LEAD SCORING CASE STUDY

GROUP MEMBERS:

MALAIKA GOVEAS
SANDHYA SAHU

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are 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.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.

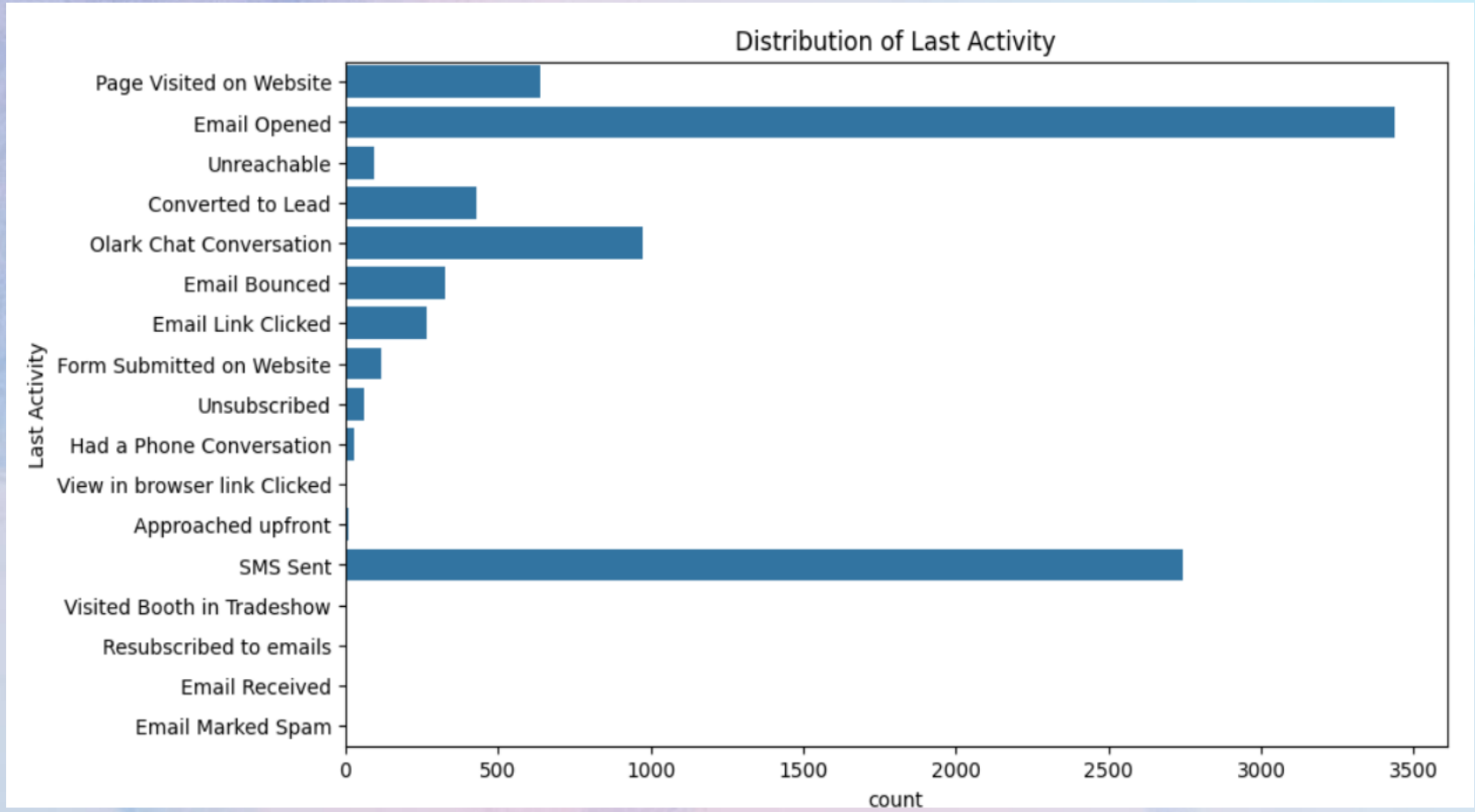
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

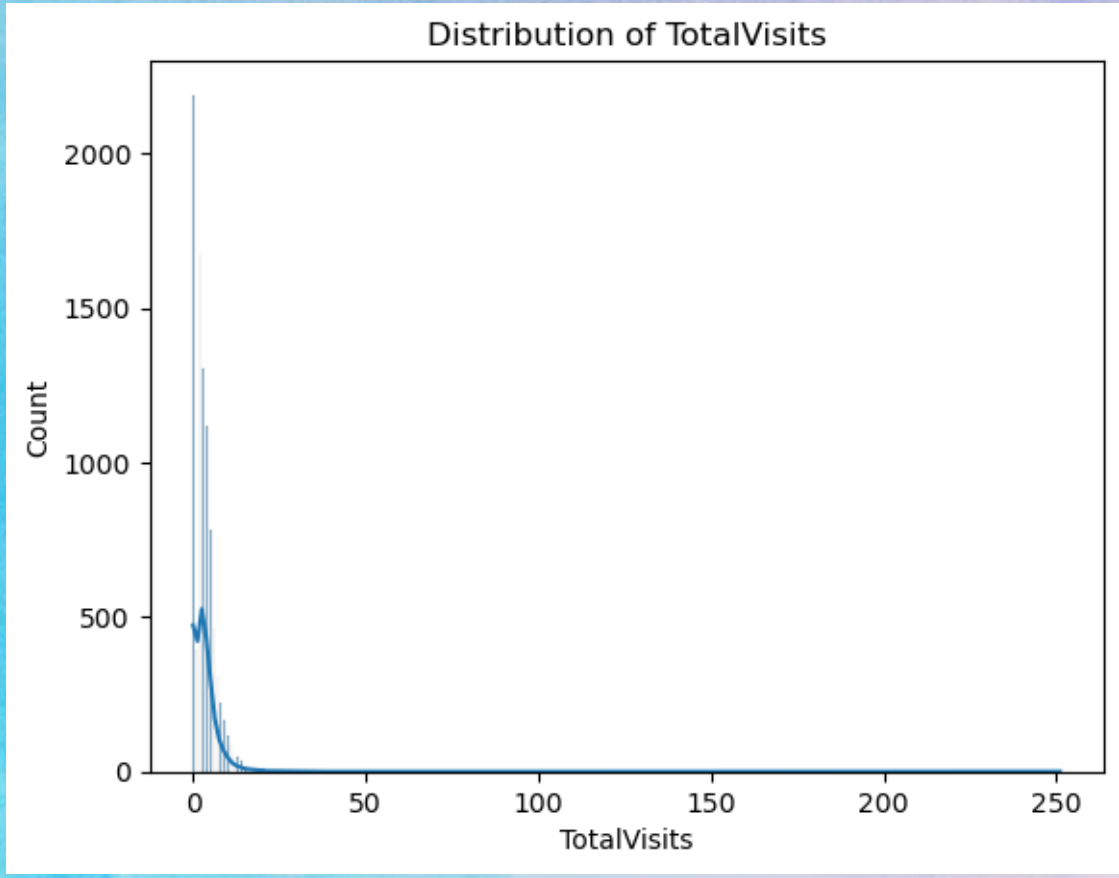
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.

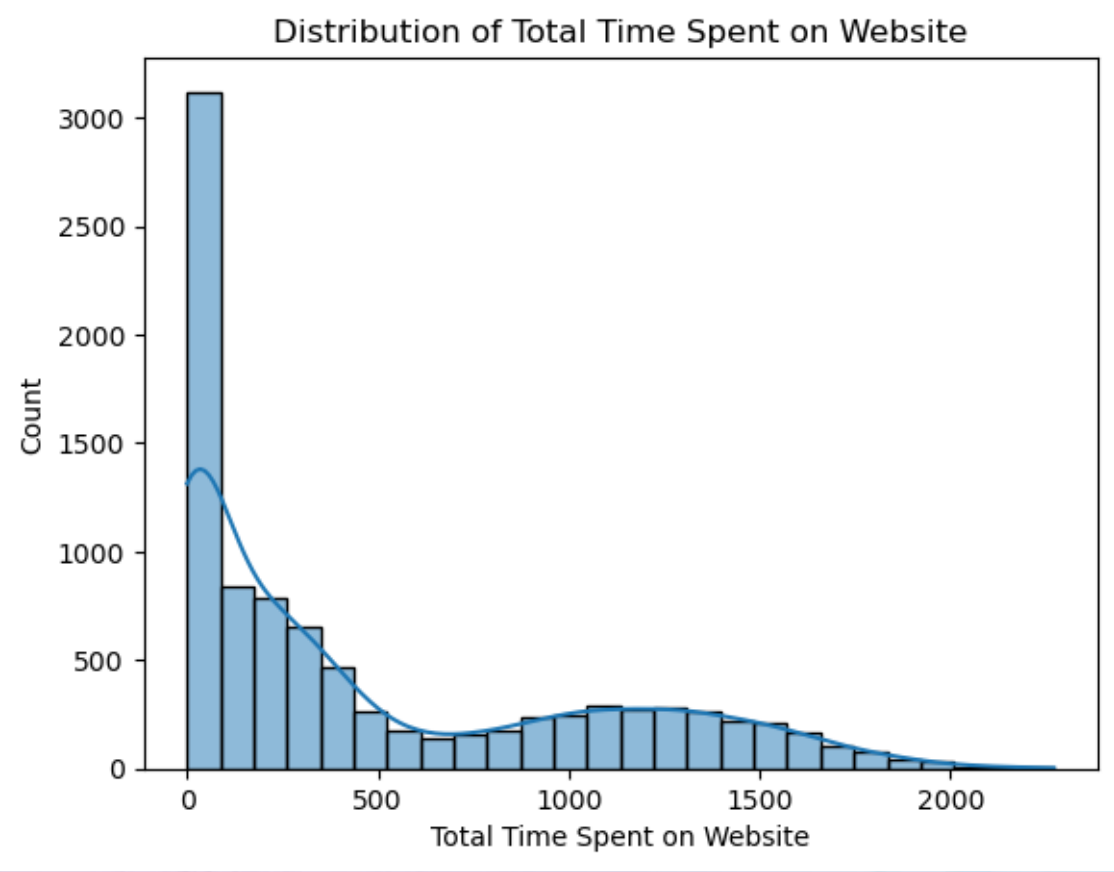
EDA

This slide shows the varied last recorded activities of leads, such as page visits or email interactions, which reflect engagement and interest levels.

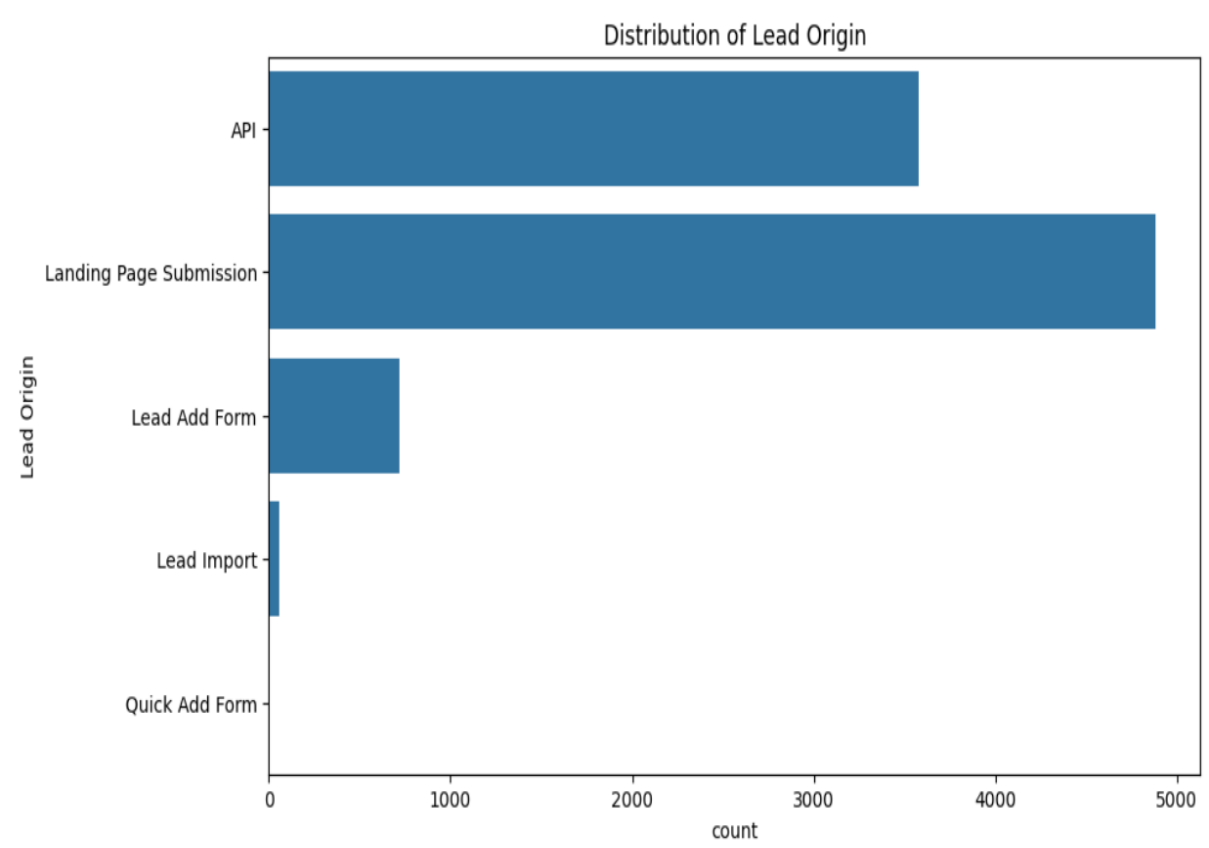
Such metrics can be pivotal in scoring leads and prioritizing follow-up actions.

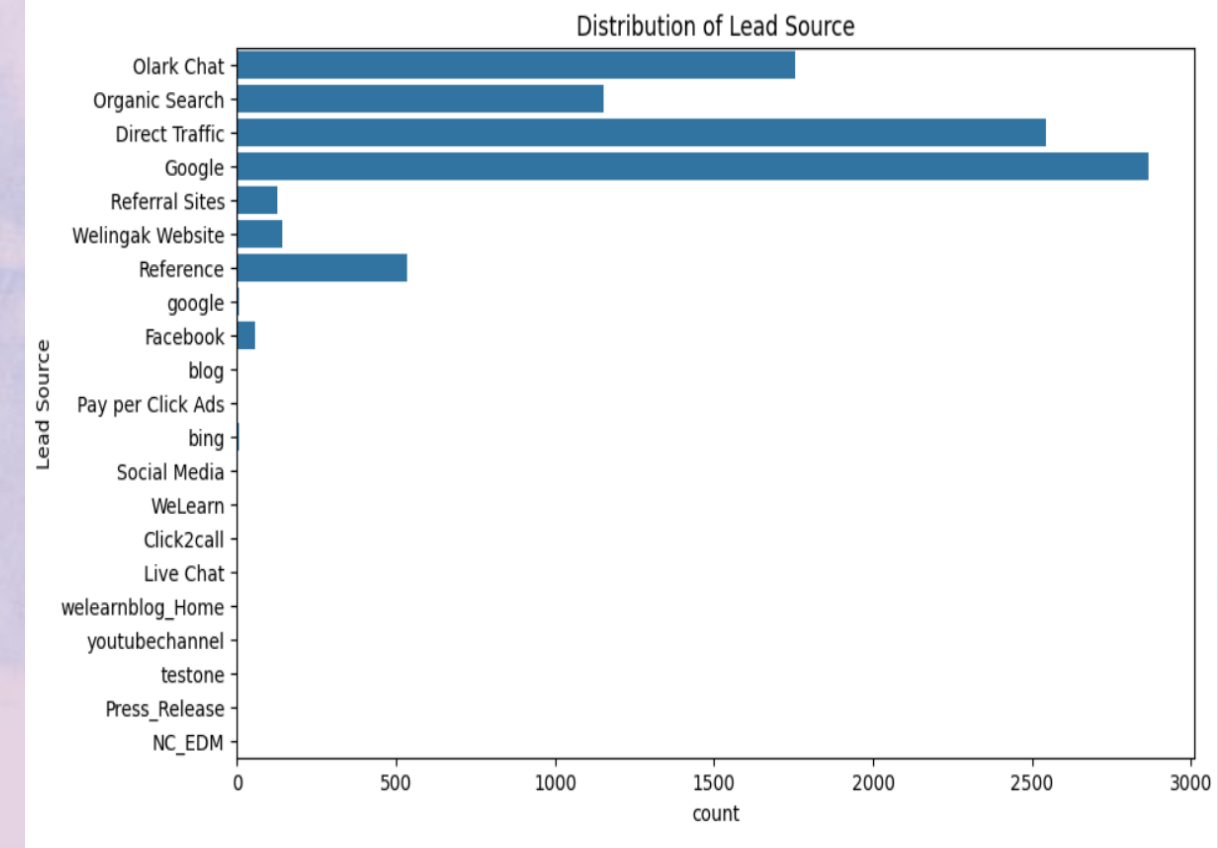






The histograms provide insights into the frequency and depth of engagement with the website, measured by the number of visits and time spent. These engagement metrics are likely to correlate with lead conversion likelihood.





The bar charts compare different lead sources and origins, indicating which channels are most effective at generating leads. Such insights can direct marketing efforts to the most productive sources.

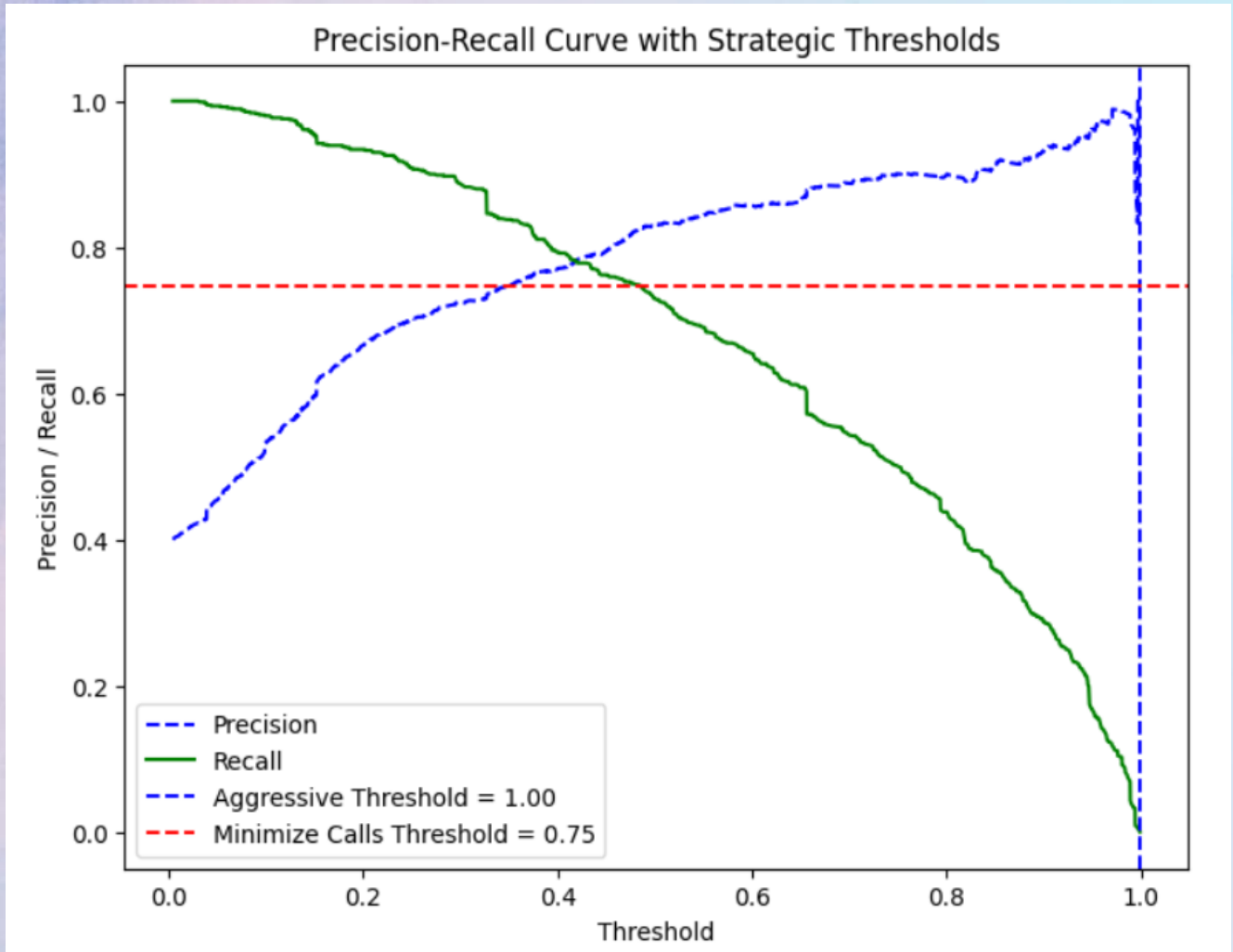
Model Building

- 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 ColumnTransformer as preprocessor to define numerical and categorical columns
- Carrying out Grid Search for Hyperparameter tuning.
- Building Model understanding and evaluating the model performance.
- Predictions on test data set
- Overall accuracy 80%

Precision-Recall Curve

Presents the precision-recall trade-off, which is critical in evaluating the model's performance in a business context.

where both false positives and false negatives carry different costs.



Conclusion

It was found that the variables that mattered the most in the potential buyers are :

(Here provided in descending order)

- The total time spent on the Website.
- Total number of visits.
- When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website
- When the last activity was:
  a. SMS
  b. Olark chat conversation
- When the lead origin is Lead add format.

These points highlight the systematic approach taken in the case study to address a common issue in sales and marketing through data analytics and predictive modeling.

THANK YOU!