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

<table>
<tr><td>Degree & Branch</td><td>B.E. Computer Science & Engineering</td><td>Semester VI</td></tr>
<tr><td>Subject Code & Name</td><td colspan="2">UCS2612 – Machine Learning Laboratory</td></tr>
<tr><td>Academic Year</td><td>2025–2026 (Even)</td><td>Batch 2023–2027</td></tr>
<tr><td>Name:</td><td>Nanda Kumar B</td><td>Roll No: 3122235001701</td></tr>
<tr><td>Due Date</td><td colspan="2">27.01.2026</td></tr>
</table>

Experiment 3: Regression Analysis using Linear and Regularized Models

1. Aim & 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.

2. Dataset

A real-world regression dataset containing numerical and categorical features related to loan applications is used. The target variable is the **loan amount sanctioned**.

Dataset reference:

- Kaggle: Predict Loan Amount Data

3. Preprocessing Steps

- The dataset was loaded into a Pandas DataFrame, and input features and target labels were separated for modeling using *train_test_split*().
- Missing values in numerical features were imputed using the median strategy with an added indicator for missingness, while categorical features were imputed with a constant value as *mean* for numerical data and *mode* for categorical data and encoded using LableEncoder.
- Numerical features were standardized using StandardScaler to ensure all features had zero mean and unit variance, allowing coefficients to be comparable across features.

4. Implementation Details

- Implemented Linear Regression, Ridge, Lasso, and Elastic Net models to predict the target variable.
- Evaluated model performance using cross-validation (CV R^2) and metrics including MAE, MSE, RMSE, and R^2 on validation and test sets.

- Visualized model results through Predicted vs Actual plots, residual plots, learning curves, and coefficient comparison charts.
- Performed hyperparameter tuning for Ridge, Lasso, and Elastic Net using GridSearchCV to identify optimal regularization parameters.

5. Visualizations

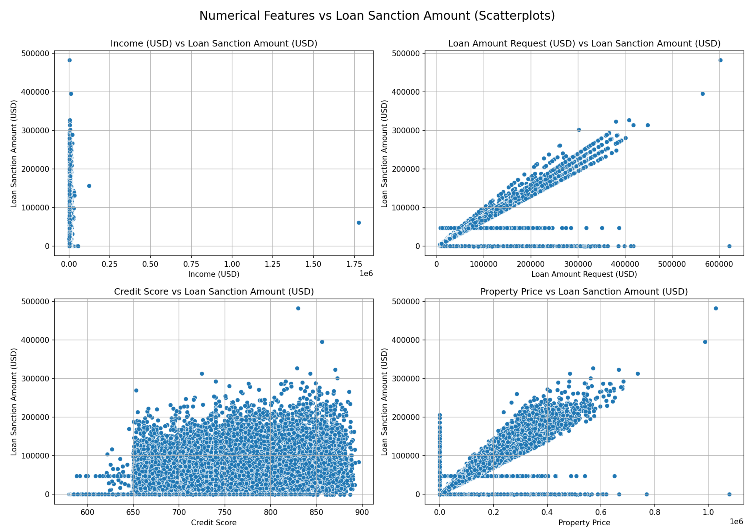

Figure 1: Scatter Plot

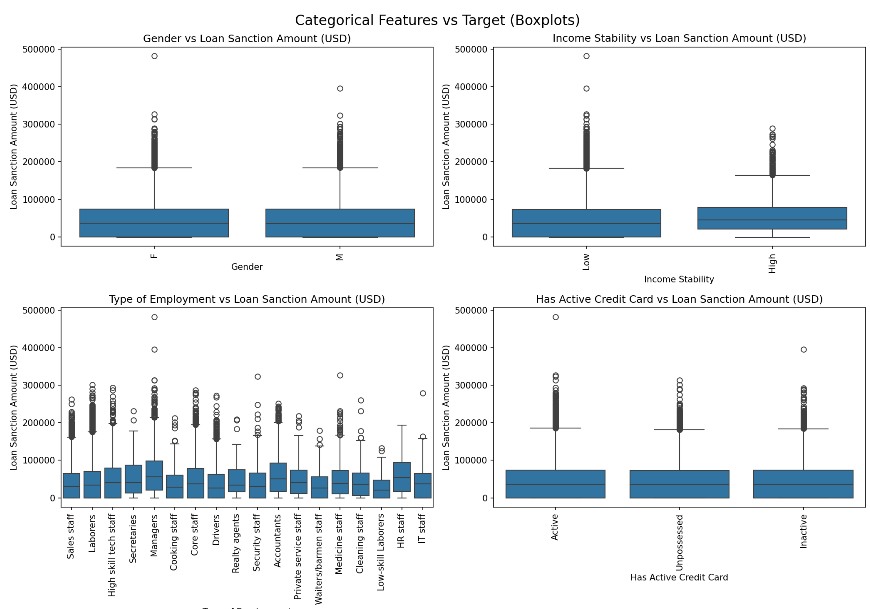

Figure 2: Box Plots for categorical features

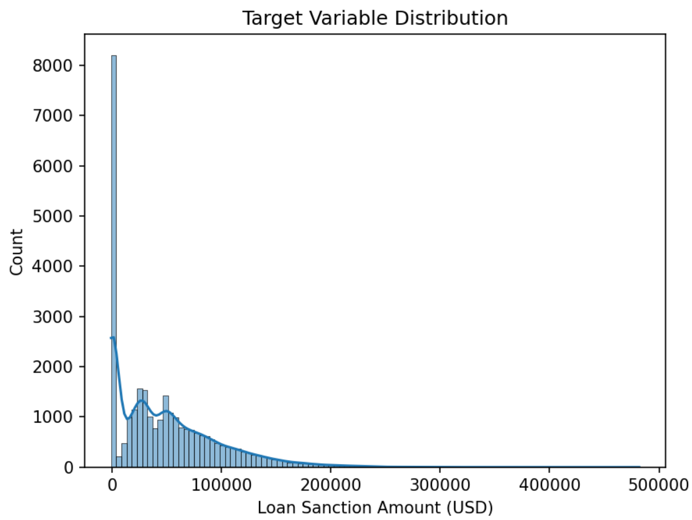

Figure 3: Target Variable distribution

5.1 Linear Regression Result

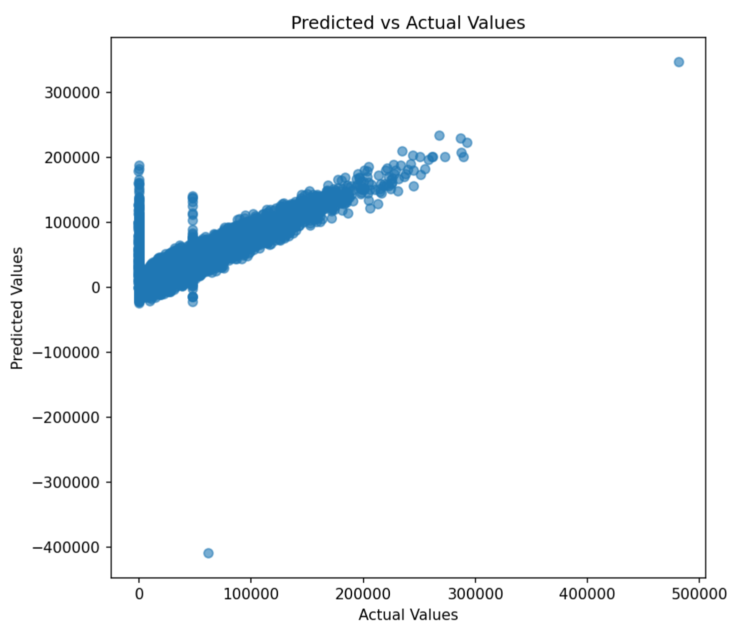

Figure 4: Linear Regression predicted vs actual

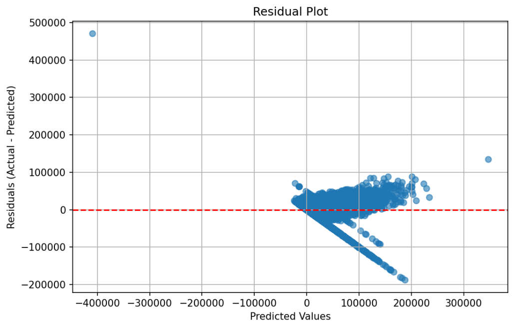

Figure 5: Residual Plot

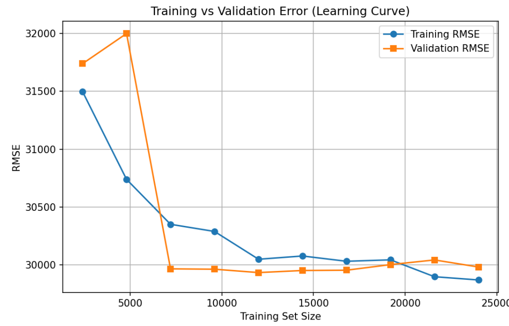

Figure 6: Learning Curve

5.2 Lasso Regression Result

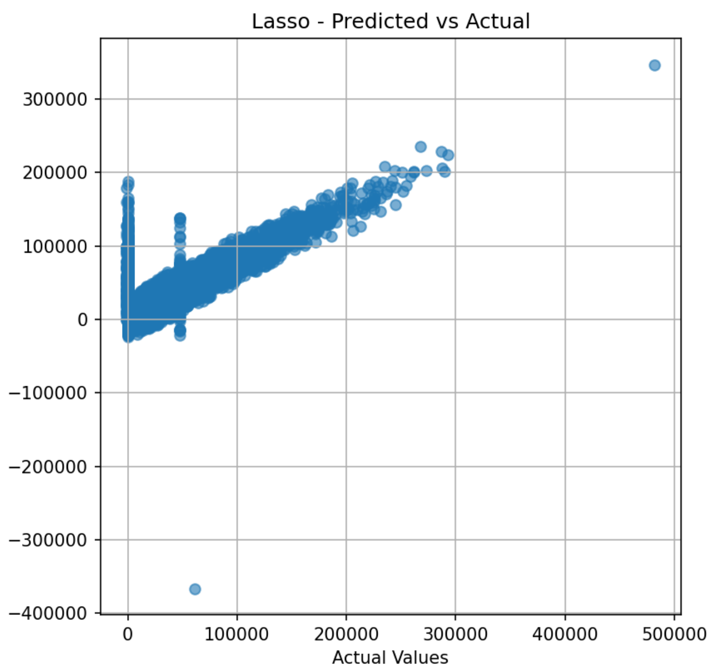

Figure 7: Lasso Regression predicted vs actual

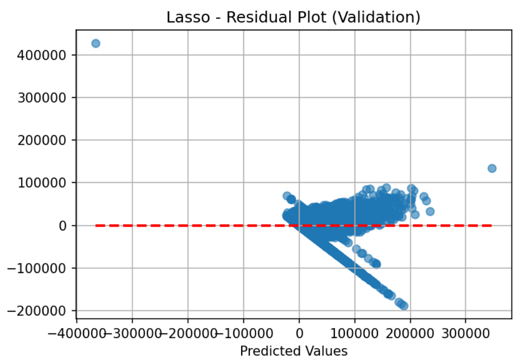

Figure 8: Lasso Residual

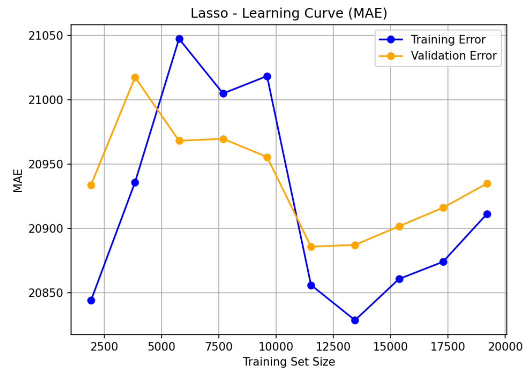

Figure 9: Lasso Learning curve

5.3 Ridge Regression Result

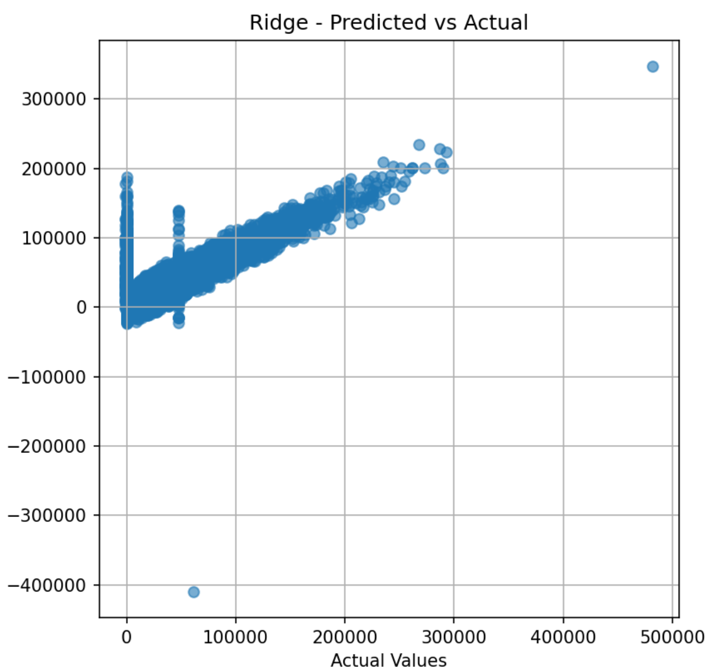

Figure 10: Ridge Regression predicted vs actual

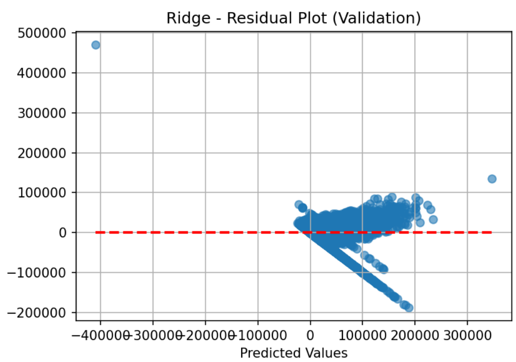

Figure 11: Ridge Residual

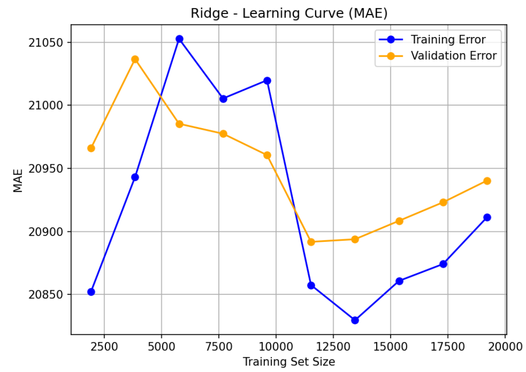

Figure 12: Ridge Learning curve

5.4 Elastic Net Regression Result

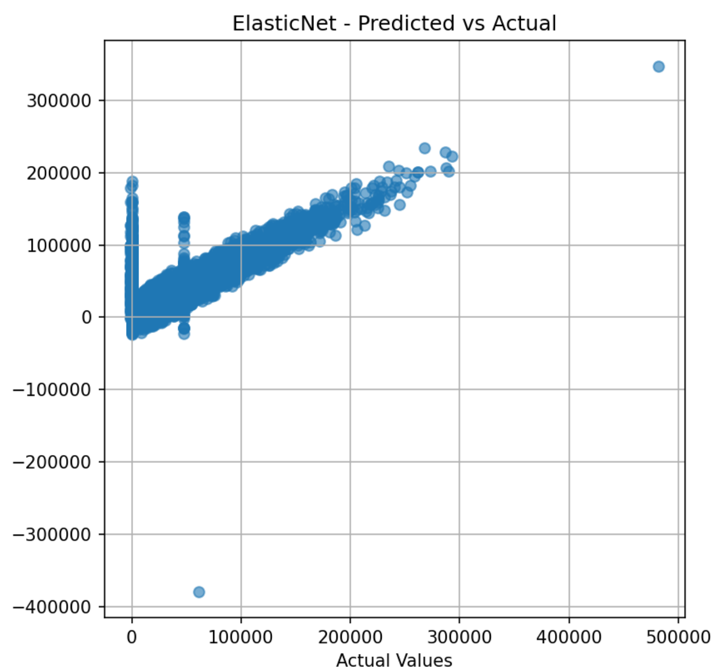

Figure 13: Elastic Net Regression predicted vs actual

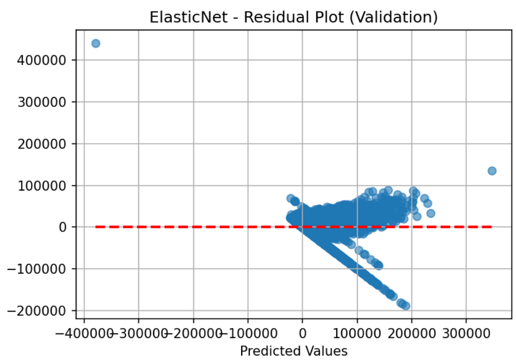

Figure 14: Elastic Net Regression Residual

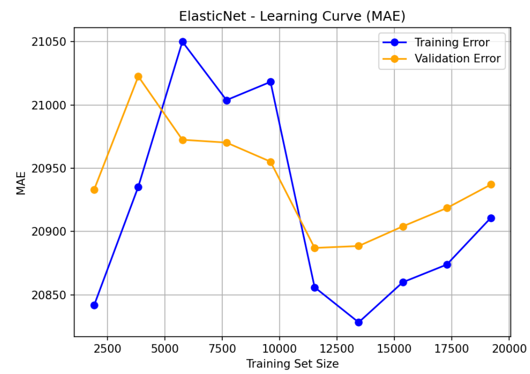

Figure 15: Elastic Net Learning curve

5.5 Coefficient comparison bar plot

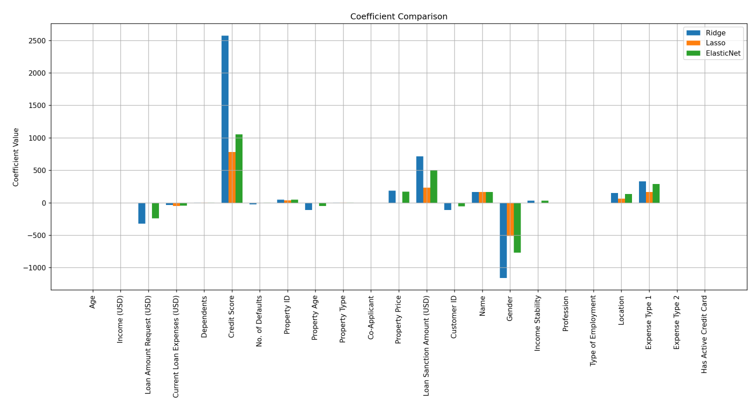

Figure 16: Coefficient comparison bar plot

6. Performance Table

6.1 Hyperparameter Tuning Results

Table 1: Hyperparameter Tuning Summary

Model	Search Method	Best Parameters	Best CV R^2
Ridge Regression	Grid / Random	α: 0.8	0.6080
Lasso Regression	Grid / Random	α: 100	0.6081
Elastic Net Regression	Grid / Random	α: 0.1 & $l1\_ratio$: 0.2	0.6081

6.2 Cross-Validation Performance (K = 5)

Table 2: Cross-Validation Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression	20722.68	906300898.71	30104.83	0.6057
Ridge Regression	20722.68	906295876.93	30104.75	0.6058
Lasso Regression	20716.18	899120095.27	29985.33	0.6089
Elastic Net Regression	20719.67	901275943.74	30021.26	0.6080

6.3 Test Set Performance Comparison

Table 3: Test Set Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression	20722.68	906300898.71	30104.83	0.6057
Ridge Regression	20722.68	906294622.44	30104.72	0.60
Lasso Regression	20722.51	906218854.32	30103.46	0.6058
Elastic Net Regression	20719.65	898050647.11	29967.49	0.6093

6.4 Effect of Regularization on Coefficients

Table 4: Coefficient Comparison

Feature	Linear	Ridge	Lasso	Elastic Net
Age	9398.73	9009.11	9203.68	8791.55
Loan Amount Request	7220.17	6435.43	6731.75	6057.56
Elastic Net Regression	5937.50	5632.21	7436.76	5362.59

7. Overfitting and Underfitting Analysis

- Difference between training and validation errors
  - A significantly lower training error compared to validation error indicates overfitting, meaning the model has memorized the training data but struggles on unseen data.
  - Similar training and validation errors suggest the model generalizes well and has achieved a good balance between bias and variance.
- Effect of regularization strength
  - Small regularization strength (weak regularization) allows the model to fit the training data closely, which may lead to overfitting.

  - Large regularization strength (strong regularization) restricts model complexity, reducing overfitting but potentially increasing bias and underfitting.

- Improvement in generalization after tuning
  - Proper tuning of regularization parameters helps achieve the optimal bias–variance trade-off, improving performance on unseen data.
  - Tuning can reduce validation error and n.

8. Bias–Variance Analysis

- Bias behavior of Linear Regression
  - Linear Regression has low bias when the underlying relationship is truly linear.
  - However, it may underfit if the data has complex nonlinear patterns, resulting in higher bias in such cases.
- Variance reduction using Ridge and Elastic Net
  - Ridge Regression reduces variance by penalizing large coefficients, making the model less sensitive to noise in the training data.
  - Elastic Net combines L1 and L2 penalties, reducing variance while also allowing some feature selection, balancing stability and flexibility.
- Feature sparsity effect in Lasso
  - Lasso (L1 regularization) drives some coefficients to exactly zero, creating a sparse model.
  - This sparsity improves interpretability and can help with feature selection, but excessive regularization may increase bias.

9. Conclusion

Linear Regression showed low bias but could overfit on correlated features. Ridge reduced variance through L2 regularization, Lasso introduced sparsity to simplify the model and aid feature selection, and Elastic Net balanced sparsity and variance reduction. Hyperparameters chosen via cross validation minimized validation error, achieving a good trade-off between accuracy and model complexity while improving generalization.

References

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