SUMMARY

The Eduaction Tech company has provided the data set with a lot of features which helps us to study the data in a sense and analyse the problem. The company have the courses for the professionals for which they advertise with different process as search engine and use different websites as well. The problem which is given as the company have the high expenditure on converting these professionals so they want to cut down the price on such and also want to increase the profit by focusing on the hot leads (i.e., the leads which have a greater chance of conversion). The company targets as 80% for the conversion rate which is initially as 39%. And also, it is given that in the converted columns have the value as 0 and 1 which basically represents non-converted and converted respectively. So, we will try to increase the 1 as possible to satisfy the demand of the CEO of the company.

By the help of the Logistic Regression Algorithm, we solve the problem and will predict the Lead Score for those who are going to converted between 0 to 100.

As we imported the necessary library by the help of that the data set is loaded and then we move forward with data sanity checking where we found that few of the columns have the selecting category which may be useful so we converted those to np.nan [Null Values] so that it can be treated if required. As we also found that due to the greater number of features the data frame is too hazy as we are not required few of the features which we realize it after reading the dictionary given which explain about the features so, finally we decide it to drop those features. After dropping the features, we merge some categories which are more in number in the features so that the data set can be easily understandable. As these things are done, we divide the data set into numerical data and the categorical data and analyses the specific features with the Exploratory Data Analysis. We also see the boxplot for the numerical columns so that we can treat the outliers by capping method. After it reached to do the mapping of the variables and creating dummies to the categorical variables so that it can run fast as the binary digits are more easily understandable. Moving forward we do the splitting of the variable as train and test where we considered 70 percent as the train data and remaining to be test and then we scale the data using the standard scaling. This helps us to get all the data into same scaling value. And as we get all the features on the same scale, we use the automatic feature selection for selecting the best feature of all using the RFE and then we do the manual section using the p-value method and the VIF. We continue model building till the value of VIF and p-value under the 5and 0.05 for it. After finalizing the model, we assumed the cut off and calculated the confusion matrix against it and then we draw the ROC curve for it. After it we draw the lines of intersecting of sensitivity, specificity and the recall we get the point of optimal cut off and then we move forward to search for precision and recall matrix and again we draw a curve which is called as threshold curve as the point of intersection of precision and recall. After it we go with test data set and again, we do the scaling as without fitting the data just by transforming it. We do the prediction using the final model and we achieve the predictive prob of conversion. Now we predict the conversion using the optimal cut-off. The prob above the optimal cut off are those which are converted as the hot lead and the below optimal cut off it is seen as non- converted. For performance checking we again do the accuracy check, sensitivity check, specificity check using the confusion matrix and finally we add column i.e., Lead score which is actually prob*100 and we are having a lead score for the professional.