Michael Petrey

End-of-Course Project

Using Data to Expand Enrollment and Recommend Students for 8th-Grade Algebra

**Background**

This project was based on the real-life example found [here](https://www.sas.com/en_us/customers/wake-forest-rolesville.html) on the SAS website.

Wake Forest and Rolesville, two middle schools in North Carolina, want to expand the number of 8th graders taking Algebra and to have a model recommend which students should take the class rather than rely on teacher recommendations.

**Preliminary Questions**

Why expand the number of students taking Algebra in 8th Grade?

“Taking algebra in eighth grade is considered a key stepping stone that allows students to progress to calculus by 12th grade. Students who take calculus in high school have a greater chance of successfully finishing a four-year college degree - particularly if they want to major in a STEM (science, technology, engineering or math) field.” [[1]](#footnote-1)

Why use a system that is not based on teacher recommendations?

From my own background as a teacher and from working at an educational non-profit, I know that all teachers have biases. Using data can help to bring these unconscious biases to light and even to address and counteract them. Having a system based purely on teacher recommendations could result in unfairly tracking students, but having an algorithm based on those same assumptions could be just as detrimental so we will have to be careful with which data to include in the model. Teachers see their students day after day and are the ones with the most intimate knowledge of their students, but that does not mean that they are well-equipped to forecast a student’s success into the future.

**Analysis Summary**

Three models combine to address the many facets of this project:

1. A Logistic Regression tree predicts the probability that a rising 8th-grade student will succeed in Algebra.
2. A discrete Markov Chain predicts the number of students in each class each year.
3. A Linear Regression model highlights the important factors in determining which students were recommended by teachers to take Algebra in 8th grade.

**Model 1: Predict students who will succeed in 8th-Grade Algebra**

The data used in this model will be the key to its success. If factors that are heavily influenced by teachers have too much weight in the model, then the model will have the same prejudices as the teachers. Of course, nearly all measurable characteristics of students are influenced by their teachers. North Carolina, like many states, has standardized End of Course (EOC) exams beginning in 3rd grade for math[[2]](#footnote-2), and that is my recommended data to use for as many students as possible. For students who have missing data, perhaps because they recently moved to the state, we will have to come up with another approach.

The tree above shows which model to use based on what data is available for students. Ideally, students would have been in the school system since at least the 3rd grade, and therefore most students would have enough EOC data to base the model only on that data. You will notice that all of the options here end in logistic regression models, which will give the probability of a student passing the EOC test for Algebra in 8th grade, and that the data going into each of the models is independent from the teachers. For students with some or all of the EOC data missing, care must be taken to gauge their readiness for Algebra with as much accuracy as possible. This will likely mean looking at these students on a case by case basis to determine whether the placement exam should be taken and how to weight it with the EOC scores.

In the end, the logistic regression models will give a probability of succeeding in 8th-Grade Algebra for each student. In this case, success is measured by passing the standardized EOC exam. The threshold is important to consider here because it will be used to determine which students proceed into Algebra in 8th grade. Because having students in a class they are not yet prepared for is a more costly error than not advancing students into 8th-Grade Algebra when they are ready for it, I would recommend a threshold of around 0.75[[3]](#footnote-3). This threshold could be different for the different models, for example the regression based on the placement exam may have a different threshold, but I would recommend scaling the results to have a consistent threshold across the different models to better make comparisons across models and students.

**Model 2: Anticipate teaching needs in high school**

With more students taking Algebra in 8th grade, more students will be taking Geometry, Algebra II, and other higher math courses in high school. A discrete Markov Chain could be used to predict the teaching needs at each grade level for math.

Some assumptions for the Markov Chain model:

* The historical percentage of students progressing from one class to the next will remain accurate even when there is an increase in students in 8th-Grade Algebra.
* Because this is a stochastic model, we are assuming that the students progressing from one class to the next are randomly selected. This may not be the case upon closer inspection, but it will suffice for predicting the overall teaching needs for each class, just not exactly which students will be in which class.
* The next class that a student will take is based only on the current class and not past classes (because the model is memoryless). Although this could be seen as a limitation of the model, it could also be helpful
* Students who fail the EOC exam for a particular class will retake that class the next year.

Unfortunately, I was not able to get the data needed from the [publicly available sets of data](http://www.ncpublicschools.org/accountability/reporting/) to determine the historical transition rates from one class to the other. The data is not detailed enough to determine the transition rates nor the exact path that students take. North Carolina refers to high school math courses as Math I, Math II, Math III, etc. but for clarity I have used more specific names here. I have based the transition rates on the stated historical rate of 97% of students passing 8th-Grade Algebra, with slightly lower passing rates as the courses get harder. Also, because not all students take or pass Algebra in 8th grade, there will be students graduating prior to taking Calculus or even possibly Pre-Calculus depending on the schools’ graduation requirements.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table of Transition Probabilities** | | To | | | | | |
| Algebra | Geometry | Algebra II | Pre-Calculus | Calculus | Graduation! |
| From | Algebra | 0.03 | 0.97 | 0 | 0 | 0 | 0 |
| Geometry | 0 | 0.03 | 0.97 | 0 | 0 | 0 |
| Algebra II | 0 | 0 | 0.04 | 0.9 | 0 | 0.06 |
| Pre-Calculus | 0 | 0 | 0 | 0.04 | 0.8 | 0.16 |
| Calculus | 0 | 0 | 0 | 0 | 0.05 | 0.95 |
| Graduation | 0 | 0 | 0 | 0 | 0 | 1 |

This is a discrete Markov Chain model, so each iteration represents a separate school year with the results being the approximate number of students in each class for each school year, beginning with the first year there is a surplus of students in 8th-Grade Algebra. An important note is that this model is also dependent on the grade of a student. For example, the total number of students taking Geometry next school year is equal to 97% of 8th-Grade Algebra students plus 97% of 9th-Grade Algebra students. This model begins with Algebra and not Pre-Algebra for two reasons: 1) the historical number of students transitioning from Pre-Algebra to Algebra will not be accurate because of the first model and 2) Algebra students will be split between middle school and high school so those teaching needs would have to be separated as well.

If there is more variability in the actual transition numbers, this same methodology could easily be transferred to a simulation which could account for that variability. For example, if between 90-99% of students in Algebra went on to take Geometry the next year, that distribution could be used in a simulation that would be run many times to give a range of possible scenarios.

**Model 3: Analyze which students were being missed in teacher recommendations**

Expanding the 8th-Grade Algebra enrollment with more students who are predicted to do well in the course is a great goal, but analyzing why those students were being missed in the first place has the potential to make monumental shifts in the teachers’ and school leaders’ beliefs. This form of analysis shows the power of data to highlight unconscious biases and tell a different story than the one within a teacher’s head. This also has the potential to make teachers feel very vulnerable or undermined, so it should be approached carefully and as a learning tool. I would strongly suggest reporting at a high level so teachers or schools do not feel singled out in the analysis.

The model itself is a linear regression model that will use a multitude of factors to determine which factors are most important in determining which students teachers recommend for 8th-Grade Algebra.

Data:

* For each student we want: overall class grade, test grades, final exam grade, class participation grade, homework grade, race, sex, whether the student receives free or reduced lunch, End of Course exam grade, student interest, parent interest, current math class.
* All of these factors are for a single year because I believe teachers do not take into account past factors, although past factors could be included because they certainly can affect teachers’ assumptions about students.
* Some of these factors may be difficult to measure, such as whether the student or parents expressed a particular interest in which math class the student would take, but that data should be included if available. Some teachers may also have more ambiguous measures for some of the factors, such as an overall feeling of the effort a student puts into the work, which may be correlated with the appearance of consistently completing homework.

Model:

A linear regression model that uses the factors listed above to predict which students are recommended by teachers to take Algebra in 8th grade. Note that the model and data should include all 7th-grade students, not just those in Pre-Algebra.

To:

Determine which factors most heavily influence whether a student is recommended to take Algebra in 8th grade.

This analysis will be able to answer questions such as:

* Are more boys than girls recommended for the higher math class?
* Are students who have free or reduced lunch less likely to be recommended?
* Do teachers consider students outside of Pre-Algebra?
* Does the model rank of importance agree with the teachers’ own ranking? Are there factors that the model considers important that teachers do not feel are important?

An important note here is that teachers recommended students who they felt would succeed in the class, but their definition of success may have been different than the definition used in the recommendation model, whether a student passes the End of Course exam or not. Perhaps teachers highly weight factors like homework completion because they view that factor as having a high importance in succeeding in the next class as well. Having discussions across the department and across schools about what success is and how to measure success is an important piece of successfully implementing this entire program.

1. https://www.sas.com/en\_us/customers/wake-forest-rolesville.html [↑](#footnote-ref-1)
2. http://www.ncpublicschools.org/accountability/reporting/ [↑](#footnote-ref-2)
3. SAS recommends a threshold of 0.7 for this use of their software models. For the first year, the school district wanted to be more conservative and set a higher threshold at 0.8 to ensure most students in the class continue to pass the EOC exam. [↑](#footnote-ref-3)