created by Yutong (Cindy) Liu

Machine Learning modelling: student performance prediction

This research project investigates the use of university students' clickstream data to predict their final result of the course. This research provides possible solutions to industry applications about how to utilise clickstream data to solve problems related to student retention improvement programs.

1 DATA PREPARATION

- **Goals**: to clean raw data (students' clickstream data from Learning Management System for one course) and merge it into the prediction label (students' final results of the course)
- **Achievements**: click data of 5341 students, 32% 'fail', 68% 'pass' (2 classes in label)

the course name is BBB | the course opened 4 times in 2013 and 2014 | student id | from -23 to 268 indicate the -th of day of the course | number of click

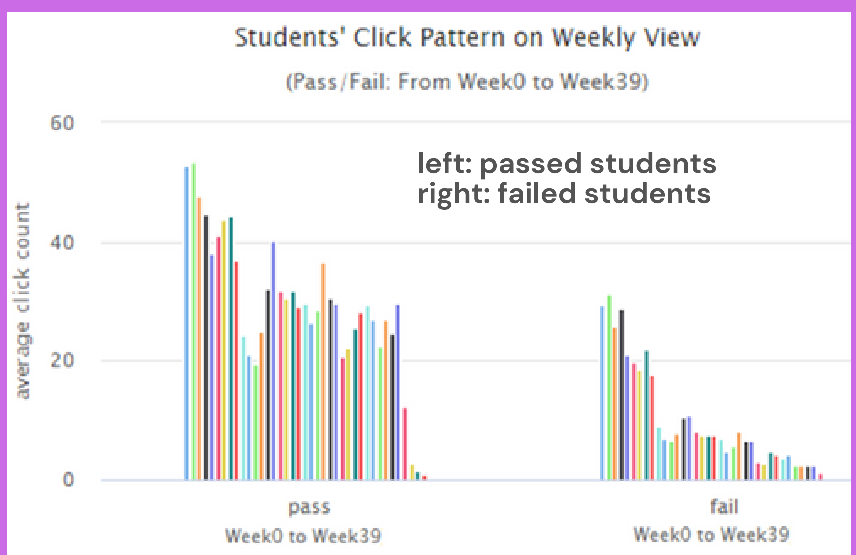
code_module	code_presentation	id_student	id_site	date	sum_click
BBB	2013J	2078479	703737	2	1
BBB	2013J	2056947	703737	2	1
BBB	2013J	2164944	703737	2	1
BBB	2013J	1411627	703737	2	1
BBB	2013J	1421720	703737	2	1
BBB	2013J	1421720	703737	2	1
BBB	2013J	1421720	703737	2	1

id_sites are groupped into 12 category types

Act1: forumng	*Act2: oucontent*	*Act3: subpage*	*Act4: homepage*
Act5: quiz	*Act6: resource*	*Act7: url*	*Act8: oucollaborate*
Act9: questionnaire	*Act10: ouelluminate*	*Act11: glossary*	*Act12: sharedsubpage*

2 EXPLORATORY ANALYSIS

- **Goals**: to explore click behaviour patterns between students who passed and failed the course
- **Achievements**: time and activity category were examined as two significant aspects that can be used to do feature engineering



Time

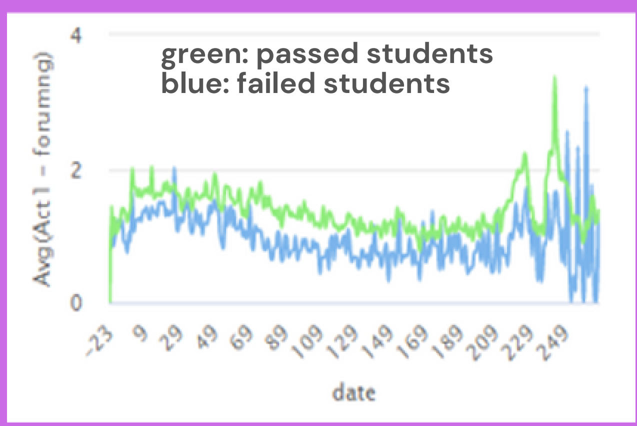
Passed and failed students have different click patterns over time.

(e.g. the left figure shows the passed and failed students' click patterns on a weekly view)

Activity category

Click behaviours on the ***forumng, oucontent, subpage, homepage, quiz*** activity categories show different patterns between students who passed and failed over time.

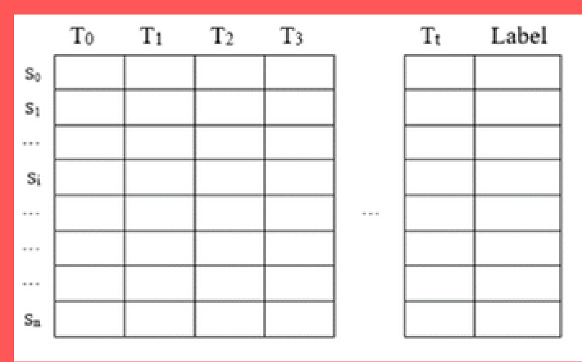
(e.g. the right figure shows students' click patterns on ***forumng***)



3 FEATURE ENGINEERING

- **Goals**: to generate features through transforming datasets
- **Achievements**: 3 strategies to generate features; each strategy involves click count aggregation by week and month, respectively. Eventually, 6 datasets were generated

Strategy 1: time periods as features

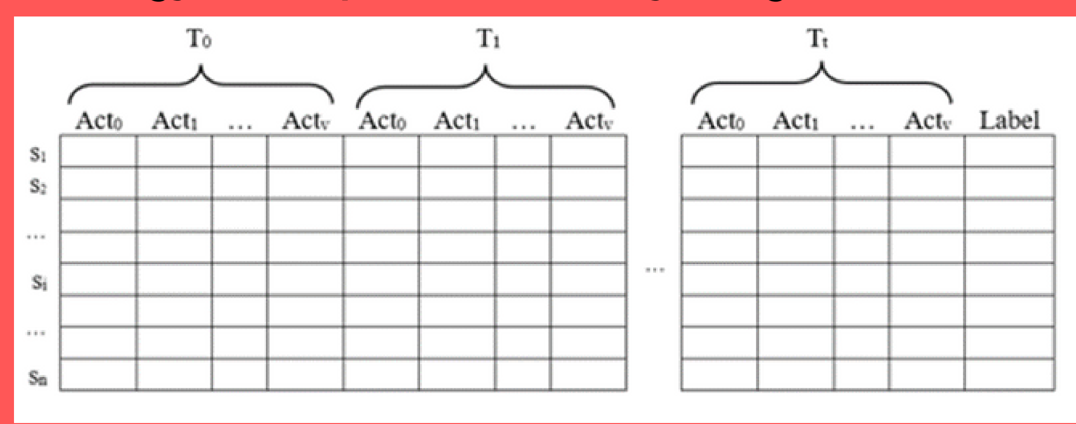


S: student T: time period

- each row indicate each student
- each column indicate click number in each time period (each week or month)

Two datasets:
- S1-WEE
- S1-MON

Strategy 2: time periods & activity categories as features



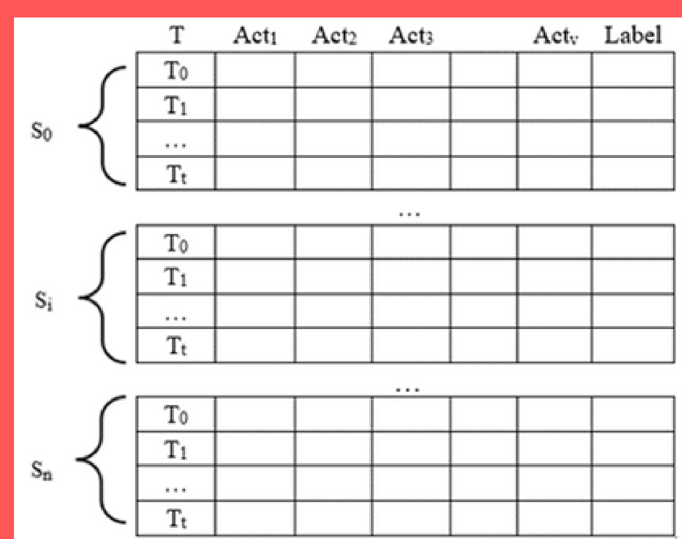
S: student T: time period Act: activity category

- each row indicate each student
- each column indicate each combination of each time period (week or month) and each activity type (12 types in total)

Two datasets:
- S2-WEE
- S2-MON

Strategy 3: panel data

Each panel represents each student; each panel is a matrix of time and activity



S: student T: time period Act: activity category

- For one panel (one student), each row indicates each time period (week or month), each columnn indicates click numbers on each activity type
- There are 5521 panels (students)

Two datasets:
- S3-WEE
- S3-MON

4 PREDICTIVE MODELLING

- **Goals**: to build predictive models
- **Achievements**: 60 models were built using
  - 6 datasets
  - 6 machine learning algorithms
  - with/without a feature selection method
  - 10-fold cross validation

Feature selection
- using all features (no feature selection method)
- using **Infomation Gain** to select features

10-fold cross validation

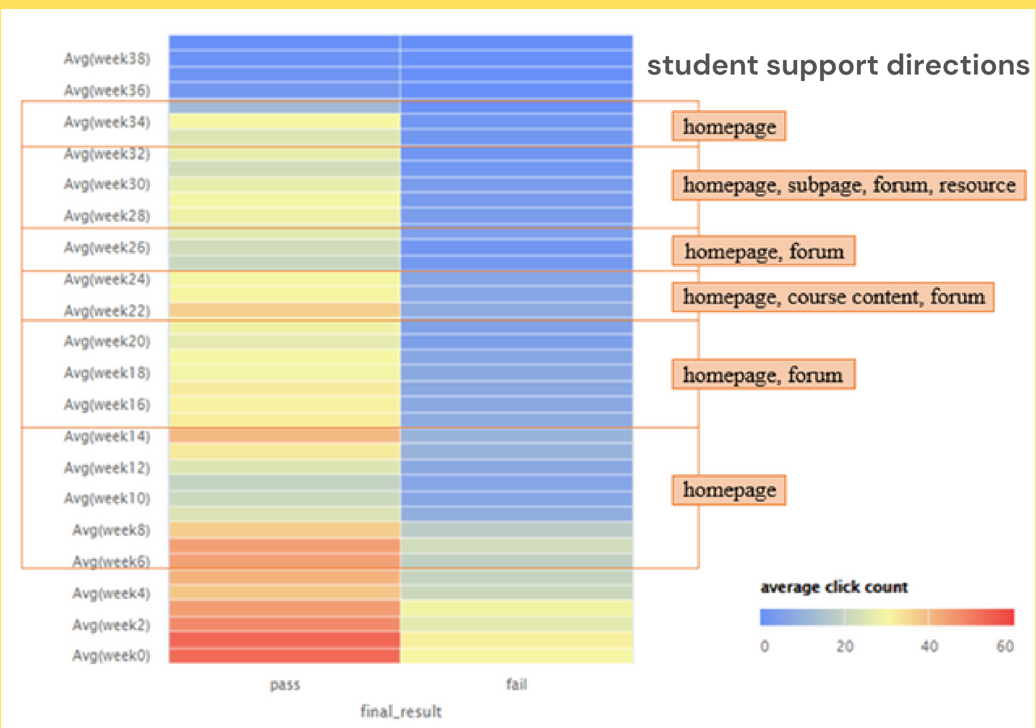
6 datasets
- S1-WEE
- S1-MON
- S2-WEE
- S2-MON
- S3-WEE
- S3-MON

Machine learning algorithms
- Logistic Regression (LR)
- K-Nearest Neighbors (k-NN)
- Random Forest (RF)
- Gradient Boosting Tree (GBT)
- 1D Convolutional Neural Network (1D-CNN)
- Long short-term memory (LSTM)

5 KEY FINDINGS

- **Goals**: to evaluate models, find the best model, analyse results
- **Achievements**:
  - models were evaluated using accuracy, F1-score, AUC
  - the best model was LSTM + S3-WEE + using all features, achieved up to accuracy of 90.22%, F1-score of 93.33% and AUC of 92.65%
  - feature engineering Strategy 3 performed the best
  - week-based performed better than month-based datasets
  - LSTM stood out among all the algorithms
  - LSTM + using all the features performed the best. Therefore, feature selection methods could be optional when using LSTM

6 INSIGHT GENERATION



Goals: to generate insights into how teachers can support students to pass the course

Achievements: by analysing the feature importance in the best model, it is found that

- the importance of time period: week 0-3, week 38-39 < weeks 4-37 < weeks 22-35
- important activity categories include homepage, subpages, forum, resources
- It is suggested to provide support to 'at-risk' students (i.e. students who are likely to fail the course) based on different activity categories in different course periods (see the left figure)