34	1	-0.14	0.081	-0.082	0.069	-0.068	-0.018	-0.025	-0.14	0.12	0.066	0.055	-0.059	-0.12
Lead Source_Welingak Website	-0.087	-0.14	1	0.076	-0.082	-0.019	0.066	-0.0028	0.029	0.1	-0.05	-0.045	-0.021	-0.022	-0.024
Last Activity_SMS Sent	0.12	0.081	0.076	1	-0.14	0.087	-0.091	0.033	-0.044	-0.087	0.049	0.27	0.049	-0.11	-0.23
What is your current occupation_Missing_CurOccu	-0.14	-0.082	-0.082	-0.14	1	-0.092	-0.13	-0.014	-0.0021	0.81	-0.25	-0.34	-0.1	-0.11	0.16
Tags_Busy	0.056	0.069	-0.019	0.087	-0.092	1	-0.029	-0.0031	-0.02	-0.11	-0.056	-0.078	-0.023	-0.025	-0.05
Tags_Closed by Horizzon	0.072	-0.068	0.066	-0.091	-0.13	-0.029	1	-0.0043	-0.027	-0.15	-0.077	-0.11	-0.032	-0.019	0.11
Tags_Lateral student	-0.011	-0.018	-0.0028	0.033	-0.014	-0.0031	-0.0043	1	-0.003	-0.016	-0.0084	-0.012	-0.0035	-0.0037	0.013
Tags_Lost to EINS	0.074	-0.025	0.029	-0.044	-0.0021	-0.02	-0.027	-0.003	1	-0.1	-0.053	-0.073	-0.022	-0.0091	0.086
Tags_Missing_tag	-0.15	-0.14	0.1	-0.087	0.81	-0.11	-0.15	-0.016	-0.1	1	-0.29	-0.4	-0.12	-0.13	0.08
Tags_Ringing	-0.064	0.12	-0.05	0.049	-0.25	-0.056	-0.077	-0.0084	-0.053	-0.29	1	-0.21	-0.062	-0.066	-0.13
Tags_Will revert after reading the email	0.26	0.066	-0.045	0.27	-0.34	-0.078	-0.11	-0.012	-0.073	-0.4	-0.21	1	-0.086	-0.087	-0.24
Tags_switched off	-0.028	0.055	-0.021	0.049	-0.1	-0.023	-0.032	-0.0035	-0.022	-0.12	-0.062	-0.086	1	-0.027	-0.035
Lead Profile_Student of SomeSchool	-0.04	-0.059	-0.022	-0.11	-0.11	-0.025	-0.019	-0.0037	-0.0091	-0.13	-0.066	-0.087	-0.027	1	0.047
Last Notable Activity_Modified	-0.12	-0.12	-0.024	-0.23	0.16	-0.05	0.11	0.013	0.086	0.08	-0.13	-0.24	-0.035	0.047	1

- We see that most of the features selected by RFE has not much Multicollinearity.
- There is significant correlation between 'What is your current occupation_Missing_CurOccu' & 'Tags_Missing_tag'. We took care of this while checking VIFs.

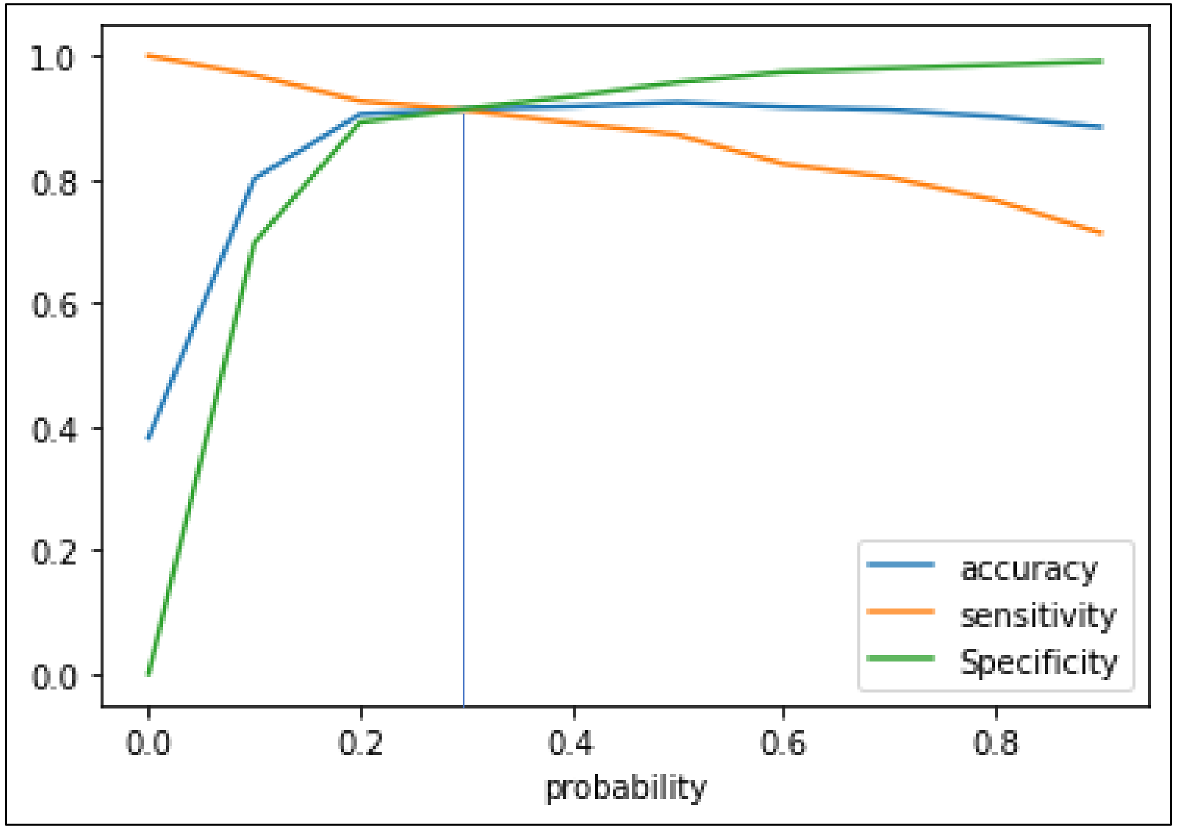
Finalized Model



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6404
Model:                            GLM   Df Residuals:                     6392
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1351.3
Date:                Tue, 13 Sep 2022   Deviance:                       2702.6
Time:                        01:53:07   Pearson chi2:                 9.78e+03
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5967
Covariance Type:            nonrobust
============================================================================================================
                                               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                       -2.0112      0.101    -19.914      0.000      -2.209      -1.813
Total Time Spent on Website                  3.7297      0.216     17.279      0.000       3.307       4.153
Page Views Per Visit                        -2.1038      0.305     -6.895      0.000      -2.702      -1.506
Lead Source_Welingak Website                 5.9704      1.028      5.809      0.000       3.956       7.985
Last Activity_SMS Sent                       2.1972      0.111     19.862      0.000       1.980       2.414
Tags_Closed by Horizzon                      7.6933      0.727     10.581      0.000       6.268       9.118
Tags_Lost to EINS                            6.7786      0.737      9.197      0.000       5.334       8.223
Tags_Ringing                                -2.9134      0.220    -13.258      0.000      -3.344      -2.483
Tags_Will revert after reading the email     4.9165      0.174     28.229      0.000       4.575       5.258
Tags_switched off                           -3.7638      0.610     -6.175      0.000      -4.958      -2.569
Lead Profile_Student of SomeSchool          -3.2757      0.786     -4.169      0.000      -4.816      -1.736
Last Notable Activity_Modified              -1.8459      0.122    -15.174      0.000      -2.084      -1.607
============================================================================================================
                                    Features   VIF
1                       Page Views Per Visit  2.30
0                Total Time Spent on Website  2.16
7   Tags_Will revert after reading the email  1.56
3                     Last Activity_SMS Sent  1.54
10           Last Notable Activity_Modified  1.28
6                               Tags_Ringing  1.21
4                    Tags_Closed by Horizzon  1.09
5                          Tags_Lost to EINS  1.05
8                          Tags_switched off  1.05
2               Lead Source_Welingak Website  1.04
9         Lead Profile_Student of SomeSchool  1.03
```

- We can see that p-value for all features are less than 0.05. And VIFs for all selected feature also below 5.
- Model Stats like likelihood is also good hence we can make this model as final.

Optimal and Final Cut off – Train Dataset

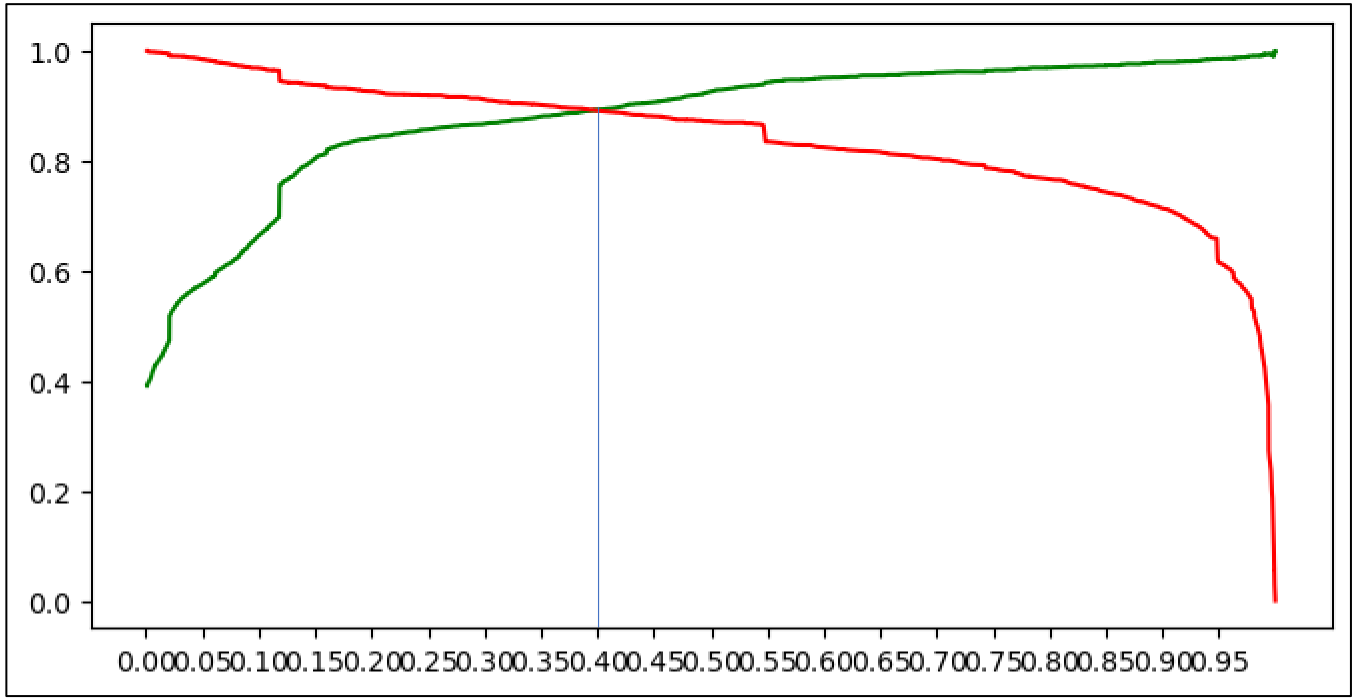


- We see from the plot that, optimal cut off at 0.3 where we have optimum value of all Accuracy, Sensitivity & Specificity.
- So, we have selected 0.3 as optimal and Final cut off.
- Confusion Matrix:

  [[3618, 340],

  [215, 2231]]
- Accuracy:- 91.33%,
- Sensitivity:- 91.21%
- Specificity:- 91.40%.

Precision and recall trade-off

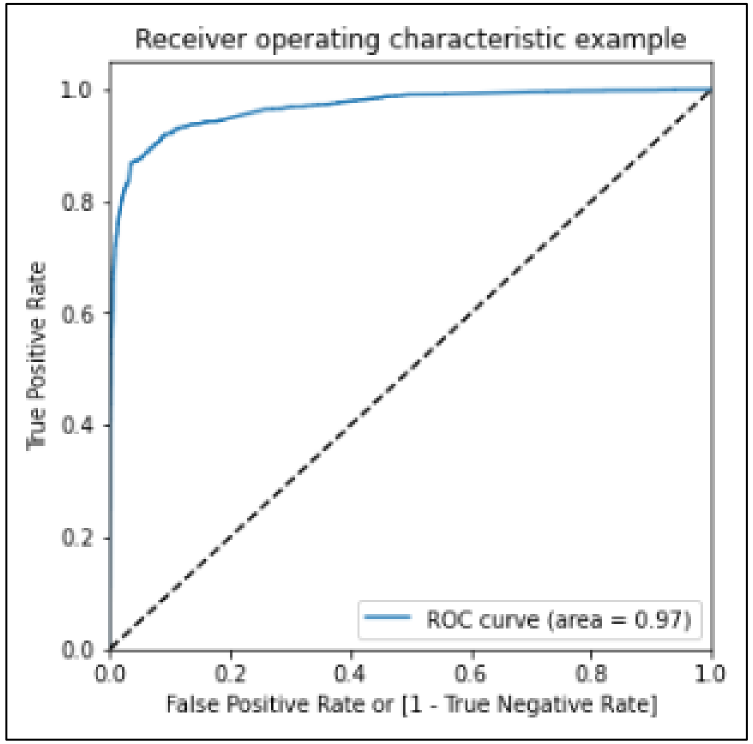


- The plot depicts an optimal cut-off 0.4 based on Precision and recall trade off
- Confusion Matrix:

  [[3698, 260],

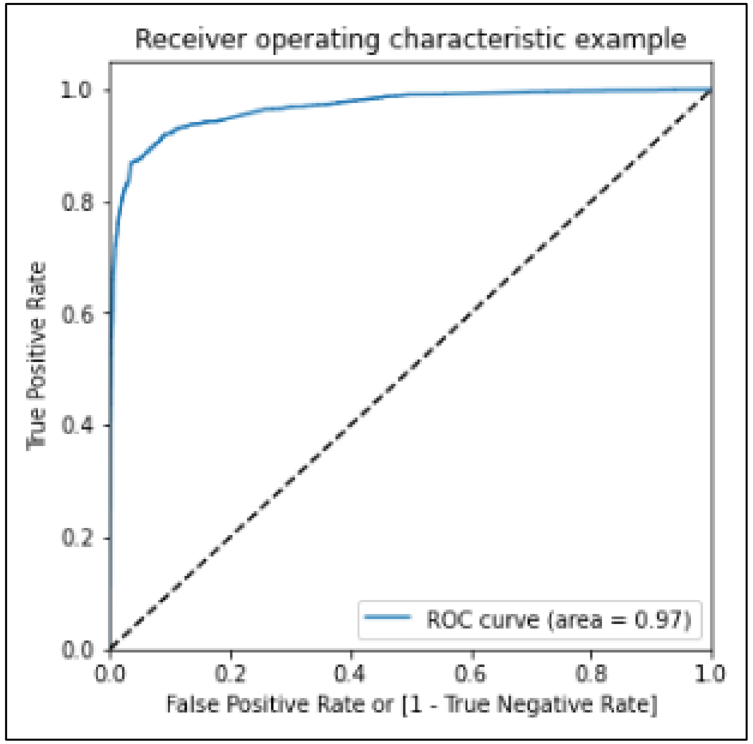
  [265, 2181]]
- Precision:- 89.34%
- Recall:- 89.16%
- F1-Score:- 89.25%

ROC with Predicted Values for both train and test data



Train Data set

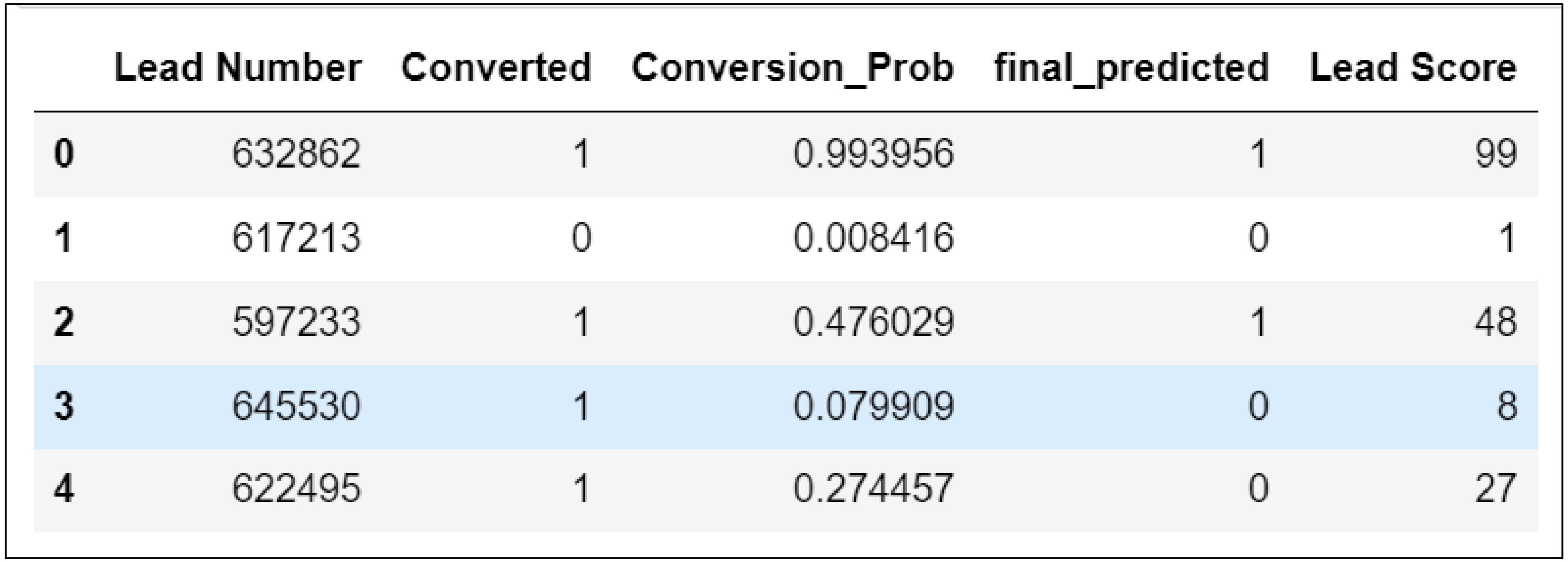
- ROC Curve should be close to 1. And we are getting a value of 0.97, indicating very good predicting value.
- Both the datasets Test and Train have the same ROC as 0.97 which is excellent.



Test Data set

Lead Score Against Lead Number for Original Dataset

Top 5 rows of data frame which is created with Lead Number, Predicted Value and their Lead Score.
For this final predicted values and Lead score calculated with 0.3 cut-off which derived from Accuracy, Sensitivity and Specificity method which looks better compare to other cut-off derived from Precision-Recall trade-off.



	Lead Number	Converted	Conversion_Prob	final_predicted	Lead Score
0	632862	1	0.993956	1	99
1	617213	0	0.008416	0	1
2	597233	1	0.476029	1	48
3	645530	1	0.079909	0	8
4	622495	1	0.274457	0	27

The entire dataset with 'Lead Number', Actual 'Converted', Predicted value through model and Lead Score

Summary of Analysis

- After this entire exercise we have made a Final Logistic Regression Model with below statistics. Cut-off calculated with both Trade-off methods

Training Dataset:

- **With 0.3 Cut-off**

Accuracy - 91.33%

Sensitivity - 91.21%

Specificity - 91.40%

- **With 0.4 cut-off**

Precision = 89.34%

Recall = 89.16%

F1-Score = 89.25%

Test Dataset:

- **With 0.3 Cut-off**

Accuracy - 91.36%

Sensitivity - 91.09%

Specificity - 91.54%

- **With 0.4 cut-off**

Precision = 90.51%

Recall = 89.42%

F1-Score = 89.96%

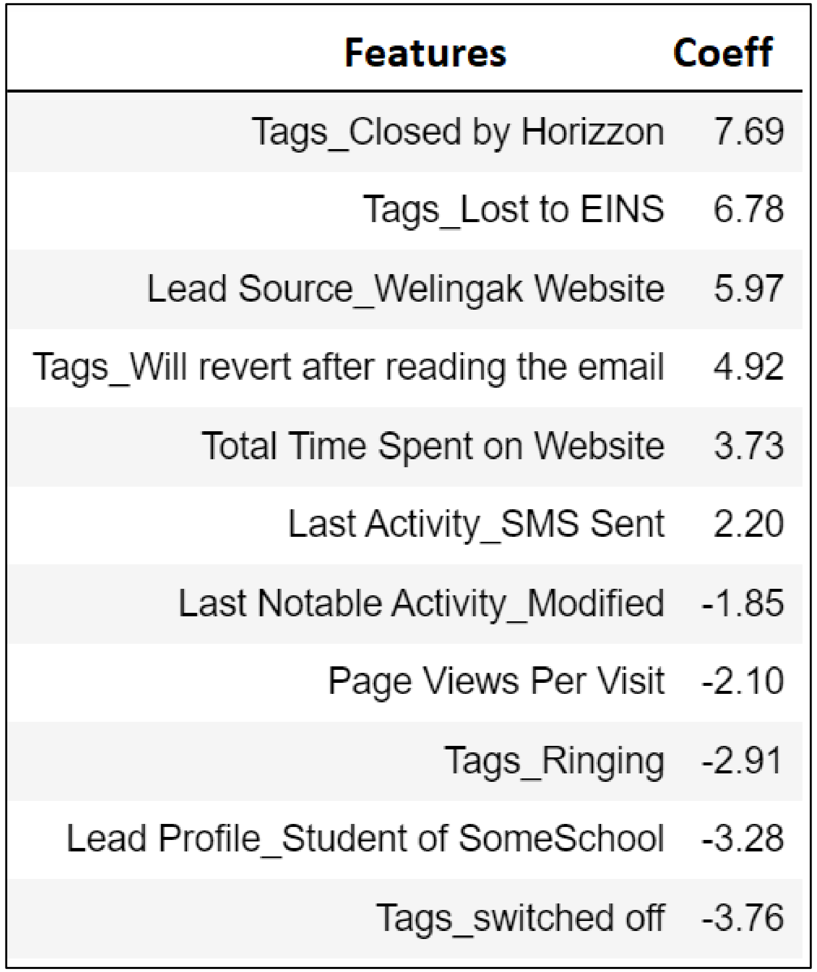
ROC Curve also we are getting 0.97 for both training and test dataset, which is close to 1, which is good predictive value.

This very good values of these metrics suggests that our model will predict all leads in better way.

Most Affecting Features

In Descending order of their Co-efficient:

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Lead Source_Welingak Website
- Tags_Will revert after reading the email
- Total Time Spent on Website
- Last Activity_sent SMS
- Last Notable Activity_Modified
- Page Views Per Visit
- Tags_Ringing
- Lead Profile_Student of SomeSchool
- Tags_switched off



Features	Coeff
Tags_Closed by Horizzon	7.69
Tags_Lost to EINS	6.78
Lead Source_Welingak Website	5.97
Tags_Will revert after reading the email	4.92
Total Time Spent on Website	3.73
Last Activity_SMS Sent	2.20
Last Notable Activity_Modified	-1.85
Page Views Per Visit	-2.10
Tags_Ringing	-2.91
Lead Profile_Student of SomeSchool	-3.28
Tags_switched off	-3.76

These features are affecting most in conversion of leads, company should focus on top 3-4 above feature more diligently so more conversion leads can be converted and revenue of company can be increased.

Conclusion

- With this Model we can predict Conversion Rate in a better way and X Education company can utilize this and from Lead Score they can identify potential buyers and can arrange call back and convince them to buy their course.
- This way X Education company can improve their Lead Conversion Rate and can achieve provided target of 80% Conversion Rate through this model.

Thank You