The impact of our simple habits on daily stress

Prof. Ying Lin

Nidhee Nishikant Patil, Syracuse University,
NY-13210. [E-mail: npatil07@syr.edu].

Video Link- https://video.syr.edu/media/t/1_lwa4bazb

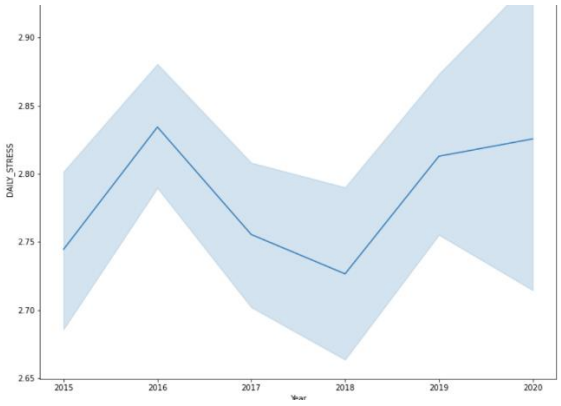
Abstract

Living a healthy lifestyle is as simple as doing things that bring you joy and satisfaction. On one person, that might entail going for a mile-long walk five times per week, consuming fast food only once per week, and spending every other day either virtually or physically with loved ones. Others may define a healthy lifestyle as preparing for and competing in two marathons annually, adhering to a ketogenic diet, and abstaining from all alcoholic beverages. Now a days daily stress in a person's life is increasing due to little or no importance for the healthy lifestyle. Healthy lifestyle may include sufficient sleep hours, meditation, sufficient salary, work-life-balance, vacation and much more. We are going to use linear regression to find the daily stress score based on this factor.

Introduction

In today's world where everything is moving faster than ever, people are trying to race along. The simple habits and interests like old time are getting more and more complex. Food habits are changing, people are relying more on fast foods or pre-meals than fresh cooked foods. Work is becoming important than family, and needs are becoming wants. In this race of life, we are leaving behind our healthy lifestyle. Healthy comprises of both physical and mental health and disturbance in any causes stress. This daily stress may not be because of major reason like illness or big problems but they can be caused by our daily habits.

Insufficient sleep can cause irritation, hormonal imbalance due to food or lack of exercise can cause stress, not taking enough break, vacation or spending time with family may case mental imbalance and many such reasons can contribute towards daily stress. I will be focusing on such factors with some physical parameters of body such as age, gender etc. to find out the most significant factors that affect stress and major to reduce it. I will be using regression to build a model having these factors as our predictors and stress as our dependent variable. Regression will give us the stress score depending upon the variability in these factors. Here, is how people have reported stress over the course of time.



Previous work

Previous works includes various methods, some that include questionnaire, and some include machine learning. Basic research was

conducted by interviewing some of the villagers talking about their habits and impact of them on their mortality rates and their lifestyle. This research was done for experiment that was supposed to be conducted on some villages to promote new sources of goods and services, improve health and nutrition practices, increase participation in community life, increase self-confidence. The research was conducted, and basic trend were realized. (Noreen M. Clark and O. Nyaga Gakuru, 2014)

Another research that was conducted were on people to study the factors of hypertension. Here, more health-related factors such as blood pressure, blood pressure, high cholesterol, pulse were used to determine its impact on hypertension. This method used unsupervised learning to determine their result. They used CNN (Convolutional Neural Network) to build a unsupervised learning model that would find its own learnings and trends could be used in future to decrease hypertension in patients. Previously WHO used logistic regression of feature selection. Saliency maps can reveal if a CNN model classifies objects based on the presence or absence of a target feature or whether it is discovering any latent features connected to the target feature. Thus, researchers can investigate the ideas that a CNN has learned by comparing the changes in the saliency maps owing to changes in the input. From their research they found that Gender, age, health (such as BMI), and comorbidities are all important factors in predicting an individual's propensity. It is also noticed that the possession of luxury goods vs capital for a living affects people differently (Md. MazharulIslam, RittikaShamsuddin, 2021).

In contrast to first research methods, I decided to use the power of Machine learning for the prediction of my result that is the effect of different variables on daily stress. And unlike second experiment/research, I decided to use supervised learning based on data collected of each individual across four years. I also chose more variable that are related to our daily habit rather than technical health parameters such as blood pressure or pulse. I decided to tale into consideration variables that we come across everyday and have full control over. My study is based on daily stress rather than major illness such as hypertension. I decided to do that as these small habits accumulate to be part of major illness and if we learn to control those, we might prevent big diseases.

Data

Information gathered from a Kaggle which was collected in the global study on work-life balance conducted in collaboration with http://www.authentic-happiness.com/ , https://www.360living.co/ , and http://www.guidebienetre.org/

The UN Sustainable Development Goals are supported by this study:

8.4 Increase global resource efficiency in production and consumption, and work to break the link between economic expansion and environmental destruction. 12.8 Make sure that everyone has access to the necessary knowledge and awareness of sustainable development and environmentally friendly lifestyles. 12.8.1 The degree to which curricula, teacher preparation programs, national education policies, and student assessments mainstream global citizenship education and education for sustainable development, including education on climate change.

This data consists of numerical as well as categorical factor. The numerical data is collected by the absolute value and some by rating/score on a scale. The dataset analyzed in this kernel contains10,000+ responses to Authentic-Happiness.com global work-life survey. This online survey includes 23 questions about the way we design our lifestyle and achieve work-life balance. These questions were asked to various individuals for four years. the data contains such variables on the study and prediction on which an individual can take effective measures to reduce the daily stress in his/her life. The data include factors like sleep hours, daily meditation, passion/hobby, vacations taken, gender, age, social network, achievements etc. Age is a categorical variable with ranges like less than 20, 21 to 35, so on and so forth. There is another categorical variable that is gender. There is total of 23 columns and 12755 rows. The data also contains a timestamp column that records the time and date the date was recorded.

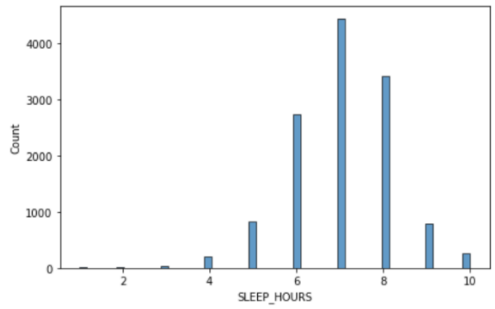
Data analysis and modelling

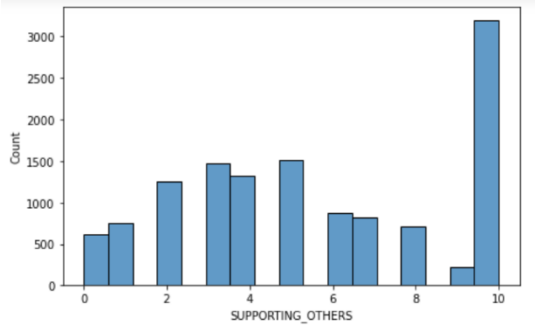
To start with, I pre-processed the data to encode the categorical variables to numerical ones. I did to include them in my machine model building process. For this purpose, I used one hot encoding method provided by scikit learn library. After converting my categorical variables to numerical variables, it looks like below.

AGE_Adults	AGE_Children/teenagers	AGE_Middle-age adults	AGE_Senior Citizens
0	0	1	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1

After the conversion the next challenge was to find out the most significant factors. For this I used three methods that could help me decrease the dimensionality of the model. The first approach was to do some exploratory analysis on the data. The descriptive statistics revealed some insights like:

a) Majority of the people have support of their closed ones, and they support others too
b) Most of the people tend o get sleep between six to eight hours.
c) More than half of the people have less or no time for their passion or hobbies





After analyzing the descriptive analysis, the second approach I used for identifying impactful factors is the correlation matrix. Correlation among the variable is used to find out the relation between our dependent and independent variables that would filter out some variables.

	FRUITS_VEGGIES	DAILY_STRESS
FRUITS_VEGGIES	1.000000	-0.095132
DAILY_STRESS	-0.095132	1.000000
PLACES_VISITED	0.248520	-0.131707
CORE_CIRCLE	0.153606	-0.115973
SUPPORTING_OTHERS	0.207907	-0.035373
SOCIAL_NETWORK	0.105804	0.012720
ACHIEVEMENT	0.166643	-0.120786
DONATION	0.200787	-0.038291
BMI_RANGE	-0.091937	0.084938
TODO_COMPLETED	0.230350	-0.166975

Here, we can see that completion of to-do tasks or the tasks they have planned contributes to the stress of a person. It's the negative correlation as more the tasks complete less the stress. By doing so we go few significant variables like with correlation as:

-0.131707: PLACES_VISITED
-0.120786: ACHIEVEMENT
#-0.166975: TODO_COMPLETED
#-0.142187: FLOW
#-0.152862: SLEEP_HOURS
#-0.135016: LIVE_VISION
#0.192353: LOST_VACATION
#0.309264: DAILY_SHOUTING
#-0.144872: SUFFICIENT_INCOME
#-0.161858: TIME_FOR_PASSION
#-0.213672: DAILY_MEDITATION

The last approach to study the importance of the variable is with the help of P-values. The lower the P-values the more is its significance to reject the null hypothesis. There I built a linear regression which in detail I will be discussing about in later section. The linear regression helped me obtain the p-values of the variable and extract the significant ones. The p-values look something like this:

	coef	std err	t	P>|t|
FRUITS_VEGGIES	-0.0115	0.008	-1.387	0.165
PLACES_VISITED	-0.0113	0.004	-3.042	0.002
CORE_CIRCLE	-0.0274	0.004	-6.334	0.000
SUPPORTING_OTHERS	0.0200	0.004	4.845	0.000
SOCIAL_NETWORK	0.0272	0.004	6.801	0.000
ACHIEVEMENT	-0.0145	0.005	-2.980	0.003
DONATION	0.0066	0.007	1.000	0.317
BMI_RANGE	0.1004	0.023	4.332	0.000
TODO_COMPLETED	-0.0267	0.005	-5.621	0.000
FLOW	-0.0323	0.006	-5.732	0.000
DAILY_STEPS	0.0073	0.004	1.792	0.073
LIVE_VISION	-0.0154	0.004	-4.143	0.000

As we can see to-do, flow and many others have very less P-values. There with the help of correlation matrix and P-values I reduced the variables from 23 to 12 for the final model.

Linear regression

I chose this particular method because regression is used to predict the numerical outcome based on other factors. Here, I have score of daily stress that goes from one to five. Number of people choose the stress level that is caused to them on the daily basis and score their other habits a well. To determine the daily stress according to the other factors I decided to use linear regression. In order to model the relationship between two variables, linear regression fits a linear equation to the observed data. The first variable is regarded as an explanatory variable, whereas the second is regarded as a dependent variable. For instance, a modeler might use a linear regression model to compare people's weights to their heights. A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

Evaluation and conclusion

To evaluate my final model, I used the value of adjusted R^2 which is one of the performance metrics of evaluation. The

adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. Here is the output of both my models. The first one contains all the variables, the second one contains the significant factors extracted by the approaches above.

OLS Regression Results

Dep. Variable:	DAILY_STRESS	R-squared:	0.207
Model:	OLS	Adj. R-squared:	0.206
Method:	Least Squares	F-statistic:	144.5
Date:	Sun, 03 Jul 2022	Prob (F-statistic):	0.00
Time:	00:12:04	Log-Likelihood:	-20659.
No. Observations:	12755	AIC:	4.137e+04
Df Residuals:	12731	BIC:	4.155e+04
Df Model:	23		
Covariance Type:	nonrobust		

We can see sudden shift in the adjusted R^2. The R square increases to a significant amount.

OLS Regression Results

Dep. Variable:	DAILY_STRESS	R-squared (uncentered):	0.812
Model:	OLS	Adj. R-squared (uncentered):	0.812
Method:	Least Squares	F-statistic:	3670.
Date:	Sun, 03 Jul 2022	Prob (F-statistic):	0.00
Time:	00:06:04	Log-Likelihood:	-17532.
No. Observations:	10204	AIC:	3.509e+04
Df Residuals:	10192	BIC:	3.517e+04
Df Model:	12		
Covariance Type:	nonrobust		

The adjusted R-square values increases from 0.2 to 0.8.

The second method od evaluation Root Mean Square Error (RMSE) value. The difference between values (sample or population values) predicted by a model or estimator and the values observed is typically measured using the root-mean-square deviation (RMSD) or root-mean-square error (RMSE).

```
#calculate RMSE
sqrt(mean_squared_error(testing_dependent, dependent_pred))
```

```
1.3546491601422996
```

As we can observe that the RMSE value is low indicating that the model is able to fit the data correctly.

In conclusion, there are number of factors that affect stress in a person's life, the most significant are meditation, sleep hours, vacation and the once mentioned above. If we over all improve these factors and the flow of our day, we can reduce the daily stress. This model gives a peek into a person's daily life habits and helps them suggest the method to reduce the stress caused on daily basis. This eventually ensures healthy mind and lead to happier life.

References

Md. Mazharul Islam, Rittika Shamsuddin, Machine learning to promote health management through lifestyle changes for hypertension patients, Volume 12, 2021, 100090, ISSN 2590-0056, https://doi.org/10.1016/j.array.2021.100090. (https://www.sciencedirect.com/science/article/pii/S2590005621000370)

Clark, N. M., & Gakuru, O. N. (2014). The Effect on Health and Self-Confidence of Participation in Collaborative Learning Activities. *Health Education & Behavior*, *41*(5), 476–484. http://www.jstor.org/stable/45088175

Ghassemi, M., Mohamed, S. Machine learning and health need better values. *npj Digit. Med.* **5,** 51 (2022). https://doi.org/10.1038/s41746-022-00595-9

Other reference:

https://datatofish.com/correlation-matrix-pandas/

https://stackabuse.com/linear-regression-in-python-with-scikit-learn/

https://towardsdatascience.com/how-to-split-a-dataset-into-training-and-testing-sets-b146b1649830

https://www.w3schools.com/python/python_ml_multiple_regression.asp

https://www.statsmodels.org/dev/generated/statsmodels.tools.tools.add_constant.html

https://www.programiz.com/python-programming/methods/list/remove

https://www.freecodecamp.org/

https://scikit-learn.org/stable/modules/classes.html#regression-metrics

https://www.statology.org/rmse-python/

https://seaborn.pydata.org/api.html

https://pandas.pydata.org/docs/reference/index.html