Machine Learning and Data Mining

COMP9417

ASSIGNMENT - 1

Group Name: Runtime tERRORs

By

Mustafa Juzar Merchant

Shubhankar Mathur

Chirag Panikkasseril Unni

Bharath Naga Chandra Surampudi

Kovid Sharma

Contents

Introduction

The Project is provided with StudentLife dataset which is collected from 48 Dartmouth university students during a period of 10 weeks using mobile apps on nexus devices. The data collected includes physical activity, audio activity, conversation start/end time, GPS location, Bluetooth data, WiFi, WiFi location, light start/end time, phone lock start/end time, phone charge start/end time.

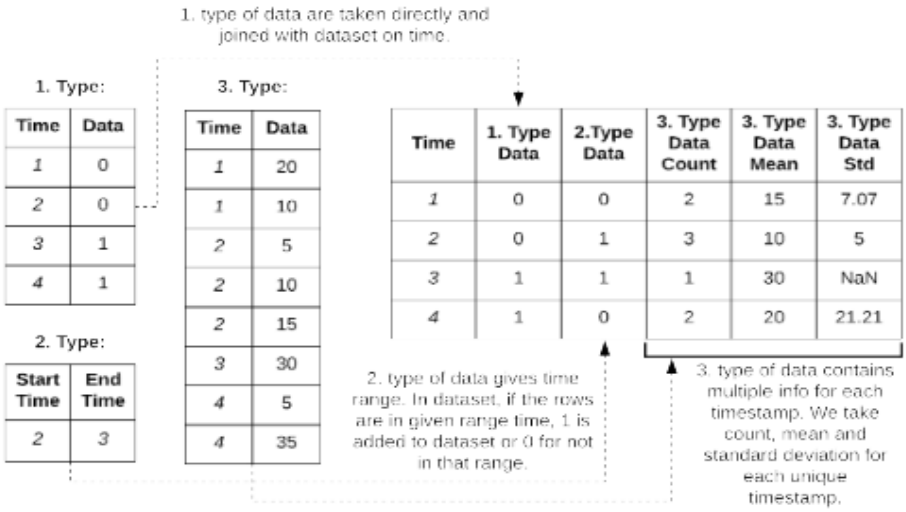
The aim of the project is to estimate flourishing scale and PANAS, two psychology-related phenomena from the collected auto sensing data. This is done by applying data cleansing and engineering new feature to build classification and regression models. We use various techniques like cross validation such as kfold to do hyper parameter tuning, to build the best model. Once the models are build, we use testing data to evaluate the models with necessary metrics.

Description of Dataset

Data cleaning and preprocessing

Data is sorted and grouped to the users participated in the research. Further the data is split according to the time with the following divisions. The preprocessing techniques on different types of data is shown below:

Type 1 data: Activity and audio data
Type 2 data: conversation, light sensor, phone charge and phone lock data
Type 3 data: Bluetooth and Wi-Fi data



Time division:

Each Users data is further split into dates in the time range given and for each date the data is further split into 3 time slots: day, evening and night.

- Day: hour >= 6:00 **and** hour < 18:00
- Evening: hour >= 18:00 **and** hour <= 23:59 (mid-night)
- Night: 0:00 to 6:00

Engineered Features:

To add new features from Bluetooth and Wi-Fi data, we did some statistical calculations to represent multiple data. Bluetooth and Wi-Fi have multiple values (signal levels) for each timestamp. Consequently, mean and standard deviation and total count were calculated for each timestamp to understand the count of nearby devices and their signal levels. After multiple data was converted to single timestamp representation by these calculations, they were added as new features.

New features have been engineered out of existing data for the preparation of the model. We applied the following on each feature:

Activity:
We have taken mode on activity inference based on the day, night and evening groups of that specific timestamp and created activity inference_mode.

Bluetooth:
We took the mode on MAC and created MAC_mode, extracted mean and standard deviation on bt_level and created bt_level_mean and bt_level_std respectively.

Conversation:
We took the difference in timestamp to produce duration of the conversation. Further calculated the average and count on the derived data thus producing duration(s)_avg_conversation and frequency_conversation respectively.

Dark:
We took the difference in timestamp to produce duration of the Dark. Further, we calculated the average and count on the derived data thus producing duration(s)_avg_dark and frequency_dark respectively

PhoneLock:
We took the difference in timestamp to produce duration of the PhoneLock. Further calculated the average and count on the derived data thus producing duration(s)_avg_PhoneLock and frequency_PhoneLock respectively.

Gps:
We applied mode on travelstate to get the most repeated activity on that user in the particular time frame. The we created feature travelstate_mode and further applied binarization where stationary =0 and moving=1.

Wifi:
We took mode on BSSID and created BSSID_mode and applied mean and standard deviation on wifi_level and created wifi_level_mean and wifi_level_std respectively.

Binarization Method:

For the Outputs in Flourishing scale and PANAS, each user's data is summed up for pre and post data, and further average of this is taken as the y for model computation. For classification the output is divided on the basis of median value which creates two balanced classes "high" and "low".

We used the scoring pattern mentioned here.
Positive Affect Score: Add the scores on items 1, 3, 5, 9, 10, 12, 14, 16, 17, and 19. Scores can range from 10 – 50, with higher scores representing higher levels of positive affect. Mean Scores: 33.3 (SD±7.2)

Negative Affect Score: Add the scores on items 2, 4, 6, 7, 8, 11, 13, 15, 18, and 20. Scores can range from 10 – 50, with lower scores representing lower levels of negative affect. Mean Score: 17.4 (SD ± 6.2)

The derived data is stored in positive_avg, negative_avg,,fs_avg for each user in uid_y.

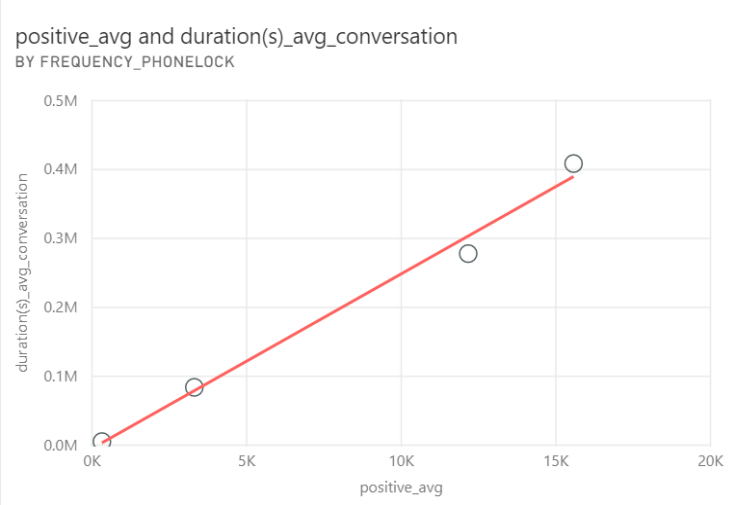
Normalization

Preprocessing is done by applying Normalization Min-Max scaler on the numerical data. We normalized the y-value by using ln(y+1) to reduce the gap between the features and the output but noticed overfitting during our evaluation, so we discarded it.

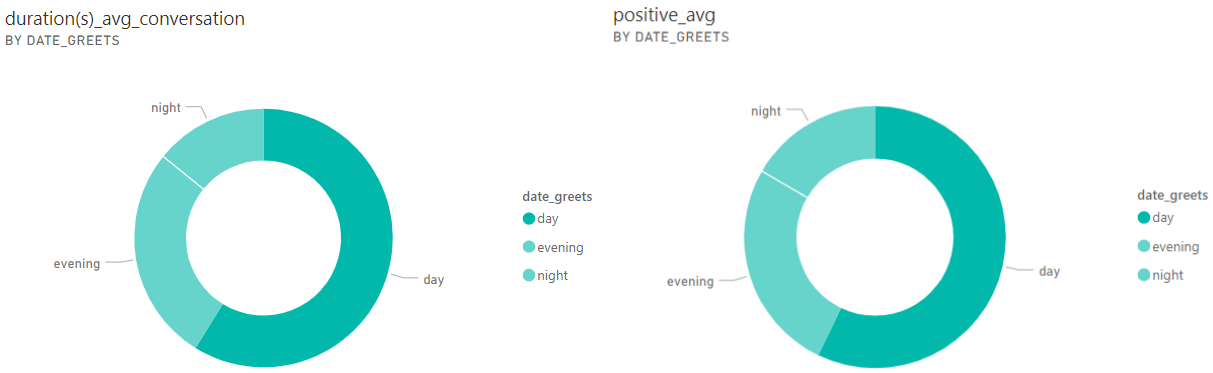
Feature Selection

We used linear regression on the full data set with the help of statsmodels.api.
Ordinary Least Square regression technique is applied from the function and we calculated the positive and negative impacts of each feature on the target y values separately. We have taken into account the corresponding coefficients and the p value which were calculated by the OLS model and the threshold was determined. We considered the threshold value 0.05 for p and picked the following features for each of the target y data as given below. The features which were under the threshold were dropped.

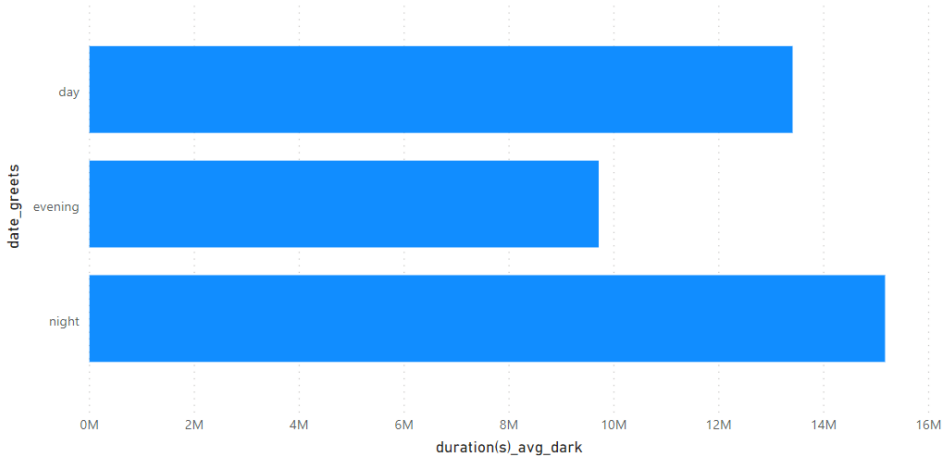
Visualizations



Duration Average Conversation and frequency conversation is directly proportional to the positive average panas indicating students with more social or non-social interactions have a positive impact on their positive average panas.



It can be seen from the above graph that students tend to have the highest amount (duration) of conversation during the day time which is also seen to have the high impact on the positive panas score.



From the above graph we observe that students tend to stay in dark even in the day time which resonates signs of depression and can impact the panas and flourishing scale scores.

Flourishing scale features:

- Wifi_level_std, wifi_level_mean : These features show strong positive and negative impact on target value.This can be deduced by saying that people studying in groups can impact flourishing scale.
- duration(s)_avg_conversation : Shows strong positive correlation with target value. People having longer conversations tend to flourish.
- BSSID_mode, MAC_mode, date_greets : Shows significant impact on the target value.

	Coefficients	p
const	40.2167	0
wifi_level_std	6.98813	2.00721e-09
wifi_level_mean	-11.6759	2.85571e-07
duration(s)_avg_conversation	9.34204	2.90039e-06
BSSID_mode	18.9525	3.3925e-06
MAC_mode	6.31611	3.43489e-06
date_greets	0.559079	0.00105021
frequency_dark	-2.33041	0.00302441
activity inference_mode	-1.22494	0.0257275
frequency_phoneLock	1.75493	0.0881894
duration(s)_avg_phoneLock	1.13835	0.0988892
travelstate_mode	-0.710631	0.110738
bt_level_std	2.82483	0.220556
bt_level_mean	2.82483	0.220556
duration(s)_avg_dark	0.398197	0.537912
frequency_conversation	-0.517981	0.694563

PANAS (Positive):

- frequency_phoneLock, MAC_mode, bt_level_mean, bt_level_std, activity inference_mode, travelstate_mode, wifi_level_mean, Date_greets: Shows strong positive correlation since people having frequent phone locks tend to be positive in life.

	Coefficients	p
const	28.7319	0
frequency_phoneLock	3.59525	3.65999e-05
MAC_mode	3.93276	0.000619263
bt_level_mean	5.69658	0.00348709
bt_level_std	5.69658	0.00348709
activity inference_mode	1.07636	0.0204178
travelstate_mode	0.854531	0.0232952
wifi_level_mean	-3.69707	0.0540621
date_greets	0.254781	0.0770996
BSSID_mode	4.41682	0.199522
frequency_dark	0.803873	0.225945
frequency_conversation	0.729909	0.512701
duration(s)_avg_dark	0.319799	0.558353
duration(s)_avg_phoneLock	-0.249513	0.668622
duration(s)_avg_conversation	0.659169	0.695664
wifi_level_std	0.215913	0.82598

PANAS (Negative):

- activity inference_mode, travelstate_mode, wifi_level_mean, bt_level_std, Bt_level_mean: shows positive correlation since people indulging in activities like being stationary can affect negativity.
- wifi_level_std, frequency_conversation, duration(s)_avg_phoneLock: These features show negative correlation since people tend to stop being social when they feel negative.

	Coefficients	p
const	24.0806	0
activity inference_mode	4.56055	3.15845e-20
wifi_level_std	-9.11724	3.0474e-18
frequency_conversation	-5.63562	1.9143e-06
duration(s)_avg_phoneLock	-2.02149	0.00107363
travelstate_mode	1.21204	0.00239686
wifi_level_mean	5.71656	0.00494699
bt_level_std	4.75459	0.0213415
bt_level_mean	4.75459	0.0213415
duration(s)_avg_dark	0.940244	0.10439
frequency_dark	1.05241	0.134607
MAC_mode	1.56971	0.196837
duration(s)_avg_conversation	-1.84543	0.301316
frequency_phoneLock	0.949333	0.302946
date_greets	-0.0950141	0.533691
BSSID_mode	-0.90641	0.803723

Model Selection

Regression models:

We applied multiple regression techniques such as Bayesian Linear Regression, Neural Network, Regression, Boosted Decision Tree Regression, Linear Regression, and Decision Forest Regression, in the data and determined the rmse values for each model. This is done to determine the best model. We concluded that Linear regression with least rmse score of 6.760949 is the best outcome from the applied regression techniques.

Algorithm	Root Mean Squared Error
Bayesian Linear Regression	6.810698
Neural Network Regression	6.766953
Boosted Decision Tree Regression	6.951471
Linear Regression	6.760949
Decision Forest Regression	7.069191

Classification Models

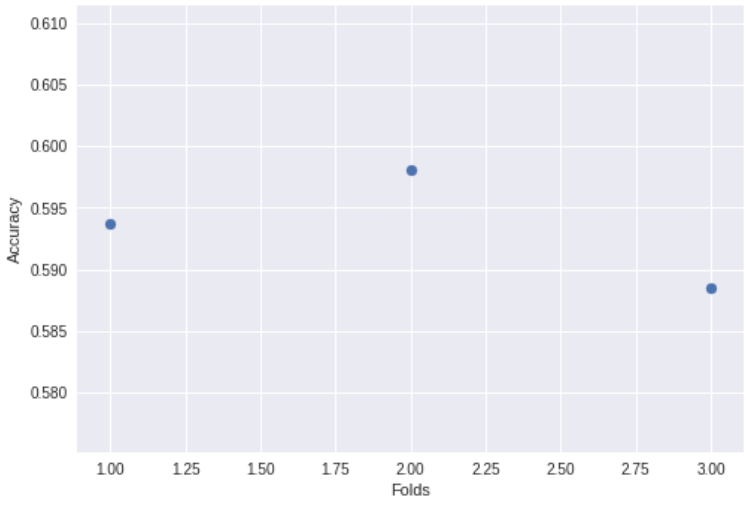
We determined cross validation scores for the above classifiers and noticed that Decision Tree, AdaBoost and Neural Net have relatively high cross validation scores. Since Ada boost can make a weak classifier better, we selected AdaBoost with Random Forest and Neural Net Classifiers to build the models.

Classifiers	Cross validation scores
Nearest Neighbors	0.524535
Linear SVM	0.520904
RBF SVM	0.530284

Gaussian Process	0.508582
Decision Tree	0.556748
Random Forest	0.549504
Neural Net	0.565809
AdaBoost	0.567925
Naive Bayes	0.527378
QDA	0.549083

Cross Validation:

We divided the dataset in three parts i.e Train set, Validation set and Test set. The train and test set were splitted on size 0.8 to 0.2 ratio respectively and the train set was used for 3-Fold Cross Validation. We chose 3 Folds because we wanted enough instances for our training model to adapt to the data which might help optimizing the model. Later the cross validation score was evaluated as a mean of all scores of different folds.Later the cross validation predictions was exposed to 20% unseen data set that we splitted and result metrics were calculated for each model. The graph Accuracy v/s Folds is shown for Random Forest Classifier to predict PANAS positive scale.



Justification for Model Selection

Linear Regression

Now we fit the linear regression model using default hyperparameters .We performed cross validation on linear regression model with all possible hyperparameters and models are fit with training data to get tuned hyperparameters .After computing the tuned hyper parameters we evaluated the cross val predictions of the model against evaluation metrics r2 score, RMSE and cross validation score on testing data.

Neural Network

Since we have high cross validation scores for Neural Network, we took the classifier to build the model. We applied GridSearchCV on hyperparameters for the classifier to get the optimal tune hyperparameters and we fitted the model using training set. Further test data was used to evaluate and produce following metrics mentioned in the results.

Adaboost

Boosting model's key is learning from the previous mistakes, e.g. misclassification data points. We used Random Forest classifier and got an accuracy of 60.39 % and then we applied Ada Boost classifier and got higher accuracy around 60.65%. we evaluated the model against the testing data and got the following evaluation metrics mentioned in the results

Results

Features selected

	Flourishing Scale	PANAS Negative	PANAS Positive
Selected and Engineered features used for modelling	Wifi_level_std, wifi_level_mean. duration(s)_avg _conversation, BSSID_mode, MAC_mode, date_greets	activity inference_mode, wifi_level_std, frequency_conversation, duration(s)_avg_phoneLock, travelstate_mode, wifi_level_mean, bt_level_std, Bt_level_mean	frequency_phoneLock, MAC_mode, bt_level_mean, bt_level_std, activity inference_mode, travelstate_mode, wifi_level_mean, Date_greets

Hyperparameters

	Boosted Random Forest	MLP Classifier	Linear Regression
Best Grid Parameters	bootstrap=True, max_depth=80, max_feature=3, min_sampled_leaf = 5, min_sample_split =8, and n_estimator=10	activation=tanh, learning_rate=invscaling, shuffle=True, and solver=lbfgs	copy_X=False, fit_intercept=True, normalize=False

Score metrics

Accuracy	Random Forest classifier	Boosted Random Forest classifier	Neural Network	Linear Regression (mean score deviation)
Flourishing scale	60.39%	60.65%	59.31%	2.44
Panas Positive	59.49%	62.59%	62.40%	2.52
Panas Negative	59.49%	60.61%	53.16%	2.56

Panas negative		
	MLP	Adaboost
Precision	0.534421	0.660299
Roc_Auc	0.543663	0.656277
F1	0.024051	0.598501

Panas positive		
	MLP	Adaboost
Precision	0.606905	0.628581
Roc_Auc	0.644753	0.672948
F1	0.488489	0.540999

Flourishing Scale		
	MLP	Adaboost
Precision	0.603568	0.6690518
Roc_Auc	0.585646	0.5899246
F1	0.710269	0.6404109

Linear Regression	
	RMSE
Flourishing Scale	6.886498
Panas Positive	5.57189
Panas Negative	6.055211

Discussion

Using continuous passive sensing data of college students, we are able to decide with 60.65% accuracy on flourishing scale and 60.61% accuracy on PANAS scale by looking at their mobile sensing data. To train this model, we used 80%training and validation instances and 20% test instances, and all of them are balanced dataset. We aimed to test boosted Random forest, MLP Classifier and Linear Regressor because continuous sensor data is time-varying and we thought that these algorithms are suitable for this dataset. We can feed the timestamp data as a series to discover correlations and dependencies between each time's feature values. In this way, it can disclose the changing behavior of users over time. The features of the input data are activity inference, Bluetooth, conversation, phone charging, phone in the dark and phone locked. Activity inference and audio inference are categorical and they are one Label Encoded to feed into the model.

We trained our model using cross validation and tuned hyper parameters to reach our best generalizable performance. In our tests, we tried to process the labels using ln(n+1) to make the features and target value very close but we can see there is clear overfitting of data. So, we tried to apply different method to reduce overfitting but none

of them worked. We think that this problem occurs because of the data that we have. We need more information and more features over expanded timestamp values to solve this problem. We trained 3 different models and as it is seen in the Score Metrics Accuracy table, all models perform equally to each other but Boosted Random Forest Classifier outperforms the other models. The reason for this may be about the dataset and characteristics of Boosting. In general, MLPClassifier is better suited for spatial data and Boosting algorithms for sequential data. Linear Regressor has considerable performance for all three predicted values. It tries to reduce the error rate to 2.43 which is not bad. Considering the lack of information and features we can further improve this regressor to predict the required values. In conclusion, our main focus was to predict three values for this assignment i.e Flourishing scale, PANAS positive and PANAS negative.

Conclusion

The results produced conclude that by using Linear regression, Adaboost and Neural Net we were able to predict the PANAS and flourishing scale from the data using above machine learning models. Out of which Adaboost model has higher accuracy than other models. We can further improve and optimize the models once we gather more features and information from the students and reduce the overfitting problem.

Reference

1. Depression. http://www.nimh.nih.gov/health/topics/depression/index.shtml

2. funf-open-sensing-framework. https://code.google.com/p/funf-open-sensing-framework/

3. StudentLife Dataset 2014. http://studentlife.cs.dartmouth.edu/

4. McLeod, S. A. (2019, May 20). What a p-value Tells You About Statistical significance. Simply Psychology. https://www.simplypsychology.org/p-value.html

5.YasinAcikmese,S.EmreAlptekin (2019). *Prediction of stress levels with LSTM and passive mobile sensors.*