# Applying Machine Learning to Predict Happiness: A case study of 20 Countries

[1]Yu Tan, [2] Charuk Singhapreecha, [3] Woraphon Yamaka
[1] [2] [3] Faculty of Economic ,Chiang Mai University,Thailand

***Abstract— Happiness is the current important research issues in psychology and social sciences, which is affecting people's daily lifestyle, work habits and thinking patterns, it also provides guidance for government policy making. However, in the current analysis of happiness, there are many challenges in variable selection and prediction. Due to the large personal differences and differences in determinants of the experience of happiness, this undoubtedly makes the modeling of happiness more difficult. Based on reviewing several academic literatures, this paper uses questionnaire data from 20 countries in World Values Survey Database, then uses machine learning methods to compare traditional regression approach with machine learning regression approach to predict happiness. Finally, some determinants variables were found. To a certain extent, the elastic net method applied in model was successfully used to predict happiness, social factors, economic factors and personal factors affect the modeling and prediction of happiness in varying degrees, which brought new opportunities for the development of related theories and practices at the study of happiness.***

**Index Terms—*Elastic net, Happiness, LASSO regression, Machine learning, Prediction, Ridge regression.***

## I. INTRODUCTION

Happiness is a concept of such fundamental importance that it has preoccupied philosophers and religions for millennia. Happiness, and how it can be maximized, has long been of interest to economists, too. Utility entered economic analysis as a close synonym of happiness; although the two concepts later deviated from one another, in recent decades economists have developed a renewed interest in happiness and ways it can directly be measured. It is conceivable that happiness is a key goal for most people; Layard's research pointed out in 2011 that this may be the ultimate goal for most people to pursue their lives [1]. Therefore, if the source of happiness is understood and the cause of happiness is found, there is no doubt that the whole society will benefit. Policy makers will consider how to help people improve their present or future happiness. In other words, policy makers are interested in which factors exert a positive influence, and it is worth supporting them. Intentionally, it is very considerable to find appropriate variables to predict happiness and set up a suitable model to predict happiness. This paper selected data of 20 countries from World Values Survey Database. Considering that the existence of individual differences may cause huge errors in prediction, therefore, the method of machine learning is introduced in this study. In the past literature, few mechanical learning methods have been used. This article will be divided into five parts. The first part is introduction, the second part is literature review, the third part is research methodology, the fourth part is empirical analysis results, and the last part is discussion.

## II. LITERATURE REVIEW

Subjective well-being (SWB) or what we called happiness refers to how people evaluate their lives and includes variables such as life satisfaction and education. Since the 1980s, there has been an extraordinary increase in research on well-being, with most researchers concur that feelings of heppiness comprise of a cognitive-evaluative factor (life satisfaction) and an affective factor (happy)[2], [3], [4].

More specifically, satisfaction is more of a cognitive rating that is outstandingly dependent on social comparisons with other major reference groups as well as the individual's desires, expectations, and hopes. In contrast, happiness is developed as an emotional state generated by positive and negative events and experiences in the life of an individual. Although there is some empirical correlation, in varying degrees, between happiness and life satisfaction, they nevertheless deviate[5],[6].

Along with other social scientists, economists have begun to study the patterns in subjective well-being data. Some of the main economics references include (1974, 1995)[7],[8], Clark and Oswald (1994)[9], Ng (1996, 1997)[10],[11], Kahneman, Wakker and Sarin(1997)[12], Winkelmann and Winkelmann (1998)[13], Di Tella and MacCulloch (1999)[14], Frey and Stutzer (2000)[15], Di Tella et al (2001, 2003)[16],[17], Blanchflower and Oswald (2004)[18], Helliwell and Putnam (2004)[19], and Frijters, Haisken-DeNew and Shields (2004)[20],[21]. In fact, psychologists and sociologists were working on such data sets before most economists paid much attention.

Occurring recent findings from such statistical happiness research include the following:

1. For individuals, money can buy for some levels of happiness to some extent. But it is useful to keep this in standpoint. Very generally, for the typical individual, a doubling of salary makes a lot less difference than life events like a good family relationship.
2. For a country, the situation is different. As far as western developed countries are concerned, some of the existing research subjects are all western developed countries. The conclusion drawn in these studies is that as people become more affluent, they do not seem to be happier. There are few related studies in developing countries.
3. The age trend of happiness is U-shaped. In existing studies, women are happier than men. The two biggest negative factors in life are unemployment and divorce. Education and happiness are related, even if the same

income is controlled, more education is happier.

4. In every industrialized country, the structure of the happiness equation has the same general format (as far as the current situation is concerned). In other words, perhaps by adding data from developing countries, the structure will change.
5. There is adaptation. Whether it is a good thing or a bad thing in life, when people get used to it, the influence will gradually weaken.
6. Relative things matter a great deal. For example, people are more concerned about the views of people who are closely connected with themselves. Furthermore, the comparison of relative income also affects happiness. Later, unfair income distribution has an impact on happiness, but it is not large.

Let us state more details: Overall satisfaction relates to specific satisfaction associated with a person's job, education, income, family, leisure time, and the like. Several social scientists have evaluated the relationship between subset satisfaction and satisfaction with life as a whole. (Andrews and Withey 1976[2]; Argyle 1989[22];Vermunt, Spaans and Zorge 1989[23]; Veenhoven 1996[24]; Kousha and Mohseni 1997[25]). Happy is a good personal evaluation of the overall quality of life and is generally considered to be the ultimate goal of life. Almost everyone wants to be happy. happy depends on many factors, including income, labor market conditions, job characteristics, health, leisure, family, social relations, safety, freedom, moral values, and many other factors [26]. The economics literature regarding happiness started with Easterlin (1974) [7]. There are three major reasons for economists to study happiness. The first is economic policy. The second reason is the effect of institutional conditions, such as the quality of governance and the size of urbanization on individual happiness. The third reason for happiness research is to understand the structure of happiness (Frey and Stutzer 2002) [27].

## III. RESEARCH METHODOLOGY

### *A. Data*

The research data used in this paper are all from the WVS database, and it is worth mentioning that the wave 6 of data is applied. Its sampling time is probably from 2012 to 2014, and random sampling method was introduced to obtain the recorded data in different countries. In terms of country selection methods, this article selects the top 30 countries (or regions) according to the ranking of the GDP of each country (or region) by the World Bank database. However, since the WVS database was not investigated in some countries, only 20 countries (or regions) were left to be selected for the study. The selected countries (or regions) are: United States, China, Taiwan, Japan, Germany, India, Brazil, Russia, South Korea, Australia. Spain, Mexico, Tukey, Netherland, Argentina, Sweden, Poland, Thailand, Norway and Nigeria. After the data is selected, we perform basic data cleaning, and finally, 33,121 observations are obtained.

### *B. Model*

Reference [2], [3] and [4] shows that feelings of happiness comprise of a cognitive-evaluative factor (life satisfaction) and an affective factor (happy). Therefore, based on the available data, I created a happiness score, the source of this happiness score is the happy or not currently and overall life satisfaction and the happiness score equal to square of happy choice plus square of overall life satisfaction. The first part, whether you are happy currently, in the questionnaire measurement, there are four options:1- Not at all happy, 2- Not very happy, 3 - Rather happy and 4 - Very happy. The second part, overall life satisfaction, it is divided into 1-10 measurement degrees, 1 means completely dissatisfied and 10 means completely satisfied. Accordingly, our goal is to predict this happiness score.

The linear regression model is:

$$Happscore_i = \beta_0 + X_i\beta_i + \varepsilon_i \quad (1)$$

Where $Happscore_i$ are happiness score,

$\beta_i$ are coefficients of the explanatory variables,

$\beta_0$ is constant term,

$\varepsilon_i$ are Estimation errors,

$i$ are the number of observations.

### *C. Apply machine learning*

From the idea of linear regression (OLS) in prediction, we would like to find the best coefficient to minimize the sum square of error:

$$Min: \sum (y - \beta_0 - x_j\beta_j)^2 \quad (2)$$

The best means "the best linear unbiased estimators". However, considering that the explanatory variables will directly have a certain correlation, it may cause the instability of the prediction, that is, the variance of unstable changes in coefficient of explanatory variables under linear regression method .

To solve this problem, firstly, Ridge regression was proposed:

$$\begin{aligned} &Min: \sum (y - \beta_0 - x_j\beta_j)^2 \\ &s.t.: \sum (\beta_0^2 + \beta_1^2 + ... + \beta_j^2) \le c^2 \end{aligned} \quad (3)$$

Or in the matrix form:

$$Min: \|y - A\beta\|_{Ols}^2 - \lambda\|\beta\|_{L2}^2 \quad (4)$$

Secondly, the least absolute shrinkage and selection operator (LASSO) was also proposed:

$$\begin{aligned} &Min: \sum (y - \beta_0 - x_j\beta_j)^2 \\ &s.t.: \sum |\beta_0| + |\beta_1| + ... + |\beta_j| \le c \end{aligned} \quad (5)$$

Or in the matrix form:

$$Min: \|y - A\beta\|_{Ols}^2 - \lambda\|\beta\|_{L1} \quad (6)$$

Sometimes, the lasso regression can cause a small bias in the model where the prediction is too dependent upon a specific variable. In these cases, elastic net is proved to better it combines the regularization of both lasso and Ridge . The strong point of that it does not easily eliminate the high collinearity coefficient. It is:

$$Min: \|y - A\beta\|_{Ols}^2 - \lambda(\alpha\|\beta\|_{L1} + (1-\alpha)\|\beta\|_{L2}^2) \quad (7)$$

*D. Empirical model*

$$Happscore = \beta_0 + \beta_1 x_1 + ... + \beta_{31} x_{31} + \varepsilon \quad (8)$$

Where : $x_1$ to $x_6$ are the importance of family, friends, leisure time, politics, work, and religion. $x_7$ is state of health. $x_8$ is social trust. $x_9$ is freedom of choice level. $x_{10}$ is feeling of social fairness. $x_{11}$ are marital status. $x_{12}$ are number of children. $x_{13}$ is current financial situation in family. $x_{14}$ is income equality. $x_{15}$ is the level of lack in enough food in the last period. $x_{16}$ is the level of lack in enough security in the last period. $x_{17}$ is the level of lack in enough medical service in the last period. $x_{18}$ is the level of lack in enough cash in the last period. $x_{19}$ is expectation in future. $x_{20}$ is the degree of national pride. $x_{21}$ are employment status. $x_{22}$ is the past financial situation in family. $x_{23}$ is income class positioning. $x_{24}$ is income. $x_{25}$ is gender. $x_{26}$ are year of birth. $x_{27}$ is the age. $x_{28}$ are the status of education. $x_{29}$ are the years of schooling. $x_{30}$ is "Whether live with parents". $x_{31}$ is the town size. $\beta_0$ is constant term and $\beta_1$ to $\beta_{31}$ are coefficients of the explanatory variables.

*E. Empirical analysis process*

First, divide the data into two parts, the training set and the test set. The training set accounted for 80% of the total data, and the training set accounted for 20% relatively.

Second, set the custom for machine learning. The repeated cross-validation method is practiced in the empirical analysis. Specifically, we subdivide the training set into 10 parts again. Nine of them are used for machine learning, and the last one is used for testing, and it is repeated five times randomly.

Finally, run the linear regression model, Ridge regression model, LASSO regression model, and elastic net model in sequence. At the same time, taking the minimum root mean square error, mean absolute error, and maximum R square as the main reference value, the best model is selected, the coefficient of the explanatory variable is found, and then the prediction model is obtained.

## IV. EMPIRICAL ANALYSIS RESULTS

After comparing the four models in table 1, we find that the elastic net model has advantages in root mean square error, mean absolute error, and R square. Therefore, we choose the elastic net model as our prediction model. From equation (7), we know that it is necessary to find the value of lambda and value of alpha. The feedback provided by machine learning is "Fitting alpha = 0.111, lambda = 0.1 on full training set."

After selecting the prediction model, try to test the prediction model using the test set data. The comparison results are as follows in table 2.

Table 1: Model comparison

Mean absolute error

| | 1st Qu. | Median | Mean | 3rd Qu. |
|---|---|---|---|---|
| LinearModel | 18.39 | 18.55 | 18.54 | 18.68 |
| Rideg | 18.41 | 18.58 | 18.58 | 18.77 |
| Lasso | 18.40 | 18.54 | 18.54 | 18.71 |
| ElasticNet | 18.38 | 18.53 | 18.54 | 18.73 |

Root mean square error

| | 1st Qu. | Median | Mean | 3rd Qu. |
|---|---|---|---|---|
| LinearModel | 23.47 | 23.63 | 23.64 | 23.79 |
| Rideg | 23.41 | 23.67 | 23.64 | 23.91 |
| Lasso | 23.47 | 23.66 | 23.64 | 23.85 |
| ElasticNet | 23.41 | 23.57 | 23.64 | 23.91 |

R squared

| | 1st Qu. | Median | Mean | 3rd Qu. |
|---|---|---|---|---|
| LinearModel | 0.34 | 0.35 | 0.35 | 0.36 |
| Rideg | 0.33 | 0.35 | 0.35 | 0.36 |
| Lasso | 0.34 | 0.34 | 0.35 | 0.36 |
| ElasticNet | 0.33 | 0.35 | 0.35 | 0.36 |

Table 2: Comparison of training set and test set

Elastic net model with training set

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | RMSE |
|---|---|---|---|---|---|---|
| -15.13 | 54.82 | 66.91 | 65.71 | 77.80 | 114.42 | 23.58 |

Elastic net model with test set

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | RMSE |
|---|---|---|---|---|---|---|
| -2.599 | 54.445 | 66.508 | 65.579 | 77.607 | 114.436 | 23.45 |

alpha=0.111 and lambda=0.1

It is not difficult to find from Table 2 that the application of the elastic net model in the training set and the test set makes our target variable, namely the happiness score, very similar. And the root mean square error on the training set is still some less than the test set, even if it is not much. We can conclude that the elastic net model is a model that is suitable to a certain degree to predict the happiness score. The coefficients of each explanatory variable are all presented in Table 3, and I will discuss the specifics in the follows:

*A. Important in life*

Among the explanatory variables in this section, there are 6 main variables. They are family, friends, leisure time, politics, work, and religion. In the original questionnaire, the options 1 to 4 represent a decrease in importance. 1 means the most important, 4 means not important. Returning to the analysis of the coefficient, the importance of family (-3.606) and leisure time (-1.119) in life has a certain effect on the improvement of happiness score. Perhaps we can say this, the more important the individual values family and leisure time, the higher the happiness score.

*B. Health and medical services*

Unsurprisingly, the better the individual's health (-7.023), the lower the feeling of unhappiness. In the same way, the more medical services (0.75) you get, the more happiness

score you get. The only difference is that the current health status has a relatively greater impact on happiness scores than the medical services received.

Table 3: Coefficients of each explanatory variable

| Variables | Coefficient | Variables | Coefficient |
|---|---|---|---|
| (Intercept) | 39.465 | national_pride3 | -3.468 |
| family | -3.606 | national_pride4 | -2.576 |
| friends | 0.327 | national_pride5 | 0.731 |
| leisuretime | -1.119 | employment_status1 | -1.281 |
| politics | 0.168 | employment_status2 | -1.150 |
| work | -0.577 | employment_status3 | -3.056 |
| religion | -0.599 | employment_status4 | -0.026 |
| health | -7.023 | employment_status5 | 1.043 |
| social_trust1 | 2.156 | employment_status6 | -0.358 |
| social_trust2 | 0.158 | employment_status7 | -3.063 |
| freedom_choice | 2.417 | employment_status8 | -3.231 |
| social_fairness | 0.372 | past_financial _situation | 0.068 |
| marital_status2 | 3.571 | income_class | -1.212 |
| marital_status3 | -4.364 | income | -0.322 |
| marital_status4 | -1.028 | gender1 | -1.648 |
| marital_status5 | -3.501 | gender2 | -0.011 |
| marital_status6 | -0.884 | age | 0.043 |
| marital_status7 | 6.849 | education1 | -3.086 |
| number_children | 0.431 | education2 | 4.587 |
| current_financial _situation | 3.878 | education3 | 1.497 |
| income_equality | 0.098 | education4 | . |
| lack_food | 1.167 | education5 | . |
| lack_security | -0.432 | education6 | 1.304 |
| lack_medical_ service | 0.750 | education7 | 0.961 |
| lack_cash | 0.284 | education8 | -1.521 |
| expectation | 0.135 | education9 | -1.149 |
| national_pride1 | 2.564 | schooling | 0.096 |
| national_pride2 | -2.069 | townsize | -0.100 |

### *C. Income and economic factors*

According to the analysis results obtained, the current financial situation (3.878) of households has a greater impact on happiness than the financial situation in the past (0.068). Therefore, it is more efficient to formulate policies to improve happiness by improving the current financial situation of households.

The next concern is income. We get similar conclusions to the literature reviewed earlier. The increase in income does not necessarily bring happiness. However, in the case of low income or insufficient income in life to pay for expenses, the impact of income on happiness is still positive.

### *D. Demographic factors*

In this part of marital status, there is no obvious evidence that marriage makes people happier. It is worth noting that the two factors of cohabitation but unmarried (3.571) and "in relationship" (6.849) have a very significant positive effect on happiness.

In this part of the employment situation, it seems that most occupations have no positive effect on the improvement of happiness. This probably means that people do not like work. But there is a career that has a positive effect on happiness, and that is housewives (1.043), which is amazing.

Regarding gender, women are indeed more likely to perceive happiness than men. And with age, happiness will increase slightly. Different education levels have different perceptions of happiness. As far as the analysis is concerned, the higher the education level, the harder it is to perceive happiness. The magic is that the longer you take in schooling, the more happiness you can perceive.

## V. DISCUSSION

This paper aims to find the most accurate model for predicting happiness scores. Therefore, the data from World Values Survey was used in our research. The top 30 countries and regions in the world ranking by GDP plan to participate in the study, but because not all countries or regions are included in the database, only 20 countries or regions were selected to participate in the study. Due to the correlation between various explanatory variables, the traditional linear regression model cannot accurately estimate the coefficients, making our predictions risky. Ridge regression, LASSO regression, and elastic net based on machine learning are applied to our research goals. By comparing the relevant statistics, we finally select the prediction model based on the elastic net model.

The main theoretical contribution of this paper is to show the advantages of machine learning in traditional statistical models to a certain extent. Especially for prediction, machine learning shows great productivity. The practical contribution of this paper is mainly to provide reference for the government to improve the happiness level of residents in the most efficient way.

## REFERENCES

[1] Layard, R., 2011. Happiness: Lessons from a New Science. Penguin, UK.
[2] Andrews, F. M., & Withey, S. B. (1976). Social indicators of well-being. New York: Plenum Press.
[3] Campbell, A. (1981). The sense of well-being in America. New York: McGraw-Hill.
[4] Emmons, R. A., & Diener, E. (1985). Factors predicting satisfaction judgment: A comparative examination. Social
[5] Tsou, M. W., & Liu, J. T. (2001). Happiness and domain satisfaction in Taiwan. Journal of Happiness Studies, 2, 269–288.
[6] Gitmez, A. S., & Morc¸o¨l, G. (1994). Socio-economic status and life satisfaction in Turkey. Social Indicators Research, 31, 77
[7] Easterlin, R. A. (1974). Does economic growth improve the human lot? Some empirical evidence. In Nations and households in economic growth (pp. 89-125). Academic Press.
[8] Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all?. Journal of Economic Behavior & Organization, 27(1), 35-47.

[9] Clark, A. E., & Oswald, A. J. (1994). Unhappiness and unemployment. The Economic Journal, 104(424), 648-659.
[10] Ng, Y. K. (1996). Happiness surveys: Some comparability issues and an exploratory survey based on just perceivable increments. Social Indicators Research, 38(1), 1-27.
[11] Ng, Y. K. (1997). A case for happiness, cardinalism, and interpersonal comparability. The Economic Journal, 107(445), 1848-1858.
[12] Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. The quarterly journal of economics, 112(2), 375-406.
[13] Winkelmann, L., & Winkelmann, R. (1998). Why are the unemployed so unhappy? Evidence from panel data. Economica, 65(257), 1-15.
[14] Di Tella, R., & MacCulloch, R. (1999). Partisan Social Happiness. Harvard Business School. mimeo.
[15] Frey, B. S., & Stutzer, A. (2000). Happiness, economy and institutions. The Economic Journal, 110(466), 918-938.
[16] Di Tella, R., MacCulloch, R. J., & Oswald, A. J. (2001). Preferences over inflation and unemployment: Evidence from surveys of happiness. American economic review, 91(1), 335-341.
[17] Tella, R. D., MacCulloch, R. J., & Oswald, A. J. (2003). The macroeconomics of happiness. Review of Economics and Statistics, 85(4), 809-827.
[18] Blanchflower, D. G., & Oswald, A. J. (2004). Well-being over time in Britain and the USA. Journal of public economics, 88(7-8), 1359-1386.
[19] Helliwell, J. F., & Putnam, R. D. (2004). The social context of well–being. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359(1449), 1435-1446.
[20] Frijters, P., Haisken-DeNew, J. P., & Shields, M. A. (2004). Investigating the patterns and determinants of life satisfaction in Germany following reunification. Journal of Human resources, 39(3), 649-674.
[21] Frijters, P., Haisken-DeNew, J. P., & Shields, M. A. (2004). Money does matter! Evidence from increasing real income and life satisfaction in East Germany following reunification. American Economic Review, 94(3), 730-740.
[22] Argyle, M. (1989). The psychology of happiness. London: Routledge.
[23] Vermunt, R., Spaans, E., & Zorge, F. (1989). Satisfaction, happiness and well-being of Dutch students. Social Indicators Research, 21, 1–33.
[24] Veenhoven, R. (1996). Developments in satisfaction research. Social Indicators Research, 37, 1–46.
[25] Kousha, M., & Mohseni, N. (1997). Predictors of life satisfaction among urban Iranian women: An exploratory analysis. Social Indicators Research, 40, 329–357.
[26] Ahn, N., Garcia, J. R., & Jimeno, J. F. (2004). The impact of unemployment on individual well-being in the EU. Europen Network of Economic Policy Research Institutes, Working paper, No: 29.
[27] Frey, B. S., & Stutzer, A. (2002b). What can economists learn from happiness research? Journal of Economic Literature, 40(2), 402–435.