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 Laboratory		
Academic year	2025-2026 (Even)	Batch:2023-2027	**Due date: 27-01-2026**

Experiment #3: Regression Analysis using Linear and Regularized Models

Aim: 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 Description:

#	Column	Non-Null Count	Dtype
0	Customer ID	30000 non-null	object
1	Name	30000 non-null	object
2	Gender	29947 non-null	object
3	Age	30000 non-null	int64
4	Income (USD)	25424 non-null	float64
5	Income Stability	28317 non-null	object
6	Profession	30000 non-null	object
7	Type of Employment	22730 non-null	object
8	Location	30000 non-null	object
9	Loan Amount Request (USD)	30000 non-null	float64
10	Current Loan Expenses (USD)	29828 non-null	float64
11	Expense Type 1	30000 non-null	object
12	Expense Type 2	30000 non-null	object
13	Dependents	27507 non-null	float64
14	Credit Score	28297 non-null	float64
15	No. of Defaults	30000 non-null	int64
16	Has Active Credit Card	28434 non-null	object
17	Property ID	30000 non-null	int64
18	Property Age	25150 non-null	float64
19	Property Type	30000 non-null	int64
20	Property Location	29644 non-null	object
21	Co-Applicant	30000 non-null	int64
22	Property Price	30000 non-null	float64
23	Loan Sanction Amount (USD)	29660 non-null	float64

Table 1: Loan Dataset Info

Libraries used:

- **Pandas & NumPy:** For data manipulation, numerical analysis, and array operations.
- **Matplotlib & Seaborn:** For data visualization, including scatter plots and heatmaps.
- **Time:** For measuring the computational time of model training and inference.

- **Scikit-Learn (Preprocessing):** For handling missing values (`SimpleImputer`), encoding categorical variables (`OrdinalEncoder`), and feature scaling (`StandardScaler`).
- **Scikit-Learn (Model Selection):** For splitting datasets, cross-validation (`KFold`, `cross_val_score`), and hyperparameter tuning (`GridSearchCV`).
- **Scikit-Learn (Algorithms):** For implementing regression models, including Linear Regression, Ridge, Lasso, ElasticNet, and Support Vector Regression (SVR).
- **Scikit-Learn (Metrics):** For evaluating regression performance using R-squared (`r2_score`), Mean Squared Error (MSE), and Mean Absolute Error (MAE).

Mathematical/theoretical description of the algorithm/objective performed:

- **Preprocessing steps:**
  - **Filtering:** Columns like `Customer ID` and `Name` dont have a significance in the loan amount prediction and hence have been removed.
  - **Handling null values:** Observed -999 used as null value in some columns and replaced with proper `np.nan` value. Removed all rows where loan sanction amount is null as it is the target variable. Replaced null values in `Dependents` column with 0 dependents and `Has Credit Card` as 0 (No credit card). Used mode imputation for Location and Gender. Finally the null values remaining columns are replaced with median of the respective column.
  - **Transformation:** We observe some columns with outliers causing a heavily left skewed distribution like `Current Loan Expenses, Income, Property Age, Loan Sanction Amount`. We apply a log transform on these features to make them more distributed and reduce the effect of outliers.
  - **Scaling:** Used `StandardScaler` to scale parameters. It transforms parameters based on the mean and variance of the dataset, resulting in a distribution with mean 0 and variance 1. This is a good Scaler for regression tasks

$$x_{scaled} = \frac{x_{original} - \mu_x}{\sigma_x}$$

- **Exploratory Data Analysis:**
  - Observed a large number of loans with loan sanction amount = 0. This symbolizes loans that have been rejected and hence are not needed for this analysis.

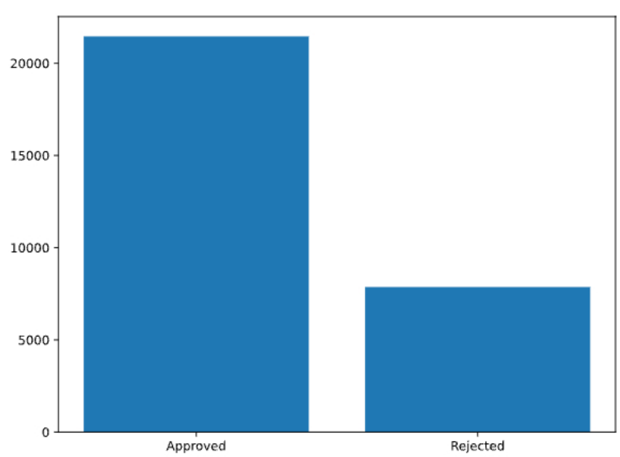

Figure 1: Rejected loans analysis

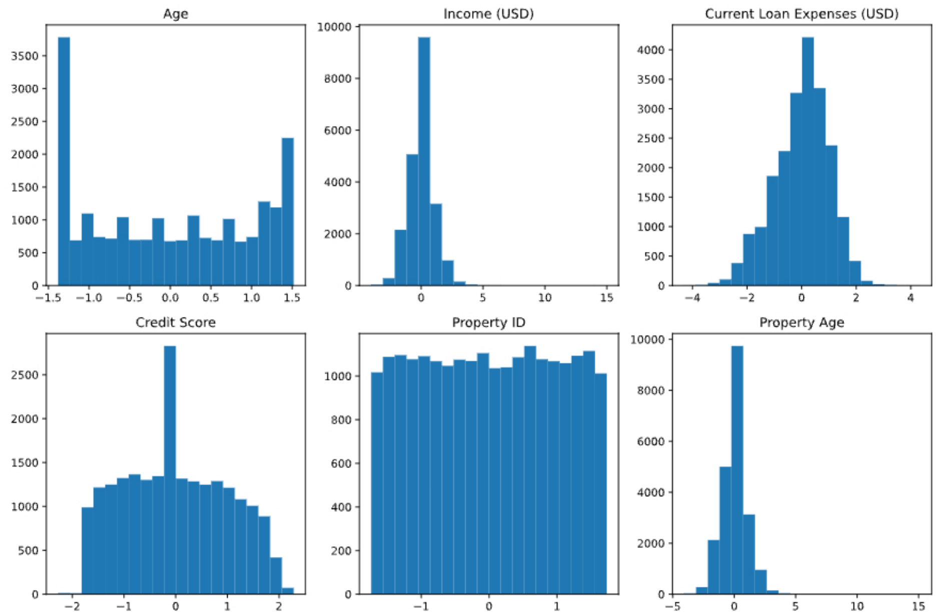

Figure 2: Distribution of features

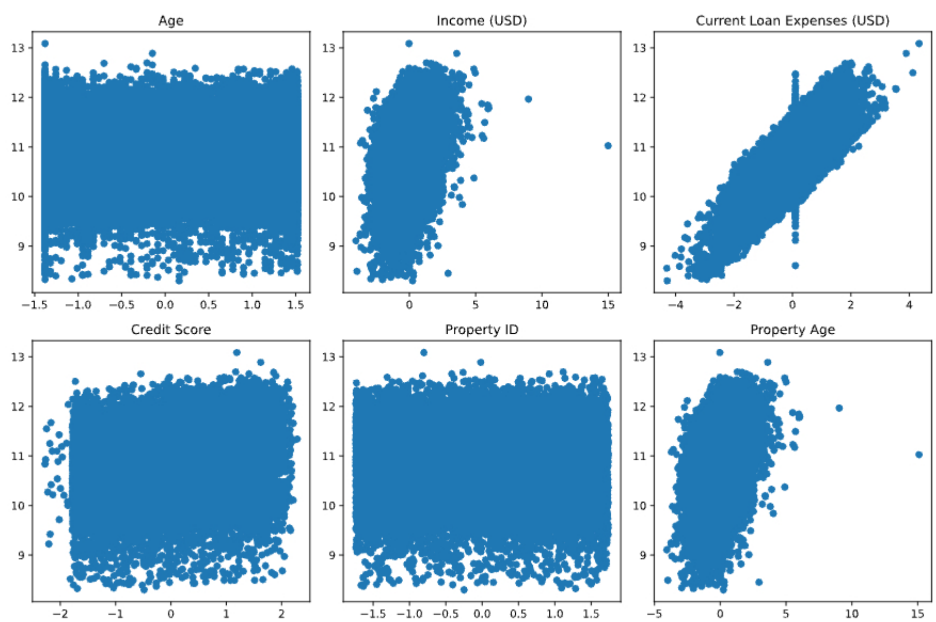

Figure 3: Analysing linear relationship

- We observe that some columns like Property Age and Income show linear relationship with target. and we also note that the columns are almost following a normal distribution

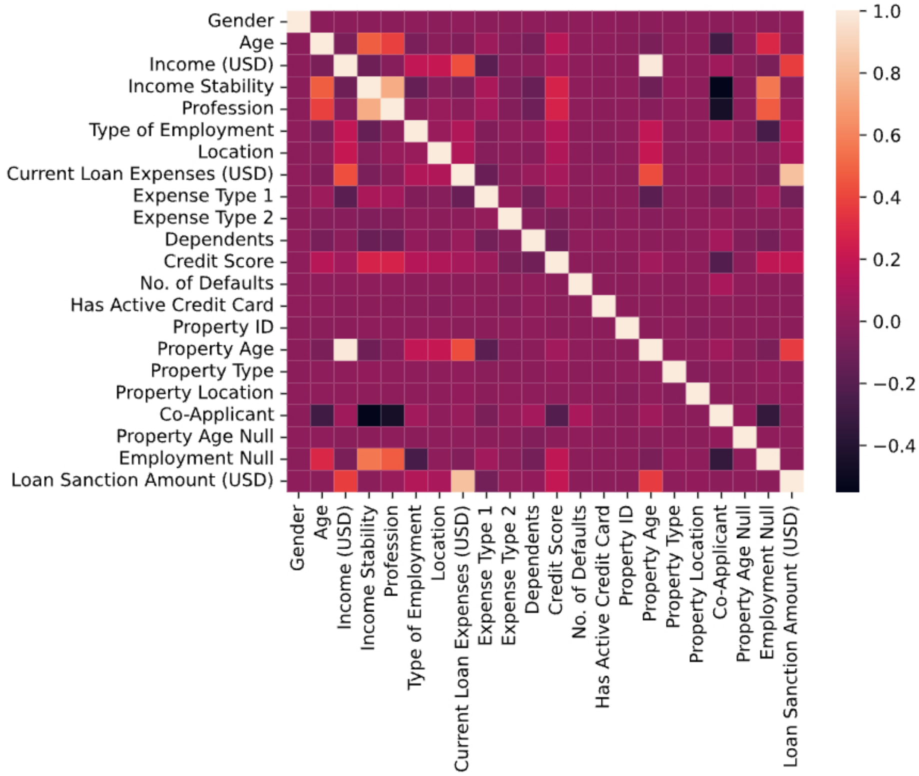

Figure 4: Heatmap

– The heatmap show strong correlation between input features Loan amount request, Income, and Property Price. This isn't ideal as this repetition might cause only one of these parameters to be considered important by the model and the other parameters may get ignored. Hence, we remove Property price and Loan amount request and Property price from input features.

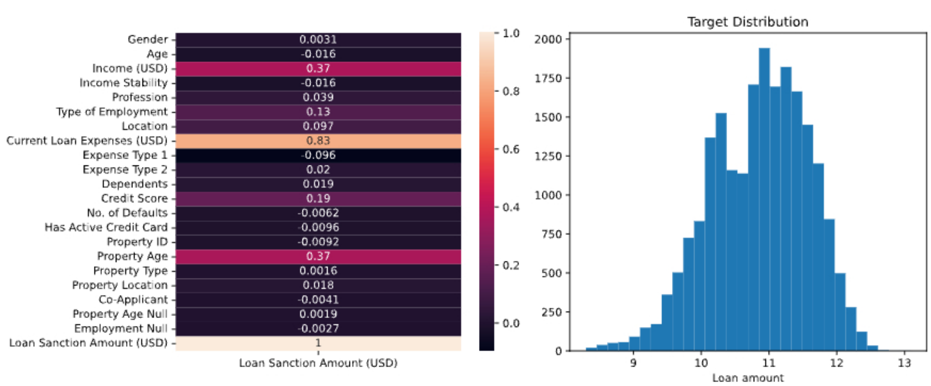

Figure 5: Target distribution analysis and correlation

– Target distribution follows a gaussian distribution (After log transform). The correlation map shows the correlation between input features and target. We observe Current Loan Expenses having the highest correlation value.

- **Performance Metrics:**

– Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

– Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

– Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

– R^2 Score:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

– Adjusted R^2 Score:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Results and Discussions:

- **Visualizations:**

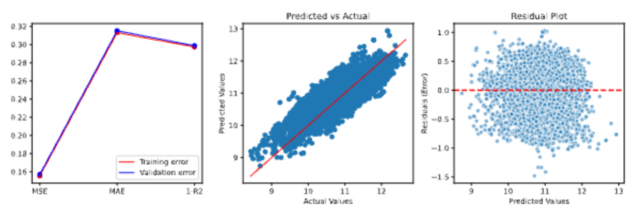

Figure 6: Linear regression results

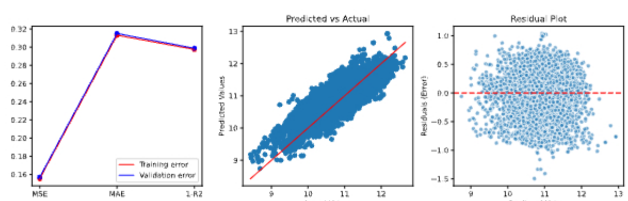

Figure 7: Lasso regression results

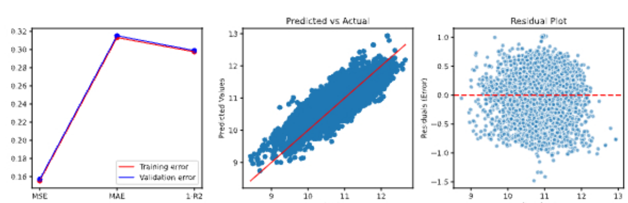

Figure 8: Ridge regression results

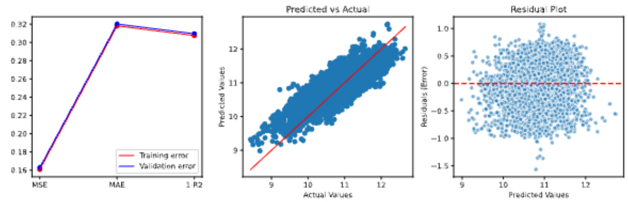

Figure 9: Elasticnet regression results

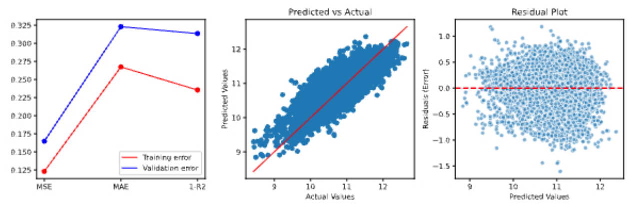

Figure 10: Support Vector regressor results

- **Hyperparameter Analysis:**

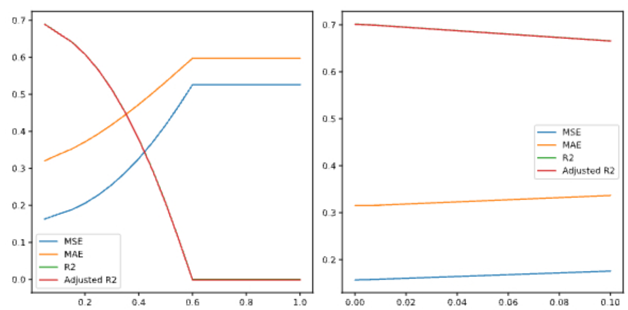

Figure 11: Lasso Regression α vs Error graph

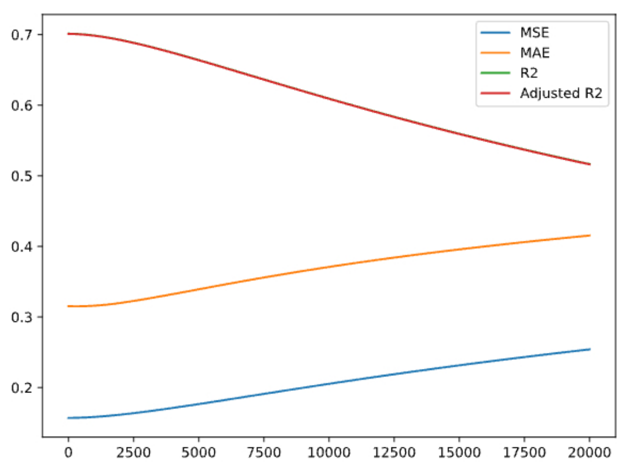

Figure 12: Ridge Regression α vs Error graph

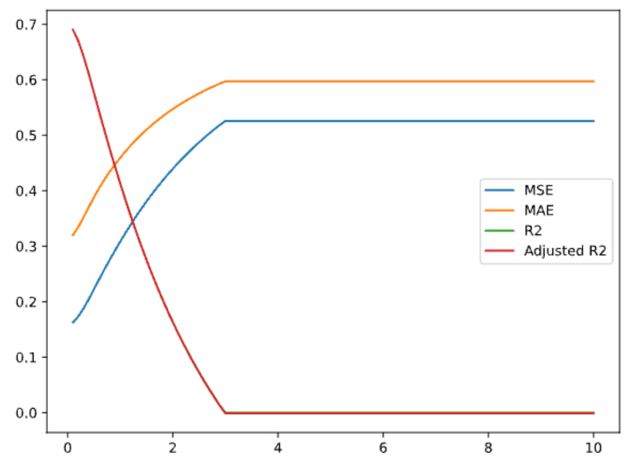

Figure 13: Elastic net Regression α vs Error graph (l1_ratio=0.5)

- **Hyperparameter Tuning Results:**

Table 2: Hyperparameter Tuning Summary

Model	Search Method	Best Parameters	Best CV R^2
Ridge Regression	Grid	$\alpha = 53.5789$	0.7015
Lasso Regression	Grid	$\alpha = 0.0016$	0.7016
Elastic Net Regression	Grid	$\alpha = 0.0016$, $l1\_ratio = 1$	0.7016
Support Vector Regressor	Grid	$C = 10$, $\gamma = scale$, $kernel = rbf$	0.6215

- **Cross-Validation Performance (K = 5):**

Table 3: Cross-Validation Performance (Training Metrics)

Model	MAE	MSE	RMSE	R^2
Linear Regression	0.3132	0.1554	0.3942	0.7026
Ridge Regression	0.3132	0.1554	0.3942	0.7026
Lasso Regression	0.3132	0.1555	0.3943	0.7024
Elastic Net Regression	0.3132	0.1555	0.3943	0.7024
SVR	0.1974	0.0759	0.2755	0.8547

- **Test Set Performance Comparison:**

Table 4: Test Set Performance (Validation Metrics)

Model	MAE	MSE	RMSE	R^2
Linear Regression	0.3152	0.1570	0.3963	0.7012
Ridge Regression	0.3152	0.1570	0.3963	0.7012
Lasso Regression	0.3151	0.1571	0.3963	0.7011
Elastic Net Regression	0.3151	0.1571	0.3963	0.7011
SVR	0.3560	0.1989	0.4460	0.6215

- **Effect of Regularization on Coefficients:**

Table 5: Coefficient Comparison

Feature	Linear	Lasso	Ridge	Elastic Net
Current Loan Expenses (USD)	0.585803	0.585048	0.585762	0.522390
Credit Score	0.081505	0.080747	0.081502	0.064288
Income (USD)	0.038214	0.015573	0.038006	0.014405
Property Age	-0.021503	0.000000	-0.021279	0.007094
Expense Type 2	0.041373	0.036730	0.041362	0.000000

Learning Practices:

- Learned pre-processing and EDA steps for a regression task.
- Learned to solve Regression tasks using Linear Regression and SVM.
- Learned to apply L1 and L2 Regularization to Linear Regression.
- Learned to tune hyperparameters for L1 and L2 Regression.