Summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

X Education, an online education company, faces challenges in converting leads into paying customers, with a typical conversion rate of only 30%. To enhance this process, the company aims to identify "Hot Leads"—those most likely to convert—by developing a predictive model that assigns lead scores. The goal is to increase the lead conversion rate to around 80%, thereby improving sales efficiency and overall revenue.

The approach to achieving this goal involves several structured steps, starting with data understanding and preprocessing. The company has a dataset of approximately **9,240 leads** and **37 features** which includes various attributes such as **Lead Source, Total Time Spent on Website, Total Visits, and Last Activity**. The first step is to clean the data by handling missing values, particularly in categorical variables where the level 'Select' is treated as a null value. This ensures that the dataset is robust and ready for analysis.

The missing values in categorical variables are replaced by **the mode** and the missing values in numerical variables are replaced with **the median** values. We are using median values instead of mean values, as most of the variables have skewed distributions.

Some of the features such as ‘specialization’, last notable activity’, ‘last activity’ and ‘lead source’ had many sub features. We did feature engineering and grouped those sub-features under one category. These along with other categorical and binary features are then converted to dummy variables for further analysis in the model.

In the numerical features Total Visits, Total Time Spent on Website, Page Views Per Visit were treated for outliers by removing top 1% outliers using quantiles.

The data is then checked for class imbalance which turned out to be 1.6 and this is moderate class imbalance.

Feature selection is also crucial, as it helps identify which attributes are most relevant to the target variable, 'Converted'. This was achieved through correlation analysis and Recursive Feature Elimination process.

Once the data is prepared, the next step is to build a logistic regression model. The dataset is split into training and testing sets, with an 70-30 ratio. The logistic regression model is then fitted to the training data, utilizing regularization techniques to prevent overfitting.

We used RFE to reduce the features to 15. With the features obtained we again look for p value and VIF to eliminate features having higher p-value. We stop once the VIF value is less than 5 and p value less than 0.05.

Model evaluation is essential to ensure its effectiveness. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score are used to assess performance on the test set. A confusion matrix provides a visual representation of the model's predictions, helping to identify areas for improvement. Additionally, the classification threshold can be adjusted to align with the company's target conversion rate, ensuring that the model meets business objectives.

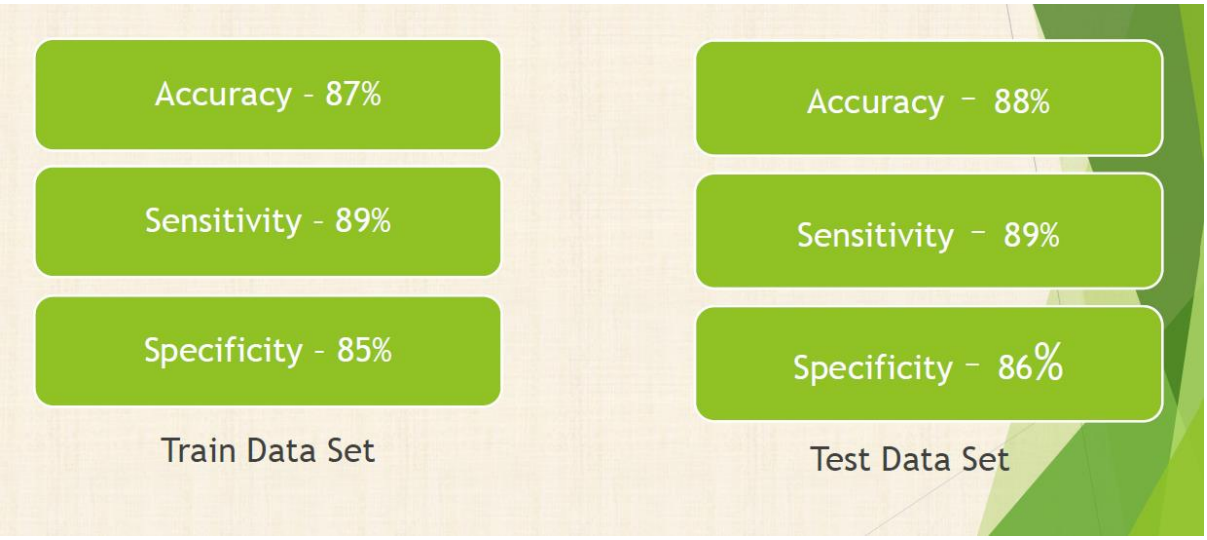
The prediction on the train dataset was done with probability 0.5 for which the accuracy was nearly 87%. Then we did a ROC to check the trade off between specificity and sensitivity.

Which was around 94% indicating the model build to be good.

Then we tried for different probability and found that 0.25 would be more appropriate

probability to balance between sensitivity and specificity and accuracy.

After training, the model predicts the probability of conversion for each lead, which is then scaled to create a lead score ranging from 0 to 100. Higher scores indicate a greater likelihood of conversion, allowing the sales team to prioritize their outreach efforts.



The implementation of this lead scoring model offers several key benefits. By focusing on leads with higher scores, the sales team can optimize their efforts, leading to improved conversion rates and more efficient use of resources. The model's adaptability allows it to respond to changing business needs, ensuring that X Education remains competitive in the market. Continuous monitoring and updating of the model with new data will further enhance its accuracy and effectiveness over time.