WGU C964

Task 2

Capstone: Readiness Score Prediction

Mark Nefzger

Student ID #001411596

10/12/2023

[Part A: Project Proposal for Business Executives 2](#_Toc147975623)

[Letter of Transmittal 2](#_Toc147975624)

[Part A: Project Recommendation 4](#_Toc147975625)

[Problem Summary 4](#_Toc147975626)

[Application Benefits 4](#_Toc147975627)

[Application Description 5](#_Toc147975628)

[Data Description 6](#_Toc147975629)

[Objectives and Hypothesis 6](#_Toc147975630)

[Methodology 7](#_Toc147975631)

[Funding Requirements 7](#_Toc147975632)

[Data Precautions 7](#_Toc147975633)

[Developer’s Expertise 7](#_Toc147975634)

[Part B: Project Proposal 8](#_Toc147975635)

[Problem Statement 8](#_Toc147975636)

[Customer Summary 9](#_Toc147975637)

[Existing System Analysis 10](#_Toc147975638)

[Data 11](#_Toc147975639)

[Project Methodology 11](#_Toc147975640)

[Project Outcomes 12](#_Toc147975641)

[Implementation Plan 13](#_Toc147975642)

[Evaluation Plan 13](#_Toc147975643)

[Resources and Costs 15](#_Toc147975644)

[Timeline and Milestones 16](#_Toc147975645)

[References 17](#_Toc147975646)

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

Mark Nefzger

Smart Solutions, LLC

Ōura Health Oy

Elektroniikkatie 10

Oulu, Finland

Dear Sir/Madam,

I am writing to submit a proposal to enhance your health monitoring systems. Our initiative, developed by an experienced machine learning expert, aims to transform lagging health indicators into leading ones, enabling individuals to make proactive health decisions.

The current approach, exemplified by the Ōura application, provides users with readiness scores based on lagging indicators. Our project leaps forward, incorporating real-time behaviors and estimated sleep cycles, predicting readiness scores using machine learning algorithms.

The total implementation cost is $19,200, including all development and deployment expenses. Mark Nefzger, with a degree in Computer Science, is an expert in machine learning.

Key benefits for Ōura Health Oy:

1. **Enhanced User Engagement**: Users will engage more with the application due to precise, real-time assessments, leading to higher customer satisfaction.
2. **Competitive Advantage**: Embrace this predictive model for a significant edge in the health monitoring industry.
3. **Improved Health Outcomes**: Empower individuals to proactively manage their health, benefiting users and public health outcomes.
4. **Revenue Growth**: Offer more value to users, potentially increasing subscription rates and expanding your user base.

Please find the detailed proposal enclosed. We believe this initiative can revolutionize health monitoring and bring substantial benefits to Ōura Health Oy. For further information, contact me at 208-681-6783 or [mark.nefzger@wgu.edu](mailto:mark.nefzger@wgu.edu).

Thank you for your time. We look forward to discussing this initiative and potential collaboration.

Sincerely,

Mark Nefzger

Computer Scientist

Smart Solutions, LLC

208-681-6783

[mark.nefzger@wgu.edu](mailto:mark.nefzger@wgu.edu)

## Part A: Project Recommendation

## Problem Summary

The current health monitoring system, powered by smart technology and the Oura application, offers valuable insights into our well-being, but it predominantly relies on lagging indicators, such as Resting Heart Rate, Heart Rate Variability Balance, and more. This approach provides a retrospective 'readiness score' that reflects the previous day's activities and sleep cycle, limiting its real-time relevance.

To enhance users' health outcomes, this project will address this limitation by introducing predictive coding through machine learning. The problem will address the absence of a predictable readiness value based on real-time behaviors and user-estimated sleep duration. Users will input their estimated sleep duration, and the algorithm will calculate a predicted readiness score.

Our goal is to provide a readiness score factoring in leading indicators such as Sleep Balance, Sleep Regularity, Previous Day Activity, and Activity Balance, alongside the conventional lagging indicators. By doing so, individuals will be empowered to make timely decisions and take actions to optimize their health and well-being.

This project seeks to bridge the gap between current health data analysis and real-time health insights, ensuring that our users can make informed choices for better health outcomes in the future.

## Application Benefits

This project's initiative to enhance the health monitoring system with predictive coding brings forth a host of benefits:

* **Enhanced User Engagement:** Users will experience a significant increase in engagement with the application. They will be drawn to its precise, real-time assessments, ultimately leading to higher customer satisfaction as they gain a deeper, more relevant understanding of their health status.
* **Competitive Advantage:** This project secures a substantial competitive advantage within the health monitoring industry by embracing this predictive model. This project's forward-looking approach sets it apart from competitors and positions it as a leader in the field.
* **Improved Health Outcomes:** This project's predictive model empowers individuals to proactively manage their health. This empowerment benefits this project's users by enabling them to make informed, timely decisions for their well-being and contributes to improved public health outcomes. A healthier population can lead to reduced healthcare costs and enhanced overall community well-being.
* **Revenue Growth:** By offering more value to this project's users through accurate, real-time insights and the potential to make better-informed health choices, this project could increase subscription rates and attract new users. This, in turn, promises significant revenue growth for this project.

In sum, this project is not just a technological enhancement; it's a strategic move that will enhance user engagement, secure a competitive edge, improve health outcomes, and fuel revenue growth, thereby benefiting both the users and the organization.

## Application Description

This application will transform the way users monitor their readiness for the day ahead using leading indicators, with a focus on predicting readiness scores based on estimated sleep cycles. Data for this project will be sourced from the Oura website (<https://cloud.ouraring.com/trends>). The ‘Readiness Score’ and ‘Total Sleep Duration’ fields will be shifted by one day to align with their leading indicators.

To prepare the data for machine learning analysis, we will undertake several crucial steps:

* Initial data exploration will reveal the mean ‘Readiness Score,’ providing a baseline for the project.
* To create a training dataset with leading indicators, all data unavailable before the user's sleep cycle will be removed.
* Additionally, any null data will be purged from the dataset.
* Various charts and diagrams will be used to identify potential correlations between data fields and the ‘Readiness Score.’

For the machine learning component, supervised regression will be chosen as the method to predict readiness scores. Options considered will include linear regression, multiple linear regression, and polynomial regression, which are commonly used regression techniques in machine learning.

## Data Description

User data is available from the Oura website for individual users. This data is available for download (<https://cloud.ouraring.com/trends>) for individual users. Date ranges are selectable allowing sufficient data to be extracted. For this project, the ‘readiness score’ will be the variable that the machine learning will be trained to predict. The other data will be evaluated to determine if it is a leading or lagging indicator of ‘readiness score’. Finally, the data that provides a leading indication will be evaluated for its predictive value in the machine learning algorithm.

## Objectives and Hypothesis

The hypothesis of this proposal is to use machine learning to provide a leading or predictive value of a user’s ‘Readiness Score’. The data from a user will be transformed and evaluated for useful leading indicators of readiness. Various machine learning techniques along with varying data will be used to train the algorithm to maximize the accuracy of the prediction.

## Methodology

This project will be developed using Agile project management methods. This will allow an iterative approach to developing the application while allowing for user feedback. Phases will be divided as follows:

* Proposal accepted. Proof of concept performed.
* Data set selected. Data will be explored using various visualization techniques for relationships. Data will be modified and transformed in preparation for a machine-learning algorithm.
* A machine learning model will be developed and assessed.
* Project will be documented including machine learning code.

## Funding Requirements

A rough order of magnitude (ROM) estimate has projected project costs to be $19,200. This includes one programmer for four weeks of programming and documentation. Additional costs for hardware and software are not expected.

## Data Precautions

* All personally identifiable information (PII) will be removed from the data.

## Developer’s Expertise

The developer holds dual Bachelor of Science Degrees in Computer Science and Electrical Engineering. Areas of expertise related to this project include:

* Machine learning applications
* Relational databases
* Python programming
* Project Management
* Engineering Management

# Part B: Project Proposal

## Problem Statement

The current landscape of health monitoring and readiness assessment for individuals relies on lagging indicators provided by smart technology, notably the Oura application. While these indicators, such as Resting Heart Rate, Heart Rate Variability Balance, Body Temperature, Recovery Index, and Sleep, offer valuable insights, they inherently reflect past data, leading to a delayed understanding of overall health. To address this limitation, this proposal aims to leverage machine learning to create predictive coding that can provide individuals with leading indicators for readiness based on real-time behaviors and estimated sleep cycles.

Currently, the Oura application offers a readiness score in the morning following sleep, which is a commendable prediction of an individual's current health state. However, the project's objective is to equip users with a readiness score that can be anticipated in advance, influenced by their real-time actions, and estimated sleep patterns. Users input their estimated sleep duration, and the algorithm calculates a predictive readiness score, effectively offering a personalized, real-time assessment of their health.

To illustrate the concept further, on a chosen date, the project will generate a predicted readiness score based on the user's provided estimated sleep duration. This development seeks to redefine the approach to health monitoring, enabling individuals to take proactive measures to improve their overall well-being and health outcomes.

The readiness score incorporates both lagging and leading indicators, further emphasizing the shift towards real-time predictive assessment. Lagging indicators include Resting Heart Rate, Heart Rate Variability Balance, Body Temperature, Recovery Index, and Sleep. Leading indicators encompass Sleep Balance, Sleep Regularity, Previous Day Activity, and Activity Balance, all offering a holistic view of an individual's health, considering not only historical data but also the latest behaviors and sleep patterns.

The challenge at hand, therefore, is to design, implement, and refine the predictive coding system that seamlessly integrates with existing smart technology and accurately delivers real-time readiness scores. This solution will require the design of machine learning algorithms, ensure data accuracy, and create a user-friendly interface that empowers individuals to make informed decisions about their health and well-being. The goal is to revolutionize health monitoring by providing individuals with a reliable, anticipatory health assessment that can guide them toward healthier and more informed lifestyle choices.

## Customer Summary

The primary customer for the problem statement described is individuals who are actively interested in monitoring and improving their health and well-being using smart technology. These individuals typically have access to wearable devices and applications that track various biological data, such as heart rate, sleep patterns, body temperature, and activity levels. They are looking for ways to gain deeper insights into their health and receive timely, predictive assessments of their readiness or overall health status.

Characteristics of the customer for this problem statement include:

**Tech-Savvy:** The customer is comfortable with and proficient in using technology, particularly wearable devices, and mobile applications for health monitoring. They may already be using or have experience with tools like the Oura application.

**Proactive Health Enthusiasts:** These individuals are proactive about their health and well-being. They are motivated to make lifestyle changes based on data-driven insights and are interested in optimizing their health outcomes.

**Data-Driven Decision-Makers:** The customer values data and wants access to advanced analytics and machine learning-driven predictions. They seek to make informed decisions about their daily activities, exercise routines, and sleep patterns based on this data.

**Personalization:** They appreciate personalized solutions and are willing to input data, such as estimated sleep duration, to receive tailored predictions about their readiness and overall health.

**Early Adopters:** These individuals are likely early adopters of new health and wellness technologies and are willing to embrace innovative solutions that go beyond traditional lagging indicators to provide predictive health insights.

The proposed solution aims to cater to this customer group by offering them the capability to move beyond retrospective health insights and transition into the realm of proactive health management, thereby helping them lead healthier and more informed lives.

## Existing System Analysis

Current solutions largely rely on descriptive analytics to provide historical data and trends. The proposed solution incorporates machine learning, enabling it to learn from user behavior and generate predictive models, thus providing more sophisticated and actionable insights.

The existing solutions focus on retrospective analysis, providing users with a summary of their health status. The proposed solution encourages proactive health management, empowering users to make informed choices and adjustments in real-time to improve their health outcomes.

The proposal allows users to input their estimated sleep duration and calculates readiness scores based on this information. This customization and personalization make the solution more adaptable to individual needs, while many existing solutions offer one-size-fits-all assessments.

Current solutions often provide health data, which may not lead to immediate action. In contrast, the proposed solution's predictive readiness scores are designed to drive actionable insights, guiding users to make choices that positively impact their health and well-being.

The proposal represents an innovative approach to health monitoring, introducing predictive coding and combining both lagging and leading indicators. Existing solutions, while valuable, typically stick to established methods of health assessment and monitoring.

The proposed solution may incorporate user feedback and data to improve its predictive models over time. This iterative approach to development and enhancement sets it apart from many static, one-time health assessment tools.

## Data

User data will be available from the Oura website for individual users. This data will be available for download (<https://cloud.ouraring.com/trends>) for individual users. Date ranges will be selectable, allowing sufficient data to be extracted. As a precaution for sensitive data, personally identifiable information (PII) will be removed. Null data will be required to be removed before training a machine learning algorithm.

As a note, because a previous day’s health activities impact the next day’s readiness score, the data may need to be shifted to align leading indicators with actual results when training the machine learning model.

## Project Methodology

The undertaking of this project will be meticulously orchestrated through the implementation of Agile project management methodologies. By adopting Agile principles, we will ensure an iterative and highly adaptive approach to the development of the application, thereby fostering the incorporation of continuous user feedback throughout the project's lifecycle. To provide a comprehensive overview of the project's structure and progression, it will be divided into distinct phases, each crucial in achieving the project's objectives:

**Proposal Will Be Accepted and Proof of Concept**: In this initial phase, the project will kick off with the formal acceptance of the proposal. Simultaneously, a proof of concept will be performed to validate the feasibility and potential success of the project. This phase will establish the project's foundation and confirm its viability.

**Data Set Selection and Exploration**: Following the project's acceptance, the next phase will be dedicated to selecting an appropriate data set. This data will be subjected to a rigorous exploration process, employing a diverse array of visualization techniques to uncover intricate relationships within the dataset. Furthermore, data will undergo meticulous modification and transformation to prepare it for subsequent application in machine learning algorithms. This phase will serve as the basis for the data-driven aspects of the project.

**Machine Learning Model Development and Assessment**: With the preprocessed data in hand, the focus will shift to the development of sophisticated machine learning models. These models will be carefully designed and fine-tuned to meet the project's specific requirements. Subsequently, rigorous assessment and validation procedures will be conducted to ensure the models' accuracy and effectiveness. This phase will be the crux of the project, where advanced algorithms and methodologies will be harnessed to extract meaningful insights from the data.

**Project Documentation and Machine Learning Code**: The final phase will center on comprehensive documentation. This will encompass all facets of the project, including the machine learning code. This documentation will provide a valuable resource for future reference, replication, and knowledge sharing.

By adhering to Agile practices and adopting this structured, phased approach, this project will not only guarantee adaptability but also ensure that it remains responsive to evolving user requirements. Throughout each of these distinct phases, the project will actively engage with stakeholders to foster collaboration and deliver an application that will meet both technical excellence and user satisfaction.

## Project Outcomes

The deliverables of this project will include the following:

* A machine learning algorithm that is trained on leading health indicators to accurately predict a ‘readiness score’. This shall be based on an estimated sleep cycle that is input by the user of the algorithm.
* A user manual that describes the installation and operation of the application sufficiently to navigate a user familiar with machine learning programming to be able to use the application (a user manual for the general user is beyond the scope of this project).

## Implementation Plan

This project will be developed using Agile project management methods. This will allow an iterative approach to developing the application while allowing for user feedback. Phases will be divided as follows:

**Phase 1** - Proposal accepted. A proof of concept will be performed to verify the project should continue.

**Phase 2** - Data will be set and selected. Data will be explored using various visualization techniques for relationships. Data will be modified and transformed in preparation for the machine learning algorithm.

**Phase 3** - The machine learning model will be developed and assessed. This is the phase where incremental testing will occur. Customer feedback will be sought and applied to the design as the design incrementally approaches completion.

**Phase 4** - The project will be documented including the machine learning code. The final product will be delivered to the customer.

## Evaluation Plan

1. Data Collection and Preprocessing

* **Data Source Verification**: Ensure that the data source is reputable, accurate, and relevant to the problem at hand. Verify that data collection methods are sound.
* **Data Preprocessing Verification**: Verify that data preprocessing steps are correctly implemented. This includes checking for missing data, confirming proper encoding of categorical variables, and validating feature scaling/normalization.

2. Model Evaluation Metrics

* **Metric Selection Verification**: Verify that the chosen evaluation metric (coefficient of determination of the prediction) is appropriate for the specific problem and the nature of the data. Ensure that the metric aligns with the goals of the project.

3. Initial Model Performance

* **Accuracy Verification**: Verify that the initial model's accuracy is calculated correctly using the selected metric (coefficient of determination of the prediction). Ensure that the code for calculating the accuracy is accurate and functioning properly.

4. Model Comparison

* **Baseline Verification**: Confirm that the baseline or previous methods used for comparison are well-documented and reproducible. Ensure that the comparison is made using consistent criteria.

5. Performance Analysis

* **Predicted vs. Actual Verification**: Verify that the predicted readiness scores are compared accurately to actual readiness scores. Ensure that the code for calculating and comparing these scores is correct.

6. Areas for Improvement

* **Data Collection Verification**: Ensure that plans to collect additional data on sleep quality, caffeine intake, alcohol consumption, and eating habits are feasible and will improve the dataset.
* **Model Adjustment Verification**: Confirm that the iterative testing process for model adjustments is well-documented and that each adjustment is validated to improve predictions.

7. Application Testing

* **Iterative Testing Verification**: Ensure that code changes made during iterative testing are properly tracked and tested systematically.
* **Dataset Adjustment Verification**: Verify that dataset adjustments made during testing are documented and that the impact on model accuracy is accurately measured.
* **Debugging Verification**: Confirm that errors or issues identified during testing are logged, traced, and resolved systematically.

8. Future Steps

* **Future Data Collection Verification**: Ensure that plans for future data collection are practical and will enhance the dataset. Verify that data collection methods are ethical and maintain data privacy.
* **Model Refinement Verification**: Document the changes made to the model and ensure that each refinement is rigorously tested and evaluated.
* **Performance Monitoring Verification**: Develop a process for ongoing performance monitoring and verification of the model's effectiveness as new data becomes available.

## Resources and Costs

|  |  |  |
| --- | --- | --- |
| Resource | Description | Cost |
| Machine Learning Developer | 40 hours for four weeks ($120/hr) | $19,200 |
| Computer | Existing resource | $0 |
| Software | Open-source Spyder, Jupyter Notebook, and other APIs | $0 |
|  | **Total Cost:** | $19,200 |

## Timeline and Milestones

|  |  |  |  |
| --- | --- | --- | --- |
| Sprint | Start | End | Tasks |
| 1 | 9/11/23 | 9/18/23 | Proposal accepted. Proof of concept performed. |
| 2 | 9/18/23 | 9/25/23 | Data set selected. Data was explored using various visualization techniques for relationships. Data was modified and transformed in preparation for a machine learning algorithm. |
| 3 | 9/24/23 | 10/2/23 | Machine learning model developed and assessed. |
| 4 | 10/2/23 | 10/13/23 | The project is documented including machine learning code. |

## References

1. Ōura Health Oy (n.d.). Oura mission. Oura. Retrieved September 4, 2023, from https://ouraring.com/about-us.

# Part C: Application

Included in this application are the following files:

* Capstone.ipynb
* Capstone.py
* requirements.txt

Data:

* oura\_2019-01-01\_2023-09-09\_trends\_Original.csv
* oura\_2019-01-01\_2023-09-09\_trends\_Shifted.csv
* oura\_2023-02-20\_2023-02-20\_trends.csv
* oura\_2023-02-21\_2023-02-21\_trends

Images:

* Figure\_1.png
* Figure\_2.png
* Figure\_3.png
* Figure\_4.png
* Figure\_5.png
* Figure\_6.png
* Figure\_7.png
* Figure\_8png
* Figure\_9.png
* Figure\_10.png
* Figure\_11.png
* Figure\_12.png

WGU C964

Task 2

Capstone: Readiness Score Prediction

Post Implementation Report

Mark Nefzger

Student ID #001411596

10/12/2023

[Part D: Post-Implementation Report 2](#_Toc148012718)

[Project Overview 2](#_Toc148012719)

[Datasets 4](#_Toc148012720)

[Data Product Code 4](#_Toc148012721)

[Hypothesis Verification 7](#_Toc148012722)

[Effective Visualization and Reporting 9](#_Toc148012723)

[Accuracy Analysis 16](#_Toc148012724)

[Application Testing 17](#_Toc148012725)

[Application Files 18](#_Toc148012726)

[User Guide 19](#_Toc148012727)

[Summation of Learning Experience 20](#_Toc148012728)

[References 21](#_Toc148012729)

# Part D: Post-Implementation Report

## Project Overview

This proposal described the creation of predictive coding to improve health outcomes for individuals using smart technology to monitor their various biological data. Currently, this data is providing a lagging indicator of overall health by providing an overall readiness score following the previous day’s activities and a sleep cycle. This proposal used machine learning to provide a leading or predictive value of this score.

Currently, individuals are provided a readiness score from the Oura application in the morning after waking up (Ōura Health Oy (n.d.). *Oura mission*. Oura. Retrieved September 4, 2023, from <https://ouraring.com/about-us>). While this is a helpful prediction of the current state of overall health, the goal of this project was to provide a predictable readiness value for individuals based on real-time behaviors and estimated sleep cycles. The estimated sleep duration is input by the user and the algorithm calculates a predicted readiness score. For this project, a typical day was selected (2/20/2023) and the predicted readiness score is output for the user based on their estimated duration of sleep.

A screenshot of a sleep test

Description automatically generatedThe readiness score is based on the following lagging indicators:

* Resting Heart Rate (detected while sleeping)
* Heart Rate Variability Balance (detected while sleeping)
* Body Temperature (detected while sleeping)
* Recovery Index (detected while sleeping)
* Sleep

The readiness score is based on the following leading indicators:

* Sleep Balance (based on the past two weeks)
* Sleep Regularity (based on the past two weeks)
* Previous Day Activity (previous day)
* Activity Balance (based on past days)

## Datasets

User data is available from the Oura website for individual users. This data is available for download (<https://cloud.ouraring.com/trends>) for individual users. Date ranges are selectable allowing sufficient data to be extracted. One disadvantage to this data extraction method is that data cannot be selected from other users. This data is only available within the company database or for the specific user. As a precaution for sensitive data, personally identifiable information (PII) was removed.

The full downloaded dataset is included with the files provided:

New Data/oura\_2019-01-01\_2023-09-09\_trends\_Original.csv

## Data Product Code

Data was obtained from the Oura website (<https://cloud.ouraring.com/trends>). Since the hypothesis of this product is to use leading indicators to predict a readiness score for the next day based on an estimated sleep cycle, both the ‘Readiness Score’ and ‘Total Sleep Duration’ fields were shifted by one day to align with their leading indicators. This data was loaded into the program as:

New Data/oura\_2019-01-01\_2023-09-09\_trends\_Shifted.csv

Once the raw data was imported, the ‘Total Sleep Duration’ and ‘Rest Time’ fields were converted from seconds to hours (3600 seconds per hour):

df[‘Total Sleep Duration’] = df[‘Total Sleep Duration’] / 3600  
df[‘Rest Time’] = df[‘Rest Time’] / 3600

The mean ‘Readiness Score’ was determined to be 77.17 using the following code:

health\_data.describe()

As this project was based on developing an algorithm using leading indicators, all data that would not be available before the user sleeping were removed from the training dataset. An example would be ‘Deep Sleep Score’ which is not available before going to sleep.

Any null data was removed from the data set. All data was transformed to a type usable to the machine learning language. The ‘Date’ field was an object datatype. This was initially transformed to a datetime datatype, but it was later determined that this field was not required to train the algorithm.

Various charts and diagrams were utilized to look for a correlation of data fields to the ‘Readiness Score’.

The algorithm used for this machine learning application was supervised learning using regression. Regression within machine learning includes linear regression, multiple linear regression, and polynomial regression (Kurama, V., & Powers, J. (2023, February 28). Regression in Machine Learning: What it Is and Examples of Different Models. Builtin.com. Retrieved September 4, 2023, from <https://builtin.com/data-science/regression-machine-learning>). These methods were used to solve this problem.

One approach to regression is called random forests. This is a classification and regression approach that displayed outstanding performance regarding prediction error on a suite of benchmark datasets (Segal, M. R. (2003, April 14). Machine Learning Benchmarks and Random Forest Regression. EScholarship - Open Access Publications from the University of California. Retrieved September 4, 2023, from <https://escholarship.org/uc/item/35x3v9t4>). This algorithm was used to train this machine learning application, but ultimately showed a poor accuracy (0.055) and was not used.

from sklearn.ensemble import RandomForestClassifier  
model = RandomForestClassifier()  
  
# Setup random seed  
np.random.seed(42)  
  
# Create the data  
df.dropna(inplace=True)  
X = df.drop(‘Readiness Score’, axis=1)  
y = df[‘Readiness Score’] #target  
  
# Split into train and test sets  
from sklearn.model\_selection import train\_test\_split  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2)

model.get\_params()

{'bootstrap': True,  
 'ccp\_alpha': 0.0,  
 'class\_weight': None,  
 'criterion': 'gini',  
 'max\_depth': None,  
 'max\_features': 'sqrt',  
 'max\_leaf\_nodes': None,  
 'max\_samples': None,  
 'min\_impurity\_decrease': 0.0,  
 'min\_samples\_leaf': 1,  
 'min\_samples\_split': 2,  
 'min\_weight\_fraction\_leaf': 0.0,  
 'n\_estimators': 100,  
 'n\_jobs': None,  
 'oob\_score': False,  
 'random\_state': None,  
 'verbose': 0,  
 'warm\_start': False}

model.fit(X\_train, y\_train);

y\_preds = model.predict(X\_test)

model.score(X\_test, y\_test)

A second algorithm was attempted using ridge regression and yielded an improved accuracy (0.389).

# Try Ridge Regression  
  
from sklearn.linear\_model import Ridge  
  
model = Ridge()  
model.fit(X\_train, y\_train)  
  
# Check the score of the model (on the test set)  
model.score(X\_test, y\_test)

0.7083864563369104

model.get\_params()

{'alpha': 1.0,  
 'copy\_X': True,  
 'fit\_intercept': True,  
 'max\_iter': None,  
 'positive': False,  
 'random\_state': None,  
 'solver': 'auto',  
 'tol': 0.0001}

Since the dataset that is logged within the Oura application is large, a relevant subset of the data was needed to be extracted. One method of regression that performs this function is called the least absolute shrinkage and selection operator (LASSO). It is an analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical regression model (Emmert-Streib, F., & Dehmer, M. (2019, January 14). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. MDPI. Retrieved September 4, 2023, from <https://www.mdpi.com/2504-4990/1/1/21>). A version of this method using the LARS algorithm was evaluated as part of this project ([Scikit-learn: Machine Learning in Python](https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011). This algorithm also yielded the highest accuracy (0.389) and was ultimately selected for this project.

from sklearn import linear\_model  
model = linear\_model.LassoLars(alpha=1.0)  
model.fit(X\_train, y\_train)  
  
# Check the score of the model (on the test set)  
model.score(X\_test, y\_test)

## Hypothesis Verification

The hypothesis of this proposal was to use machine learning to provide a leading or predictive value of a user’s ‘Readiness Score’. The data from a user was transformed and evaluated for useful leading indicators of readiness. The following datasets were selected to train the machine learning algorithm: Total Sleep Duration, Temperature Trend Deviation, Activity Score, Rest Time, Previous Night Score, Sleep Balance Score, Previous Day Activity Score, and Activity Balance Score.

Various linear regression techniques were used to train the algorithm. The final selected algorithm was based on the least absolute shrinkage and selection operator (LASSO) and yielded machine learning code that accurately predicts user readiness.

## Effective Visualization and Reporting

This project focused on the Readiness Score. The following graph shows a histogram of readiness scores for the entire dataset:

(df[‘Readiness Score’].hist(figsize=(10, 10)))

<Axes: >

A graph with a number of blue squares

Description automatically generated with medium confidence

The below lists the data names, non-null count, and datatypes:

health\_data.info()

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 970 entries, 0 to 969  
Data columns (total 54 columns):  
 # Column Non-Null Count Dtype   
--- ------ -------------- -----   
 0 date 970 non-null object   
 1 Sleep Score 898 non-null float64  
 2 Total Sleep Score 898 non-null float64  
 3 REM Sleep Score 898 non-null float64  
 4 Deep Sleep Score 898 non-null float64  
 5 Sleep Efficiency Score 898 non-null float64  
 6 Restfulness Score 898 non-null float64  
 7 Sleep Latency Score 898 non-null float64  
 8 Sleep Timin Score 898 non-null float64  
 9 Total Sleep Duration 896 non-null float64  
 10 Total Bedtime 896 non-null float64  
 11 Awake Time 896 non-null float64  
 12 REM Sleep Duration 896 non-null float64  
 13 Light Sleep Duration 896 non-null float64  
 14 Deep Sleep Duration 896 non-null float64  
 15 Restless Sleep 896 non-null float64  
 16 Sleep Efficiency 896 non-null float64  
 17 Sleep Latency 896 non-null float64  
 18 Sleep Timing 896 non-null float64  
 19 Bedtime Start 896 non-null object   
 20 Bedtime End 896 non-null object   
 21 Average Resting Heart Rate 895 non-null float64  
 22 Lowest Resting Heart Rate 895 non-null float64  
 23 Average HRV 895 non-null float64  
 24 Temperature Deviation (°C) 883 non-null float64  
 25 Temperature Trend Deviation 866 non-null float64  
 26 Respiratory Rate 896 non-null float64  
 27 Activity Score 970 non-null int64   
 28 Stay Active Score 968 non-null float64  
 29 Move Every Hour Score 968 non-null float64  
 30 Meet Daily Targets Score 968 non-null float64  
 31 Training Frequency Score 968 non-null float64  
 32 Training Volume Score 968 non-null float64  
 33 Activity Burn 970 non-null int64   
 34 Total Burn 970 non-null int64   
 35 Steps 970 non-null int64   
 36 Equivalent Walking Distance 970 non-null int64   
 37 Inactive Time 970 non-null int64   
 38 Rest Time 970 non-null int64   
 39 Low Activity Time 970 non-null int64   
 40 Medium Activity Time 970 non-null int64   
 41 High Activity Time 970 non-null int64   
 42 Non-wear Time 970 non-null int64   
 43 Average MET 970 non-null float64  
 44 Long Periods of Inactivity 970 non-null int64   
 45 Readiness Score 898 non-null float64  
 46 Previous Night Score 888 non-null float64  
 47 Sleep Balance Score 892 non-null float64  
 48 Previous Day Activity Score 881 non-null float64  
 49 Activity Balance Score 887 non-null float64  
 50 Temperature Score 895 non-null float64  
 51 Resting Heart Rate Score 898 non-null float64  
 52 HRV Balance Score 887 non-null float64  
 53 Recovery Index Score 898 non-null float64  
dtypes: float64(39), int64(12), object(3)  
memory usage: 409.3+ KB

The above data was analyzed using graphical methods. Since the hypothesis of this project was that ‘Total Sleep Duration’ would be input by the user to determine a predicted ‘Readiness Score’, this was investigated first:

# 1. Prepare data  
x = df[‘Total Sleep Duration’]  
y = df[‘Readiness Score’]  
  
# 2. Setup plot  
fig, ax = plt.subplots(figsize=(10, 10))  
  
# 3. Plot data  
ax.scatter(x,y)  
  
# 4. Customize plot  
ax.set(title=‘Total Sleep vs. Readiness’,   
 xlabel=‘Total Sleep Duration’,  
 ylabel=‘Readiness Score’)  
  
# 5. Save and show  
fig.savefig(‘Figures/Figure\_1.png’)

A chart with blue dots

Description automatically generated

From this graph, as ‘Total Sleep Duration’ increases, ‘Readiness Score’ also generally increases. Other input fields were evaluated similarly.

‘Rest Time’ had a similar plot that generally shows a higher readiness with an improving ‘Rest Time’.

A diagram of blue dots

Description automatically generated

‘Move Every Hour’ seems to have little correlation to the readiness score and was not used in the final dataset:

A graph of blue lines

Description automatically generated

Using the graphical visualization methods above, various machine learning algorithms were trained. This was an iterative process that yielded an algorithm that met the hypothesis. The final dataset was determined to be:

df.drop(df.columns[[0,1,2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,26,28,29,30,31,32,33,34,35,36,37,39,40,41,42,43,44,50,51,52,53]], axis=1, inplace=True)

df.info()

<class 'pandas.core.frame.DataFrame'>

Index: 819 entries, 8 to 895

Data columns (total 9 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Total Sleep Duration 819 non-null float64

1 Temperature Trend Deviation 819 non-null float64

2 Activity Score 819 non-null float64

3 Rest Time 819 non-null float64

4 Readiness Score 819 non-null float64

5 Previous Night Score 819 non-null float64

6 Sleep Balance Score 819 non-null float64

7 Previous Day Activity Score 819 non-null float64

8 Activity Balance Score 819 non-null float64

dtypes: float64(9)

memory usage: 64.0 KB

## Accuracy Analysis

Accuracy was checked using the following command:

# Check the score of the model (on the test set)

model.score(X\_test, y\_test) #Coefficient of determination of the prediction R^2

This returned the coefficient of determination of the prediction and resulted in an accuracy of 0.38925080239747933 ([Scikit-learn: Machine Learning in Python](https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.).

For a typical day, the same amount of sleep as occurred was entered into the machine learning model resulting in a readiness score of 79 (see below). The actual readiness score was 85.

A screenshot of a computer

Description automatically generated

While predicted and actual readiness scores are within ten percent and demonstrate that the predictive model is working, the differences could be due to factors such as quality of sleep. Other factors may include caffeine intake, alcohol consumption, or eating too close to sleep. These might be areas where future improvements to the machine learning dataset could be improved.

## Application Testing

The application was tested iteratively. Code was added and run methodically. At several points, the dataset used to train the model was adjusted to check for accuracy improvements. Additionally, there were errors identified during this process that required systematic debugging.

Changing the dataset and the model used to train the machine learning algorithm resulted in an improved prediction of readiness scores.

## Application Files

This application was developed in the Spyder environment on a Windows 10 platform. It requires:

* Windows 10
* Spyder
* Pandas
* Numpy
* Streamlit
* Matplotlib
* Sklearn

Application files are organized as follows (Capstone.docx is not an application file):

A screenshot of a computer

Description automatically generated

## User Guide

The following steps are required to execute this program:

1. Ensure the following is loaded onto your Windows 10 computer:
   1. Anaconda (or miniconda3)
   2. Spyder
   3. Jupyterlab
   4. Pandas
   5. Numpy
   6. Streamlit
   7. Matplotlib
   8. Plotly
   9. Sckit-learn
2. Ensure the .csv files are in a subfolder titled ‘New Data’.
3. Ensure the .png files are in a subfolder titled ‘Figures’.
4. At the Anaconda prompt, navigate to the directory with the Capstone.py file.
5. Ensure the conda environment is activated.
6. To run the program, enter ‘streamlit run Capstone.py’ from the prompt.
7. A new window will be generated on the default browser.
8. At the bottom of the window is a slider bar that allows the user to select an estimated sleep. The program will then display a predicted readiness score.

## Summation of Learning Experience

As a certified project manager, I was able to effectively manage the scope and schedule for this project. This is important to ensure that scope creep does not enter the project and the project can be delivered on time and within budget.

To complete this project, additional learning resources were required. Examples include knowledge repositories and videos. With computer science evolving rapidly, the ability to research and understand new areas is increasingly becoming a valuable skill.

As this is my last project to complete my degree, I have had time to reflect on this journey. WGU has been a very pleasant experience. It has reinforced and taught me new skills that are highly valuable no matter which direction I take from here. I highly recommend WGU to other individuals. Thank you.

## References

1. Ōura Health Oy (n.d.). Oura mission. Oura. Retrieved September 4, 2023, from https://ouraring.com/about-us.
2. Kurama, V., & Powers, J. (2023, February 28). Regression in Machine Learning: What it Is and Examples of Different Models. Builtin.com. Retrieved September 4, 2023, from <https://builtin.com/data-science/regression-machine-learning>
3. Segal, M. R. (2003, April 14). Machine Learning Benchmarks and Random Forest Regression. EScholarship.org - Open Access Publications from the University of California. Retrieved September 4, 2023, from <https://escholarship.org/uc/item/35x3v9t4>
4. Emmert-Streib, F., & Dehmer, M. (2019, January 14). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. MDPI. Retrieved September 4, 2023, from <https://www.mdpi.com/2504-4990/1/1/21>
5. [Scikit-learn: Machine Learning in Python](https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.