The Analysis of Happiness Indicators Based On RPTBXS Framework

Haotian Xiao

Hongyi Honor College, Wuhan University

Abstract

This article studies the impact of income and using technology and social media on the happiness level by the proposed RPTBXS framework with the data of the General Social Survey(GSS). It can be concluded that the increase of income will not increase happiness when it is high; thas a significant impact on the happiness level, especially when income is at a high level; The use of social media will also significantly benefit those happinese level are at low or high. The result suggests that government establish a sound psychological guidance system to ensure smooth channels for psychological counselling and high group recognition.

Keywords: Happy, PDS, ordered logistic regression, instrumental variable, XGBOOST

1. Introduction

All human endeavors are aimed at achieving happiness. Intuitively, there are many factors that influence an individual's level of happiness. Up to date, researchers have identified a variety of factors that contribute to an individual's overall sense of well-being. For wealth, Silvio et al(2013) propose that poverty –as well as other adverse situations– has an undermining effect on happiness, Bao and Le(2021) suggest that daily wealth returns positively affect the changes in happiness sentiment. As for equality, Alberto et al(2004) prove that individuals have a lower tendency to report themselves happy when inequality is high. For inflation and unemployment, Tella et al(2001) demonstrate that reported well-being is strongly correlated with inflation and unemployment by a panel analysis of nations. For technology, Graham and Nikolova(2013) find that technology access is positive but with diminishing marginal returns for general, with signs of increased stress and anger among cohorts for whom access to the technologies is new. For education, Hartog and Oosterbeek(1998) find that the group with a non-vocational intermediate level education score highest on happiness. For air pollution, Zhang et al(2022) propose that the well-being and happiness level of people who are male, young , less educated, urban resident and live in eastern or western China are more sensitive to the pollution.

When it comes to the research methods, other than the regression methods in traditional econometrics, many literature use machine learning to study this subject. In Gabriele(2022), machine learning techniques were used to identify the correlates of quality of life. In Gorden et al(2021),

machine learning shows promise in distinguishing bipolar from unipolar disorders.

Although a lot of studies have explored the indicators of happiness level by various approaches, they still need to be improved in the following three aspects.

Initially, the method to select the control variables is extremely important. Control variables are used to mitigate interference from confounders on causal effect estimates in multiple regression analysis. But it usually has no structural explanation. Even valid control variables are often associated with other unobserved (or unobservable) factors, which make their marginal effects unexplained from a causal inference perspective (Westreich and Greenland, 2013; Keele et al, 2020).

Additionally, how to select and test the validity of instrumental variables are always the focus of econometricians, especially in the problem when need to use logistic or probit regression in the second stage. How to select iv and effectively and correctly test the validity of the iv matters.

Ultimately, most machine learning models that have been applied to the predict or optimize are indeed a black box. That is, they are actually hard to be explained to humans. Generally speaking, the lack of interpretability in machine learning will unfortunately undermine confidence in the built predictive models and outcomes. In response to the challenges inherited in classical machine learning, it is a must to put efforts into explainable machine learning approaches through creating intelligible explanations of the model prediction mechanism and the variable importance.

To fill in the research gap and further explore the indicators of happiness, this article focuses on data GSS and proposes a new analysis framework RPTBXS. When it comes to data cleaning, random forest in mice is adopted to fill in the NAs with features whose NAs exceed percentage 30 deleted. As for feature engineering, post-double-selection is used to select the control variables. Logit regression, Ordered logit regression, TSLS are compared and a R package "instruments" is utilized to test the validity of iv. By and large, an effective machine learning model BO-XGBOOST-SHAP is established to further explore the correlation and support the conclusions previously obtained.

The rest of this paper is organized as follows. Section 2 and 3 introduces the process of data cleaning and the result of exploratory analysis. Section 4 and 5 performs the relationship between happiness and technology, social media use. Section 6 concludes the result of study and Section 7 discusses the drawbacks and limitations.

2. Data Preprocessing

2.1. Data background

The General Social Survey (GSS) is NORC's longest-running project and one of its most influential. GSS data are frequently used in newspaper, magazine, and journal articles and by legislators, policymakers, and educators. GSS topics include national spending priorities, marijuana use, crime, intergroup relations, social and economic life, lifestyle, civil liberties, subjective well-being, and con-

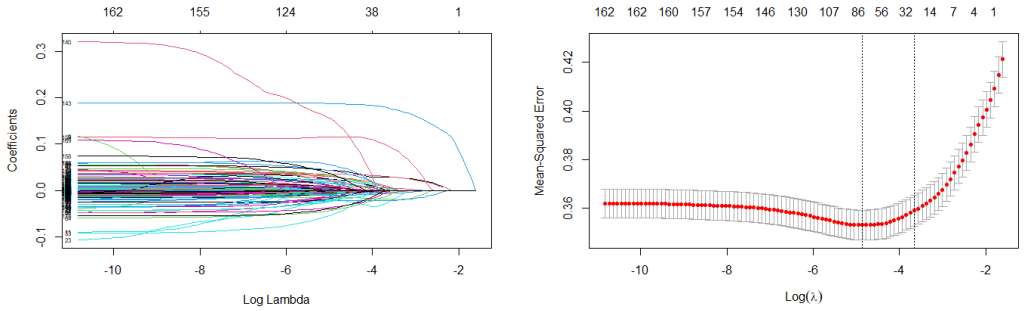
fidence in institutions. Since 1988, the GSS has also collected data on sexual behavior, including number of sex partners, frequency of intercourse, and extramarital relationships.

2.2. Missing Values Handling

Firstly, the GSS data for the required years are imported and features with more than percentage 30 missing values are selected for deletion due to their excessive missingness and lack of explanatory power and credibility. Secondly, the missing values are filled in using the random forest method in the mice package to obtain a complete sample with no missing values. The random forest method was chosen because many of the features are integer values and using other methods such as lasso padding would fill in fractional numbers and affect subsequent processing.

2.3. Feature Engineering

The post-double-selection(Belloni (2014)) is adopted to select the feature. Lasso regressions are conducted with happiness as the dependent variable, the indicator and other features explored as regressors, and indicator as the dependent variable and the remaining features as regressors. The Lasso estimator results in a sparse solution, i.e. many of the regressors have coefficients of 0. On this basis, the control variables selected in the subsequent operations include those selected in the two Lasso regressions.



(a) LASSO

(b) FIND LAMBDA

Figure 1: LASSO

3. Exploratory Analysis

3.1. Happiness and Income

The three-year GSS dataset is merged, features with more than percentage 30 missing values are removed, and the remaining data are populated with random forest. Then, lasso regression are used to find variables associated with HAPPY and INCOME was found to be associated with HAPPY and there are no features associated with technology or media. An exploratory analysis of

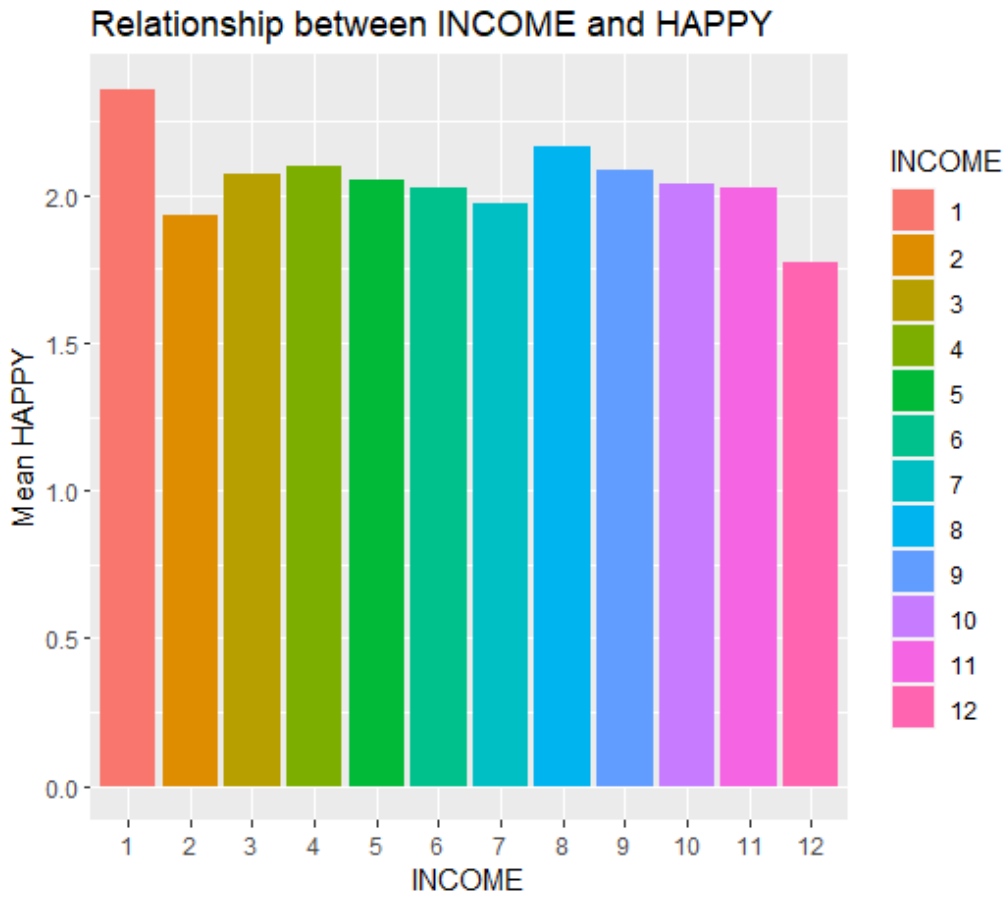
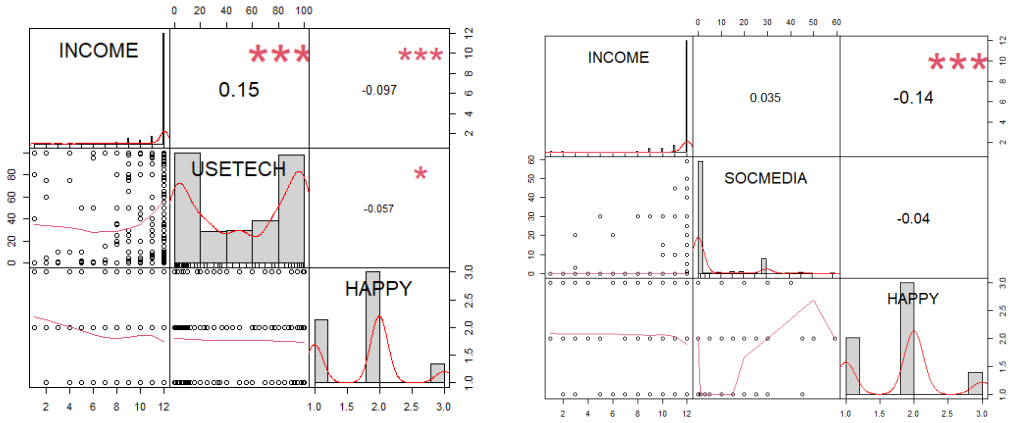


Figure 2: HAPPY and INCOME

the relationship between INCOME and HAPPY is shown in Figure 1. It is found that the sample with an INCOME of 1 is very small and could be treated as noise, on the basis that at lower incomes, happiness levels increased as income increased, but at higher incomes, income is negatively correlated with happiness levels. It is hypothesized that at a better standard of living, the correlation between happiness levels and income is weak because at this time other features may be have larger impact on happiness.

3.2. Happiness, Income and Technology, Social Media

To analyse the relationship between happiness, income and technology, the GSS2014 dataset is used, remove the sample with the missing feature USETECH and the features with more than percentage 30 missing values, and fill the remaining data with random forest. For the social media analysis, the GSS2016 dataset is used, with the missing feature INTWKENM removed and other operations the same. Correlation plot is shown in fig2.



(a) HAPPY, INCOME and TECH

(b) HAPPY, INCOME and MEDIA

Figure 3: EXPLOTARY ANALYSIS

4. Technology and Happy

4.1. Preprocessing

By the PDS process, the control variables selected are WORKDIFF, USEDUP, COWRKHLP, INTETHN, HEFINFO, JOBSECOK, RACLIVE. So the base model is :

$$HAPPY_i = \alpha + \beta X_i + \gamma USETECH_i + \theta R_i + \varepsilon$$

where i represents the individual, X_i represents the control variables, R_i is obtained in the model to represent regional fixed effect.

4.2. Logit Regression

First divide the happiness level into low, middle and high represent 1,2,3, and do logit regression respectively. From the regression result the USETECH is significant when happiness level is at high. However, the regression can only roughly explain the relationship between happiness and technology and needs more research. See fig4.

4.3. Ordered Logit Regression

Then do the ordered logit regression which seems more suitable for the happy has three different level. The ordered logit regression successfully pass the parallel test and likelihood ratio test, however unfortunately from chi-test the USETECH is not significant. See fig5.

4.4. TSLS

In order to further study this problem, turn to TSLS and choose IV 'LEARNNEW'. From intuition, at this technological era, if people learn something new at work, it is likely to be correlated

	Low Happy	Middle Happy	High Happy
USETECH (logit)	0.00030024	2.8053e-04	-0.00058077**
	(0.375081)	(0.44157)	(0.0068760)
REGION FE	√	√	√
USETECH (2SLS)	0.014710*	-0.009913	-0.005887
	(0.057889)	(0.1567)	(0.597006)
Model AIC	1482.6	1667.6	747.27
IV Valid	√	√	√
SOCMEDIA (logit)	5.1022e-05	-0.00036171	0.00031069
	(0.96978)	(0.8116)	(0.7405)
REGION FE	√	√	√
SOCMEDIA (2SLS)	-0.12412*	-0.001513	0.26512**
	(0.0425)	(0.978)	(0.005432)
Model AIC	1045.6	1265.1	580.14
IV Valid	√	√	√

The effect of technology & media use on happiness (logit & 2SLS)

Figure 4: LOGIT and 2SLS

with technology, or, he uses technology to learn it. Then I test the validity of this IV by package 'instruments' in R and do IV-LOGIT regression. From the result, the IV is valid, but the USETECH is a little significant when happiness level is low, and not significant in other cases. See fig4.

4.5. INCOME Impact

Add four interaction features, respectively, INC1,INC2,INC3,INC4, which is calculated by:

$$INCOME1 = \begin{cases} 1 & INCOME = 1, 2, 3 \\ 0 & ELSE \end{cases}$$

$$\dots$$

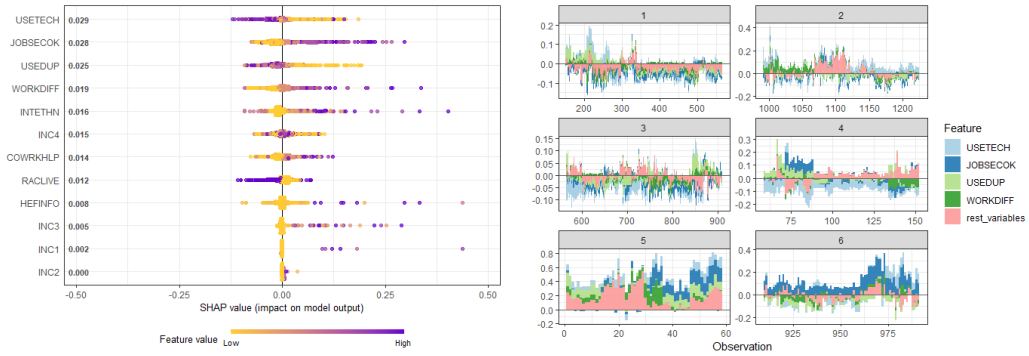
$$INCOME4 = \begin{cases} 1 \; INCOME = 10, 11, 12 \\ 0 \; ELSE \end{cases}$$

$$INCi = INCOMEi \times USETECH$$

Use this four features to replace USETECH and do logit regression with whether at high happiness level. The result shows that high income has a significant impact on the happiness level, see fig6.

4.6. Machine Learning

Finally, to fully address this problem, I establish a BO-XGBOOST-SHAP explainable machine learning model. The regression model is xgboost, and use Bayesian Optimization to find the best hyperparameters and use SHAP to enhance the explainability of the model and show the shapley value, that is, the importance of each feature. The result of machine learning shows that USETECH indeed has significant impact on the happiness level, and high income group influences the result. If income is lower than 10, it has little impact on the happiness degree. See fig7.



(a) SHAP PLOT (b) FORCE PLOT

Figure 5: MACHINE LEARNING RESULT

5. Media and Happiness

5.1. Analysis

5.1.1. Preprocessing

By the PDS process, the control variables selected are HARASS5, TUMBLR, MARCOHAB, PARTNRS5, SATFIN. So the base model is :

$$HAPPY_i = \alpha + \beta X_i + \gamma SOCMEDIA_i + \theta R_i + \varepsilon$$

where i represents the individual, X_i represents the control variables, R_i is obtained in the model to represent regional fixed effect.

5.1.2. Logit Regression

Initially divide the happiness level into low, middle and high represent 1,2,3, and do logit regression respectively. From the regression result the SOCMEDIA is not significant at any level. So apparently it needs more research. See fig4.

	COEFFICIENT	Pr (>chi)	PARALLEL TEST	LIKELIHOOD RATIO TEST
TECH	-0.00279*	(0.0603678)	0.44	0
MEDIA	-0.0005975	(0.92311)	0.47	0.0002999499

Ologit Model

Figure 6: OLOGIT Model

	COEFFICIENT	Std	T-Value	Pr (> | t |)
INC1	0.00047494	0.00140455	0.3381	0.7353126
INC2	-0.00076689	0.00138795	-0.5525	0.5806855
INC3	0.00104373	0.00081805	1.2759	0.2022434
INC4	-0.00062347	0.00021567	-2.8908	0.0039116**

Different Income Impact

Figure 7: OLOGIT Model

5.1.3. Ordered Logit Regression

In addition do the ordered logit regression which seems more suitable for the happy has three different level. The ordered logit regression successfully pass the parallel test and likelihood ratio test, however unfortunately from chi-test the SOCMEDIA is not significant. See fig5.

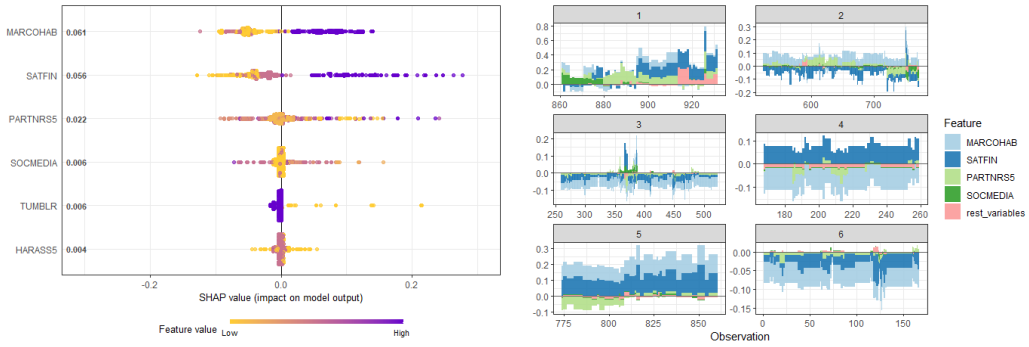
5.1.4. TSLS

Aimed at solving the problem, I turn to TSLS and create a special instrumental variable, which is the sum of all social software features except for PINTERST. Intuitively, it can represent how much one utilize the social media so tightly correlate with SOCMEDIA. The result shows the validity of IV and SOCMEDIA is significant when happiness level is low and high. See fig4.

5.1.5. Machine Learning

Ultimately use the same explainable machine learning model to explore the correlation between social media and happiness. From the result it shows that social media has impact on the level of

happiness but not very significant. See fig 8.



(a) SHAP PLOT (b) FORCE PLOT

Figure 8: MACHINE LEARNING RESULT

5.2. Mechanism Policy Implications

Research has shown that when happiness levels are low, i.e. when subjects are dissatisfied with the status quo, they choose to place their emotions on external forces according to the empathy effect. The use of technology is well suited to this psychology. Interdependence theory suggests that people generally seek out interpersonal interactions that offer the greatest reward value at the least cost. Social media, as a technology platform that is less accessible and more attainable, and which satisfies the two most basic human needs of social affiliation and stimulation, is more likely to be chosen by subjects as an emotional support. In a time of high stress and rising anxiety, it is important for the government to use the mass media as a guide to create a positive and healthy trend in public opinion and to improve the general well-being of the population.

In particular, technology has a more significant effect on subjects' happiness when their happiness level is low and their wealth level is high. This is due to their higher psychological disparity and the individual conditions for acquiring more advanced technology. A high level of psychological disparity and a high level of social resource appropriation can increase the social vulnerability of this group and, accordingly, government attention to this group should be appropriately increased.

When happiness levels are high, according to the ABC (Attitude, Behaviour, Cognition) theory in psychology, attitudes are an acquired disposition and set the stage for behavioural change in subjects. Groups with high levels of well-being are more self-aware and less susceptible to negative external influences; they have higher levels of inward satisfaction and a more developed attitude system, and increased frequency of interaction with the outside world positively increases their sense of entitlement, thus creating positive feedback. In response to this positive feedback, the government does not need to intervene much. From another perspective, social media can only have a positive effect on the well-being of groups in society once they have generally reached a high threshold.

To sum up, the government should do a good job in guiding public opinion and creating a positive and clean media trend; continue to improve social security work to enhance people's happiness on an individual basis; and establish a sound psychological guidance system to ensure smooth channels for psychological counselling and high group recognition.

6. Conclusion

Generally, the result of this article corresponds to the intuition. As for income, the marginal utility of it is decreasing when it gets higher. When it comes to technology, the logit model, ordered logit model and machine learning show that it has a significant impact on the happiness level, especially when income is at a high level. For the social media, the TSLS shows that it has a signifanct influence on happiness level, especially when happy is very low or high. The highlight of this article is that both instrument variables constructed in the model pass the IV validity test, which may contribute to research in the future.

7. Some Exploration

7.1. Drawback and Limitations

When analysing the impression factor of happiness, I came across these problems:

- Dealing With Missing Values

  Almost every feature in the GSS data has more than percentage 30 missing values, so it has bothered me for a long time how to deal with these missing values. It doesn't seem reasonable to remove all the missing values, so I choose to use the mice package and compare the effects of various methods of padding and finally choose random forest padding. Whether filling in the NA has a bad influnce on the result remains vague.

- The Selection of Control Variables

  The choice of control variables has always been a difficult issue in econometric research. I have read a lot of literature and there are some differences in the control variables they choose. Because I have done many projects on machine learning feature engineering, I have tried to use lasso regression to filter variables, and I have also discovered the PDS control variable selection method by reading the literature. However, whether this method is suitable for this problem or not is to be studied in depth.

- Dealing with Insignificant Result

  When running regressions, unfortunately, most of my regressions are not significant. Having been unable to get significant result through familiar econometric knowledge, reluctantly, I use

machine learning to solve the problem. It is also partly because I have not found a way to use ordered logit and probit regression models in two-stage least squares.

7.2. The Definition of Technology and Social Media

In this article, the "percentage of time spent using electronics" is used to represent technology, and the instrumental variable for technology in 2SLS, "learning something new on the job", is set because it is highly likely that the new thing learned on the job is new technology, and most likely It is also likely to be the use of technology to learn something new. The use of "time spent using the internet on weekends" was used to represent social media use, as social media use on weekdays is likely to be for work reasons and does not have strong explanatory power, and the instrumental variable for social media in the 2SLS was set, i.e. whether or not major online social software and platforms are used.

7.3. Causality

This econometric methods this articile adopt address potential endogeneity, thus I can make the casual inference. IV is used to represent the technology and social media in 2SLS. For technology, "LEARNNEW" is tightly correlated with technology because when people learn something new at work, he is learning technology or he is likely to use technology to learn it. The exogeneity of this iv can be proved by iv test in the code. For social media, the sum of using various social network platfroms can represent social media use, because when people use the social media, he is likely to use those platforms. The exogenity of this iv can also be proved by iv test in the code.

References

Alberto Alesina, Rafael Di Tella, Robert MacCulloch, Inequality and happiness: are Europeans and Americans different? Journal of Public Economics, Volume 88, Issues 9–10, 2004.

Silvio Borrero, Ana Bolena Escobar, Aura María Cortés, Luis Carlos Maya, Poor and distressed, but happy: situational and cultural moderators of the relationship between wealth and happiness, Estudios Gerenciales, Volume 29, Issue 126, 2013.

A. Alesina, R. Di Tella, and R. MacCulloch. Inequality and happiness: are europeans and americans different? Journal of Public Economics, 88(9-10):2009–2042, 2004.

R. Di Tella, R.J. MacCulloch, and A.J. Oswald. Preferences over inflation and unemployment: Evidence from surveys of happiness. American economic review, 91(1):335–341, 2001.

C. Graham and M. Nikolova. Does access to information technology make people happier? insights from well-being surveys from around the world. The Journal of Socio-Economics, 44:126–139, 2013.

J. Hartog and H. Oosterbeek. Health, wealth and happiness: Why pursue a higher education? Economics of Education Review, 17(3):245–256, 1998.

Guanglai Zhang, Yayun Ren, Yanni Yu, Liguo Zhang, The impact of air pollution on individual subjective well-being: Evidence from China, Journal of Cleaner Production, Volume 336, 2022.

Gabriele Prati, Correlates of quality of life, happiness and life satisfaction among European adults older than 50 years: A machine-learning approach, Archives of Gerontology and Geriatrics, Volume 103, 2022.

Gordon Parker, Michael J. Spoelma, Gabriela Tavella, Martin Alda, Tomas Hajek, David L. Dunner, Claire O'Donovan, Janusz K. Rybakowski, Joseph F. Goldberg, Adam Bayes, Verinder Sharma, Philip Boyce, Vijaya Manicavasagar, Differentiating mania/hypomania from happiness using a machine learning analytic approach, Journal of Affective Disorders, Volume 281, 2021.

Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol. 2013 Feb 15;177(4):292-8. doi: 10.1093/aje/kws412. Epub 2013 Jan 30. PMID: 23371353; PMCID: PMC3626058.

Page, L. C., Lenard, M. A., Keele, L. (2020). The Design of Clustered Observational Studies in Education. AERA Open, 6(3). https://doi.org/10.1177/2332858420954401.

Jeremy Petch, Shuang Di, Walter Nelson, Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology, Canadian Journal of Cardiology, Volume 38, Issue 2, 2022.

Yue Pan, Limao Zhang, Zhenzhen Yan, May O. Lwin, Miroslaw J. Skibniewski, Discovering optimal strategies for mitigating COVID-19 spread using machine learning: Experience from Asia, Sustainable Cities and Society, Volume 75, 2021.

Yuxuan Shen, Yue Pan, BIM-supported automatic energy performance analysis for green building design using explainable machine learning and multi-objective optimization, Applied Energy, Volume 333, 2023.