Lead Score Case Study

By:
Archana
Ankit Agrawal
G Aabhay

Lead Score Case Study for X Education

Business Goal:

X Education sells online courses to industry professionals. The company markets its courses on websites and search engines like google.

When these individuals arrive at the websites, they may peruse the course materials, complete the registration form, or see some videos. These folks are categorized as leads when they complete a form with their phone number or email address. Additionally, the business receives leads from previous recommendations. After obtaining these leads, sales team members begin contacting customers by phone, email, etc. Some of the leads convert during this process, but the majority do not. At X education, the lead conversion rate is typically 30%.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.
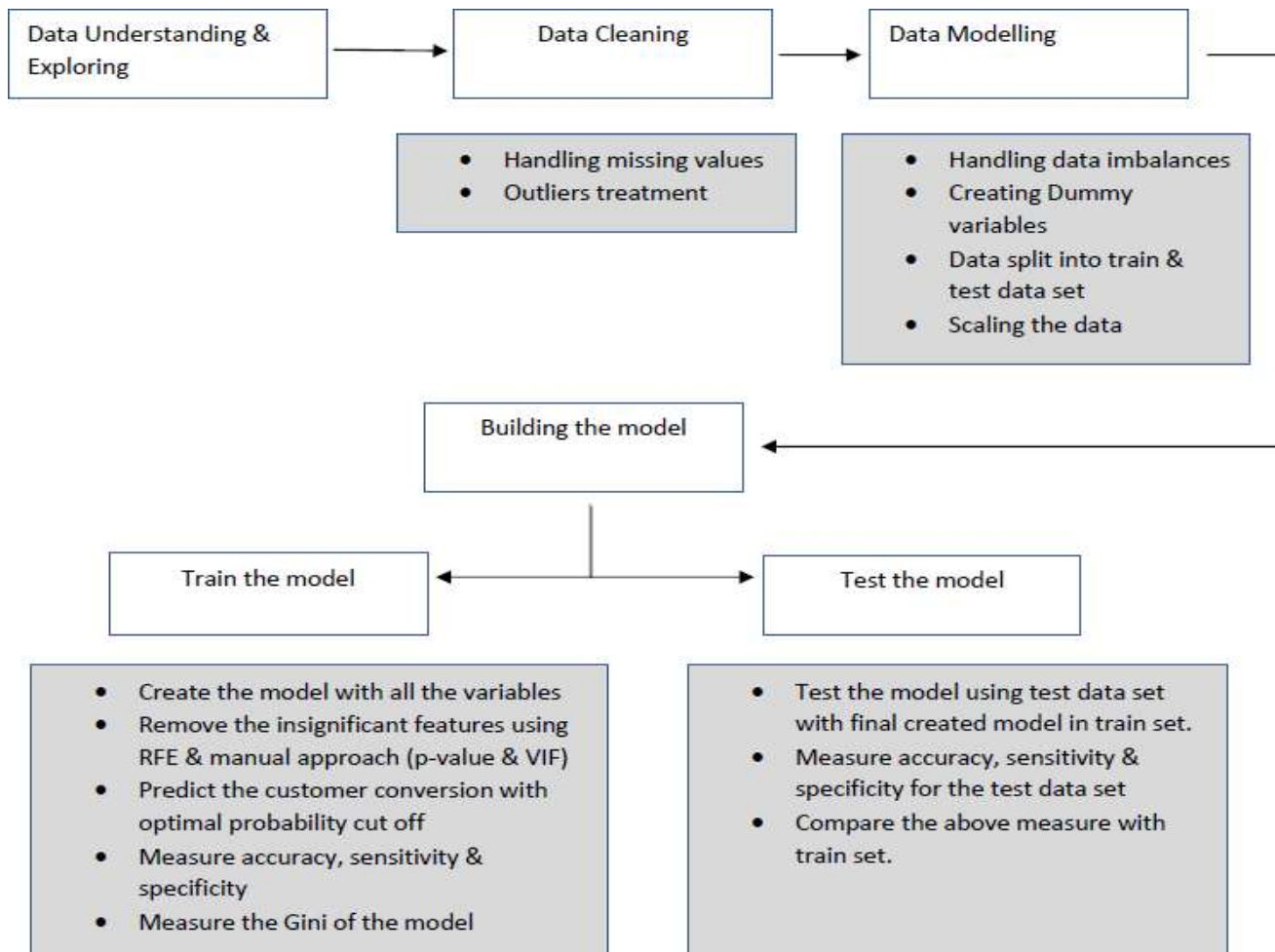
Problem Statement:

Identify the sets of leads of X Education company so that the lead conversion rate should go up so the sales team of the company could focus more on communicating with the potential leads (Hot leads) rather than approaching each and every customer which in turn would help them in paying more attention to the customers that are more likely to purchase their product/course.

The company needs a model where in it assigns a lead score to each of the leads such that the customers with high lead score have a higher conversion rate & the customer with lower lead score have a lower conversion rate.

Strategy:

➢ Source the data for Analysis

➢ Clean & Prepare the data

➢ Exploratory Data Analysis

➢ Splitting the data into Train & Test data set

➢ Scaling the Numerical parameters

➢ Building Logistic Regression model & Calculating Lead score

➢ Evaluating the Model – Accuracy, Specificity & Sensitivity (or) precision & Recall

➢ Test the model using test data set with final created model in train set and compare different metrics ~ Accuracy, Specificity & Sensitivity.

Analysis Approach:

```
┌─────────────────────────┐      ┌─────────────────────────┐      ┌─────────────────────────┐
│ Data Understanding &    │ ───> │ Data Cleaning           │ ───> │ Data Modelling          │
│ Exploring               │      │                         │      │                         │
└─────────────────────────┘      └─────────────────────────┘      └─────────────────────────┘
```

Data Cleaning:
- Handling missing values
- Outliers treatment

Data Modelling:
- Handling data imbalances
- Creating Dummy variables
- Data split into train & test data set
- Scaling the data

```
┌─────────────────────────┐
│ Building the model      │
└─────────────────────────┘
```

```
┌─────────────────────────┐              ┌─────────────────────────┐
│ Train the model         │ <──────────> │ Test the model          │
└─────────────────────────┘              └─────────────────────────┘
```

Train the model:
- Create the model with all the variables
- Remove the insignificant features using RFE & manual approach (p-value & VIF)
- Predict the customer conversion with optimal probability cut off
- Measure accuracy, sensitivity & specificity
- Measure the Gini of the model

Test the model:
- Test the model using test data set with final created model in train set.
- Measure accuracy, sensitivity & specificity for the test data set
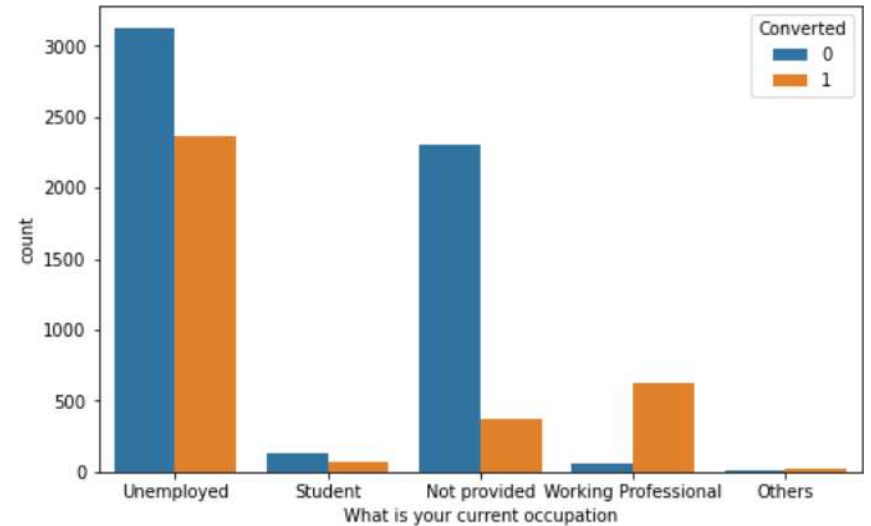- Compare the above measure with train set.

Exploratory Data Analysis:

- With increase in Total time spent on website the conversion rate also increases.
- As the customer spends more time on the X Education website more is the chance of him opting for the course.
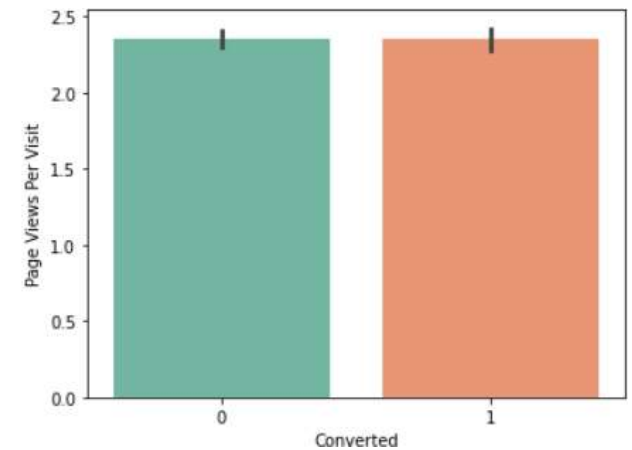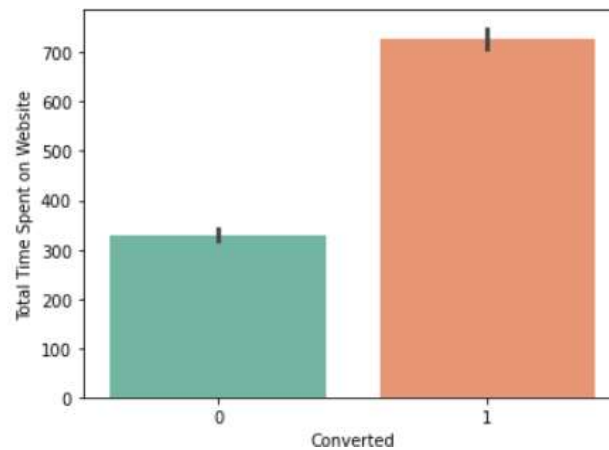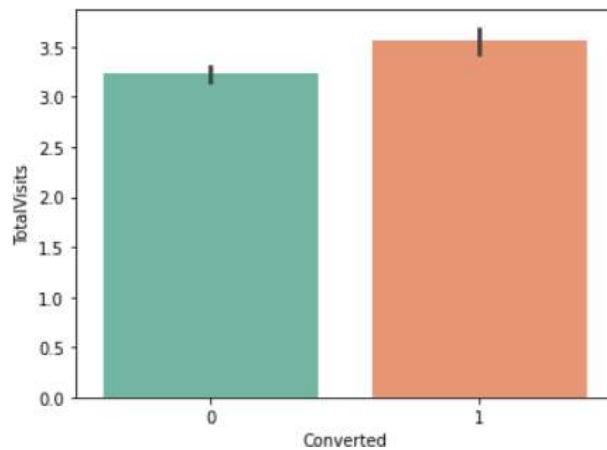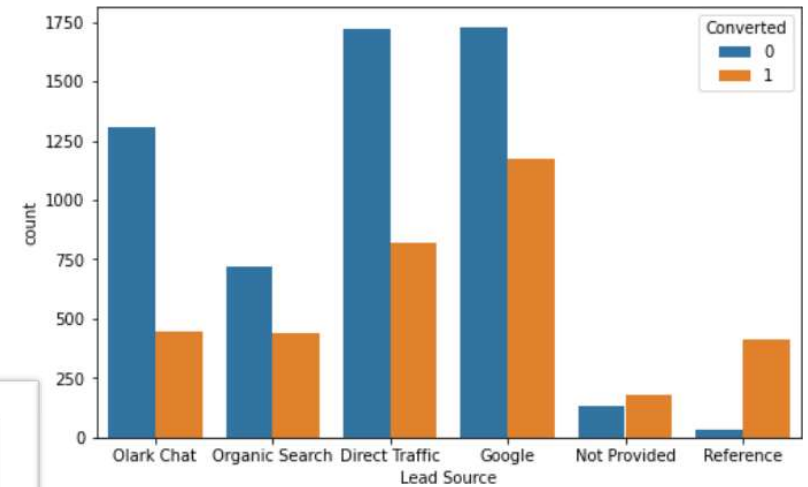
- More conversion has happened among the people who are Unemployed & followed by Working professionals
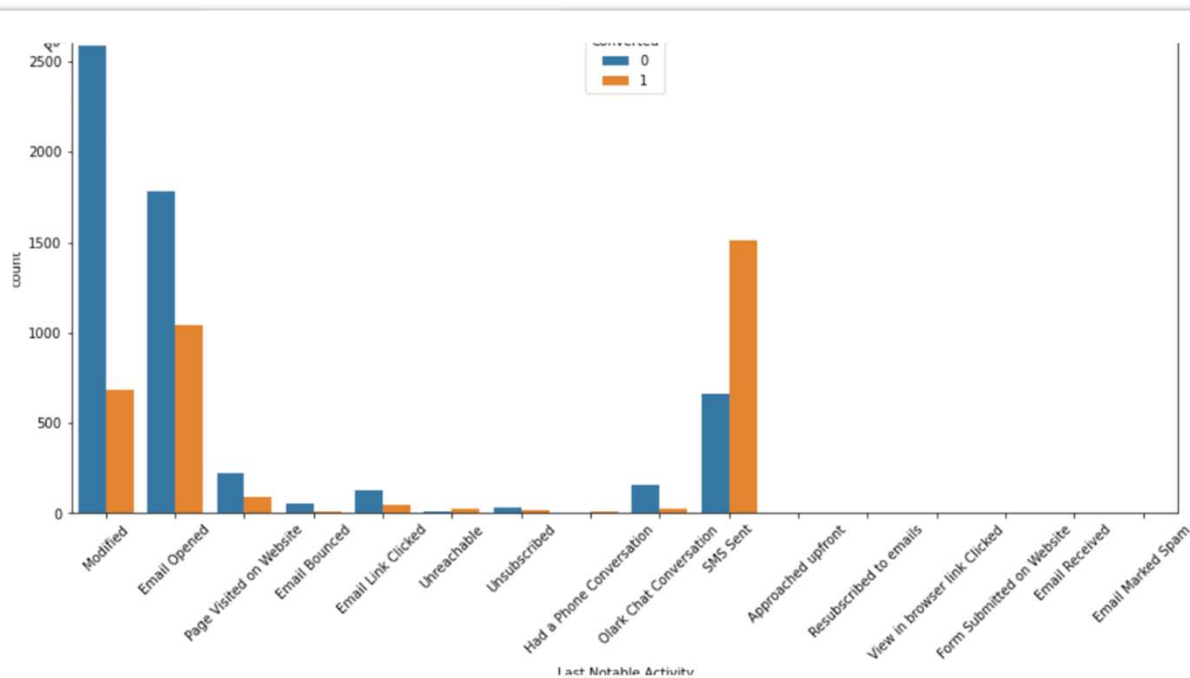

- In Lead source maximum conversion happened from Landing Page Submission, followed by API & Lead Add Form
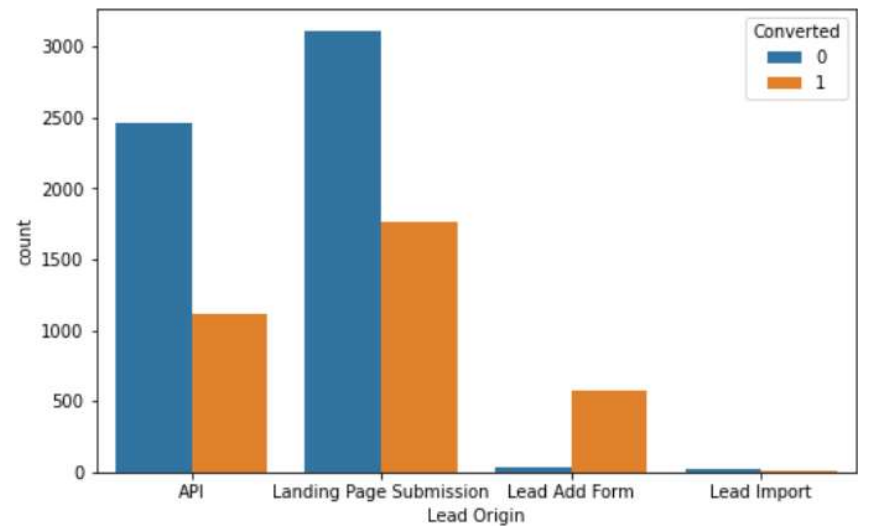
- Major conversion in lead source is from Google, followed by Direct Search, Olark chat

- Major conversion has occurred when the Last Notable Activity is 'SMS Sent'.

Variables impacting the conversion rate achieved from model

- Total Time Spent on Website
- Current Occupation – Unemployed
- Current Occupation – Working professional
- Current Occupation – Student
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity – Unreachable
- Last Notable Activity – SMS Sent
- Total Visits
- Lead Origin – Lead Add Form
- Lead Source – Olark Chat
- Last Activity – Converted to lead
- Last Activity – Olark Chat Conversation
- Last Activity – Email Bounced

Model Evaluation (Accuracy, sensitivity & Specificity) – Train Data set

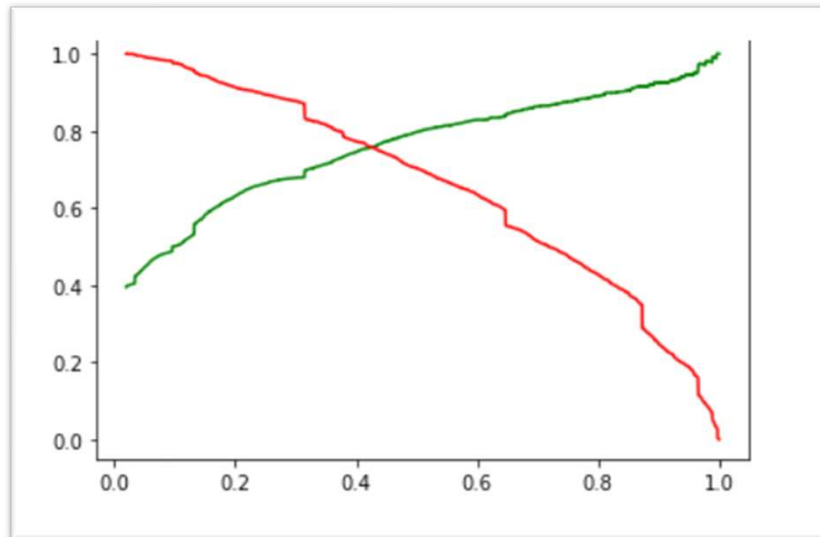The graph depicts an optimal cutoff of 0.35 based on Accuracy, sensitivity & Specificity
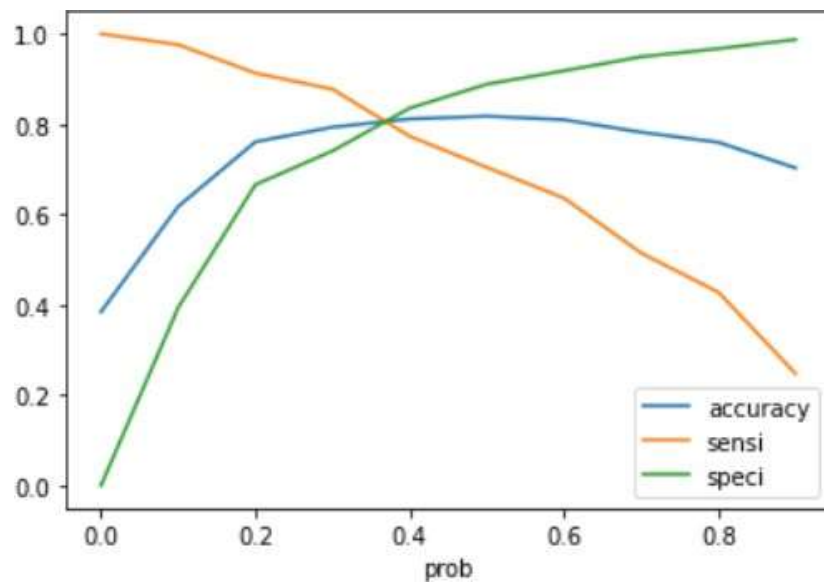
Confusion Matrix

3117	801
449	1998

- Accuracy – 80%
- Sensitivity – 82%
- Specificity – 80%

Model Evaluation (Precision & Recall) - Train Data set

The graph depicts an optimal cut off of 0.41 based on precision & cutoff

Confusion matrix

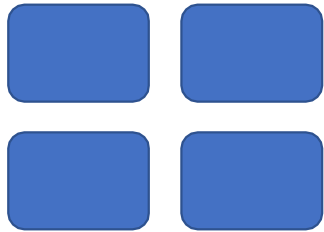|	
3296	622
563	1884

- Precision – 75%
- Recall – 77%

Model Evaluation – Test Data set

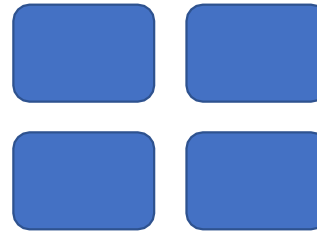- Accuracy, Sensitivity & Specificity

    Confusion matrix

    - Accuracy – 80%
    - Sensitivity – 81%
    - Specificity – 80%

- Precision & Recall

    Confusion matrix

    - Precision – 74%
    - Recall – 76%

➤ Accuracy, Sensitivity & Specificity values for Test data set are approximately closer to that of the Train data set.

➤ Precision & Recall values for Test data set are approximately closer to that of the Train data set.

Conclusion

- The top 4 variables that contribute for lead getting converted are:
    - ➤ Total Time Spent on Website
    - ➤ Lead Origin – Lead Add Form
    - ➤ Current Profession – Working Professional
    - ➤ Last Notable Activity – Had a Phone Conversation

Business recommendation for higher conversion rate:

Highly likely to be converted leads:

1. Lead score more than 71
2. Current occupation – Working professional, Student & Others
3. Lead source Olark Chat
4. Last notable Activity – SMS sent, Had a phone conversation

Less likely to be converted leads:

1. Lead score less than 13
2. Last Activity – Email bounced
3. Last Activity – Converted to lead, Olark chat conversation