SUMMARY

About Leads and Data Set

X Education aims to improve its lead conversion rate, which is currently around 30%. The company receives numerous leads daily through various channels, including their website and referrals. However, only a fraction of these leads convert into sales. To enhance efficiency, X Education wants to identify "Hot Leads"—the most promising prospects among the leads they acquire. By focusing their sales team's efforts on these high-potential leads, the company hopes to increase its conversion rate. This targeted approach should help streamline their sales process, reduce wasted efforts, and ultimately boost the success rate of their lead conversion efforts.

Summary

Introduction to Python Libraries and their functionalities :

Python libraries such as NumPy, Pandas, Matplotlib, Seaborn, scikit-learn, and Warnings play crucial roles in data analysis and visualization. **NumPy** provides powerful array and matrix operations, essential for numerical computations, while **Pandas** offers flexible data structures like DataFrames for efficient data manipulation and analysis. **Matplotlib** is widely used for creating static, animated, and interactive visualizations, helping to represent data graphically. **Seaborn**, built on top of Matplotlib, enhances statistical visualizations with a high-level interface for more attractive and informative graphics. **scikit-learn** is pivotal for implementing machine learning algorithms, offering tools for classification, regression, and clustering. Lastly, the **Warnings** module helps manage and display warning messages about potential issues in code execution, allowing developers to address them without halting program execution. Together, these libraries form a comprehensive toolkit for effective data analysis and machine learning in Python.

Dataset and Data Loading :

Handling the dataset from X Education, which consists of 9,240 rows and 37 columns, involves a systematic approach to ensure data quality and model efficacy. The first step is defining the problem statement by clearly specifying the input features and the output variable, such as predicting student performance based on various predictors. Data gathering follows, where the dataset is imported using Pandas.

Next, Exploratory Data Analysis (EDA) is crucial. This includes univariate analysis to explore individual feature distributions, bivariate and multivariate analyses to uncover relationships and patterns among features. Handling missing values is essential; columns or rows with excessive NULL values might need imputation or removal. Additionally, addressing outliers is important to prevent distortion of analysis and model performance. Data types of columns must be encoded properly for machine learning algorithms to process, while skewness should be corrected to normalize distributions. Scaling features ensures that all variables contribute equally to the model.

Feature engineering and selection involve creating new features or refining existing ones to enhance model performance. Finally, model training and evaluation are performed, where different algorithms are tested and assessed for accuracy, precision, recall, and other metrics, ensuring the model is robust and generalizable. This comprehensive approach prepares the dataset for effective predictive modeling.

Some Required points to do here is :

- Problem statement- define input and output variables.
- Data gathering (read the data
- Exploratory data analysis(EDA)
  - Univariate, Bivariate, Multivariate Analysis
  - Missing values handling
  - Datatype of columns(Encoding)
  - Outliers handling
  - Skewness
  - Scaling
- Feature engineering and selection
- Model training and evaluation.

Training and Prediction Process :

For X Education's dataset, the training and prediction process using logistic regression involves several key steps. Start by importing and preprocessing the data, including handling missing values and encoding categorical variables. Split the dataset into training and test sets. Train the logistic regression model on the training data and use it to make predictions on the test set. Evaluate the model's performance using metrics like accuracy and F1 score, and refine the model based on these results to enhance predictive accuracy.

Model Evaluation :

Model evaluation for the logistic regression model involves using scikit-learn metrics to gauge performance. The **accuracy** metric shows the proportion of correct predictions; here, the model achieved 89% accuracy on the training set and 88% on the testing set, indicating strong performance and good generalization to unseen data. The **classification report** provides a comprehensive view, including precision, recall, and F1 score for each class. **Precision** measures the accuracy of positive predictions, **recall** indicates the model's ability to capture all relevant instances, and the **F1 score** balances precision and recall into a single metric. Together, these metrics highlight the model's effectiveness and areas for improvement, ensuring it performs well across different aspects of classification.

At the end some **Insights** I got is :

- I dropped some useless column which are just increasing the accuracy in generally.
- The X – Education sector most demanded in the hype of knowledge as the no of customers come here to increase their knowledge.
- The most valuable features are the type of medium that they are using to get the knowledge mostly came to upgrad and mostly are here to get employed in the area.