Predictive Models Developments For Calorie Consumption

Qiyang Wang, Andrew Cao, Dichuan Zheng ,Yiru Fang

STAT 527 Final Project

Dec 6, 2024

Agenda

- Background
- Data Exploring
- Model selection and improvement
- Further Exploration of the Models
- Analyzing Method & Result
- Conclusion

Background

- **Reason**: Healthy living and exercise have become increasingly popular among teenagers.
- **Curiosity**: Exploring novel insights into the relationship between calories burned and various influencing factors.
- **Objective**: Designing a unique approach to analyze the relationship between calories burned and other factors and developing a predictive model to provide actionable insights.

Algorithm

- ***Full Model and Null Model with Stepwise Selection Regression :***

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

- ***LASSO Regression :***

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

- ***Ridge Regression :***

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

- ***Elastic Net Regression***

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

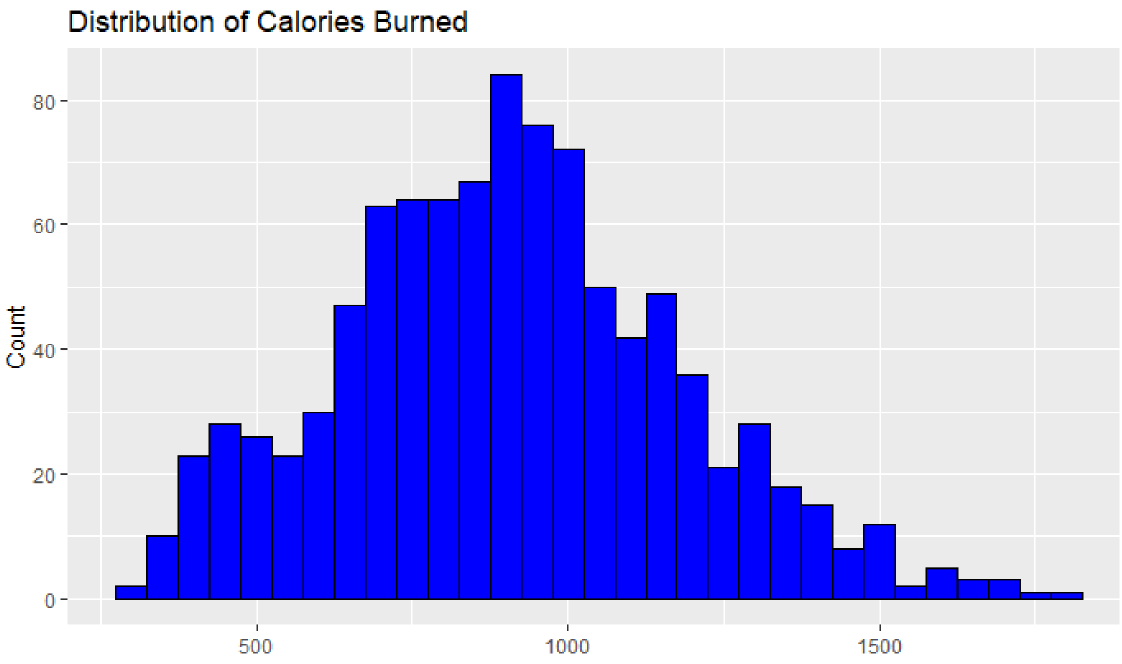
Data Background

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.

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).

Data Engineering



Analysis of Calories Burned

- Data is close to a normal distribution with slight right skew.
- Most values fall between 800–1400 calories.
- A few higher values suggest possible outliers or unique cases.

Insights and Next Steps

- Data is approximately normal, suitable for modeling.
- Check for potential outliers in higher calorie range.
- Proceed with model adjustments

Correlation Analysis

Height..m.	0.086348051	Max_BPM	0.002090016
Avg_BPM	0.339658667	Resting_BPM	0.016517951
Session_Duration..hours.	0.908140376	Calories_Burned	1.000000000
Fat_Percentage	-0.597615248	Water_Intake..liters.	0.356930683
Workout_Frequency..days.week.	0.576150125	Experience_Level	0.694129448
BMI	0.059760826		

Key Insights:

- Strong predictors: Session Duration (0.91), Experience Level (0.69).
- Moderate predictors: Workout Frequency (0.58), Water Intake (0.36).
- Negative predictor: Fat Percentage (-0.60).

Next Steps:

- Fit regression model including top predictors.
- Evaluate and refine feature importance.

AIC and BIC Criteria Used

- **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.

- ***AIC (Akaike Information Criterion)***: Allows for more flexibility by balancing goodness-of-fit and model complexity.

- ***BIC (Bayesian Information Criterion)***: Prefers simpler models by applying a stronger penalty for adding predictors.

Stepwise Selection

Key Steps:

- ***Data Split***:
  - 70% for training, 30% for testing.
- ***Models***:
  - ***Full Model***: Includes all predictors, serving as the upper limit.
  - ***Null Model***: Includes only the intercept, serving as the lower limit.
- ***Stepwise Selection***:
  - **Principle**:
    - Iteratively adds or removes predictors based on their contribution to the model's fit.
    - Balances model complexity with predictive performance.

Stepwise selected model-BIC

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

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

KeyMetrics:

- Residual Standard Error(RSE):39.42
- Multiple R-squared: 0.9795
- Adjusted R-squared: 0.9794

Stepwise selected model-AIC

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

The table below shows the coefficients estimated by the regression:

Predictor	**Coefficient**	**Std. Error**	**p-value**	**Significance**
Intercept	-881.9137	32.2916	$< 2e\text{-}16$	***
Session_Duration_hours	715.9314	4.4036	$< 2e\text{-}16$	***
Avg_BPM	6.3562	0.1048	$< 2e\text{-}16$	***
GenderMale	85.1221	3.7317	$< 2e\text{-}16$	***
Age	-3.4842	0.1241	$< 2e\text{-}16$	***
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

Why AIC Model Was Selected

Model Comparison:

- **AIC Model**:
  - AIC: **6942.609**
  - BIC: **6978.797**
  - Retains more predictors
- **BIC Model**:
  - AIC: **6943.891**
  - BIC: **6971.033**
  - Simpler but may omit relevant variables.

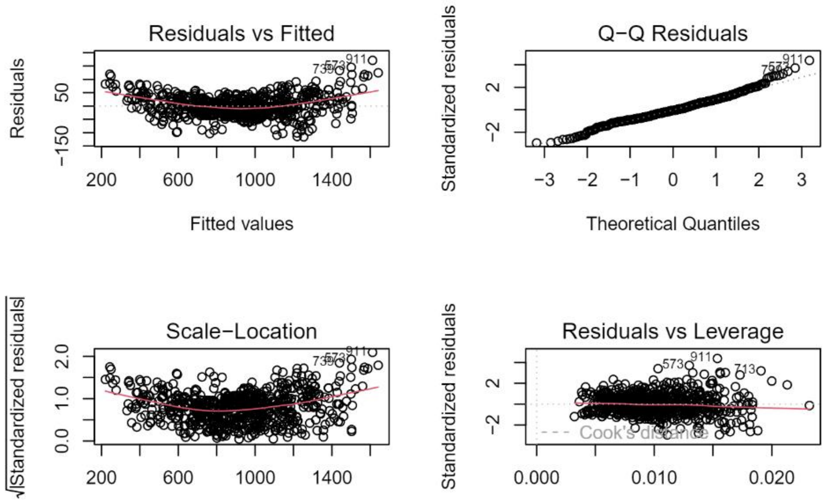
Why AIC?

- **Predictive Accuracy**: Better for new data, aligns with modeling goals.
- **Retains Key Variables**
- **Balance**: Optimizes fit while managing complexity.

Key Insights:

- AIC model explains **97.97% of variance** in calories burned.
- Top Predictors: **Session Duration**, **Heart Rate**, **Gender**, **Age**.

Selected Regression



Residuals vs Fitted:
- Slight curve suggests possible non-linear relationships.
- Consider adding non-linear terms or transformations.

Q-Q Plot:
- Residuals deviate at tails, indicating slight non-normality.
- Transform response variable if necessary.

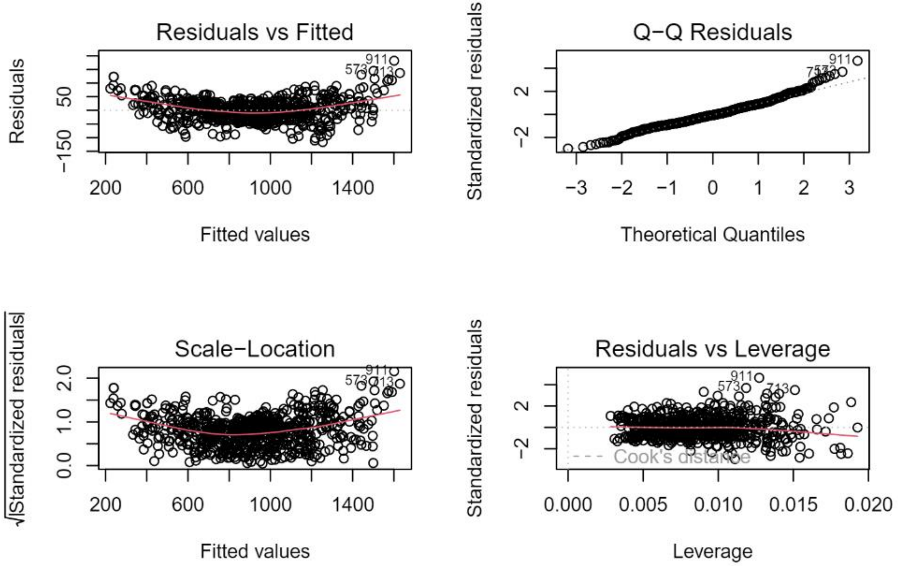
Scale-Location:
- Variance appears mostly constant (homoscedasticity).
- No major issues observed.

Residuals vs Leverage:
- Few high-leverage points (e.g., 757, 110).
- Investigate these points for potential influence.

Transformation added : Log Transformation, Square Root Transformation, Quadratic Term, Polynomial Terms

Log Transformation

Calories_Burned = -4508.068 + 715.524*Session_Duration..hours + 911.802*log(Avg_BPM) + 88.797 * Gender + -3.509 * Age + 18.963*log(Resting_BPM)



- **Residuals vs Fitted**:
Linearity slightly improved but deviations persist

- **Q-Q Plot**:
Better normality alignment, but tails show minor bias

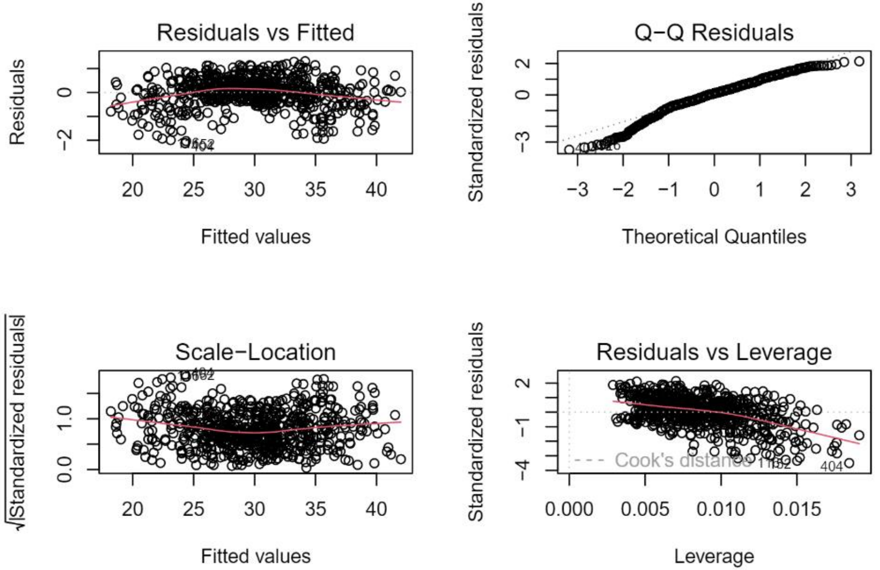
- **Scale-Location**:
Variance mostly constant, slight heteroscedasticity remains

- **Residuals vs Leverage**:
High-leverage points detected but within acceptable limits

- **Next step** :
Introduce **square root transformation** for response variable to further address residual bias and improve fit.

Square Root Transformation

Sqrt(Calories_Burned) = 0.7633 + 12.13*Session_Duration..hours + 0.1053*Avg_BPM + 1.466 * Gender + -0.05649 * Age + 9.879e-5*Resting_BPM



- **Residuals vs Fitted**:
Suggests partial **non-linearity** remains in the model

- **Q-Q Plot**:
Residuals align better with normality.

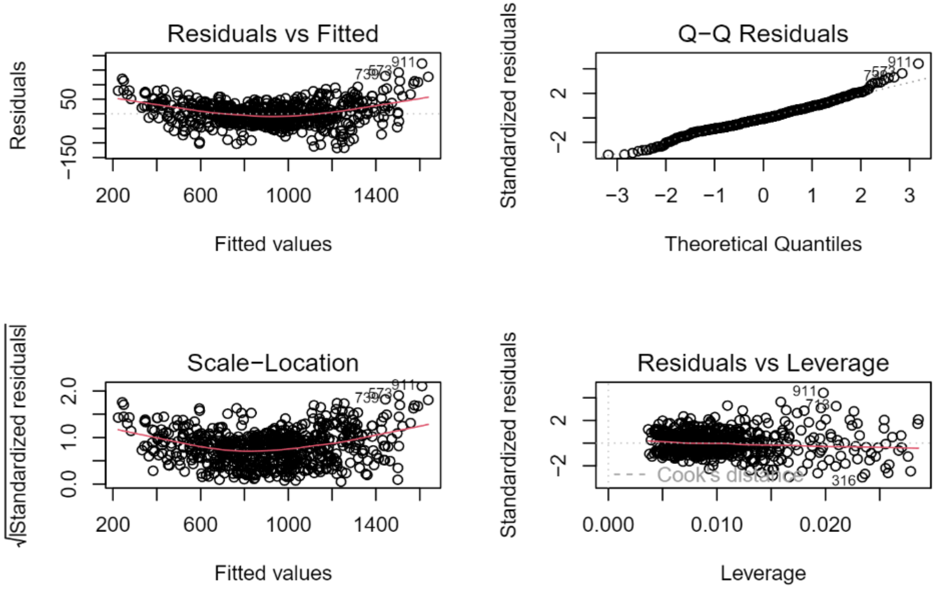
- **Scale-Location**:
Variance appears stable but not perfect

- **Residuals vs Leverage**:
High-leverage points are minimal and manageable

- **Next step** :
Introduce **quadratic terms** to address residual non-linearity

Quadratic Term

Calories_Burned = -830.5902 + 701.9309 * Session_Duration..hours + 5.4142 * (Session_Duration..hours^2) + 6.3581 * Avg_BPM + 88.9679 * Gender – 3.4956 * Age + 0.3104 * Resting_BPM



- **Residuals vs Fitted**:

Partial non-linearity remains despite quadratic term addition.

- **Q-Q Plot**:

Residuals closer to normal distribution but with slight deviation at tails.

- **Scale-Location**:

Variance is still not perfectly constant.

- **Residuals vs Leverage**:

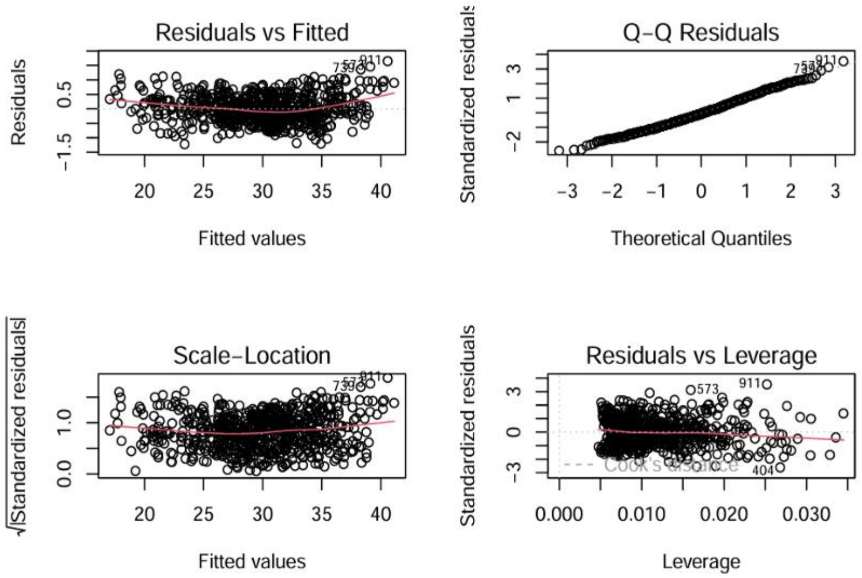
High-leverage points detected but do not exceed Cook's Distance threshold.

Next Steps:

- Apply **square root transformation** to the response variable

Square Root Polynomial Terms

Sqrt(Calories_Burned) = 31.146584 + 108.577001 * poly(Session_Duration..hours,2)1 –10.049548 * poly*Session_Duration..hours.,2)2 + 39.147791* poly(Avg_BPM,2)1 – 1.425752 * Gender – 0.056822 * Resting_BPM



Residual Analysis:

- Residuals vs Fitted: Deviations caused by a few specific points.

- Q-Q Plot: Points in the tails disrupt normality.

High-Leverage Points:

- Cook's Distance plot indicates influential points

- These points may disproportionately affect model performance.

Next Steps:

- Outlier Removal

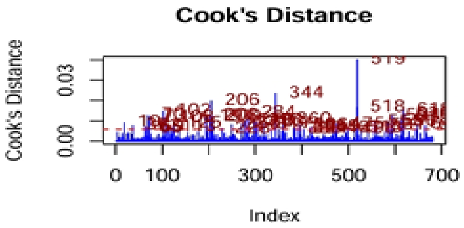
Outlier and Influential Points

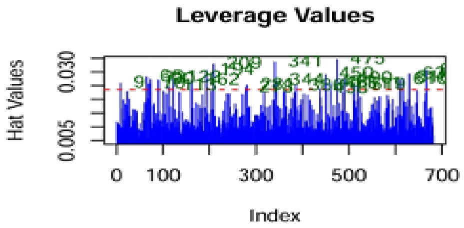
Formula:

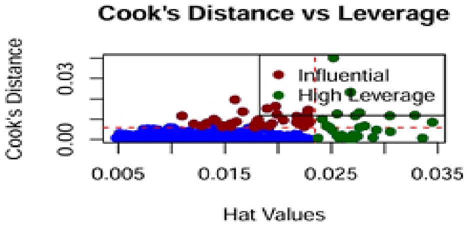
$$h_{ii} = \mathbf{x}_i \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i^\top$$

Where: - $\mathbf{x}_i$: The row vector of predictor values for observation i. - $\mathbf{X}$: The matrix of all predictor values.

Threshold: Points with $h_{ii} > \frac{2p}{n}$ are considered high leverage.





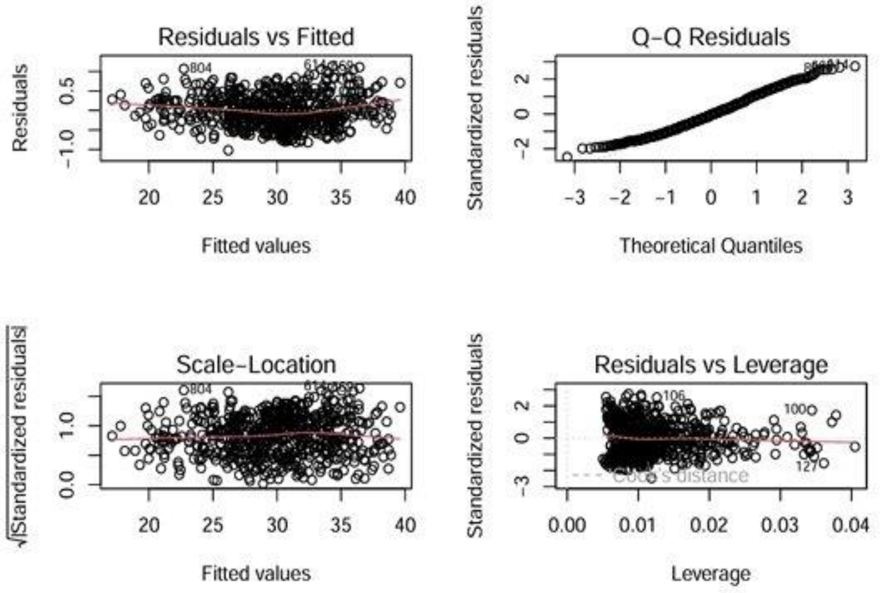


Key Insights:

- **Influential Points (Cook's Distance)**:
  - Points like **519**, **518**, **344**, **278**, and **284** exceed the Cook's Distance threshold, significantly influencing the model.
- **High Leverage Points**:
  - Observations like **9**, **70**, **113**, **209**, and **519** have high leverage, indicating they are far from average predictor values.
- **Overlap**:
  - Points **519**, **344**, and **278** are both **influential** and **high leverage**, requiring further investigation.

Debiased Square Root Polynomial Regression

Sqrt(Calories_Burned) = 31.146584 + 108.577001 * poly(Session_Duration..hours.,2)1 – 10.049548 * poly(Session_Duration..hours.,2)2 + 39.147791 * poly(Avg_BPM, 2)1 – 1.692661 * poly(Avg_BPM, 2)2 +1.425752 * Gender – 0.056822 * Age + 0.002505 * Resting_BPM



- Removing outliers has significantly improved model diagnostics.
- The updated model better satisfies assumptions and provides more reliable predictions.

Debiased Polynomial Regression with a Transformation

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

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	

Prediction Results and Model Evaluation

VIF Interpretation

- VIF = 1: No collinearity.
- 1 < VIF <= 5: Moderate collinearity (acceptable).
- VIF > 5: High collinearity (requires attention).
- VIF > 10: Severe collinearity (predictor showed be reconsidered).

Observation:

- VIF values are all below 5.
- Moderate collinearity.
- Model is relatively stable.
- Predictors are not highly correlated.

Variable	**GVIF**	**Df**	$\mathbf{GVIF}^{1/(2*Df)}$
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

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

Model Advantages and Limitations

Advantages:

- **Accurate Predictions**: Captures complex patterns with transformations (\sqrt{} , log) and polynomial terms.
- **Comprehensive Insights**: Analyzes key factors like session duration, avg BPM, gender, and age.
- **Handles Non-Linearity**: Polynomial terms model non-linear relationships effectively.
- **Robust and Reliable**: Outlier removal improves stability and generalizability.
- **Better Model Fit**: Meets key regression assumptions (linearity, normality, homoscedasticity).

Limitations:

- **Generalizability**: Performance may drop with unseen data.
- **Complexity**: Transformations and higher-order terms are harder to interpret.
- **Residual Issues**: Slight normality deviations remain.
- **Risk of Overfitting**: Especially with small datasets.
- **Limited Interactions**: Does not include predictor interaction effects.

Further Exploration of the Models

***Challenges with Current Model**:*

- Overfitting: High risk due to complex transformations and polynomial terms.
- Limited Generalizability: Performance may decline on new or unseen data.

***Explore Regularization Techniques**:*

- Prevents overfitting by controlling model complexity.
- Improves performance and robustness on unseen data.
- Balances model simplicity and predictive accuracy.

Considered Regularization Methods: LASSO, Ridge, Elastic Net

Introduction:

LASSO:

1. Adds an L1 penalty ($\lambda \sum |\beta i|$)
2. Shrinking some coefficients to zero
3. Creating a simpler model

Ridge:

1. Adds an L2 penalty ($\lambda \sum \beta i\hat{}2$)
2. Effective for multicollinearity and reducing overfitting
3. Creating a more complex model

Elastic Net:

1. Combines L1 and L2 penalties ($\lambda(\alpha \sum |\beta i| + (1-\alpha) \sum \beta i\hat{}2)$),
2. Balancing Ridge's regularization with LASSO's variable selection.

Insight:

- **Elastic Net:** Preferred for higher prediction accuracy and capturing complex factor interactions.
- **LASSO:** Recommended for identifying critical factors affecting Calories Burned, only keep significant variables.

Assumption:

As our dataset is high dimensional and large in size, Elastic Net and LASSO may better than Ridge.

Result of Models

Ridge has hightest optimal λ, suggesting the model may depending too much on penalty function.

LASSO has lowest optimal λ, meaning that the model depending less on penalty function than Ridge,

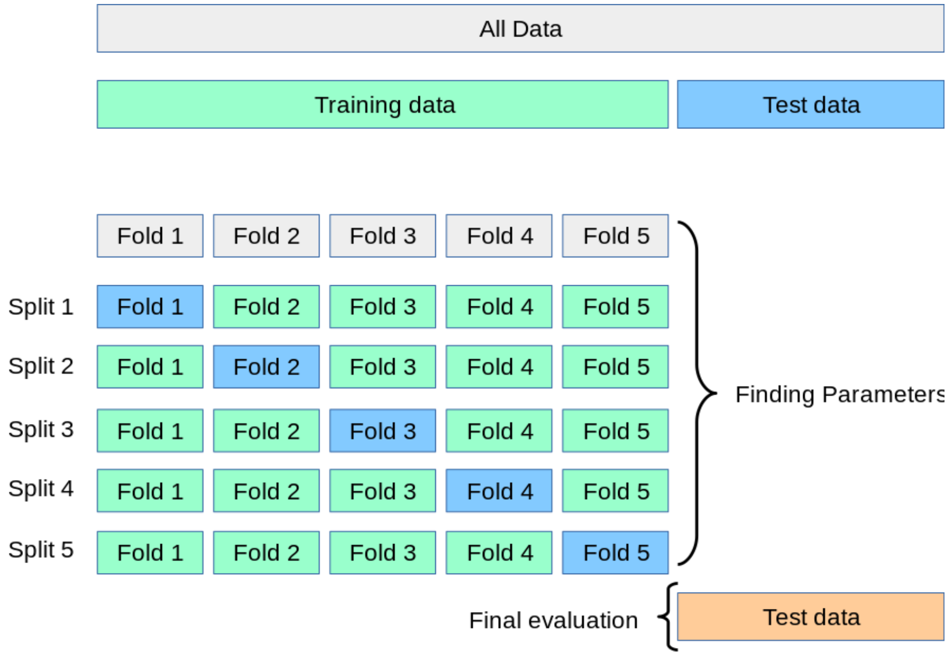
Model	**Optimal λ**	**Key Feature**
LASSO	0.00257	Shrinks coefficients to exactly zero
Ridge	0.09033	Shrinks coefficients, no selection
Elastic Net	0.00427	Combines LASSO and Ridge properties

Cross-Validation

Split the Data:
Divided the dataset is into k subsets (of approximately equal size).

Training and Testing:
Train the model on k−1 folds and tested on the remaining fold. Repeat This process k times,

Compute the Average Performance:
Calculate the performance metric for each iteration.



Optimal lambda Parameters Through Cross-Validation

Optimal λ by Cross Validation 10 folds:

- **LASSO:** 0.0234
- **Ridge:** 0.0903
- **Elastic Net:** 0.00469

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

Model Evaluation on Test Data and Discussion

Model Evaluation (Test MSE):

- **LASSO:** 0.02348011
- **Elastic Net:** 0.02360708
- **Ridge:** 0.03692044

The MSE results proved our assumption.

Future Directions:

- Explore interactions among predictors.
- Investigate advanced non-linear models for enhanced predictive power.

Conclusion

- The study explores factors influencing **Calories_Burned** and develops predictive models under rigorous testing.
- **Square Root Polynomial regression**:
- Captures non-linear relationships for variables like **Session_Duration**, **Avg_BPM**, and **Gender**.
- Practical for analyzing interactions but risks **overfitting** and **multicollinearity**.

Conclusion

- **To address these challenges:**
- **LASSO**:
  - Simplifies models by selecting key predictors.
  - Ideal for interpretability and prioritizing significant variables.
- **Elastic Net**:
  - Balances variable selection and multicollinearity handling.
  - Preferred for robust predictions, especially with correlated predictors.

Polynomial regression is suited for exploratory analysis; **LASSO** for interpretable models; **Elastic Net** for robust predictions.

Further directions include validating models on independent datasets, examining interaction effects, and applying advanced machine learning techniques.

Reference

https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset/data

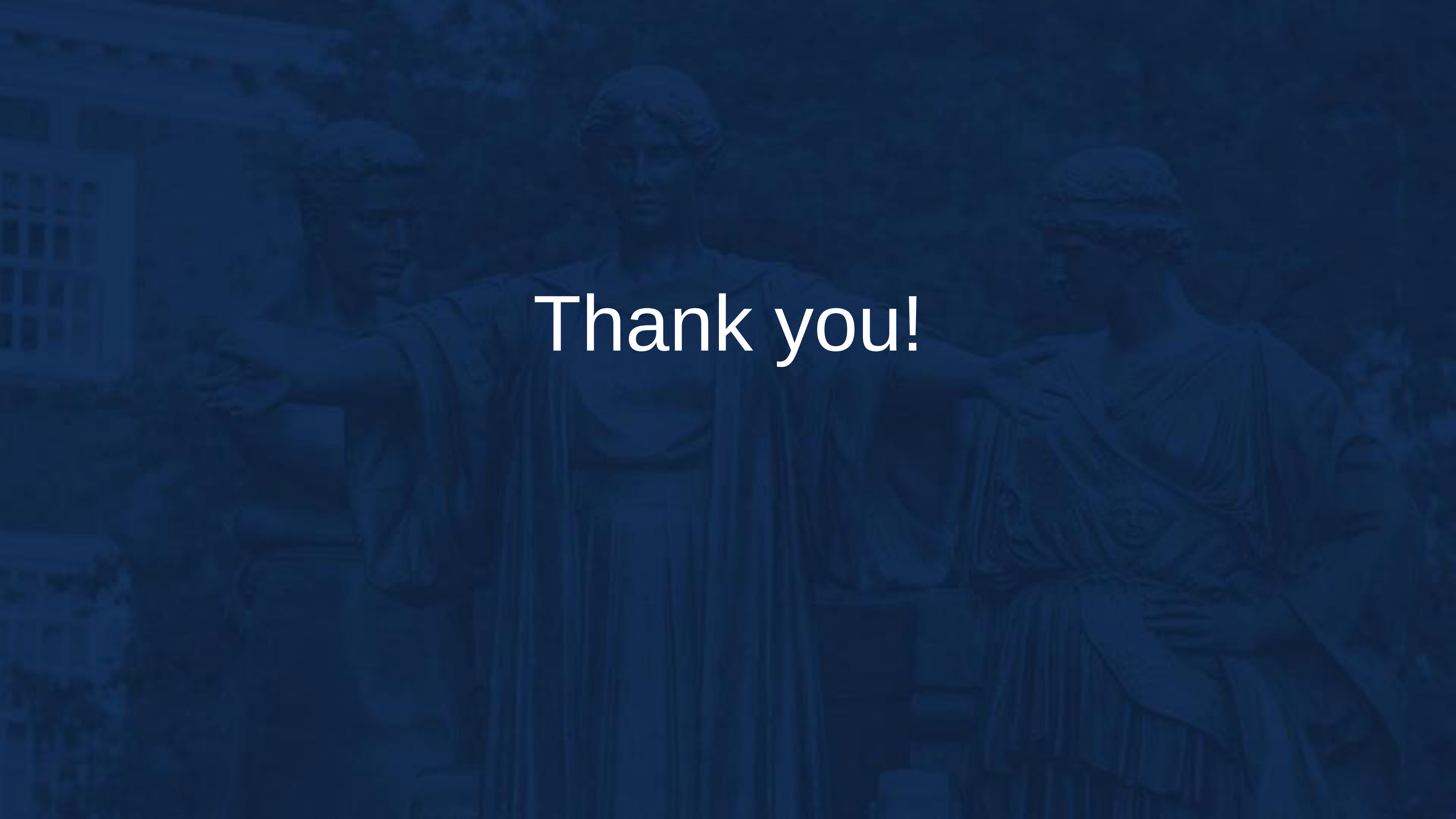
Github Source

https://github.com/WQY497/Predictive-Models-development-for-calorie-consumptions-

“ ”

Q & A

“ ”


Thank you!