**ABB\_Assessment - Loan Prediction**

1. **Dataset Understanding**: I start by thoroughly reviewing the dataset in either Excel or Python, identifying the columns and their significance.
2. **Descriptive Analysis**: I use the describe() function to explore the numeric columns in the dataset and understand their basic statistics.
3. **Handling Missing Data**: I check for any null values in the numeric columns and replace them with the mean values, ensuring minimal impact on the data's integrity.
4. **Outlier Detection and Removal**: Using boxplots, I identify outliers in the dataset, particularly in the Item\_Visibility column. I apply the Interquartile Range (IQR) technique to remove outliers, replacing them with the upper limit values.
5. **Exploratory Data Analysis (EDA)**: After cleaning the data, I move on to performing EDA, where I use various visualizations to identify relationships and patterns within the dataset.
6. **Model Selection**: Following the EDA, I proceed to select the best predictive model for the dataset. For this, I use XGBoost, a powerful algorithm for regression tasks.

**Model Training and Evaluation Process**:

* **Data Splitting**: I divide the dataset into features (X) and target variable (y), and then split the data into training and testing sets using train\_test\_split for model evaluation.
* **Feature Scaling**: The features are scaled using StandardScaler to ensure consistency across the data and improve model performance.
* **Model Training with XGBoost**: I train an XGBoost model with tuned parameters like learning rate, max depth, and others to optimize its predictive power. Hyperparameter tuning is done using GridSearchCV for better accuracy.
* **Model Evaluation**: After training the model, I predict the target variable and evaluate its performance using metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² score to measure how well the model fits the data.

**Handling Test Data and Predictions**:

* **Preprocessing Test Data**: For the test dataset, I perform label encoding on categorical variables to convert them into numeric format, enabling compatibility with the model.
* **Prediction**: I use the best-trained model to predict sales for the test data.
* **Submission Creation**: Finally, I prepare a submission DataFrame with the predicted sales values, ensuring the test data's original identifiers are maintained.