Lead Scoring Case Study

X Education Sales Analysis

PROBLEM STATEMENT

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%.

Goal

Goal is to help X Education Company 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 a higher lead score have a higher conversion chance and the customers with a 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%.

Goal is to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Overall Approach

Step 1: Importing the dataset

Import the dataset "**Leads.csv**" provided into the Jupiter notebook, the dataset comprises of 9247 rows and 37 features.

Step 2: EDA

Performing Exploratory Data Analysis on the dataset provided which includes the below

- Null Handling: Features that contain null values more than 35 % have been dropped, also rows where there were considerable count of nulls have been dropped to clean the data
- Dropping Insignificant features : Features which don't add any value to our analysis and also features that have only one value dominating have been dropped

Overall Approach

Step 3: Data Preparation

- Dummy Variables: Dummy variables were created for categorical features
- Binary mapping (0 and 1) has been done for the features with binary flags(”No” and “Yes”)
- In the final steps, features used for Dummy variable creation have been dropped.

Step 4: Test – Train Split

- Data has been split into Test and Train data sets considering the feature “Converted” as the target variable, following are the parameters used

  train_size=0.7

  test_size=0.3

  random_state=100

Overall Approach

Step 5: Feature Scaling

Feature Scaling has been done on the continuous variables to bring them into same scale, MinMaxScaler has been used for the Scaling.

Step 6: Checking Correlations

Correlations have been checked among the variables of the dataset.

Step 7: Model Building

Initial training model has been built using all features available in the train dataset and key factors have been observed

Step 8: Feature Elimination using RFE

Initial feature elimination has been done using RFE, and below steps are followed

- Train model has been fitted with the features selected by using RFE
- Features are eliminated step by step ,checking p-value and VIF and models are fitted after eliminating each feature.
- Arrived at the final train model where all features have p-value <0.05 and VIF < 5

Overall Approach

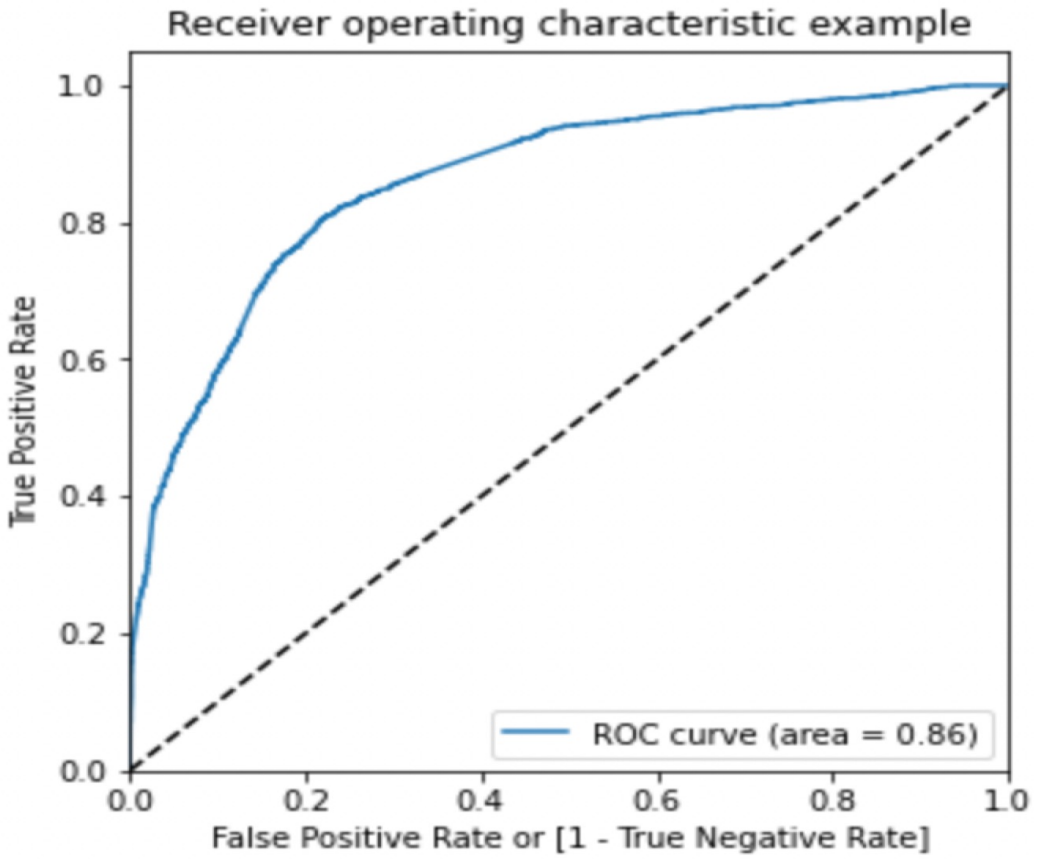
Step 9: Model Evaluation

- Probability of the target have been predicted using the final train model.
- Taking an arbitrary cut-off value of 0.5 ,predicted values have been calculated using probabilities
- Confusion matrix has been built to check the True Positive , True Negative, False Positive and false Negative values.
- Accuracy of the model has been calculated which was found to be 79%.
- Metrics beyond simple accuracy like Sensitivity, Specificity , Precision, Recall have been calculated

Overall Approach

Step 10: ROC Curve:

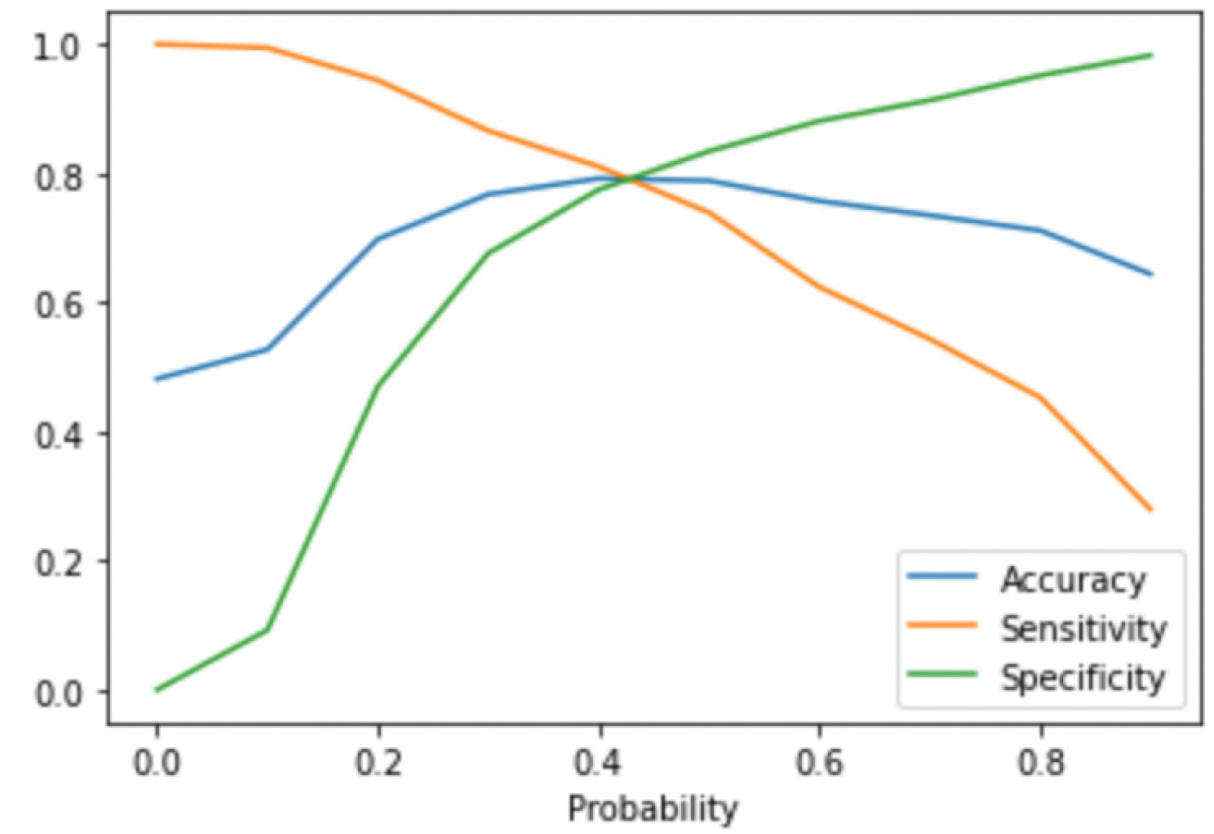
ROC curve has been plotted to calculate area under the curve which was found to be 0.86 and was good enough.



Overall approach

Finding Optimal Cut-off

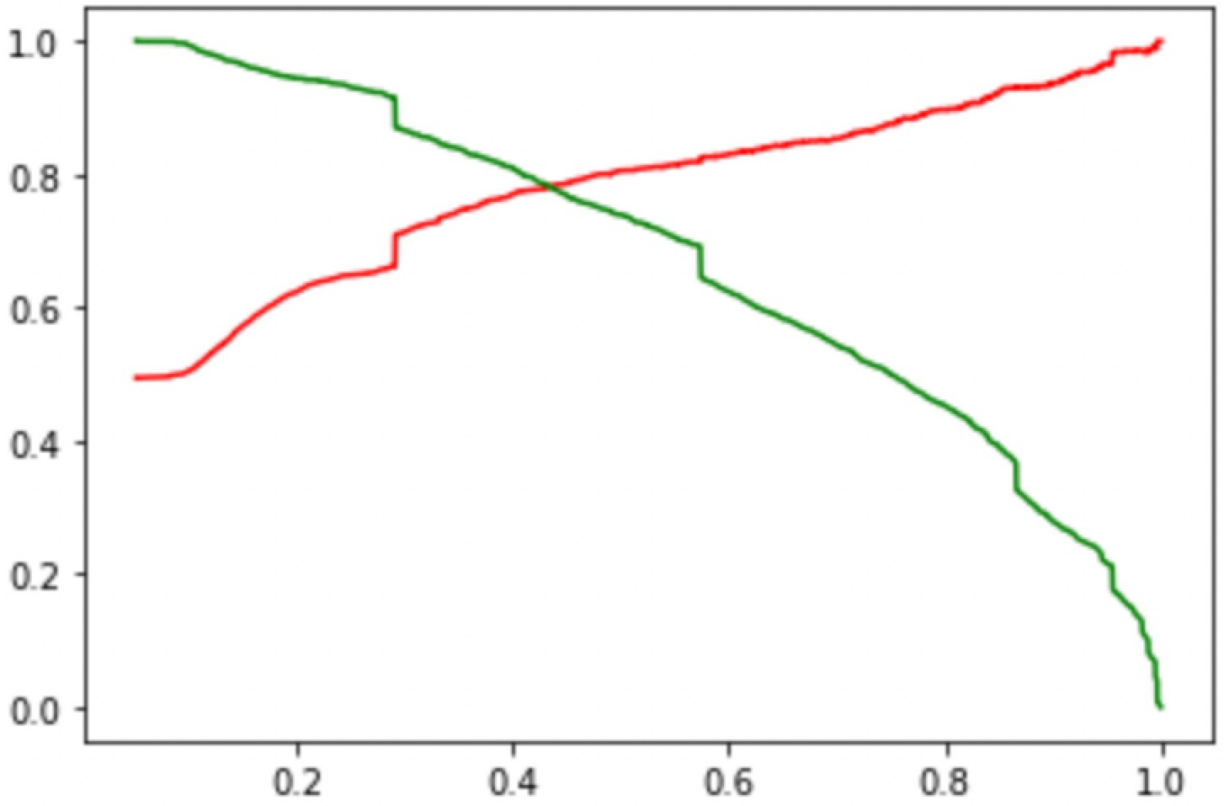
Optimal cut-off has been arrived at 0.42 by plotting Accuracy, Sensitivity and Specificity against various probabilities between 0 and 1.



Overall Approach

Precision and Recall Trade-off

Optimal cutoff 0.42 was verified by plotting the Precision and Recall curves against different Thresholds.



Overall Approach

Step 11: Predictions on the test set

- Feature Scaling on the test set
- Probability prediction on test set using final train model
- Calculating predicted values using threshold 0.42 on probabilities of test set.
- Calculating metrics like Confusion matrix, Accuracy, Sensitivity, Specificity, Precision and Recall on test set.
- Accuracy on test set was found to be 78.45% which was close to that of train dataset.

Final Results

- The below features were found to be contributing mor towards lead getting converted
- Do not Email= 0
- TotalVisits
- Total Time Spent on Website
- Lead Origin = “Lead Add Form”
- Lead Source = “Olark Chat”
- Lead Source= “Welingak Website”
- Last Activity = “Had a Phone Conversation”
- Last Activity = “SMS Sent”
- What is your current occupation <> Student
- What is your current occupation <> Unemployed
- Last Notable Activity = “Unreachable”

Recommendations

- Professionals who are not students, who are employed, who frequently visit the website and spend time on the website must be targeted.
- Leads whose Origin is ‘Lead Add Form’ and Sources are ‘Olark Chat’ and ‘Welingak Website’ tend to convert more.
- Can contact the Leads using Phone, SMS and Email as it was found that those with Last Activity was ‘Had a phone conversation’, ‘Sent SMS’ also who have set Do not Email option as ‘No’ tend to get converted.

Thank You!