Lead Scoring Case Study

SUCHINDRAN VARADARAJ

UMANG MALHOTRA

VIKASH KR. TRIPATHI

Problem Statement

Problem Statement

Background:

- X Education company sells online courses and gets its initial leads from various modes of online marketing.
- The company wishes to identify Hot Leads, who have high chances of conversion.
- It would help sales team focus on nurturing hot leads rather than all available ones.
- Their present lead conversion rate is 30% and their target is 80%.

Goal:

- An ML model to assign lead score from 0 to 100 to each lead.
- Higher the lead score, higher are the chances of lead conversion.

Analysis Approach

Data Analysis

Data Cleaning:

- Drop features with >40% null values (including 'Select')
- Impute remaining categorical null values with 'Not Provided' string
  - Imputing with mode would impact data distribution in such columns
- Impute remaining numerical null values with respective median

Exploratory Data Analysis:

- Univariate Analysis: Histograms, Boxplots and Bar plots to see distribution and remove outliers and single category variables
- Bivariate/Multivariate Analysis: Boxplots, Heatmap and Stacked Bar plots to visualize correlation of various features with lead conversion.

ML Modeling

Feature Engineering:

- Dummy variables creation for categorical features
  - This gives higher number of attributes to model on compared to Label Encoding.

Model Building:

- Min-Max Scaling to bring all numerical values in similar range (sklearn).
  - Scaling make numerical data better fit and helps improve model accuracy.
- Feature selection using RFE on Logistic Regression (sklearn).
- Statistical summary based on the recommended features from RFE (statsmodels).
  - Feature elimination based on p-values and VIF scores of different features.

Model Evaluation

Training Dataset:

- Evaluation of model on training data with various metrics (sklearn)
  - Accuracy Score: 81.57% & ROC-AUC Score: 88.97%
  - Precision: 79% accurate on conversion
  - Recall: 71% accurate on conversion
  - Area under ROC curve: 89%

Test Dataset:

- Evaluation of model on test data with various metrics (sklearn)
  - Accuracy Score: 81.85%
  - Precision: 80% accurate on conversion
  - Recall: 71% accurate on conversion
  - Confusion Matrix: TP: 762, TN: 1507

Lead Scores & Key Points

Generating Scores:

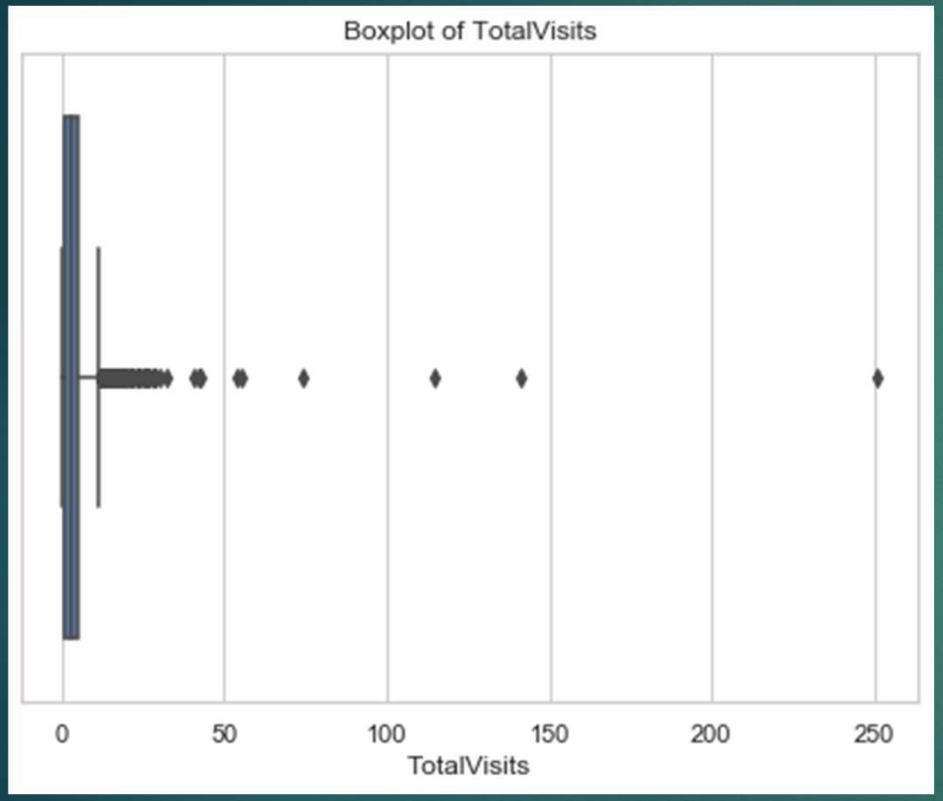
- Based on test data probability information, scores fall between 0 to 100.
  - Scores beyond 80 can be considered the leads with highest chances of conversion.
  - Score between 50 to 80 can be considered the leads with moderate chances.
  - Score below 50 can be considered the leads with lowest chances of conversion.

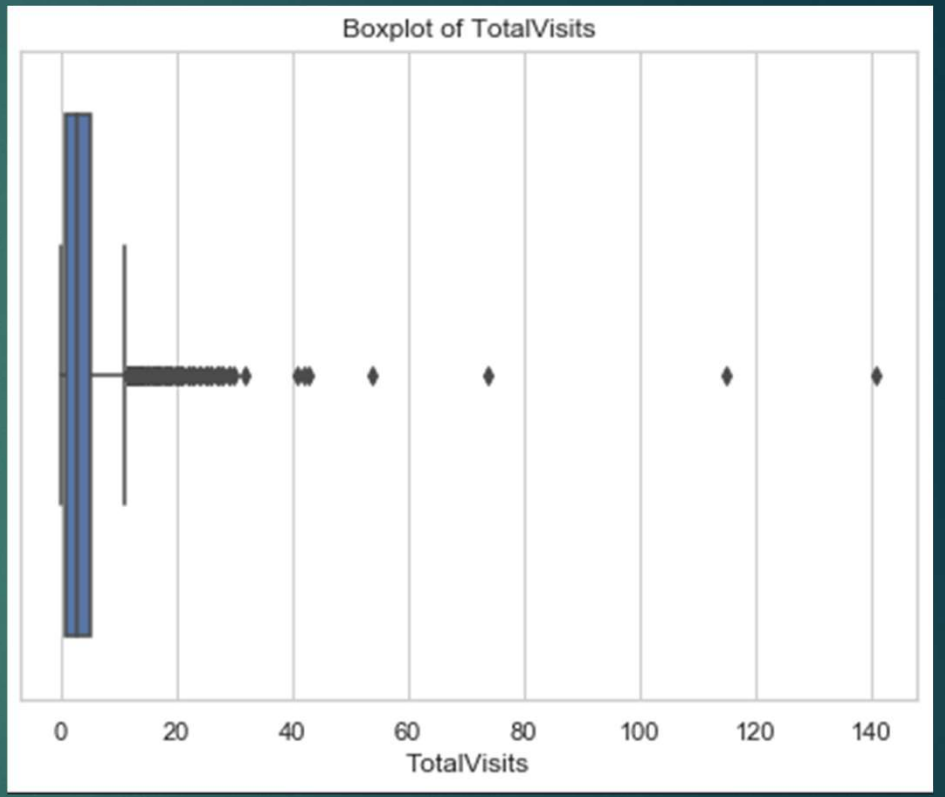
Key Points:

- Identified a few (among others) top predictors contributing in lead conversion.
  - Welingak Website as Lead Source
  - Total Time Spent on Website
  - Working Professionals
  - Lead Add Form as Lead Origin
  - Phone Conversation as Last Notable Activity

Important Visualizations

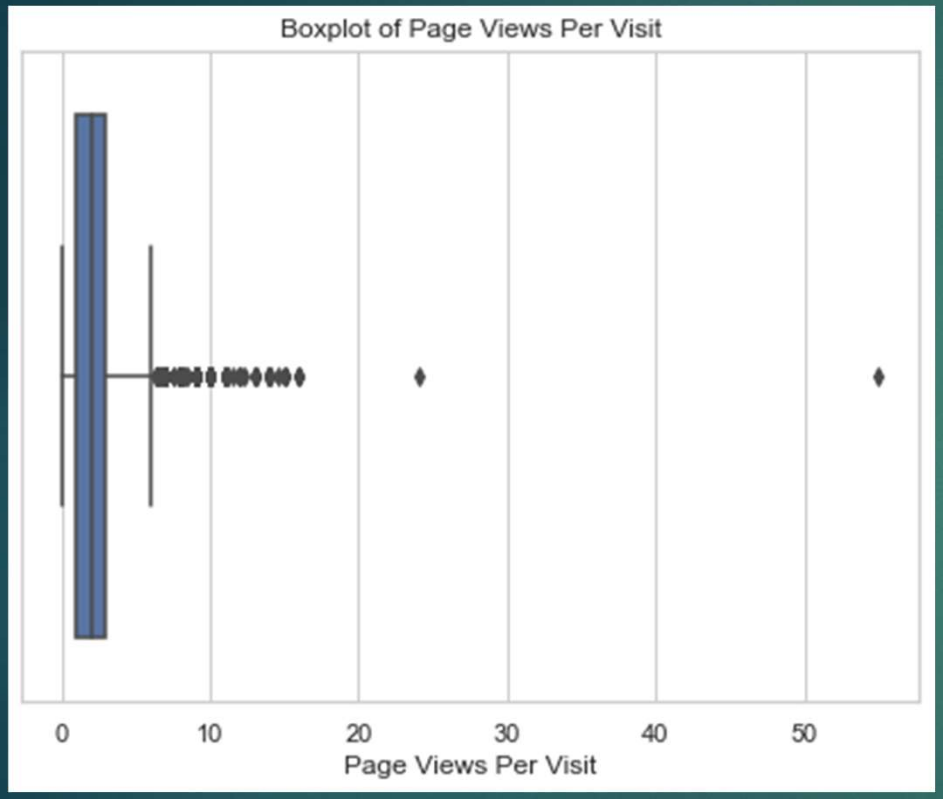
Univariate Analysis – Outlier: 'TotalVisits'

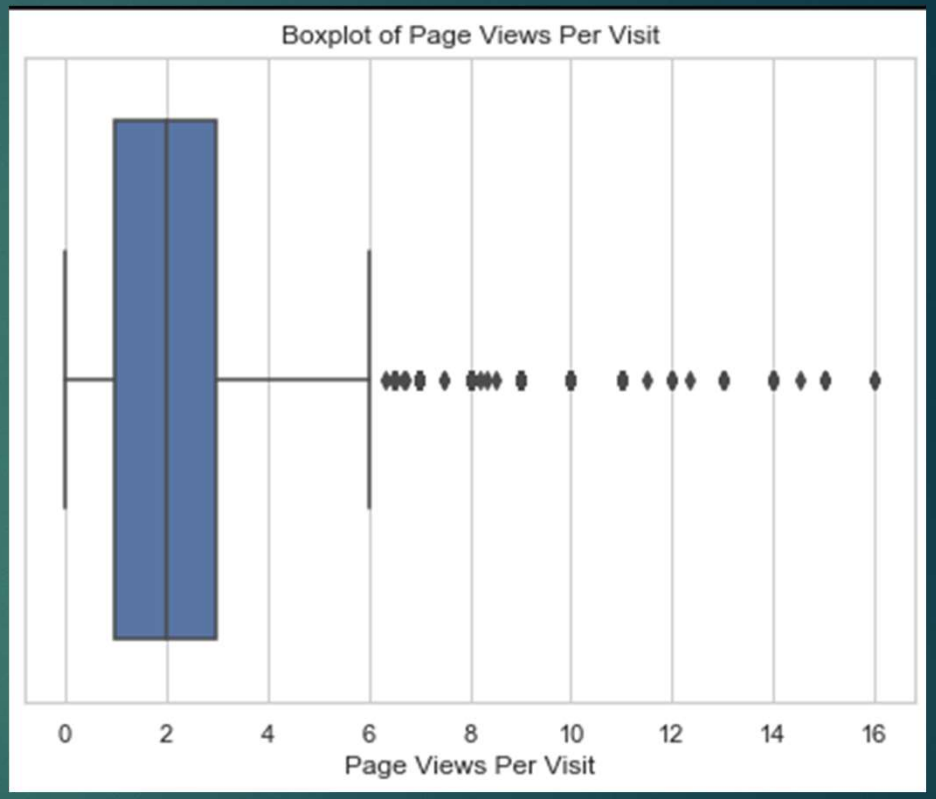




'TotalVisits' feature indicated presence of outliers, which got treated well by limiting the values within 150.

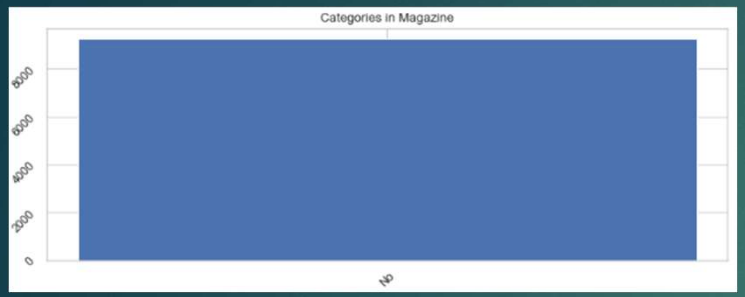
Univariate Analysis – Outlier: 'Page Views Per Visit'

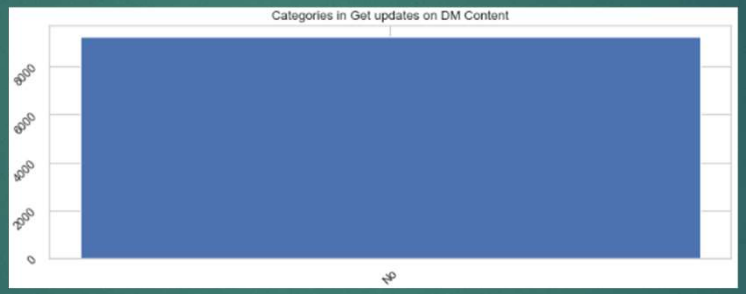




'Page Views Per Visit' feature has some outliers too, which were treated by keeping the values within 20.

Univariate Analysis – Single Category Features

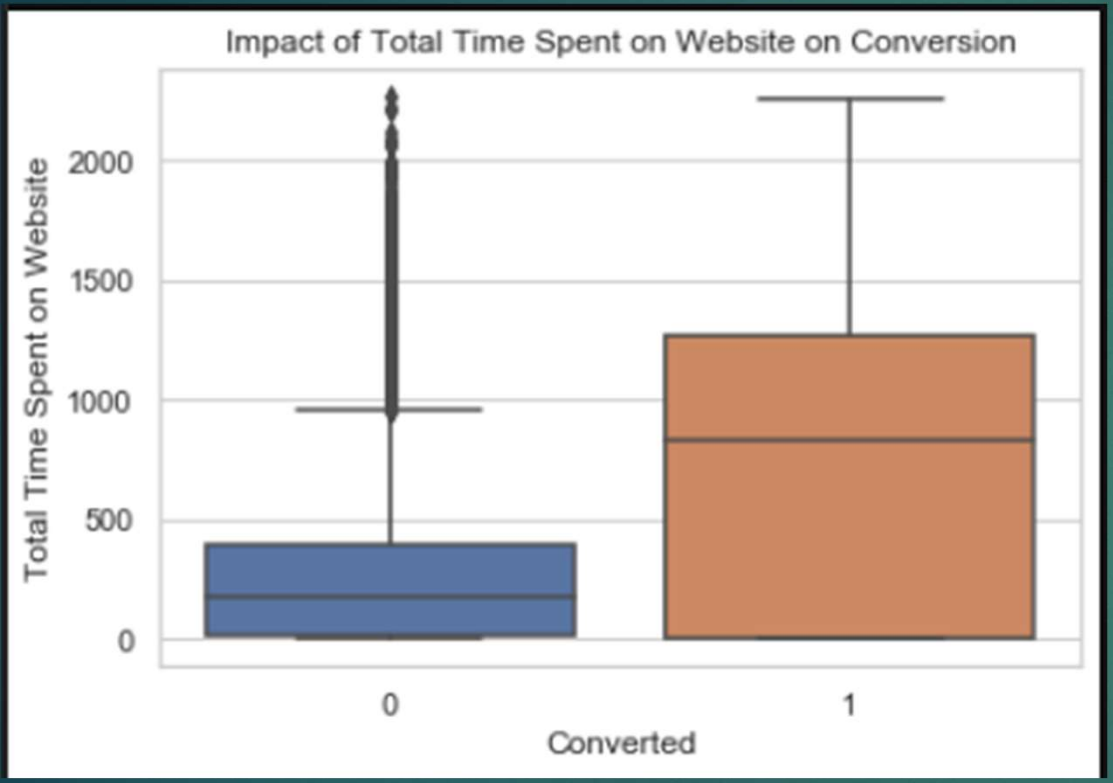


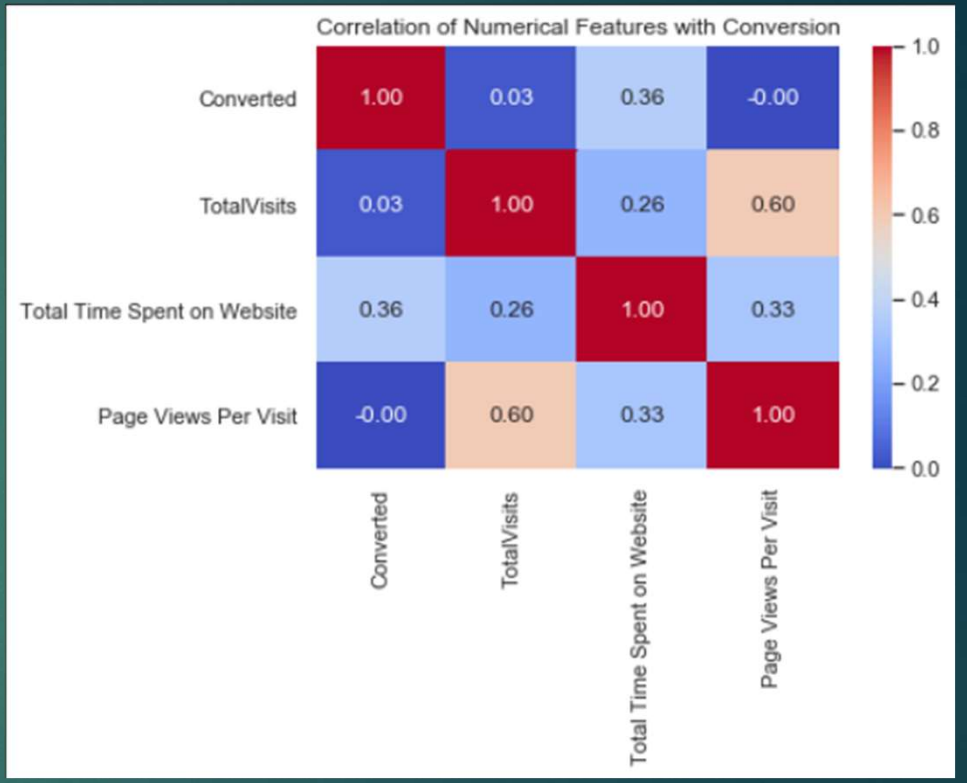


Some categorical features had only one categorical value. These would not make sense for the ML modeling, so dropped.

- 'Magazine',
- 'Receive More Updates About Our Courses',
- 'Update m on Supply Chain Content',
- 'Get updates on DM Content',
- 'I agree to pay the amount through cheque'

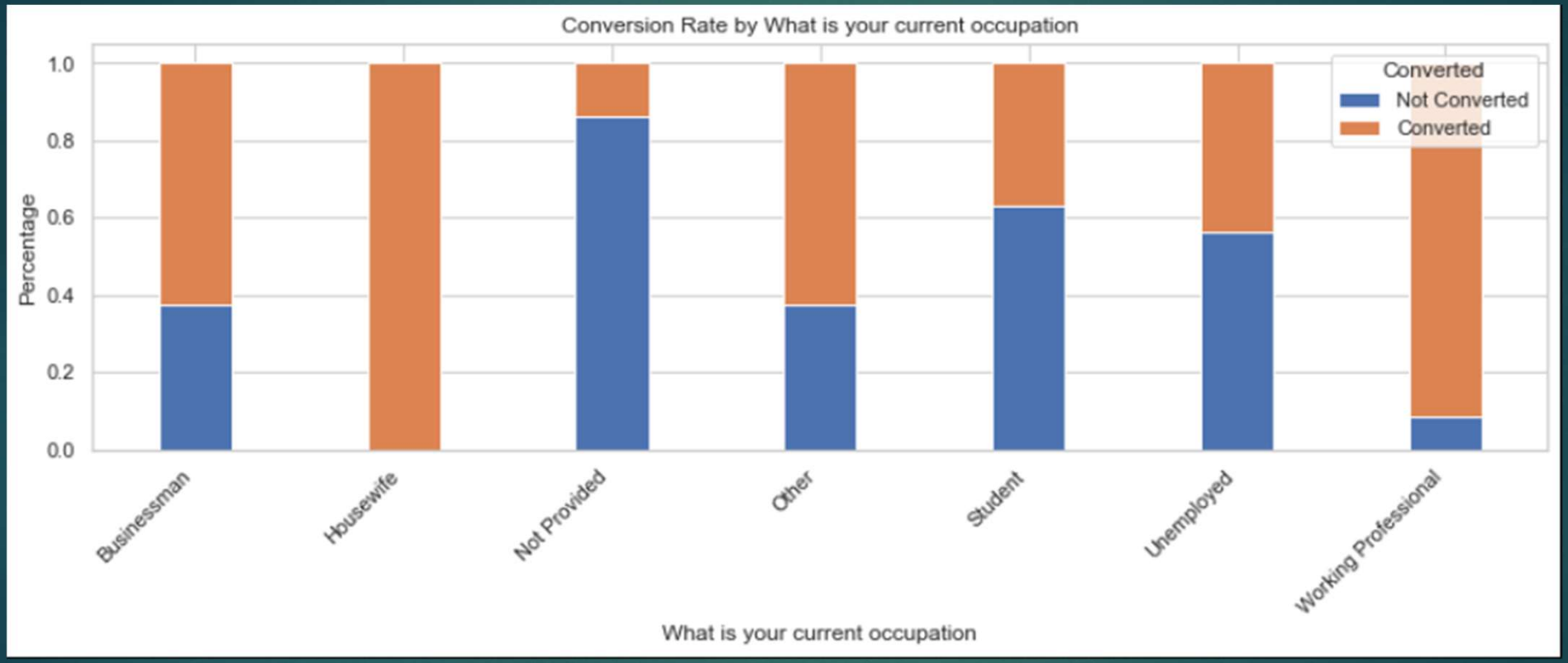
Bivariate Analysis – 'Total Time Spent on Website'





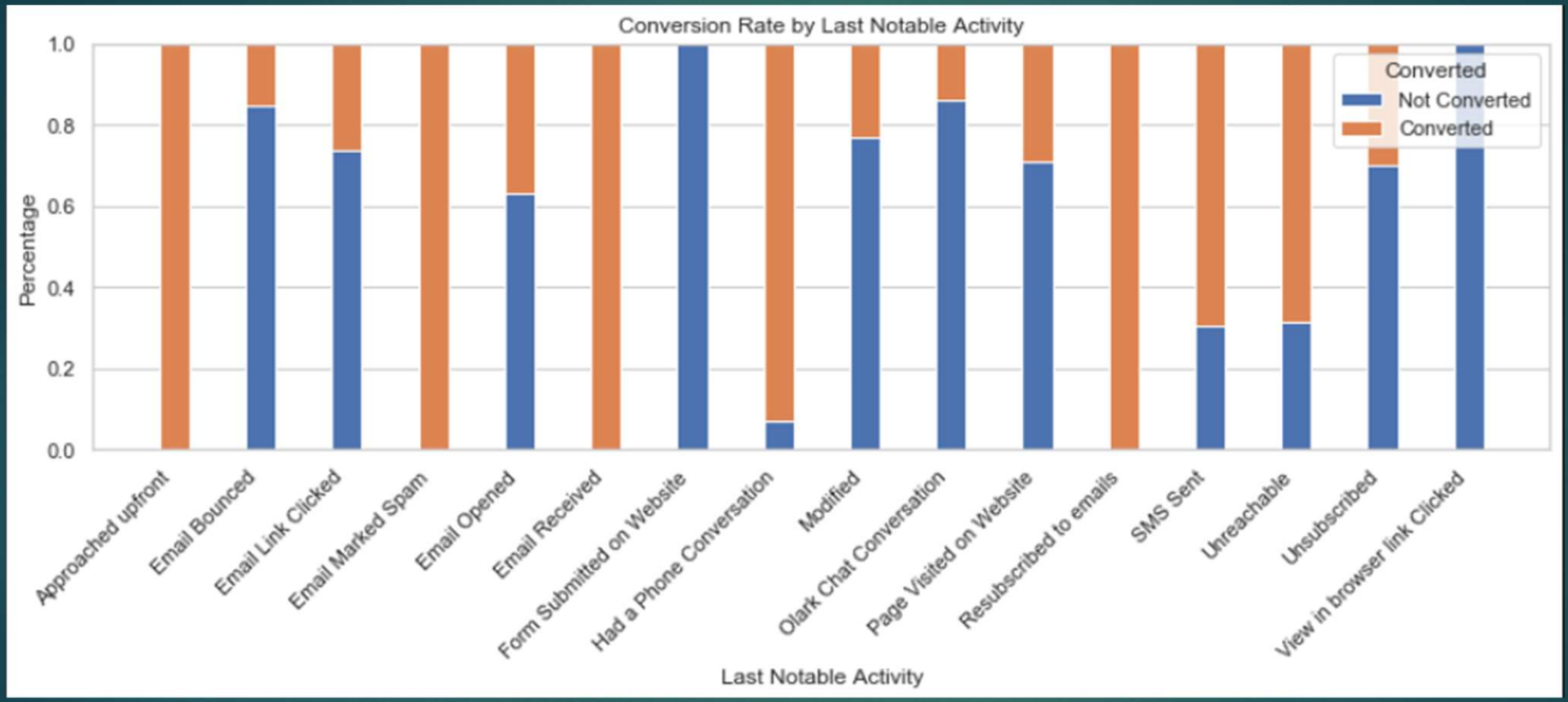
'Total Time Spent on Website' feature is seen having a good correlation with the target 'Converted'.

Bivariate Analysis – 'Occupation'



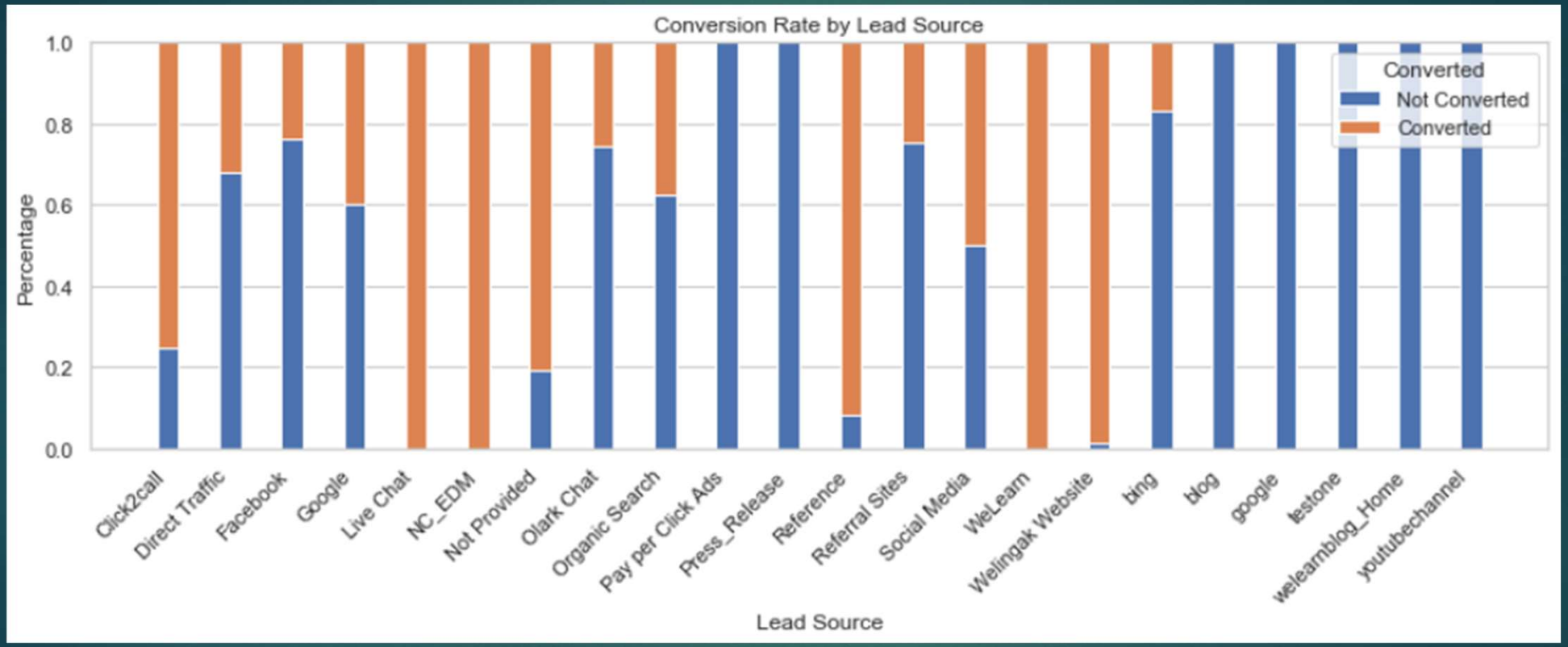
'Working Professionals' are observed to have quite high chances of lead conversion.
Housewife and Businessman also indicate a good correlation with target 'Converted'.

Bivariate Analysis – 'Last Notable Activity'



'Leads who had a phone conversation lately' have quite fair chances of lead conversion.
Among others, 'SMS Sent' & 'Resubscribed to emails' also indicate a good correlation with 'Converted'.

Bivariate Analysis – 'Lead Source'



Leads who approached from Welingak website, WeLearn or References show a good chance of lead conversion.
Leads who were found on 'Live chat' or 'Click2call' options, also indicate good correlation with 'Converted'.

Results in Business Terms

Business Insights

Key Findings:

- The most influencing factors behind a lead conversion are total visits to and time spent on website.
- Leads visiting Welingak website have high chances of getting converted.
- Direct traffic on X Education website and platforms such as google or facebook have negative influence on the lead conversion.
- Working professionals among occupation categories are the highest potential leads to buy a course with X Education.
- Student and Unemployed leads have fairly good influence on lead conversion.
- Leads whose recent engagement has been a phone call have high chances of conversion.
- Leads with recent interaction in form of SMS have fairly good chance of conversion.
- Leads who didn't opt for email, are very less likely to get converted.
- Leads contacting through Olark chat platform have negative impact on the lead conversion.

Business Recommendations

Key Actions:

- Enhance website content and user experience to encourage longer visits and more interactions.
- Prioritize marketing efforts for leads visiting the Welingak website.

- Shift marketing budget away from direct traffic sources (Google/Facebook) to more productive channels.
- Develop targeted campaigns for working professionals as they show the highest conversion potential.

- Prioritize working professionals and students/unemployed leads for proactive outreach.
- Use automated engagement for leads with lower likelihoods.

- Prioritize phone calls for recent interactions as they have the highest conversion likelihood.
- Increase SMS engagement strategies for nurturing leads.

- Encourage leads to opt-in for emails through incentives.
- Personalize email campaigns to improve conversion rates.

- Reassess the effectiveness of Olark chat and explore alternative communication tools.
- Identify reasons behind the negative impact and refine chat engagement strategies.

Thank You