SUMMARY REPORT

An education company 'X Education' sells online courses to the professionals. We need to a logistic model to select the leads that are most likely to become the paying customers. The present conversion rate of the customers is about 30%. The company's CEO wants us to increase the lead conversion rate to be around 80%.

1. Read and understand the data:

We have a dataset with around 9200 data points with various attributes. The attributes are of integer, float and string data types. Some of these attributes have null values in them. The target variable in the dataset is 'Converted'.

2. Data cleaning:

The unwanted variables for the analysis and the variables that are highly imbalanced are dropped. There are some variables that have null values in them. The variables with null percentage greater than 35% are removed. The variables Lead Profile, How did you hear about X Education and Specialization are having a category 'Select' in them. The null values and Select category in these variables are replaced with 'Not Mentioned'. The null values in the other columns are imputed with their mode values.

3. Eda:

The numerical and categorical variables are analyzed with the target variable. On this analysis, we found the variables that contribute to the lead conversion. Also the numerical variables had outliers in them which are handled well.

4. Data preparation:

The binary variables with Yes/No has been converted to 1/0. Dummy variables for the categorical variables have been created.

5. Train-test split:

The entire dataset is split as training and test data set in the ratio 70:30.

6. Feature scaling:

The variables in the dataset are changed to comparable scale using the Standardization method.

7. Model building:

The top 15 features are selected using the RFE method and model is build by removing the unwanted variables manually by checking the p-values and VIFs. A model with 13 variables whose p-values and VIFs are within range is created. Using this model, the probabilities for the lead to be converted as hot leads is calculated. A predicted value of 0 or 1 is obtained by using the threshold value as 0.5.

A confusion matrix is obtained by using the predicted values. Then the model is evaluated by the metrics Accuracy, Sensitivity, Specificity, Precision and Recall. Then an optimal threshold value was obtained by using the Accuracy, Sensitivity and Specificity graph which is 0.3. On using the threshold 0.3, we obtained a model with a lead conversion rate of about 84%.

8. Model Evaluation:

An optimal threshold value was obtained by using the Accuracy, Sensitivity and Specificity graph which is 0.3. On using the threshold value on the test set, we found that the model created has lead conversion rate of about 84%.

9. Conclusion and recommendations:

The features in our final model are listed with their coefficients as follows.

1. Do Not Email : -1.413358
2. Total Time Spent on Website : 1.094459
3. Lead Origin_Lead Add Form : 3.113291
4. Lead Source_Olark Chat : 1.226329
5. Lead Source_Welingak Website : 3.325618
6. Last Activity_Others : 1.115349
7. Last Activity_SMS Sent : 1.387078
8. What is your current occupation_Unknown : -0.938224
9. What is your current occupation_Working Professional : 2.257155
10. Lead Profile_Lateral Student : 2.693674
11. Lead Profile_Potential Lead : 1.613543
12. Lead Profile_Student of SomeSchool : -2.311967
13. Last Notable Activity_Modified : -0.934799

We can increase the potential leads by concentrating on the above features.

Top 3 important features:

1. Lead Source_Welingak Website
2. Lead Origin_Lead Add Form
3. Lead Profile_Lateral Student

Evaluation Metrics on train data:

- Accuracy - 0.82
- Sensitivity - 0.86
- Specificity - 0.79
- Precision - 0.71
- Recall - 0.86

Evaluation Metrics on test data:

- Accuracy - 0.81
- Sensitivity - 0.84
- Specificity - 0.78
- Precision - 0.71
- Recall - 0.85

Lead Conversion Rate:

- The target lead conversion rate to be around 85% as expected by the CEO of X Education. Therefore we have created a good model.