Lead Scoring Case Study Summary

In order, for X Education to achieve high lead conversion rate through all possible leads generated, below steps were taken,

- **Data Cleaning**
  Replaced value ‘Select’ in dataset with nan, as given in problem statement.
  Dropped columns having missing values more than 40%.
  Replaced missing values with median/mode.
  Less < 2% rows were having missing data. Dropped these rows.
- **EDA**
  Univariate/Bivariate/Multivariate analysis done.
  Lead number column dropped as it was just number unique to each lead.
  Basis EDA, ‘What matters most to you in choosing a course’, ‘Search’, ‘X Education Forums’, ‘Digital Advertisement’, ‘Magazine’, ‘Newspaper’, ‘Newspaper Article’, ‘Through Recommendations’, ‘Country’, ‘City’, ‘Receive More Updates About Our Courses’, ‘I agree to pay the amount through cheque’, ‘Update me on Supply Chain Content’, ‘Tags’, ‘Get updates on DM Content’ columns were dropped as most of the values in these columns were monotonous values.
  Do not see any linear strong relationship in multivariate analysis.
- **Data Preparation**
  Assigned binary values to columns 'Do Not Call', 'Do Not Email', 'A free copy of Mastering The Interview'.
  Dummy variables were created for 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Last Notable Activity'.
- **Train test data split**
  Data split into 70% - 30% as train - test respectively
- **Feature Scaling**
  Scaled 'TotalVisits','Total Time Spent on Website', 'Page Views Per Visit' these variables in data.
- **Model Building**
  20 Features selected using RFE for model Building.
  Assessed model using statsmodel GLM for selected 20 features
  Calculated VIF for all selected features.
  Basis P value being higher, dropped “What is your current occupation_Housewife” column.
  Rebuild model for remaining features.
- **Model Evaluation**
  Predicted values on train dataset.
  Created actual vs predicted conversion flag.
  Created converted column assuming optimal cutoff as 0.5.
  Analyzed confusion matrix.
  Calculated accuracy, sensitivity, specificity of model and found sensitivity was low.

- **Finding optimal cutoff**
  88% area was covered under ROC.
  Created columns with different cutoffs and calculated accuracy, sensitivity and specificity for all of them.
  Line chart plotted of accuracy, specificity and sensitivity against probability to find maximum cutoff and found value as 0.37.
- **Calculate accuracy, sensitivity, specificity**
  Calculated final accuracy, sensitivity and specificity using predicted calculated with optimal cutoff.
- **Precision/Recall and its Tradeoff**
  Calculated precision and recall using confusion matrix as well as sklearn utilities
  Plotted precision and recall tradeoff.
- **Making Prediction on test data**
  Scaled test data.
  Predicted values on test data.
  Predicted final conversion rate values using optimal cutoff on test data.
  Calculated accuracy, sensitivity and specificity on test data.
- **Comparing Train and test data results**
  Compared accuracy, sensitivity and specificity of test and train data and they were meeting ballpark of 80% rate expected by X Education
- **Finding features contributing to conversion**
  Below features contributed to conversion rate most,
  Total Time Spent on Website
  Last Activity_Had a Phone Conversation
  Lead Source_Welingak Website
  What is your current occupation_Working Professional
  Lead Origin_Lead Add Form