Title: Reproducibility Report for Concrete Compressive Strength Data Mining Analysis

The objective of this analysis is to understand the key factors of Concrete Compressive Strength.

1. Data Description Dataset:

- Source: UCI Machine Learning Repository

- Variables: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age, Concrete Compressive Strength

Preprocessing Steps:

- Missing value handling: None (dataset is complete)

- Train/test split: 87% training, 13% testing

- Transformations:

    - o Raw predictors: Original feature values used without modification.

    - o Normalization: Features were scaled using min-max normalization to the range [0,1].

    - o Log transformation: Features were transformed using log(x + 1) to reduce skewness and stabilize variance.

2. Methods

Gradient Descent Algorithm:

- Pseudocode:

```
Initialize parameters (weights w, bias b)
Repeat until convergence or max iterations:
    y_pred = w * x + b
    error = y_pred – y
    compute gradients, dw, db
    update weights, w
    update bias, b
    compute MSE
    check if gradients are less than epsilon
return final weights, bias, MSE, and VE (1 – MSE/var(y))
```

- Parameters:

- Univariate models:

  o Normalized predictors: learning rate = 0.1, epsilon = 1e-6, max iterations = 500,000

  o Raw predictors: learning rate = 1e-6, epsilon = 1e-3, max iterations = 500,000

- Multivariate models:

  o Normalized predictors: learning rate = 0.1, epsilon = 1e-6, max iterations = 1,000,000

  o Raw predictors: learning rate = 1e-9, epsilon = 1e-3, max iterations = 1,000,000

Regression Analysis:

Library: statsmodels.api

Function call: sm.OLS(y_train, sm.add_constant(x_train)).fit()

Model types tested:

- Raw predictors

- Normalized predictors

- Log-transformed predictors (n.log1p(X))

Implementation:

- Python (NumPy, Statsmodels)

- Data loaded with custom CSV reader function.

- Train/test split: rows 501-630 used for testing, remainder for training.

3. Results

Gradient Descent:

- Final weights:

Table 1

Univariate Model with Normalized Predictors

Feature	Training MSE	Training VE	Testing MSE	Testing Ve
Cement	204.283	0.263	269.543	-0.186
Blast Furnace Slag	270.117	0.026	300.019	-0.32
Fly Ash	265.631	0.042	394.761	-0.737
Water	255.821	0.078	256.665	-0.13
Superplasticizer	248.748	0.103	193.494	0.148
Coarse Aggregate	272.756	0.017	270.513	-0.191
Fine Aggregate	271.087	0.023	287.513	-0.265
Age (day)	243.052	0.124	295.692	-0.301

Table 2

Univariate Model with Raw Predictors

Feature	Training MSE	Training VE	Testing MSE	Testing Ve
Cement	224.759	0.19	253.009	-0.114
Blast Furnace Slag	466.168	-0.681	493.796	-1.173
Fly Ash	374.284	-0.349	188.471	0.17
Water	290.068	-0.046	297.784	-0.311
Superplasticizer	251.616	0.093	177.811	0.217
Coarse Aggregate	280.68	-0.012	202.509	-0.336
Fine Aggregate	282.305	-0.018	299.584	-0.319
Age (day)	245.534	0.115	288.964	-0.272

Table 3

Multivariate Models

Predictor Type	Training MSE	Training VE	Testing MSE	Testing Ve
Normalized Predictors	104.15	0.625	146.937	0.353
Raw Predictors	137.807	0.503	107.509	0.527

Table 4

Multivariate Models with statsmodels.api

Predictor Type	Training MSE	Training VE	Testing MSE	Testing Ve
Normalized Predictors	104.15	0.625	146.937	0.353
Raw Predictors	104.15	0.625	141.038	0.379
Log-transformed Predictors	55.786	0.799	54.684	0.759

- Loss over iterations:

Figure 1

MSE Loss Curve for Normalized Predictors

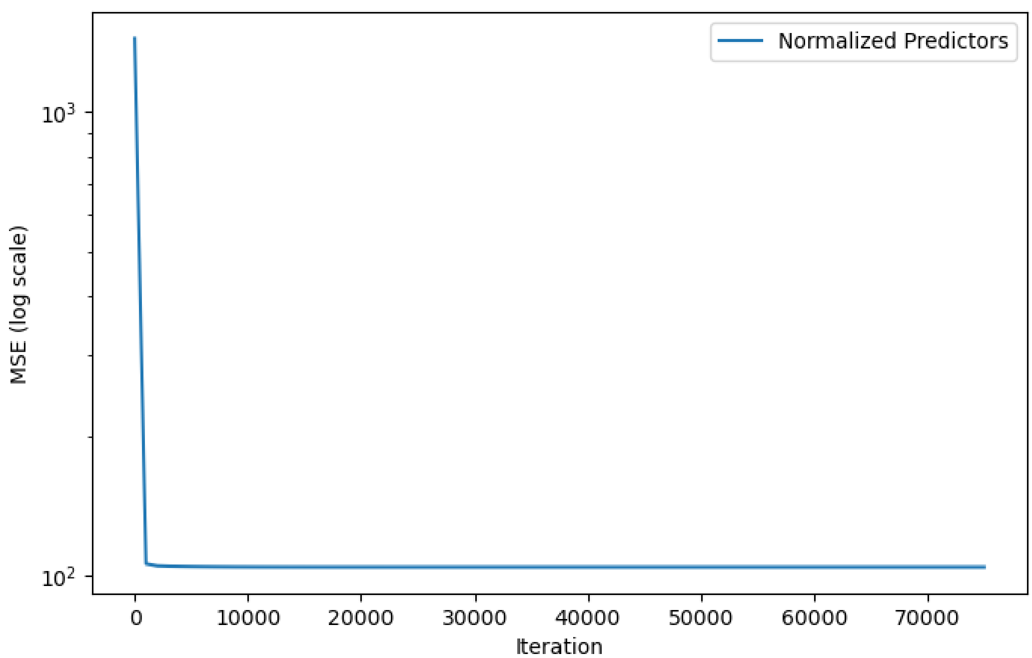


Figure 2

MSE Loss Curve for Normalized Predictors from Iteration 10000 Onward

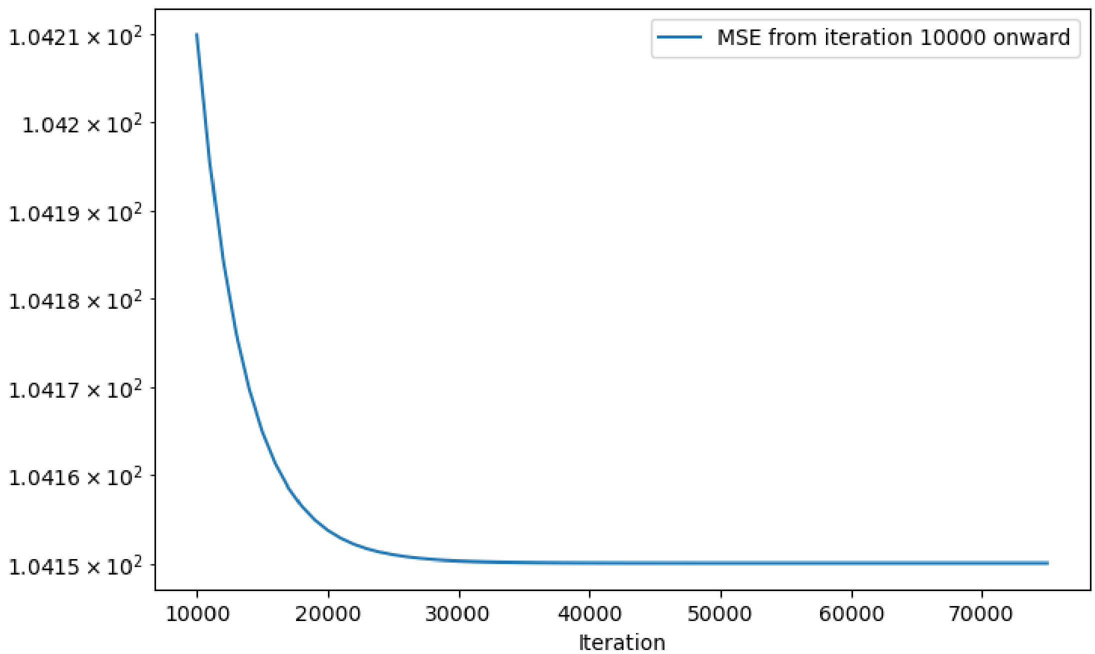


Figure 3

MSE Loss Curve for Raw Predictors

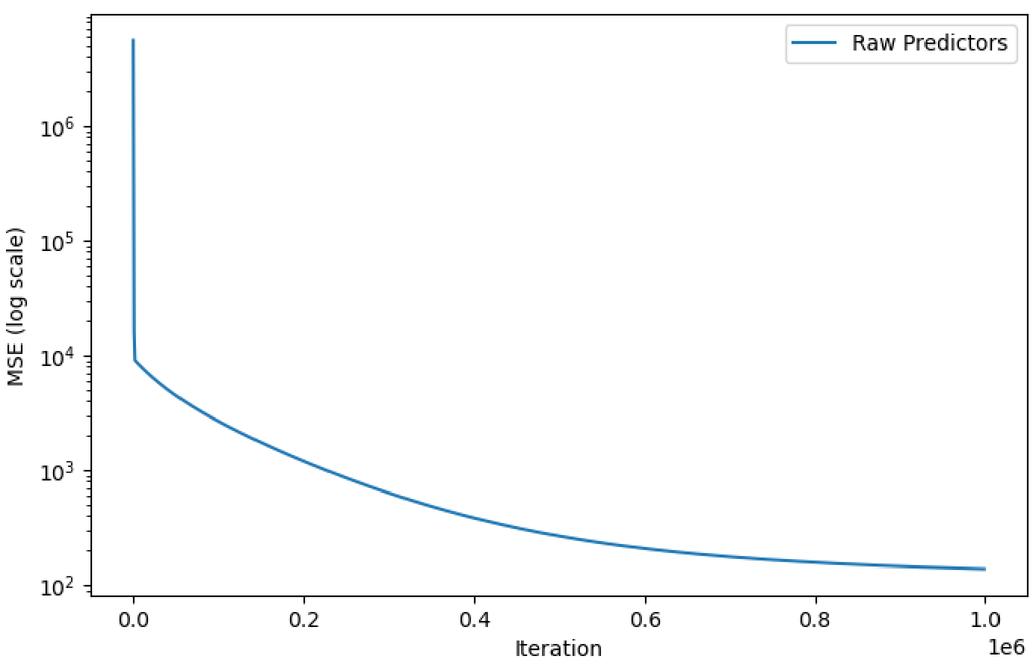
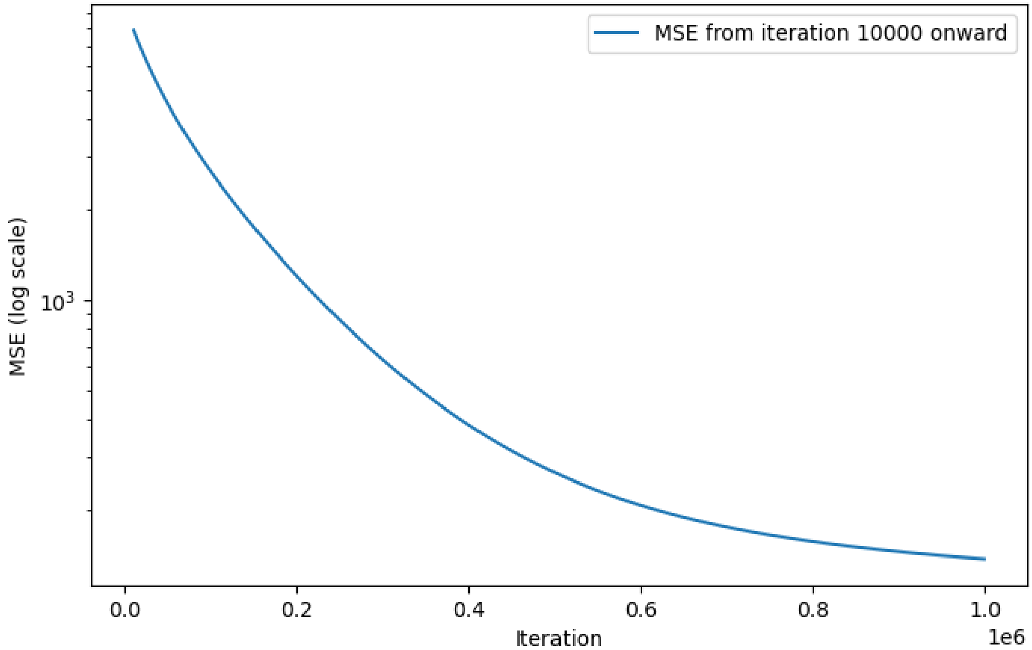


Figure 4

MSE Loss Curve for Raw Predictors from Iteration 10000 Onward



Regression Analysis:

- p-Values:

Table 1

Multivariate Model with Raw Predictors

Feature	p-Value
Cement	3.30E-39
Blast Furnace Slag	1.86E-25
Fly Ash	7.27E-13
Water	1.91E-03
Superplasticizer	2.36E-01
Coarse Aggregate	5.99E-03
Fine Aggregate	3.34E-03
Age (day)	6.98E-74

Table 2

Multivariate Model with Normalized Predictors

Feature	p-Value
Cement	3.30E-39
Blast Furnace Slag	1.86E-25
Fly Ash	7.27E-13
Water	1.91E-03
Superplasticizer	2.36E-01
Coarse Aggregate	5.99E-03
Fine Aggregate	3.34E-03
Age (day)	6.98E-74

Table 3

Multivariate Model with Log-transformed Predictors

Feature	p-Value
Cement	3.79E-84
Blast Furnace Slag	2.37E-35
Fly Ash	3.42E-01
Water	1.17E-20
Superplasticizer	7.13E-06
Coarse Aggregate	4.21E-01
Fine Aggregate	7.98E-02
Age (day)	8.69E-193

4. Discussion

Interpretation:

- Using both gradient descent and regression analyses, Cement, Age and Blast Furnace Slag consistently appear as the strongest indicators of concrete compressive strength, as indicated by their low p-values and large coefficients in the models using normalized predictors.

- Water and Superplasticizer show moderate influence as their effects can vary depending on whether raw, normalized, or log-transformed predictors are used.

- Fly Ash, Coarse Aggregate, and Fine Aggregate have weaker or more variable effects on compressive strength, suggesting that their impact may depend on interactions with other components or non-linear relationships.

- The log-transformed predictors in the multivariate regression produced the highest variance explained and the lowest MSE, indicating that transforming skewed variables improves model performance.
- The MSE loss curves show that normalized predictors converge faster and more smoothly compared to raw predictors.

Strengths and Limitations:

- Strengths:
    - o Normalization allowed gradient descent to converge efficiently and improved interpretability of the models.
    - o The Multivariate models produced positive variance explained values on the testing data, indicating that the models were able to capture underlying relationships in the data.
    - o The multivariate models provided consistent identification of key predictors across different approaches, supporting the idea that these variables are influential, regardless of preprocessing.
- Limitations:
    - o Nearly all univariate models yielded a negative variance explained on testing data, showing that individual features are poor predictors of compressive strength.
    - o The multivariate model using gradient descent and normal predictors and the regressions using normal and raw predictors showed a steep drop-off in variance explained between training and testing, which is a strong indicator of overfitting.
    - o Substantial discrepancies between raw, normalized, and log-transformed models suggest that predictor interpretations depend heavily on the modelling approach.
    - o The linear modeling framework may not fully capture non-linear effects or interactions, which could restrict interpretability.

5. Appendices

https://github.com/tiff123poof/DataMiningProject1.git