

LEADS SCORING CASE STUDY

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 has given a ballpark of the target lead conversion rate to be around 80%

Approach:

The following technical steps are used: -

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 null values and imputing them with appropriate methods
 - We used mode imputation for categorical columns.

- We used median imputation for numerical columns, if there is skewness in the data.
- We have dropped the columns which are highly imbalanced/skewed.
- We have grouped the values of columns with low frequency into “Others”.

3. Data Visualization and Outliers Treatment:

- We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
- 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/heatmap to identify the columns which are correlated.
- We have detected the outliers and treated them by removing the value above 99%ile and below 1%ile.

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.

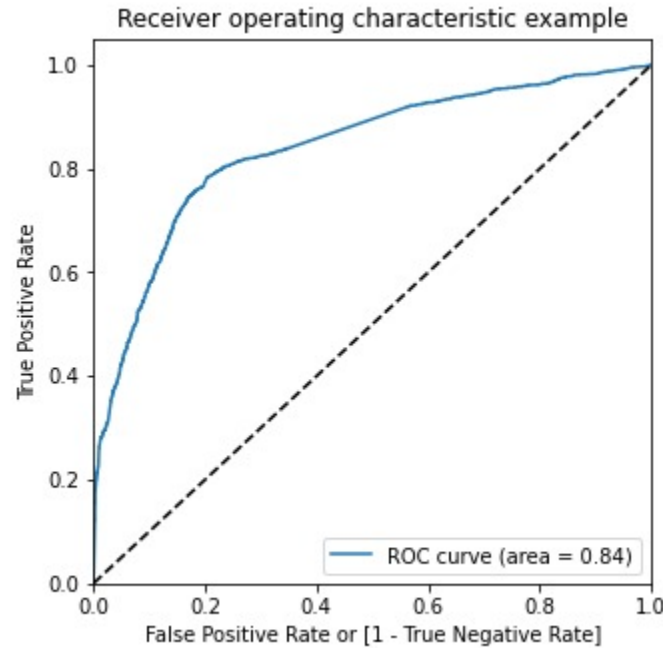
- Columns which have only two levels “Yes” and “No” were converted to numerical using binary mapping.
- Columns which have more than two levels were converted to dummies using `pd.get_dummies` function.

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. In this step we made the model stable by using stats library, where we checked the p-values to be less than 0.05 and vif values to be under 5. Variance inflation factor(vif) is used to treat the multicollinearity.

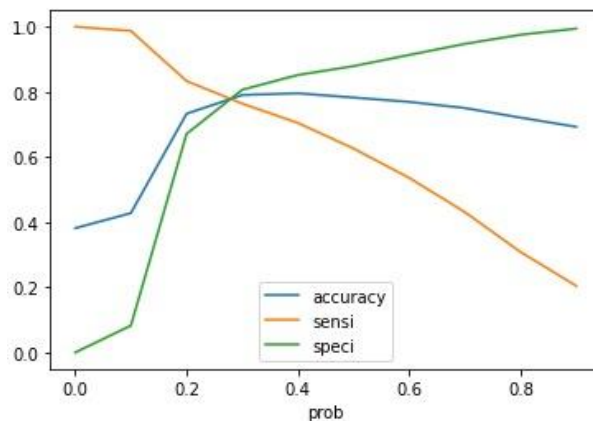
Once the stable model was created, we predicted probabilities on the train set and created a new column predicted with 1 if probability is greater than .5 else 0.

We calculated the confusion matrix on this predicted column to the actual converted column. We also calculated the metrics sensitivity, specificity, precision, recall and accuracy. We also plotted roc curve to find the area under the curve.



6. Model Evaluation on Train Set:

- In the step 5 we took 0.5 as the cut-of. To confirm that it was the best cut, we calculated the probabilities with different cut-offs.
- With probabilities from 0.0 to 0.9, we calculated the 3 metrics -accuracy, sensitivity and specificity.
- To make predictions on the train dataset, optimum cutoff of 0.25 was found from the intersection of sensitivity, specificity and accuracy as shown in below figure:



- 7. Predictions on Test Dataset:** 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: 77.76%
- Sensitivity: 80.51%
- Specificity: 76.06%
- Precision: 67%
- Recall: 80%

Test Data:

- Accuracy: 77.66%
- Sensitivity: 79.47%
- Specificity: 76.50%
- Precision: 68.49%
- Recall: 79.47%

8. 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.

Below variables helps in driving the hot leads-

Total Time Spent on Website
Lead Origin_Lead Add Form
Lead Origin_Lead Import
What is your current occupation_Student
What is your current occupation_Working Professional
Specialization_Banking, Investment And Insurance
Specialization_Business Administration
Specialization_E-COMMERCE
Specialization_International Business
Specialization_Management Specializations
Specialization_Media and Advertising
Lead Source_Direct Traffic
Lead Source_Google
Lead Source_Organic Search
Lead Source_Referral Sites
Lead Source_Welingak Website