Lead Scoring Case Study – Summary

1. Objective

The objective of the Lead Scoring case study is to improve the sales activity by increasing the lead conversion rate of persons showing interest in the online courses of X-Education.

2. Approach

Data of existing leads is analysed to understand the features that impact the conversion of leads. Since the Lead Conversion is indicated as 1 (Converted) and 0 (Not Converted), Logistic Regression model is considered. The final output is to assign a Lead Score to each lead to help decide if it is a Hot lead. This will help the sales team's decision to call the lead to facilitate the conversion.

3. Logistic Regression Model

a. Data Analysis and Cleaning

The initial data had 37 fields and 9240 records. The initial lead conversion rate was 38%.

Actions were performed as part of data cleaning

- Checks for duplicate records: None found
- Checks for Null values: Removal of columns or rows as part of Null Value treatment
- Checks for data imbalance/skewness
- Outlier treatments
- Combining Multiple Categorical Values

b. EDA

- Univariate analysis on the target variable "Converted" for Lead conversion rate and other variables for data cleaning and preparation
- Bi-Variate analysis on categorical and numerical variable using Bar Plots, Box Plots and Heat Maps to understand relationships between variables

c. Data Preparation for modelling

i. Data Transformation:

- Boolean values of Yes and No converted as '0' and '1'.
- Dummy Variables by encoding the categorical data

ii. Test Train Split:

- The data set split into test and train sets with a proportion of 70-30 for model prediction and evaluation.

iii. Feature Rescaling

Scaling done for numeric values using the MinMax Scaler.

d. Building the model on Train data set

- The Logistic Regression models built on the train data using the Python packages of statsmodels and sklearn.
- The RFE approach taken to create the initial model of 15 most significant variable from an initial of 40 variables
- The VIF values taken to eliminate variables with highest corelation.

e. Model Evaluation

- Using the Confusion Matrix, the accuracy, recall and precision values calculated to measure the model

- Using the iterative process to come up with the best model
- The ROC Curve and AUC further confirms the model
- The coefficients of the variables indicate the significance in the model
- The final cut-off point calculated.

f. Final evaluation on Test Data

- The final model applied on test data set and predictions evaluated based on the metrices and model confirmed.

g. Final set of important variables impacting the lead conversion

- Tags, Lead Origin and Total Time Spent on Website

h. Lead score calculation and Cut Off

- Lead score, calculated from the predicted conversion probability, helps to identify Hot Leads with values above a defined Cut-off.
- Probability Cut-Off recommended is 0.4 to meet the problem statement

4. Conclusion

This case study helps in implementing most aspects of EDA and Logistic Regression modelling. It also facilitates at understanding business challenges and using the appropriate metrics to solve it.