




Predictive Analytics for Tertiary Learners in New Zealand Who Are at Risk of Dropping Out of Education

Wenying Xu¹, Scott Luo¹, Stephanie Hacksley¹, Tim Trewinnard¹,
Stuart Cambridge², and Syen Jien Nik¹ 

¹ Jade Software Corporation, Christchurch, New Zealand
snik@jadeworld.com

² Tertiary Education Commission, Wellington, New Zealand

Abstract. This industry showcase covers a proof-of-concept predictive model in the education sector of New Zealand. Jade Software worked with New Zealand's Tertiary Education Commission on research to find out how to predict the likelihood of learners dropping out. Our model informs the implementation of intervention programs to support learners in completing their qualifications. The goal of this research is to identify a common data set across multiple types of tertiary education organizations and develop predictive models using the data set. We found that the Single Data Return is a viable data source to form a base model. By comparing the area under the receiver operator characteristic curve, we show that additional data sources, including the attendance data and the learner's results, are helpful in improving model performance. We also developed an interactive dashboard to facilitate estimating the return on investment for intervention programs and the optimal intervention threshold.

Keywords: Predictive modelling · Tertiary Education · Learner success

1 Background

The primary objective of the Jade Learner Retention Model (JLRM) is to reduce the levels of learner dropout in New Zealand's tertiary education organizations (TEOs). The model gives TEOs a solution that enables them to better direct their investment in intervention programs and reduce learner dropout rates. Implementing the JLRM enables TEOs to create targeted intervention programs for learners who are at greater risk of dropping out. This facilitates a higher success rate and a greater return on Investment (ROI) for intervention programs.

Jade Software and the Tertiary Education Commission (TEC) collaborated on research in order to:

- Analyze the TEO data to create a common data set. This enables the JLRM to be applicable across most TEOs in New Zealand.
- Evaluate the JLRM performance across multiple organizations.
- Assess the sensitivity of JLRM to additional data sets.

2 Research Approach

2.1 Defining Learner Dropout

We needed a meaningful definition of learner dropout that was consistent across multiple TEOs. For this research, we used the TEC's definition of a learner not completing their qualification within:

- 2 years for Level 1 to 3 qualifications
- 4 years for Level 4 to 7 non-degree qualifications
- 6 years for Level 7 and higher degree qualifications [1].

2.2 Defining Input Data and Predictors

To develop a common predictive model for most TEOs in New Zealand, we needed to find and use a consistent data set that was being captured by all TEOs.

Single Data Return (SDR)

The SDR was identified as the only common data set available. It's a set of data items that are specifically required by the Ministry of Education and the TEC for funding, monitoring performance, publishing performance information, and statistical reporting purposes [2]. TEOs in New Zealand are required to submit an SDR three times a year.

Using the SDR data, we developed up to 100 variables, including demographics and other measures. Some of these variables were extracted directly from the SDR data set. Three examples of the variables are equivalent full-time student value for the current qualification, highest secondary school qualification, and ethnicity code. Other variables were calculated and aggregated from data points within the SDR [3]. For example, percentage of course completion so far for the current qualification, number of credits passed so far for the current qualification, and total number of completed past qualifications.

To create training and evaluation data sets for the JLRM, we separated the timeline of the SDR data into four components. Figure 1 shows the construction of the data for Level 7 and higher qualifications. Learners whose latest qualification fall between 2009 and 2011 are included in the Training data set, while learners whose latest qualification fall between 2011 and 2012 are included in the Evaluation data set. The four defined components for Level 7 and higher qualifications are:

- Profile: 2007-2009. For every learner we calculated characteristics about their qualifications from the past two years. To achieve this, we left a two-year gap in the timeline between 2007 and 2009 so even the learners with a qualification at the beginning of 2009 would have the same two-year profile period.
- Training: 2009-2011. We used two years of data to train the JLRM. All learners with a qualification between 2009 and 2011 were used to train our model.
- Evaluation: 2011-2012. We evaluated our model by applying the trained model on the 2011 data. All learners with a qualification in 2011 were used to evaluate our model.

- Dropout: 2012-2018. We left a six-year gap in the timeline to allow for dropout to happen that the model needed to predict. Even learners with a qualification at the end of 2012 had the same six-year dropout period.

Profile	Training	Evaluation	Dropout
2007	2009	2011	2012
			2018

Fig. 1. The data components for Level 7 and higher qualifications. We separated the timeline of the SDR data into four components to create training and evaluation data sets.

We only needed a four-year gap in the dropout component for Level 4 to 7 qualifications. This means the Profile, Training, and Evaluation components shifted two years closer to 2018 to include the latest available data. Similarly, the three components for Level 1 to 3 qualifications shifted by another two years closer to 2018.

Additional Data

To supplement the SDR data, we explored additional data that would be consistently available across multiple TEOs. Such data was limited due to the TEOs operating independently of each other outside the common need to submit the SDR data. Attendance and results data were the only two additional data sources that were considered to have potential consistency across multiple TEOs.

Data Combinations

Together with the TEC, we decided to use the data from four TEOs to train and evaluate models. The TEOs selected comprise one vocational college, one university, and two polytechnics. As per our non-disclosure agreement with the TEC, the names of the TEOs are not reported.

From these selected TEOs, three combinations of data were used to train and evaluate the JLRM:

- SDR data only for all four TEOs – This was to determine the model performance for more than one TEO using the standard SDR data set.
- SDR and attendance data for one TEO – This was to evaluate the benefit of using attendance data as an additional data source.
- SDR and results data for one TEO – This was to evaluate the benefit of using results data as an additional data source.

TEC provided SDR data for all four TEOs. We were also given attendance data from one TEO which we appended to their SDR data set. Similarly, we appended the results data from another TEO to their SDR data and evaluated the model performance. Table 1 shows the data sets we received from the four TEOs.

Table 1. Collection of data sources in addition to the SDR data we have for the four TEOs.

TEO (TEO type)	SDR	Attendance	Results
TEO1 (Vocational college)	Yes	Yes	No
TEO2 (Polytechnic)	Yes	No	Approximately 60% recorded
TEO3 (Polytechnic)	Yes	No	No
TEO4 (University)	Yes	No	No

Model Selection and Optimization

Using only SDR data, we built our models on the scikit-learn, h2o, TensorFlow, and R platforms. We explored the Random Forest, Gradient Boosting Machine, Extreme Gradient Boosting, Support Vector Machine, Deep Learning Classifier, and Logistic Regressions models.

We ran several iterations of data transformation and modelling. Gradient Boosting Machine on the h2o platform was confirmed as the preferred model approach. This is because it provided the optimal balance between computational speed, ease of use, ease of interpretation, and model performance.

Having selected the Gradient Boosting Machine, we tuned the hyperparameters in h2o using grid search to find optimal model performance. The hyperparameters we have in the grid search are number of trees, maximum tree depth, learning rate, learning rate reduction, row sampling rate, and column sampling rate [4].

We used cross-validation to quantify the model performance to make sure the model was not overfitted. We applied the cross-validated Gradient Boosting Machine on the Evaluation data set and reported the area under the Receiver Operating Characteristic curve (AUR).

We repeated the process of hyperparameters tuning, cross-validation, and evaluation for all the three data combinations listed above.

2.3 Model Results and Output

The AUR for all four TEOs using only the SDR data is 0.83. This AUR shows that the JLRM is a useful and valuable model for predicting learner dropout across several types of TEOs, using just the SDR data. The eight most important variables are percentage of course completion so far for the current qualification, percentage of credits failed so far for the current qualification, full time indicator, number of credits passed so far for the current qualification, which TEO a learner belongs to, sum of tuition (FTE teaching) and vacation weeks for the current qualification, first character from Funding Category code in SDR, and self-learning hours required for the current qualification.

Including Additional Data

Using only the TEO1's data, adding learner attendance lifts the AUR of the model to 0.86. The AUR also improved slightly by adding the learner's results to the SDR data for TEO2. The improvement is smaller than the attendance data because only 60% of

TEO2’s learners have results recorded against them. Table 2 shows the AUR for the three data combinations.

Table 2. This table shows the AUR for three data combinations. The AUR for all four TEOs using only the SDR data is 0.83. The AUR improved slightly by adding either learner attendance or learner’s results to the SDR data.

Data set	AUR
Four TEOs with SDR only	0.83
TEO1 – SDR only, without attendance	0.80
TEO1 – With attendance	0.86
TEO2 – SDR only, without results	0.79
TEO2 – SDR only, with results	0.80

Return on Investment (ROI) of Intervention

Equation (1) is used to facilitate the TEC in estimating the ROI for the programs and the optimal intervention threshold:

$$ROI = \sum_k return * k * f_1 * f_2 - \left(running\ costs + \sum_l intervention\ cost * l \right) \quad (1)$$

where

- *k* is the total number of true positives included in the intervention threshold,
- *return* is the average value of retaining a learner,
- *f₁* is the percentage likelihood of an intervention succeeding,
- *f₂* is the percentage likelihood that a learner will take up an intervention,
- *l* is the total number of learners included in the intervention threshold,
- *running costs* include annual licence fees and marketing costs,
- *intervention cost* is the average cost of intervening with a learner.

These inputs enable the JLRM to indicate the intervention threshold that would provide the highest ROI. Alternatively, the intervention threshold can be guided by the allocated budget for interventions. Equation (1) can then be used to indicate the ROI at the budgeted intervention threshold.

For the ROI calculation to be accessible by the TEC, we developed a dashboard in Shiny that interacts directly with the Evaluation data set using the inputs above. We use interactive sliders to control the inputs above to correspond to the interventions, while *k* is given by the total number of true positives as predicted by the JLRM within the Evaluation data set. Figure 2 provides a screenshot of the graph outputs within the dashboard.

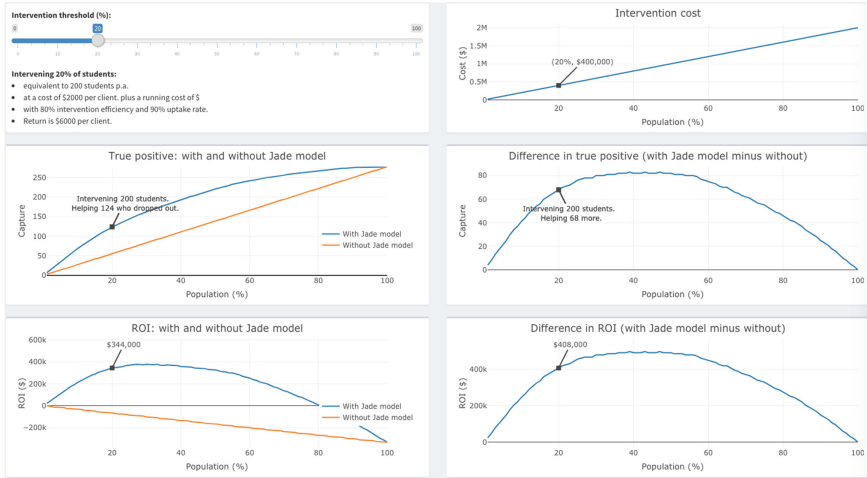


Fig. 2. A screenshot of the Shiny dashboard we created to facilitate the TEC estimating the ROI for the intervention, and the optimal intervention threshold.

3 Conclusions and Discussion

The JLRM provides a way for the TEC or individual TEOs to better target learners that are at a high risk of dropping out. An AUR of between 0.79 and 0.86 across multiple TEOs gives us confidence that the JLRM model would be effective in standardizing how intervention programs are targeted. Any of the task value, framing, or personal value intervention programs can be applied [5]. Alternatively, it can be a tiered approach - for example, less intensive interventions for mid- to high-risk learners and more intensive interventions for high-risk learners.

There are multiple options for implementing the JLRM into the operational process of a TEO or the TEC. When considering the options, we noted that:

- By using SDR data, the TEC can provide a multi-TEO service and direct the output of the JLRM to all, or a subset of, TEOs.
- Individual TEOs could adopt the model internally and improve model performance by including extra data sets.

In all scenarios, Jade would give TEC the output from JLRM, either through a self-service portal or an integration with an existing learner management system.

3.1 How Could the Model Be Improved?

Additional Data Sets

We showed that model performance improves with adding data that isn't present in the SDR data set. As part of the research, we looked at more data points collected by TEOs, including:

- Online course logins and results – These weren't used because they're too new and don't provide enough data to develop models.
- Campus Wi-Fi logs – These aren't consistently recorded or stored over time.

Consistently collecting data about online courses and Wi-Fi use can further improve the JLRM.

Cohort Targeting

The performance of the JLRM against multiple learner cohorts, compared to a single countrywide model, hasn't been assessed. Multiple models might outperform one model in the areas of TEO, level of qualification, and area of study.

However, using cohort-based models would create increased complexity. This is because the cohorts can overlap, thereby generating multiple risk scores for an individual learner.

Data Cadence

The JLRM used SDR data that was submitted at the end of each semester. Increasing the data cadence to run the model with data available during the semester might improve model performance. It can also reduce the time between a learner becoming at risk and an intervention being initiated.

References

1. Tertiary Education Commission: Educational Performance Indicators: Definitions and Methodology For Institutes of Technology and Polytechnics, Private Training Establishments, Universities and Wānanga Version 1.1
2. SDR. <https://www.tec.govt.nz/funding/funding-and-performance/reporting/sdr/>. Accessed 05 June 2019
3. Ministry of Education and Tertiary Education Commission: Single Data Return. A Manual for Tertiary Education Organisations and Student Management System Developers. Version 1.0 (2018)
4. h2o Gradient Boosting Machine. <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html>. Accessed 05 June 2019
5. Harackiewicz, J.M., Priniski, S.J.: Improving student outcomes in higher education: the science of targeted intervention. *Annu. Rev. Psychol.* **69**, 409–435 (2018)