Data Analysis Mini-Project

Tomasz Petecki

Introduction

Migraine affects over a billion people globally, representing one of the leading causes of short-term disability. Despite the widespread impact, many individuals with migraines fail to manage their attacks effectively, often waiting until symptoms become unbearable before taking action. This project aims to shift migraine care from a reactive to a predictive approach using advanced machine learning techniques, specifically Long Short-Term Memory (LSTM) models, to forecast migraine attacks.

By leveraging signals such as, sleep patterns, screen time, and hydration levels, one might seeks to predict when a migraine is likely to occur. The goal is to develop a tool that empowers users by providing insights into their unique migraine trigger patterns. Our project aims to improve life quality for millions of people and create a solution that can scale globally (provided the models work).

Brief summary of what was achieved: Collected data from wearables was cleaned and preprocessed in order to learn the migraine patterns from the data. Two methods were emplyed - Random Forest and Long Short-Term Memory (LSTM) neural networ.

Data Collection and Pre-processing

Collection and content

I used a dataset available on Kaggle - Migraine Dataset from Wearable Devices. The quality and source of the data is challenging to assess as there were no information about the origin of data. Dataset contains 11,879 entries from 100 users, in the following columns: `user_id`, `date`, `sleep_hours`, `mood_level`, `stress_level`, `hydration_level`, `screen_time`, `migraine_occurrence`, `migraine_severity`. The only continuous data in the set is pertains to the *hours of sleep* and *screen time*, the rest of the columns is either self explainatory or represents categorical data – levels from 1 to 5. Raw data is presented in the table below:

```
  user_id      date sleep_hours mood_level stress_level hydration_level
1       1 1/15/2024         7.8          3            2               2
2       1 1/16/2024         6.6          4            1               2
3       1 1/17/2024         8.5          4            2               2
4       1 1/18/2024         7.5          3            2               3
5       1 1/19/2024         9.0          3            2               1
6       1 1/20/2024         8.8          4            1               2
  screen_time migraine_occurrence migraine_severity
1         4.7                   1                 1
2         3.2                   1                 1
3         4.7                   1                 2
4         3.8                   1                 3
5         6.8                   1                 2
6         7.2                   1                 2
```

Preprocessing

Although the data was claimed to be clean, a standard preprocessing procedures were conducted among the others I cleaned the data by removing potential missing values. Also, I added target features (occurrence of a migraine tomorrow), added some cyclical features (in order to capture weekly or monthly cycles, day-of-the-week and day-of-the-month features were added as sinusoidal signals) for improved learning of the LSTM network. For more insight, I encourage to look at the git repo where the preprocessing functions and feature-specific functions were included.

In the case of the Random Forest no further preprocessing was required, while in the case of the LSTM the data had to be truncated as each user contributed a different number of rows.

Exploratory Data Analysis

Part of EDA included looking at multiple graphs to spot some clear patterns. To this end, throughout the analysis functions from the module `utils` in the git repo were incorporated.

Major part of the Exploratory Data Analysis was to learn from what is the relation between the different variables, for this reason I firstly plotted some raw data plots with highlighted migraine episodes, to hopefully spot some trends.

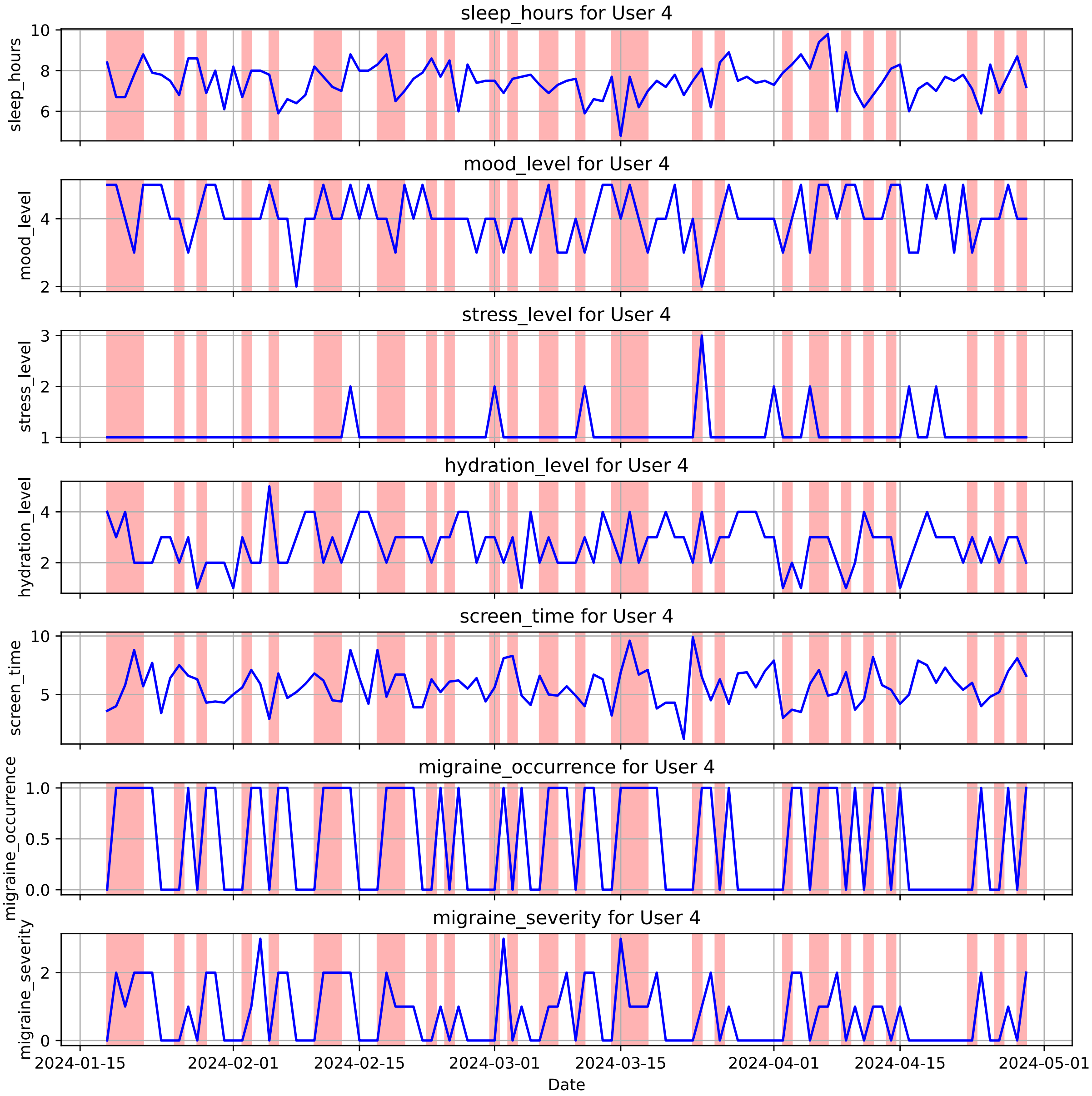


Figure 1: Raw data plot with migraine episodes one day ahead highlighted in red

Then to get the necessary relations in the data, I binned the data and plotted bunch of heatmaps. Also, I conducted a small regression analsysis. I was looking to find the regression coefficients that describe the linear relationships between multiple variables in the dataset. The linear regression model can be written in matrix form as:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Where $\mathbf{y}$ is the vector of dependent variables (e.g., `sleep_hours`, `mood_level`, etc.), $\mathbf{X}$ is the matrix of independent variables β is the vector of regression coefficients, ϵ is the vector of residuals (errors).

The vector of regression coefficients β is estimated by solving the following normal equation:

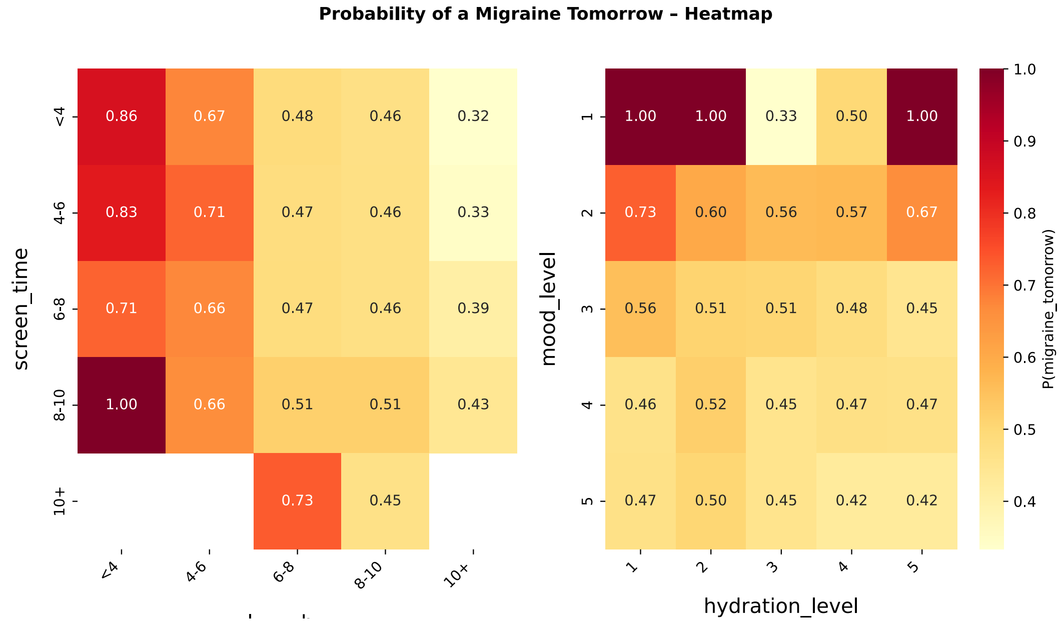
$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

In the case of pairwise regressions, each variable in the dataset is treated as a column in $\mathbf{X}$, and we compute the regression coefficient β_i for each variable i by solving this equation for each pair of variables.The results are shown in the table below:

```
                   sleep_hours  mood_level stress_level hydration_level
sleep_hours                 NA  0.29508610  -0.02716927     0.004760224
mood_level         0.295086096          NA  -0.25709731     0.020112713
stress_level      -0.027169268 -0.25709731           NA    -0.021095476
hydration_level    0.004760224  0.02011271  -0.02109548              NA
screen_time        0.038861980 -0.01964706   0.13856428     0.019139884
migraine_severity -0.126147972 -0.10465736   0.04520702    -0.282351106
                   screen_time migraine_severity
sleep_hours        0.038861980      -0.126147972
mood_level        -0.019647059      -0.104657361
stress_level       0.138564278       0.045207020
hydration_level    0.019139884      -0.282351106
screen_time                 NA       0.008138504
migraine_severity  0.008138504                NA
```

From the table it should be quite evident that the regression coefficients in the data are quite low, which is somewhat dissapointing. One of the more correlated variables with migraine severity is turned out to be hydration. However, we might want to look at this from a slightly different angle.

The heatmaps, although they contain some of the information from this table, are quite informative when it comes to the “correlations” in the data. In the heatmap below, one can see the probability of the migraine tomorrow over some binned data. They were constructed for pooled data from all users.



Here, for example in the first heatmap, we see that migraine probability is greatest in the regions of least sleep hours. The same goes for other variables except maybe hydration level, which is a surprise taking into consideration its high correlation with migraine severity. So maybe hydration level aggravate migraines that were already bound to happen, but hydration level alone is a weak predictor of the occurrence of a migraine episode the next day?

Learning From Data

We used two following methods to learn from the data: Random Forest and LSTM network. Both were tested but the accuracy of classification was rather poor in both cases. Nevertheless, following good reporting practices, I include both in my report.

Random Forest

In the context of regression or classification tasks, the Random Forest algorithm constructs a set of B decision trees, where each tree is built using a bootstrap sample from the data. Each decision tree T_b makes a prediction $\hat{y}_b$, and the final prediction $\hat{y}$ is obtained by aggregating the predictions from all trees. For regression, the aggregation is typically done by averaging the predictions:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b$$

For classification, the aggregation is done by majority voting:

$$\hat{y} = \text{mode}\left(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_B\right)$$

Random Forest reduces variance by averaging predictions from multiple models. Without all the gory details about how decision trees work, this method is best suited for high-dimensional, non-stationary data, which the migraine prediction task satisfied. In this case, Random Forest was trained on the migraine data with an additional features of the "days since start", also `hydration_level` was removed from features as it significantly reduced the accuracy of the model. The following table presents the results of learning from the migraine data:

```r
rep_rf = read.csv(rf_report)
print(rep_rf)
```

```
             X precision    recall  f1.score      support
1            0 0.5854430 0.6135987 0.5991903 1206.0000000
2            1 0.5732601 0.5443478 0.5584300 1150.0000000
3     accuracy 0.5797963 0.5797963 0.5797963    0.5797963
4    macro avg 0.5793516 0.5789732 0.5788101 2356.0000000
5 weighted avg 0.5794963 0.5797963 0.5792945 2356.0000000
```

Long Short-Term Memory (LSTM)

The LSTM model is a type of recurrent neural network (RNN) designed to handle sequences of data, such as time-series data. At each time step t, the LSTM takes the input x_t and updates its hidden state h_t and cell state c_t based on the previous states h_{t-1} and c_{t-1}. The key equations that govern an LSTM unit are:

The LSTM learns to capture long-range dependencies in sequential data by selectively forgetting, updating, and outputting information at each time step. The following table presents the report for our LSTM setup (for detailed description consult the git hub repo). The model in question was trained using all the features, and additional sinusoidal features which improved the prediction accuracy of the model. This report was generated for a model trained on 23-day long sequences, as this is the model used for the app. However, one must acknowledge that the we had to truncated data to include only 90 day long sequences. Great improvements were observed by training on longer sequences.

```r
lstm_rep = read.csv(lstm_report)
print(lstm_rep)
```

```
             X precision    recall  f1.score      support
1            0 0.5693548 0.4744624 0.5175953  744.0000000
2            1 0.4414286 0.5364583 0.4843260  576.0000000
3     accuracy 0.5015152 0.5015152 0.5015152    0.5015152
4    macro avg 0.5053917 0.5054603 0.5009607 1320.0000000
5 weighted avg 0.5135325 0.5015152 0.5030778 1320.0000000
```

Conclusions

Although the models were used in the migraine prediction app that was submitted as a part of the project - the overall accuracy of the models is slightly better than chance (Random Forest – 0.57, LSTM - 0.50). Therefore, their clinical significance is nul. Albeit, this project provided valuable insights into the complexities of predicting migraine occurrences. The low accuracy emphasized the challenge of modeling a multifactorial medical event, where numerous unaccounted variables may influence the outcome.

One key takeaway was the importance of data quality. The dataset, sourced from Kaggle, lacked detailed information on its origins, which made it difficult to assess its reliability. Additionally, although the dataset contained over 11,000 entries, the data was not structured in a way that allowed the models to easily capture meaningful patterns. Feature engineering played a significant role, particularly for the LSTM model, where cyclical features (e.g., day of the week) were added to capture temporal dependencies. Despite these efforts, the models were unable to meaningfully predict migraines, likely due to the limited and possibly noisy data.

Another lesson learned was that while Random Forest and LSTM are powerful models, they are not a one-size-fits-all solution, especially for complex time-series tasks like migraine prediction. Future work could focus on incorporating more diverse data sources, improving feature engineering, and exploring alternative models that might be better suited for capturing the intricate patterns in medical data. While the results were underwhelming, the project provided a strong foundation for future exploration in the realm of predictive modeling for healthcare. The challenges I faced underscore the importance of high-quality, well-structured data in building reliable predictive models. Nevertheless, the current academic reports on migraine prediction analysis underscore that migraine episodes are notoriously unpredictable - one study reported moderate accuracy (best 0.62, achieved by random forest) of machine learning models in the migraine prediction task [Stubberud et al. (2023)]. So hey, we are not completely lagging behind.