Sri Sivasubramaniya Nadar College of Engineering, Chennai

(Autonomous Institution under Anna University)

Degree & Branch	5 years Integrated M.Tech CSE	**Semester**	V
Subject Code & Name	ICS1512 – Machine Learning Algorithms Laboratory		
Academic Year	2025–2026 (Odd Semester)	**Batch**	2023–2028
Name	Pravin G	**Reg No**	3122237001041

Experiment # 2: Loan Amount Prediction using Linear Regression

Aim:

To apply Linear Regression to predict the loan amount sanctioned to users using the dataset provided.

Libraries used:

- Numpy
- Pandas
- Scipy
- Scikit-Learn
- Matplotlib.pyplot

Description of the objective performed

- **Data Preparation:** Loaded dataset using kagglehub.dataset download() and converted it into a Pandas DataFrame.
- **Exploratory Data Analysis (EDA):**
  - Performed Numerical Column analysis using histogram and pdf
  - Performed Categorical column analysis using One way ANOVA test
  - Visualized Missing Values
  - Visualized distributions and relationships using:
    - plt.hist() for histograms
    - plt.scatter() for 2D scatter plots
    - sns.heatmap() for feature correlation matrix
- **Data Preprocessing :**

  - Handled Missing Values
  - Outlier Treatment.
  - Encoding categorical column values
  - Standardize

- **Modeling**
  - K-Fold cross validation
  - Model Fitting

- **Evaluation and Visualization**
  - Metrics MSE,MAE,RMSE,R^2
  - Visualization Actual vs Predicted Plot,Residual Plot, Bar Plot of Feature Coefficients

Mathematical Description

Model Equation

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

where:

- $\hat{y}$ is the predicted output,
- β_0 is the intercept (bias),
- β_i are the model coefficients,
- $\mathbf{X}$ is the input feature matrix,
- $\boldsymbol{\beta}$ is the coefficient vector.

Cost Function (Mean Squared Error)

$$J(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 \tag{2}$$

where:

- y_i is the actual output,
- $\hat{y}_i$ is the predicted output for the i-th observation,
- n is the number of training examples.

Code :

Feature Separation and Train Test Split

```
X = train_encoded.drop(columns=["Loan Sanction Amount (USD)"])
y = train_encoded["Loan Sanction Amount (USD)"]
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2)
```

K-Fold Cross Validation

```
lr = LinearRegression()
kf = KFold(n_splits=5, shuffle=True, random_state=42)
for fold, (train_idx, val_idx) in enumerate(kf.split(X_train), 1):
    X_t, X_v = X_train.iloc[train_idx], X_train.iloc[val_idx]
    y_t, y_v = y_train.iloc[train_idx], y_train.iloc[val_idx]

    lr.fit(X_t, y_t)
    preds = lr.predict(X_v)
```

Plots Included

Actual Vs Predicted Plot

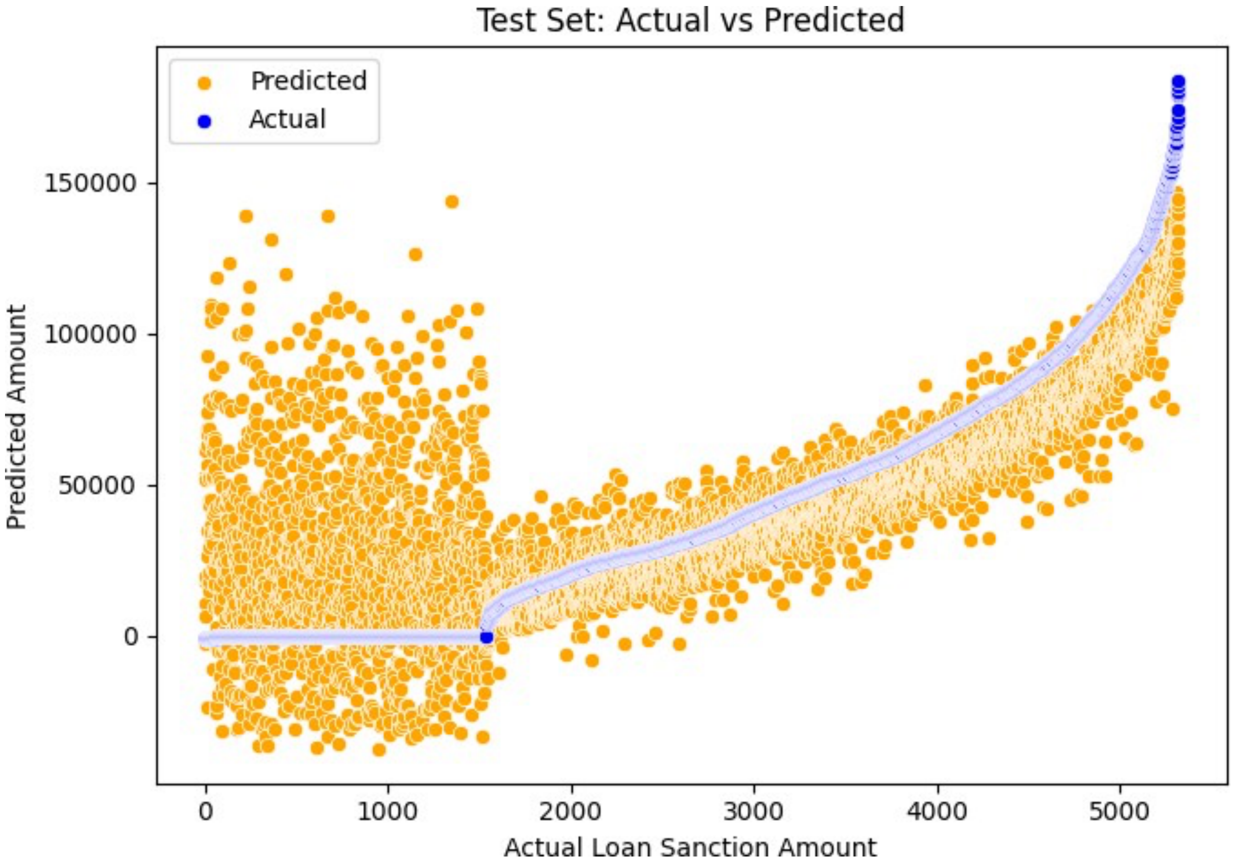

Figure 1: Actual Vs Predicted

Residual Plot

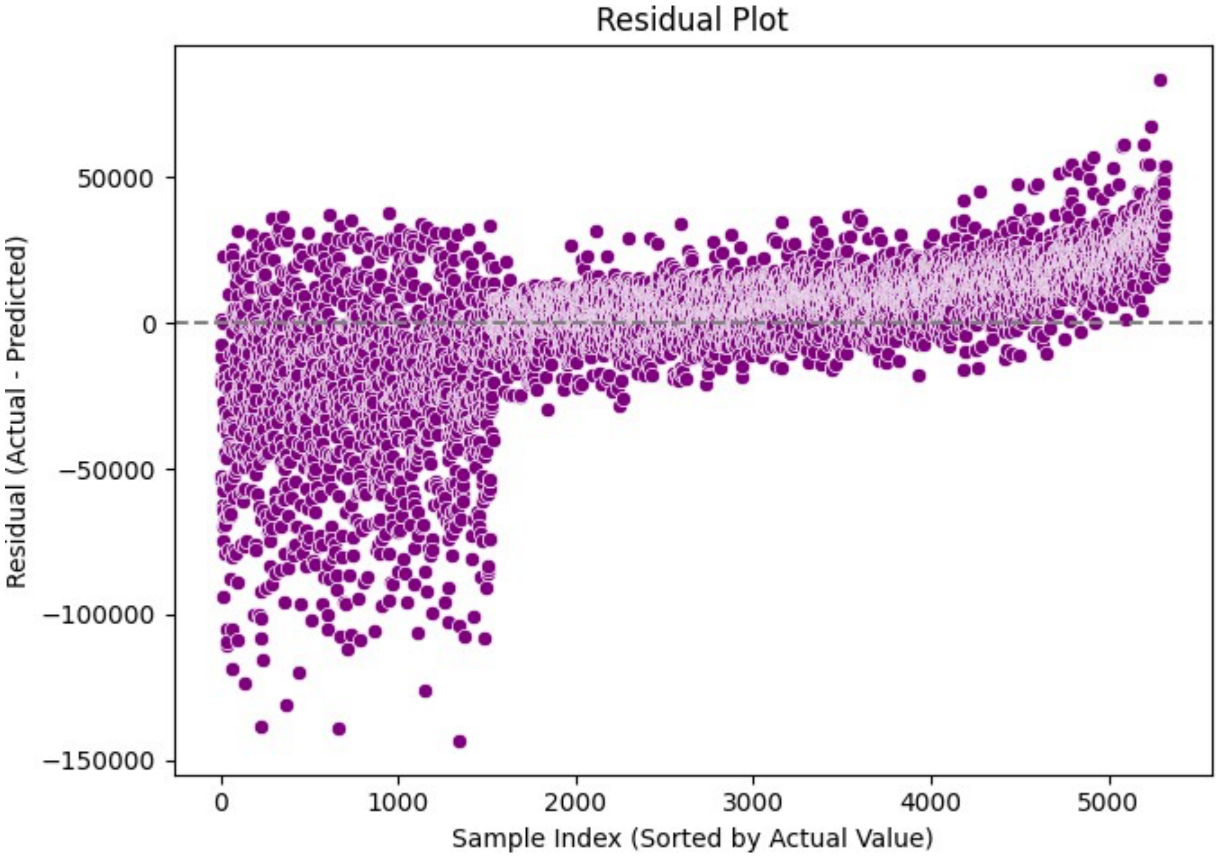

Figure 2: Residual Plot

Bar Plot

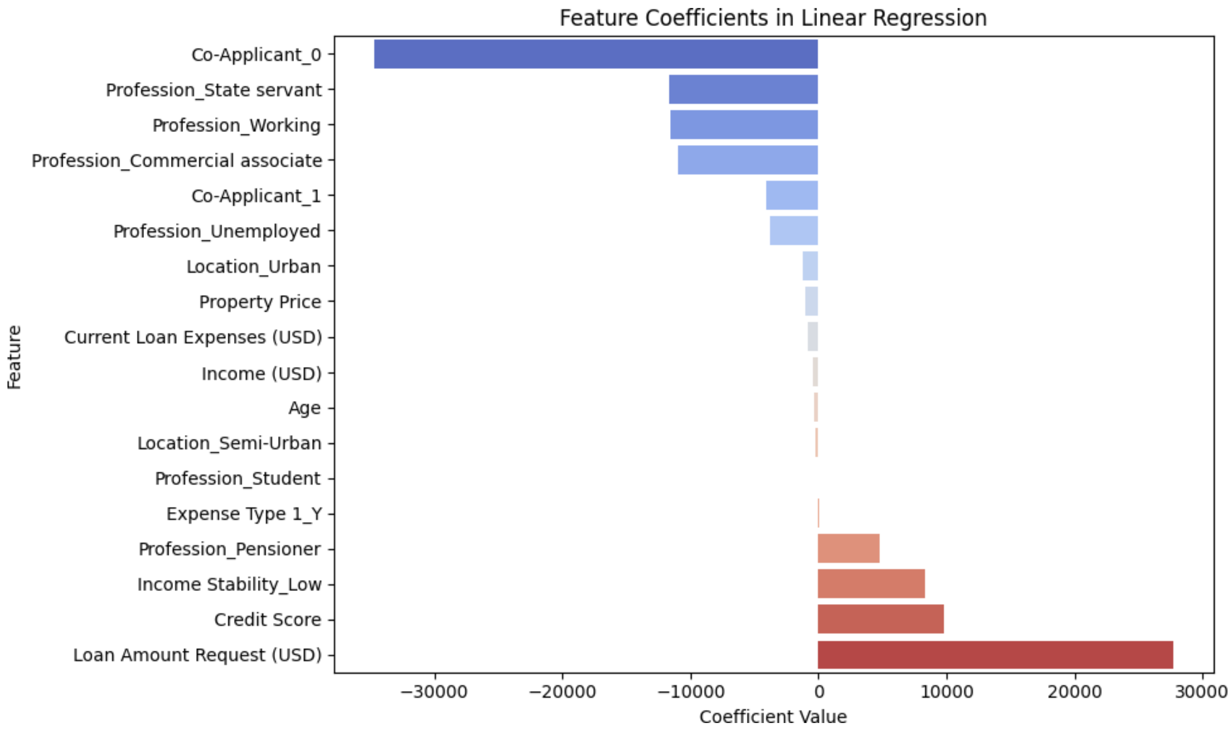

Figure 3: Bar Plot

Result Tables:

Fold	MAE	MSE	RMSE	R^2 Score
Fold 1	17414.54	617243258.78	31207	0.61
Fold 2	17826.16	637814227.93	31207	0.61
Fold 3	17646.17	618047080.23	31363	0.62
Fold 4	17989.57	658150344.53	31342	0.60
Fold 5	18052.41	650893584.19	31359	0.61
Average	17785.77	636429699.13	31342	0.61

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

Description	**Result**
Dataset Size (after preprocessing)	26585
Train/Test Split Ratio	80-20
Feature(s) Used for Prediction	Age,Loan Amount Request, Current Loan Expenses, Credit Score, Property Price, Income Stability, Profession, Location, Expense Type 1, Co-Applicant
Model Used	Linear Regression
Cross-Validation Used?	Yes
If Yes, Number of Folds (*K*)	5
Reference to CV Results Table	Table 1
Mean Absolute Error (MAE) on Test Set	17785.77
Mean Squared Error (MSE) on Test Set	636429699.13
Root Mean Squared Error (RMSE) on Test Set	31342
R^2 Score on Test Set	0.61
Most Influential Feature(s)	Loan Amount Request
Observations from Residual Plot	A Strong diagonal line indicating model might be underfitting, Spread of residuals indicates
Interpretation of Predicted vs Actual Plot	Scatter plot is not tightly packed indicating moderate to high variance, Some predicted values are negative
Any Overfitting or Underfitting Observed?	Yes Underfitting
If Yes, Brief Justification (e.g., training vs test error, residual patterns)	Many points lie far from the ideal diagonal line in the actual vs predicted line

Table 2: Summary of Results for Loan Amount Prediction

Learning Outcomes:

- Gained practical experience in data preprocessing including handling missing values and outliers.
- Understand how to train & evaluate a linear regression model.
- Learned the importance of various evaluation metrics (MAE,MSE,RMSE,R^2).