Lead Scoring Summary Report

An education company named X Education sells online courses to industry professionals, their current lead conversion rate is 30% which is very poor. CEO wanted to improve their lead conversion rate, where our objective is to use Logistic Regression model to improve the lead score and create a model which can have a lead conversion rate of 80%.

Firstly, we inspected the data and found data needs to be cleaned. For which, we dropped the irrelevant columns. We also removed columns which were skewed giving bias information. Categorical columns were combined into one. Also, some outliers in few columns were handled by capping.

After which we performed EDA on our cleaned dataset where we found some useful information which can be very useful for the sales team. Customers having higher average time spent on websites were more likely to convert. Also, customers coming from references has higher conversion ratio. Regarding conversion rate, it is maximum for Leads from Add forms. Moreover, it is observed that although people visiting websites who are Unemployed is high, conversion rate is maximum for Working professionals. It is noted that people with Last activity as Email sent or SMS sent has higher conversion rates. It is also seen that people who belongs to Mumbai have higher count, so company should also start focusing on other Cities.

We created Dummy variables for categorical variables to perform further modelling after which we split the test and train (7:3). On Train dataset we did rescaling of features where in our case we used Standardization Scaler method.

Next step was to perform modelling using the function GLM () from statsmodel library. This model contained all the variables, some of which had insignificant coefficients. Hence, some of these variables were removed first based on RFE. Variables having a high p-value were removed one-by-one, because they were insignificant. After this, the VIF was checked of all the remaining variables. After performing logistic regression and dropping variables one at a time, we reached a level where all the P values were almost 0 and VIF was under control. We can use this model for predictions on Train data set. To do this we chose an arbitrary cut-off value of 0.5 and getting probability score.

To assess this model, we used confusion matrix to get the performance of our model. Using this model, we calculated the accuracy which was 80.6% but Sensitivity was very low around 66.8% and so we need to change our cut-off point. For this we had to plot a ROC curve which suggested us that threshold value should be 0.35. By using this value, we got an overall accuracy of 80.6% and Sensitivity (Recall) of 80.5% and Specificity of 80.7% which was pretty good. We applied this model on our Test data set and achieved accuracy of 80.8% and Sensitivity 80.4%. Thus, we have an efficient and balanced model which will be helpful for sales team to get conversion rate of around 80%.