Sri Sivasubramaniya Nadar College of Engineering, Chennai
(An autonomous Institution affiliated to Anna University)

Degree & Branch	B.E. Computer Science & Engineering	Semester	V
Subject Code & Name	ICS1512 & Machine Learning Algorithms Laboratory		
Academic year	2025-2026 (Odd)	Batch:2023-2028	**Due date:25-07-25**

Experiment 2

Aim: To Apply Linear Regression to predict the loan amount sanctioned to users using the dataset provided. Visualize and interpret the results to gain insights into the model performance.

Libraries used:

- pandas
- numpy
- matplotlib.pyplot
- seaborn
- sklearn.linear_model.LinearRegression
- sklearn.model_selection.StratifiedKFold
- sklearn.metrics.mean_absolute_error
- sklearn.metrics.mean_squared_error
- sklearn.metrics.r2_score
- sklearn.preprocessing.LabelEncoder
- sklearn.preprocessing.StandardScaler
- pandas.get_dummies

theoritical description of the algorithm:

Cross-Validation Strategy

A 5-fold stratified cross-validation approach was used. This ensures that each fold maintains a balanced distribution of the target variable, leading to a more reliable and unbiased performance estimation.

Data Preparation

The training dataset was preprocessed by cleaning missing values, removing irrelevant columns, and performing feature engineering to ensure consistent and useful inputs for the model.

Feature Engineering

Custom features such as the Loan-to-Income ratio, Total Expenses-to-Income ratio, and Loan-to-Value (LTV) ratio were created. These derived metrics provided insights into the customer's financial standing and credit risk.

Data Cleaning

Columns like Customer ID, Property ID and Name, which had no predictive value, were removed. Missing values in numerical columns were imputed with the mean, while categorical variables were filled with the mode.

Encoding and Scaling

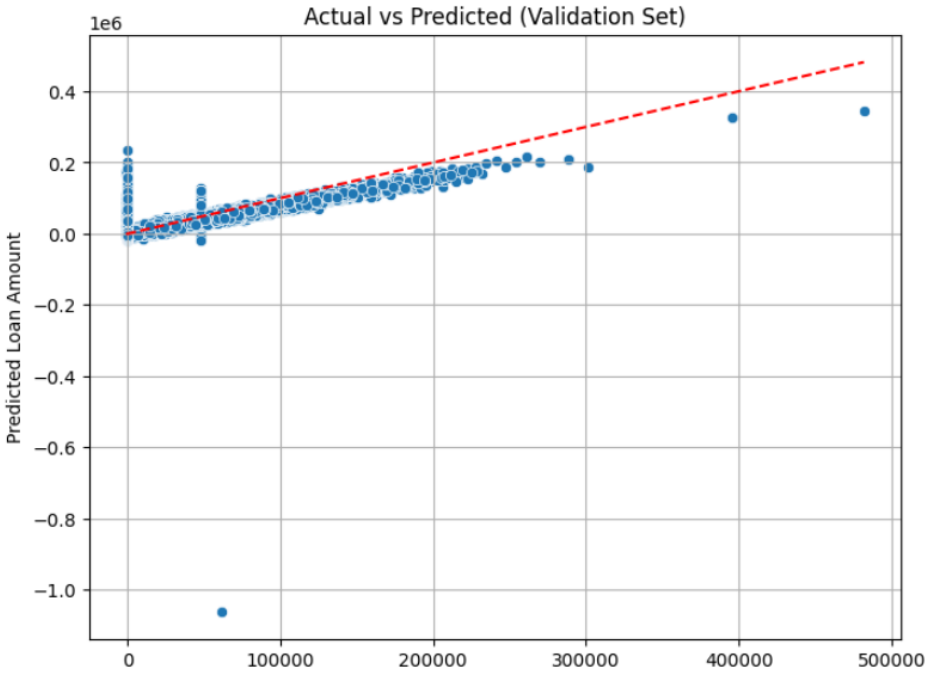
Categorical features were encoded using Label Encoding. Numerical features were standardized using `StandardScaler` to bring them to a common scale.

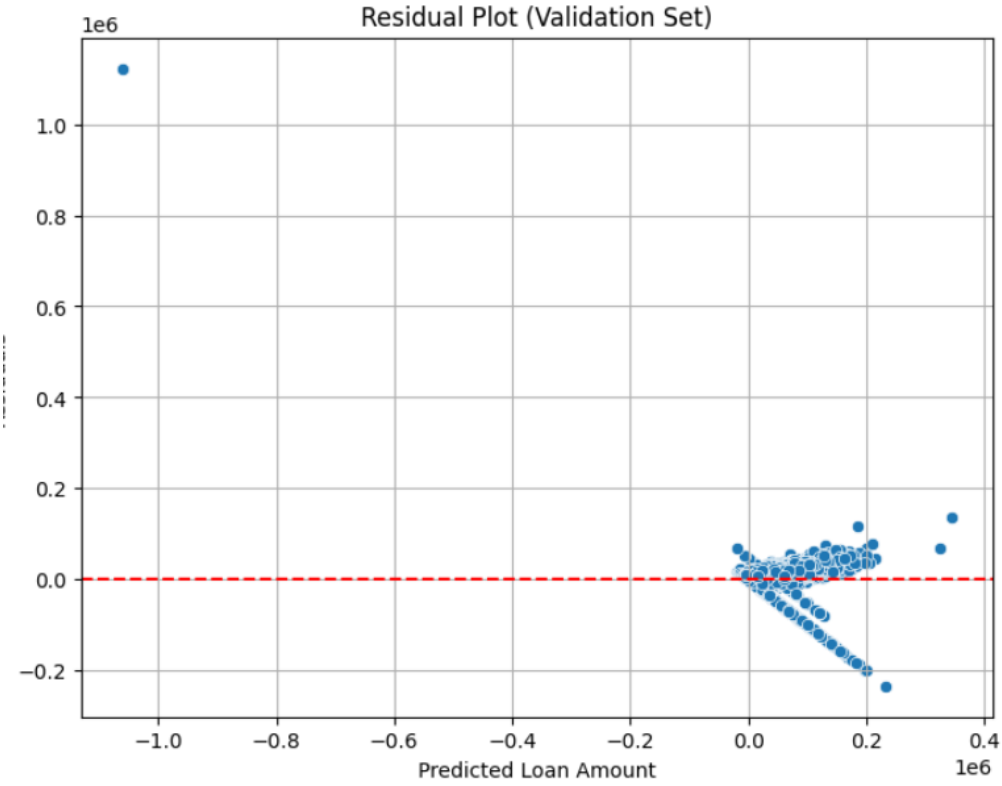
Model Training and Evaluation

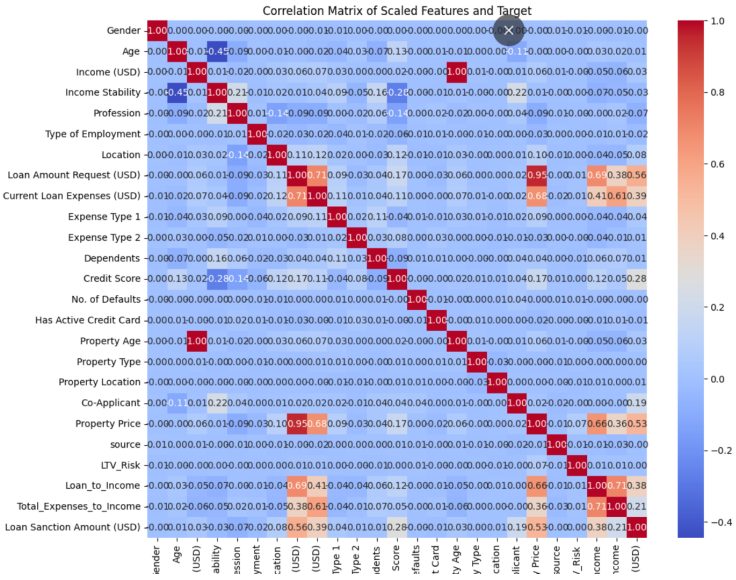
A Linear Regression model was trained. Evaluation included MAE, MSE, RMSE, R^2, and Adjusted R^2, along with Actual vs Predicted and Residual plots for visual interpretation.

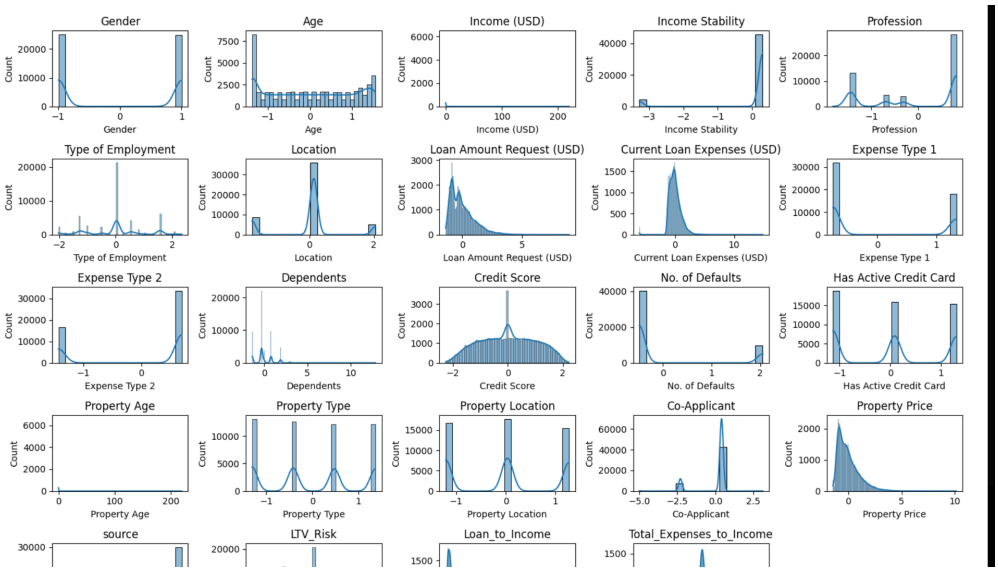
Code Implementation:

Screenshots of Output:




Residual Plot (Validation Set)
1e6
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
1e6
Predicted Loan Amount


Correlation Matrix of Scaled Features and Target
Gender
Age
Income (USD)
Income Stability
Profession
Type of Employment
Location
Loan Amount Request (USD)
Current Loan Expenses (USD)
Expense Type 1
Expense Type 2
Dependents
Credit Score
No. of Defaults
Has Active Credit Card
Property Age
Property Type
Property Location
Co-Applicant
Property Price
source
LTV_Risk
Loan_to_Income
Total_Expenses_to_Income
Loan Sanction Amount (USD)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



Result and Discussions:

Fold	MAE	MSE	RMSE	R^2 Score	Adjusted R^2 Score
Fold 1	21855.65	9.77×10^8	31254.23	0.5738	0.5722
Fold 2	21883.08	9.78×10^8	31275.97	0.5830	0.5814
Fold 3	21608.39	9.60×10^8	30987.54	0.5681	0.5664
Fold 4	21668.17	9.67×10^8	31094.08	0.5834	0.5818
Fold 5	21957.28	1.19×10^9	34491.73	0.4855	0.4835
Average	**21794.51**	1.01×10^9	**31820.71**	**0.5588**	**0.5571**

Table 1: Cross-validation metrics across 5 folds including Adjusted R^2 Score

Table 2: Summary of Results for Loan Amount Prediction

Description	**Student's Result**
Dataset Size (after preprocessing)	50000 rows × 25 columns
Train/Test Split Ratio	5-Fold Stratified Cross-Validation
Feature(s) Used for Prediction	All encoded and scaled features (excluding target)
Model Used	Linear Regression
Cross-Validation Used? (Yes/No)	Yes
If Yes, Number of Folds	5
Reference to CV Results Table	Table 1
Mean Absolute Error (MAE) on Test Set	21,794.51 USD
Mean Squared Error (MSE) on Test Set	1.01×10^9 USD^2
Root Mean Squared Error (RMSE) on Test Set	31,820.71 USD
R^2 Score on Test Set	0.5588
Adjusted R^2 Score on Test Set	0.5571
Most Influential Feature(s)	`Loan Amount Request (USD)`, `Income (USD)`, `Loan_to_Income`
Observations from Residual Plot	Random scatter around zero with some spread at higher predictions
Interpretation of Predicted vs Actual Plot	Positive correlation, with underestimation at very high loan amounts
Any Overfitting or Underfitting Observed?	Mild underfitting
If Yes, Brief Justification	$R^2 \sim 0.56$, residual variance at higher values, indicating underestimation of extreme loan values

Performance Analysis

The model's performance was assessed using MAE, MSE, RMSE, and both R^2 and Adjusted R^2 scores across five folds of cross-validation. Key findings:

- The average MAE of ∼21,795 USD indicates the typical prediction error magnitude.
- MSE averages to 1.01×10^9, penalizing larger errors more strongly.
- RMSE is around 31,821 USD, consistent with MAE but more sensitive to high outliers.
- The R^2 score improved to 0.56 (compared to earlier runs), meaning the model explains about 56% of variance in sanctioned loan amounts.
- Adjusted R^2 is nearly the same (0.557), confirming stable generalization without excessive predictors.

Learning Practices:

- Learned to apply Linear Regression for predicting continuous loan amounts.
- Gained practical experience with stratified cross-validation to ensure balanced evaluation.

- Practiced preprocessing: handling missing data, encoding categorical features, and scaling numerical ones.
- Learned to evaluate regression models using MAE, RMSE, and R^2.
- Interpreted residual and prediction plots to assess underfitting patterns.