Title of Final Project: Child Mind Institute - Detect Sleep States

Section: 52745

Student Name: Thomas Wynn

Student UT EID: ttw483

Date: 10/30/23

[0] Goal of the Project

We aim to predict sleep/wakeup times from wrist-worn accelerometer time series data. We have training data with labeled sleep onset/wake-up times, and our solution will be evaluated on hidden test data. However, this problem is not truly a time series forecasting/prediction problem, but rather a time series segmentation/change-point detection problem. This is because we do not need to make predictions in real time, but can see the entire time series before making any predictions. We have two explanatory variables: ENMO (Euclidean Norm Minus One of all accelerometer signals, rounded to zero) and z-angle (a metric derived from individual accelerometer components, commonly used in sleep detection, and referring to the angle of the arm relative to the vertical axis of the body). We will create cool, interactive visualizations throughout our model testing and EDA to visualize how sleep states can be predicted using machine learning.

[1] Familiarizing the Dataset

In train_series, there are 127,946,340 rows and 5 columns. Train_series is data containing 277 unique series appended to each other on the first axis. Each series represents a continuous period in which a child has worn an accelerometer on their wrist. A recording is taken every 5 seconds, and 1 series can last multiple days, hence a large number of rows.

In train_events there are 14,508 rows and 5 columns. Train_events is data containing wakeup/onset (of sleep) events. Each event is associated with a step, time stamp, and most importantly, the series_id that the event is associated with.

The column names of train_series are 'series_id', 'step', 'timestamp', 'anglez', and 'enmo'. The datatypes of each column are string, 32-bit integer, string (will convert to datetime), 32-bit float, and 32-bit float, respectively.

Train events have column names: 'series_id', 'night', 'event', 'step', and 'timestamp' which are string, 64-bit integer, string, 64-bit float, and string (will convert to datetime), respectively.

‘Series_id’ is nominal data (the hexadecimal string is used as a series’ unique identifier) There are 277 unique series_ids, but a lot of these series have missing/invalid data, as we will see later. ‘Step’ indicates the index of a measurement within a unique series ID (ratio data). The statistics for the step columns are range: 1,433,879, median: 234,519, mean: 254,804.8, and standard deviation: 177,893. The units for step are unitless since step is just an index of a measurement for a given series. ‘Timestamp’ is a timestamp string in YYYY-MM-DDTHH:MM:SSZ format (ratio data). We will convert this to a datetime format type using pandas. ‘Anglez’ is a measurement of the accelerometer’s angle relative to the body’s vertical axis it is measured in radians (interval data). The statistics for the step columns are range: 1,433,879, median: 234,519, mean: 254,804.8, and standard deviation: 177,893. ‘Enmo’ is a measure of angular acceleration measured in radians/s^2. The statistics for the ‘enmo’ columns are range: 11.43 median: 0.0017, mean: 0.004, and standard deviation: 0.01.

[2] Data Wrangling

In train_events, 240 series are missing event data. If we train a machine learning model on these data, we could increase our probability of getting false negatives, since the event data (wakeup/sleep times) simply aren't there. For further exploratory data analysis and model building, we decided to drop these series. This leaves a total of 37 series that we can explore.

Since the data is split up into train_series and train_events, it would be nice if we could make one data frame where the data has a column that is labeled 0 for sleep and 1 for awake states. This would make it easier to visualize the data and train machine-learning models on it. In this step, we also converted the Date column to a datetime object column.

[3] Preliminary Statistical Analysis

Based on the correlation matrix of the cleaned data, we found that enmo and awake had the highest correlation at 0.23, while the other correlation pairs were nearly 0. Therefore, we decided to focus our statistical analysis on this pair. It does not make sense to use a quadratic, cubic, or any other polyfit curve to fit our data since our target variable is binary. We will use a logistic regression instead.

We plotted autocorrelation plots for the enmo and anglez data of one series. The enmo plot showed autocorrelation at lower lags, indicating that the data may be white noise. The anglez plot showed that the anglez data was also nearly white noise.

Since the target column is binary, we ran a logistic regression model on the explanatory variables (anglez and enmo) to predict the target variable (awake). The logistic regression model for enmo showed that there were very low enmo values, even in the awake state. The model also showed that low enmo values usually indicated awake states. Therefore, we may be able to exploit the relationship between low enmo and awake states.

The logistic regression model for the anglez explanatory variable did not provide any insights into the relationship between anglez and awake states.

[4] Narration (In bullet points)

- Goal: Predict sleep/wakeup times given wrist-worn accelerometer data time series data.
- Dataset:
    - train_series: 127,946,340 rows, 5 columns, containing data from 277 unique children
    - train_events: 14,508 rows, 5 columns, containing wakeup/onset of sleep events
- Data Wrangling:
    - Dropped 240 series from train_events that were missing event data.
    - Combined train_series and train_events into a single dataframe with a column labeled 0 for sleep and 1 for awake states.
- Preliminary Statistical Analysis:
    - Enmo and awake had the highest correlation at 0.23.
    - The enmo and anglez data of one series were nearly white noise.
    - A logistic regression model for enmo showed that low enmo values usually indicated awake states.
    - A logistic regression model for the anglez explanatory variable did not provide any insights into the relationship between anglez and awake states.

Overall, the findings suggest that enmo is a more promising explanatory variable for predicting sleep/wakeup times than anglez. Future work could focus on developing a more accurate machine learning model to predict sleep/wakeup times using enmo data.