Objective: The primary objective of this case study was to develop a predictive lead scoring logistic regression model for X Education, an online education platform, to enhance its sales and marketing efficiency. The goal was to prioritize leads based on their likelihood of conversion, enabling the sales team to focus on high-potential prospects.

Methodology:

1. **Data Collection & Exploration:**

   - The first step involved collecting and exploring the data, which included information on leads such as demographics, past behaviour, interaction with X Education content, and previous engagement with the sales team.
   - Key variables included demographic details (age, location), behavioural data (website visits, email interactions), and past purchase history.

2. **Exploratory Data Analysis:**

   - This process includes handling missing values, creating categorical variables, and scaling numerical features.
   - Data was also balanced to ensure that the model is not biased towards the majority class and all data have almost the same scale to build the model.

3. **Feature Selection:**

   - Features were selected based on their correlation with the target variable (lead conversion). We used correlation matrices for this selection.

4. **Model Building:**

   - Three models were built, and we kept building the models till the time p-values were almost equivalent to 0 and VIF values are below 3
   - The third model was our final model
   - Each model was evaluated based on key performance metrics such as accuracy, precision, recall, and the area under the ROC curve (AUC-ROC).

5. **Model Evaluation:**

   - The final model was selected based on a combination of metrics, with a particular emphasis on the precision and recall, as these directly impact the effectiveness of the lead scoring system.
   - The chosen model was further fine-tuned using hyperparameter optimization techniques to improve its performance.

6. **Implementation & Testing:**

   - The lead scoring model was deployed in a test environment to evaluate its performance in real-time scenarios.

Learnings:

1. **Quality of data is very important:**

   - The importance of quality data is highly recommended in the entire process. Inaccurate or incomplete data can lead to skewed results, impacting the effectiveness of the lead scoring model.

2. **Proper Feature Selection is important:**

   - Proper feature selection significantly influence performance. To make this work, it is really important that we understand what the business is all about and what are the key factors which can impact the business.

3. **Continuous Learning and regular updates to Model:**

   - Building a lead scoring model is not a one-time activity. Continuous monitoring, learning from new data, and iterating on the model is necessary to keep it relevant and effective.

4. **Cross-functional Collaboration:**

   - Collaboration with different teams, such as sales and marketing, is vital for understanding the practical challenges and ensuring that the model aligns with business needs.

Conclusion: The lead scoring model developed for X-Education demonstrated the potential to significantly improve lead prioritization and conversion rates. The learnings from this case study underscore the importance of data quality, model interpretability, and continuous iteration to maintain the effectiveness of predictive models in dynamic business environments.