Lead Scoring Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. Many professionals who are interested in the courses of land on their website and browse for courses. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. The company markets its courses on several websites and search engines like Google platform. When these people fill up a form providing their email address or phone number, they marked to be a lead. The company also gets leads through past referrals. Once these leads are potential, employees from the sales team start making calls, writing emails & connected with them via different channels etc. This process of leads get converted while most do not. The typical lead conversion rate at X education is around 30%. There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e. constantly communicating, educating the leads about the product etc.) in order to get a higher lead conversion. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need 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 has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study:

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. 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 Summary:

Step1: **Reading and Understanding Data**.

Read and analyse the data.

Step2: **Data Cleaning and Preparation**

Dropped the variables for high percentage of NULL values in them. This step also included imputing the missing values as and where required with median values in case of numerical

variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

Step3: **Data Analysis**

Exploratory Data Analysis of the data set started in this process & the data is oriented. There are 3 variables that were identified to have only one value in all rows. These variables were dropped.

Step4: **Creating Dummy Variables**

Creating dummy data for the categorical variables.

Step5: **Test Train Split**

In this process, it divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling

We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step7: **Feature selection using RFE**:

Using the Recursive Feature Elimination (RFE), we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
Finally, we arrived at the 15 most significant variables. The VIF's for these variables were also found to be good.
We then created the data frame having the converted probability values with an initial assumption that a probability value of more than 0.5 means 1 else 0.
Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.
We also calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand the reliability of model.

Step8: Plotting the ROC Curve

We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 89% which further solidified the reliability of the model.

Step9: Finding the Optimal Cut-off Point

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cut-off point. The cut-off point was found out to be 0.37
Based on the new value we could observe that close to 80% values were rightly predicted by the model.
We could also observe the new values of the 'accuracy=81%, 'sensitivity=80%', 'specificity=82%'.
Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step10: Computing the Precision and Recall metrics

We also found out the Precision and Recall metrics values came out to be 79% and 71% respectively on the train data set.
Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.42

Step11: **Making Predictions on Test Set**

Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 81%; Sensitivity=80%; Specificity= 82.2%.