SUMMARY

The “Lead Scoring Case Study Using Logistic Regression” focuses on developing a data-driven model to identify high-potential leads for X Education, an online education provider. The business problem involved a low conversion rate of about 30%, meaning that most leads generated through website forms were not converting into paying customers. To enhance sales efficiency, the goal was to assign each lead a score from 0 to 100, allowing the sales team to focus on “hot leads” with the highest likelihood of conversion.

Data Preparation and Cleaning
The process began with importing and inspecting the dataset to understand its structure and quality. Several quality issues were identified during cleaning — missing values, categorical inconsistencies (such as “Select” values acting as nulls), and outliers. Missing categorical values were imputed with the mode, while numeric ones used the median. Columns with over 70% missing data were dropped. Dummy variables were then created to prepare the dataset for modeling. Numeric features were scaled to ensure uniform contribution to predictions, and multicollinearity was checked using the Variance Inflation Factor (VIF).

Exploratory Data Analysis (EDA)
EDA revealed several key business patterns:

- Lead Source: Leads from Google searches, direct website traffic, and referral sources showed higher conversion likelihoods.
- Lead Origin: API and landing page submissions performed better than manual lead forms.
- Specialization: Courses related to HR, finance, and marketing showed higher conversion rates than other streams.
- Behavioral Traits: Leads spending more time on the website and engaging via emails (opened, clicked, or SMS interactions) were more likely to convert.
- Occupation and Geography: Working professionals, especially from India and the UAE, were stronger prospects than students or unemployed individuals.

Model Building and Evaluation
A logistic regression model was built to predict conversion probabilities. Recursive Feature Elimination and statistical significance testing guided feature selection. The model was trained on 70% of the data and tested on 30%, achieving an accuracy of about 80% and sensitivity/specificity of 77%. The ROC curve helped identify an optimal cutoff probability of 0.42 for classifying hot leads. Precision and recall balance ensured efficient use of sales resources.

Learnings and Insights
Time spent on the website, lead source type, and occupation were major influencers of conversion.

Email and SMS interactions were strong behavioral signals for prioritizing leads.

The model successfully improved the targeting strategy, reducing wasted sales effort.
The project showcased how logistic regression can translate business challenges into actionable insights, boosting lead conversion rates through intelligent prioritization.