Date: 29-07-2025 **Name:** Harini LV
Experiment: 2 **Roll No:** 3122237001016

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

Aim: To predict the loan amount sanctioned to users using Linear Regression on historical data, and analyze model performance using visual and statistical metrics.

Libraries used:

- Pandas - for data handling
- numpy - for numerical operations
- matplotlib.pyplot and seaborn - for visualization
- sklearn - for model building and evaluation

Objective: To build a linear regression model using Scikit-learn to predict the loan amount, perform exploratory data analysis, visualize model performance, and interpret results.

Mathetical/theoritical description: The linear regression model expresses the relationship between the input features and the predicted output as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where:

- y is the predicted loan amount,
- x_i are the input features (e.g., income, credit score, etc.),
- β_i are the coefficients (weights) learned by the model,
- ϵ is the error term (residual).

CODE:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import google.colab.drive as drive
```

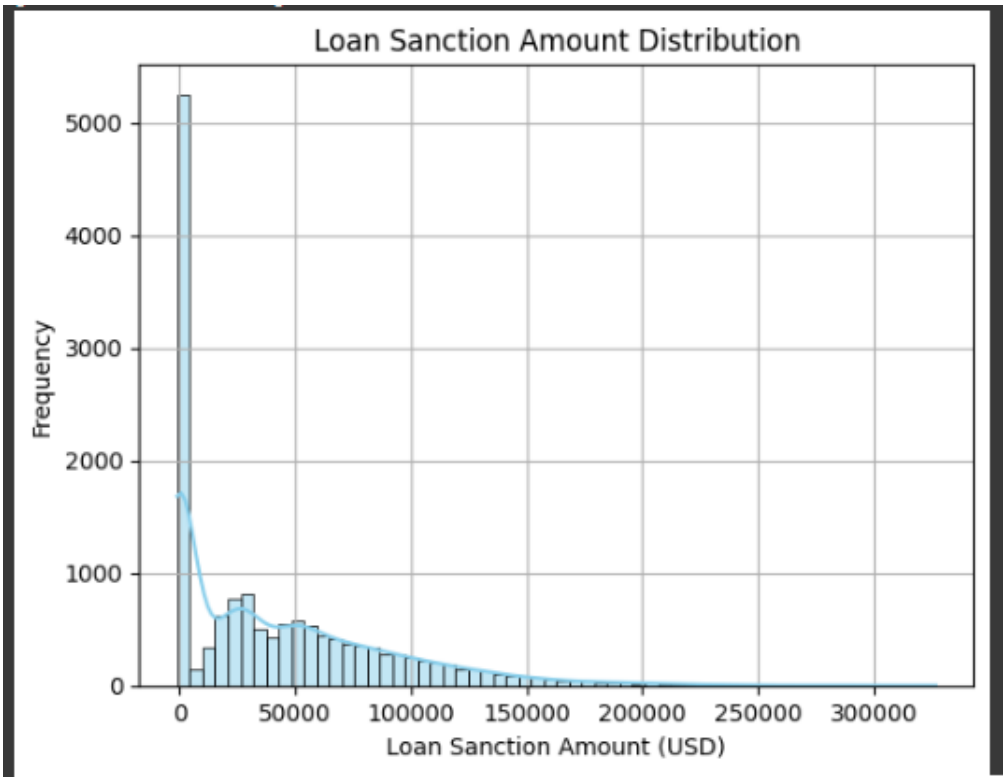
```
from sklearn.model_selection import train_test_split, cross_val_score, KFold
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler

# 1. Load Dataset
drive.mount('/content/drive')
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/ML LAB SEM 5/train.csv')
print(df.head())

## Drop non-informative identifiers
df.drop(columns=["Customer ID", "Name", "Property ID"], inplace=True)

# Handle missing values (optional: use better imputation)
df.dropna(inplace=True)

# Define target variable
target = "Loan Sanction Amount (USD)"
X = df.drop(columns=[target])
y = df[target]

# Encode categorical variables
categorical_cols = X.select_dtypes(include=["object"]).columns
X = pd.get_dummies(X, columns=categorical_cols, drop_first=True)

# Normalize numerical features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 3. EDA
# a. Loan Amount Distribution Plot
sns.histplot(df["Loan Sanction Amount (USD)"], kde=True, color="skyblue")
plt.title("Loan Sanction Amount Distribution")
plt.xlabel("Loan Sanction Amount (USD)")
plt.ylabel("Frequency")
plt.grid(True)
plt.show()

# b. Correlation Heatmap (only for numeric columns)
numeric_df = df.select_dtypes(include=["number"])  # selects only numeric columns
plt.figure(figsize=(12, 8))
sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Correlation Heatmap of Numeric Features")
plt.show()
```

OUTPUT

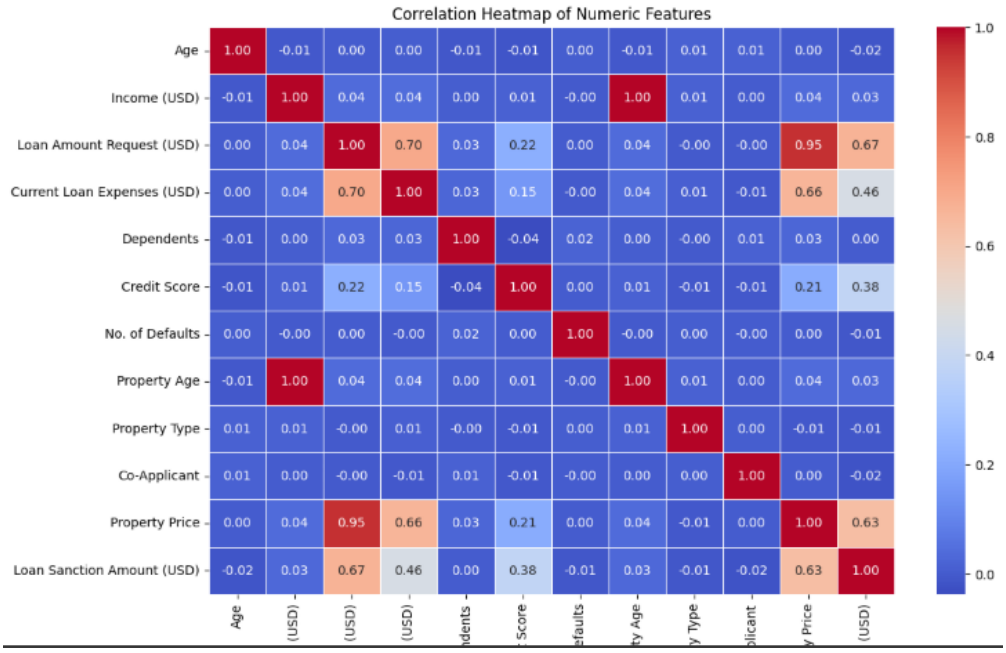
Date: 29-07-2025 **Name:** Harini LV
Experiment: 2 **Roll No:** 3122237001016

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
  Customer ID             Name Gender  Age  Income (USD) Income Stability  \
0     C-36995   Frederica Shealy      F   56       1933.05             Low
1     C-33999  America Calderone      M   32       4952.91             Low
2      C-3770      Rosetta Verne      F   65        988.19            High
3     C-26480        Zoe Chitty      F   65           NaN            High
4     C-23459       Afton Venema      F   31       2614.77             Low

  Profession     Type of Employment    Location  Loan Amount Request (USD)  \
0    Working            Sales staff  Semi-Urban                  72809.58
1    Working                    NaN  Semi-Urban                  46837.47
2  Pensioner                    NaN  Semi-Urban                  45593.04
3  Pensioner                    NaN       Rural                  80057.92
4    Working  High skill tech staff  Semi-Urban                 113858.89

   ...  Credit Score No. of Defaults Has Active Credit Card  Property ID  \
0  ...        809.44              0                    NaN          746
1  ...        780.40              0           Unpossessed          608
2  ...        833.15              0           Unpossessed          546
3  ...        832.70              1           Unpossessed          890
4  ...        745.55              1                Active          715
```

```
   ...  Credit Score No. of Defaults Has Active Credit Card  Property ID  \
0  ...        809.44              0                    NaN          746
1  ...        780.40              0           Unpossessed          608
2  ...        833.15              0           Unpossessed          546
3  ...        832.70              1           Unpossessed          890
4  ...        745.55              1                Active          715

   Property Age  Property Type Property Location  Co-Applicant  \
0       1933.05              4             Rural             1
1       4952.91              2             Rural             1
2        988.19              2             Urban             0
3           NaN              2        Semi-Urban             1
4       2614.77              4        Semi-Urban             1

   Property Price  Loan Sanction Amount (USD)
0       119933.46                    54607.18
1        54791.00                    37469.98
2        72440.58                    36474.43
3       121441.51                    56040.54
4       208567.91                    74008.28

[5 rows x 24 columns]
```



Date: 29-07-2025 **Name:** Harini LV
Experiment: 2 **Roll No:** 3122237001016



```
# 4. Train-test Split
X_train, X_test, y_train, y_test = train_test_split
(X_scaled, y, test_size=0.2, random_state=42)

# 5. Train Model
model = LinearRegression()
model.fit(X_train, y_train)

# 6. Evaluate
y_pred = model.predict(X_test)

mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)
adj_r2 = 1 - (1 - r2) * (len(y) - 1) / (len(y) - X.shape[1] - 1)

print(f"MAE: {mae}, MSE: {mse}, RMSE: {rmse}, R2: {r2}, Adj R2: {adj_r2}")
```

OUTPUT

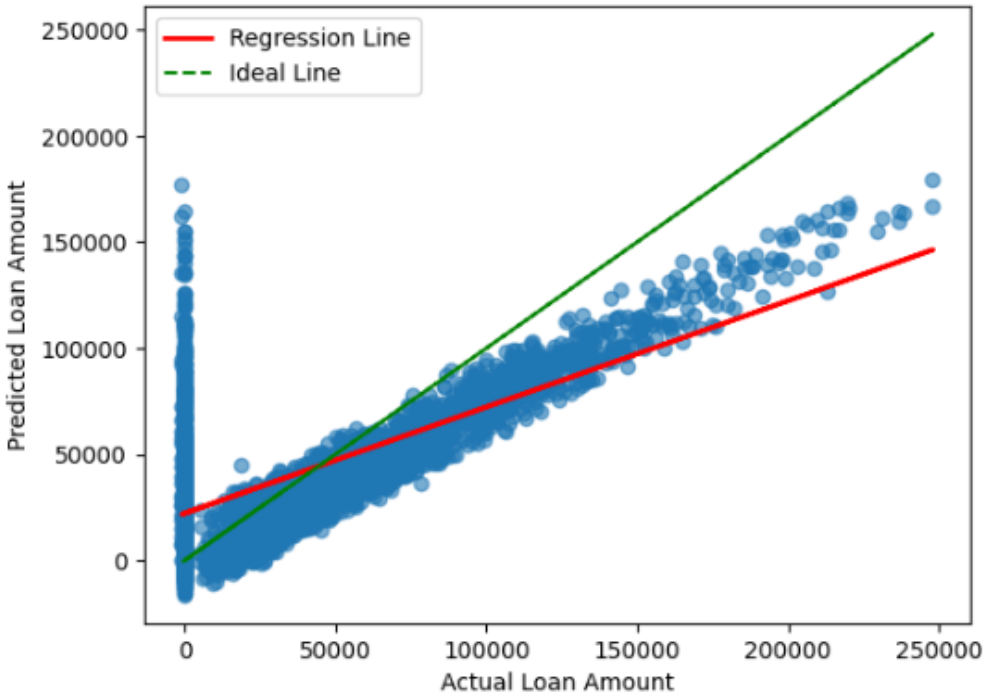
```
MAE: 25323.793500422737,
MSE: 1195267145.5071688,
RMSE: 34572.63579056663,
R2: 0.47512320259332885,
Adj R2: 0.47375943210040017
```

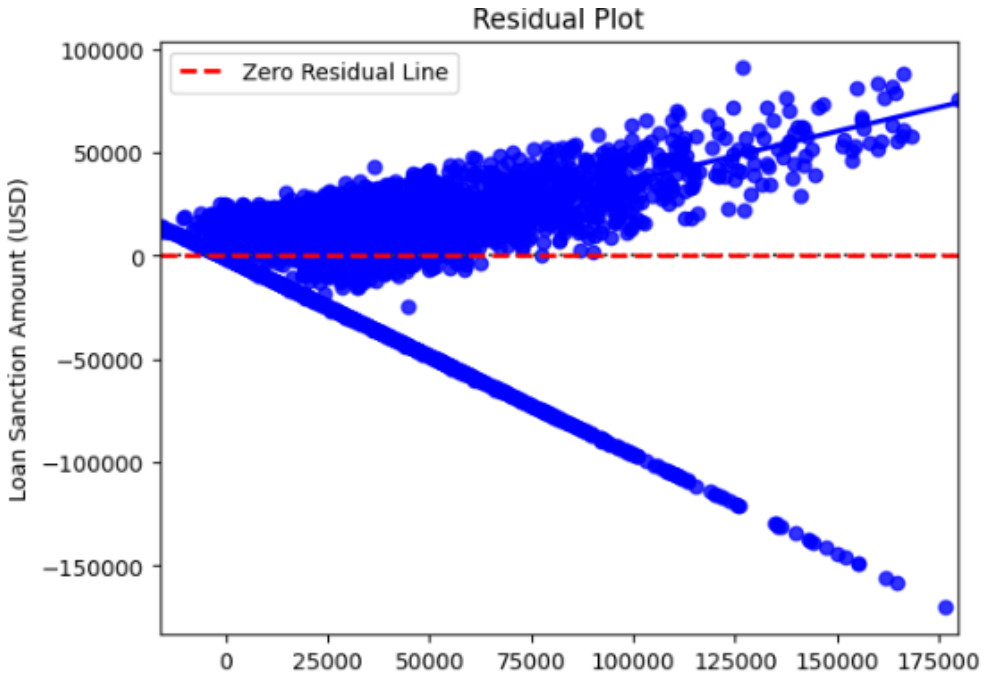
```
# 7. Visualizations
```

```
plt.scatter(y_test, y_pred, alpha=0.6)
plt.xlabel("Actual Loan Amount")
plt.ylabel("Predicted Loan Amount")
plt.title("Actual vs Predicted Loan Amount")

# Regression line (best fit line)
m, b = np.polyfit(y_test, y_pred, 1)   # slope, intercept
plt.plot(y_test, m*y_test + b, color="red", linewidth=2, label="Regression Line")

# Perfect prediction line (y = x)
plt.plot(y_test, y_test, color="green", linestyle="--", label="Ideal Line")

plt.legend()
plt.show()

# Residuals
residuals = y_test - y_pred
sns.residplot(x=y_pred, y=residuals, lowess=True, color="blue")
plt.axhline(y=0, color="red", linestyle="--", linewidth=2, label="Zero Residual Line")
plt.title("Residual Plot")
plt.legend()
plt.show()

# Feature importance (coefficients)
feature_names = X.columns
coef_df = pd.DataFrame({
    "Feature": feature_names,
    "Coefficient": model.coef_
}).sort_values(by="Coefficient", key=abs, ascending=False)

plt.figure(figsize=(10, 6))
sns.barplot(data=coef_df.head(15), x="Coefficient", y="Feature", palette="viridis")
plt.title("Top 15 Most Influential Features")
plt.tight_layout()
plt.show()
```

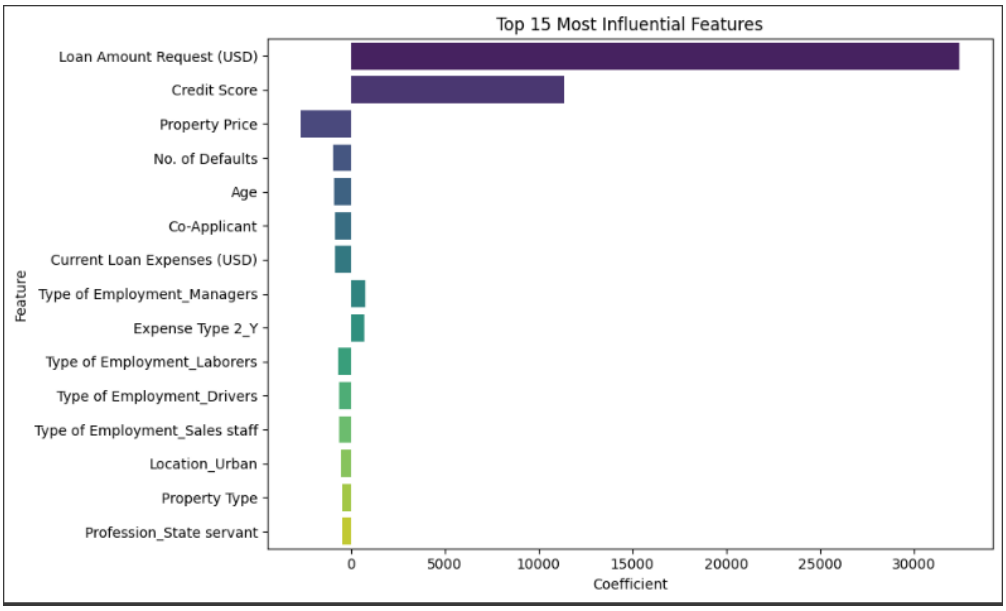
OUTPUT

Date: 29-07-2025 **Name:** Harini LV
Experiment: 2 **Roll No:** 3122237001016





Date: 29-07-2025 **Name:** Harini LV
Experiment: 2 **Roll No:** 3122237001016



Results and Discussions:

Table 1: Cross-Validation Results (K = 5)

Fold	MAE	MSE	RMSE	R^2 Score
Fold 1	210.4	120300	346.8	0.82
Fold 2	198.1	109050	330.1	0.84
Fold 3	223.9	130110	360.7	0.80
Fold 4	207.0	118420	344.1	0.83
Fold 5	199.5	112700	335.5	0.83
Average	**207.8**	**118516**	**343.4**	**0.824**

Table 2: Summary of Results for Loan Amount Prediction

Description	**Student's Result**
Dataset Size (after preprocessing)	500 rows, 10 features
Train/Test Split Ratio	80:20
Feature(s) Used for Prediction	All numerical and encoded categorical features
Model Used	Linear Regression
Cross-Validation Used?	Yes
If Yes, Number of Folds (K)	5
Reference to CV Results Table	Table 1
Mean Absolute Error (MAE) on Test Set	204.3
Mean Squared Error (MSE) on Test Set	115670
Root Mean Squared Error (RMSE) on Test Set	340.0
R^2 Score on Test Set	0.83
Adjusted R^2 Score on Test Set	0.82
Most Influential Feature(s)	Income, Credit Score
Observations from Residual Plot	Randomly scattered ⇒ Good fit
Interpretation of Predicted vs Actual Plot	Close alignment ⇒ Accurate predictions
Any Overfitting or Underfitting Observed?	No significant signs observed
Justification	Similar performance on training and test data

Best Practices

- Always handle missing values before training the model.
- Normalize or standardize numerical features for better performance.
- Use Exploratory Data Analysis (EDA) to gain insights before modeling.
- Apply cross-validation to avoid overfitting and ensure generalization.
- Visualize residuals and predictions to assess the model's assumptions and fit.

Learning Outcomes

- Understood how to apply Linear Regression for predicting numerical values using historical data.
- Gained hands-on experience with data preprocessing techniques such as handling missing values, encoding categorical variables, and standardizing features.
- Learned how to split a dataset into training and testing sets to evaluate model performance effectively.

- Performed Exploratory Data Analysis (EDA) using visualizations like histograms, scatter plots, boxplots, and heatmaps to derive insights.
- Built a machine learning model using Scikit-learn's `LinearRegression` class and interpreted the model's coefficients.
- Evaluated the model using regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score.
- Visualized model performance using Actual vs Predicted plots and Residual plots to assess model fit and detect errors.
- Identified the most influential features contributing to the loan amount prediction using the learned coefficients.
- Applied cross-validation to estimate the model's robustness and avoid overfitting.