Applied and Computational Mathematics
2018; 7(4): 197-202
http://www.sciencepublishinggroup.com/j/acm
doi: 10.11648/j.acm.20180704.13
ISSN: 2328-5605 (Print); ISSN: 2328-5613 (Online)

Subjective Well-Being Prediction Using Data Mining Techniques: Evidence from Chinese General Social Survey

Leibao Zhang[1], Yanli Fan[1, *], Wenyu Zhang[2], Shuai Zhang[2]

[1]School of Public Finance and Taxation, Zhejiang University of Finance and Economics, Hangzhou, China
[2]School of Information, Zhejiang University of Finance and Economics, Hangzhou, China

Email address:
la_bor@zufe.edu.cn (Leibao Zhang), fanyanli@zufe.edu.cn (Yanli Fan), wyzhang@e.ntu.edu.sg (Wenyu Zhang), zhangshuai@zufe.edu.cn (Shuai Zhang)

[*]Corresponding author

To cite this article:
Leibao Zhang, Yanli Fan, Wenyu Zhang, Shuai Zhang. Subjective Well-Being Prediction Using Data Mining Techniques: Evidence from Chinese General Social Survey. *Applied and Computational Mathematics*. Vol. 7, No. 4, 2018, pp. 197-202. doi: 10.11648/j.acm.20180704.13

Received: August 2, 2018; **Accepted**: August 18, 2018; **Published**: September 18, 2018

Abstract: Data mining techniques have attracted increasing attentions recently and played more and more important roles in various domains. However, few studies have used these prevalent techniques to explore the rules of subjective well-being for individuals. In this study, a prevalent data mining method, XGBoost, is applied to predict the subjective well-being according to various predictive factors. Feature selection step is implemented to further improve the prediction results and reduce the computational complex based on the importance calculated by XGBoost. An authoritative academic database, Chinese General Social Survey, is used for providing an evidence for classification prediction performance. Moreover, five benchmark models, i.e., logistic regression, support vector machine, decision tree, random forest, and gradient boosting decision tree, are used for comparative analysis based on three evaluation metrics, Accuracy, AUC and F-score. The experimental results indicate that XGBoost outperforms other benchmark models, and feature selection step can improve the prediction performance and reduce the computational time to some extent. In reality, using data mining methods can deeply explore the rule of subjective well-being based on various predictive features, and provide an overwhelming support for improving subjective well-being. Therefore, the methods used in this study are effective and the results provide a support for making society more harmonious.

Keywords: Subjective Well-Being, Data Mining, XGBoost, Classification Prediction

1. Introduction

Eudaemonics is one of the popular disciplines for studying the rule of generating and pursuing well-being. With the rapid development of current social economy, the diversified demand of people causing the factors of well-being are not limited to economy. Others, such as occupation development, housing condition, social equity, and social security, are playing an increasingly important part in subjective well-being. Therefore, the research on well-being attracts intensively attention in recent years [1-8].

The development and application of big data and data mining techniques promotes the research and practice in many fields. Additionally, subjective well-being is increasingly valued by government and all sectors of society; many social surveys on well-being are conducted, which provide a good basis for studying the rule of subjective well-being. Therefore, exploring subjective well-being with data mining techniques has a promising nature and it provides a scientific basis for improving subjective well-being and livelihood of individuals.

In this study, we initially use feature selection methods to reduce high dimension features. After that, some redundant and irrelevant features are removed to improve the accuracy of subsequent classification prediction and cut down computational time. Moreover, an advanced machine learning classification model, XGBoost [9], is applied to predict subjective well-being and obtains a superior prediction results. Application of data mining techniques in exploring the rule of subjective well-being provides a solid foundation for decision makers to make reasonable decisions and ensure the livelihood of individuals. Furthermore, this study paves a new way for exploring subjective well-being and occupies with a promising nature.

The remainder of this study is structured as follows. Section 2 reviewed the related work about data mining techniques and subjective well-being. Section 3 elaborates the description of dataset, data preprocessing, and the methods used in this study. Section 4 illustrates the experimental results and gives a detailed analysis about exploratory data analysis and comparative results. In Section 5, a conclusion is drawn and some future work is provided for further research.

2. Related Work

Data mining techniques and machine learning methods are extensively used with the arrival of big data era. These methods are widely applied in production and living domain, such as medical treatment [10], trade and commerce [11], and credit scoring [12-13], promoting the development and innovation in these fields.

In recent years, subjective well-being has aroused extensive attention from all sectors of society. More and more researchers have attempted to explore the rule of subjective well-being to provide theoretical basis for improving subjective well-being of individuals and creating a more harmonious and stable society. Dong & Eun-Kyoung [14] used rural community samples and found there are ethnic differences in the reliance on religiosity/spirituality between Elderly Whites, African Americans, and Native Americans; and they indicated that religiosity/spirituality is significantly associated with subjective well-being. Dong [15] applied a sample from Virginia and North Carolina to explore the effect of Carolina spirituality, religiousness, and social support on subjective well-being of rural elderly individuals. Income, closely related to living standards, is always attached great importance to subjective well-being. Luhmann et al. [16] used bivariate latent state-trait models to explore the relation between cognitive well-being and income and concluded that the relationship varies from individual differences. Besides, Liang & Wang [17] studied subjective well-being from prudential value, and obtained effective result using linear regression. Qian & Knoester [18] adopted dataset from 2006 Chinese General Social Survey to study the relationship between parents status and subjective well-being among currently married individuals. Moreover, some researches have explored the subjective well-being from different countries [19], social origins [20], and generations [21].

Although much has been done about subjective well-being, few researchers have explored it using data mining techniques. With the emergence of interdisciplinary research, it is promising to study the well-being from a data mining aspect. Otoiu et al. [22] selected the key variables for well-being and used clustering analysis method to obtain the optimal classification; final results are compared to three well-known and well-documented indicators of well-being, HDI, LPI and HPI. Wang et al. [23] adopted Naïve Bayes, Decision Tree and Rule Induction to predict happiness and well-being, and concluded that Rule Induction outperformed others, comparatively. In this study, a machine learning algorithm, XGBoost, is used to predict subjective well-being and compare with other five widely used classification models; the experimental results demonstrated that it is superior to other comparative methods. Therefore, the applied methods are efficient and effective, and meaningful for studying subjective well-being of individuals.

3. Data Preprocessing and Modeling Methods

In this section, the source and composition of the used data, data preprocessing and data mining methods are elaborate. The general framework of the methods is presented as Figure 1.

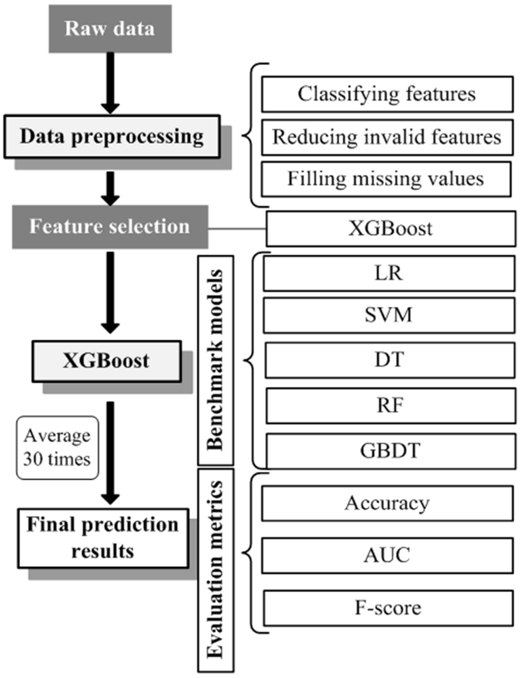

Figure 1. *The general framework of the method.*

3.1. Description of the Dataset

With the development of economy and the improvement of living standard, people are no longer limited to satisfy the basic living demand according to Maslow's hierarchy of needs. Accordingly, more and more researchers have focused on the rule of subjective well-being. The data used in this study, 2015 Chinese General Social Survey (CGSS), is from an authoritative academic database, Chinese National Survey Data Archive (CNSDA), http://cnsda.ruc.edu.cn/.

2015 CGSS covers 478 villages from 28 provinces (municipalities) and contains 10968 efficient samples. The survey is divided into 6 parts corresponding to different functions. Part A and B investigate all of the samples and include almost all endogenous and exogenous variables for social livelihood research. Part C and D investigate 1/6 of all samples, and Part E and F only covers 1/3, mainly for international comparison. Therefore, the data from Part A and B, containing 10968 samples, are adopted to predict the rule of subjective well-being.

Table 1. *The description of representative features for classification models.*

Features	Description	Number of values
a36	Subjective well-being	2
a7a	Highest education level	15
a8a	Total income	0-9999990
a15	Physical condition	6
a16	Frequency of health problem	6
a17	Frequency of depression	6
a18	Residence registration	8
a313	Learn to charge in spare time	6
a35	Social equity	6
a431	Current social classes	11
a432	Social classes before 10 years	11
a433	Social classes after 10 years	11
a54	Reason of no work	11
b1	Economic status compared to contemporary	4
b163	Satisfaction of housing security	0-100
b164	Satisfaction of social management	0-100
b165	Satisfaction of labor employment security	0-100
b166	Satisfaction of social security	0-100
b167	Satisfaction of social services	0-100
b169	Satisfaction of infrastructure	0-100

3.2. Data Processing

Real world data always contains missing values or irregular values, so that data preprocessing is indispensable for improving prediction performance. The main data preprocessing in this study is depicted as following parts.

(1) In CGSS, subjective well-being is classified as “completely satisfied”, “generally satisfied”, “a little dissatisfied”, “totally dissatisfied”, and “unanswered”. To better conduct classification prediction, “completely satisfied” and “generally satisfied” are merged as “satisfied”, noted as “1”; “a little dissatisfied” and “totally dissatisfied” are merged as “dissatisfied”, noted as “0”; samples with “unanswerable” are canceled.

(2) The features with missing value exceeding 50% are canceled, such as some open-ended questions; and some redundant features, such as ID card number, are also deleted to reduce the effect on the prediction accuracy and efficiency of the classification model.

(3) Missing values are filled according to feature types. As for categorical features, a new dimension is generated or fill the missing values according to corresponding features. For example, “contract signing” is related to “occupation condition”, so the missing values in “contract signing” are filled according to “occupation condition”. As for numerical features, average value of the corresponding feature is used to fill the missing values and “inapplicable” is set as “0”.

After data preprocessing, the used data contains 9347 samples, 217 features, and some representative features are described in Table 1 because of the space limit.

3.3. Modeling Methods

In this study, XGBoost, which is the extension to Gradient Boosting Decision Tree (GBDT) [24], is applied as the primary model. XGBoost, recently proposed by Chen & Guestrin [9], has achieved increasingly attraction for its satisfactory performance in many fields. Compared to GBDT, the objective function is expanded by Taylor expansion in XGBoost, enabling to accelerate the convergence speed while training the model. Additionally, a penalty term is added to control the construction of model and prevent overfitting problem.

Moreover, five commonly used data mining methods, logistic regression (LR) [25], support vector machine (SVM) [26], decision tree (DT) [27], random forest (RF) [24], and GBDT, are used to predict subjective well-being for benchmarking analysis. All of data used in experiment is divided into two parts, training set and testing set, with the proportion of 8:2. Six models, including primary model and five comparative models, are trained in training set and then used for predicting subjective well-being in testing set; the prediction results are compared to real subjective well-being to measure the performance of each model. In order to get rid of the contingency in a single experiment, the experiments are repeated for 30 times and the average prediction results of 30 times are recognized as the final prediction results.

4. Experimental Results and Analysis

This section is consisted of two parts. One is data exploratory analysis which analyzes correlation and importance of features. Another is prediction results and performance analysis, which elaborates the comparison between the primary model and other benchmarks. All of these experiments are implemented by python Version 3.6.

4.1. Exploratory Data Analysis

In order to analyze basic structure of the data and explore the correlation and importance of all features, exploratory data analysis is implemented in this sub-section. It mainly consists of two parts, correlation analysis and importance analysis.

Correlation is used to measure the degree of correlation between two features. Top 15 features correlated with subjective well-being are depicted in Figure 2. According to Figure 2, some features can improve subjective well-being, such as satisfaction of social security; but others, such as frequency of depression, can reduce subjective well-being. Regardless of the effect direction, current social classes, frequency of depression, and social classes after 10 years are the top three features correlated with subjective well-being most. According to correlation analysis, priori knowledge of the used data is obtained. Moreover, it helps us explore potential structures and get the most intuition of the data.

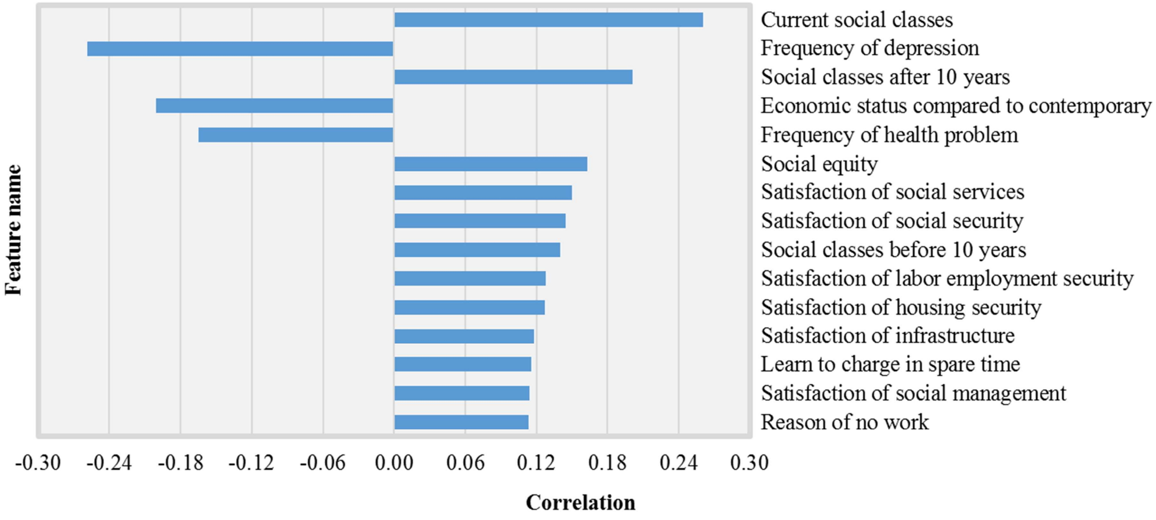

Figure 2. *Top 15 features correlated with subjective well-being.*

Another part of exploratory data analysis is importance analysis, which indicates the importance of each feature for prediction results. XGBoost algorithm is applied to calculate each importance and the results are used in further feature selection to improve the calculation efficiency and prediction performance. Top 15 features with their importance are shown in Figure 3. Social equity, frequency of depression and class of family economic status are three features most contributed to subjective well-being. Compared to correlation results, the result of each feature varies, so using correlation analysis for feature selection remains to be discussed. XGBoost is applied in this study for feature selection and obtaining a good result.

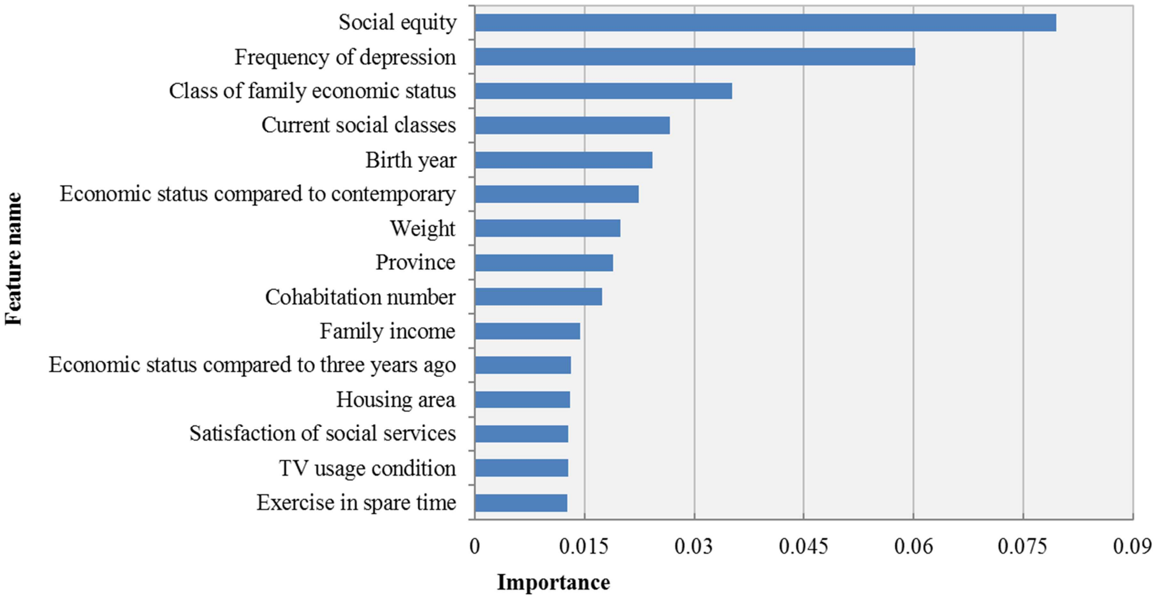

Figure 3. *Top 15 features with their importances by XGBoost.*

4.2. Prediction Results and Performance Analysis

In this sub-section, the comparative results between primary model and five benchmarks are described. In order to eliminate the constraint from a single evaluation metric, three recognized evaluation metrics, Accuracy, AUC, and F-score, are used to evaluate the prediction performance comprehensively. Accuracy indicates the ratio between the number of samples correctly classified and the total number of samples for a given testing set. AUC is the area under Receiver Operating Characteristic (ROC) curve which indicates the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative sample [28]. F-score is a value combined with precision and recall and possesses comprehensive power in valuating classification performance.

As shown in Table 2, XGBoost outperforms other five comparative methods before feature selection according to all of three evaluation metrics. To further elaborate the performance of feature selection, the experiment is implemented after reducing 43 features based on their importance. Comparative results between primary model and benchmarks after feature selection are presented in Table 2.

Table 2. *Comparative results between primary model and benchmarks before and after feature selection.*

Algorithm	Accuracy		AUC		F-score	
	before	after	before	after	before	after
LR	0.9170	0.9174	0.8372	0.8406	0.9557	0.9559
SVM	0.9145	0.9155	0.8227	0.8233	0.9552	0.9557
DT	0.8733	0.8739	0.6297	0.6312	0.9302	0.9305
RF	0.9169	0.9181	0.8608	0.8630	0.9564	0.9570
GBDT	0.9260	0.9363	0.8805	0.8807	0.9605	0.9606
XGBoost	0.9268	0.9268	0.8830	0.8830	0.9610	0.9610

After feature selection, XGBoost is still superior to other comparative methods based on AUC and F-score, but a little weaker than GBDT based on Accuracy. Overall, feature selection help to improve, or at least keep the prediction performance for these methods, but it can reduce the computational time to some extent. Detailed description is shown in Figure 4.

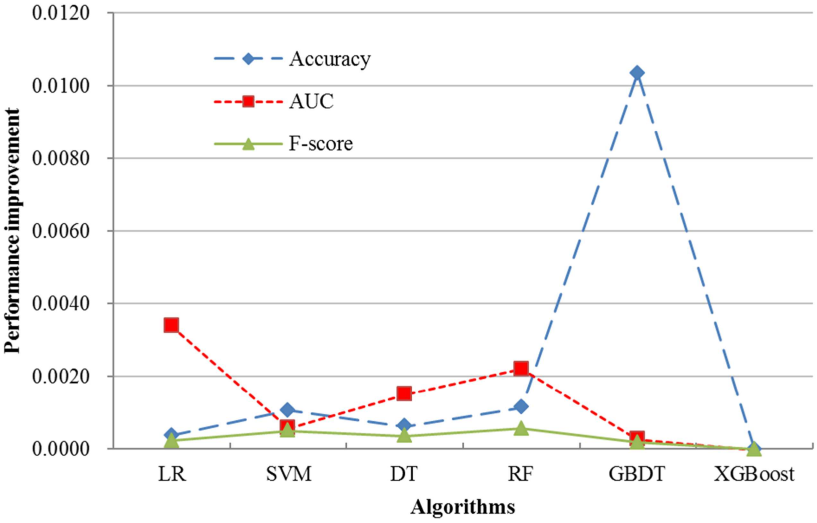

Figure 4. *Performance improvement after feature selection.*

5. Conclusions

Subjective well-being is a prevalent research concern for which it influences everyone in the society and has aroused wide attention from government, stuffs, and every individual. With the development of economy and the improvement of living quality of individuals, the greater diversity of factors will impact subjective well-being. So some effective data mining techniques are indispensable for mining the rule of subjection well-being.

In this study, the data from CGSS is used to predict the subjective well-being of each individual. A prevalent data mining method, XGBoost, is used for feature selection and classification prediction. The obtained final prediction results are compared to five benchmarks, LR, SVM, DT, RF, and GBDT based on Accuracy, AUC, and F-score. The experiment results indicate that the primary model, XGBoost, is superior to other benchmark models. Moreover, feature selection can improve, at least keep the performance of classification prediction, and reduce the computational time to some extent.

In future work, other feature selection methods will be used to further improve the prediction performance. In addition, some ensemble models will be considered to predict subjective well-being, so that it can provide a stronger support for government and all society about how to improve subjective well-being of individuals and make the society more harmonious.

Acknowledgements

This work was supported by the Fund from Xinmiao Talent Project of Zhejiang Province, China. (No. 2018R414050).

References

[1] Gilboa, I., & Schmeidler, D. (2001). A cognitive model of individual well-being. *Social Choice and Welfare*, 18 (2), 269-288.

[2] Liu, J., Xiong, M., & Su, Y. (2012). National sense of happiness in the economic growth period: a study based on CGSS data. *Social Sciences in China*, 12, 82-102.

[3] Zhao, W. (2012). Economic inequality, status perceptions, and subjective well-being in China's transitional economy. *Research in Social Stratification and Mobility*, 30 (4), 433-450.

[4] Hu, A. (2013). Public sector employment, relative deprivation and happiness in adult urban Chinese employees. *Health Promot Int*, 28 (3), 477-486.

[5] Liu, J., Xiong, M., & Su, Y. (2013). National happiness at a time of economic growth: a tracking study based on CGSS data. *Social Sciences in China*, 34 (4), 20-37.

[6] Cheng, Z. (2014). The effects of employee involvement and participation on subjective wellbeing: evidence from urban China. *Social Indicators Research*, 118 (2), 457-483.

[7] Applasamy, V., Gamboa, R. A., Al-Atabi, M., & Namasivayam, S. (2014). Measuring happiness in academic environment: a case study of the school of engineering at taylor's university (Malaysia). *Procedia-Social and Behavioral Sciences*, 123, 106-112.

[8] Coverdale, G. E., & Long, A. F. (2015). Emotional wellbeing and mental health: an exploration into health promotion in young people and families. *Perspectives in Public Health*, 135 (1), 27-36.

[9] Chen, T., & Guestrin, C. (2016). Xgboost: a scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, August 13-17, pp. 785-794.

[10] Kasperczuk, A., & Dardzińska, A. (2016). Comparative evaluation of the different data mining techniques used for the medical database. *Acta Mechanica Et Automatica*, 10 (3), 233-238.

[11] Salehan, M., & Dan, J. K. (2016). Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics. *Decision Support Systems*, 81, 30-40.

[12] Wang, D., Zhang, Z., Bai, R., & Mao, Y. (2017). A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. *Journal of Computational and Applied Mathematics*, 329, 307-321.

[13] He, H. L., Zhang, W. Y., & Zhang, S. (2018). A novel ensemble method for credit scoring: adaption of different imbalance ratios. *Expert Systems with Applications*, 98, 105-117.

[14] Dong, P. Y., & Eun-Kyoung, O. L. (2004). Religiousness/spirituality and subjective well-being among rural elderly Whites, African Americans, and Native Americans. *Journal of Human Behavior in the Social Environment*, 10 (1), 191-211.

[15] Dong, P. Y. (2006). Factors affecting subjective well-being for rural elderly individuals. *Journal of Religion and Spirituality in Social Work Social Thought*, 25 (2), 59-75.

[16] Luhmann, M., Schimmack, U., & Eid, M. (2011). Stability and variability in the relationship between subjective well-being and income. *Journal of Research in Personality*, 45 (2), 186-197.

[17] Liang, Y., & Wang, P. (2014). Influence of prudential value on the subjective well-being of Chinese urban-rural residents. *Social Indicators Research*, 118 (3), 1249-1267.

[18] Qian, Y., & Knoester, C. (2015). Parental status and subjective well-being among currently married individuals in China. *Journal of Family Issues*, 36 (10), 1351-1376.

[19] Oshio, T., Nozaki, K., & Kobayashi, M. (2011). Relative income and happiness in Asia: evidence from nationwide surveys in China, Japan, and Korea. *Social Indicators Research*, 104 (3), 351-367.

[20] Zhang, Z., & Treiman, D. J. (2013). Social origins, hukou conversion, and the wellbeing of urban residents in contemporary China. *Social Science Research*, 42 (1), 71-89.

[21] Cheng, Z., Wang, H., & Smyth, R. (2014). Happiness and job satisfaction in urban China: a comparative study of two generations of migrants and urban locals. *Urban Studies*, 51 (10), 2160-2184.

[22] Otoiu, A., Titan, E., & Dumitrescu, R. (2014). Are the variables used in building composite indicators of well-being relevant? Validating composite indexes of well-being. *Ecological Indicators*, 46, 575-585.

[23] Wang, Y., Wu, Y., & He, W. (2016). Development of classification models for predicting happiness: a data mining approach. *International Journal of Digital Content Technology and its Applications*, 10 (3), 1-10.

[24] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29 (5), 1189-1232.

[25] Hand, D. J., & Kelly, M. G. (2002). Superscorecards. *Ima Journal of Management Mathematics*, 13 (4), 273-281.

[26] Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural network: a market comparative study. *Decision Support Systems*, 37 (4), 543-558.

[27] Li, X., Ying, W., Tuo, J., & Li, B. (2004). Applications of classification trees to consumer credit scoring methods in commercial banks. *In Proccedings of IEEE International Conference on Systems*, Man and Cybernetics, Hague, Netherlands, October 10-13, Vol. 5, pp. 4112-4117.

[28] Fawcett, T. (2004). ROC graphs: notes and practical considerations for researchers. *Pattern Recognition Letters*, 31 (8), 1-38.