Lead Scoring Case Study

By: Vipul Jadhav and Rohit Venkat

Summary

Problem Statement:

- We have started this case study with keeping the problem statement in mind. An education company named X Education sells online courses to industry professionals.
- X Education needs help in selecting the most promising leads that are most likely to get converted.
- The company needs a model where a lead score is assigned to a lead such that a customer with higher lead score is has a higher chance of conversion rate.
- The lead conversion should be 80%

Data Preparation:

- We have been provided with the data in “Leads.csv” file.
- We checked datatypes and missing values in the data. We have dropped columns having more than 45% missing values.
- Then we started handling missing values, in this step we have imputed missing values with respective categories for categorical columns and for numerical columns we had less % of missing values, so we dropped these.
- We did some feature engineering on some of the categorical variables, we have reduced the number of levels in the categorical column.
- Then we had to handle outliers in the numerical columns. For that we only kept values which are between 0.1% to 0.99%.
- Later we dropped all the un-necessary columns.

Exploratory Data Analysis:

- For EDA we started first with all the numerical columns.
- We saw the effect of all the numerical columns on conversion rate.
- We clearly saw the effect of “Total time spent on Website” column on the conversion rate.
- Then we saw the effect of all the categorical columns on conversion rate.
- We got some important insights from these plots. Few of them are mentioned below:
- For Lead Origin Maximum Conversion happened in landing Page Submission
- Major Conversion in Lead Source is from Google.
- Unemployed people have higher conversion rate.

Dummy Creation and Splitting the data into train and test:

- For Model building first we created dummy variables for all the categorical columns.
- Then we have dropped columns where information is not provided.
- After dummy variable creation we had to create dependent and Independent variables which is X and y.
- The next step is splitting data into Training and Testing set. We have used 70:30 ratio for train and test split.

Model Building:

- After splitting the data into train and test set, we started with building the model on the train data.
- Before building the model, it is important to scale all the numerical columns. For scaling we have used Standard Scaling technique.
- After scaling we started with building the model. We have used Recursive Feature Elimination (RFE) technique to get our top 15 variables.
- We have then checked significance of the variables i.e. p-value and also checked if there is any multi-collinearity i.e. VIF for all the variables.

- We have dropped variable with p-value more than 0.05 and VIF more than 5.
- In our final model, we have only 8 variables.

Model Evaluation:

- For model evaluation we first took 0.5 as the probability cut-off value.
- Later we calculated metrics like accuracy, precision and recall.
- For 0.5 cut-off value we got the model accuracy of 78.66%
- When we plotted graph for Accuracy, Sensitivity and Specificity, we found the optimal cut-off to be 0.3.
- But for 0.3 cut-off the model accuracy was still 79.48%. As per the requirement we needed 80% model accuracy. So we further adjusted the cut-off to 0.35.
- For 0.35 cut-off value our model was able give 80% accuracy.
- We've observed below values for our final model on the train set:

1. **Accuracy: 80%**
2. **Sensitivity: 72%**
3. **Specificity: 85%**
4. **Precision: 74%**
5. **Recall: 72%**

Making Predictions on the Test Set:

- Now its time to test our model. We will use our final model to make predictions on the test data.
- After using our model to predict on the test data we have observed the below values.

1. **Accuracy: 80%**
2. **Sensitivity: 74%**
3. **Specificity: 83%**
4. **Precision: 73%**
5. **Recall: 74%**

As we can see overall accuracy on both train and test set is 80%, this means our model is very good.