Lead Scoring Case Study Summary:

Problem Description:

An education company named X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor and is around 30%. X Education needs help with building a logistic regression model so as 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 CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach:

Reading & understanding the data:

✓ Checked the shape of the data

✓ Data types for each column

✓ Got the descriptive statistics for the numerical columns

✓ Did basic research to get better understanding of the domain

Data Cleaning:

✓ Converted 'Select' values to null values.

✓ Missing value treatment:

Exploratory Data Analysis

- ✓ Did basic EDA and identified very interesting patterns in the data.

Data Preparation:

- ✓ Created dummy variables the categorical columns with more than 2 categories using the pd.get dummies function
- ✓ Performed a 70-30 spilt the leads dataset into Train and Test respectively
- ✓ ✓ Performed feature scaling using the standard scaler.

Model Building:

✓ We shortlisted the top 15 features using the Recursive Feature Elimination (RFE) technique to build our first model.

Model Evaluation:

✓ We also calculated the metrics sensitivity, specificity, precision, and accuracy. ✓ To make predictions on the train dataset, optimum cut-off of 0.34 was found from the intersection of sensitivity, specificity and accuracy

Predictions on the Test Set:

Final Observations:

Below are the predictor variables that we used in our final model and their relative importance:

1.lead origin_lead_add form

2.whats is your current occupation_working professional

3.Last_activity_SMS sent

4.Total time spent on website

5.Lead_source_olark chat

6.Last_activity_email opened