Summary Report Lead Scoring Case Study

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.

Data cleaning part consisted of converting all the columns and rows to lowercase for uniformity of data. We also checked for converted and non-converted leads in our dataset to check if it being bias. Later we checked for missing values and dropped the variables having missing values more than 30%. For missing value treatment of few columns which had a value 'Select' was as good as missing values but rather than dropping those variables we categorised a new value in column as select and these variables could play an important role later in model building.

Missing values were treated by categorizing lesser values and missing values together as others to avoid data loss and improper imputation of missing values. Later we differentiated the numerical and categorical values and checked for the variance among their values and low variance were dropped. Outliers were treated with the help of percentiles and boxplot. We checked for converted and non-converted values of categorical variables. We did binary encoding for categorical variables and checked for correlation matrix and dropped highly correlated variables.

For Model, we split the data in train and test with 70:30 ratio and defined x and y variables. We did feature scaling for numerical variables using StandardScaler and fit transform. We build the Logistic Regression model and used RFE and manual feature elimination to eliminate the feature with P-Value more than 0.05 and VIF more than 5. We also check for metrics like Confusion matrix and Accuracy Score. We also checked for Sensitivity, Specificity, false positive rate, positive predictive value and Negative predictive value.

We found the ROC curve, Optimal Cut-off Point, accuracy sensitivity and specificity plot, Precision and recall tradeoff. We checked for Test and Train - Accuracy, Precision and Recall. We found the Probability Cut-Off and Projected Leads for optimum utilization of model in future.