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 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. Cleaning data: The data was cleaned by replacing the 'select' value by Null and then removing the columns having Null values > = 30%. After that, Remaining columns which had high percent of null values, the nulls were replaced by 'Not provided' . After that, those columns which had only one category dominance and those columns which were not important as per business understanding were dropped.

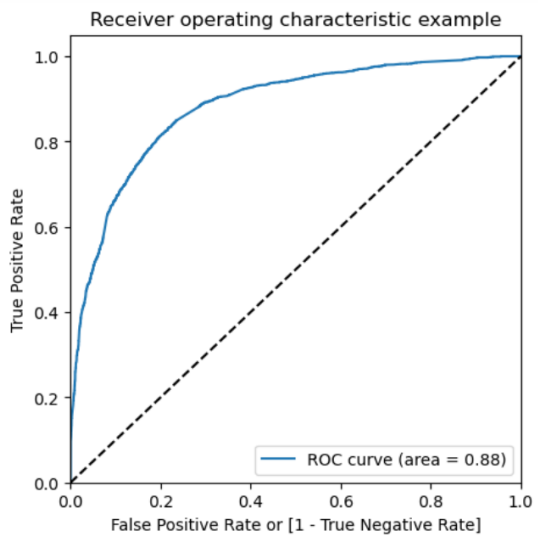
2. EDA - univariate and Bivariate analysis for all the continuous and categorical variables was done. Following inferences were made which indicated higher conversion rate-

    Customers with Lead origin - 'landing page submission' Lead Source- google
    Unemployed category

3. Data preparation- The dummy variables were created and the corresponding original columns were dropped. The data was splitted into test and train in the ration 70-30 % . The continuous columns were scaled using min-max method.

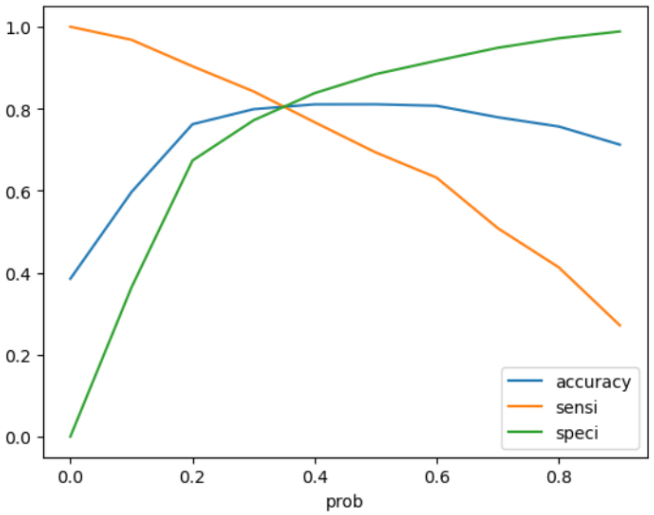
4. Model Building- Using rfe, the 15 features were selected and then first model was built on them. After 4 iterations of removing features with high p value and vif value, the final model was selected with accuracy of 0.88

5. Model Evaluation- Different metric like sensitivity, specificity and precision were computed , first by fixing the cutoff at 0.5 and 0.36 later.



The ROC Curve should be a value close to 1. We are getting a value of 0.88 indicating a good predictive model.

To make predictions on the train dataset, optimum cutoff of 0.38 was found from the intersection of sensitivity, specificity and accuracy as shown in below figure.



From the curve above, 0.38 is the optimum point to take it as a cutoff probability.

Conclusion

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 : 81.0 %

Sensitivity : 75.1%

Specificity : 84.7%

Precision Score: 75.5%

Recall Score: 75.1%

Test Data :

Accuracy : 81.0%

Sensitivity : 74.0%

Specificity : 85.0%

Precision Score: 73.8%

Recall Score: 74.0%