Lead Score data analysis and model creation to find the model which can find major leads is a very good example how in today's scenario organisations are enhancing their sales. In this we have been provided with the raw data which was first being analyzed for their data types, columns etc.
Then the process of data cleaning where the first problem was to handle the ‘select’ property which is resolved by using null values and then dropping those rows. Then dropped the columns with a higher percentage of null values. One more hurdle was to handle the null value of specialization which has to be handled manually.
Next step was to do Exploratory Data Analysis where first the binning of the columns data are done with less percentages of values into one. Then Outlier treatment of data is done and the columns are checked if it is required or not, if not then columns are dropped.
Next Step is Data Preparation in which the columns with boolean values are converted to numbers and the categorical columns are converted to dummy columns.
After the above steps our data is ready for doing the next step which is splitting the data to train and test. Then the scaling of the data is done so that the dataset can be treated in one scale.
Model Building is the step from where our main business starts and has created the model first with statsmodel and check for the coefficient and p-value. After that used the RFE from sklearn to create the model initially with 15 columns.
Accessing the model with p-value and dropping the columns and doing the model evaluation. Creating a dataframe with the actual converted flag and the predicted probabilities
After that Confusion matrix is used to check the accuracy and specificity and alternatively drop the columns which have more RFE value. This process is repeated till the specificity is more than 80%.
Then finally checking for the ROC curve which helped to check how much of our model is efficient. One problem here was to find the optimum value which is resolved by finding the optimum value of ROC by method provided in the python notebook where the optimum probability came as 0.4.
Then done Differentiating Converted and Not-Converted based on new cut-off. prediction done on the test data set where specificity came as 88% which tells our model is effective.
Finally got the parameters which are to be looked out for more productivity which are

- Do Not Email
- Monitoring Lead Origin_Lead Add Form
- What is your current occupation_Working Professional