Lead Scoring Case Study

Summary Report

Problem Statement:

To help 'X Education' to increase their lead conversion rate by identifying the hot leads who are most likely to be converted. To be specific, the company wants to build a model which has a lead score for each lead. The customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The lead conversion rate is expected to be around 80%.

Solution Approach:

The solution to the given problem has been met by using the logistic regression technique. The concepts of specificity and sensitivity, precision and recall are also leveraged to arrive at the final solution. There are various steps involved in arriving at the final solution which are as follows:

1. **Data Understanding**: Understanding the data thoroughly, its types, the values *statistics* and the total count of values present. Checking for *null values* are also performed at this step.
2. **Data Cleaning**: The columns having **high null** percentage, **highly skewed** columns and the columns with not much **information** about the target variables are all dropped as part of the data cleaning process. For some columns, even though there are more null values they seemed as potential candidates for predictions and hence are handled by the **imputation** technique (*Mean, Median and Mode*). The **binomial** categorical columns are converted into numerical values (*0 and 1*) for having same scale and better interpretation.
3. **Data Visualization: Univariate analysis** on the numerical variables has been carried out to identify the outliers in the data. These **outliers** are then handled by the **capping** technique
4. **Data Preparation:** In order to leverage the logistic regression model using the '**statsmodel'** library, the data are expected to be numerical. Hence all the categorical variables are undergone **one-hot encoding** (*dummy variable creation*) to achieve this. Some of the variables appeared to be **highly correlated** to each other which was identified using the **correlation matrix** and heat-map visualization. These are noted for future interpretation. The data then is split into **train** and **test** at a **70:30** ratio, so that we can train our model using the train data and predict on the test data. The data is scaled using '**Standard Scalar'** technique and then the target variable is separated from the predictor variables.

5. **Modelling:** After dummy variable creation, quite a number of columns has been added to the dataset and hence to select only those variable which may have high impact on the target, the **RFE** approach is used. The features selected by this approach are then trained using the logistic regression using stats models. Based on the **p-value** and **VIF** values, the columns are dropped one by one and the model is re-trained until the final model with optimum p-value and VIF are reached.
6. **Model Evaluation:** The prediction is carried out first on the train data with the final model. The best metrics (*confusion matrix, accuracy, sensitivity, specificity and F1-score*) are used to evaluate the model after the **prediction probabilities** are converted to final prediction using a random **cut-off** for the probability. Since the metrics evaluated the model to be less effective, the **optimal cut-off** was chosen using the **ROC curve**. Then with this optimal cut-off and model, the test data predictions are found out and then evaluated using the metrics and found to be much effective.
7. **Precision-Recall:** As another approach, the **precision and recall** are also calculated for the final model and the optimal cut-off has been found out using the **trade-off plot**. The metrics (*confusion matrix, accuracy, precision, recall and F1-score*) are then calculated for this optimal value and found to be as effective as the previous approach (*sensitivity and specificity*).
8. **Final Result:** The **lead scores** are then calculated for each of the leads on the test data with the final model having the expected high conversion rate. With this lead score, the **hot leads** (*with high lead scores*) are obtained which can be shared to the sales team to draft out various marketing strategies based on these leads' preferences.

Summary:

- The equation for Logistic Regression based on the final model `Model 3` is:
$$Converted = 0.9303 \times const + 1.1188 \times TotalTimeSpentonWebsite + 3.2419 \times LeadAddForm - 0.6192 \times DirectTraffic + 1.2087 \times OlarkChat + 2.5373 \times WelingakWebsite - 1.5277 \times Student - 1.5044 \times Unemployed + 1.0724 \times WorkingProfessional - 1.3279 \times Last\_EmailBounced - 0.8449 \times Last\_OlarkChatConversation + 0.9830 \times Last\_SMSSent - 0.8698 \times Modified + 2.5536 \times Unreachable$$

- The Optimal cut-off chosen is `0.38`
- Model Accuracy: `0.7740458015267175`
- Sensitivty: `0.8373983739837398`
- Specificity: `0.7104994903160041`
- F1-Score: `0.7877629063097515`

- After Precision-Recall Tradeoff, the cut-off for probability is `0.4`
- Model Accuracy: `0.7786259541984732`
- Precision: `0.7720515361744301`
- Recall: `0.7916666666666666`
- F1-Score: `0.7877629063097515`