Summary

Problem Statement

X Education, an online education company, receives a large number of leads through various sources, but its lead conversion rate is quite low at approximately 30%. The company has tasked us with developing a model to assign a lead score to each lead, ensuring that customers with higher lead scores have a better chance of conversion. The CEO's goal is to achieve a lead conversion rate of around 80% by identifying the most potential leads, also known as 'Hot Leads'.

Goal of the Case Study

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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Step by step approach

Data Cleaning

- Columns with >30% missing values were dropped.
- Columns with only one unique value were dropped.
- Missing values were imputed In a few columns
- Columns with skewed values were dropped
- Outliers in the numerical columns were balanced by capping the values
- Created new category (others) where the columns could not be dropped

EDA

- Started with Univariate analysis, and understood data imbalance
    - Data Imbalance Ratio - 0.62
- Performed Bivariate analysis for categorical and numerical variables
    - Time spend on website shows positive impact on lead conversion
    - Working Professionals show a great conversion rate

Model Building

Data Preparation

- Created dummy features (one-hot encoded) for categorical variables
- Scaled the data for model building using Standardization

Model building

- Split the data into test and train datasets
- Used RFE to reduce variables from 48 to 15. To make DataFrame lean.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05 and VIF > 5
- Total 3 models were built and we finally found the results satisfactory with No sign of multicollinearity or high P-value.
- logm3 was selected as final logistic regression model with 13 variables. This was used for making prediction on train and test set.

Model Evaluation

- Confusion matrix was made and an optimal cut off point of 0.35 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity, precision and Recall all around 80%.
- These results were inline with the target given by the CEO i.e. 80% conversion rate
- Lead score was assigned to train and test data
- We are getting the same result on the test set with the Area under the ROC curve as 0.88. That means we have a good predictive model

Results

The evaluation shows that the values of the performance metrics on the test set are very close to those of the train set. Hence, we can conclude that our model is performing very well with predicting the data.

- The model achieved a sensitivity of 80.27% on the test dataset with a cut-off threshold of 0.35.
- Sensitivity here reflects how many leads the model correctly identifies out of all potential converting leads.
- The model also reached an accuracy of 80.27%, aligning with the study's objectives.