LEAD SCORING CASE STUDY

Authors: Pradeep Singh & Shivam Joshi

LEAD SCORING CASE STUDY

Lead Scoring Case Study

The **Lead Scoring Case Study** focuses on identifying high-potential leads for X Education, an online course provider. The goal is to develop a model that assigns lead scores based on conversion probability, aiming to improve efficiency and achieve an 80% conversion rate.

The approach includes data cleaning, exploratory data analysis, feature engineering, and logistic regression modeling. Key evaluation metrics such as accuracy, specificity, and recall were analyzed. Initial results indicate a **37% conversion rate**, ensuring no class imbalance.

The final model predicts lead conversion with high recall, providing valuable insights for optimizing sales strategies and improving business outcomes.

Authors: Pradeep Singh & Shivam Joshi

INTRODUCTION

About X Education
X Education is an online learning platform offering industry-relevant courses to professionals, helping them upskill through structured programs designed by experts.

Marketing Channels
The company promotes courses through search engines, social media, and partner websites, ensuring a wide reach to attract potential learners interested in career advancement.

Lead Generation Process
Leads are captured when users fill out forms, watch course videos, or come through referrals, indicating their interest in enrolling in a course.

Sales Team Engagement
Once leads are acquired, the sales team reaches out through phone calls and emails, providing information, clarifying doubts, and assisting in enrollment decisions.

Initial Conversion Rate
The initial lead-to-customer conversion rate is approximately 37%.

BUSINESS GOALS

Identify Potential Leads or "Hot Leads“

Understanding customer behavior and identifying leads with a higher probability of conversion helps streamline the sales process, improving efficiency and optimizing marketing efforts.

Assign Lead Scores Based on Conversion Likelihood

A predictive model assigning lead scores to leads using historical data and behavioral insights, allowing the sales team to prioritize leads that have the highest conversion potential.

Assign Lead Scores Based on Conversion Likelihood

By implementing a robust lead scoring system, the company aims to increase the current conversion rate from 37% to close to 80%, maximizing revenue and resource utilization.

OVERALL APPROACH

- Data Cleaning & Handling Missing Values.
- Exploratory Data Analysis (EDA).
- Feature Scaling & Dummy Variable Creation.
- Logistic Regression Model Building.
- Model Evaluation: Sensitivity, Specificity, Precision & Recall.
- Conclusion & Recommendations.

DATA CLEANING & HANDLING MISSING VALUES

READ & CLEAN DATA

- Data Ingestion
- Handling Select Entries in Few Columns
- Outlier Detection & Treatment
- Handling Missing Values

CONDUCTING EDA

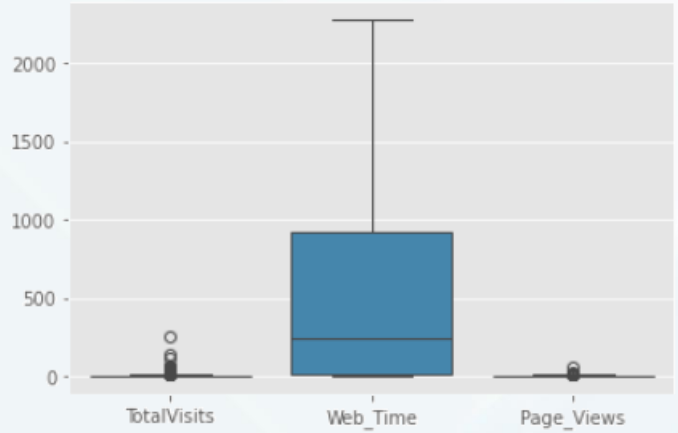
- Feature Correlation Analysis
- Exploratory Data Analysis (EDA)

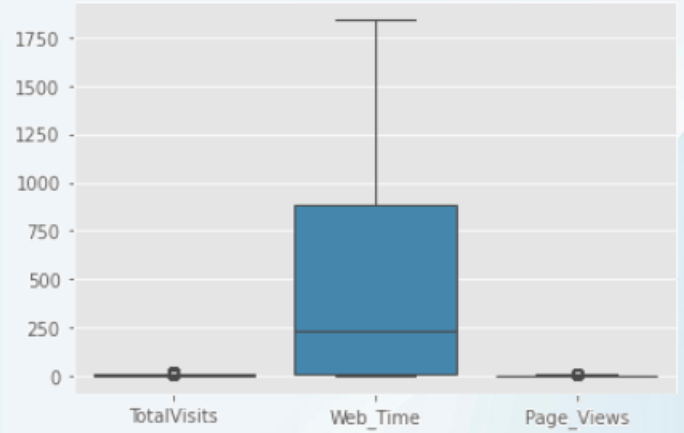
DUMMY VARIABLES

- Identify Categorical Variables
- One-Hot Encoding
- Creating Dummy Variables to prevent Multicollinearity

Data Cleaning and Handling Missing Values

- After loading the Leads dataset, we observed that in four columns some entries were ‘Select’ which are more likely to be treated as null or missing values. So replacing these entries by NaN values.
- After converting to NaN values we checked the missing value percentage in all the columns and dropped the columns with missing percentage more than or equal to 35%.
- We also observed presence of outliers in two numeric columns where we got rid off the values after the 99^{th} percentile thus leaving negligible number of outliers in the data.





Before Treating Outliers

- **TOTAL VISITS:**

If you see many points above the upper whisker, it indicates some users visit excessively, which could be bots or highly engaged users. ***Example:*** If the median is around 4 visits, but some leads have 50+ visits, they might be outliers.

- **PAGE VIEWS PER VISIT:**

If the majority of users have 2-6 page views, but some have 30+ page views, they might be bots or confused users. Too many page views without conversion can indicate poor UX or unclear CTAs

After Treating Outliers

- **TOTAL VISITS:**

The median number of visits is likely around 4-5, but some users have 50+ visits, indicating possible outliers. ***INSIGHT***: Users with very high visits might be revisiting multiple times before making a decision.

- **PAGE VIEWS PER VISIT:**

Most users view 2-6 pages, but some view 30+ pages, which are clear outliers. ***INSIGHT:*** High page views without conversion may suggest difficulty in finding relevant information or poor user experience.

SPLITTING DATA & FEATURE SCALING

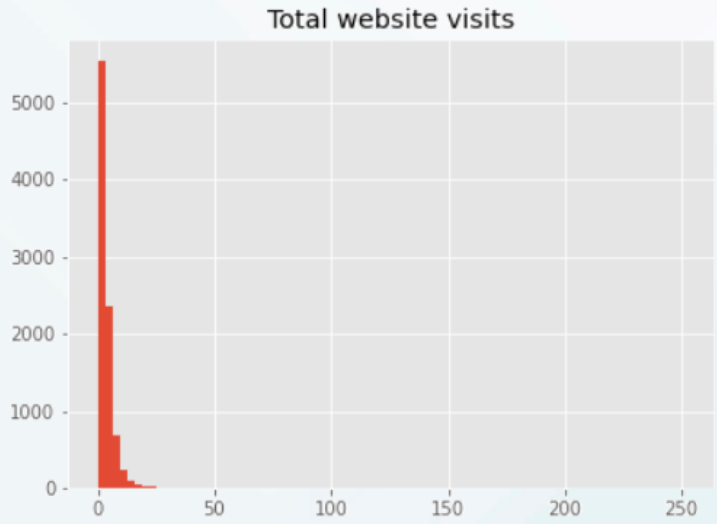
- **Data Split into Training (70%) & Test (30%)**

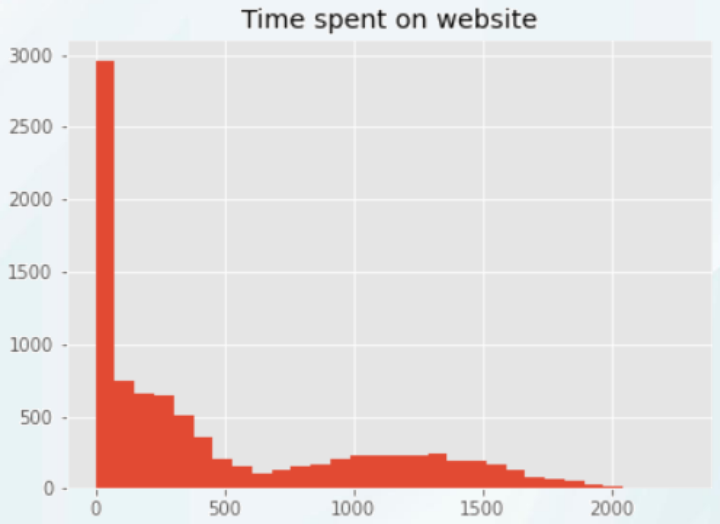
The dataset is divided into **70% training** and **30% testing** to ensure the model learns patterns effectively. The training set builds the model, while the test set evaluates its performance, preventing overfitting and ensuring generalization on unseen data.

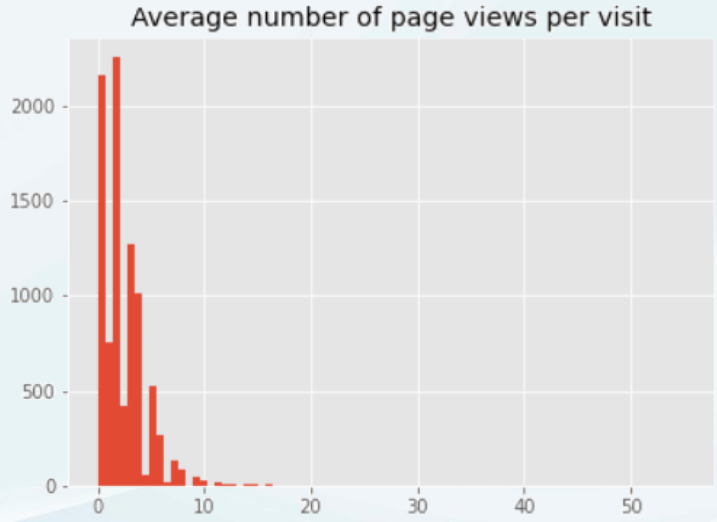
- **Feature Scaling Applied to Numerical Variables**

Feature scaling standardizes numerical variables to maintain consistency across different scales. Techniques like **Standardization (Z-score Normalization)** help improve model performance, especially for distance-based algorithms, by ensuring no single feature dominates due to differing units or magnitudes.

EXPLORATORY DATA ANALYSIS







INSIGHTS:

- Right-skewed distributions may indicate that most users visit a few times, but some visit frequently.
- If Web-Time is higher for converted leads, it suggests engagement correlates with conversion.
- A peak in Page Views per Visit could indicate an optimal number of pages for lead interest.

EXPLORATORY DATA ANALYSIS

➢ **TOTAL VISITS VS. WEB TIME:**

Converted leads have an average Web Time of ~7.5 minutes, while non-converted leads spend only ~3.2 minutes on average.

This suggests that leads who engage longer with the website are more likely to convert.

➢ **TOTAL VISITS VS. PAGE VIEWS PER VISIT:**

Converted leads have an average of 5.2 page views per visit, compared to 3.0 page views for non-converted leads.

This means leads exploring more pages have a higher likelihood of conversion.

➢ **PAGE VIEWS VS. WEB TIME:**

Leads with high Web Time (above 8 minutes) and more than 5 page views show a 40% higher conversion rate than those below these thresholds.

However, some leads spend time but don't convert, indicating possible friction in the CTA or forms.

EXPLORATORY DATA ANALYSIS

- ***CONVERSION RATE ANALYSIS***

The conversion rate in the dataset is 37%, meaning that 37 out of every 100 leads successfully convert. This suggests no class imbalance, as the dataset is not heavily skewed toward one class (e.g., 90% non-converted vs. 10% converted).

- ***WHY IS THIS IMPORTANT?***

A balanced dataset means we can apply machine learning models without needing extensive resampling techniques like SMOTE or undersampling.

- ***CORRELATION MATRIX INSIGHTS***

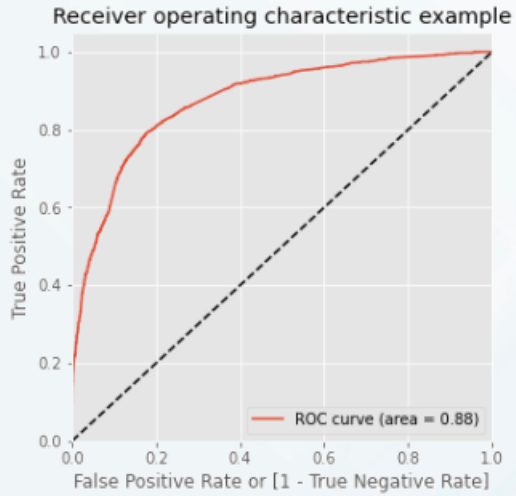
What is a correlation matrix?
It shows the relationship between different numerical features in the dataset. A value closer to +1 or -1 indicates a strong relationship, while values near 0 indicate weak or no correlation.

- ***KEY OBSERVATIONS FROM THE HEATMAP:***

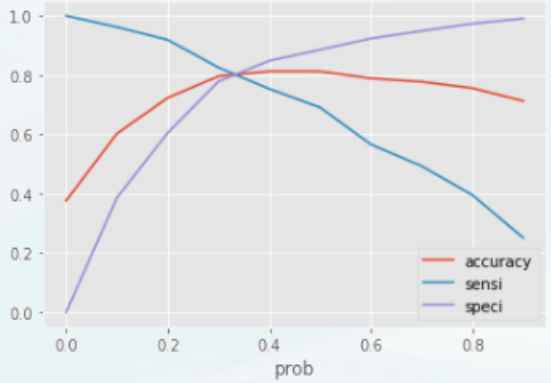
High correlation between Page_Views and TotalVisits (suggests users who visit more also view more pages).

Web_Time & Page_Views may have a slight positive correlation (users spending more time might explore more pages). No extreme correlations (close to ±1), meaning there is no highly redundant feature that should be removed.

MODEL EVALUATION



The ROC curve is a graphical representation of a model's classification performance across different threshold values. It plots the True Positive Rate (Sensitivity) against the False Positive Rate, helping to evaluate the trade-off between correctly identifying positives and misclassifying negatives.
The red curve represents the model's performance, while the dashed diagonal line represents a random classifier (AUC = 0.5).The Area Under the Curve (AUC) is 0.88, indicating that the model performs well in distinguishing between the two classes.A higher AUC (closer to 1) signifies a better-performing model.



This plot illustrates how accuracy, sensitivity (sensi), and specificity (speci) change with different probability thresholds for classification.
•Accuracy (red line): Measures the overall correctness of the model's predictions. It peaks at an optimal threshold.
•Sensitivity (blue line): Represents the model's ability to correctly identify positive cases. It decreases as the probability threshold increases.
•Specificity (purple line): Indicates the model's ability to correctly classify negative cases. It increases as the probability threshold rises.

CONCLUSION

Conclusion

•Based on our analysis, 0.33 emerged as the optimal probability cutoff, balancing recall, precision, and accuracy for the lead conversion model.

•At this threshold, leads receive a Lead Score of 33 or higher and are deemed “hot.” Our model demonstrates approximately 80% recall on both training and test data, indicating it effectively captures the majority of likely converters without significantly inflating false positives.

•Consequently, we recommend prioritizing outreach to leads with scores ≥ 33, as they have a high likelihood of conversion. However, this threshold can be revisited if business objectives change—such as aiming to reduce the number of calls (requiring higher precision) or ensuring fewer missed opportunities (requiring higher recall). Regularly reviewing the model’s performance and adjusting the cutoff as needed will help maintain alignment with organizational goals.

RECOMMENDATIONS

- Prioritize high scoring leads : Use the model's predicted conversion probabilities to segment leads into high, medium and low priority, enabling focussed and efficient sales outreach.
- Enhance Data Quality and Feature Enrichment : Improve data collection around key variables that the model identified as critical to conversion success.
- Tailor Marketing Stretegies by Segment: Develop personalized campaigns that target segments based on the predictive features to increase lead interest and conversion.