Supervised Learning Mid Project

Mecchia Alessandro

November 12, 2024

1 Dataset Analysis and Preparation

1.1 Dataset Exploration

The dataset has 600 instances with missing values in 'Daily Calories Consumed', 'Daily Caloric Surplus/Deficit', 'Weight Change', and 'Work Sector'. Features differ in scale, so scaling is necessary. 'Daily Calories Consumed' is skewed, and implausible values were found in 'Age'.

1.2 Handling Missing Values

Missing values were imputed using kNN to maintain relationships between features without data loss.

1.3 Encoding and Scaling

- **One-Hot Encoding**: Applied to 'Smoking', 'Gender', and 'Work Sector'.
- **Ordinal Encoding**: Used for ordered features 'Sleep Quality' and 'Physical Activity Level'.
- **Scaling**: Standard scaling for normally distributed features, Min-Max for skewed features.

1.4 Dimensionality Reduction

I applied univariate (ANOVA), iterative (RFE), and model-based (Random Forest) methods. Consistently selected features, like **BMR** and **Daily Calories Consumed**, were retained. We chose **dataset_reduced_combined** for its balance between reduced dimensionality and RMSE.

2 Linear Regression Models

2.1 Polynomial Model (With and Without BMR)

The polynomial model (degree 2) with BMR performed better on the test set than the model without BMR. The additional complexity of the polynomial model allowed it to capture more intricate relationships, narrowing the gap between training and test performance.

2.2 Learning Curves for Batch Gradient Descent

With BMR, batch gradient descent used a lower learning rate and took more iterations to reach optimal performance, achieving a low, stable RMSE. The model without BMR converged faster but had a higher final RMSE, indicating less accurate data adaptation.

2.3 Mini-Batch Gradient Descent

With BMR, mini-batch gradient descent achieved a stable, low RMSE within 1000 iterations, demonstrating strong generalization. Without BMR, the model converged more slowly and with a higher RMSE, reflecting reduced predictive accuracy.

3 Impact of Training Set Size (With and Without BMR)

3.1 Standard Linear, Batch Gradient and Mini-Batch

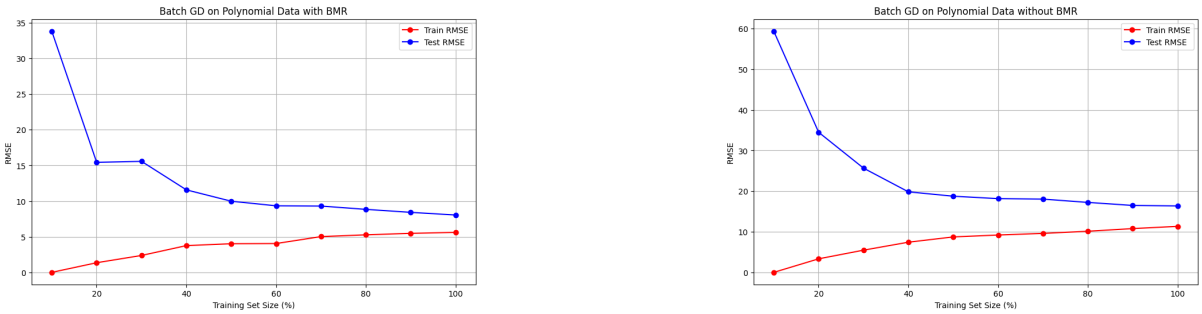
- **With BMR**: Minimal train-test RMSE gap, indicating good generalization.
- **Without BMR**: Larger gap, suggesting underfitting due to limited data complexity capture.

3.2 Polynomial Model with BMR Augmentation

With BMR, both train and test RMSE decrease as the training set size grows, with test RMSE stabilizing around 5 and train RMSE around 3, reflecting effective data pattern capture.

3.3 Polynomial Model without BMR

Without BMR, a significant gap between train and test RMSE persists, indicating limited generalization.



(a) With BMR

(b) Without BMR

Figure 1: Batch Gradient Descent on Polynomial Data with and without BMR

4 Regularization Analysis

4.1 Lasso Regularization (With and Without BMR)

Lasso with BMR performed well, emphasizing BMR's importance. Moderate alpha values minimized test RMSE, balancing feature selection with performance. Without BMR, the model's performance dropped, underscoring BMR's stabilizing role.

4.2 Ridge Regularization (With and Without BMR)

Ridge with BMR improved generalization at higher alpha values, balancing overfitting and robustness on test sets. Without BMR, Ridge still performed better than other models without BMR, though it underperformed relative to Ridge with BMR.

5 Gender-Based Analysis

5.1 Learning Curves and Model Performance by Gender

- **Rapid Convergence**: Both male and female models stabilized near zero MSE within 20 iterations.

- **Similar Performance Across Genders**: Nearly identical curves suggest equal predictive value for both genders.

5.2 Learning Rate Comparison for Gender Models

- **Optimal Eta at 0.03**: An η of 0.03 achieved rapid MSE reduction for both genders.

5.3 Effect of BMR on Gender-Specific Models

Including BMR significantly lowered RMSE and increased R^2 for both genders. Models without BMR struggled to capture variance effectively, confirming BMR's importance in weight prediction accuracy across genders.

6 Performance Comparison by Gender

I compared gender-stratified models (Task 4) with the global approach (Task 3) on **dataset_reduced** to evaluate predictive accuracy across genders.

6.1 Performance Summary with and without BMR

Without BMR, all models had higher RMSE and lower R^2, underscoring BMR's importance for accurate predictions and variance explanation.

6.2 Model Insights

- **BMR's Impact**: Including BMR nearly halved RMSE and significantly boosted R^2.
- **Top Model**: Linear Regression with BMR showed the lowest RMSE and highest R^2, especially for females.
- **Batch Gradient Descent**: Effective with BMR, though slightly behind Linear Regression.
- **Best Gender-Specific Models**: Female Linear Regression and Mini-Batch Gradient Descent for both genders performed best.
- **Worst Model**: Batch Gradient Descent without BMR performed weakest.

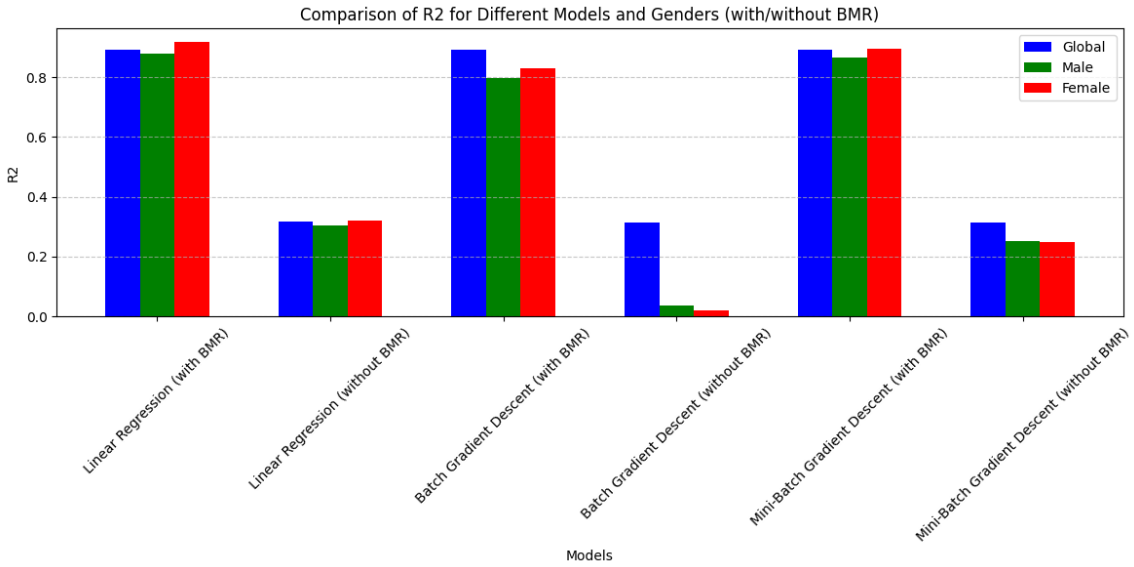


Figure 2: Overall Model Performance across Different Configurations

6.3 Conclusion

BMR is essential for reliable predictions across both genders. Linear Regression with BMR was the top performer, with Gradient Descent models also benefiting from BMR.