Summary

Details of Approach and Learnings

- Data Sanity, Data Cleaning, Correlation, Data Preparation

  - ✓ Browse through the data dictionary and understanding the domain / variables.
  - ✓ Import / Load the data sets
  - ✓ Check the shape, info, column names, data types of the data set, check for null, empty, missing values
  - ✓ Drop columns where missing value is greater than 30-40%
  - ✓ Drop columns than have high proportions of 'Select' since they are same as null
  - ✓ Create dummy features wherever applicable for categorical variables
  - ✓ Impute / remove missing values and standardize the values where needed
  - ✓ Check for outliers and skewed them
  - ✓ Perform univariate analysis
  - ✓ Perform bivariate analysis
  - ✓ Converted variables with Yes / No flag to 1 / 0
  - ✓ Created dummy feature for categorical variables with multiple levels

- Started with model building steps. Test-Train Split with 70-30 ratio
- Scaling Feature – Used standard scaler for skewed numerical outliers. Calculated initial lead conversion rate of 38%

- Feature Selection – Used RFE for feature selection to identify most significant features
- Model Building – Assessed model using statsmodel library using sm.GLM function.
- Model Evaluation –
  - ✓ Simultaneously iterated through features selection using RFE.
  - ✓ Dropped columns with high P-values.
  - ✓ Checked VIF value of feature variables.
  - ✓ Dropped columns with VIF values greater than 5.
  - ✓ Once all features have VIF value less than 5 we can proceed with making predictions using the model.
- Model Prediction –
  - ✓ Find out metrics value for sensitivity, specificity, false positive rate, positive predictive value, Negative predictive value.
  - ✓ Plot the ROC curve to get tradeoff between sensitivity and specificity. To get accuracy of the test.
  - ✓ Plot accuracy sensitivity and specificity for various probabilities to get cut off probability
  - ✓ Calculate conversion rate.
  - ✓ Calculate Precision and Recall tradeoff.
  - ✓ Complete model prediction on the test set
- Model Summary - Conversion rate is almost 90% which is a huge improvement from the initial 30% and we are achieving target lead conversion rate as was expected by CEO of X Education company.