LEAD SCORE CASE STUDY

Select leads that are most likely to convert into paying customers

Submitted By:
Shabd Shashank
Gargi Singh

Problem Statement and Objective

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- Typical lead conversion rate at X education is around 30%.
- Company wishes to identify the most potential leads, also known as 'Hot Leads'.
- Sales team can then focus on these 'Hot Leads' for bettering the conversion rate.

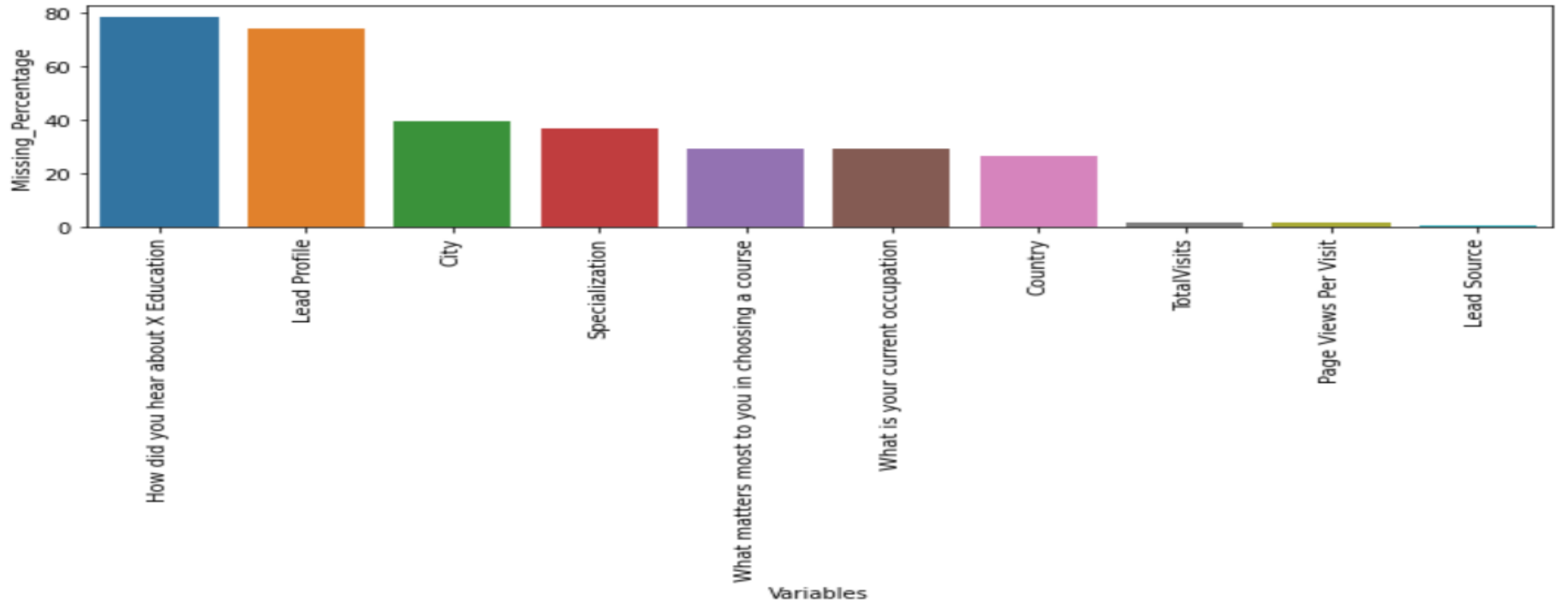
Objective

- Categorize the leads and identify hot leads.
- For the above, build a logistic regression model

Analysis Approach

1. Data cleaning and manipulation:
   - Check the data for duplicate values
   - Check the data for any missing values and treat them accordingly, either by imputing or removing the column if large amount of data is missing.
   - Check and handle outliers in the data.
2. Univariate and Bivariate analysis:
   - Analyze important features using various plots like:
     - Countplot
     - Boxplot
   - Create a heatmap of correlation coefficients
3. Data preparation:
   - Create dummies of the variables with multiple options.
   - Encode binary categorical variables to 0s and 1s.
4. Model Building:
   - Split the data into train and test sets.
   - Scale the train set, if necessary.
   - Apply RFE to select the top 15 variables.
   - Classification technique: Use Logistic Regression to build the model and predict hot leads.
5. Model Evaluation:
   - Check various scores of the model like, precision, recall etc.
6. Interpret the model and Business recommendations

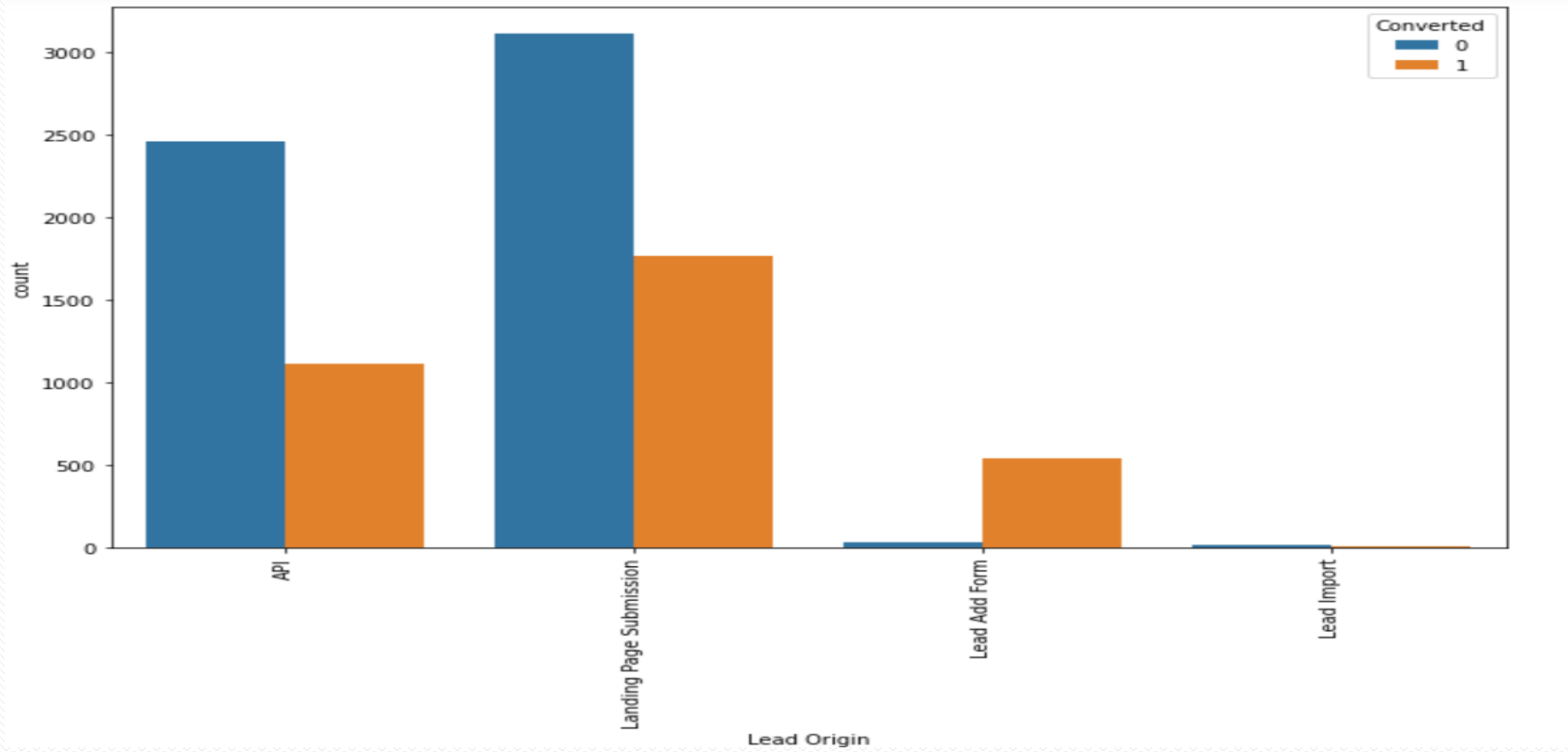
Data Cleanliness Check



❖ *Observations:*

1. Few variables have very high missing values (>60%). These variables can directly be dropped.
2. Other categorical variables can be imputed with mode of the variable
3. Variables with very less missing values (<=1%), rows can be directly dropped.

Univariate Analysis

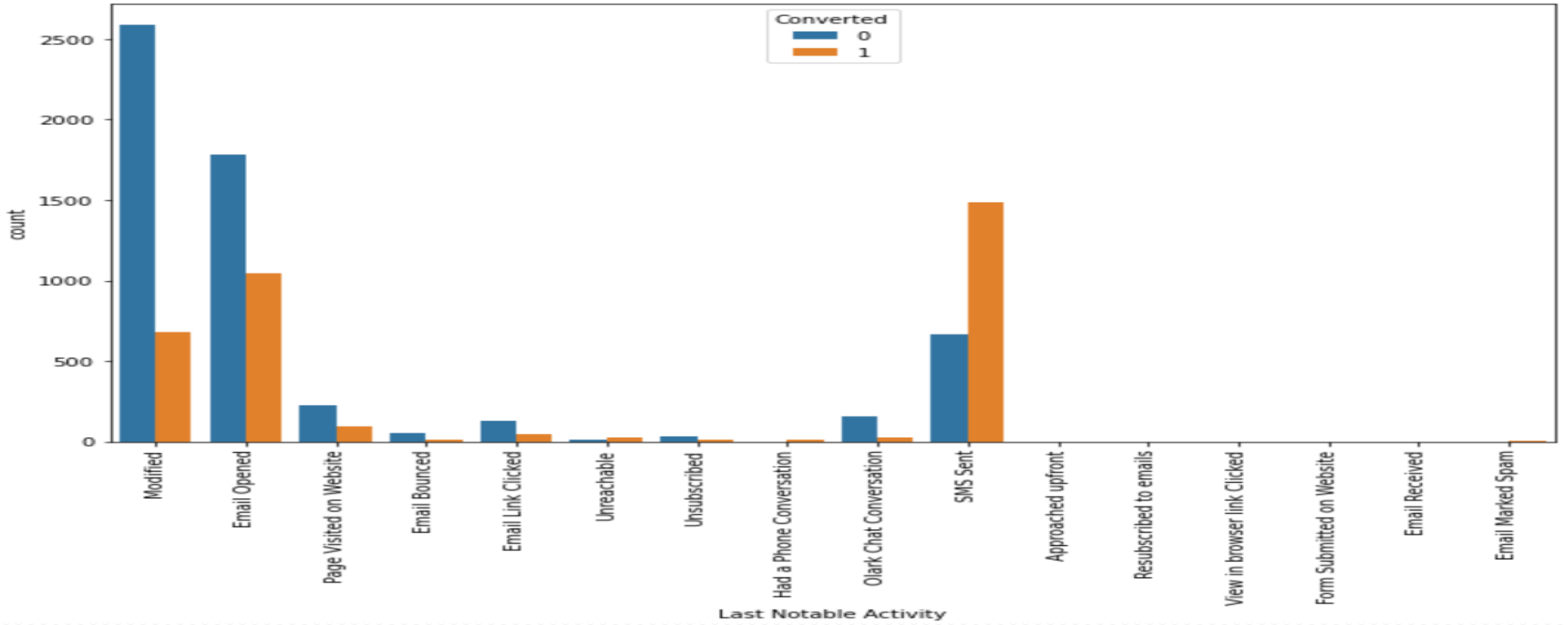


LEAD ORIGIN

- *Observations:*

1. When the lean origination is Lead Add Form – the chances of conversion is very high.
2. When the lead origination is Landing Page Submission – the chances of conversion is low.

Univariate Analysis



LAST NOTABLE ACTIVITY

- *Observations:*

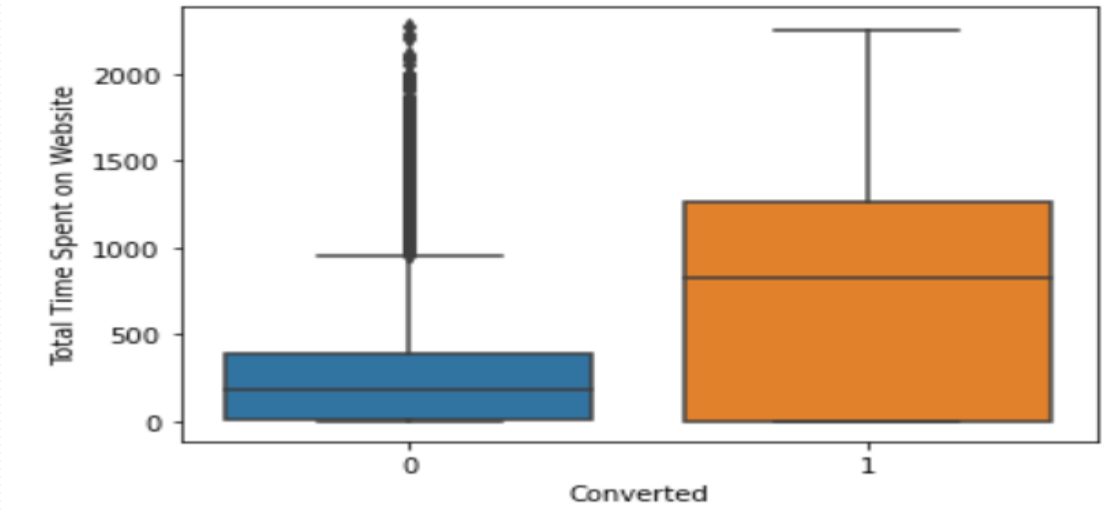
1. When the last notable activity was SMS sent – the conversion ratio is very high
2. When the last notable activity was Modified - the conversion ratio is very poor.

Univariate Analysis

TOTAL TIME SPENT ON WEBSITE

❖ *Observations:*

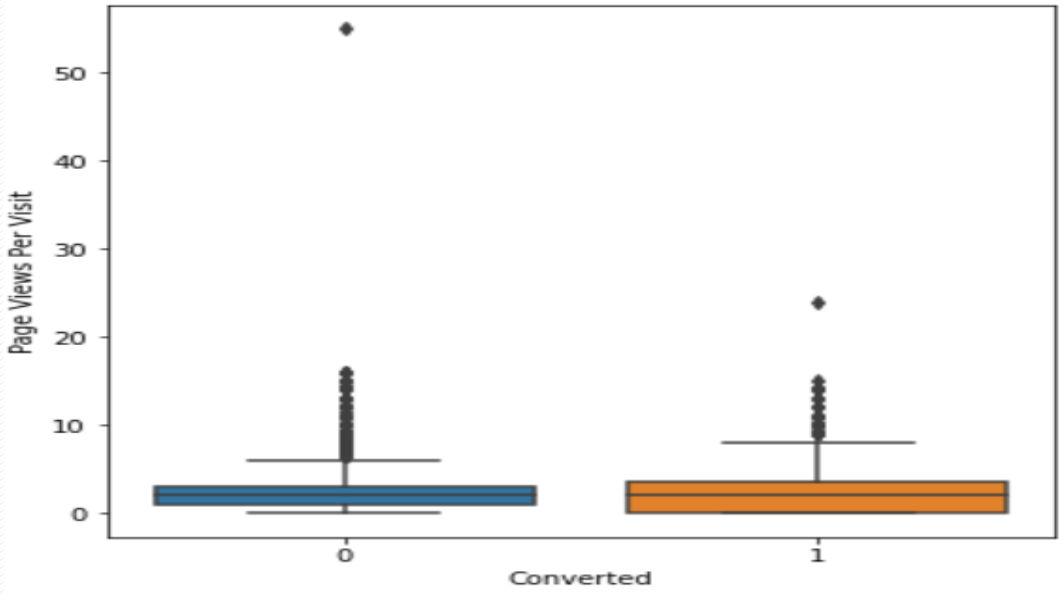
1. More is the time spent on the website, the higher is the likelihood of conversion.
2. Business should focus on improving their website UI to engage the customer on the website more.



PAGE VIEWS PER VISIT

❖ *Observations:*

1. Median page views per visit for both the categories are very similar.
2. Converted category has more page views per visits, which again boils down to engaging the customers more on the website.

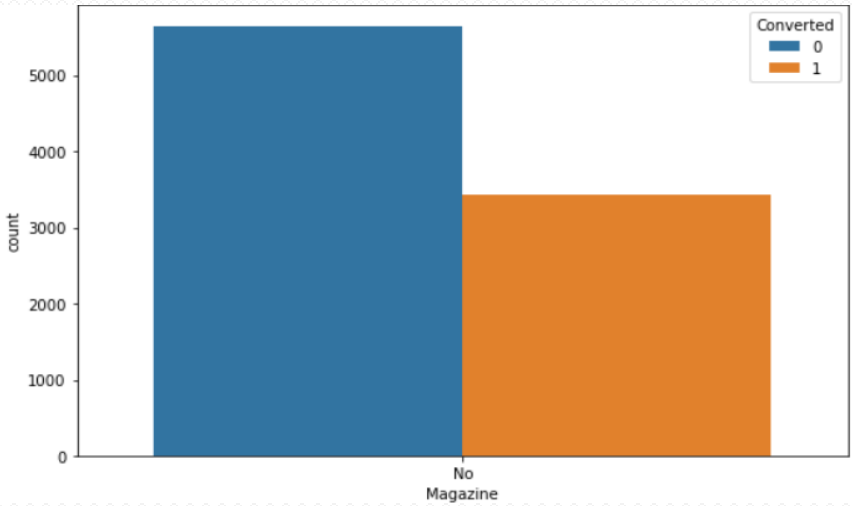


Univariate Analysis

MAGAZINE

❖ *Observations:*

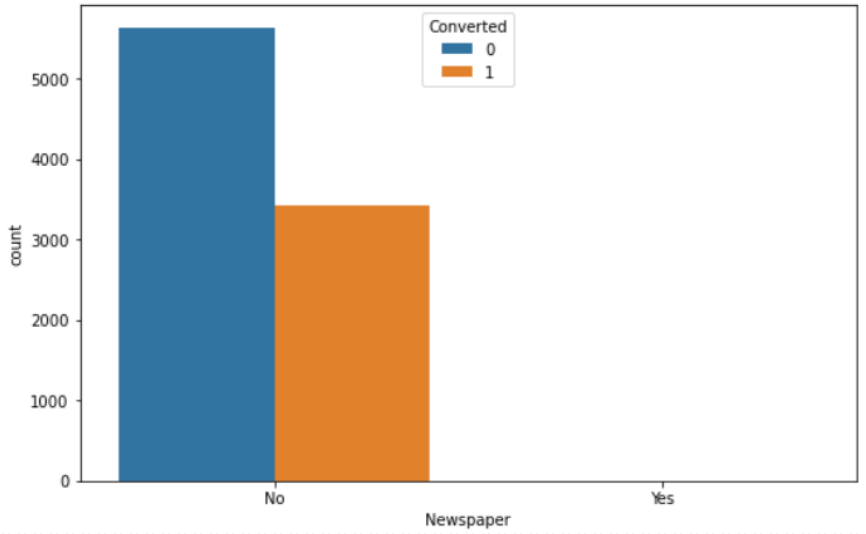
1. Data is highly skewed – we have only “No” category in the data.



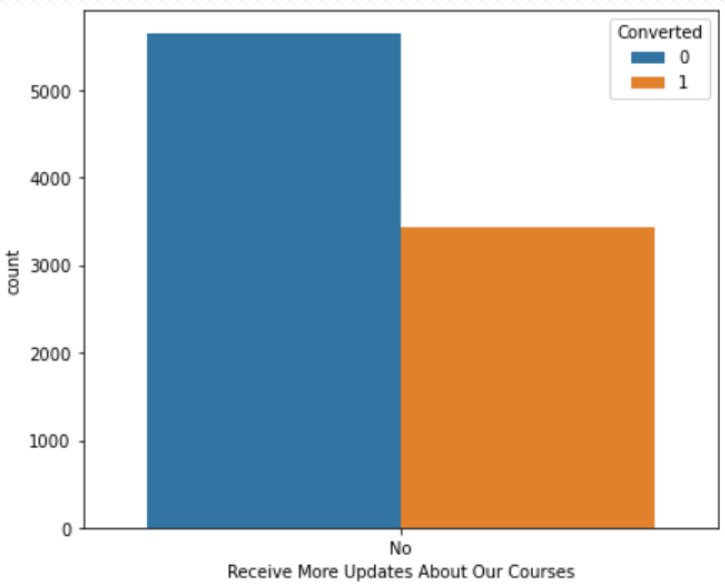
NEWSPAPER

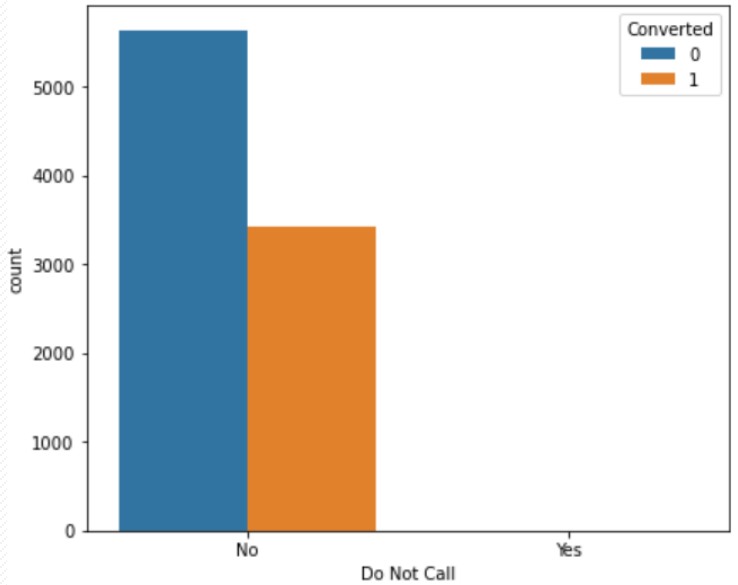
❖ *Observations:*

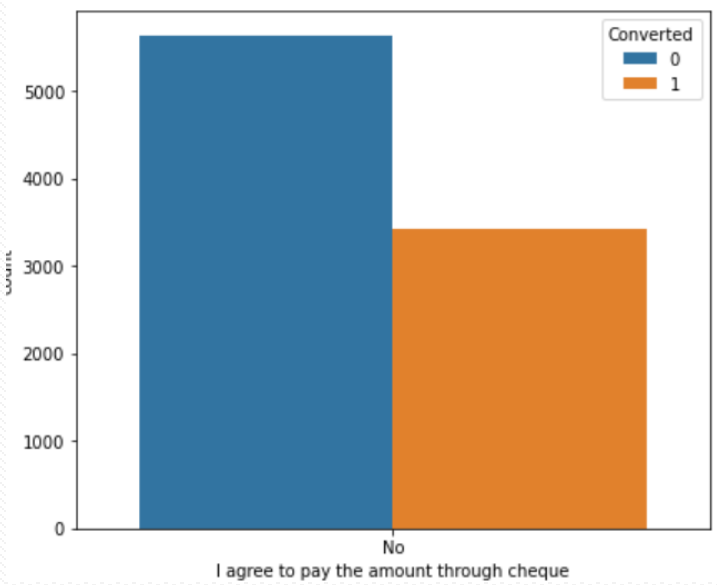
1. Data is highly skewed – we have mostly “No” category in the Newspaper.



Univariate Analysis







RECEIVE MORE UPDATES ABOUT OUR COURSES

- *Observations:*

1. Data is highly skewed – we have only "No" category

DO NOT CALL

- *Observations:*

1. Data is highly skewed – we have only "No" category

I AGREE TO PAY THE AMOUNT THROUGH CHEQUE

- *Observations:*

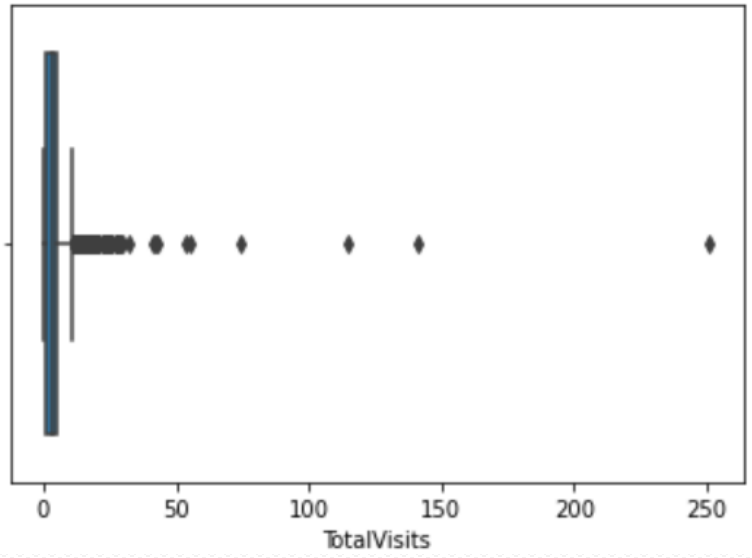
1. Data is highly skewed – we have only "No" category

Correlation Table

❖ *Observations:*

1. We note that “Total tine spent on Website” is highly correlated with “Converted” variable.
2. “Page Views per Visit” is highly correlated with “Total time spent on Website”.

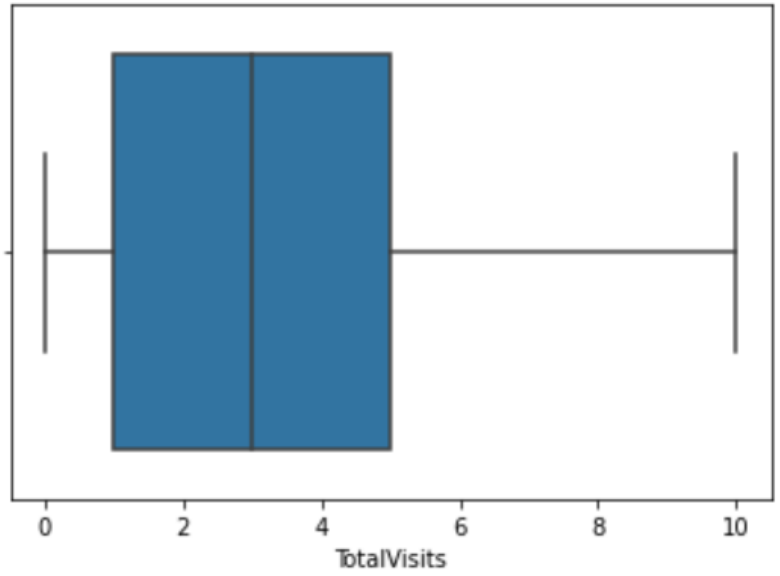
Outlier Analysis

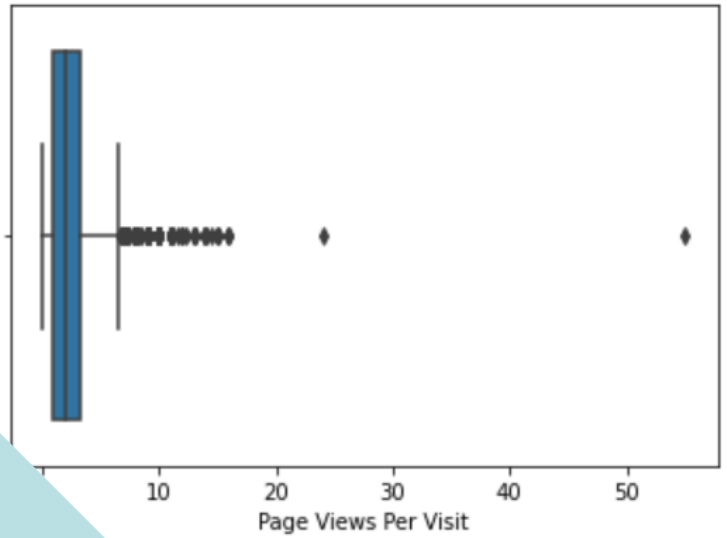


TOTAL VISITS

- *Observations:*

1. Few Outliers are noted – we cap the data at 95th percentile.

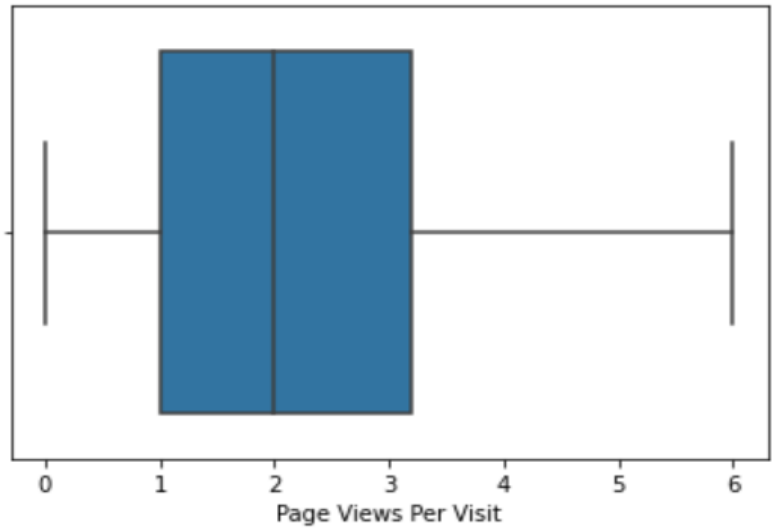




PAGE VIEWS PER VISIT

- *Observations:*

1. Few Outliers are noted – we cap the data at 95th percentile.



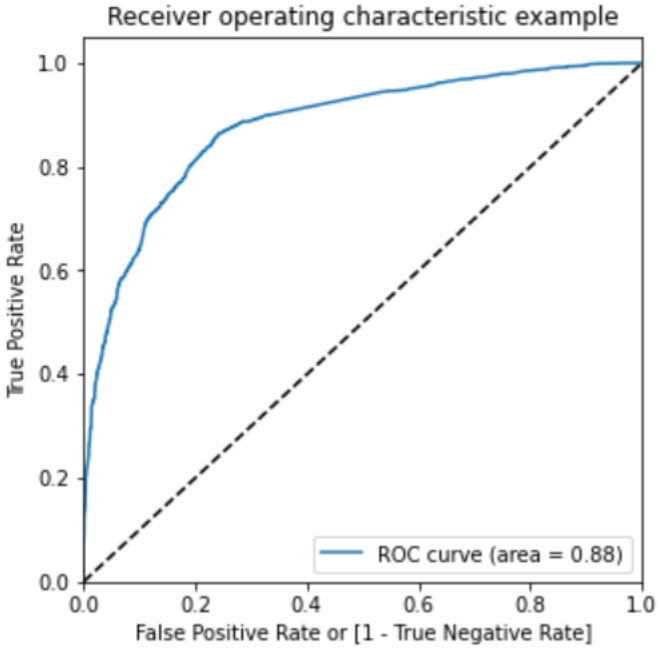
Model Building and Evaluation

- Split the data into train test sets
- Scale the variables in train dataset
- Using RFE select 15 variables to build the model on.
- Build the logistic regression model and follow the below:
  - If the p-value is high, drop the variable and rebuild the model.
  - Continue the above till all p-values are low.

ROC CURVE

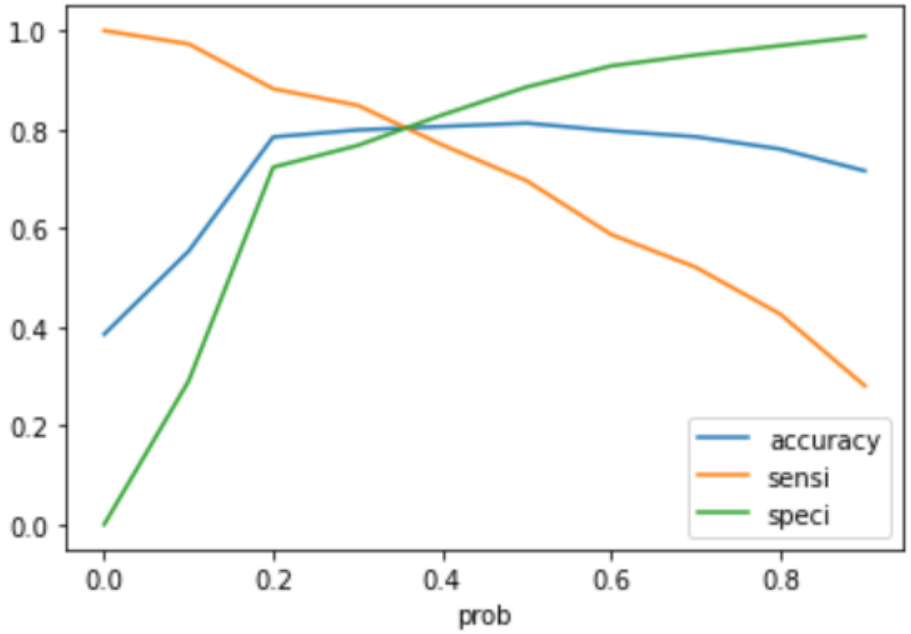
❖ *Observations:*

1. The ROC curve looks acceptable.
2. Area under the curve = 0.88



Model Building and Evaluation

Optimal Cutoff



❖ *Observations:*

1. Cutoff = 0.35 seems like a good option to predict the values on.
2. Overall accuracy of 80.40% is achieved using 0.35 as cut off for Converted probability.

❖ Using 0.35 as the cutoff, we find the metrics for the model evaluation:

Train Set:

- Overall Accuracy: 80.40%
- Sensitivity / Recall : 81.89%
- Precision: 71.41%

Test Set:

- Overall Accuracy: 79.76%
- Sensitivity / Recall : 80.89%
- Precision: 68.85%

Interpretation

- The business should focus on the leads who are being sourced from Welingak website.
- The total time spent on the website is directly proportional to the lead being hot. The business should focus on spending to make website UI more engaging.
- When the last notable activity was:
  - Conversation on phone.
  - Chat on Olark Chat
- People who have their occupation as “Working Professional” should be focused more upon.
- When the lead is sourced through references – Business can create more incentives and campaigns to encourage references from existing customers.

THANK YOU