Lead Scoring Case Study

By Namitha Murugesh and Sunila R K

Problem Statement

- X education sells online courses to industry professionals.
- These course are marketed on several websites and search engines like Google.
- When people fill up a form providing their email address or phone number, they are classified to be a lead.
- Sales team contacts these leads through emails and calls aiming to convert them to customers.
- The typical lead conversion rate at X education is around 30%
- How to identify “HOT LEADS” by giving them a score

The Goals

1. Build a logistic regression model to assign a lead score between 0 and 100, higher the score more successful will the sales team be if it pursues this lead.
2. Top three variables which contribute most towards the probability of a lead getting converted
3. Top 3 categorical/dummy variables in the model.
4. Strategy for the company to minimize phone calls after it reaches its target for a quarter before the deadline.

Data Inspection and Cleaning

1. Data set has 36 columns and 900 records
2. Many columns have null values and ‘Select’(considered as equivalent to null)
3. Columns with more than 40% null values are dropped. Others are dealt with individually either by imputing them with median or mode.
4. Outliers are soft capped.
5. Columns with ‘Yes’ and ‘No’ are converted to 0 and 1
6. Skewed Columns are dealt with by combining the categorical values into one ‘Combined’ Value
7. Dummy variables are created for Categorical Values using one hot encoding.

Logistic Regression Model

Using the logistic regression model the leads can be given a score.

- Data is split into train and test data with 70 to 30 ratio
- Data is scaled using a standard scaler
- The correlation are checked to drop the highly correlated columns
- Recursive Feature Elimination is used to select features.
- The models are bulit repeating by dropping the features that have high p value. The Variance Inflation Factor score is also simultaneously checked to make sure their value is kept low.

Metrics

Metrics Accuracy,Sensitivity,Specificity

True Positive Rate = Recall = Sensitivity = True Positive /(Actual Positive) = True Positive/(True Positive +False Negative).

This measure indicates what portion of the positive cases got classified correctly.For the model built on training data it is 83% and for test data it is 79%

Specificity = True Negative/(True Negative False Positive)

This measure indicates what portion of the negativecases got classified correctly.For the model built on training data it is 75% and for test data it is 80%

Precision score is 70% on train data and 72% on test data.

False Positive Rate = 1-Specificity

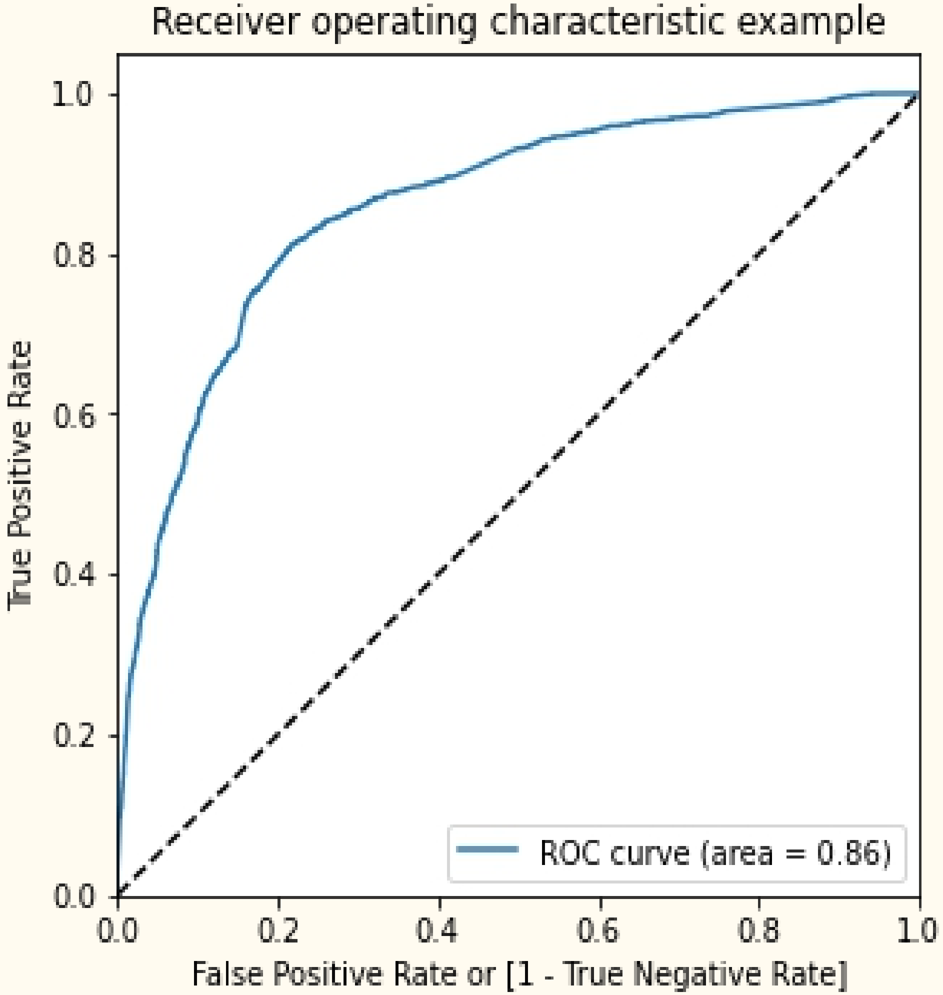
Metrics ROC /AUC

1. ROC Curve is a graph plotted with True positive rate vs False Positive rate.

2. The area under the ROC curve is 0.86

3. This value is a good score as it is close to one.

4. This measure denotes how well the classifier works in distinguishing between negative and positive cases.

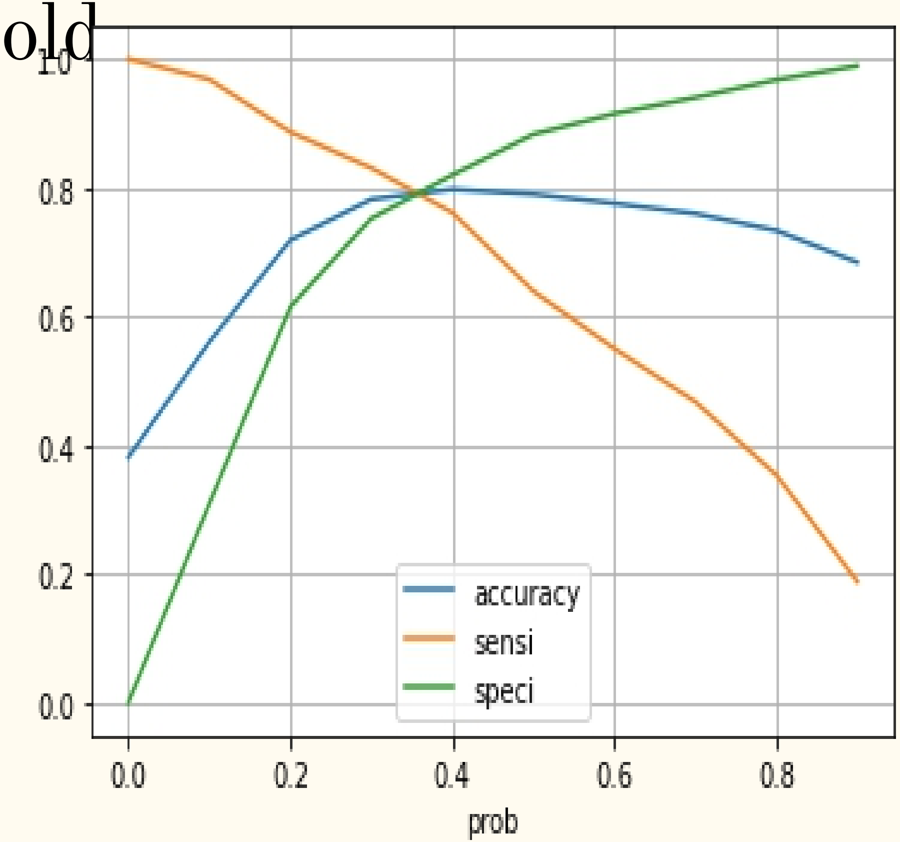


Finding Optimal Threshold

1. The curves for accuracy , sensitivity and specificity are plotted for a range of thresholds.

2. The optimal threshold is said to be the one where the three curves meet.

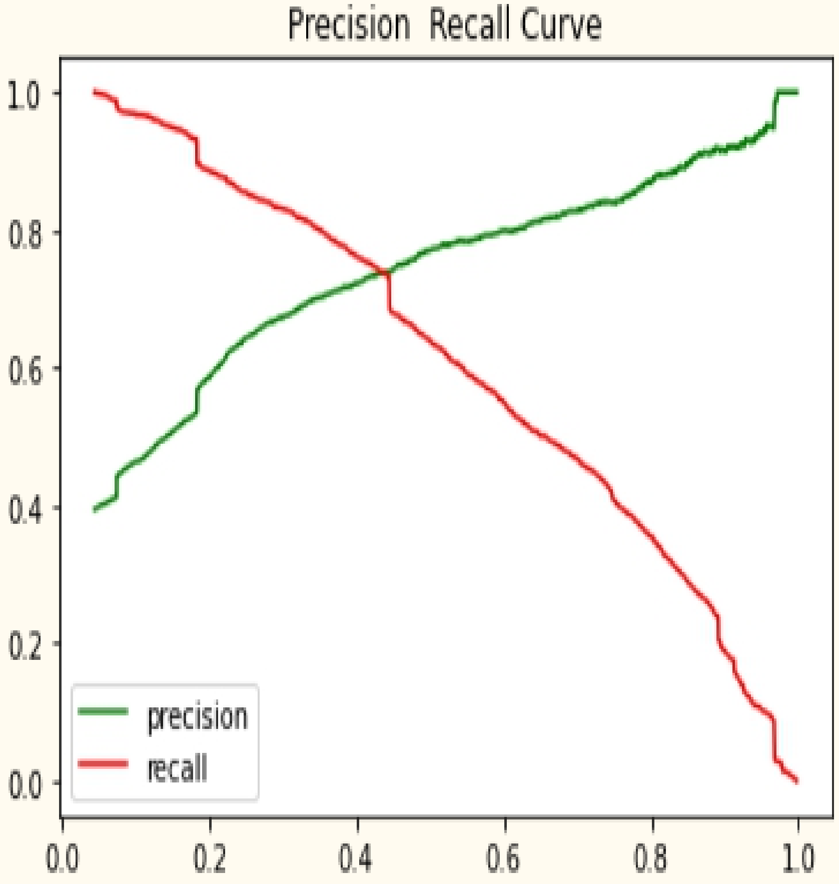
3. This is around 3.5



Precision Recall Curve

1. This graph shows a trade off between precision and recall

2. The trade off point is 0.40 approx. So any prospect lead whose conversion probability more than 40% can be chosen as hot leads.



Lead Score Calculation

- Using the logistic Regression Model the probability of the lead getting converted was given.
- Using this probability a score is assigned to each lead.
- Higher the score the lead is hotter.
- This lead can be used to prioritize the efforts taken by the sales team.

Conclusions

Top three variables in the model which contribute most towards the probability of a lead getting converted

1. Lead Origin
2. Last Activity
3. Do Not Email

Top 3 categorical/dummy variables in the model which should be focused the most

1. Lead Origin_Lead Add Form
2. Last Activity_SMS Sent
3. Last Notable Activity_Modified

Strategy to utilize by Interns to make the marketing more aggressive

Since the variable with the highest coefficient in our model is

Lead Origin_Lead Add Form,

Last Activity_SMS Sent, and

Last Notable Activity_Modified

the customers who have filled a form, sent an SMS or have any activity on the website can be prioritized based on the score assigned by the model.

Strategy to minimize the rate of useless phone calls

If all the leads based on Lead Origin_Lead Add Form, Last Activity_SMS Sent, and Last Notable Activity_Modified that is the customers who have filled a form, sent an SMS or have any activity on the website have been converted then the features **Total Time Spent on website, Specialization_Finance Management, Specialization_Operations Management** are also important.

Since our final model shows that people who have spent more time on the website, people from finance management and operation management are likely to be converted. Among these people who haven't yet converted can be targeted based on their lead score.