## Summary of Lead Score Assignment

### Approach to the Analysis:

We began the analysis by exploring and cleaning the dataset. This involved handling missing values, removing redundant or irrelevant columns, and encoding categorical variables. For instance, we filled missing entries in key columns like 'Lead Quality' with 'Unknown' and dropped features like 'Prospect ID' and 'Lead Number' and a few others whilst referring to the data dictionary, which didn't contribute to the model.

We also identified columns with excessive missing data, such as 'Tags' and 'Asymmetrique' features, and decided to drop them. Categorical variables were converted to dummy variables

After preparing the data, we split it into training and testing sets (70% for training, 30% for testing). The logistic regression model was then trained, and its performance was evaluated using metrics such as accuracy, precision, recall, and F1 score. A confusion matrix was also generated to visualize the classification results.

### Key Findings:

The logistic regression model showed strong performance, achieving an **accuracy of 85.1%** on the test set. The precision was **82.2%**, meaning that the majority of the leads predicted to convert did so. The recall stood at **77.2%**, indicating that the model captured a good portion of actual conversions. With an **F1 score of 0.80**, the model was well-balanced between precision and recall, making it effective in real-world use.

Several features contributed significantly to lead conversion:

- **Total Time Spent on Website**: The more time a lead spent on the website, the higher the likelihood of conversion, reflecting their interest and engagement.
- **Lead Source (Google)**: Leads that arrived through Google searches had a higher probability of converting, emphasizing the importance of organic search traffic.
- **Email Interactions**: Leads who interacted with emails were more likely to convert, indicating that email marketing is an essential strategy.

### Business Insights and Recommendations:

1. **Invest in High-Performing Channels**: The analysis highlighted Google as a critical source for high-quality leads. Similarly, the effectiveness of email marketing should be leveraged further by refining email content and frequency to keep leads engaged.
2. **Adjust Strategy Based on Business Needs**: During high-pressure times, like when interns are hired for two months, lowering the lead conversion threshold (from 0.5 to 0.4 or 0.3) would allow the team to capture a wider range of potential leads. Conversely, during low-pressure times (after hitting quarterly targets), increasing the threshold (ex: to 0.7) would focus efforts on the most likely conversions, reducing unnecessary outreach.
3. **Prioritize High-Scoring Leads**: The lead scoring system derived from this model provides a clear way to prioritize sales efforts. By focusing on leads with higher scores, the team can efficiently allocate resources, improving overall conversion rates.

**Summary of Lead Score Assignment**

## Learnings:

The project underscored the importance of thorough data preprocessing. Managing missing values, eliminating irrelevant columns, and correctly encoding categorical variables had a substantial impact on the model's effectiveness. Logistic regression proved to be a solid choice for predicting lead conversion, thanks to its simplicity and interpretability.

Another key takeaway was the flexibility offered by logistic regression models. By adjusting the classification threshold, X Education can dynamically adapt its strategy to different business needs, balancing aggressive lead conversion with more conservative approaches when required. Overall, this assignment highlighted how data-driven decision-making can lead to more efficient sales processes and better business outcomes.