Predictive Models Developments For Calorie Consumption

Qiyang Wang, Andrew Cao, Yiru Fang, Dichuan Zheng

December 9, 2024

1 Abstract

This study utilizes a comprehensive dataset that captures gym members' exercise routines, physiological attributes, and fitness metrics. The dataset contains 973 entries, including key performance indicators such as heart rate, calories burned, workout duration, and body composition metrics like BMI and fat percentage. Additionally, it incorporates demographic data (e.g., age, gender) and behavioral factors (e.g., workout frequency, water intake), enabling an in-depth analysis of the relationships between exercise habits and calorie expenditure.

The study aims to provide actionable insights into how various factors influence workout intensity, endurance, and overall health. To achieve this goal, statistical testing is conducted to identify the most significant variables in the dataset. A practical machine learning algorithm was developed, which is capable of making predictions for a larger population based on these significant variables. The findings not only facilitate personalized fitness recommendations but also contribute to a broader understanding of fitness trends and strategies for optimizing individual health outcomes.

During the prediction stage, supplementary modeling is incorporated. To address potential multicollinearity and improve model stability, three regularized regression techniques are employed: Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Ridge Regression, and Elastic Net Regression. These methods apply penalties to regression coefficients to prevent overfitting and mitigate the effects of correlated predictors.

2 Data Background

Our dataset was retrieved from Seyed Vala Khorasani through Kaggle. The data have 973 observations and 15 variables. The response is the amount of calories burned. The other variables are predictors. The variables are described in the following:

1. **Age**: the age of the individuals in years.
2. **Gender**: the gender of the individuals, classified as "Male" or "Female".
3. **Weight (kg)**: the weight of the individuals in kilograms.
4. **Height (m)**: the height of the individuals in meters.
5. **Max_BPM**: the maximum heart rate (beats per minute) recorded during a workout session.
6. **Avg_BPM**: the average heart rate (beats per minute) recorded during a workout session.
7. **Resting_BPM**: the heart rate (beats per minute) of the individuals at rest.
8. **Session_Duration (hours)**: the total duration of each workout session in hours.
9. **Calories_Burned**: the total number of calories burned during a workout session.
10. **Workout_Type**: the type of exercise performed, classified as "Yoga", "HIIT", "Cardio", or "Strength".
11. **Fat_Percentage**: the body fat percentage of the individuals.
12. **Water_Intake (liters)**: the total water intake of the individuals in liters.
13. **Workout_Frequency (days/week)**: the number of days per week that individuals engage in workouts.
14. **Experience_Level**: the self-reported experience level of the individuals, rated on a scale from 1 (beginner) to 3 (advanced).
15. **BMI**: the body mass index (BMI) of the individuals, calculated as weight (kg) divided by the square of height (m).

3 Algorithm

3.1 Debiased Polynomial Regression with a Transform

To develop a predictive model for **Calories_Burned**, a full model and a null model were used as starting points for stepwise selection to determine the optimal set of predictors. And then we introduce transformation terms and variables reduction.

$$\begin{aligned}\sqrt{\text{Calories\_Burned}} =& \beta_0 + \beta_1 \cdot \text{poly(Session\_Duration, 2)} + \beta_2 \cdot \text{poly(Avg\_BPM, 2)} \\ &+ \beta_3 \cdot \text{Gender} + \beta_4 \cdot \text{Age} + \beta_5 \cdot \text{Resting\_BPM} + \epsilon\end{aligned}$$

Steps

1. **Data Split**: The dataset was split into 70% training and 30% testing data for model evaluation.
2. **Full Model**: This model included all predictors and served as the upper limit for stepwise selection.
3. **Null Model**: This model only included the intercept, serving as the lower limit for stepwise selection.
4. **Stepwise Selection**:
   - **BIC**: Bayesian Information Criterion was used to select a simpler model by penalizing complexity.
   - **AIC**: Akaike Information Criterion was used to select a more flexible model by balancing fit and complexity.
5. **Model adjustments**: To improve the model's performance and address issues observed in the diagnostic plots, several transformations and higher-order terms were introduced. The square root of the response variable was applied to address skewness, while logarithmic transformation was used for predictors like Age to linearize relationships. Additionally, the square root of Resting BPM was included to reduce non-linear effects. Second-degree polynomial terms for Session Duration and Avg BPM were added to capture curvature in the data, ensuring a better fit and improved interpretability of the model.

Stepwise Selection Principle: This algorithm iteratively adds or removes predictors based on criteria like AIC or BIC to optimize model fit while avoiding overfitting.

AIC and BIC Formulas:

- **AIC**:

  $$\text{AIC} = 2k - 2\ln(\hat{L})$$

  where k is the number of parameters in the model and $\hat{L}$ is the maximized value of the likelihood function for the model.

- **BIC**:

  $$\text{BIC} = k\ln(n) - 2\ln(\hat{L})$$

  where k is the number of parameters, n is the number of observations, and $\hat{L}$ is the maximized likelihood.

3.2 Least Absolute Shrinkage and Selection Operator (LASSO) Regression

Definition: LASSO (Least Absolute Shrinkage and Selection Operator) adds a penalty L_1 to the regression loss function, shrinking some coefficients to exactly zero, effectively performing variable selection.

Formula:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Where:

- λ: Regularization parameter controlling the strength of the penalty.
- $\sum_{j=1}^{p} |\beta_j|$: L_1-norm penalty.

3.3 Ridge Regression

Definition: Ridge regression adds an L_2 penalty to the loss function, shrinking coefficients towards zero but not exactly zero. It is ideal for handling multicollinearity but does not perform variable selection.

Formula:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Where:

- $\sum_{j=1}^{p} \beta_j^2$: L_2-norm penalty.

3.4 Elastic Net Regression

Formula:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right\}$$

Where:

- α: Mixing parameter ($\alpha = 1$ for LASSO, $\alpha = 0$ for Ridge).
- Both L_1-norm and L_2-norm penalties are applied.

4 Model Optimization for Stepwise_AIC model

To develop a predictive model for **Calories_Burned**, we started with a full model, including all available predictors, and a null model, which only included the intercept. The dataset was divided into 70% training and 30% testing sets for model evaluation. Stepwise selection was applied to determine the optimal set of predictors, using AIC and BIC as criteria. AIC allowed us to balance model fit and complexity, potentially retaining more predictors for better performance, while BIC favored simpler models by penalizing complexity more heavily.

4.1 comparing AIC based model with BIC based model

4.1.1 Model Formula-Stepwise Selection Based on BIC

$$\begin{aligned}\text{Calories\_Burned} = &-820.8356 + 715.5878 \cdot \text{Session\_Duration\_hours} + 6.3692 \cdot \text{Avg\_BPM} \\ &+ 88.9493 \cdot \text{GenderMale} - 3.4951 \cdot \text{Age}\end{aligned}$$

The table below shows the coefficients estimated by the regression:

Predictor	**Coefficient**	**Std. Error**	**t-value**	**p-value**	**Significance**
Intercept	-820.8356	16.8076	-48.84	< 2e-16	***
Session_Duration_hours	715.5878	4.4116	162.21	< 2e-16	***
Avg_BPM	6.3692	0.1046	60.89	< 2e-16	***
GenderMale	88.9493	3.0310	29.35	< 2e-16	***
Age	-3.4951	0.1243	-28.13	< 2e-16	***

Table 1: BIC Regression Coefficients

Key Metrics

- **Residual Standard Error (RSE)**: 39.42
- **Multiple R-squared**: 0.9795
- **Adjusted R-squared**: 0.9794
- **F-statistic**: 8085 on 4 and 676 degrees of freedom

Interpretation: This model explains 97.95 % of the variance in Calories_Burned and identifies session duration, average BPM, gender, and age as significant predictors. It provides a reliable tool for estimating calorie expenditure and optimizing workout strategies.

4.1.2 Model Formula-Stepwise Selection Based on AIC

The AIC-optimized linear regression model for predicting `Calories_Burned` is:

$$\begin{aligned}\text{Calories\_Burned} = &-881.9137 + 715.9314 \cdot \text{Session\_Duration\_hours} + 6.3562 \cdot \text{Avg\_BPM} \\ &+ 85.1221 \cdot \text{GenderMale} - 3.4842 \cdot \text{Age} + 25.1815 \cdot \text{Height\_m} \\ &+ 0.3322 \cdot \text{Resting\_BPM}\end{aligned}$$

The table below shows the coefficients estimated by the regression:

Predictor	**Coefficient**	**Std. Error**	**p-value**	**Significance**
Intercept	-881.9137	32.2916	< 2e-16	***
Session_Duration_hours	715.9314	4.4036	< 2e-16	***
Avg_BPM	6.3562	0.1048	< 2e-16	***
GenderMale	85.1221	3.7317	< 2e-16	***
Age	-3.4842	0.1241	< 2e-16	***
Height_m	25.1815	14.4701	0.0823	.
Resting_BPM	0.3322	0.2121	0.1178	

Table 2: AIC Regression Coefficients

Key Metrics

- **Residual Standard Error (RSE)**: 39.32
- **R-squared**: 0.9797

Interpretation: The AIC optimized model explains 97.97 % of the variance in calories burned. Key predictors such as session duration, heart rate, gender, and age provide actionable insights for optimizing fitness plans.

- **Stepwise Model (BIC)**:
  - AIC: 6943.891
  - BIC: 6971.033
- **Stepwise Model (AIC)**:
  - AIC: 6942.609
  - BIC: 6978.797

Hence, we selected the AIC-based model because it prioritizes predictive accuracy, which aligns with our goal of building a reliable predictive model. Although the BIC-based model simplifies the model by penalizing complexity more heavily, this trade-off results in the omission of potentially relevant predictors, such as Height_m and Resting_BPM, which could improve the model's performance when making predictions on new data.

4.2 Transformation added by Diagnostic Plots

After selecting the initial model, we examined the Diagnostic Plots to evaluate key assumptions of linear regression, including linearity, normality of residuals, and homoscedasticity. The plots (See figure 1) indicated that the current model failed to fully satisfy the assumptions of linearity and normality. To address these issues, we applied transformations to the variables. Specifically, polynomial terms were used to improve the model's performance and adherence to these assumptions.

- **Log Transformation**: The response variable Session (Intercept, Session_Duration..hours, log(AVG_BPM), gender, age, log(Resting_BPM)) was transformed using the natural logarithm, $\log(Y)$, to stabilize variance and improve normality.
- **Square Root Transformation**: The predictor variable (Intercept, Session_Duration..hours, Avg_BPM, GenderMale, Age, Resting_BPM) was transformed using $\sqrt{X}$, as it displayed non-linear relationships with the response variable.
- **Quadratic Term**: The predictor variable (Intercept, Session_Duration..hours, I(Session_Duration..hours.$\hat{2}$), Avg_BPM, GenderMale, Age, and Resting_BPM) was transformed using X^2, as it also displayed non-linear relationships with the response variable.

- **Polynomial Terms**: Higher-order polynomial terms (X^2) were introduced for the predictor variables (Intercept, poly(Session_Duration..hours., 2), poly(Avg_BPM,2), Gender, Age, Resting_BPM) and $\sqrt{X}$ was introduced to predictor to capture non-linear relationships that were evident in the residual plots.

These transformations improved the model's adherence to linear regression assumptions, as confirmed by the revised Diagnostic Plots. The final model demonstrates better residual randomness, reduced heteroscedasticity, and improved normality of residuals, ensuring its reliability for prediction and inference.

The final model was developed after addressing influential points and systematically introducing transformations, including logarithmic (log), square root ($\sqrt{\cdot}$), and polynomial terms (poly), to address issues identified in the diagnostic plots and capture non-linear relationships. The best-performing model is expressed as:

$$\sqrt{\text{Calories\_Burned}} = \beta_0+\beta_1\cdot\text{poly}(\text{Session\_Duration}, 2)+\beta_2\cdot\text{poly}(\text{Avg\_BPM}, 2)+\beta_3\cdot\text{Gender}+\beta_4\cdot\text{Age}+\beta_5\cdot\text{Resting\_BPM}+\epsilon$$

where:

- $\sqrt{\text{Calories\_Burned}}$: Square root transformation of the response variable, applied to address heteroscedasticity and stabilize variance.
- $\text{poly}(\text{Session\_Duration}, 2)$: A second-degree polynomial term for `Session_Duration` to model its non-linear relationship with the response variable.
- $\text{poly}(\text{Avg\_BPM}, 2)$: A second-degree polynomial term for `Avg_BPM` to account for its non-linear effect.

This model effectively captures the relationships between predictors and the response variable, while ensuring compliance with linear regression assumptions. Diagnostic plots of the final model confirm improved residual randomness, normality, and homoscedasticity, indicating a robust and reliable model.

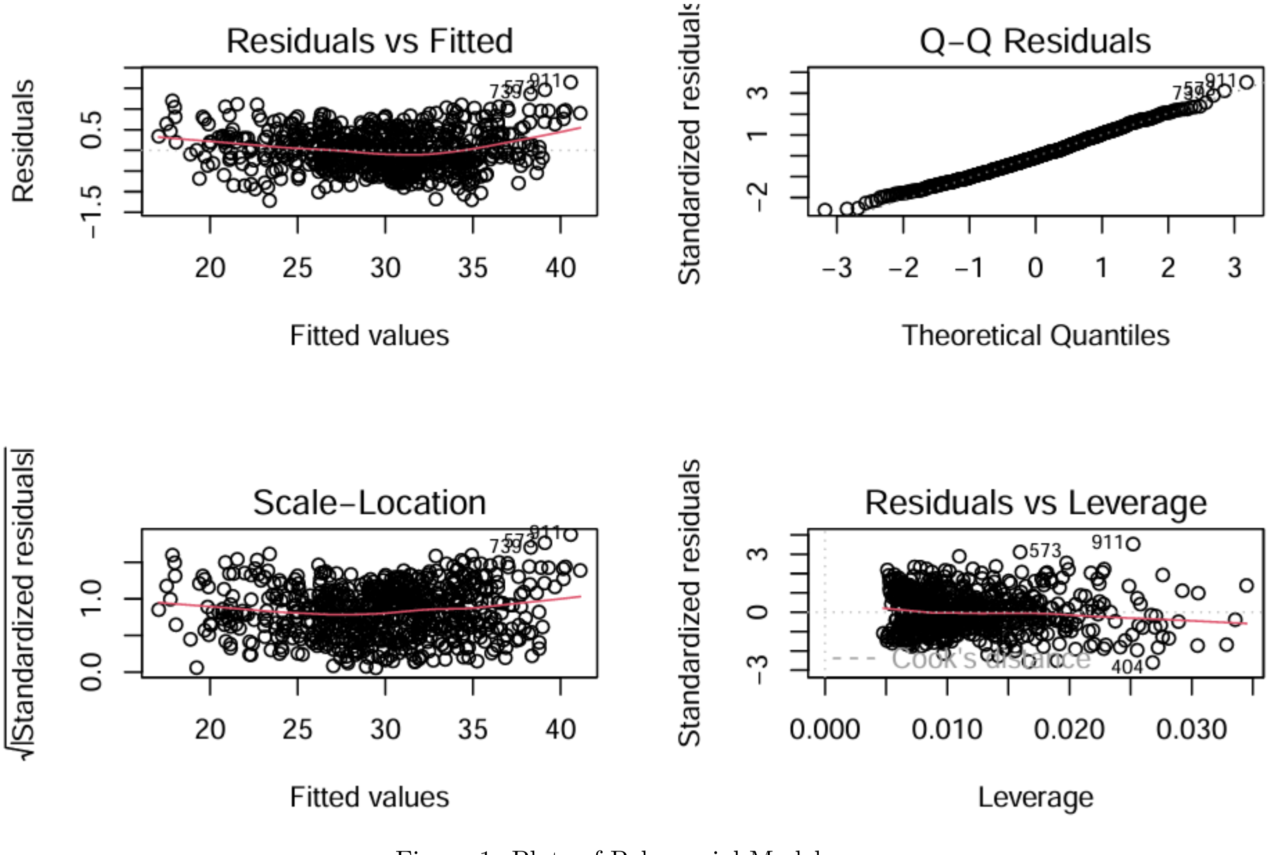

Figure 1: Plots of Polynomial Model.

4.3 Outlier and Influential Points

In transformation and diagnostic plots stage, we attempt linear, log, square root, quadratic, and polynomial transformation to the model. However, none of them reached our standard. (These attempts are shown in Appendix section 10). Hence, we realize that outliers and influential points can have a significant impact on the model's performance. Based on polynomial model, we implement diagnostic plots and leverage techniques such as Cook's distance to detect influential observations. To evaluate the impact of specific data points on the regression model, we analyzed Cook's Distance and Leverage Values. These diagnostics help to identify points that may unduly influence the model's estimates or predictions.

Our key findings include that influential points, such as 519, 518, 344, 278, and 284, exceed the Cook's Distance threshold, suggesting significant influence on the model. High leverage points, including 9, 70, 113, 209, and 519, are far from the average predictor values. Additionally, observations like 519, 344, and 278 are both influential and high leverage, making them critical for further review. However, these points may distort the model, leading to biased coefficients or poor predictions. Our next steps involve examining the data to determine whether these points are valid or represent errors and re-fitting the model to test the impact of including or excluding these points.

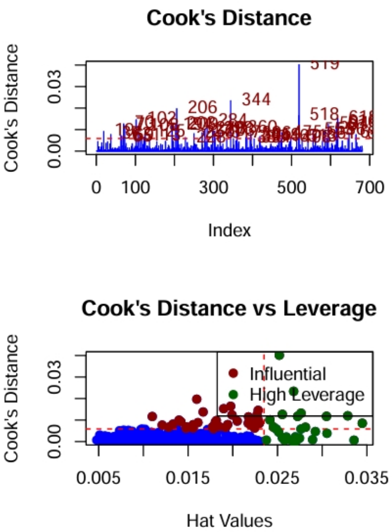

Figure 2: Plots of Cook's Distance, Leverage values, and Cook's Distance vs Leverage.

Interpretation

: These diagnostic plots indicate a well-fitted model. From these plots, we can see that the assumptions of linearity, homoscedasticity, normality of residuals, and influential points are met. Hence, we conclude that the model is capable for interpretation and prediction. To evaluate the model's performance, the Mean Squared Error (MSE) was calculated for both the training and test datasets. In which we received Training MSE: 0.1708961 and Test MSE: 0.2630758. This indicates the model is in a good performance.

- **Residuals vs. Fitted Values**: The residuals now show a more random scatter around the zero line, indicating that the linearity assumption has improved.
- **Q-Q Plot**: The standardized residuals align more closely with the theoretical quantiles, suggesting that the normality of residuals has improved.
- **Scale-Location Plot**: The spread of standardized residuals is more consistent across fitted values, addressing previous issues with heteroscedasticity.
- **Residuals vs. Leverage**: Fewer high-leverage points are present, and Cook's distance flags no influential points in the refitted model, indicating that the model is more robust.

By comparing these plots to those of the model that included outliers, we observe a significant improvement in the model diagnostics. The removal of influential points and the introduction of variable transformations have ensured better compliance with the assumptions of linear regression, enhancing the reliability and predictive accuracy of the model.

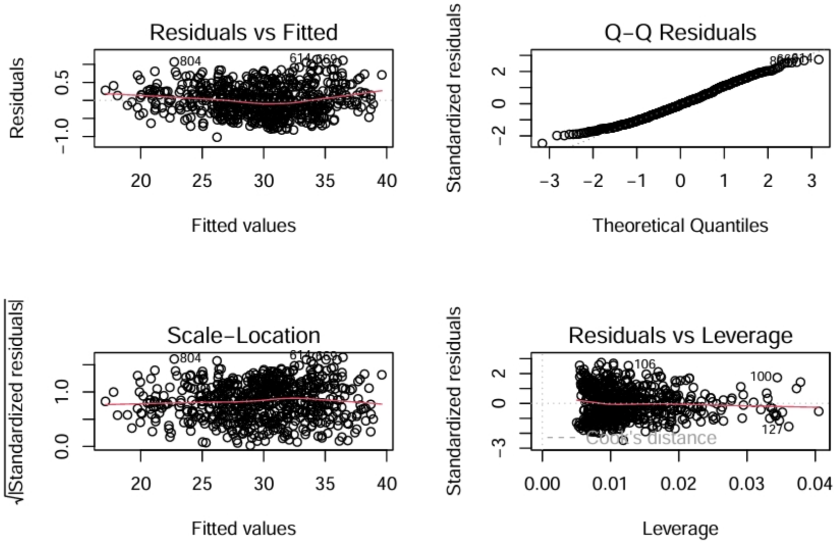

Figure 3: Plots of Removing Influential Points and Re-fit the Model.

5 Result

5.1 Debiased Polynomial Regression with a Transform

The final model was built after addressing influential points and including quadratic terms to capture non-linear relationships. The model formula is:

$$\begin{aligned}\sqrt{\text{Calories\_Burned}} = &\beta_0 + \beta_1 \cdot \text{poly(Session\_Duration, 2)} + \beta_2 \cdot \text{poly(Avg\_BPM, 2)} \\ &+ \beta_3 \cdot \text{Gender} + \beta_4 \cdot \text{Age} + \beta_5 \cdot \text{Resting\_BPM} + \epsilon\end{aligned}$$

The coefficients of the final model, along with their standard errors, t-values, and significance levels, are shown in Table 3.

Table 3: Estimated Coefficients of the Final Model

Variable	**Estimate**	**Std. Error**	**t-value**	**p-value**	**Significance**
Intercept	31.136	0.156	200.116	$< 2 \times 10^{-16}$	***
poly(Session_Duration, 2)1	95.539	0.417	228.837	$< 2 \times 10^{-16}$	***
poly(Session_Duration, 2)2	-8.590	0.418	-20.544	$< 2 \times 10^{-16}$	***
poly(Avg_BPM, 2)1	37.033	0.419	88.348	$< 2 \times 10^{-16}$	***
poly(Avg_BPM, 2)2	-1.566	0.417	-3.756	0.000189	***
GenderMale	1.417	0.033	42.705	$< 2 \times 10^{-16}$	***
Age	-0.056	0.001	-41.770	$< 2 \times 10^{-16}$	***
Resting_BPM	0.002	0.002	0.816	0.415	

5.2 Prediction Results and Model Evaluation

The final polynomial regression model achieved the following Mean Squared Error (MSE) values:

- **Training MSE:** 0.1708961
- **Test MSE:** 0.2630758

Interpretation: The training MSE of 0.1708961 indicates that the model fits the training data well. However, the test MSE of 0.2630758 is notably higher, suggesting that the model may be overfitting the training data. This is likely due to the complexity introduced by the polynomial terms (`Session_Duration` and `Avg_BPM`), which capture non-linear relationships but may reduce generalizability to new data.

While the predictors such as `Session_Duration` and `Avg_BPM` were significant (p-values < 0.001), as shown in Table 3, the model's performance on unseen data could benefit from techniques that balance complexity and generalizability.

To address potential overfitting and improve test performance, regularization techniques such as LASSO and Elastic Net can be applied. These methods shrink coefficients to prevent overfitting while retaining important predictors. Elastic Net, in particular, is well-suited for handling the multicollinearity present in this dataset, especially between `Session_Duration` and `Avg_BPM`.

Variance Inflation Factor (VIF): Table 4 summarizes the VIF values for the predictors in the final model. All GVIF are below 5, indicating moderate collinearity and a relatively stable model.

Table 4: Variance Inflation Factor (VIF) for Predictors

Variable	**GVIF**	**Df**	**GVIF**$^{1/(2*Df)}$
poly(Session_Duration, 2)	1.017347	2	1.004309
poly(Avg_BPM, 2)	1.020214	2	1.005016
Gender	1.005533	1	1.002763
Age	1.005985	1	1.002988
Resting_BPM	1.008961	1	1.004471

Interpretation: we can see that the GVIF are all below 5, indicating moderate collinearity. This suggests that the model is relatively stable and the predictors are not highly correlated. However, to further improve model performance and interpretability, we can explore regularization techniques such as LASSO, Ridge, and Elastic Net regression.

6 Model Advantages and Limitations

6.1 Advantages

The primary objective of our model is to explore the influence of various factors on `Calories_Burned` and improve the accuracy of predictions. The following advantages demonstrate the model's capabilities:

- **Predictive Power**: By incorporating transformations ($\sqrt{\cdot}$, $\log(\cdot)$) and polynomial terms, the model accurately captures complex relationships, improving predictions of `Calories_Burned`.
- **Comprehensive Factor Analysis**: The inclusion of predictors such as `Session_Duration`, `Avg_BPM`, `Gender`, `Age`, and `Resting_BPM` enables a detailed analysis of their individual contributions to energy expenditure.
- **Non-Linear Relationships**: Polynomial terms for key variables effectively capture non-linear patterns that are critical for understanding the factors influencing `Calories_Burned`.
- **Robustness**: Removing influential points and addressing outliers reduces the model's sensitivity to extreme values, ensuring more stable and generalizable results.
- **Improved Diagnostics**: The final model satisfies key regression assumptions (linearity, normality, and homoscedasticity), as confirmed by diagnostic plots.

6.2 Limitations

While the model provides valuable insights, it has certain limitations that must be acknowledged:

- **Generalizability to New Data**: The model is tailored to the current dataset, and its performance on new or unseen data may be limited if relationships between factors differ.
- **Interpretation Complexity**: Transformations and polynomial terms, while enhancing predictive performance, make it harder to interpret the direct effects of some predictors.
- **Residual Deviations**: Despite improvements, residuals show minor deviations from normality, especially in the tails, which could affect the accuracy of interval estimates.
- **Potential Overfitting**: With higher-order terms and multiple transformations, the model risks overfitting, particularly with smaller datasets.
- **Exclusion of Additional Factors**: The model focuses on the selected predictors but does not account for potential interaction effects or other unmeasured factors that may also influence `Calories_Burned`.

6.3 Potential Solution: Regularization Methods

Given the limitations of our current model, particularly the risk of overfitting and the lack of generalizability to new data, we propose exploring regularization techniques such as Ridge, Lasso, or Elastic Net regression. These methods introduce a penalty term to the loss function, which helps to constrain the model complexity and improve its performance on unseen data.

7 Considered Regularization Methods: LASSO, Ridge, Elastic Net

- **Ridge Regression**: Ridge adds an L_2 penalty ($\lambda \sum \beta_i^2$) to the loss function, which shrinks all regression coefficients towards zero. It is particularly effective in handling multicollinearity and reducing overfitting by discouraging large coefficients.
- **Lasso Regression**: Lasso incorporates an L_1 penalty ($\lambda \sum |\beta_i|$), which not only shrinks coefficients but also performs variable selection by driving some coefficients to exactly zero. This can result in a simpler, more interpretable model.
- **Elastic Net Regression**:This approach is particularly useful when predictors are highly correlated. The L_1-norm penalty ($\sum |\beta_j|$) promotes sparsity by selecting a subset of predictors (Lasso), while the L_2-norm penalty ($\sum \beta_j^2$) discourages large coefficients to improve model stability (Ridge). The mixing parameter α controls the balance between Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$). By blending these two methods, Elastic Net balances regularization and variable selection effectively, especially in datasets with multicollinearity.

These regularization methods can provide the following benefits:

- **Improved Generalizability**: By reducing the influence of overly complex relationships, these methods ensure better performance on new datasets.
- **Handling Multicollinearity**: Ridge and Elastic Net are effective when predictors are correlated, which can improve the model's stability.
- **Variable Selection**: Lasso and Elastic Net automatically select important predictors, simplifying the model while retaining predictive power.

In future work, we recommend these methods using cross-validation to evaluate their performance on prediction accuracy, interpretability, and generalizability. This analysis would help in identifying the most robust approach for predicting `Calories_Burned`.

7.1 Implementation of Regularization Methods

To investigate alternative approaches for improving model performance and addressing overfitting, we directly applied the `glmnet` package to implement three regularization methods: LASSO, Ridge, and Elastic Net regression. Without applying cross-validation, we obtained the results by fitting the regularization models to the data using default parameters. The optimal λ values and key features of each method are summarized in Table.

Table 5: Comparison of Model Properties

Model	**Optimal λ**	**Key Feature**
LASSO	0.00257	Shrinks coefficients to exactly zero
Ridge	0.09033	Shrinks coefficients, no selection
Elastic Net	0.00427	Combines LASSO and Ridge properties

7.2 Optimal λ Parameters Through Cross-Validation

The optimal lambda parameters were determined through cross-validation as follows:
Cross-validation is a widely used technique to evaluate and optimize model performance, particularly in scenarios where overfitting is a concern. In our analysis, we used cross-validation to determine the optimal λ parameter for each regularization method (LASSO, Ridge, and Elastic Net), ensuring that the models generalize well to unseen data.
Cross-Validation Implementation: Cross-validation is essential for evaluating model performance and selecting optimal hyperparameters, such as the λ penalty in regularization methods. It helps prevent overfitting by testing the model on unseen data and ensures better generalization to new datasets. Cross-validation works by splitting the data into k folds. The model is trained on $k - 1$ folds and validated on the remaining fold. This process is repeated k times, and the average validation error is used to select the optimal λ. This approach provides a robust estimate of model performance and helps balance complexity and accuracy.
By using cross-validation, we ensured that the λ values chosen for LASSO, Ridge, and Elastic Net provide a balance between model complexity and predictive accuracy. The results of cross-validation, including the optimal λ values

for 10 folds, are summarized in Table below.

Model	**Optimal λ**
LASSO	0.002344007
Ridge	0.09032699
Elastic Net	0.004688015

Table 6: Optimal Lambda Parameters for Models

7.3 Model Evaluation on Test Data

The steps for model evaluation are as follows: First, normalize the test data by scaling numeric variables to match the training data. Second, create a design matrix using `model.matrix()` to prepare the predictors in the same format as the training data. Third, predict and calculate the Mean Squared Error (MSE) for each model using the optimal λ:

- **LASSO** uses λ_{min} for sparse models.
- **Ridge** reduces multicollinearity impact without variable selection.
- **Elastic Net** balances LASSO and Ridge properties.

The MSE formula is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Finally, the MSE values for each model are compared to determine the best-performing approach.

```
## LASSO Test MSE: 0.02348011
## Ridge Test MSE: 0.03692044
## Elastic Net Test MSE: 0.02360708
```

8 Discussion

The study focuses on understanding how various factors influence Calories Burned and on building a model to predict it. Both LASSO (MSE = 0.02348011) and Elastic Net (MSE = 0.02360708) perform well with minimal difference, while Ridge regression (MSE = 0.03692044) performs slightly worse. Key insights reveal that Elastic Net is best suited for balancing prediction accuracy and capturing relationships among correlated factors like Session Duration and Avg_BPM. LASSO simplifies the model by selecting key factors, thereby helping identify the most influential variables on Calories Burned. In terms of recommendations, Elastic Net is preferred when prediction accuracy is the priority as it captures complex factor interactions while retaining key variables. However, if the primary goal is to identify the most critical factors affecting Calories Burned, LASSO offers a simpler and more interpretable solution. Future studies could explore interactions among predictors or advanced non-linear models to further enhance predictions.

9 Conclusion

This study aimed to explore the relationship between various factors and `Calories_Burned`, and to develop a predictive model under rigorous tests. The initial polynomial regression model effectively captured non-linear relationships and provided valuable insights into the influence of factors such as `Session_Duration`, `Avg_BPM`, and `Gender` on `Calories_Burned`. This model is particularly practical for understanding complex variable interactions and identifying significant predictors; however, its complexity introduced risks of overfitting and multicollinearity. To address these challenges, we applied LASSO and Elastic Net regularization. LASSO simplified the model by selecting key predictors, making it ideal for identifying the most critical factors influencing `Calories_Burned`, while Elastic Net balanced variable selection with multicollinearity handling, offering a robust framework for accurate predictions. The polynomial regression model is valuable for exploratory analysis and understanding complex relationships, whereas LASSO is preferred for interpretable models and factor prioritization. For robust prediction tasks, Elastic Net is better suited, particularly when predictors are correlated. Future work could validate these models on independent datasets, incorporate interaction effects, or leverage advanced machine learning methods to enhance prediction accuracy and model robustness. methods to further enhance prediction accuracy and model robustness.

10 Appendix

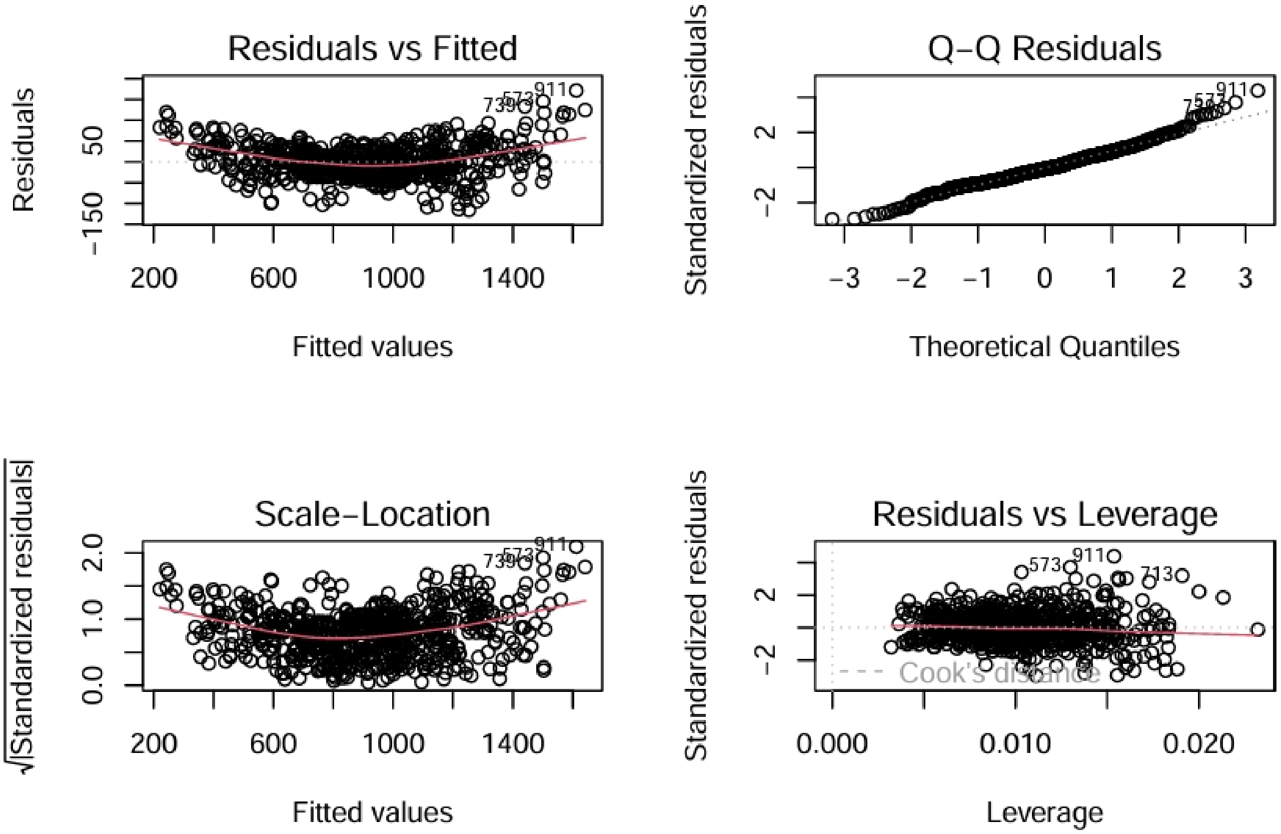

Figure 4: Diagnostic Plots Of Linear Regressionl.

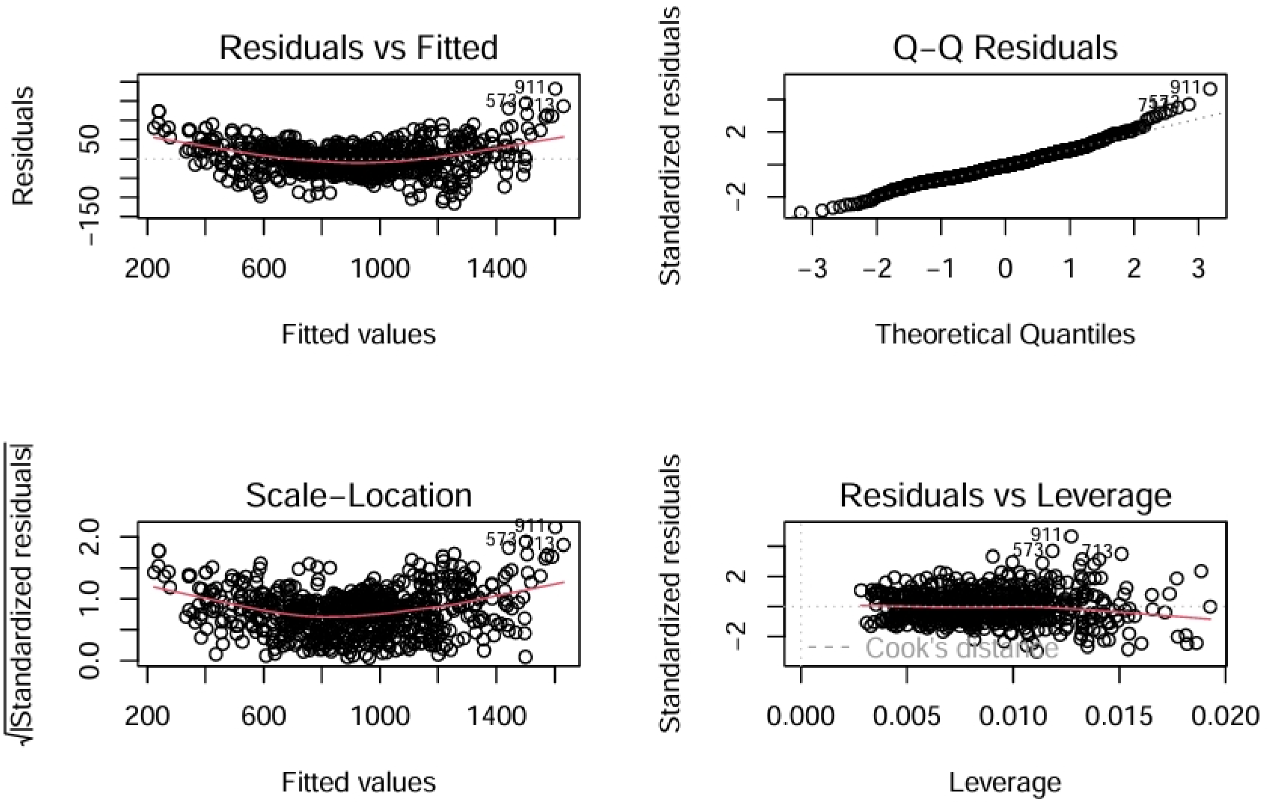

Figure 5: Diagnostic Plots Of Log Transformation.

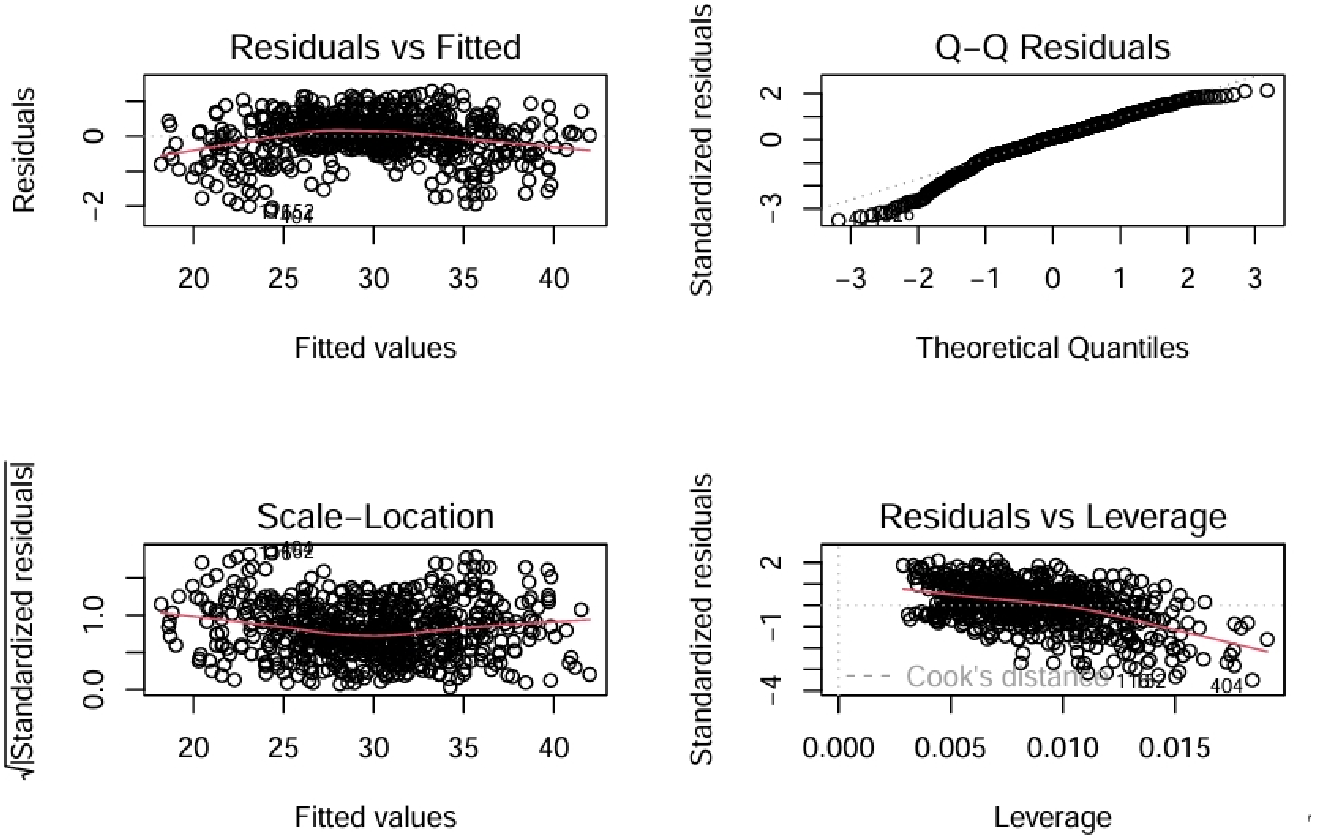

Figure 6: Diagnostic Plots Of Square Root Transformation.

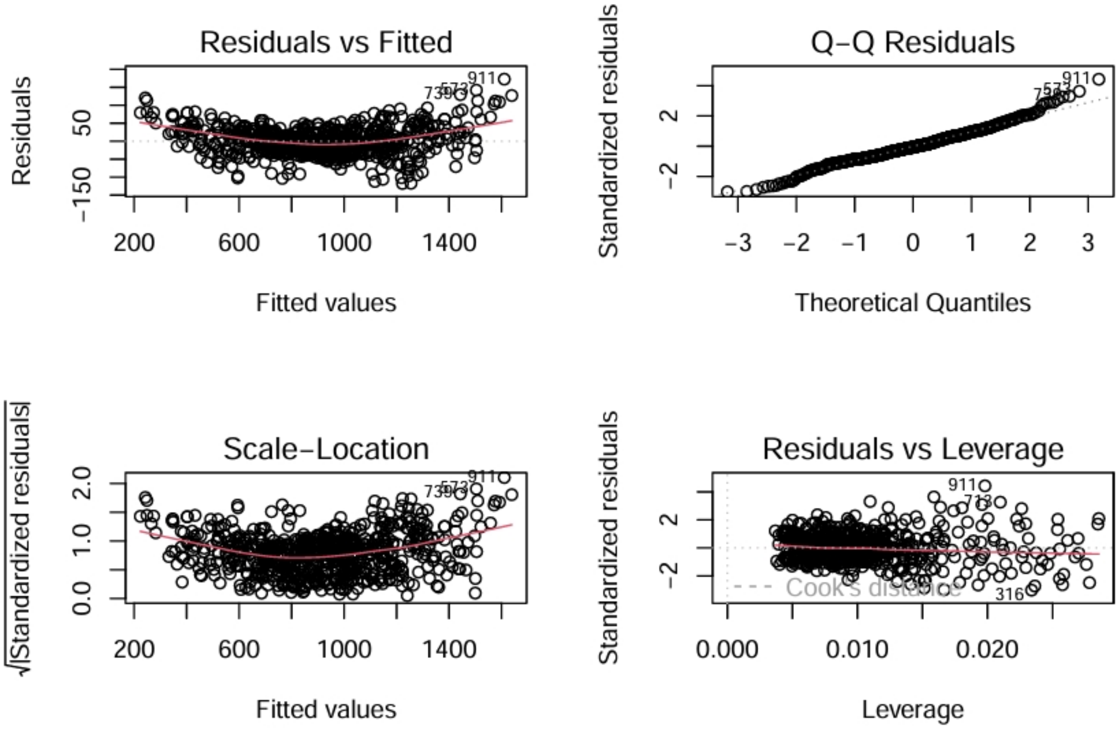

Figure 7: Diagnostic Plots Of Quadratic Transformation.