Deconstructing the Oura Ring Sleep and Readiness Algorithms with Comparisons to the Fitbit Charge 4 Sleep Score Features and Algorithm

Paul Richardson

Spring 2022

paulmrichardson@lewisu.edu

Abstract

The Fitbit Charge 4 and Oura Ring products provide sleep scores that provide insight into sleep quality. With these competing products providing the same types of insights with similar sensors, this project aims to explain the similarities and differences between the sleep scores provided by the Oura Ring and Fitbit Charge. This study also models the Oura Ring readiness score to identify the features of the readiness score that most influence the readiness score. It also seeks to validate the hypothesis that removing the Oura Ring for training should not adversely impact the Oura Ring readiness score. It concludes with a cursory look at how intuition compares to the Oura Ring readiness score. To accomplish these objectives Oura Ring and Fitbit Charge 4 data is analyzed for comparisons and is modeled to represent the sleep and readiness scores from the Oura Ring and Fitbit Charge 4. Feature importance for the models is analyzed using model coefficients, permutation feature importance, and Shapley Additive Explanations. Additionally, it uses a survey to capture user readiness scores to compare the Oura Ring device score. In conclusion, we identify that the Oura Ring and Fitbit Charge 4 produce similar quantitative sleep scores. However, the devices can produce considerably different scores for the same user on the same day. This difference in sleep scores is attributed to how the two devices interpret the underlying sensor data. A key difference in the scores is how the two devices determine Rapid Eye Movement (REM) and Deep Sleep. To maximize the Oura Ring sleep score, prioritize total sleep, deep sleep, and minimize sleep latency. To maximize the Oura Ring readiness score, maximize the HRV Balance Score, Sleep Balance Score, and Previous Night Score. Removing the Oura Ring for training should not adversely impact a user's readiness score. The Oura Ring readiness score can not replace intuition concerning how ready a user is to face the day.

Index Terms - Fitbit, Fitness Tracker, Machine Learning, Oura Ring, Sleep Score, Readiness Score Fitbit Charge 4, Permutation Feature Importance, Shapley Additive Explanations

I. INTRODUCTION

Fitness trackers presume to provide actionable insights into an individual's health. These trackers host a plethora of sensors and contain propriety algorithms to quantify sleep, activity level, and general readiness to take on the day, with new capabilities routinely being added. These devices are generally equipped with accelerometers to track motion, altimeters to track altitude changes, and optical sensors to track heart rate. In addition, these devices use these sensors and proprietary algorithms to provide sleep and readiness scores intended to give users insights into the quality of their sleep and provide guidance on how to approach their day. The Oura Ring is a finger-worn tracker that has provided sleep scores since its launch on Kickstarter in 2015[1]. The Fitbit Charge (Heart Rate) is a wrists worn tracker introduced in 2015. Fitbit introduced the sleep score as an indicator of a user's sleep quality in 2019. With these two products now competing with the same types of insights with similar sensors, this project aims to explain the similarities and differences between the sleep scores provided by the Oura Ring and Fitbit Charge. See Table I for a full list of Fitbit [2] and Oura Ring[3] sensors.

TABLE I

SENSORS FOR FITBIT CHARGE 4 AND OURA RINGS

Sensor	Monitors	Fitbit Charge 4	Oura Ring Gen 2	Oura Ring Gen 3
3-Axis Accelerometer	Motion	X	X	X
Infrared sensors	Heart rate and oxygen saturation	X	X	X
Red sensors	Heart rate and oxygen saturation	X	X	-
Green sensors	Heart rate	X	X	-
Temperature	Body temperature	X	X	X
Altimeters	Changes in altitude	X	X	X

In November of 2021, Fitbit announced the availability of the readiness score introduced to *provide personalized workout intensity and recovery recommendations based on your body*[4]. Unfortunately, access to the Fitbit readiness score is limited to newer devices and a premium account. However, the Oura Ring offers a readiness score composed of 8 different measures presented through its iPhone application without any indication of the level of influence of these contributors on a user's readiness score. In this work, we identify the data contributing to the readiness score to identify the features of the readiness score that most influence a user's score. Additionally, given that the Oura Ring is worn on a user's finger, it presents some challenges for the safety of the device and the user when worn during some physical activities. For example, the Oura Ring is endorsed basketball players such as Chris Paul[**?**] and CrossFit champion Tia-Clair Toomey[6] who participate in weightlifting that, could harm the ring, and basketball poses a safety risk to the user's fingers when wearing a ring. This paper examines the validity of the readiness scores for users who choose not to wear the ring during training or competition by evaluating the influence of each of the features that compose the Oura readiness score. Our effort hypothesizes that a user's Heart Rate Variability (HRV) is a more influential feature capturing the impact of physical activities when a user elects not to wear the ring. HRV is a measure of the variation in time between heartbeats. This variance measure of the time between heartbeats is generally accepted as a measure of various stressors on the autonomic nervous system(ANS). The greater the time between beats suggests a lower stress level. This measure also suggests that the lower the variance, the higher the user's stress level; more information about HRV is in the reference section[7]. Additionally, a part of this effort was a minimal look at subjective versus quantified readiness measures. The capability of a device to quantify a person's physical readiness sparks a curiosity concerning how users interpret their readiness vs. the physical device. This paper builds on the introduction and, provides an additional description of the project objectives—reviews how the data was collected and prepared for analysis. Walks through the methods used to address the project objectives and presents and discusses the results along with the ethical considerations of this work.

II. Discussion of Related Work

Several papers have been published discussing the accuracy of sleep trackers in tracking sleep. In one paper written by Fernando Moreno-Pino et al. (2019) titled *Validation of Fitbit Charge 2 and Fitbit Alta HR Against Polysomnography for Assessing Sleep in Adults with Obstructive Sleep Apnea* the accuracy of two Fitbit trackers were tested against a polysomnography (PSG) device. PSG is a comprehensive test used for sleep studies. These devices are bulky and obtrusive, which drives some of the appeal of a small unobtrusive device that may produce similar results. However, the measures (results) were statistically different for all measures except rapid eye movement (REM) in this study. The study concluded that the Fitbit wearables had acceptable sensitivity but poor specificity, which means that these devices are promising but not ready for clinical use[8]. Testing users with a sleep disorder is a challenging study for a consumer-grade device. In a paper by Shahab Haghayegh et al. (2019) titled textitAccuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis these researchers used the results of 22 studies pitting PSG against Fitbit tackers for accuracy and also concluded that the Fitbit trackers are promising, economical and suitable for gross estimates however their level of specificity makes them inferior to PSG [9]. The aim of the research for this paper does not approach device accuracy but a comparison of sleep scores between two devices that measure sleep and provide sleep scores. One researcher on Youtube has formulated a possible algorithm for the Oura Ring Readiness Score. The work done by Rob Ter Horst, known as The Quantified Scientist on his YouTube channel, suggests a viable model representing the Oura Ring readiness model[10]. In Ter Horst's work, the researcher uses underlying Oura data to fit a linear model to represent the Oura readiness score. He then validates this model with data from other users. In this study, we approach building a model for the readiness score in the same manner. We will also use Rob Ter Horst's work for model evaluation comparisons developed in this study. The work done by Rob Ter Horst and this study identify feature importance with coefficients; this work goes further by using additional methods to evaluate feature importance. Additionally, this work addresses the impacts of not wearing the Oura during training and how to maximize a user's readiness score.

III. Project Description & Objectives

With the Oura Ring and Fitbit Charge now both providing sleep scores, this has motivated the following objectives:

- To determine if the sleeps scores from the Fitbit Charge 4 and the Oura Ring provide similar scores for a single user wearing the devices simultaneously.
- To describe and compare the underlying features for the sleep scores provided by the Fitbit Charge 4 and Oura Ring, highlighting the similarities and differences in the sleep score features.
- To develop a model for the Oura sleep score providing analysis on feature importance to inform users on how to maximize their sleep scores.

This work also:

- Integrates the Oura readiness score to develop a model for the Oura readiness score providing analysis on feature importance to inform users on how to maximize their sleep scores.
- Provides a glimpse into how a user's intuition compares to the Oura Ring readiness score with the development of a simple survey and a summary of the survey results.

This project did not attempt to measure the accuracy of the sensors or the methods used to calculate and identify sleep stages, heart rate variability, or any of the other possible measures from the devices. However, this project does highlight how the interpretation of sleep stages highlights a crucial difference between the Oura Ring and Fitbit sleep scores.

IV. Data

A. Data Collection

The researcher recorded the data for this project with the Fitbit Charge 4 and the Oura Ring and an online form. The Fitbit Charge 4 tracker recorded data from January 1, 2019, to December 31, 2021. The Oura Ring data was collected by a second generation and third generation devices. The Oura Ring generation 2 device recorded data from January 1, 2019, to November 2021. The generation 3 device recorded data from November 2021 to December 31, 2021. The data was collected via the export functionality from the Oura Ring and Fitbit user portals. The Fitbit data in this analysis is limited to the features necessary for the Sleep Score only. The Oura Ring data was limited to the Sleep Score and Readiness Score features.

The Readiness Score from Fitbit requires newer or different hardware than the Charge 4 and a premium account. An online survey was created to track and compare subjective scores to the actual hardware scores for a user's readiness intuition vs. the devices readiness score.

B. Data Description

The data collected from the Oura Ring and Fitbit devices represents a single user wearing both devices simultaneously since 2020. A validation dataset from a different Oura Ring user is also used in this project to validate the models for the sleep and readiness scores.

The Oura Ring data consisted of 731 rows and 54 columns of daily data from January 1, 2019, to December 31, 2021. The quantitative scores are comprised of multiple scores generated from the sensors to contribute to the overall scores. Table II includes an example of single record from the dataset of the columns of interest that comprise the quantitative scores for sleep and readiness. Additionally, we have provided columns of interest used in the comparison of the Oura Ring and Fitbit trackers. The Oura Ring validation data set consisted of 518 rows and 54 columns of daily data from October 7, 2020 to April 24, 2022.

TABLE II

OURA RING DATASET GENERAL INFORMATION

Category	Data Columns	Sample Record ((Negative) 0 to 100 (Positive)
Sleep	Sleep Score	80
Sleep	Total Sleep Score	70
Sleep	REM Sleep Score	72
Sleep	Deep Sleep Score	58
Sleep	Sleep Latency Score	78
Sleep	Sleep Timing Score	50
Readiness	Readiness Score	84.0
Readiness	Resting Heart Rate Score	100
Readiness	HRV Balance Score	94
Readiness	Temperature Score	100
Readiness	Recovery Index Score	72
Readiness	Sleep Balance Score	100
Readiness	Previous Day Activity Score	90
Readiness	Activity Score	82

The Fitbit data consisted of 716 rows and 9 columns of daily data from January 1, 2020, to December 31, 2021. The quantitative sleep score is comprised of multiple scores generated from the sensors to contribute to the overall score. Table III includes an example of single record from the dataset of the columns of interest that comprise the quantitative score for sleep.

TABLE III
FITBIT DATASET GENERAL INFORMATION

Data Columns	Sample Record
Sleep Score	87
Overall Score	21.0
Composition Score	20
Revitalization Score	46

TABLE IV
OURA RING SLEEP FACTORS GENERAL INFORMATION

Data Columns	Sample Record
Total Sleep Score	70
REM Sleep Score	72
Deep Sleep Score	58
Sleep Latency Score	78
Sleep Timing Score	50
Restfulness Score	82

C. Data Preparation

Data preparation to directly compare the device's sleep data required data to be removed, columns of interest selected, and dates converted to the same formats. The datasets were then joined into a single dataset using an inner join on the date field. The Oura Ring and Fitbit handle data collected from naps differently. The Fitbit creates separate records for naps that do not impact the overnight sleep score. The Oura Ring will log naps and boost the users' readiness and sleep scores with decreasing levels of impact for higher scores[11]. The Fitbit data contained 12 separate records for naps; these records, both the overnight entries and the nap entries, were removed. The Oura data included 8 of the 12 entries found in the Fitbit data by date; these 8 entries from the Oura Ring data were also removed. The data was narrowed down to features identified in the Table II and Table III. The timestamp record for the Fitbit was truncated down to the YYYY-MM-DD format and renamed to Date. The Oura Ring date data was converted to the YYYY-MM-DD format. The Oura measure of deep sleep duration was divided by 60 to convert the unit of measure to minutes, similar to the Fitbit.

The columns were renamed to replace spaces and data types were converted to integers to prepare the data to model the sleep and readiness data. This process resulted in none of the data being removed for the sleep data. For the sleep validation set this process resulted in 340 records. For the readiness data this process resulted in 14 records being removed. For the readiness data validation step, this process resulted in 353 records. The features of interest for the models are captured in the Oura table. A challenge for fitting the data to the readiness model is that HRV Balance Score is a measure that was introduced as part of the readiness score equation in late 2019 / early 2020. The HRV Balance Score reflects the users HRV trend as a rolling average over a two week period in relation to a three month baseline[12]. The Oura Ring also provides a readiness score when values are unavailable. Only records containing all 7 factors identified for Readiness in Table II required for the Readiness Score were used to develop the model.

V. Methods

The sleep scores from the Oura Ring and Fitbit Charge 4 and some of their underlying data were examined for insights into similarities and differences in how the devices generate sleep scores. This analysis of the sleep scores for a single user measured the correlation between the sleep scores and the features of the sleep scores. The sleep scores from the Fitbit and the Oura Ring are both integers from 1 to 100, with the higher number being more positive. The factors for the Oura Ring and Fitbit Sleep Scores are not one for one. The Fitbit sleep score is derived from three scores. The Fitbit Sleep algorithm, factors, and coefficients are documented as the following:

$$SleepScore = (DurationScore/50) + (QualityScore/25) + (RestorationScore/25)$$

The Fitbit Duration score is a measure of sleep duration. The Fitbit Quality Score is comprised of Deep Sleep Score and REM Sleep Score. Deep Sleep and Rapid Eye Movement (REM) are more formally divided into REM Sleep and Non REM Sleep (NREM). The impacts of these stages of sleep is beyond the scope of this paper. However, how the Fitbit and Oura Ring determine REM and Deep Sleep impact the overall sleep score. The Fitbit Restoration Score is determined by the Resting Heart Rate Score and Restfulness Score to measure how peaceful or agitated a user was during their sleep. Additional information about the algorithm and factors can be found on the Fitbit website[11][13]. For comparison, the seven factors from the Oura Ring are mapped to simulate the Fitbit sleep factors. The Fitbit scores were refactored to match the Oura Ring measures; see Table V.

Once these variables were derived to match scope and units of measure they were used for comparisons. This comparison and mapping works well for illustration but is not an exact mapping back to the Oura Sleep factors. The Oura Resting Heart Rate Score is not a factor in the Oura Sleep score. It is used here only to map for comparisons with the device scores. To confirm this work, analysis of how the devices handle measuring deep sleep was also done to measure the correlation for this factor between the devices.

TABLE V

DERIVED FACTORS & ALIGNED UNITS OF MEASURE

Score	Equation
Oura Duration Score	Oura Duration Score = Oura Duration Score
Fitbit Duration Score	Fitbit Duration Score = (Fitbit Duration Score / 50) * 100
Oura Quality Score	Oura Quality Score = (Oura REM Sleep Score + Oura Deep Sleep Score) / 200
Fitbit Quality Score	Fitbit Quality Score = (Fitbit Quality Score / 25) * 100
Oura Restoration Score	Oura Restoration Score = (Oura Restfulness Score + Oura Resting Heart Rate Score) / 200
Fitbit Restoration Score	Fitbit Restoration Score = (Fitbit Restoration Score / 25) * 100

Prior to deriving the sleep and readiness models, we look at the validation dataset compared to the primary data set. This analysis provides some insight into the difference between the users and the general applicability of the fitted models.

To derive the sleep and readiness models, the factors identified in the device applications (iPhone apps) and their associated websites were used to determine the data and features to use for the models. The final models were developed in Python. Several models were initially evaluated using Orange 3[14] and AutoML[15] solutions (Lazy Predict[16] and TPOT[17]), and several iterations of hyperparameter tuning in Python. The derived models were then validated for performance against the validation data set. The feature importance of the models was then evaluated using model coefficients, permutation feature importance[18], Shapley Additive Explanations (SHAP)[19]. The models were evaluated using the r-squared, mean squared error, root mean squared error, and max error. When evaluating the regression metrics, the continuous values from the model were rounded to the nearest whole number to match the units of the sleep scores. TPOT was used in the additional evaluation of models, but the resulting models were simply over-complicated with no gain in performance. However, TPOT did have an iteration of results that used Polynomial Features that was incorporated to improve the metrics for the final Readiness model.

Also, a part of this effort was a minimal look at subjective versus quantified readiness measures. For ten days data was captured every morning and each evening to capture readiness score. Every morning the perceived Readiness Score was captured. Each evening the actual readiness score was captured with a reflection. The reflection represented a comparison between the two scores measured with the following responses 1) Device and my score are about the same 2) my score was more reflective of my readiness or 3) my devices score was more reflective of my readiness. This effort was completed by the researcher only.

This work incorporates permutation feature importance and Shapley Additive Explanations to identify the relationships between the features and the target variable. According to the documentation *permutation* *feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled 1. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature* [18]. Shapley Additive Explanations are based on the Shapely value which is based on game theory. This application takes the theory behind trying to quantify the contribution of players to results to features in a model. According to the documentation *SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions* [19].

VI. Results

A. Sleep Score Analysis

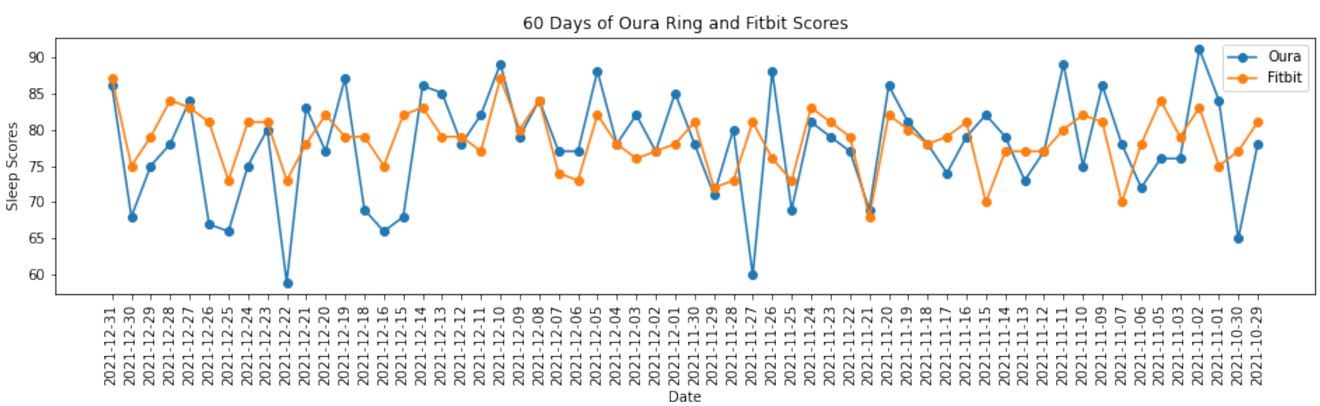

Fig. 1. Daily Scores over 60 day period for the Oura Ring and Fitbit Charge 4

The Sleep Scores between the Oura Ring and Fitbit generally follow the same trend (figure 1), with many scores having very marginal difference between the measures. For example, the number of scores with a difference of 5 or less is 398 or 60% of the scores. If/when the range is increased to 7 or less, it accounts for 74% of the scores.

The overall correlation between the two devices sleep scores is a moderate 57% (figure 2). The correlation values between the Oura derived variables mapped to the Fitbit variables indicate very low correlation with the exception of the Duration Score. This observation of low correlation between the mapped variables also supports that the devices vary greatly in interpreting sleep stages (Table VI).

The standard deviation in the sleep delta scores is 4.17 representing minimal spread between the data. However, there are instances of divergence between the scores and the trends. To illustrate how the devices might be interpreting factors differently, we can look at the deltas between the Fitbit factors and

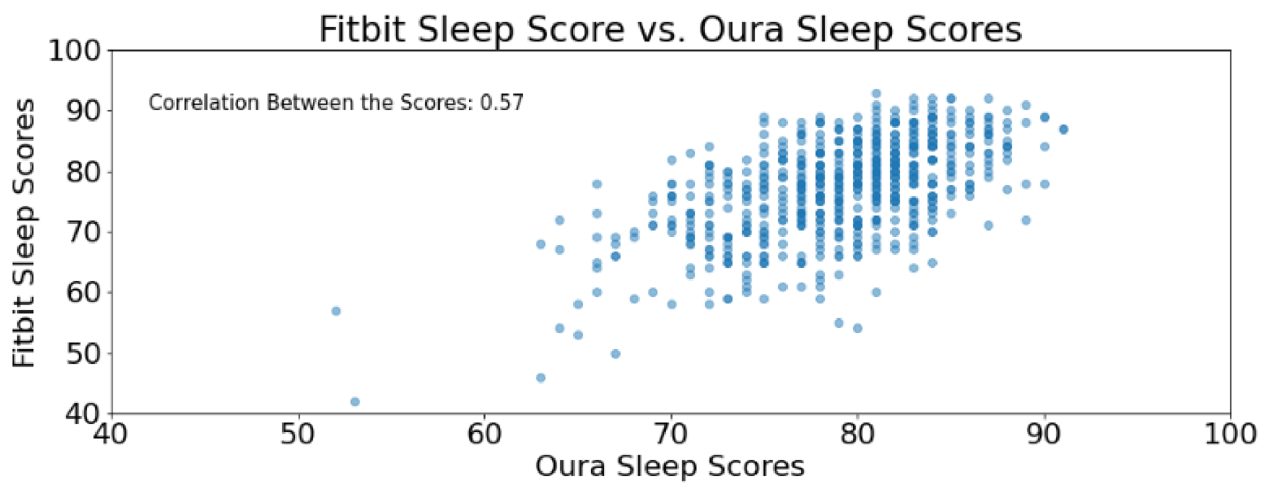

Fig. 2. Fitbit Sleep Scores vs. Oura Sleep Scores

TABLE VI

CORRELATION METRICS

Factor	Correlation
Revitalization Score	0.13
Composition Score	0.20
Duration Score	0.61
Sleep Score	0.57

the derived Oura Ring factors. The difference between the Duration and Quality scores is minimal, with standard deviations of 7 and 6. However, a much more significant difference between the two devices exists for the Quality Score with a standard deviation of 13. All of the scores also appear to have extreme scores. I do not classify these as outliers in this case, as the scores are real and not a misrepresentation. See figure 3.

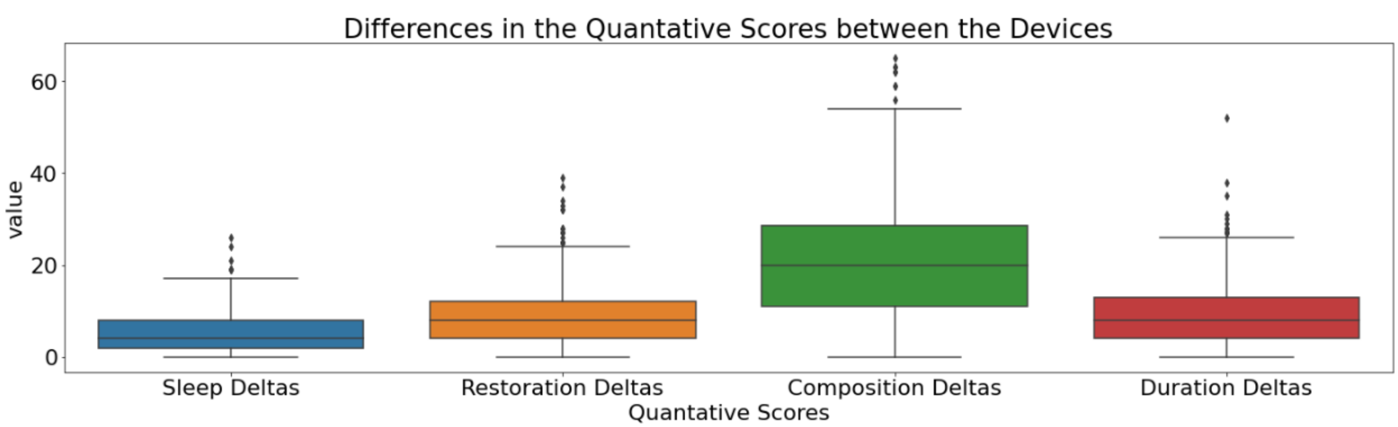

Fig. 3. Differences between the Quantitative Scores between the Devices.

The devices treat the data they collect differently from their sensors and have proprietary algorithms generating scores. The information is pretty clear that both devices are playing in the same ballpark but have different interpretations and application of the data. Of these measures, the two devices can vary greatly and frequently when assessing sleep stages REM and Deep sleep for the Quality Score. This differences is best illustrated in the deep sleep measure for the devices. Depending on how the devices deciphers REM and Non REM (Deep Sleep) is critical to the overall results for the Quality score. Looking at Deep sleep we can see in our that the correlation for deep sleep is only 17% between the devices.

B. Validation Set Analysis

The validation dataset is from another user. The data is applied only to validate the final models and is not incorporated into the training steps of the model development. The dataset was prepared like the primary dataset and applied directly to the model for evaluation without splitting the data. The validation set sleep scores median value of 71 with the interquartile range (IQR) with scores between 67 and 77. The data used for training and testing are a bit higher with a median score of 78 with an IQR between 72 and 83. Indicating that the sleep scores are similar but that the validation set tends to lean towards scores in the lower 70 range vs higher scores that tend towards values in closer to 80. See figure 4.

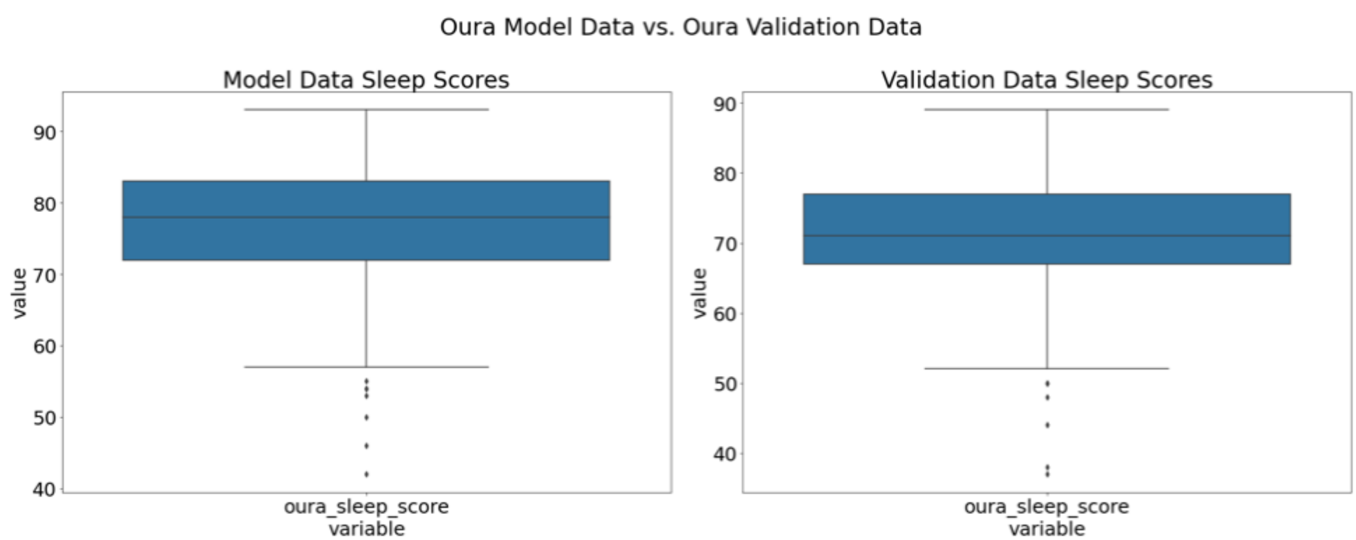

Fig. 4. Boxplot of Model and Validation Sleep Score

C. Sleep Score Model

After identifying viable modeling approaches, a Linear Regression model was used for the final model development. The model metrics evaluated against the validation data are r-squared of 0.999, MSE of 0.065, RMSE of 0.254, and a Max Error of 1. R-squared values of 0.999 appear to confirm that the published documentation for the factors aligns with the data provided in the application. See Figure 5 to see how the models visually performed and their metrics against the test and validation datasets.

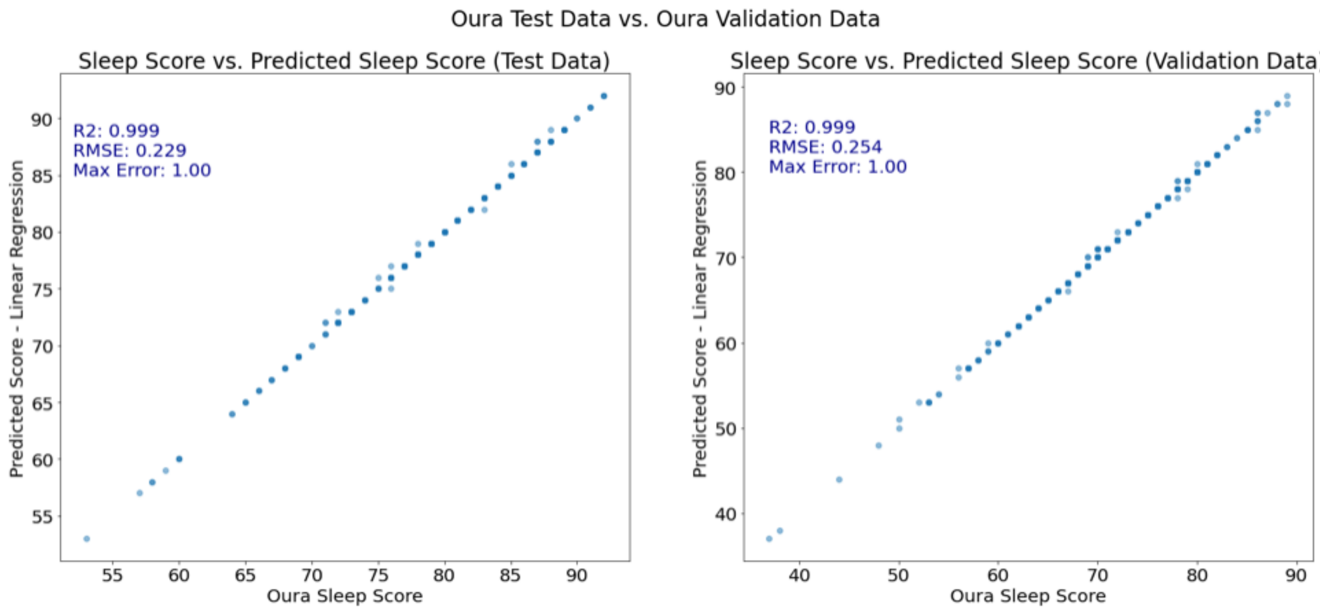

Fig. 5. Oura Sleep Score model results and metrics

The resulting coefficients of the model and their values indicate that the most influential factor is Total Sleep (.35), followed by Restfulness (.15). See Figure 6. When compared to the Fitbit algorithm both give the most weighting to Sleep Duration however Fitbit weights it at 50% of the score vs. 35% of the score for the Oura Ring. The Fitbit Quality score (combination or REM and DEEP sleep) are 25% of the score. The Oura Ring REM Score and Deep Sleep Score are 20% of the sleep score. The other 25% of the Fitbit score is the Restoration score comprised of Restfulness Score and Resting Heart Rate Score. The Oura Ring does not include resting heart rate as part of this sleep algorithm. The Oura Ring attributes 15% of the sleep algorithm score to the Restfulness Score. The Oura Rings attribute the rest of the 30% of the score split between Sleep Timing Score and Sleep Latency Score.

It is noted that Resting Heart Rate is handled differently between the two devices. Fitbit uses resting heart rate to inform its sleep score while Oura Ring uses the metric for its readiness score. The model coefficients inform us about the level of influence of each of these features on the score. When this model was evaluated with permutation feature importance against the measures for r2, MAE, RMSE and MSE, the Oura Total Sleep Score aligned with the story of its highest-valued coefficient. However, Oura Restfulness Score fell to the bottom across all measures. Oura Sleep Latency Score was second across all measures except for MAE where it was third. The Oura Deep Sleep was third across all measures except for MAE where it was second. The Shapley Additive Explanation (SHAP) summary plot also highlights

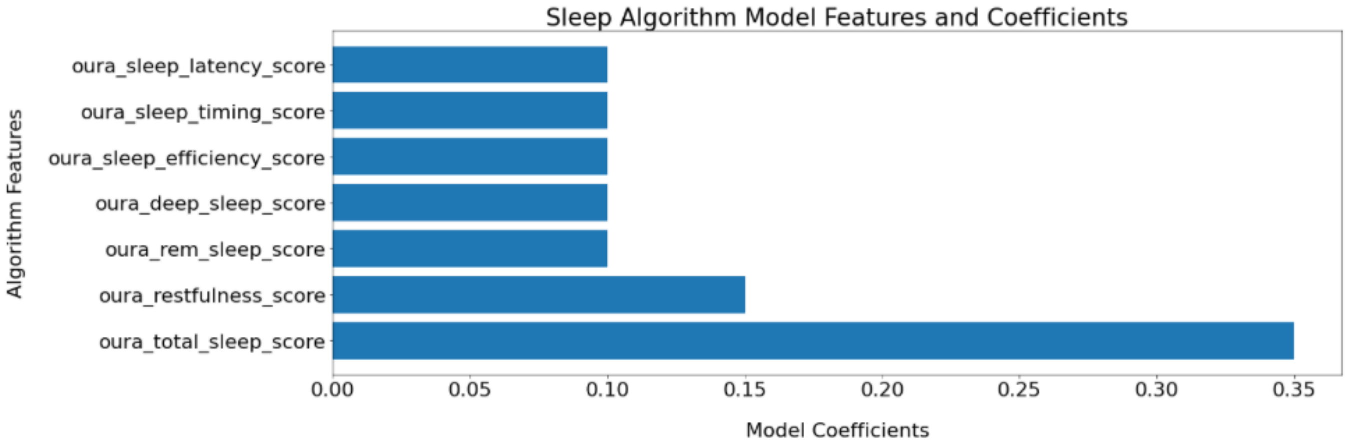

Fig. 6. Sleep Scores Factors and Model Coefficient Values

this relationship. Low and high scores for the Oura Total Sleep Score can significantly impact the overall score. The plot also indicates that the Oura Deep Sleep Score, Oura Sleep Latency Score and the Oura Sleep Timing have relatively the same impact and influence for high and low scores. Also interesting is that the Oura Restfulness score with the second-highest coefficient is also noted as having the least impact or importance on the overall Sleep Score. See figure 7 for the SHAP plot summary.

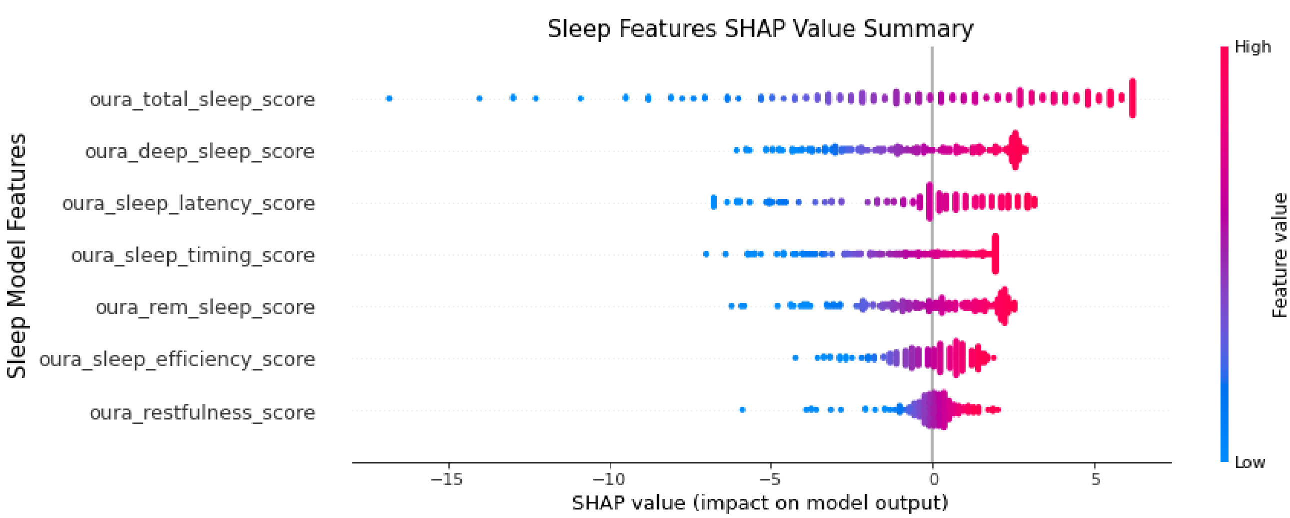

Fig. 7. Beeswarm plot, ranked by mean absolute SHAP Value

D. Readiness Score Model

After identifying viable modeling approaches a linear regression model was used for the final model development. The model metrics evaluated against the validation data are r-squared of 0.996, MSE of 0.233, RMSE of 0.483 and a Max Error of 2. The model is a linear regression model that incorporates

TABLE VII

REGRESSION MODEL EVALUATION METRICS: SLEEP

Source	Model	MSE	RMSE	MAE	R2
Orange 3	AdaBoost	0.021	0.144	0.021	1.00
Orange 3	Linear Regression	0.093	0.306	0.260	0.998
Orange 3	Stochastic Gradient Descent	0.096	0.309	0.260	0.998
Lazy Predict	LinearRegression	None	0.28	None	1.00
Lazy Predict	LinearSVC	None	0.28	None	1.00
Lazy Predict	RidgeCV	None	0.28	None	1.00
Lazy Predict	GeneralizedLinearRegressor	None	3.09	None	0.83
Lazy Predict	AdaBoostRegressor	None	2.88	None	0.85

feature transform of 2 degrees of Polynomial Features. See Figure 8 to see how the models visually performed and their metrics against the test and validation datasets

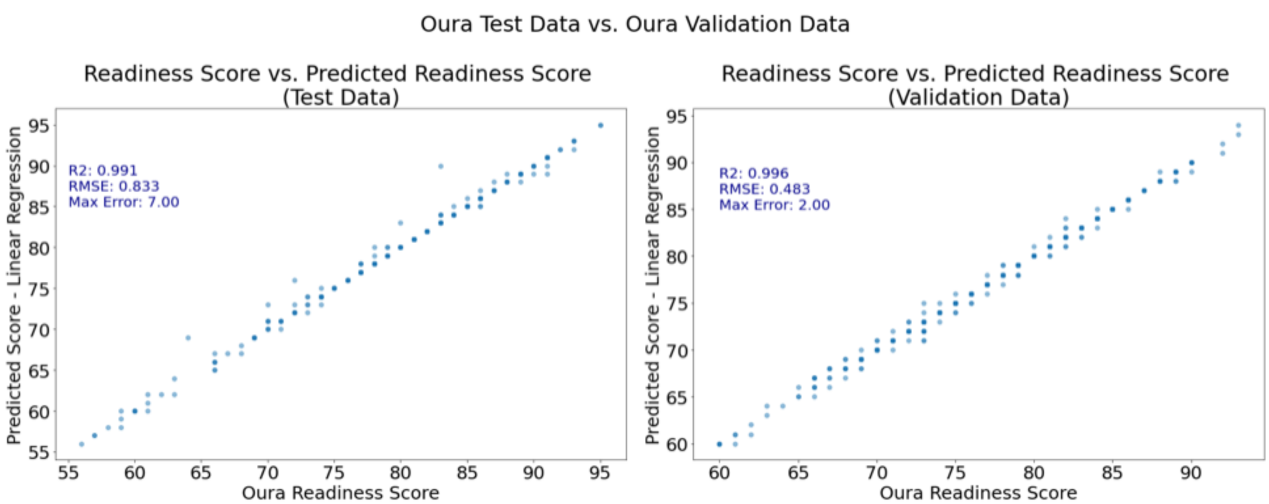

Fig. 8. Oura Sleep Score model results and metrics

The additional complexity introduced with the polynomial features makes using the coefficients directly and illustratively a challenge. For illustrative and comparison purposes we will be using the coefficients from the work by the linear model proposed by Rob Ter Horst, known as The Quantified Scientist or Rob Ter Horst, on his YouTube site as the Oura Ring Readiness Model[10]. This model performed with the following metrics against the validation set; r-squared of 0.996, MSE of 0.281, RMSE of 0.530 and a Max Error of 3. Rob's work suggest that the most influential coefficient is the Sleep Balance Score (.25) with Previous Night Score, Resting Heart Rate Score and HRV Balance Score all in second (.15). See Figure 9 for all of the coefficients from this Readiness model.

The referenced model provided some grounding into the coefficient weightings. When the model created

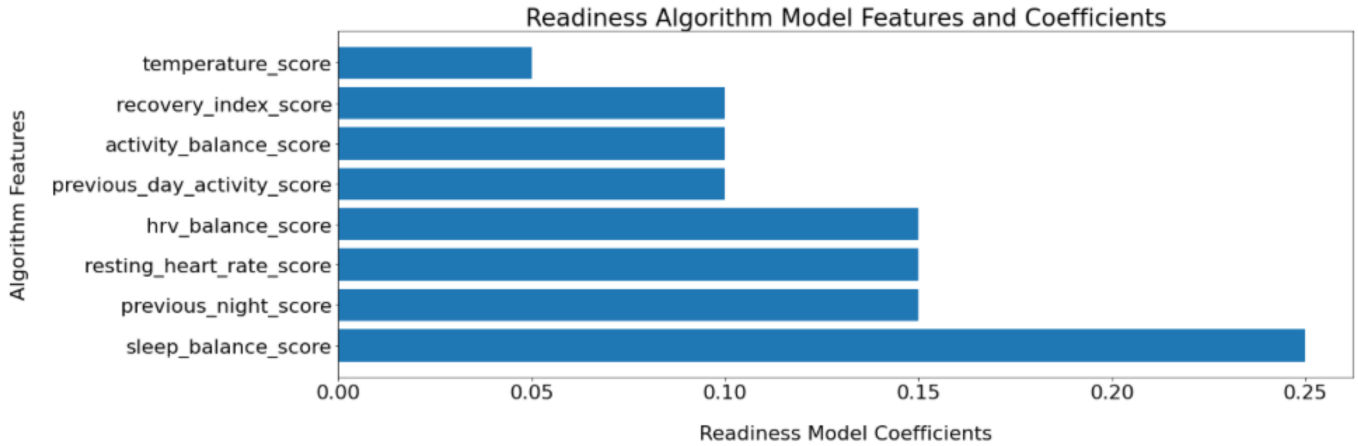

Fig. 9. Readiness Scores Factors and Model Coefficient Values from Rob Ter Horst

by this project is evaluated with permutation feature importance against the measures for r2, MAE, RMSE and MSE the Oura Sleep Balance Score aligned with its level of influence as the largest coefficient. The other factors for the other scores line up with the coefficient values. The permutation feature importance does highlight that HRV Balance Score is generally the next most influential to the Readiness Score. The Shapley Additive Explanation (SHAP) summary plot (Figure 10) also illustrates the influence of the Sleep Balance Score and the HRV Balance Score however it suggests that the HRV Balance Score as the most important feature. Also indicated by the coefficient values, the permutations feature importance scores and the SHAP summary plot is that the Activity Balance Score is a feature close to being the least important feature.

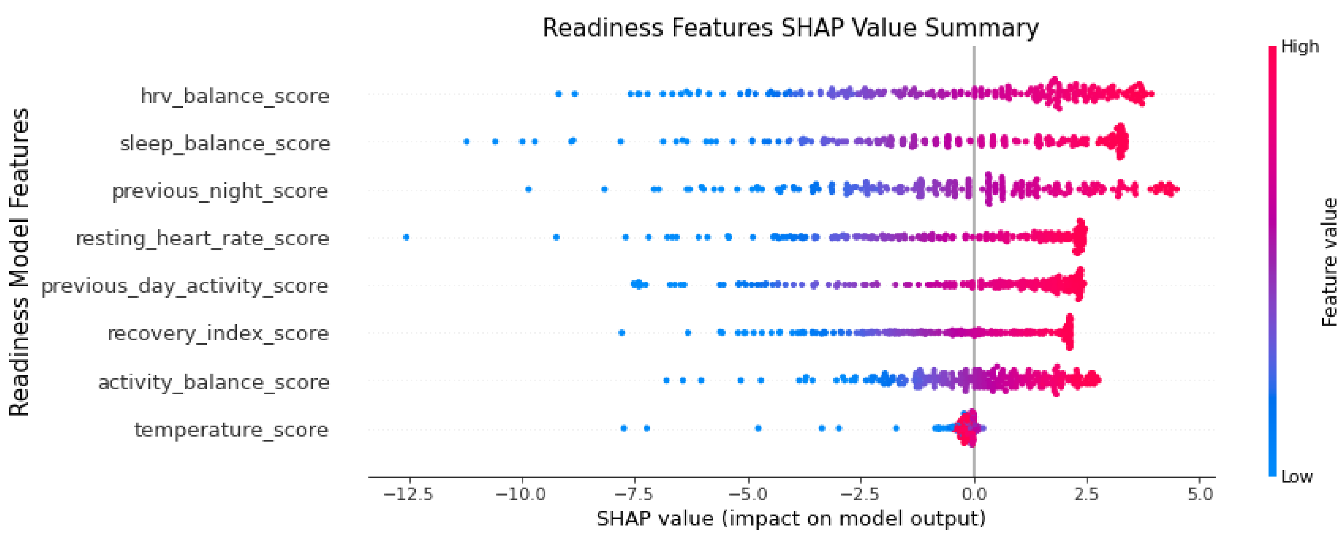

Fig. 10. Beeswarm plot, ranked by mean absolute SHAP Value

E. Subjective Readiness Scores

After ten days of capturing subjective readiness scores and comparing them to the device scores it provides a glimpse into how intuition and the devices scores compare. Fifty percent of the observations found that my perception of my readiness and the device score were about the same. Thirty percent of the time the device scores was a better reflection. These three scores were all score where the researcher felt pretty good at the start of the day and realized they were being conservative. Twenty percent of the time the researcher found that the subjective score was a much better reflection of personal readiness. These scores were both days that saw the impact of allergies that impacted the researchers subjective readiness score but according to the device were both days with mid 80 scores (85, 83). See 11

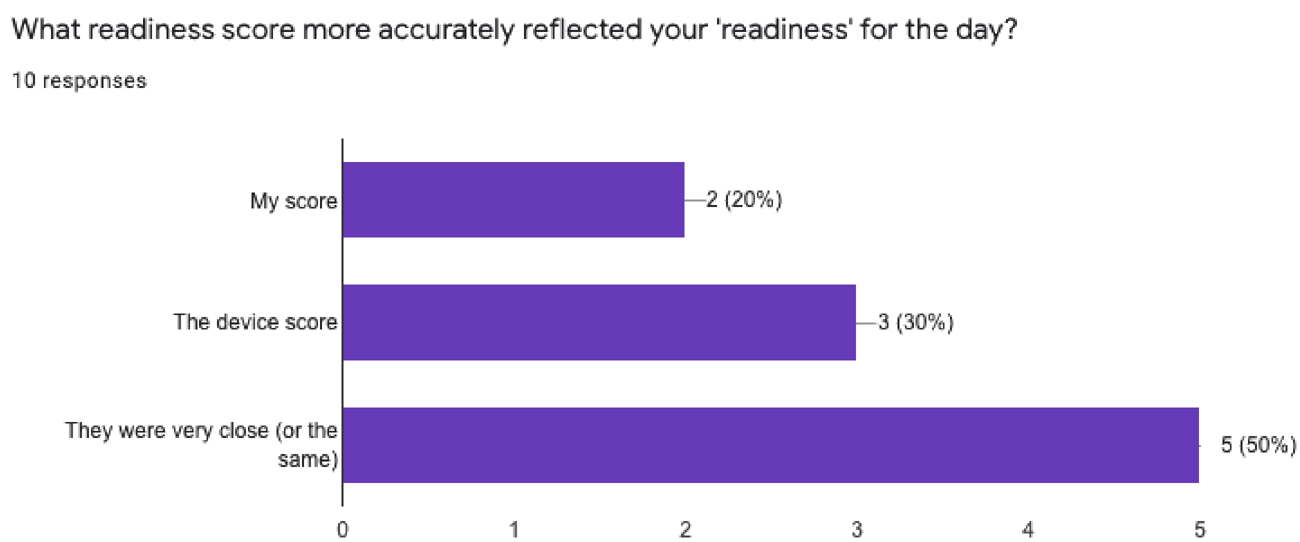

Fig. 11. Subjective Readiness Scores Compared to Device Scores

VII. Discussion of results

The wearable devices track similar data and produce similar quantitative scores. These devices can produce similar results at times, but generally, the results are very different. For example, a user would have to accept a +/- difference of about 7 points on the score before the results correspond 74% of the time. This research suggests that the primary difference in the sleep scores is determined by how the devices determine a user's sleep stages. These two devices varied greatly, as illustrated by deep sleep duration. The composition of the sleep models has some similarities. However, they differ regarding feature importance, and the Oura Ring sleep score has differing inputs than the Fitbit sleep model. The Oura Ring sleep model coefficients suggest that Oura Total Sleep is the most noteworthy feature, followed by the Oura Restfulness Score. However, in application, according to the results of permutation feature importance and the SHAP values, users should focus on getting plenty of sleep and creating an environment and a set of habits that improve their ability to fall asleep faster. The combined steps should maximize deep sleep and provide the best strategy for optimum sleep scores. The readiness model results highlight that activity plays a role in the overall Readiness metric, but its feature importance is low comparatively. Also, the results suggest that HRV Balance Score and Resting Heart Rate Score are high impact features for the model. The relative importance of these scores suggests that when a user removes their Oura Ring for activities that may damage the ring or are not safe (for example...playing basketball with a ring) not having the activity logged with the device is not going to significantly impact your score. However, a person's overall activity level will have a high level of influence on their Readiness Score. Therefore, when considering the overall interactions, the results suggest that if a person manages stress (mental and psychical) and routinely gets a good night's sleep, their readiness levels should be consistently good. The devices provide a lot of data; however, an individual's intuition may still be the best judge in managing readiness for the day. This works was a very small window of insight into work that could warrant additional observations in helping consumers use these products effectively. Also a result of this model development, it is clear that either different models or rule-based values are used to derive readiness scores when not all features are available. This work also suggests that the personalization of the scores is based on individual sensor scores and static algorithm(s) as opposed to the algorithms learning and adapting to the user.

VIII. Conclusions

This research deconstructed the quantitative scores for readiness and sleep using machine learning algorithms for the Oura Ring. Using the derived algorithms, we identified feature importance for these quantitative scores and considerations for Oura Ring users when interpreting their scores. This paper also compares the Fitbit fitness tracker Sleep Score and its features for comparisons and illustrative purposes between the products. This work highlighted some similarities and key differences between the devices. It also highlight the how to maximize your sleep and readiness scores based on representations of the proposed algorithms that govern these devices.

References

[1] "Oura Ring. Improve Sleep. Perform Better." Kickstarter, 21 Feb. 2016, https://www.kickstarter.com/projects/oura/oura-ring-improve-sleep-perform-better.

[2] "Fitbit Charge 4 User Manual." Fitbit,

[3] "Meet gen3 – Oura Help." Oura Ring, https://support.ouraring.com/hc/en-us/articles/4409072131091-Meet-Gen3.

[4] "Daily Readiness Score with Fitbit Premium." Daily Readiness Score, https://www.fitbit.com/global/us/technology/daily-readiness-score#: :text=What%20is%20Daily%20Readiness%20Score,how%20energized%20you%20feel%20today.

[5] "Chris Paul ." Ouraring.com, https://ouraring.com/meet-the-community.

[6] Godwin, Josh. "Oura Q&A with Crossfit Champion & Olympian Tia-Clair Toomey." The Pulse Blog, 5 Nov. 2021, https://ouraring.com/blog/oura-qa-with-crossfit-champion-olympian-tia-clair-toomey/.

[7] Staff, Harvard Health Publishing. "Heart Rate Variability: How It Might Indicate Well-Being." Harvard Health, 1 Dec. 2021, https://www.health.harvard.edu/blog/heart-rate-variability-new-way-track-well-2017112212789.

[8] Moreno-Pino, Fernando, et al. "Validation of Fitbit Charge 2 and Fitbit Alta HR against Polysomnography for Assessing Sleep in Adults with Obstructive Sleep Apnea." Journal of Clinical Sleep Medicine, 15 Nov. 2019, https://jcsm.aasm.org/doi/full/10.5664/jcsm.8032.

[9] Haghayegh1, Shahab, et al. "Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis." Journal of Medical Internet Research, JMIR Publications Inc., Toronto, Canada, https://www.jmir.org/2019/11/e16273/.

[10] R. Ter Horst, "Oura Ring Review: Readiness Score Decoded," YouTube. [Online]. Available: https://www.youtube.com/watch?v=6u9vGYo/-tPA&list=PLrpJQJ3tlvknGZp8kVr40kCUWAQ1YJI3s&index=8. [Accessed: 20-Mar-2022].

[11] "Use Nap Detection - Oura Help." Oura Ring, https://support.ouraring.com/hc/en-us/articles/1500009653181-Use-Nap-Detection.

[12] Hale, Tom, et al. "What Is HRV Balance & How to Increase It: Oura Ring." The Pulse Blog, 2 June 2021, https://ouraring.com/blog/hrv-balance/.

[13] "Fitbit Help." Fitbit MyHelp, https://help.fitbit.com/articles/en_US/Help_article/2439.htm#calculated.

[14] Bioinformatics Laboratory, University of Ljubljana. "Data Mining." Orange Data Mining - Data Mining, https://orangedatamining.com/.

[15] "AutoAI." Wikipedia, Wikimedia Foundation, 19 Feb. 2022, https://en.wikipedia.org/wiki/AutoAI.

[16] "Welcome to Lazy Predict's Documentation!" Welcome to Lazy Predict's Documentation! - Lazy Predict 0.2.9 Documentation, https://lazypredict.readthedocs.io/en/latest/.

[17] Olson, Randal S. TPOT, http://epistasislab.github.io/tpot/.

[18] "4.2. Permutation Feature Importance." Scikit, https://scikit-learn.org/stable/modules/permutation_importance.html

[19] "Welcome to the Shap Documentation." Welcome to the SHAP Documentation - SHAP Latest Documentation, https://shap.readthedocs.io/en/latest/index.html.

IX. Appendix

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
LinearSVR	1.00	1.00	0.28	0.02
LassoCV	1.00	1.00	0.28	0.11
Ridge	1.00	1.00	0.28	0.01
RANSACRegressor	1.00	1.00	0.28	0.02
RidgeCV	1.00	1.00	0.28	0.01
BayesianRidge	1.00	1.00	0.28	0.01
LarsCV	1.00	1.00	0.28	0.03
Lars	1.00	1.00	0.28	0.02
LassoLarsIC	1.00	1.00	0.28	0.02
LassoLarsCV	1.00	1.00	0.28	0.03
LinearRegression	1.00	1.00	0.28	0.01
TransformedTargetRegressor	1.00	1.00	0.28	0.01
HuberRegressor	1.00	1.00	0.28	0.04
ElasticNetCV	1.00	1.00	0.28	0.07
SGDRegressor	1.00	1.00	0.29	0.01
PassiveAggressiveRegressor	1.00	1.00	0.39	0.01
PoissonRegressor	0.99	0.99	0.60	0.01
GradientBoostingRegressor	0.96	0.96	1.51	0.09
OrthogonalMatchingPursuitCV	0.96	0.96	1.51	0.02
HistGradientBoostingRegressor	0.96	0.96	1.52	1.85
LGBMRegressor	0.96	0.96	1.54	0.52
ExtraTreesRegressor	0.94	0.94	1.77	0.21
XGBRegressor	0.92	0.92	2.11	0.34
RandomForestRegressor	0.92	0.92	2.15	0.27
KNeighborsRegressor	0.90	0.91	2.28	0.02
Lasso	0.90	0.90	2.36	0.02
BaggingRegressor	0.89	0.90	2.40	0.04
SVR	0.88	0.89	2.53	0.03
NuSVR	0.88	0.88	2.57	0.02
ElasticNet	0.86	0.86	2.76	0.01
AdaBoostRegressor	0.85	0.85	2.88	0.11
GammaRegressor	0.83	0.84	3.00	0.01
GeneralizedLinearRegressor	0.82	0.83	3.09	0.01
TweedieRegressor	0.82	0.83	3.09	0.01
DecisionTreeRegressor	0.79	0.80	3.38	0.01
ExtraTreeRegressor	0.73	0.74	3.83	0.02
OrthogonalMatchingPursuit	0.51	0.52	5.19	0.01
DummyRegressor	−0.06	−0.03	7.61	0.01
LassoLars	−0.06	−0.03	7.61	0.01
GaussianProcessRegressor	−4.15	−3.99	16.77	0.04
MLPRegressor	−6.65	−6.42	20.43	0.47
KernelRidge	−106.79	−103.46	76.69	0.02

Fig. 12. Lazy Predict Scores for Sleep Models

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
SGDRegressor	0.98	0.98	1.10	0.02
Lars	0.98	0.98	1.10	0.02
LinearRegression	0.98	0.98	1.10	0.02
TransformedTargetRegressor	0.98	0.98	1.10	0.02
LassoLarsIC	0.98	0.98	1.10	0.02
LassoLarsCV	0.98	0.98	1.10	0.04
LarsCV	0.98	0.98	1.10	0.04
BayesianRidge	0.98	0.98	1.10	0.02
Ridge	0.98	0.98	1.10	0.02
RidgeCV	0.98	0.98	1.10	0.02
LassoCV	0.98	0.98	1.10	0.11
ElasticNetCV	0.98	0.98	1.10	0.11
RANSACRegressor	0.98	0.98	1.20	0.02
PoissonRegressor	0.98	0.98	1.22	0.02
LinearSVR	0.97	0.97	1.40	0.02
HuberRegressor	0.97	0.97	1.42	0.04
PassiveAggressiveRegressor	0.97	0.97	1.43	0.02
GradientBoostingRegressor	0.95	0.95	1.93	0.13
XGBRegressor	0.94	0.94	2.10	0.52
LGBMRegressor	0.94	0.94	2.16	0.36
HistGradientBoostingRegressor	0.93	0.94	2.21	2.73
Lasso	0.92	0.92	2.41	0.02
ExtraTreesRegressor	0.91	0.91	2.58	0.25
KNeighborsRegressor	0.89	0.90	2.80	0.02
SVR	0.89	0.89	2.87	0.03
RandomForestRegressor	0.89	0.89	2.89	0.35
NuSVR	0.88	0.89	2.94	0.04
ElasticNet	0.88	0.88	3.03	0.02
BaggingRegressor	0.87	0.88	3.04	0.05
GammaRegressor	0.84	0.85	3.39	0.02
TweedieRegressor	0.84	0.85	3.44	0.02
GeneralizedLinearRegressor	0.84	0.85	3.44	0.02
OrthogonalMatchingPursuitCV	0.83	0.83	3.56	0.03
AdaBoostRegressor	0.80	0.81	3.84	0.13
DecisionTreeRegressor	0.70	0.71	4.70	0.02
ExtraTreeRegressor	0.66	0.67	5.00	0.02
OrthogonalMatchingPursuit	0.37	0.39	6.83	0.02
DummyRegressor	−0.04	−0.00	8.75	0.01
LassoLars	−0.04	−0.00	8.75	0.02
MLPRegressor	−4.31	−4.12	19.79	0.71
GaussianProcessRegressor	−5.56	−5.33	22.00	0.05
KernelRidge	−83.12	−80.11	78.78	0.03

Fig. 13. Lazy Predict Scores for Readiness Models

```
Permutation Feature Importance for Sleep Score using scores
-r2
-neg_mean_absolute_error
-neg_root_mean_squared_error
-neg_mean_squared_error


r2
  oura_total_sleep_score          0.764   +/- 0.045
  oura_sleep_latency_score        0.216   +/- 0.012
  oura_deep_sleep_score           0.212   +/- 0.013
  oura_sleep_timing_score         0.163   +/- 0.010
  oura_rem_sleep_score            0.133   +/- 0.009
  oura_sleep_efficiency_score     0.051   +/- 0.004
  oura_restfulness_score          0.032   +/- 0.002


neg_mean_absolute_error
  oura_total_sleep_score          4.838   +/- 0.168
  oura_deep_sleep_score           2.506   +/- 0.122
  oura_sleep_latency_score        2.318   +/- 0.079
  oura_sleep_timing_score         2.053   +/- 0.084
  oura_rem_sleep_score            1.881   +/- 0.086
  oura_sleep_efficiency_score     1.092   +/- 0.048
  oura_restfulness_score          0.703   +/- 0.029


neg_root_mean_squared_error
  oura_total_sleep_score          6.277   +/- 0.196
  oura_sleep_latency_score        3.215   +/- 0.100
  oura_deep_sleep_score           3.179   +/- 0.110
  oura_sleep_timing_score         2.758   +/- 0.096
  oura_rem_sleep_score            2.463   +/- 0.090
  oura_sleep_efficiency_score     1.433   +/- 0.060
  oura_restfulness_score          1.082   +/- 0.039


neg_mean_squared_error
  oura_total_sleep_score          43.005   +/- 2.558
  oura_sleep_latency_score        12.172   +/- 0.693
  oura_deep_sleep_score           11.926   +/- 0.758
  oura_sleep_timing_score         9.184   +/- 0.585
  oura_rem_sleep_score            7.476   +/- 0.496
  oura_sleep_efficiency_score     2.870   +/- 0.204
  oura_restfulness_score          1.786   +/- 0.106
```

Fig. 14. Feature Permutation Score

```
Permutation Feature Importance for Readiness Score using scores
-r2
-neg_mean_absolute_error
-neg_root_mean_squared_error
-neg_mean_squared_error


r2
  sleep_balance_score             0.256   +/- 0.021
  hrv_balance_score               0.234   +/- 0.017
  previous_night_score            0.190   +/- 0.012
  resting_heart_rate_score        0.180   +/- 0.013
  previous_day_activity_score     0.176   +/- 0.014
  recovery_index_score            0.116   +/- 0.006
  activity_balance_score          0.079   +/- 0.006
  temperature_score               0.027   +/- 0.003


neg_mean_absolute_error
  sleep_balance_score             3.008   +/- 0.170
  hrv_balance_score               2.934   +/- 0.117
  previous_night_score            2.687   +/- 0.117
  resting_heart_rate_score        2.416   +/- 0.111
  previous_day_activity_score     2.299   +/- 0.124
  recovery_index_score            2.007   +/- 0.074
  activity_balance_score          1.567   +/- 0.074
  temperature_score               0.204   +/- 0.024


neg_root_mean_squared_error
  sleep_balance_score             3.679   +/- 0.181
  hrv_balance_score               3.489   +/- 0.148
  previous_night_score            3.075   +/- 0.120
  resting_heart_rate_score        2.979   +/- 0.130
  previous_day_activity_score     2.938   +/- 0.148
  recovery_index_score            2.264   +/- 0.079
  activity_balance_score          1.771   +/- 0.089
  temperature_score               0.825   +/- 0.079


neg_mean_squared_error
  sleep_balance_score             19.611   +/- 1.614
  hrv_balance_score               17.930   +/- 1.276
  previous_night_score            14.524   +/- 0.931
  resting_heart_rate_score        13.790   +/- 0.978
  previous_day_activity_score     13.480   +/- 1.099
  recovery_index_score            8.854   +/- 0.484
  activity_balance_score          6.055   +/- 0.457
  temperature_score               2.042   +/- 0.253
```

Fig. 15. Feature Permutation Score