MODELING SLEEP EFFICIENCY

Nimeesha Kolari
Cupertino High School, Inspirit AI
Cupertino, California, USA

ABSTRACT

I’ve always been interested in sleep and why some people tend to sleep better than others. I think it’s important to recognize what could be affecting a person’s sleep and what they can do to improve it. The goal of my research question, “How much do alcohol, exercise, and smoking affect sleep?” is to show how factors such as these can affect the quality of one’s sleep. I found a data set consisting of many factors that can affect sleep, as well as sleep quality. Using this data set, I built a model that allows you to input values for factors and come out with a sleep quality score. My model is a regression model, as it makes predictions through numerical data given to it. I tested out different models, settling on the most accurate one, Linear Regression, with a mean absolute error of 0.051.

INTRODUCTION

As a young child, I often had trouble falling asleep, even after a long day. I ate well and exercised, so I had no idea what was causing this. This made me really interested in sleep and in the reason behind my lack of sleep. In recent years, many smartwatches have tools to monitor sleep, which can be very helpful for people like me, people who struggle with sleep. However, a lot of these smartwatches don’t show the users how their lack of sleep could be connected to outside factors. The purpose of my project is to predict sleep efficiency/duration using other given qualities such as alcohol, exercise, and smoking, forming the research question, “How much do alcohol, exercise, and smoking affect sleep?”. People can play around with the model, and see how not drinking or smoking could potentially increase their sleep efficiency. With this new information they can make lifestyle changes to improve their sleep quality. This problem is supervised, as it has labeled data and is predicting. It is also a regression- the model predicts a numerical value. The purpose of the model is for the inputs to predict sleep efficiency, a value between 0 and 100.

BACKGROUND

While trying to figure out what I wanted to research about, I looked at a few research papers with similar ideas to the research question I started with- “What outside factors affect sleep?”. I used the papers to help me narrow down my research idea. Searching the word “sleep efficiency” led me to many articles about caffeine and alcohol, including “Caffeine: Sleep and Daytime Sleepiness”[1] and “Alcohol and Sleep I: Effects on Normal Sleep.”[2] The number of articles made me hypothesize that part of the reason so much of the population has trouble sleeping is because of their consumption of caffeine and alcohol. This helped me identify my final research question, “How much do alcohol, exercise, and smoking affect sleep?”.

DATASET

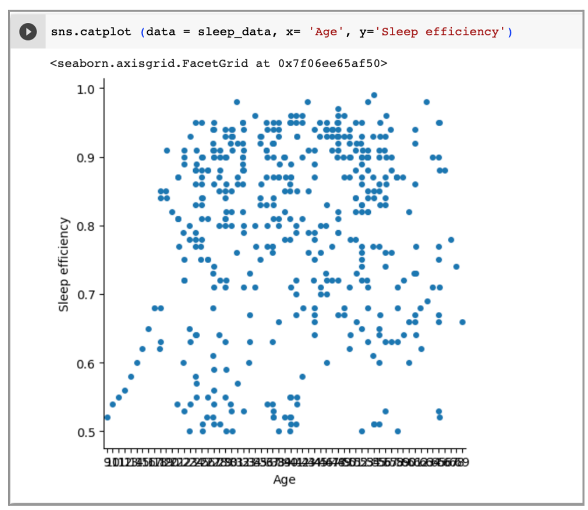

Figure 1. Catplot of Age vs Sleep efficiency. After picking a dataset, I played around with the different variables in graphs.

I found my dataset on kaggle by searching up key words like “sleep efficiency” and “smoking”. Through that I was able to find a dataset called “Sleep Efficiency Dataset”[3] consisting of 452 subjects, with information like age, gender, bedtime, wakeup time, sleep efficiency, and the types of sleep (REM, deep, and light). Kaggle defines sleep efficiency as “a measure of the proportion of time spent in bed that is actually spent asleep." The "sleep efficiency" piece of data was given as a score, a numerical value between 0 and 100.

```
from datetime import datetime
datetime_object = datetime.strptime(sleep_data['Bedtime'][0], '%Y-%m-%d %H:%M:%S')
print(datetime_object)
time = datetime_object.time()
print(time)
```

Figure 2. Converting the times to datetime objects to make them simple numerical values.

In order to input data into an AI model, it must be numerical. The learning algorithm can only understand numerical values, so we take the categorical values and make them numerical. The majority of the data was numerical, but gender, smoking status, bedtime and wakeup time were not. The times were a bit complicated, written with both dates and the time, like "2021-03-06 01:00:00". Working with my mentor, I made this a datetime object, taking just the time out (Figure 2). Then, I was able to use a simple equation to make it a numerical value instead of just the time - for example turning 7:00 to 7.0. Using LabelEncoder, I was also able to convert gender and smoking status to numerical values - with "1" symbolizing "male" and "yes", and "0" symbolizing "female" and "no".

```
[17] pd.isnull(sleep_data).sum()

ID                        0
Age                       0
Gender                    0
Bedtime                   0
Wakeup time               0
Sleep duration            0
Sleep efficiency          0
REM sleep percentage      0
Deep sleep percentage     0
Light sleep percentage    0
Awakenings               20
Caffeine consumption     25
Alcohol consumption      14
Smoking status            0
Exercise frequency        6
dtype: int64
```

Figure 3. Pandas function "isnull()" displays missing values in 4 different columns.

The model needs a full dataset to make accurate predictions. After everything was made numerical, I checked for missing values in my dataset - using the pandas function "isnull()". As you can see in the picture, a few values were missing in 4 of the columns. It wasn't a huge amount that could greatly affect the results, but significant enough to change something. Using the "fillna()" function, we replaced the empty values with the mean of each respective column. After the data preprocessing, my data was all numerical and there were no missing values, so I saved it as a new csv, ready to train and test data.

MODELS AND METHODOLOGY

I first identified "sleep efficiency" as the y-vector, dropping it from the x-matrix to prevent it from skewing the model. I also dropped the "ID" column, as it doesn't have any effect on sleep efficiency. The other columns all can affect the y-vector, so I kept them. After this, I split the data into testing and training, training on 80% of the data and testing on the remaining 20%. In order to test the efficiency of data, the model must be trained first. Splitting the data allows us to prevent overfitting the model and to evaluate it more accurately.

```
# STEP 1: Initialization
from sklearn.linear_model import Ridge
import numpy as np
# STEP 2: Training
clf = Ridge(alpha=1.0)
clf.fit(X_train, y_train)
# STEP 3: Prediction
predictions=clf.predict(X_test)
# STEP 4: Evaluation
mean_absolute_error(y_test, predictions)
```

Figure 4. Testing the Ridge Model - 4 steps: Initialization, training, predicting, and finally evaluating with the mean absolute error (MAE) metric.

I tested out 4 different models - Linear Regression, the Random Forest Regressor, Ridge Regressor, and Support Vector Machine (SVM) Regressor. These models are from sklearn, a machine learning library written in Python. The Linear Regression model predicts using linear approximation. The Random Forest Regressor builds decision trees, using averaging to predict accurately. The Ridge Regressor generally helps avoid overfitting, and is used when the model is more complex to simplify it. The SVM Regressor fits data points over a space instead of a singular line, which can be helpful with complex models.

After importing each model in, I passed in "X_train" and "y_train" and trained using the fit function. I then passed in "X_test" to the predict function, and used "y_test" to evaluate each model's prediction with the Mean Absolute Error metric.

RESULTS

Table 1. Regression Metrics for Tested Models.

	MSE	MAE	RMSE
Linear Regression	0.00391	0.051001	0.062532
Random Forest Regressor	0.00597	0.064441	0.077276
Ridge Regressor	0.00391	0.051043	0.062567
SVM Regressor	0.00631	0.066399	0.079487

Some metrics used to evaluate how well a regression model is working are Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Mean Squared Error calculates the average squared error between the original and predicted values. Mean Absolute Error is the total error over all of the samples. Root Mean Squared Error is the square root of the Mean Squared Error, and is less sensitive to outliers.

In order to figure out which model was most accurate, I calculated these metrics for each of my four models. The Linear Regression model had the lowest metrics, just slightly lower than the Ridge Regressor (Table 1). This surprised me because earlier when I was inputting the data into different plots (Figure 1), it didn't seem very linear. The accuracy of the model also surprised me. When I first tested out the mean squared error for predicting sleep efficiency, I accidentally left the sleep efficiency of the data set in the training data. Because of this, the accuracy was extremely high, almost perfect. Even once I removed the sleep efficiency from the training data and retested, the regression metrics were still quite high, which was unexpected. I was impressed by the accuracy of the model.

CONCLUSIONS

With my research question, "How much do alcohol, exercise, and smoking affect sleep?", I was able to build a model that predicts sleep efficiency using a linear regression model. This model performs extremely well when tested with various accuracy metrics. The majority of this project was driven by my own curiosity - my own interests about how outside factors affect sleep. However, after building this model, I think it would also be interesting for other people to input their values for these factors into the model. This way, they can see how changing some of these elements would help them improve their sleep and hopefully implement those changes.

Acknowledgments

Thank you to Serena Hughes, my Inspirit AI mentor, for helping me through the entire process, including all the coding, data preprocessing, testing out models, and completing this research paper.

References

[1] Roehrs, T. and Roth, T. (2008) 'Caffeine: Sleep and daytime sleepiness', *Sleep Medicine Reviews*, 12(2), pp. 153–162. doi:10.1016/j.smrv.2007.07.004.

[2] Ebrahim, I.O. *et al.* (2013) 'Alcohol and sleep i: Effects on normal sleep', *Alcoholism: Clinical and Experimental Research*, 37(4). doi:10.1111/acer.12006. Psychosom Med 2001;63:49–55.

[3] Equilibriumm (2023) *Sleep efficiency dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency (Accessed: 11 August 2023)