Leading Score Summary

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 has appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score 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:

Inferring from the problem description that the problem at hand is a classification problem, we choose using logistic regression to calculate the lead rate.

The steps taken to resolve this issue are listed below.

1. Data Reading and Understanding:

When attempting to comprehend how the data seemed and felt, the following was what we observed:

- The number of rows and columns
- Data types for each of the columns
- Checking the data's appearance in the first few rows and its distribution.
- Checking for any duplication.

2. Data Cleaning:

Here, we checked the dataset for discrepancies.

- Examining any necessary corrections to the column names
- Identifying the null values and imputing them using the appropriate method
  - For categorical columns, mode imputation was used.
  - In numerical columns where there is no skewness in the data, mean imputation was applied.
  - If the data are skewed, we performed median imputation for numerical columns.

3. Data Visualization and Outliers Treatment:

- To determine which category columns made the most sense, we used univariate analysis. We then eliminated any columns whose variance was close to zero.
- To see how categorical columns differ from the Converted column, we performed bivariate analysis on the data.
- To determine whether there were any outliers in the data, we performed univariate analysis on numerical columns by drawing box plots.
- To determine how the leads are related to these columns, we conducted bivariate analysis on numerical columns with the Converted column.
- In this phase, we plotted the correlation matrix to see which columns are correlated and applied the IQR approach to deal with the outliers in the data set.

4. Feature Scaling:

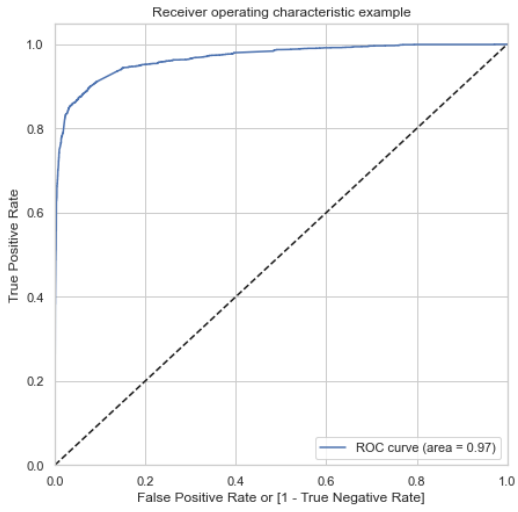
We had no outliers at this point, and our data was very clean. We are aware that the input parameters for logistic regression are numbers. Therefore, we changed the columns that had categories to numbers.

- Binary mapping was used to transform columns with just the "Yes" and "No" options to numbers.
- pd.get dummies were used to convert the columns with more than two levels into dummies.

The data now only included numerical columns and dummy variables. All numerical columns were rescaled using the standard Scaler approach before we started developing the model.

5. Model Building:

Recursive feature elimination technique was used to eliminate features, and a model was then constructed using the attributes that remained.



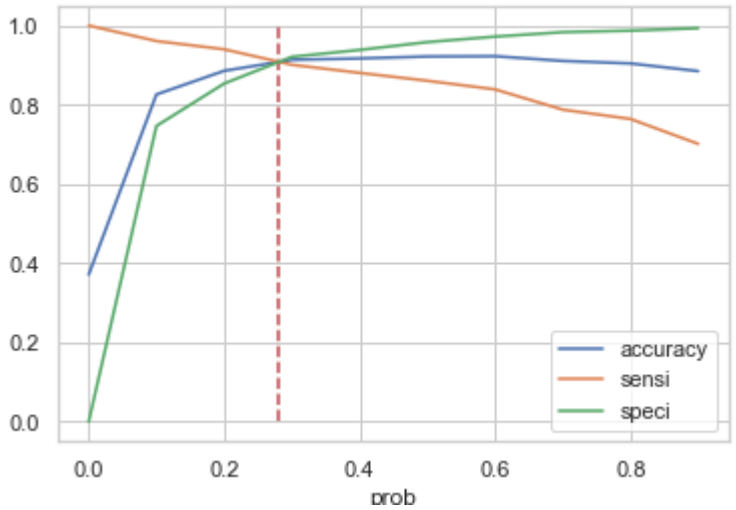
RFE uses the model accuracy to determine which qualities (and combinations of attributes) are most helpful in predicting the target attribute.

Using the stats package, we stabilised the model in this stage by ensuring that the p-values and VIF values were both less than 0.05. To address the multi-collinearity, the variance inflation factor (VIF) is applied.

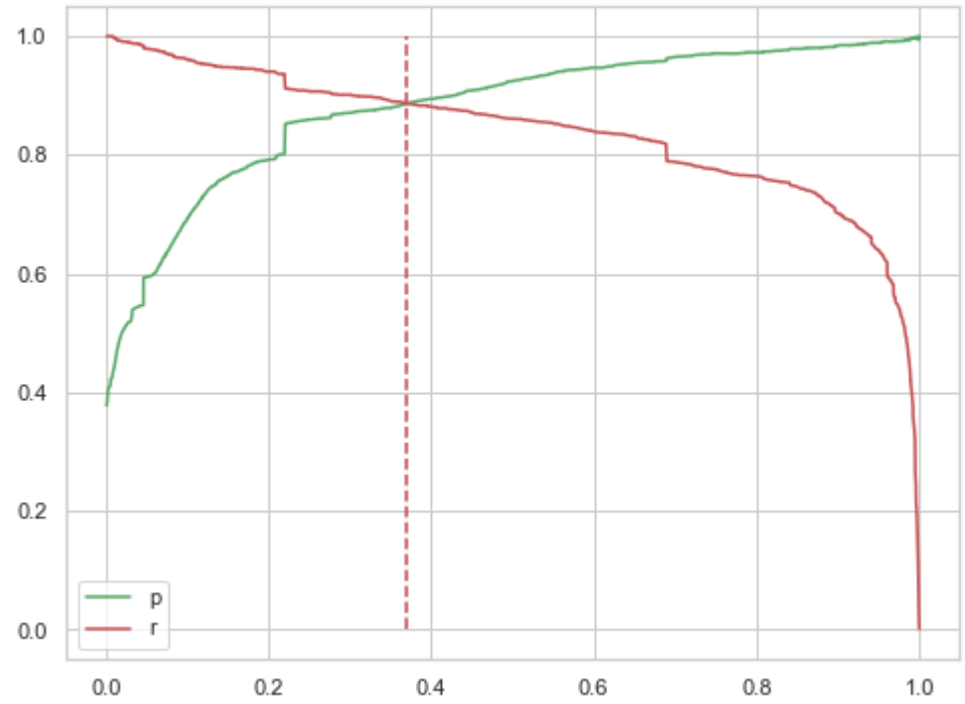
We predicted probabilities on the train set after creating a stable model, and we created a new column predicted with 1 if probability is greater than.5, otherwise 0.

6. Model Evaluation on Train Set:

- In step 5, we used 0.5 as the cut-of value. We estimated the probabilities using several cut-offs in order to verify that it was the optimal cut.
- We calculated the three metrics—accuracy, sensitivity, and specificity—with probabilities ranging from 0.0 to 0.9.
- The intersection of sensitivity, specificity, and accuracy yielded an optimal cut-off of 0.28 for predictions on the train dataset, as shown in the figure below:



- The best cut-off was considered while making predictions on the test dataset based on the precision recall graph of the train dataset, as illustrated in the image below:



- We observe that the trade-off between precision and recall is 0.37. Therefore, it is feasible for us to say that any Prospect Lead with a Conversion Probability of more than 37% is a hot Lead.

7. Predictions on Test Set:

We predicted the data on the test data set after deciding on the ideal cut-off and calculating the metrics on the train set. The observations are as follows:

Train Data:

- Accuracy = 91.25%
- Sensitivity: 90.23%
- Specificity: 91.85%

Test Data:

- Accuracy: 92.06%
- Sensitivity: 88.95%
- Specificity: 94.04%

8. Final Observations:

The Conversion Rate appears to be accurately predicted by the Model. We should be able to assist the educational company in choosing the Hot Leads or the Leads that have the best chance of success.

Driving the hot leads is helped by the below variables.