Team Member: Siyi Wei, Zixin Lin, Chenhao Wang, Hao Liang

Team Name: Superpotato

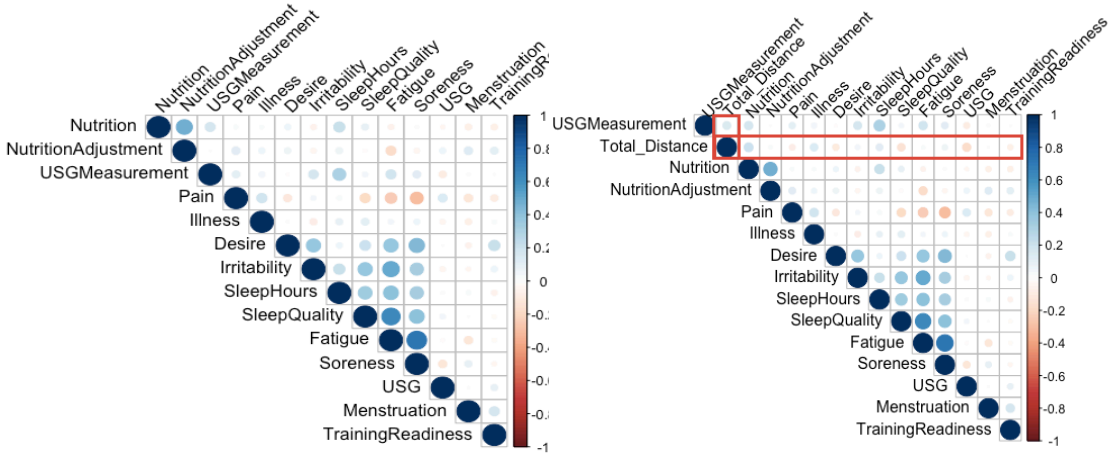
Brief Introduction:

Our goal is to analyze the dataset to generate a faigue measurement for each athlete as a reference for coach before each game.

Steps:

1. Cleaning the data and choose the valid data as training inputs. Found the most accurate and related measurements to describe the fatigue level.
2. Combing the objective data and subjective data to construct a more accurate model than RPE by deep neural network
3. Validating the model by cross-validation. Determine the accuracy of the model.

Data Preprocessing




```
> summary(fit, test="Hotelling-Lawley")
                            Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
Training_Sets$Total_Distance   1          0.28436   2.2748     14    112 0.008987 **
Residuals                    125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Data observation:

We want to see a low correlation of input features and a high correlation between the input features and the output variables(Total_Distance). The correlation matrix shows the correct pattern in input features. But not that much for Total_Distance. We decide to move on at this point since there might be hiden Causality.

The correlation matrix also show the correlation of subjective data and objective data. The objective data like sleep time and USG are highly correlated them self. However the correlation between objective data and subjective data is significant low. We made a conclusion that the subjective data is not accurate compare to objective data. That is the reason we choose valid distance instead of RPE score, which is measured by atheletes themselves.

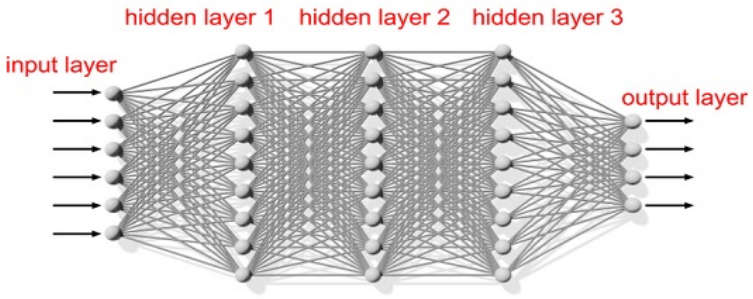
Data choosing:

There are several ways to gain the distance of the athletes that move per game by the GPS dataset. We dropped longitude and altitude since the accuracy is not precise and it does not reveal energy cost. After the observation of acceleration and the research of acceleration sensor. We made a hypothesis that the simplest move will trigger the sensor and generate the data. This is not accutate as well since it will overestimate energy cost. In the end, we choose the speed to calculate the valid distance as our measurement of the fatigue level for athletes in specific game. Since fatigue is a significant factor in that it affects biomechanical aspects of long-distance movement.(Christina, K. A., White, S. C., & Gilchrist, L. A. (2001)).

Final training sets:

There are 127 valid observations. The mean of the output is 1054.09 and the standard deviation is 456.07. The input features include Fatigue, Soreness, Desire, Irritability, SleepHours, SleepQuality, Pain, Illness, Menstruation, Nutrition, NutritionAdjustment, USGMeasurement, USG, TrainingReadiness. The output variable is the Total_Distance, which is calculated by the speed per second. All the observations were scaled and quantified. The NA variables were masked by the mean of this feature.

Model Construction



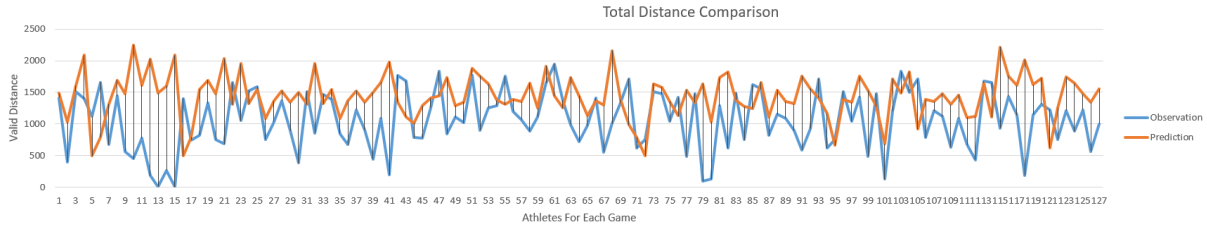
deep neural network

We use DNN regressor build in tensorflow. This is a powerful regression tool especially for regression problem with large features. Which is perfect for the amount of features we have in the training dataset.

We have 12 layers in the neural network. Each layer has around 60 neurons. We also use Adagrad as the optimizer. The model has 15000 training steps with batch size to be 10. The loss for final step is 776005.6. This is a huge loss but we want to predict the data and see the pattern of our prediction.

Model Validation

We validate the model by cross validation. The model has 1000 validation steps with batch size to be 10. The average loss for each predict is 763.45. The mean of the prediction is 1434.63 and the standard deviation is 347.55. The loss is acceptable and the prediction's pattern is close the the observation.



Problem Found

1: The input features are lack of objective features. Most of the features are subjective, and they are causing a huge loss to our prediction. We did a T-test about the RPE score and the valid distance for each athlete. We conclude there is no significant relationship between those two variables.

2: We need more observations. 127 observations are not enough for the DNN training. We saw a decrease on the training loss but the is stops after 8800 steps. We decide to increase the layers and the neurons but it did not help too much. We are not confident enough to our model since the lack of observations led us shrink our test sets.

3: The outliers are significantly affect the model. We need to investigate more on thoses outliers but we guess this problem is caused by lack of observations.

Conclusion

There is a relationship between the features collected before the game and the valid distance for each athlete take in game. Our model could help the coach to found the fatigue level of each athlete and decide which athlete should go for first round and manage their energy cost.

Reference:

Christina, K. A., White, S. C., & Gilchrist, L. A. (2001). Effect of localized muscle fatigue on vertical ground reaction forces and ankle joint motion during running. Human Movement Science, 20(3), 257-276.