An Analysis of Health & Lifestyle Factors Associated with Sleep Quality

Kevin Kosewick and Emma Hardin-Parker

2024-04-29

# 1. Summary/Abstract (1)

For our project, we obtained a dataset from Kaggle.com containing 373 observations of 13 variables from a study on different factors that may affect sleep quality. Variables of interest from the study included physical activity level, age, sex, occupation, stress, gender, BMI, blood pressure, and sleep duration. In [RStudio](https://posit.co/download/rstudio-desktop/) (version 4.3.1), we used a combination of packages to create models suitable for answering our primary question: which variables are most important for determining an individual’s sleep quality and what effect do they have? After conducting an exploratory data analysis (EDA) we hypothesized that stress, sleep duration, and occupation would have the largest impact on sleep quality. We believed that stress and sleep quality would be negatively correlated, sleep duration and quality would be positively correlated, and that jobs with longer or more variable hours would negatively affect sleep quality.

After creating linear models for each variable’s impact on sleep quality and fitting a Random Forest model with Cross-Validation to test variable importance, we determined that sleep duration, stress level, and age were the most important factors. Our hypothesis for sleep duration and stress was supported; however, we did not correctly predict that age would be one of the most important predictors for sleep quality. Our hypotheses regarding the effects of sleep duration, stress, and occupation were all supported as well. Overall, we determined that an individual’s duration of sleep, stress level, and age are all key factors that can impact sleep quality.

# 2. Introduction (2)

## 2.1 Background Information

The data was obtained from kaggle.com at this [link](https://www.kaggle.com/datasets/henryshan/sleep-health-and-lifestyle/data). The data contains observations from a study on individuals of varying sexes and age regarding their sleep quality and different variables that may affect this. We are unsure of where this data set came from and how the data were collected; the publisher on Kaggle has yet to respond to a comment with this question posted on March 5th 2024. Regardless, it contains interesting variables that have been shown in previous studies to have an effect on sleep quality.

For instance, in Sun et al. (2015), a sample of Chinese individuals showed a correlation between obesity and worsened sleep quality in men but not women. We are interested in seeing if this trend in present in our dataset; however, given that most of the observations are of men, we expect to see a correlation between lower sleep scores and obesity in our data. Furthermore, it has been shown that age and gender have a notable effect on sleep quality (Madrid-Valero et al. 2016). Women seem to experience a deterioration in sleep quality as they age; however, while this trend is still present in men, it is notably less consistent and can vary dramatically between individuals. We expect to see similar trends in our dataset, but as stated previously, our observations contain significantly more men than women. Finally, we wish to test whether or not certain occupational industries have different effects on sleep quality. It has been shown that managerial positions tend to have the worst sleep quality among civilian sector workers while 24 jobs that have rotating shifts tend to have the worst sleep quality (Luckhaupt et al. 2010). Our dataset contains observations of mostly white-collar/service-based jobs, so we expect to find similar trends upon analysis.

Additionally, based on the findings of Sun et. al that obesity can negatively impact sleep quality, we also expect that higher levels of blood pressure will result in poorer sleep quality. The American Sleep Foundation states that both short and long sleep duration increase risk of hypertension in adults (Li & Shang, 2021). This draws interesting focus to the circular nature of sleep quality and hypertension. Those with hypertension also face a higher risk of insomnia, a disorder affecting the ability to fall asleep, maintain sleep, and sleep without continuous disturbance (Liu et. al, 2022). Short sleep duration is also positively associated with stress level (APA, 2013). Survey findings indicate that the average American sleeps 6.7 hours per night, which is well below the minimum recommended duration seven to nine hours (APA, 2013). Similar to the relationship between hypertension and sleep quality, as stress level increases, sleep duration decreases thereby causing poorer sleep quality. Poorer sleep quality then leads to higher stress, greater risk of hypertension, and shorter sleep duration.

Engaging in physical activity has been suggested to improve both sleep quality and sleep duration. Regular moderate to intense physical activity stimulates an increase of melatonin production, a hormone that regulates sleep-wake cycles (Alnawwar et. al, 2023). Additionally, moderate to intense exercise has shown to have a positive effect on those with insomnia (Alnawwar et. al, 2023). Physical activity also reduces stress and improves both sleep quality and sleep efficiency (Alnawwar et. al, 2023). However, studies also show that frequent high-intensity workouts could lead to poorer sleep quality among people with and without insomnia (Alnawwar et. al, 2023). Studies indicate that consistent moderate exercise three to four times a week (approximately 150 minutes total) is the optimal weekly regimen to improve sleep quality and reduce stress. We expect to see similar trends with our data where increased stress levels negatively impact sleep quality. We expect to see a positive association between increasing physical activity levels and sleep quality until the point of “too much” activity.

# 3. Materials and Methods (3)

## 3.1 Data and Processing

The dataset originally contained 373 observations for 13 variables covering a broad spectrum of health and lifestyle variables associated with sleep quality. The curators of the dataset did not provide information on how or where this data was collected, so we solely relied on the codebook for variable definitions. We have added this codebook to the supplementary materials file for ease of access but you can also find it on the Kaggle page for this dataset. There were no missing values or erroneous variables in the dataset that we needed to remove. The “Blood.Pressure” variable was transformed to reflect the individual systolic and diastolic blood pressure measurements from each subject. We then chose to add an additional variable, labeled ‘cat\_bp’ to reflect the categorical blood pressure status of a subject based on the [American Heart Association guidelines](https://www.heart.org/en/health-topics/high-blood-pressure). Additionally we created a variable named ‘StepsGroup’ to determine the categorical activity level of subjects in the dataset based on these guidelines provided by [10,000 steps](https://www.10000steps.org.au/articles/healthy-lifestyles/counting-steps/#:~:text=The%20following%20pedometer%20indices%20have%20been%20developed%20to,day%205%20Highly%20active%20is%20more%20than%2012%2C500) Finally, we created an additional categorical variable called ‘PhysicalActivityGroup’ to categorically represent the level of physical activity reported by a subject. The levels are differentiated by 30 minute intervals of weekly physical activity. The outcome of interest is noted as “Quality.of.Sleep” that we will reference as Sleep Quality or Quality of Sleep throughout this report.

## 3.2 Variables included in Analyses

Given that there were only 17 variables after final data transformation, simple linear regression models were performed to determine baseline associations between variables and Sleep Quality (the outcome of interest). These regression models were created to determine which variables had the largest impact on Sleep Quality and drove the rest of the analyses. Simple linear regression models were fitted for the following variables: BMI, cat\_bp, Stress Level, Physical Activity Level, Sleep Duration, Gender, Age, and Occupation. The other variables were determined to have too weak of a relationship after our EDA and were therefore left out of the analyses.

The Occupation variable was then transformed to group the various occupations into the following groups: Healthcare, Education, Engineering, Business/Finance, and Science. Prior to performing any subsequent analyses, we removed the original versions of the variables we transformed: systolic, diastolic, daily steps, occupation, Physical.Activity.Level, Heart.Rate, and AgeGroup.

The objective of our analysis was to determine which of these variables are most important for determining an individual’s sleep quality and what specific effect they have. After conducting an EDA we hypothesized that stress, sleep duration, and occupation would have the largest impact on sleep quality. We believed that stress and sleep quality would be negatively correlated, sleep duration and quality would be positively correlated, and that jobs with longer or more variable hours would negatively affect sleep quality.

## 3.3 Model Development

Random Forest Models and 5-fold Cross-Validation were considered to be the “best” fit for our outcome of interest, Quality of Sleep. The data were not split into train/test subsets as the data contains less than 400 observations and many of the values were unique in comparison to the rest of the data. Prior to constructing this model, a colinearity plot was constructed to determine any presence of colinearity in our data.

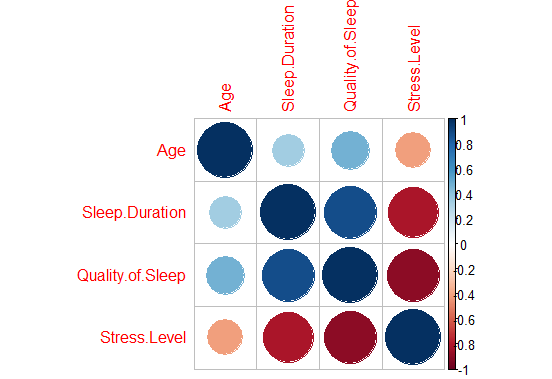


Figure 1: A colinearity plot demonstrating the colinearity between our variables.

Stress and Sleep Quality had an absolute value of 0.9 on the correlation scale. Sleep Duration and Stress were also seen to be strongly correlated. Stress was still included in our models due to its well-studied impact on sleep; however, we found it important to note the strong linear relationship between Stress and Sleep Quality.

## 3.4 Defining the Models

We fit GLMs for each predictor in order to determine their effect on sleep quality. Each model only used a single predictor as the only aim was to quantify and understand the impact of each for our outcome of interest. For our Random Forest model, we used mtry (with a range of 1-7) and min\_n (with a range of 1-21) as tuning parameters to create 300 different decision trees. Mtry controls how many predictors a decision tree can consider at any given point in time. It adds randomness to the decision tree creation process and ensures that all of the trees do not look the same. Min\_n specifies the minimum number of samples that should be present in the leaf node after splitting a node. In our case, this means that if any terminal node has more than 21 observations and is not a pure node, we can split it further into subnodes. This was all done using 5-fold Cross-Validation repeated 5 times in order to train the data to predict on an unseen dataset.

## 3.5 Evaluation of Models

Root Mean Squared Error (RMSE) is a common metric used to evaluate regression models. The RMSE is formally defined as the square root of the mean square of all error and is defined by the following formula:

It is important to consider that while RMSE is scale dependent, common practice notes that low RMSE values indicate stronger model performance.

R-squared (R²) is a statistical measure used in regression analyses. It quantifies the proportion of the variance in a dependent variable that can be explained by an independent variable within a regression model and is calculated with the following formula:

Null models provide a baseline for the evaluation of a model’s performance. They generate predictions for an outcome of interest without using any predictor variables. Instead, they simply average the values of the outcome of interest to make these “predictions”. We used both RMSE and R² values to evaluate the performance of our model compared to a “null model”.

## 3.6 Software Used for Analyses

This analysis was conducted under R version 4.3.1 on a MacOS operating system. The following R packages were used in the development of these analyses: here, skimr, broom, tidyverse, ggplot2, dplyr, corrplot, ranger, and vip. All processing and analysis code can be found in the Supplementary Material file.

# 4. Results (4)

## 4.1 Outcomes of Interest

The ‘Quality of Sleep’ variable has been selected as our outcome of interest. It is important to note that this variable is a subjective measure of a participant’s self-reported sleep quality. Figure 2 shows that most subjects reported a sleep quality score of 8 (out of 10).

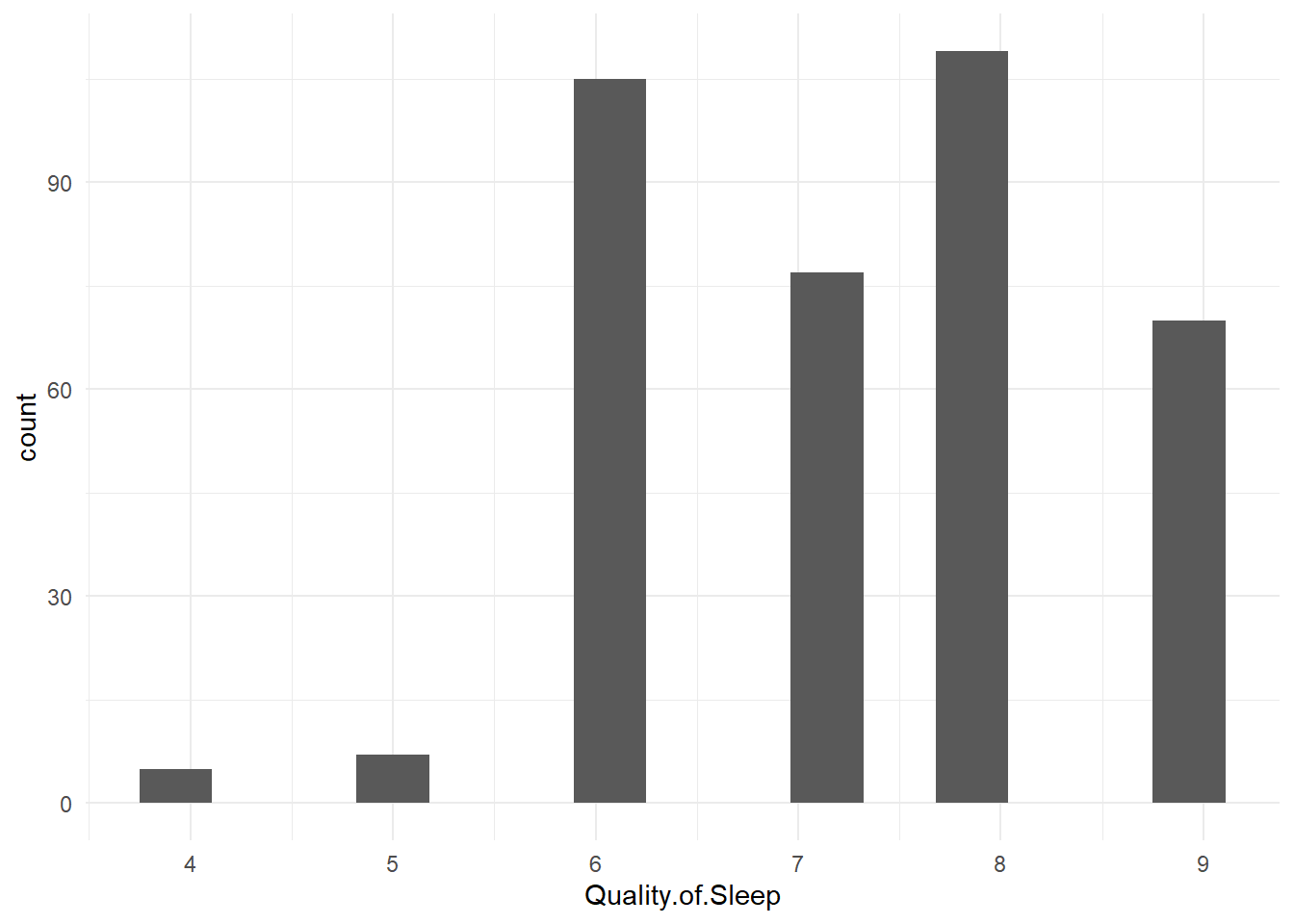


Figure 2: Distribution of Sleep Quality Scores in the dataset.

Our Random Forest Model built with 5-fold Cross-Validation explored different combinations of predictors to determine which parameters gave the most accurate predictions of sleep quality.

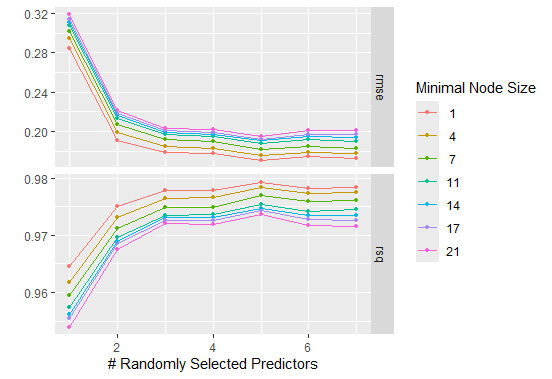


Figure 3: The results generated from our RF model. RMSE increases when more predictors are added to the model. The R-squared value also decreases as more predictors are added to the model.

Figure 3 shows that as we include more predictors in our model, RMSE increases slightly while the R-squared value decreases. This indicates that some variables in our dataset are more important for predicting sleep quality than the others. Based off of our variable importance graph, we can assume that stress level, sleep duration, and age are the factors that carry the most weight when predicting sleep quality.

We went on to evaluate how well our RF model predicted values by creating an observed vs. predicted plot as seen in figure 4.

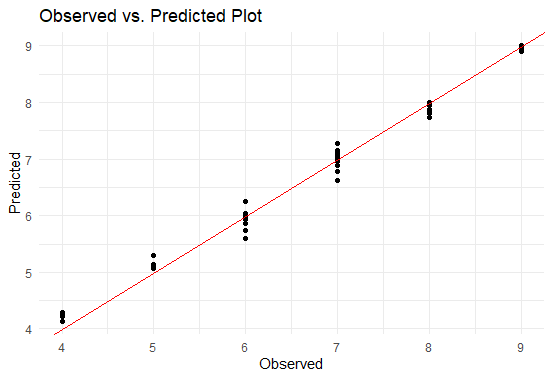


Figure 4: A comparison of the actual values for sleep quality in our dataset (observed) compared to the values predicted by our fitted RF model.

While the predicted values don’t align perfectly with the red “ideal value” line, they fall very close overall, indicating that our model can sufficiently predict sleep quality using unseen data.

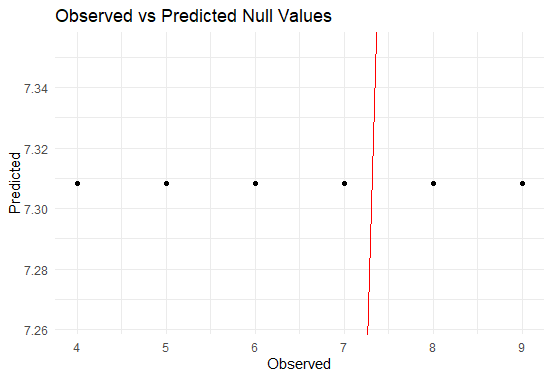


Figure 5: The observed vs. predicted values for our “Null model”.

We can see in figure 5 that the null model only predicts values that are the average of our observed values. As expected, these predictions are far from our actual observed values.

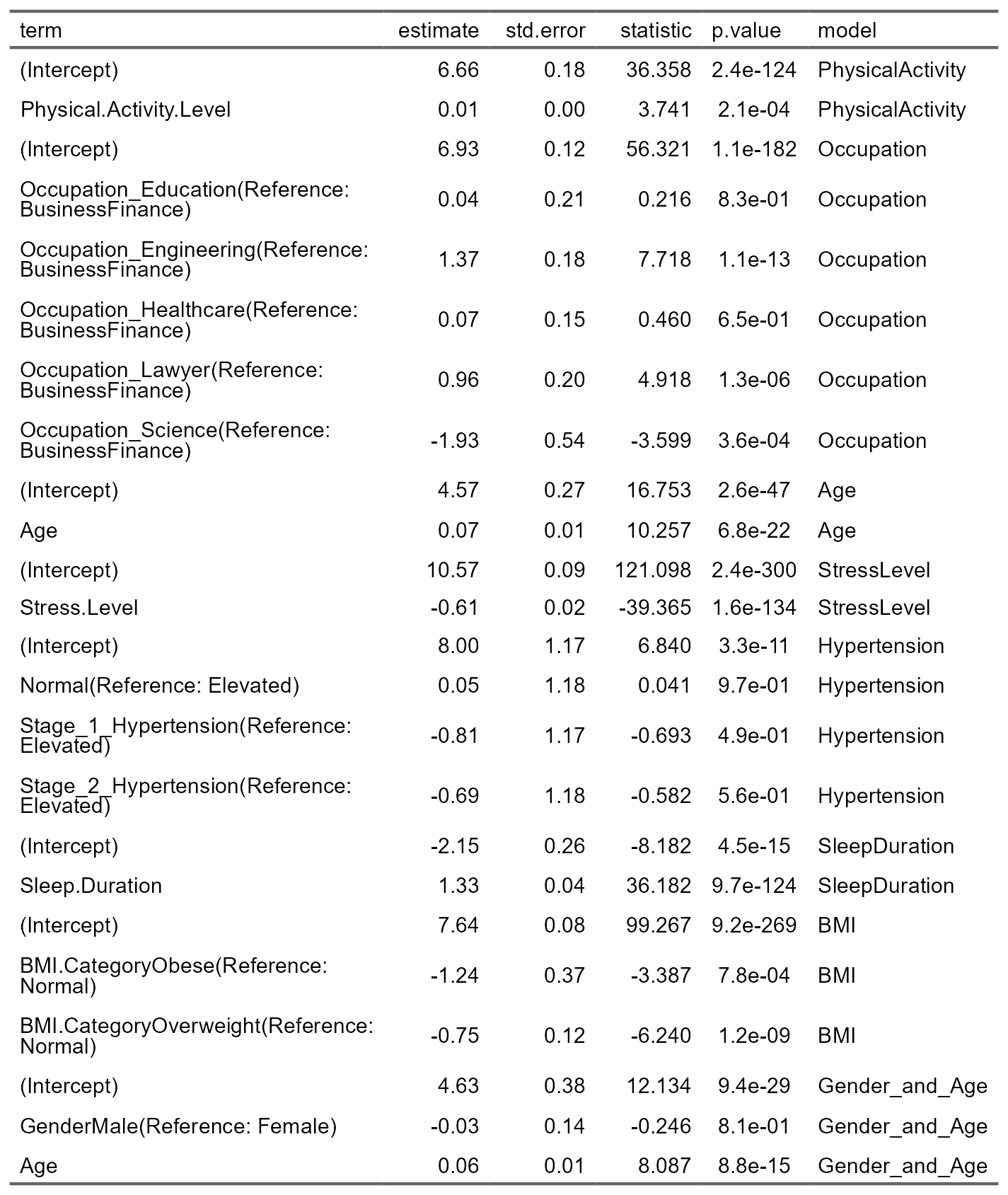


Table 1: The results of our GLMs. The “term” column identifies the category/variable being analyzed. The “model” column identifies the predictors used to fit each model. The “statistic” column is a relative measure of the strength of the relationship between each predictor and sleep quality. The “estimate” column shows the predicted change in sleep quality score as a result of each associated predictor. Each reference category that is used for categorical variables is included next to the term in parentheses (e.g. Reference: Female).

We believed that stress and sleep quality would be negatively correlated, sleep duration and quality would be positively correlated, and that jobs with longer or more variable hours would negatively affect sleep quality. As seen in table 1, we found that all of our assumptions were supported. The “intercept” values under the term column represents what the expected sleep quality score would be if a continuous predictor was at a value of 0. For categorical variables, it represents the reference category that the others are being compared to.

## 4.2 Machine Learning Models

A variable importance plot (VIP) (figure 6) was constructed using the “best” set of parameters (as guided by our RF model) to determine which variables had the strongest impact on sleep quality.

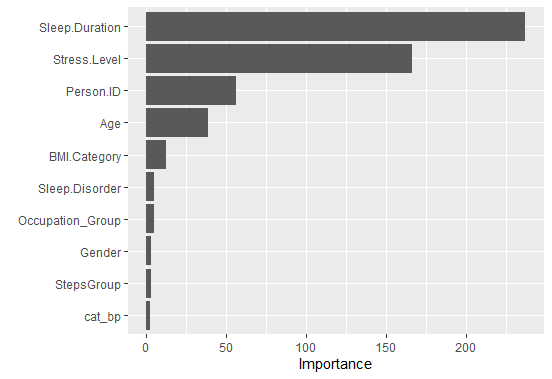


Figure 6: The ranking of each variable’s importance in determining sleep quality according to our Random Forest model.

As seen in figure 6, Sleep Duration, Stress Level, and Age had the highest levels of importance in our model. Physical Activity Level was excluded entirely, which was to be expected given our exploratory data analysis and simple linear regression models.

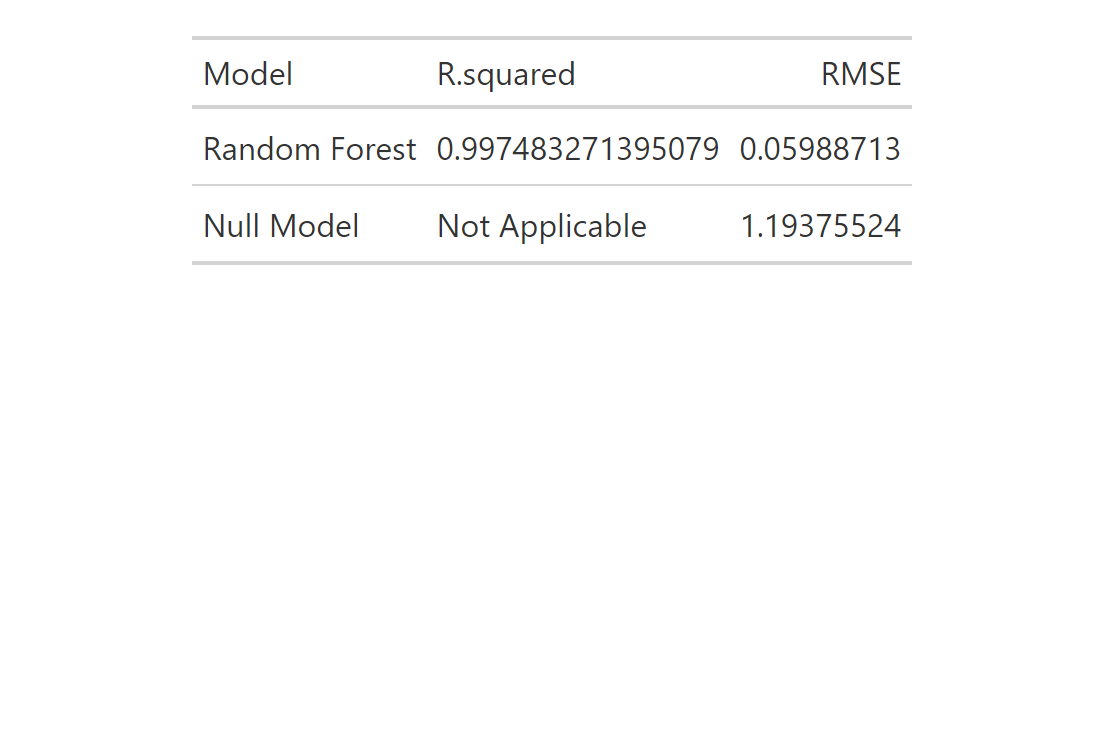


Table 2: A comparison of the RMSE and R-squared values generated by the null and RF models. The null model has no R-squared value due to its method of generating predictions.

As seen in table 2, our null model was found to have an RMSE of 1.193; compared to our best parameter RF model’s RMSE of 0.059, this is relatively high. The R-squared value doesn’t exist for the null model since there is technically no variation between the observed and predicted values.

# 5. Discussion (5)

## 5.1 Summary

The differences between the null model and RF model’s metrics show that our RF model is better suited for predicting sleep quality on unseen data than a model using no predictors. The R-squared value of 0.997 is extremely high; this shows that a large amount of variation is explained using our RF model. The variable importance plot we created answers our questions about which variables are most important for predicting sleep quality. As we can see, sleep duration, stress level, and age were the top three predictors for sleep quality in our data. This reflects the relationships we saw in our colinearity plot as well as the clear correlations we found between the variables in our EDA. We did not expect to see age so high on the list; this could be due to us grouping ages together in our EDA and therefore misleading us in our preliminary data exploration. We did not use the groups when creating our models as we learned that grouping continuous variables is a poor practice as we progressed in our MADA course.

Through our linear models, we determined that our hypotheses regarding each variable’s effect on quality of sleep were supported. The variables determined to be the most important by our RF model had the following effects on sleep quality: sleep duration was positively correlated, age was positively correlated, and stress level was negatively correlated.

## 5.2 Strengths and Limitations

The main limitations of our models stemmed from the data itself. As stated in our introduction section, we have been unable to identify exactly where the data were collected and how it was gathered. The values for stress level and sleep duration were also based on a subjective rating given by each participant in the study; this could introduce individual bias to the data and prevent an objective evaluation of the effect of these variables. An inherent limitation in RF models is difficulty of interpretation; we attempted to minimize this by creating easy to read tables and graphs for the results. We chose not to split our data into a train/test split due to the small number of observations and presence of rare values in our data. We believe that fitting using CV minimized the downsides of this. Future analyses would benefit from more detailed meta data and a better understanding of the experimental design that generated our observations.

The strengths of our analysis lie in the easily interpretation of our GLMs and the predictive power of an RF model. We were able to clearly display and quantify the relationships between our predictors and sleep quality, answering our question about each variable’s effect on our outcome of interest. We were also able to create a powerful RF model that explained nearly all of the variation in our data as well as a ranking of each variable’s impact on sleep quality. Fitting our RF model with CV allowed us to train the model to create predictions for “unseen” data as well.

## 5.3 Conclusion

Overall, while our study limitations prevent us from generalizing the findings of this analysis broadly, we were able to clearly answer our questions about the data. Our findings supported our hypotheses about the effects of each predictor on sleep quality. However, our hypothesis about which variables would be the most important was not supported by the analysis. Instead, we determined that stress level, sleep duration, and age were the most important predictors for our outcome of interest. Ultimately, the results indicate that the sleep quality of this study’s participants is impacted disproportionately by an assortment of variables dealing with physical, mental, and environmental factors.

# 6. References