

## LEAD SCORING SUMMARY REPORT

By Chandan Singh, Diwakar Mishra and Shreeharsh Pande

---

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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%.

So, the company wants to know:

- Leads that are most likely to convert into paying customers.
- Expecting the target lead conversion rate to be around 80%

We have used the following steps in order to prepare this logistic regression model:

### **Step 1: Loading and Cleaning Data**

- Import Leads.csv file and read the dataset
- Inspection of the dataframe
- Cleaning the dataframe
- Check the Null values, Shape, other columns/rows and duplicate rows

**Explanation:** We dropped the variable that a percentage more than 35% of NULL values in them. This step also includes imputing the missing values as and checking for duplicates. Few of the null values were changed to 'not provided' so as not to 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'.

### **Step 2: EDA**

- Univariate Analysis
- Working with Numerical Data
- Dropping columns which are not relevant
- Univariate and Bivariate Analysis and their visualisations
- Handling and checking outliers

**Explanation:** We have used EDA in order to check the condition of our leads data and found lot of elements are irrelevant which belongs to categorical variables. However numerical variables were looks good. It is understandable from the above EDA that there are many elements that have very little data and so will be of less relevance to our analysis.

### **Step 3: Creating Dummy variables**

**Explanation:** The dummy variables were created and later the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

#### **Step 4: Split train-test and Scaling**

**Explanation:** The split was done at 70% and 30% for train and test data respectively.

#### **Step 5: Feature selection and Model building**

- Extract top n features with RFE
- Improve the model further inspecting adjusted R-squared, VIF and p-values
- Build final model
- validate linear regression assumptions, residuals

**Explanation:** 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 p-value (The variables with  $VIF < 5$  and  $p\text{-value} < 0.05$  were kept).

#### **Step 6: Predicting target variable using final model**

#### **Step 7 : Model evaluation**

- Confusion matrix
- Sensitivity
- Specificity

**Explanation:** 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.

#### **Step 8 : Optimise Cut off (ROC Curve)**

**Explanation:** The area under **ROC curve is 0.88** which is a very good value.

#### **Step 9 : Prediction on Test set**

- Confusion matrix
- Sensitivity
- Specificity

**Explanation:** Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

#### **Step 10 : Precision-Recall**

- Precision and recall tradeoff

#### **Step 11 : Final Prediction on Test set**

**Explanation:** This method was also used to recheck and a **cut off of 0.41** was found with **Precision around 73%** and **recall around 76%** on the test data frame.