Lead Scoring Case Study Summary

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 Interpretation:

- ✓ Accuracy at threshold 0.4 is highest.
- ✓ A common way to visualize the trade-offs of different thresholds is by using an ROC curve, a plot of the true positive rate (true positives/ total positives) versus the false positive rate (false positives /total negatives) for all possible choices of thresholds.
- ✓ A model with good classification accuracy should have significantly more true positives than false positives at all thresholds.
- ✓ The optimum position for roc curve is towards the top left corner where the specificity and sensitivity are at optimum levels.

Area Under the Curve (AUC)

- ✓ The area under the ROC curve quantifies model classification accuracy, the higher the area, the greater the disparity between true and false positives, and the stronger the model in classifying members of the training dataset.

- ✓ An area of 0.5 corresponds to a model that performs no better than random classification and a good classifier stays as far away from that as possible. An area of 1 is ideal.

- ✓ The closer the AUC to 1 the better.

Recommendation:

- ✓ It was found that the variables that mattered the most in the potential leads are (In descending order):

- ✓ The total time spend on the Website.

- ✓ Total number of visits.

- ✓ When the lead source was:

    - Google
    - Direct traffic
    - Organic search
    - Welingak website

- ✓ When the last activity was:
    - SMS

- ✓ Olark chat conversation

- ✓ When the lead origin is Lead add format.

- ✓ When their current occupation is as a working professional.

- ✓ Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

- ✓ All variables have a good value of VIF. So, we need not drop any more variables and we can proceed with making predictions using this model only.