Modeling Student Outcomes

Chris Malec

Springboard DSC Track
Capstone Project #1
July, 2019

Introduction

Graduating High School is an important part of a young person's education. There are a large number of pathways to a successful life after this point, but many require at least a high school degree or GED. Young adults face more challenges when trying to enter the workforce without a high school diploma. Therefore, school districts, city, state, and even federal governments have an interest in keeping students on track to graduate.

Sometimes, a student in danger of dropping out may be clear to their teachers, but other times a student may just stop coming to school with little warning. Or perhaps a student does not fit what a teacher, counselor, or administrator considers 'a dropout.' In this project, we aim to try to use data to assist schools and districts in identifying students with a high potential for dropping out.

By using data to create a model to predict whether or not a student will dropout, we hope to direct interventions toward particular students to increase graduation rates and hopefully smooth over a potential bump in the student's path to success.

Approach

The approach taken in this project is treating this as a classification problem, meaning that the information collected about each student becomes a set of 'features' that allow a model to predict if a student should fall into the 'dropout' category or the 'non-dropout' category. Obviously, if successful, this approach could expand to more categories to try to find why students are not reaching their academic potential, however, decreasing real dropout rates is a higher priority than most other problems that may be treated with this data set.

It is possible that the model may identify effective interventions by finding some program reported by the counselor or administrator that is a good predictor of dropping out or not dropping out. This is not a stated aim of the current project, so even if it may be a helpful side-effect, our goal here is to identify students who are likely to dropout. Effective interventions would be an important next step.

Data Acquisition and Data Wrangling

A particularly difficult task is deciding how to determine whether a student should be classified as a dropout or not. Since this is central to the question at hand, we looked at

a few options. At first, we decided to look at the high school transcript to see if students had dropped out by that time. This column had about 20,000 out of 4 million students dropout, a rate of 0.5%. However, high schools are known to overestimate their graduation rate because it is tied to funding (LA Times story: Numbers Game).

We therefore chose to use the data column that recorded a dropout event as reported by parents, students, and the school. This label revealed a dropout rate closer to 12%, which is more in line with other NCES data, given that about 85% of students graduate from high school in 4 years. It also makes sense that students may only be labeled as a dropout on their transcript if they did so in their senior year, before being removed from the school roll.

Exploratory Analysis and Inferential Statistics

Several questions came to mind when we began to explore the data, particularly from my teaching experience. We were curious as to how many students had dropped out in this study, as well as what problems teachers and principals felt plagued the school.

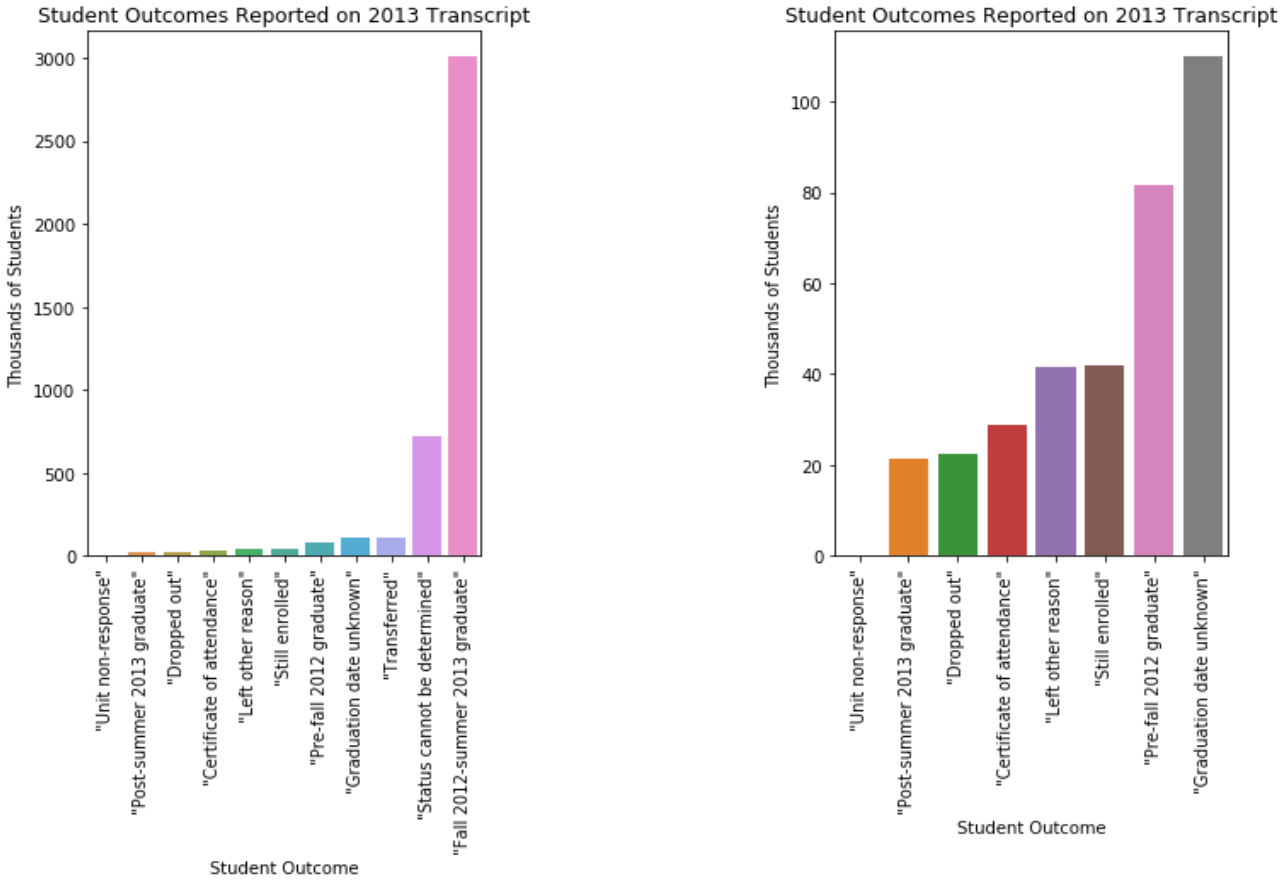

Figure 1: Information on student transcripts four years after starting the ninth grade.

The first question was just what did the proportion of student outcomes look like? We made several graphs from the outcomes entered on students' transcripts, as well as a question of whether or not they had ever dropped out of high school by their senior year.
sometimes going on to graduate.

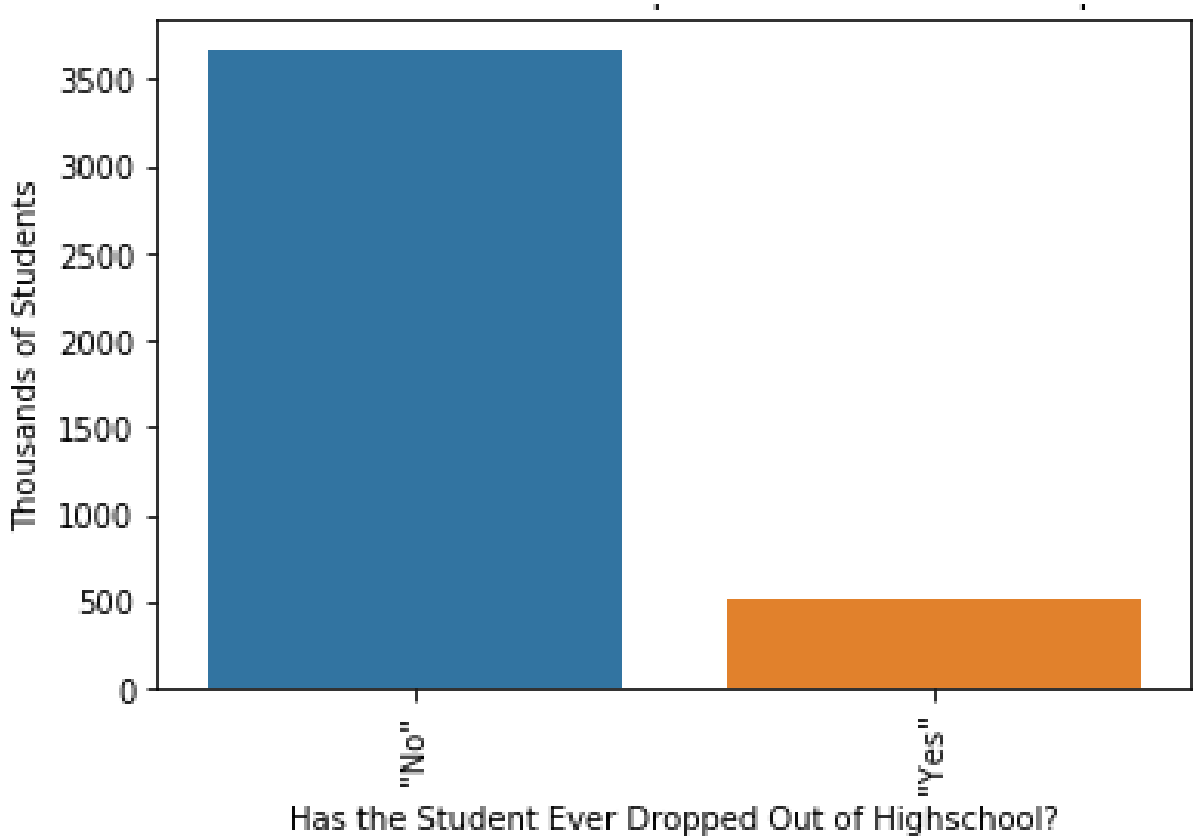

Figure 2: Whether or not students have ever dropped out of high school, this question is used to label the data into 'dropout' and 'non-dropout' classes.

These bar graphs show that the percentage of students with the second label for dropout is much higher than the percentage of students with the first label for dropout. Since the rate of predicted dropout is so low in the first case, we choose the second one as our label, as it agrees better with other available data.

The number of students who ever dropped out over the course of the study grows from 400 thousand, three years after the study begins, to over 500 thousand when the cohort graduates. Seven years after the study begins, nearly 800 thousand of the roughly 4.3 million ninth graders had dropped out of high school at some point. This is summarized in the graph below and demonstrates that the number of students who dropped out of high school is fairly large and stays large after their freshman year.

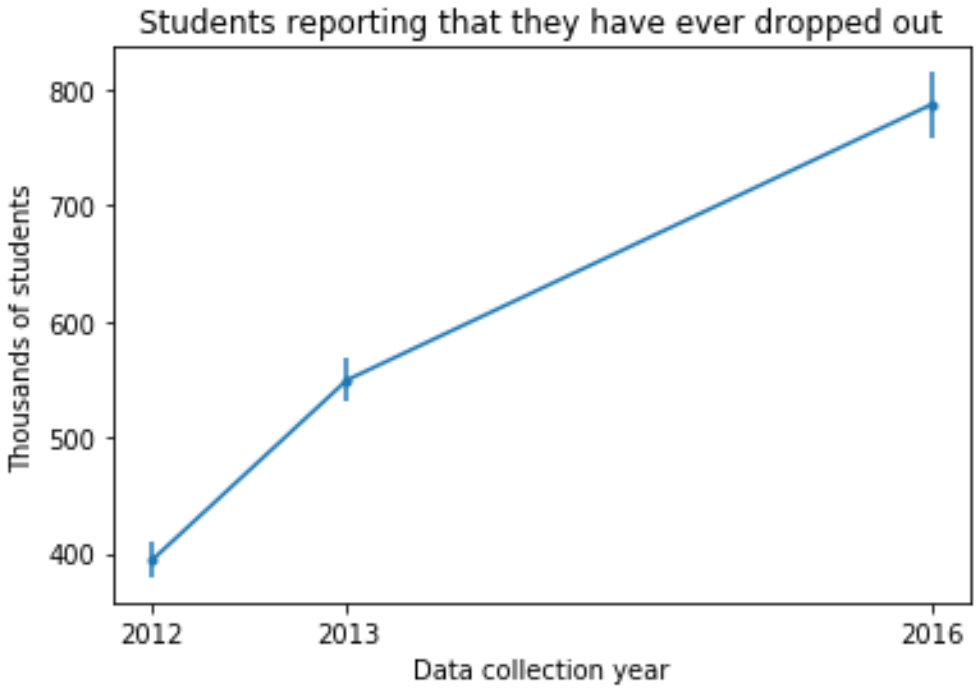

Figure 3: The number of students reporting that they have ever dropped out of high school vs time.

Statistical tests were conducted on the data to look for statistically significant differences between the proportion of students that dropped out and those that did not drop out. For discrete variables, the mean proportion of all students corresponding to each category and dropout status was calculated along with a standard error, and then the difference between the two groups was found along with a p-value.

Category ↓ / Class →	Dropout	Non-dropout	Difference (Dropout - Non-dropout)
Male	57.1% ± 0.2%	49.5% ± 0.06%	7.6% (p-value < 0.001)
Female	42.9% ± 0.2%	50.5% ± 0.06%	-7.6% (p-value < 0.001)
Feel safe at school	24.8% ± 0.2%	31.2% ± 0.06%	-6.4% (p-value < 0.001)
Don’t feel safe at school	12.0% ± 0.1%	7.22% ± 0.01%	4.8% (p-value < 0.001)

Table 1: Selected statistical tests on the dataset

For example, as summarized in Table 1, males are 57% of the dropout population compared to 43% female. Also, how students answered a question about school safety also had a statistically significant difference between the dropout and non-dropout group.

We considered other questions, which can be seen in my storytelling notebook.

Baseline Model

The survey data includes responses from students, teachers, counselors, administrators, and parents. Each row of data represents one student. There are 23,503 students appearing in the data set, that were selected to give a representative sampling of the U.S. population. Each student has 2,719 columns associated with survey data, not counting columns with weights and imputation flags. After removing columns with a large number of missing values, and creating dummy variables for the discrete variables (usually multiple choice questions), the number of features grew to 4,345.

Once the features have been encoded in a numpy array and the labels have been transformed to a numpy vector, we import some necessary sci-kit learn modules in order to build Logistic Regression models.

Normally models are selected by searching a parameter space and selecting the parameters that produce the highest training accuracy. In our current problem, this would result in a poor model. Since the majority of students do not drop out, guessing that all students do not drop out would lead to a high training accuracy (~90%), but completely fail to meet our goal.

Therefore, we use different metrics, specifically precision and recall. Both precision and recall are with respect to a class being predicted, in this case, being a member of the ‘dropout’ class. Precision addresses how many of the predicted dropouts are in fact dropouts. Recall addresses how many of the known dropouts were predicted. Obviously having both 100% precision and recall would be ideal, but in reality there is often a trade-off between the two.

To have the greatest impact on graduation rates, optimizing for recall would be the best course of action, as this would capture the greatest number of students at risk for dropping out. However, if the precision is too low, the model’s impact could be blunted.

One way to get 100% recall is by guessing all students will dropout, and therefore any intervention would be applied across the board to all students. This would both result in wasted resources, and may lead to the perception that the applied interventions are useless since they aren't applicable to the majority of students.

For the purpose of a baseline model, a balance between precision and recall is chosen by choosing parameters that optimize the f1 score. The f1 score is the harmonic mean between precision and recall, therefore creating models that balance the two. The parameter space of a logistic regression model is not extensive. The major parameters to adjust are the regularization C and the type of penalty applied (L1, L2, or Elastic net). We used a GridSearchCV object to build models with selected values of the parameters. In general, the L1 penalty fared much better than the L2. Therefore, we did not pursue the Elastic net penalty, since this allows a mixture of L1 and L2. We figured that we would just tune the Elastic net to be L1. Additionally, the Elastic net is only available for certain logistic regression solvers.

The baseline model achieved a precision of 0.32 and a recall of 0.40. This means that 32 percent of possible dropouts are detected, and 40 percent of those selected are potential dropouts. A major advantage of logistic regression is that it can be used to output probabilities, which allows the model to show the probability of a student dropping out. In addition, the model can express the top predictors for dropping out for students in general. An example of this is shown in Figure 4.

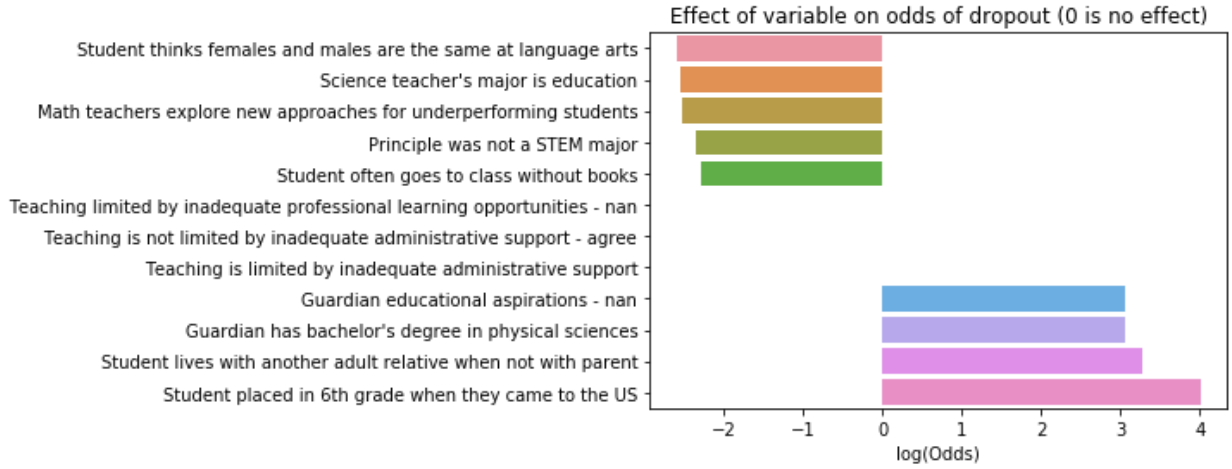

Figure 4: Graph showing some of the top predictors for classifying a student into the dropout class (positive values) or non-dropout class (negative values)

We can see that many of the top predictors are not necessarily obvious, but many of them can fit into our usual perception of who will drop out of school. ‘Student thinks females and males are the same at language arts’ and ‘Math teachers explore new approaches for underperforming students’ are not the first factors most of us might think of for predicting a student to stay in school, but they imply a mindset that effort is important and students and teachers have the ability to improve. Likewise ‘student lives with another relative when not with parent’ and ‘student placed in 6th grade when they came to the US’ suggest students that have more difficulties or less stability than many of their peers.

However, not all of the categories fall into an obvious narrative. For example, ‘student often goes to class without books’ does not sound much like a student who will go on to graduate and ‘guardian has a bachelor degree in the physical sciences’ doesn’t sound like a predictor of a student dropping out. Not shown in Figure 4 is the predictor ‘student has never been absent from school for a month or more’ because it dwarfs the other predictors of dropping out. We will discuss reasons why features do not always conform with our expectations later in this analysis.

Extended Model

A major difficulty of the dataset is that the classes are imbalanced. Imbalanced classes occur when there are inherently many more training examples in one class than the other, and it is very common with data in the real world. Loan defaults and credit card fraud are too areas where the classes are so imbalanced that an entirely different approach known as ‘anomaly detection’ might become necessary.

Since there are more non-dropouts than dropouts, the model will tend to fit better to characteristics that predict non-dropouts than dropouts, which lowers the rate at which dropout cases are identified while maintaining a high model accuracy. There are a variety of ways to deal with this, and in extending the model, we investigate several of these methods.

Precision and recall are good metrics for this type of data. A high precision model means that most of the students that we label as dropouts are in fact dropouts. This goal makes interventions more targeted by concentrating on the students who will most likely dropout. A high recall model means that our model labels most of the possible dropout students. This goal ensures that we reach as many students as possible who are likely to dropout.

It is possible to have a high precision, low recall model and vice versa. Therefore, we will look at both in evaluating our models.

Changing the cutoff and Precision-Recall curves

An easy way to change the precision and recall of a model is to change the probability cutoff used to predict which class a student falls in. By using a variety of cutoffs, we can construct what is known as a precision-recall curve which displays all the possible precision and recall pairs within a model.

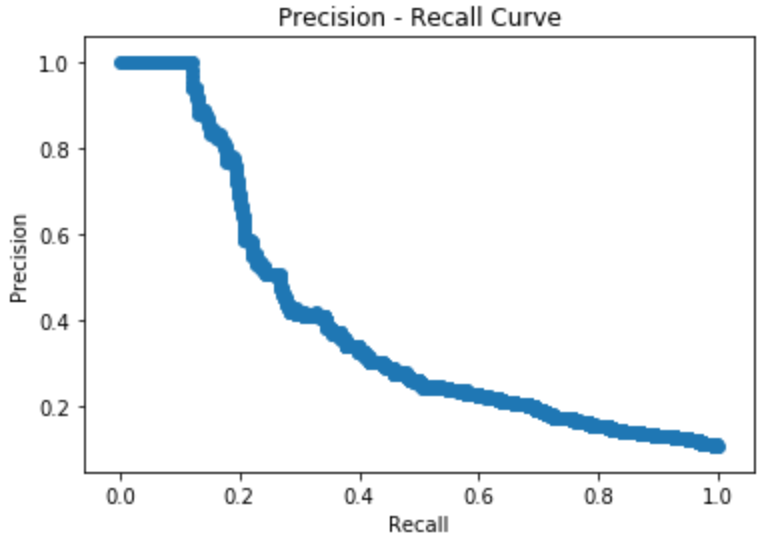

Figure 5: Precision - Recall curve of the baseline model.

The precision-recall curve for the baseline model, as seen in figure 5, reveals that we have a pretty rapid loss in precision to make gains in recall. This is not ideal, so we will explore some methods to improve this curve.

Class weights

A very straightforward method is to take advantage of the fact that sklearn's LogisticRegression object has a 'class_weight' parameter that can be set to 'balanced'. This adds a weight to each training example that is the inverse of its class frequency. This means that each training example in the non-dropout class will be weighted less and each dropout training example will be weighted more. This does indeed increase the recall of our model.

Under- and Over-sampling

These methods try to explicitly even out the balance in the training examples. With under-sampling, we randomly remove examples from the training set in the majority class to create two classes with a roughly equal number of examples. An advantage of this method is it decreases the size of your data set and the model trains faster, a disadvantage is that you end up with a lot less data which may lead to models with high bias.

Over-sampling relies on taking bootstrap samples (sampling with replacement) of the minority class to increase the number of training examples that correspond to dropouts. This method replicates the performance of using the class_weight parameter.

Changing the Score Function

The sklearn library has many different scoring functions besides accuracy. We can evaluate our models by either precision or recall, and select the one that optimizes these scores. Since we want at least some balance between precision and recall, it makes sense to use the F1 score, or the harmonic mean of precision and recall.

We can see from the precision-recall curve of the F1 score optimized model, as shown in figure 6, that there isn't as sharp a loss in precision for gains in recall.

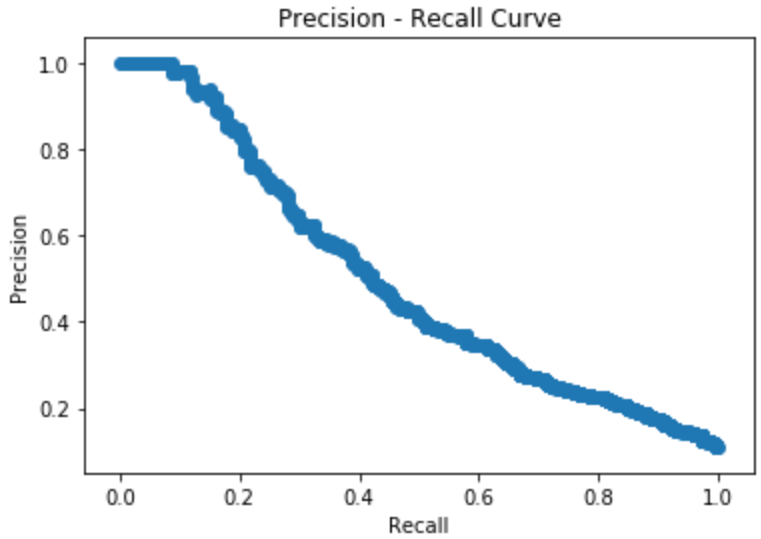

Figure 6: Precision-recall curve of F1 optimized model.

Finally, to obtain yet more flexibility in how we select models, we can use the f-beta score (defined as $F_\beta = (1+\beta^2)\frac{precision \cdot recall}{(\beta^2 precision) + recall}$), which allows us to set a parameter 'beta' to control the importance of precision and recall.

We generated a number of models and selected the ones with the best f-beta score for different values of beta. We can see in figure 7 below that the ones with a low beta have high precision and low recall, while the ones with higher beta have low precision and high recall.

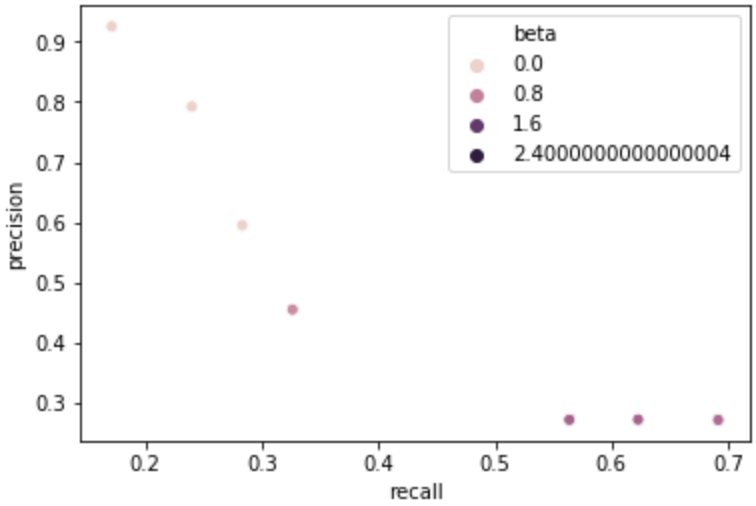

Figure 7: Precision and recall values for models selected with different values of fbeta

Findings

We summarize our findings for different models below. We can see that various techniques lead to different rates of precision and recall. By balancing the training set with respect to examples in the dropout and non-dropout class we can increase either the precision or recall from the baseline, but it seems difficult to increase both to any appreciable degree.

Model	Precision	Recall
Baseline	**0.32**	**0.40**
Balanced Class Weights	**0.28**	**0.47**
Random Undersampling	**0.20**	**0.62**
Random Oversampling	**0.30**	**0.45**
F1 Score Optimization	**0.56**	**0.29**
Balanced Class weights and F1 Score Optimization	**0.27**	**0.69**

Table 2: Precision and Recall results for different model selection methods

This model identifies students who are at risk of dropping out, however, the strongest predictors of dropping out or not dropping out do not necessarily lie in line with our expectations of what good predictors should be. For example, not strong predictor of not dropping out is regularly forgetting to bring your books to class, and a strong predictor of dropping out is not being absent from school for more than a month before starting High School. This could mean one of two things, either our intuition for who is likely to dropout is very wrong, or there are strong interactions between features in the model.

In reality, probably both are true to some extent. There are many factors that go into people's decisions, both within and outside their control. It is unlikely that hundreds of thousands of students every year fall into the categories described by a handful of easily understood variables. Likewise, the dataset is from survey data, there was little to no effort to create a set of independent features to train a machine learning model, it was designed to try to answer questions the researchers felt may be interesting. A strong predictor for dropout may have a high correlation with a strong predictor for non-dropout, thus we would see two features as strong predictors when they in fact cancel themselves out.

Therefore, we conclude that the model has shown some success in predicting possible dropout cases, however further work is required before we try to use the model to predict why a student may drop out, or important risk factors.

Future Work

A major avenue for future work would be to gather non-public use data. The NCES makes the current dataset available to anyone, with a number of columns labeled with a '-5' to mark them suppressed. There are 887 suppressed variables in the public use data file. Non-public use data requires an approval process, but with that process, more specific data is possible since school and student level data can be combined. In addition, specific subsets of students can be studied since the strata involved in the sampling of the original data would be known. Therefore, much more specific predictions could be made from a subset of the data with a student population that more closely resembles a specific school.

The reason that much of this information is suppressed in the public-use files, or at least one reason, is that such information can be aggregated to identify individual students, administrators, parents, teachers, schools, etc. In addition, having detailed racial, ethnic, and geographic information about the students contained in the data opens the possibility of bias in creating the model. By bias in this case, we mean predicting a greater or lesser probability of dropout for a particular group that does not match the actual rate of dropout for that group.

On the one hand, such information can have more downside than upside since not everyone with access to the data might have the same goals. On the other hand, if there is some type of implicit bias caused by other features in the data, there is no way to know. So with more detailed information it is possible to confirm that particular groups are not being singled out by the model.

Another route to take is to completely change the approach of the model. Since the number of dropout cases is relatively rare in some schools, an anomaly detection algorithm may farer better than the logistic regression shown here. We believe that variations on the present model would best serve a school that had higher dropout rates. When dropouts decreased, perhaps with the help of the model, the regression model would have an increasingly difficult time detecting the smaller and smaller minority of potential dropouts. Therefore, developing an anomaly detection paradigm would provide the best value for those schools where dropouts are relatively rare.

Client Recommendations

An important decision with this data product is the precision/recall tradeoff. Even as the model improves, there will likely still be at least some tradeoff to be considered. A couple examples will hopefully illustrate how to the decision should be weighed. We look at what a 25/80 precision recall split looks like, as well as a 40/50.

A school of 725 students will have approximately 100 potential dropouts if the school conforms to the national average. 80% recall means that 80 out of the 100 dropouts are identified, and 25% precision means that 320 students are identified as potential dropouts. This is nearly 45% rather than 15% of the student population. This would be a good route for a school if the intervention were aimed at entire grades. The model in this case would help schools and districts focus on which risk factors their interventions should focus on.

On the other hand 50% recall means that 50 out of 100 dropouts are identified, and 40% precision means that 125 students are identified as potential dropouts. These are more manageable numbers for a targeted intervention meant to treat students in small groups. An even more precise model would focus on a handful of students and may attempt to identify the students most at risk of dropping out.

A possible solution, and one we strongly recommend is to include humans as part of the workflow. By combining human scanning, and knowledge of students as individuals, the best of both worlds can be achieved. Therefore, we could attempt a high recall model to narrow down the list of possible dropouts, and use teachers, counselors, and administrators familiar with the student to increase the precision.