**Name: Nikhil Sanjay Shinde**

**Roll No: TYITB120**

**Assignment No. 2**

**Aim:**

Implementation of Linear Regression on Loan Dataset.

**Objective:**

To implement and evaluate a Linear Regression model using Python to predict loan amounts based on different features in the dataset.

**Theory:**

Linear Regression is a fundamental supervised learning algorithm that assumes a linear relationship between the dependent variable (target) and independent variables (features). It is widely used in predictive modeling to estimate continuous values. In this project, the goal is to preprocess the loan dataset, apply a linear regression model, validate its performance using error metrics, and interpret the results based on the predictive power of different features such as income and credit history.

**Simple Linear Regression**

Simple Linear Regression aims to establish a linear relationship between an independent variable (input) and a dependent variable (output). The relationship can be represented by a straight-line equation where one variable influences the other.

In this model:

* The dependent variable is what we want to predict.
* The independent variable is what we use to make predictions.
* The intercept is the constant term that determines the starting point of the line.
* The slope defines the degree of impact the independent variable has on the dependent variable.

The objective of linear regression is to find the best-fitting line that minimizes the difference between the actual values and the predicted values.

**Estimating Coefficients**

The coefficients in the linear regression model are determined using statistical calculations based on the dataset. The slope is estimated by assessing how much the dependent variable changes with respect to the independent variable. Once the slope is obtained, the intercept can be computed easily using the average values of the dataset.

**Making Predictions**

After determining the coefficients, we can use the linear regression equation to predict values for new input data. By substituting different values into the equation, we can estimate the expected output values. A graphical representation helps visualize how well the model fits the data points.

**Assessing Model Performance**

The accuracy of a linear regression model is often evaluated using the Root Mean Squared Error (RMSE). RMSE measures the average magnitude of prediction errors and provides insight into how well the model generalizes to new data. A lower RMSE indicates better performance.

**Importance of Linear Regression**

Linear Regression is widely used in various fields such as economics, finance, healthcare, and engineering. It helps in forecasting trends, identifying relationships, and making data-driven decisions. The simplicity of the model makes it a great starting point for understanding machine learning concepts.

**Importance of Linear Regression:**

* **Predictive Analysis**: Useful for estimating continuous targets like loan amounts based on applicant data.
* **Feature Relationships**: Highlights how variables such as income and credit history influence loan amounts.
* **Efficiency**: Simple and computationally light algorithm suitable for small to medium datasets.
* **Baseline Model**: Serves as a foundation for comparing more complex regression models.

**Dataset:**

The dataset used is train\_u6lujuX\_CVtuZ9i.csv and includes the following relevant columns:

* **ApplicantIncome**: Income of the primary applicant.
* **CoapplicantIncome**: Income of the co-applicant.
* **LoanAmount**: Target variable representing the requested loan amount.
* **Loan\_Amount\_Term**: Duration of the loan in months.
* **Credit\_History**: Credit background of the borrower.
* **Property\_Area**: Area type of the property (Urban, Semiurban, or Rural).

**Steps of Implementation:**

**1. Importing Libraries**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error, r2\_score

**2. Loading the Dataset**

df = pd.read\_csv("train\_u6lujuX\_CVtuZ9i.csv")

df.head()

df.info()

df.describe()

**3. Data Preprocessing**

# Handling missing values

df['LoanAmount'].fillna(df['LoanAmount'].median(), inplace=True)

df['Credit\_History'].fillna(df['Credit\_History'].mode()[0], inplace=True)

df['Loan\_Amount\_Term'].fillna(df['Loan\_Amount\_Term'].mode()[0], inplace=True)

# Dropping rows with missing target values

df.dropna(subset=['LoanAmount'], inplace=True)

# Encoding categorical variables

df = pd.get\_dummies(df, columns=['Property\_Area'], drop\_first=True)

# Feature and target definition

X = df[['ApplicantIncome', 'CoapplicantIncome', 'Credit\_History', 'Loan\_Amount\_Term',

'Property\_Area\_Semiurban', 'Property\_Area\_Urban']]

y = df['LoanAmount']

# Splitting the dataset

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

**4. Training the Model**

model = LinearRegression()

model.fit(X\_train, y\_train)

**5. Making Predictions**

y\_pred = model.predict(X\_test)

**6. Model Evaluation**

mae = mean\_absolute\_error(y\_test, y\_pred)

mse = mean\_squared\_error(y\_test, y\_pred)

rmse = np.sqrt(mse)

r2 = r2\_score(y\_test, y\_pred)

print(f"MAE: {mae}")

print(f"MSE: {mse}")

print(f"RMSE: {rmse}")

print(f"R^2 Score: {r2}")

**7. Visualization of Results**

plt.figure(figsize=(8,6))

plt.scatter(y\_test, y\_pred, color='blue')

plt.xlabel("Actual Loan Amount")

plt.ylabel("Predicted Loan Amount")

plt.title("Actual vs Predicted Loan Amount")

plt.plot([y\_test.min(), y\_test.max()], [y\_test.min(), y\_test.max()], 'k--', lw=2)

plt.show()

**Conclusion:**

The linear regression model was effectively implemented to predict loan amounts based on features like applicant income, coapplicant income, credit history, loan term, and property area. The model demonstrated good predictive capability with evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which helped quantify the error between actual and predicted values. The R² score reflected how well the independent variables could explain the variance in loan amounts. A scatter plot comparing actual versus predicted loan amounts visually reinforced the model's accuracy and performance.

**References:**

* [GeeksforGeeks - Linear Regression](https://www.geeksforgeeks.org/ml-linear-regression/)
* <https://github.com/mrnik89/ML_Lab/blob/main/Assignment2_ML.ipynb>