

Lead Scoring Model – Case Study Summary Report

Objective

The objective of this project was to develop a predictive lead scoring model for X Education, an online education platform, to support and enhance its sales conversion strategy. By assigning a probability score to each incoming lead, the model enables the sales team to efficiently prioritize follow-ups, reduce resource wastage, and increase the overall conversion rate, aligning with the company's target of achieving 80% conversion efficiency for high-potential leads.

Project Approach

The project followed a structured and iterative approach comprising five key phases:

1. Data Audit and Cleaning

The initial dataset underwent a comprehensive quality assessment. Redundant identifiers (Prospect ID, Lead Number) were dropped. Features with a high proportion of missing values (over 70%) were removed to maintain data integrity. Remaining null values were treated using context-appropriate imputation strategies.

2. Exploratory Data Analysis (EDA)

EDA provided valuable insights into lead behaviour and engagement.

- Leads from the **Welingak website** showed the highest conversion rates.
- Channels like **Direct Traffic** and the **Lead Add Form** emerged as high-yield sources.
- The "Country" column was removed due to lack of variability (predominantly "India"), which offered minimal predictive value.

3. Data Transformation

To ensure compatibility with modelling algorithms, categorical variables were transformed using **one-hot encoding**, while numerical variables were **standardized**. These transformations helped in improving algorithm performance and interpretability.

4. Feature Selection

A **Recursive Feature Elimination (RFE)** approach, using logistic regression as the estimator, was adopted to identify the most predictive variables. Additional refinement was achieved by evaluating **p-values** and **Variance Inflation Factors (VIF)** to address multicollinearity and improve model stability.

5. Model Development and Evaluation

A **Logistic Regression** model was trained due to its simplicity, transparency, and suitability for binary classification tasks. The model's performance was validated using multiple metrics:

- **Accuracy:** 92.78% (Test Set)
- **Sensitivity:** 91.98%
- **Specificity:** 93.26%
- **ROC-AUC:** 0.97

These results indicate strong discriminatory capability and robustness across both training and testing datasets.

An **optimal probability threshold** was selected post-modelling to ensure the right trade-off between lead acquisition (sensitivity) and operational efficiency (specificity), in line with business goals.

Key Learnings

- **Data Quality Drives Accuracy:** Effective preprocessing and cleansing were instrumental in achieving reliable results.
 - **Feature Engineering Matters:** Proper selection and transformation of features significantly impacted the model's performance and interpretability.
 - **Evaluation Should Be Multi-Dimensional:** Reliance on multiple performance metrics provided a more holistic understanding of model strengths and trade-offs.
 - **Business Alignment is Crucial:** The determination of cut-off scores and prioritization logic was guided by business use cases, ensuring that the solution was both actionable and scalable.
-

Conclusion and Next Steps

The project successfully delivered a lead scoring model that enables X Education to prioritize outreach and improve conversion efficiency. The solution has immediate applicability in driving sales productivity and can serve as the foundation for a more advanced, data-driven customer engagement pipeline.

Moving forward, it is recommended to explore more complex machine learning techniques such as **Random Forests** or **Gradient Boosting** to further enhance predictive performance. Additionally, integrating real-time scoring and feedback loops could further improve lead management processes and business outcomes.