**Supervised Learning:**

**Loan Default Prediction**

**Business Use Case:**

Loan Default Prediction: Banks need to minimize financial risk by predicting which customers are likely to default on loans. Accurate predictions can lead to better loan approval processes and risk management.

**Approach:**

● Algorithm:

Logistic Regression, Random Forest and Gradient Boosting.

● Steps:

1. Data Preprocessing: Cleaning and preparing the data, handling missing values, and encoding categorical variables.

2. Feature Engineering: Identifying important features and creating new ones if necessary.

3. Model Training: Splitting data into training and test sets, training models, and tuning hyperparameters.

4. Model Evaluation: Evaluating model performance using metrics like accuracy, precision, recall, and F1-score.

**Data Understanding:**

* Data Source: This dataset has been taken from Coursera's Loan Default Prediction Challenge
* The dataset contains 255,347 rows and 18 columns in total.
* **Key Data Dictionary:**

1. Loan ID: A unique identifier for each loan.
2. Age: The age of the borrower.
3. Income: The annual income of the borrower.
4. Credit Score: The credit score of the borrower.
5. Loan Amount: The amount of money being borrowed.
6. Interest Rate: The interest rate for the loan.
7. Loan Term: The term length of the loan in months.
8. Repayment Status: Indicate whether the loan defaulted or not.

**Data Preparation:**

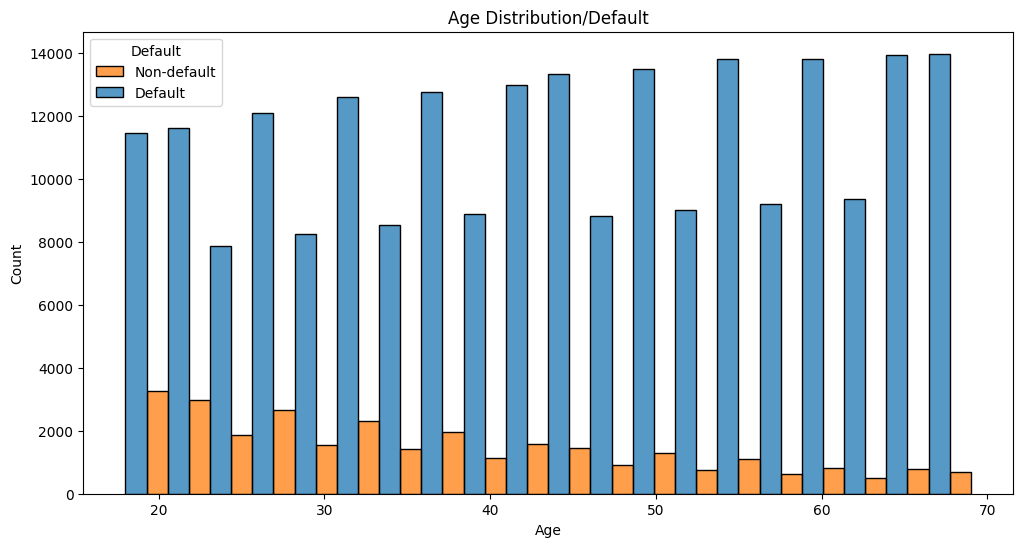
Packages: NumPy, Pandas, Matplotlib, Seaborn,Sklearn.

**Data Cleansing:**

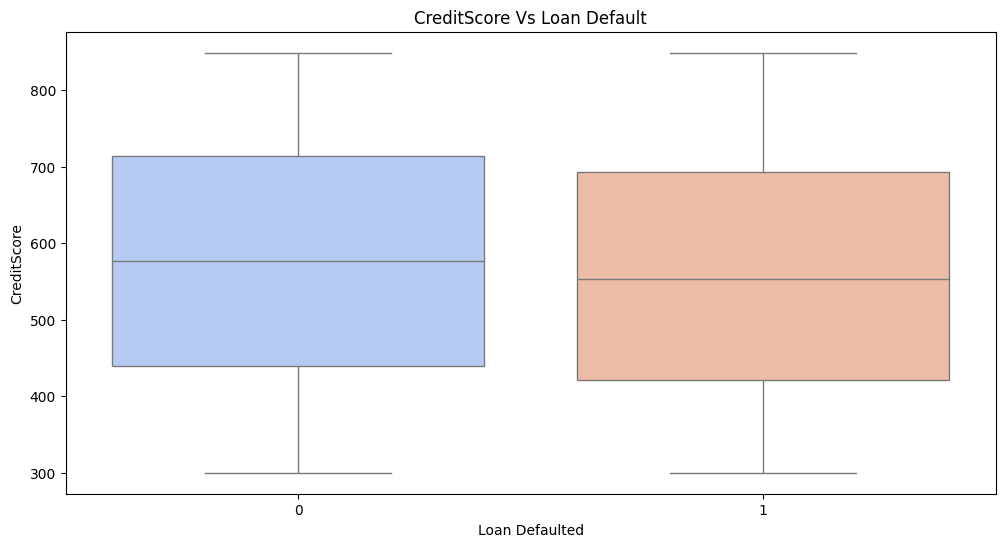
* There are no null values in the data.

**Exploratory Data Analysis:**

**Age Distribution VS Default**

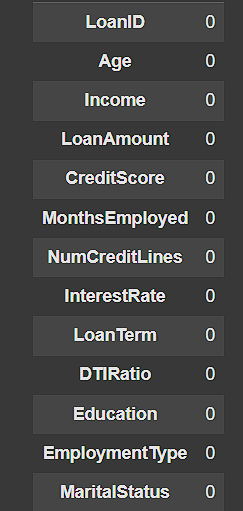


**Credit Score VS Default**

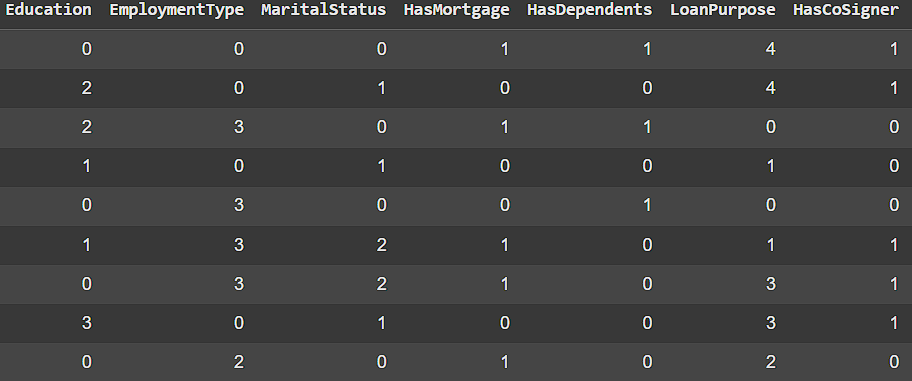


**Preprocessing:**

**Checking for Null Values**



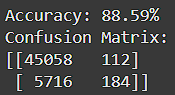
**Label Encoding**



**Models:**

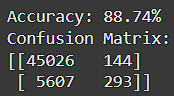
**Logistic Regression**

Logistic regression is a statistical and machine learning technique used for binary classification problems. It predicts the probability that a given input belongs to one of two classes. Despite its name, logistic regression is a classification algorithm, not a regression algorithm.



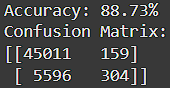
**Random Forest Classifier**

A Random Forest Classifier is an ensemble machine learning algorithm that combines the predictions of multiple decision trees to achieve better performance and robustness. It is commonly used for classification tasks and can handle both categorical and numerical data.



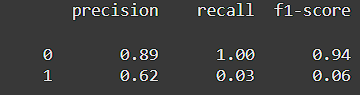
**Gradient Boosting Classifier**

A Gradient Boosting Classifier (GBC) is a powerful machine learning algorithm used for classification tasks. It builds an ensemble of decision trees in a sequential manner, where each tree corrects the errors of its predecessor. Gradient Boosting is known for its high predictive accuracy and flexibility.

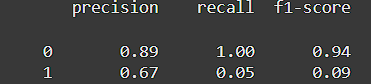


**Model Evaluation:**

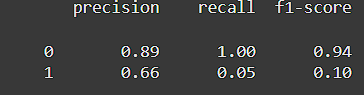
**Logistic Regression**



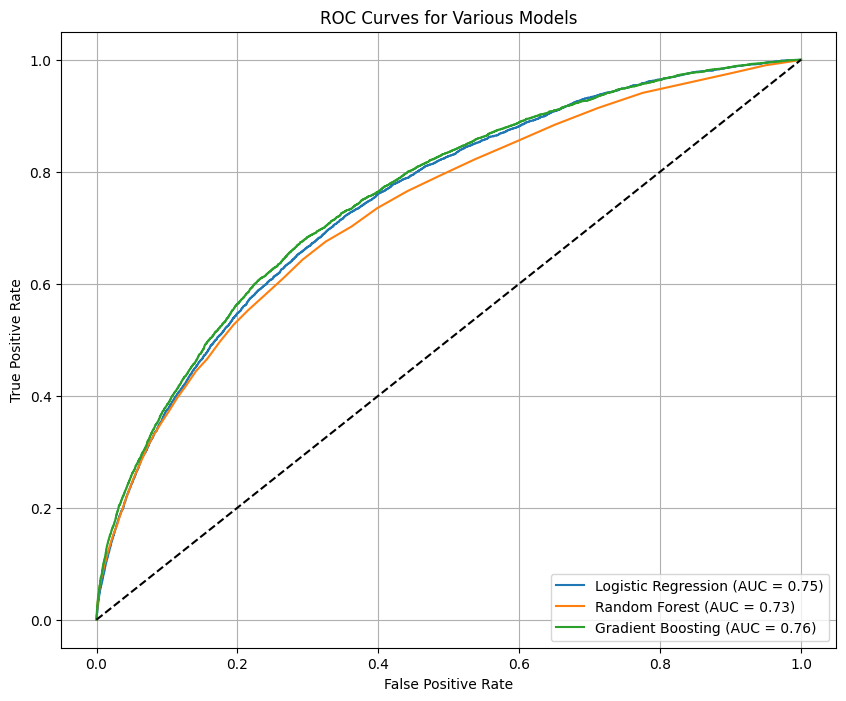
**Random Forest Classifier**



**Gradient Boosting Classifier**



**ROC-AUC Score**



**Justification:**

**Model Selection**

1.Gradient Boosting achieves the highest AUC (0.76), indicating it is slightly better at distinguishing between the classes than the other models.

2.Logistic Regression performs slightly better than Random Forest, with an AUC of 0.75 compared to 0.73.

3.The diagonal dashed line represents a random classifier (AUC = 0.5). All models perform significantly better than random guessing.

**Conclusion:**

Gradient Boosting is the best-performing model in this comparison.