Ruth Maina

DSC 680 - Applied Data Science

Professor Amirfarrokh Iranitalab

September 21, 2024

Predicting Calories Burned by the Human Body

Final Report White Paper.

Topic: Predicting Calories Burned by the Human Body

Background/History:

The topic of health, food calories and weight loss is one that has continued to be of high interest worldwide and here in America. According to the Center of Disease Control (CDC), about forty percent of the American population is obese. So many diseases are because of being overweight.

Business Problem:

Predicting calories burned by the human body would help curb the obesity pandemic since weight loss is directly correlated to calories burned. The research findings would help shed light on high-ranking variables that potentially would be of highest impact than other variables, thus enlightening individuals on areas of focus *(refer to appendix for real life application).* Knowing this would eventually help curb health problems that are associated with obesity.

Data Explanation:

Dataset:

The data sourced from Kaggle contains 15000 rows of health bio information. Below are Columns in the dataset:

- User_ID – Unique ID for each row
- Gender – contains male or female
- Age – contains ages 20 till 79
- Height – contains height in centimeters
- Weight – contains weight in kilograms
- Duration – contains exercise duration in minutes
- Heart_Rate – contains heart rate beats per minute
- Body_Temp – contains body temperature
- Calories – contains calories burned

Data preparation steps

- I loaded the csv data into a data frame and performed cleanup activities such as checking for null values and unwanted characters. I also checked for missing values to enforce data integrity. All the aforementioned activities yield clean data, which impacts model performance positively.

- I dropped the ID column to avoid noise in the model. The column is a unique identifier and does not contribute any meaningful information
- No outliers are observed from the scatter plots which is a good thing, as they can negatively impact the accuracy of the model.
- I added dummy variables for Gender since it is a categorical variable thus not quantifiable
- Also performed feature scaling by applying a standard scaler to the data to ensure the features are in a comparable scale
- I split the data into training and test data and used Calories as the dependent variable, and all others as independent variables.
- Chose a linear model since its best suited for prediction of a target variable based on one or more input features, which is my exact use case.
- I used R2 and RMSE scores since these are good indicators of how well a model is doing - they quantify the overall performance and predictive power of the models.

Methods:

- I first viewed summary statistics to get a general feel of the data
- I pared the variables and visualized them via scatterplots - Calories with Height, Calories with Weight, Calories with Age which showed no correlation since no linear relationship or pattern observed from the shape of the distribution. All other variables indicated positive correlation. However, I kept all variables for further analysis since in general they are all relevant to the subject and they do contain data that could be meaningful to the outcome

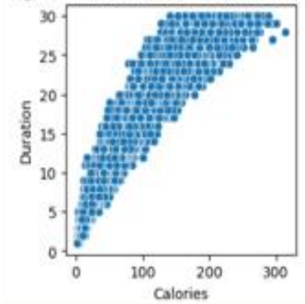


Figure 1 - Plot of Calories versus Duration

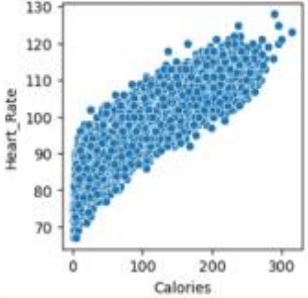


Figure 2 - Calories versus Heart_Rate

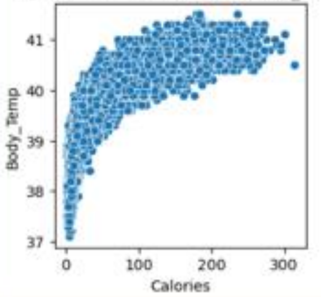


Figure 3 - Calories versus Body_Temp

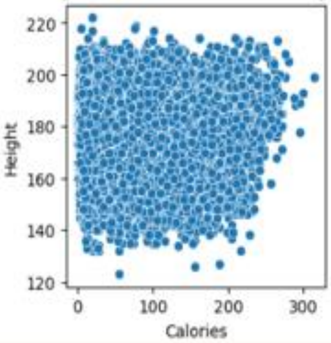


Figure 4 - Calories versus Height

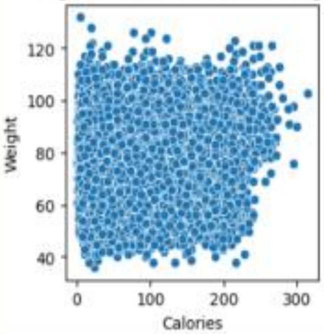


Figure 5 - Calories versus Weight

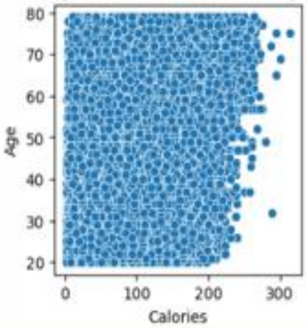


Figure 6 - Calories versus Age

Analysis:

- R2 scores were high for the model at 96.73% and 96.67 indicating a good model that predicts calories as close to 100% as possible. The R2 score of the training set was slightly better than the test set noted by the slightly higher percentage, indicating the train set sample yielded better results than the test set
- RMSE score of 11.51 was reasonable compared to the scale of the target/Calories variable range of 1-314. This was a good indication as well since calorie values are within a reasonable range
- Plot of predicted and actual values indicated good model performance:

✓ Plot of predicted training data versus actual indicated a linear relationship – the diagonal shape/distribution of the data points was a good indication of the model learning the training data very well.

✓ Plot of predicted test data versus actual also indicated a linear relationship – the diagonal shape/distribution of the data points was a good indication of the model performing very well in relation to unseen data (test data)

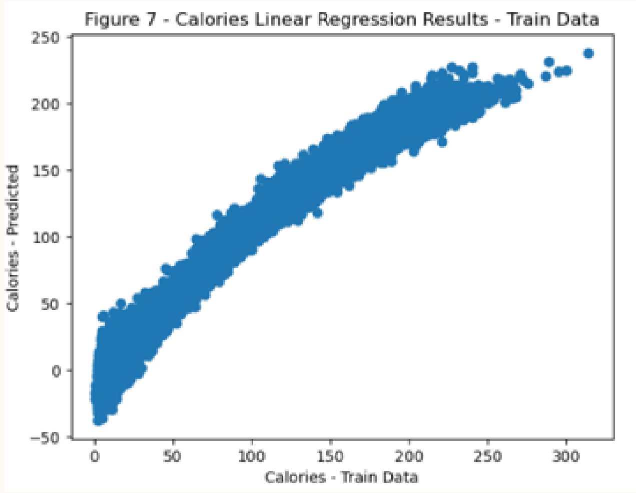


Figure 7 - Calories Linear Regression Results - Train Data

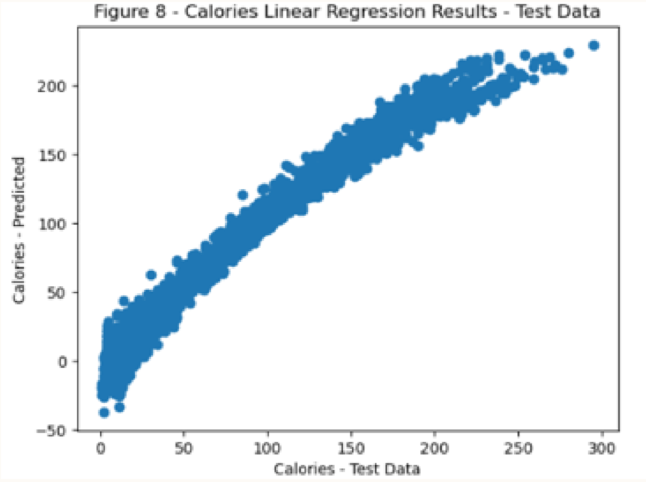


Figure 8 - Calories Linear Regression Results - Test Data

Conclusion:

- The project is an eye opener on the positive impact of exercise duration to the human body, highlights how increased heart rate and increased body temperature causes more calories to be burned. All these factors are directly related to weight loss, which in turn helps curb health problems that are associated with obesity.

Assumptions:

- Every human body is different thus would react differently to the variables in scope and would yield slightly different results. These details would need to be communicated to the users benefiting from the model.

Limitations:

- This is an independent prediction that does not allow for human testing or validation of results. This can be overcome by accumulating larger datasets that allow for bigger samples to enable smaller margins of error.

Challenges:

- The dataset is sourced from a Kaggle thus might not be reputable source for health information which could pose an ethical concern that it’s not coming from well-known and trusted health sources such as Mayo Clinic or even the CDC. To mitigate this, the model itself could be run on any set of true data to test its versatility. An example of such could be data collected from an individual's health tracker, from an exercise machine tracker, and so on.

Future Uses/Additional Applications:

- The model may be used for academic research, it could be used by clinicians for offering recommendations to individuals, and by health-conscious individuals who gather data via such means as health trackers, exercise equipment, and so on. I personally will use this model for my own data collected since 2014 on my fitness tracker. I could also analyze week by week data to see how best to increase my calories burned – I would work on a variable at a time to see which give me best results. For example, I know Heart Rate, Temperature and Duration are the top ones, I would workout at increased temperature (such as at noon time versus cooler morning temperature), to see which data gives me most calories burned. Same goes for duration (this is common knowledge on relation to calories, but I can test it anyway using personal data for personal exploration. Heart Rate I will also analyze further for optimal output. My personal goal would be to optimize weight loss. *Please refer to Appendix for a sample data collection template.*
- The model may be used by individuals seeking to maintain overall good health and thus is not limited only to those seeking to lose weight. It may be tweaked to give optimal numbers for the variables in scope.

Recommendations:

- A recommendation can be made to scale down/eliminate some features and to only focus on the select few with the highest correlation. As noted earlier, Height Weight and Age features did not indicate a linear relationship thus my recommendation would be to remove these from the model one at a time/iteratively while re-reviewing the metrics to gauge positive or negative impact on the model. More robust feature selection mechanisms can be used to solidify the model more by utilizing the best performers. Nevertheless, the model is currently ready for deployment since the metrics are high thus indicating a solid performance.

Implementation Plan:

- The model will first be deployed into lower testing/sandbox environments for validation before it goes live into production. The Site Reliability teams, or production support will be trained on the model details and will be given a troubleshooting runbook to enable model support is covered post deployment. This will ensure that any breaks or unexpected behavior is addressed promptly.

Ethical Assessment:

- It is imperative that the predicted outcome is accurate since putting incorrect recommendation or providing less than precise correlation could pose ethical implications. The core data from Kaggle data source is not a medical website and thus not ideal for health information. Therefore, end users should be informed for transparency.
- Privacy laws is another consideration as the data pertains to health information. Disclosing information that can be a direct identifier or personal in nature is not permitted.

Questions and Answers:

1. How was the data gathered?
   The data was collected from individuals who opted-in to share part of their results during annual wellness check-up with their doctors, as part of a clinical study.
2. If this is real human data, did the data owners approve of this use?
   Yes the individuals approved by opting-in and full disclosure occurred regarding the intent and how the data was going to be used
3. How can you ensure privacy is adhered to?
   Our networks are fully firewall secured, and two-way authentication is enabled for access, which is also restricted to individuals on a need-to-know basis.
4. What data retention policy will be applied?
   Our intent is to keep the data for five years, after which it will go through destruction process (complete purge)
5. What policies we're consulted to ensure compliance with health regulations?

- *HIPAA (Health Insurance Portability and Accountability Act) was consulted to ensure patient data safety*
- *Centers for Disease Control and Prevention (CDC) was consulted for additional reference*

6. How was the sampling done?

   The 80/20 Pareto Principle was used to split the data into training and test datasets accordingly – 80% used for training the model mainly to learn the correlation and the relationship between the data, and 20% used for model validation.

7. Was any of the gathered data not used?

   The entire dataset was used, except for the identifier column

8. How do you plan to use the model?

   The model will be made available to healthy conscious persons interested in the subject, and for academic research

9. Why was the model chosen over others?

   The linear model is best suited for prediction of a target variable based on one or more input features, which is my exact use case

10. Can the model be applied to all age groups?

    Yes this model may be applied to all age groups

GitHub Code link: https://github.com/ruthmaina2022/Data-Science-Portfolio/blob/main/Applied%20Data%20Science%20Projects/Ruth%20Maina_DSC680_Assignment%204.1_Project%201_Milestone%203_Final%20Report%20Code.ipynb

References:

JHA, Muskan (2021,8.21). *Calories Burned Prediction.*

https://www.kaggle.com/code/muskanjha/calories-burnt-prediction/input

Pneumol, J. (2022 Nov 25). *Linear and logistic regression models: when to use and how to interpret them?*

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9747134/#:~:text=Linear%20regression%20is%20used%20for,or%20a%20mix%20of%20both.

National Center for Health Statistics (n.d). *Health, United States – Data Finder*
https://www.cdc.gov/nchs/hus/data-finder.htm.

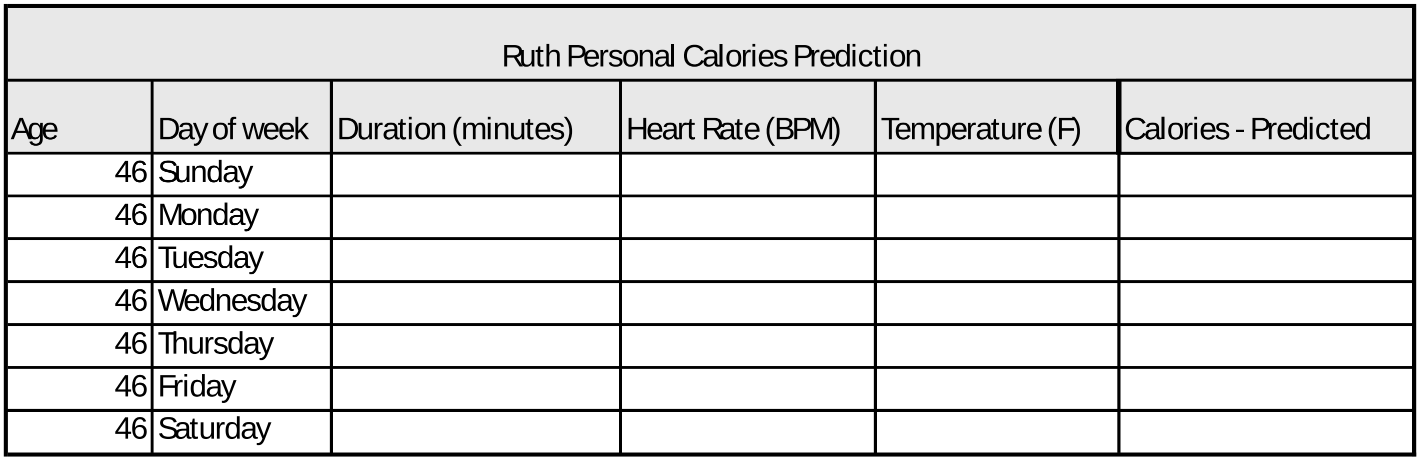
US Department of Health and Human Services
https://www.hhs.gov/hipaa/index.html

IBM (2021, Dec 7). *Linear regression.*

https://www.ibm.com/docs/el/spss-statistics/beta?topic=features-linear-regression

Appendix:

Below is a sample data collection template for individual use - this is one I would use as noted earlier, from data collected via a personal tracker or an exercise machine. This sample is specifically for a week-by-week analysis, with the main purpose to see how best to increase calories burned for increased weight loss.



Ruth Personal Calories Prediction					
Age	Day of week	Duration (minutes)	Heart Rate (BPM)	Temperature (F)	Calories - Predicted
46	Sunday				
46	Monday				
46	Tuesday				
46	Wednesday				
46	Thursday				
46	Friday				
46	Saturday				