Lead Score Case study

BY:

AMIT KULKARNI

ASHUTOSH KUMAR

SWARANGI ARVIKAR

Problem Statement

❑ X Education provides online courses tailored for industry professionals.

❑ Although X Education generates a significant number of leads, its conversion rate is low; for instance, out of 100 daily leads, only about 30 convert.

❑ To enhance efficiency, the company aims to pinpoint the leads with the highest potential, referred to as ‘Hot Leads’.

❑ By accurately identifying these 'Hot Leads,' the conversion rate is expected to improve, allowing the sales team to focus their efforts on engaging with these promising leads instead of contacting every prospect.

Business Objective

❑ X Education seeks to identify the most promising leads.

❑ To achieve this, they plan to develop a model that can accurately pinpoint hot leads.

❑ Once developed, this model will be deployed for ongoing use.

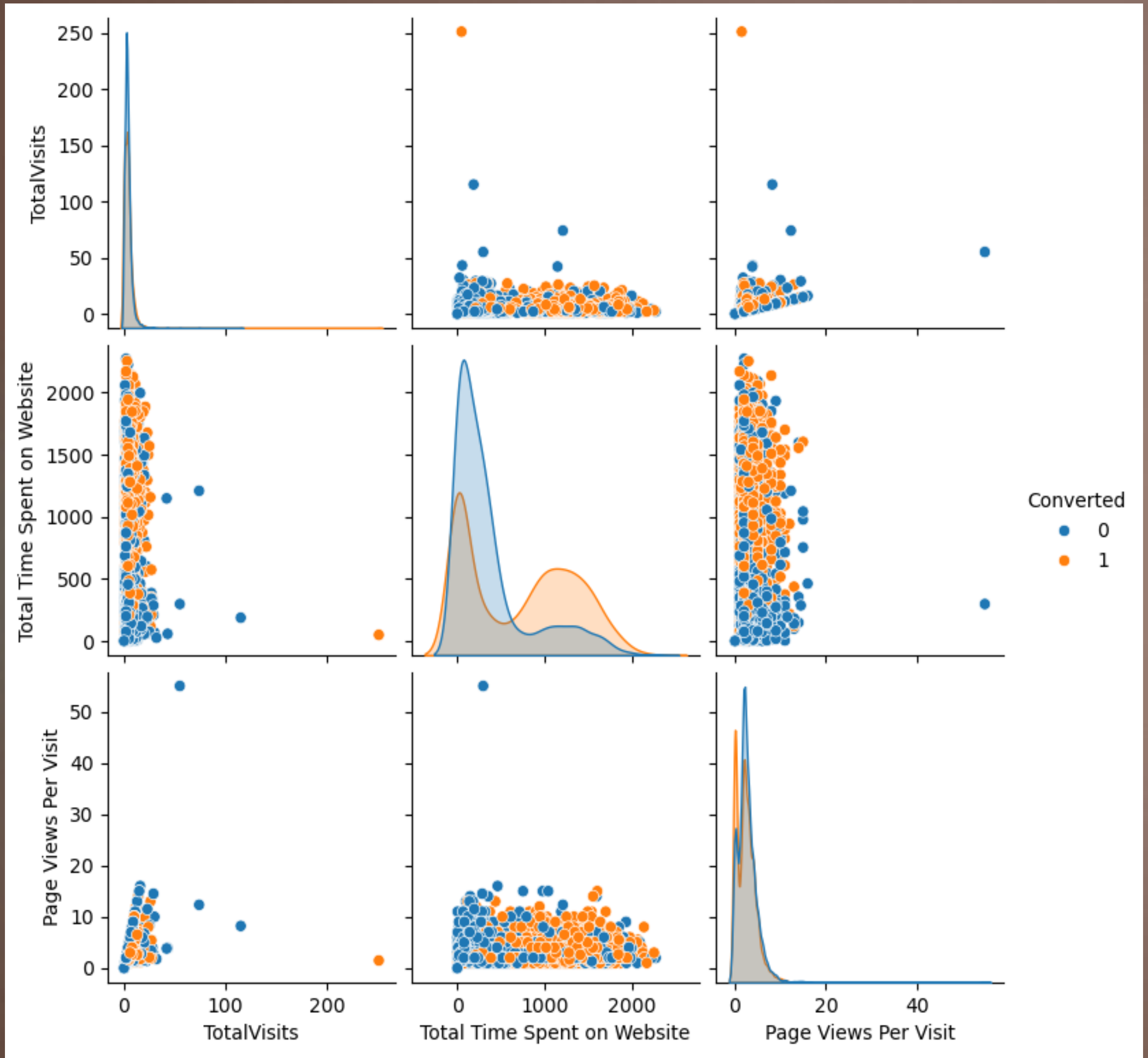
Solution Methodology

- ❑ **Data Cleaning and Manipulation:** Ensuring data quality and appropriate formatting for analysis.
- ❑ **Exploratory Data Analysis (EDA):**
  - **Univariate Data Analysis:** Examining the distribution of variables.
  - **Bivariate Data Analysis:** Analyzing correlation coefficients between variables.
- ❑ **Feature Scaling & Encoding:** Standardizing the data and converting categorical variables into numerical format.
- ❑ **Classification Technique:** Implementing logistic regression for model development and prediction.
- ❑ **Model Validation:** Assessing the accuracy and reliability of the model.
- ❑ **Model Presentation:** Demonstrating the model’s capabilities and results.
- ❑ **Conclusions:** Providing insights and actionable strategies based on the model's findings.

Data Manipulation

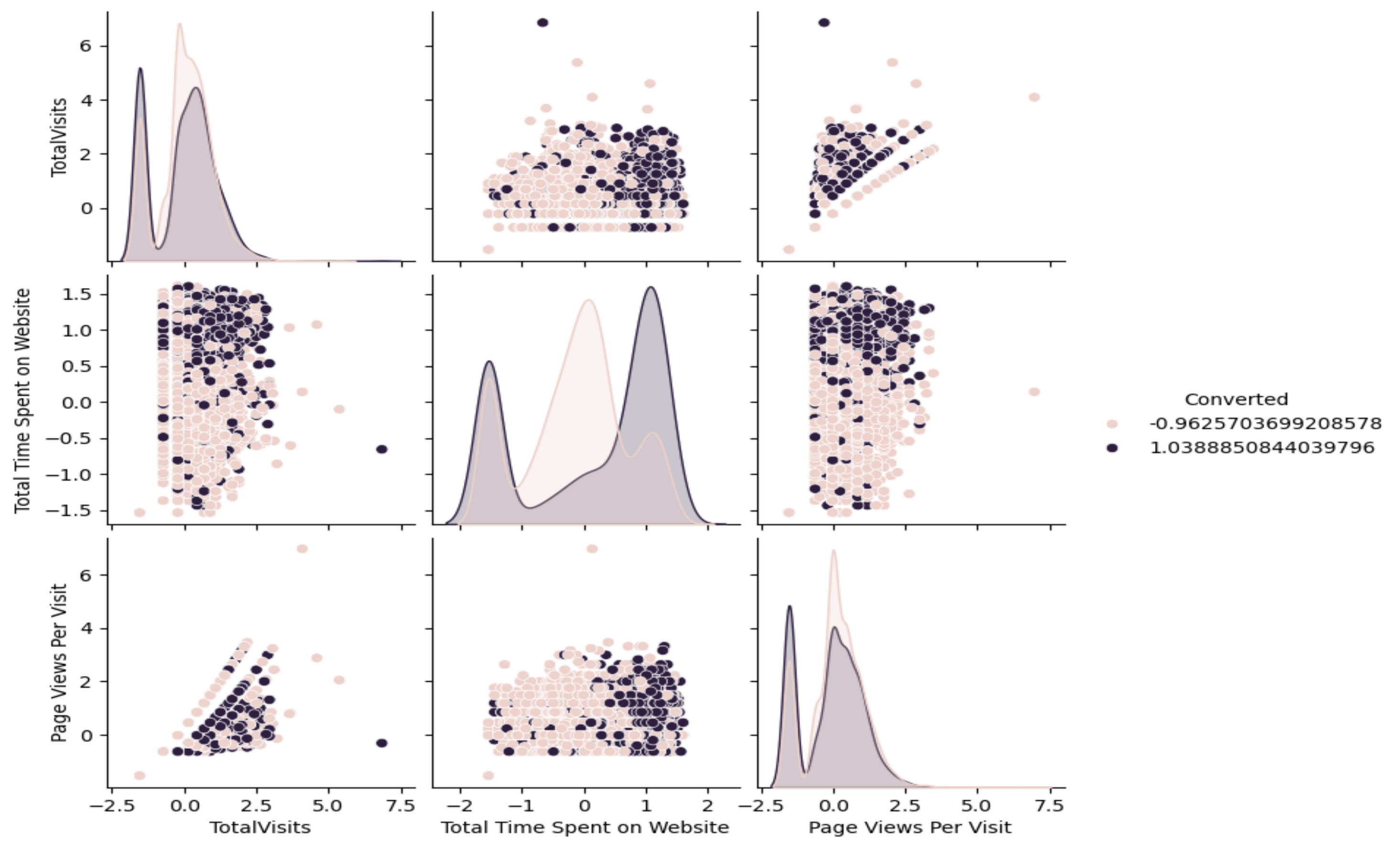
- The dataset, comprises 9240 entries and 37 attributes related to lead scoring information.
- Columns with over 3000 missing entries were removed, reducing attributes from 37 to 31, to retain only useful information.
- Columns with negligible variability, such as Country and City, were dropped.
- The variables Prospect ID and Lead Number were dropped from the analysis as they were not useful.
- Columns with a single predominant value that offered little analytical value, such as Do Not Call and Magazine, were dropped.
- Certain attributes like What matters most to you in choosing a course showed highly skewed distributions which were not useful and hence were excluded from further analysis.
- The resultant clean dataset contains 6373 entries, representing approximately 69% of the original data.

Exploratory Data Analysis



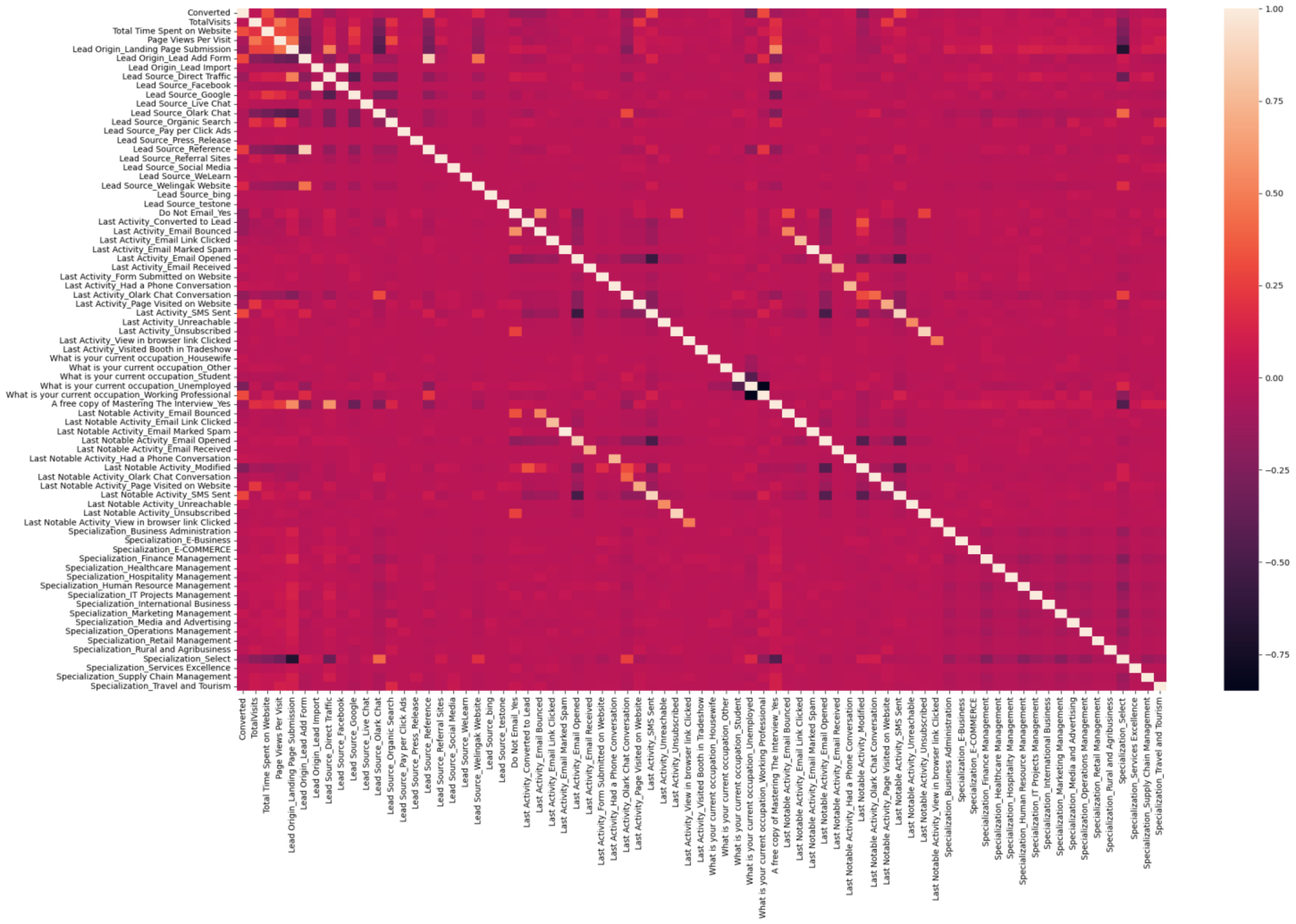
- ❑ **Time on Site:** Leads who spend more time on the website are more likely to convert.
- ❑ **Engagement Levels:** Higher total visits and page views don't strongly correlate with conversion, suggesting quality of interaction over quantity is key.
- ❑ **User Behavior:** Density plots indicate that successful conversions typically involve longer website sessions, emphasizing the need to improve user engagement.

Normalized Relationships and Distribution



The visualization shows pairplot graphs from a dataset where features like TotalVisits, Total Time Spent on Website, and Page Views Per Visit were normalized using a Power Transformer.

Analysing Correlation



- ❑ The heatmap's color scale ranges from -1 to 1, where light side of red indicates strong positive correlations and black shows no correlation between variables.
- ❑ The diagonal white line represents a perfect correlation (value of 1) for each variable with itself, while off-diagonal cells reveal the correlations between different variables.
- ❑ Blocks of lighter red in the heatmap highlight groups of variables that share stronger relationships, potentially impacting each other significantly.
- ❑ Areas with darker shades suggest weaker correlations, indicating less direct or insignificant relationships between certain variables.

Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 6373
- Total Columns for Analysis: 75

MODEL BUILDING

❑ Splitting the Data into Training and Testing Sets

❑ Data is split in a 70:30 ratio for training and testing.

❑ Numeric features are standardized for better model performance.

❑ Using RFE, the top 15 features are selected.

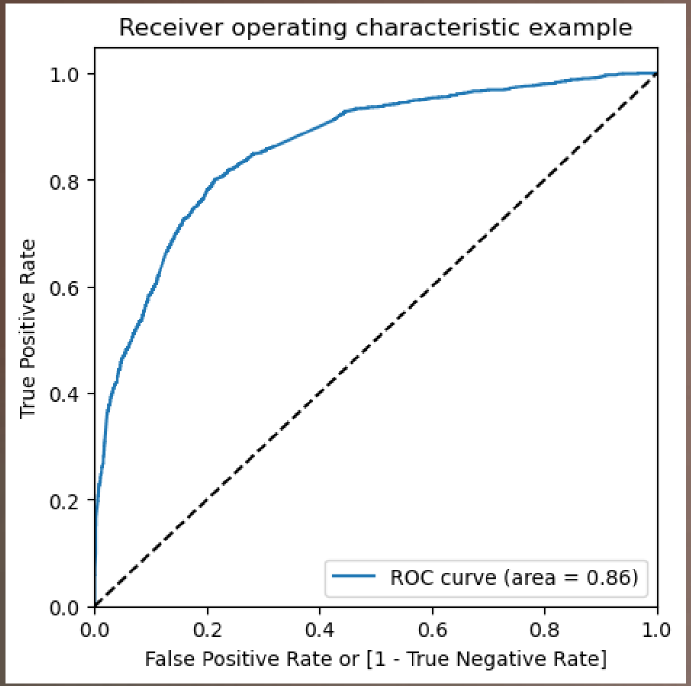
❑ Variables with p-values > 0.05 and VIF > 5 are removed to optimize the model.

❑ Predictions are evaluated on the test dataset, yielding a confusion matrix:

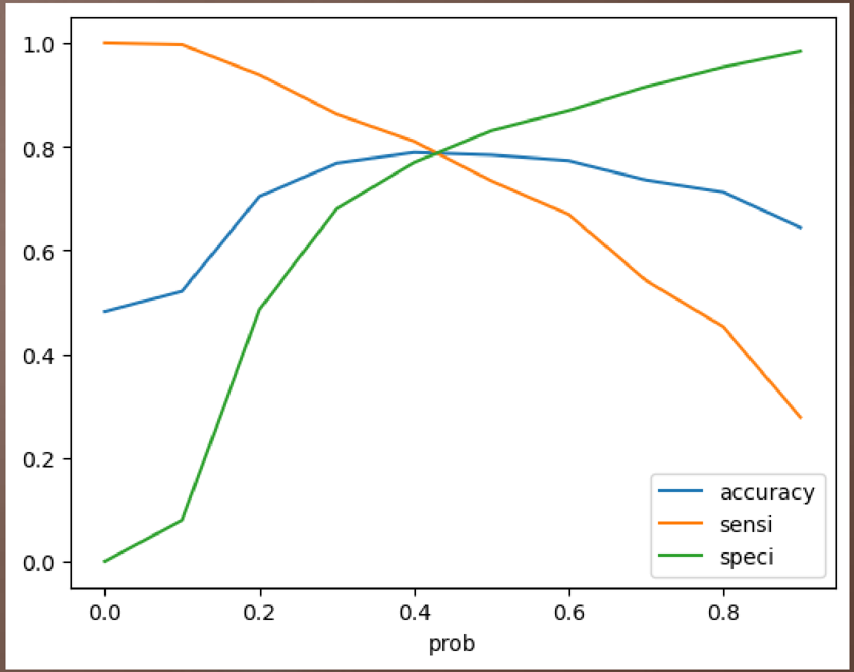
- True Positives: 1577
- True Negatives: 1921
- False Positives: 391
- False Negatives: 572

❑ The model achieves an overall accuracy of 78.4%.

Model Performance Optimization



The ROC curve, with an AUC of 0.86, indicates that the model effectively distinguishes between positive and negative classes.



The second graph shows that model accuracy peaks at a probability threshold of about 0.4, suggesting this is the most balanced point for correctly predicting both classes.

Predicting Test Data Set

After evaluating classification metrics and probability thresholds prediction made on the test data set:

- ❑ We assessed the model's performance by calculating accuracy, precision, and recall.
- ❑ The results were promising, with an accuracy of 0.78, and both precision and recall at approximately 0.77.
- ❑ These figures indicate that the model performs reliably across different measures.
- ❑ Overall, the model demonstrates stability with robust accuracy and recall rates.

Conclusion

This study has revealed key factors significantly influencing potential buyers, ranked by their impact:

❑ **Total Time Spent on the Website:** Most crucial in converting leads.Total Number of Visits: Indicates repeated interest and engagement.

❑ **Lead Source:** Particularly effective when originating from Olark Chat, and the Welingak website.

❑ **Last Activity:** Most influential activities include SMS and Olark chat conversations.

❑ **Lead Origin:** Strongly correlates when the lead is sourced from the 'Lead Add Form’.

❑ **Current Occupation:** Leads identified as unemployed and student show a higher likelihood of conversion.

With these findings, X Education can strategically focus its efforts to effectively engage and convert potential buyers into course participants, thereby enhancing their market success.