Lead Score – Case Study Summary

The analysis of any data or business problem has to begin with an understanding and being thorough with the data.

We started with going through and understanding what each variable in the data contains and signifies. The null values in the data were taken care of at the beginning, where all the columns with more than 35% null values were dropped outright. It was ascertained whether imputing the columns made sense or not, and those that seems reasonable and necessary for the analysis were retained and imputed aptly.

Then we moved onto the EDA step, looking at the impact of some key variables on the target column using various graphs. The conversion rates w.r.t like Lead source, origin, Employment status, country, etc were seen and appropriate inferences drawn about the data. For example, referrals turn into converted leads most of the times, but unemployed leads are less likely to convert.

After having catered to the outliers and encoding the categorical variables appropriately by the use of Utility Encoder from SkLearn, we moved onto dividing the data between Train and Test data, so that we could move onto modelling the data.

Recursive Feature Elimination was used to boil the number of variables down to 15. We used the Statsmodel library for modelling the data using Logistic Regression.

Post training the first model, the P – values and Variance Inflation Factors were noted and the variables with higher P – Values were removed and if the P – Values are alright, we removed the columns with high VIFs. This is so because VIF denotes the amount of correlation that a particular variable has with the rest of the columns or features. And if it high, then the variance in this variable can be easily explained by using just the rest of the columns, making this particular variable redundant.

After 7 such iterations, we arrived at an apt enough model with good P and VIF values, we decided to use it.

Post model selection the threshold for the cut-off has to be carefully chosen, because in a binary classification problem like this one, the threshold value can cause a great impact on metrics like Specificity, Sensitivity, Positive rates and the values of the Confusion Matrix. After having calculated Accuracy, Sensitivity and Specificity of all the predictions made with all of the Cut-offs from 0.1 to 0.9, it was clearly visible that the 0.38 was a good cut-off point.

Using this threshold, we moved ahead to make the predictions and achieved the desired result of approx. 80% target, which reinforced of the model being good. After making the predictions on the test set and arriving at similar results we got the below mentioned metrics from the model

Accuracy – 74.1%

Sensitivity – 79.7%

Specificity – 70.8%

Precision – 60.9%

Recall – 79.7%

Submitted by –

Harshit Goyal

Chetna Joshi

Ankit Kumar