Summary Report: Lead Scoring Case Study

The lead scoring case study for X Education aimed to enhance their lead conversion rate, currently at 30%, by developing a logistic regression model that assigns a lead score between 0 and 100. This score enables prioritization of "hot leads" for the sales team, targeting a conversion rate of 80%. Here's how the assignment was approached and the key learnings:

Approach

1. **Problem Understanding**:
    - The objective was to identify key attributes influencing lead conversion and develop a predictive model to optimize the sales funnel.
2. **Data Understanding and Cleaning**:
    - A dataset of ~9,000 leads was analysed.
    - Columns with more than 30% missing data (e.g., "Country" and "City") were dropped.
    - Missing values in the "Specialization" column were imputed as "Not Available."
    - Outliers in numerical variables (e.g., "Total Visits" and "Page Views Per Visit") were capped at the 95th percentile.
3. **Exploratory Data Analysis (EDA)**:
    - Key patterns were identified, such as high conversion rates for specific lead sources (e.g., Google and Reference) and demographic groups (e.g., Working Professionals).
    - Correlation analysis revealed moderate relationships among variables, guiding feature selection.
4. **Model Development**:
    - Logistic regression was chosen due to its interpretability.
    - Categorical variables were converted into dummy variables, and numerical features were scaled using MinMaxScaler.
    - Recursive Feature Elimination (RFE) and statistical significance testing (p-values and VIF) refined the feature set.
    - The optimal model (Model-4) was evaluated with metrics such as sensitivity, specificity, and accuracy. Adjusting the cut-off score to 0.4 improved performance:
        - Sensitivity: 0.82
        - Specificity: 0.77
        - Accuracy: 0.79
5. **Insights and Business Recommendations**:

- High engagement (e.g., time spent on the website) and specific activities (e.g., phone conversations) were the strongest predictors of conversion.
- Channels like "Lead Add Form" and "Olark Chat" were identified as key sources for high-quality leads.
- Recommendations included focusing on high-engagement leads, optimizing content, prioritizing productive communication channels, and reducing efforts on low-converting segments.

Key Learnings

1. **Data Preparation is Critical**:
   - The cleaning and imputation of missing values were pivotal in ensuring a robust dataset for modelling.
2. **Importance of Feature Selection**:
   - Recursive feature elimination combined with statistical tests allowed us to select impactful variables while avoiding multicollinearity.
3. **Model Evaluation**:
   - Metrics like sensitivity and specificity provided a balanced evaluation of model performance, highlighting its ability to minimize both false negatives and false positives.
4. **Actionable Business Insights**:
   - The analysis uncovered specific variables that are directly actionable, enabling targeted marketing strategies and efficient resource allocation.
5. **Interplay Between Data and Business Goals**:
   - Tailoring the model and recommendations to align with business objectives (e.g., 80% conversion target) demonstrated the importance of domain context in data science.

By applying these methods and learnings, the case study successfully provided a roadmap for X Education to achieve its desired growth, optimize sales efforts, and maximize ROI.