**Report**

#### ****Approach Taken****

**Data Preprocessing:**

* 1. The data was cleaned by handling duplicates and missing values. Numeric conversion for specific columns, such as 'Cibil Score,' was performed to ensure consistency.
  2. Categorical and numerical columns were separated for further analysis.
  3. Columns which had more than 80% of missing data were dropped
  4. Column which had less than 10% of missing data those rows were removed
  5. Analyzed the column which are not required for model training and dropped them
  6. Few column were having only on type of data inside were removed as they don’t play any role in model training

**Exploratory Data Analysis (EDA):**

* 1. Histograms and Kernel Density Estimates (KDEs) were plotted for numerical columns to understand the distribution of data.
  2. Correlation of numerical columns with target column were calculated for feature selection
  3. Value counts for categorical and numerical columns were analyzed to determine the frequency distribution of categories.
  4. Encoded the categorical column using Labelencoder
  5. Calculated the correlation of categorical column with target column
  6. Selected the relevant column required for model training by analyzing the correlation of columns with target column

**Model Building:**

* 1. Trained the model with several classification techniques including-LogisticClassifier ,Naive Bayes Classifier,SVC,Gradient Boosting Classifier and XGradient Classifier.
  2. Three best working models were selected: Random Forest, Gradient Boosting, and XGBoost.
  3. Randomized search was used to tune hyperparameters for each model, optimizing for F1 score through cross-validation.
  4. The models were evaluated on both training and validation datasets using accuracy and F1 score.

**Model Selection:**

* 1. The performance of the models on training and validation data was compared.
  2. The Random Forest model was selected as the best model based on its ability to balance training and validation accuracy, minimizing the risk of overfitting.

**Model Saving:**

* 1. The best model, along with the label encoders, was saved for future predictions.

#### ****Insights and Conclusions from Data****

### Data Overview:

The dataset contains 10,000 rows and 55 columns with various types of information such as:

* **Identification and Application Details**: Dealer ID, Application Login Date, Branch Names.
* **Customer Personal Information**: First Name, Last Name, Mobile, Email, Gender, DOB, Aadhar Verification status, CIBIL Score, Address Type, Marital Status.
* **Asset Information**: Total Asset Cost, Asset Category, Applied Amount, Asset Model, Primary Asset Details.
* **Employment Information**: Employer Name, Employer Type, Employment Constitution, and Employment Type.
* **Miscellaneous**: Social Media interactions, Digital Age, and Name Match Scores.

### Key Insights:

**Missing Data**: Several columns have missing values:

* + HDB Branch State, Last Name, Employer details, and Asset details have considerable missing values.
  + CIBIL Score has about 4,297 missing values (around 43%).
  + Total Asset Cost and related details are missing for about half of the dataset.

**Numeric Columns**:

* + The "Applied Amount" ranges widely.
  + The "CIBIL Score" needs cleaning as it's currently a mix of strings and numbers.

**Categorical Columns**:

* + Gender, Marital Status, and Employment types seem important for analysis.

### Conclusion:

The dataset is rich but has missing values in several columns, particularly related to assets and personal details like CIBIL scores and employer information. This may impact modeling efforts, so handling missing data should be a priority. There's also potential for cleaning and feature engineering in categorical and numeric data for deeper analysis.

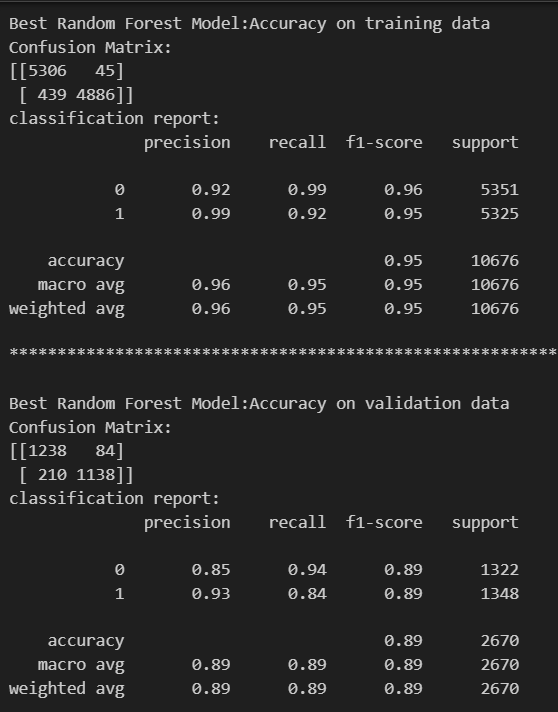
The categorical variables had distinct distributions, indicating their potential impact on the model's predictive capabilities.

The Random Forest model showed strong generalization ability, making it the most suitable model for unseen data compared to Gradient Boosting and XGBoost.

#### ****Performance on Train Data Set****

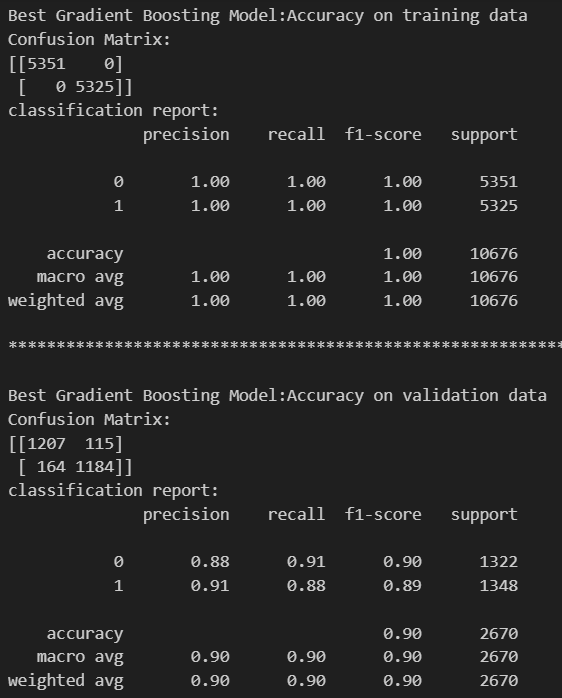
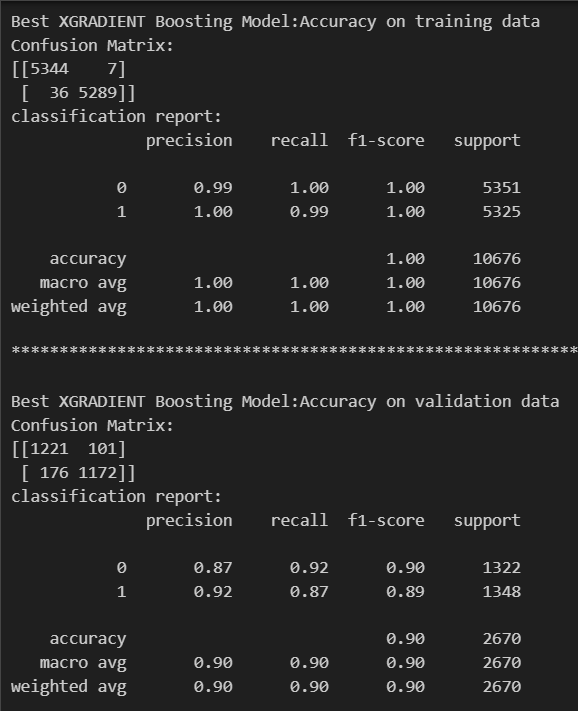
**Random Forest:**

* + The Random Forest model achieved a balance between training and validation accuracy, indicating minimal overfitting.
  + Performance metrics on the training set indicated high accuracy, and similar performance on the validation set confirmed the model's robustness.



**Gradient Boosting & XGBoost:**

* Both models performed well on the validation set but showed signs of overfitting due to slightly higher training accuracy.
* These models were less favored due to potential risks when generalizing to new data.

#### ****Performance Metrics****

* **Accuracy:** The models were evaluated on their accuracy, with Random Forest showing consistent performance across both training and validation sets.
* **F1 Score:** This was the primary metric used during hyperparameter tuning, ensuring that the models balanced precision and recall effectively.