Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned 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%.

Solution Summary:

Step1: Reading and Understanding Data:

Initial data seems to have 37 columns & 9240 entries.

Most of the variables are categorical variables. Few Numerical variables are also given.

Step2: Data Cleaning:

- Variables with more than 20 percent null values are dropped. 11 variables were dropped.
- Columns with unique values are dropped as it does not contribute to the modelling.
- Categorical columns with does not have much variation in the values(Imbalance data) & its distribution are dropped as it does not help in model building. 12 such columns were dropped.
- Next, the rows with null values are removed.
- At last, 17% of data were during data cleaning. Remaining data will be used for Model building.

Step3: Data Transformation:

Changed the binary variables into '0' and '1'

Step4: Dummy Variables Creation:

- We created dummy variables for the categorical variables.
- Some variables have select option which means the lead does not give any details to those variables. It is almost same as null value. So, we can delete those data.
- Removed all the repeated and redundant variables

Step5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70- 30% values.

Step6: Feature Rescaling:

- Min Max Scaling is used to scale the numerical variables.
- Correlation heatmap is drawn to check the correlations among the variables.

Step7: Model Building:

- Using the Recursive Feature Elimination, 15 top important features were selected.
- Using the statistics generated, out of 15 features, the features with p-value<0.05 & VIF<5 were chosen for model building.
- Finally, 13 most significant variables were found. The VIF's for these variables were also found to be good.
- For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
- We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 86% which further solidified the of the model.
- Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.38.
- Based on the probability cutoff the below parameters are obtained.
    - Accuracy: 0.76
    - Sensitivity : 0.82
    - Specificity : 0.72
    - Precision : 0.69
    - Recall : 0.82

Step 8: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 76% on the final predicted model.
- Good value of sensitivity of our model will help to select the most promising leads.
- Features which contribute more towards the probability of a lead getting converted are:

    Total Time Spent on Website

    Lead Source_Welingak Website

    Lead Source_Reference