

Lead Scoring case study Summary

Problem Statement:

X Education offers online courses to industry professionals. X Education needs help in selecting the prospective leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein you 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 leads are then contacted by the sales team to convert them into customers. The typical lead conversion rate at X Education is around 30%. The CEO, in particular, has given a target of lead conversion rate to be around 80%.

Approach:

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:

1. Data Understanding and Cleaning:

- Number of rows and columns and checking data spread
- Data types of each columns
- Checking for duplicates, if any.
- Checking for any column names correction
- Checking for null values and imputing them with appropriate methods

2. Data Analysis:

Started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

3. Creating Dummy Variables:

we went on with creating dummy data for the categorical variables. Removed all the repeated and redundant variables

4. Test-Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

5. Feature Rescaling:

We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

6. **Model Building:**

- Using the Recursive Feature Elimination we went ahead and selected the 15 top important features.
- Using the statistics generated, we tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
- Then, checked if 80% cases are correctly predicted based on the converted column.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
- Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.44.
- Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 78.7%; Sensitivity=79%; Specificity= 78%.

Conclusion:

- The most numbers of leads are from India and in terms of city highest number are from Mumbai.
- First, sort out the best prospects from the leads you have generated. 'Total Visits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.
- The leads are joined course for Better Career Prospects, most of having Specialization from Finance Management. Leads from HR, Finance & marketing management specializations are high probability to convert.
- Talking to last notable Activity, making improvement in customer engagement through email & calls will help to convert leads. As the leads which are opening email have high probability to convert.