Leading Scoring 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%

Goal:

- Build a logistic regression model 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations..

Solution summary

1. Step 1 -2: Importing library and data and inspec data
2. Step 3: Data cleaning
    - There few column with ‘Select’ value which mean the leads did not select option. We changed it to null value.
    - Identifying Missing Values, Drop columns having Null values greater than 35%
    - Then we do Categorical and Attributes Analysis
    - Finally we drop some imbalance data colums

3. Step 4: Data Preparation
    - Converting some binary variables (Yes/No) to 0/1 then creating dummies and dropping the first column and adding the results to the master dataframe
4. Step 5: Test-Train Split: Devide dataset into test train with proportion of 70 and 30%.
5. Step 6: Rescaling the feature
6. Step 7: Building model
    - Using RFE technique to remove attributes and build a model for the remaining attributes.
    - Making the model stable by using StatsModels library, we check the p-values are less than 0.05 and VIF value is under 5.
    - Once the stable model is created, we will predict probabilities on the train set and create new columns predicted.
    - Calculate the confusion matrix on the predicted columns to the actual converted column. Also, calculate the metrics sensitivity, specificity, recall and accuracy. And plotting ROC Curve.
    - Prediction on Test Set