

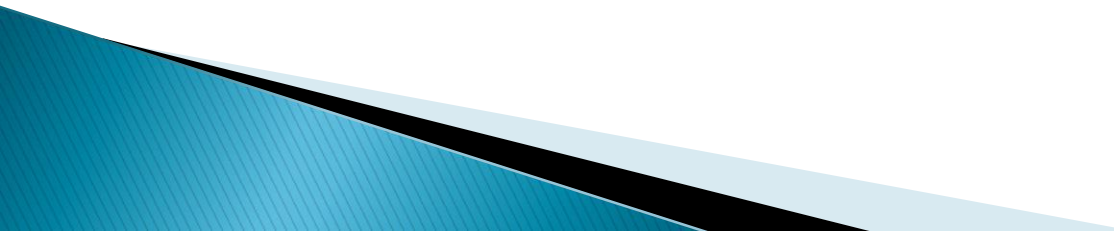
LEAD SCORING CASE STUDY

LOGISTIC REGRESSION

PROBLEM STATEMENT

- ▶ The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

GOALS OF THE CASE STUDY

- ▶ **Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.**
- 

CLEANING DATA

The data was partially clean except for few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to “not provided” so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to “India”, ‘outside India’ and ‘not provided’ .

EDA

- ▶ A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.

DUMMY VARIABLES

- ▶ The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler

TRAIN TEST SPLIT

- ▶ The split was done at 70% and 30% for train and test data respectively.

MODEL BUILDING

- ▶ Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and pvalues.

MODEL EVALUATION

- ▶ A confusion matrix was made. Later on the optimum cut off value(using ROC curve)was used to find the accuracy, sensitivity, and specificity which came to be around 80% each.

PREDICTION

- ▶ Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy sensitivity and specificity of 80%.

PRECISION RECALL

- ▶ **This method was also used to recheck and a cut off of 0.41 was found with precision around 73.24% and recall around 76.61% on the test data frame.**

FINAL OBSERVATION

- ▶ Let us compare the values obtained for Train & Test:
- ▶ Train Data: Accuracy : 92.29%
- ▶ Sensitivity : 91.70%
- ▶ Specificity : 92.66%
- ▶ Test Data:
- ▶ Accuracy : 92.78%
- ▶ Sensitivity : 91.98%
- ▶ Specificity : 93.26%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model