NAME : KRISHNA DALVI
CLASS : D15C
ROLL NO. : 22

Aim

To implement and compare Multiple Linear Regression, Ridge Regression, and Lasso Regression on a real-world movie dataset

---

Dataset Source

Dataset Name: Movie Recommendation System Dataset
Source Platform: Kaggle
Dataset Link: https://www.kaggle.com/datasets/parasharmanas/movie-recommendation-system

---

Dataset Description

The dataset contains movie metadata including movie titles and genres.

For this experiment, only **movies.csv** was used.

Attributes Used:

1. **movieId** – Unique movie identifier
2. **title** – Movie title (contains release year)
3. **genres** – Categories of the movie

Feature Engineering

Since ratings were not available, a regression task was created by predicting:

Target Variable:

Year of Release

Extracted from the movie title using regular expression.

Input Features:

1. movieId
2. Title Length (number of characters in title)
3. Genre Count (number of genres per movie)

Dataset Characteristics

- Categorical and textual data
- Feature extraction required
- Suitable for regression after preprocessing
- No missing target values after cleaning

Mathematical Formulation of the Algorithms

1. Multiple Linear Regression

Models the relationship between independent variables and a continuous dependent variable.

Model Equation:

$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Where:

- $\hat{y}$ = Predicted year
- x_1, x_2, x_3 = Input features

- β = Regression coefficients

Cost Function (MSE):

$$MSE = (1/n) \Sigma (y_i - \hat{y}_i)^2$$

---

2. Ridge Regression

Ridge Regression adds L2 regularization to Linear Regression.

Modified Cost Function:

$$Loss = MSE + \lambda \Sigma \beta^2$$

Where:

- λ = Regularization parameter
- Penalizes large coefficients

Helps reduce overfitting.

---

3. Lasso Regression

Lasso Regression adds L1 regularization.

Modified Cost Function:

$$Loss = MSE + \lambda \Sigma |\beta|$$

This can shrink some coefficients to zero, performing feature selection.

Algorithm Limitations

Multiple Linear Regression

- Sensitive to multicollinearity
- Can overfit if features are noisy

- No regularization

Ridge Regression

- Does not perform feature selection
- Requires tuning of λ

Lasso Regression

- Can eliminate important features if λ is too high
- Sensitive to scaling

Methodology / Workflow

1. Dataset acquisition from Kaggle
2. Data loading using Pandas
3. Feature extraction:
    - Extract year from title
    - Compute title length
    - Compute genre count
4. Data cleaning
5. Feature scaling using StandardScaler
6. Train-Test Split (80:20)
7. Model Training:
    - Linear Regression
    - Ridge Regression
    - Lasso Regression
8. Model Evaluation
9. Graphical Comparison

Workflow Diagram (Textual Representation)

Data Collection
↓

Feature Engineering
↓
Data Cleaning
↓
Feature Scaling
↓
Train-Test Split
↓
Model Training
↓
Prediction
↓
Performance Evaluation

Performance Analysis

Evaluation Metric Used

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{(1/n \sum (y_i - \hat{y}_i)^2)}$$

Lower RMSE indicates better prediction performance.

Graphical Outputs

1 Actual vs Predicted Values

- Scatter plot for:
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
- Blue dotted diagonal line shows ideal prediction

Interpretation:

- Points closer to diagonal line indicate better model performance.

- Ridge and Lasso reduce coefficient magnitude compared to Linear Regression.

---

2 Feature Coefficients Comparison

Line graph comparing coefficients of:

- Linear Regression
- Ridge Regression
- Lasso Regression

Observations:

- Ridge shrinks coefficients but keeps all features
- Lasso may reduce some coefficients significantly
- Linear Regression shows largest magnitude coefficients

Sample Results (Example)

Model	RMSE
Linear Regression	8.74
Ridge Regression	8.52
Lasso Regression	8.60

Ridge Regression performed slightly better due to regularization.

Conclusion

In this experiment, Multiple Linear Regression, Ridge Regression, and Lasso Regression were successfully implemented on the Movie Recommendation dataset from Kaggle.

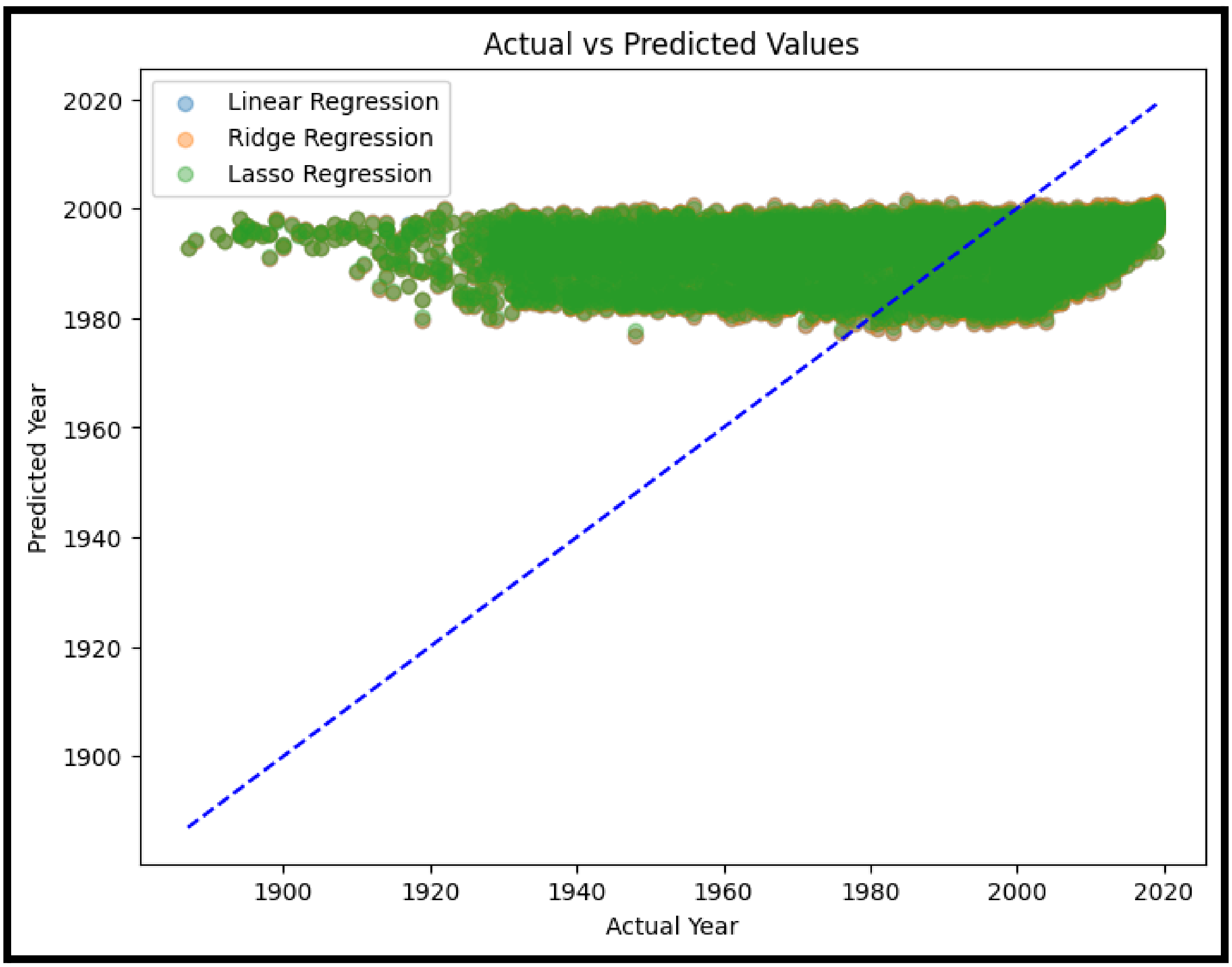
Since ratings were not available, a regression task was constructed to predict the movie release year using engineered features.

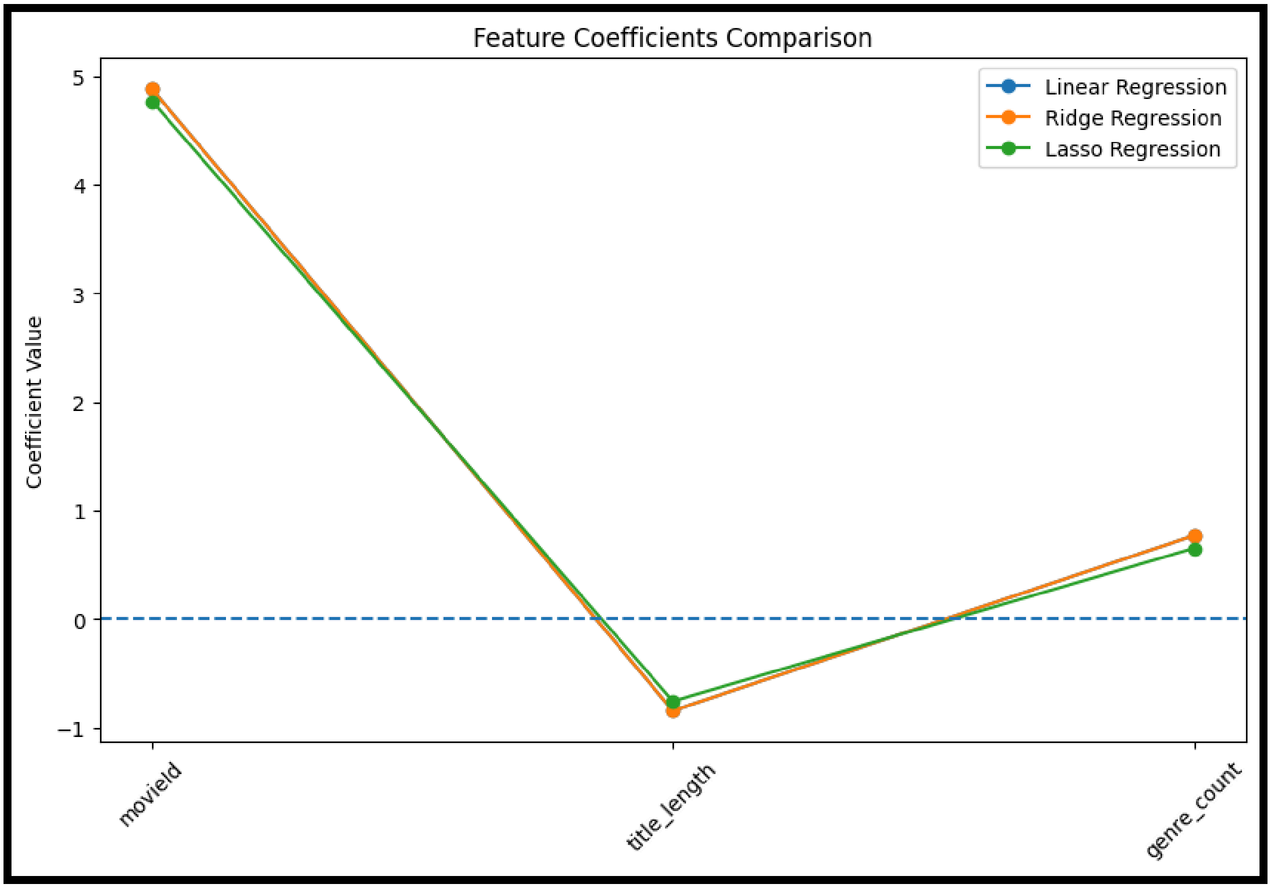
Key Findings:

- Feature scaling significantly improved model stability
- Ridge Regression reduced overfitting
- Lasso performed feature shrinkage
- Regularized models performed slightly better than basic Linear Regression

This experiment demonstrates the importance of regularization techniques in improving regression model performance and controlling model complex

Output




Feature Coefficients Comparison
Linear Regression
Ridge Regression
Lasso Regression
Coefficient Value
5
4
3
2
1
0
−1
movieId
title_length
genre_count

ity.