Personalized Whoop Analysis and Predictions

Introduction to Data Science Deliverable 1

Topu Saha

September 26th 2023

Introduction

As a collegiate Cross Country/Track and Field athlete at New Jersey City University, understanding my body is key to being a successful athlete. This is why back in 2021 I decided to wear the Whoop fitness band 24/7 to gain insights pertaining to my body and performance. Whoop is a fitness-tracking band unlike any of the other fitness wearables with the goal of collecting data that actually matters and producing insights to understand what is good for your body and what is bad. It does this by being on your body at all times and collecting around "50-100 megabytes of data on a person per day." [1] The data contains measurements for sleep and exercises throughout the day and more. With all this data, it can analyze trends and answer questions such as what the user does to improve sleep, have a better workout, and more.

I stopped using Whoop because of its negative impact on my mental state during training. Recieving alerts that my body isn't optimal and having to compete in a race is not the best way to have a postive mental state. The insights I was receiving were not very helpful as well. This allows for some data analysis to spot trends and patterns myself instead of having a pre-determined computer algorithm set by Whoop analyzing them for me. This also allows for the recreating of the Whoop algorithm that identifies how ready you are to train by using some regression algorithms and understanding how they work to better make decisiosn for myself.

Understanding the Data

The following figure contains the conlumns and definitions of Physiological Cycle cvs file that was used.

Physiological Cycle

Field Name	Description
Cycle start time	Timestamp of when the physiological cycle started, in the timezone the cycle started
Cycle end time	Timestamp of when the physiological cycle started, in the timezone the cycle started
Cycle timezone	Timezone that the physiological cycle took place, relative to UTC
Recovery Score %	Recovery Score from the physiological cycle, which reflects how well prepared your body is to take on Strain and is a measure of your body's "return to baseline" after a stressor, as a percentage
Resting Heart Rate (bpm)	Resting Heart Rate calculated from sleep, in beats per minute
Heart Rate Variability (ms)	Heart Rate Variability calculated from sleep, in milliseconds
Skin Temp (celsius)	Skin temperature captured during sleep, in celsius. Only available with WHOOP 4.0+
Blood Oxygen %	The oxygen level of your blood captured during sleep, measured as a percentage. Referred to as SpO2. Only available for WHOOP 4.0+
Day Strain	Strain accumulated during physiological cycle
Energy burned (cal)	Calories burned during physiological cycle, in calories
Max HR (bpm)	Maximum HR during physiological cycle, in beats per minute

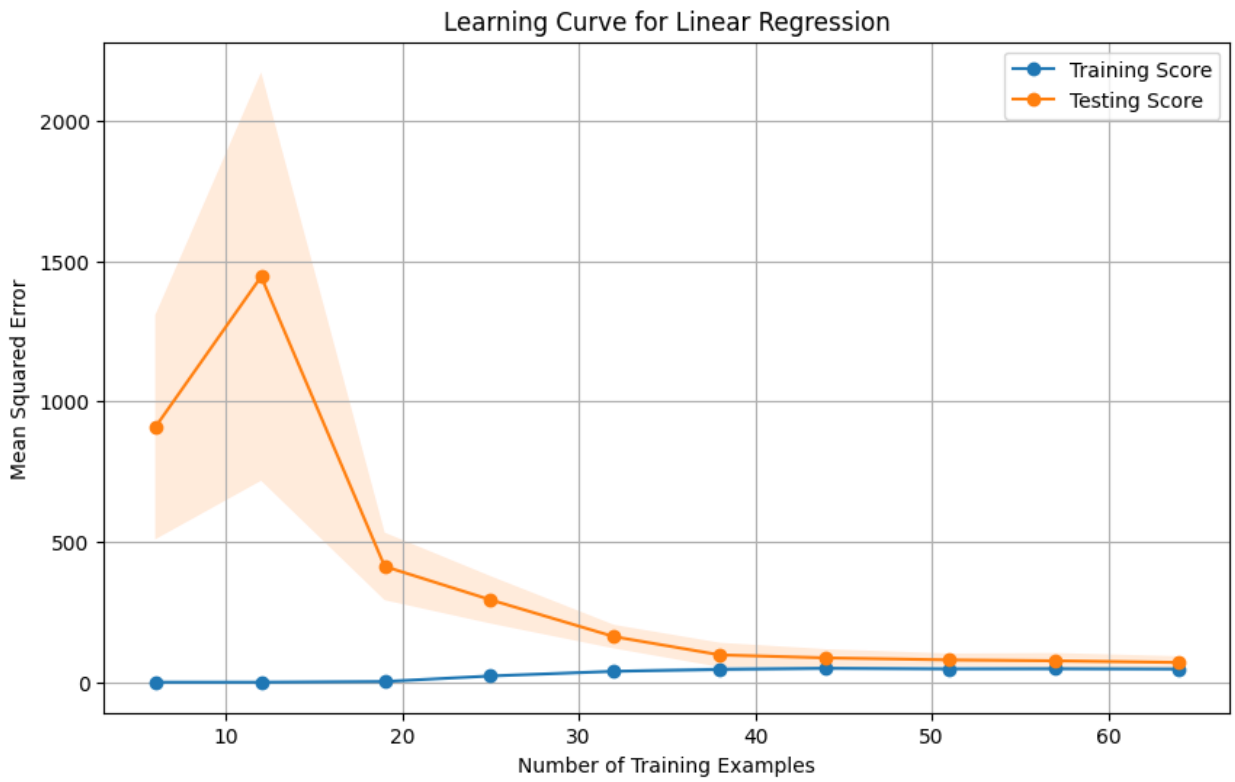
Average HR (bpm)	Average HR during physiological cycle, in beats per minute
Sleep onset	Timestamp of the start of the sleep, in the timezone the cycle started
Wake onset	Timestamp of the end of the sleep, in the timezone the cycle started
Sleep performance %	Hours of Sleep divided by Sleep Need
Respiratory rate (rpm)	Respiratory rate captured during Sleep, in respirations per minute
Asleep duration (min)	Time asleep (time in bed minus time awake), in minutes
In bed duration (min)	Time in bed (time awake plus time asleep), in minutes
Light sleep duration (min)	Time in Light sleep, in minutes
Deep (SWS) duration (min)	Time in Deep sleep, in minutes
REM duration (min)	Time in REM sleep, in minutes
Awake duration (min)	Time awake while in bed, in minutes
Sleep need (min)	The amount of Sleep Needed going into this physiological cycle, which factors in recent strain, sleep debt, and recent naps, in minutes
Sleep debt (min)	The amount of sleep debt accrued going into this physiological cycle, in minutes
Sleep Efficiency %	A measure of sleep quality; the percentage of time in bed actually asleep, as a percentage

My dataset consists of 104 rows × 26 columns which is roughly a three month period of data collection. After dropping features my Whoop device did not have and creating new features like day of the week and recovery category to provide value to the recovery scores my dataset consist of 102 rows × 25 columns.

Regression: Recreating the Recovery Algorithm Results

Linear Reggression

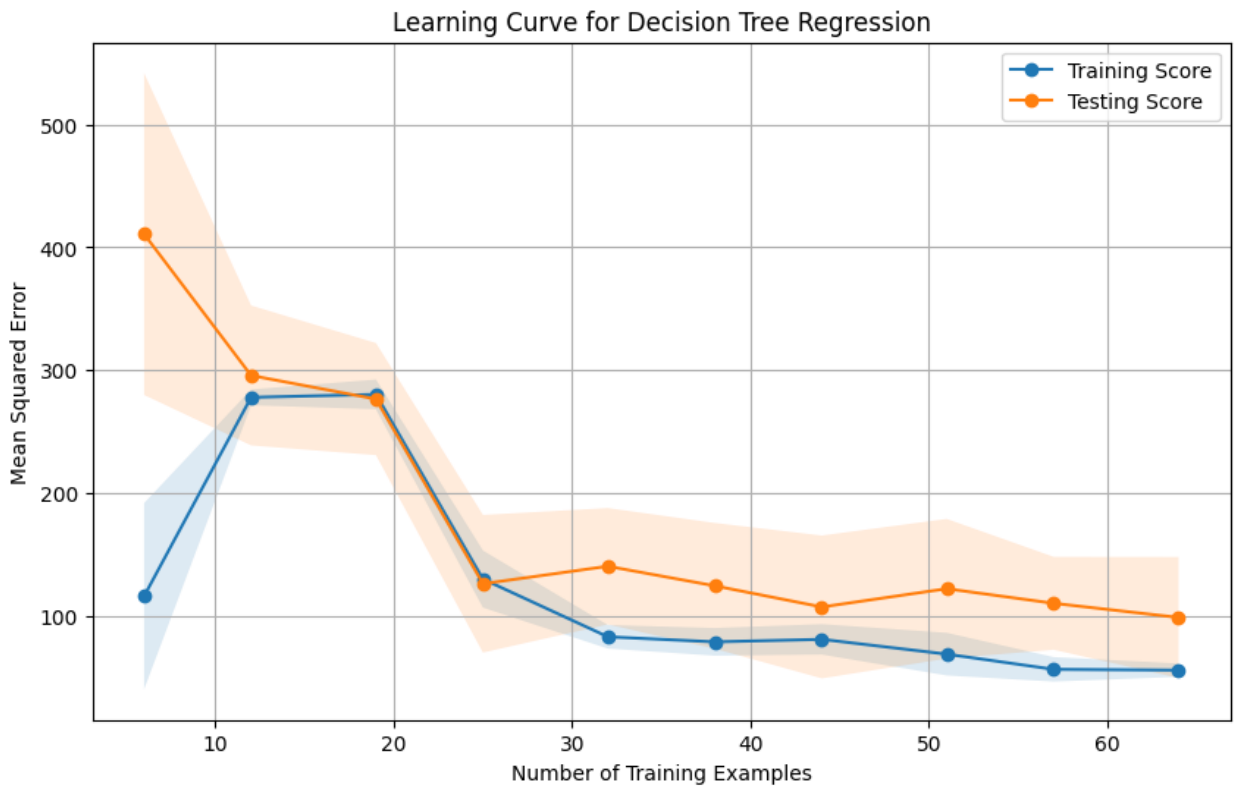
- Params_grid = {'fit_intercept' : [True, False], 'positive' : [True, False]}
- Best Hyperparameters: {'fit_intercept': False, 'positive': False}
- Results
    - MAE: 7.266689747098346
    - MSE: 89.33115074769962
    - RMSE: 9.451515791009378



Decision Tree Reggression

- Params_grid = { 'max_depth': [None, 10, 20, 30], 'min_samples_leaf': [1, 2, 4, 8], 'min_samples_split': [2, 5, 10, 20]}

- Best Hyperparameters: {'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 20}
- Results
    - MAE: 8.640065000359119
    - MSE: 93.97044125606018
    - RMSE: 9.693835219151406



Conclusion

When training my linear regression model, I tuned the hyperparamters fit intercept and positive which both can take a value of ether True or False. Fit intercept decides if the model should include a intercept term. It is applied when there is reason to believe that the independent variable has a non-zero value when the dependent variables are all set to zero. In this case, fit intercept is set to False, meaning an intercept was not included. When all the dependent values in the dataset is zero there is no default value the recovery score can be. Positive is also set to false which is used when there is a strong belief that the dependent variables does not positively affect the independent variable. The reason it is false is because not all of the dependent variables have a positive affect on the Recovery score. For example, the more strain you put on your body or the more calroies you burn the more tired you can expect to feel the next day.

For Decision Tree Reggression I used pre-training prunning. Pre-training prunning is when I set the type of pruning I want to stop the decision tree for growing out of control during the training phase. The hyperparamters I chose are max depth, min samples leaf, and min samples split. The max depth parameter controls the maximum depth a decision tree can go and the best one was set to none which makes sense since there is a lot of complexity within the data. So allowing the tree to grow to its maximum length gives it the best chance to understand the underlying pattern and reduce bias. Now we also want a balance between how complex the model is which is why the minimum amount of sample leafs was set to 4. This controls the complexity and improves generalization to reduce the chances of overfitting. Now the minimum sample split was set to 20 meaning the minimum number of samples required to split a node is set to 20. This reduces the depth of the tree which explains why max dept can be none. Combining these two pruning techniques before training controls how the tree is grown.

When comparing the algorithms to each other both are doing well. The measurements of MAE. MSE, and RMSE are all low meaning that the predicted values are close to the actual values. When choosing a winner, the linear regression out preforms the decision tree regression since its error rates are smaller. On average the linear regression is 7 units off the actual value and the decision tree is 8 units off. In the scenario if I wake up and I am shown a recovery score of 86 for the day through my linear regression and my actual score is 79, there will not be a big difference within how I feel throughout the day. The error rate is minimal enough to be disregarded in this case.

When looking at the learning curves for linear regression it is seen that the training and testing scores seems close to converging when the training is using around 60% of the total data. Similarly, the decision tree shows that pattern as well even though it does not fully coverage.

This means if more data was provided the decision tree would be able able to preform better. With that said, on the Whoop website it is stated that after a month of consistent use there algorithm will be able to provide insights and fully activate its predictive capabilities. [2] In my model I am using the standard 80/20 split for training and testing. Providing it more then a months worth of data to find the underlying pattern within the dataset.

Future Improvements and Next Steps

As shown with the learning curve, the best route of action to take to find more of a pattern is to collect more data. Through my exploratory data analysis on my github, I found that this algorithm is mostly trained on when I am pushing myself as a student athlete. I should collect more data when I am not training or taking a rest day to really dig deep into the patterns. One technique to see if there are any changes in the decision tree is post-training pruning techniques such as Cost Complexity Pruning to further control the size of the tree. Although, because of the error rates being low as it is and for the scenario it is functional, there is no need to further tune the model for better metrics, but there might be a need to further tune the model to ensure it is working well when deployed and analyzing new data.

With this said the goal of the assignment is not deployment but to discover the hidden patterns. Next steps would be to visualize the decision tree and linear regression to see what actionable steps I can take to increase my recovery scores. The issue I ran into this is an AttributeError when I tried to visualize the decision tree using graphviz to understand its logic. This error can be caused for multiple reasons such as my packages not being updated, or simply my notebook settings not allowing me to visulize the image.

Once this issue is resolved, and the visualizations are able to render, I can continue to build a classification model to further analyze the trends. From personal experience having a recovery score above 60 means I can push myself in training, building a model to find the trend between the dependent variables to optimize my lifestyle to achieve a score above that range will be beneficial. This can be done using a decision tree classifier or logistic regression. Main reason this was not implemented in this paper was due to time constraints.

References

1 - https://www.whoop.com/us/en/thelocker/podcast-74-story-of-whoop/

2 - https://www.whoop.com/us/en/