Summary

This analysis is done for X Education and to find the most promising leads or variables that are most likely to convert. The data provided to us gives lot of information among which we have find the most vital variable to find which will help in elevating the conversion rate. Lead Scoring case study has been done using logistic regression model to meet the constraints as per business requirement.

The following are the steps to be carried out:

1)**Reading and Understanding the data:**

Read and analyze the data find the meaning of the variables

2) Data Cleaning

We drop the variables with high null values in them, and other variables which were either has unique variable or of least importance. Outliers were identified. Furthermore the variables with only one value were also dropped

3) Data Analysis (EDA)

We did univariate and multivariate analysis to understand how the data is oriented and which of the variables are of importance and if any can be dropped

4)Creating Dummy Variables

We went on with creating dummy data for the categorical variables. Step5: Test Train Split: The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

5)Feature Rescaling

We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

6)Feature selection using RFE:

Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

7)Model Evaluation

Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model. We also calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 88% which further solidified the of the model.

8)Finding optimal Cutoff:

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point

The cutoff point was found out to be 0.35

9) Making prediction on Train Set:

Based on the new value we could observe that close to 80% values were rightly predicted by the model. We could also observe the new values of the 'accuracy=81%, 'sensitivity=79.51%', 'specificity=81%'. Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Furthermore calculating Precision and Recall metrics values came out to be 75% and 76% respectively on the train data set. Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.4

10) Making Predictions on Test Set

Then we implemented the learning's to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 79%; Sensitivity=77%; Specificity= 81%.

We can conclude saying that the model is good

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 79%, 77% and 81% which are approximately closer to the respective values calculated using trained set.
- Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.