Rahul Rajeev
07/03/2023
Predictive Analytics
DSC630-T301 (2237-1)

Case Study Analysis

"Detecting Anomalous Activity Patterns"

by D. Holcomb, M.S., George Washington University

Introduction:

FitBit and other personal trackers have become increasingly popular throughout the recent decade. Especially with the extremely restricted living style during the pandemic, many were finding the need to be more responsible and aware of their physical health. Data scientists, statisticians, and medical experts have been using this data to detect relationships in complex time series data. The relationships can be related to daily life patterns and detecting deviations from these patterns can be important to be able to predict future behaviors based on the user's FitBit history. For example, being able to predict an ill medical patients improving or deteriorating conditions can lead to automated motivational messages being sent and avoid prior pitfalls in their future goals. In summary, the studies focus on finding a baseline behavior profile to predict future activities and to also identify atypical activities from this established baseline. Examples of atypical activities would be variations in the sleep schedule and how these variations can add up to a change in the levels of workouts and high-activity moments for the future. For instance, when a user wouldn't have much sleep on consecutive days, it could define how long or how effective their workouts would be. The goal of this case study was to identify these anomalous behavior studies and try to predict real-time data using the model used in this case study.

Rahul Rajeev
07/03/2023
Predictive Analytics
DSC630-T301 (2237-1)

Dataset

The data was collected from Mechanical Turk, an Amazon crowdsourcing marketplace, and was openly available on Zenodo to analyze. The dataset itself was spread across thirty total anonymized users of the FitBit and their tracking applications and was recorded from March 12, 2016 through May 12, 2016. To encapsulate the data, it was stored in two monthly datasets, the first was eleven and the second was eighteen csv files. The data included minute-level output for physical activity, heart rate, and sleep monitoring. There were only three spreadsheets existing across both months: daily activity, sleep duration, and weight log. Sleep was combined on a per-day basis and weight data was not factored in because of how sparse it was.

Methods and Results:

Data preparation

The study used Python and Python libraries such as Pandas, NumPy, SciPy, and Matplotlib to discover and visualize relationships between anomalous activity patterns. Starting with preprocessing, Pandas was used to clean the data, delete records for when the FitBit data was empty (the user wasn't wearing the watch), and group the resultant data frames by user id and date. In terms of data analysis, NumPy and Sci-kit Learn were used to establish average behavior, identify outliers, and identify/correlate activities. The goal was to categorize users into groups of similar activity and sleep patterns to find unique users as well as create a model to predict behaviors of other users who might share the same lifestyle.

Rahul Rajeev
07/03/2023
Predictive Analytics
DSC630-T301 (2237-1)

Solution

To identify indicators of anomalous behavior, Holcomb plotted features against each other and found that steps and distance were highly correlated and active minutes weren't correlated with quality sleep ratio. After, to begin grouping users in each of these different categories, Venn Diagrams identified "Best" or "Worst" users comparing the individual features in steps, distance, sleep, and active minutes. As a result, two very interesting users were discovered. One had high steps, distance, and sleep but had the worst amount of active minutes. Another user was high on steps, distance, but was the worse sleep. Understanding these extreme cases was important to understanding how different users could be grouped. Which led to the use of unsupervised clustering as the preferred modeling method for this project.

Modeling Techniques and Results

As mentioned previously, it seemed that K Means clustering would be the best choice for predictive analysis for this project. Holcomb created a K means model using three clusters as the optimal choice after using the elbow method. The elbow method and a visualization weren't present on the Jupyter Notebook file itself unfortunately. Three clusters, while it didn't seem like the right choice, made sense considering that there were only thirty points to choose from. The clustering was done on the four values plotted in the Venn Diagrams: steps, distance, active minutes, and sleep. Holcomb made a note of not being able to see 4 dimensional points so she visualized a 3D space instead to show the clusters. She also referred to using PCA and t-SNE for dimensionality reduction if there were more time for the project to develop. But the dimensionality reduction would no longer be human-understandable labels as she stated it would "obfuscate the context of the axes" and put it on a higher dimensional plane (Holcomb, 2018).

The results of the clusters presented a clear difference between the “Best” and “Worst” users and also presented a third cluster which Holcomb appropriately labeled “Average” users, those of whom have the average values for each of the features in the dataset. The previous named users in the “Best”/”Worst” categories from the Venn Diagrams were also correctly labeled in the cluster model. As far as metrics of evaluation, there weren’t any present in the notebook file or in the paper. Since K means is not used to classify per say since there aren’t existing labels on the groups of users, analyzing accuracy from this model wouldn’t necessarily be the greatest way to see whether the model fits well. However, metrics such as silhouette coefficients, homogeneity score, v measure, and completeness score are all possible accuracy measures that could’ve been included with the report to get a numerical understanding for the clustering (Valle, 2021). Just as an example, the homogeneity score would show whether clusters are homogeneous or that they have points that belong to one cluster only. Since the data set only had about thirty points, I would expect high homogeneity.

Conclusion:

In conclusion, the FitBit data has numerous potential uses and being able to predict a user’s behavior based on their past history could prove useful once applied to more special cases as mentioned before such as assisting medically ill users in terms of recommending workouts and sending automated motivational messages. Holcomb brings up an interesting argument in being able to group users based on their sleep duration, active minutes, steps, and distance traveled by using K Means in order to have a better understanding of how behavioral patterns differ by clusters of users.

Rahul Rajeev
07/03/2023
Predictive Analytics
DSC630-T301 (2237-1)

Interpretations of Results

Amidst the normal behavioral patterns of having average amounts of steps, distance, sleep duration, and active minutes, there were two different anomalous behavior patterns:

1. Users that have a high step/distance count, but strangely low active minutes
2. Users that excel in all active categories, but get very poor sleep

These two anomalous behavior patterns could reveal potential indicators for deviations from regular behavior patterns or can help find other far more interesting indicators in behavior prediction.

Actionable Consequences

The case study was conducted upon a small anonymized set of users who have allowed for their activity to be tracked and used for studies, most likely have also accepted a certain terms of conditions. Due to this, I don’t believe that there are any actionable consequences for this study and its results. The anonymity was preserved and the data was taken with the knowledge of the users. In terms of perhaps applying the results and model to real-time data, that could lead to some consequences.

For example, since this was a very small subset of data, applying the current model could ignore other anomalous behavior and other clusters that may exist. This could lead to wrong recommendations being sent or in the worst case scenario, incorrect prescriptions for medically ill users. For that reason, I believe it might be better to first apply the K Means clusterisng on a larger subset of data and then apply the real-time data to it after. For more details about recommendations for future cases and use, check the recommendations section below.

Rahul Rajeev
07/03/2023
Predictive Analytics
DSC630-T301 (2237-1)

Future Recommendations

Holcomb made note of a few ways to improve her model and case study results should it be worked on in the future. One of which was introducing PCA and t-SNE for dimensionality reduction which was discussed during the model creation section of the paper. Later in the appropriately named “Recommendations” section of the paper, she discusses a couple adjustments and here is a list:

1. More data, either having more users to look at or data over a longer period of time
2. More features in the data, weight measurements, gender, geographic location, weather, local holidays; all of which can be attributed to anomalous behavior
3. Users that actively participate in the study, and therefore, have complete records
4. Redistributing the sleep data from a per-day to a per-sleep-session to also account for other times of day of sleep.

Rahul Rajeev
07/03/2023
Predictive Analytics
DSC630-T301 (2237-1)

References

Bennet, D. T. (2022, March 27). *Five fascinating case studies in predictive analysis.* Medium. https://medium.com/@DeepthiTabithaBennet/five-fascinating-case-studies-in-predictive-analysis-b9bf15216ff7

Holcomb, D. (2018, August 16). *Analysis of Fitbit data: Detecting anomalous activity patterns*. Zenodo. https://zenodo.org/record/1346296#.YkAPJ-dBy5d

Purta, R., Mattingly, S., Song, L., Lizardo, O., Hachen, D., Poellabauer, C., and Striegel, A. (2016). Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits, *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, September 12-16, 2016, Heidelberg, Germany. DOI: 10.1145/2971763.2971767

Sathyanarayana ,A., Joty, S., Fernandez-Luque,L., et al. (2016). Sleep Quality Prediction From Wearable Data Using Deep Learning. Eysenbach G, ed. *JMIR mHealth and uHealth*. 2016;4(4):e125. doi:10.2196/mhealth.6562.

Wright, S., Brown, H., Tyish S., Collier, S., and Sandberg, K. (2017). How consumer physical activity monitors could transform human physiology research. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 2017, 312, 3, R358--R367.