# Happiness Index Prediction Using Hybrid Regression Model

Yuning Han[1], Yichen Shao[2](✉), and Yazhuo Zhang[3,4]

[1] School of Public Affairs and Administration, University of Electronic Science and Technology of China, Chengdu 611731, China

[2] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
syc_bupt@bupt.edu.cn

[3] School of Natural Science, University of International Business and Economics, Beijing, China

[4] University of Manchester, Manchester M13 9LP, UK

**Abstract.** In the field of social science, research on happiness combines multiple subjects like philosophy, psychology, sociology and economics, and plays an essential role in national health. In this report, with the publicly available questionnaire results, we select multiple sets of variables, including individual variables family variables, social attitudes, to predict its evaluation of happiness. During data preprocessing, imputation, and outlier processing, binning and the one-hot encoder were applied to stabilize and rationalize the data. As for the prediction algorithms, Extreme Gradient Boosting (XGBoost), CatBoost, and Gradient Boosting Regressor are put into use, and then we put forward the weighted average methods to fuse the models and reach the final results. In the final results, we discovered that all the features are significantly related to the happiness index, in which being depressed for a long time, lack of exercise, or being in a toxic social environment may all have severe unfavorable effects on the level of happiness. It is worth stressing that such research on happiness can help optimize the allocation of resources and the application of certain policies in order to raise the domestic happiness level, which is of great impact on both the economy and our whole society.

**Keywords:** Happiness prediction · XGBoost · CatBoost · Gradient Boosting · Model fusion with weighted averaging

## 1 Introduction

Life is nonlinear and full of setbacks, but having a high level of happiness can make our life smoother as well as make the society more harmonious and vibrant. Many factors contribute to the level of happiness, including personality, self-esteem [1, 2], mental and physical state [2], sleep quality [2, 3], family environment, social interactions [2, 4, 5], social security, public service and so on. Therefore, a proper analysis of happiness index may shed some light on improving these kinds of factors. In this research, we have joined

Y. Han, Y. Shao and Y. Zhang—The authors are equally contributed.

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 76–87, 2023.
https://doi.org/10.2991/978-94-6463-198-2_9

a competition about predicting happiness index. In addition to prediction, we also focus on investigating the correlations among happiness and behavioral factors like exercise and social activities, as well as depression, health and public service.

Prior studies about happiness put emphasis on different aspects. Research by Jaques et al. shows that a positive mental state is beneficial to resistance to depression. They implement Gaussian Mixture Models and ensemble classification to predict happiness. It turns out that SVM and RF classifiers can provide the best results [4]. Another research done by them compares three formulations of Multi-task Learning and builds up optimal models to predict people's well-being, with the hope of finding the treatment for mental illness. It finally concludes that all of the three formulations have their strengths and drawbacks [6]. In terms of an individual's working life, research by Piyanuch et al. points out that positive mental and physical health contributes to better attitudes towards work, which can eventually improve productivity in study or work. They practice several algorithms, including KNN, Multi-Layer Perceptron, Decision Tree, and Naïve Bayes. The Decision Tree with Random Over sampler is revealed to be the best method in the end [7]. As for machine learning algorithms, there are a variety of studies on boosting models. Research about comparing gradient boosting algorithms illustrates that in general, CatBoost performs best in terms of the average accuracy, while other algorithms like XGBoost and LightGBM don't have statistically significant differences between each other [8].

Our primary purpose of this research is to evaluate and make predictions about the happiness index with different models, and also do some regression analysis to explore the relationship between the happiness index and the factors mentioned above. In this study, we use a dataset from one of the questionnaires of the Chinese General Social Survey (CGSS) project, of which the respondents are Chinese people from all walks of life. As is partly shown in Table 1, the dataset is described by both demographics and general information, with a total of 138 features and 1 target variable, namely the happiness index. The degrees of the happiness index vary from 1 to 5, with 1 representing the lowest level, and 5 representing the highest level. We divide the features into 3 different aspects: individual variables, family variables, and social attitudes.

As the dataset has a large number of dimensions and instances, we apply 3 machine learning algorithms called XGBoost, CatBoost and Gradient Boosting to predict the happiness index and compare their performances. In this paper, we state the contribution as follow. We aim to improve three basic model by weighted averaging fusion and the proposed method achieves the best performance on the test set, which shows the effectiveness of the weighted fusion method. The results indicate that XGBoost performs slightly better than others in terms of the mean squared error scores. Further, by using the reciprocal of the average of the scores as a weight matrix, and applying weighted average method to combine the prediction results of the 3 boosting models, the mean squared error of the test set successfully reduces. By now, the ranking of our final result in the competition is 145 out of 8558 participating groups. Besides, the linear regression analysis on different features and happiness also shows the significance of happiness for individuals and even the entire society.

The 3 columns in the table represent variable names, questions in the survey, and subjects of the question, respectively. But the table is slightly altered to fit in the paper,

**Table 1.** Questions in the Survey

| Name | Question | Subject |
|---|---|---|
| survey_type | Sample type | countryside/city |
| province/city/county/time | Location and time of interview | |
| gender | Gender | |
| birth | Year of birth | interviewee, spouse, father, mother |
| nationality | Nationality | |
| religion | Religious belief and the frequency of attending religious activities | |
| edu | Highest level of education | interviewee, spouse, father, mother |
| edu_status | Education status | |
| edu_yr | Year of receiving a diploma of the highest education level | |
| income | Income of last year | interviewee, spouse |
| political | Political outlook | interviewee, spouse, father, mother |
| join_party | Year of joining the party | |
| floor_area | Floor area of the house you currently live in | |
| property | The property right of the house you live in | yourself, spouse, child, parent, spouse's parent, child's spouse, other relatives, other people, |
| height/weight | Height and weight | |
| health | How do you feel about your current health condition? | |
| health_problem | The frequency that your health problems affect your work and life | |
| depression | The frequency of feeling depressed in the past four weeks | |
| hukou | Registered residence and its location | interviewee, spouse |

(*continued*)

**Table 1.** (*continued*)

| Name | Question | Subject |
|---|---|---|
| media | The frequency of using media in the past year | newspapers, magazines, broadcast, television, the Internet, customized message on mobile phone |
| leisure | The frequency of having leisure in the past year | TV or DVD, movies, shopping, reading, cultural activities, parties, music, exercising, sport competitions, handcrafts, surfing the Internet |
| leisure_time | The frequency that you spend your leisure time on: socializing | socializing, relaxing, learning |
| social | The frequency that you socialize with your neighbor/friends | |
| social_outing | In the past year, how many nights did you spend out of home because of vacations or visiting friends? | |
| equity | In general, do you think the society is fair? | |
| happiness | In general, do you think your life is happy? | |
| class | Which social class do you think that you belong to? | at present, 10 years ago,<br>10 years later, when you are 14 |
| work_exper | Your working experience and current state | interviewee, spouse |
| work_status | Your current working status | interviewee, spouse, father, mother |
| work_yr | How many years have you been working since your first job? | |
| work_type | The type of your current work | interviewee, spouse |
| work_manage | The working management of your current job | |
| insur | Do you currently participate in the social security program? | (basic/commercial) medical & endowment insurance |

(*continued*)

**Table 1.** (*continued*)

| Name | Question | Subject |
|---|---|---|
| family_income | The total income of your family last year | |
| family_m | How many people live together with you? | |
| family_status | How is your family economic status? | |
| house | How many houses do your family own? | |
| car | Do your family have a car? | |
| invest | Is your family engaged in the following investment activities? | stock, fund, bond, futures, warrant, speculation in real estates, foreign exchange |
| son/daughter | The number of your sons/daughters | |
| minor_child | The number of your children under 18 | |
| marital | Marital status | |
| marital_1st | The year of your first marriage | |
| marital_now | The year when you and your spouse get married | |
| status_peer | Among people of your age, where is your socioeconomic status? | |
| status_3_before | Compared to 3 years ago, how have your socioeconomic status changed? | |
| view | How often do your views on important issues meet that of the public? | |
| inc_ability | Considering your ability and working status, do you think your current income is reasonable? | |
| inc_exp | How much do you think your income should be to meet your expectation? | |

(*continued*)

**Table 1.** (*continued*)

| Name | Question | Subject |
|---|---|---|
| trust | How do you trust others in normal social connections not involving financial interests directly? | neighbors, villagers, relatives, colleagues, acquaintances, former classmates, fellow townsmen, people you have same spare-time/religious/charitable activities with, strangers |
| neighbor_ familiarity | How familiar are you with your neighbors? | |
| public_service | How satisfied are you with different kinds of public services? | education, health care, housing protection, social management, employment, social security, assistance for the disadvantaged people, culture & sports, infrastructure |

for some questions have been asked multiple times to different people or subjects. For example, there are questions about the birth year of not only the interviewee himself (herself) but also his (her) spouse and parents. There are also questions asking about the frequency of using all kinds of media, including newspapers, magazines, broadcasts, television, the Internet, and customized messages on mobile phones. Limited by the length of paper, we merge the questions which are similar in form but differ in subject and show the subjects in column "subject". The names are also changed accordingly after the fusion, for example, "media" was originally media 1–6, but the sequence numbers are removed to suit the altered table. As for the non-mergeable questions, the names are the same as the variable names, and the subjects are void.

## 2 Method

### 2.1 Data Preprocessing

Since the dataset is complicated with a considerable number of selective questions, it is rather important to impute the data before embarking upon using it. There are different kinds of missing data, and we deal with each of them separately. We delete those values which are illogical or not corresponded properly, select those with special conditions and fill them artificially, use the median to fill up for all the negative values, and fill in the rest rational missing value with $-1$.

After the imputation, we plot the boxplot of part of the continuous variables, find the outliers, and delete them. With the processed data, we apply some feature engineering. We combine height and weight into BMI, and transform "the time when joining the party" into "whether in the party now" so that the value is either 1 or $-1$. The accuracy of the outcome isn't any better after deleting some data that are not independent, so we stick with the former dataset.

To make the dataset suit our algorithm better, we take some further steps. First, for the continuous variables, we take equal frequency (or distance) binning, separate and stabilize them. Second, we deal with the three columns (invest_other/edu_other/property_other) containing text data by classifying the texts into several general categories according to their meanings and further transferring the strings into ordinal type which is integer. Third, we apply the same preprocessing to the test set, then combine the test set and the training set, take out the features whose value had no obvious linear relation with the meaning of the feature, and apply one-hot encoder, before finally reseparating the training set and the test set.

## 2.2 Introduction of Algorithms

**1) XGB regression**

First introduced by Chen et al., XGBoost is outstanding for its extraordinary efficiency and relatively high accuracy [9]. If the generation of a weak prediction model for each step of the boosting algorithm is based on the Gradient direction of the loss function, it is called Gradient boosting. XGBoost algorithm uses a forward stepwise additive model, while no longer calculating a coefficient after generating weak learners in each iteration.

The model goes as follows:

$$F_T(X) = \sum\nolimits_{m=0}^{T} f_m(x). \tag{1}$$

XGBoost algorithm realizes the generation of the weak learner by optimizing structured loss function (loss function with regular term added can reduce the risk of overfitting), and XGBoost algorithm does not adopt the searching method, but directly uses the first and second derivative value of loss function. The performance of the algorithm is improved by pre-sorting and weighted quantile and other techniques.

**2) CatBoost regression**

CatBoost is a Boosting algorithm developed by Russian Y Andex in 2017, it is an improved realization under the framework of Gradient Boosting Decision Tree (the GBDT) algorithm [10], based on oblivious trees algorithm with fewer parameters, category variables support, and higher accuracy. CatBoost is composed of categorical and Boost, which makes it possible to process categorical features efficiently and reasonably. In addition, it deals with Gradient bias and Prediction shift, in order to improve the accuracy and generalization ability of the algorithm.

When adapting CatBoost, high model quality can be achieved without tuning, and very good results can be achieved by using the default parameters, which reduce the time spent on tuning. For the model support category variables, pre-processing of non-categorical variables in not necessary. Within the model, a new gradient lifting mechanism is proposed to reduce overfitting to improve accuracy. In addition, the model can provide fast prediction, making it easier to deploy models fast and efficiently even for very demanding tasks.

**3) Gradient boosting regressor**

Gradient Boosting is a machine learning technique for regression and classification problems. It integrates weak prediction models, typically decision trees, to produce a strong prediction model. The method builds the weak model in stages and builds the weak model by optimizing an arbitrary differentiable loss function in each stage. Gradient Boosting the GBDT is a forward distribution algorithm reducing the loss through fitting the previous residual each time.

The GBDT is an algorithm of efficiency and high accuracy, it is able to flexibly handle various types of data, including continuous and discrete data. The accuracy of prediction can be relatively high in the case of relatively little adjustment time. But at the same time, for the dependence between weak learners, it is sometimes difficult to train data in parallel [11].

The key is to the GBDT is fitting a regression tree by using the value of the negative gradient of the loss function in the current model as the approximate value of the residual in the ascending tree algorithm for regression problems. For example, training data $x_i$, $y_i$ in the $m$-th round, first we need to calculate the residual r$_{mi}$ of $f_{m-1}(x)$ as the data of round m fitting, where

$$r_{mi} = -\left[\frac{\delta L(y, f(x_i))}{f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{2}$$

**4) Model fusion with weighted averaging**

It is possible to improve machine learning performance by fusing multiple different models. This method is widely used in various machine learning competitions, and it is also the key to sprint the top at the critical moment of the competition. Models can be fused from different perspectives such as model results, the model itself, and sample sets.

From the perspective of algorithms, the most commonly used fusion approach is the weighted averaging method. That is, candidate results and result scores generated from different recommendation algorithms are further weighted by an ensemble to generate the final recommendation ranking results [12].

$$H(x) = \sum \omega_i \cdot h_i(x), \tag{3}$$

where $\omega_i \geq 0$, $\sum_{i=1}^{T} \omega_i = 1$.

## 3 Results and Discussion

As is shown in Table 2, after adapting the three models (XGBoost, CatBoost, and GBDT), we apply 5-fold cross-validation, adapting mean squared error (MSE) as the scoring index, which provides us with 5 MSE scores for each model. Then model fusion is adapted following Eq. (3). Since the lower the MSE is, the better the algorithm performs, we take the average score of these 5 figures for every model, and calculate its reciprocal

**Table 2.** Results of the Prediction Models on the Test Set

| | XGB Regressor | CatBoost Regressor | Gradient Boosting Regressor | Weighted Average of The Models |
|---|---|---|---|---|
| mean squared error | 0.46876 | 0.47027 | 0.47472 | 0.46704 |

in order to determine the weights for each model, so that the best algorithm weights the most. The function goes as below:

$$\omega_i = \frac{\frac{1}{\overline{mse_i}}}{\sum_{k=1}^{3} \frac{1}{\overline{mse_k}}}, \tag{4}$$

where $\overline{mse_i}$ represents for the average MSE score of model $i$.

Decreasing the MSE to about 0.467, the fusion model shows better performance than all the basic models, which is quite satisfactory.

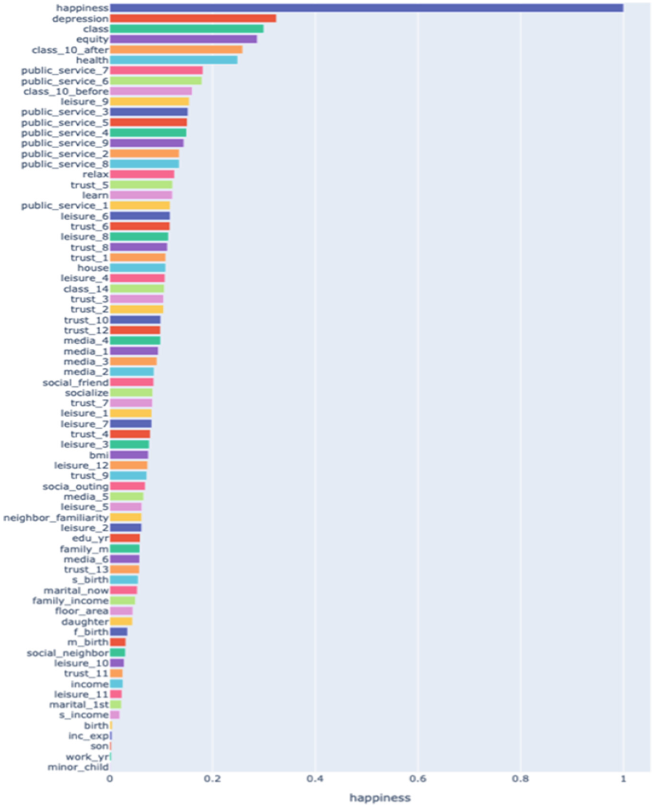

**Fig. 1.** Correlation Ranking of Numerical Features with Happiness Index

Figure 1 illustrates the absolute value of correlation coefficients among happiness and numerical features arranged in rank. It can be seen that depression, self-awareness, equity, satisfaction with different public services as well as exercise and social interactions are all strongly related to happiness. It is also worth noting that being in a depressed mood for a long time, taking less exercise, or being in a negative social relationship may all have an adverse impact on happiness (Table 4).

We choose several important features, which are shown in Table 3, as predictors $X_i$, and happiness index as dependent variable y, to explore the relationship between them. Table 2 shows the overall accuracy of the linear regression model. Although the MSE of regression analysis on the training set is slightly higher than other prediction models, we perform it with only 8 features, which indicates that these features can explain happiness in a relatively good way.

Besides, the formula is

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}. \tag{5}$$

R-squared explains the variance fraction of the model. The F-statistic and its corresponding P-value illustrate that at least one predictor is useful, which means that the overall significance test of the regression equation has passed.

Table 3 shows the results of the test on the null hypothesis. As we can see, all the $P$-values are $< 0.05$, indicating all the features are significantly related to the happiness

**Table 3.** Overall Accuracy of the Regression Model

| Quantity | Value |
|---|---|
| mean squared error | 0.530 |
| R-squared | 0.208 |
| F-statistic | 299.5 |
| prob (F-statistic) | 0.00 |

**Table 4.** Coefficients' Accuracy of the Regression Model

| | *coef* | *std err* | *t* | $P > \|t\|$ |
|---|---|---|---|---|
| intercept | 2.1473 | 0.074 | 28.986 | 0.000 |
| depression | 0.1967 | 0.010 | 19.634 | 0.000 |
| equity | 0.1894 | 0.009 | 22.286 | 0.000 |
| health | 0.0903 | 0.009 | 10.474 | 0.000 |
| social security | 0.0047 | 0.000 | 10.011 | 0.000 |
| socialize | 0.0248 | 0.008 | 3.093 | 0.002 |
| exercise | −0.0498 | 0.006 | −8.969 | 0.000 |
| play with relatives | −0.0487 | 0.011 | −4.456 | 0.000 |

index [13]. However, maybe there are correlations among predictors, which may cause problems.

## 4 Conclusion

In the paper, we begin with a dataset that is too huge to apply common algorithms on, impute and reshape it before finally making it an operatable series of data. To find out the features with the greatest influence on people's sense of happiness, we try out several algorithms and reach the best result by combining three different boosting models and taking their weighted average, which decreases the MSE to 0.467. We also analyze the relationship among numerous features, which are correlated with the happiness index. As our study shows, there is a complex relevance between one's sense of happiness and social attitudes, the family situation as well as personal life, among which depression, self-awareness, social equity, satisfaction with different public services as well as exercise and social interactions are especially influential on one's contentment. Though the discoveries are but the tip of an iceberg, they show great potential in various fields, including making policies, preventing mental illness, improving productivity, and so on.

## References

1. Cheng, H., & Furnham, A. (2003). Personality, self-esteem, and demographic predictions of happiness and depression. Personality and individual differences, 34(6), 921-942.
2. Sano, A. (2015). Measuring college students sleep, stress and mental health with wearable sensors and mobile phones (Doctoral dissertation, PhD thesis, MIT).
3. Sano, A., Amy, Z. Y., McHill, A. W., Phillips, A. J., Taylor, S., Jaques, N., ... & Picard, R. W. (2015, August). Prediction of happy-sad mood from daily behaviors and previous sleep history. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)(pp. 6796–6799). IEEE.
4. Jaques, N., Taylor, S., Azaria, A., Ghandeharioun, A., Sano, A., & Picard, R. (2015, September). Predicting students' happiness from physiology, phone, mobility, and behavioral data. In 2015 International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 222–228). IEEE.
5. Peirce, R. S., Frone, M. R., Russell, M., Cooper, M. L., & Mudar, P. (2000). A longitudinal model of social contact, social support, depression, and alcohol use. Health Psychology, 19(1), 28.
6. Jaques, N., Taylor, S., Nosakhare, E., Sano, A., & Picard, R. (2016, December). Multi-task learning for predicting health, stress, and happiness. In NIPS Workshop on Machine Learning for Healthcare.
7. Chaipornkaew, P., & Prexawanprasut, T. (2019, October). A Prediction Model for Human Happiness Using Machine Learning Techniques. In 2019 5th International Conference on Science in Information Technology (ICSITech) (pp. 33–37). IEEE.
8. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937-1967.
9. Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4–2, 2015, 1(4): 1–4.
10. Hancock J T, Khoshgoftaar T M. CatBoost for big data: an interdisciplinary review[J]. Journal of big data, 2020, 7(1): 1-45.

11. Prettenhofer P, Louppe G. Gradient boosted regression trees in scikit-learn[J]. 2014.
12. Lu S, Hwang Y, Khabibrakhmanov I, et al. Machine learning based multi-physical-model blending for enhancing renewable energy forecast-improvement via situation dependent error correction[C]//2015 European control conference (ECC). IEEE, 2015: 283–290.
13. Alan O. Sykes, An Introduction to Regression Analysis, in Chicago Lectures in Law and Economics, Eric A. Posner ed., New York: Foundation Press, 2000.