Lead Scoring Analysis for X Education

# Introduction

X Education, an online course provider, aims to improve its lead conversion rate from 30% to 80% by identifying high-potential leads, also known as ‘Hot Leads’. The goal of this analysis is to develop a predictive model that assigns a lead score, enabling the sales team to focus on the most promising prospects.

# Data Preprocessing

The dataset contains approximately 9000 records with attributes such as Lead Source, Total Time Spent on Website, Total Visits, and Last Activity. The target variable, 'Converted', indicates whether a lead was successfully converted (1) or not (0). Key preprocessing steps include:

## Handling Missing Values:

* + 1. Columns with more than 3000 missing values were dropped.
    2. Categorical variables with a ‘Select’ option were removed as they provided no meaningful information.

## Categorical Encoding:

* + 1. One-hot encoding was applied to categorical features such as Lead Source and Last Activity.

## Feature Scaling:

* + 1. MinMax scaling was applied to numeric features (Total Visits, Page Views per Visit, Total Time Spent on Website) for better model performance.

# Exploratory Data Analysis (EDA)

* 1. A correlation heatmap revealed that Total Time Spent on Website had the highest correlation with lead conversion.
  2. Certain Lead Sources, such as Google and Direct Traffic, had higher conversion rates compared to others.
  3. Leads who engaged more frequently (e.g., email interactions, video watching) were more likely to convert.
  4. Unnecessary variables like City and Country were dropped as they showed minimal impact on lead conversion.

# Model Building

* 1. A Logistic Regression model was chosen due to its interpretability and effectiveness in binary classification.
  2. Features were selected iteratively based on p-values and Variance Inflation Factor (VIF) to reduce multicollinearity.
  3. The final model was trained using 70% of the dataset, with 30% used for testing.

# Model Evaluation

The model’s performance was assessed using:

* 1. **Confusion Matrix:** To evaluate false positives and false negatives.
  2. **Accuracy:** 79% on the test set.
  3. **Precision and Recall:** Ensuring a balance between identifying actual conversions while minimizing false predictions.
  4. **AUC-ROC Score:** 0.88, indicating strong model performance.
  5. The optimal probability threshold for lead classification was determined to be 0.42, balancing precision and recall.

# Lead Scoring & Business Interpretation

* 1. Leads were assigned a probability score indicating their likelihood of conversion.
  2. A cutoff of **0.42** ensured that the company could focus on high- quality leads.
  3. The model enables the sales team to **prioritize** efforts on leads most likely to convert, reducing wasted resources on low-potential leads.

# Key Learnings & Recommendations

* 1. **Effective Lead Segmentation:** High engagement levels (time spent, interactions) are strong indicators of conversion likelihood.
  2. **Optimized Sales Efforts:** By focusing on high-scoring leads, the sales team can work more efficiently, increasing conversions.
  3. **Future Improvements:** Exploring other models like Random Forest or Gradient Boosting for further accuracy improvements.
  4. **Additional Data Integration:** Incorporating demographic or behavioural data could enhance predictive power.

# Conclusion

The developed lead scoring model provides a structured approach for identifying high-potential leads, allowing X Education to improve its conversion rate and allocate sales efforts effectively. By implementing this data-driven strategy, the company can significantly enhance its marketing and sales efficiency, ultimately driving higher revenue and customer acquisition.