LEAD SCORE CASE STUDY

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 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

Business Goal

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we 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%.

Below are the steps that we followed to find out the leads that are most likely to convert into paying customers:

1. As a first Step, we began with reading and understanding the data such as Null values, Outliers etc.
    - There were columns which had more than 45% null values so, it's better to drop them.
    - We have also dropped some derived and skewed columns, which won't help us in building the accurate model.
    - As there were also some outliers in the data, so it's better to not include those outliers in our analysis.
2. In the next Step, we need to see the trends in data by performing some univariate and Bivariate analysis.
    - API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable. Lead Add Form has more than 90% conversion rate but count of leads are not very high.
    - Leads spending more time on the websitehave more converted rate.
    - Leads having specialization in Finance, HR and Marketing have good converted rate.
    - Working Professionals going for the course have high chances of joining it.

3. Next, Model Building.

    - After creating dummy variables of the categorical columns, we have done the scaling and split the data set in two parts (test and train data sets)
    - Then we performed RFE to find best 25 columns for the model.
    - Then by continually performing the logistic model using GLM for checking P value and VIF for multi collinearity. We have started dropping columns depending on the P value and VIF value. Then final model was ready with 12 columns.

4. Next, Predicting the Model.

    - First, we have predicted the model on Train data set. After predicting the model, we have checked the confusion matrix and accuracy which comes around 80%.
    - Then we checked the ROC curve and plot the accuracy, sensitivity and specificity for various probabilities to check the cutoff value which comes as 0.29.
    - Finally, we have done the same tasks as scaling, add constant and predict the model on Test data set.

5. Verifying the model

    - The model does not over-fit, as there is not much difference between the accuracy of Test and Train model
    - The model is simple enough to be understood
    - The model is built using significant features.
    - The VIF value is under 5 & the p value is under 0.05 for each feature
    - The accuracy, sensitivity and specificity of our model after test are at least 80%(+- 3% between all 3 parameters)

Learnings:

- Identification of Junk/Null values – For e.g. In the given dataset, Missing values were present in the form of “Select” because of system design architecture. These values meant that the user had made no selection in those fields. So, we should handle such cases.
- Combine categories which are in very less percentage.
- It is important to drop columns having skewed data as it sways the model heavily towards its direction makes the model incapable of predicting the results correctly.
- We should fix the random state so that our final model variables would remain consistent every time we run it.