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%.

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%.

Goals of the Case Study:

1. Build a logistic regression model to assign a lead score between 0 and 100 to ea ch of the leads which can be used by the company to target potential leads. A high er score would mean that the lead is hot, i.e., is most likely to convert whereas a l ower 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 sh ould be able to adjust to if the company's requirement changes in the future so yo u 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. Als o, make sure you include this in your final PPT where you'll make recommendation s.

Analysis Approach

Below are few of the considerations / assumptions we have made during this analysis:

- Many the features had a level called ‘Select’, we have imputed that with null values
- >35% of missing values have been dropped, we have also checked these features before dropping, we did not find them significant to our model
- We have created a new category (such as ‘Others’) in few features to combine lower valued data points or data points that had similar behaviour to avoid creating unnecessary dummy variables. We have also checked for duplicate data levels as well.
- Categorical variables with very less variance of data distribution (highly skewed data points) are either imputed or dropped based on their corresponding conversion rate.
- Numerical variables have gone through outlier treatment.
- Train and Test data sets are split with a ratio of 70-30, ‘Converted’ feature has become our ‘y’ variable while the rest are under X.
- We have used ‘Standard Scaler ()’ to rescale the numeric features which aids in the total accuracy of model building.
- With the help of heatmap we have checked for multicollinearity between the variables to remove highly correlated variables to improve further efficiency of model.
- Using RFE with an estimator of ‘Logistic Regression ()’ along with max of 20 steps we have further refined features.
- P-value has been checked for all features using the model summary, features with higher than 0.05 p-value have been dropped from the model, VIFs are also checked for models and are kept below 2.00
- Arbitrary cut off point of 0.5 was chosen initially which we then changed to 0.3 after plotting the sensitivity, specificity and accuracy plot along with ROC curve.
- We have used the Predicted probability *100 as the formula for lead score.
- Model is performing well with both test and train data sets, recall score or Sensitivity is above ~80% in both scenarios.

Conclusions:

- Top features that are impacting the conversion rate positively are:
    - Lead sources originating from 'Welingak Website'
    - Customers that are working professionals tend to join the courses compared to other occupations

    - Lead sources originating via 'Reference'

- If the customer has opted 'Do not email' option they are most likely uninterested in the program and we can avoid wasting time on such leads.
- There are few areas where we see room for improvement in terms of conversion rate:
    - Under the feature 'Lead Origin' we have categories such as 'Landing Page Submission' and 'API', here even though we have a decent amount of leads volume they are not getting converted efficiently, team can dive deep into this.
    - Lead source is an important feature during our initial EDA it is found that both 'Google' and 'Direct Traffic' is generating very similar amount of lead volume however the 'Direct Traffic' leads' conversion is not up to the mark, perhaps sales team can probably find an alternative approach to these leads.
    - Sales team also has to relook at leads that are sourcing through 'Olark Chat' to understand if the process is user friendly or not because conversion rate is low compared to its volume.