Capstone Project Report

Drew Matheson

10/25/2017

An Introduction to the Problem

In New Zealand, National Standards are a national benchmark for student progress throughout the primary school years. The standards set clear expectations that students need to meet in reading, writing, and mathematics in the first eight years at school. At the end of each year of schooling teachers use a range of student achievement evidence to judge whether a student is 'At', 'Below', or 'Above' the expected National Standard.

According to Ministry of Education data from 2015, 84.5% of students after one year at school are achieving 'At' or 'Above' the National Standard and this drops to 74.2% of students at the end of Year 3 [(Ministry of Education)] (http://www.educationcounts.govt.nz/statistics/schooling/national-standards/National_Standards).

Is it possible that some of the students who are judged as being 'Below' the expected National Standard after 3 years could have been 'At' if they had been identified after 1 year and had some intervention put in place to support or accelerate their progress?

The purpose of this analysis is to identify key early warning signs for students at risk of not achieving the National Standard at the end of Year 3, and therefore taking action to support their progress and learning so that they may have the best chance of achieving 'At' the National Standard.

Client

New Zealand school teachers and leaders are the intended client.

New Zealand school teachers and leaders strive to provide an educational experience that caters for all students and supports them all to achieve educational success. In this case 'educational success' is measured by achieving 'At' or 'Above' the expected National Standard.

The assessment and achievement data used in this analysis is already being collected by teachers and schools. If we are able to use this data to identify students that are at risk of achieving 'Below' the National Standard then teachers will be able to action additional support and targeted teaching to ensure these students are more likely to attain the expected National Standard.

The Data

All data in this project will come from one school with a roll size of 485 students. There is data for approximately 150 students who have been assessed at both time 1 (end of 1 year at school) and time 2 (end of Year 3). All data will be anonymized.

The data used in this project come from 2 different assessment tools used in New Zealand primary schools.

The first is the JAM (Junior Assessment in Mathematics) assessment tool which is used to assess children in mathematics in the first two years of school (Year 1 and Year 2). The JAM tool assess students in 11 aspects (domains) of mathematical knowledge and strategies. For this analysis we will use the first 7 domains which are the number strategies and number knowledge aspects. We will use the JAM results from the end of one year at school.

The second assessment is an overall teacher judgement (OTJ) of student progress and achievement up to a particular point in time, made using the National Standards. For this analysis we will use the mathematics National Standards OTJ from the end of Year 3.

Other categorical data included are student gender and ethnicity.

Important Fields in the Data

The 7 domains from the JAM Assessment are key parts of this data set. For each domain a student is marked as being at a certain stage, from Stage 1 (S1) to Stage 5 (S5). Stage 1 is the most basic and Stage 5 the more complex or higher level thinking.

They domains are as follows:

- Additive Strategies - how a student solves addition and subtraction problems
- Numeral Identification - reading numerals
- Forward Number Sequence - counting forwards and identifying the number that comes after a given
- Backwards Number Sequence - counting backwards and identifying the number that comes before a given
- Fraction Knowledge - identifying fractions of shapes and reading fraction symbols
- Place Value Knowledge - understanding place value and groupings of numbers
- Basic Facts Recall - quick recall of basic addition and subtraction facts

The National Standards OTJ which is assigned to each student indicates if they have achieved the expected standard or not. Students are judged as being 'Well Below', 'Below', 'At' or 'Above' the expected National Standard. For this project we will be separating these students into one of two groups: those achieving the standard ('At' and 'Above'), and those who are not achieving the standard ('Below' and 'Well Below'). This is our dependent variable that we hope to predict using the other independent variables.

Challenges in the data set

The main limitation with this data set is the small size. The initial data set which included all students from the school had 485 observations. Due to the fact that to we wanted to create a predictive model that would predict if students would achieve 'At' the standard after 3 years at school, we had to filter the data set to include only those students who had already reached that 3 years at school milestone. Once we did this, and removed any incomplete cases the tidy data set was reduced to 147 observations. This small data set will present some challenges in the machine learning and modelling stage of the project.

Approach to solving this problem

The JAM assessment from the end of 1 year at school marks a student at a stage (from Stage 0 - Stage 5) for each of the 7 modules in the assessment.

The National Standard judgement from the End of Year 3 will mark a student as either 'Below', 'At', or 'Above'.

An appropriate regression model will be applied to see how we might predict the National Standard judgement for a student at the End of Year 3 using the JAM assessment results from the end of 1 year at school.

The Data Wrangling

The data came in as 2 csv files exported from the school management system. The way in which the school management system exported the data meant that the data was not tidy and did require some wrangling.

The first was a file containing the (anonymous) JAM test results of the school of 489 pupils.

In the original csv file there were 69 columns. Each aspect of the JAM assessment took up 6 columns, one for each year from 2012 to 2017. All students are assessed with this tool twice, once after 1 year at school and again after 2 years at school, therefore most students had values in two columns (two years) for each aspect (e.g. they were assessed in 2012 and then 2013) and there were many empty cells in the data frame.

Because we were interested in only the assessment results from after one year at school, and we were not interested in the year (date) in which they were assessed, the first part of the data wrangling was to unite the 6 columns for each aspect of the assessment and extract the first instance of a stage being assigned (i.e the first time they were assessed: after 1 year at school).

The second csv file contained the the OTJ (Overall Teacher Judgement) for all 489 pupils after 3 Years at school. This indicated if a student was judged as achieving 'Well Below', 'Below', 'At', or 'Above the expected National Standard. Again, this data had 6 columns (one for each year from 2012 to 2017) and each observation had a judgement in only one of these 6 columns. We needed to unite these columns and then bind this data frame with the JAM results data frame. Because some of the students had not yet reached this milestone there was missing data.

The next step was to filter the data to only include those students who had reached the 3 Years at school milestone. To do this we filtered the data frame to only include those who were Year 4, 5, or 6. Next, we created a binary variable from the OTJ variable so we could use this for our logistic regression models. The goal of the project is to identify those students who will not be achieving at the National Standard so the values binary values for OTJ were split into: Below or Well Below = 0, and At or Above = 1.

Finally, we used the complete.cases function to ensure that we only kept those observations with data for each variable, and wrote the csv file of a complete and tidy data set.

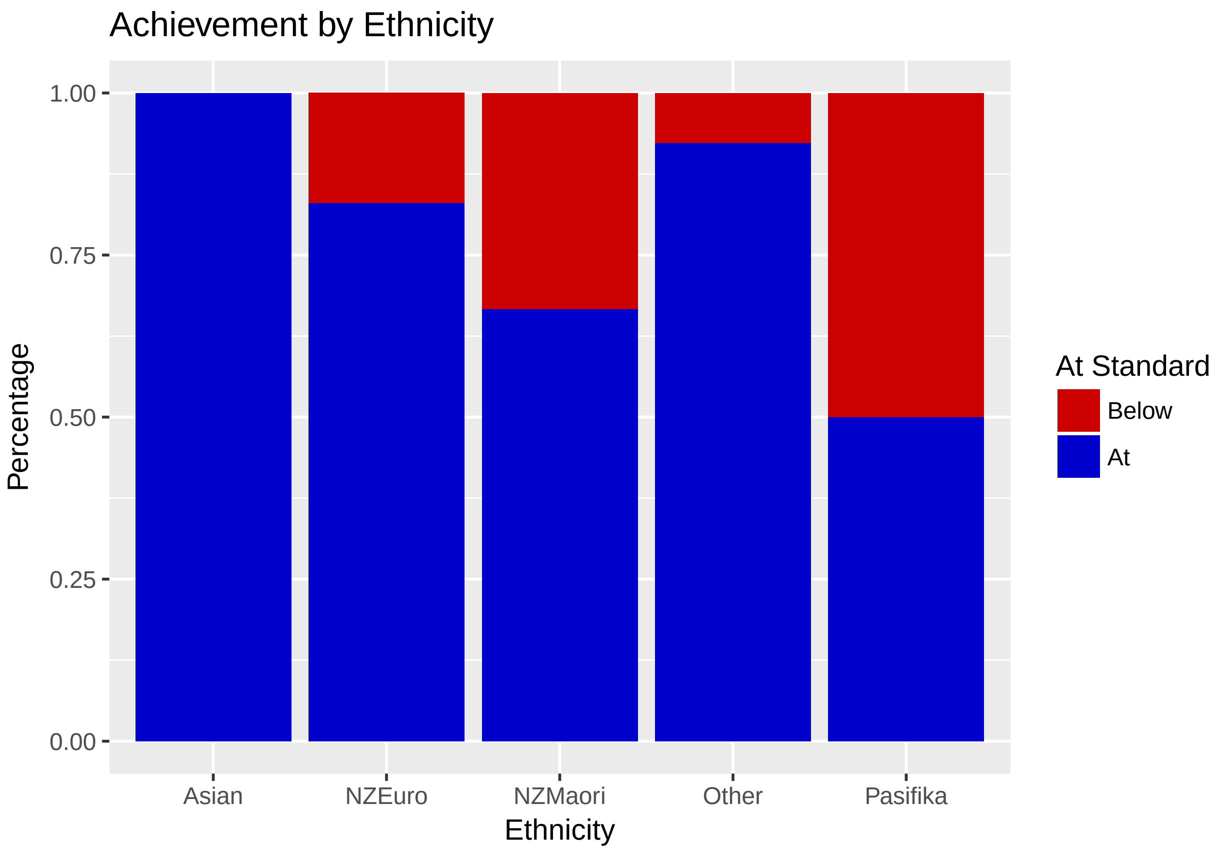
Statistical Analysis and Exploratory Data Analysis

Percentage of students achieving 'At'

The first step was to identify the percentage of students in the data set who were achieving 'At' or 'Above' the standard. 28 students (19%) were achieving 'Below' or 'Well Below', 119 students (81%) were achieving 'At' or 'Above'.

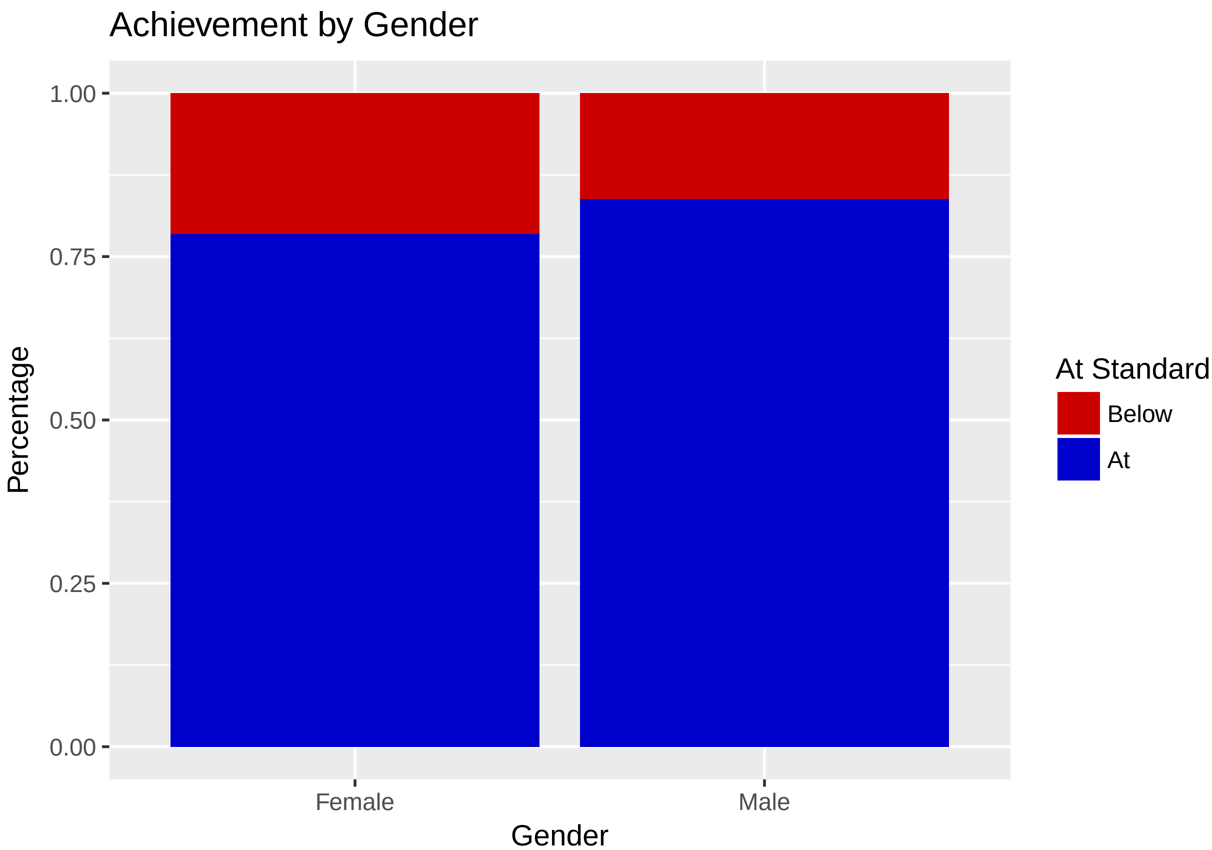
Investigating achievement within ethnic groups

I grouped the data by the 'ethnic' variable and then took the mean of the 'At' variable, as this is a binary variable this gave the percentage of each ethnic group that was achieving the standard. This showed that NZ Maori and Pasifika students were more likely to achieve below the standard. 33% of NZ Maori students were below the standard, as were 50% of Pasifika students. This was in contrast to NZ Euro students of which 17% were achieving below. This suggests that NZ Maori and Pasifika students are much more likely to achieve below the standard. I plotted this as follows:



Investigating achievement within gender groups

Data was grouped by the ‘gender’ variable and then the mean of the ‘At’ variable was calculated, as this is a binary variable this gave the percentage of each ethnic group that was achieving the standard. This showed that 21% of female students and 16% of male students were achieving below the standard. Female students are slightly to achieve below the standard. I plotted this as follows:



Investigating achievement of students within each JAM assessment domain.

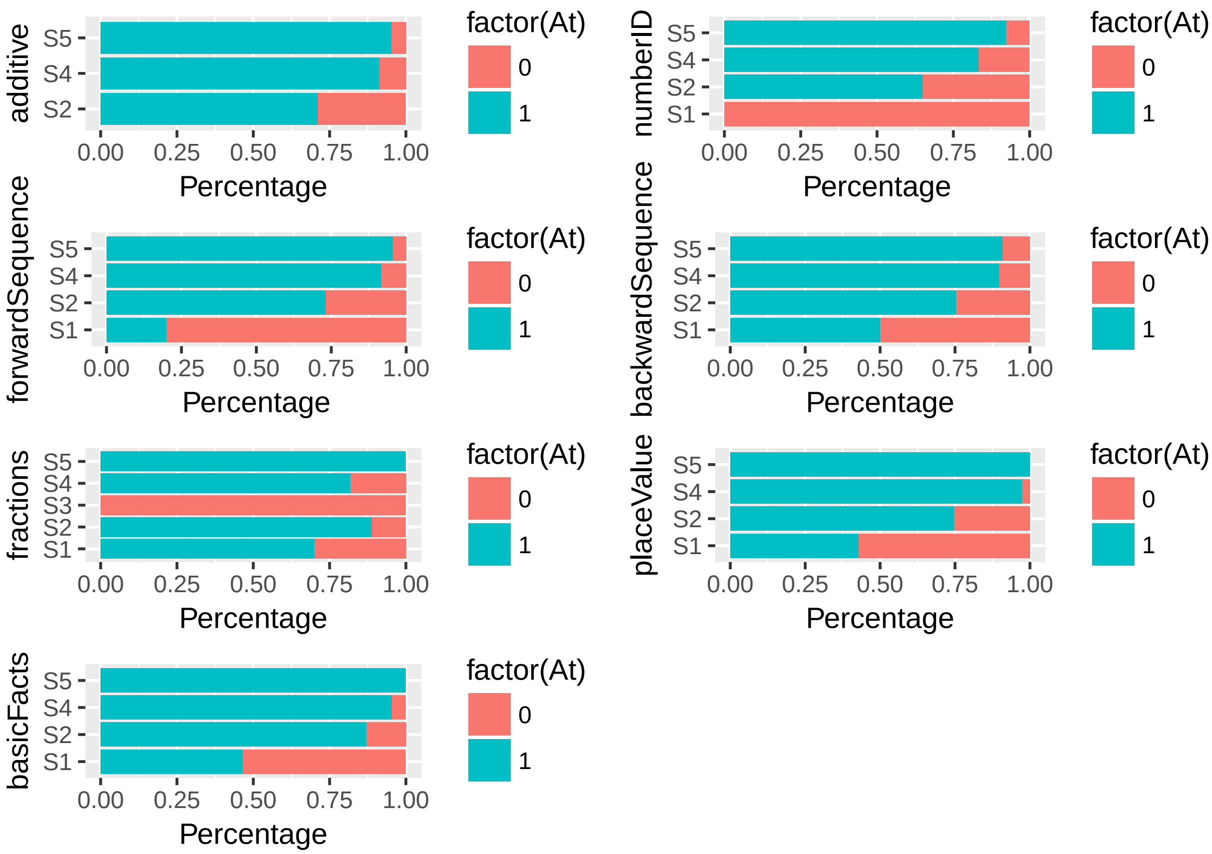
Data was grouped by each of the 7 JAM assessment domain variables and then the mean of the 'At' variable was calculated. This gave the percentage of students at each stage within each domain that was achieving the standard.

Findings

As expected, with nearly every domain from the JAM assessment, the higher stage that a student is assessed as being at the more likely it is that they will be achieving 'At' the standard. The one exception is the 'fractions' domain, there was not a clear trend in this domain.

Bar Plots

To visualize this I plotted each domain in a combined plot. Each individual bar plot showing the percentage of students at each stage of that domain who are judged 'At' or 'Below' the standard.

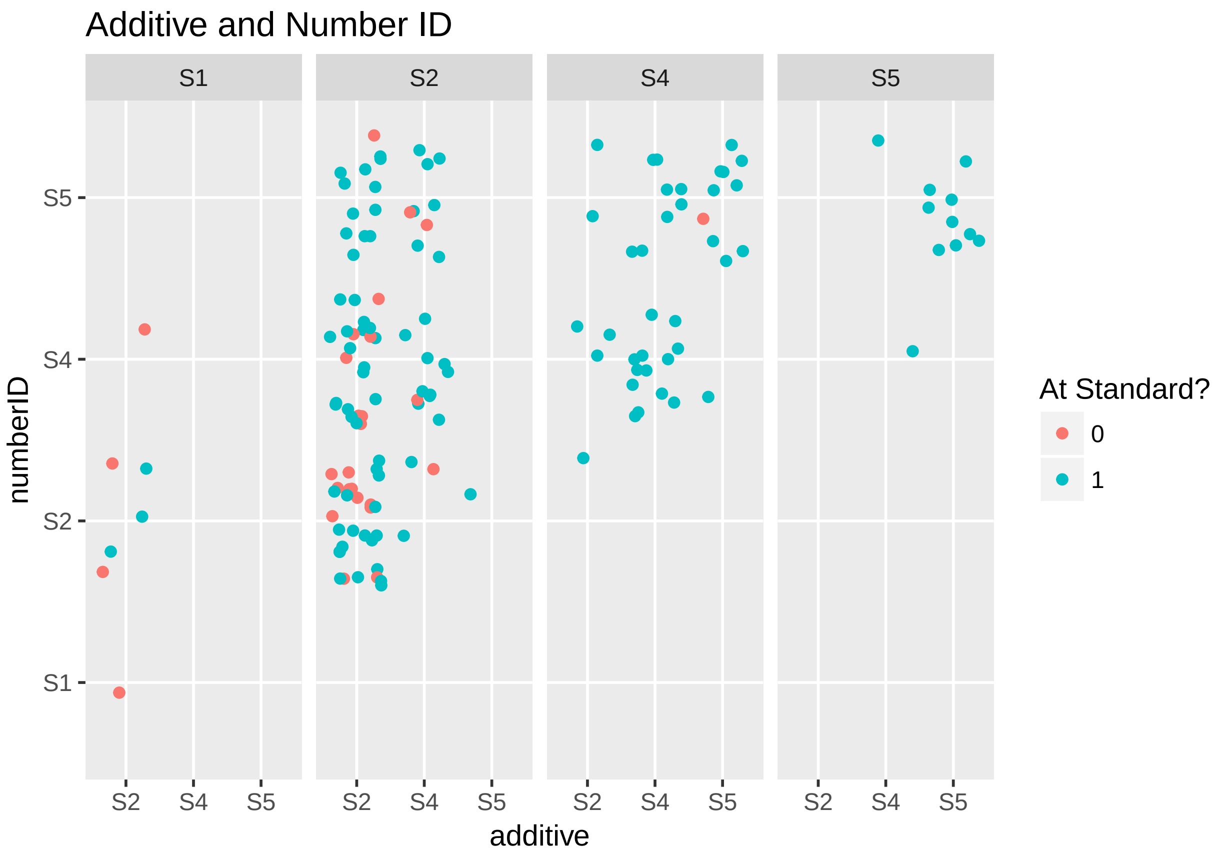


Interpreting the plot

The plots show visually that the higher the stage the more likely that a student will be ‘At’ the standard. The ‘numberID’, ‘placeValue’, ‘forwardSequence’, and ‘basicFacts’ domains have the largest proportion of students who were at stage 1 (S1) and then judged as ‘Below’ the standard. Over half of the students who were Stage 1 in these domains did not reach the standard.

Scatter Plots

I used facet grid scatter plots to visualize three variables and colored the points to show which observations achieved ‘At’ or ‘Below’ standard.


Additive and Number ID
S1
S2
S4
S5
numberID
additive
At Standard?
0
1


Forward Sequence and Number ID by Place Value
S1
S2
S4
S5
numberID
forwardSequence
At Standard?
0
1

Interpreting Scatter Plots

As expected the plots clearly show that the higher the stage in each domain the more likely that a student will achieve 'At' the standard. There is not a clear indication of which variables have the greatest impact on the outcome at this stage.

Machine Learning

The Problem

Can we use the JAM assessment results from the end of 1 year at school to train a model to predict the National Standard judgement for a student at the End of Year 3?

This is a supervised classification problem with a binary outcome: Will a student achieve 'At' (1) or 'Below' (0) the National Standard after 3 years at school?

The independent variables are:

- Gender
- Ethnicity
- Additive Strategies stage (from the JAM Assessment)
- Numeral Identification (from the JAM Assessment)

- Forward Number Sequence (from the JAM Assessment)
- Backwards Number Sequence (from the JAM Assessment)
- Fraction Knowledge (from the JAM Assessment)
- Place Value Knowledge (from the JAM Assessment)
- Basic Facts Recall (from the JAM Assessment)

The dependent variable is:

- At Standard (0 or 1)

Baseline Model

If we assumed that all students were achieving 'At' standard then we would have an accuracy of 81%. This is based on 119 of the 147 students being judged as 'At' standard. We will be looking to better this prediction accuracy in our machine learning models.

Binary Logistic Regression

The first approach was to create a binary logistic regression model that would predict if a student will achieve 'At' the standard using the independent variables outlined above.

Preparing the data

The first step was to import the tidy data set and then introduce factors for each of the independent variables in preparation for logistic regression.

For each of the models we used the select() function to create a model frame for the model selecting only the variables from the data set required for that model. For example:

```
ModelFrameA <-select(Targets, At, year, ethnic, gender, additive, numberID, forwardSequence,
                backwardSequence, fractions, placeValue, basicFacts)

ModelA <- glm(At ~ .,
              family = binomial,
              data = ModelFrameA)
```

The Models

We first created a glm model using all of the independent variables. This returned a warning message suggesting that perfect separation had occurred. We then systematically worked through a combinations of the independent variables creating another 18 glm models.

Results

The results for these glm models were not good. For each of the combinations of independent variables the model had one of three undesirable outcomes: * A warning message suggesting perfect separation
* Coefficients with very little significance * Null effect

It was decided that trying to create a CART model would be an appropriate next step for trying to predict student National Standard achievement using the JAM assessment data.

Decision Tree - CART

The approach here was to create a CART model using all of the independent variables and then use a Monte Carlo cross-validation to select the minbucket parameter that would give our model the most accurate prediction rate.

Preparing the data

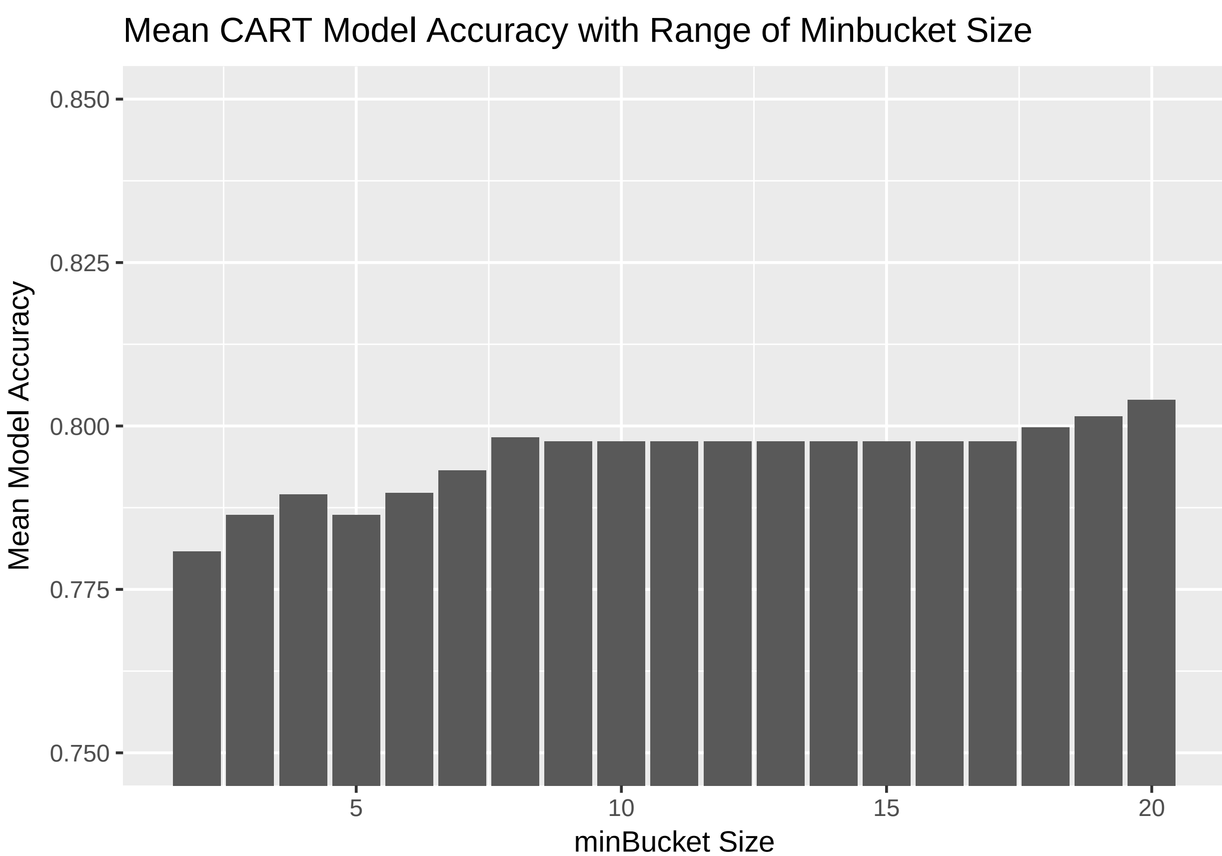
Again the first step was to import the tidy data set and then introduce factors for each of the independent variables in preparation for our CART model.

The Approach

1. Create a simple function to create list with 100 items, each item with 100 random numbers between 1:147 (the number of observations in our data set) to create training and test splits for Monte Carlo CV.
2. Create a function to take a list of 100 random numbers (between 1:147) and split data into train (100 observations) and test sets (47 observations). Run CART model 19 times with each split, each time using different minbucket size from 2:20. The function will return a list of 100 items, each is list of 19 model accuracy results from running CART model with a different minbucket size.
3. Take the mean model accuracy for each of the different minbucket sized CART models.

The Results

The results of this process did provide us with a solution, although it was not entirely satisfactory. After the Monte Carlo cross-validation the CART model with minbucket size 20 had the largest mean model accuracy.



Running the predict() function with this model on our complete data set returned a model accuracy of 82%. This was only a very small improvement on our baseline of 81%.

The decision tree produced by the model was very simple and only had one split. The model said that if a student achieved Stage 1 in the basic facts domain of the JAM assessment after 1 year at school, then they would not achieve 'At' the standard after 3 years at school. The reason for the simplified decision tree was that the cross-validation suggested that the best minbucket size was 20. Because our data set was small (147 observations) the minbucket size of 20 was proportionally large and had a significant effect. See figure below:

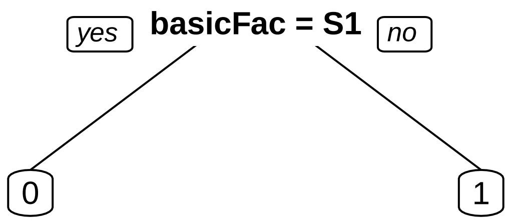


Figure 1: CART Model Decision Tree

Discussion

The primary reason for the rather unsatisfactory results of the logistic regression and CART models in this project was the small data set. The 147 observations simply was not enough to train an effective and accurate predictive model. During the logistic regression modelling it also become apparent that it was highly likely there was multicolinearity within the variables.

The variable that CART model identified as the sole predictor in the decision tree was the students basic facts knowledge. This was also identified in the logistic regression models as one of the variables with some significance, albeit not much. To put too much emphasis on this one aspect of the JAM assessment would be too simple a solution.

In education we know there are many factors that enable students to be successful learners in mathematics. Part of it what makes learners successful is the ability to recall some basic facts to enable them to efficiently and confidently solve problems. However, just being able to recall an equation from memory does not demonstrate that one has deep conceptual understanding of numbers or what they represent. Students also need to have a good understanding of the number sequence, place value and all of the other domains of mathematical knowledge that are assessed and recorded in the JAM assessment. Beyond the number knowledge domains, we also know that for students to be successful learners in mathematics they need to be able to make connections between different concepts and explain, reason, and justify their mathematical thinking.

Recommendations

As discussed, due to the small data set used in this project we were unable to create a predictive model that we could confidently recommend to use to identify students who might be at risk of not being 'At' the National Standard for mathematics after 3 years at school. However we can make some recommendations:

- According to our model the basic facts domain did appear to be the most significant indicator, therefore we recommend that teachers review how they are currently teaching basic facts knowledge and how they might change their classroom programme to support children to improve in this area.

- During the exploratory data analysis we identified that NZ Maori and Pasifika students are less likely to be ‘At’ the standard. We recommend that teachers reflect upon how they can develop their practice to ensure these students have the best possible opportunity to reach the standard.
- We also recommend that further work could be done in this area. If a much larger data set was available with assessment and achievement data from students across multiple New Zealand schools there could more exploration and predictive modelling completed on a much larger scale. This could provide educators with valuable insights into which students are most at risk of not achieving the National Standard, and allow them to put in place earlier interventions to support their learning.

Acknowledgement

I would like to offer a huge thank you Goran S. Milovanovic for sharing his expertise, and his guidance throughout this project.