Analyzing Significant Features Using Predictive Modeling for Digital Learning Platform

Wei-cheng Chen, Cheng Cheng, Wenying Huang, Matthew A. Lanham

Purdue University Krannert School of Management

chen1614@purdue.edu; cheng436@purdue.edu; huang814@purdue.edu; lanhamm@purdue.edu

Abstract

The rise of digital platforms and its prevalence in teaching and learning brings up the problem of lacking interaction and mentorship in distance learning. Perceivant, as a prominent educational technology company, needs to upgrade the online platform previously designed for KSU WELL 100 course to achieve seamless integration between old curriculum and new system. In this research we analyze the hidden behavior patterns in students' academic data and develop predictive models to identify influential features of the platform that affects engagement of students and staffs. Our solution enables decision-makers to optimize online learning management system for student success.

Introduction

Improving student academic performance and facilitating teaching and learning are the most crucial functionalities of online learning platform. However, uncontrollable large class size, inaccessible course delivery method, ineffective teaching method, and other potential factors results in student low participation and performance. To diagnose the existing online learning tools, we need to determine important features and factors in the learning flow that measure student performance and affect learning experiences.

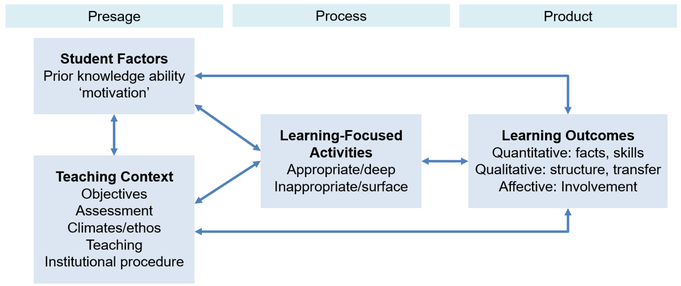

Figure 1. The 3P model of teaching and learning (Biggs, 2003)

Research questions:

- How does features other than scores such as lecture type, instructors, and length of study time contribute to student success?
- Which machine learning algorithms can most accurately identify student outcome, based on student performance and other interaction features?

Literature Review

Most prevailing online-learning analytics researches consider aggregation of different features in pattern discovery and classification methods like decision trees, logistic, and support vector machine regression in predictive modelling.

Study	RF	DL	GBM	GLM	SVM	Naive
(Kabakchieva, 2013)						x
(Pedro, 2015)	x				x	x
(Fazel, 2016)				x		x
(Aaron, 2018)				x	x	
Our Study	x	x	x	x		x

Table 1. Model comparison of digital learning platforms

The novelty that distinguishes our algorithm from others is that we used cross-validation and trained both classification and regression models.

Methodology

Data

Data is extracted from Perceivant's database that contain 3 student grades detail, assessment response log, and pageview of the online guided learning. We also did some exploratory analysis to find out what each dataset about

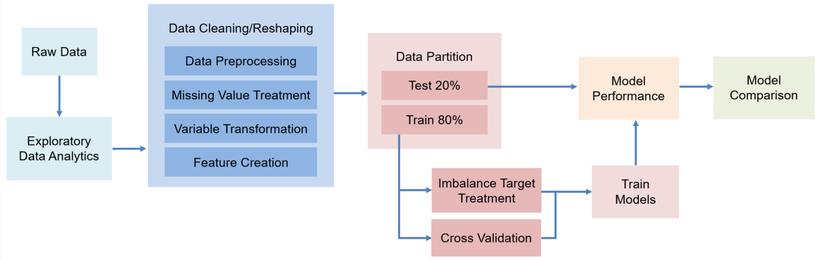

Figure 2. Methodology Flow Chart

Data Preparation

- Converted features to its appropriate type.
- Created new features that sums up the score of quizzes, assessment, projects etc. of each student
- Merged the different students grades together based on student's unique id
- Joined duration column from pageview dataset that shows total time of students reading e-book
- Replaced missing values as 0s
- Created dummies for categorical variables, removed high correlation features and linear combos
- Normalize each numeric column to avoid bias
- Generated target columns to specify pass/fail based on Grade (%)

Model Design using H2O

- Converted data into H2O cluster
- Partitioned the data into **80-20%** train-test
- Performed **5-fold cross-validation** to avoid over fitting the model
- Loaded SMOTE package to **up-sample** the number of students, who are at risk of failing, to achieve a balanced ratio between the targets

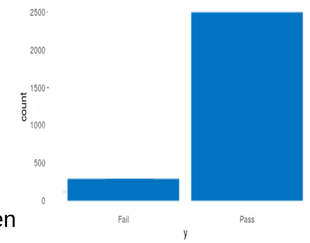

Figure 3. Imbalance Target Distribution

Methodology (Approach) Selection

We used 5 classification algorithms to generate predictions for student success and failure. Each model will generate a confusion matrix that shows the accuracy of identifying a student who fails or pass based on selected features. We will select the best model based on the highest specificity rate, which measures the accuracy of correctly identifying students who are most likely to fail.

Results

We applied specificity to measure each algorithm's performance. Using crossing validation, we confirmed that our models were not overfitting the data (except Naïve Bayes model). Among the algorithm we deployed, Generalized Linear Model (GLM) yielded stronger predictive capabilities.

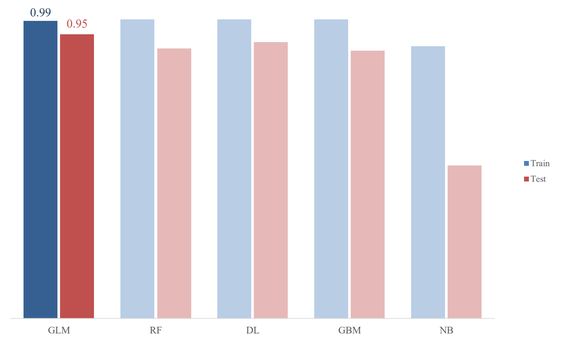

Figure 4. Model Evaluation with Specificity

Among all the features we created and applied, we found 'Total project score', 'Total quiz score', 'Total assessment score', 'Total guided learning score', 'Final exam score' and 'Total Discussion score' have the significant effect on student success.

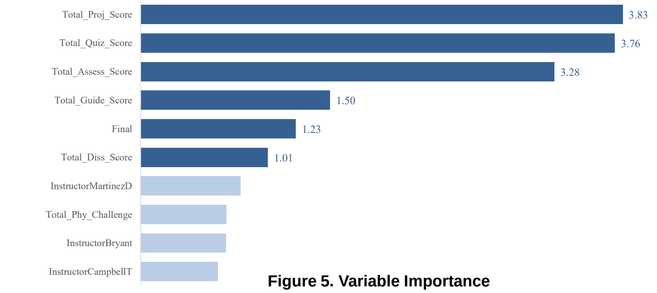

Figure 5. Variable Importance

Conclusions

Using our model, decision-makers can identify student success based on student's current performance. Although features such as lecture type, instructors, and length of study time can contribute to student success, these features don't have significant impact. Various type of assessments like 'Total project score', 'Total quiz score', etc. can significantly affect a student's outcome. In our study, Generalized Linear Model (GLM) has the best performance, which resulted 0.95 specificity on test set. By applying our model, decision-makers can identify students who may fail the course in advance and take actions to alert them and help them to pass the course.

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project.