created by Yutong (Cindy) Liu

Machine Learning modelling: student performance prediction

DATA TRANSFORMATION

- The raw data (from Open University) is transformed to demostrate students' click behaviours in LMS (Learning Management System) during a course and their final results of the course
- The data involve 5521 students - 32% failed, 68% passed students

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 | the 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

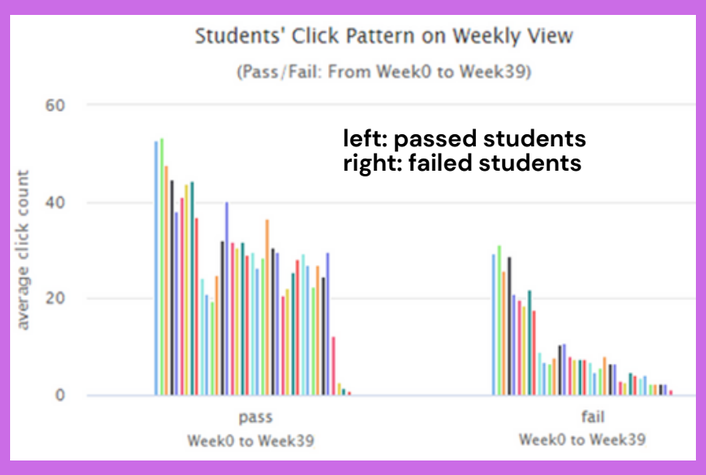
id_sites are groupped into 12 category types

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

1

EXPLORATORY ANALYSIS

- Explore click behaviour patterns between students who passed and failed the course
- It is found that time and activity category are two significant aspects that can be used to demonstrate informative patterns



Time

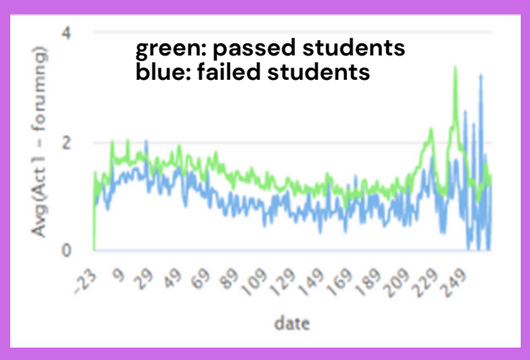
Passed and failed students have different click patterns over time.

(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***)

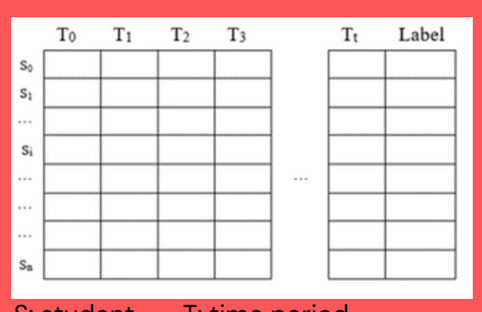


2

FEATURE ENGINEERING

- 3 strategies of feature enginnering are developed based on time and category type
- 2 time granularity sizes are used - week and month
- 6 datasets are generated (3 strategies * 2 time granularity sizes)

Strategy 1: time - based features



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

Two datasets:
- S1-WEE
- S1-MON

S: student T: time period

3

PREDICTIVE MODELLING

- 60 models are built using 6 datasets (generated in the last step) and 6 machine learning algorithms, along with feature selection method and 10-fold cross validation. Models are evaluated using accuracy, F1-score, AUC
- **As a result, the best model is LSTM & S3-WEE & using all features; the model achieves the accuracy of 89.25% (+/- 0.97%), F1-score of 92.71% (+/- 0.62%) and AUC of 91.28% (+/- 1.37%)**

Feature selection
- using all features
- using infomation gain to select features

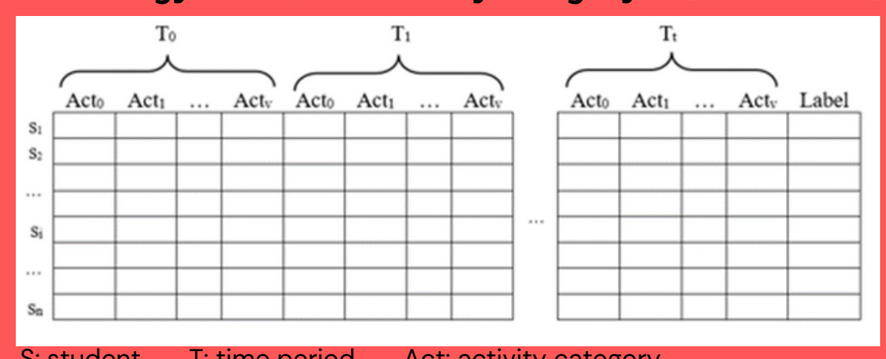
10-fold cross validation

Datasets
- S1-WEE
- S1-MON
- S2-WEE
- S2-MON
- S3-WEE
- S3-MON

Machine learning algorithms
- Logistic Regression
- k-NN
- Random Forest
- Gredient Boosting Tree
- 1D-CNN
- LSTM

4

Strategy 2: time and activity category-based features



S: student T: time period Act: activity category

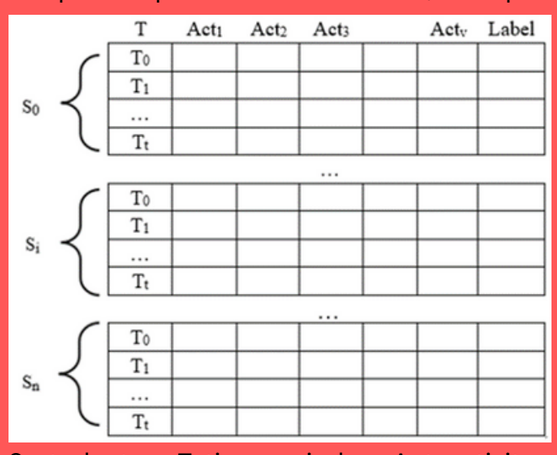
- each row indicate each student
- each column indicate each combination of each time period (week or month) and each activity type (12 types in total)

Two datasets:
- S2-WEE
- S2-MON

5

Strategy 3: panel data

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



- 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)

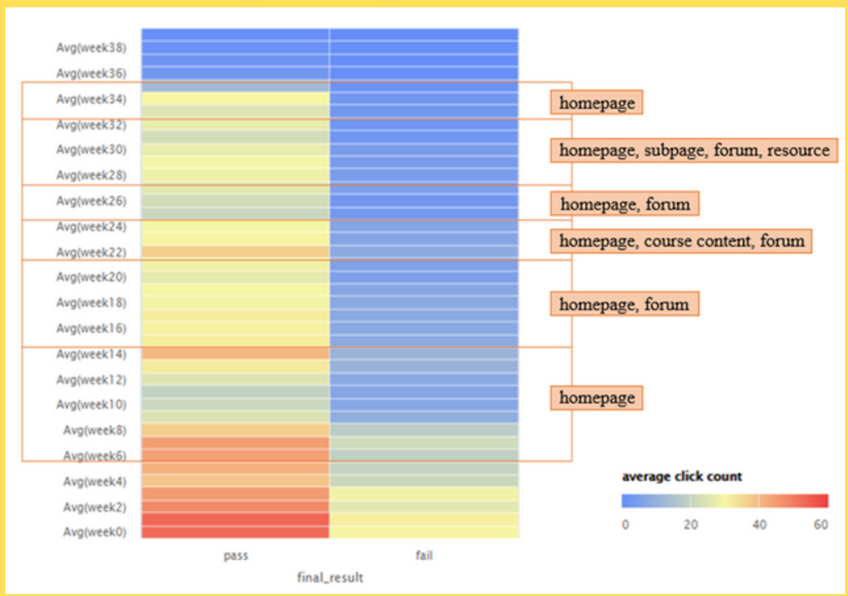
Two datasets:
- S3-WEE
- S3-MON

S: student T: time period Act: activity category

6

KEY FINDINGS

In this classification case,
- feature enginneering Strategy 3 (panel data) performs the best;
- WEEK is better than MONTH as a time granularity size;
- LSTM model performs the best among the six algorithms;
- feature selection method is optional when using LSTM



INSIGHT GENERATION

The feature importance in the best model are analysed:
- **week 0-3** and **week 38-39** are least important to predict students' performance
- **weeks 4-37** is the significant period to predict students' performance
- **weeks 22-35** is the most significant period to predict students' performance
- Important activity categories to predict students' performance are **homepage, subpages, forum, resources**

According to the feature importance analysis, teachers are suggested to provide support to 'at-risk' student (student who are likely to fail the course) based on **different activity categories in different time periods (see the left figure)**