Preserving Student Mental Health – A Data Mining Analysis on Student Stress Level

Jing Wang, Yuan Li

Abstract

College student mental health has been getting more and more attention recently. Despite many solutions have been proposed, this is not an easy-to-solve problem. For this project, we are seeking to build a prediction system that can tell when students are easily get stressed out, and thus render possibility to certain interventions which would help college students to maintain a healthy mental condition. We would also expect this project to be a starting point in terms of helping people to better understand College student stress patterns as well.

1 Introduction

The students in college seems live in the tower of ivory and enjoy the peaceful life. But actually the students are facing different issues, such as assignments, projects, courses, the academic challenges, the uncertain future career and the peer to peer pressure. Mental health is a hot topic in every university. According to the Association for University and College Counseling Center Directors survey of counseling center directors, 95% of college counseling center directors surveyed said the number of students with significant psychological problems is a growing concern in their center or on campus. Seventy percent of directors believe that the number of students with severe psychological problems on their campus has increased in the past year. But the mental problem is easy to be overlooked by students and is detected at very late phase (Henriques 2014).

Is there a solution to predict the mental health status? If we can figure out a solution to provide the students mental health in advance, it will help students and schools to support the students who have the potential possibility of mental problem to avoid or relieve the mental diseases. For this project, we are planning to build up a system to predict the students mental health status.

2 Problem Definition and Data

2.1 Problem

It is not uncommon that college students nowadays are overwhelmed by heavy workload and a competitive job market. However, the mental status of college students seem not to get enough attention from the public, and not many of them would seek professional help before things are get really worse. So we want to build this prediction system, with the capacity to predict when college students would easily get stressed out, and thus would let schools and families of college students to take some measures before students get to stressed out.

2.2 Data

In this project we use StudentLife Dataset, which is collected by research group of Dartmouth College. The data set contains a wide variety of data including sensor data, EMA data, survey responses and educational data. Among the large amount of data, We mainly focus on data concerning student mental condition such as Stress Level, Enthusiasm, Calm, etc. and their daily behavioral data, including Sleeping Hours, Working hours, Exercise, Social, etc.

After diving into the data, we can summarize the data shape as follow. There are sixty students in total who participated in this research project. The sample number of sixty seems rather small. However, for each participant, we can get data covering more than 50 different features. To sum up, the data can be categorized to nine big categories, which are application usage, calendar, call log, dinning status, education information, EMA, sensing data, sms, and survey results.

In total, there are 3.05GB data presented by the

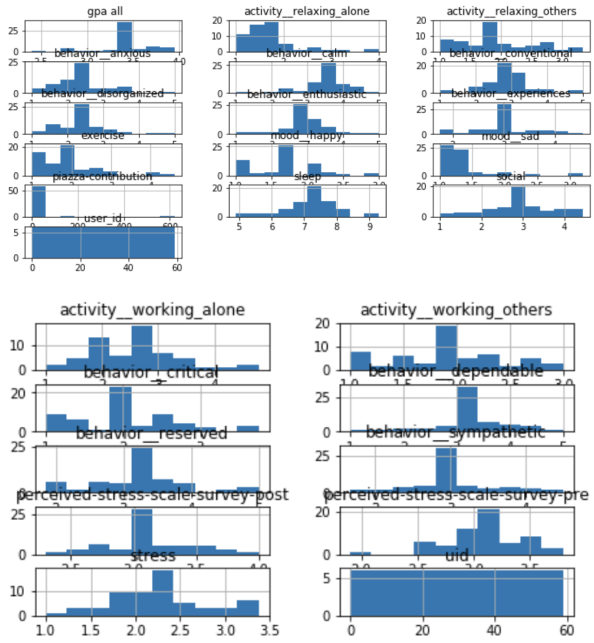


Figure 1: Histogram Figures for All Feature Data

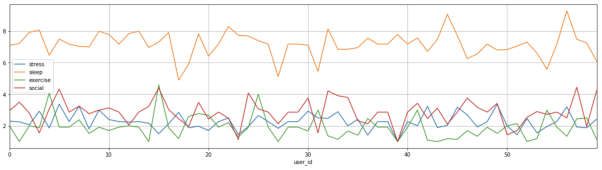


Figure 2: Trend Figure for Part of Data

Dartmouth research group. In each category, different subcategory information is stored. For example, eight different survey forms are listed in the "Survey" data set, including big five survey, loneliness survey, perceive stress survey and so on. We conclude that it is not feasible for us to analysis all the data in the data set, and decide to narrow down the scope of our project by selecting only data from EMA, Education and Survey, which we believe would include all the information that we find interesting and also are pertaining to our project goal. After consolidation, the data set has a data size of 25 columns and 60 rows. Below are histograms (Figure 1) showing the distributions of our feature variable, and the changing trend of partial feature data among different participants.

Missing values seem to be a major problem for us; otherwise, the data set is relatively clean. In order to fix that, we decide to handle the situation in a case-by-case manner. For example, we fill in the missing values in Survey data set with average values, but we fill in Education data set with zero when the majority is dominated by zero. In this way, the data set becomes much cleaner for us to conduct further analysis.

3 Related Work

There don't reveal much work on this area that have been done by using data mining methods before. But we find some papers about the connections between mental health and the physical activities. The paper "Predicting Students Physical Activity and Health-Related Well-Being: A Prospective Cross-Domain Investigation of Motivation Across School Physical Education and Exercise Settings" shows the physical education and exercise settings impact the health positively (Philip Tyson el. 2010). The another paper "Predicting Physical Activity and Healthy Nutrition Behaviors Using Social Cognitive Theory: Cross-Sectional Survey among Undergraduate Students in Chongqing, China." demonstrates how to use the social cognitive theory to predict the activities and healthy nutrition behaviors (Xu X el. 2017). The last paper "Autonomous Motivation Predicts 7-Day Physical Activity in Hong Kong Students" indicates the research of autonomous motivation predicts by seven days' physical activities in Hong Kong Students (Amy S. Ha el. 2015). All of the articles provide the theory of the system we plan to build.

Besides the social science area, some data mining related research have been found. Studies within the domain - the Educational Data Mining (EDM) have been shown a diversified orientation. Regression is one of the common practice in terms of predicting educational attributes such as the academic performances of students based on several behavioral factors. In the study by Abdous, He Yen (Abdous, He Yen, 2012), which focus on using data mining for predicting relationships between online activities of students and their final grades. They used Agglomerative Clustering based on student similarity, and classifies students based on characteristics of their online interactions and behavior. Other than common practice of data mining, deep learning methodologies have shown to be more and more common in the domain of Educational data mining. For example, Mao, Lin and Chi (Mao, Lin Chi, 2018) focus on student models for intervention, and utilized a combined approach of BKT and LSTM. Previous related works can give us some guidance on how to conduct our

own project.

4 Methodology

Since our data set contains quite a limited number of records – only 60 students were involved in the study – it would be very hard for us to use these 60 students as sample to predict other student's stress level. Therefore, we decided to switch our study focus a bit, from predicting the stress level of each individual student, to predict the overall trend of students stress level as a whole. We would utilize the time series skills to conduct further research.

First, we would summarize the stress value and average it based on the participants amount. The next step we would analyze the average stress value change trend by time. Then we would use the Granger Causality Analysis to detect the causality between stress and other features. This is because we are interested in seeing how students stress pattern in general is formed, and where would it lead to. Hopefully we would find out the most significant feature that impact stress level.

Meanwhile, we would build an Autoregressive(AR) Model to predict student stress level based on their historical stress data. In order to predict student stress in a more accurate manner, we also introduce auto-correlation function (ACF) to measure the coefficient of correlation between student stress values in a time series. The ACF of a time series is given by[7]:

$$Corr(y_t, y_{t-k})$$

The Autoregressive (AR) Model with lag length between 1 to 7 can be formulated as the form:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + ... + \beta_k y_{t-k} + \epsilon_t$$

where:
y_{t-k} is student stress level measured at time period $t - k$

In order to compare the prediction performance among different models, we also used a Moving Average (MA) model and an Integrated (ARIMA) model to compare and evaluate the prediction results. A q^{th} order Moving Average (MA) model, denoted by MA(q) is formulated as:

$$x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q}$$

where:
w_t are identically, independently distributed, each with normal distribution having mean 0 and the same variance

Source codes can be found on github: https://github.com/yuanlii/SI671_CollegeStudentStress.git

5 Evaluation Baseline

Accordingly, we made some adjustments to our evaluation metrics as well, by mainly looking at the RMSE and RSS scores of the predictive values. In order to provide some guidance and reference for later evaluation, we set two simple baselines for our prediction system.

Baseline 1: Average student stress level

The average stress level during the surveyed periods is 1.305 (student stress level ranging from 1 to 5, with 5 the most stressful). We would use student average stress as a reference value for later evaluation of our method.

Baseline 2: Naive Bayes Classification result

We also built a Naive Bayes classifier to fit our data set. It turns out the performance of Naive Bayes Classifier is not that bad, reaching 0.75 accuracy in terms of predicting and classifying student stress level per day. The result would be helpful for us to evaluate our own prediction model later.

6 Project Results

6.1 Granger Causality Analysis results

6.1.1 Stress and Sleeping hours

Before looking into the analysis, we are curious about whether the increase of stress level would have effect on students sleeping hours, or it is the other way round, that sleeping hours would influence students stress level. If we simply look at the trend plot(figure 3), we can hardly tell which one is the leading factor, but it seems that students sleeping hours and stress level demonstrate a negative correlation, especially during the midterm of surveyed period, stress level and sleep hours apparently display quite opposite directions.

However, the result of our Granger Causality analysis shows that, it is in fact students' stress level would demonstrate an significant influence on their sleeping hours with p-value = 0.0547 when lag = 5. In comparison, the p-value is only 0.182 under the hypothesis that sleeping hours would have effect on student stress level.

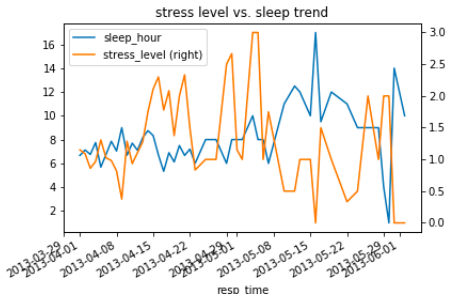


Figure 3: Time Series of student stress and sleeping hours

```
Granger Causality
number of lags (no zero) 4
ssr based F test:         F=2.4572  , p=0.0636  , df_denom=35, df_num=4
ssr based chi2 test:   chi2=12.3561 , p=0.0149  , df=4
likelihood ratio test: chi2=10.8901 , p=0.0278  , df=4
parameter F test:         F=2.4572  , p=0.0636  , df_denom=35, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=2.4501  , p=0.0547  , df_denom=32, df_num=5
ssr based chi2 test:   chi2=16.4617 , p=0.0056  , df=5
likelihood ratio test: chi2=13.9377 , p=0.0160  , df=5
parameter F test:         F=2.4501  , p=0.0547  , df_denom=32, df_num=5

Granger Causality
number of lags (no zero) 6
ssr based F test:         F=2.1921  , p=0.0727  , df_denom=29, df_num=6
ssr based chi2 test:   chi2=19.0485 , p=0.0041  , df=6
likelihood ratio test: chi2=15.7080 , p=0.0154  , df=6
parameter F test:         F=2.1921  , p=0.0727  , df_denom=29, df_num=6
```

Figure 4: Granger Causality result (student stress and sleeping hours)

6.1.2 Stress and Exercise

Our results also show that students' stress level would Granger Cause exercise. In other words, students' work out hours display an increase soon after they get stressed out. When time lag = 1, the granger-causal relationship seems to be the most significant with p-value = 0.0026. The result is interesting for it shows that students seem to exercise more as they get more stressful.

Figure 5: Time series of stress and exercise

6.1.3 Stress and Social

The plot (Figure 5) shows the interactions of student stress level and the number of people they socialize with. It seems that social display a leading influence on student stress. However, the Granger Causality results actually show that student stress in fact have a more leading influence on social number, with p-value reaching the smallest of 0.0443 when lag is 4. We may interpret that students would usually seek social support not long after they get stressed out.

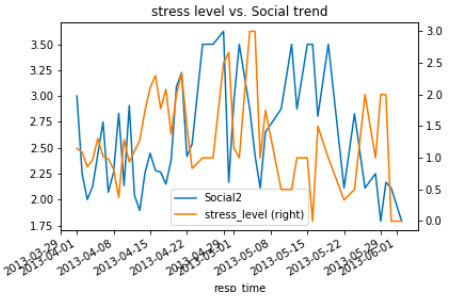


Figure 6: Time series of stress and exercise

6.1.4 Stress and Working hours

Our analysis shows that working hours would "granger cause" stress level of students, with p-value = 0.0077, which is quite significant result. We can interpret from the analysis that, students stress level tend to peak after their working hours peak. Based on this result, the best of alleviating student stress seems to be reducing their working hours.

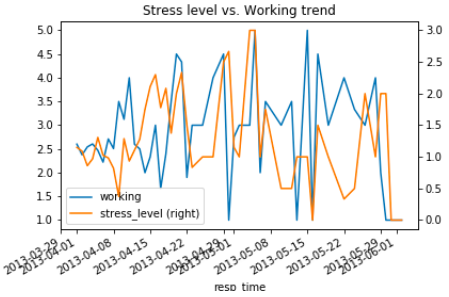


Figure 7: Time series of stress and working

```
Granger Causality
number of lags (no zero) 3
ssr based F test:         F=4.5928  , p=0.0077  , df_denom=38, df_num=3
ssr based chi2 test:   chi2=16.3165 , p=0.0010  , df=3
likelihood ratio test: chi2=13.9224 , p=0.0030  , df=3
parameter F test:         F=4.5928  , p=0.0077  , df_denom=38, df_num=3
```

Figure 8: Granger Causality result (stress and working hours)

6.1.5 Stress and Enthusiasm

With an hypothesis that student stress would have a leading influence on student enthusiasm scale, we can get the smallest p-value of 0.0114 when lag = 2. This shows that its significant that stress

level would Granger Cause enthusiasm level. It is interesting to see that students' enthusiasm level tend to increase as they get more stressed out.

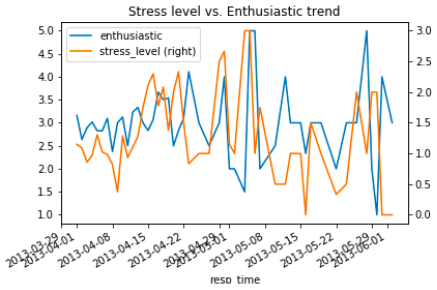


Figure 9: Time series of stress and enthusiasm

6.1.6 Stress and Calmness

The result shows that stress level would have leading influence on calmness level. With lag = 5, the p-value reaches the smallest of 0.0178, which display a significant Granger causal relationship between student stress level and calmness. It seems that not long after the students feel stressed out (with an approximate lag unit of 5), they tend to get more calm as well.

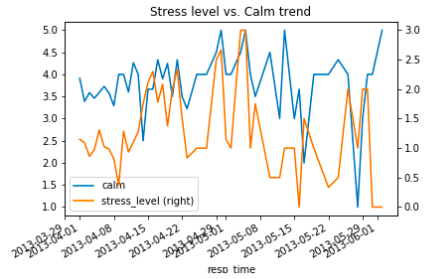


Figure 10: Time series of stress and calmness

```
Granger Causality
number of lags (no zero) 5
ssr based F test:         F=3.2358  , p=0.0178  , df_denom=32, df_num=5
ssr based chi2 test:   chi2=21.7404 , p=0.0006  , df=5
likelihood ratio test: chi2=17.5949 , p=0.0035  , df=5
parameter F test:         F=3.2358  , p=0.0178  , df_denom=32, df_num=5
```

Figure 11: Time series of stress and calmness

6.2 Prediction on student stress

6.2.1 Stability & Auto-correlation testing

We first calculated the moving average values of student stress level with a sliding window size of 3. This helps create a smoothed version of the original data (Figure 12). It seems that students stress level would peak during the midterm of the surveyed period, and it reaches another peak almost at the end of the survey period.

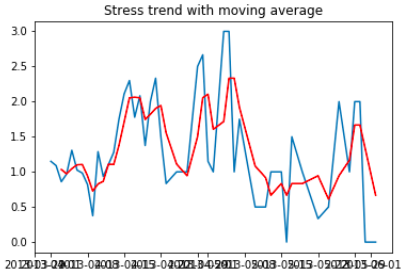
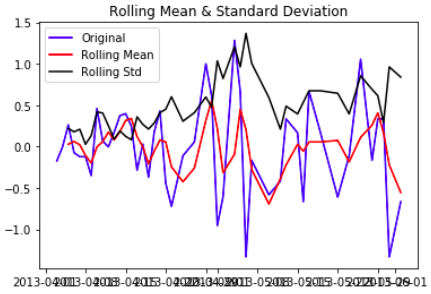


Figure 12: Moving Average plot on student stress

The rolling mean and rolling standard deviation of the time series data (Figure 13) looks much more smooth than the original stress time series data (ts_stress). However, it is hard for us to tell how stable the time series actually are, if we simply rely on visual observations. In order to test the stability of stress data, we conducted a Dickey-Fuller Test on its moving average differences, and get a small p-value of 0.024. Based on this, we can conclude that the ts_stress is mostly stable.




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.133863
p-value                          0.024147
#Lags Used                       6.000000
Number of Observations Used     39.000000
Critical Value (1%)             -3.610400
Critical Value (5%)             -2.939109
Critical Value (10%)            -2.608063
dtype: float64
```

Figure 13: Dickey-Fuller Test on moving average

In comparison to Moving Average, we also calculated the Exponential Weighted Moving Average of the original time series data. This is because Moving Average would set an undifferentiated weights to historical records. If we want to predict the next status of students stress level, we would like to assign higher weight to the most recent moving average values.

We set the size of sliding window to be 3, and compute the exponential weighted average of the

ts_stress. The small p-value from the Dickey-Fuller Test also indicates that the exponential weighted average is largely stable.

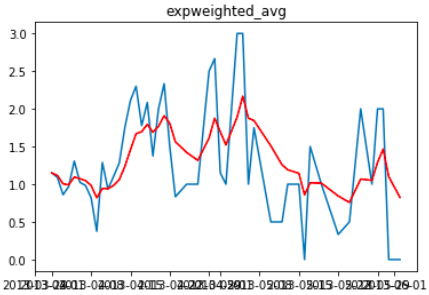


Figure 14: Exponential Weighted Moving Average of student stress

We also used Auto Correlation Function (ACF) to measure the correlation between ts_stress and itself. When lag is smaller than 7, the time series is positively correlated with itself.

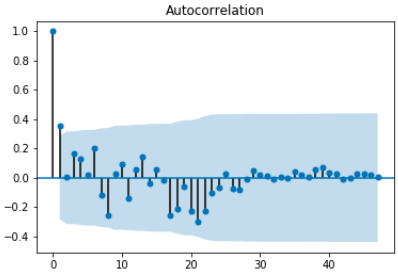


Figure 15: Auto-correlation of student stress

6.2.2 Fitted Models on student stress

The auto correlation plot (Figure 11) shows that before lag = 7, ts_stress is positively correlated with itself. So we built an Auto-regressive Model (AR) to fit the time series data, with parameter setting lag = 7. Below is the fitted values (red line) as compared to the original stress data (blue line), and the AR model would result in residual sum of squares (RSS) of 17.383. It seems that the AR model pretty much capture the trend of students stress level.

We would also like to if other models would outperform the AR model, so we built an Moving Average Model (MA) to see if the results get any better. It turned out the RSS value of MA model is higher than AR model, which reaches an RSS value of 22.210. It can also be implied from the plot that, MA model fails to capture some of the trends of ts_stress.

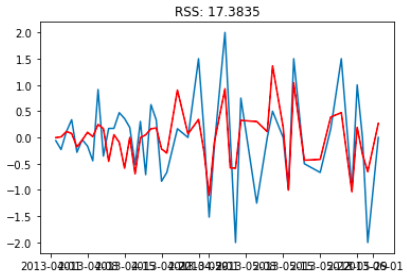


Figure 16: Auto-regressive (AR) Model

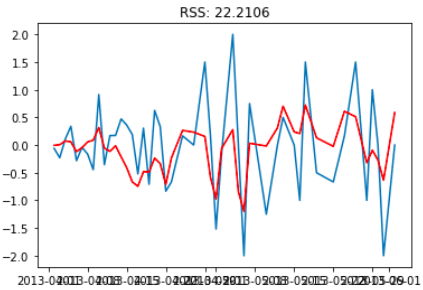


Figure 17: Moving Average (MA) Model

If we applied an integrated approach, by combining Moving Average and Auto Regressive model, it actually shows an even better result with RSS reduced to 14.129. The fitted values (red line) of model can capture most of the trend features of the ts_stress data.

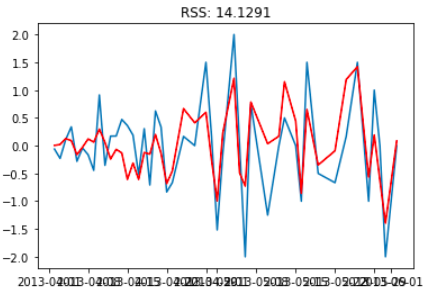


Figure 18: Integrated (ARIMA) Model

To sum up, among all the models that we have tried to fit the time series data, the performances are ranked as: ARIMA (RSS:14.129), AR (RSS:17.383), MA (RSS:22.210). An integrated method in fact outperforms each of the single method, and fit the original data pretty well.

6.2.3 Model Prediction on student stress

In order to reduce the effect of extreme values, we first log-transform our original ts_stress time series data for model prediction. It turns out

that AR model would yield the best performance among all, with parameters set as lag = 1 and sliding window size = 2. The comparison results are as follow: AR (RMSE: 0.6704), ARIMA (RMSE:1.0134) , MA (RMSE: 1.0209). Moreover, the fitted values by AR model seem to be much better than by MA model, and it shows an increasing tendency of student stress level as approaching the end of survey period, which we assume to be near the end of the semester.

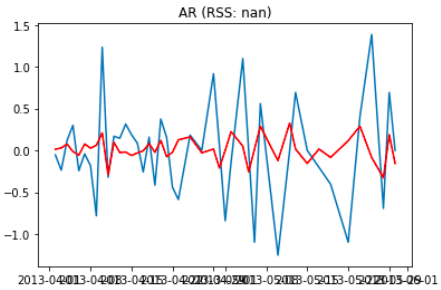


Figure 19: AR Model on student stress (lag=1)

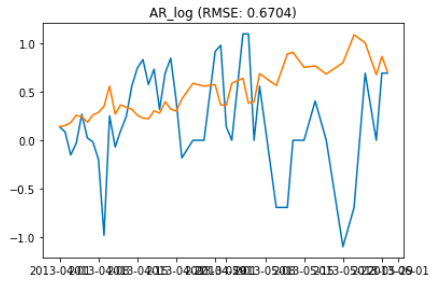


Figure 20: AR Model on student stress (log transformed)

6.3 Evaluation

Below we list the first ten records of stress level predicted by our selected model, as compared to the original student stress values. After comparing the performance of the three models, we select AR model as our final choice, as it returns the smallest RMSE score among all of 1.357. And we also list the comparison of the three models in terms of RMSE and RSS values, from which we can see that the AR model that we built outperforms the MA model and ARIMA model.

According to baseline 1, the average stress level of students is 1.305, and the average stress level predicted by our AR model is 1.89 – which is slightly higher than the baseline. However, our model predicts that among all surveyed days, there would be about 17% of days that students would feel stress, which is in fact much higher than the baseline result that 6.25% of days among all that students actually feel stress.

resp_time	actuall_stress	predict_stress
2013-04-01	1.150000	1.150000
2013-04-02	1.089286	1.163522
2013-04-03	0.859848	1.196921
2013-04-04	0.968750	1.342573
2013-04-05	1.307212	1.434915
2013-04-06	1.025000	1.291886
2013-04-07	0.982993	1.281957
2013-04-08	0.819444	1.403226
2013-04-09	0.375000	1.594809
2013-04-10	1.287500	2.319956

Figure 21: First ten predictions by AR Model

Metrics	AR	MA	ARIMA
RMSE	1.357	2.958	2.830
RSS	82.936	393.905	360.455

Table 1: Model Evaluation Results

In order to compare the results of Naive Bayes classifier, we also compute the accuracy rate of our AR model, which in fact yields an accuracy of 0.73. In other words, our prediction system successfully predicts about 73% of all surveyed days in terms of whether the students would feel stress or non-stress.

7 Discussion

In this project, we have used three different models to fit and predict student stress level. Among all models, Auto-regressive (AR) model in fact beats the Moving Average (MA) model and the Integrated model (ARIMA). However, the prediction results are in fact not quite satisfactory, and we have reflected that this can be due to several reasons. First, we used mean stress level to represent the general stress of all students, and so our system would fail to differentiate among each individual student. This can cause problem in accuracy, since in most days, student stress level are below 2.5 if we normalize the stress level of all students. This may to some extent also explain the

classification result. Although the performance of classification seems to be not that bad, this might be due to the fact in the majority of days throughout the surveyed periods, students would in fact be recognized as non-stress by taking our approach.

Another problem is that our prediction system can only predict student stress level based on their previous stress records, which fails to take into account other behavioral factors. Also, with limited records in this data set, the time series data (ts_stress) may not be stable enough for use to make accurate predictions in terms of their stress status.

8 Things left undone

Since we believe that our prediction accuracy is compromised by the limited number of participants involved in the survey, one way that we think of that could potentially improve our prediction accuracy would be to collect more data on a individual level. More student participants are desirable to train our model in order to make more precise prediction on student stress. Also, the original data set contains quite a lot of features covering almost every aspects of student life, such as GPA, assignment dues, etc. We would expect many of the features would be very useful if they can be incorporated into our prediction system. However, currently we simply use stress data from a previous period of time to predict the future, which does not considering the advantages of many useful predictive features. For the next step, we would also want to explore the possibilities of incorporating those insightful features into our prediction system as well.

9 Work Plan

In this project, we represent student stress data into a time series, and conduct Granger Causality on stress and other behavioral and spiritual data from students. However, what we initially planned to do is to build a Vector Space Model, and use similarity metrics such as Cosine Similarity, Jaccard distance, and Euclidean Distance, etc. to predict other students' stress level based on those students that we have already had stress data. The detail steps that we planned to take include the following:

- First, we would define each students as a vector, based on the measurements that would capture a students' activities and daily behavior of all kinds.
- The next step we would take similarity as a criteria to measure the relevance of one another of the students. Some of the similarity metrics that we considered include Cosine Similarity, Jaccard distance, Euclidean Distance, etc.
- We also anticipated that we would need to apply Value Decomposition (SVD) or Principal Component Analysis (PCA) to reduce the dimensions of the data set, since there are too many features in the original data set, which might cause over-fitting problem.

However, since we realize that the limited number of student participants in our dataset will be a major hinderance for us to use this method, we decided to switch our direction a bit. Since it would be too hard for us to predict stress level of each individual student, we combined student records together and use the average stress to represent the general stress level of the students. In this sense, we no longer able to predict stress per person, but instead, we can predict stress level per day.

10 Multi-person Team Justification

Originally the Student Life project provides tremendous data. The total size of the data is 3.05G and the data includes 2081 files. The categories of the data include App usage, calendar, call log, dinning, education, EMA, sensing, SMS, and survey. It is difficult for one person to analyze the whole data in the limited time. We both like this project and would like to dig out some clew between the stress and student life so we decide to work on the project together. Additionally, we discuss and analyze the data from different views. It help us to analyze the project more thoroughly.

11 Other Things We Tried

Previous baselines that we came up with: (to predict stress level of individual student)

- Baseline 1 & Evaluation:

  We built a random forest regression model on the data, with intention to evaluate the predictive effectiveness of our system. The mean stress scale predicted using the random forest regressor is roughly 3, which is quite close

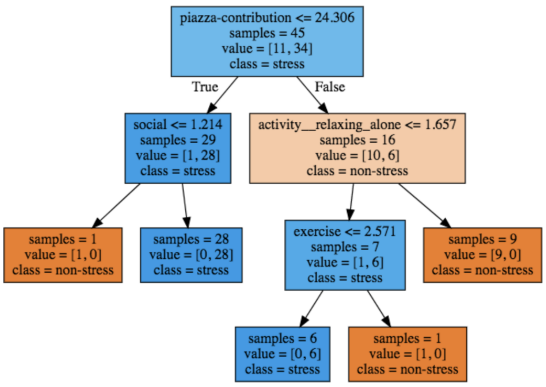


Figure 22: The Decision Tree Classifier

to but slightly smaller than the mean stress scale from original data set (3.086). However, the accuracy of the random forest regression model is pretty low (only roughly about 0 percent), To handle the missing data, I convert NA values to mean stress scale value. And if I convert the missing value to 0, and regression model would yield an higher accuracy (about 40%) yet the predicted value (about 2) is even far from 3.086.

- Baseline 2 & Evaluation:

  We also built Decision Tree Classifier (Figure 22) and Random Forest Classifier to compare our results. 1) Decision tree Classifier Result shows that only 60% of students are classified as ”stress” by the decision tree classifier, which is not perform very well based on our baseline, since its much smaller than what we would expect as the 75% of students are perceived as stress in the original data set. 2) Random Forest Classifier We built a Random Forest Classifier to see if the results are getting better compared to a single decision tree. In this case, 73% of the students are classified to be ”stress”, which is pretty close to our baseline – 75% of students are perceived as ”stress”. It’s slightly smaller than the 75% baseline though, probably because I convert the NA values to 0.

12 References

[1] Gregg Henriques (2014) *The College Student Mental Health Crisis*, Psychology Today.

[2] Amy S. Ha* and Johan Y.Y. Ng (2015) *Autonomous Motivation Predicts 7-Day Physical Activity in Hong Kong Students*, APPLIED PSYCHOLOGY: HEALTH AND WELL-BEING, 2015, 7 (2), 214229 doi:10.1111/aphw.12045

[3] Xu X, Pu Y, Sharma M, Rao Y, Cai Y, Zhao Y (2017) *Predicting Physical Activity and Healthy Nutrition Behaviors Using Social Cognitive Theory: Cross-Sectional Survey among Undergraduate Students in Chongqing, China.*, Int J Environ Res Public Health. 2017 Nov 5;14(11). pii: E1346. doi: 10.3390/ijerph14111346.

[4] Philip Tyson, Kelly Wilson, Diane Crone, Richard Brailsford Keith Laws (2010) *Physical activity and mental health in a student population*, Journal of Mental Health, 19:6, 492-499, DOI: 10.3109/09638230902968308

[5] Abdous, M., He, W., & Yen, C. J. (2012)*Using Data Mining for Predicting Relationships Between Online Question Theme and Final Grade*, Educational Technology & Society, 15(3), 77-88

[6] Ye Mao, Chen Lin, Min Chi (2018) *Deep Learning vs. Bayesian Knowledge Tracing: Student Models for Interventions*, Journal of Educational Data Mining. Page 28-54

[7] Penn State University, *Applied Time Series Analysis*, Retrieved from https://onlinecourses.science.psu.edu/stat510/node/48/