Name: Keshav K S
Reg. No: 3122235001067
Class: CSE- B

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	**23.1.2026**		

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 amount sanctioned**.

Dataset reference:

- Kaggle: Predict Loan Amount Data

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 Parameters 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 Parameter tuning using 5-Fold Cross-Validation
9. Evaluate all models using regression metrics

Required Visualizations

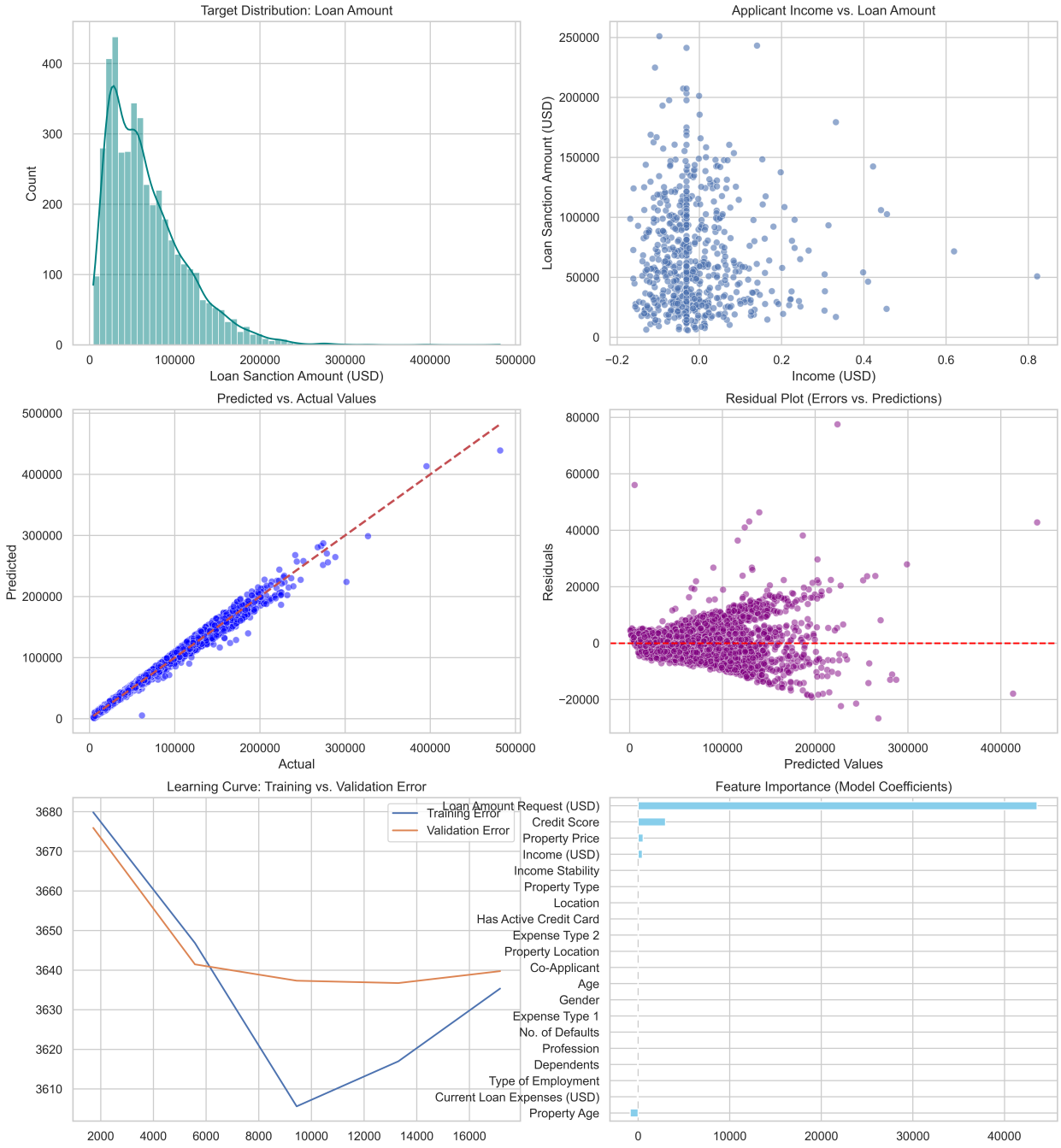

Figure 1: Comprehensive Regression Analysis: Target Distribution, Feature Correlations, Residual Analysis, Learning Curves, and Model Coefficients.

Performance Metrics to be Reported

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

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

Parameter Tuning Results

Table 1: Parameter Tuning Summary

Model	Search Method	Best Parameters	Best CV R^2
Ridge Regression	Grid	alpha: 1	0.9852
Lasso Regression	Grid	alpha: 10	0.9856
Elastic Net Regression	Grid	alpha: 0.1 l1ratio: 0.8	0.9834

Cross-Validation Performance (K = 5)

Table 2: Cross-Validation Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression	3609.00058	2.764272e+07	5257.634035	0.9861091
Ridge Regression	3608.588185	2.764016e+07	5257.391122	0.9861102
Lasso Regression	3605.700848	2.762616e+07	5256.059731	0.9861173
Elastic Net Regression	3825.689632	3.179339e+07	5638.563133	0.984025

Table 3: Test Set Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression	3960.60	34931250.26	5910.27	0.9834
Ridge Regression	3736.13	31082233.67	5575.14	0.9852
Lasso Regression	3719.83	30323919.93	5506.72	0.9856
Elastic Net Regression	3960.60	34931250.26	5910.27	0.9834

Test Set Performance Comparison

Effect of Regularization on Coefficients

Table 4: Coefficient Comparison

Feature	Linear	Ridge	Lasso	Elastic Net
Gender	17.580328	-13.749662	-3.332303	17.580328
Age	2.577235	-8.053297	-0.000000	2.577235
Income (USD)	14.711576	467.419995	-0.000000	14.711576
Income Stability	113.774970	96.178853	83.734410	113.774970
Profession	-62.116131	-49.089305	-34.995102	-62.116131
Type of Employment	-95.298883	-77.458892	-67.525299	-95.298883
Location	64.721397	72.108777	59.081410	64.721397
Loan Amount Request (USD)	37114.264323	43555.327303	43567.138220	37114.264323
Current Loan Expenses (USD)	724.552055	-105.200195	-87.614799	724.552055
Expense Type 1	1.649791	-40.258424	-35.701416	1.649791
Expense Type 2	11.295738	31.770559	19.856662	11.295738
Dependents	-58.948498	-62.172947	-51.819072	-58.948498
Credit Score	3030.738136	2989.115500	2980.104096	3030.738136
No. of Defaults	-54.441014	-42.373181	-32.993159	-54.441014
Has Active Credit Card	81.524001	46.118142	36.565360	81.524001
Property Age	0.559413	-858.754419	-0.000000	0.559413
Property Type	96.813176	92.930411	82.030963	96.813176
Property Location	27.391578	1.153002	-0.000000	27.391578
Co-Applicant	-14.548111	-6.096593	-0.000000	-14.548111
Property Price	5968.459789	545.564251	502.370322	5968.459789

Overfitting and Underfitting Analysis

- **Difference between training and validation errors:** Training error measures how well the model fits the data it was trained on, while validation error evaluates performance on unseen data. A very low training error with a high validation error indicates *overfitting*, whereas high errors on both training and validation sets indicate *underfitting*.
- **Effect of regularization strength:** Increasing the regularization parameter (λ or α) in Ridge, Lasso, and Elastic Net penalizes large coefficients and reduces model complexity. Low regularization behaves similarly to Linear Regression and may overfit, while very high regularization can oversimplify the model, leading to underfitting. An optimal intermediate value is required.
- **Improvement in generalization after tuning:** After tuning the regularization strength using validation data, the gap between training and validation error decreased. This indicates improved generalization.

Bias–Variance Analysis

- **Bias behavior of Linear Regression:** Linear Regression assumes a strictly linear relationship between features and the target variable. This simplicity results in relatively high bias, when the true relationship is more complex.
- **Variance reduction using Ridge and Elastic Net:** Ridge Regression reduces variance by preventing any single feature from dominating the model. Elastic Net combines both Ridge and Lasso penalties giving the advantages of the both.
- **Feature sparsity effect in Lasso:** Lasso Regression introduces sparsity by driving some coefficients exactly to zero. This effectively performs feature selection.

Conclusion

Using this lab experiment, Linear Regression is used as a baseline model. Ridge reduces the error of variation and lasso for feature selection. Elastic Net gives the best balance by combining the strengths of Ridge and Lasso.

References

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