Sleep and Physical Performance Analysis: A case study of Collegiate Women's Basketball Players

Data Imputations

Srishti Sharma
School of Engineering and Applied Sciences
Ahmedabad University
Ahmedabad, India
srishti.s1@ahduni.edu.in

***Abstract*—Over the decade, there has been a considerable amount of research done in analyzing the impact of sleep and exercise on the performance and recovery of an athlete. The *long term objective* of this research is to make use of the Machine Learning paradigm to create a predictive model that would help a coach determine whether a player is fit to play in the forthcoming match on the basis of the underlying patterns and correlations between sleep, training load, cognitive state information of the athlete and their performance. In this term, the objective in this project duration is to perform necessary data imputations as a part of preprocessing of data.**

Keywords—Sports Data Analytics, Basketball, Sleep, Exercise, Injury, Recovery, Data Imputations

I. Introduction

An athlete undergoes both cognitive and physical strain in their life due to the training load imposed upon them, their emotional and physical states, academic routine or regular routine strain, sleep etc. The performance of an athlete is highly dependent upon the quality and quantity of sleep an athlete gets. Sleep as well as exercise has a profound impact on the recovery time for an athlete from an injury.

The research conducted in this domain so far has been concluded on the basis of statistical models or on the basis of correlation drawn between the measured parameters. Issue with statistical models is that they try to fit a generalized equation emphasizing less on the underlying individual patterns. There is no holistic research done for example there is a statistical study of impact of sleep on shooting accuracy of a player or impact of cardiac rhythm on athletic performance. There is a need of a holistic research done taking into account all training load, cognitive stress (emotional and physical state), sleep impacts, physiological recordings in routine etc. There is no research carried out to study the impacts of sleep and exercise on the recovery time from an injury for an athlete.

The aim is examination of the interaction of sleep, stress and training load of athlete and its impact on performance and recovery. The problem statement will be tested and validated over the data collected for 17 women's basketball team players. The data will be gathered by tracking their sleep using a Whoop wearable strap (sleep monitoring & physiological markers), recording their training load by the strength coaches and the athletes and a short recovery and short stress questionnaire that would help evaluate their emotional and physical state.

The data collection period began during the final three weeks of offseason training and continued through a six-week preseason training schedule. During this time the athletes wore the WHOOP straps continuously during the collection period allowing for consistent monitoring of data throughout the day.

We experiment with a few data imputation techniques based on global and local statistics to the missing values in the dataset. Substitution using global mean value, analysis of the most dominant attribute over target attribute, imputation using k-means clustering based on dominant attribute, k-NN algorithm for imputations are experimented and the results are presented in section 3. Section describes a review of existing literature. Section also includes the details about the dataset and imputation techniques. Section 4 concludes the paper.

II. Literature Review

Sleep has a profound effect on the athletic as well as cognitive performance of the athletes. It is believed that a good sleep helps the athletes recover fast from the cognitive load they undergo [1]. Also, sleep plays an important role in helping the athletes recover from the routine strain thereby improving their performance such as their response time [2].

A study of the soccer players [3] suggested how a degradation in the athletic performance and recovery was directly related to the consumption of substances such as caffeine, alcohol, deprivation of the right quality of the right quantity of sleep, travel fatigue etc. The impact of cardiac rhythm on the athletic performance was studied by analyzing the impacts of deprived sleep due to night soccer matches [2]. It was reported that a deprivation of sleep for 36 long hours would lead to degradation of tolerance to a long duration of exercise of training load [4]. It was also concluded that sleep extension resulted in a better shooting rate and accuracy of basketball players [5].

A week-long sleep deprivation and irregular sleep habits would lead to degradation in attention and alertness of the players [6]. As per recommendations by the National Sleep Foundation, wearable devices were the most appropriate ones to measure the sleep activity or the heart rate measurements of players [7].

III. Experiment and Results

A. Dataset

To study the impacts of sleep and exercise on the athletic performance and recovery from injuries of the players, an experiment is to be conducted over 17 women's basketball team players from Sacred Heart University. For all these players, their information related to sleep, training load and cognitive stress is recorded using the following:

1. Sleep Monitoring using a WHOOP Strap – This is a wearable devices that helps track the sleep and recovery patterns of the player. It records the data in three categories: attributes dependent on the cardiovascular strain and exertion, other one depending on the resting heart rate and the sleep hours, sleep consistency, sleep score etc. and the last category is related to the physiological markers such as heart rate, respiratory rate, sleep etc.
2. Short Recovery Short Stress Questionnaire – A set of eight questions on emotional and physical state of the athletes.
3. Training Load – Time and the intensity of the metabolic activities of the athletes is computed in the form a score every week and is recorded in the dataset.

The data collection period began during the final three weeks of offseason training and continued through a six- week preseason training schedule. During this time the athletes wore the WHOOP straps continuously during the collection period allowing for consistent monitoring of data throughout the day.

B. Discussion

The data gathered from these three sources is then to be preprocessed which include imputing the missing values that arise due to any fault with the device. Next step is the reduction of dimensionality of the data as the numbers of attributes taken into consideration are too many and hence extraction of the most relevant attributes from the attribute set is an important step. The final phase will be studying the impact of sleep and exercise on the athletic performance of the athletes as well as their recovery from injuries. This predictive analysis will thereby help the coach determine whether a player will be able to play well in a forthcoming match.

As a part of the project, we used various techniques for imputing the missing values.

1. *Global mean substitution:* Each missing value of an attribute is substituted by the global average value of the entire attribute (column)

2. *Local mean substitution:* Techniques like k-means algorithm and k-NN algorithms were used for imputing values.

a. *K-means Clustering algorithm:* was applied over the entire dataset and then the mean of an attribute in each cluster was used to impute the missing values of that particular attribute in that particular cluster.

b. *Single attribute based K-means Clustering:* For filling missing values, K-means clustering is performed over the dataset fitting using the attribute whose values are to be imputed. Now for each missing value in cluster, the mean of that attribute values in cluster are averaged and substituted.

c. *Permutation based feature importance + Attribute based k-Means Clustering:* Fit a model for predicting target variable using independent variables and store the predictions. For each independent attribute, considering one attribute at a time, shuffle values of only that attribute rows and predict target variable using the fitted model. Calculate the loss suffered from shuffling. The independent variable for which there was maximum loss incurred due to shuffling is the most important attribute for particular target attribute considered. K- means is applied over the dataset fitting it on basis of the most dominant attribute as obtained from feature importance algorithm. Various combinations of dominant attributes and number of clusters were tried.

d. *K-NN Algorithm:* kNN algorithm was applied over the entire dataset. For a missing valued tuple, k nearest tuple were found and the missing tuple value was filled using the average value of that attribute values of all near neighbors found.

e. *Permutation based feature importance + Attribute based k-NN based Imputations:* kNN algorithm was applied over the entire dataset fitting using the most dominant attribute. For a missing valued tuple, k nearest tuple were found on basis of the most dominant attribute and the missing tuple value was filled using the average value of that attribute values of all near neighbors found.

3. *Linear Regression based Imputations: Each variable is modelled as a function of other variables that acts as a regressor and the missing values are predicted using this regressor.*

a. *Deterministic Linear Regression:* In Deterministic Linear Regression Imputation, we replace the missing data with the values predicted in our regression model and repeat this process for each variable.

b. *Stochastic Linear Regression:* Some normally distributed noise with a mean of zero and the variance equal to the standard error of regression estimates to add uncertainty back to the imputed variable values.

c. *Iterative Imputer:* Iterative imputation refers to a process where each feature is modeled as a function of the other features, e.g. a regression problem where missing values are predicted. Each feature is imputed sequentially, one after the other, allowing prior imputed values to be used as part of a model in predicting subsequent features. It is iterative because this process is repeated multiple times; allowing ever improved estimates of missing values to be calculated as missing values across all features are estimated.

The metrics used for drawing a comparison between these data imputation techniques were:

1. Raw Bias: This is the difference between the expected value and the predicted value. The permissible range of raw bias must be close to 0 [10]
2. Percentage Bias: This is the percentage of difference between the expected and predicted value divided by the predicted value. The upper bound set of percentage bias is 5% [10].

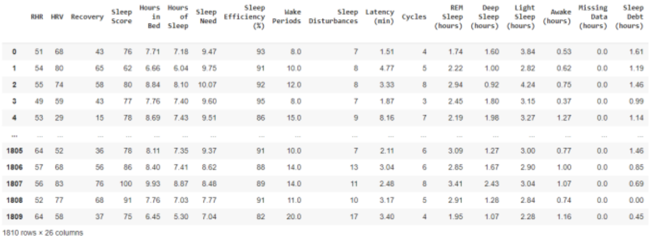
C. Results

A real time dataset collected for an experiment to be conducted over 17 women's basketball team players from Sacred Heart University.

For imputations, Whoop Strap Data is considered:

I. There are 1810 records (14 weeks of data)

II. Attributes: 26 (RHR, HRV, Sleep Score, Hours of Sleep, Wake Periods, Sleep Consistency, Respiration Rate, etc.)



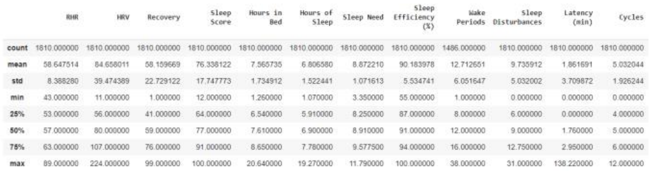
	RHR	HRV	Recovery	Sleep Score	Hours in Bed	Hours of Sleep	Sleep Need	Sleep Efficiency (%)	Wake Periods	Sleep Disturbances	Latency (min)	Cycles	REM Sleep (hours)	Deep Sleep (hours)	Light Sleep (hours)	Awake (hours)	Missing Data (hours)	Sleep Debt (hours)
0	51	68	43	76	7.71	7.18	9.47	93	8.0	7	1.51	4	1.74	1.60	3.84	0.53	0.0	1.61
1	54	80	65	62	6.66	6.04	9.75	91	10.0	8	4.77	5	2.22	1.00	2.82	0.62	0.0	1.19
2	55	74	58	80	8.84	8.10	10.07	92	12.0	8	3.33	8	2.94	0.92	4.24	0.75	0.0	1.46
3	49	59	43	77	7.76	7.40	9.60	95	8.0	7	1.87	3	2.45	1.80	3.15	0.37	0.0	0.99
4	53	29	15	78	8.69	7.43	9.51	86	15.0	9	8.16	7	2.19	1.98	3.27	1.27	0.0	1.14
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1805	64	52	36	78	8.11	7.35	9.37	91	10.0	7	2.11	6	3.09	1.27	3.00	0.77	0.0	1.46
1806	57	68	56	86	8.40	7.41	8.62	88	14.0	13	3.04	6	2.85	1.67	2.90	1.00	0.0	0.85
1807	56	83	76	100	9.93	8.87	8.48	89	14.0	11	2.48	8	3.41	2.43	3.04	1.07	0.0	0.69
1808	52	77	68	91	7.76	7.03	7.77	91	11.0	10	3.17	5	2.91	1.28	2.84	0.74	0.0	0.00
1809	64	58	37	75	6.45	5.30	7.04	82	20.0	17	3.40	4	1.95	1.07	2.28	1.16	0.0	0.45

1810 rows × 26 columns

Figure 1: Dataset

Step 1: Identifying Attributes with Missing Values

Wake Periods, Sleep Consistency, Respiratory Rate, Total Cycle Sleep Time (hours), REM Percentage, Deep Sleep Percentage, Restorative Sleep (hours), Restorative Sleep (%)

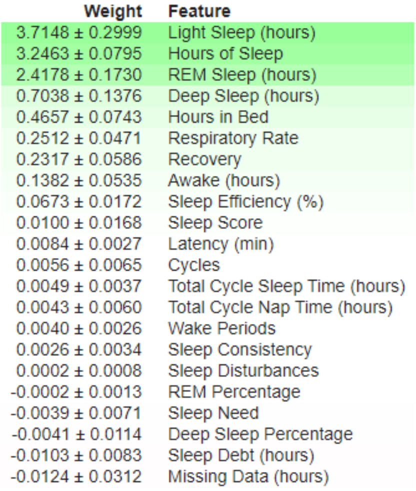


	RHR	HRV	Recovery	Sleep Score	Hours in Bed	Hours of Sleep	Sleep Need	Sleep Efficiency (%)	Wake Periods	Sleep Disturbances	Latency (min)	Cycles
count	1810.000000	1810.000000	1810.000000	1810.000000	1810.000000	1810.000000	1810.000000	1810.000000	1486.000000	1810.000000	1810.000000	1810.000000
mean	58.647514	84.658011	58.159669	76.338122	7.565735	6.806580	8.872210	90.183978	12.712651	9.735912	1.861691	5.032044
std	8.388280	39.474389	22.729122	17.747773	1.734912	1.522441	1.071613	5.534741	6.051647	5.032002	3.709872	1.926244
min	43.000000	11.000000	1.000000	12.000000	1.260000	1.070000	3.350000	55.000000	1.000000	0.000000	0.000000	0.000000
25%	53.000000	56.000000	41.000000	64.000000	6.540000	5.910000	8.250000	87.000000	8.000000	6.000000	0.000000	4.000000
50%	57.000000	80.000000	59.000000	77.000000	7.610000	6.900000	8.910000	91.000000	12.000000	9.000000	1.760000	5.000000
75%	63.000000	107.000000	76.000000	91.000000	8.650000	7.780000	9.577500	94.000000	16.000000	12.750000	2.950000	6.000000
max	89.000000	224.000000	99.000000	100.000000	20.640000	19.270000	11.790000	100.000000	38.000000	31.000000	138.220000	12.000000

Figure 2: Dataset Description

Step 2: RHR is a complete column. I selected RHR and randomly deleted 82 records. This was my new dataset. This was done so as to test the correctness of imputation techniques with original data

Next, permutation based feature importance was performed over the dataset to find features most dominant on the RHR attribute. Fit a model for predicting target variable using independent variables and store the predictions. For each independent attribute, considering one attribute at a time, shuffle values of only that attribute rows and predict target variable using the fitted model. Calculate the loss suffered from shuffling. The independent variable for which there was maximum loss incurred due to shuffling is the most important attribute for particular target attribute considered.



Weight	Feature
3.7148 ± 0.2999	Light Sleep (hours)
3.2463 ± 0.0795	Hours of Sleep
2.4178 ± 0.1730	REM Sleep (hours)
0.7038 ± 0.1376	Deep Sleep (hours)
0.4657 ± 0.0743	Hours in Bed
0.2512 ± 0.0471	Respiratory Rate
0.2317 ± 0.0586	Recovery
0.1382 ± 0.0535	Awake (hours)
0.0673 ± 0.0172	Sleep Efficiency (%)
0.0100 ± 0.0168	Sleep Score
0.0084 ± 0.0027	Latency (min)
0.0056 ± 0.0065	Cycles
0.0049 ± 0.0037	Total Cycle Sleep Time (hours)
0.0043 ± 0.0060	Total Cycle Nap Time (hours)
0.0040 ± 0.0026	Wake Periods
0.0026 ± 0.0034	Sleep Consistency
0.0002 ± 0.0008	Sleep Disturbances
-0.0002 ± 0.0013	REM Percentage
-0.0039 ± 0.0071	Sleep Need
-0.0041 ± 0.0114	Deep Sleep Percentage
-0.0103 ± 0.0083	Sleep Debt (hours)
-0.0124 ± 0.0312	Missing Data (hours)

Figure 3: Important Features for RHR Attribute

Data Imputation Techniques Experimented

1. Imputation using Global Mean
2. Imputation using Local Mean
a. k-Means Clustering based Imputations on entire dataset
b. RHR based k-Means Clustering
c. light-sleep (hours) based k-Means Clustering
3. Imputation using Linear Regression
a. Deterministic Linear Regression based Imputations
b. Stochastic Linear Regression based Imputations
c. Iterative Imputer based Imputations

Results for Raw Bias:

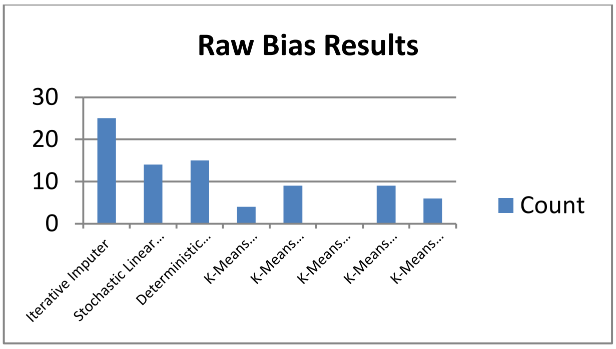


Figure 4: Raw Bias based results

Results for Percentage Bias:

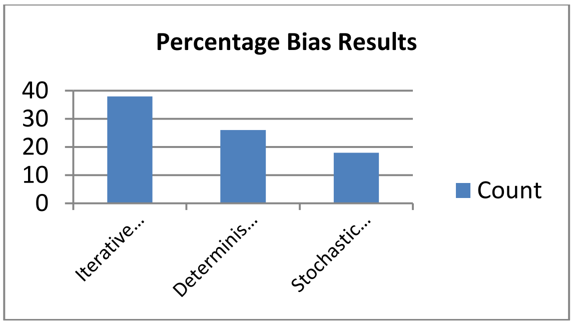


Figure 5: Percentage Bias based results

IV. CONCLUSION

Permutation based feature importance is an effective technique for finding the most dominant attribute for a given target attribute whose values are to be imputed. Data Imputation using Multivariate Regression using combination of important attributes selected worked well. Linear Regression, Stochastic Linear Regression as well as Iterative Imputer (MICE imputer) performed well on the dataset. Iterative Imputer performed the best on the basis of the raw bias and the percentage bias metrics.

REFERENCES

1. Vyazovskiy, V.V., 2015. Sleep, recovery, and metaregulation: explaining the benefits of sleep. Nature and science of sleep, 7, p.171.
2. Fullagar, H.H., Skorski, S., Duffield, R., Hammes, D., Coutts, A.J. and Meyer, T., 2015. Sleep and
3. athletic performance: the effects of sleep loss on exercise performance, and physiological and cognitive responses to exercise. Sports medicine, 45(2), pp.161-186.

4. Nédélec, M., Halson, S., Abaidia, A.E., Ahmaidi, S. and Dupont, G., 2015. Stress, sleep and recovery in elite soccer: a critical review of the literature. Sports Medicine, 45(10), pp.1387-1400.
5. Martin, B.J., 1981. Effect of sleep deprivation on tolerance of prolonged exercise. European journal of applied physiology and occupational physiology, 47(4), pp.345-354.
6. Mah, C.D., Mah, K.E., Kezirian, E.J. and Dement, W.C., 2011. The effects of sleep extension on the athletic performance of collegiate basketball players. Sleep, 34(7), pp.943-950.
7. Al-Kandari, S., Alsalem, A., Al-Mutairi, S., Al-Lumai, D., Dawoud, A and Moussa, M., 2017. Association between sleep hygiene awareness and practice with sleep quality among Kuwait University students. Sleep health, 3(5), pp.342-347.
8. Demirtas, H., Freels, S.A. and Yucel, R.M., 2008. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, *78*(1), pp.69-84.