Report
Midterm Project

Supervised Learning - A.Y. 2024-2025

Carlo Schillaci

12 November 2024

1. Introduction

The project aims to build predictive models to estimate the final weight of subjects participating in a dietary study. The dataset includes variables on demographic details, caloric intake, physical activity, and stress levels. The analysis explores the dataset, addresses quality issues, applies linear regression, and compares model performances under various conditions.

2. Data Exploration and Quality Assessment

-Dataset Overview:

The dataset contains records of participants in a dietary study with variables including demographics (e.g., age, gender), lifestyle factors (e.g., smoking status, physical activity), physiological measures (e.g., BMR, caloric intake), and outcome variables related to weight change. Key target variable is "`Final Weight (lbs)`," which this project aims to predict.

-Data Quality Analysis:

Missing Values:

The dataset had several missing values, notably in `Gender`,`Daily Calories Consumed`, `Daily Caloric Surplus/Deficit`, `Weight Change (lbs)`**.** Since `Gender` is essential for later stratified analyses, an imputation strategy was necessary. For non-critical categorical variables (like `Smoking` and `Work Sector`), straightforward replacements like "No" (for smoking) and "Unknown" (for missing work sector) were applied to simplify preprocessing without introducing bias.

-**Data Cleaning and Imputation**:

Missing `Gender` values were imputed using KNN and regression-based imputation methods, comparing their distributions to maintain consistency with the original data. For `Daily Caloric Surplus/Deficit`, since it represents the difference between `Daily Calories Consumed` and `BMR`, missing values were calculated directly, reducing the number of missing entries significantly.

A hybrid imputation approach was used for other missing numerical values, balancing accuracy and preserving data integrity, given the dataset's modest size.

-Outlier Detection and Feature Analysis:

Key features were visually inspected using histograms and box plots to identify skewness and outliers.

Age had some unrealistic values (e.g., negative values and overly high ages) that were removed to improve data integrity.

Outliers in features like `Daily Calories Consumed` and `Weight Change (lbs)` were retained initially, as extreme values can be relevant for modeling weight-related behaviors. Removing them would risk biasing the model and reducing its ability to generalize to diverse individuals.

3. Data Preparation for Linear Regression

The dataset's numerical features were analyzed for skewness. Most followed a normal distribution but slight skewness was observed in some features, particularly Daily Calories Consumed and Weight Change (lbs). While these deviations were minor, scaling was deemed necessary to ensure consistency in the Linear Regression model due to the broad range of values across features.

-Numerical Feature Transformation

Since the features have varied ranges (e.g., age in years, BMR in thousands of calories), standardization was applied. Standardization centers each feature to a mean of zero and scales to unit variance, optimizing model convergence and preventing features with large ranges from dominating the regression.

-**Categorical Feature Encoding**

Each categorical feature was encoded based on its type and relevance to the model:

Binary Variables (Gender, Smoking): Encoded as 0 and 1 to retain interpretability and simplicity.

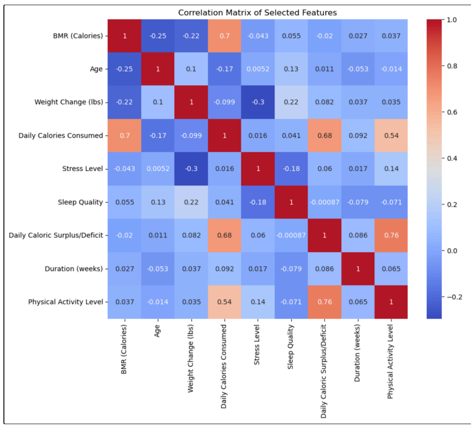
Ordinal Variables (Physical Activity Level, Sleep Quality): Ordinal encoding was applied to maintain the inherent order within these categories, enhancing the model's understanding of the relationships.

Nominal Variables (Work Sector): One-hot encoding was used to avoid imposing any ordinal relationship, preserving information without introducing bias.

-Feature Selection for Dimensionality Reduction & Correlation Analysis

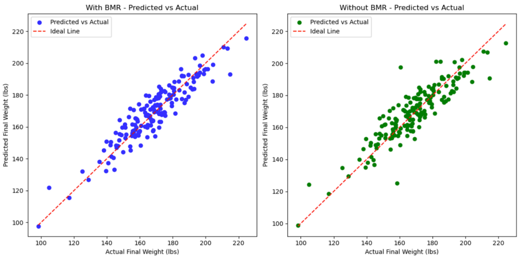
Three feature selection methods (ANOVA, RFE and Model-Based Selection) were applied to identify the most influential variables to include on `dataset_reduced`, which will be exploited for the supervised task This feature selection ensures model simplicity while maintaining predictive power, focusing on factors most likely to impact the `Final Weight(lbs)` prediction.

The correlation matrix helps, other than providing a visual confirmation that the selected features have meaningful interactions. detect redundancy among selected features. If two features are highly correlated, they may provide overlapping information to the model and reveal indirect relationships with the target variable, like with `BMR (Calories)` and `Daily Calories Consumed.`



4. **Predictive Modeling(s) for Final Weight**

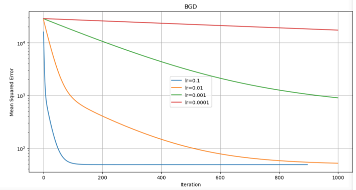
“The predictive models were developed under two scenarios: one with BMR information and another excluding BMR”



Linear Regression (Sklearn): Implemented for a baseline performance reference. The reason the model works without BMR, even though its an essential feature, is because `Daily Calories Consumed` (DCC) acts as a proxy for BMR. Their high correlation (0.7), while it might lead to Multicollinearity, still maximizes prediction accuracy, on top of providing complementary information.

In fact, even though DCC may act as a proxy, the model performs worse than it would if it had BMR too.

But generally, the tighter clustering suggests that the model has learned a strong, accurate relationship between the features and Final Weight. The high R^2 means that whether you have BMR or DCC, you can predict well the target feature, since more than 80% of the variance in the dependent variable is explained by the model



Gradient Descent Approaches:

a) Batch Gradient Descent: Optimized learning rate and iterations to ensure model convergence. At the start of training, the model’s weights are typically initialized to zero or small random values. Since the model has no learned information yet, its initial predictions will generally be poor.

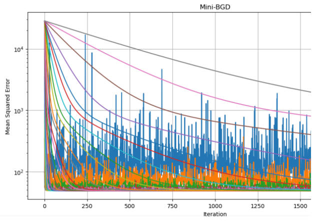
The model converges faster without BMR at the same learning rate (0.1), but the error is greater and the model is less accurate

b)Mini-Batch Gradient Descent: Used mini-batches for training to improve computational efficiency. Different batch sizes and learning rates were evaluated for convergence.

Medium Batch Sizes show moderate fluctuations and steady progress toward lower error levels, with faster convergence than large batch sizes.

Best results, with and without BMR, are given by a high l.r. (0.1) and med. batch sizes (64 and 16).

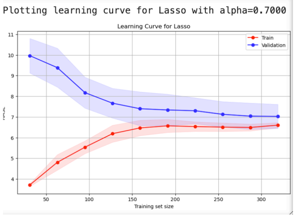
The model remains more accurate and stable with BMR.



BGD with Polynomial Features: it considers the complete dataset instead of the reduced one, with the addiction of interaction and polynomial features (dataset_augmented). Increasing the number of parameters makes it harder for the model to converge within a reasonable number of iterations, leading to overfitting.

5. **Regularization Models**

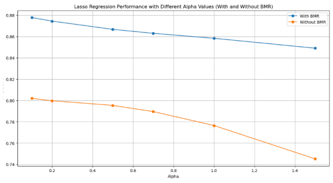
Exploiting the augmented dataset, we want to improve model generalization, reduce overfitting, and handle multicollinearity among features, and for that we can use, exploring varying α values, regularization:



```
Plotting learning curve for Lasso with alpha=0.7000
```

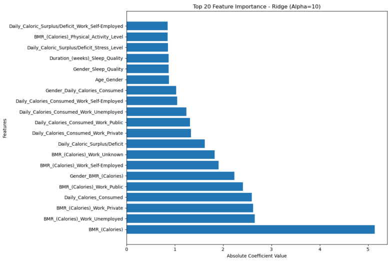
Lasso Regression: Assessing feature importance and model performance. Lasso’s sparsity helped identify the most critical predictors. Alpha=0.8 strikes a balance by retaining meaningful features and interactions while keeping high accuracy, simplifying the model and potentially improving generalization.

Without BMR, the model shifts focus to features associated with caloric intake and caloric balance. This shift highlights how Lasso adjusts feature importance depending on available information.



Again, the results are consistent with the other models above. Daily Caloric Surplus/Deficit and Daily Calories Consumed become critical when BMR isn’t available, indicating that these caloric metrics are directly tied to weight management.

Ridge Regression: for robust generalization. Provides more stable predictions. Ridge doesn’t zero out features like Lasso; instead, it makes less important features smaller but keeps them in the model. With a high alpha like 10 we can see that the training RMSE is consistently low, the regularization is strong enough and the model becomes less sensitive to variations in the training data, leading to a stable training curve. The model generalizes well since the gap is pretty narrow still. Like in the Lasso, the importance of BMR is substituted by Daily Calories when missing. What is suprising is that now the Work variable gained much influence as an interaction term paired with the main influential features.



The categorical values of the Work feature in dataset_reduced were broadly categorized, without details about the specific nature of the job. In line with this, different feature selection methods did not highlight Work categories as highly important. However, Ridge regression, which retains all features and explores interactions, assigned some importance to interactions . These interactions might reflect indirect correlations rather than true lifestyle differences tied to work environment. Therefore, while Work interactions appeared moderately relevant in the Ridge model, these findings should be viewed cautiously.

6. **Gender-Stratified Analysis**

Two linear regression models (mini-batch)were developed independently for male and female subjects, focusing on whether stratifying by gender would improve performance.

Including BMR improved model performance, significantly for women, less for men

Men achieved a higher R² overall, suggesting that the included features (especially BMR) may be more predictive of final weight in males than in women under these conditions.

Men tended to perform well with smaller batch sizes and moderate learning rates, while women saw improvements with larger batch sizes, especially when BMR was included. The variability in optimal batch sizes might reflect different noise tolerances in each gender's data structure.

7. **Comparative Performance Analysis**

The gender-stratified models (task 4) were compared to the global model (task 3a, 3b, 3c) using test data for each gender. The models were evaluated using R² scores and RMSE.

Outcome: The results are fairly consistent across both gender-specific and global models, with Task 4 showing more variance between male and female results compared to the global models. This consistency reaffirms that all models perform relatively well, but some variations exist when BMR is included or excluded, and based on the batch approach used.

Impact of BMR:

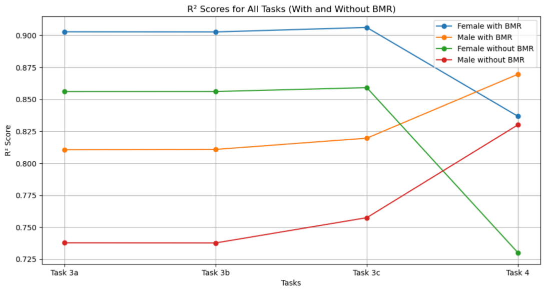
Across both genders, models that include BMR consistently show higher R² scores and lower RMSE compared to those without BMR. This confirms that BMR is a significant predictor of final weight, and excluding it reduces the predictive power of the models. The improvement with BMR is more pronounced for women, as seen by the sharper drop in R² and the increase in RMSE when BMR is excluded.

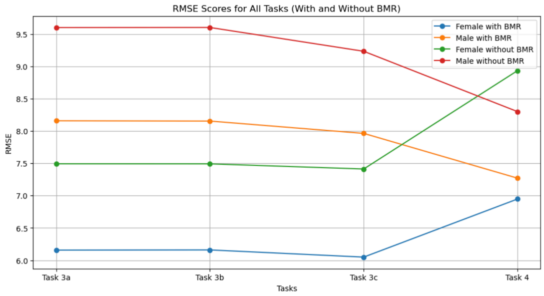
For men, the models perform relatively well even without BMR, especially in approaches like Task 4.

Comparison Across Approaches

Tasks 3a, 3b, and 3c display minor performance variations, indicating that these 'global approaches' are comparably effective for predicting weight outcomes. Among these, Task 3c (Mini-Batch Gradient Descent) demonstrates slightly better results with faster convergence and reduced noise, as evidenced by marginally improved R² and RMSE scores.

In contrast, Task 4 models, which are gender-stratified, show more variability in R² and RMSE scores between genders. This difference could suggest that Task 4's approach, potentially involving more complex feature interactions or specific gender nuances, has greater sensitivity to the inclusion or exclusion of BMR. When BMR is absent, Task 4 models seem to struggle more with generalization, particularly for women predictions, as shown by increased RMSE and decreased R² scores.





8. **Conclusion**

The importance of BMR

When we included BMR, the models performed slightly better. BMR is a key factor that reflects a person's metabolism—how many calories they burn naturally. Including it makes a lot of sense since it directly impacts weight. It helped both global and stratified models predict weight more accurately.

Without BMR, the models still worked, but they weren't as accurate."Daily Calories Consumed" (DCC) replaced BMR's importance as a feature thanks to the high-correlation between the two features. However, the model's predictions were generally less precise.

Complexity of the prediction

The similarity in results across methods suggests that the dataset is best suited to a straightforward linear approach. The relationships between features and the target are well-captured by a simple linear model, making advanced methods unnecessary in this case. This reinforces the robustness of the model and the dataset's linear characteristics, allowing us to choose the most efficient modeling approach without sacrificing performance.

Both Lasso and Ridge identified caloric metrics as crucial predictors when BMR is missing, validating the adequacy of a linear approach for capturing these core relationships without requiring complex modeling.

Which model is the best

A global linear model with Mini-Batch Gradient Descent is optimal, particularly with BMR included, as it balances accuracy and efficiency without introducing unnecessary complexity, especially if the dataset gets expanded with more data points. Gender-specific models add value for deeper insights but should be used with caution when feature availability is limited.