MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

DATA ENABLED STUDENT SUPPORT (DESS)
How our system was developed

The DESS system predictive and cluster models

Overview

The key aim of the Data Enabled Student Support (DESS) system is to use data-driven methods to identify at-risk students and to inform Massey's Student Achievement team which students are in need of interventions in order to improve successful outcomes.

The DESS system uses machine learning and a variety of data inputs in order to determine the risk status of undergraduate students based on the model's predictive outputs in respect to their qualification completion.

Apart from attempting to predict the risk status of each student, the DESS system also uses a data mining technique which automatically groups similar students into several clusters which enables the Massey's Student Achievement team to cater in a customised way to each one of the distinct groups in most effective ways.

The DESS system can therefore be summarised as comprising of two key components:

1. The predictive model that identifies at-risk students, together with an associated probability of the risk.
2. The cluster model that is used for gaining insights into what kind of a student-cohort group a student belongs to in order to initiate a suitable intervention if needed.

How was DESS developed?

The data for building the models was sourced from Massey's Student Management System and Stream databases. The models were built using historic data describing various behaviours and outcomes of previous students who have studied at Massey University. The machine learning techniques were then used to uncover patterns within the data footprints of previous cohorts of students which could then be applied for making predictions about the current students. The data features and source system tables for both models are included in **Appendix A**.

Predictive model

The predictive models were developed using a variety of data variables. The models use data describing learner characteristics like gender, age, nationality, full-time study status and the basis for University admission. We combine these demographic data with academic performance data describing each student such as the current grade average and how far off it is from the student's peers as well as how many courses a student might have failed to complete. Additionally, we integrate data describing learner behaviours around online engagement on Stream like the number of forum posts created and read. The models also take into account the amount of learning content accessed on Stream as well as attempted quizzes and the number of courses from which a student has withdrawn and a student's outstanding fee status (if any). Finally, the models also consider the type of a qualification that a student is enrolled in as well as the number of credits that are associated with the qualification.

The variables listed above were chosen through an iterative experimental process. Numerous predictive models were developed and tested for their accuracy. The effectiveness of each variable towards predicting at-risk students was analysed using specific tools and the most useful and meaningful variables were retained in the final predictive model. The predictive models were developed using state-of-the-art machine learning algorithms currently available. The predictive accuracy of the current model is approximately 85% which is within the expected range for a domain of this nature.

Cluster model

The cluster model groups the current cohort of students into six categories based on common attributes. The data used for clustering students is similar to the data used for predictive models; however, some additional student variables are included. Some of these are: ethnicity, NZ citizenship status, length of enrolment, percent of the qualification completed as well as the predicted risk status of a student generated by the predictive model.

Numerous cluster models were generated during the experimental phase. The final selection of variables included in the cluster model was decided both by the stakeholders' requirements for facilitating effective intervention strategies and by the coherence of the models in describing each cluster group with a sufficient level of distinction and meaningfulness. Meaningfulness of the clusters was determined by examining the characteristics of each cluster. If sufficient structure emerged in the clusters enabling a distinctive narrative to be devised for each cluster, then this was determined as a valid and useful cluster model.

DESS system validation

The methodology for both the predictive and the cluster models were peer-reviewed. The steps taken to build the models, as well as the data used to generate them, together with the final accuracies were analysed by an expert in machine learning who was external to the project, and was validated.

DESS predictive model data features and source systems

The data used in the DESS predictive model can be divided into four general categories. The tables below describe each category, the names of the features, examples of their values as well as the source database systems from which they originate. SMS is Massey's Student Management System.

1. Learner characteristics data

FEATURE NAME	FEATURE VALUE TYPES	DATA SOURCE
CITIZENSHIP COUNTRY DESCRIPTION	Nationality	SMS
BASIS FOR ADMISSION DESCRIPTION	NCEA, adult admission etc	SMS
DEBTOR	Does the student owe fees?	SMS
HAS PREVIOUS TERTIARY STUDY	Yes/No	SMS
HIGHEST SCHOOL QUALIFICATION DESCRIPTION		SMS
CURRENT FULL TIME STATUS	Full-time/part-time	SMS
CURRENT STUDENT MODE	On-campus/distance/block	SMS
CURRENT PRIOR ACTIVITY DESCRIPTION	What was the primary activity that the student was engaged in, in the previous year	SMS
AGE DESCRIPTION	Current student age	SMS
GENDER	Male/female/another gender	SMS

2. Learner behaviour data

FEATURE NAME	FEATURE VALUE TYPES	DATA SOURCE
PAPERS WITHDRAWN FOR STUDENT ACADEMIC YEAR	Number of paper that a student has withdrawn from in a given year	SMS
ONLINE LEARNING SUBMITTED ASSIGNMENT ZSCORE	Student's Z-score calculation in respect to the average cohort grade across all courses enrolled in for number of assignments submitted	Stream
ONLINE LEARNING PAGES VIEWED COUNT ZSCORE	Student's Z-score calculation in respect to the average cohort grade across all courses enrolled in for the amount of online Stream content accessed	Stream
ONLINE LEARNING QUIZ TAKEN COUNT ZSCORE	Student's Z-score calculation in respect to the average cohort grade across all courses enrolled in for the number of quizzes taken	Stream
ONLINE LEARNING FORUM POST CREATED COUNT ZSCORE	Student's Z-score calculation in respect to the average cohort grade across all courses enrolled in for the number of forum posts created	Stream
ONLINE LEARNING FORUM POST READ COUNT ZSCORE	Student's Z-score calculation in respect to the average cohort grade across all courses enrolled in for the number of forum posts read	Stream

3. Academic performance data

FEATURE NAME	FEATURE VALUE TYPES	DATA SOURCE
GRADE MARK MEAN	Student's mean grade for current academic year courses	SMS
GRADE MARK MIN	Student's minimum grade for current academic year courses	SMS
GRADE MARK MAX	Student's maximum grade for current academic year courses	SMS
GRADE MARK DEVIATION FROM CLASS MEAN	Student's Z-score calculation in respect to the average cohort grade across all courses enrolled in	SMS
PAPERS FAILED FOR STUDENT ACADEMIC YEAR	Number of courses that a student has not success-fully completed	SMS
GRADE MARK MEAN	Student's mean grade for current academic year courses	SMS

4. Programme characteristics data

FEATURE NAME	FEATURE VALUE TYPES	DATA SOURCE
PROGRAMME TITLE	Name of qualification	SMS
PROGRAMME CREDITS REQUIRED	360, 180 etc.	SMS

DESS cluster model data features and source systems

FEATURE NAME	FEATURE VALUE TYPES	DATA SOURCE
ETHNICITY	European, Māori, Asian etc.	SMS
NZ CITIZENSHIP STATUS	Domestic/International	SMS
LENGTH OF ENROLMENT	Years since initial enrolment year	SMS
% QUALIFICATION COMPLETED	Percent of the student's qualification completed	SMS
PREDICTED RISK STATUS OF MODEL	Predicted probability of the student's qualification completion	DESS Predictive Model
CURRENT FULL TIME STATUS	full-time/part-time	SMS
PAPER COUNT	The number of courses that a student is enrolled in	SMS
CURRENT STUDENT MODE	On-campus/distance/block	SMS
GRADE MARK MEAN	Student's mean grade for current academic year courses	SMS
GENDER	male/female/another gender	SMS
AGE DESCRIPTION	current student age	SMS
BASIS FOR ADMISSION DESCRIPTION	NCEA, adult admission etc.	SMS
CURRENT PRIOR ACTIVITY DESCRIPTION	What was the primary activity that the student was engaged in, in the previous year	SMS

DESS METHODOLOGY