Predicting the Age of Fossils Using a Regression Model: AIBD Assign-2

Code link- AIBD-ASSIGN 2

Name: Jilla Sai Mahima
Roll number: BE22B047

Introduction

In this assignment, we will predict the fossil's age based on eight features (uranium lead ratio, carbon 14 ratio, radioactive decay series, stratigraphic layer depth, inclusion of other fossils, isotopic composition, fossil size, and fossil weight).
Using these features, we will build a regression model that predicts the age of the fossil.

Steps we will follow to build the model-

1. **Data Preprocessing**
2. **Model Selection**
3. **Training model (With and without regularization)**
4. **Hyperparameter tuning and Cross-Validation**
5. **Performance Metrics**
6. **Predictions and Sample Results**
7. **Conclusion**

To ensure model reliability, we test on two different train-test splits(80-20 & 70-30) and apply hyperparameter tuning & cross-validation to improve performance.

---

1. Data Preprocessing

Data preprocessing in machine learning involves cleaning, transforming, and preparing raw data to make it suitable for analysis and model training, ensuring optimal performance and accuracy.

The steps involved are

- Check for missing values because void spaces can affect the model's performance.
  - No missing values are found

- Getting the summary of the dataset helps us to understand the data distribution, spot outliers and check scaling issues.

- Defining the features and target variable.
  - Here, our target variable is “age,” and the remaining all are features.

---

2. Model Selection

Since the dataset provided is a continuous variable (age), regression models are the best option for predicting it.

We evaluate two models:

- **Linear Regression** (Without Regularization) - Baseline model for comparison
- **Lasso Regression** (Linear regression with L1 Regularization) - Reduces overfitting & selects important features

---

3. Training Models (With & Without Regularization)

Splitting the dataset using two different ratios-

- 80-20 Split → More training data, better generalization
- 70-30 Split → More text data, better performance.

Why Lasso Regression?
Lasso adds L1 regularization, which can shrink some coefficients to zero → Feature Selection and reduced overfitting.

---

4. Hyperparameter Tuning & Cross-Validation

Why tune hyperparameters?

Lasso has an **alpha parameter** (λ) that controls **regularization strength**. A higher alpha means more regularization, leading to fewer selected features.

Defining Cross-Validation Methods

- **K-fold(5)** → Splits the data into 5 parts, trains on 4, tests on 1 (repeats 5 times).
- **Shuffle-Split(80-20)** → Randomly splits data multiple times for a better estimate.

Finding the Best Alpha

To optimize the performance of the Lasso Regression model, we performed hyperparameter tuning by identifying the best value of the regularization parameter alpha using cross-validation.

We tested a range of alpha values from 0.001 to 1000 (on a logarithmic scale) and evaluated each using two different cross-validation strategies, i.e, K-fold, ShuffleSplit CV.

For each alpha, the average R^2 score (coefficient of determination) was computed. The alpha with the highest average R^2 was selected as the best alpha for each strategy.

```
Final models trained successfully!
 Lasso (80-20 Split) - Alpha: 10.0000
 Lasso (70-30 Split) - Alpha: 2.1544
```

- **80-20 Split**
    - **Alpha: 10.0000**
    - With more training data, a higher alpha (stronger regularization) helped control overfitting and simplified the model.
- **70-30 Split**
    - **Alpha: 2.1544**
    - With slightly less training data, the model performed better with weaker regularization, allowing it to retain more features.

---

5. Performance Metrics

To evaluate the performance of our trained regression models- Lasso Regression and Linear Regression, we used three standard regression metrics: R^2 Score, Mean Squared Error(MSE), and Mean Absolute Error(MAE). These were applied to both the 80-20 and 70-30 train-test splits.

Unlike classification tasks, where accuracy is an appropriate metric, our problem involves predicting continuous values(the age of fossils). Hence, we rely on regression-specific metrics that assess how closely our model's predictions match the actual values. The results are-

```
Lasso (80-20) Model Performance
  R² Score: 0.9203
  Mean Squared Error (MSE): 18892354.6629
  Mean Absolute Error (MAE): 3571.8555

Lasso (70-30) Model Performance
  R² Score: 0.9207
  Mean Squared Error (MSE): 18873262.2572
  Mean Absolute Error (MAE): 3582.1036

Linear Regression (80-20) Model Performance
  R² Score: 0.9205
  Mean Squared Error (MSE): 18860648.5184
  Mean Absolute Error (MAE): 3569.6696

Linear Regression (70-30) Model Performance
  R² Score: 0.9207
  Mean Squared Error (MSE): 18869409.7971
  Mean Absolute Error (MAE): 3581.6544
```

Key insights are-

- Both models show **very close performance** with $R^2 \approx 0.92$, indicating high accuracy.
- Lasso did not significantly outperform Linear Regression, suggesting **minimal benefit from regularization** in this dataset.
- Train-test splits (80-20 vs 70-30) had a **negligible impact**, demonstrating **model stability and generalization**

---

6. Predictions & Sample Results

To compare model behavior beyond just performance metrics, we generated predictions using both **Lasso** and **Linear Regression** models on a subset of the test data for both data splits (80-20 and 70-30).

- **Lasso (80-20 Split)** and **Linear Regression (80-20 Split)** produced similar outputs, with minor variations due to Lasso's feature regularization.
- **Lasso (70-30 Split)** and **Linear Regression (70-30 Split)** also had closely aligned predictions, reflecting the consistency and reliability of both models in capturing the

underlying data patterns.

Observation:
Despite the Lasso model applying regularization (shrinking some coefficients), its predictions remain remarkably close to those of standard linear regression, confirming that Lasso maintained predictive accuracy while also simplifying the model.

Final Takeaway

Lasso Regression is useful for reducing overfitting and selecting important features, but Linear Regression can still perform well when regularization is not needed.

THANK YOU!!!