

Leading Score Summary

Problem Description

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.

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%.

X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. **A model is required to be built wherein a lead score is assigned** to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead **conversion rate to be around 80%**.

From above problem description we conclude that the above problem is the classification problem, hence we choose logistic Regression to calculate the Lead rate.

Below are the steps followed to solve this problem

1. Data Reading and Understanding:

Here we tried to get the look and feel of the data, we observed following things

- Number of rows and columns
- Data types of each columns
- Checking first few rows how data looks
- Checking how the data is spread.
- Checking for duplicates, if any.

2. Data Cleaning:

Here we checked for discrepancies in the dataset

- Checking for any column names correction
- Checking for null values and imputing them with appropriate methods

3. Data Visualization and Outliers Treatment:

- We performed univariate analysis on categorical column to see which columns makes more sense and removed those columns whose variance is nearly zero.
- We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
- We performed univariate analysis on numerical columns by plotting box plots to see are there any outliers in the data or not.
- We performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.
- We have used IQR method to treat the outliers in the data set.
- In this step we also plotted the correlation matrix to identify the columns which are correlated.

4. Feature Scaling

At this stage our data was very clean and no outliers. We know that logistic regression takes the input parameters as numerical values. Hence, we converted all the categorical columns to numerical.

Now, the data contained only numerical columns and dummy variables. Before proceeding for model building, we have rescaled all numerical columns by using standard Scaler method.

5. Model Building

We have used Recursive Feature Elimination Technique to remove attributes and built a model on those attributes that remain. RFE uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute

After finalizing the optimum cutoff and calculating the metrics on train set, we predicted the data on test data set. Below are the observations:

Train Data:

- *Accuracy : 76 %*
- *Sensitivity : 84 %*
- *Specificity : 69%*

Test Data:

- *Accuracy : 75 %*
- *Sensitivity : 80 %*
- *Specificity : 70 %*

6. Final Observations

The Model seems to predict the Conversion Rate very well. We should be able to help the education company select the most promising Leads or the Hot Leads.