Summary

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.

Solution Approach :

1. Importing the data and inspecting the data frame
2. **Clean and prepare the data**
    - Handling missing values
    - Replace "Select" with null. Drop columns with >40% nulls.
    - Impute categorical and numerical values.
    - Remove outliers.
    - Standardize data.
3. **Exploratory Data Analysis**.
   Performed Univariate, Bivariate and multivariate columns. Columns with only one unique or all unique rows were dropped
4. **Data Preparation for model building:**
    - **Creating Dummy Variables** : Created dummy data for the categorical variables.
    - **Data Split:** Data set was split into two sets, train and test, at a ratio of 70:30
    - **Data Scaling:** Numerical Columns were scaled using MinMax scaling to convert them into normalized form
5. **Model Building:**
    - Recursive Feature Elimination (RFE) technique was used to select only the top 15 columns
    - This was further optimized to give a better model based on statsmodel parameter (p-value) and VIF
    - Model 3 looks stable after three iteration with:
        - ✓ significant p-values within the threshold (p-values < 0.05) and
        - ✓ No sign of multicollinearity with VIFs less than 5
6. **Finding the Optimal Cutoff Point :**
    - An arbitrary probability of 0.5 was selected which gave a poor balance between specificity and sensitivity
    - Accuracy, Specificity and sensitivity were plotted and their intersection at 0.37 was used as probability threshold

    - Precision and Recall tradeoff were also calculated which gave a threshold of 0.41.
    - However, metrics were better balanced with Accuracy, Specificity and sensitivity intersection, so 0.37 was selected
7. **ROC Curve**: ROC curve was also plotted which gave area under curve as 0.88 out of 1.
8. **Making Predictions on Test Set:** Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 81.2%; Sensitivity=81.3%; Specificity= 81.1%.