Regression Analysis of Sleep Style Dataset

Spring 2024 Assignment
DATA 301 – Introduction to Data Science

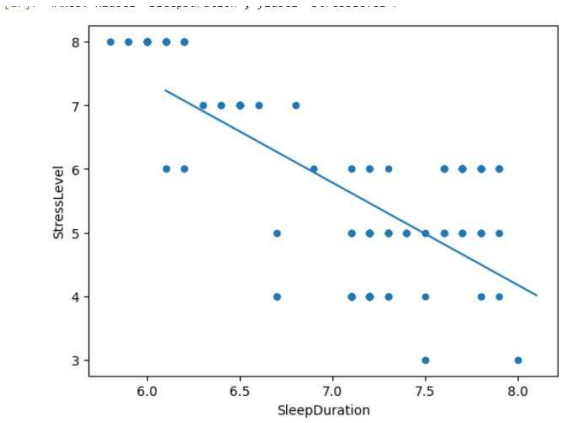
Written by Tara Rajagopalan

1. Summary

This document describes the exploratory data analysis of the Sleep Health Lifestyle data set found on Kaggle data website. The results from two different models, namely Linear Regression and K-Nearest Nearest Neighbors are compared side-by-side. Further, K-Fold Cross Validation and hyper-parameter tuning were leveraged to accurately predict Stress Level based on Sleep Quality and Sleep Duration. Finally, K-Nearest neighbors classification was used to determine the BMI category for a person based on their Sleep Quality and Sleep Duration

2. Linear Regression

The relationship between duration of sleep and stress levels was studied using Linear Regression.



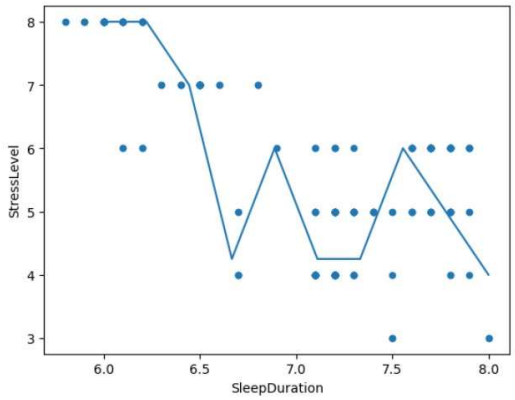
Above, a linear regression model that predicts a person's Stress Level based on their Duration of Sleep (Number of Hours of Sleep Received) was built, along with a scatter plot and fitted line as a visual representation.

As shown in the model above, the fitted line is sloping downwards. The stress levels tend to be higher when someone has had less hours of sleep, while the stress levels tend to be lower when someone has had more hours of sleep. Thus, this visual shows that there is a relationship between Sleep Duration and Stress Levels.

3. K-Nearest Neighbors Regression Function

A k-nearest neighbors regression model with 4 neighbors was built to predict Stress Levels based on Sleep Duration in the span between 6 – 8 hours.

[31]: <Axes: xlabel='SleepDuration', ylabel='StressLevel'>



The visual demonstrates the relationship between Sleep Duration and Stress Levels. However, this visual is different from the linear regression model as this visual displays the resulting Stress Levels of individuals in a piecewise manner because the goal is to predict the Stress Levels using the idea of nearest neighbors.

For example, the Stress Levels for those whose sleep duration is between 6.0 – 6.5 hours is 8 as those data points in the similar sleep duration period have the same 4 nearest neighbors when predicting stress levels. Another example is that those with a sleep duration in the range between 7.0 – 7.5 hours all have a predicted Stress Level of approximately 4.5, as demonstrated by the fitted line. Data points near the 7.0 – 7.5 hour Sleep Duration range have the same nearest neighbors, which is why the Stress Levels are predicted to be approximately 4.5 for all those data points, making the stress level prediction constant for this range of sleep duration periods.

3.1 Comparing the Results of Stress Levels from Sleep Duration: Linear Regression Vs. K Nearest Neighbors

The K Nearest Neighbors model to predict Stress Levels could potentially be a more accurate depiction, compared to the Linear Regression Model.

The linear Regression Model does a great job at conveying the general idea that the more hours a person spend sleeping, the lower their stress levels will be. The line shows a clear negative correlation between Sleep Duration and Stress Levels.

However, the fitted line in the Linear Regression model fails to accurately show that data points in a similar Sleep Duration range have similar Stress Levels. While the individual data points in the linear model all lie next to each other, the fitted line does not group them together appropriately for analysis. The K Nearest Neighbors model is more effective for this as it groups the data in a piecewise manner based on similarity of Stress Levels and Sleep Duration using K Nearest Neighbors. From the K Nearest Neighbors model, conveys that those who sleep a

certain number of hours will have a very similar Stress Level, which is how this model is more accurate.

3.2 RMSE & MAE for Sleep Duration & Stress Level

The data was fitted by creating a pipeline that included K-Neighbors Regressor with n-neighbors = 4. The Skicit package was then used to calculate the Mean Absolute Error (MAE) and the Root Mean Square error (RMSE).
As a result, RMSE had a value of 0.64. The RMSE of 0.64 means that the K Neighbors model prediction of Stress Levels based on Duration of Sleep is off by an average of 0.64. This implies that the using K Nearest Neighbors model to fit the data on Stress Levels may not be the most accurate.

As shown by the Mean Absolute Error, the average variance between Predicted Stress Levels and Actual Stress Levels is 0.43. The MAE of 0.43 is not terrible, however the K-Neighbors model definitely does not predict Stress Levels perfectly.

4. K-Fold Cross Validation

Now, 2 variables were used to predict Stress Level. The Sleep Duration and Sleep Quality variables from the data set were used to make predictions on Stress Level.

The train data set comprised of the first 200 rows of the original data set. The data was then randomized, and 50% of the random data went into train, and 50% of the random data went into val. The train data set was used to predict val. A pipeline was then made using K Nearest Neighbors value of 4, and the pipeline was fitted using the train data. The *predict* function on the pipeline was used to predict the y values on the validation data sets.

The Root Mean Square error was then computed, and resulted in a value of 0.61. This means that the average difference between the actual Stress Level and the predicted Stress Level is around 0.61. 0.61 is a relatively high RMSE, so this model does not necessarily predict the Stress Level precisely from Sleep Duration and Sleep Duration.

Basically, the idea of cross validation is to use the first half of the training dataset to predict the 2nd half of the data, or use the 2nd half of the training dataset to predict the first half of the training dataset.

In the photograph above, the train data (1st half) was used to predict the validation data (2nd half) and in turn, the validation data(2nd half) was used to predict the train data (1st half). The RMSE for both was obtained. Their RMSE is not the same, as an RMSE of 0.61 was obtained when using train to predict validation, and an RMSE of 0.525 when using validation to predict

train. This means that using validation to predict train for Stress Level is more accurate than using train to predict validation for the Stress Level Variable.

4.1 Cross Validation

Apparently, using half of the data in cross validation to predict Stress Levels does not necessarily give accurate results, so K Fold Cross Validation Method was used alternatively to try and predict Stress Level.

Four folds were made, where 3 sub-samples are used as training data, and 1 sub-sample represents validation, which is the data that will be predicted. The RMSE was computed after using K-Fold Cross Validation to predict Stress Level, which was 0.67. Surprisingly, the RMSE was way higher compared to using cross validation.

However, when using 100 folds instead of just 4 folds, the CV changed from CV = 4 to CV = 100, and as a result, the RMSE had significantly lowered to 0.40. This RMSE was significantly lower than the RMSE found for cross validation.

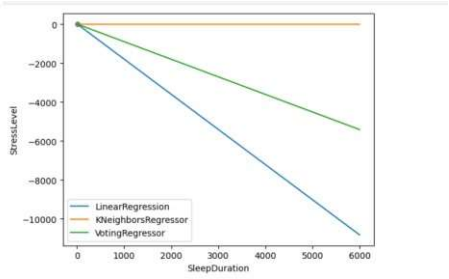
In conclusion, at least 100 Folds are needed for the model to accurately predict Stress Levels based on Sleep Quality and Sleep Duration.

5. Model Selection and Hyperparameter Tuning

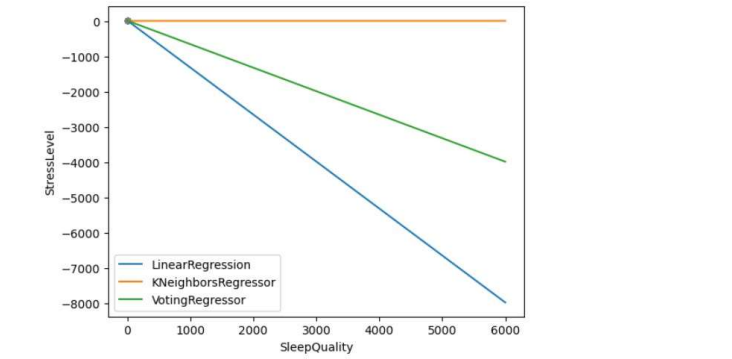
Hyperparameter Tuning was now used to determine the number of nearest neighbors that will minimize the Mean Square Error in order to accurately predict Stress Level from Sleep Quality and Sleep Duration.

A K-Fold Cross Validation model was fitted with 100 folds, as specified by cv = 100. As a result, the number of neighbors requied to get the most accurate prediction of StressLevel is 16 neighbors, as the MSE is minimized at 16 neighbors.

Lastly, a loadings plot with 16 neighbors was made.



The first loading plot above represents the relationship between SleepDuration and StressLevel.



[ ]:

The second loading plot represents the relationship between sleep quality and stress level.

The loading plot displays that as Sleep Quality and Sleep Duration increases, Stress Levels tend to decrease. These are shown from Linear Regression and Voting Regression methods. However, the K Neighbors Regressor Line is completely straight, which does not seem to be an accurate depiction between the variables.

6. K-Nearest Neighbors for Classification

For classification, the BMI category that a person would fall under would be predicted based on their how long they sleep (Sleep Duration) and the quality of their sleep (Sleep Quality).

A 5 neighbors model was fitted with the X variables being Sleep Duration and Sleep Quality, and Y variable being “BMI Category”.

First, the probabilities for each BMI category were predicted using sleep Duration of 5.9 Hours and Sleep Quality of 6. The results show that 80% of these people were overweight, and 20% of these people were normal in weight. According to value counts, the majority of people with these sleep statistics were overweight.

```
]:  BMICategory
    Overweight    3
    Normal        2
```

When using 10 nearest neighbors to predict BMI Category from Sleep Duration and Sleep Quality, 90% of the people were overweight and only 10% of these people were normal in weight.

When utilizing 10 neighbors for prediction, Value Counts showed that 7 of the people were overweight and 3 of the people were of normal weight.

```
[9]: BMICategory
     Overweight    7
     Normal        3
```

6.1 Using 100 Nearest Neighbors

```
]: BMICategory
   Overweight    66
   Normal        33
   Obese          1
```

6.2 Sleep Duration of 5.9 hours and Sleep Quality of 6.

A model was fitted to predict the BMI categories that people fall under when they have a sleep duration of 5.9 hours and a Sleep Quality rating of 6. The findings show that 66% of these people fall under the Obese BMI category.

```
[36]: BMICategory
      Normal           58
      Overweight       33
      Normal Weight     7
      Obese             2
```

6.3 Sleep duration of 8.1hours with Sleep Duration and Sleep Quality of 9

A model was fitted to predict the BMI categories that people fall under when they have a

sleep duration of 8.1 hours and a sleep quality of 9 hours. 100 nearest neighbors were used, as more data was obtained for each of the categories. The findings show that 58% of the people were normal in weight, and 2% of these people were obese. This shows that those who get more sleep are generally normal in weight.

The comparisons when using 100 nearest neighbors show that those who sleep for more hours and have better quality sleep tend to fall under the normal weight BMI category. However, those who sleep for less hours and have poor quality of sleep tend to fall under the Obese BMI category.

Earlier, predictions were done using 5 nearest neighbors and 10 nearest neighbors. However, the analysis using 100 nearest neighbors is a more accurate depiction of how sleep quality, sleep duration, and the BMI category that a person would fall under is related.

6.4 Precision, Recall, and F1 Score

Now, precision and recall were calculated for the three BMI Categories: Normal, Overweight, and Obese.

The BMI Category of Normal has a precision score of 0.799 and a recall score of 1.0. The precision score of 0.799 means that approximately 80% of the observations that were predicted to fall under the Normal BMI category fell under the Normal BMI category correctly. The recall score of 1.0 means that 100% of the observations that actually fell under the Normal BMI category were predicted to be in the Normal BMI Category.

The BMI Category of Obese has a precision score of 1.0 and a recall score of 0.4. The precision score of 1.0 means that approximately 100% of the observations that were predicted to fall under the Obese BMI category fell under the Obese BMI category correctly. The recall score of 0.4 means that 40% of the observations that actually fell under the Obese BMI category were predicted to be in the Obese BMI Category.

The BMI Category of Overweight has a precision score of 0.9 and a recall score of 0.736. The precision score of 0.9 implies that approximately 90% of the observations that were predicted to fall under the Overweight BMI category fell under the Overweight BMI category correctly. The recall score of 0.736 means that 73.6% of the observations that actually fell under the Overweight BMI category were predicted to be in the Overweight BMI Category.

The F1 scores were computed for each of the BMI categories to get a general idea of their accuracy.

As a result, f1 scores were collected for each of the BMI categories. The Normal BMI category has an f1 score of 0.88, and the Overweight BMI category has an f1 score of 0.81.

These scores imply that the predictions for the Normal and Overweight BMI categories are relatively accurate, however, either the precision or recall score for one of these should be slightly higher to ensure more accuracy.

Unfortunately, the f1 score for the Obese category is around 57%, which is pretty low in accuracy compared to Normal and Overweight BMI. This means that either the precision or recall score is significantly low. In order to increase the f1 score, the precision or recall score must be increased.

Hyperparameter tuning was then used to obtain a K value that will lead to the best f1 Score.

As a result, using 3 neighbors will give the highest f1 Score. This means that using 3 neighbors will most accurately predict the BMI category that a person falls under based on the variables Sleep Quality, Sleep Duration, and Stress Level.