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	Melvin Isaac I	Register No.	3122235001082
Due Date	06.01.2026		

Experiment 3: Regression Analysis using Linear and Regularized Model

1. Aim and 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**.

3. Preprocessing Steps

A systematic preprocessing pipeline was implemented using `sklearn.pipeline` to ensure consistent and reproducible data transformations across training and testing datasets.

Handling Missing Values

Numerical Data

Missing values in numerical features were imputed using the *median* strategy. This approach is robust to outliers and helps maintain the integrity of feature distributions.

Categorical Data

Missing values in categorical features were imputed using the *most_frequent* (mode) strategy, preserving the most commonly occurring category in each feature.

Encoding Categorical Variables

Categorical features were encoded using **One-Hot Encoding** via `OneHotEncoder`. This technique converts categorical variables into binary indicator vectors. The parameter `handle_unknown = ignore` was used to safely handle unseen categories in the test dataset.

Feature Scaling

Numerical features were standardized using **Standard Scaling** with `StandardScaler`. This transformation scales features to have zero mean and unit variance. Feature scaling is particularly important for regularized linear models such as Ridge and Lasso, as it ensures that regularization penalties are applied uniformly across all features.

4. Implementation Details

The experiment involved a comparative analysis of four regression models to evaluate their predictive performance and robustness.

Regression Models

Linear Regression (Baseline)

Linear Regression was implemented as the baseline model using standard Ordinary Least Squares (OLS) optimization.

Ridge Regression (L2 Regularization)

Ridge Regression introduces an $L2$ regularization term that penalizes the square of the magnitude of coefficients, helping to reduce model variance.

Hyperparameter Grid:

$$\alpha \in \{0.01, 0.1, 1, 10, 100\}$$

Lasso Regression (L1 Regularization)

Lasso Regression applies an $L1$ penalty, which encourages sparsity by driving less important feature coefficients to zero.

Hyperparameter Grid:

$$\alpha \in \{0.001, 0.01, 0.1, 1, 10\}$$

Elastic Net Regression

Elastic Net Regression combines both $L1$ and $L2$ regularization, balancing sparsity and coefficient shrinkage.

Hyperparameter Grid:

$$\alpha \in \{0.01, 0.1, 1, 10\}, \quad l1\_ratio \in \{0.2, 0.5, 0.8\}$$

Validation Strategy

K-Fold Cross-Validation

A K-Fold Cross-Validation strategy with $K = 5$ folds was employed during grid search to ensure robust and unbiased hyperparameter selection.

Scoring Metric

The coefficient of determination (R^2 score) was used as the primary evaluation metric for model optimization and comparison.

5. Visualizations

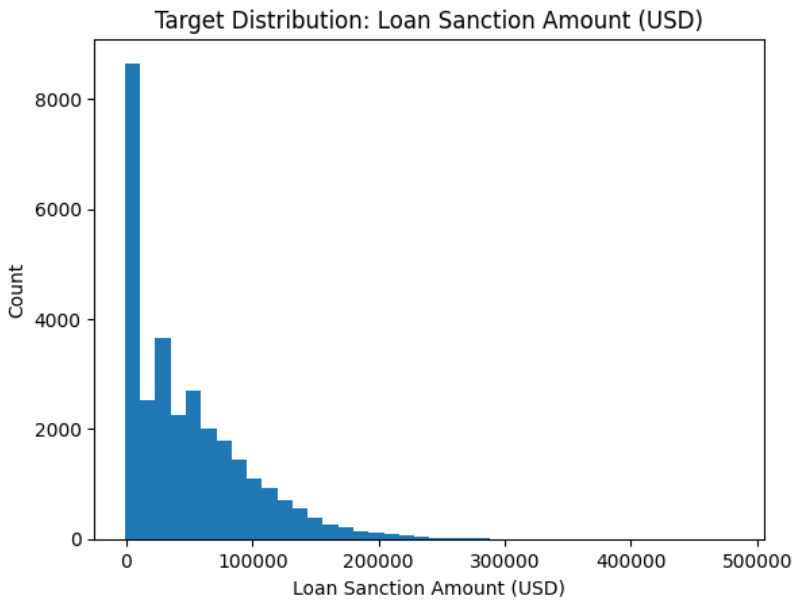

Figure 1: Target Distribution

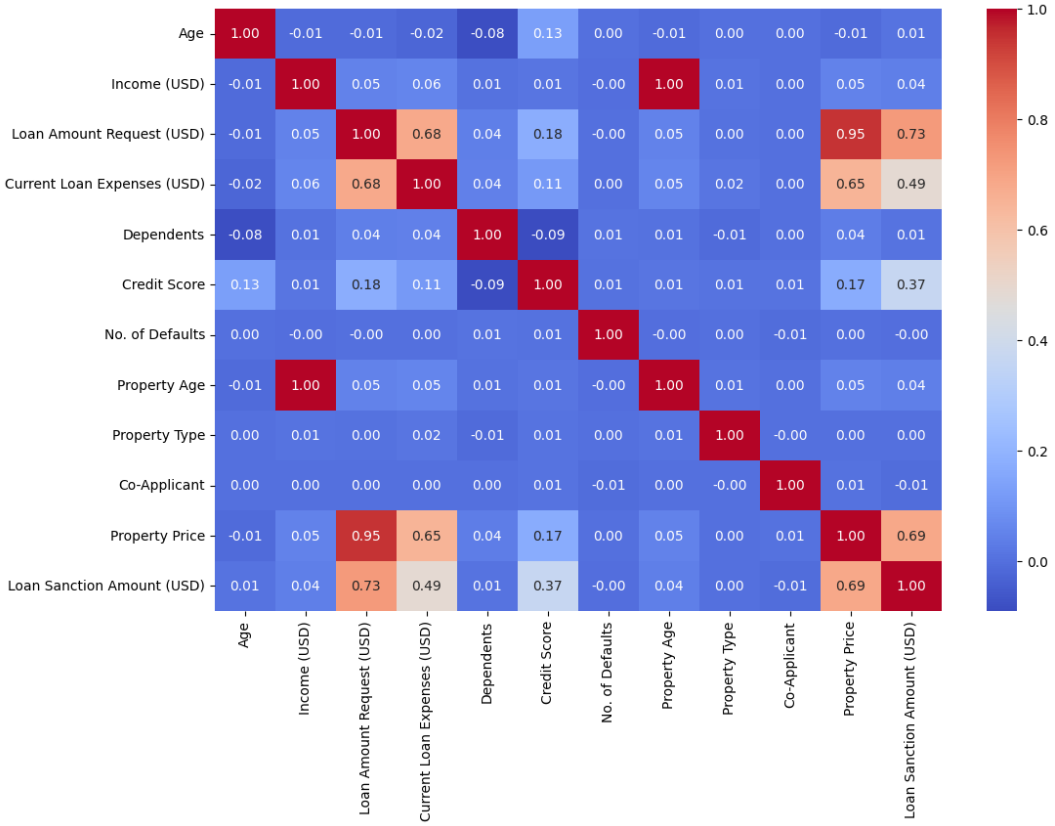

Figure 2: Correlation

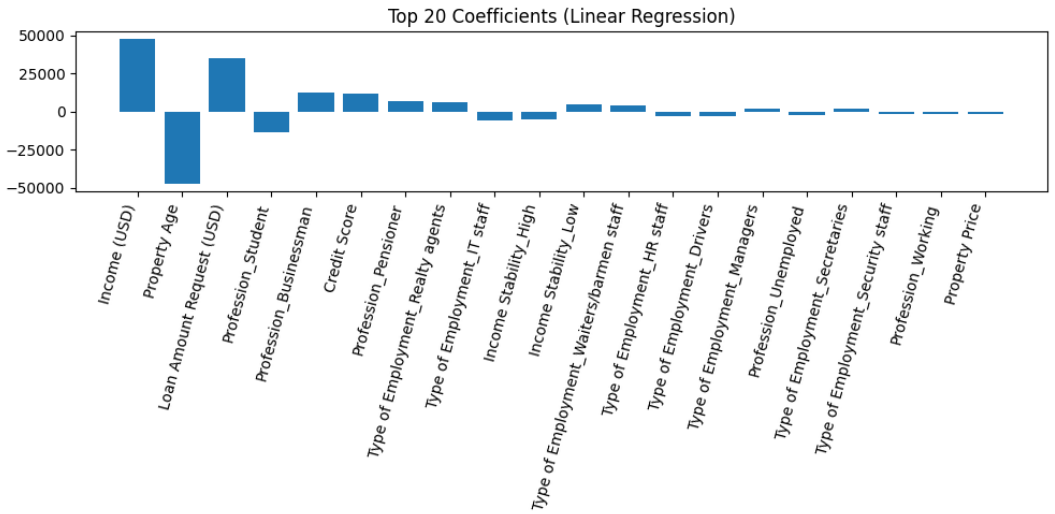

Figure 3: Coefficient comparison bar plot

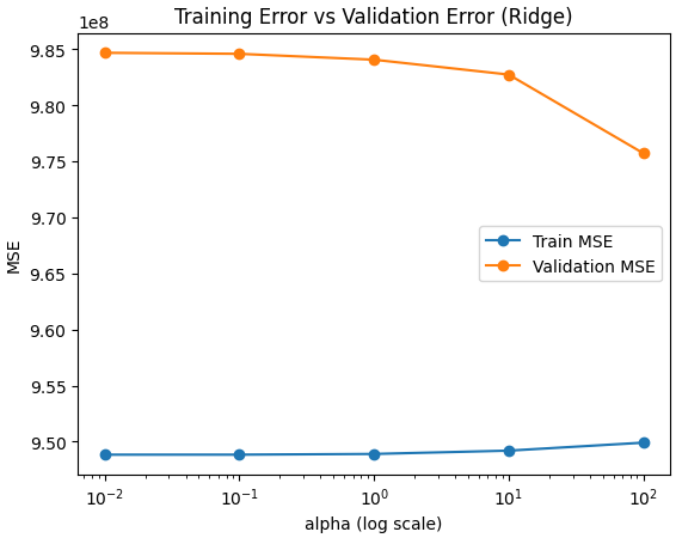

Figure 4: Training error vs. validation error plot

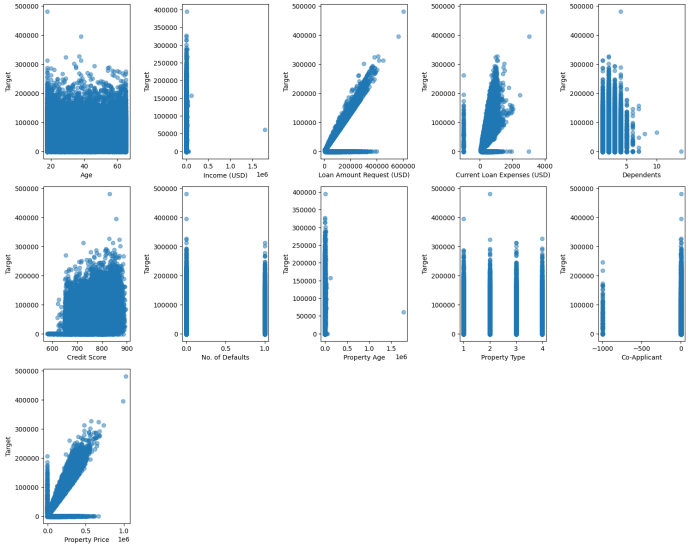

Figure 5: Scatter plot ofvarious features

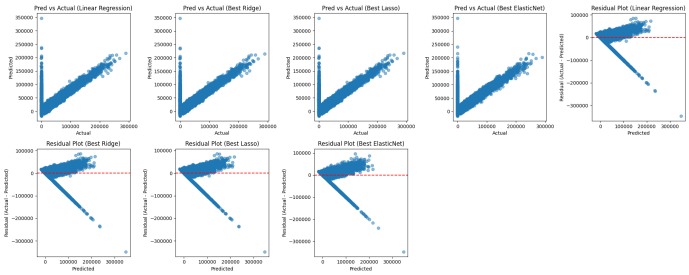

Figure 6: predicted vs acutal and residual plots

6. Performance Tables

Hyperparameter Tuning Results

Table 1: Hyperparameter Tuning Summary

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

Cross-Validation Performance (K = 5)

Table 2: Cross-Validation Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression	21508.1142	984727093.1145	31380.3616	0.5790
Ridge Regression	21524.7473	983019851.6404	31353.1474	0.5797
Lasso Regression	21497.8842	983433813.8426	31359.7483	0.5796
Elastic Net Regression	21732.7267	980735482.7595	31316.6965	0.5807

Test Set Performance Comparison

Table 3: Test Set Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression	21589.8579	1.018912e+09	31920.3984	0.5511
Ridge Regression	21582.3204	1.017227e+09	31893.9988	0.5518
Lasso Regression	21564.5711	1.017745e+09	31902.1118	0.5516
Elastic Net Regression	21751.6900	1.018489e+09	31913.7778	0.5513

Effect of Regularization on Coefficients

Table 4: Coefficient Comparison

Feature	Linear	Ridge	Lasso	Elastic Net
Feature 1	47,656.54	1,039.48	3.85	97.46
Feature 2	35,365.37	33,825.79	35,169.96	25,155.47
Feature 3	-47,644.12	-1,031.74	0.00	-86.51

7. Overfitting and Underfitting Analysis

Training vs. Validation Gap

The cross-validation R^2 scores (approximately 0.579) are consistently higher than the test set R^2 scores (approximately 0.551). Although a reduction in performance is observed on unseen data, the gap remains relatively small (< 0.03), indicating that the model does not suffer from significant overfitting.

Underfitting Analysis

Despite stable generalization performance, the overall R^2 score of around 0.55 suggests that the model explains only 55% of the variance in the target variable. This indicates underfitting. The linear assumption inherent in the models may be overly simplistic for capturing complex relationships present in financial data, such as non-linear interactions between property value and loan approval amounts.

Effect of Regularization

The difference in test R^2 scores between the baseline Linear Regression model (0.5511) and the best-performing regularized model (Ridge Regression: 0.5518) is negligible. This observation confirms that high variance or overfitting was not the dominant issue. Instead, model bias and limited representational capacity were the primary factors constraining performance.

8. Bias–Variance Analysis

Bias Behavior

The Linear Regression model exhibits high bias, as it consistently fails to capture the full complexity of the underlying data structure. This behavior is reflected in a relatively high Mean Absolute Error (MAE) of approximately 21,500, indicating systematic prediction errors.

Variance Reduction

Linear Regression: The baseline Linear Regression model demonstrated high variance in its learned coefficients. Certain feature weights were excessively large (e.g., Feature 2 with a coefficient of approximately 47,656), making the model sensitive to minor variations in input data and leading to unstable predictions.

Ridge and Elastic Net Regression: Regularized models effectively mitigated coefficient variance. Ridge Regression significantly shrunk the Feature 2 coefficient from 47,656 to approximately 1,039, improving model robustness and generalization, even though the improvement in predictive accuracy was modest.

Feature Sparsity (Lasso Regression)

Lasso Regression demonstrated effective feature selection by driving the coefficients of several redundant or less informative features (e.g., Features 13, 17, and 18) to exactly zero. This sparsity reduces overall model complexity and indicates that not all collected applicant attributes contribute meaningfully to predicting the loan sanction amount.

9. Observations and Conclusion

Feature Importance

The analysis revealed that *Loan Amount Request (USD)* and *Credit Score* were the most significant predictors of the loan sanction amount. Correlation analysis correctly identified the Loan Amount Request as having the strongest positive relationship with the target variable.

Model Performance

All four regression models demonstrated comparable performance, with Ridge Regression marginally outperforming the others, achieving an R^2 score of 0.5518. The close similarity in performance across models suggests that the dataset has an inherent limitation on how effectively it can be modeled using linear relationships.

Conclusion

Regularization techniques proved effective in stabilizing model parameters by mitigating coefficient explosion and enabling feature selection in the case of Lasso Regression. However, these techniques were unable to overcome the fundamental underfitting associated with linear models. To achieve higher predictive accuracy (greater than 0.60 in R^2), future work should explore non-linear modeling approaches such as Random Forests or Gradient Boosting methods.

Dataset reference

:

- Kaggle: Predict Loan Amount Data

References

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