

# Lead Scoring Case Study

## Summary Report

An analysis has been done for X Education company to select the most promising leads joining their online courses coming through various sources. Using the dataset provided, a logistic regression model was built that assigned a score to each of the leads. Leads who scored high are most likely to get converted.

The different steps followed in the analysis are:

### 1. Data cleaning:

- Redundant variables which will not be required for further analysis and model building were dropped.
- Converting 'Select' values to null values.
- Calculated null percentage of the variables
- Dropped variables having null percentage greater than 40%
- Imputed missing values for columns having less than 40% null values with suitable function: mean/median/mode.
- Checked for other error in numerical columns (such as negative sign, special character etc.)

### 2. EDA:

- Checked for data imbalance using our target variable 'Converted'. We saw there was fair distribution of data (68% converted & 32% not-converted)
- Performed Univariate Analysis for both Continuous and Categorical variables.
- Performed Bivariate Analysis with respect to Target variable.

### 3. Data preparation:

- Created dummy variables for categorical variables
- Dropped off original categorical columns after dummies are created

- Concatenated dummy dataframe with original dataframe.
- Train-Test split of dataset is done.
- Rescaling of continuous variables done using StandardScaler
- Feature selection was done using RFE (Recursive Feature Elimination)

#### **4. Model Building:**

- Created models using statsmodel
- Observed the p-values of the feature variables.
- Variables with high p-values were dropped.
- The above 3 steps were repeated until we got feature variables with lower p-values
- VIF scores were checked and variables were dropped having VIF score > 5

#### **5. Predictions on Train data:**

- Derived probabilities of each leads.
- Assigned Predicted score to each leads based on probability cut-off of 0.5
- Created confusion matrix and calculated model accuracy, sensitivity, specificity, positive predicted value, negative predicted value and false positive rate.
- Plotted a ROC curve and found the area under curve to be 0.87
- Finding optimal cut-off point by calculating accuracy, sensitivity, specificity for different cut-offs & plotting a graph to find the optimal cut-off to be 0.36
- Calculated Lead score for each leads.

#### **6. Model Evaluation:**

- Checked model parameters (accuracy[79.78%], sensitivity[79.76%], specificity[79.79%], positive predicted value[71.2%], negative

predicted value[86.29%] and false positive rate[20%]) using confusion matrix after finding optimal cut-off probability.

- Calculated precision and recall using sklearn
- Plotted precision-recall curve

## **7. Prediction on Test data:**

- Scaling of continuous variables using StandardScaler
- Assigned final feature variables to test data (X\_test).
- Added prospect ID as index
- Appending predicted probability and Prospect ID
- Added a final predicted column with scores of 0 or 1 based on optimal cut-off
- Calculated lead scores.
- Created confusion matrix and calculated model accuracy[79.51% ~ 80%], sensitivity[78.76%], specificity[79.93%].