Summary

This case study of X Education is analyzed to have an insight on better conversions of leads. The data provided had information about different activities and modes through which leads are found. Those variables are used for analysis. Exploratory data analysis and logistic regression model has been implemented.

Problem statement:

X Education wants to select most promising leads that can be converted to paying customers. Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, google searches etc. The company has had 30% conversion rate.

Objective:

The company requires a model to be built for selecting most promising leads. Lead score to be given to each lead such that it indicates how promising the lead could be. The model to be built in lead conversion rate around 80% or more.

Steps followed:

1. **Data import**
    - Read csv file and imported libraries
    - Glancing data and column structure

2. **Data cleaning**
    - Null values treated through imputation and dropping columns with high null values
    - Case error and repeated values are replaced
    - Outliers are found and treated

3. **Exploratory Data Analysis**
    - Analyzed all categorical values and numerical values
    - Visualized various variables with converted variable
    - Found that people spending more time in the websites are promising leads
    - SMS messages have high impact
    - References have good conversion rates

4. **Data preparation for model building**
    - Dummy variables created for all categorical data

5. **Train-Test split**
    - The data set is split into 70-30 as train and test data

6. **Feature scaling**
    - Numerical values are scaled through MinMaxScaler

7. **Model building**
    - First logistic regression model is built with all columns with GLM method
    - Variables are checked and dropped through RFE
    - Models were built to check the alteration made and its impact on the model
    - Manually dropped columns that had high P- values
    - Checked VIF to remove columns that are highly correlated
    - Predicting with training set

8. **Model evaluation**
    - Confusion matrix prepared to evaluate
    - Accuracy, Sensitivity, Specificity were measured as 80.9%, 77.6%, 82.9% respectively.
    - ROC curve was plotted
    - Precision and recall were also measured and plotted, it measured as 73.4% and 77.6%
    - Threshold was selected keeping all the metrics as the base

9. Making predictions on test set

- Scaled test set numerical variables
- Predicted using test set
- Accuracy and Sensitivity, Specificity were measured as 80.1%, 75.5%, 83.1% respectively
- Precision and recall were measured as 74.4% and 75.5%.