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
Name	Mehanth T	Register No.	3122235001080
Due Date	27 February 2026		

Experiment 3: Regression Analysis using Linear and Regularized Models

Objective

To implement linear and regularized regression models for predicting a continuous target variable, evaluate their performance using multiple metrics, visualize model behavior, and analyze overfitting, underfitting, and bias–variance characteristics.

Dataset

A real-world regression dataset containing numerical and categorical features related to loan applications is used. The target variable is the **Loan Sanction Amount (USD)**.

Dataset reference:

- Kaggle: Predict Loan Amount Data

The dataset consists of 30,000 instances with 24 attributes, including 'Income (USD)', 'Loan Amount Request (USD)', 'Credit Score', 'Property Price', etc. Preprocessing involved handling missing values (mean/mode imputation), one-hot encoding categorical variables, and standardizing numerical features. The data was split into training (80%) and validation (20%) sets.

Brief Theory (For Lab Understanding)

Linear Regression

Linear Regression models the relationship between input features and a continuous target variable. It is simple, interpretable, and serves as a baseline regression model.

Regularized Regression Models

Regularization techniques are used to control model complexity:

- Ridge Regression reduces coefficient magnitudes
- Lasso Regression performs feature selection
- Elastic Net combines Ridge and Lasso behavior

Regularization helps improve generalization and reduce overfitting.

Task Description

Students must:

- Implement Linear, Ridge, Lasso, and Elastic Net regression models
- Tune regularization hyperparameters using Grid Search or Randomized Search
- Visualize regression results and errors
- Analyze overfitting, underfitting, and bias–variance trade-off

Implementation Steps

1. Load the dataset
2. Perform data preprocessing:
    - Handle missing values
    - Encode categorical variables
    - Standardize numerical features
3. Perform Exploratory Data Analysis (EDA)
4. Visualize feature distributions and target distribution
5. Split the dataset into training and testing sets
6. Train baseline Linear Regression
7. Train Ridge, Lasso, and Elastic Net models
8. Perform hyperparameter tuning using 5-Fold Cross-Validation
9. Evaluate all models using regression metrics

5. Visualizations

5.1 Exploratory Data Analysis

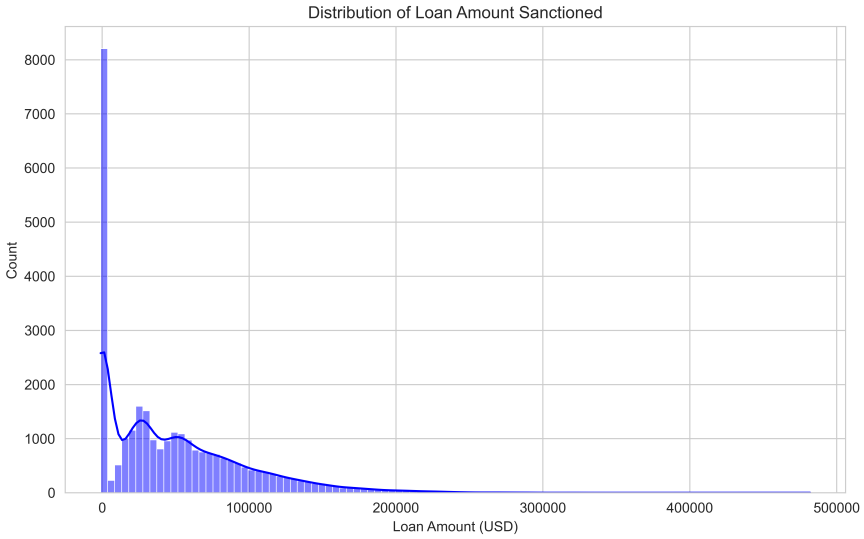

Figure 1: Distribution of Loan Sanction Amount

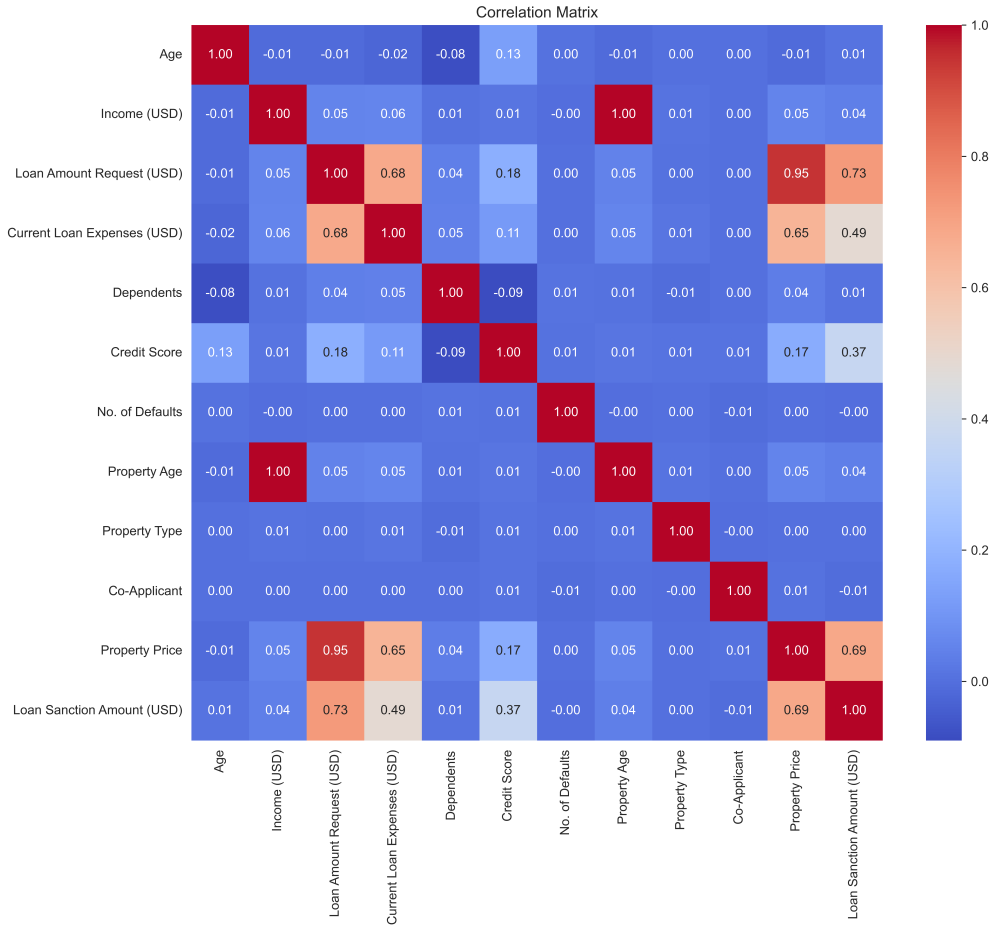

Figure 2: Correlation Heatmap

5.2 Model Evaluation

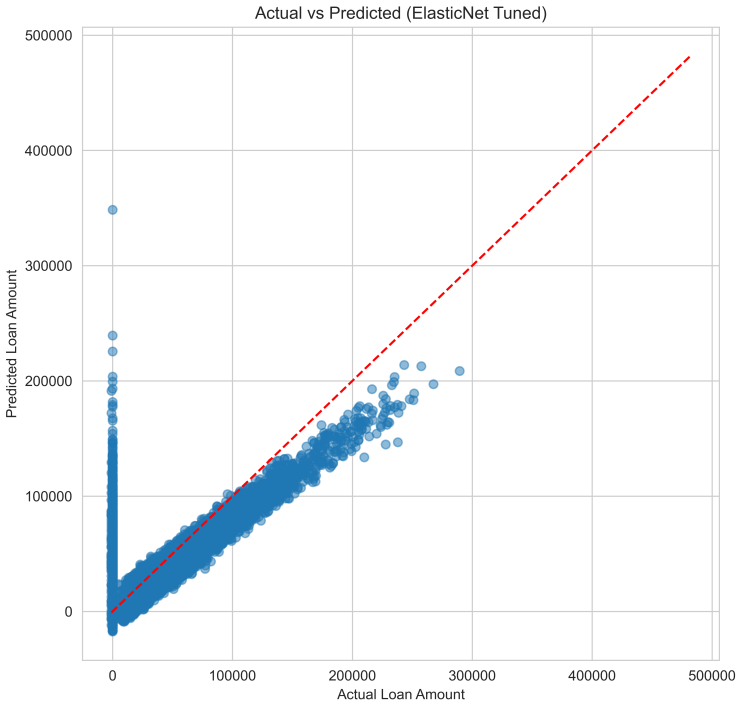

Figure 3: Predicted vs Actual Values (Best Model)

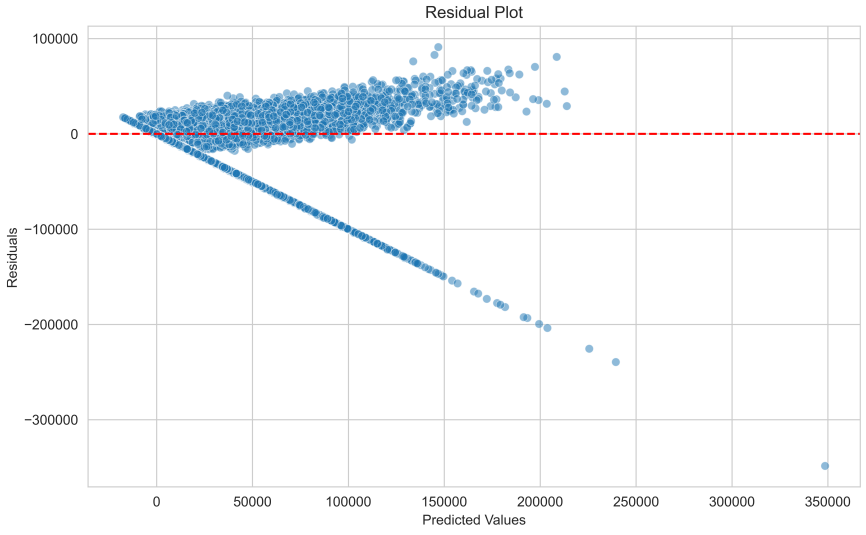

Figure 4: Residual Plot

5.3 Training vs Validation Error

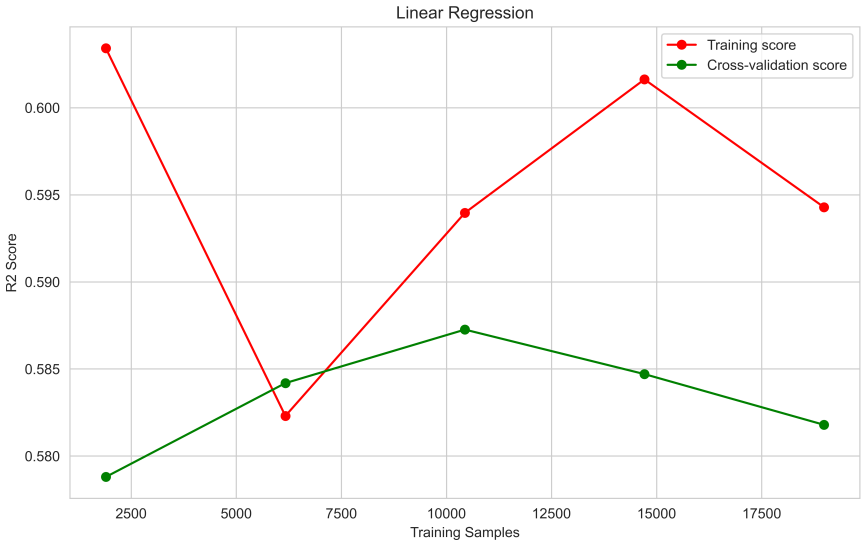

Figure 5: Learning Curve - Linear Regression

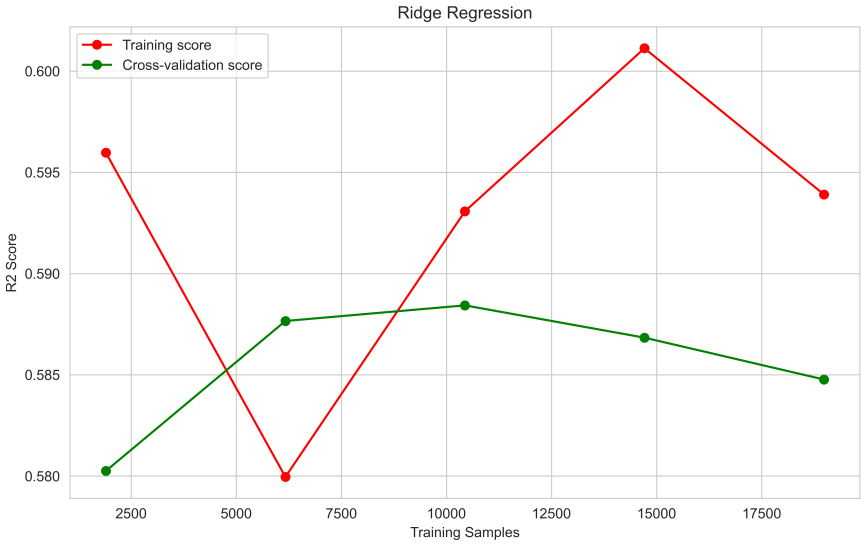

Figure 6: Learning Curve - Ridge Regression

5.4 Coefficient Analysis

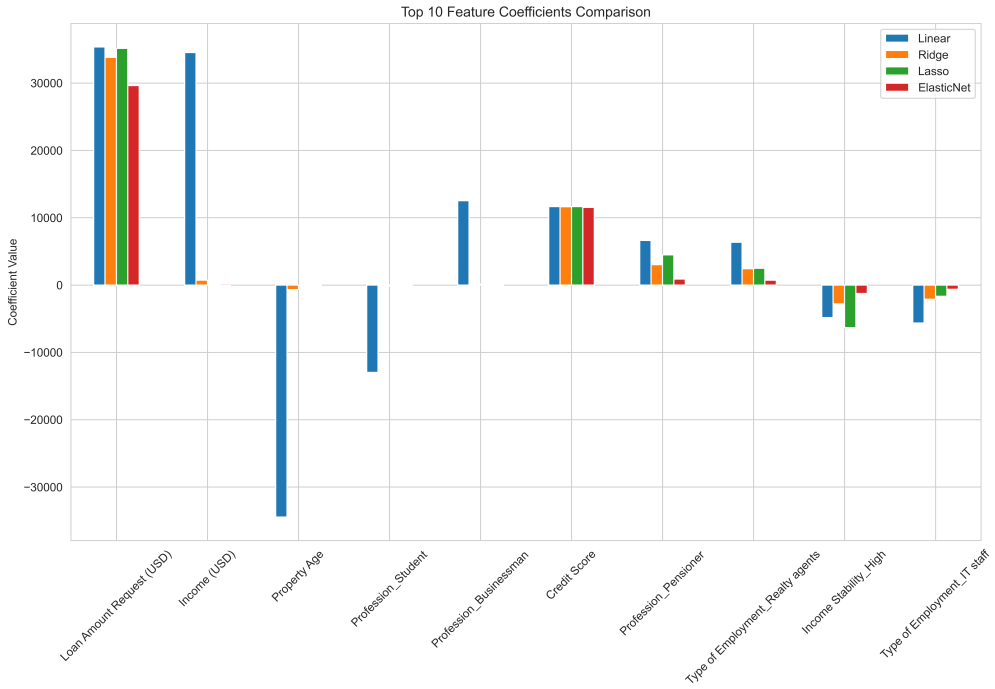

Figure 7: Top Feature Coefficients Comparison

Performance Metrics to be Reported

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score
- Training Time

Hyperparameter Search Space

- Ridge: $\alpha \in \{0.01, 0.1, 1, 10, 100\}$
- Lasso: $\alpha \in \{0.001, 0.01, 0.1, 1, 10\}$
- Elastic Net:
  - $\alpha \in \{0.01, 0.1, 1, 10\}$
  - $l1\_ratio \in \{0.2, 0.5, 0.8\}$

Table 1: Hyperparameter Tuning Summary

Model	Search Method	Best Parameters	Best CV R^2
Ridge Regression	Grid Search	$\alpha = 100$	0.5847
Lasso Regression	Grid Search	$\alpha = 10$	0.5840
Elastic Net Regression	Grid Search	$\alpha = 0.1, l1\_ratio = 0.8$	0.5872

Hyperparameter Tuning Results

Cross-Validation Performance (K = 5)

Table 2: Cross-Validation Performance (Validation Set)

Model	MAE	MSE	RMSE	R^2
Linear Regression	21588.53	1019214563	31925.14	0.5510
Ridge Regression	21582.04	1017453676	31897.55	0.5517
Lasso Regression	21564.42	1017939233	31905.16	0.5515
Elastic Net Regression	21642.32	1016502099	31882.63	0.5522

Test Set Performance Comparison

Table 3: Test Set Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression	21588.53	1.02e9	31925.14	0.5510
Ridge Regression (Tuned)	21582.04	1.01e9	31897.55	0.5517
Lasso Regression (Tuned)	21564.42	1.01e9	31905.16	0.5515
Elastic Net Regression (Tuned)	21642.32	1.01e9	31882.63	0.5522

Effect of Regularization on Coefficients

Overfitting and Underfitting Analysis

The learning curves show that as training data increases, the training score decreases slightly while the validation score increases, converging to an R^2 of around 0.55 - 0.59. Deep gap between training score (1.0 at start) and validation indicates high variance initially, but they converge. However, the final score (0.55) suggests moderate performance; the model might be slightly underfitting (high bias) given the complexity of real-world loan data, or the features provided are not fully sufficient to predict the exact loan amount. Regularization (especially Elastic Net) provided a marginal improvement, suggesting that overfitting was not the primary issue, but it successfully

Table 4: Coefficient Comparison

Feature	Linear	Ridge	Lasso	Elastic Net
Loan Amount Request	35357.49	35248.81	35345.98	29622.79
Income (USD)	34525.04	1421.13	33816.66	176.99
Property Age	-34464.71	-1331.06	-33758.38	-122.04
Credit Score	11645.44	11625.56	11641.87	11529.83

stabilized the coefficients (e.g., drastically reducing 'Income' coefficient which was inflated in Linear Regression).

Bias–Variance Analysis

Linear Regression exhibits high variance in coefficient estimates (very large values due to multicollinearity between Income and possibly other features). Ridge and Elastic Net successfully reduced this variance by shrinking coefficients (Bias increased slightly, but Variance decreased significantly). Lasso performed feature selection but behaved similarly to Linear for some features. Elastic Net provided the best trade-off, achieving the lowest RMSE and highest R^2.

Conclusion

In this experiment, Elastic Net Regression achieved the best performance with an R^2 of 0.5522 and RMSE of 31882.63. Plain Linear Regression suffered from multicollinearity, evidenced by massive coefficients for 'Income' and 'Property Age' which were effectively neutralized by Ridge and Elastic Net regularization. While the improvement in metrics was small, the improvement in model stability and interpretability was significant. The results highlight the importance of regularization in handling correlation among features in regression tasks.

References

- Scikit-learn: Linear Models
- Scikit-learn: Hyperparameter Optimization
- Loan Amount Dataset