Lead Scoring Case Study Summary

Problem Statement:

An EdTech company named as X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company requires us to build a logistic regression model wherein you 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach:

We were given a data set (Leads.csv) which contains data of incoming Leads. Dataset includes many features like Lead origin, their last notable activity, their occupation, Specialization etc.

- **Reading and understanding the data set**
- **Data cleaning** –
    i. There were many columns which had Null values in them. Also, there were certain columns which has "Select" value indicating no selection was made.
    ii. We have dropped all the columns which had Null value percentage more than 45% except Lead Quality.
    iii. For the remaining columns with less than 45% missing values we imputed them accordingly.
- **Data Analysis** –
    i. We performed exploratory data analysis of categorical and numerical features.
    ii. Performed outlier analysis on numerical variables and treated them by removing the 99% quantiles data.
- **Data Preparation** –
    i. We created dummy variables for all the categorical variables having categories more than 2.
    ii. Performed feature scaling on numerical values using **StandardScaler** method of Scikit learn.
- **Train-Test data split** – Divided the original data set into Train and Test data set with **70-30%** values respectively.
- **Model Building –**
    i. We created a model with all the 59 variables.
    ii. Further, we used RFE method to achieve the final model initially with 20 high relevance features of data set. With the help of statistics generated, we

recursively calculated VIFs and noted p-values of features in order to remove the insignificant features from the dataset.

iii. We achieved the required statistics in 4th model which has 18 significant variables.

iv. Calculated Accuracy, Sensitivity, Specificity and other metrics with final model for train data set.

- **Plotting ROC curve** – ROC curve shows the trade-off between sensitivity and specificity. With our model we are getting a good value of 0.94 indicating a good predictive model.
- **Finding Optimal Cutoff point** –
  i. We plotted probability graph with "**Accuracy-Sensitivity-Specificity**" values obtained from final model. The intersection of these values gives the optimal cutoff point which comes out to be **0.38** for our model.
  ii. Based on this cutoff value we calculated the "**Lead Score**" between **0 and 100**.
  iii. Also, we predicted the **Conversion rate** which comes out to be **83%**, i.e. our model has achieved the target conversion rate given by CEO of X education.
  iv. We have also plotted **Precision-Recall trade-off** curve which has given **optimum cutoff value as 0.48**.
- **Predictions on Test data set** – In this step we have validated the final model on test data set and calculated evaluation metrics for test data.

Evaluation metrics

Train Data Set	Test Data Set
Accuracy – 86.68	Accuracy – 86.23
Sensitivity – 83.85	Sensitivity – 83.25
Specificity – 88.41	Specificity – 88.09
Precision – 81.50	Precision – 81.43
Recall – 83.92	Recall – 83.31

Conclusions:

1. Actual conversion ratio was **38.02%,** predicted conversion ratio is **83.25%.**
2. The final model has 18 features predicting the target variable.
3. Optimum cut-off was obtained from Sensitivity-Specificity-Accuracy plot and value was chosen as **0.38**.
4. **Accuracy, Sensitivity and Specificity** values of test data set are around **87%, 83% and 88%** respectively.