Music Enabled Running

“Can music improve by using Music”

Mai Linh Luong
m.luong@student.fontys.nl

Arwen
a.voorn@student.fontys.nl

Mitchel Kuijpers
m.kuypers@student.fontys.nl

Filip Vangelov
f.vangelov@student.fontys.nl

I.INTRODUCTION

During the Minor program in Applied Data Science, four aspiring researchers took the task analyzing multiple generated from foot sensors in a setup described in Janssen (2018). The sensors record while test runners listen to music from Spotify. This begs the question “Can music improve running?”. This project covers extensive research on the different datasets generated by the sensors and other devices that were used in data gathering.

II.APPROACH

The approach taken is based on the IBM Data Science Methodology, going through the phases of Business Understanding, Data Preparation, Modeling, Evaluation and Deployment.

III.DATA

The data gathered by two recreational runners, wearing RunScribe Pro footpods and dedicated waist-worm iPhone 11 smartphones, which they took with them during their usual outdoor running sessions. The datasets contain stride and foot placement data from the footpods, the music listened to by the runners, the music features, like energy and artist as supplied by the Spotify API, phone locations, phone activity, and general session information. During the project, additional new data was supplied, which was used to further progress the project. The first two datasets consisted of the footpod, sessions and music data. The phone data was not yet necessary for that part of the project. The two cleaned datasets for the first and second test runner were made with the help of multiple supplied CSV-files. The dataset for the first test runner consists of 29 columns and 59,457 rows and mostly contains object, float64, int64 and bool data types. The dataset for the second test runner consists of 20 columns and only 6,381 rows.

After creating the cleaned datasets during the exploratory data analysis, new data was supplied. This dataset was already cleaned and contained additional columns about the foot pod data, such as symmetry angles. The new dataset for the first test runner contained 245,725 rows and 95 columns. The dataset for the second test runner contained 17,840 rows.

IV.GOALS

The goal of this project is to assess whether there is a correlation between running quality and music.

V.EXPERIMENT

Initially, most of the focus was on cleaning the raw sensor data. We use several data cleaning methods and use classification by clustering to filter outliers that compromise the quality. K-means was used for creating clusters based on parameters such as foot-contact-time and impact force. By the “Elbow rule”, we determined the preferred number of clusters and removed the ones with the fewest data points. To determine the test person was running or walking, a step frequency column was added. The lower the frequency, the lower the test person moved, the more likely the test person was walking than running. With the help of boxplot, outliers for the frequency were removed. This way, only running-rows of the dataset are being saved. Before each run the test person had to put on some sensors and turn on the app on the phone in order to save the running data. This sometimes produced short, canceled test runs, ending up as a new session in the dataset. And sometimes data was added much later to an outdated session. To filter out the wrong sessions, we set a song count. Hereby, a session where less than 4 songs or more than 10 songs were played, considered a “bad” session. Bad sessions were removed. Another way to filter out bad sessions was to use the session

duration to filter out short and unrealistically long sessions.

With this cleaned data, and the newly supplied dataset, we set out to find correlations between the foot data attributes and the music. We used Correlation Analysis, Extreme Gradient Boosting, Self-Organizing Maps and developed a map covering the running session for better understanding using kepler gl.

We performed a correlation analysis on both test persons to find running metrics correlating with music. From trying both Pearson and Spearman correlation, we concluded that the impact symmetry has the highest correlation with music for test person 2, test person 1 did not show any significant correlation.

Extreme Gradient Boosting (XGB) was applied on the cleaned data to predict the impact symmetry. We extracted a feature importance list from the XGB which shows what features have an effect on the impact symmetry. This method was evaluated by using 5-fold cross validation, to decide the best parameters.

VI.RESULTS

For test person 2 the XGB produced a MSE of 0.196. This may appear small but with a value range of [-0.24 & 0.173]. The coefficient of determination (R2) score, which is the proportion of the variation of the dependent variable that is predictable from the independent variable measured by a percentage. Close to 100% means a high correlation. The actual score was -0.9%, meaning that there is no relation between the XGB prediction and the impact symmetry score.

Experiments with the Self Organizing Map (SOM) did not lead to much success even though there were interesting results. In a SOM, each data point in the data set recognizes themselves by competing for representation. SOM mapping starts from initializing the weight vectors. From there a sample is selected randomly and the map of weight vectors is searched to find which weight best represents that sample. Each weight vector has neighboring weights that are close to it. Therefore, the weights of the vectors did not tell much about, for instance, the danceability of the running sessions of the test person.

VII. CONCLUSION AND RECOMMENDATIONS

The project Music Enabled Running was quite unique compared to other projects. Instead of creating and having a clear goal, this project focused more on the scientific statement whether or not there is a correlation between running behavior and music.

In general, it is assumed that a lack of symmetry in the running parameters results in injury. One of the things that would be considered as bad running is when the impact symmetry is high. Stomping hard on the ground while running could also lead to serious injuries.

One of the things that could’ve been done with more time, is to collect additional data in a different trial. One person would run with music and the other person would run without music. Or let one test person run one session with music and one session without music. This way, a comparison can be made giving a more clearer view whether or not there is an improvement in the running performance.

More data from test person 2 could change the results of the correlation analysis. Comparing test person 2 with 17,840 rows to test person 1 with 245,724 rows, a case could be made that the correlations are because of the low data volume. A strong statistical and mathematical understanding and evaluation is needed to decide the meaning and validity of the model output.