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

Goals
Help X education select the most promising leads, i.e. the leads that are most likely to convert into paying customers. Build a model to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

1. 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.

2. 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 Methodology

The first step in our assignment was to understand the given data set at hand. The data set was named Leads.csv. It had 9240 rows and 37 columns. The data set contained information about the various leads that an Education Firm X Education had received for its courses. Each Lead was identified by a Lead Number and it had information about characteristics of lead such as - Source of the Lead, Origin of Lead, Demographical characteristics of prospect – country and City , personal traits and preferences of the prospect like - Occupation, Specialization and some personal preferences.

Also it was given that the data had Select as value for various columns which were to be treated as missing values. The data had a lot of null values. In the data cleaning part, we removed all the columns that had >35 % null values. We also removed all columns that had no variation and columns that had little variation. We also handled outliers and also

Then we visualized the various numerical (continuous) and categorical variables versus whether the lead was converted to understand what impact the continuous and categorical variables had on lead conversion. We then spilt the data into test and train sets such that tests had 30 % of data and train had 70 %.

Then we took train data set and we went through creation of dummy variables for the categorical variables and normalized the continuous variables. We got a total of 43 variables in all.

After this we ran RFE to select the top 30 variables for our model. We chose a higher number of variables because the data has a lot of gaps and to get better accuracy taking more variables made sense.

Once we chose the variable we used statsmodel GLM based Logistic Regression builder to build our model. Once we built our model, we analysed insignificant variables in model (p-value >0.5) and VIF >7) and removed them one by one and in each step rebuilt our model

After we reached a stage with all model variables having p-value <0.5 and VIF<7 we arrived at our final model.

We then evaluated our model on various parameters such as accuracy, sensitivity, specificity, precision , recall and analysed the metrics. We also plotted the ROC Curve to understand how much was the area under curve was 0.84. Then using Precision Recall Trade of we arrived at a cut-off of 0.5.

We then evaluated the same model on our test data. The model accuracy with test data was also found to be 81 %. We also analysed the sensitivity, specificity and ROC Curve for test data,

Learnings

The case study was essential to us in getting a hands on experience of solving a real life data science problem with Logistic Regression. We demonstrated how using Logistic Regression we can build a model to enable Prediction of Probability of Conversion of a Sales Lead given some dependent parameters. We also got to apply model evaluation metrics such as Sensitivity, Specificity, Accuracy, Precision, Recall, AUC, ROC etc. We also learned how we could tune our model parameters to enable a different prediction behavior depending on our needs.