LEAD SCORING CASE STUDY

by

Shivabasav Aursang

Ritik Patel

Akash Adrashannavar

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- Though X Education gets a lot of leads, its lead conversion rate is very poor. For example, if they acquire 100 leads in a day, only 30 of them get converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- 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.

AGENDA

- To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- We will also try to check if certain adjustments or recommendations can be made to ensure that we are suggesting the right business solutions to tackle future problems as well.

PROBLEM SOLVING APPROACH

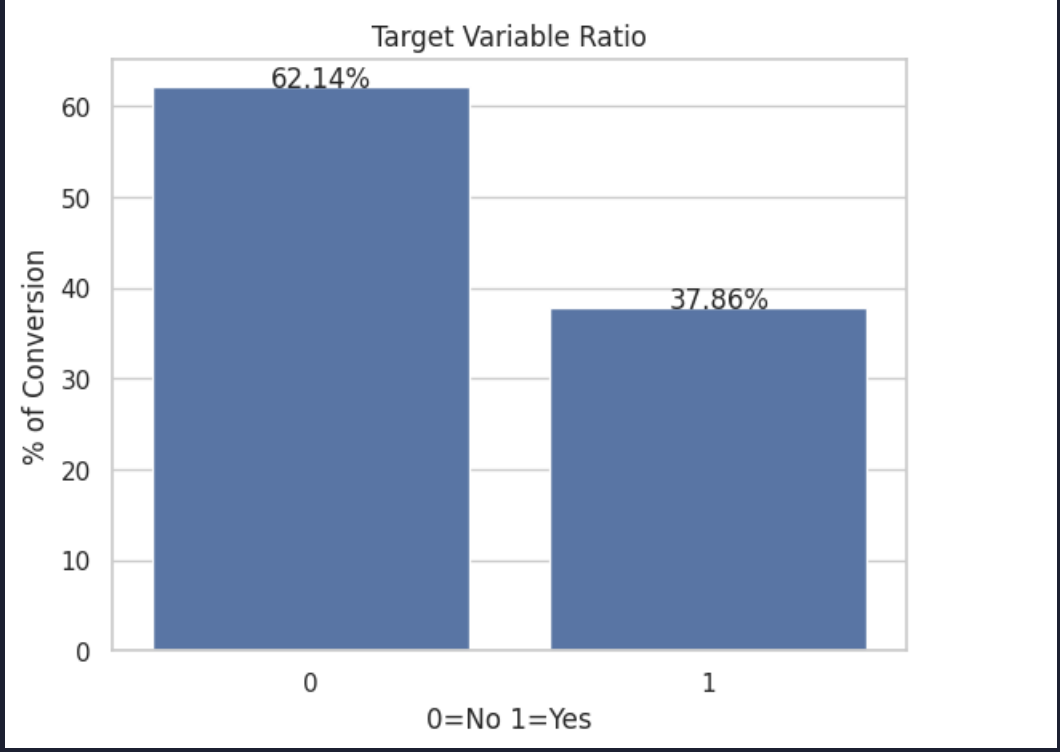
- Source the data for analysis
- Clean and prepare the data by dealing with missing and duplicate values and checking for outliers.
- Perform Exploratory Data Analysis by conducting both Univariate and Bivariate Analysis.
- Feature Scaling and creation of Dummy Variables
- Split the data into test & train dataset
- Model Building
- Model Evaluation – specificity & sensitivity or precision-recall
- Making predictions on the test dataset.

DATA PRE-PROCESSING

- Data is in CSV format, with 9240 rows and 37 columns.
- There are 4 columns with the float data type, 2 as integer, and the rest belong to the object data type.
- It has been observed that a lot of columns have the value ‘Select’ in them. They are being converted to Null values to give us a true picture of the data.
- Columns with more than 30% of null values are dropped as a significant number of entries are missing.
- Similarly, rows with less than 3% of null values are removed as they are fewer in number, and removing them will yield a better analysis.

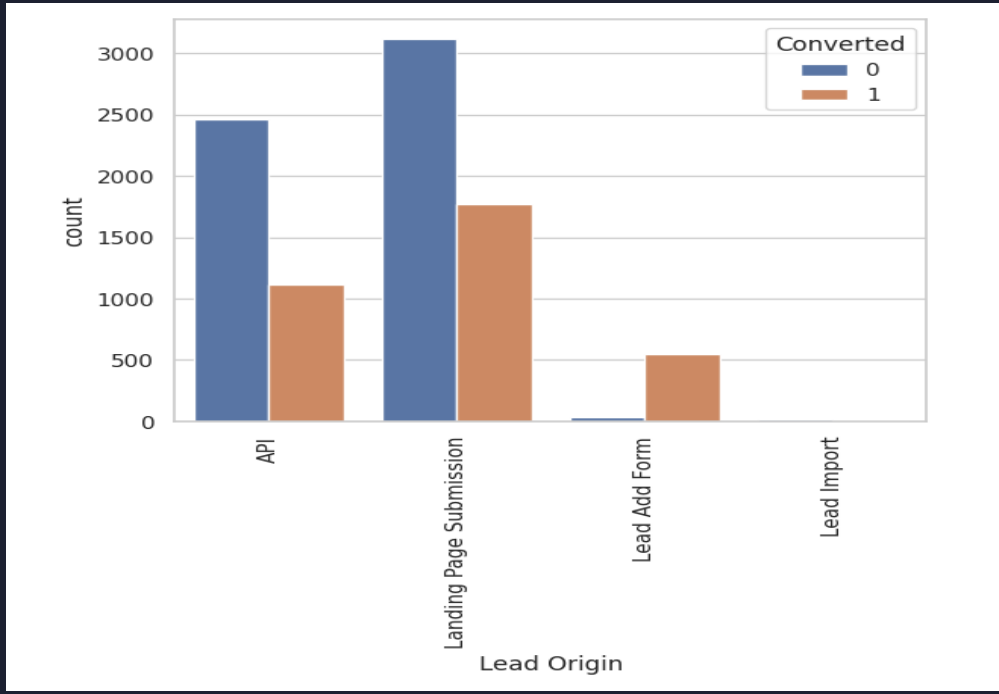
- In the ‘Country’ column, 25 percent of data is missing, and out of the existing data, 95% of data shows ‘*India’*. This won't provide significant insights to our analysis, so it's better we drop it.
- In the ‘What is your current occupation’ column, the missing value was imputed using its Mode – ‘*Unemployed* ‘
- ‘Prospect ID’ and ‘Lead Number’ denote the ID numbers of the contacted people. Hence, they can be dropped as we cannot analyze using them.
- Columns where more than 90% of the data denote a single value have been dropped as high skewness in data would not lead to a fair analysis.
- Outliers in Numerical columns such as 'Total Visits’ and 'Page Views Per Visit’ were treated by removing values above the 97^{th} and 99^{th} percentile respectively.

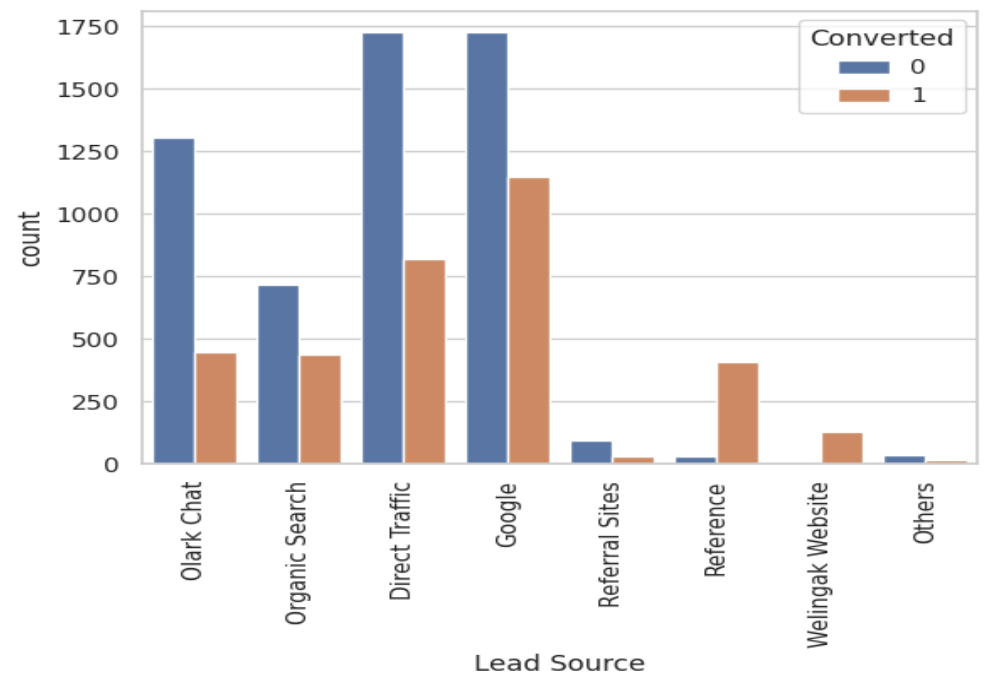
EXPLORATORY DATA ANALYSIS



We can observe that around 37% of the leads have been converted into customers.

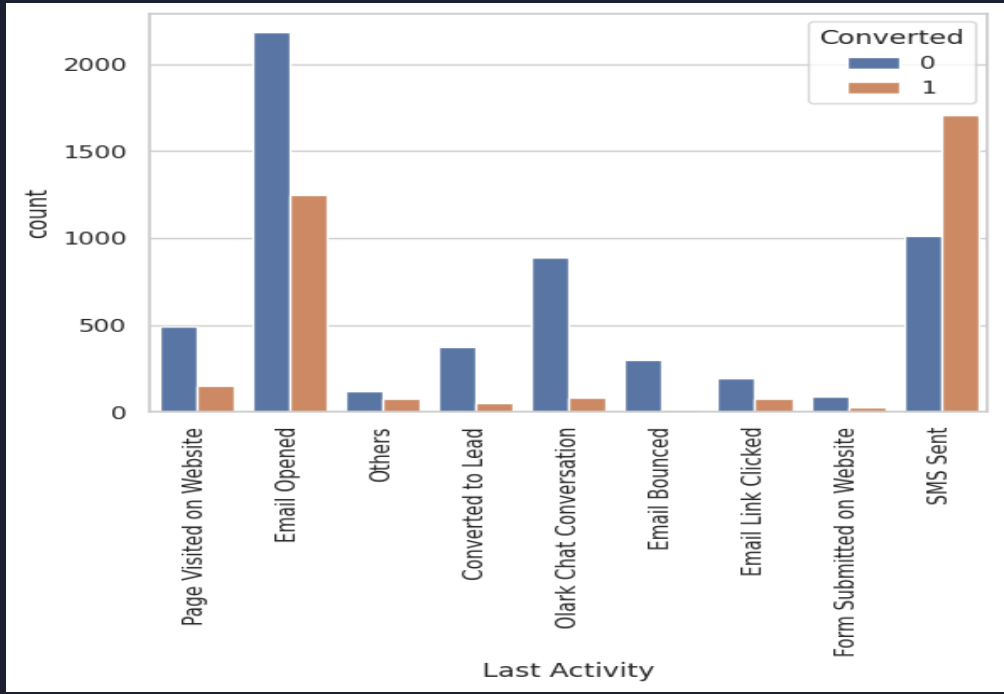
CATEGORICAL VARIABLES



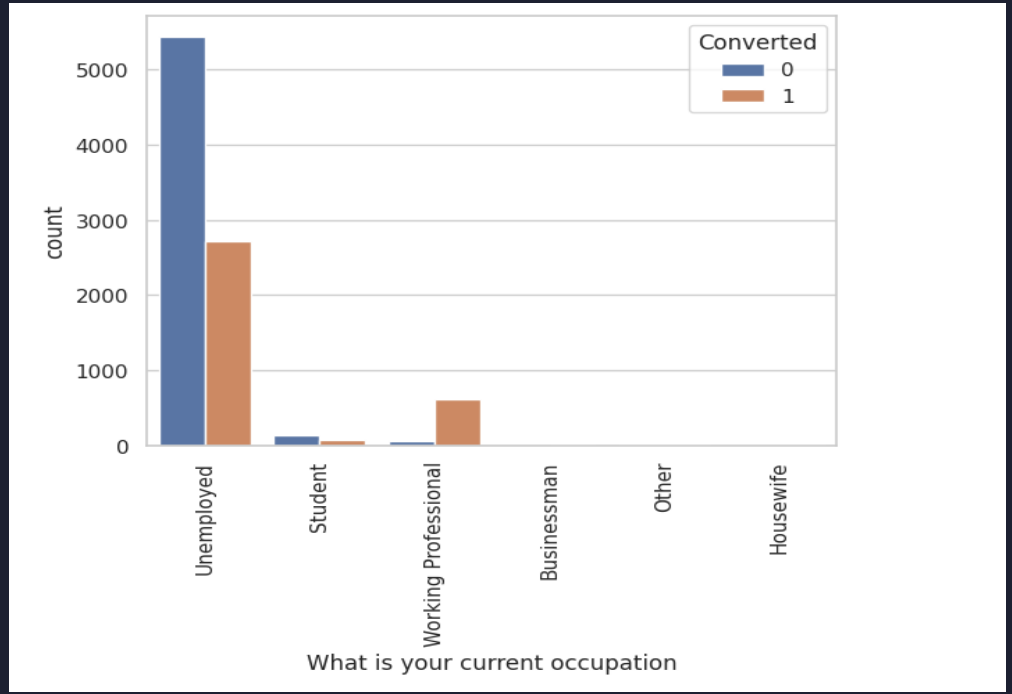


`API` and `Landing page` have brought in maximum leads, but the leads from `Lead Add Form` are not high but their conversion is phenomenal.

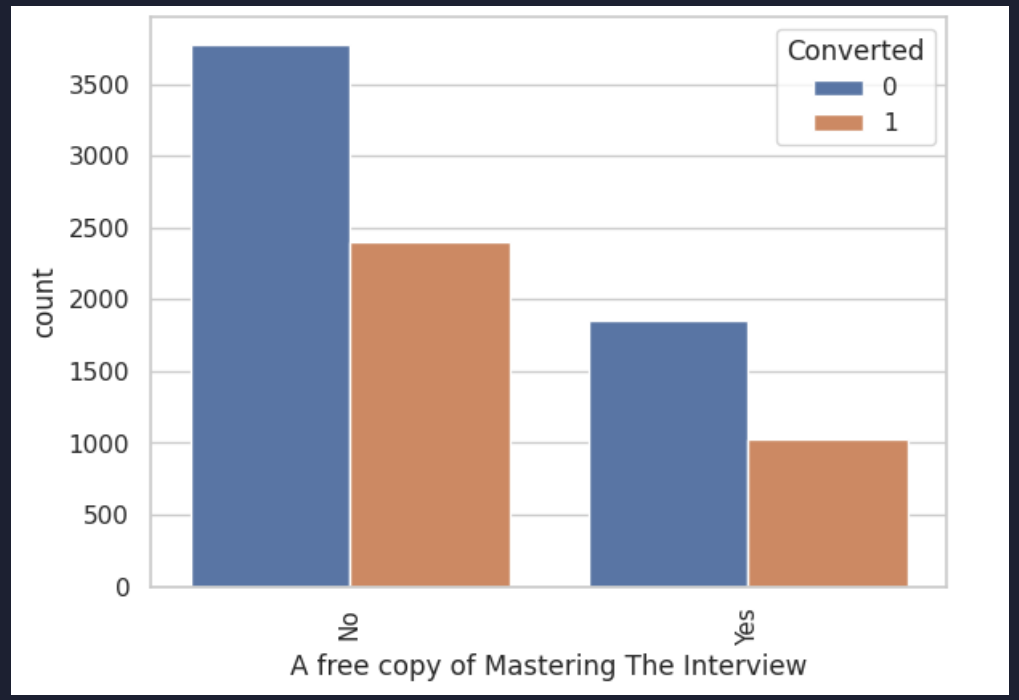
Maximum leads are brought in by `Direct Traffic` and `Google`, whereas `Welingak` and `Reference` have the best conversion rate.

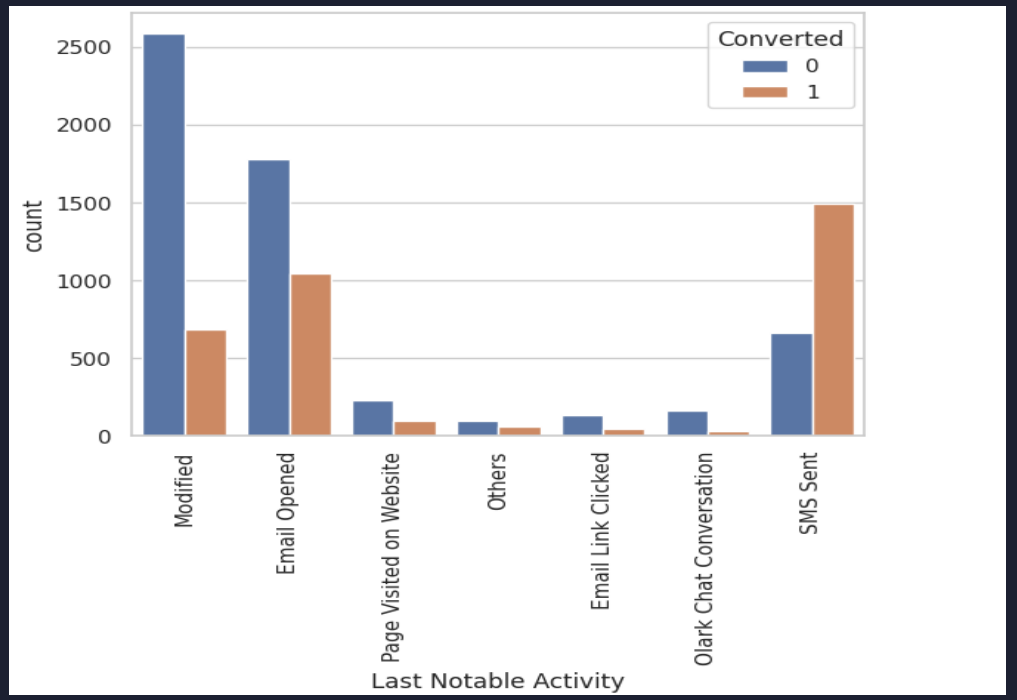


The conversion rate of `SMS Sent` is out of the roof, while measures can be taken to convert leads from email.



The conversion rate of `Working Professionals` is above par, they can be reached out more, while strategies to derive customers from the `Unemployed` sector can be executed.

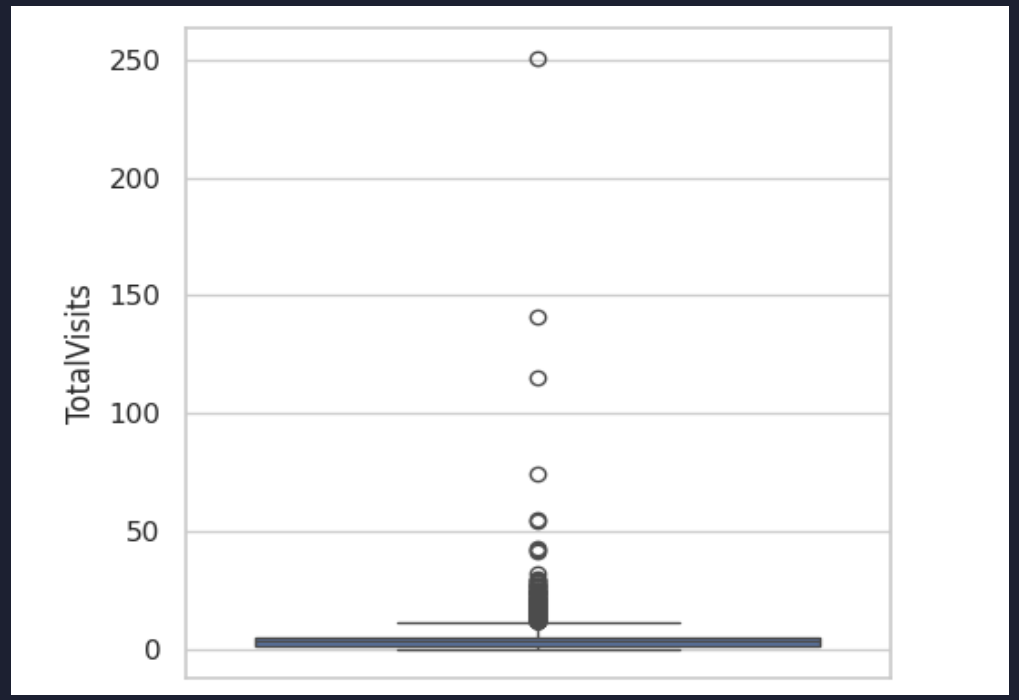




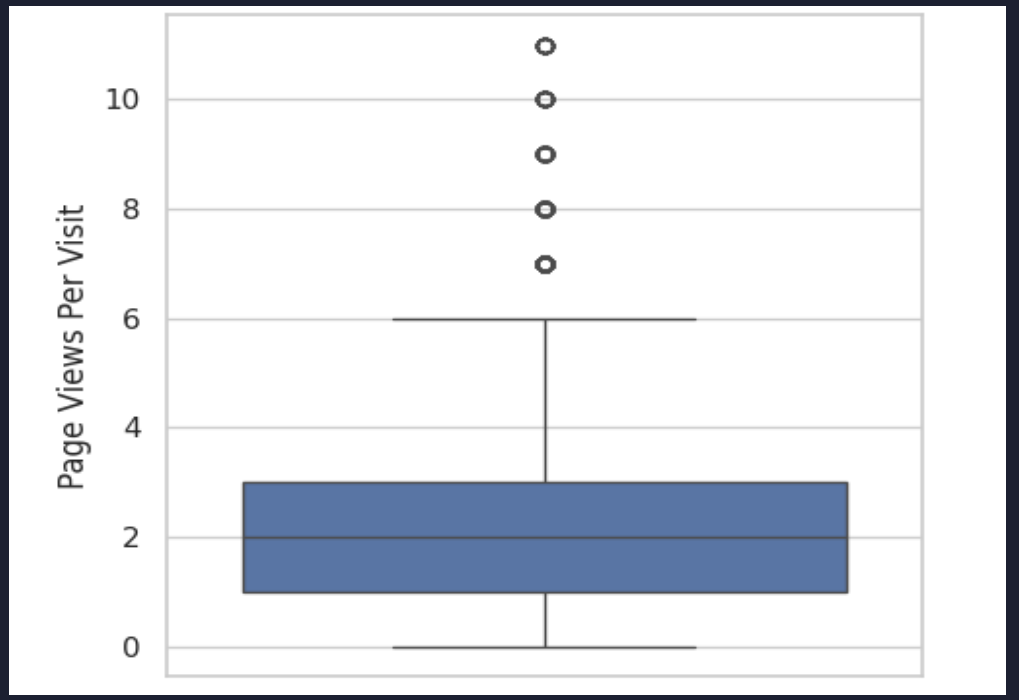
More leads are generated from those who do not ask for a free copy of Mastering Interviews. They can be focused on conversion.

As seen earlier, `SMS Sent` has the highest conversion, we could focus more on `Email Opened` and `Modified`

OUTLIERS IN NUMERICAL COLUMNS

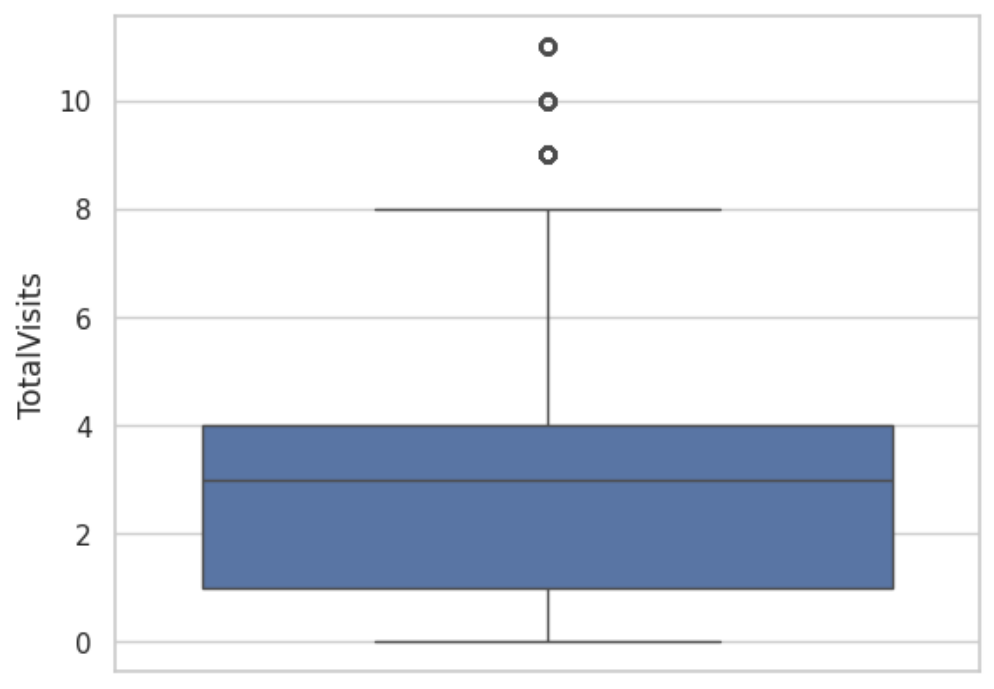


Removing values above the 97th percentile.

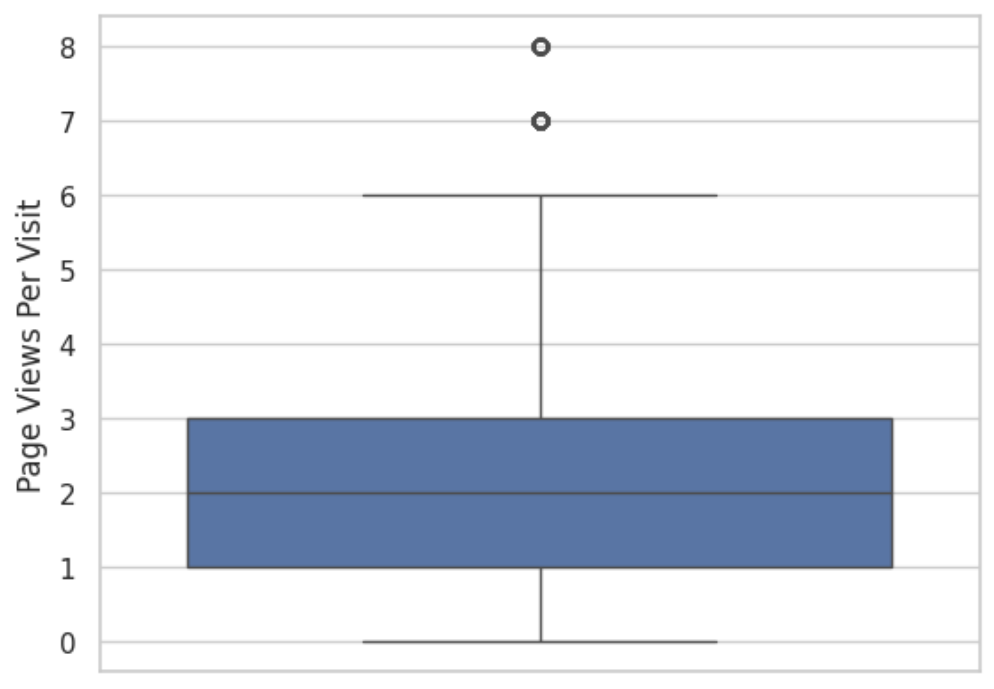


Removing values above the 99th percentile.

NUMERICAL COLUMNS AFTER CORRECTION



Removing values above the 97th percentile.



Removing values above the 99th percentile.

HEATMAP

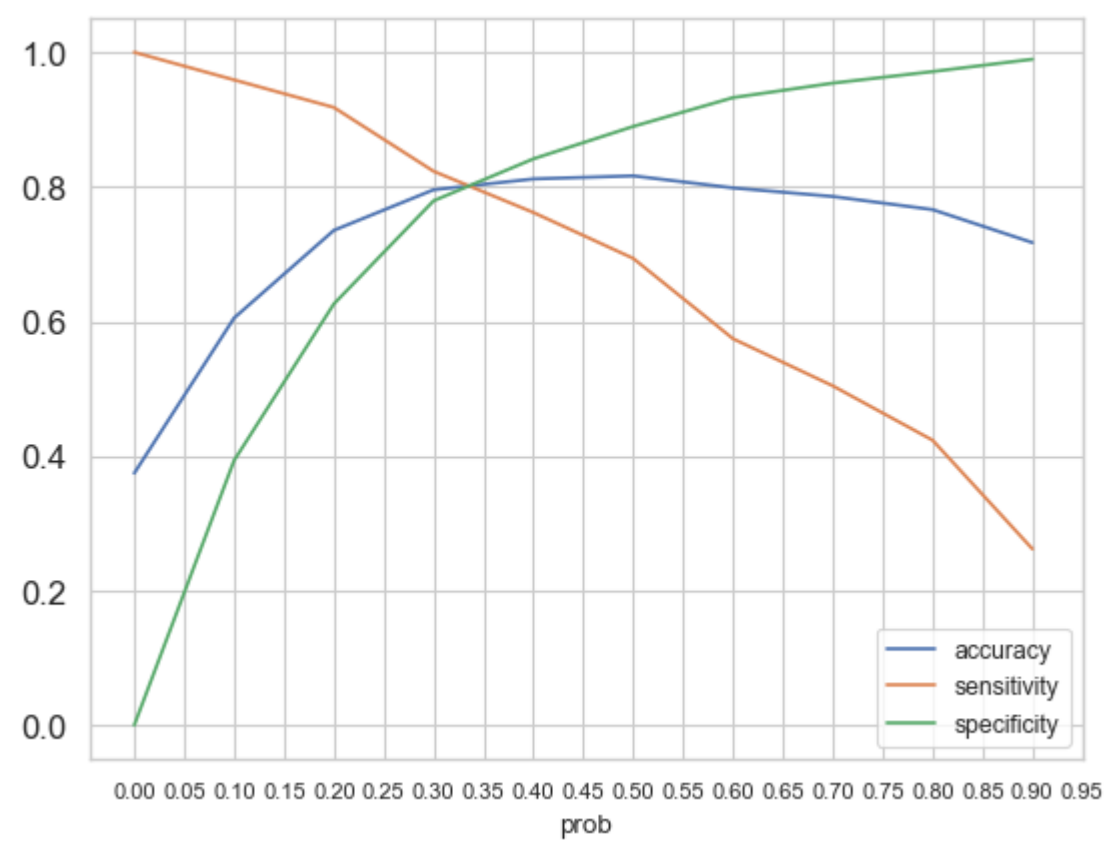
- There is a very strong correlation between `Total Visits` and `Page Views per Visit`, which is around 0.77

- The correlation between `Converted` and `Page Views Per Visit` is negligible and negative.

MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 80:20 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 10
- Predictions on test data set
- Overall accuracy 80%

MODEL EVALUATION (TRAIN DATASET)

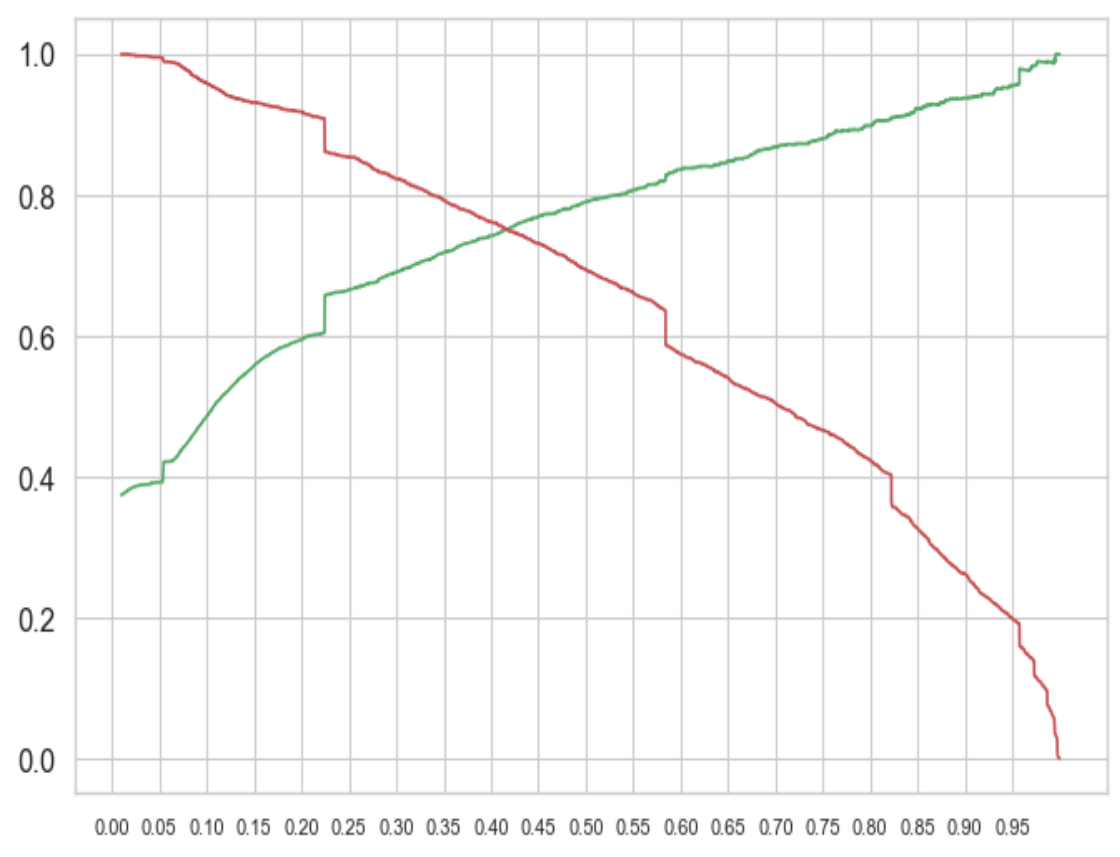


Accuracy - 80.26%

Sensitivity - 80.67%

Specificity - 80.01%

PRECISION AND RECALL



Precision - 75.37%

Recall - 74.84

RECOMMENDATIONS

- Initiatives should be taken to get more leads from Lead Add Form as their conversion rate is above par.
- The website can be made more interactive to engage customers and increase the overall time spent on the website, which will indeed increase the conversion as observed in the EDA.
- Working Professionals can be targeted more, as they have the spending capacity and willingness to upskill themselves which can also be seen in the analysis.
- Students can be avoided as they are already enrolled in a course and might not purchase a course for industry professionals this early in their careers.
- Loyalty programs can be initiated for our existing customers to get more references from them as the conversion rate is very high.
- Similarly, leads whose last Notable Activity was SMS and who visit the website repeatedly on a regular basis can be nudged towards conversion.

CONCLUSION

- Our model development process involved several crucial steps to ensure accuracy and reliability. Initially, we applied one hot encoding to convert categorical variables into a suitable format for modeling.
- Subsequently, we utilized Recursive Feature Elimination (RFE) to identify the most significant features, enhancing our model's predictive capability.
- Building on this foundation, we constructed logistic regression models and refined them iteratively based on improving p-values. This iterative process enabled us to fine-tune the models for optimal performance.
- This approach led to an impressive accuracy rate of around 80% and a precision rate of around 75%, showcasing the effectiveness of our method in predicting lead conversions.
- To validate our model's performance, we rigorously tested it on unseen data, confirming its reliability and applicability in real-world scenarios.
- Overall, our systematic approach and meticulous attention to detail have yielded a robust predictive model capable of providing valuable insights for enhancing lead conversion rates and optimizing resource allocation strategies.

THANK YOU!