NOVEMBER 26, 2019

UNSW
AUSTRALIA

COMP9417 PROJECT

STUDENT LIFE AND MENTAL HEALTH

SHIZUKA HAYASHI
KESHUO LIN
XINYUE CHEN
YIHENG QUAN
NA LIU
UNIVERSITY OF NEW SOUTH WALES

Table of Contents

Introduction

Recent research on university students shows that there are a lot of students suffering from mental health problems. Studies also shows that mental health problems can negatively influence student's academic performance and their relationships with family and friends. (Duffy, 2019)

The aim of our project is to find the best model to predict student's mental well-beings after final exams, based on their lifestyles during a semester and investigate the relationship between their lifestyles and mental health.

Datasets

Description

- Run
  This dataset contains the relative frequency of running in all activities for each participant every week. The index of the dataset is uid and the columns are the week of semester.
- Walk
  This dataset contains the relative frequency of walking in all activities for each participant every week. The index of the dataset is uid and the columns are the week of semester.
- Noise
  This dataset contains the relative frequency that each participant being surrounded by noisy environment every week. The index of the dataset is uid and the columns are the week of semester.
- Conversation_freq
  This dataset contains the total frequency of conversation for each participant every week. The index of the dataset is uid and the columns are the week of semester.
- Conversation_time
  This dataset contains the total conversation time for each participant each week. The index of the dataset is uid and the columns are the week of semester.
- Dark_freq
  This dataset contains the frequency of each participant's phone at a dark environment for more than one hour. The frequency is calculated every week of semester. The index of the dataset is uid and the columns are the week of semester.
- Dark_time
  This dataset contains the total time that each participant's phone at a dark environment which is calculated every week. The index of the dataset is uid and the columns are the week of semester.
- Call_log
  This dataset contains the number of time that a participant receives a phone call each week. The index of the dataset is uid and the columns are the week of semester.
- SMS
  This dataset contains the number of messages received by each participant for each week. The index of the dataset is uid and the columns are the week of semester.
- Social
  This dataset contains the frequency of each participant's social activities

- Event
  This dataset contains the number of events which each participant attendant.

Binarization Method

Binarization of flourishing score

The participants are classified into 2 classes according to their post score for a flourishing scale. A total score of flourishing scores is evaluated for each participant and their total score is binarized with a threshold equal to 44. This threshold is chosen because the mean of flourishing total scores for people aged 20-29 years old is approximately equalled to 43.29. The participants with a total score higher than 44 are labelled as 'High' class and other as 'Low' class (Hone, 2013).

Binarization of Positive affect schedule score

The participants are classified into 2 classes according to their post score for a flourishing scale. The total score of positive scores is evaluated for each participant and the participants with a total score higher than 29 are labelled as 'High' class and other as 'Low' class. According to the report (Watson, 1988), the mean of PANAS positive score was 33.3 however, 'excited' is deleted from the dataset given to us, we need to subtract the average score for 'excited' from the mean to be the threshold. The average score for the missing question was found in an article and which is equal to 2.72 so that total values smaller than and equal to 29 is classified into 'low' class and total larger than 29 is classified as 'high' class. (Serafini, 2016)

Binarization of Negative affect schedule score

The participants are classified into 2 classes according to their post score for a flourishing scale. A total score of positive scores is evaluated for each participant and the participants with a total score higher than 15 are labelled as 'High' class and other as 'Low' class. Similarly, for a negative score, the average score for 'ashamed' was subtracted from the mean of the total negative score which is 17.4 given in the report. (Watson, 1988) According to the article, the average score for 'Ashamed' was 2.1. (Serafini, 2016)

Methods

Pre-Processing:

Missing values:

There are missing values in inputs datasets such as activity data with few days missing. The easiest solution to this is to ignore the missing parts however it may result in losing some important information. To fill missing values, we have used function called interpolate in pandas using linear method. Generally, the missing values occurs during semester or at the end of semester, however there are some missing values at the beginning. The missing values at the beginnings are filled by interpolating backward also using linear method.

Feature Extraction:

Feature extraction is reduction of dimensionality of high dimensional data to create useful information for statistical modellings.

There are three different ways to extract feature:

- Calculating the proportion of the whole term
- Using MapReduce to count frequency with specific situation

- Collecting data of every weeks to form time series

Description of each feature extraction

- Activity
  The proportions of each activity for each participant are recorded. The proportion is used because the frequencies of activity recorded are different for each participant and thus it is reasonable to consider the proportions instead of frequencies

- Audio
  The raw dataset of audio is state of each timestamp. We count the total frequency of 'noise' and calculate the proportion of it among other audio for each week.

- Conversation
  Conversation_time - The total duration in unit of timestamp of conversation for each participant for each week are extracted from original dataset.
  Conversation_freq – The total number of rows in original dataset for each week is equal to the frequency of conversations.
- Dark
  Dark_time - The total duration in unit of timestamp of phone in dark environment for each week are extracted from original dataset.
  Dark_freq – The total number of rows in original dataset for each week is equal to the frequency of phone in dark environment.
- Event
  Calculate the frequency of each participant for ten weeks.
- Social
  Calculate the number of social activities which each tester participates.
- SMS
  For LSTM, we extract each week the number of messages received by each participant. For KNN and Random Forest, we sum all of the number of messages within 10 weeks.
- Call_log
  For LSTM, we extract each week the number of phone calls received by each participant. For KNN and Random Forest, we sum all of the number of phone calls within 10 weeks.

Evaluation metrics

- For three of the following models, accuracy, precision, recall, f1 and ROC-AUC scores are computed to compare their performances in classification. The model with the highest score will be considered as the best model to classify participant's flourishing scores, positive scores or negative scores into 'high' and 'low' groups.
- These scores are the most commonly used evaluation metrics for representing the performance of classification. Those are calculated under the optimized condition for each of the models. There are some models with a non-consistent score among those 5 scores and we will discuss this problem in the latter part of this report.

Method 1: Long-Short Term Memory Network

Long-short term memory network is a special kind of recursive neural network that can learn the long-term dependencies. Since our data shows the trends throughout the

semester, it is reasonable to assume that predicting post scores for PANAS and flourishing score are sequence prediction problem. By using LSTM, we aimed to effectively learn the week to week patterns of data and achieve higher accuracy on classification. (Qin, 2019)

Steps

1. Fit LSTM model using data from each week as one layer
2. Each epoch, plot the loss and accuracy for train and validation set
3. Repeat for 10 weeks
4. Check if the loss is minimised and accuracy is maximised for both train and validation set

Evaluation metrics:

- Cross entropy loss is calculated after every batch training and this is minimized by adaptive moment estimation optimizer. Cross entropy loss is used here because the output from LSTM is the prediction probability of binary classes and this loss function can compute the uncertainty of our predictions based on the predicted probability and the true label.

Optimisation

- Adam optimiser is used as it is computationally efficient and requires small memory space.
- Train and validation loss and accuracy are plotted for every epoch to check whether the model is minimising loss and maximising accuracy for validation set.

Feature Importance

- Feature importance for each feature is calculated by comparing the accuracy of the model with 9 features and one feature taken away from 9 features under the same random seed. If the accuracy decreases for 8 features, then the one taken from the model is important. Similarly, if the accuracy increases, then the one taken is not important. This is repeated for 10 epochs and the average of differences in accuracy for the model with taking away one feature and model with all features are calculated. (Hooker, 2018)

Method 2: K-Nearest Neighbours with LSTM

K-Nearest Neighbours can be used to solve both regression and classification problems. In this project, K-Nearest Neighbours was used to the classification which classifies the participants into the classes described above.

Steps

1. Fit recursive feature eliminations to find the important features for KNN score prediction.
2. Find optimal number K, by using ROC and AUC scores
3. Use top five important variables to fit KNN.

Feature Selections

- Recursive feature elimination is used for selecting the top five important features for predicting each of the scores. In this project, we computed 100 iterations of RFE with Random Forest classifier and only the iteration with accuracy higher than 50% was saved for feature selection. After 100 iterations, the 5 most frequently selected

features are chosen for the final model. Feature selection is done before because the accuracy of KNN is easily influenced by using the wrong features.

Optimisation

- K-fold Cross-Validation was used together with ROC and AUC score as the dataset is too small and choosing k – neighbor only using one iteration of ROC and AUC score may give a biased solution. In k-fold cross-validation, the available datasets are separated into k small subsets and leave-one-out cross-validation was repeated k times. The average of ROC and AUC score was computed to select the k-neighbor with the largest score.

Method 3: Random Forest

Random Forest is an ensemble learning algorithm that can be used for both classification and regression. In this project, it was used as a classification to classify the participants into 'High' and 'Low' classes.

Steps

1. Fit validation curve with 3-fold cross validation for 4 parameters, n_estimators, max_depth, min_samples_split and min_samples_leaf to find the optimal parameters.
2. Use the above optimal parameters to fit Random Forest.
3. Find the importance of each features.

Optimisation

- The validation curve was created for each of the following four parameters, the number of estimators, maximum depth, the minimum number of samples required before splitting and the minimum number of samples required in each leaf.
- The optimal number for each parameter were chosen separately. Generally, the optimal number is which gives the largest accuracy score in cross-validation. However, for selecting maximum depth, the large number with a really small improvement in accuracy score is omitted to avoid overfitting.

Feature Importance

- By using the importance_ method embedded in sklearn to find the importance rate for each of the features. This method returns the importance rate which adds up to 1 and the feature with a higher importance rate is more important.

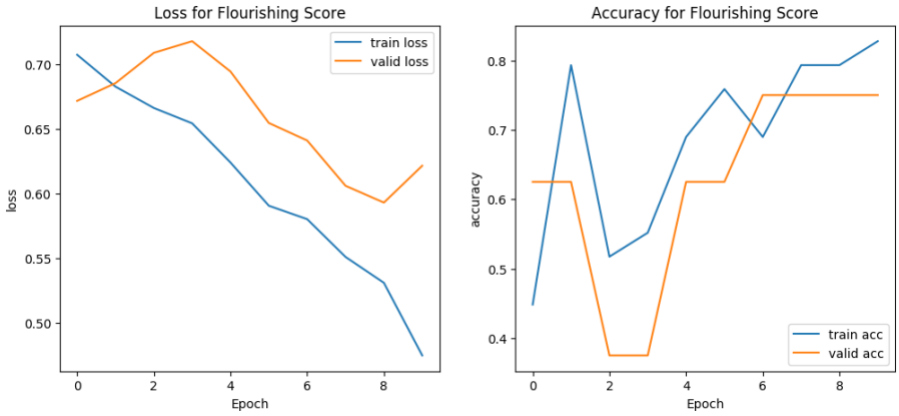
Result

Model 1: Long-Short Term Memory

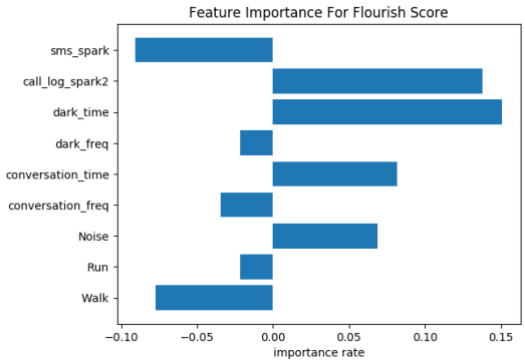
Flourishing Score

Optimisation

The left-hand side and right-hand side of the figure below show the loss and accuracy after n epochs respectively. As shown in the graph, both train loss and validation loss are converged to lower value which indicates that the LSTM neural network finds optimal numbers for hyperparameters. As the loss converges, the accuracies are also reaching 1 for both train and validation set, which also shows that the model is optimized. However, after the 8th epoch, the validation loss increased slightly, maybe indicating the overfitting of the model.



Feature Importance



From the bar plot to the left shows that 4 of the features are important for predicting flourishing scores. The features with importance rate smaller than 0 means that they decrease the accuracy rate if they are included into training samples. Thus, these features are less related to flourish scores.

The important features are ‘call_log’, ‘dark_time’, ‘conversation_time’ and ‘noise’.

Evaluation Metric

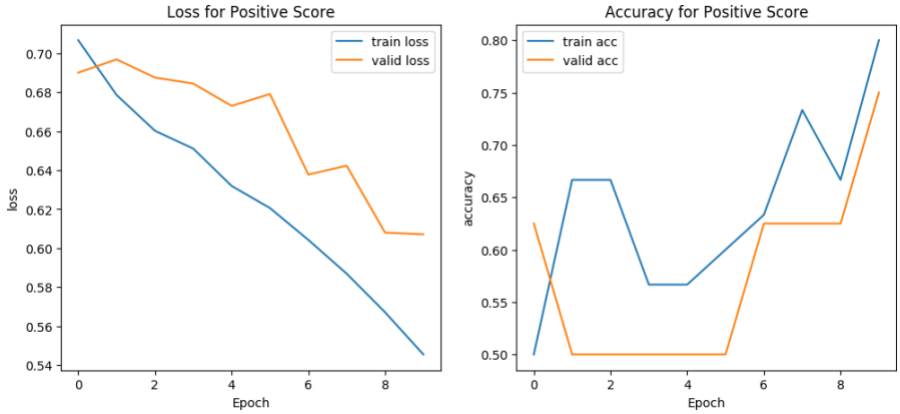
```
Accuracy: 0.7325
Precision: 0.7999999999999998
Recall: 0.7813333333333331
f1score: 0.7898181818181819
roc and auc score: 0.7123333333333334
```

Positive Score

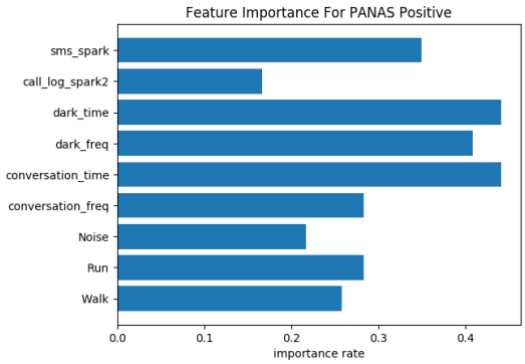
Optimisation

Similarly, the left-hand side and right-hand side of the figure below show the loss and accuracy after n epochs respectively. As shown in the graph, both train loss and validation loss are converged to lower value which indicates that the LSTM neural network finds optimal numbers for hyperparameters. As the loss converges, the accuracies are also

reaching to 1 for both train and validation set, which also shows that the model is optimised.



Feature Importance



For classifying positive score, the bar plot to the left shows that every feature is influential to positive scores. The 5 most important features are 'dark_time', 'conversation_time', 'dark_freq' and 'sms'.

Evaluation Metric

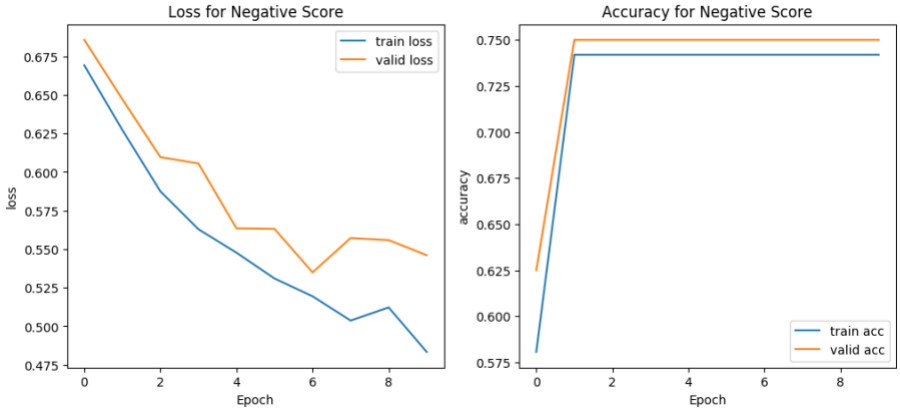
```
Accuracy: 0.8494897959183674
Precision: 0.9336734693877551
Recall: 0.8091836734693877
f1score: 0.8610139293812763
roc and auc score: 0.8717687074829931
```

Negative Score

Optimisation

Similarly, the left-hand side and right-hand side of the figure below show the loss and accuracy after n epochs respectively. As shown in the graph, the validation loss is decreasing together with training loss and both validation and training accuracies stay constant. This

shows that the model is optimised.



Feature Importance

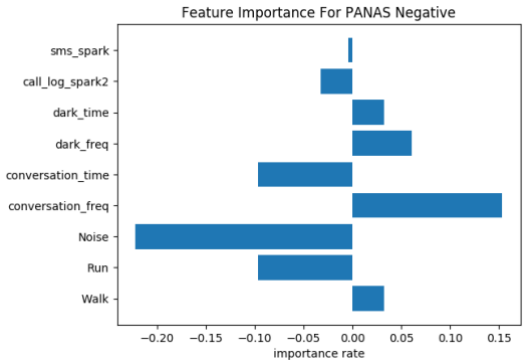


Figure to the left shows the feature importance rate for predicting negative scores using LSTM. As we can see from the bar plot, the importance rate of 'conversation_freq' is the highest and followed by 'dark_freq', 'dark_time' and 'walk'.

Evaluation Metric

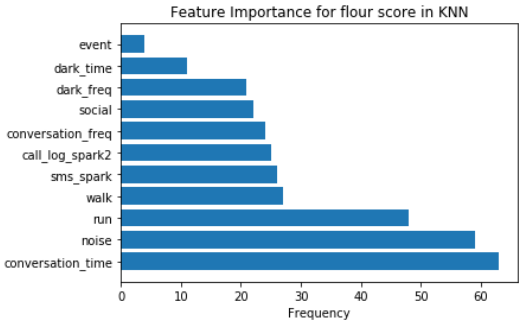
```
Accuracy: 0.75
Precision: 0.9583333333333334
Recall: 0.7708333333333334
f1score: 0.8511904761904762
roc and auc score: 0.6666666666666667
```

Model 2: K-Nearest Neighbours

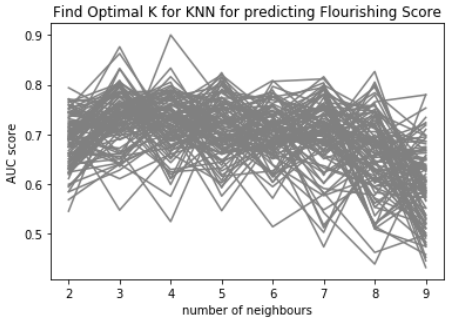
Flourishing Score

Feature Selection

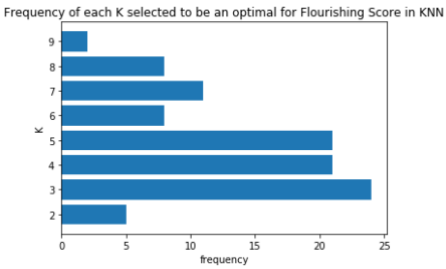
This figure shows that 'conversation_time', 'noise', 'run', 'walk' and 'sms' are the top 5 importance features. We used these 5 features for optimisation, training and validation described below.



Optimisation



In the figure to the left shows AUC score of KNN trained with features selected above. As can be seen from the figure, the maximum AUC score is around 3.



The figure to the right shows the frequency of each K selected to be an optimal in 100 iterations in bar plot. As this figure shows that 3 is the most frequently selected K, we will use 3 as optimal K for KNN to classify participant's flourishing scores

Evaluation Metric

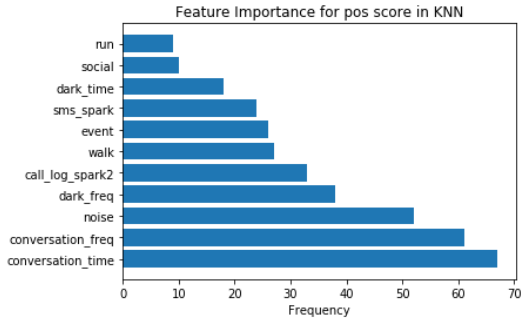
	Score for Flourishing
acc	0.656250
precision	0.735690
recall	0.656250
fscore	0.650692
roc_auc	0.674917

From precision and recall scores, we can say that the model has less type one error than type two error.

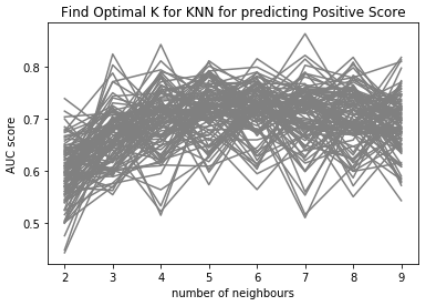
Positive Score

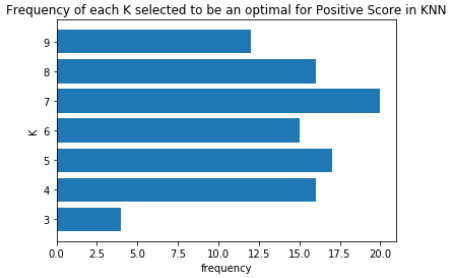
Feature Importance

As can be seen from the figure to the right, 'conversation_freq' , 'conversation_time', 'noise', 'dark_freq', 'call_log' are the 5 most important features for classifying using KNN and these are used for training the model.



Optimisation





Similar to the optimisation for flourishing score, we plotted the frequency of each K selected to be an optimal K and the figure above shows that 9 is the optimal K for classifying positive scores in KNN.

Evaluation Metric

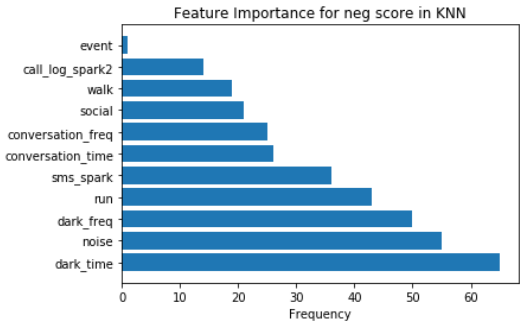
Score for PANAS Positive	
acc	0.667500
precision	0.710074
recall	0.667500
fscore	0.650755
roc_auc	0.662071

From precision and recall scores, we can say that the model has less type one error than type two error.

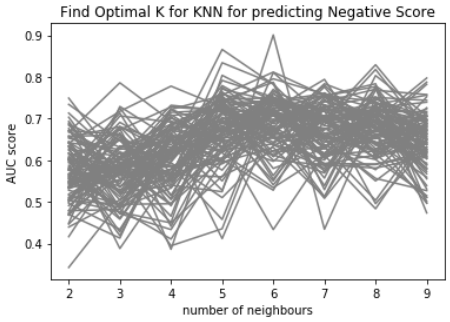
Negative Score

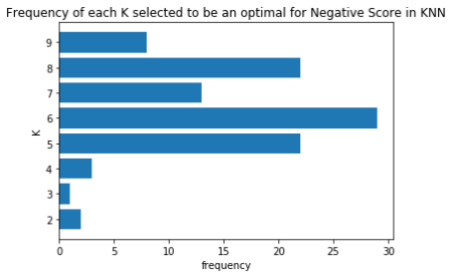
Feature Importance

Similar to predicting positive score, we selected the 5 most important features, 'dark_time', 'dark_freq' , 'noise', 'run' and 'sms' for training the model



Optimisation





Similar to the optimisation for flourishing score and positive score, we can see that 2 is the optimal K for classifying negative score in KNN.

Evaluation Metric

Score for PANAS Negative	
acc	0.690000
precision	0.664357
recall	0.701250
fscore	0.662478
roc_auc	0.571393

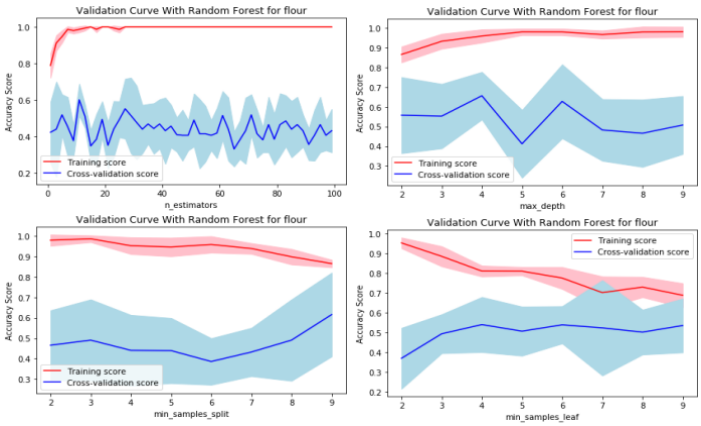
From precision and recall scores, we can say that the model has less type two error than type one error.

Model 3: Random Forest

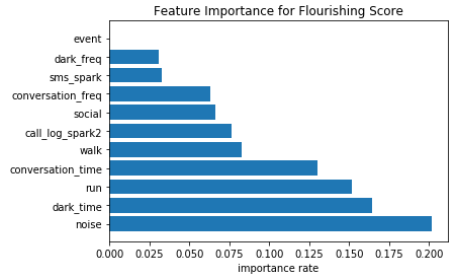
Flourishing Score

Optimisation

From the figure to the right, the number with highest cross-validation score are selected to be optimal parameter for random forest. N-estimators is the number of trees in random forest, and it is chosen to be 17, maximum depth is 4, minimum samples split is 9, and minimum samples in leaf node is 9.



Feature Importance



The features with higher importance rate are more important than others.

Evaluation Metric

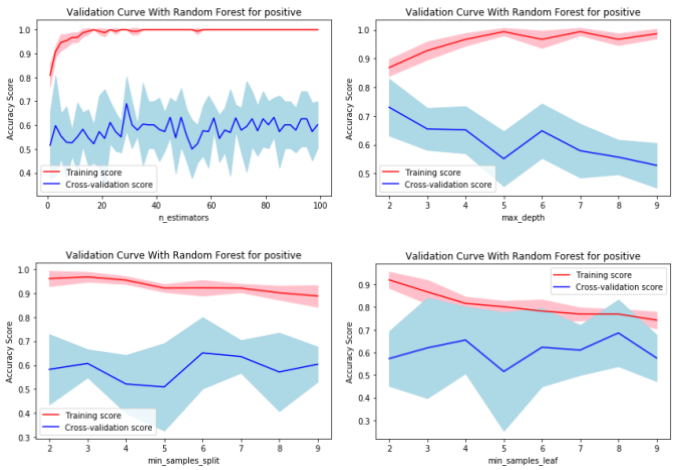
Scores for Flourishing	
accuracy	0.500000
precision	0.833333
recall	0.500000
fscore	0.500000
auc	0.666667

The model has high precision rate with low recall rate which indicates that the model have high type two error.

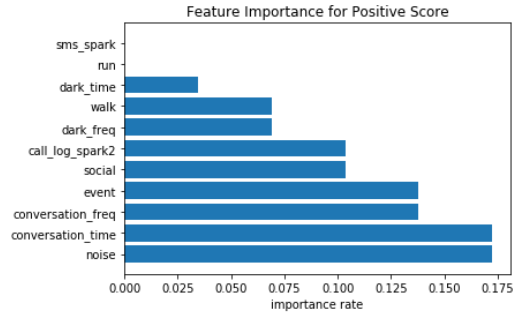
Positive Score

Optimisation

Similar to parameter selection for flourishing score, the parameters with highest cross validation score are chosen. N-estimators is 30, maximum depth is 2, minimum samples split is 6, and minimum samples in leaf node is 8.



Feature Importance



The higher the importance rate indicates the feature is more important.

Evaluation Metric

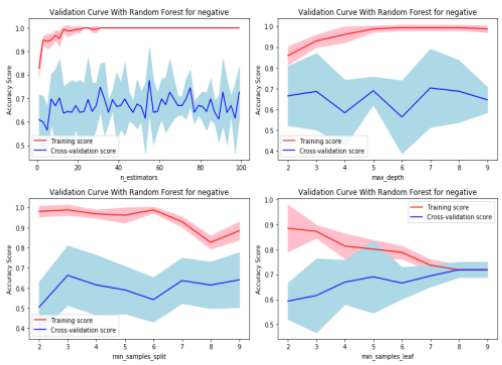
	Scores for Positive
accuracy	0.625000
precision	0.656250
recall	0.625000
fscore	0.630952
auc	0.633333

All the scores are similar, which means that the model have similar number of type one and type two error.

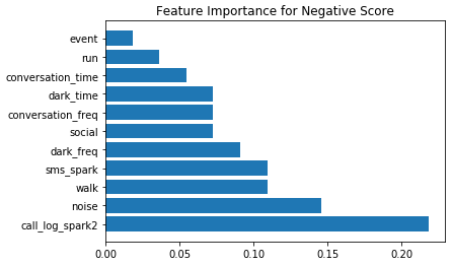
Negative Score

Optimisation

Similar to parameter selection for flourishing score, the parameters with highest cross validation score are chosen. N-estimators is 58, maximum depth is 5, minimum samples split is 3, and minimum samples in leaf node is 9.



Feature Importance



The features with higher importance rate are more important than others.

Evaluation Metric

	Scores for Negative
accuracy	0.875000
precision	0.765625
recall	0.875000
fscore	0.816667
auc	0.500000

As you can see from above figure, the precision and all scores except for AUC scores are high. This indicate that the data is imbalanced. There only 11 sample with negative score higher than the threshold but there are 28 below the threshold.

Discussion

Comparison

Model Performances

From the evaluation metrics, the model using LSTM is considered to be the best model among the 3 methods, to predict participant's emotions. LSTM is fed with panel data rather than section data to achieve neural network with time series. Section data, the data containing information about the whole semester with no time series are used for KNN and Random Forest. As a result, LSTM got the highest scores for every evaluation metrics.
To compare KNN and Random Forest, we need to compare the evaluation metrics for flourishing, positive and negative scores.
For the flourishing score, KNN has all the scores higher than random forest other than the precision score. This indicates that random forest achieves lower type one error but has a higher type two error than KNN.
For Positive score, KNN has achieved better performance than Random forest as KNN has higher values for every evaluation metric.
For the negative score, Random forest has higher accuracy, precision and recall rate than KNN however it has a lower AUC score which was caused by the imbalanced dataset.

Trade-off Between 3 methods

Although LSTM got the highest accuracy for the prediction, there are some drawbacks to the model. LSTM is much more difficult to implement and also difficult to find the optimal number for hidden dimension and output dimension. In addition, unlike KNN and random forest, LSTM is hard to calculate feature importance as LSTM takes time dimension into account.
The advantage of KNN is the ease of implementation however it cannot deal with a large dataset because it is computationally expensive. KNN is also easily influenced by non-informative features and noisy data. Random forest, on the other hand, will not be largely influenced by non-informative features and missing values. It is computationally cheap and able to consider many more features and situations than KNN and LSTM. However, KNN and random forest are not good at solving the problem with time series. (Arora, 2018)

Discovery

For this project, LSTM has a high accuracy of predicting student's mental health at the end of the semester by using their lifestyles during the semester. This indicates that student's flourishing scores and PANAS scores are changing according to their changing in lifestyle during semesters. Their lifestyles can be affected by many factors such as the amount of assignments, mid-semester breaks, and final exams. Features importance are slightly different for each model however as LSTM gives the highest scores, it is reasonable for us to consider the feature importance given by the LSTM model.
By looking at the feature importance for the LSTM model, call log, dark time, conversation time and noise are important features for the flourishing score.
The figures below show the correlation between these features with the flourishing score. Dark_time is negatively correlated with a flourishing score indicates that the longer a phone in a dark environment, the lower the flourishing score. The features other than dark_time are positively correlated with the flourishing score. This reveals that a person with a longer

time in a noisy environment and longer time having a conversation with others, as well as more frequently receiving phone calls will have a higher flourishing score.

	dark_time	flourishing score
dark_time	1.0000	-0.0561
flourishing score	-0.0561	1.0000

	noise	flourishing score
noise	1.000000	0.388669
flourishing score	0.388669	1.000000

	conversation_time	flourishing score
conversation_time	1.000000	0.268085
flourishing score	0.268085	1.000000

	call_log	flourishing score
call_log	1.000000	0.006472
flourishing score	0.006472	1.000000

On the other hand, all features used in the model for positive score are important, so we choose top 5 important features to see their correlation with positive score. As can be seen from the figure below, all features except for frequency of receiving SMS are positively correlated with PANAS positive scores.

	dark_freq	positive score
dark_freq	1.00000	0.15477
positive score	0.15477	1.00000

	conversation_time	positive score
conversation_time	1.0000	0.4953
positive score	0.4953	1.0000

	sms	positive score
sms	1.000000	-0.103163
positive score	-0.103163	1.000000

	dark_time	positive score
dark_time	1.000000	0.204682
positive score	0.204682	1.000000

	conversation_freq	positive score
conversation_freq	1.000000	0.423779
positive score	0.423779	1.000000

For negative score, dark time, dark frequency, conversation time and walk are the important features. From the correlation matrix below, it can be seen that conversation time and walk are negatively correlated with PANAS negative scores and dark frequency and time are positively correlated.

	dark_freq	negative score
dark_freq	1.000000	0.225181
negative score	0.225181	1.000000

	conversation_time	negative score
conversation_time	1.000000	-0.071896
negative score	-0.071896	1.000000

	dark_time	negative score
dark_time	1.000000	0.190943
negative score	0.190943	1.000000

	walk	negative score
walk	1.000000	-0.149681
negative score	-0.149681	1.000000

Conclusion

Overall, our project has successfully demonstrated the best model for predicting the participant's mental health and has identified the relationship between lifestyle and mental well-being. LSTM model with missing values filled by a linear method is our best model for the prediction. The result shows that the flourishing score, which is the summary of self-perceived success, is highly positively correlated with a person's conversation time with others and the frequency of receiving calls. A person who favors a noisy environment is also likely to have high flourishing scores. In contrast, a person staying in a dark environment for a long time tend to have a lower flourishing score. Similarly, the analysis for PANAS positive score also shows a person having a lot of conversation with others have high positive scores.

Besides, the amount of walking and conversation are negatively correlated with a negative score. This indicates that a person having more social activities will have fewer negative emotions. Both dark frequency and time have a positive correlation between positive and negative score, but they are negatively correlated with the flourishing score. This means the self-confidence rate can be negatively affected by staying in a dark environment for longer hours, but their mood cannot be distinguished by them. Further investigation can be done in the future by gathering more student to participates the program to get more samples for analysis so that will allow us to get a more precise picture of the problem that students with mental illness have.

Bibliography

Arora, A., 2018. Why Random Forests can't predict trends and how to overcome this problem?. *[online] Medium. Available at: https://medium.com/datadriveninvestor/why-wont-time-series-data-and-random-forests-work-very-well-together-3c9f7b271631 .*

Duffy, A., 2019. *University student mental health care is at the tipping point.* [Online] Available at: http://www.citethisforme.com/cite/sources/websiteautociteeval

Hone, L., 2013. Psychometric Properties of the Flourishing Scale in a New Zealand Sample. *Social Indicators Research,* 119(2), pp. 1031-1045.

Hooker, S. E. D. K. P.-J. &. K. B., 2018. A Benchmark for Interpretability Methods in Deep Neural Networks.

Qin, Z. C. C. a. G. X., 2019. Prediction of Air Quality Based on KNN-LSTM.. *Journal of Physics: Conference Series,* p. 1237.

Serafini, K. M.-M. B. N. C. H. K. a. C. K., 2016. Psychometric properties of the Positive and Negative Affect Schedule (PANAS) in a heterogeneous sample of substance users.. *The American Journal of Drug and Alcohol Abuse.*

Watson, D. C. L. A. &. T. A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology,* p. 1063.