Lead Scoring Case Study

Aaron Alex

Introduction

Business Understanding

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.

Problem Statement

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. The typical lead conversion rate at X education is around **30%**.

To make this process more efficient, the company wishes to identify the most potential leads, also known as '<u>Hot Leads</u>'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective

X Education has appointed us to help them 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 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%**.

ML Workflow

1. **Data Import**
2. **Data Inspection**
3. **Data Understanding**

   The target variable is **Converted**

4. **Data Cleaning**
   - Checking for duplicates
   - Treating wrong values. (Ex. *Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value*)
   - Handling features with missing values
   - Treating outliers in the data

5. **Exploratory Data Analysis**
   - Univariate Analysis – Categorical Variables
   - Univariate Analysis – Numerical Variables

6. **Data Preparation**
    - Converting categorical variables (yes/no) to binary variables.
    - Creating dummy variables. (One-hot encoding)
    - Feature Scaling

7. **Model Building**

    We have used **Logistic Regression** to build our model, assign lead scores and make predictions.

8. **Model Evaluation**
9. **Model Presentation**
10. **Conclusion and Recommendations**

Inferences from Data Cleaning

- **The following columns were dropped because they had a high percentage of missing values:**
  - *How did you hear about X Education*
  - *Lead Profile*
  - *Asymmetrique Activity Score*
  - *Asymmetrique Profile Index*
  - *Asymmetrique Activity Index*
  - *Asymmetrique Profile Score*
  - *Tags*
  - *What matters most to you in choosing a course*
  - *What is your current occupation*
  - *Country*

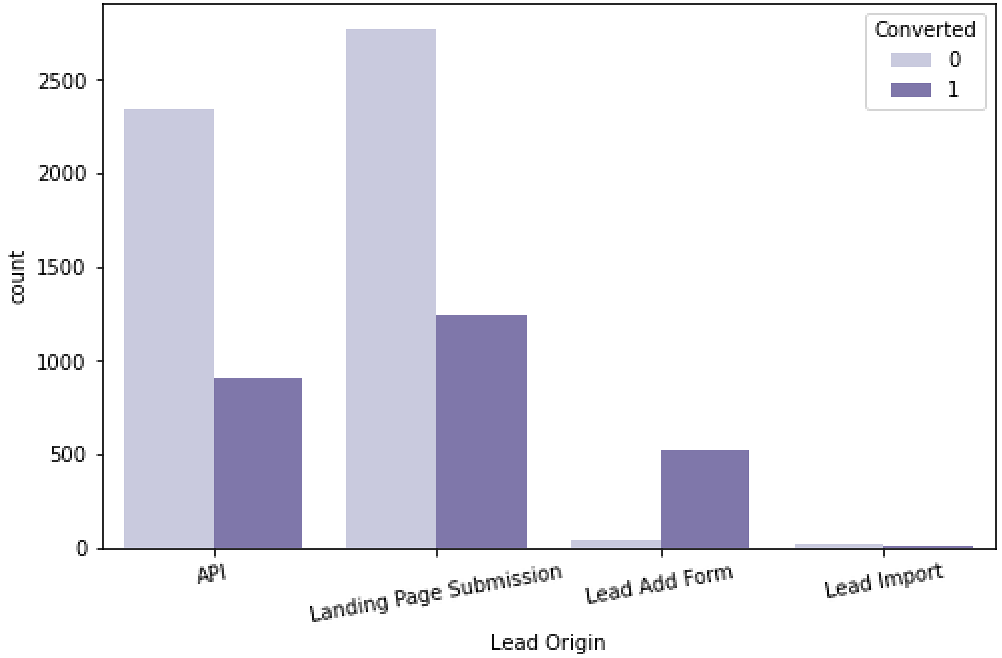
Inferences from EDA

- **Class Imbalance**

  Our data is close to a 60/40 distribution therefore it isn't severely imbalanced. We did not need to employ other methods to balance our data.

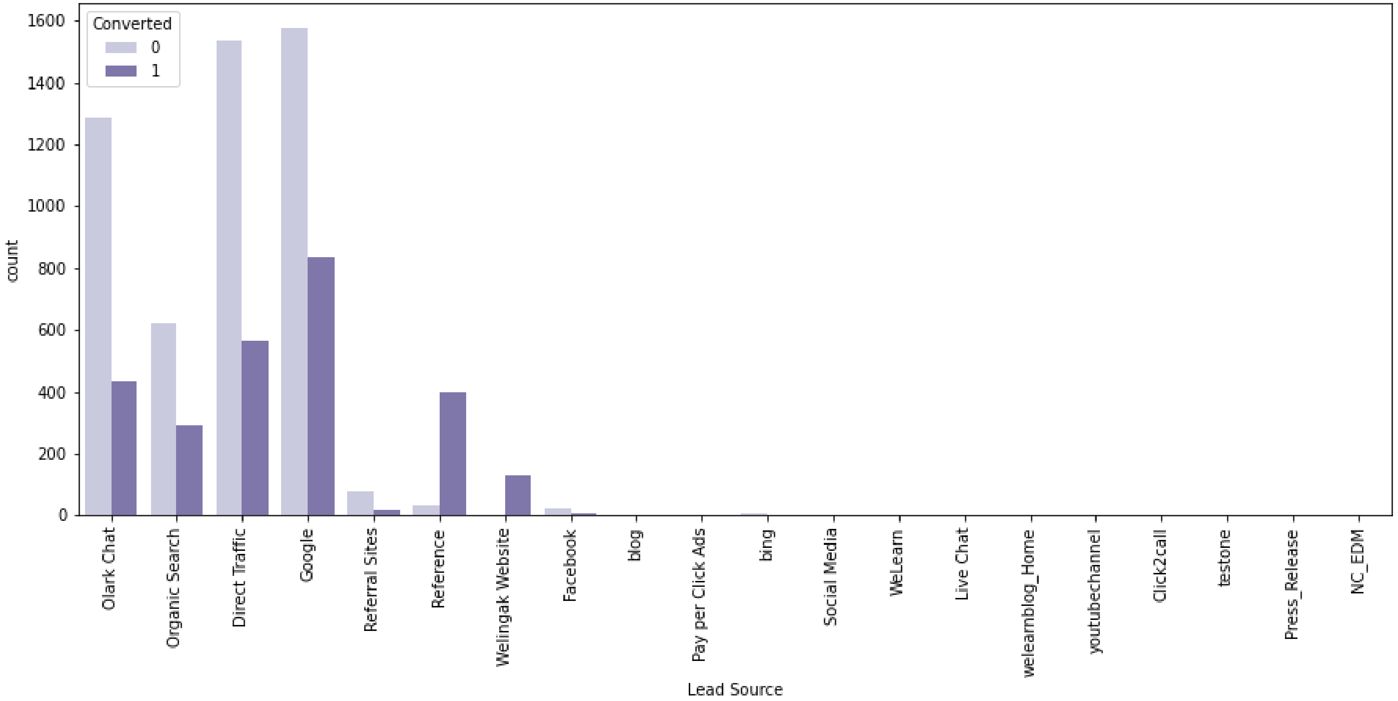
- **Lead Origin**

  Leads originating from the Lead Add Form have the highest conversion rate at 94%. However, the total count of leads originating through this channel is less.



- **Lead Source**

  Leads coming from References and Welingak website are low in count but have high conversion rates. The most number of leads are obtained from Google.

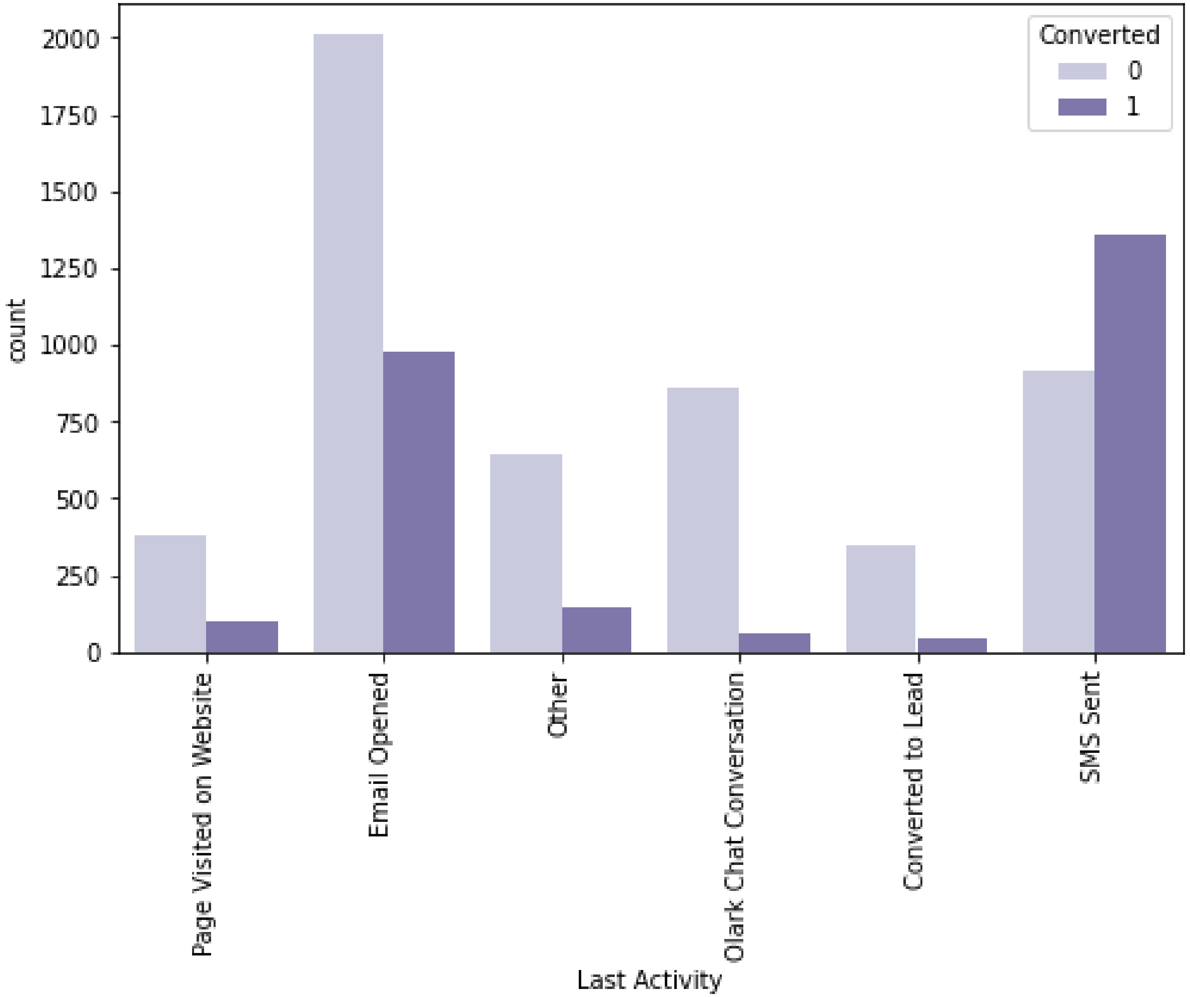


- **Do Not Email & Do Not Call**

  Majority of the leads have opted to not receive calls or emails from the company.

- **Last Activity**

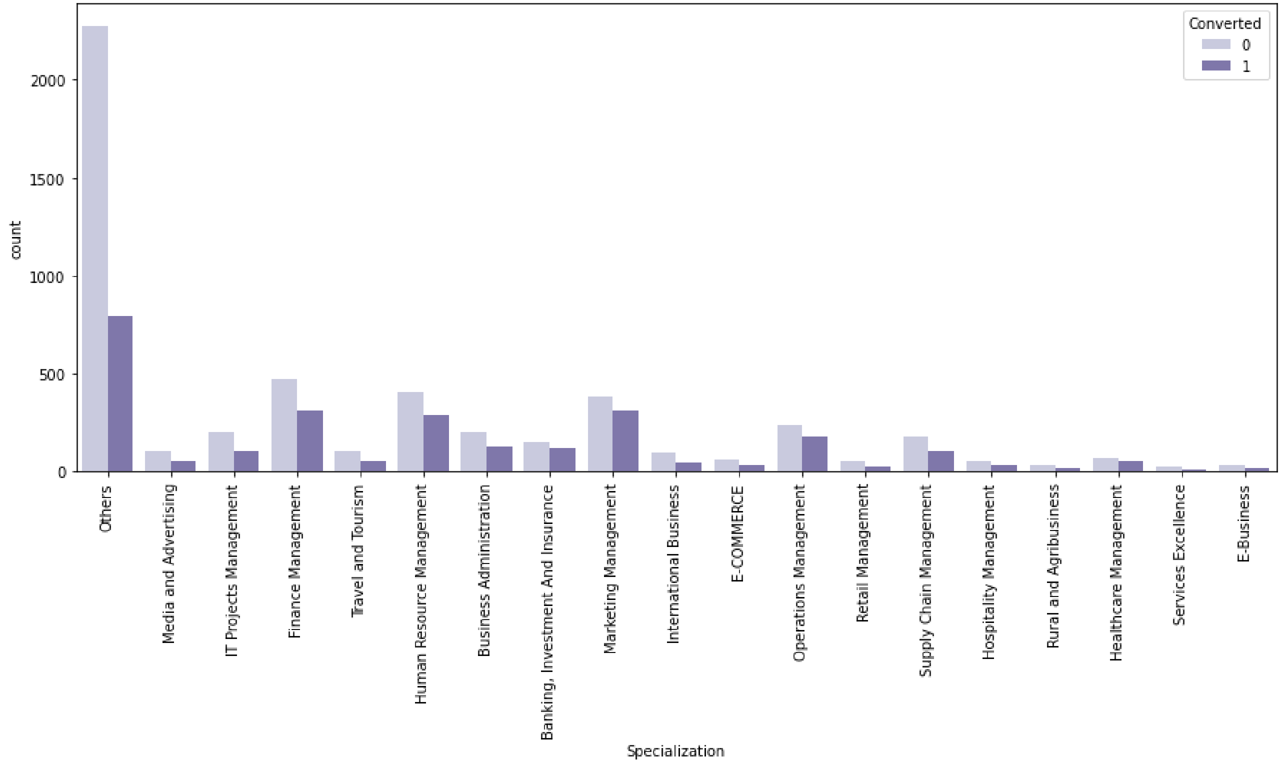
  The conversion rate for leads whose last activity was sending an sms was the highest at 60%



Specialization

Conversion rates for the following categories are higher (>40%)

- Human Resources Management
- Marketing Management
- Banking, Investment and Insurance
- Operations Management
- Healthcare Management



- **Advertising Channels**

  Almost none of the leads have seen advertisements through the following channels:

  Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations

- **Receive more updates about our courses**

  All leads have opted out of receiving more updates about our courses.

- **Update me on Supply Chain Content**

  All the leads have opted out of updates on Supply Chain Content.

- **Get updates on DM Content**

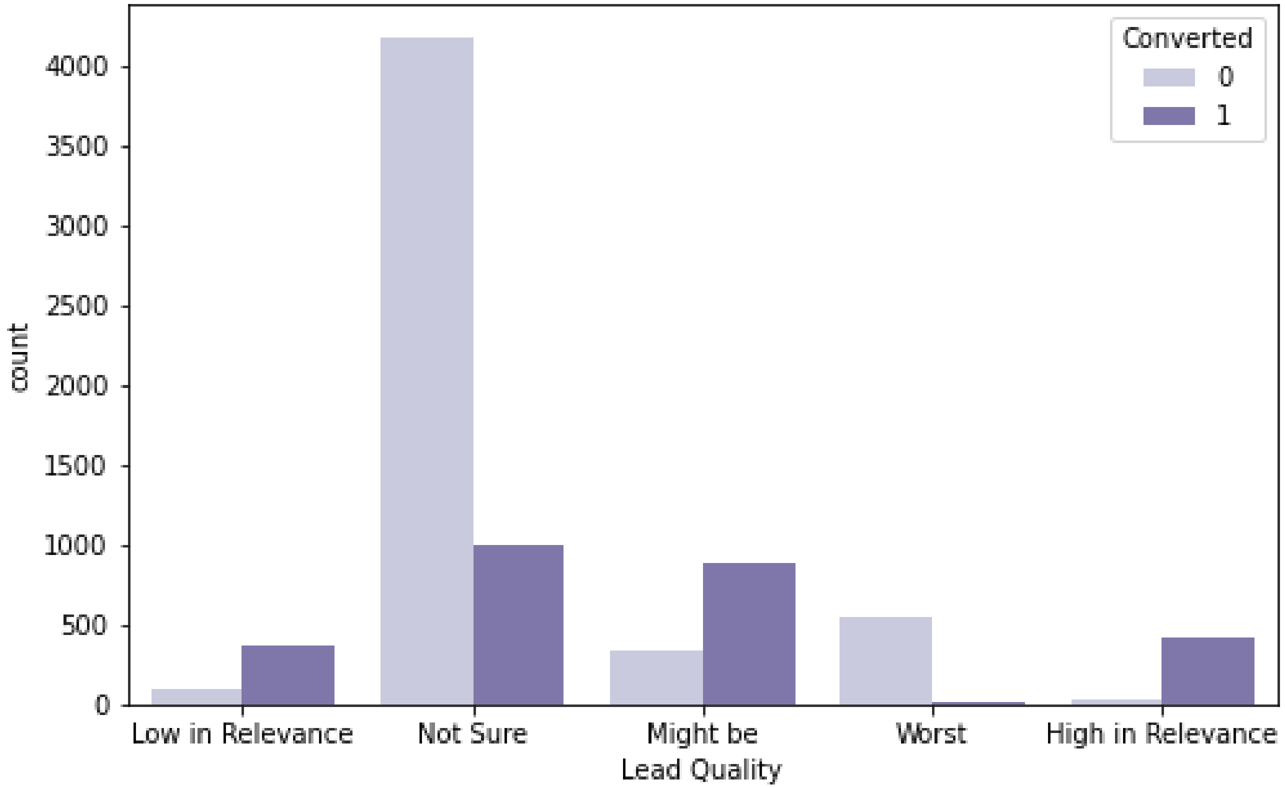
  All the leads have opted out of updates on Supply Chain Content.

- **I agree to pay the amount through cheque**

  None of the leads have agreet to pay the amount through cheque.

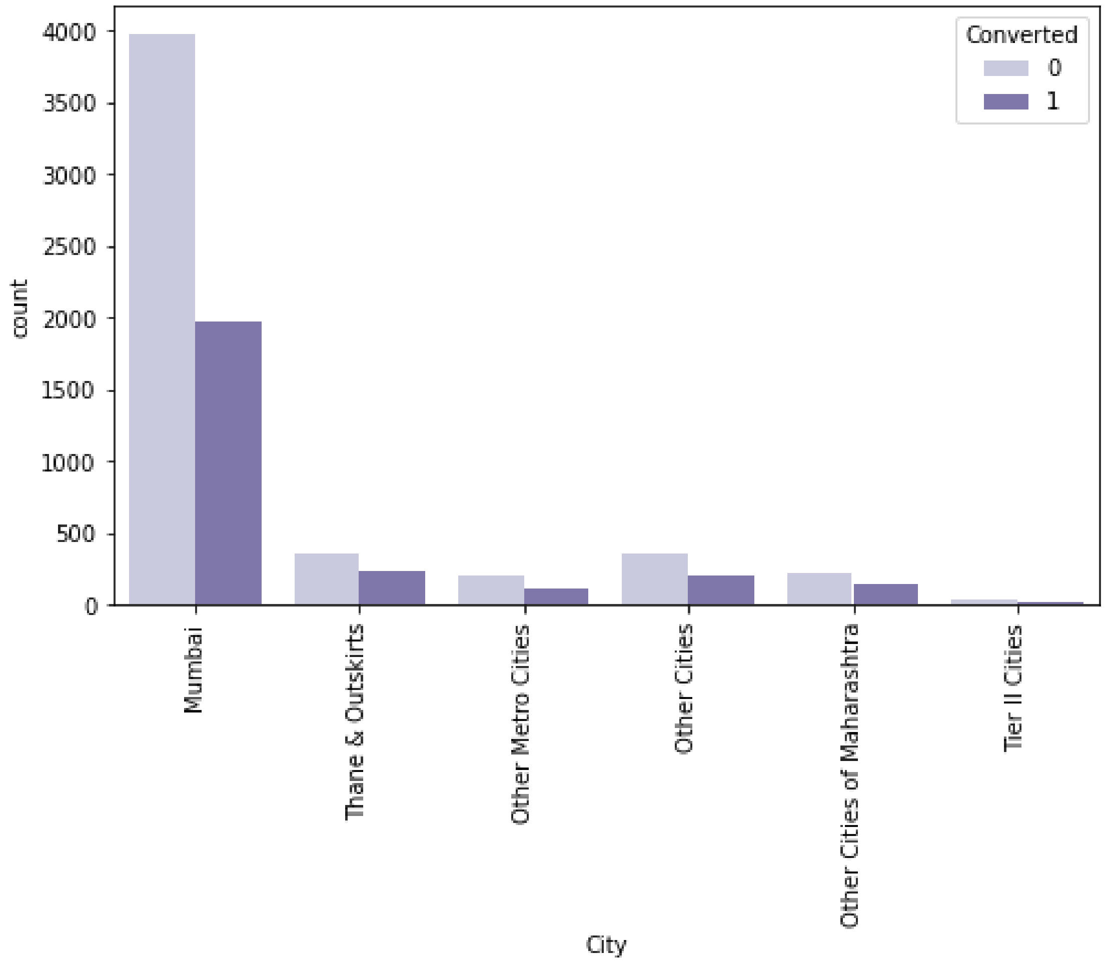
- **Lead Quality**

  High in Relevance leads have the highest conversion rate at 80%



- **City**

  Mumbai has the highest count of leads. Thane and Outskirts have the highest conversion rate at 40%



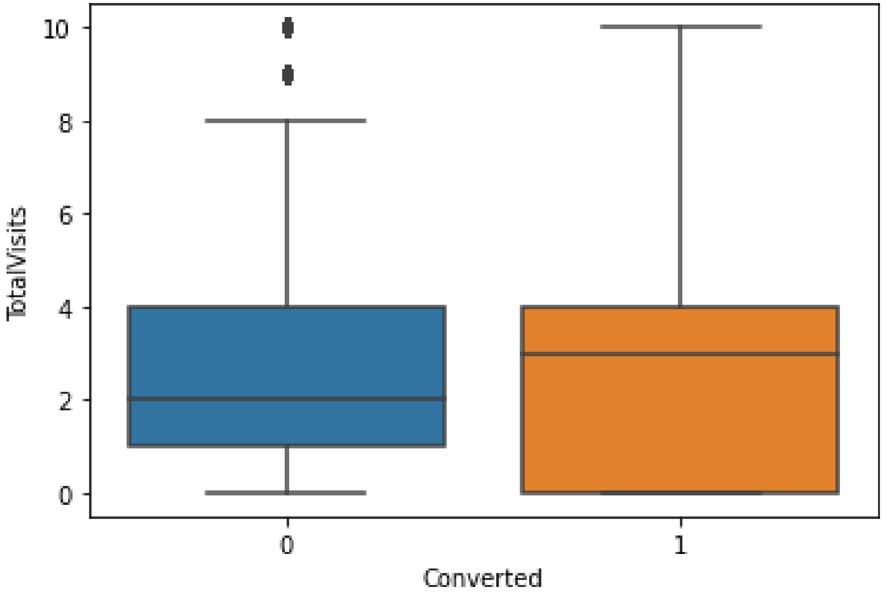
- **A free copy of Mastering the Interview**

  Most of our leads have not opted for the free copy of Mastering The Interview. Conversion rate for leads who have not opted for the free copy is slightly higher than the leads who opted for it.

- **TotalVisits**

  The median TotalVisits for both converted and non converted leads lie between 2 and 4.

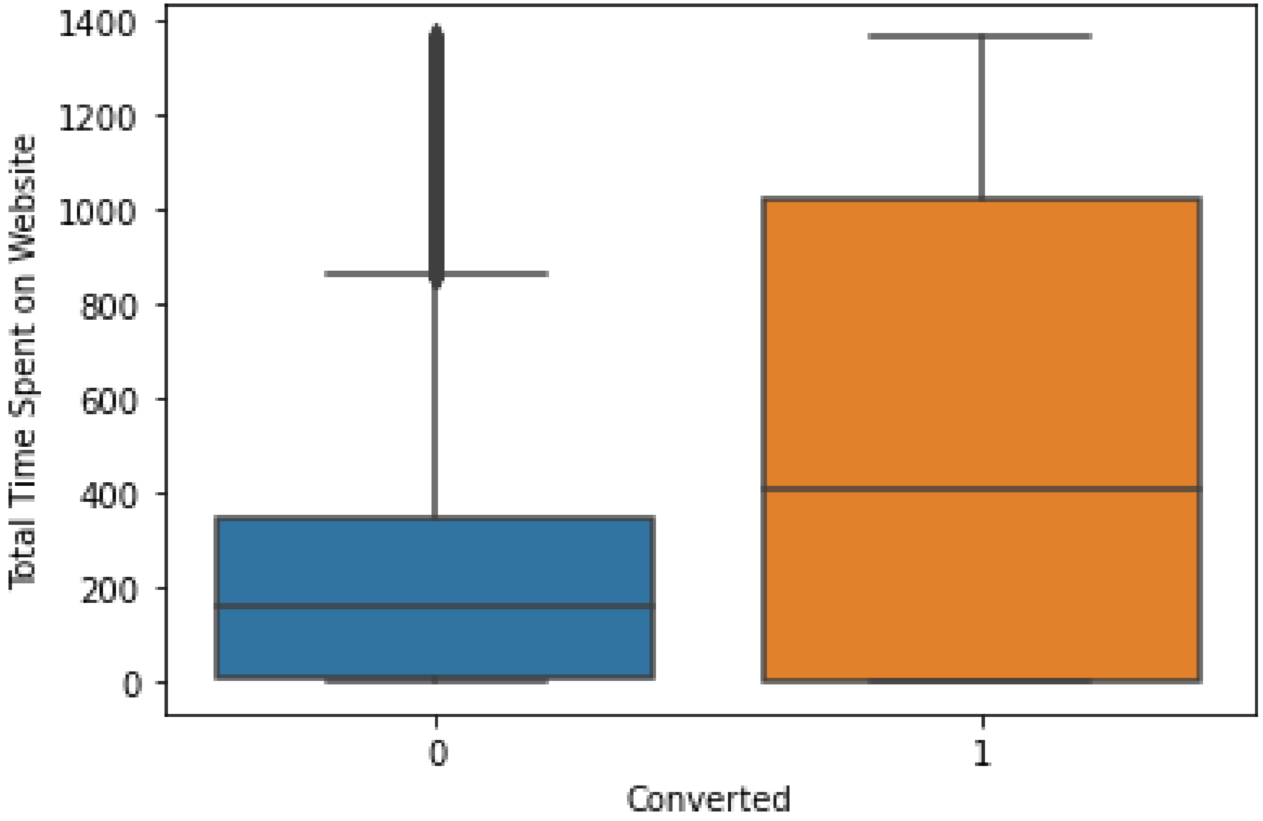
  Nothing conclusive can be derived from this feature



- **Total Time Spent on Website**

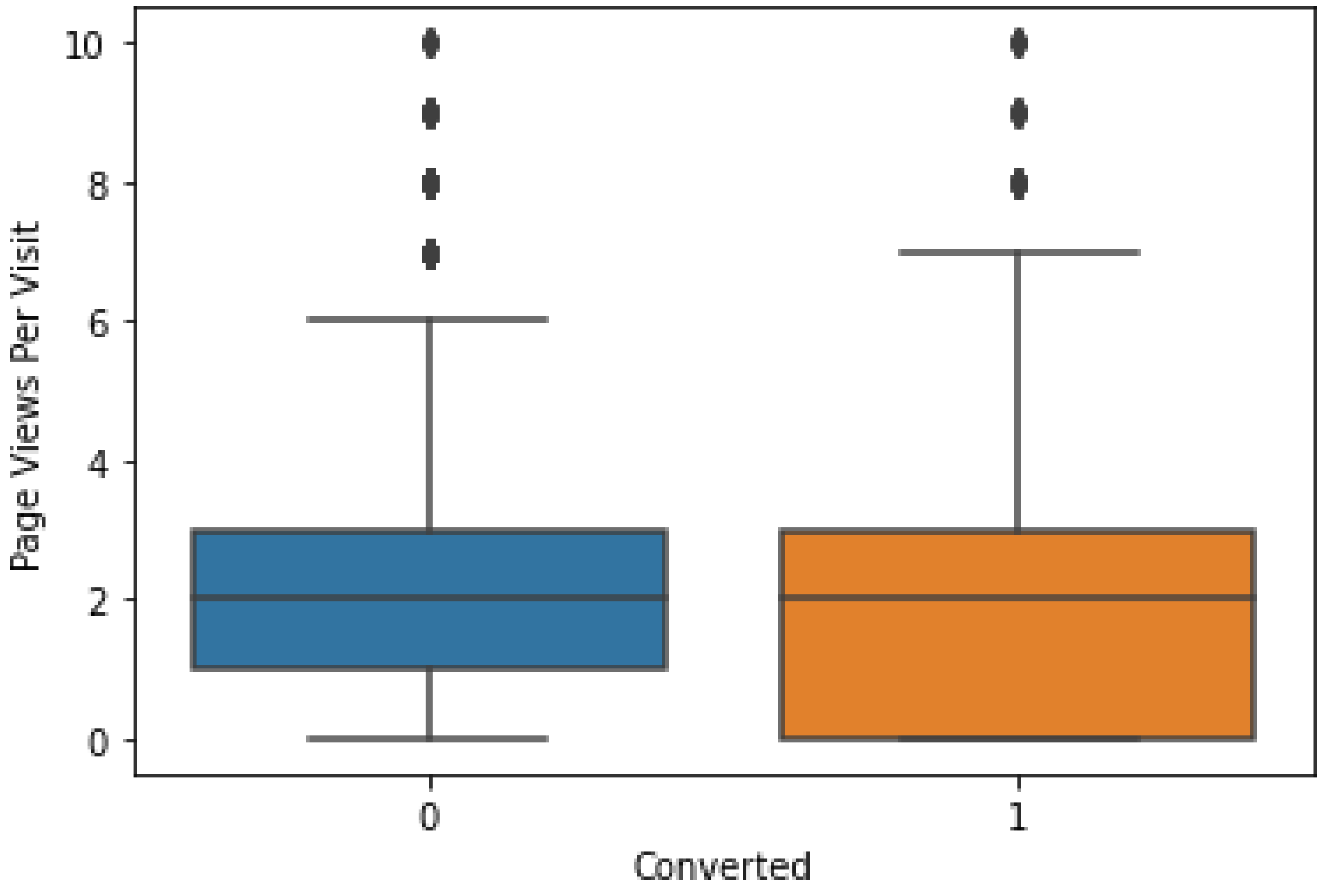
  There is greater variance in the time that converted leads spent on the website.

  The median time spent by leads that converted is higher than leads that did not convert. Leads that spent more time on our site are likelier to convert.



- **Page Views Per Visit**

  The median value for Page Views Per Visit is similar for both converted and non converted leads. Nothing conclusive can be derived from this feature.



Results of EDA

Many of the features were not adding any information to our model, hence we drop them from our model for further analysis. The following columns were dropped:

- *Lead Number*
- *Search*
- *Magazine*
- *Newspaper Article*
- *X Education Forums*
- *Newspaper*
- *Digital Advertisement*
- *Through Recommendations*
- *Receive More Updates About Our Courses*
- *Update me on Supply Chain Content*
- *Get updates on DM Content*
- *I agree to pay the amount through cheque*
- *Last Notable Activity*

Data Preparation

- Converting some categorical variables to binary variables
- Creating dummy variables
- Scaling numerical features

Model Building

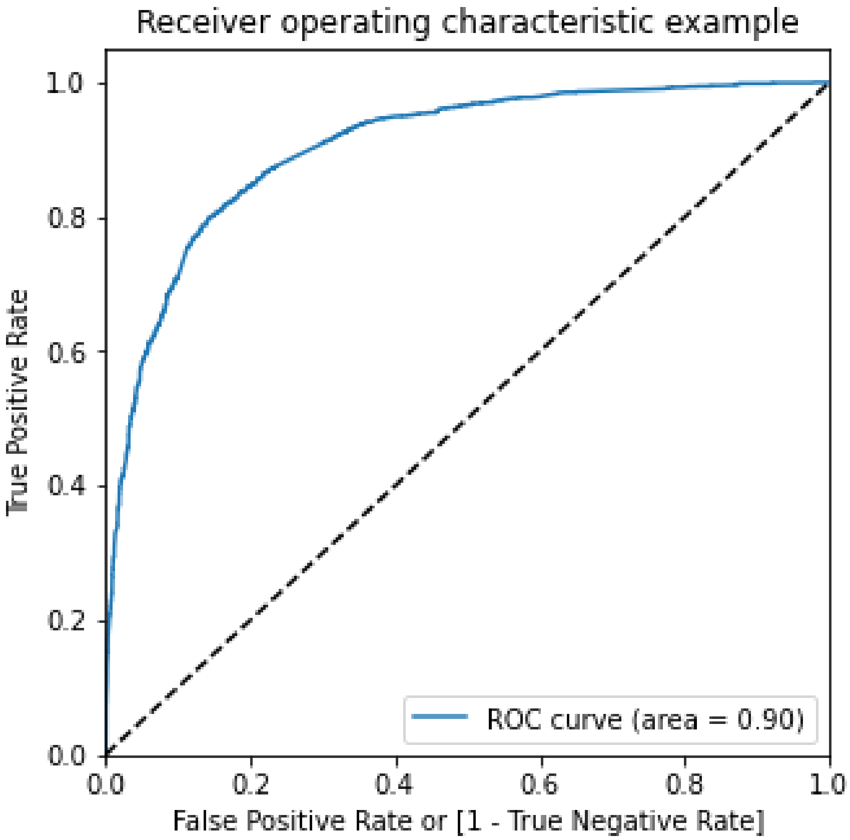
- **Separating X and y**

- **Performing Train-Test split.**
  We have chosen 70-30 ratio.

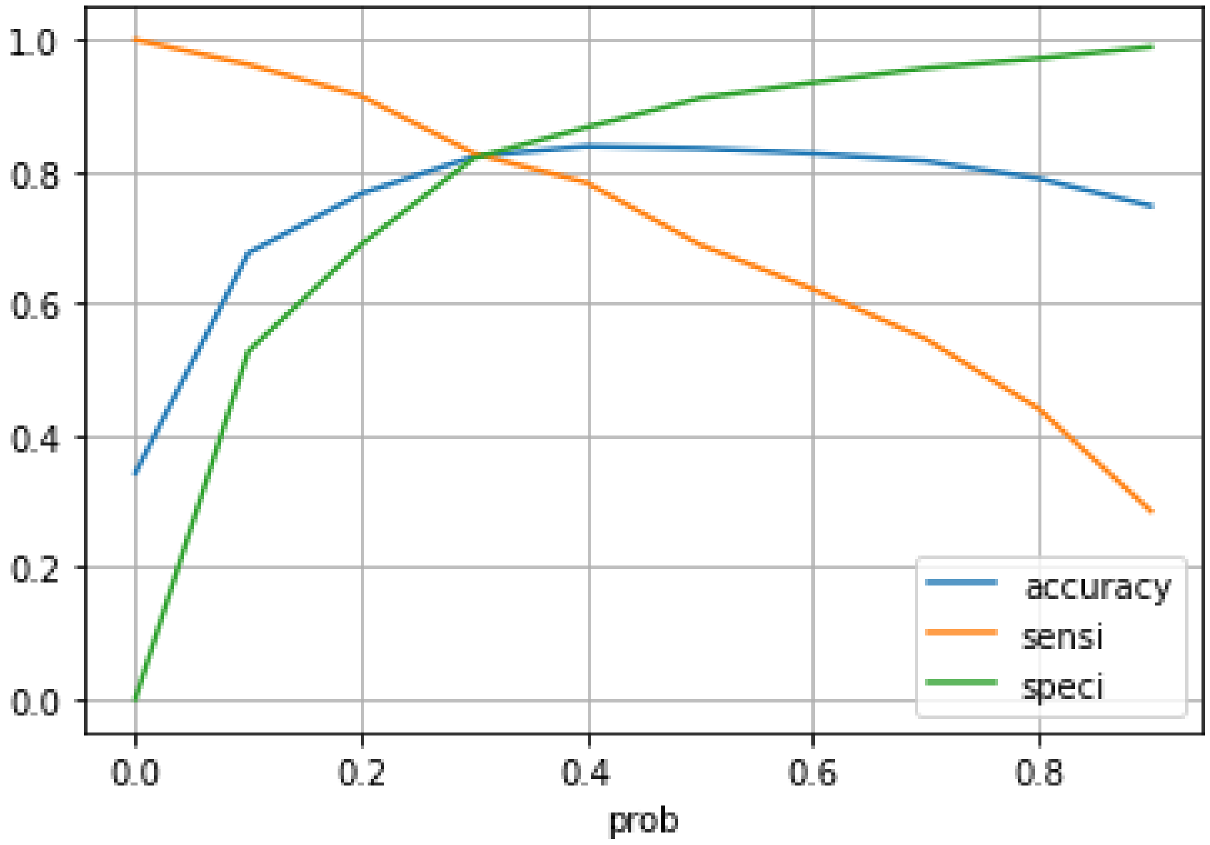
- **Using RFE for feature selection.** **(Coarse Tuning)**
  Running the RFE with 20 variables as the output.

- **Fine Tuning the model**
  Removing variables having p-values greater than 0.05 and VIF-values greater than 5

- **Selecting optimal cut-off point**
  - Plotting the ROC curve to observe the trade-off between specificity and sensitivity. The **area under the curve is 0.90**. An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability.



- Plotting accuracy, sensitivity and specificity for various probability cut-offs. The optimal cut-off value is where we get balanced sensitivity and specificity. From the curve, **0.3 is the optimal cut-off point**.



Model Evaluation