Lead Scoring Case Study – Summary

We build a multi class logistic regression model for X Education, an online education company using python libraries - Pandas, Numpy, Matplotlib.plotly, Seaborn, Statsmodels and sklearn. Data was observed, cleaned and analysed using imputations, deletion, outlier treatments, data imbalance, Univariate and Multivariate analysis. The data initially showed 9240 rows & 37 columns (30 columns were object/string type balance numerical). Converted as Target column. Initial conversion rate at 38%. 5 Columns with no variance , columns with more than 30% missing values or having select as value were deleted. Columns with 2% - 30% missing values were imputed with other / unknown. Rows of variables such as TotalVisits with less than 2% missing values were deleted. Categorical variables such as 'Country' having large number of unique value but having less than 8% contribution were merged as Other. For numerical columns Boxplots were made and outliers were analyzed. After data cleaning splitting, dummy creation and scaling operations were carried out. Train test was done as 70% - 30%. Train set numerical variables were fit and scaled using standardization. Test set numerical variables were only scaled using standardization. Dummy variables were created for categorical columns and one was dropped from each. Correlations were done. Columns with correlations >|0.70| were deleted. Now 17 columns were left. RFE was used to keep 12 columns and rest were dropped. P-value < 0.02 and VIF < 5 was used as cutoff to keep remaining variables thus, 3 more were deleted after a few iterations. Total 9 variables were left for model to be trained upon. Logistic Regression was used to teach the model. Confusion matrix was made on the predicted and actual conversions at cutoff probability of 0.5 initially. Recall, precision and accuracy were calculated for multiple cutoff probability. Final cutoff was set at 0.4 while optimizing the model for Recall. Finally 'Score' column was added.

Model Performance

On Train Set: Recall = 75.4%, Accuracy = 77%,

Precision = 68% and F1 Score = 71.5%.

On Test Set: Recall = 75%, Accuracy = 76%,

Precision = 68% and F1 Score = 71.8%.

Most important Variables for Model:

1) Do Not Email_No
2) Occupation_Unemployed
3) Lead Source_Google
4) Lead Source_Direct Traffic
5) Last Notable Activity_Modified
6) Last Activity_SMS Sent
7) Lead Source_Organic Search
8) Last Activity_Olark Chat Conversation

9) Total Time Spent on Website

Final F1 score was 0.71 with recall of 0.75, precision of 0.68 and accuracy of 0.76.