Lead Scoring Case Study

Summary report

- The objective of this analysis is to identify ways to attract more leads to join X Education courses. The analysis was conducted by examining data related to the customers' website visit, time spent on the site, site source, conversion rate, etc. The analysis steps included data cleaning, exploratory data analysis (EDA), and creation of dummy variables, train-test splitting, model building, model evaluation, prediction, and precision-recall analysis.
- The data was partially clean but had a few null values that were replaced with 'Unknown' to avoid losing data. Irrelevant elements in categorical variables were also identified and removed. Numeric values were found to be good with very few outliers. The split was done at 70% and 30% for train and test data, respectively.
- Model building involved the use of Recursive Feature Elimination (RFE) to obtain the top relevant variables. The variables with VIF < 5 and p-value < 0.05 were then kept by removing the rest manually. A confusion matrix was used to evaluate the model. The optimum cut off value was identified using the ROC curve. Accuracy, sensitivity, and specificity were found to be around 80% each.
- Prediction was done on the test data frame with an optimum cut off of 0.4, resulting in accuracy of 79%, sensitivity of 64%, and specificity of 88%. Specificity-Sensitivity analysis was also conducted, and a cut off of 0.4 was found, resulting in precision around 77% and recall around 64% on the test data frame.
- In conclusion, X Education can focus on the top 10 relevant variables identified in the analysis and work on improving their website to attract more industry professionals. The analysis shows that with an optimum cut off value of 0.4, X Education can achieve an accuracy of 79%, sensitivity of 64%, and specificity of 88%for predicting customers likely to join their courses.
- Feature like **Lead Origin_Lead Add Form, Last Notable Activity_Had a Phone Conversation and Total Time Spent on Website** have high coefficients which means that these variables should be target more as they significantly improve the conversion probability. On the other hand , **Last Notable Activity_Olark Chat Conversation, Last Notable Activity_Email Bounced and Specialization_Unknown** have high negative coefficients which means such leads should be targeted at last as they have very low conversion rate.