PMBA 8358 Term Project

Bellabeat is a high-tech company that manufactures health-focused smart products. They offer different smart devices that collect data on activity, sleep, stress, and reproductive health to empower women with knowledge about their own health and habits.

The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions.

**Objective:** In this project, you will analyze one of Bellabeat’s products, **the Bellabeat app**, a smart device which records Bellabeat users’ activity data.

It’s no secret that sleep is important to our overall health, but a lack of sleep could have substantial effects on our heart, a new study shows. Your goal is to develop a predictive model to understand which activity variables are associated with sleep quality, or sleep hours. You will analyze the trend of activity levels of poor sleepers and propose efficient communication strategies to remind them of engaging in physical activities.

There are three major tasks in this project:

1. Data Visualization
2. Develop a predictive model to predict sleep quality
3. Propose a strategy to encourage Bellabeat users to engage in physical exercises frequently

**Data**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Id | User ID |
| ActivityDate | Date when the fitness activities were recorded |
| Total Steps | Total number of walking steps |
| Total Distance | Total distance of the walking steps (Total Steps\* distance/step) |
| 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 |
| PCT\_FairlyActiveMinutes | Percentage of FairlyActiveMinutes to the sum of VeryActiveMinutes, FairlyActiveMinutes, Lightly ActiveMinutes and SedentaryMinutes |
| PCT\_LightlyActiveMinutes | Percentage of LightlyActiveMinutes to the sum of VeryActiveMinutes, FairlyActiveMinutes, Lightly ActiveMinutes and SedentaryMinutes |
| PCT\_SedentaryMinutes | Percentage of SedentaryMinutes to the sum of VeryActiveMinutes, FairlyActiveMinutes, Lightly ActiveMinutes and SedentaryMinutes |
| BMI | Body Mass Index |
| Calories | Total calories burned |
| Total Hours Sleep | Number of sleeping hours per day |
| Sleep Quality | 1: poor quality with sleep hours < 5 hours per day  0: good quality with sleep hours >= 5 hours per day |
| Weekday | 1: Sunday; 2: Monday; 3: Tuesday; 4: Wednesday  5: Thursday; 6: Friday; 7: Saturday |

**Evaluation Criteria**

* **Appropriate use of technology**: this involves correct implementations of technology for data visualization and model fitting.
* **Appropriateness of analysis:** this involves efficiently extract information from data through visualization and select appropriate predictive models to fit the data.
* **Clarity and relevance of research objective:** analysis is connected to the goal of this project such as justifying the selections of the predicting variables for the model and strategic recommendations.
* **Communication of findings:** effectively interpret the analysis outcomes and propose strategies based on the insights obtained from the analysis.

**Written Document**

A report with maximum 10 pages (excluding Appendix) including the following sections:

1. *Introduction*: briefly describe the problem and goal of this projet.
2. *Data visualization*:

* Create plots to visualize the relationship between **each** of the following 5 variables with ***Sleep Quality***:

*VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes,* and *Calories*

* Discuss what you find from the visualizations.
* Use the **entire** dataset for your visualization

1. *Major analytical methodology*:

* You may select one of the following predictive models to predict poor quality sleepers: (1) classification tree (2) logistic regression model
* Your target variable should be **Sleep Quality**
* Use the 5 variables in the Data Visualization section as the predicting variables for your model
* Since this is supervised learning, you will need to split the model into Training and Test datasets. Keep the 70% as the training and 30% as the test data. **Fit your model to the training data**.
* Present your fitted model and interpret the results. (If you choose the classification tree as your methodology, you also need to present a plot of your classification tree.) For example, how the signs of the estimated coefficients are associated with poor sleep. Which rules will lead to poor sleep (outcome=1) ?
* Model performance assessment: evaluate your model prediction performance based on the **test data**. Present the evaluation metrics such as Error Matrix and ROC curve. Comment on the prediction performance of your model.

1. *Strategic Recommendations*:

* For each category of Sleep Quality, create plots to visualize the trends of the following 4 variables from Monday to Sunday:

PCT\_VeryActiveMinutes, PCT\_FairlyActiveMinutes, PCT\_LightlyActiveMinutes, and PCT\_SedentaryMinutes

* Based on the insights from the trends and your fitted model, design strategies to communicate with poor sleepers and encourage them to engage in physical activities.

1. *Appendix*:

* include screenshots of each step in Rattle R for model building and evaluation.