X Education – Lead Scoring Case Study

Detection of Hot Leads to concentrate more of marketing efforts on them, improving conversion rates for X Education

- CHARANRAJ N

Table of Contents

Background of X Education Company

- 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.
- 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%.

Problem Statement & Objective of the Study

Problem Statement:

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we 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 has given a ballpark of the target lead conversion rate to be around 80%.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are grouped based on their propensity or likelihood to convert.
- This results in a focused group of hot leads.

Better Communication

- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

Since we have a target of 80% conversion rate, we would want to obtain a high sensitivity in obtaining hot leads.

Analysis Approach

- **Data Cleaning**:Loading Data Set, understanding & cleaning data
- **EDA**:Check imbalance, Univariate & Bivariate analysis
- **Data Preparation**:Dummy variables, test-train split, feature scaling
- **Model Building**:RFE for top 15 feature, Manual Feature Reduction & finalizing model
- **Model Evaluation**:Confusion matrix, Cutoff Selection, assigning Lead Score
- **Predictions on Test Data**:Compare train vs test metrics, Assign Lead Score and get top features
- **Recommendation**:Suggest top 3 features to focus for higher conversion & areas for improvement

Data Cleaning and Preparation:

- Select level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

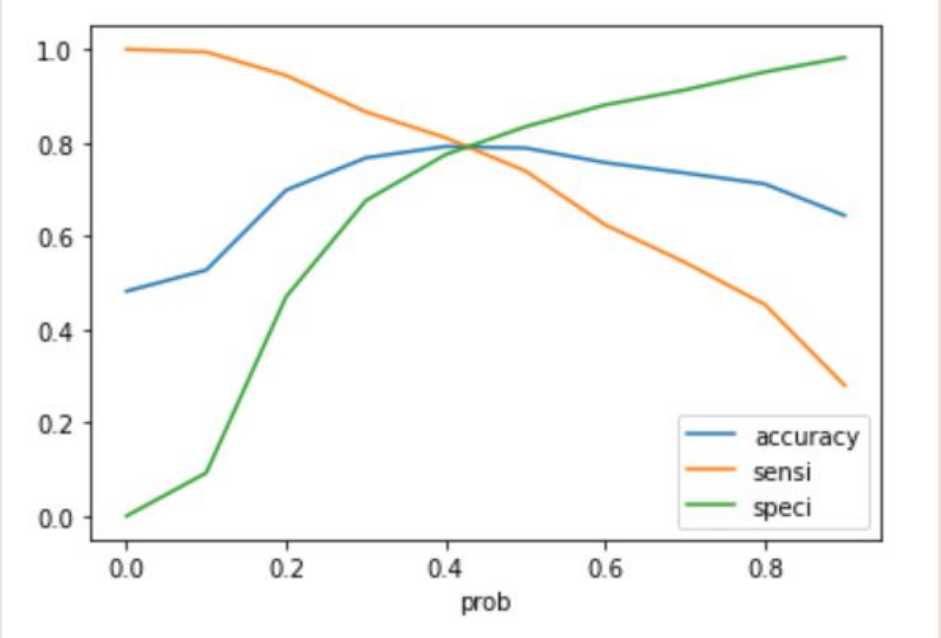
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in TotalVisits and Page Views Per Visit were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Other cleaning activities were performed to ensure data quality and accuracy.
- Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google)
- Splitting Train & Test Sets
- 70:30 % ratio was chosen for the split
- Feature scaling
- Standardization method was used to scale the features
- Checking the correlations
- Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

Model Building
Feature Selection

- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome
- Pre RFE – 48 columns & Post RFE – 15 columns
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 4 looks stable after four iteration with:
- significant p-values within the threshold (p-values < 0.05) and
- No sign of multicollinearity with VIFs less than 5
- Hence, logm4 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

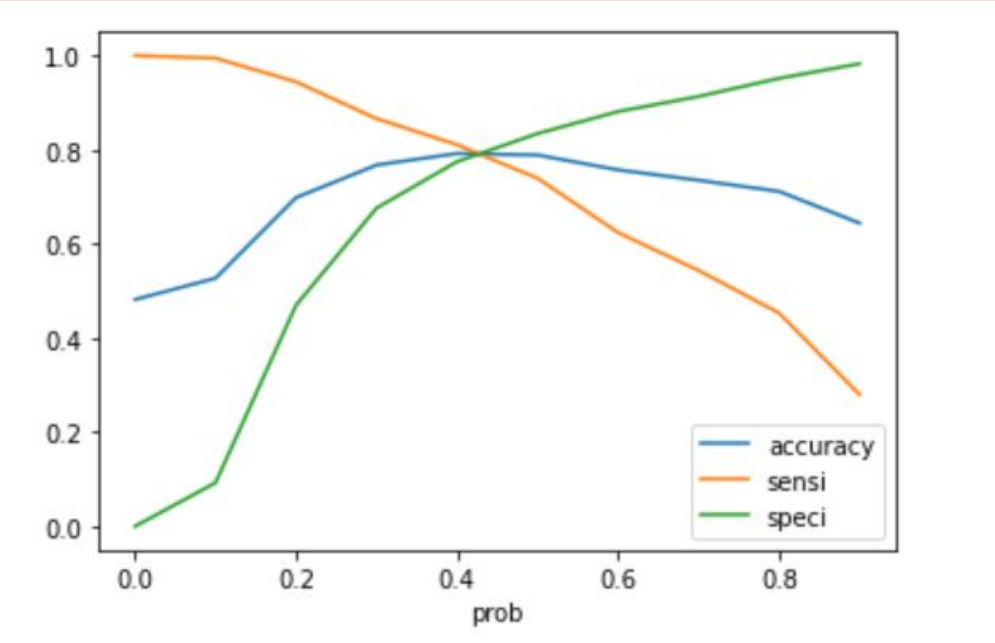
Model Evaluation
Train Data Set
Confusion Matrix & Evaluation Metrics with 0.42 as cutoff

	prob	accuracy	sensi	speci
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.527012	0.994416	0.092561
0.2	0.2	0.698274	0.944160	0.469723
0.3	0.3	0.767541	0.865984	0.676038
0.4	0.4	0.791975	0.810610	0.774654
0.5	0.5	0.788612	0.739414	0.834343
0.6	0.6	0.757229	0.624011	0.881055
0.7	0.7	0.735037	0.543509	0.913062
0.8	0.8	0.711500	0.453234	0.951557
0.9	0.9	0.644026	0.279665	0.982699



Model Evaluation

ROC Curve – Train Data Set



Recommendation based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.
- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
  - TotalVisits :11.1489
  - Total Time Spent on Website :4.4223
  - Lead Origin_Lead Add Form :4.2051
  - Last Notable Activity_Unreachable :2.7846
  - Last Activity_Had a Phone Conversation :2.7552
  - Lead Source_Welingak Website :2.1526
  - Lead Source_Olark Chat :1.4526
  - Last Activity_SMS Sent :1.1856
- We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:
  - Do Not Email_Yes :-1.5037
  - What is your current occupation_Student :-2.3578
  - What is your current occupation_Unemployed :-2.5445


Thank you