Lead Scoring Case Study

Building Logistic Regression Model

Submitted BY-:
Aditya Anand (adityaanand47@gmail.com)
&
Hrithik Chand (hrithiknovember@gmail.com)

On
12 November 2021

Why we are doing the case study?

The analysis is done for X Education to focus on those leads who are more likely to join the course. This in return increase the profitability of company and helps in saving expenditure. This is done by building a logistic regression model where we pass a complete data (demographic, Professional etc.) of lead and will filter the group of lead into potential and non potential leads and hence the company will only focus on potential leads rather than wasting time on non potential leads. It help in saving time and making the company to generate more revenue in less time and spending less effort.

The steps which we follow to do the aforesaid requirement.

Step 1: Data Cleaning and Imputing missing Value

The dataset were cleaned by dropping the unnecessary columns, columns having unique value and columns which have more than 40% null values. For categorical feature, missing value is imputed by mode. For columns which have very less null values handled by dropping the rows contains null values. Outlier in numerical feature handled by dropping the value greater than 99 percentile. While exploring the data, we noticed some feature contains value as 'Select' which means the data is not provided by the lead and this is handled by replacing the value 'Select' by null value.

Step 2: Data Preparation & Analysis

Before directly jumping into model building we need to prepare the data for model. This is done by converting the binary variable to 0/1 and dropping columns which have very skewed data.

After preparation, we have done the EDA on all feature with respect to our target variable to see the influence on dependent variable on target variable.

The result of EDA was:

Conversion rate is approx. 38% which means our data has not imbalanced. Features like Lead origin, Lead Source, Do not email, Last Activity, Specialisation, Current occupation, Tags, Last notable activity and total time spent has strong influence on target variable.

Step 3: Model Building & Evaluation

First we have created dummy variables for all categorical feature, then we split the data into train & test in 70-30 ratio. After doing the initial steps we have scaled the numerical feature and started our model building process. As we started with large number of dependent feature and if we build model using all feature then it means we are over training our model and our model accuracy fails in test set. To get over this we have selected top 15 features contributing for our model using RFE (Recursive Feature Elimination) also known as coarse tuning. After selecting top 15 feature we started analysing each feature statistically and manually checking the significance of each feature and dropped insignificant feature and those who is having high VIFs (Variation Inflation Factor) as these feature is pretty well explained by other feature. Then at-last after doing manual and coarse tuning we have our best fitted model with us.

After building model, we started the prediction on test set with some randomly assigned cut off probability. To find the optimal cut off we used the ROC curve & printed the data-frame which contain accuracy, sensitivity & specificity against a set of cut off probability. After observation, we noticed that at 0.3 cut off we have all the parameters at its peak value so we take 0.3 as our cut off probability. After taking 0.3 threshold we made predictions train set and check the accuracy, sensitivity and specificity. Then we started our prediction on test set and again check all the evaluation parameters. Based on the result obtain our model seems to be a good fit model with high accuracy.

Train Set: Accuracy : 88.63% Sensitivity : 88.08% Specificity : 88.97%

Test Set: Accuracy : 88.46% Sensitivity : 88.43% Specificity : 88.47%

At-last we also check the precision-recall for our model, Precision : 83.04%

Recall : 88.08%

Step 4: Final conclusion:

Most important feature for model :

- The total time spend on the Website.
- When the lead origin is Lead add format.
- When the current occupation was:
  - Unemployed
  - Working Professional
- When the last notable activity was:
  - SMS sent
  - Olark chat conversation

By focusing on these variable X Education Forms can grow and convert most lead.