CS5228 Project Report

Group 17 - Workload Distribution

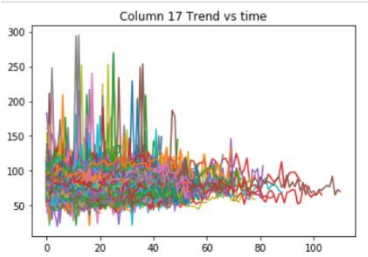
Chan Jun Wei (e0374282) : Model Training, Model Tuning, Feature Extraction, Feature Selection
Chua Chin Siang (e0403439) : Data Preprocessing, Feature Extraction, Model Training
Tey Shi Wei (e0403430) : Data Exploration, Data Preprocessing, Model Training
Wang Zihao (e0404053) **: MIA**

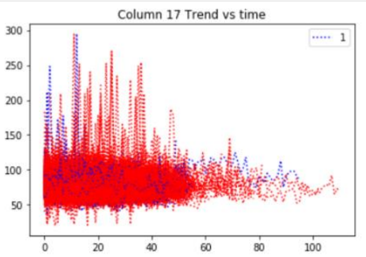
Key Words

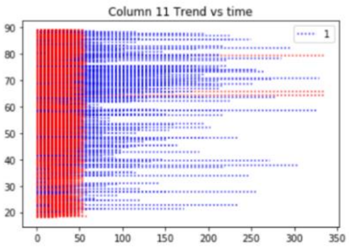
Column: The column of the given data loaded from .npy.
Feature: Characteristics of the data.

Data Exploration

P1: Trend vs Time – Each column is investigated, and each column has different features that could contribute to the predictions. For example, column 17 data has high variations and column 11 is sensitive to time length.







P2: Correlation between columns which is not all NaN and expected label – We found out that when some columns have value which is not NaN, the expected label tends to be 1. In other words, after replacing NaN with a number, data standard deviation will correlate with the label.

Column	Label 0	Label 1	Difference
5	30.78%	63.54%	-32.76%
13	46.68%	77.08%	-30.40%
18	50.74%	78.18%	-27.44%
22	48.72%	77.04%	-28.32%

Data Preprocessing

NaN Handling – We set all the NaN to the average value for the column across all the training data. We tried setting all the NaN to 0 but it yields worse result.

Features Extraction

TSFresh library is used to extract features of time series data. Given a multivariable time series, TSFresh will return many time series characteristics. Based on P1, each column has many different features that contribute to the predictions, TSFresh enables us to consider more characteristics of the time series data. TSFresh considers characteristic including standard deviation, variance, mean, count above mean, which have also considered P2.

We used extract_features to get the features and used select_features to remove irrelevant features. The output results are stored as parquet(.gzip) files for model training later.

Features Selection

Column Selection

We believe that if an irrelevant column is removed, the model will have better performance. As trying all combination of the columns would cost very long time (2^40 combinations), we employ random search technique with a fixed validation set. However, most result is inconsistent because most of the combinations that score well in the validation set, score worse in Kaggle Public Score.

Nevertheless, removing column 18 consistently give a good result, thus it is chosen in the final submission.

Feature Selection Technique

The features extracted by TSFresh might be irrelevant for the data. Therefore, feature selection techniques are employed to increase the model AUC. Several feature selection techniques were tried including Sklearn Variance Threshold, Sklearn SelectFromModel and some feature selection techniques from Skfeature. SelectFromModel using ExtraTreesClassifier was found to have good performances, thus it is chosen to select individual features in most of the submissions.

We tried running feature selection on individual feature as well as running feature selection on combined data. They are shown in result table.

Model Training

LightGBM, XgBoost, Tensorflow Neural Network, Sklearn Gradient Boosting are tested. LightGBM yields the best result.

Model Tuning

Sklearn Random Search CV and Hyperopt are experimented. Hyperopt is chosen due to its speed performance. The best parameters found by Hyperopt is {'boosting_type': 'dart', 'metric': 'auc', 'colsample_bytree': 0.7912493015275733, 'learning_rate': 0.1825037992404684, 'min_child_samples': 95, 'num_leaves': 48, 'reg_alpha': 0.5753421176471192, 'reg_lambda': 0.6231841891673146, 'subsample_for_bin': 20000}.

Table 1 Exploration Result Table, validation AUC score is not recorded.

Column	Fill NaN	Preprocessing	Feature Selection	Tuning	Model	Public Score	Private Score
All but 18	**Average**	**TSFresh**	**Extra Tree Classifier**	**Hyperopt**	**LightGBM**	**0.96169**	**0.94081**
All but 18, 27	Average	TSFresh	Extra Tree Classifier	None	LightGBM	0.95695	0.9365
All but 27	Average	TSFresh	Extra Tree Classifier Select for Each Feature	Hyperopt	LightGBM	0.95909	0.93639
All but 18	Average	TSFresh	Extra Tree Classifier	None	LightGBM	0.96069	0.93373
All but 27	Average	TSFresh	Extra Tree Classifier Select for Each Feature	None	LightGBM	0.96063	0.93336
All but 18	Average	TSFresh	Extra Tree Classifier	None	XGBoost	0.95403	0.93057
All but 18	Average	TSFresh	Extra Tree Classifier	None	Gradient Boosting	0.95056	0.92814
All but 27	Average	TSFresh	Extra Tree Classifier	None	LightGBM	0.95609	0.92755
All but 18	Average	TSFresh	None	None	Gradient Boosting	0.94368	0.9187
All	Average	TSFresh	None	None	Gradient Boosting	0.93715	0.91264
All but 2 and 18	Average	TSFresh	None	None	Gradient Boosting	0.93338	0.90609
5, 22	Average	TSFresh	None	None	Gradient Boosting	0.91056	0.87796
18, 22	Average	TSFresh	None	None	Gradient Boosting	0.91187	0.87456
22	Average	TSFresh	None	None	Gradient Boosting	0.91283	0.87368
22	Average	TSFresh	None	None	XGBoost	0.90624	0.86369
18	Average	TSFresh	None	None	Gradient Boosting	0.90419	0.86143
5, 13	Average	TSFresh	None	None	Gradient Boosting	0.90916	0.86097
5	0	TSFresh	None	None	Gradient Boosting	0.88856	0.85067
5, 33	Average	TSFresh	None	None	Gradient Boosting	0.89429	0.84817
5	Average	TSFresh	None	None	Gradient Boosting	0.89372	0.84773
13	Average	TSFresh	None	None	Gradient Boosting	0.85919	0.80678
33	0	TSFresh	None	None	Gradient Boosting	0.84482	0.78877
26	Average	TSFresh	None	None	Gradient Boosting	0.8261	0.78346
0	Average	TSFresh	None	None	Gradient Boosting	0.82475	0.78263
2	0	TSFresh	None	None	Gradient Boosting	0.82657	0.77792
35	Average	TSFresh	None	None	Gradient Boosting	0.8246	0.77493
11	0	TSFresh	None	None	Gradient Boosting	0.82215	0.7724
All	Average	TSFresh	FCBF	None	Gradient Boosting	0.77254	0.25281
2,3,6,8,11,12,13,15,17,18,22,24, 26,29,31,33,35,37,39	0	None	None	None	Tensorflow Neural Network	0.90039	0.85684
22	0	None	None	None	Tensorflow Neural Network	0.89307	0.83614
22	0	None	None	None	Tensorflow Neural Network	0.8971	0.86794
2	0	None	None	None	Tensorflow Neural Network	0.84987	0.80797
22	Average	TSFresh	None	Random Search Optimization	LightGBM	0.8483	0.8206