IS-467 FUNDAMENTALS OF DATA SCIENCE

HAPPINESS FACTORS

DONGYU XU
SHANGJUN LIU
FIONA BAENZIGER

1. ABSTRACT

[create an abstract - 250 words max]

2. PROBLEM DESCRIPTION

Many young people are finding themselves are puzzled by mental health problems. In an American study, the number of students seen with depression doubled over a 13-year period and the number of suicidal students tripled. University student surveys confirm the high rates of mental health issues on campus, as well as their seriousness. A survey of 51 US postsecondary institutions that yielded over 28,000 student responses revealed that 45% of respondents felt "unhappy", 50% felt "overwhelmed by anxiety", 30% felt "depressed" and 7% had seriously considered suicide in the previous 12 months. A systematic review of 11 articles on students with mental health issues identified anxiety, depression, eating disorders, self-harm, obsessive-compulsive disorder, and psychotic disorders as the most common problems [1]. This problem is spreading worldwide. There is a need to understand the issues these people face in their young life e so that they can be supported in living a healthy psychological condition.

The question is, what are the factors that contribute to overall happiness? People's emotion is influenced by several reason. The data mining study can point out a general idea about what factors is the most significant. By knowing this, psychologists can acquire more understanding about how the emotions generated. Or even diagnose the psychological condition of an observation.

We designed this study to analyze how views on life can be important indicators of certain mental health issues and potential solutions. With set of psychological attributes, we want to explore the key factors that influence people's happiness. By applying the result of analysis, people who suffered by depressed emotions or mental health issue can find a way to get through it.

Data mining, the science of extracting useful knowledge from such huge data repositories, has emerged as a young and interdisciplinary field in computer science. Data mining techniques have been widely applied to problems in industry, science, engineering and government, and it is widely believed that data mining will have profound impact on our society. [2] There are three data mining methods, which are association rule mining, classification and prediction, and clustering.

Classification is a kind of supervised learning technique which classify data into predefined class labels. It is one of the most useful methods in data mining to build classification models from a dataset. The used Classification techniques commonly build models that are used to predict future data trends. The algorithms for data classification used in this study are clustering and decision tree classification technique.

3. ABOUT THE DATASET

The dataset used in the study original comes from a Slovakian study in 2013. The researchers asked students of the Statistics class at FSEV UK. These young people were between 15 to 30

years old. The original questionnaire was in Slovak language and was later translated into English.

The original dataset consists of 860 rows and 150 columns (139 integers and 11 categorical). [illegible] weight, height, and education status.

The variables can be split into eight groups which are music preferences, movie preferences, hobbies & interests, phobias, health habits, personality traits, views on life & opinions, spending habits, and demographics.

4. DATA PRE-PROCESSING

In order to answer the question, what are the factors that contribute to overall happiness? We set the Likert scale question "I am 100% happy with my life" as dependent variable, Happiness. And the independent variables are narrow down to the "view on life & opinions" group specifically, which includes 56 variables in total. All of the variables are coded on Likert scale.

First, there is missing value in the dataset could be cleaned. Observations including missing value were removed from the study since the number of those observation is less than 5% of the total number of cases. The missing data can simply be deleted without any significant ramifications. [3] And because of the nature of the dataset, there is no outlier detected.

Second, the dependent variable, happiness, is binned for decision tree analysis. Unbalanced datasets are common in many real-world domains. A decision tree learned from unbalanced data typically creates a critical challenge related to the high misclassification rate of the minority class. [4] The original distribution is unbalanced in 5 levels of. Some options like "1" and "2" only take 1.46% and 5.72% of the total data. (Figure 1)

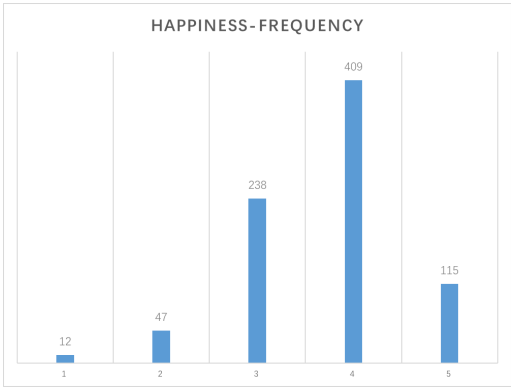

Figure 1: Frequency chart of Happiness

To solve the unbalanced problem, the variable was binned in to a binary variable with two level called "happy" and "unhappy". The first three group of observations who are strongly disagree, disagree and neutral are binned into "unhappy". Observations in agree and strongly agree were binned into "happy". The distribution after binning are shown in Figure 2. Each of those two levels take 36.18% and 63.82% of data. This result is balanced to do further analysis.

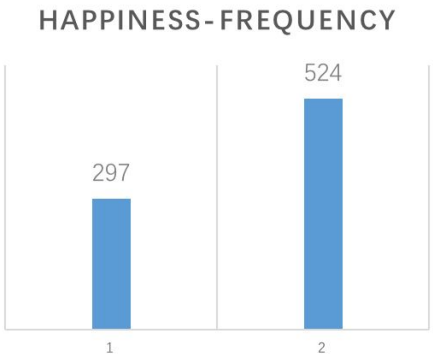

Figure 2: Frequency bar chart of Happiness_recoded

5. PEARSON CORRELATION ANALYSIS

The Pearson Correlation is a measure of linear correlation between pairs of variables. The Pearson correlation analysis was conducted in order to identify if they are strongly correlated with each other.

A correlation value greater than 0.7 is considered a high correlation between two variables. Table 1 shows the pairs of variables with a value greater than 0.7. Pearson's correlation coefficient varies from -1 to 1. The absolute correlation value is reported since a negative correlation is as important as a positive correlation.

Variable 1	Variable 2	Pearson Correlation	Sig. (2-tailed)
Appearenceandgestures	Health	0.95	0
Assertiveness	Loneliness	-0.8	0
Assertiveness	Newenvironment	0.84	0
Borrowedstuff	Prioritisingworkload	0.78	0
Changingthepast	Selfcriticism	0.89	0
Charity	God	0.8	0
Cheatinginschool	Criminaldamage	0.83	0
Children	Empathy	0.89	0
Children	Numberoffriends	0.76	0
Compassiontoanimals	Empathy	0.98	0
Decisionmaking	Writingnotes	0.92	0
Elections	Dailyevents	0.97	0
Energylevels	Changingthepast	-0.79	0
Energylevels	Dreams	0.84	0
Energylevels	Fake	-0.9	0
Energylevels	Hypochondria	-0.94	0
Energylevels	Knowingtherightpeople	0.97	0
Energylevels	Moodswings	-0.98	0
Energylevels	Selfcriticism	-0.82	0
Findinglostvaluables	Cheatinginschool	-0.76	0
Findinglostvaluables	Friendsversusmoney	0.82	0
Findinglostvaluables	Giving	0.9	0
Gettingangry	Achievements	0.8	0
Gettingangry	Changingthepast	0.77	0
Gettingangry	Health	0.75	0
Gettingangry	Lossofinterest	0.88	0
Gettingup	Cheatinginschool	0.98	0
Gettingup	Workaholism	-0.72	0

Giving	Workaholism	0.9	0
God	Friendsversusmoney	0.73	0
Health	Decisionmaking	0.73	0
Health	Writingnotes	0.85	0
Interestsorhobbies	Children	0.8	0
Interestsorhobbies	Decisionmaking	-0.83	0
Keepingpromises	Prioritisingworkload	0.89	0
Knowingtherightpeople	Funniness	0.88	0
Knowingtherightpeople	Health	0.73	0
Lifestruggles	Children	0.98	0
Lifestruggles	Gettingangry	0.9	0
Lifestruggles	Giving	0.78	0
Lifestruggles	Loneliness	0.75	0
Lifestruggles	Newenvironment	-0.78	0
Lifestruggles	Writingnotes	0.93	0
Moodswings	Hypochondria	0.8	0
Moodswings	Selfcriticism	0.94	0
Moodswings	Waiting	-0.88	0
Newenvironment	Health	-0.9	0
Newenvironment	Loneliness	-0.76	0
Numberoffriends	Fake	-0.77	0
Numberoffriends	Giving	0.87	0
Numberoffriends	Selfcriticism	-0.9	0
Parentsadvice	Prioritisingworkload	0.89	0
Parentsadvice	Thinkingahead	0.96	0
Personality	Changingthepast	-0.94	0
Personality	Socializing	0.92	0
Publicspeaking	Assertiveness	-0.79	0
Publicspeaking	Moodswings	0.9	0
Smallbigdogs	Cheatinginschool	0.8	0
Smallbigdogs	Writingnotes	-0.85	0

Table 1: Summary of Correlation Analysis

REFERENCES

[1] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[2] Han J., Kamber M., and Pie J. Data Mining Concepts and Techniques. 3rd edition, Morgan Kaufmann Publishers. 2011.

[3] Harrell, Frank E. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, 2001.

[4] Lee, J., & Zhu, D. (2011). When Costs Are Unequal and Unknown: A Subtree Grafting Approach for Unbalanced Data Classification. Decision Sciences, 42(4), 803-829. doi:10.1111/j.1540-5915.2011.00332.x