Summary – Lead Scoring Case Study

As a part of the Lead Scoring case study, X education company wants to find the promising leads. They have provided us the data of customers/students who are visiting their site and how much they are spending time and what are the conversion rate.

The current conversion rate is 30% which is quite low so we have task to analyse the data and we need to create model through which we can make predictions to the order to 80% Lead conversion.

We have followed the below steps and have done the basic analysis of the given set:

Data Understanding and Cleaning -

- ✓ We have treated null values and drop invalid and redundant columns.
- ✓ Dropping the columns which are having null values greater than 30%

EDA -

- ✓ Identified the relevant data columns which can factor in accurate prediction and found the insignificant variables.
- ✓ Identifying the relationship and distribution of column data using graphs
- ✓ Finding and Handling outliers
- ✓ Found correlations using heatmap

Data Pre-processing

- ✓ encoding of the categorical data into dummy variables

Model Building and Tuning

- ✓ Splitting the data into train and test data in ratio of 70:30.
- ✓ Feature scaling done by using StandardScaler function
- ✓ RFE was done to find the top 15 significant variables.
- ✓ Model building using Generalized Linear Model (GLM).

Model Evaluation

- ✓ Evaluation started by creating confusion matrix
- ✓ We have graphed the ROC curve which is helpful in quantifying the model accuracy in case of classification.

Prediction

- ✓ Predictions on test data set and accuracy of the predictions is 88.67% with an optimum cut off as 0.4

Model Interpretation

- ✓ After running model on the test, we have found that accuracy, sensitivity and specificity are above 88%.
- ✓ Precision and Recall was also calculated.
- ✓ We have used the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

To conclude, it is found that the below conditions are mattered the most

1. The total time they are spending on the Website.
2. How much time they are spending on the website?
3. When the lead source was from Google, Direct Traffic, Organic search and Welingak website.
4. When the last activity was from SMS and Olark cha conversation
5. When the lead origin is Lead add format.
6. Working professional are most likely to buy the course.