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%.**

Approach:

From the above problem description, we conclude that the above problem is the classification problem, hencewe 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 the following things

- Number of rows and columns
- Data types of each column
- Checking the first few rows of 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
  - ✓ We used mode imputation for categorical columns.
  - ✓ We used mean imputation for numerical columns, if there is no skewness in data.
  - ✓ We used median imputation for numerical columns, if there is skewness in the data.

3. Data Visualization and Outliers Treatment:

- We performed univariate analysis on a categorical column to see which columns make moresense 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 whether there were any outliers in the data.
- We performed bivariate analysis on numerical columns with Converted columns to see howthe leads are related to these columns.
- We have used the 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

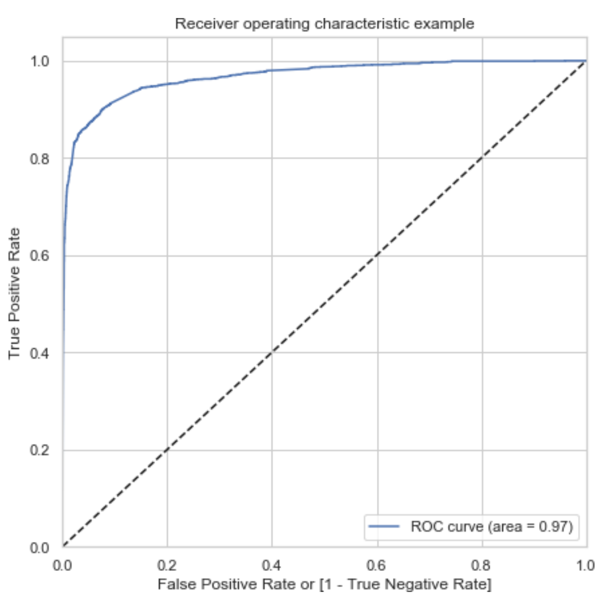
At this stage, our data was very clean and had no outliers. We know that logistic regression takes theinput parameters as numerical values. Hence, we converted all the categorical columns to numerical ones.

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

Now, the data contained only numerical columns and dummy variables. Before proceeding with model building, we rescaled all numerical columns by using the 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.



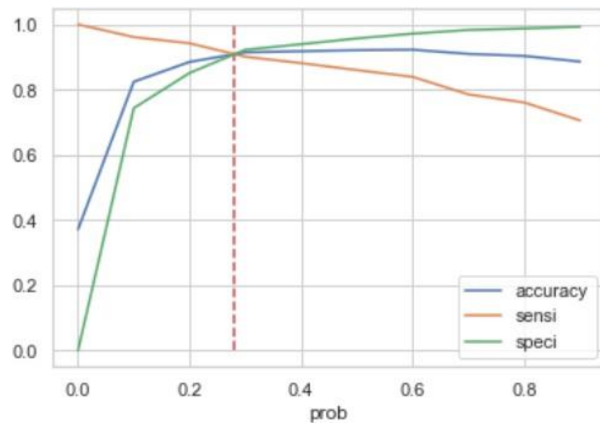
In this step, we made the model stable by using stats library, where we checked the p-values to be less than 0.05 and vif values to be under 5. Variance inflation factor(vif) is used to treat the multi-collinearity.

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.

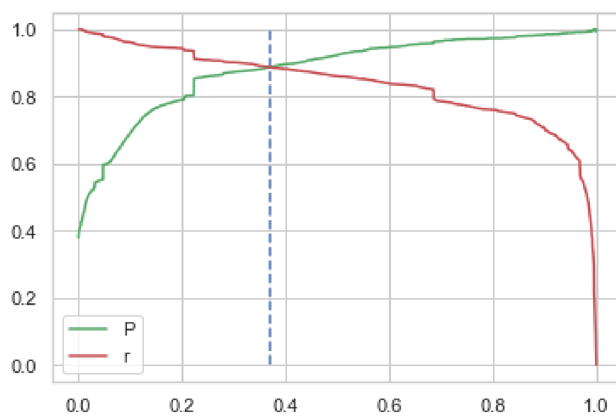
We calculated the confusion matrix on this predicted column to the actual converted column. We also calculated the metrics sensitivity, specificity, precision, recall, and accuracy. We also plotted the roc curve tofind the area under the curve.

6. Model Evaluation on Train Set

- In step 5 we took 0.5 as the cut-off. 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, an optimum cutoff of 0.28 was found from the intersectionof sensitivity, specificity, and accuracy as shown in below figure:



- To make predictions on the test dataset, optimum cutoff was considered as obtained from Precision recall graph of the train dataset as shown below figure:



- We can observe that 0.37 is the tradeoff between Precision and Recall. Thus we can safely choose to consider any Prospect Lead with a Conversion Probability higher than 37 % to be a hot Lead

7. Predictions on Test Set

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

Train Data:

- Accuracy: 78.57%
- Sensitivity: 81.02%
- Specificity: 77.06%

Test Data:

- Accuracy: 77.67%
- Sensitivity: 80.18%
- Specificity: 76.02%

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