Sleep Efficiency and Number of Awakenings analysis with Statistical Learning approaches

Nicolas Anselmi, Andrea Arici, Francesco Corrini

University of Bergamo

Keywords: Statistical Learning, Sleep Efficiency, Awakenings, Lasso, Ridge Regression, Stepwise, Random Forest, Poisson Regression

Abstract

It's well known that insufficient sleep leads to different problems such as cardiovascular disease, obesity or decrease in productivity[1]. Although some people spend too little time sleeping, others don't get as much sleep as they would like, having a negative impact on their life. Sleep efficiency is the percentage of time spent asleep compared to time spent in bed and it's useful to understand if a person is able to sleep the right amount of time. The aim of this work is to study how different factors affect sleep efficiency using statistical learning approaches and find a model to make predictions. In this project will be also analyzed how these factors can influence the number of awakenings a person has during the sleep time.

1 Introduction

One of the most time consuming activity in people life is sleeping. It is common for many people to give up hours of sleep to spend more time working, studying or having fun. However, this practice can lead to many problems, such as cardiovascular problems, obesity, diabetes and decrease of productivity[1]. There are also other factors that affect how much time a person sleeps. A research team from the University of Oxfordshire did a study aimed to investigate the impact of lifestyle factors on sleep quality. A part of this study was collecting data from real person, using a combination of self-reported surveys, actigraphy and polysomnography which is a sleep monitoring technique. This data are available on Kaggle[2]. The aim of this work is, using the provided dataset, to make inference on the "sleep efficiency" variable, which is a measure of the proportion of time in bed spent asleep, obtained dividing the time spent asleep by the total time spent in bed. Another goal is to estimate a model that can make prediction on sleep efficiency. Another way to measure sleep quality is to count the number of awakenings during bed time. In fact, waking up many times during the night can affect a person's performance the next day. In this work a model to make predictions on the number of awakenings will be presented. Statistical learning approaches are used to achieve these three goals[3].

Every feature in this dataset is from a different subject and there aren't multiple measurements on the same person. Age and gender are also recorded. The raw dataset contain fifteen attributes and 468 observations. The "Bedtime" and "Wakeup time" features indicate when each subject goes to bed and wakes up each day and the “Sleep duration” feature records the total amount of time each subject slept in hours. The “REM sleep percentage”, “Deep sleep percentage”, and “Light sleep percentage” features indicate the amount of time each subject spent in each stage of sleep. Additionally, the dataset includes information about each subject’s caffeine and alcohol consumption in the 24 hours prior to bedtime, their smoking status, and their exercise frequency. “Bedtime” and “Wakeup time” features has been removed from the initial dataset: the difference between them is used (“Sleep duration” feature). Furthermore the feature “Light sleep percentage” has been removed from the dataset because there is a linear dependency with “REM sleep percentage” and “Deep sleep percentage”.

The observations that included missing values has been removed and the two categorical variables “Gender” and “Smoking status” are treated as follows: zero in “Gender” means the female sex and one means male sex and, where the “Smoking status” is zero, it means that the subject doesn’t smoke, one in the opposite case.

In the end the edited dataset has 388 data points and ten features (eleven considering sleep efficiency). Features frequency have been plotted in Figure 1 and Figure 2. Frequency of the two response variables (“Awakenings” and “Sleep efficiency”) is shown in Figure 3. In order to understand the correlation relationship between all the features in the dataset, the correlation matrix has been calculated and presented in Figure 4. Features description is summarized in Table 1.

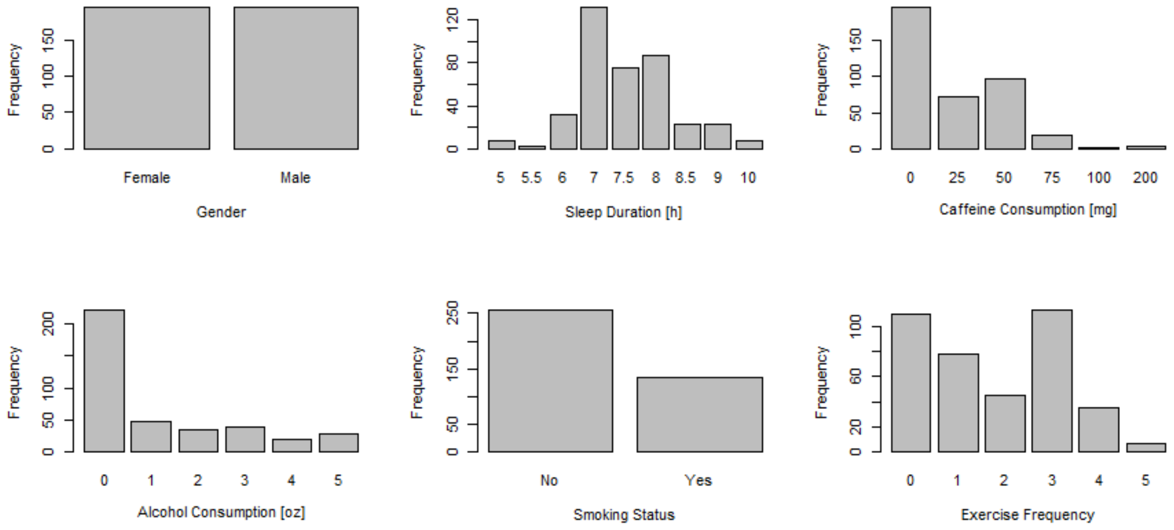

Figure 1: Frequency of non-continues features represented using barplots

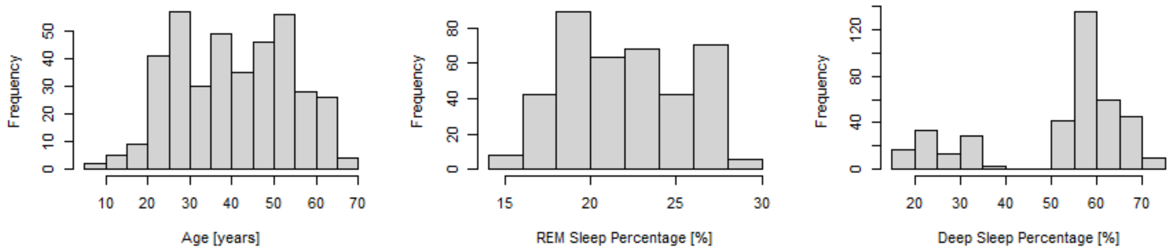

Figure 2: Frequency of continues features represented using histograms

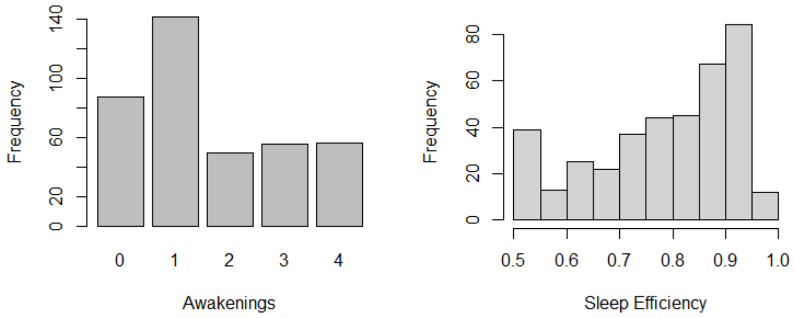

Figure 3: Frequency of awakenings and sleep efficiency, the objects of this study

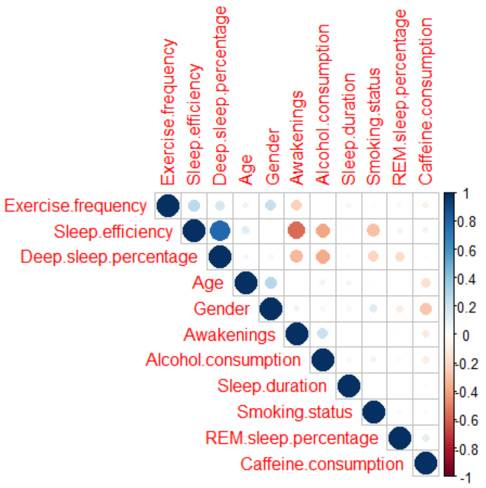

Figure 4: Correlation matrix of the features in the dataset. None of these is hard correlated with others, so none has to be removed

Feauture	Description	Measure unit	Name in dataset
Sleep efficiency	Ratio of time asleep on total time spent in bed	-	Sleep.efficiency
Awakenings	Number of awakenings during the sleep time	-	Awakenings
Age	Age of the person taken in exam	years	Age
Gender	Sex of the person taken in exam: male or female	-	Gender
Sleep duration	Total hours of actual sleep	hours	Sleep.duration
REM sleep percentage	Ratio of the time in REM phase on total sleep time	-	REM.sleep.percentage
Deep sleep percentage	Ratio of the time in deep sleep phase on total sleep time	-	Deep.sleep.percentage
Caffeine consumption	Quantity of caffeine assumed in the 24 hours before the measurements	$[mg]$	Caffeine.consumption
Alcohol consumption	Quantity of alcohol consumed in the 24 hours prior to bedtime	$[oz]$	Alcohol.consumption
Smoking status	Whether or not the test subject smokes	-	Smoking.status
Exercise Frequency	Number of times the test subject exercises each week	-	Exercise.frequency

Table 1: Summary of features in the dataset

2 Methodology

As shown in the introduction, the first aim of this work is to understand which features affect sleep efficiency the most. In order to do this, methods for selection and regularization for linear models are used, in particular stepwise regression and lasso regularization.

The first method taken into account is stepwise selection. There are two version of this approach: forward stepwise and backword stepwise. Forward stepwise starts estimating the null model M_0 that is the mean of the response variable. Having p predictors, at the iteration k it estimates $p - k + 1$ linear regression models using $k - 1$ predictors selected in the past iteration and one from the features that aren't already chosen. The M_k model is selected among them using Akaike Information Criterion (AIC). In the end $p + 1$ model are estimated: the best is selected using AIC. Backward stepwise starts estimating the full model M_p with all predictors. At iteration k, $p - k + 1$ models are estimated removing one of the $k + 1$ predictors from M_{k+1}: the M_k model is the one among these that minimize AIC. As in forward stepwise, the best among the $p + 1$ estimated models is selected using AIC. Finally, best model between forward and backward approach is selected using AIC. Stepwise selection is then combined with a bootstrap simulation approach: at every bootstrap iteration, the stepwise algorithm is performed on a resampled dataset, counting which features the model picks. In the end, if a predictor was included in less than half models it will be discard[4]. Remaining features are used to estimate a linear regression model to make inference on. The model is validated using a bootstrap approach computing test root mean squared error (RMSE) and R^2 at every bootstrap iteration in order to have a distribution of it. Also the coefficients distributions are measured to see which of them are statistically significant, providing a confidence interval.

The second method proposed is lasso regularization. With this approach coefficients are estimated solving the following optimization problem

$$\hat{\theta} = arg \min_{\theta}[(Y - X\theta)^T(Y - X\theta) + \lambda\|\theta\|_1] \tag{1}$$

where $\hat{\theta}$ is the $k \times 1$ vector of estimated coefficients, Y is a $N \times 1$ vector containing values of the response variable, X is the $N \times k$ matrix containing all features values and λ is a penalization coefficient. The greater the value of λ the greater will be the shrinkage of the model parameters, reducing some of them to zero, making it a good approach for features selection. The optimal λ is selected performing 10-fold cross-validation, selecting the one that minimize MSE. As in stepwise approach, a bootstrap simulation is performed to understand which features are relevant, discarding those which has been included in less than half of models. During this simulation the model is validated computing RMSE and the R^2.

In the end stepwise and lasso results are compared to understand which is best to make inference. In particular, the best model is chosen analyzing the trade-off between model's performance (less RMSE and higher R^2) and complexity (lower number of features).

Second goal of this study is to make prediction on sleep efficiency. For this purpose more complex methods like ridge regression and ensemble methods (random forest and bagging) are considered.

The first approach taken into account is ridge regression, an approach similar to lasso that use the following objective function:

$$\hat{\theta} = arg \min_{\theta}[(Y - X\theta)^T(Y - X\theta) + \lambda\|\theta\|_2^2] \tag{2}$$

Using norm-2 instead of norm-1 in the penalization term makes the coefficients to become smaller but not zero. This means that all features will be included in the final model. Thus, ridge regression could lead to better performance, in sense of less RMSE and higher R^2. Similar to lasso, the best λ is chosen with 10-fold cross-validation and bootstrap is used to compute RMSE and R^2 distribution.

The second approach considered is random forest. Random forest is an ensemble method which consist in using bootstrap simulation to estimate different trees over different (resampled) dataset and take as prediction the average of predictions of all the trees in the model. This approach tends to reduce the variance respect to classical single tree estimation approach. Random forest also use only a subset of features in every iteration to avoid estimating similar trees. In this study the subset has dimension $p/2$ where p is the number of features. In this work is also considered bagging approach, which is like random forest but using all predictors for every tree.

All these approaches will be evaluated using RMSE and R^2. For prediction is useful to choose the model that minimize RMSE, even if it's too complex to understand. However, results are compared with linear models to understand if it's necessary and useful to use more complex approaches. In the end predictions are made on a test dataset. Punctual confidence interval at 95% on predictions are provided and coverage (the percentage of test data that is inside the estimated interval) is computed. A good model should have the coverage equal to the confidence level.

Last goal of this work is to estimate a model to make prediction on the number of awakenings. To achieve this Poisson regression[5] is used. Poisson regression model takes the form

$$log(E(y|x)) = log(\lambda) = \alpha + \beta' x \tag{3}$$

where x is a $p \times 1$ vector of independet features, y the response variable value and λ the mean of the Poisson distribution that is given by

$$p(y|x;\theta) = \frac{\lambda^y}{y!}e^{-\lambda} = \frac{e^{y\theta' x}e^{-e^{\theta' x}}}{y!} \tag{4}$$

where $\theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$. The likelihood function L can be written as follow:

$$L(\theta|X,Y) = \prod_{i=1}^{N} p(y_i|x_i,\theta) \tag{5}$$

where N is the number of observation, X is the $N \times p$ matrix containing all features values and Y is the $N \times 1$ vector containing the response variable. $\hat{\theta}$ can be computed maximizing the likelihood function. Instead, an easier way to compute $\hat{\theta}$ is to maximizing $log(L)$ function.

The optimization problem can be rewritted as

$$\hat{\theta} = arg \max_{\theta} \sum_{i=1}^{N} (y_i \theta' x_i - e^{\theta' x_i}) \tag{6}$$

The model is validated using bootstrap simulation to provide a confidence interval of the coefficients $\hat{\theta}$. A test dataset is used to understand if the model can make good predictions on the number of awakenings.

3 Results Analysis

3.1 Sleep efficiency inference results

Stepwise approach starts with a bootstrap simulation that count how many times features are included in the estimated models. Results of this computation are in Table 2. “Gender” and “sleep duration” features are included in less then half of bootstrap iteration and so they are discarded. Remaining features are used to estimate a linear regression model. This model is validated using a bootstrap simulation: results of this are in Table 3. All predictors are statistically significant. RMSE is 0.0623 and R^2 is equal to 80.9%.

Feature name	Feature frequency in SW	Feature frequency in lasso
Age	97 %	99.8 %
Gender	16.7 %	64.9 %
Sleep.duration	18.1 %	72 %
REM.sleep.percentage	100 %	100 %
Deep.sleep.percentage	100 %	100 %
Awakenings	100 %	100 %
Caffeine.consumption	69.5 %	96.7 %
Alcohol.consumption	83.9 %	97.6 %
Smoking.status	100 %	100 %
Exercise.frequency	85 %	98.6 %

Table 2: Features frequency in models estimated during bootstrap simulation for stepwise and lasso approach. Features that are selected in less than the 50% of models are discarded

Lasso approach starts computing the optimal λ using cross-validation technique. Results of this are reported in Figure 5. Using optimal λ (0.000704), bootstrap simulation is applied similar to stepwise to count how much times a regressor is picked by lasso approach. Results of this are in Table 2. RMSE of the final model is 0.0626, R^2 is 78.3%, and the coefficients value is showed in Table 4.

Feature name	Coefficient	Lower CI.	Upper CI
Intercept	0.363	0.294	0.429
Age	0.000957	0.000393	0.00148
REM.sleep.percentage	0.00665	0.00461	0.00866
Deep.sleep.percentage	0.00556	0.00496	0.00615
Awakenings	-0.0319	-0.0373	-0.0264
Caffeine.consumption	0.000241	0.0000244	0.000498
Alcohol.consumption	-0.00623	-0.0117	-0.000841
Smoking.status	-0.0455	-0.0611	-0.0303
Exercise.frequency	0.00638	0.00158	0.0112

Table 3: Stepwise regression results: first column shows predictors that are in the final model, second, third and fourth coloumns are the coefficients' values and their interval of confidence at 95%

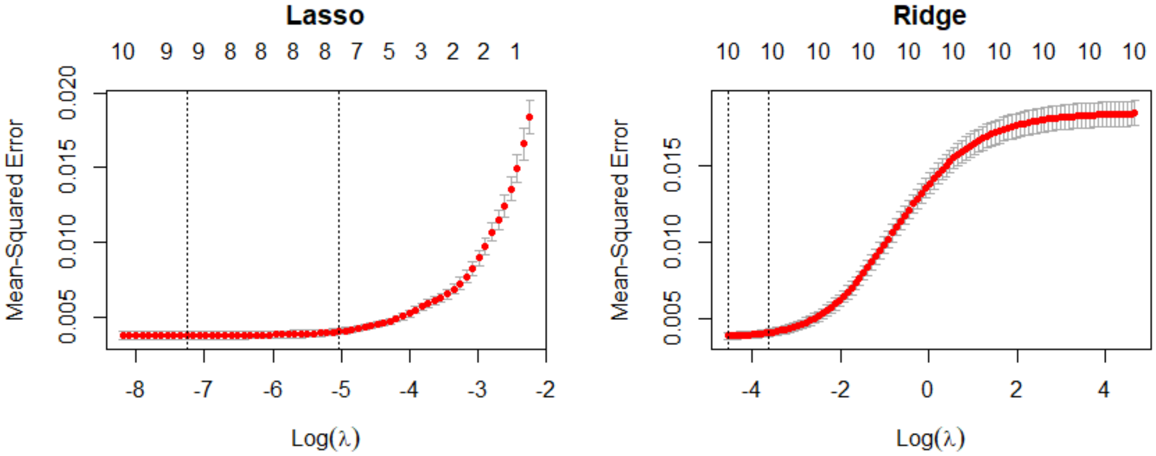

Figure 5: Test MSE of lasso and ridge regression approaches in function of λ. The selected λ is the one that minimize MSE. The values above the two plots represents the number of features selected at the current value of λ.

Feature name	Coefficient
Age	0.000571
Gender	0.00634
Sleep.duration	0.000532
REM.sleep.percentage	0.00685
Deep.sleep.percentage	0.00549
Awakenings	-0.0340
Caffeine.consumption	0.000150
Alcohol.consumption	-0.00756
Smoking.status	-0.0437
Exercise.frequency	0.00657

Table 4: Coefficients of the final linear model estimated with lasso approach

Stepwise and lasso approach give similar results in terms of performance. Stepwise model is selected because it has less features and so it has an easier interpretation. Analyzing the coefficients of the model, it appears that older people has a little bit better sleep efficiency. Also doing physical exercising increase the response variable. Smoking and drinking alcohol have a negative impact on sleep efficiency. As expected, the number of awakenings during bedtime decrease significantly sleep efficiency while high value of REM and deep sleep increase it. It's strange that caffeine consumption has a positive coefficient (also with lasso approach): it can be that subject who answered the survey assume the same amount of caffeine every day and so their metabolism is used to those doses and so consuming them doesn't affect their sleep efficiency.

3.2 Sleep efficiency prediction results

The first approach considered to build a prediction model is ridge regression. The optimal value of λ obtained by cross-validation is 0.0107, as shown in Figure 5. Test root mean squared error computed by the bootstrap simulation is 0.0629, with an R^2 of 78.4%.

Random forest approach gives a significant decrease in test RMSE, about 0.0432 (30% less respect to ridge regression) and an R^2 equal to 89.3%. The number of trees that gives it is 150, as shown in Figure 6. Bagging method leads to similar results $R^2 = 86.7\%$, with a test RMSE a little bit more (0.0473). Because of this random forest model is chosen to make predictions. The confidence intervals at 95% were calculated for every point in the test data by taking the quantile of the predictions' distribution. Then for each prediction it was verified whether it was within the confidence interval in order to get the coverage of the predictions. The coverage is 90%, which is very close to the expected one of the 95% confidence interval. However, confidence interval are too large (average range of prediction is large 0.29), so prediction aren't very accurate. In the end, these model aren't very useful to make prediction.

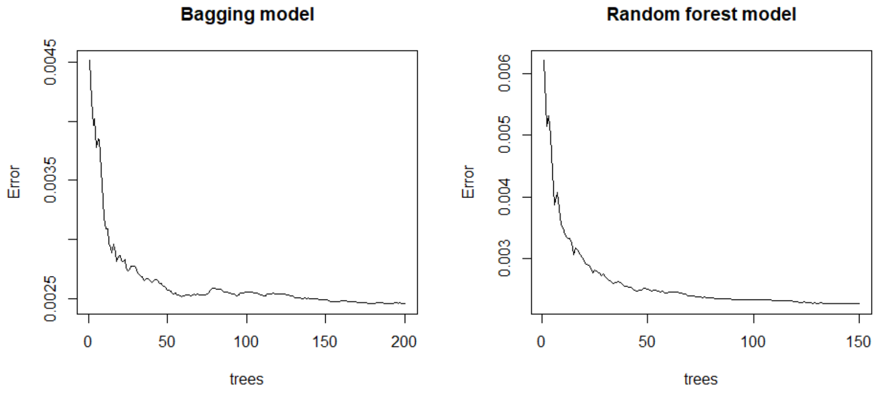

Figure 6: Random forest and bagging test MSE in function of the number of tree estimated

3.3 Awakenings prediction results

Results of bootstrap simulation for Poisson regression model are in Table 5. The only features that are statistically significant are “sleep efficiency”, “deep sleep percentage” and “smoking status”. The remaining features are used to estimate a new Poisson regression model. Coefficients of this model are in Table 6.

Feature name	Coefficients	CI down	CI up	Significance
Intercept	3.366	2.112	4.740	Yes
Age	0.00507	-0.00300	0.0131	No
Gender	0.168	-0.0621	0.414	No
Sleep.duration	-0.0151	-0.136	0.0926	No
Sleep.efficiency	-6.213	-7.630	-4.840	Yes
REM.sleep.percentage	0.0329	-0.0000472	0.0698	No
Deep.sleep.percentage	0.0227	0.0125	0.0335	Yes
Caffeine.consumption	-0.000474	-0.00544	0.00348	No
Alcohol.consumption	-0.00772	-0.0752	0.0620	No
Smoking.status	-0.450	-0.721	-0.170	Yes
Exercise.frequency	-0.0603	-0.133	0.0173	No

Table 5: Table shows coefficients (column 2) of the poisson regression validated with bootstrap approach. The third and forth columns are the coefficients’ interval of confidence at 95%. Last column shows if the coefficient is statistically different from zero

Feature name	Coefficients
Intercept	3.886
Sleep.efficiency	-5.575
Deep.sleep.percentage	0.0189
Smoking.status	-0.442

Table 6: Coefficients of the final poisson regression model with only the most relevant features

Results shows that increasing sleep efficiency decrease the average number of awakenings, while a greater percentage of deep sleep increase them. Also, a person who smokes has less average awkenings that one who does.

The model null deviance is 377, while the residual deviance is 250. Performing χ^2 hypothesis test with 3 degrees of freedom on the difference between them lead to a p-value that is less then 0.001. Thus, the model is able to make better predictions than the null model. On the other hand, Figure 7 shows that the model isn’t able to make good predictions on a test dataset. This may be due to the fact that the occurrence of one awakening isn’t independent from another for the same person (for example, the first awakening of a person can influence the second). That means that awakenings can’t be modeled with a Poisson distribution and so a Poisson regression model can’t be used to make predictions.

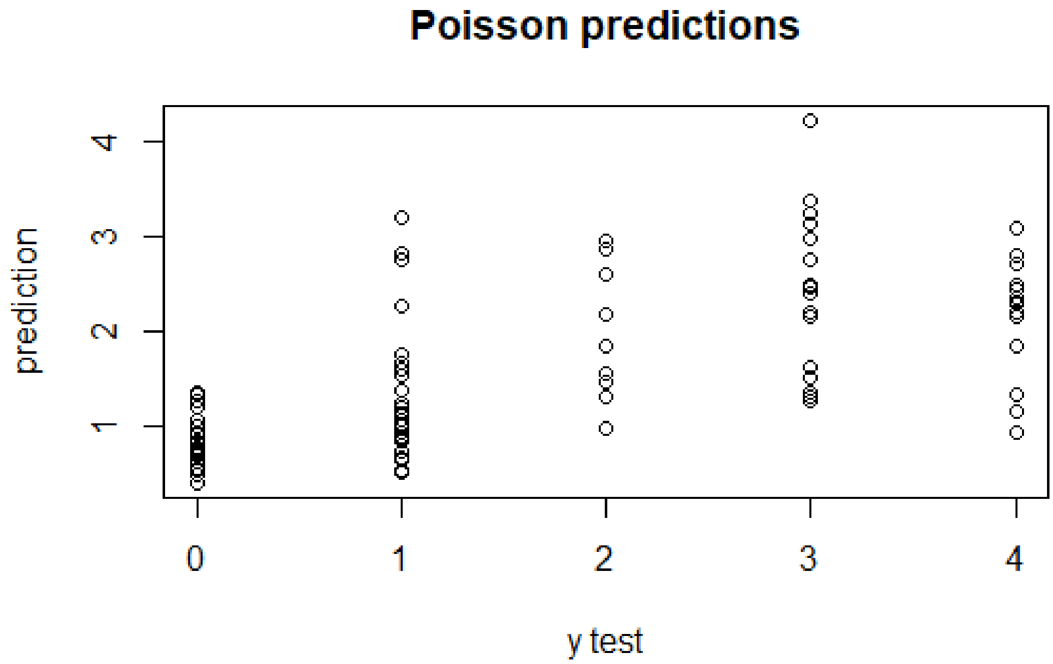

Figure 7: Poisson regression model's predictions on a test dataset compared to real value. It's easy to see that a lot of prediction are wrong.

4 Conclusions

In conclusion, stepwise and lasso both performed well in feature selection allowing analyzing the resultant model to make inference. It's clear that smoking and drinking are bad habits that has negative impact on sleep efficiency, while regular physical exercise increase it. Ensemble methods didn't performed very well in predicting sleep efficiency on test data: the confidence interval are too large to make accurate predictions. The ridge regression wasn't good either for inference (no features are shrunken) and predictions (MSE was higher respect to random forest and bagging). In the end, Poisson regression model isn't good for prediction of the number of awakenings.

References

[1] Sleep and Health, Division of Sleep Medicine at Harvard Medical School, https://sleep.hms.harvard.edu/education-training/public-education/sleep-and-health-education-program/sleep-health-education-86

[2] Sleep Efficiency Dataset, University of Oxfordshire, https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency

[3] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R (Second Edition), Springer.

[4] Raposo LM, Arruda MB, de Brindeiro RM, Nobre FF. Lopinavir Resistance Classification with Imbalanced Data Using Probabilistic Neural Networks. J Med Syst. 2016 Mar;40(3):69. doi: 10.1007/s10916-015-0428-7. Epub 2016 Jan 6. PMID: 26733278.

[5] Yates, Roy D; Goodman, David J (2014). Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers (2nd ed.). Hoboken, NJ: Wiley. ISBN 978-0-471-45259-1.