Experiment-2

Loan Amount Prediction using Linear Regression

Name: Vigneshwaran S

Reg no:3122237001059

1. Aim

To develop and evaluate a Linear Regression model for predicting loan amounts using Python, Scikit-learn, and Matplotlib.

2. Libraries Used

- **pandas** – For data loading and preprocessing
- **numpy** – For numerical operations
- **matplotlib** & **seaborn** – For data visualization
- **scikit-learn** – For machine learning (Linear Regression, preprocessing, model evaluation)

3. Objective

- Handle missing values and encode categorical variables
- Perform exploratory data analysis (EDA)
- Train and test a Linear Regression model
- Evaluate performance using MSE, MAE, and R^2 score
- Visualize results like predicted vs actual values and feature importance

4. Mathematical Description

Linear Regression Equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Where:

- y = Target variable (Loan Amount)
- $x_1, x_2, ..., x_n$ = Input features
- β_0 = Intercept
- β_i = Coefficients
- ϵ = Error term

5. Code,

Github Link,

6. Included Plots

- Actual vs Predicted Loan Amount (Scatter Plot)
- Feature Importance (Bar Plot)
- Histogram
- Boxplot
- Heatmap (correlation)
- Residual Plot

7. Results Tables

Description	Result
Dataset Size	28,734
Train/Test Split Ratio	70/30
Feature(s) Used for Prediction	Age,Income,Property_price,Property_age,Current loan expenses,Total_income,credit_score
Model Used	Linear Regression
Cross-Validation Used? (Yes/No)	No

If Yes, Number of Folds (K)	-
Mean Absolute Error (MAE) on Test Set	31522.8658
Mean Squared Error (MSE) on Test Set	898067220.05
Root Mean Squared Error (RMSE) on Test Set	
R2 Score on Test Set	0.5632
Observations from Residual Plot	Residuals are mostly centered around 0, few outliers observed
Interpretation of Predicted vs Actual Plot	Predictions follow the actual values fairly well; slight deviations for high loan amounts
Any Overfitting or Underfitting Observed?	Slight underfitting

8. Best Practices

- Handle missing values carefully to avoid data loss
- Scale numeric features to improve model performance
- Split dataset into train/test/validation sets for fair evaluation
- Use visualization to interpret model behavior

9. Learning Outcomes

- Learned how to preprocess and clean data for ML
- Implemented Linear Regression using Scikit-learn
- Evaluated model performance using key metrics
- Visualized insights using Matplotlib