Lead Scoring Case Study Summary Report

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:

Approach: From above problem description we conclude that the above problem is the classification problem, hence we choose logistic Regression to calculate the Lead rate.

Step1: Reading and Understanding Data.

Read and analyse the data.

Step2: Data Cleaning:

We dropped the variables that had high percentage of NULL values in them and imputed the missing values as and where required with mode values in case of categorical variables. This step also identified features that are not required/useful for our analysis and dropped them.

Step3: Data Analysis

Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, we merged the low / no conversion sub categorical groups to one. In this step we also plotted the correlation matrix to identify the columns which are correlated.

Step4: Creating Dummy Variables

Then went on with the creation of dummy data for the categorical 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

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, we went ahead and selected the 15 top important features. In this step we made the model stable by using stats library, where we checked the p-values to be less than 0.05 and vif values to be under 5. Variance inflation factor(vif) is used to treat the multicollinearity. Finally, we arrived at the 12 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 and we had 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 how reliable the model is.

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 96% which further solidified the of the model.

Step9: Finding the Optimal Cutoff 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 cutoff point. The cutoff point was found out to be 0.55 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=91%, ‘sensitivity=90.6%’, ’specificity=92.2%’.

Step10: Computing the Precision and Recall metrics

we also found out the Precision and Recall metrics values came out to be 89% and 92.2% respectively on the train data set. Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.52

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 91.1%; Sensitivity=90%; Specificity= 92.3%.

Conclusion:

Top 3 variables that contributing to convert a lead are:

* TotalVisits * Tags_Lost to EINS * Lead Origin_Lead Add Form

Top 3 variables that need improvement to convert a lead are:

* Tags_switched off * Tags_Ringing * Last Notable Activity_Olark Chat Conversation

Learnings we gathered from this assignment are as below:

a. How to handle missing value and scaling the features in a data set.
b. How to create dummy variables/labels on categorical columns.
c. Functions to perform repetitive steps can help in building a modular code. This also help in reusability of the code.
d. How to use python libraries to perform logistic regression on selected features. (RFE is efficient technique to identify key features to start building model).
e. How to determine ideal optimal cutoff from the confusion metrics(Accuracy, sensitivity, specificity) and to choose best model based on balanced sensitivity and specificity.
f. Finally, how to solve a problem with team effort.