Lead Score Case Study Summary

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'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

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 a higher lead score have a higher conversion chance and the customers with a 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%.

Solution:

The following steps are used to analyze data and finding the solution,

1. Read and understand the data:

Data is uploaded and understood the characteristics of data, such as shape, number of missing values, types of variables etc.

2. Data Cleaning:

After checking all the null values, dropped the columns with high percentages of null values and imputed missing values with our logics.

3. EDA:

Did EDA to the data set and understood the variables with the help of data visualization. Identified outliers attached with some columns and treated the outliers according to logic. By analyzing, we found that some columns are not providing any such information for the analysis and we dropped these columns.

4. Creating Dummy Variables:

For categorical variables, we created dummy variables. This helped us to understand the data more.

5. Train – Test Split:

Then we split data in to two, Train set & Test set with a ratio of 70%-30%.

6. Scaling of data:

Then we used Standard Scaler to scale the data. Which helped us to convert the units of data in to a common scale.

7. Feature Selection Using RFE:

Using RFE we selected the columns with important features which are contributing more to the model analysis. And we looked for better significant model by checking p-values and VIF values. The features with high p-values and VIF values are eliminated to get significant model. At last we got 11 significant features that contributes more to the analysis.

Using this data we performed prediction on this training data set. And we looked for converted probability and selected arbitrary cut-off point to make prediction.

8. Making the Confusion matrix:

With these data we made confusion matrix and found the Accuracy, Sensitivity, Specificity.

9. Plotting ROC Curve:

By plotting ROC Curve we got the value for area under ROC curve as 0.80, which is good.

10. Finding optimal cut-off point:

By plotting Accuracy, Sensitivity, Specificity we got our optimal cut-off point is in the middle of 0.3 & 0.4. So we took 0.35 as our optimum cut-off point. Based on this we found Accuracy: 76.1 %, Sensitivity: 60.8%, Specificity: 85.8 %

11. Precision and Recall:

By computing precision and recall we got the values as,

Precision: 73.1% and Recall: 60.6%.

12. Prediction on Test model:

By using the above information, we made prediction on test set and got Accuracy: 75.5 %, Sensitivity: 57.3 % & Specificity: 86%. Also we found the hot leads IDs so that organization can concentrate on this specific leads to get them to converted.