**JEREMY PETERSON DUTOYA**

**PR. MENG-SHEN HSIEH**

**PMBA-8358-OLA: DATA-DRIVEN STRATEGIES FOR BUSINESS**

**TERM PROJECT: THE BELLABEAT APP**

Logo

Description automatically generated

Monday, March 13th, 2023

1. **INTRODUCTION**

Sleep is a fundamental aspect of human health and well-being, yet it is often overlooked and undervalued. In our fast-paced world, it can be tempting to sacrifice sleep in favor of work, socializing, or other activities. However, the truth is that getting a good quality of sleep is essential for both the body and brain to function optimally.

During sleep, the body repairs and restores itself, allowing the immune system to function properly and promoting the growth and repair of tissues and organs. Sleep also plays a critical role in cognitive function, including memory consolidation, learning, and problem-solving. Inadequate or poor-quality sleep can lead to a range of health problems, including obesity, diabetes, cardiovascular disease, and mental health issues.

As we know, regular physical activity is also essential for overall health and well-being. Exercise has been shown to improve physical fitness, reduce the risk of chronic diseases, and enhance mental health. But what about its impact on sleep quality? Can regular exercise help us get better sleep?

In this project, we aim to explore the relationship between sport and sleep quality. We will examine the effects of different intensity of exercise on aspects of sleep, including duration, latency, efficiency, and subjective quality. To create a predictive model for sleep quality, we will specifically examine the information gathered by the Bellabeat app, which tracks users' levels of activity and quality of sleep. In order to help users improve their sleep patterns and general health, we will work to understand the activity variables that are connected to poor and good sleep. Based on our examination of the trends in activity levels among poor sleepers, we will then suggest tactics to encourage Bellabeat users to engage in physical activity. In this study, we seek to advance our knowledge of the connection between exercise and the quality of sleep as well as offer useful advice for enhancing sleep practices by using the best practices we have learned during the class.

By analyzing the impact of sport on sleep quality, we hope to provide valuable insights into how individuals can optimize their sleep habits and overall health. The findings of this study could have important implications for public health policy, exercise prescription, and sleep interventions. We believe that this research will contribute to a better understanding of the importance of sleep and exercise for human health and well-being.

1. **DATA VIZUALIZATION**

Before performing any analysis, it is crucial to understand the data that we are working with. In this first part of the project, I am focusing on 5 variables that we are exploring to see their relationship with the quality of sleep. The variables are “VeryActiveMinutes”, “Fairly Active Minutes”, “LightlyActiveMinutes”, “SedentaryMinutes”, and “Calories”. These variables represent the amount of time spent on high, fairly, and lightly impact activities, inactivity, and total calories burned. These variables are important to consider as they may have a significant impact on sleep quality.

The target variable that we are interested in is Sleep Quality, which is a binary variable. A Sleep Quality score of 1 indicates poor quality, while a score of 0 indicates good quality with sleep hours >= 5 hours per day. By examining the relationship between these variables separately and Sleep Quality, we hope to gain insight into the factors that affect sleep quality and ultimately improve the sleep habits of individuals. Therefore, it can be insightful to carefully analyze and visualize the data to understand the relationship between these variables and Sleep Quality:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| VeryActiveMinutes | Minutes on high impact activities |
| FairlyActiveMinutes | Minutes on fairly impact activities |
| LightlyActiveMinutes | Minutes on lightly impact activities |
| SedentaryMinutes | Minutes recorded on inactivity |
| PCT\_VeryActiveMinutes | Percentage of VeryActiveMinutes to the sum of VeryActiveMinutes, FairlyActiveMinutes, Lightly ActiveMinutes and SedentaryMinutes |
| Calories | Total calories burned |
| Sleep Quality | 1: poor quality with sleep hours < 5 hours per day  0: good quality with sleep hours >= 5 hours per day |

By creating visualizations of these variables to highlight their correlation with sleep quality, we discovered some interesting insights that shed light on the importance of physical activity for a good night's rest. Specifically, I have found out that individuals who engage in any level of activity for a minimum amount of time tend to have a much better sleep quality compared to those who are sedentary for too long. Even a light intensity activity can help transforming the quality of your sleep. I will provide a detailed analysis of my findings and the implications of these results before diving in more detailed analysis with the prediction model.

Chart, histogram

Description automatically generated

*DATA VIZUALIZATION #1 – Relationship between Sedentary Minutes and Quality of Sleep*

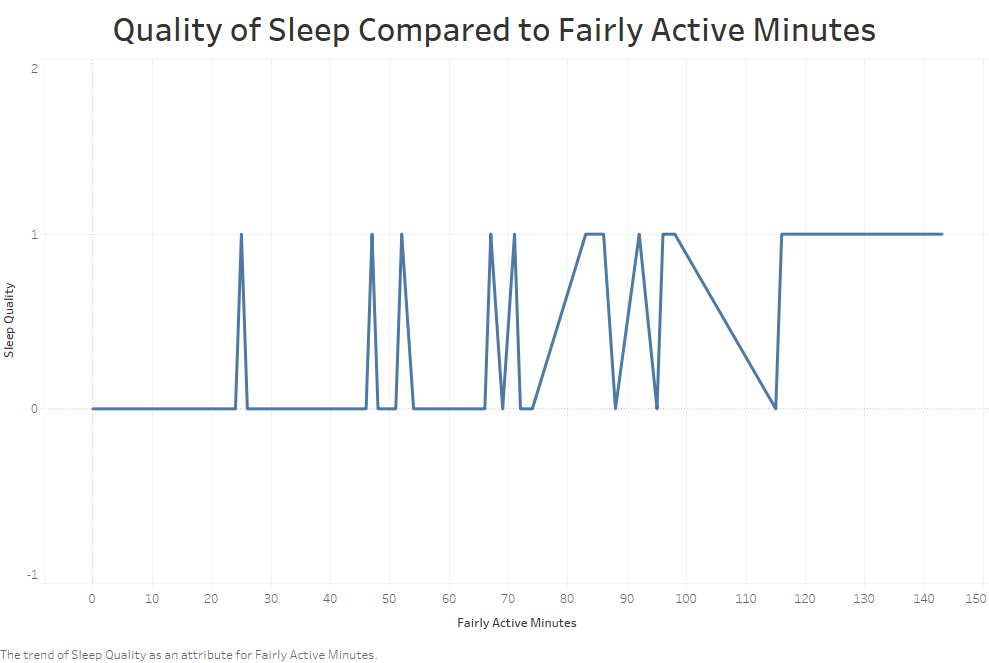
This first data visualization revealed interesting insights regarding the relationship between the duration of sedentary activities and sleep quality. We can observe that a sedentary level of inactivity for up to 688 minutes or approximately 11.5 hours negatively impact sleep quality. From 0 to 688 minutes of inactivity, the sleep quality is and remain good. However, beyond this threshold, sleep quality began to fluctuate between good and poor. Notably, past 1062 minutes or 17 hours of inactivity, the sleep quality remained consistently poor. Our findings suggest that while inactivity may not necessarily harm sleep quality, extended periods of inactivity exceeding 11 hours are resulting in an uncertain quality of sleep while very important periods of inactivity exceeding 17 hours are more likely to result in poor sleep quality.

Chart, histogram

Description automatically generated

*DATA VIZUALIZATION #2 – Relationship between Lightly Active Minutes and Quality of Sleep*

The second visualization in our analysis explores the potential relationship between the duration of light activity and sleep quality. Although the findings are not definitive, some trends can still be observed. Notably, the visualization suggests that spending between 34 and 170 minutes engaged in light activity may result in varying sleep quality outcomes, with some individuals experiencing good quality sleep while others experience poor quality sleep. However, beyond 283 minutes of light activity, a good quality of sleep is consistently and always observed. Within the 170–283-minute range, there are 4 instances where poor quality sleep was recorded, but it is still very likely that individuals within this range will experience good quality sleep. We can then extend our expected good quality sleep range to make start at 170 minutes, approx. 3 hours. These findings suggest that the duration of light activity may have some influence on sleep quality, spending 3 hours on light activity have a positive impact on the quality of sleep, although other factors are also at play such as the finding highlighted in the first data visualization.



*DATA VIZUALIZATION #3 – Relationship between Fairly Active Minutes and Quality of Sleep*

The third visualization explores the relationship between "FairlyActiveMinutes" and "Sleep Quality". Despite the fluctuating nature of the line, we can still draw meaningful insights from this visualization. We observe that between 0 and approximately 45 minutes of fairly active activity, the quality of sleep is consistently good, with the exception of one instance. Between 45 and 96 minutes, the quality of sleep fluctuates between good and poor. Beyond 96 minutes, or an hour and a half, the quality of sleep is more likely to be poor. The visualization highlights the surprising finding that excessive fairly active activity can lead to poor sleep quality. The optimal duration for fairly active activity is found to be less than 45 minutes to ensure good quality sleep.

Chart, histogram

Description automatically generated

*DATA VIZUALIZATION #4 – Relationship between Very Active Minutes and Quality of Sleep*

The fourth visualization in our analysis examines the relationship between the quality of sleep and the duration of high-impact activities. Notably, this visualization features the lowest number of instances of poor sleep quality, with only five observations recorded. This finding suggests that maintaining an active lifestyle, including some high-intensity activity, is likely to result in good sleep quality. Furthermore, we observed that exceeding 22 minutes of very active activity resulted in only three instances of poor sleep quality. Thus, we can infer that engaging in more than 22 minutes of high-intensity activity is likely to result in a good quality of sleep, with over 80 minutes of such activity resulting in good sleep quality in most cases.

The fifth visualization in our analysis explores the correlation between the total calories burned and the quality of sleep. Despite the fluctuations of the line, we can still draw some observations from this visualization. Between 0 and 1401 calories burned, we see a consistently good quality of sleep. However, it should be noted that most people burn over 1800 calories per day without exercise, and this number increases for those who are more active. As a result, it is challenging to interpret this curve and derive any concrete conclusions. Beyond 3500 calories burned, it is highly likely to result in a good quality of sleep. Nevertheless, the fluctuating nature of the curve between 1400 and 3500 makes it difficult to make definitive conclusions and provide recommendations. Interestingly, this visualization shows that there is no clear relationship between these two variables. We can see that both good and poor quality of sleep occur across a wide range of total calories burned, and there are no clear trends or patterns to be observed. This suggests that other factors, such as the duration and intensity of physical activity, may have a greater impact on sleep quality than the total number of calories burned. Therefore, it is important to consider multiple variables when analyzing the factors that influence sleep quality.

Chart, histogram

Description automatically generated

*DATA VIZUALIZATION #5 – Relationship between Calories burned and Quality of Sleep*

Through our five visualizations, we have explored the relationship between several variables taken alone and the target variable, which is the quality of sleep. From our analysis, we have identified some important insights.

First, we have found that a long period of inactivity, lasting more than 11 hours, can negatively impact the quality of sleep. On the other hand, spending too much time on fairly active activities for over 45 minutes can also result in poor quality of sleep. However, spending between 34 and 283 minutes on light activities and over 22 minutes on high impact activities will most likely result in a good quality of sleep.

Second, we have found that the relationship between the total calories burned, and the quality of sleep is not clear, and it is hard to draw any conclusions from this visualization. Suggesting that type of activity has a higher impact on quality of sleep than the number of quality burned.

Given these insights, it can be beneficial to now focus on a classification tree analysis to combine these variables and identify which ones have the most significant influence on the quality of sleep. By building a classification tree model, we can determine the most relevant variables, their optimal thresholds, and their impact on the quality of sleep. This can lead to a more accurate prediction of the quality of sleep based on a combination of factors and can help individuals make better lifestyle choices to improve their sleep quality.

1. **MAJOR ANALYTICAL METHODOLOGY**

In the case of predicting poor quality sleepers, as the target variable is binary, we have two options - logistic regression or classification tree. While logistic regression models are popular, I chose to based my work on the use of a classification tree as they have several advantages, particularly in terms of interpretation of results. Let me explain, classification trees use a hierarchical structure to divide the data into smaller groups based on the predictor variables, creating decision rules that lead to the outcome variable. This allows for easy interpretation of the results and identification of the most important predictors for poor sleep. In addition, the resulting decision tree provides a visual representation of the decision-making process, making it easy to understand and communicate to others that might not have the same background in terms of analytics.

With the 5 variables in the data visualization section as predictors for the model, a classification tree will help us identify the most important predictors for poor sleep and their associated decision rules. To evaluate the model, we will split the data into training and test datasets (70/30), fit the model to the training data, and assess its performance using evaluation metrics such as an error matrix and ROC curve.

The classification tree analysis reveals that there are three key factors that lead to a good quality of sleep, that is the 3 variables used when splitting the tree: sedentary minutes less than 1010, fairly active minutes of activity less than 19 minutes (62%), and total calories burned greater than or equal to 1790 (25%) and again a test on sedentaryMinutes. In contrast, poor quality sleep outcomes are associated with sedentary minutes greater than or equal to 1010. These findings provide important insights for individuals seeking to improve their quality of sleep. For instance, by limiting sedentary time, increasing moderate intensity activity, and aiming for a higher number of calories burned, individuals may be able to improve the quality of their sleep. Furthermore, the analysis highlights the importance of taking a holistic approach to health and well-being, where multiple lifestyle factors are considered together to optimize outcomes.

Diagram

Description automatically generated

We can also access the rules responsible for splitting the tree and see in a more detailed way what variables are responsible for a poor-quality sleep and on the other hand, the parameters facilitating a good quality sleep.

Graphical user interface, text, application

Description automatically generated

*Image #2: Rules for splitting the classification tree.*

After analyzing both the classification tree and the rules responsible for splitting the tree, we can conclude that the most important variable responsible for a poor quality sleep a level of sedentary activity higher than 1010 minutes, that is approximately 17 hours (rule 3), or sedentary minutes higher than 683.5 minutes with a total calories burned less than 1790 and fairly active minutes higher than 19 minutes (rules 23 accounting for only 3% of the training set).

Now let’s discuss the performance of the model. As part of assessing the performance of the classification tree model, I have generated an error matrix and a ROC curve on the test set after partitioning the dataset into a 70/30 split between the training set and testing set. The area under the ROC curve for our model is 0.7695, indicating that the model has a good degree of predictive power. The averaged class error obtained in the error matrix is 20.9%, the overall error of 4.9%. An overall error of 4.9% in the error matrix of a classification tree model means that the model correctly predicted the target variable (quality of sleep) 95.1% of the time on the test set. In other words, the model has an accuracy of 95.1%, which is a very good performance. Overall, these findings demonstrate that this classification tree model is a useful and reliable tool for predicting the quality of sleep based on these variables, but further refinement may be needed to improve its predictive power. Screenshots of these results are also available in the appendix of the project.

Graphical user interface, text, application, email

Description automatically generatedChart, line chart

Description automatically generated

1. **STRATEGIC RECOMMENDATIONS**

Based on the analysis conducted, we can suggest strategies to communicate with poor sleepers and encourage them to engage in some level of physical activities, engaging in even light level of activity limiting time of sedentary activity has proven to increase chances to have a good quality sleep by 90%. Firstly, it is important to advise them to reduce their sedentary time to less than 1010 minutes per day, as this is a key factor influencing their sleep quality. Secondly, they should aim for less than 19 minutes of fairly active activity to ensure good quality of sleep. Thirdly, they should aim to burn a total of 1790 or more calories per day. These strategies should be communicated in a clear and concise manner, highlighting the benefits of physical activity on sleep quality.

Additionally, it may be helpful to suggest specific types of physical activities that can be incorporated into daily routines, such as walking or yoga. It is important to emphasize that these strategies are based on the analysis of data and may be personalized to each individual's needs and lifestyle. As part of our effort to develop more effective recommendations, I have generated supplementary visualizations to better understand the trends of four key variables over the course of a week. Specifically, I have examined the patterns of PCT\_VeryActiveMinutes, PCT\_FairlyActiveMinutes, PCT\_LightlyActiveMinutes, and PCT\_SedentaryMinutes for each category of Sleep Quality from Monday to Sunday. By analyzing these trends, we have gained further insights that will inform the development of a comprehensive plan for achieving optimal sleep quality.

Chart, scatter chart

Description automatically generated

Timeline

Description automatically generated

The initial two visualizations are geared towards identifying the optimal percentage of daily activity levels required to achieve a high quality of sleep. Based on our analysis, it is recommended that an individual should maintain a sedentary level of activity for approximately 72% of their day, engage in light activity for approximately 24%, and very active and fairly active activities for 3% and 2% of the day, respectively, in order to attain a good quality sleep.

Chart

Description automatically generated

Based on this visualization, which demonstrates the relationship between the percentage of different levels of activity and the total hours of sleep, adhering to the aforementioned percentages results in the most suitable and optimal average number of hours of sleep. By following the 72-23-3-2% guideline, the average number of hours of good quality sleep obtained is 7.75-7.50-7.30-7.58 hours. This suggests that adhering to these percentages can help individuals achieve an optimal quality of sleep for a desirable optimal duration considering other lifestyle responsibilities, it is all about balance.

Chart, line chart

Description automatically generated

As depicted by the fourth visualization, the slopes of the curves representing the percentage of physical activity remain constant throughout the week for all four categories of activity. Therefore, to maintain a good quality of sleep, one should sustain a consistent level of activity for each category throughout the week.

To conclude on my strategic recommendations and report, one engage in light level of activity daily without even noticing. It includes activities such as walking, climbing stairs that we all do without focusing, while fairly active activities could involve cycling or low-intensity workouts. Examples of very active activities could include running, high-intensity workouts or playing sports such as basketball or soccer. Incorporating activities such as walking or cycling into a daily routine can significantly increase the chances of achieving a good quality of sleep. Additionally, engaging in high-intensity activities or playing sports a few times, a week can also contribute to maintaining a consistent level of activity and, ultimately, improving sleep quality. Based on the percentages suggested in this analysis, it is feasible to achieve a good quality sleep by incorporating additional light level of activities into one's routine. For instance, taking a walk with a friend or by listening to music for around 20-25 minutes in addition to the light level of activity involved in daily responsibilities can assist in achieving this objective.

Additionally, taking some time to reflect on personal aspirations and goals is beneficial for mental health, and physical activity is an effective method to support mental health. It is recommended to engage in higher levels of activities (very active and fairly active) for up to 40 minutes (24+15), which can be achieved through activities such as a brief run or a friendly game of basketball or soccer. I can only testify, sport and soccer has changed my life for the better, and I have never had a poor-quality night of sleep when practicing. Find activities that align with personal interests and preferences, enjoy them, and watch the quality of your sleep improve.

1. **APPENDIX**

Graphical user interface, table

Description automatically generated



Step 1: Load the Fitness\_Record.csv data to Rattle R. Check the Partition Box and set the Splitting Rule 70/30. Assign the Target variable “SleepQuality” and input variables. Click Execute.

Graphical user interface, text, application

Description automatically generatedStep 2: Click the Model Tab. Keep the Standard Values of the model tuning parameters as Pr. Hsieh suggested in the Q&A forum of the class. Click Execute.



Graphical user interface, application

Description automatically generated



Step 3 : Still in the Model Tab. Click the “Draw” button to display the classification tree. Click Execute. Go to RStudio to find the tree under the tab “Plot”.

Graphical user interface, application

Description automatically generated



Graphical user interface, text, application

Description automatically generated

Step 4: Still in the Model Tab. Click the “Rules” button to display the rules of splitting the tree and discover what variables are significant in splitting the classification tree and influence the target variable.

Graphical user interface, text, application, email

Description automatically generated



Step 5: Click on “Evaluate”. Use the test data to generate error matrix.

After executing the step, we can see the error matrix below:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated



Step 6: Still on the “Evaluate Tab”. Click ROC and Execute. You will have the following message.



Go to RStudio to find the ROC curve under the tab “Plot”.

Chart, line chart

Description automatically generated