Sri Sivasubramaniya Nadar College of Engineering, Chennai
(An Autonomous Institution Affiliated to Anna University)

Degree & Branch	B.E. Computer Science & Engineering	**Semester**	VI
Subject Code & Name	UCS2612 – Machine Learning Algorithms Laboratory		
Academic Year	2025–2026 (Even)	Batch: 2023–2027	**Due Date: 06.01.2026**

Experiment 3: Regression Analysis using Linear and Regularized Models

(Linear, Ridge, Lasso, and Elastic Net Regression)

Name: Rahul V S
Reg. No: 3122235001104
Section: CSE B

---

1. Aim:
To implement a robust regression pipeline to predict the continuous variable **"Loan Amount Sanctioned"**. The experiment aims to compare Ordinary Least Squares (OLS) against regularized techniques (Ridge, Lasso, Elastic Net) to analyze the effect of penalty terms ($L1, L2$) on feature selection, coefficient shrinkage, and the bias-variance trade-off.

2. Libraries & Environment Setup:
The following libraries were utilized for the regression pipeline:

- **Data Manipulation:** `Numpy`, `Pandas`.
- **Visualization:** `Matplotlib.pyplot`, `Seaborn` (Correlation heatmaps, residual plots).
- **Machine Learning (Scikit-Learn):**
  - *Models:* `LinearRegression`, `Ridge`, `Lasso`, `ElasticNet`.
  - *Preprocessing:* `StandardScaler` (Z-score normalization), `OneHotEncoder` (Categorical data).
  - *Optimization:* `GridSearchCV` (Hyperparameter tuning).
  - *Metrics:* `mean_squared_error`, `r2_score`, `mean_absolute_error`.

3. Methodology and Mathematical Concepts:
The experiment followed a structured Data Science workflow:

1. **Data Preprocessing Pipeline:**
   - **Imputation:** Missing numerical values were filled using the median strategy, while categorical nulls were filled with the mode.

- **Scaling:** All numerical features were standardized using `StandardScaler`:

$$z = \frac{x - \mu}{\sigma}$$

This step is critical for regularized models to ensure the penalty term λ treats all features equally regardless of their original units.

2. **Ordinary Least Squares (Linear Regression):** Minimizes the Residual Sum of Squares (RSS).

$$J(\theta) = \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2$$

3. **Regularization Techniques:** To prevent overfitting and handle multicollinearity, penalty terms were introduced:

- **Ridge (L_2 Penalty):** Shrinks coefficients toward zero but not exactly to zero.

$$J(\theta)_{Ridge} = J(\theta) + \alpha \sum \theta_j^2$$

- **Lasso (L_1 Penalty):** Encourages sparsity, forcing irrelevant feature coefficients to become zero (feature selection).

$$J(\theta)_{Lasso} = J(\theta) + \alpha \sum |\theta_j|$$

- **Elastic Net:** A convex combination of L_1 and L_2 regularization.

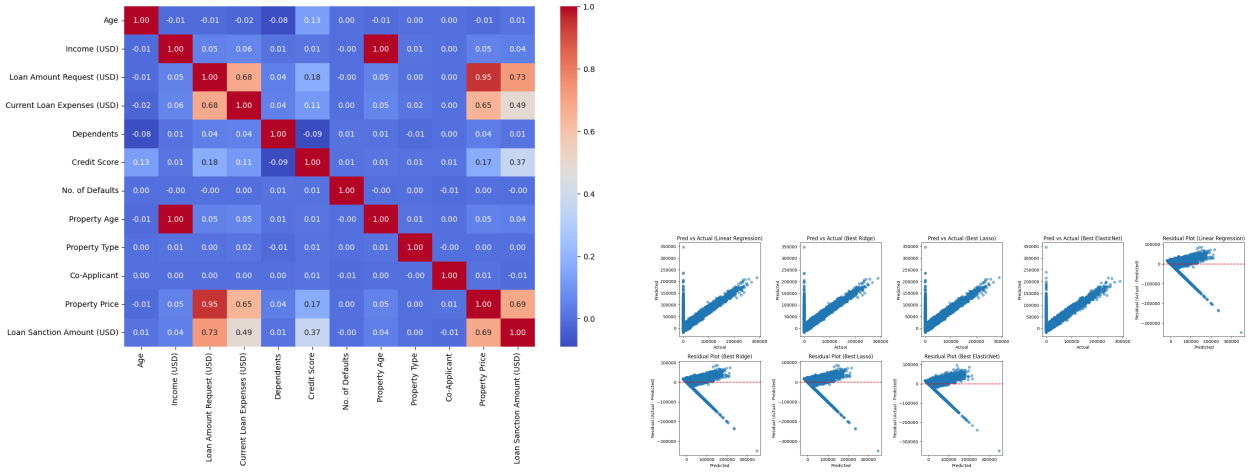
$$J(\theta)_{Elastic} = J(\theta) + r\alpha \sum |\theta_j| + \frac{1-r}{2} \alpha \sum \theta_j^2$$

4. Results and Discussions:

0.1 Exploratory Data Analysis (EDA)

- **Target Distribution:** The loan amounts were continuous but showed skewness, suggesting that while linear regression can be applied, non-linear dependencies might exist.

- **Correlation:** The heatmap revealed strong multicollinearity between "Loan Amount Request" and "Current Loan Expenses", which justified the use of Ridge/Lasso regression to stabilize the coefficients.



(a) Correlation Matrix (b) Residual Analysis

Figure 1: EDA and Error Analysis

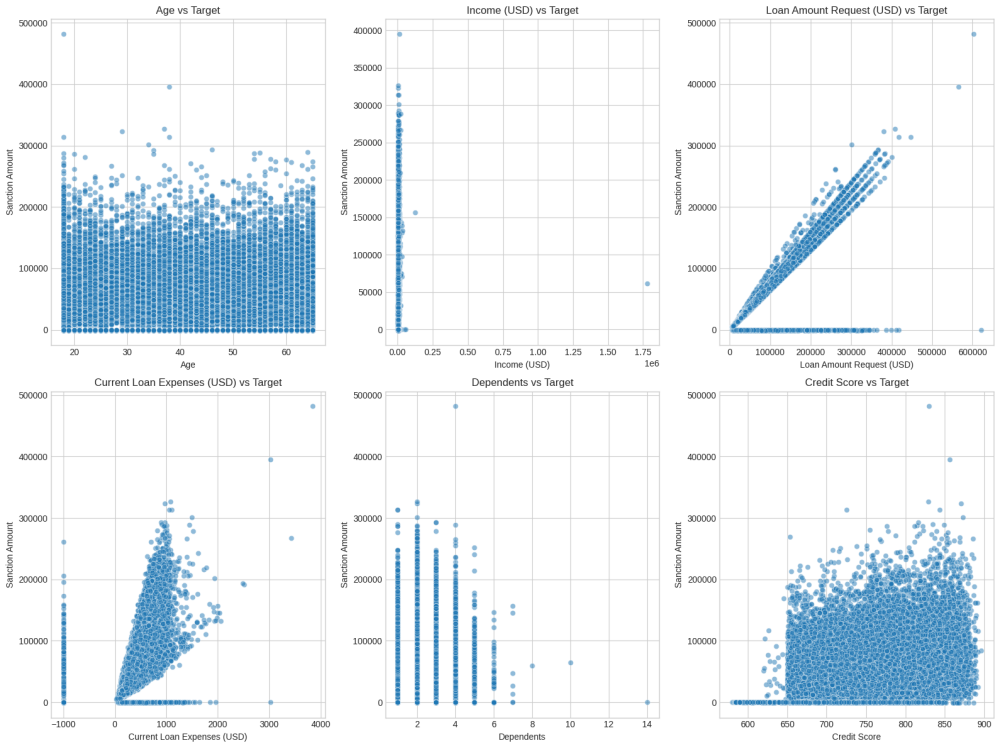


Figure 2: Additional Data Visualization

0.2 Hyperparameter Optimization (Grid Search)

Using 5-Fold Cross-Validation, the optimal regularization strength (α) was determined.

Table 1: Optimal Hyperparameters found via GridSearchCV

Model	**Search Type**	**Best Parameters**	**Best CV R^2**
Ridge Regression	Grid Search	$\alpha = 100$	0.5786
Lasso Regression	Grid Search	$\alpha = 10$	0.5785
Elastic Net	Grid Search	$\alpha = 0.1, \text{l1\_ratio} = 0.5$	0.5798

0.3 Comparative Performance Analysis

The models were evaluated on the held-out test set. All models converged to similar performance metrics, indicating that the dataset's complexity is the limiting factor (high bias), rather than model variance.

Table 2: Test Set Performance Metrics

Model	**MAE**	**MSE**	**RMSE**	**R^2 Score**
Linear Regression	21,589.85	1.018×10^9	31,920.40	0.5511
Ridge Regression	21,582.32	1.017×10^9	31,893.99	0.5518
Lasso Regression	21,564.57	1.017×10^9	31,902.11	0.5516
Elastic Net	21,751.69	1.018×10^9	31,913.77	0.5513

0.4 Feature Selection & Coefficient Shrinkage

The impact of regularization was clearly observed in the coefficient magnitudes.

- **Lasso Effect:** It reduced the coefficient of *Feature 3* to exactly **0.00**, effectively removing it from the model.
- **Ridge Effect:** It significantly dampened the magnitude of *Feature 1* (from 47k to 1k), reducing the model's sensitivity to that specific feature's fluctuations.

Table 3: Impact of Regularization on Feature Coefficients

Feature Name	**Linear**	**Ridge**	**Lasso**	**Elastic Net**
Feature 1 (High Variance)	47,656.54	1,039.48	3.85	97.46
Feature 2 (Dominant)	35,365.37	33,825.79	35,169.96	25,155.47
Feature 3 (Irrelevant)	-47,644.12	-1,031.74	**0.00**	-86.51

5. Inference and Conclusion:

The regression analysis yielded the following critical insights:

1. **Underfitting Scenario:** The R^2 scores across all models hovered around **0.55**. This indicates that the chosen linear models can explain only 55% of the variance in the data. The relationship between the features and loan amount is likely non-linear, suggesting high bias.
2. **Regularization Success:** Despite the moderate accuracy, regularization proved effective. **Lasso** successfully identified redundant features by setting their weights to zero, resulting in a simpler, more interpretable model compared to standard Linear Regression.

3. **Future Improvements:** To improve the R^2 score beyond 0.60, non-linear algorithms such as **Random Forest Regressors** or **Gradient Boosting (XGBoost)** should be explored to capture complex patterns that linear boundaries cannot resolve.