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:

1: Reading and Understanding Data.

2: Data Cleaning:

- •Handling ‘Select’ values in some columns
- •Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- •Dropping the columns having more than 40% as missing value
- •After checking for the value counts for some of the object type variables, we find some of the features do not have enough variance and were dropped, like, ‘Country’, ‘City’
- •Replacing NaNvalues with ‘Not Specified’ wherever it made business sense and combining low frequency values of some categorical variables together under ‘Others’
- •Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.

3: Outlier Treatment:

- •Remove top & bottom 1% of the Column Outlier values, wherever applicable

4: Dummy variable creation of categorical variables for modelling

- Splitting into binary variables for features with binary possible values
- One Hot Encoding for variables with multiple options

5: Exploratory Data Analysis:

- Performed numerical and categorical variable analysis to understand the general trend and relation between the variables

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

7: Feature Rescaling and Modelling:
We used the Standard Scaler to scale the original numerical variables.

8: Feature selection using RFE:

- Using the Recursive Feature Elimination, we selected the 20 top important features and created our base model. Using the Stats Model, we recursively tried looking at the P-values and VIF in order to select the most significant values that should be present and dropped the insignificant values.
- 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’ metrics to understand how reliable the model is.

8: 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 97% which further solidified the reliability of the model.

9: Finding the Optimal Cut Off 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 cut-off point. The cut-off point was found out to be 0.3.

Based on the new value we could observe that close to 92% values were rightly predicted by the model.

We could also observe the new values of the ‘Accuracy: 92.88%’, ‘Sensitivity: 91.82%’ and ‘Specificity: 93.53%’.

10: Computing the Precision and Recall metrics
We also found out the Precision and Recall metrics values came out to be 87% and 92% respectively on the train data set.
Based on the Precision and Recall trade off, we got a cut off value of approximately 0.4.

11: 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 92.88%, Sensitivity=91.69%; Specificity= 92.76%.

12: Conclusion

- Test set is having accuracy, recall/sensitivity in an acceptable range.
- In business terms, our model is having stability an accuracy with adaptive environment skills. This means it will adjust with the company’s requirement changes made in coming future.
- Top features for good conversion rate:
    - Tags_Closedby Horizzon
    - Tags_Lostto EINS
    - Tags_Willrevert after reading the email