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:	Piranow c	Roll No:	3122235001096
Due Date	27-01-2026		

Experiment 3: Regression Analysis using Linear and Regularized Models

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 Description

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

- Dataset loaded into a Pandas DataFrame and features separated.
- Numerical missing values imputed using median with missing indicators.
- Categorical features imputed using constant value ("Missing") and encoded using OneHotEncoder.
- Numerical features standardized using StandardScaler.
- Train–validation–test split performed to avoid data leakage.

4. Implementation Details

- Implemented Linear Regression, Ridge, Lasso, and Elastic Net models.
- Hyperparameter tuning using GridSearchCV.
- Performance evaluated using CV R^2, MAE, MSE, RMSE, and R^2.
- Visualized predicted vs actual, residuals, learning curves, and coefficients.

5. Visualizations

5.1 Exploratory Data Analysis (EDA)

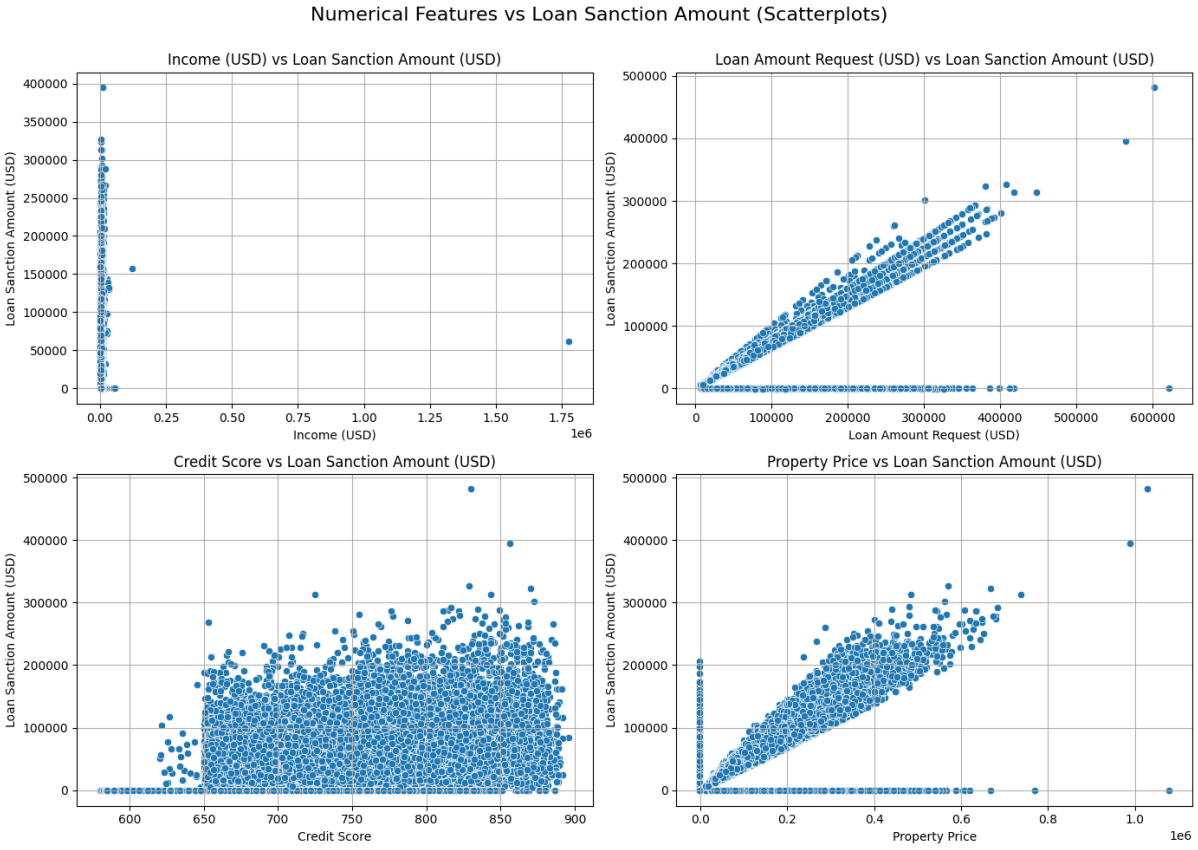

Figure 1: Numerical Features vs Loan Sanction Amount (Scatterplots)

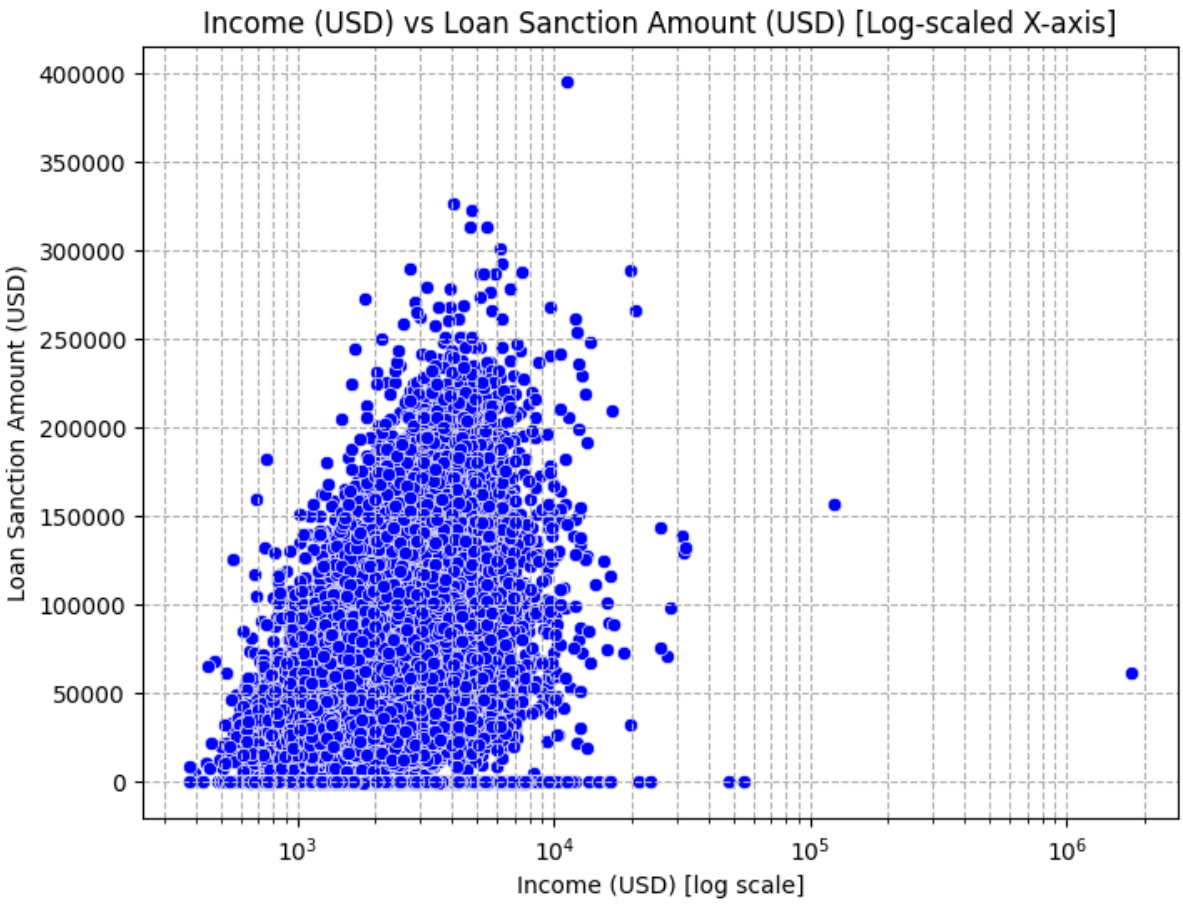

Figure 2: Income (USD) vs Loan Sanction Amount (USD) [Log-scaled X-axis]

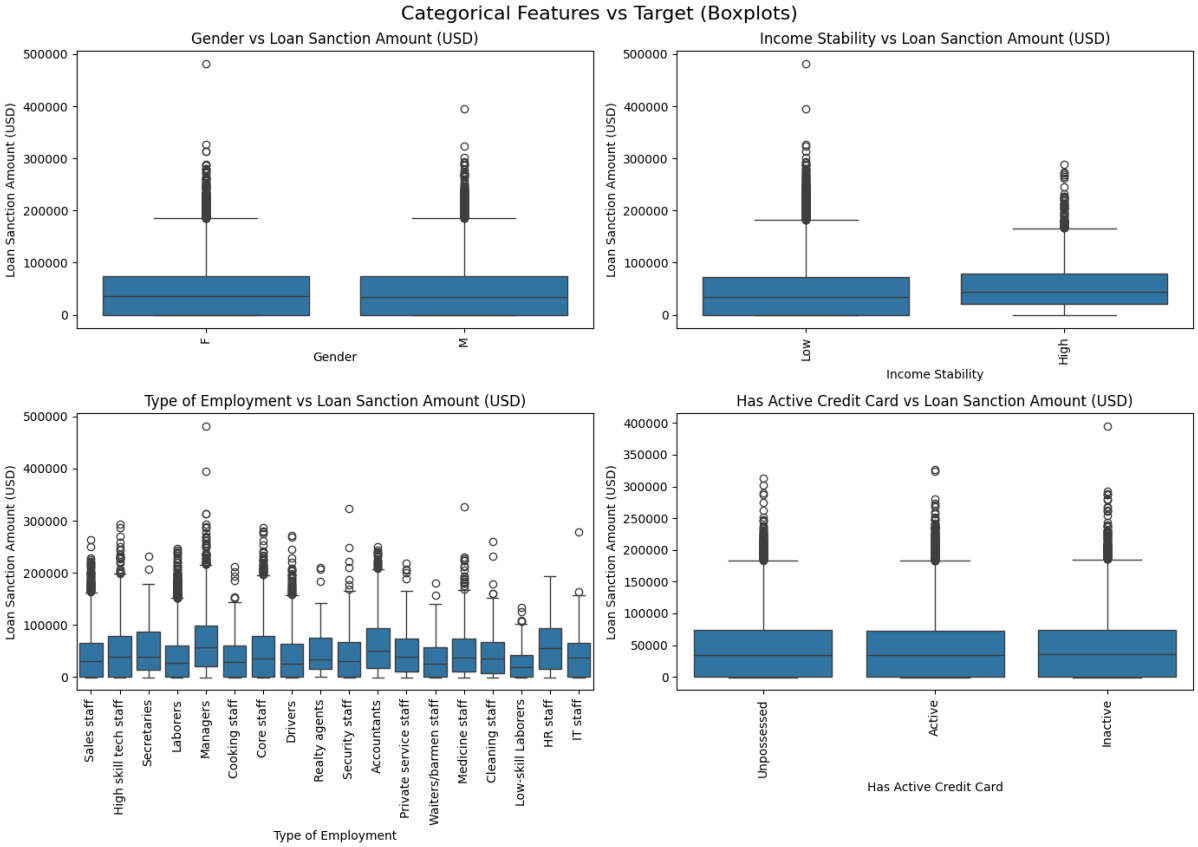

Figure 3: Categorical Features vs Loan Sanction Amount (Boxplots)

5.2 Linear Regression Results

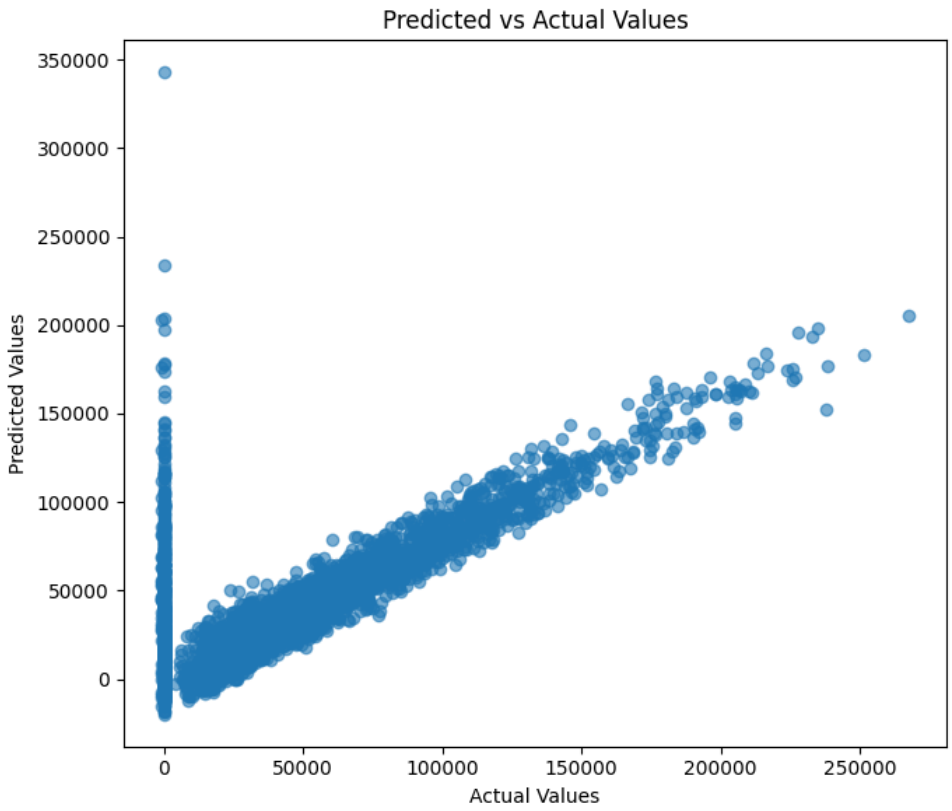

Figure 4: Predicted vs Actual Values (Linear Regression)

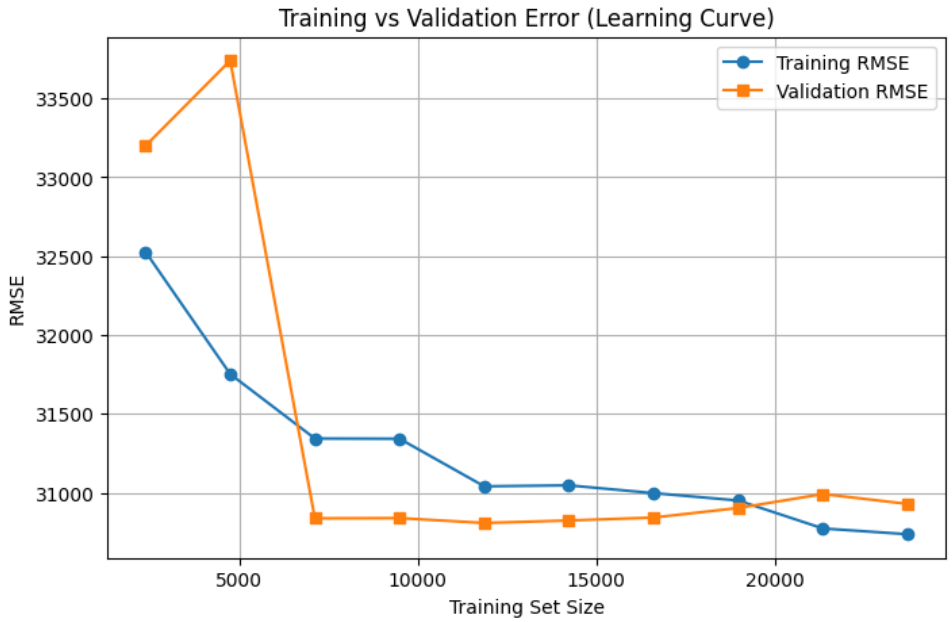

Figure 5: Training vs Validation Error (Learning Curve)

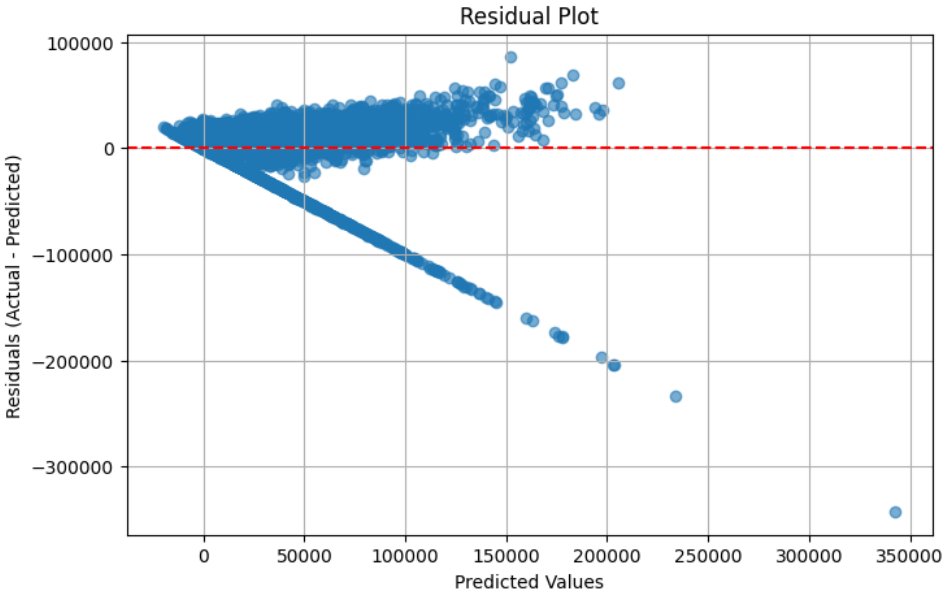

Figure 6: Residual Plot (Linear Regression)

5.3 Ridge Regression Results

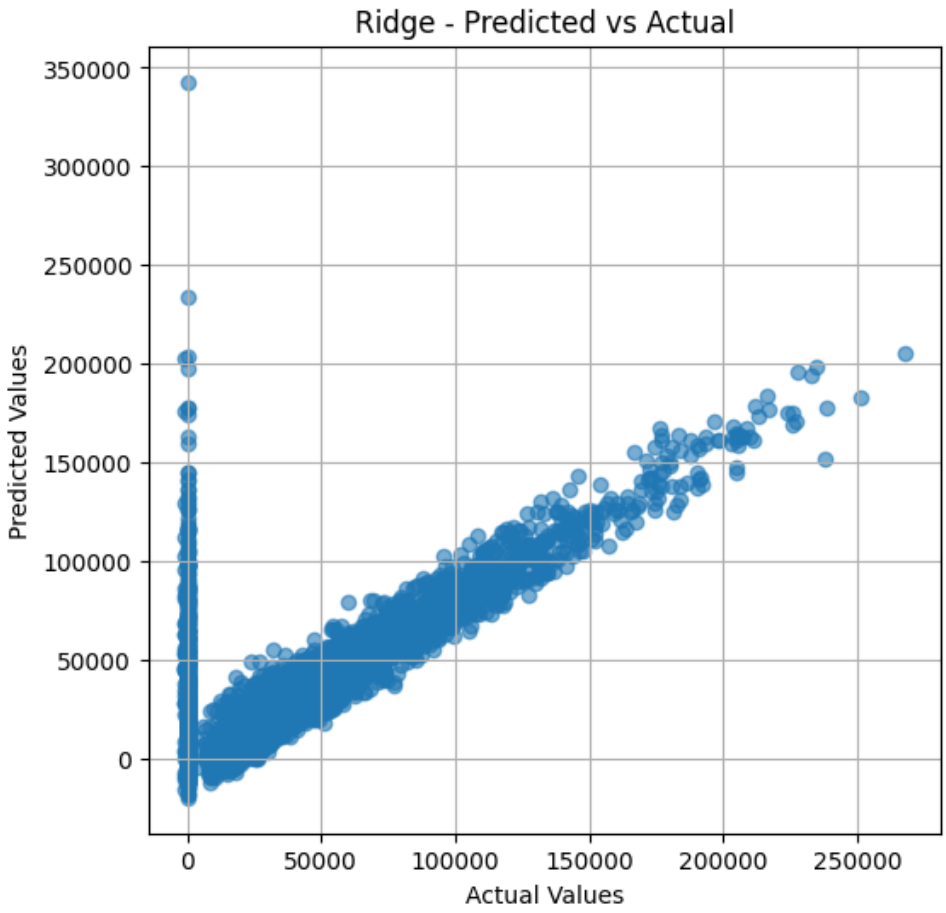

Figure 7: Predicted vs Actual Values (Ridge Regression)

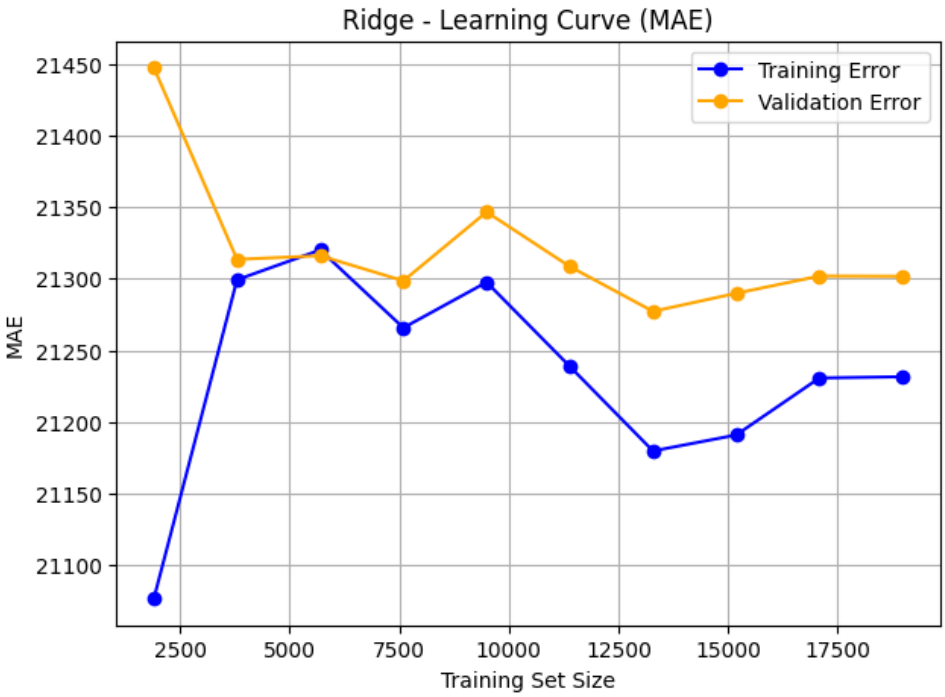

Figure 8: Ridge Regression Learning Curve (MAE)

5.4 Lasso Regression Results

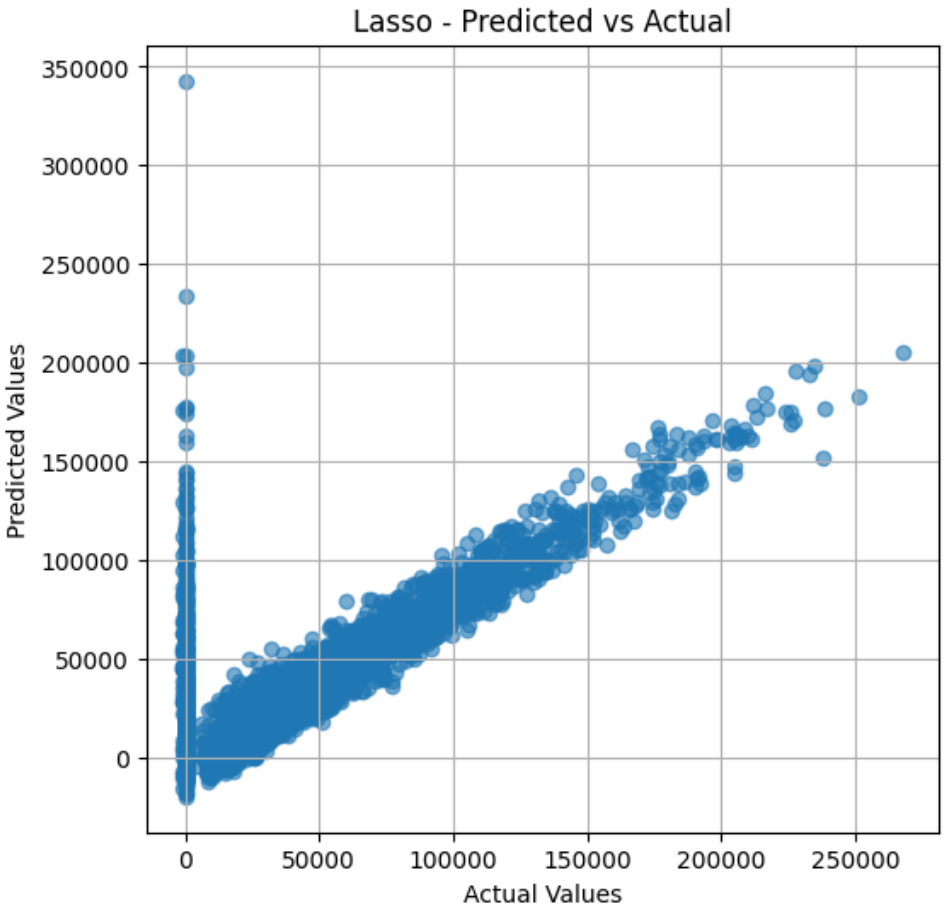

Figure 9: Predicted vs Actual Values (Lasso Regression)

5.5 ElasticNet Regression Results

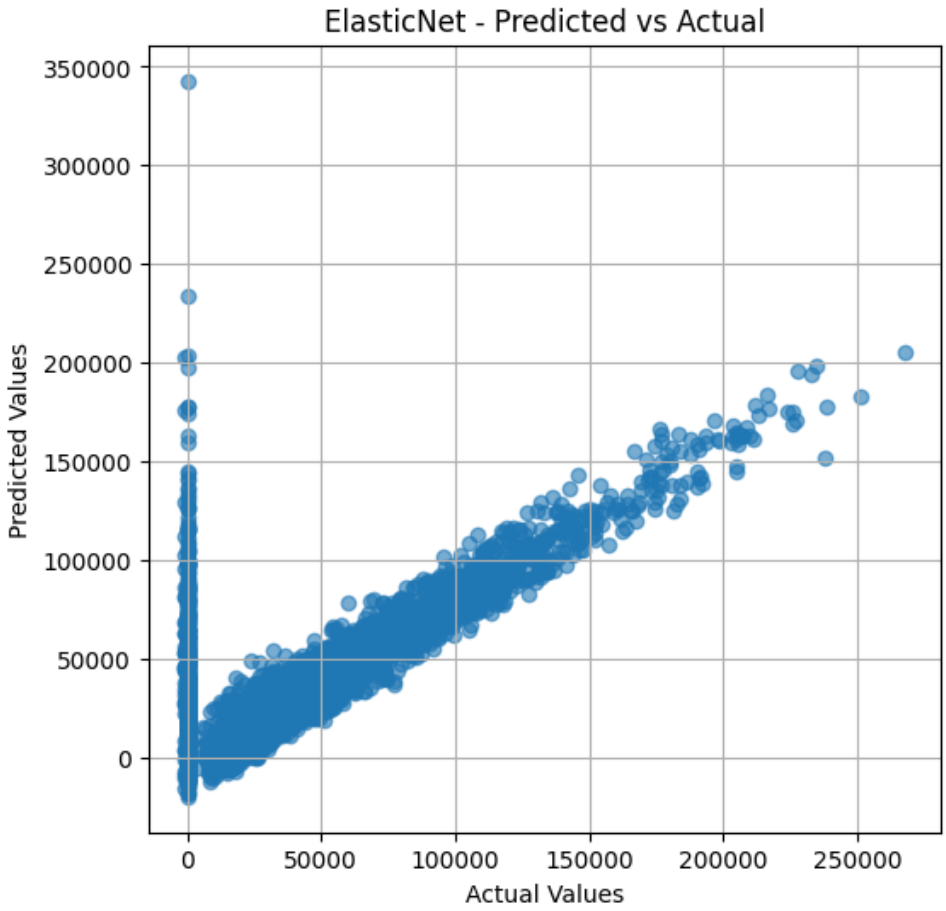

Figure 10: Predicted vs Actual Values (ElasticNet Regression)

5. Visualizations (Continued)

5.6 Lasso Regression Learning Curve

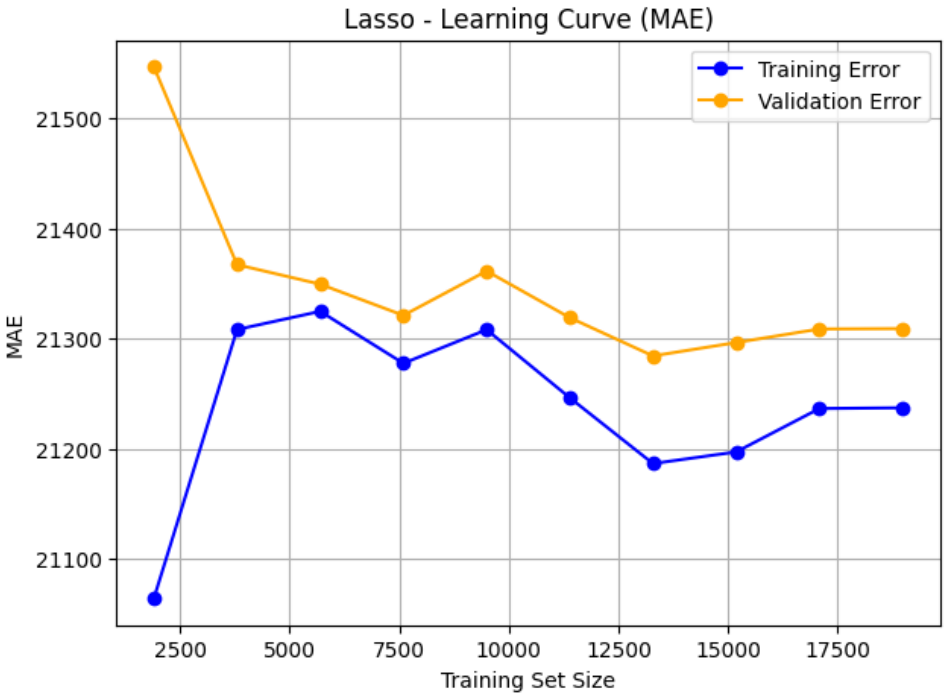

Figure 11: Lasso Regression Learning Curve (MAE)

5.7 ElasticNet Regression Learning Curve

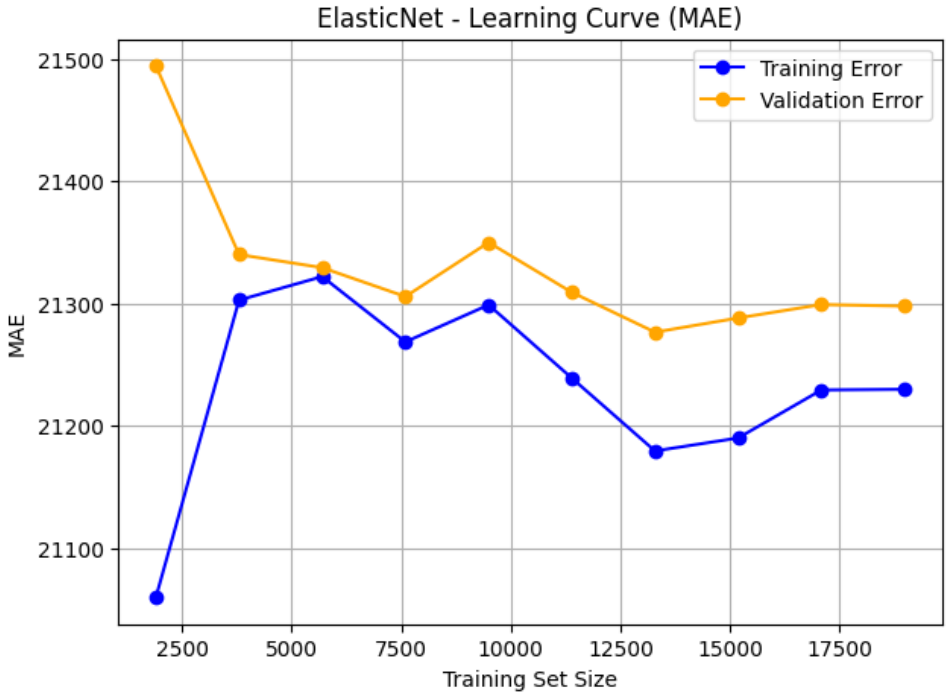

Figure 12: ElasticNet Regression Learning Curve (MAE)

5.8 Coefficient Comparison Across Models

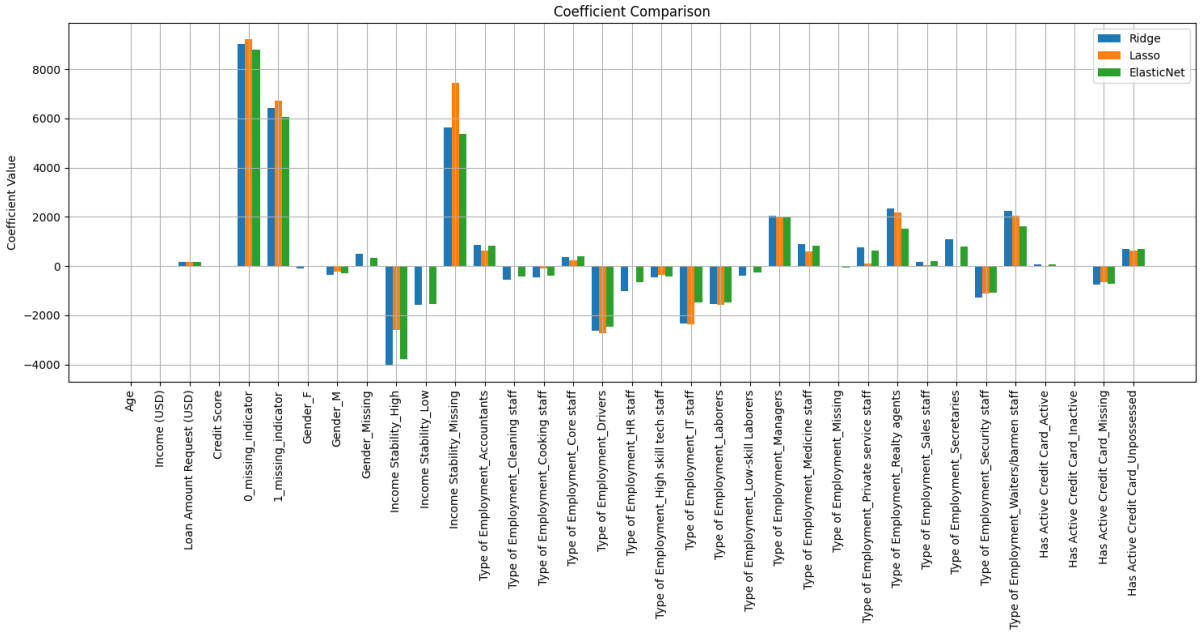

Figure 13: Coefficient Comparison for Ridge, Lasso, and ElasticNet Regression Models

6. Performance Tables

6.1 Hyperparameter Tuning Results

Model	Search Method	Best Parameters	Best CV R^2
Ridge Regression	Grid / Random	$\alpha = 100$	0.5889
Lasso Regression	Grid / Random	$\alpha = 10$	0.5891
Elastic Net	Grid / Random	$\alpha = 0.01$, $l1 = 0.2$	0.5890

Table 1: Hyperparameter Tuning Summary

6.2 Cross-Validation Performance (K = 5)

Model	MAE	MSE	RMSE	R^2
Linear Regression	21103.11	1002429332.22	31661.16	0.5445
Ridge Regression	21086.35	1001589383.17	31647.90	0.5449
Lasso Regression	21089.94	1001445007.21	31645.62	0.5449
Elastic Net	21080.35	1001466603.48	31645.96	0.5449

Table 2: Cross-Validation Performance

6.3 Test Set Performance Comparison

Model	MAE	MSE	RMSE	R^2
Linear Regression	21418.73	988649077.30	31442.79	0.5772
Ridge Regression	21424.25	988873190.77	31446.35	0.5771
Lasso Regression	21426.30	988395979.97	31438.77	0.5773
Elastic Net	21427.69	989141953.56	31450.63	0.5770

Table 3: Test Set Performance

6.4 Effect of Regularization on Coefficients

Feature	Linear	Ridge	Lasso	Elastic Net
Age	9398.7376	9009.1165	9203.6845	8791.5549
Loan Amount Request	7220.1758	6435.4392	6731.7544	6057.5620
Income Stability	5937.5006	5632.2180	7436.7697	5362.5975

Table 4: Coefficient Comparison

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 narrow the gap between training and validation errors, indicating better generalization.

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