Summary

This analysis is done for X Education where the business demand is to find ways to get more candidates to join their courses. The basic data provided, gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.
Business Objective includes building a logistic regression model to assign lead score value between 0 to 100 to each of the leads which can be used by the company to earmark potential leads.

These are the steps used for achieving the agenda:-

1. Understanding the dataset

1. We began by basic steps like data reading, checking numeric columns, checking for different data types available and shape of the data file.
2. We dropped a few of the redundant and not so necessary columns.
3. We checked for the percentage of null values and then dropped those exceeding 45% of null values.
4. Then we did some imputation and visualization on the remaining columns containing lesser null values.
5. Here we found a few of the columns containing an attribute ''Select" that hasn't been selected by candidates while filling the form ,so we replaced all of those with null and then did the necessary immupatation.
6. We also found a few of the columns which were part of the sales team job and hence these were not essential for our model ,hence we dropped these columns.
7. We also dropped most of the skewed columns.
8. Then we checked for numerical variables and plotted a heatmap to observe collinearity checked for outliers by plotting box plots and checking different quartiles for rise of count i.e. whether it is gradual or not. Did the outlier treatments as well and are visualized.

2.Data Transformation and Data Preparation:

- We Changed the multicategory labels into dummy variables and binary variables into ‘0’ and ‘1’.
- The dummy variables were created and later on the dummies with ‘not provided’ elements were removed and dropped the original columns.
- We Split the dataset into train and test dataset into 70-30 ratio and scaled the dataset.

3.Model Building:

- Firstly,we used both statsmodel and then RFE was done with the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
- Then we after achieving right p values and vif we proceeded with the prediction of the leads.

4.Model Evaluation:

- A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity.
- The accuracy was around 80%, specificity was 82%, sensitivity around 75% for both train and test set.
- Predictions were made now on the test set and predicted values were recorded.
- We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is. and it was showing similar predictions to the training set.

5.Conclusion:

After trying a few of the models we finally arrived at a model where all the correlated values and high p values were removed.

i.In the Test set we had accuracy, recall/sensitivity in an acceptable range.

ii.In business terms, our model is having stability and accuracy with adaptive environment skills. Means it will adjust with the company's requirement changes made in coming future.

iii.Top features for good conversion rate:

1.Lead Origin_Lead Add Form
2.Total Time Spent on Website
3.Lead Source_Welingak Website
4. Specialization_Banking, Investment And Insurance

So, in the end we satisfied the agenda of the analysis and provided necessary steps that can be followed by the company to achieve its target.