Summary Report: Lead Score Analysis – Anusha N

Project Overview

The goal of this project was to help X Education—a digital learning platform—improve its lead conversion rate, which stood at just 30%. Our task was to build a logistic regression•based prediction model that could assign a score to each lead, thereby identifying “Hot Leads” with a high likelihood of converting. This model aims to enable the sales team to focus their efforts more efficiently and improve overall conversion rates, ideally pushing it closer to 80%. Logistic regression was selected for its interpretability, scalability, and simplicity from a business perspective.

Process Followed

Data Understanding & Cleaning

We began with a dataset of 9,240 leads and 37 features, covering user behavior, marketing channels, and engagement. We identified key issues: missing values, redundant IDs, and placeholder categories like “Select.” Variables with over 45% missing data (e.g., Lead Quality, Asymmetrique scores) were dropped. Categorical features were cleaned and standardized, and binary columns with negligible variation were removed. This ensured that only informative features remained.

Feature Engineering

We created two new variables:
• Engagement_Score: a composite metric combining visit frequency, page views, and time spent on the site.
• Visit_Page_Ratio: page views per visit normalized.
Categorical variables were one•hot encoded, and binary “Yes/No” fields were mapped to 1/0. All numeric features were scaled for uniformity.

Exploratory Data Analysis (EDA)

We explored feature distributions, correlations, and segmentation by conversion status. Tags like “Will revert after reading the email” and activities like “SMS Sent” showed strong positive correlation with conversions. Outliers in numeric features were capped using the IQR method to preserve distribution shape without skewing results.

Model Building

Using Recursive Feature Elimination (RFE), we selected the top 15 features contributing most to conversion. The logistic regression model was then refined using p•values from statsmodels to remove statistically insignificant or unstable predictors. The final model retained interpretable features directly linked to business logic and lead behavior.

Model Evaluation

The model achieved a pseudo R^2 score of ~0.59, indicating strong explanatory power. It performed well on evaluation metrics including ROC•AUC and confusion matrix accuracy. The model's coefficients were intuitive—Tags, Last Activity, and Lead Source emerged as powerful predictors, offering clear insights to stakeholders.

Key Learnings

- Cleaning and transforming categorical data was foundational.
- Smart feature engineering (Engagement Score) significantly boosted model relevance.
- Logistic regression remains a strong baseline model when supported by high•quality features.
- Behavioral signals like Tags and Last Activity outperformed demographic attributes.

Challenges

- Managing high•cardinality categorical variables without overfitting.
- Ensuring model accuracy while keeping it interpretable for business use.
- Carefully handling missing values to avoid introducing bias.

Conclusion

This lead scoring model is a game•changer for X Education. It enables the business to focus efforts on the most promising leads, boosting efficiency and potentially improving conversion rates significantly. The project emphasizes the importance of clean data, relevant features, and balancing simplicity with impact. Even with a basic model like logistic regression, when applied thoughtfully, the business outcomes can be substantial.