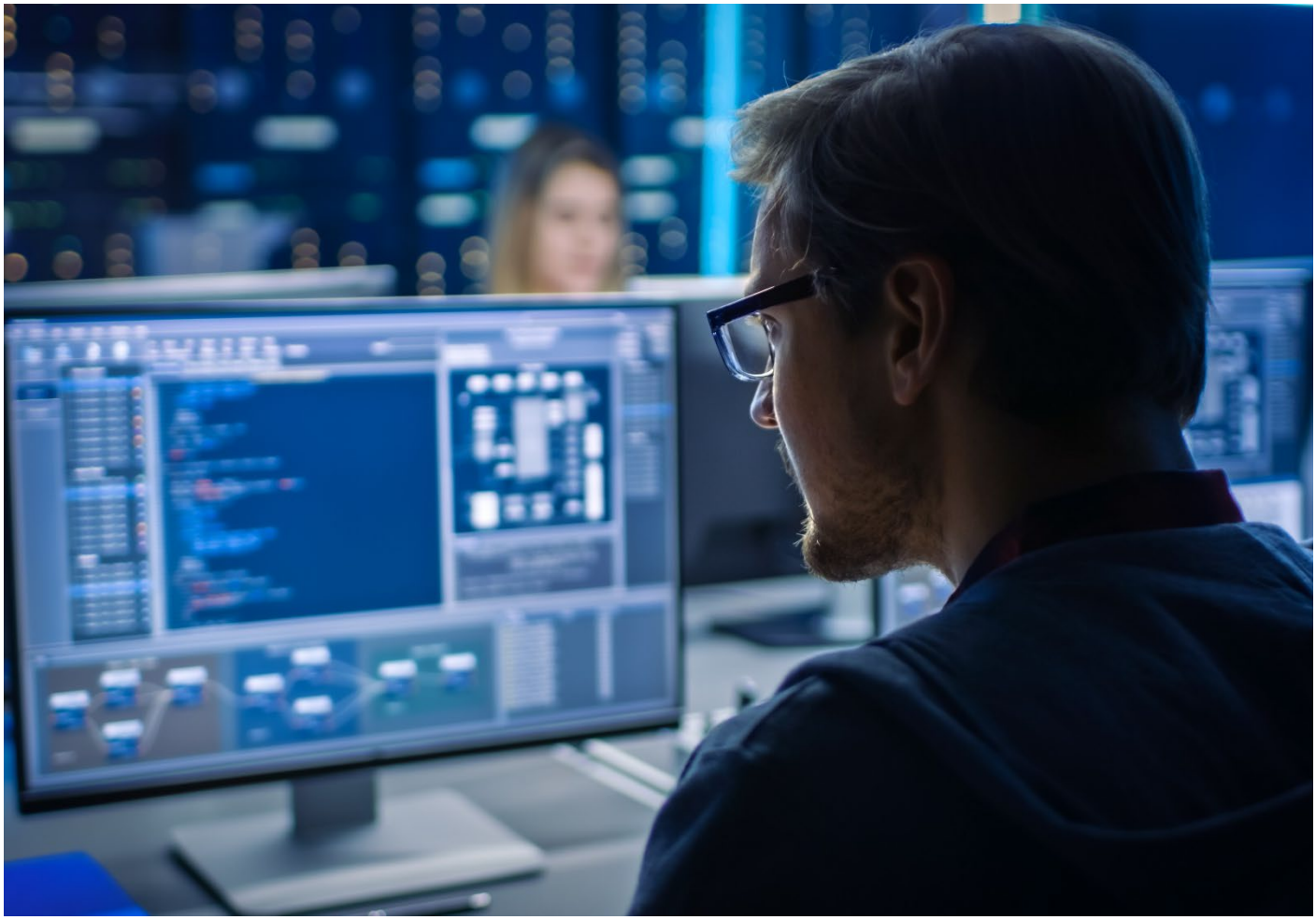
TECHNICAL PAPER

NCVER

# Evaluating machine learning for projecting completion rates for VET programs

**Michelle Hall**
**Melinda Lees**
**Cameron Serich**
National Centre for Vocational Education Research

**Richard Hunt**
Deloitte Australia



INFORMING + INFLUENCING
THE AUSTRALIAN VET SECTOR

## Publisher's note

The views and opinions expressed in this document are those of NCVER and do not necessarily reflect the views of the Australian Government, state and territory governments, or Deloitte. Any interpretation of data is the responsibility of the author/project team.

To find other material of interest, search VOCEDplus (the UNESCO/NCVER international database <http://www.voced.edu.au>) using the following keywords: Comparative analysis; Completion; Evaluation; Outcomes; Participation; Statistical method; Vocational education and training.



This document should be attributed as Hall, M, Lees, M, Serich, C & Hunt, R 2023, *Evaluating machine learning for projecting completion rates for VET programs*, NCVER, Adelaide, 2023.

This work has been produced by NCVER on behalf of the Australian Government and state and territory governments, with funding provided through the Australian Government Department of Employment and Workplace Relations.

COVER IMAGE: GETTY IMAGES
ISBN 978-1-922801-11-1
TD/TNC 151.06

Published by NCVER, ABN 87 007 967 311
Level 5, 60 Light Square, Adelaide SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

**Phone** +61 8 8230 8400 **Email** ncver@ncver.edu.au
Web <**https**://www.ncver.edu.au> <https://www.lsay.edu.au>

**Follow us:** <https://twitter.com/ncver> <https://www.linkedin.com/company/ncver>

# About the research

## *Evaluating machine learning for projecting completion rates for VET programs*

Michelle Hall, Melinda Lees and Cameron Serich, NCVER, and Richard Hunt, Deloitte Australia

This paper summarises exploratory analysis undertaken to evaluate the effectiveness of using machine learning approaches to calculate projected completion rates for vocational education and training (VET) programs, and compares this with the current approach used at the National Centre for Vocational Education Research (NCVER) — Markov chains methodology.

NCVER publishes annual observed VET qualification completion rates for qualifications that commenced four years prior to the most recent data collection period, based on the assumption that sufficient time has passed for all students who intended to complete their qualification to have done so. Projected rates are published for the more recent years, as the actual completion rates cannot be known until enough time has passed for the qualifications to be completed and the outcomes reported to NCVER.

While the Markov chains methodology currently used by NCVER has demonstrated that it is reliable, with predictions aligning well with the actual rates of completion for historical estimates, it has not been reviewed for some time and it does have some limitations. The evaluation of machine learning techniques for predicting VET program completion rates was undertaken to overcome some of these limitations and with a view to improving our current predictions.

This report includes:

- an overview of the methodologies: Markov chains and two machine learning algorithms that were applied to predict completion rates for VET programs (XGBoost and CatBoost)
- a comparison of the accuracy of the predictions generated by both methodologies
- an evaluation of the relative strengths and limitations of both methodologies.

## Key messages

- For the 2016 commencing cohort, the completion rate predictions using machine learning algorithms were generally more accurate than the rates achieved using Markov chains methodology. When evaluated against *actual published* completion rates:
  - The 'XGBoost' machine learning approach produced the most accurate predictions overall, with a high level of recall and precision.
  - The 'XGBoost' machine learning approach also had fewer instances where the prediction for a training attribute deviated from the actual completion rate by more than three percentage points, as compared with the Markov chains methodology.
- Both projection approaches have strengths and limitations:
  - The key advantage of Markov chains theory is that the projected rates are calculated from a three-year period of recent enrolments (and their transitions between enrolment states), without requiring the full history of all qualification enrolments. That said, a key limitation of this methodology is the 12-month delay before projected rates can be calculated, the reason

being that the calculation of the transitional probabilities that form the basis for the completion rate projection for a given year relies on data that includes the following year.

- Markov chains projected completion rates for VET qualifications commencing in the most recent years are overinflated (particularly the current year projections). The alignment of projections to actual rates improves as time passes and as more records reach their final state of 'completed' or 'discontinued'.
- One of the key advantages anticipated by the adoption of a machine learning model for predictions is the timeliness of the predictions. The machine learning model is anticipated to allow projections to be calculated for a new cohort as soon as the enrolment data are received from the various training providers. However, this method relies on a four-year window of historical training activity data to train the model.
- While the results from the machine learning model demonstrate how accurately the model can generate projected rates for the 2016 commencing year, the model's ability to consistently make accurate predictions for other commencing years is as yet untested.
- Due to the significant disruption to the VET sector from the COVID-19 pandemic, it is not clear whether the assumptions underlying either methodology remain valid for the years where training may have been disrupted by the pandemic.

Simon Walker
Managing Director, NCVER

# Acknowledgements

The authors would like to thank the Deloitte Consulting Data & AI team for their input into this project and report. Thanks also goes to the extended project team at NCVER for their involvement.

# Contents

# Tables and figures

## Tables



## Figures

# Introduction

A vocational education and training (VET) qualification completion rate is simply defined as the proportion of VET qualifications that commenced in a given year that are eventually completed. Determining a completion rate requires information on when a student commences a qualification and, ultimately, when a student exits (that is, successfully completes or discontinues). The time taken for a student to exit a VET qualification varies according to factors such as the Australian Qualifications Framework (AQF) level and mode of study.

Observed and projected VET qualification completion rates are published annually by NCVER. Observed 'actual' completion rates are only reported for qualifications that commenced four years prior to the most recent data collection period, based on the assumption that enough time has passed for all students who intended to complete their qualification to have done so.[1] The rates for more recent years are estimated using projection methodology and take into account students who have not yet had sufficient time to complete their qualification. As time passes and program enrolments are re-categorised as 'completed' or 'discontinued', the alignment of projections to actual rates will improve.

The current methodology for calculating projected completion rates, which has been used by NCVER for some time, is presented in detail in Mark and Karmel (2010). The approach uses information about program enrolments over a three-year window (centred on the year of interest), along with the theory of absorbing Markov chains, to derive the probability that a commencing VET program enrolment will eventually be completed.

Markov chains theory offers an advantage, in that it has the property whereby the probability of an entity 'transitioning' from one state to another in successive time periods is not dependent on past transitions. This means we can use knowledge of the 'state' of qualification enrolments across successive years to predict the long-term completion rates without having the full history of all qualification enrolments. Another advantage of the methodology is that it can be readily applied to subsets of the data based on particular student demographics or attributes of the training. This method has been shown to be reliable and it aligns well with actual rates of completion for historical estimates; however, reliability and alignment take time to emerge.

Despite these strengths, the current methodology has some key limitations:

- The Markov chains methodology requires a three-year window of data, centred on the year of interest. This means that projected rates cannot be calculated for a commencing cohort until data are available for the following year.
- The projected completion rates are overinflated for the VET qualifications that commenced in the most recent years (particularly the current year projections).
- The Markov chains methodology has not been reviewed for some time (NCVER 2016). The methodology was introduced at a time when it was difficult to track students (and students within qualifications and registered training organisations [RTOs]) over time in total VET activity (TVA) data. However, with the introduction of the unique student identifier (USI) in 2015, the enduring problem of predicting future completion rates for those enrolments in the current year might be resolved by establishing alternative methods for estimating transition probabilities.

1 Students studying part-time may not complete in four years, but the numbers for this cohort are small enough not to affect calculated rates.

The purpose of this technical paper is to outline exploratory work that has been conducted to evaluate the effectiveness of using machine learning approaches for calculating projected completion rates for VET programs. The actual completion rate will be used as a baseline comparison for evaluation.

# Projection methodologies

## The National VET Provider and National VET in Schools collections

The TVA dataset contains data records from both the National VET Provider and VET in Schools collections. Data are submitted to NCVER annually and contain the most recent information on program enrolments. The data include information on:

- program enrolment and completion events (for example, when they commence or complete)[2]
- the student (for example, demographics)
- the program (for example, the qualification being completed and the level of education)
- the training provider (for example, provider type).

While the TVA dataset is essentially cross-sectional by year, it contains enough information to match data over several years for individual VET students and the qualifications they undertake. Obtaining such a longitudinal dataset allows the use of Markov chains methodology to then calculate projected completion rates.

## Markov chains methodology

The absorbing Markov chains method provides a means of modelling objects that progress through one or more transient states before reaching one or more final (absorbing) states. This methodological approach calculates the probabilities of objects transitioning between the different states and can be used to calculate the probability of an object eventually reaching one of the 'absorbing' states. Markov chains have applications in a wide variety of areas, including science, economics and mathematics. For a theoretical review see Isaacson and Madsen (1976).

NCVER has been using Markov chains methodology to project completion rates for VET qualifications for several years. The methodology draws on data on VET qualification participation, where a training episode is defined as a particular student enrolling in a particular program, in consecutive years. If the student has enrolments in non-consecutive years, these are treated as separate training episodes.[3] Using this methodology, the lifetime of training can be expressed in terms of the probability of transitioning between one state (for example, continuing) to another state (for example, completed).

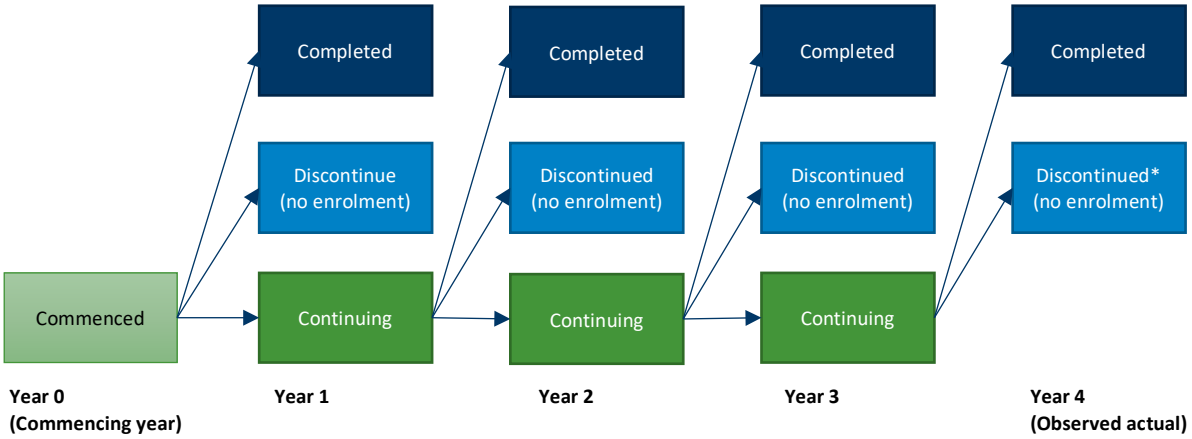
In the Markov chains formula, qualifications can belong to one of four states in a given year:

- commenced
- continuing: they had an enrolment in the previous year and have an enrolment in the current year
- discontinued: they had an enrolment in the previous year and do not have an enrolment in the current year
- completed: there is a record of their completion.

2 Enrolments considered to be withdrawn/dropped out are derived with the following methodology: where no record exists for that program enrolment in the subsequent collection and no award has been issued for that program enrolment.

3 In 2021, 'registered training organisation' was removed as a mandatory linking requirement in NCVER's VET completion rate methodology. This change accounts for transfers of students between RTOs and RTO restructures and provides the ability to match these records from commencement to completion.

**Figure 1 Definition of the states of a VET qualification in the model used to calculate VET qualification completion rates**




Once the qualification has been completed, the state will not change again, so completion and discontinuing are referred to as 'absorbing' or 'final' states. The Markov chains methodology estimates the proportion of qualifications that will eventually reach the completed state, that is, the projected completion rate. The projected completion rate for a given year is based on recent longitudinal data, which include the year for which the completion rate is being projected, the previous year and the following year. For more detailed information on the Markov chains methodology in this context, refer to Mark and Karmel (2010) and McDonald (2018).

A key limitation of the Markov chains methodology is the 12-month delay before projected rates can be calculated. This is because calculating the transitional probabilities that form the basis for the completion rate projection for a given year relies on data from the following year. That is, rates for the most recent year cannot be estimated since there is no following year to provide data.

Another limitation of the Markov chains methodology is that it overestimates projected completion rates for enrolments commencing in the most recent collection year. This is due to:

- the time lag associated with data submissions; there is a delay in reporting completions, meaning that completions occurring each year might take a year or more to be reported
- continuing and discontinued enrolments are not explicitly captured in VET statistics and must be inferred from the data by the presence or absence of an enrolment in the following year.

In practice, this means that the number of continuing enrolments is overestimated, while the number of discontinued enrolments is underestimated. As time passes and subsequent collections are submitted, some of these 'continuing' program enrolments will be re-categorised as 'completed' or 'discontinued' and the alignment of projections to actual rates will improve.

# Machine learning methodology

## Why use machine learning?

While the Markov chains method has shown that it performs fairly reliably when a few years of qualifications data are available, it does require data that are at least one year behind the actual educational situation. Given the 12-month delay before projections can be made with the Markov chains approach, along with its overinflation of projected completion rates calculated for the most recent years, NCVER conducted an evaluation of machine learning techniques, with the aim of improving on the timeliness and accuracy of the Markov chains model results.

One of the key advantages anticipated by using a machine learning model for predictions is the timeliness of the predictions - it is anticipated that the machine learning model will allow projected rates to be calculated in the first year; that is, as soon as enrolment data are received from the various RTOs.

Intellify Pty Ltd (now Deloitte Australia) assisted with our exploration of these options through the application of machine learning approaches to TVA data.

## Algorithms

Machine learning refers to a variety of automated methods whose objective is to uncover patterns in a set of data observations (Murphy 2012). Classification is one form of machine learning, whereby patterns are used to differentiate between two or more categories of interest. Classification methods are a form of supervised machine learning because the categories are made explicit in the example data that are used to 'train' the classification model.

The first step in machine learning classification is to train a model by exposing it to example records from each category of interest. The goal of training is to identify patterns among the features (also known as attributes) of the example records that distinguish between the categories. Following training, the model is evaluated on a set of new records, with the category labels removed. The model's prediction of the category to which each new record belongs is then compared with the true category labels. If the model's performance is satisfactory, it can be used to make predictions for new data records where the categories are not yet known.

Many machine learning algorithms exist and can be applied to different types of machine learning problems. Gradient boosting is a machine learning technique based on decision trees that originated with Friedman (2001, 2002) and over the years it has become a popular technique used for classification as well as regression tasks. The XGBoost algorithm (which is a specific and effective implementation of the gradient boosting approach) has been successfully used for solving classification problems.

Another popular family of techniques include neural networks and deep learning. These methods aim to solve classification problems by representing information at multiple levels of abstraction. In such layered architectures, important features are amplified relative to features that are less useful for distinguishing between the classes of interest (LeCun, Bengio & Hinton 2015).

One useful way of evaluating the popularity and effectiveness of various machine learning alorithms is via the website 'Kaggle' (a machine learning and data science community). Kaggle hosts competitions in machine learning for monetary prizes, meaning that participants are highly motivated to find and use the best algorithms.

Chen and Guestrin (2016) note:

> Among the 29 challenge winning solutions published at Kaggle's blog during 2015, 17 solutions used XGBoost … For comparison, the second most popular method, deep neural nets, was used in 11 solutions.

There is some evidence that deep neural networks have become more popular in Kaggle competitions, but in general it appears they only equal rather than significantly exceed the effectiveness of gradient boosting algorithms generally.

For the purpose of this study, gradient boosting and neural network approaches display relatively similar characteristics. They both require all features to be coded specifically as numeric values, and as noted later in this report interpolation techniques can be used when individual data items are missing or unavailable.

Neural network algorithms do require scaled input data, which is not, on the whole, a requirement of a gradient boosting algorithms; however, this scaling generally happens transparently.

Generally speaking, the training process for a gradient boosting algorithm is considerably simpler to implement and monitor than that for a neural network.

It should also be noted that it can be more difficult to understand why a given neural network model would make the prediction that it does, compared with a gradient boosting model, since the inner workings of the neural network can be convoluted. However, modern tools are available that allow machine learning practitioners to understand which features have driven the model to give the individual predictions (for example, 'SHAP', Lundberg & Lee 2017). These tools can be used for both neural networks and gradient boosting models.

In this study, a gradient boosting algorithm was selected for evaluation over a neural network, given the considerably simpler implementation and faster training.

## Experimental process

When using machine learning, it is important that the training and validation data are as representative as possible of the data that the final model will use for its predictions. The data used for the model development (training and validation) were selected from the 2016 cohort of TVA data and the attributes, as reported at enrolment, were extracted (for example, student residence as at the time of enrolment, ignoring any changes to state of residence during students' training). Furthermore, the data were de-identified to preserve each student's anonymity.

After four years, it could be expected that almost all of the students have either completed or discontinued their programs, so the situation for each student as at 2021 was used to identify the long-term outcome for the student.

These data were then randomly divided into a training sample (80%) and validation sample (20%). Data from the validation samples were not included when training the models, and evaluation of the models was undertaken using the validation data only. A variety of models were built from the training data and their performance was evaluated and compared using the validation samples.

It is important to note that, while NCVER currently publishes completion rates for VET qualifications (training package qualifications and accredited qualifications), TVA data also capture enrolments in other types of VET programs (training package skillsets and accredited courses). This experimental process was conducted on a dataset that included all program types.

Note that the later section 'The COVID-19 pandemic' details the impact that the pandemic may have had on data quality; however, it was felt that most students from the 2016 commencing cohort would have completed or withdrawn from their studies by the start of the pandemic, so it seems unlikely the pandemic had a significant effect on these results.

These data are of course collected from a wide variety of RTOs, each of which has different approaches to collecting and storing their data. Although NCVER goes to great lengths to ensure consistency of definitions and that all data are collected, there are inevitably some areas where one or more data elements might not be available. A variety of mean-value interpolation techniques were used to interpolate those missing data elements – these are detailed in appendix B.

Two variants of a gradient boosting algorithm were examined for this study: XGBoost (Chen & Guestrin 2016) and CatBoost (Prokhorenkova et al. 2018).

As with many new machine learning algorithms, gradient boosting algorithms possess ‘hyper-parameters’, whose value can impact on the accuracy of the model to a greater or lesser extent. Machine learning practitioners often use one of a variety of hyper-parameter ‘tuning’ algorithms to identify a set of hyper-parameters that seem to allow the model to represent the training data most accurately, and in this case the ‘Optuna’ package was used to perform hyper-parameter tuning (Akiba et al. 2019).

The experiments were performed on a DataBricks cluster running on Microsoft Azure using the ‘scikit-learn’ infrastructure (Pedregosa et al. 2011).

# Comparing the accuracy of projected completion rates

This section presents a comparison of the performance of the Markov chains and machine learning projection models described previously. All results presented here for the Markov chains model relate to the 2016 commencing cohort, as calculated from the 2017 collection; and all results presented for the machine learning models are for the validation subset of the 2016 commencing cohort, as calculated from the 2016 collection. As noted earlier, one of the key limitations of the Markov chains method is the 12-month delay before projected rates can be calculated.

Based on data for the 2016 commencing cohort, the Markov chains methodology provides overinflated projections for the VET qualifications commencing in the most recent years, whereas the XGBoost and CatBoost models appear to perform very well. The accuracy of the Markov chains projected rates improves as time passes (not shown in this report).

As noted in the previous section, the machine learning models were trained on data that included training package qualifications, accredited qualifications, training package skillsets and accredited courses. To allow comparison with the Markov chains methodology, the results presented here show predictions for training package qualifications and accredited qualifications only. These results for the 2016 metrics are determined from the validation data only, and do not include the training data (table 1).

**Table 1 Comparison of overall results for Markov chains and machine learning methodological approaches, 2016 commencing cohort (%)**

| | Published actual completion rate | Predicted completion rate | Precision | Recall |
|---|---|---|---|---|
| **Markov chains** | 43.8 | 46.9 | na | na |
| **XGBoost** | 43.8 | 44.9 | 78.5 | 78.7 |
| **CatBoost** | 43.8 | 45.1 | 77.9 | 78.1 |

Note: The published actual completion rate is based on the 2021 collection (NCVER 2022).

The Markov chains predicted completion rate (NCVER 2018) and the machine learning predicted completion rates are based on the projected completion rate for the 2016 commencing cohort. For Markov chains, this rate was calculated from the 2017 collection (i.e. data from the year after commencement). For both machine learning models, the rate was calculated from the 2016 collection (i.e. data from the commencing year only).

These results indicate only small differences between the performance of the machine learning algorithms, but the XGBoost algorithm does perform slightly better than the CatBoost algorithm. XGBoost also outperforms the Markov chains methodology, with a result closer to the actual completion rate (1.1 percentage-point difference compared with the actual rate as opposed to 3.1 percentage-point difference for Markov chains).[4] The higher percentage-point difference for the Markov chains method reflects the overinflation typical of projections calculated in the first year using this method. With a further year of delay, the projected rate of the Markov chains improves to a 0.9 percentage-point difference when calculated from the 2018 collection and includes data from the two years after commencement (projected rate 42.9; NCVER 2019). Based on these results, XGBoost was chosen as the preferred model.

Note that the term 'Precision' used in table 1 refers to the proportion of 'True Positives' among the 'Predicted Positives'; that is, the proportion of all predicted completions that were correctly identified

4 The 2016 Markov chain projected completion rate of 46.9 was first published in the VET program-completion rates 2016 publication: <https://www.voced.edu.au/content/ngv%3A80349>.

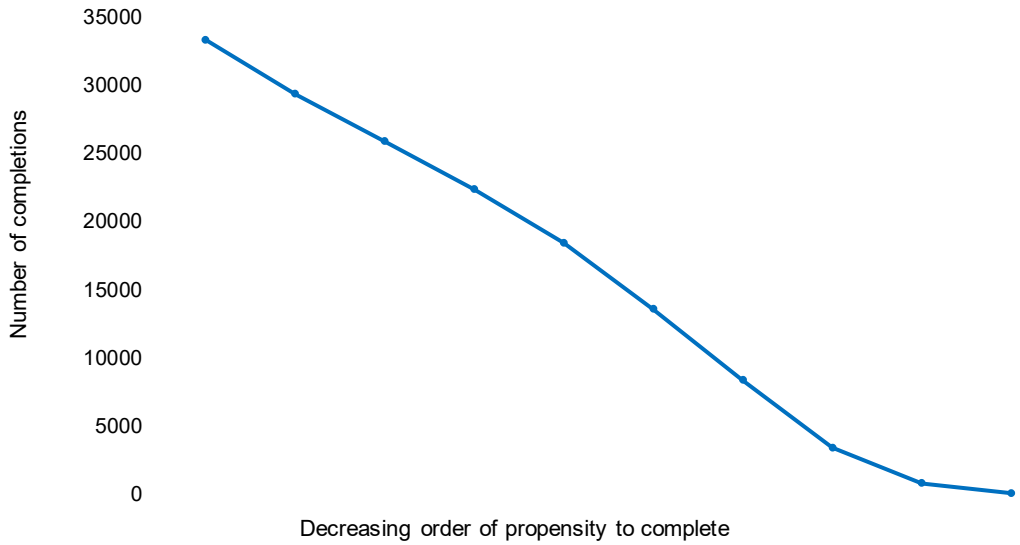
by the algorithm. The 'Recall' is the proportion of all qualifications that were actually completed that were correctly identified by the algorithm.

There is a characteristic trade-off between precision and recall: a model would have perfect recall but very poor precision if it predicted that every qualification enrolment would result in a completion. Such a model would correctly identify every actual completion; however, it would also have a high rate of false positives because every non-completed qualification would be misclassified as a completion. On the other hand, a model would have near-perfect precision but very low recall if it only predicted completions for the very few enrolments that were overwhelmingly likely to result in completion but missed many other enrolments that also resulted in a completion.

From table 1 it can be seen that precision and recall are well balanced for both the XGBoost and CatBoost models, indicating that these models are not biased towards false positives over false negatives (or vice versa). This is desirable for the completion rates model because the purpose of the model is to estimate the average completion rate for the cohort. As before, the XGBoost model slightly outperforms the CatBoost model in terms of both precision and recall.

A 'lift' chart can be used to visualise the effectiveness of the model in predicting the correct outcomes, shown in figure 2 for the XGBoost model. The lift chart demonstrates how accurate the model is in generating predictions for the validation data only (which the model has not 'seen'). It clearly shows that the group of enrolments with the lowest predicted propensity to complete had the fewest actual completions, the group of enrolments with the highest predicted propensity to complete had the highest number of actual completions, and so forth.

**Figure 2 'Lift chart' showing alignment between the propensity to complete according to the XGBoost model predictions and actual qualification completions**




Note: The 'lift chart' compares the model predictions with actual published completion rates for each data decile. Deciles are descending orders of 'probability of completion'. Model predictions are based on the projected completion rate for the 2016 commencing cohort, while the actual published completion rates are based on the 2021 collection (NCVER 2022).

Table 2 presents predicted and actual completion rates split by various training characteristics. There were only four instances where the completion rate predicted by the XGBoost model deviated by more than three percentage points from the actual completion rate. This compares with nine instances where the completion rate predicted by the Markov chains model deviated by more than three percentage points from the actual completion rate.

For the XGBoost predictions, the greatest deviations from the actual completion rate occurred for: funding source — *international-fee-for-service* (8.1 percentage-point difference); training provider type — *enterprise providers* (5.3 percentage-point difference) and *community education providers* (5.2 percentage-point difference).

For the Markov chains model predictions, the greatest deviations from the actual completion rate occurred for: training provider type — *schools* (6.8 percentage-point difference) and *community education providers* (6.4 percentage-point difference); funding source — *international fee-for-service* (6.1 percentage-point difference); and qualification level — *certificate IV* (5.1 percentage-point difference).

**Table 2 Comparison of Markov chains and machine learning results by training characteristics (%)**

| | Published actual completion rate (a) | Markov chains predicted (b) | XGBoost predicted (c) | Percentage-point difference between (a) & (b) | Percentage-point difference between (a) & (c) |
|---|---|---|---|---|---|
| **Qualification level** | | | | | |
| Diploma or above | 45.5 | 48.5 | 46.3 | 3.0 | 0.8 |
| Certificate IV | 48.8 | 53.9 | 50.3 | 5.1 | 1.5 |
| Certificate III | 45.9 | 48.2 | 47.6 | 2.3 | 1.7 |
| Certificate II | 40.5 | 44.0 | 41.7 | 3.5 | 1.2 |
| Certificate I | 29.9 | 31.7 | 28.2 | 1.8 | -1.7 |
| **Funding source** | | | | | |
| Government-funded | 47.3 | 49.7 | 48.8 | 2.4 | 1.5 |
| Domestic fee-for-service | 36.0 | 39.3 | 35.5 | 3.3 | -0.5 |
| International fee-for-service | 64.2 | 70.3 | 72.3 | 6.1 | 8.1 |
| **State/territory of training delivery location** | | | | | |
| New South Wales | 44.8 | 49.1 | 46.3 | 4.3 | 1.5 |
| Victoria | 40.0 | 44.6 | 40.2 | 4.6 | 0.2 |
| Queensland | 45.1 | 47.1 | 47.5 | 2.0 | 2.4 |
| South Australia | 39.2 | 39.2 | 35.7 | 0.0 | -3.5 |
| Western Australia | 46.9 | 47.8 | 48.4 | 0.9 | 1.5 |
| Tasmania | 39.0 | 40.2 | 39.1 | 1.2 | 0.1 |
| Northern Territory | 38.1 | 40.0 | 35.5 | 1.9 | -2.6 |
| Australian Capital Territory | 46.1 | 47.6 | 47.6 | 1.5 | 1.5 |
| **Training provider type** | | | | | |
| TAFEs | 42.8 | 43.2 | 41.7 | 0.4 | -1.1 |
| Universities | 46.3 | 48.6 | 48.2 | 2.3 | 1.9 |
| Schools | 46.7 | 53.5 | 49.0 | 6.8 | 2.3 |
| Community education providers | 41.0 | 47.4 | 46.2 | 6.4 | 5.2 |
| Enterprise providers | 53.1 | 52.0 | 58.4 | -1.1 | 5.3 |
| Private training providers | 43.6 | 47.5 | 45.3 | 3.9 | 1.7 |
| **Total** | **43.8** | **46.9** | **44.9** | **3.1** | **1.1** |

Note: The published actual completion rate is based on the 2021 collection (NCVER 2022).

The Markov chains predicted completion rate (NCVER 2018) and the machine learning predicted completion rates are based on the projected completion rate for the 2016 commencing cohort. For Markov chains, this rate was calculated from the 2017 collection (i.e. data from the year after commencement). For both machine learning models, the rate was calculated from the 2016 collection (i.e. data from the commencing year only).

## Limitations of the analysis

The XGBoost results demonstrate how accurately the model can generate projected rates for the 2016 collection year validation data. To understand the model's ability to consistently make accurate predictions, the model will need to be re-run on future cohorts (for example, the 2017 commencing cohort) and comparisons made against actual data (for example, from the 2021 collection).

As noted, the Markov chains method has previously been applied to VET training package qualifications and accredited qualifications. A second aim of this project was to explore whether the machine learning methodology could be applied to all nationally recognised VET programs; that is, training package skillsets and accredited courses, in addition to the training package qualifications and accredited qualifications currently modelled. While the machine learning model was trained on all accredited training (qualifications, courses and skillsets), the outputs for non-qualifications was not fully explored, with further investigation needed.

It is also important to note that the machine learning approach described in this technical paper is experimental, and it is not yet possible to evaluate its applicability to future training activity, once the disruptions of the COVID-19 pandemic no longer apply. The possible impact of the pandemic on the model is explored further in this report.

# Other considerations and future directions

The previous section presented an evaluation of the two machine learning algorithms — XGBoost and CatBoost — by comparison with the existing Markov chains methodology for predicting completion rates for VET qualifications. This section discusses some considerations associated with the use of machine learning algorithms, including the importance of data quality monitoring, model maintenance, the responsible use of machine learning, possibilities for refining the model/s in the future, and potential impacts of the COVID-19 pandemic on model performance.

## Data quality and data drift

Before data are analysed, they should be carefully examined and compared with the reference data (for example, 2016 data) to identify evidence of statistical deviations from the norm. While some deviation would be normal — particularly if the student counts are low — significant deviations in larger student counts should be examined carefully.

Data drift occurs when the statistical properties of the predictors change. For instance, the completion rate of a particular program might be quite different from one year to the next. In other words, the element we are trying to predict may change.

However, it might also be the case that a given program may change from being primarily male-only to a more balanced male—female combination. Since there can be differences in completion rates between genders, a change to the completion rate for the course overall may result. For this reason, any differences in the completion rates between the reference dataset and the new dataset will be analysed based on:

- age
- gender
- program of study
- state of residence.

## Model drift

Model drift occurs when the underlying relationships between the various data fields used in the model change. For instance, if new program requirements were enacted for some programs of study and these affected a student's likelihood of completing the program of study, then this would be an example of the existing machine learning model no longer representing the relationships in the data. Solutions for the problem of model drift can involve retraining the model or, in some cases, using a different model algorithm.

The accuracy of the model will be regularly assessed and if model drift is detected (as distinct from data drift), then model retraining or model rebuilding activities will be undertaken to retain the accuracy of the predictions.

## Algorithmic fairness

In recent years, as the adoption of machine learning algorithms has become common, the fairness of these algorithms has become a topic of considerable interest.

The Commonwealth Department of Industry, Science, Energy and Resources has published Australia's 'Artificial Intelligence Ethics Framework', which consists of eight ethics principles. The fairness principle states that:

> Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

This includes concerns regarding discrimination based on gender, race, disability and age.

For the models described in this paper, only aggregate results are released for public consumption.

The XGBoost algorithm selected does provide differing predicted completion rates based on sensitive variables such as gender and age. To evaluate the fairness of the algorithm we would need to understand the extent to which the use of the predictions might amount to 'unfair discrimination'.

For instance, suppose that the completion rates for Indigenous students were found to be significantly lower for a particular course in a particular training organisation. In these circumstances, advice should be sought before any action is taken.

Since the fairness of an algorithm depends on the intended purpose of the predictions from the algorithm, it is the responsibility of the users of the predictions to ensure that such use does not amount to 'unfair discrimination'.

## Refining the algorithms

As described above, there are numerous reasons why it may be desirable to refine the algorithms.

Over time, the data on enrolments may 'drift' and exhibit relationships that are not reflected in the 2016 cohort data — the data used to train this model.

New sources of funding or different types of training may emerge, while adjustments in government policy may mean that the entry requirements for certain courses may change. Another change might be, for example, the provision of additional support for particular student groups. Any variations such as these have the potential for the existing model to underperform and require refinement.

Model performance could be further improved with the inclusion of additional behavioural variables, separate from the administrative total VET activity data. It may be useful to look at the other data features available from the Australian Census or other data sources.

Retraining the model should in any case be undertaken when performance drops, although a larger exercise could be conducted to re-examine the suitability of the existing set of features, to develop new features for the model, or even to investigate new algorithms, given that the area of machine learning is rapidly changing and new algorithms are becoming available quite frequently.

## The COVID-19 pandemic

The projection methodologies explored in this technical paper draw upon recent historical data trends to make inferences about current or future training patterns. The Markov chains methodology relies on the assumption that the transitional probabilities of recent cohorts are likely to be similar to the transitional probabilities of the current cohort (for which a projection is being made). In the case of machine learning, the assumption is that the factors affecting completion and non-completion, and the extent to which those factors affect them, are similar for the current cohort and the cohort on which the model was 'trained'.

In the context of the COVID-19 pandemic, it is not clear whether these assumptions are reasonable, because training behaviour is likely to have been substantially disrupted. New factors, with unknown and possibly transient impacts on VET activity, which may not be captured in NCVER data, are likely to exist (for example, eligibility for government support).

It is important to note that the machine learning approach described in this technical paper is experimental, and it is not yet possible to evaluate whether it will be generalisable to future training activity, once the disruptions of the pandemic no longer apply.

# References


Akiba, T, Sano, S, Yanase, T, Ohta, T & Koyama, M 2019, 'Optuna: a next-generation hyperparameter optimization framework', viewed 19 February 2023, <https://dl.acm.org/doi/10.1145/3292500.3330701>.

Chen, T & Guestrin, C 2016, 'XGBoost: a scalable tree boosting system', viewed July 2022, <https://doi.org/10.1145/2939672.2939785>.

Friedman, JH 2001, 'Greedy function approximation: a gradient boosting machine', *The Annals of Statistics*, vol.29, no.5, pp.1189—232, viewed 19 February 2023, <http://www.jstor.org/stable/2699986>.

Friedman, JH 2002, 'Stochastic gradient boosting', *Computational Statistics & Data Analysis*, vol.38, issue 428, pp.367—78, viewed 19 February 2023, <https://doi.org/10.1016/S0167-9473(01)00065-2>.

Isaacson, DL & Madsen, RW 1976, *Markov chains: theory and applications*, John Wiley & Sons, New York.

LeCun, Y, Benigo, Y & Hinton, G 2015, 'Deep Learning', *Nature*, vol.521, no.7553, pp.436—444.

Lundberg, SM & Lee, S 2017, 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems* (NIPS'17), vol.30, Curran Associates Inc., Red Hook, NY, pp.6639—49.

Mark, K & Karmel, T 2010, *The likelihood of completing a VET qualification: a model-based approach*, NCVER, Adelaide, viewed 12 April 2021, <https://www.ncver.edu.au/research-and-statistics/publications/all-publications/the-likelihood-of-completing-a-vet-qualification-a-model-based-approach>.

McDonald, B 2018, *Total VET program completion rates*, NCVER, Adelaide, viewed 19 February 2023, <https://www.ncver.edu.au/research-and-statistics/publications/all-publications/total-vet-program-completion-rates>.

Murphy, KP 2012, *Machine learning: a probabilistic perspective*, MIT Press, Cambridge, MA.

NCVER (National Centre for Vocational Education Research) 2016, *VET program completion rates: an evaluation of the current method*, NCVER, Adelaide, viewed 19 February 2023, <https://www.ncver.edu.au/publications/publications/all-publications/vet-qualificationcompletion-rates-an-evaluation-of-the-current-method>.

——2018, *VET program completion rates 2016*, NCVER, Adelaide, viewed December 2022, <https://apo.org.au/sites/default/files/resource-files/2018-08/apo-nid186476_1.pdf>.

——2019, *VET program completion rates 2017*, NCVER Adelaide, viewed December 2022, <https://www.voced.edu.au/content/ngv%3A84181>.

——2022, *VET qualification completion rates 2021*, NCVER, Adelaide, viewed December 2022, https://www.ncver.edu.au/research-and-statistics/publications/all-publications/vet-qualification completion-rates-2021.

Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, Vanderplas, J, Passos, A, Cournapeau, D, Brucher, M, Perrot, M & Duchesnay, E 2011, 'Scikit-learn: machine learning in [P}ython', *Journal of Machine Learning Research*, vol.12, pp.2825—30.

Prokhorenkova, L, Gusev, G, Vorobev, A, Dorogush, A & Gulin, A 2018, 'CatBoost: unbiased boosting with categorical features, Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), Curran Associates Inc., Red Hook, NY, pp.6639—49.

# Appendix

## Appendix A: Data requirements for calculating projections

### Data requirements for Markov chains

To calculate the projected completion rate for the 2016 commencing cohort, the Markov chains approach uses transitional probabilities, as calculated from the projection year (2016), the following year (2017) and the preceding year (2015). This methodology uses a three-year window of recent historical data to calculate transitional probabilities. Data from the following year are required for the calculations, which means there is a 12-month delay before a projection can be made for any given projection year.

For the Markov chains methodology, the data sources used to calculate the projected completion rates for annual cohorts commencing in 2016 are provided in table A1.

**Table A1 Data sources: Markov chains**

| Projection cohort | Source collection | Reference cohort/s |
|---|---|---|
| 2016 | TVA Collection 2015 | TVA accredited qualification program enrolments only |
| | TVA Collection 2016 | TVA accredited qualification program enrolments only |
| | TVA Collection 2017 | TVA accredited qualification program enrolments only |

Note: TVA accredited qualification program enrolments only includes training package qualifications and accredited qualifications (excludes all skillsets, accredited courses and non-nationally recognised programs).

### Data requirements for machine learning

To calculate the projected completion rate for the 2016 commencing cohort, the machine learning methodology uses data from the 2016 commencing cohort. A four-year period of historical data is required to train the model before projections can be made. Data on the terminating information (discontinued or completed) are not used to calculate the projected rate; they are only used for training and evaluation. Machine learning model performance can be evaluated using different metrics, including (but not limited to) accuracy, precision and recall.

The model also uses historical estimates from the Student Outcomes Survey (SOS) at the program level and at the training organisation level (refer to appendix C).

The rates projected for the validation subset of the 2016 commencing cohort, along with the evaluation metrics, give a preliminary indication of the robustness of the application of machine learning in predicting program completions.

For the machine learning methodology, the data sources used to calculate projected completion rates for annual cohorts commencing in 2016 are provided in table A2.

**Table A2 Data sources: machine learning**

| Projection cohort | Source collection | Reference cohort/s |
|---|---|---|
| 2016 | TVA Collection 2016 | TVA accredited qualification program enrolments only |

Note: TVA accredited qualification program enrolments only include training package qualifications and accredited qualifications (excludes all other skillsets, accredited courses and non-nationally recognised programs).

# Appendix B: Interpolation techniques

Missing values in the data are common. The machine learning and statistical communities have long identified a very large number of methods for handling these.

One of the simplest and yet usually effective techniques was used for this project.

If missing values were found in a column, then an average value from the training data was calculated and this was substituted for the missing value.

For some columns, however, this method was modified to provide a slightly more sophisticated approach.

Student age at commencement was one such column. Rather than substitute an average figure for the student's age, the algorithm used was to separately calculate an age for each program. These figures were then used for both training and future predictions. Further, if there was no average value for a particular program, then an average by training provider type was used.

Features from the Student Outcomes Survey were also used. Where these were missing, an average was calculated across the training data, broken down by qualification, field of education and training package.

The birth month is another feature used in the model. If this was missing, then the most common birth month was substituted (which was June).

For categorical features in the data, where there were missing values, a new category was set up to represent the missing values (if this was not already defined in the data).

# Appendix C: Predictor variables included in the XGBoost machine learning model

**Table C1 Predictor variables included in the XGBoost machine learning model**

| **TVA collected variables** |
|---|
| Student age at commencement |
| ABS ANZSCO code that identifies the expected occupational outcome |
| Whether a student is undertaking some training under an apprenticeship/traineeship contract |
| Whether a student is currently enrolled in secondary school |
| Student's Indigenous status |
| The country where the student was born |
| Program identifier (not accounting for program supersession) |
| Whether a student has declared they have a disability, or disabilities |
| Program field of education (FOE) at the narrow level (4-digit), which is one part of the Australian Standard Classification of Education (ASCED), ABS catalogue no.1272.0, 2001 |
| Funding source for the program enrolment (i.e. government funded, domestic fee-for-service, international fee-for-service) |
| The highest level of education, including post-compulsory education, a student had successfully completed before commencing training |
| Labour force status at commencement |
| Main language other than English spoken at home by the student |
| A student's self-assessment of their level of ability to speak English |
| Whether the student had any offshore full-fee-paying activity |
| Level of education for the program (e.g. diploma or higher, cert. IV, cert. III etc.) |
| Remoteness of student residence (i.e. major city, inner regional, outer regional etc.) |
| Socio-Economic Index for Areas (SEIFA) – relative socio-economic advantage and disadvantage |
| Student gender |
| The state or territory where the client resides at commencement of the program |
| The state or territory where the training has been delivered at commencement of the program |
| Type of training provider (e.g. TAFE, private training provider, university etc.) |
| Training package identifier |
| Identifies the type of qualification (i.e. training package qualification, accredited qualification) |
| **Derived variables** |
| Student birth month |
| Whether the student was born in Australia |
| A measure of the enrolled program popularity, derived using counts of program enrolments |
| Whether the state or territory of training delivery was the same as the student's state of residence |
| **Student Outcome Survey (SOS) estimates – historical by program** |
| Proportion of graduates who achieved their main reason for undertaking their training |
| Proportion of graduates who were enrolled in further study after training |
| Proportion of graduates who had improved employment circumstances after training |
| Proportion of graduates who received at least one job-related benefit from their training |
| Proportion of graduates who were employed after training |
| Proportion of graduates who were not employed before training, but employed after training |
| Proportion of graduates who recommend their training provider |
| Proportion of graduates who recommend their program |
| Proportion of graduates who found their training relevant to their job after training |
| Proportion of graduates who improved their problem-solving skills |
| Proportion of graduates who were satisfied with the teaching |

| |
|---|
| Proportion of graduates who improved their writing skills |
| Proportion of graduates who were employed at a higher skill level following their training |
| **Student Outcome Survey (SOS) estimates – historical by training provider** |
| Proportion of graduates who achieved their main reason for undertaking their training |
| Proportion of graduates who were enrolled in further study after training |
| Proportion of graduates who had improved employment circumstances after training |
| Proportion of graduates who received at least one job-related benefit from their training |
| Proportion of graduates who were employed after training |
| Proportion of graduates who were not employed before training, but employed after training |
| Proportion of graduates who recommend their training provider |
| Proportion of graduates who recommend their program |
| Proportion of graduates who found their training relevant to their job after training |
| Proportion of graduates who improved their problem-solving skills |
| Proportion of graduates who were satisfied with the teaching |
| Proportion of graduates who improved their writing skills |
| Proportion of graduates who were employed at a higher skill level following their training |
| Proportion of graduates who were satisfied with their assessment |

Note: Student Outcome estimates are based on the 2017 National Student Outcomes Survey, which is an annual survey of students who completed their vocational education and training (VET) in Australia during the previous calendar year.

**National Centre for Vocational Education Research**

Level 5, 60 Light Square, Adelaide, SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia
**Phone** +61 8 8230 8400 **Email** ncver@ncver.edu.au
**Web** <https://www.ncver.edu.au> <http://www.lsay.edu.au>

**Follow us:** <https://twitter.com/ncver> <https://www.linkedin.com/company/ncver>