Lead scoring assignment – Summary report

Problem statement:

- ❑ An education company X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- ❑ The typical lead conversion rate at X education is around 30%.

Business Objective:

- ❑ X Education need to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- ❑ Target lead conversion rate is around 80%.

Methodology:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

Solution Summary & learnings:

1) **Data Understanding & exploration**
   - o Read the data and the dataset contained 9240 rows and 37 columns
   - o There were no duplicates in the data, but few columns had > 25% null values
   - o Datatypes of the column - 4 float64, 3 int64, 30 object (7 Numeric and 30 categorical columns)

2) **Data cleaning**
   - o Dropped the columns not useful for analysis like Record index/unique IDs
   - o Replaced 'Select' with NaN as select indicates no options were selected
   - o Treated the missing value
     - a) Dropped the columns with missing values > 50%
     - b) Column by column analysis and imputation with apt means like median or with Others/Not specified etc wherever the data is not available, but cannot be imputed with other entries
   - o Outlier treatment

     a) Removed top 5% of the column outlier values considering difference between 95% quantile and max value

     b) No capping done at bottom 5% as the value is same as min value for all columns

3) **Exploratory Data Analysis**
   - o Analysis of both categorical and numerical variables were done to identify the critical variables impacting lead conversion and dropped other variables which will not add value to the model
   - o Correlation matrix was also built to check for correlation between variables and multicollinearity

4) **Data Preparation**
   - o Before dummy variable creation, some of the categorical variables have Yes/No as responses. Converted them into 1s and 0s
   - o Created dummy variables for categorical columns and concatenated with original dataset. Dropped original columns after dummy variable creation
   - o The data is split into train and test set [70: 30 respectively]
   - o Used standard scaler for scaling the numerical columns
   - o Fit and transform done for training data

- Conversion rate is around 39%. This is neither exactly 'balanced' nor heavily imbalanced. Hence have not done any special treatment for this dataset.

5) **Model building**
   - Generalized Linear Model (GLM) from Stats models library is used to build logistics regression model
   - Considering a large number of variables in the dataset and insignificance of many towards building the model, initial feature selection was made through RFE and selected 15 features
   - From the top 15 features, final variable selection was made manually by running iterations to eliminate features considering p-values and VIF
   - Final model built has all p-values almost zero and VIF < 2

6) **Model Evaluation**
   - The model built using selected features underwent prediction in training set first
   - Selected an arbitrary cut-off probability point of 0.5 to find the predicted labels
   - Through evaluation metrics, results obtained were:
     Accuracy – 82%, Sensitivity – 71%, Specificity – 89%, Precision – 80%
   - All the metrics are above 80% except for Sensitivity which is 71%. This may be due to arbitrarily chosen cut-off of 0.5.

7) **Drawing ROC curve**
   - The ROC curve is used to evaluate the performance of the model built. It shows the trade-off between true positive rate (TPR) and false positive rate (FPR) across different classification thresholds.
   - The area under the curve of the ROC is 0.89 which is quite good

8) **Finding optimal cut off point**
   - From the curve, 0.37 is the optimal cutoff probability arrived at.
   - The evaluation metrics with 0.37 cut off shows improvement in sensitivity which is 79% now
     Accuracy – 81%, Sensitivity – 79%, Specificity – 82%, Precision – 74%

9) **Prediction on test data**
   - Made prediction on test dataset with 0.37 cut off. Evaluation metrics showed results almost same as training dataset
   - Final prediction on conversion rate is close to 80%. Hence model built seem to do a pretty good job

   Relative Feature importance

   Visualized 12 variables sorted in order and showing the positive / negative impact on the conversion probability

<u>Conclusion:</u>

- ❑ The final model built has all p-values as almost zero and VIFs are also very low. Hence there is hardly any multicollinearity

- ❑ The metrics show a pretty good achievement of Accuracy: 81%, Sensitivity: 79%, Specificity: 82% & Precision: 74%. The metrics for training dataset is also almost same

- ❑ The lead score calculated on both training & test dataset shows a conversion rate on final predicted model around 80%

- ❑ Hence the overall model built seems good

Recommendations:

- Since 'API' and 'Landing Page Submission' generate the most leads but have less conversion rates, target to increase the conversion rate from these origins. Also focus on increasing leads generated using 'Lead Add Form' as the conversation rate is pretty high at 93%
- To improve overall lead conversion rate, focus on improving lead conversion of olark chat and generate more leads from welingak website as the conversion rate is too high at 99%
- Company should focus on working professionals as they are easier to convert with good conversion rate of 92%
- Focus on total time spent on website as it has a positive correlation with conversion. Also, attention should be given to the prospects whose last activity is 'SMS sent' as 63% of conversions are happening based on it