Lead Scoring Case Study Summary

Problem Statement:

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.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

The company requires to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Analysis approach:

➢ **Loading and Understanding Data:**

- The data provided is loaded and dimensions of the data set is determined, the data provided have rows and columns.
- The data set columns are then understood by checking all the data types of variables present and the checked for missing or absurd values in the dataset.

➢ **Data Cleaning:**

- After checking null values, we found many null values in certain columns and some values like “SELECT” which made no sense from our analysis perspective.
- The columns having >50% of null values are dropped and those having large absurd values including null and “SELECT” are dropped too.
- Columns having highly skewed values e.g. 90% of similar values were dropped as it might overfit our model, if that feature is selected.
- Outliers in the dataset were capped after visualizing the boxplot of that numerical column and were capped to either 99/95th percentile.

➢ **Data Visualization:**

- Count plot for categorical variables and if some data issues are found, that is rectified.
- Similarly, box plot is used for checking and visualizing distribution of continuous variables.
- Initial inferences are derived based on the plots as well.

- **Data Preparation:**

- Data is prepared as a pre modelling step, the categorical binary variables are mapped to numeric binary as 0 and 1 (No and Yes respectively).
- Creation of N-1 dummy variables for parent variables having N level categories.
- Scaling all the continuous variables using “STANDARD SCALER” so that all the variable are on same scale irrespective of their original unit and magnitude.
- Creating train and test data by test_train split, we have chosen a 70:30 (train test split) ratio.
- X_train,X_test were created from above steps having all the independent features and Y_train and Y_Test with the target variable.

- **Model Creation:**

- Feature selection is done first by automated RFE approach, accompanied by manual selection based on p-value (<0.05) and VIF score (<3).
- The logistic model is fit and transformed on the train set and the predicted probabilities are then tested by forming a “confusion matrix” with 0.5 as the cutoff value.
- Various model check parameters (precision, accuracy, recall etc) are then checked for different cutoff values from 0 to 1.
- ROC curve is then plotted, and the intersection point is taken as the optimal cut off point to predict if the probability needs to be converted to successful lead or not.
- lead scores are then calculated by multiplying the conversion probabilities by 100. Here, higher lead scores have higher probability of getting converted i.e., they are hot leads.

- **Model Predictions:**

- The area under ROC curve is around 0.86 which is pretty good.
- Optimal threshold value is found by the intersection of plot between accuracy, sensitivity and specificity which came out to be 0.3.
- Probabilities above 0.3 are converted to successful lead conversion i.e value 1.
- Lead Score is created based on percentage of the predicted probability.
- By predicted conversion rate plotted we conclude that Lead Score of >= 60 are considered as hot leads.