Digital Health Report

Alban Bogdani, Kathapet Nawongs

Thursday 9th November, 2023

LSTMs Part 1 - Data Preparation for Sequential Predictions

Firstly, we opened the file 20220422hourheartbeatmerged.csv and started exploring. From the exploration we realized that this database has the data of 9 persons measured every hour for almost one month each. The table below shows some summary data about the dataset:

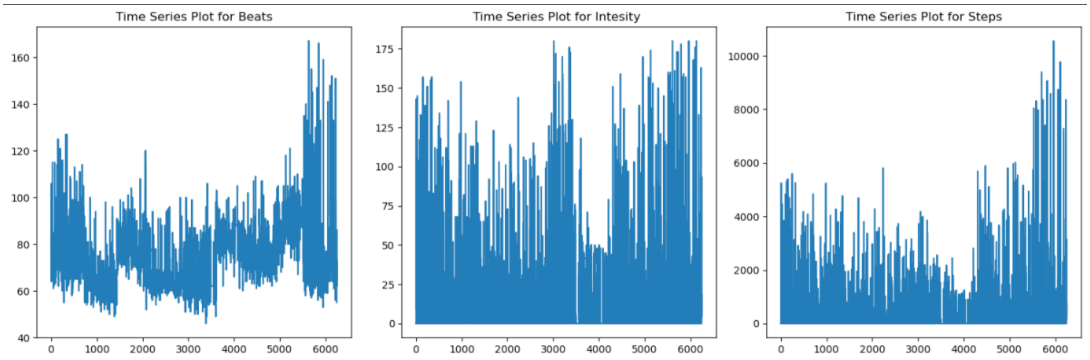
ID	Rows	Count	Period
2022484408	2-733	731	12.04-12.05
4388161847	734-1441	707	13.04-12.05
4558609924	1442-2173	731	12.04-12.05
5553957443	2174-2905	731	12.04-12.05
5577150313	2906-3613	707	12.04-11.05
6117666160	3614-4201	587	15.04-09.05
6962181067	4202-4933	731	12.04-12.05
7007744171	4934-5521	587	12.04-06.05
8877689391	5522-6253	731	12.04-12.05

Count represent the total number of data points for each person. Rows represent the row number where the data points for each person starts and ends.

Problem i. How to the trends and patterns in the 2022_04_22_hour_heartbeat_merged.csv dataset compare to those observed in the Hourly_merged.csv dataset?

Actually, it's not easy to compare this two datasets, as Hourly_merged.csv dataset has 22100 data points and 2022_04_22_hour_heartbeat_merged.csv dataset has much less data points than the other dataset, 6253 rows. The two variables that are in common between two datasets are Intensity and Steps. By observing the time series graph and the histograms, the trends and patterns look similar.

Problem ii. Based on the time series plots of the twos elected features, can we identify any recurring patterns or anomalies?



From the Intensity and Steps plot we realize that the datapoints around the interval 3600 - 4200 have relatively low values compared to the other datapoints and this datapoints interval corresponds with the user with ID: 6117666160. So this means that this user had a low physical activity. If we jump to the Beats rate, this user has relatively average values. If we compare the Beats and Steps plots we can see a linear positive correlation between Steps and beats. Specifically, the last user has done more steps and also he has the highest Beat rates. After the last user, comes the first user who also had relatively high number of steps and thus a higher beat rate.

Problem iii. After handling missing values, how does the distribution of the dataset change? Are the significant shifts in mean or variance?

In the dataset, we discovered that there were only 12 missing values in total a very insignificant number, so even after replacing the missing values, the Mean and Variance didn't change at all.

Problem vi. How does the scaling of training and testing data separately impact the range and distribution of the scaled values? If needed, compare it with scaled values without splitting the data.

Problem v. How might scaling the entire dataset(training+ testing) introduce potential biases?