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

Degree & Branch	B.E. Computer Science & Engineering	Semester	V
Subject Code & Name	ICS1512 & Machine Learning Algorithms Laboratory		
Academic Year	2025-2026 (Odd)	Batch: 2023-2028	**Due Date:**

Experiment 2: Loan Amount Prediction Using Linear Regression

1 Aim

To develop and evaluate a machine learning model in Python for predicting loan amounts based on applicants' financial and credit-related attributes, employing Linear Regression and K-Fold Cross-Validation for model assessment.

2 Libraries Used

- NumPy
- Pandas
- Matplotlib
- Scikit-learn
- Seaborn

3 Objectives

- Preprocess the loan dataset by handling missing values, encoding categorical variables, selecting relevant features, and partitioning data into training, validation, and testing sets.
- Build and evaluate a Linear Regression model using K-Fold Cross-Validation; analyze metrics such as MSE, RMSE, MAE, and R^2; and interpret residuals and prediction plots to gauge model effectiveness.

4 Mathematical / Theoretical Background

4.1 Handling Missing Values

Missing values can degrade model performance by:

- Distorting statistical summaries,
- Causing errors during model training,
- Introducing bias in predictions.

Therefore, it is essential to detect and address missing data before modeling.

Common approaches include:

- Imputing missing values with mean, median, or mode using pandas' `fillna()` method.
- Dropping columns with excessive missing data that have limited relevance, to simplify the dataset and reduce noise.

4.2 Label Encoding

Machine learning models require numerical input; thus, categorical features (e.g., "Yes"/"No", "Graduate"/"Not Graduate") must be converted:

- Binary categories can be mapped directly to numeric codes (e.g., Yes = 1, No = 0), ensuring compatibility with most algorithms.
- For features with multiple categories, **one-hot encoding** is preferred to avoid imposing artificial ordinal relationships.

4.3 Plotting and Visualization

Visual tools help reveal data characteristics, such as correlations and outliers:

- **Heatmap:** Illustrates correlations between numeric features through color gradients, aiding feature selection.
- **Histogram:** Displays frequency distributions, helping identify skewness or gaps.
- **Boxplot:** Summarizes data distribution, highlights medians, quartiles, and outliers.

4.4 Standardization

- Standardization rescales features to zero mean and unit variance, useful when variables have different units or scales.
- Formula:
$$z = \frac{x - \mu}{\sigma}$$
where x is the original value, μ the mean, and σ the standard deviation.

This step enhances model convergence and performance.

5 Code

width=!,height=!,pages=-

6 Included Plots

- **Heatmap**: Shows feature correlations to identify strong linear relationships.
- **Boxplot**: Visualizes distribution, central tendency, and detects outliers.
- **Scatter Plot**: Displays relationships between pairs of numerical variables.
- **Histogram**: Illustrates the distribution of data points across bins.

7 Best Practices Followed

- **Consistent Preprocessing**: Applied uniform cleaning, encoding, and scaling of features to improve model generalization.
- **Robust Validation**: Employed 5-fold cross-validation to assess model stability across different data splits, minimizing overfitting risk.

8 Learning Outcomes

- **Comprehensive Pipeline Knowledge**: Gained practical experience with data exploration, preprocessing, model training, evaluation, and visualization.
- **Model Evaluation Insights**: Learned to interpret MAE, MSE, RMSE, and R^2 metrics, and to analyze residual plots for diagnosing model fit quality.

GitHub Repository: https://github.com/Thamizhmathibharathi/project.git

9 Results Table

Table 1: Model Summary: Loan Amount Prediction

Field	Details
Project Description	Predicting sanctioned loan amounts using applicant income, credit, and asset information.
Dataset Size (post-preprocessing)	30,000 records
Train/Test Split	80:20 (test_size=0.2)
Features Used	Age, Income Stability, Loan Amount Requested, Dependents, Credit Score, Number of Defaults, Active Credit Card Status, Property Location, Co-Applicant Status
Model	Linear Regression
Cross-Validation	Yes (5 folds)
Mean Absolute Error (MAE)	13,803.42
Mean Squared Error (MSE)	527,445,271.77
Root Mean Squared Error (RMSE)	622,966.18
R^2 **Score**	0.71
Adjusted R^2 **Score**	Not calculated
Key Influential Features	Loan Amount Requested and Income Stability (highest positive coefficients)
Residual Plot Observations	Residuals mostly evenly spread with slight underestimation at higher values
Predicted vs Actual Plot Interpretation	Shows overall upward trend; deviations increase with larger loan amounts
Overfitting / Underfitting	Minor underfitting observed
Rationale	Training and CV scores are close, indicating no severe overfitting. Residual distribution suggests model may miss some complex patterns.

Table 2: Cross-Validation Results (K = 5)

Fold	**MAE**	**MSE**	**RMSE**	R^2 **Score**
Fold 1	13,803.42	527,445,271.77	22,966.18	0.69
Fold 2	13,779.90	493,351,608.13	22,211.52	0.70
Fold 3	14,030.71	544,801,753.47	23,340.99	0.67
Fold 4	14,044.93	513,654,615.80	22,663.95	0.70
Fold 5	13,347.16	440,761,214.18	20,994.31	0.73
Average	**13,801.23**	**504,002,892.67**	**22,435.39**	**0.70**