12-2018

Data mining approach to the identification of at-risk students

Li Chin HO
Singapore Management University, lichin.ho.2016@mitb.smu.edu.sg

Kyong Jin SHIM
Singapore Management University, kjshim@smu.edu.sg

Citation

HO, Li Chin and SHIM, Kyong Jin. Data mining approach to the identification of at-risk students. (2018). *2018 IEEE International Conference on Big Data (Big Data): Seattle, December 10-13: Proceedings.* 5333-5335.

Available at: https://ink.library.smu.edu.sg/sis_research/4339

2018 IEEE International Conference on Big Data (Big Data)
https://doi.org/10.1109/BigData.2018.8622495

Data Mining Approach to the Identification of At-Risk Students

Li Chin Ho
School of Information Systems
Singapore Management University
Singapore
lichin.ho.2016@mitb.smu.edu.sg

Kyong Jin Shim
School of Information Systems
Singapore Management University
Singapore
kjshim@smu.edu.sg

Abstract—**In recent years, the use of digital tools and technologies in educational institutions are continuing to generate large amounts of digital traces of student learning behavior. This study presents a proof-of-concept analytics system that can detect at-risk students along their learning journey. Educators can benefit from the early detection of at-risk students by understanding factors which may lead to failure or drop-out. Further, educators can devise appropriate intervention measures before the students drop out of the course. Our system was built using SAS® Enterprise Miner (EM) and SAS® JMP Pro.**

Keywords—learning analytics, at-risk students, learning management systems, educational data mining

I. Introduction

Increasing usage of digital platforms such as Learning Management Systems (LMS) in educational institutions have created a massive gold mine for educators wanting to systematically analyze student learning behavior. Such systems capture high granularity data concerning how students interact with instructors as well as learning materials, and it enables advanced analytics which can reveal insights into why, when and how students fail or succeed in learning modules (Figure 1).

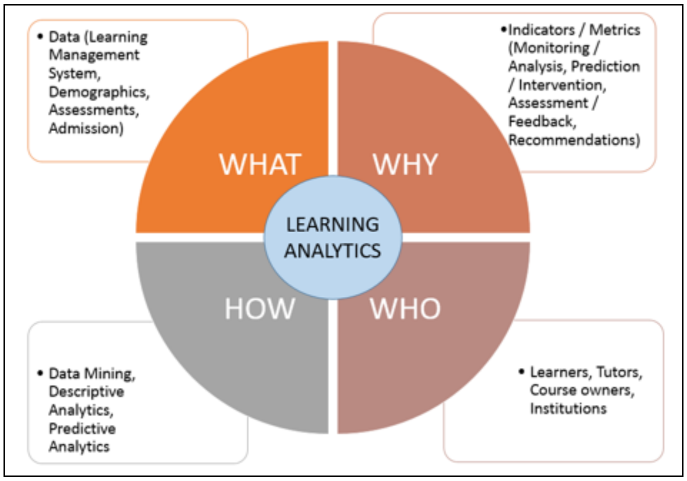

• Data (Learning Management System, Demographics, Assessments, Admission)
•Indicators / Metrics (Monitoring / Analysis, Prediction / Intervention, Assessment / Feedback, Recommendations)
WHAT
WHY
LEARNING ANALYTICS
HOW
WHO
• Data Mining, Descriptive Analytics, Predictive Analytics
• Learners, Tutors, Course owners, Institutions

Fig. 1. Learning Analytics Framework

In this study, we develop a proof-of-concept analytics system using SAS® Enterprise Miner (EM) and SAS® JMP Pro. The system aims to analyze the Open University's dataset [1].

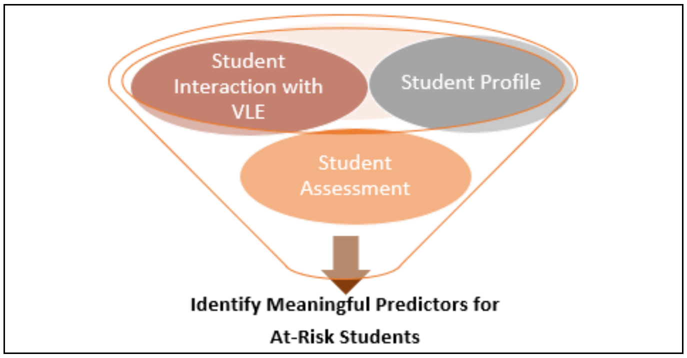

Student Interaction with VLE
Student Profile
Student Assessment
Identify Meaningful Predictors for At-Risk Students

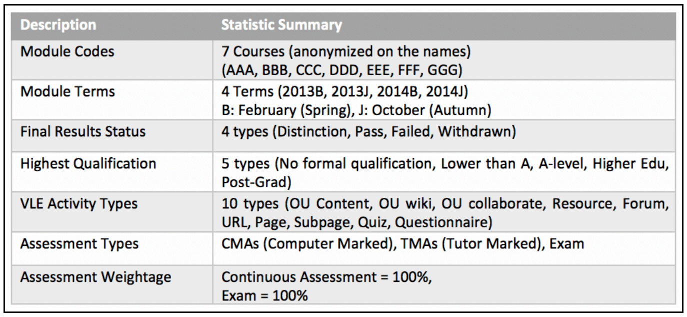
Fig. 2. Open University Learning Analytics Dataset

As shown in Figure 2, the dataset contains information about different learning modules, student profiles, students' interactions with Virtual Learning Environment (VLE), and students' assessment information.

The Open University offers modules for distance learning students – both undergraduate and post-graduate. This study seeks to develop methods for identifying and predicting students that are about to fail or drop out of learning modules. Early detection of such students will allow educators to employ appropriate intervention measures so as to guide the students to stay on track towards successful completion of the modules.

II. Dataset Overview

As shown in Figure 3, the dataset contains: 1) student profile data, 2) student activity data, 3) learning context data (module info). The dataset contains information about over 25,000 students and their online interactions. Students' ddemographics information includes students' past education info such as their highest educational qualification (e.g. A-Level, Diploma, Degree, etc.), gender, age, religion and so forth. Assessment data include students' assessment scores from online quizzes, exams, etc. VLE data indicate whether and how often students interacted with the online learning platform's resources such as forums.



Description	Statistic Summary
Module Codes	7 Courses (anonymized on the names) (AAA, BBB, CCC, DDD, EEE, FFF, GGG)
Module Terms	4 Terms (2013B, 2013J, 2014B, 2014J) B: February (Spring), J: October (Autumn)
Final Results Status	4 types (Distinction, Pass, Failed, Withdrawn)
Highest Qualification	5 types (No formal qualification, Lower than A, A-level, Higher Edu, Post-Grad)
VLE Activity Types	10 types (OU Content, OU wiki, OU collaborate, Resource, Forum, URL, Page, Subpage, Quiz, Questionnaire)
Assessment Types	CMAs (Computer Marked), TMAs (Tutor Marked), Exam
Assessment Weightage	Continuous Assessment = 100%, Exam = 100%

Fig. 3. Open University Learning Analytics Dataset

In this study, we focus on seven modules over four terms. We analyze students' online interactions including their assessment results and other significant learning behavior such as frequency with which the students access learning materials and their eventual module outcome (e.g. distinction, pass, fail, withdrawal).

III. Analysis & Findings

SAS® Enterprise Miner (EM) and SAS® JMP are the two main software programs used in the data mining tasks and predictive modelling in this study. SAS® JMP is used mainly to extract, transform and perform the exploratory data analysis task. SAS® Enterprise Miner (EM) is used to build predictive models.

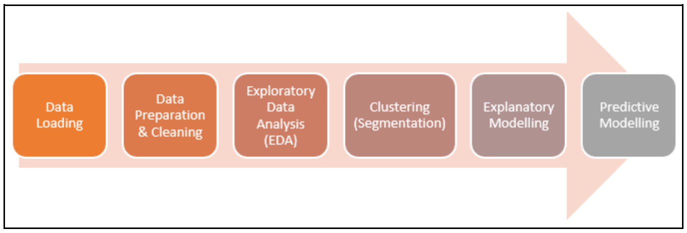

Data Loading
Data Preparation & Cleaning
Exploratory Data Analysis (EDA)
Clustering (Segmentation)
Explanatory Modelling
Predictive Modelling

Fig. 4. Analysis Workflow

We also built predictive models using Python 3.5 to cross check the predictive analytics results with the results obtained from SAS® Enterprise Miner (EM). Figure 4 shows our analysis workflow. We used Python scikit-learn package for building predictive models [2]. Figure 5 shows the details of the tools used in our analysis.

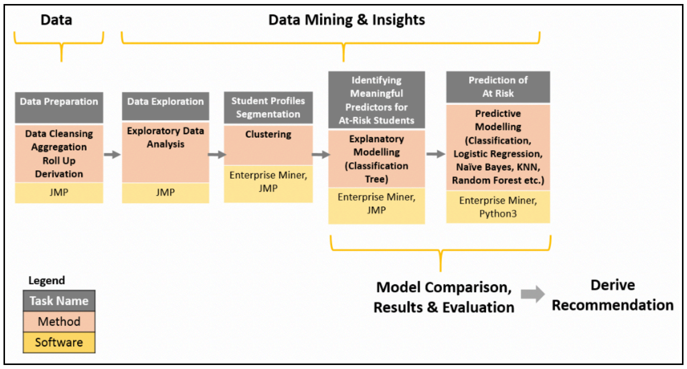
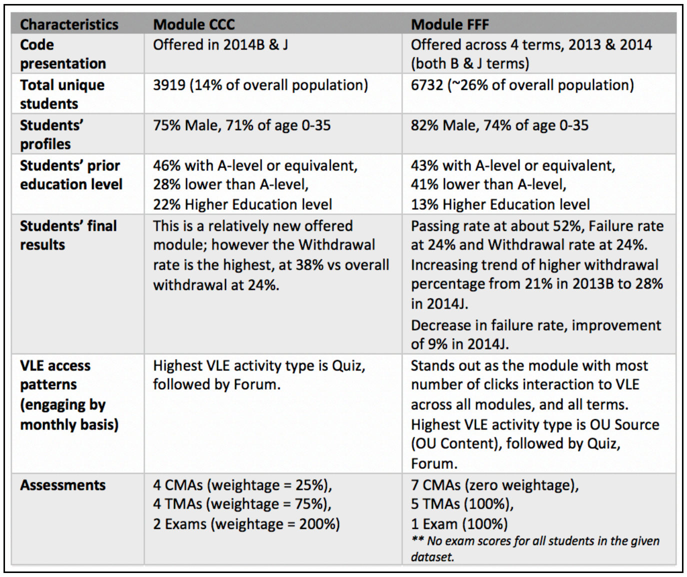

Data
Data Mining & Insights
Data Preparation
Data Cleansing Aggregation Roll Up Derivation
JMP
Data Exploration
Exploratory Data Analysis
JMP
Student Profiles Segmentation
Clustering
Enterprise Miner, JMP
Identifying Meaningful Predictors for At-Risk Students
Explanatory Modelling (Classification Tree)
Enterprise Miner, JMP
Prediction of At Risk
Predictive Modelling (Classification, Logistic Regression, Naïve Bayes, KNN, Random Forest etc.)
Enterprise Miner, Python3
Model Comparison, Results & Evaluation
Derive Recommendation
Legend
Task Name
Method
Software

Fig. 5. Analysis Approach Overview

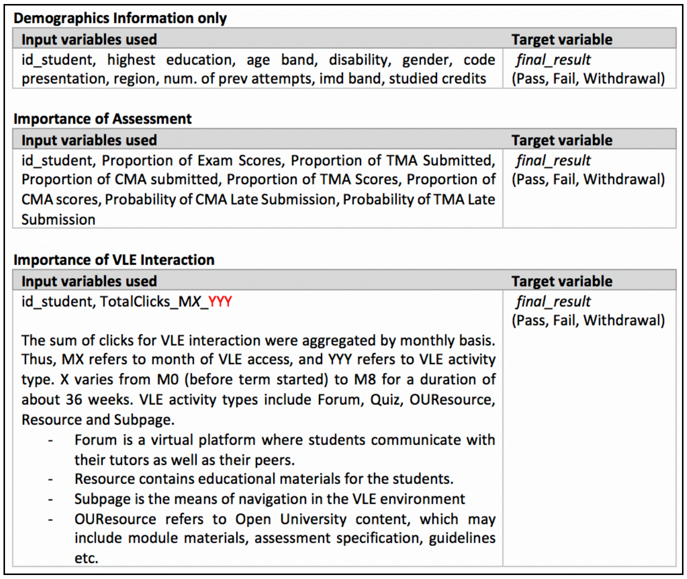
We identified two modules to be associated with at-risk students: 1) CCC has the highest withdrawal among all students, 2) FFF (offered across all terms and with large student intake size). Figure 6 shows details of the two modules.



Characteristics	Module CCC	Module FFF
Code presentation	Offered in 2014B & J	Offered across 4 terms, 2013 & 2014 (both B & J terms)
Total unique students	3919 (14% of overall population)	6732 (~26% of overall population)
Students' profiles	75% Male, 71% of age 0-35	82% Male, 74% of age 0-35
Students' prior education level	46% with A-level or equivalent, 28% lower than A-level, 22% Higher Education level	43% with A-level or equivalent, 41% lower than A-level, 13% Higher Education level
Students' final results	This is a relatively new offered module; however the Withdrawal rate is the highest, at 38% vs overall withdrawal at 24%.	Passing rate at about 52%, Failure rate at 24% and Withdrawal rate at 24%. Increasing trend of higher withdrawal percentage from 21% in 2013B to 28% in 2014J. Decrease in failure rate, improvement of 9% in 2014J.
VLE access patterns (engaging by monthly basis)	Highest VLE activity type is Quiz, followed by Forum.	Stands out as the module with most number of clicks interaction to VLE across all modules, and all terms. Highest VLE activity type is OU Source (OU Content), followed by Quiz, Forum.
Assessments	4 CMAs (weightage = 25%), 4 TMAs (weightage = 75%), 2 Exams (weightage = 200%)	7 CMAs (zero weightage), 5 TMAs (100%), 1 Exam (100%) *** No exam scores for all students in the given dataset.*

Fig. 6. Footprint Characteristics of Module CCC and Module FFF

We attempt to predict module outcome for students enrolled in module CCC and module FFF. The variables used as predictors to build the predictive models are shown in Figure 7.



Demographics Information only

Input variables used	Target variable
id_student, highest education, age band, disability, gender, code presentation, region, num. of prev attempts, imd band, studied credits	*final_result* (Pass, Fail, Withdrawal)

Importance of Assessment

Input variables used	Target variable
id_student, Proportion of Exam Scores, Proportion of TMA Submitted, Proportion of CMA submitted, Proportion of TMA Scores, Proportion of CMA scores, Probability of CMA Late Submission, Probability of TMA Late Submission	*final_result* (Pass, Fail, Withdrawal)

Importance of VLE Interaction

Input variables used	Target variable
id_student, TotalClicks_MX_YYY The sum of clicks for VLE interaction were aggregated by monthly basis. Thus, MX refers to month of VLE access, and YYY refers to VLE activity type. X varies from M0 (before term started) to M8 for a duration of about 36 weeks. VLE activity types include Forum, Quiz, OUResource, Resource and Subpage. - Forum is a virtual platform where students communicate with their tutors as well as their peers. - Resource contains educational materials for the students. - Subpage is the means of navigation in the VLE environment - OUResource refers to Open University content, which may include module materials, assessment specification, guidelines etc.	*final_result* (Pass, Fail, Withdrawal)

Fig. 7. Footprint Characteristics of Module CCC and Module FFF

For each of the areas, the dataset is partitioned into training (80%) and testing (20%) sets. The training set is used to estimate model parameters, and the testing set is the part that assesses and validates the predictive ability of the models.

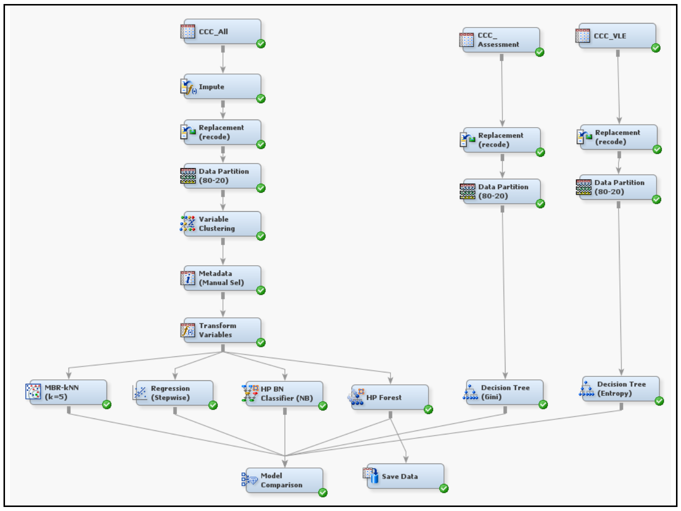
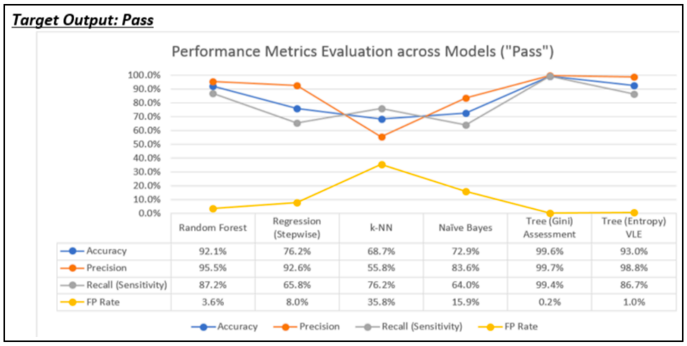


Fig. 8. Building Predictive Models Using SAS® Enterprise Miner (EM) – Module CCC

As shown in Figure 8, three sets of data are used to build the predictive models.

1. All data including student demographics, assessment and VLE data. This dataset is used to build four predictive models: KNN, Regression, Naïve Bayes, and Random Forest.
2. Students' assessment data. This dataset is used to build Decision Tree models.
3. VLE clicks including students' VLE engagement. This dataset is used to build Decision Tree models.

These models consider different properties of data and complement one another. Each model independently classifies each student into one of the following classes: Pass, Fail, Withdrawn. The final prediction decision is done by combining the outcomes of all the models.

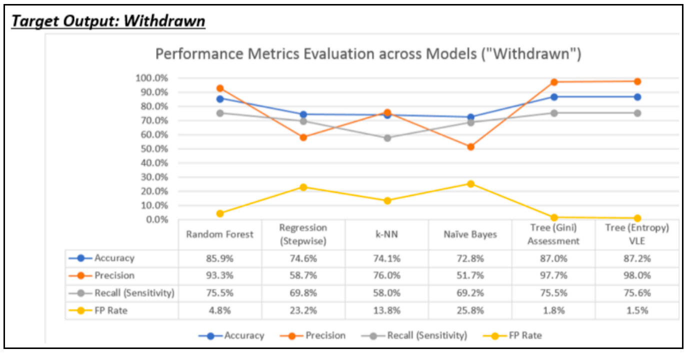


	Random Forest	Regression (Stepwise)	k-NN	Naïve Bayes	Tree (Gini) Assessment	Tree (Entropy) VLE
Accuracy	92.1%	76.2%	68.7%	72.9%	99.6%	93.0%
Precision	95.5%	92.6%	55.8%	83.6%	99.7%	98.8%
Recall (Sensitivity)	87.2%	65.8%	76.2%	64.0%	99.4%	86.7%
FP Rate	3.6%	8.0%	35.8%	15.9%	0.2%	1.0%

Fig. 9. Predicting "Pass" Outcome for Module CCC

Figure 9 shows the results of predicting "Pass" outcome for Module CCC. The Random Forest model achieved 92.1% accuracy with 95.5% precision and 87.2% recall values. The Decision Tree model using only the assessment data achieved the highest accuracy of all models. The Decision Tree model using the VLE data also produced comparable prediction results.

The best performing model for predicting "Pass" outcome for Module FFF is the Decision Tree model using only the assessment data. It achieved overall 94.5% accuracy with 97% precision and 92.8% recall values.



	Random Forest	Regression (Stepwise)	k-NN	Naïve Bayes	Tree (Gini) Assessment	Tree (Entropy) VLE
Accuracy	85.9%	74.6%	74.1%	72.8%	87.0%	87.2%
Precision	93.3%	58.7%	76.0%	51.7%	97.7%	98.0%
Recall (Sensitivity)	75.5%	69.8%	58.0%	69.2%	75.5%	75.6%
FP Rate	4.8%	23.2%	13.8%	25.8%	1.8%	1.5%

Fig. 10. Predicting "Withdrawn" Outcome for Module CCC

Figure 10 shows the results of predicting "Withdrawn" outcome for Module CCC. The best performing model is the Decision Tree model using only the assessment data. It achieved 87% accuracy with 97.7% precision and 75.5% recall values. The Decision Tree model using the VLE data also produced comparable prediction results.

The best performing model for predicting "Withdrawn" outcome for Module FFF is the Decision Tree model using only the assessment data. It achieved overall 84.2% accuracy with 60.6% precision and 70.3% recall values.

For both Module CCC and Module FFF, predicting the "Fail" outcome proved to be challenging. For Module CCC, the best performing model (Decision Tree using only the assessment data) achieved 86.6% accuracy with 35.3% precision and 86.9% recall values. For Module FFF, the Decision Tree model using only the assessment data achieved 79.4% accuracy with 58.7% precision and 56.3% recall values.

IV. Conclusion & Future Directions

Our analysis results show that the students' assessment results as well as online interaction with learning resources serve as good predictors for their eventual module outcome. Predicting "Fail" outcome proved to be challenging. We plan to look into students' learning footprints that can distinguish "fail" and "withdrawn" cases clearly and incorporate it into prediction. We plan to extend this study to real-life university settings. We plan to include both offline assessment data and online assessment data towards predicting at-risk students. The significant predictors for pass, fail and withdrawal cases identified in this study provide useful insights to educators in designing proper intervention measures to help students stay on track.

References

[1] Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017).

[2] Scikit-learn: https://scikit-learn.org