December 2025

Lead Scoring Case Study

Logistic Regression–Based Lead Conversion Prediction

Presentation

Prepared By:
Banoth Pavan

Business Problem Statement

- X Education is an organization which provides online courses for industry professional. The company marks its courses on several popular websites like google.

- X Education wants to select most promising leads that can be converted to paying customers.

- Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, google searches etc.

- The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions.

Objective of the Project

- Build a predictive model to score leads
- Identify key factors influencing conversion
- Optimize calling strategy based on business needs
- Balance conversion rate and sales effort

Strategy

- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Assign a lead score for each leads
- Test the model on train set
- Evaluate model by different measures and metrics
- Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation

Data Cleaning & Preprocessing

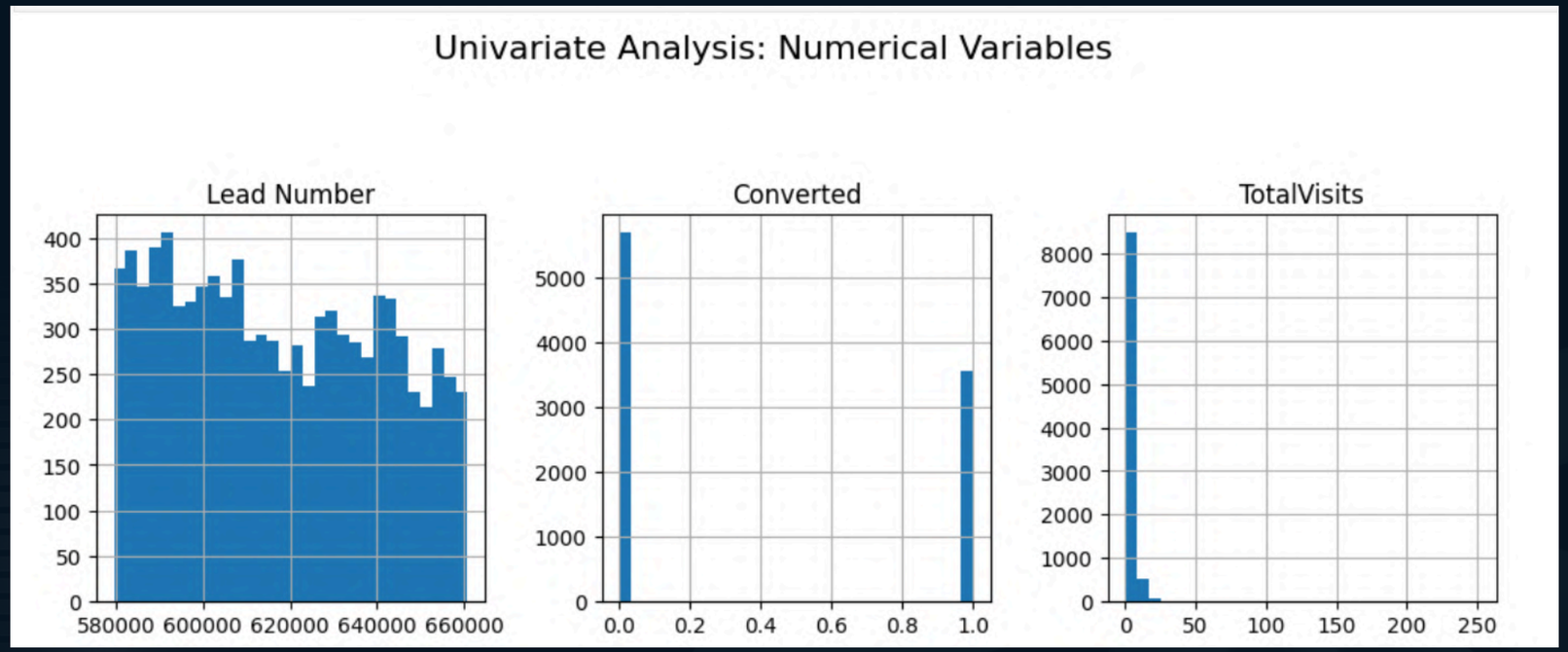
- Handled missing values appropriately
- Dropped irrelevant/high-missing columns
- Converted categorical variables into dummy variables
- Removed redundant dummy variables (reference category)
- Ensured data consistency

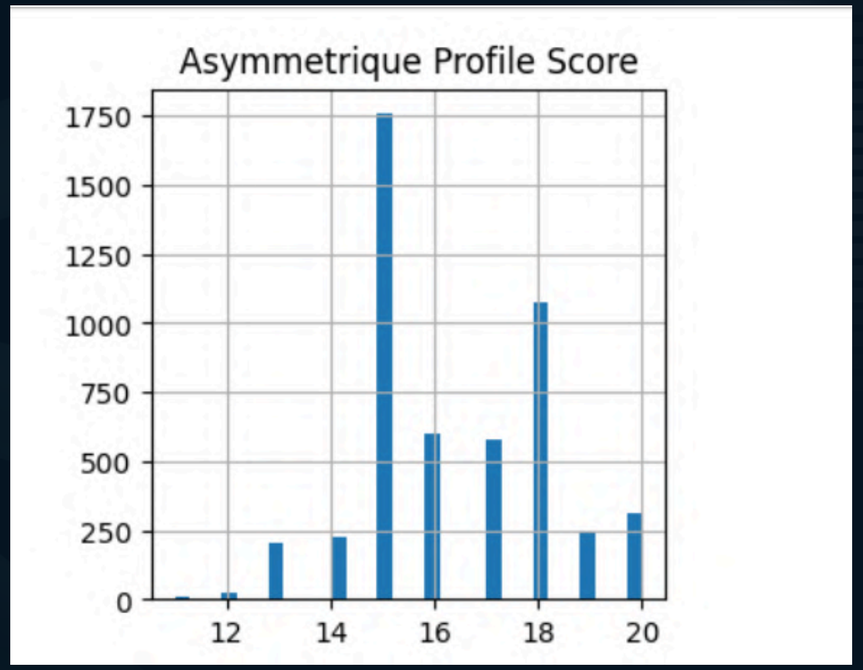
	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9240.000000	9240.000000	9240.000000
mean	3.217424	487.698268	2.255105
std	2.860471	548.021466	1.779471
min	0.000000	0.000000	0.000000
25%	1.000000	12.000000	1.000000
50%	3.000000	248.000000	2.000000
75%	5.000000	936.000000	3.000000
max	11.000000	2272.000000	6.000000

Exploratory Data Analysis (EDA)

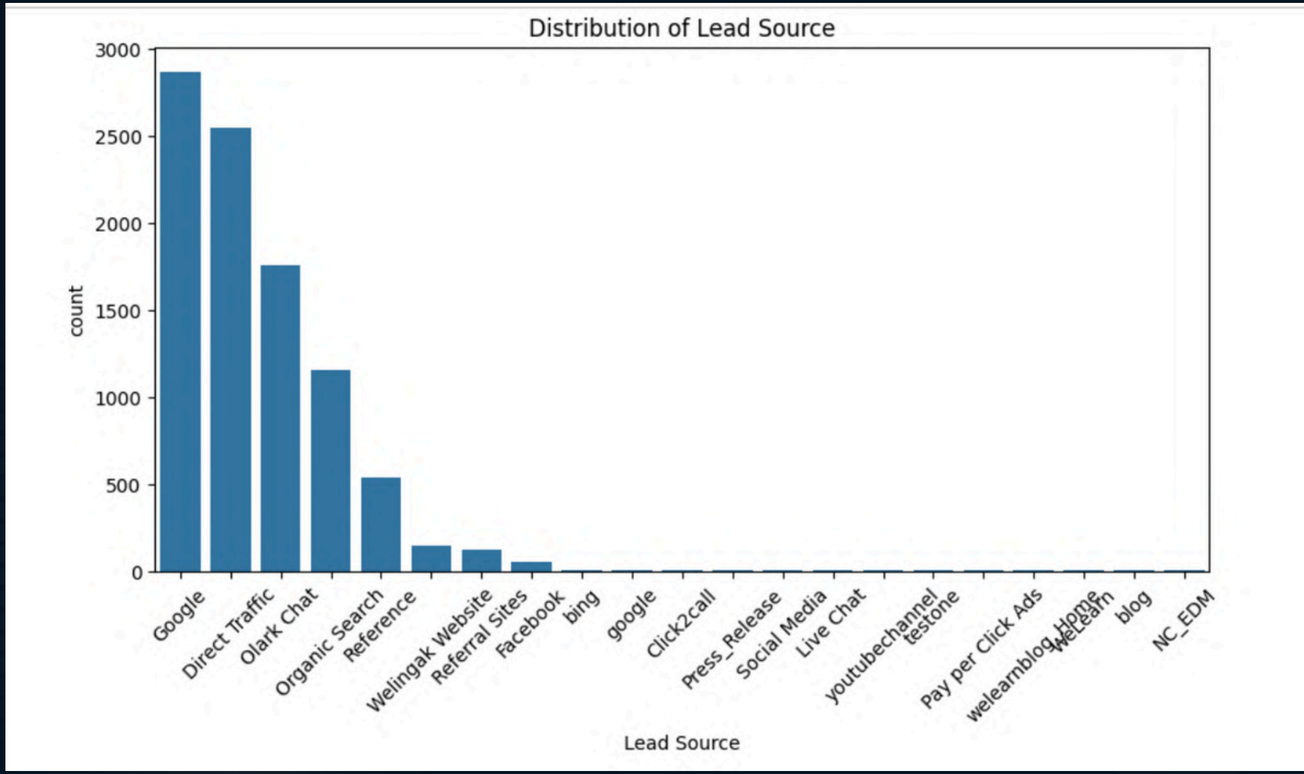
- Conversion rate varies significantly across lead sources
- Certain user activities strongly indicate conversion intent
- Converted categorical variables into dummy variables
- High imbalance between converted and non-converted leads
- EDA guided feature selection

Exploratory Data Analysis (EDA)



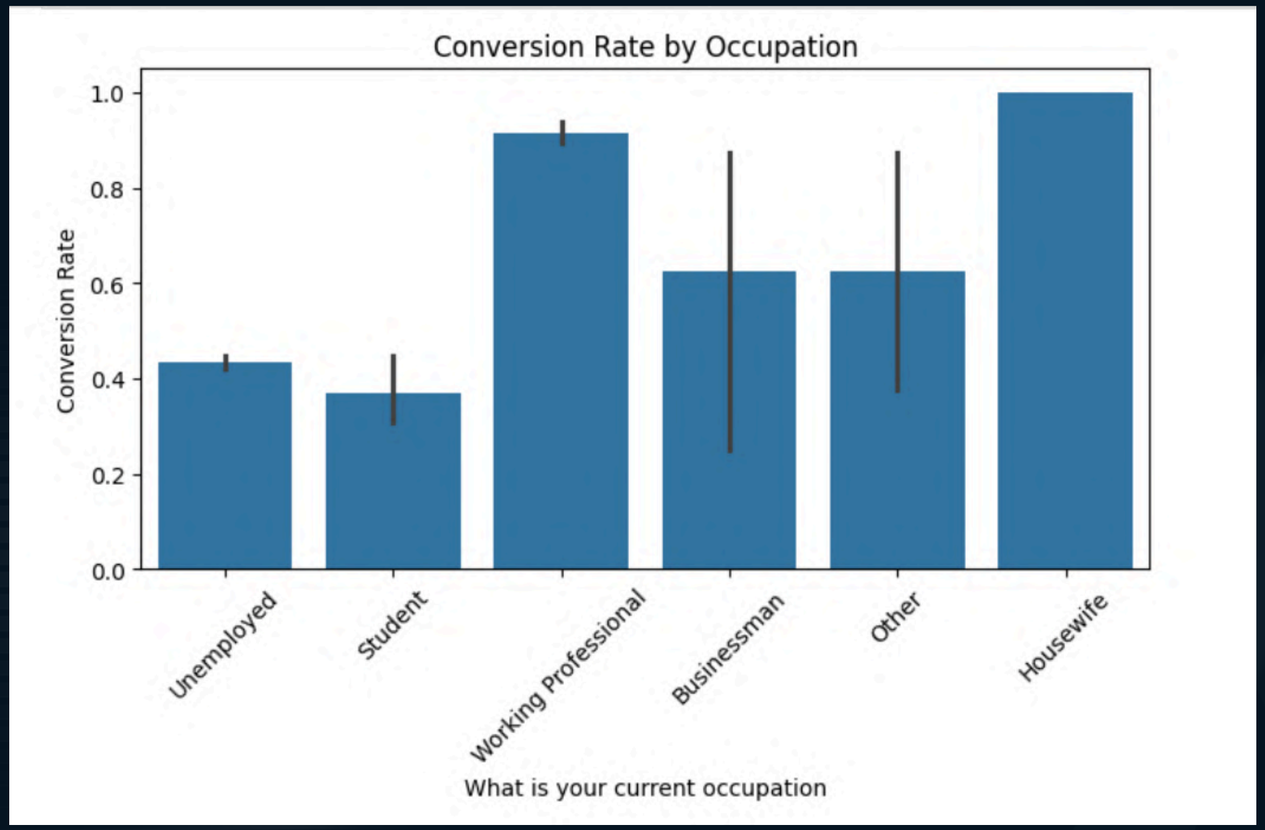


Exploratory Data Analysis (EDA)



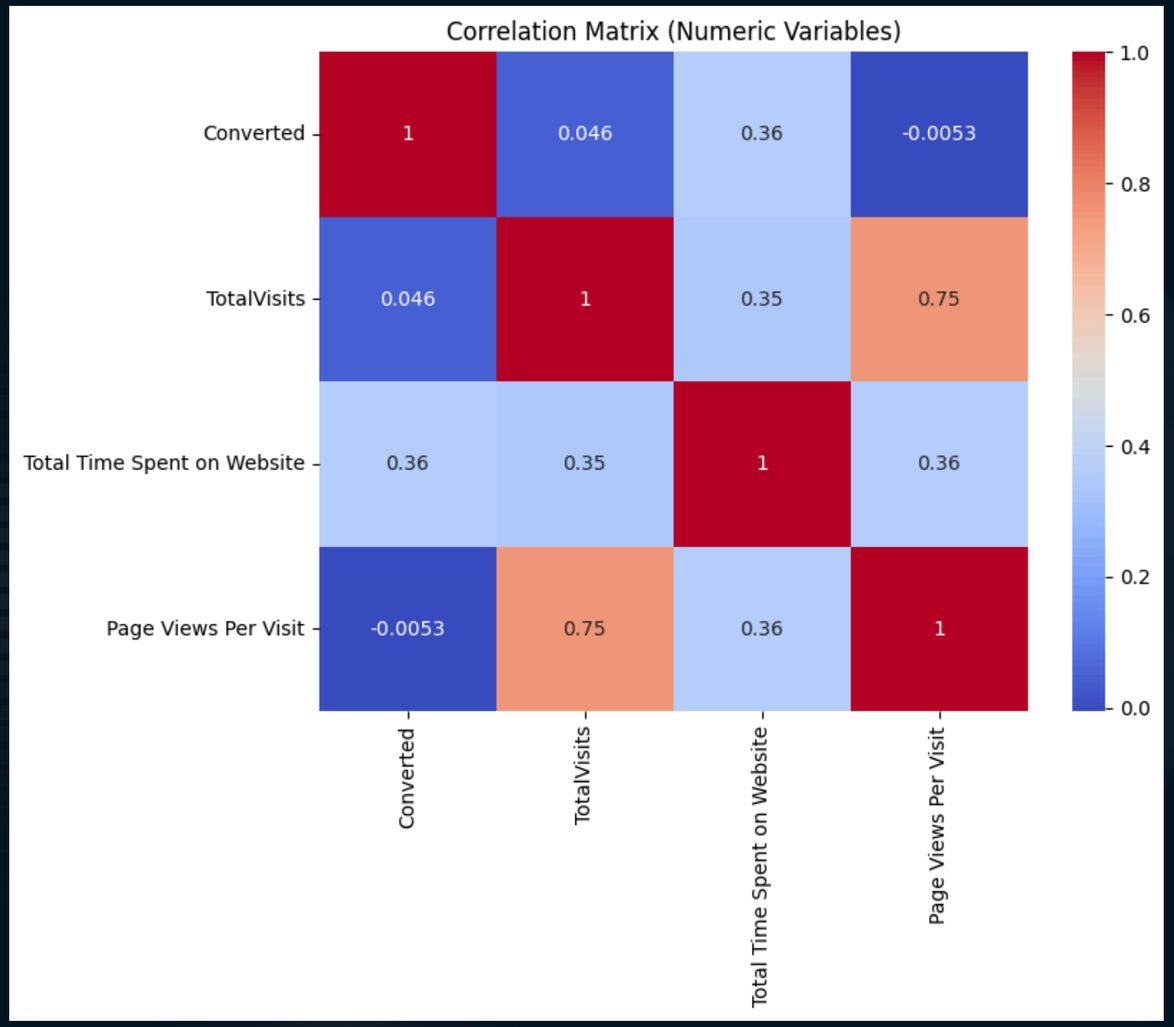
- Google and Direct Traffic are the dominant lead sources, contributing the highest number of leads, which shows strong dependence on search and direct brand awareness.

Exploratory Data Analysis (EDA)



- Housewives and working professionals show the highest conversion rates, indicating they are the most likely segments to convert and should be prioritized for targeted marketing efforts.

Exploratory Data Analysis (EDA)



- Conversion shows a moderate positive correlation with Total Time Spent on Website, indicating that users who spend more time on the site are more likely to convert.

Feature Selection & Multicollinearity

- Used Variance Inflation Factor (VIF) to detect multicollinearity
- Iteratively removed high-VIF features
- Considered p-values for statistical significance
- Retained only stable, interpretable predictors

Model Building

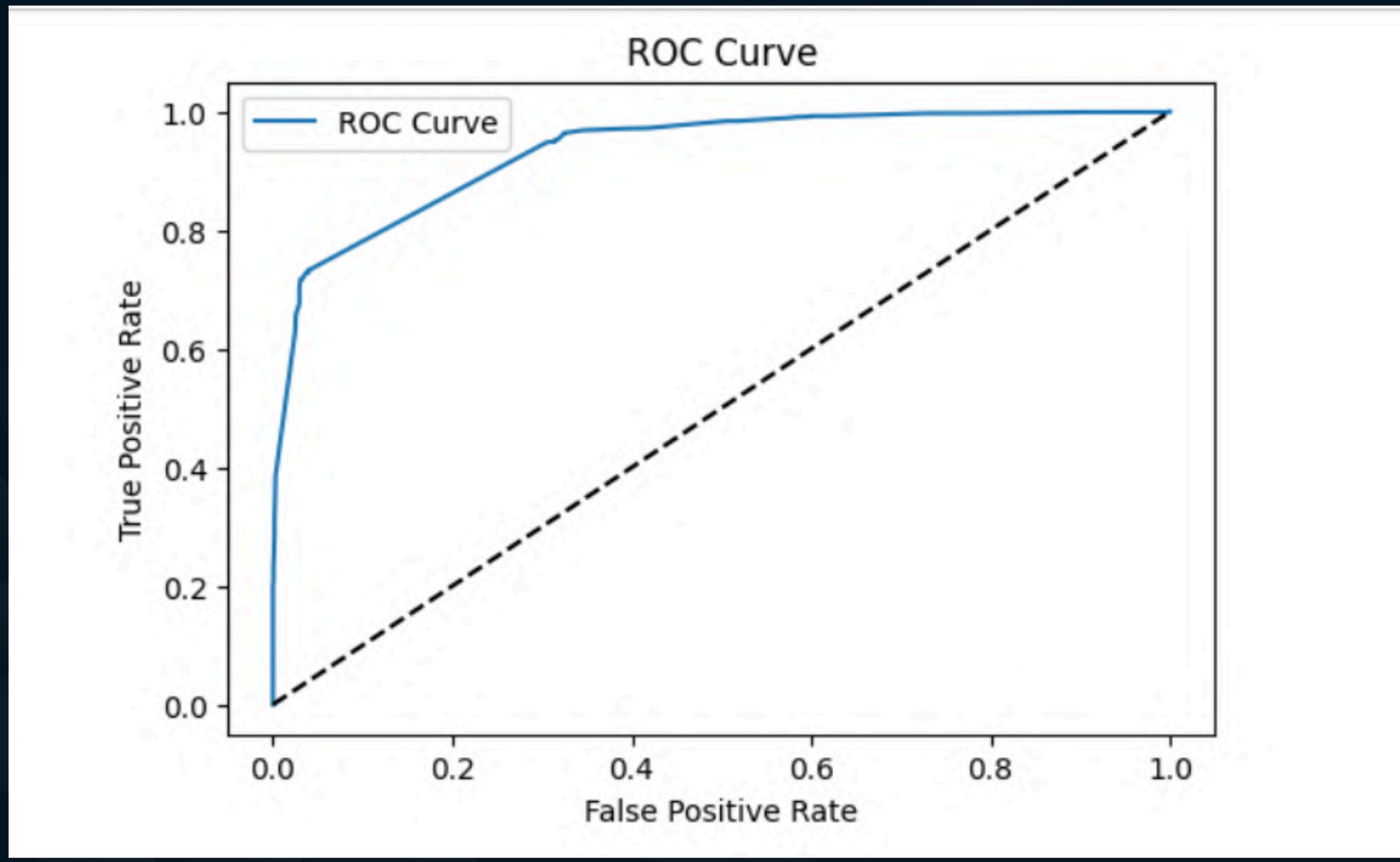
- Algorithm used: Logistic Regression
- Reason for selection:
    1) High interpretability
    2) Suitable for probability prediction
    3) Business-friendly explanation
- Model refined through multiple iterations

Model Evaluation

- Metrics used:
  1) Accuracy
  2) Precision
  3) Recall
  4) ROC-AUC

- Explain simply:
  → Precision → avoids useless calls
  → Recall → captures maximum potential leads

Model Evaluation



- The ROC curve stays well above the diagonal, showing that the model effectively distinguishes between converted and non-converted leads.
- A high true positive rate is achieved at low false positive rates, which is desirable for minimizing unnecessary follow-ups.
- Overall curve shape suggests strong model performance, indicating good classification capability.

Strategy (Business Scenarios)

- Scenario 1: Aggressive Calling Phase
  → Objective: Maximize conversions
  → Strategy: Lower cutoff
  → Result: Higher recall → more leads contacted

- Scenario 2: Conservative Phase
  → Objective: Reduce unnecessary calls
  → Strategy: Higher cutoff
  → Result: Higher precision → fewer, high-quality calls

Final Conclusion & Impact

→ Successfully built an interpretable lead scoring model

→ Enabled smarter sales prioritization

→ Reduced wasted effort while improving conversions

→ Model supports flexible business decision-making

Lead Scoring

December 2025

Thank You

Prepared By:
Banoth Pavan