Experiment 2: Loan Amount Prediction using Linear Regression

Name:Darshan Parthasarathy Reg No: 3122237001009

1. Aim

To apply Linear Regression to predict the loan amount sanctioned to users using historical data and evaluate the performance of the model.

2. Libraries Used

- `pandas`
- `numpy`
- `matplotlib`
- `seaborn`
- `sklearn.linear_model`
- `sklearn.preprocessing`
- `sklearn.model_selection`
- `sklearn.metrics`

3. Objective

To build, train, evaluate, and visualize a Linear Regression model that predicts the sanctioned loan amount based on features like income, credit score, property price, etc.

4. Mathematical Description

Linear Regression assumes a linear relationship between input variables (X) and the target variable (Y). The model can be represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:

- Y: Predicted loan sanction amount
- β_0: Intercept
- $\beta_1, \beta_2, \ldots, \beta_n$: Feature coefficients
- ϵ: Residual error term

5. Code with Plot

(Kindly refer the submitted Jupyter Notebook: `ML_LAB_2.ipynb`)

6. Included Plots

- Histogram of Loan Sanction Amount
- Boxplot of Loan Sanction Amount
- Correlation Heatmap
- Scatter Plot (Income vs Loan Sanction Amount)
- Predicted vs Actual Plot
- Residual Distribution Plot
- Bar Plot of Feature Coefficients

7. Results Tables

Table 1: Cross-Validation Results (K = 5)

Fold	MAE	MSE	RMSE	R2 Score
Fold 1	21995.18	973945638.50	31208.10	0.475
Fold 2	21896.12	1134334566.00	33679.89	0.408
Fold 3	21801.21	924002070.00	30397.40	0.519
Fold 4	21493.91	906409168.20	30106.63	0.529
Fold 5	21688.81	913357744.20	30221.81	0.530
Average	21775.05	970409943.80	31122.77	0.492

Table 2: Summary of Results for Loan Amount Prediction

Description	Student's Result
Dataset Size (after preprocessing)	4380 rows
Train/Test/Validation Split	60% / 20% / 20%
Feature(s) Used for Prediction	All except Customer ID, Name
Model Used	Linear Regression
Cross-Validation Used?	Yes
Number of Folds (K)	5
Reference to CV Table	Table 1
MAE on Test Set	21573.53
MSE on Test Set	924251037.88
RMSE on Test Set	30399.35 (approx)
R2 Score on Test Set	0.530
Adjusted R2 Score on Test Set	0.525 (approx)
Most Influential Feature(s)	Income (USD), Credit Score, Property Price
Observations from Residual Plot	Roughly normal distribution
Predicted vs Actual Plot	Most points lie near the diagonal line
Overfitting/Underfitting Observed?	No major overfitting; train test error

8. Best Practices

- Filled missing values using statistical techniques
- Encoded categorical variables numerically
- Standardized numeric features to normalize ranges
- Visualized all necessary plots for EDA and performance
- Evaluated performance with metrics and cross-validation

9. Learning Outcomes

- Learned how to prepare and clean real-world datasets
- Understood application of Linear Regression for prediction
- Visualized and interpreted model accuracy using multiple plots
- Applied K-Fold validation to assess model generalization