Identifying Students at High Risk of Not Graduating High School on Time Using Machine Learning

Akshay Oza
Carnegie Mellon University
aoza@andrew.cmu.edu

Ankit Ramchandani
Carnegie Mellon University
aramchan@andrew.cmu.edu

John Fang
Carnegie Mellon University
johnfang@andrew.cmu.edu

Min Lee
Carnegie Mellon University
minhunl@andrew.cmu.edu

Tianying Chu
Carnegie Mellon University
tianying@andrew.cmu.edu

1. Executive Summary

Graduating high school is an important milestone for students, and research has shown that failing to graduate on time is correlated with various negative social, economic, and personal consequences [8] [20] [26]. Despite the negative consequences, one in five students in the United States fails to graduate on time or even graduate at all. Identifying such students at high risk of not graduating on time is challenging because many different factors could contribute to it, and schools usually have limited resources to continually assess at-risk students.

The goal of this project is to create a machine learning solution for efficiently and equitably predicting which high school students are at highest risk of not graduating on time as assessed at the end of grade 10. Our policy goals for this project include efficiency, equity, transparency, and customizability. We want our solution to be efficient and equitable so that schools can spend their resources on students who actually need help without discriminating among students of different groups. We also want our solution to be transparent so schools can get a high level idea of how our model is making decisions and decide to place trust accordingly. Finally, we want our solution to be highly customizable so we can specifically tailor our models to schools' needs, instead of giving all schools the same model.

To achieve these goals, we use data provided by the Muskingum Valley Education Service Center (MVESC) of 14 school districts from 2006-2015 to predict risk scores for students who have just completed their grade 10 based on 350+ features like past grades, GPA, attendance, reasons for absences, extracurricular involvement, discipline incidents, demographical information, etc. We measure efficiency (precision) when predicting the top 5% highest risk students and compare the probabilities of falsely denying intervention (false negative rate) between different groups to assess fairness over ethnicities and school districts. Furthermore, our evaluation process to validate our results exactly mirrors the deployment setting as we ensure that there is sufficient temporal gap (2 years) between the training and validation set.

We train several machine learning models like random forests, decision trees, logistic regression, neural networks, etc over a large hyperparameter space to ensure we get high efficiency and fairness. We also compare our models' results with 3 baselines in which we simply rank students by their GPA, discipline incidents, and absence rate and predict the top 5% as highest risk. These baselines are meant to capture some of the non machine learning based methods that schools already use or can easily use. To ensure transparency, we also provide methods to view both global and local feature importances of all models so policymakers can understand both the overall behavior of the model and also the reasons for predicting a certain risk score for a certain student.

Moreover, consistent with our policy goal of ensuring customizability, we give maximum flexibility to policymakers in letting them decide their desired trade-off between efficiency and equity and recommending the model to use accordingly. In this report, we highlight three example choices policymakers may have to balance the aforementioned trade-off (i.e. most stable efficiency with reasonable fairness, most stable with fairness over ethnicities, most stable fairness over school districts), and present results accordingly.

Finally, we show that our models almost always perform better than baselines. This strongly suggests that our machine learning based solution has a potential to improve current methods used in identifying at-risk students. We also provide future directions like recommending a field trial design to verify the usefulness of our models in the real world.

2. Background and Introduction

Despite governmental and educational initiatives, graduating high school is still a major issue that still needs further work. 17% of students do not graduate high school on time as per 2018 statistics. Moreover, according to the recent NCES 'Condition of Education report' [15], the range of high school graduation rates for states is between 74% to 94%. This reflects that not all states have equal resources to increase their graduation rate at the same pace as the high performing states. These statistics are even worse for students of color, students with disabilities, and LGBT students. Not graduating high school on time is *correlated* with major economic, social and emotional consequences for individuals.

As for economic costs, according to the U.S. Bureau of Labor Statistics, high school dropouts are nearly three times more likely to be unemployed than college graduates. The Alliance for Excellent Education report published in 2011 showed that high school dropouts earn a national average of $8000 less annually than high school graduates. This gap is expected to increase further if high school graduation rates are not improved. Georgetown University's Center on Education and the Workforce [8] found in their recent study that more than 65% of job openings through 2020 would require at least a bachelor's or an associate degree, further decreasing the chance of high school dropouts to earn a reasonable livelihood.

In terms of social and emotional impact, recent findings published by Education Data show that over half of high school dropouts are on public assistance. Moreover, young female students who drop out of high school are nine times more likely to become single mothers. Nearly 83% of incarcerated persons are high school dropouts, and 60% of those who don't graduate high school are at a higher risk of being rearrested for repeat criminal activity. A paper on mental health consequences associated with dropping out of high school found that students who don't graduate high school are more likely to be depressed than high school graduates [20].

Although we cannot fully establish direct causation, lack of graduation is still correlated with significant negative factors as shown by the above statistics. In light of the same, many high schools provide extra support to the students in need. However, sometimes such students are identified by mere intuition or limited academic factors like only GPA, and by the time many students are provided any help, it is already too late. According to Neild and Balfanz [26], dropping out is a process, and students exhibit identifiable warnings at least one to three years before they drop out. Their study based on Philadelphia public schools found that most of the students who dropped out did so by the end of grade 10. Their further analysis showed that the students who dropped out had a 45% chance of dropping out if they had reached ninth grade, a 34% chance if they had reached tenth grade, a 23% chance if they had reached eleventh grade, and a 16% chance if they had reached twelfth grade. This indicates that the risk of a student dropping out or not graduating on time might be significantly reduced by adequately examining early warnings signs and taking appropriate action to ensure that the student stays in school longer. Moreover, since high schools cannot provide the necessary support to every student due to various resource constraints, there is a necessity to adopt a more efficient and transparent approach for doing this task.

To this end, we build a machine learning model with high efficiency and fairness to supplement educators' domain knowledge to provide an early warning about which students need targeted interventions. As described in more detail in Section 4.3, we try to meet our policy goals of efficiency, equity, transparency, and customizability (more details about why these goals were chosen is provided in 4.3). We use academic, behavioral, and demographical data provided by Muskingum Valley Education Service Centre (MVESC) for 14 school districts in Ohio to predict the highest risk students of not graduating on time at the end of grade 10. Keeping in mind that high schools have limited resources, we expect our model to help educators identify students who most need extra support, thus enhancing their chances of graduating on-time.

3. Related Work

In this section, we discuss prior work at the intersection of educational research and applied machine learning in education. We reviewed educational research to get an understanding of which features are indicative of late graduation. Building upon prior work on applied machine learning in education, we then specify how these features can be modeled to predict risk of late graduation and provide adequate and early interventions.

3.1. Factors Related to Dropout/Late Graduation

Educational research has described factors that are relevant in predicting students with high risk of dropping out or late graduation [2, 5, 29]. Balfanz et al. [1] analyzed longitudinal data of 13,000 students from 1996 to 2004, and described that students who have low behaviour scores and failures in math and English are less likely to graduate. Similarly, Burke showed that low attendance and grade point average (GPA) (e.g. attendance lower than 80% in eighth or ninth grade, GPA below 2.0 in eighth or ninth grade) are key indicators of dropouts using data of 6118 students from the Oregon school district in 2007-2008 [6].

Overall, factors that are correlated with drop-out and late graduation include exam scores/grades [22], behavioral aspects (e.g. low attendance, unsatisfactory behaviors in classes [1, 12, 21]), and social factors (e.g. single parent

[10], changing school multiple times [11], absence of close friends [12]). Among various factors, GPA and attendance are widely considered as the strongest factors to predict students at risk of not graduating on time [5, 29].

3.2. Predicting Dropout & Late Graduation

One important element of making preventive efforts is to identify students at highest risk of dropping out or late graduate. Such evidence will inform the systemic operation of the data-based diagnosis and allocation of necessary resources for at-risk students [1, 4]. Researchers and policymakers have described various approaches of an early warning system that leverages student information at school to predict whether students are on the right track for graduation [16, 7]. Prior work on an early warning system can be categorized into either rule-based or machine learning-based approaches, which we discuss next.

3.2.1 Rule-based Approaches

One intuitive approach of predicting students with high risks of dropping out or late graduation is to derive a heuristic rule with factors described in Section 3.1. Hardgrave et al. [14] utilized the grade point average (GPA) to represent students with high-risk (< 3.0), questionable-risk ($3.0 - 3.30$), and no-risk (> 3.30) of academic success. Similarly, the National High School Center (NHSC), funded by the U.S. Department of Education, recommends step-by-step instructions on how schools can identify which students are on track to graduate or not through indicators on absence rate, course failures, GPA [16]. The NHSC provides a template of Microsoft Excel that indicates a flag if one of indicators are below a certain threshold (e.g. missing 10% of instructional time, one failed course, GPA under 2.0, and two or more Fs) [16]. However, leveraging such rule-based approaches provides highly variable performance, and does not generalize well.

3.2.2 Machine Learning-based Solutions

With the increases in computing power and available data, researchers have explored the applicability of machine learning algorithms to predict students with high risk of dropping out and late graduation. Prior work extracted various features related to drop-out and late graduation (Section 3.1), and trained traditional machine learning algorithms, such as random forests [9, 17, 3], logistic regression [32, 30, 17, 25], support vector machines [19], and decision trees [32, 17] to predict at-risk students. Tamhane et al. [32] show that effective use of feature selection and machine learning can help predict risk of poor academic performance in high school as early as grade 5. Lakkaraju et al. [17] propose using risk score estimates as a primary outcome rather than just binary classification, as schools have varying resources annually and would benefit more from an algorithm which could provide them with a ranked list of students at risk.

Although prior works have shown promising results with high accuracy on predicting students with high risks of dropping out and late graduation (e.g. above 80% accuracy [25, 32, 30, 23, 9]), their evaluation methods do not have a realistic setup [17]. Specifically, prior work with high performance [25, 32, 30, 23, 9] has applied k-fold cross-validation instead of temporal validation. This causes over-optimistic estimates of performance because data from the future is included in training a model, and thus such models are likely to perform much worse when deployed. In this work, we present a validation process that considers temporal aspects of data and provide more realistic evaluation of a machine learning model to predict students with high-risks of late graduation.

In addition, our work presents a transparent machine learning model with LIME [28] that learns a locally interpretable model around prediction and presents policy makers values of important features for analysis. We also describe how such a model can be utilized to provide customized policy recommendations based on policymakers' needs (e.g. balancing efficiency of predicting high-risk students and equity over different ethnicity and school districts) instead of deploying a single model on all schools.

4. Problem Formulation

In this section, we will make our high level idea of helping schools increase their 4-year graduation rate into a concrete and actionable machine learning problem.

4.1. Goal Setting

Our team initially planned to identify students who need an intervention to graduate on time and will also respond positively to available intervention methods. This goal initially seemed reasonable because it attempts to predict not just students at high risk of not graduating on time, but the subset of those students who will also respond positively to the available intervention methods. We thought that it could be more efficient to invest resources in students who will most likely benefit from them because not all students will benefit from the resources schools can offer. However, after some data exploration, we found that only around 20% of high school students in our dataset received an academic intervention, implying that we would not be able to determine how the other 80% students would respond to an intervention. We also could not determine whether all academic interventions were given with the goal of making students graduate on time as discussed in Section 4.2. Therefore, due to such constraints posed by the data, we chose to explore the simpler goal of identifying students at the highest risk of not graduating on time. Specifically, we decided to

rank students from high to low risk using a predicted risk score. This allows educators to review risk scores and decide which students would benefit the most from the available resources.

4.2. Action Determination

From the data, we found 25 different academic intervention programs that schools used to help their students. Of these 25 programs, high school students were only enrolled in five. Although we do not have additional information about the specifics of these programs, we have identified and named those five programs here:

1. Intervention During Summer
2. Intervention During Regular School
3. Reading 180 (Online Reading Software)
4. System 44 (Online Reading Software)
5. Guided Reading (Small Group Instruction)

We expect that our model could be used to inform schools which students should be selected for which intervention programs. For example, the most risky students will perhaps need an intervention during the summer, while the less risky ones can be enrolled in a guided reading section.

It is important to note that these program names seem quite broad and each category could encompass a lot of different interventions. For example, we do not know what "intervention during regular school" means, and how serious is that intervention. It could include something as simple as a short meeting with an advisor, or something like a special class. It is also unclear whether these interventions are specifically designed to help students graduate on time, or if they just supplemental programs students can engage in to understand concepts more deeply. Although these questions remain unclear and there is no good way to answer them except asking the schools directly, these are the academic intervention programs schools in our dataset enroll their high school students in.

4.3. Policy Goals

We want our final solution to meet four policy goals that we decided beforehand. We state the policy goals below and explain their importance. In Section 6.7, we clearly describe how we met those goals.

1. **Efficiency.** We want our final model to have high efficiency. More specifically, we want to ensure that most students predicted as high risk students by our model actually are high risk students. To measure our success on this goal, we want our model to almost always performs better than current methods/baselines used by schools.

2. **Equity.** In addition to high efficiency, we want to ensure that our model is also fair across different ethnicities and school districts. We are interested in fairness over ethnicities because many AI algorithms in the past have been called out for being unfair over ethnicity [24], and we want to ensure our algorithm is not among them. We are also interested in fairness over school districts because our data exploration shows that many school districts collect very little data, naturally putting their students at a disadvantage. More details about data collection differences among school districts are provided in Section 5.4.

3. **Transparency.** To make our model more trustworthy, we want to help policymakers interpret our model results by providing relevant feature importance scores. We aim to provide both global (as measured over the full validation set) and local (as measured over just one student) feature importances to maximize interpretability.

4. **Customizability.** To maximize flexibility, we want to make our pipeline and the final model as customizable as possible to the needs of policymakers. For example, we earlier mentioned that we plan to make our models both efficient and equitable, but we would like to let policymakers choose the exact trade-off between the two goals.

4.4. Analytical Formulation

In a sentence, we intend to identify 5% of students at the end of grade 10, who are at the highest risk of not graduating on time (i.e. within 4 years) so appropriate interventions can be provided early. We now break down the above sentence to justify some of the choices we made.

First, note that choosing to identify 5% students was based on an assumption that most schools can provide meaningful interventions to up to 5% of students. This of course will vary for school to school, and the final deployed model can be easily adjusted based on those needs.

Second, we chose to provide recommendations at the end of grade 10 because of the following two reasons. Firstly, we have two years of high school data for students by the end of grade 10, and this will help us use more high school information of the student as features of our model, making predictions more reliable. Secondly, there are still at least two years remaining after end of grade 10, which should be sufficient time for schools to intervene effectively on students.

Third, the reason for predicting risk scores of students instead of something else is described in Section 4.1.

4.5. Solution Overview

We design, implement and evaluate many machine learning models to predict a risk score of a student of not graduating high school on time (i.e. 4 years). We then predict the top 5% students with the highest risk scores as ones who need intervention. We evaluate our models on both efficiency and equity, and ensure that our validation process exactly mimics the deployment setting as described in Sections 6.5 and 6.4. We also offer significant flexibility to policymakers to choose how to manage the trade-off between efficiency and equity by allowing them to choose the model that best fits their needs instead of recommending a one-size-fits-all model.

We also offer methods to policymakers to understand their chosen model better by both providing global (over the entire validation set) and local (per student) feature importances to help them both understand how the model is taking decisions overall and why a specific student was assigned a specific risk score. To compute local feature importances, we use a popular method called LIME [28].

We provide full details of our solution in Section 9.

5. Data Description

In this section, we describe the data available to us in detail and share results of data exploration.

5.1. General Description

We primarily used a dataset provided by the Muskingum Valley Educational Service Center (MVESC) that contains student records relating to academic, demographic and behavioral factors from 2006-15 of 14 school districts in Ohio. Specifically, we have access to grades, attendance, discipline behavior, extracurriculars, past interventions and demographic records of over 30,000 high school students.

Our data is collected from 21 different high schools in Ohio. Some quick preliminary exploration shows that the dataset has information about grades for 72.3% students. Also, according to our definition of labels discussed in Section A.2, 93.4% of students graduated in 4 years, and 1.76% students dropped out.

5.2. Sufficiency

Recall that we have 10 years of data (2006-2015). As fully explained in 6.4, the temporal gap that needs to be maintained between training and validation set is that of 2 years. Briefly, this is because we are predicting at the end of grade 10, and it takes 2 years since then to graduate on time. Since it takes 2 years to graduate after finishing grade 10 and our data provided ends in the 2015 academic year, we have labels for students who finished grade 10 in or before the 2013 academic year. Also, since the temporal gap between training and validation set is 2 years and we have data starting from the 2006 academic year, the earliest validation set we can use will be that of 2008 academic year (and the corresponding training set will be of 2006 academic year). Altogether, we have data that can be used for validation from academic years 2008 to 2013, meaning that we can use 6 validation sets overall, which are sufficient for strong evaluation. Therefore, we have enough data to reliably perform validation.

Label	GPA	Absences / Year	# of Disabilities	Disciplines Incidents Per Year	# of Extracurriculars
1 (did not graduate)	1.78	11.34	0.29	1.95	0.46
0 (graduated)	2.86	7.65	0.14	0.72	1.77

Table 1: Statistics about an average student for each label

5.3. Feature Exploration

We first do some exploration to find any trends between the provided features that could help guide our solution.

Table 1 shows statistics of an average students belonging to each label. The process of computing labels is described fully in Section A.2. The "GPA" column shows the average final year GPA of students in different classes. An important observation to make here is that students who do not graduate on time tend to have a significantly lower GPA than students who do. The variation in GPA also confirms our intuition that past GPA could be an important feature in the model. The "Absences/Year" column shows the average number of yearly absences by different groups of students. Note that students who do not graduated on time have very high absence rate. Another important observation is that on average 0.29 students who do not graduate on time tend to have some disability. This indicates that students who graduate late also tend to face natural difficulties. Furthermore, note that students who graduate on time tend to be more involved in extracurricular activities. We performed this label-wise analysis with 15 variables, but only a subset is shown in Table 1.

Figure 1 shows the relationship between few interesting variables in a label-agnostic manner. It can be seen that students who have a high GPA tend to have a low absence rate. Furthermore, students who have low absence rate get lower number of academic interventions, perhaps indicating that schools currently prescribe interventions to students who skip school more often. Similarly, note that students who have a high GPA usually also receive low number of academic interventions.

5.4. Bias and Fairness Exploration

When using machine learning models to influence public policy, it is important that we are aware of bias in our models against different classes of people. Besides the more commonly known classes like ethnicity, one class that is more unique to our problem is school district of a student.

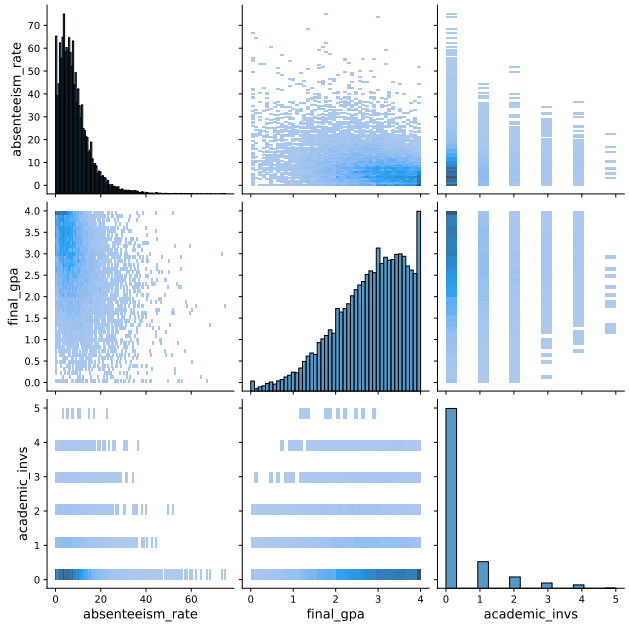


Figure 1: Scatterplot matrix showing two dimensional histograms for pairs of variables. Darker regions show more concentration of data points

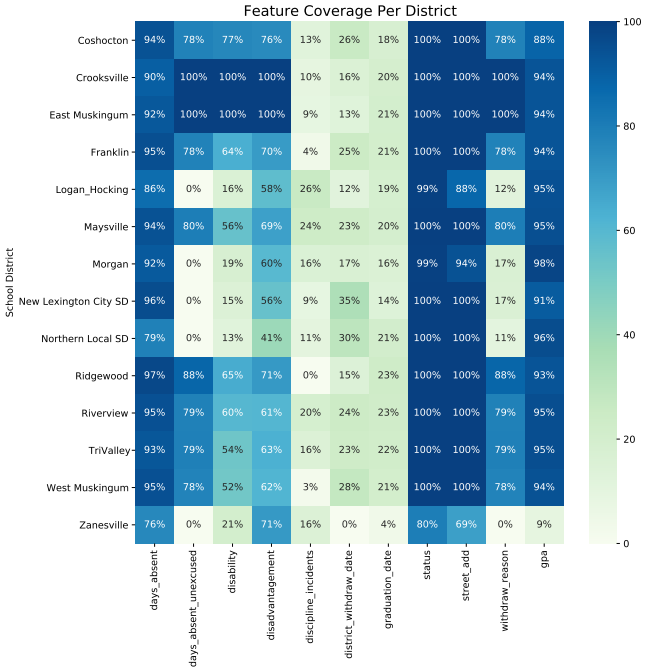


Figure 2: Coverage of Raw Features Per District. The values inside each cell show the percentage of non-null features for the corresponding school district

One can imagine that different areas of the state do not have access to the same funding and resources, and thus may be disproportionately affected by machine learning methods.

Figure 2 shows the coverage of important raw features (i.e. percentage of non-null records) for each school district. This information can help us understand if certain attributes are not recorded by certain school districts, and how constructing downstream features based on those attributes may unfairly impact students for who that data is not present. We observe that there is good coverage for `days_absent`, `status`, and `street_add` across all districts. GPA, a potentially important predictor, has good coverage for most districts, but very poor coverage for Zanesville where only 9% of records have an associated high school GPA. Note that important attributes like `graduation_date` and `discipline_incidents` have low coverage across many districts. For some districts, information about `withdraw_reason` is also recorded sparsely, which could put students of those districts at a slight disadvantage since `withdraw_reason` can help us determine if a student graduated or not. This analysis also shows that deciding how to deal with the null values based on the caveats attached with each attribute would be important in determining the success of the model.

6. Solution Details

Having formulated the problem concretely and described the available data concisely, we now provide the details of our solution. The link to our well documented code that implements all the details described in this section is https://github.com/dssg/mlpolicylab_fall20_schools3.

6.1. Prediction Units, Labels and Features

Since we want to predict risk of not graduating on time for students after they finish grade 10, our prediction unit is naturally a student who just finished grade 10 in a given academic year (for example: student with ID 153 in grade 10 in academic year 2012).

Note that the available data does not provide clear information about the graduation status of a student. So, we have to construct our own method to create labels. The exact details of how we do it are lengthy and fully provided in Section A.2. For purposes of this section, it suffices to say that we create a binary label. A "0" indicates that the student graduated in 4 years, while "1" indicates that the student did not graduate within 4 years. Therefore, we treat risk score prediction as a binary classification problem.

As for features, we use various different features to predict whether a student graduated on time or not. In the following, we briefly talk about the different types of features

we use, but the full list of features used is in Table 2:

- **Academic.** This type of features include past GPAs, number of classes taken, GPA in percentile form relative to students in the same school or district, etc.
- **Attendance.** This type of features include number of absences, number of unexcused absences, number of absences due to different types of reason (example: number of absences due to "medical emergencies"), etc.
- **Intervention.** This type of features include number of interventions received of each type (for example: number of interventions received of type "academic").
- **Standardized Test Scores.** This type of features include the results of OAA/OGT standardized tests conducted in Ohio from grades 6 to 8.
- **General Information.** This type of features include age, disability status, gender, number of discipline incidents,

Note that we also parsed certain more features from the American Community Survey (ACS) like the education and income status of the neighborhoods in which a given student lived. However, we did not observe any performance gain in any of our models, so we dropped the ACS features to save computation costs.

In summary, using the features and labels described above, our model predicts a risk score for every prediction unit (i.e. every student who finished grade 10 in a given year). Assuming that schools can intervene on 5% of all students, we recommend intervening on the 5% highest risk students. This 5% threshold is easy to change based on policymakers' needs.

6.2. Feature Pre-processing

We employ multiple pre-processing methods to both extract high performance out of our models and convert the data to a format most models can ingest. These methods are described below in the order they are applied:

1. **Discretization.** In this step, we convert all categorical variables to dummy variables. More specifically, if a categorical variable takes values from n categories, then we convert it to a one-hot vector of $n+1$ categories, so that the i^{th} value is 1 if the variable takes the i^{th} category while all other values are 0. The $(n+1)^{th}$ index is 1 when that variable's value is null/missing.
2. **Imputation.** Having handled all missing values for categorical variables in the previous step, we here handle missing values of continuous variables. Most often, continuous variables are imputed with their mean value. In some cases, they are also imputed by zeros in case there is a reason to believe that a null value for a certain variable corresponds to zero. For example, the number of discipline incidents per year is imputed with a zero because discipline incidents are only recorded for students who have committed them in our data. Table 2 shows how each feature is exactly imputed. Also, an additional variable is added in the model to denote whether the data in any field was missing or not, so that our model can use lack of data as an input feature too.
3. **Standardization.** Finally, we standardize (subtract mean and divide by standard deviation) all *continuous* variables because certain models like neural networks benefit greatly from such standardization. Note that dummy variables created in the first step of discretization are not affected by this step.

We take extreme care during the pre-processing period to ensure that the training data does not mistakenly use any information from the validation set while the imputation or standardization step. Therefore, we ensure that all imputation happens separately on the training and validation data, so that imputation methods like the mean do not leak any information. We also ensure that standardization on the validation/test set happens using the mean and standard deviation of the training set so that one feature value refers to the same original value in both the training and validation sets. If we use different means and standard deviations in the training and validation set, then such correspondence will not hold between same values in the training and validation sets.

6.3. Models & Baselines

We test out a variety of different machine learning models to ensure thoroughness. We list the types of models tried below, and the exact hyperparameter configurations used of all models in Section A.3.

- Random Forests
- Decision Trees
- Logistic Regression
- K Nearest Neighbors
- Neural Networks (Multi-layer Perceptrons)
- Kernelized Support Vector Machines

In addition to the machine learning models used, we also construct intuitive baselines to better compare our performance against current approaches used by schools to identify high risk students. Specifically we construct three baselines as shown below:

- **GPA Baseline:** For the GPA baseline, we simply rank students according to their current year (grade 10) GPA

and classify students with the lowest 5% GPA as highest risk students.

- **Absenteeism Baseline:** In this case, we classify 5% students with the highest absence rate (i.e. number of absences per year since grade 6 / number of academic years since grade 6) as highest risk students.
- **Discipline Baseline:** Finally, we classify 5% students with the highest discipline incidents in the current year (grade 10) as highest risk students.

6.4. Validation Process

Since in the real world, our project will be evaluated in the present given training data of the past, we do temporal validation. In other words, the validation set will have more recent dates than the training set. Furthermore, in a real application setting, when predicting the risk scores for the batch that just finished 10^{th} grade, the most recent training data we would have will be of the 10^{th} grade batch 2 years ago since more recent batches would still be in high school and would not have not graduated. To mimic this scenario, we ensure that appropriate temporal gap (i.e. 2 years) exists between all training and validation sets. Also, to maximize data used for training, for a given validation dataset, we use all training data from 2006 up to 2 years before the validation year. This is more clearly explained using an example shown in Figure 3.

All six validation sets that we used are shown in Figure 6. As explained, we validate over the batch of students who finish grade 10 in 2008, 2009, 2010, 2011, 2012, and 2013. For each validation set, we use training data of batch of students who finished grade 10 in the end of 2006, 2006-07, 2006-08, 2006-09, 2006-10, and 2006-11 respectively. Notice that the more recent validation sets use more amount of training data. Therefore, we expect the models to perform better on more recent validation sets.

6.5. Evaluation Metrics

Recall that our policy goals involved ensuring efficiency and fairness as described in Section 4.3. To measure efficiency, we naturally use precision at the highest 5% risk students. In non-technical terms, precision measures the percentage of students who actually do not graduate within 4 years over all students which our model predicts will not graduate in 4 years.

To measure fairness over ethnicities, we use the ratio of false negative rate (FNR) of our model among Blacks to the FNR among Whites. In non-technical terms, FNR is the probability of falsely denying intervention to a student who actually needed it. So, we want the ratio of FNRs among Blacks and Whites to ideally equal 1, meaning that the probability that we falsely deny help to a black student is equal to the probability that we falsely deny help to a white student.

To measure fairness over school districts, we again use the ratio of the median FNR for students who go to school districts that collect little data to the median FNR for students who go to school districts that collect enough data. Using the same logic as before, if this ratio is 1, then the median FNR of students that go to school districts that collect little data is equal to the median FNR of students that go to school districts that collect enough data. To clearly define "districts that collect little data", we say that districts that have less than 50% coverage for at least over 50% of features shown in Figure 2 collect little data. These districts turn out to be Zanesville, New Lexington City SD, Logan Hocking, Morgan, Northern Local SD.

6.6. Evaluation over Multiple Datasets

Note that the metrics defined in the previous section 6.5 are metrics that can be calculated over one validation set. However, as described in Section 6.4, we have multiple different validation sets, and in this section, we discuss how we combine the different evaluation metrics over all validation sets. Combining these metrics is useful because it helps us generate just a single scalar number for each model, allowing different models to be easily compared.

To find the most stable model in terms of efficiency, note that we cannot naively take the mean of precision at top 5% for all validation sets because each validation set uses a different amount of training data. It would be instead better to take a weighted mean so that we give more weight to the precision obtained over validation sets that have larger corresponding training sets, since we expect our model to be more reliable when more training data is given. Note that giving more weight to results of models with more training data also consequently gives more weight to more recent validation sets since more recent validation sets have more training data as explained in Section 6.4, and such a property will be desirable since the model will only be deployed in the future.

To explain how we find the most stable model (while considering differing amounts of training data), let $\forall_i (T_i, V_i)$ be all pairs of corresponding training and validation sets. Let $w_i' = |T_i| / \sum_j |T_j|$ and $w_i = w_i' / \sum_j w_j'$ where $|T_i|$ is just the size of the training set T_i. If the top 5% precision for the validation set V_i is p_i, then we define the "stability score" of a model over all training and validation sets as $\sum_i w_i \cdot p_i$. The model with the highest "stability score" is considered as the most stable model.

To combine fairness metrics computed over different validation sets, we do not use the weighted average method shown above based on amount of training data because we want our models to be fair regardless of how much training data is used. Therefore, using the more straightforward average suffices in this case. Recall that fairness over both school districts and ethnicities is measured as the ratio of

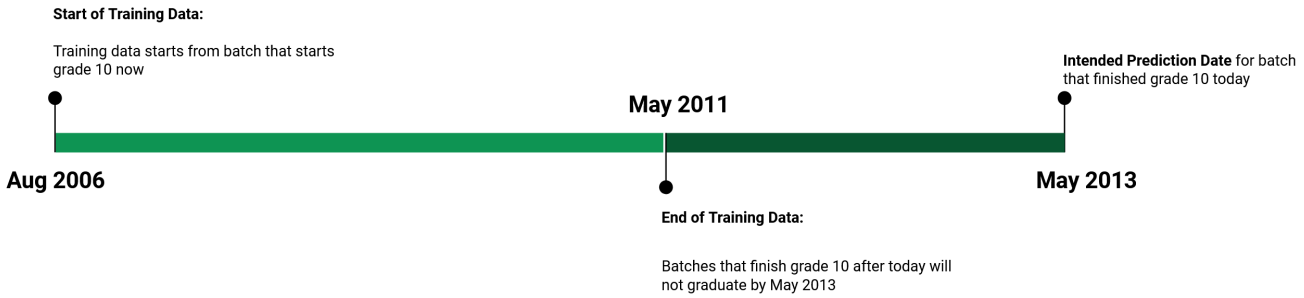


Figure 3: To predict the risk scores of students who finished grade 10 in May 2013, we use all training data upto May 2011 because students who started grade 10 after May 2011 would not have graduated by May 2013. Thus, the validation procedure exactly mimics the deployment setting.

false negative rates among two groups, and the value of 1 indicates that the model is perfectly fair. So, if the FNR ratio is r_i for the validation set V_i and there are n validation sets, then we simply compute $1/n \cdot \sum_{i \in [n]} |1 - r_i|$ to compute the overall fairness score of a model over all validation sets. Naturally, the model with most stable fairness will have the minimum possible value of the sum above. We use the same method to compute overall fairness over both school districts and ethnicities.

6.7. Meeting Policy Goals

In this section, we describe how we tried to meet the 4 policy goals described in Section 4.3. In Section 6.5, we describe how our evaluation metrics are designed to evaluate models for efficiency and fairness. So, we here focus specifically on transparency and customizability.

To ensure transparency, policymakers will be able to see the features that are most important for the model in determining the risk score for a particular student. This will allow policymakers to trust the model more because they will know how it makes decisions overall. However, just providing such global/overall feature importances may not be sufficient in all cases. For example, global feature importance results cannot inform policymakers about why the model predicted a certain risk score for a specific student. To maximize transparency, we also provide feature importance results for a particular prediction using a popular method called LIME [28]. Figure 4 shows an example use of such local feature importance results.

We take the next policy goal, customizability, quite seriously. We try to make our solution customizable in two ways. Firstly, we follow very strong software engineering practices in our code. For example, we extensively use principles of object oriented programming like inheritance, polymorphism, abstraction, and encapsulation throughout our pipeline. This allows our code to be very flexible and reusable. We also ensure that our entire code can be fully controlled by configuration files. Therefore, we can adapt

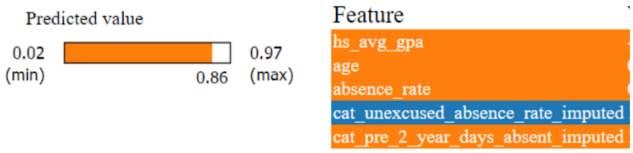


Figure 4: An example of obtaining local feature importances to explain why a particular student received a particular risk score. For example, the student in this case received a risk score of 0.86 because of factors like average high school GPA (`hs_avg_gpa`), age (`age`), and average number of absences per year (`absence_rate`). The other two important features include whether the unexcused absence rate was missing or not (`cat_unexcused_absence_rate_imputed`) and whether the number of days the student was absent two years ago was missing or not (`cat_pre_2_year_days_absent_imputed`)

to various needs of our policymakers just by a change in configuration files, making our solution be highly customizable. Secondly, we present policymakers with different choices of models and allow them to pick the one that best fits their needs as explained in Section 7. These different choices essentially allow them to pick the right trade-off between equity and efficiency according to their needs.

7. Evaluation & Policy Recommendations

Since we essentially provide a menu of choices for policymakers to choose from, we cannot discuss evaluation results and policy recommendations separately. In this section, we discuss the various choices we offer to policymakers so they get maximum flexibility over how best to decide the trade-off between efficiency and equity. We offer them three default choices listed below:

1. Most Stable Efficiency with "Reasonable" Fairness **(Most Recommended Choice)**

2. Most Stable Fairness over Ethnicities
3. Most Stable Fairness over School Districts

We discuss the models that we recommend for each of these three choices below, but we want to note that our pipeline is quite flexible and policymakers are welcome to design more custom options in addition to the three default ones provided.

7.1. Most Stable Efficiency with "Reasonable" Fairness

We strongly recommend this model to policymakers because it both achieves the most stable efficiency and reasonable fairness over ethnicities and school districts (exact method of choosing the model with most stable efficiency is described in Section 6.6). The evaluation metrics of this model are shown in 5a.

From the graph, note that the model has relatively stable precision over all validation years, and almost always beats all baselines (except in 2008 possibly due to having least data to train on, and in 2012 where it is a close second to the GPA baseline). Note that its precision is higher in 2012 and 2013 compared to previous years, indicating that using more training data is helping the model.

As for fairness, note that the model is quite fair over districts especially in 2011 and 2013 (we do not have sufficient data to calculate fairness over districts before 2011). It's fairness reduces in 2012, but it is not unacceptably low. Moreover, the model is relatively fair over ethnicities too. As can be seen, it becomes more fair as more data is provided and is quite fair in 2013, the most recent validation set. This is a positive sign because whenever the model will be deployed, it will be provided with even more data.

7.2. Most Stable Fairness over Ethnicities

If policymakers prioritize fairness among different ethnicities over efficiency, then we would recommend the model highlighted in Figure 5b. This model has lower precision and doesn't necessarily beat the best baseline, but performs reasonably well in the most recent two validation sets.

In terms of fairness, we can see that the ratio of probabilities of falsely denying intervention between ethnicities consistently hovers around 1, indicating that this model is mostly fair over ethnicities. For fairness across districts, the ratio remains close to 1 but trends away in the more recent years, but it still remains close to 1 in absolute terms.

7.3. Most Stable Fairness over School Districts

Finally, if policymakers prioritize fairness among different districts, then we would recommend the model shown in Figure 5c. Of the three recommended models, this one tends to perform in the middle in terms of precision and, promisingly, the precision also has an upward trend in more recent years.

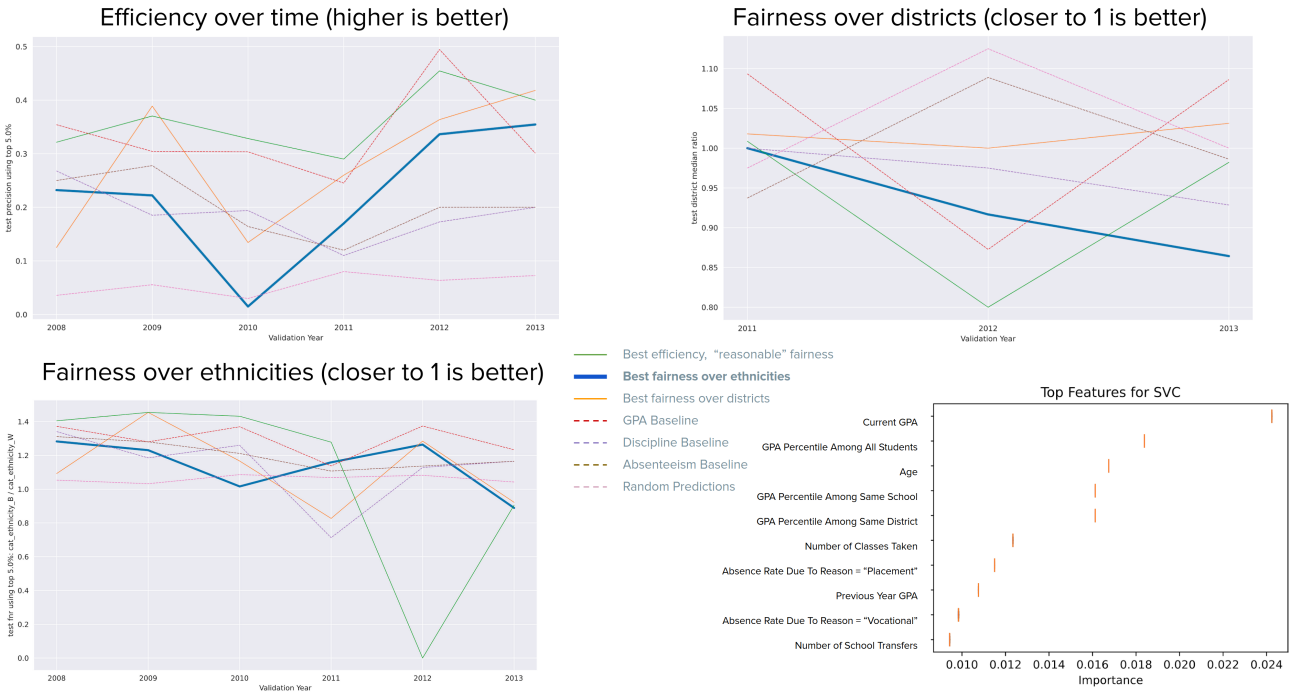
When looking at the fairness metrics, we can see that this model achieves a ratio close to 1 across districts, and this result is quite consistent across the validation sets. Note also that it is more fair than all baselines as well. As for ethnicities, the fairness ratio is less consistent, but mostly oscillates near a ratio of 1, which means that it achieves reasonable fairness across ethnicities too.

7.4. Other Policy Recommendations

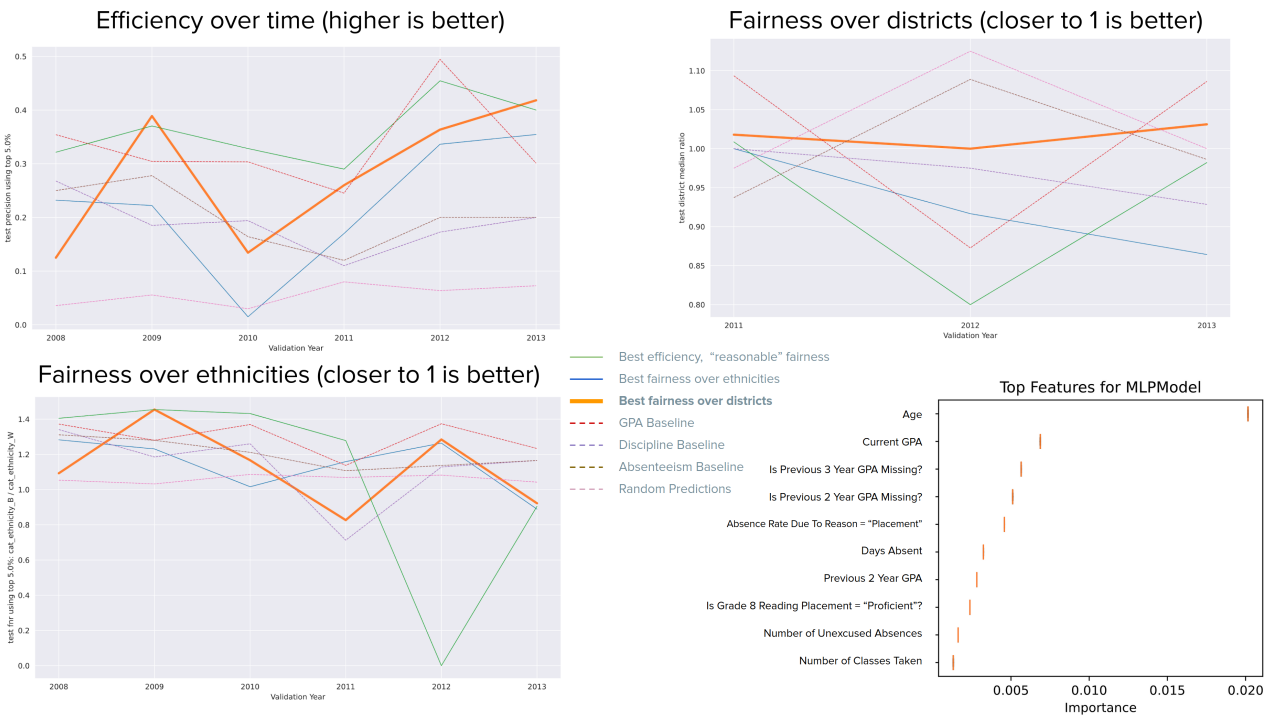
In addition to which models policymakers can use, we also have other important recommendations listed below:

1. **More Data Collection:** Since our models have shown promising results compared to current baselines, it may be worth collecting more complete data across many different school districts in the US. Some characteristics of such a dataset are described below:
   (a) **Completeness:** As described in Section 10, the current data has many missing attributes, and imputing them potentially adds bias in models. Therefore, an attempt should be made to recover as complete data as possible. Also, the current dataset does not have clear information about graduation status of a student, and we have to infer that based on intuitive rules. We think machine learning models will greatly benefit if the input data has very clear indication of whether a student graduated or not because such information directly serves as the label of the model.
   (b) **Diversity:** Currently, the data only comes from one state, which may be non-ideal. If data can be collected from multiple different places, then we could be more confident in its generalizability.
   (c) **Breadth:** A new dataset can possibly also record other factors that may influence students' graduation status, such as students' physical health, mental health, and their social relations inside and outside school. Such features could help models increase their predictive performance.
2. **Model Usage.** Although our model does quite well compared to existing baselines especially on the most recent validation set, it is important to note that in absolute terms, our best model's precision is only around 40%. This means that about 60% of students classified as high risk by our model may not actually be of that high-risk. Therefore, we do not recommend solely relying upon our model's risk scores to decide which students should be intervened upon. Instead, we recommend using our model's risk scores as one factor to augment the current methods used in deciding which

(a) All evaluation metrics plotted over time with feature importance scores of the model with most stable efficiency with "reasonable" fairness



(b) All evaluation metrics plotted over time with feature importance scores of the model with most stable fairness over ethnicities



(c) All evaluation metrics plotted over time with feature importance scores of the model with most stable fairness over districts

Figure 5: All evaluation metrics plotted over time with feature importances for all recommended models

students should be intervened upon. Throughout our project, we have placed significant emphasis on transparency, so it is relatively easy for policymakers to decide when to place more weight on our model's predictions and when to possibly discount them. We recommend that policymakers make use of our pipeline's ability to provide feature importances per student to assess how much to trust the model's predictions in determining whether a student should be given intervention. This process can also be automated in which a non machine learning based algorithm can simply look at the top feature importances used in every generating the risk score for every student, and decide how seriously should the schools trust each prediction based on how sensible the feature importances are.

3. **Adding more interventions.** As described in Section 4.2, currently it seems that schools only provide 5 academic interventions to high school students, which are listed in that section. These interventions seem quite broad, and it is not clear if these interventions are designed with the goal of helping students graduate on time. With the current intervention names, it seems possible that some of them are meant to provide just a more detailed understanding to students about certain topics, and not necessarily help them graduate on time directly. Therefore, it is paramount that schools invest time and resources in designing interventions with the specific aim of helping students graduate on time, otherwise no machine learning model can really help improve the final graduation rates.

8. Discussion

8.1. GPA is a strong baseline

We observe in figures 5a, 5b, 5c that GPA baseline is a quite strong baseline, and it is hard for models to significantly outperform it. Even when outperformed, it is only outperformed by less that 0.1. This observation seems intuitive because GPA directly influences graduation outcomes: if the student's GPA is too low, she may not graduate on time because the low GPA could be caused by failing in a required class, which would need to be repeated. Therefore, it is natural that the GPA baseline would be quite strong to beat as it is directly tied with the final outcome. Also, note that we consider GPA to determine graduation status for students who we cannot use any other more reliable information to determine their graduation status. This is explained more in Section A.2. Although the GPA used in the label is the final high school GPA instead of grade 10

GPA which is used for the baseline, the final GPA is heavily linked to grade 10 GPA due to GPA being cumulative in nature, which may again make the cause GPA baseline to be very strong. An important question for financially constrained schools will be whether machine learning models provide enough value over the GPA baseline to justify having to invest in their developmental and investment costs. Although we are hopeful that models will perform better with more data, it remains to be seen if they would perform significantly better than the GPA baseline.

8.2. Change in Data Behavior in 2012

We have observed that there could be some strange issues in the 2012 data that affect many baselines and models. For example, as seen in Figure 5a, GPA baseline performs exceedingly well in 2012, the model makes no error in classifying any black student, and the fairness over districts reduces strangely. We tried hard to understand what could be causing such issues like looking at the distribution of certain important features and labels, and checking our code to ensure there is no bug. However, despite significant effort, we were not able to find any clear cause that could explain the kind of behavior we observe in 2012. We also tried finding any regulation change in Ohio schools in 2012 that could explain the change, but could not find anything relevant. We also observed that other teams working with the same data had changes in their metrics in 2012 too, which further confirmed our suspicion that the issue is in the underlying data. We believe that a conversation with MVESC and the administrators of school districts that we have the data for could help uncover possible regulation/data collection changes in 2012, which could further provide important insights in how that data should be handled in our model.

8.3. Interpreting Feature Importances

In figures 5a, 5b, and 5c, we show the top ten most important features in the for each of our recommended models. We can see that each model gives high importance to features that we believe make sense and are intuitive, such as grade, disciplinary instances, and absences. High grades, low absences, and low disciplinary incidents are all characteristic of a good student. Thus, it makes sense that each of these features are highly important for predicting graduation. The other top features also make sense for helping predict timely graduation, although for less obvious reasons. For example, age is rated highly by all three models. Although the majority of students in a particular grade are likely to be similarly aged, students with a history of poor academic performance may have had to repeat previous grades, and therefore have higher age than their peers. These students would be likely to dropout/graduate late given their history, and therefore age is an important feature. Another important but less intuitive feature was missingness flags, which could indicate the presence of abnormalities in a student's education such as transferring schools. These abnormalities could be caused by personal circumstances that are less than ideal for students' educational environments. Overall, we believe our feature importances results help us be confident that our models are reasonable.

9. Field Trial Design

In this section, we propose a method to design a field trial using our model to test its effectiveness. For this design, we assume that the main goal of schools is to increase 4-year graduation rates of their students. We also assume that schools can help a total 5% of students at the end of grade 10, which is consistent with our assumptions made during analytical formulation phase.

For the field trial, we suggest dividing all students who just finished grade 10 into 2 groups randomly of equal size. Let us call these groups control and treatment. For the control group, schools should use their current methods (i.e. the method they would have used had our model not existed) to identify the 5% students who they would intervene on. For the treatment group, schools should use our machine learning model to either directly pick the 5% students who should be intervened or use our model together with their domain knowledge to pick the 5% students. In other words, they should pick the 5% students who should be intervened upon using the method they would use when our model's risk scores are available. This could include totally trusting our model predictions, or using the model predictions to aid in their decision making. Since the control and treatment group both have 50% students and we have chosen 5% students from each group, we would have picked 5% overall students who would be intervened upon.

After following the regular intervention methods on these 5% students, schools should compare the final graduation rate of both the control and the treatment group. When the graduation rates are available for both groups, a hypothesis test can be conducted at the 5% significance level with the null hypothesis being that the graduation rates of the control and treatment group are the same and the alternative being that the graduation rates of the treatment group are higher. If we can reject this null hypothesis at the chosen significance level, then schools can be confident that our model provides statistically significant value to them in increasing their graduation rates.

We recommend that schools perform the above mentioned field trial for at least two batches of grade 10 students. This means that the field trial will take at least 3 years to complete because each batch takes at least 2 years to graduate after interventions are provided at the end of grade 10. The field trial is extremely crucial to ensure that schools are benefiting positively by using our proposed model

10. Limitations & Future Work

In this section, we discuss some limitations of our work caused either by the data provided or the analysis method employed. We also discuss some suggestions for future work to overcome the limitations mentioned.

10.1. Data Limitations

As for data, low data collection by some schools can cause suboptimal model performance. For example, we do not have any grades for 28% of students and 35% of districts do not provide any data before 2011. The absence of features forces us to impute missing values by certain rule based methods like mean, median, zero, etc. Such rule based methods often become sources of bias in the entire pipeline, and significantly contribute to suboptimal performance.

Apart from missing data, note that we do not even have clear information about which students graduated from high school and which did not. This prompted us to again construct certain rule-based methods to decide how to decide whether students graduated, which is explained fully in Section A.2. Although the rule based methods were backed by strong domain understanding, they still are not likely to be perfect and therefore could induce additional bias in our analysis.

Finally, we only have accredited data of 14 school districts in Ohio. However, there are a total of 13,000 school districts in the United States, and 611 just in Ohio. So, we do not have any reason to believe that the performance of our models will generalize to other states in the US, or even in other districts of Ohio. This could be a major barrier in implementing such a system more widely.

10.2. Analysis Limitations

As for limitations of our analysis method, recall that we currently recommend on intervening on the top 5% students at the highest risk of not graduating high school in time. Even if schools intervene on these 5% students, we may not still be able to increase on-time graduation rate because the riskiest students may not be the ones who respond most positively to the interventions that schools can offer. So, if the sole objective of schools is to increase graduation rates because those could be linked to high rating and high funding, we cannot be sure how effective would interventions on highest risk students may be. Admittedly, schools could be more efficient by intervening on students who are most likely to benefit, but such students are quite difficult to identify.

In addition, our method currently does not recommend which types of interventions should be applied to which students. This is mainly because of two reasons, both of which arise due to insufficient/improper data. One, only around 20% of high school students in our dataset received an academic intervention, which is not enough to support such an analysis because we would have to discard the other 80% of students from our analysis. Two, currently, high school students are only offered only one of five academic interventions. These intervention categories are fairly broad and can encompass multiple different interventions under one term. For example, one such category is "intervention during summer", which could mean a lot of different things. Therefore, it would be challenging to build a model that recommends the type of intervention a student should get because data is not granular enough.

10.3. Future work

To improve upon our work, we feel the following can be great directions for future work.

1. **Data Collection**. As described in Section 7.4, we strongly recommend collecting more data with the aforementioned characteristics of completeness, diversity, and breadth. This is especially important since our work has shown machine learning models can actually add more value in helping schools intervene on high risk students compared to current baselines. We provide more details about how more data can be collected in Section 7.4.

2. **Field Trial.** As discussed in Section 10.2, the current models only help identify students with highest risk of late graduation, but not the students most likely to graduate on time if intervened upon. Therefore, it is quite important to design a field trial to study the exact effects of the proposed model on graduation rates of schools. We provide more details of the design of the field trial in Section 9.

3. **Advanced Models**. We could also use some more advanced machine learning models to predict graduation risk in a better way. Since the models we think can be used in predicting graduation risks have never been tried before on social policy problems before to the best of our knowledge, we believe that using such models is not just "future work", but an important avenue for further research. Therefore, we talk about the specifics of this point in Section 11.

11. Future Avenues of Research

While working on this project, we realized that we needed many heterogeneous features to predict the risk of not graduating on time effectively. In this context, "heterogeneous" means a combination of continuous, discrete, time dependent, and time independent features. However, some machine learning models like multi-layer perceptrons are not directly designed for handling such different types

of heterogeneous features for the following reasons. For example, such models do not know that categories are exclusively assigned (i.e. one prediction unit can be exactly assigned one category) and that any two dummy variables corresponding to different categories of the same field are related. Moreover, in some cases, it may be beneficial to embed domain knowledge into our machine learning models by informing them that certain features measure similar kinds of information and should therefore not be treated as completely independent to each other. For example, all GPA based features are related/dependent because they strictly measure academic performance of a student, and the model could benefit from knowing about such dependence.

In light of the above analysis, we hypothesize that certain advanced deep learning models used to predict click rates on ads may find a very strong use in public policy problems. Such models are specifically optimized to handle thousands of heterogeneous features and capture feature interactions explicitly. They could also be used to embed knowledge of features that are related to each other. Since interactions are captured explicitly in the model, feature groups with high interactions could also be clearly identified and studied to possibly further enhance domain knowledge [27]. We have also identified some exact architectures that could be used in public policy problems such as DeepFM [13], xDeepFM [18], AutoInt [31], and Deep & Cross Network [33].

References

[1] R. Balfanz, L. Herzog, and D. J. Mac Iver. Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, 42(4):223–235, 2007.

[2] S. Battin-Pearson, M. D. Newcomb, R. D. Abbott, K. G. Hill, R. F. Catalano, and J. D. Hawkins. Predictors of early high school dropout: A test of five theories. *Journal of educational psychology*, 92(3):568, 2000.

[3] C. Beaulac and J. S. Rosenthal. Predicting university students' academic success and major using random forests. *Research in Higher Education*, 60(7):1048–1064, 2019.

[4] T. Blount. Dropout prevention: Recommendations for school counselors. *Journal of School Counseling*, 10(16):n16, 2012.

[5] A. J. Bowers, R. Sprott, and S. A. Taff. Do we know who will drop out? a review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, pages 77–100, 2012.

[6] A. Burke. Early identification of high school graduation outcomes in oregon leadership network schools. rel 2015-079. *Regional Educational Laboratory Northwest*, 2015.

[7] B. Carl, J. T. Richardson, E. Cheng, H. Kim, and R. H. Meyer. Theory and application of early warning systems for high school and beyond. *Journal of Education for Students Placed at Risk (JESPAR)*, 18(1):29–49, 2013.

[8] A. P. Carnevale, N. Smith, and J. Strohl. Recovery: Job growth and education requirements through 2020. washington, dc: Georgetown university center on education and the workforce, 2013.

[9] J. Y. Chung and S. Lee. Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96:346–353, 2019.

[10] R. G. Croninger and V. E. Lee. Social capital and dropping out of high school: Benefits to at-risk students of teachers' support and guidance. *Teachers college record*, 2001.

[11] B. Dalton, E. Glennie, and S. J. Ingels. Late high school dropouts: Characteristics, experiences, and changes across cohorts. descriptive analysis report. nces 2009-307. *National Center for Education Statistics*, 2009.

[12] R. B. Ekstrom et al. Who drops out of high school and why? findings from a national study. *Teachers College Record*, 87(3):356–73, 1986.

[13] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*, 2017.

[14] B. C. Hardgrave, R. L. Wilson, and K. A. Walstrom. Predicting graduate student success: A comparison of neural networks and traditional techniques. *Computers & Operations Research*, 21(3):249–263, 1994.

[15] B. Hussar, J. Zhang, S. Hein, K. Wang, A. Roberts, J. Cui, M. Smith, F. B. Mann, A. Barmer, and R. Dilig. The condition of education 2020. nces 2020-144. *National Center for Education Statistics*, 2020.

[16] L. Kennelly and M. Monrad. Approaches to dropout prevention: Heeding early warning signs with appropriate interventions. *American Institutes for Research*, 2007.

[17] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1909–1918, 2015.

[18] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, and G. Sun. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1754–1763, 2018.

[19] S. N. Liao, D. Zingaro, K. Thai, C. Alvarado, W. G. Griswold, and L. Porter. A robust machine learning technique to predict low-performing students. *ACM Transactions on Computing Education (TOCE)*, 19(3):1–19, 2019.

[20] J. H. Liem, C. O. Dillon, and S. Gore. Mental health consequences associated with dropping out of high school. 2001.

[21] J. L. Mahoney. School extracurricular activity participation as a moderator in the development of antisocial patterns. *Child development*, 71(2):502–516, 2000.

[22] G. J. Marchant and S. E. Paulson. The relationship of high school graduation exams to graduation rates and sat scores. *education policy analysis archives*, 13:6, 2005.

[23] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura. Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124, 2016.

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[25] P. A. Murtaugh, L. D. Burns, and J. Schuster. Predicting the retention of university students. *Research in higher education*, 40(3):355–371, 1999.

[26] R. Neild and R. Balfanz. The dimensions and characteristics of philadelphia's dropout crisis, 2000-2005. *Philadelphia: Philadelphia Youth Transitions Collaborative. Retrieved January*, 1:2008, 2006.

[27] A. Ramchandani, C. Fan, and A. Mostafavi. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. *IEEE Access*, 8:159915–159930, 2020.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[29] R. W. Rumberger and S. A. Lim. Why students drop out of school: A review of 25 years of research. 2008.

[30] D. Sansone. Beyond early warning indicators: high school dropout and machine learning. *Oxford Bulletin of Economics and Statistics*, 81(2):456–485, 2019.

[31] W. Song, C. Shi, Z. Xiao, Z. Duan, Y. Xu, M. Zhang, and J. Tang. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1161–1170, 2019.

[32] A. Tamhane, S. Ikbal, B. Sengupta, M. Duggirala, and J. Appleton. Predicting student risks through longitudinal analysis. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1544–1552, 2014.

[33] R. Wang, B. Fu, G. Fu, and M. Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*, pages 1–7. 2017.

Appendix A. Extended Experiment Details

A.1. Feature List

See Table 2.

A.2. Label Definition

The binary label for our problem is 1 if the student did not graduate on time and 0 if the student did graduate on time. Failing to graduate on time can happen for a variety of reasons. Thus, to help explain the rationale behind how we defined the "graduation" label, we first categorize students into five disjoint classes that represent different scenarios that may occur: `on_time_grads`, `late_grads`, `dropout`, `high_gpa_grad`, and `low_gpa_no_grad`.

Firstly, the `on_time_grads` class represents students that we think graduated on time. We put a student in this class if the following three conditions hold:

- the number of calendar years spent in high school equals the number of grades the student completed
- the student was in grade 12 in the last year on record
- at least one of the following conditions holds:
  - the `withdrawal_reason` field explicitly confirms that a student withdrew due to graduation
  - the `status` field of the student says "graduate"
  - the student has a clear `graduation_date` available in grade 12

Secondly, `late_grads` represents students that we think graduated but not within 4 years. We designate students this class if they meet both the following conditions:

- a `graduation_date` is available which matches his/her last school year on record
- the number of calendar years the student spent in high school were more than the number of grades on record

Thirdly, the `dropout` class is self-explanatory and we place students in this class if their `withdrawal_reason` field indicated as such.

Finally, all remaining students are designated as either `low_gpa_no_grad` if their high school GPA is below 15$^{\text{th}}$ percentile (which is 2.25) or `high_gpa_grad` otherwise.

Given these five classes, we can define our label. We define a student has having not graduated on time (label = 1) if they belong to the `dropout`, `late_grads` or `low_gpa_no_grad` classes. We also define as a student has having graduated on time (label = 0) if they belong to the remaining classes: `on_time_grad` and `high_gpa_grad`.

A.3. Model Grid

Below, we list the configurations run for each model. The parameter names in the left column correspond to the sklearn parameter names for all models besides multi-layered perceptrons. Note that several models have two grids. This is because the purpose of the initial grid was to explore highly different spaces of the hyperparameter space to get a sense of the kind of values that work well for our model. After we got the understanding of the kind of values that are working well for our problem, we updated our hyerparameter grid to be more focused around values that work well, so we can get maximum performance out of our models. In simpler words, the purpose of the initial grid was to explore the hyperparameter space, and the purpose of the updated grid was to exploit the best regions of the hyperparameter space found earlier.

Logistic Regression:

Features from all_snapshots table	**Imputation**
number of absences per year since grade 6	0, missingness flag
number of unexcused absences per year since grade 6	0, missingness flag
gender	missingness as category
age	N/A (has 100% coverage)
ethnicity	missingness as category
school name	missingness as category
district	missingness as category
disability	missingness as category
disadvantagement	missingness as category
academic disadvantagement	false
economic disadvantagement	false
limited english (i.e. is English speaking skill limited?)	"N", missingness flag
zipcode	missingness as category
cumulative discipline incidents	0, missingness flag
discipline incidents for current year	0, missingness flag
discipline incidents for previous year	0, missingness flag
discipline incidents for 2 years ago	0, missingness flag
days absent for current year	0, missingness flag
days absent for previous year	0, missingness flag
days absent for 2 years ago	0, missingness flag
absence rate	0, missingness flag
unexcused absence rate	0, missingness flag
absence rate percentile	mean, missingness flag
number of school transfers since grade 6	1, missingness flag
Features from oaaogt table	**Imputation**
reading placement rating in grades 6, 7, 8	missingness as category
math placement rating in grades 6, 7, 9	missingness as category
writing placement rating in grades 6, 7, 10	missingness as category
6th grade ctz placement rating	missingness as category
6th grade science placement rating	missingness as category
8th grade reading placement rating	missingness as category
8th grade math placement rating	missingness as category
8th grade science placement rating	missingness as category
8th grade social studies placement rating	missingness as category
Features from high_school_gpa table	**Imputation**
number of classes taken in current year	mean, missingness flag
GPA for current year	mean, missingness flag
GPA for previous year	mean, missingness flag
GPA for 2 years ago	mean, missingness flag
GPA for 3 years ago	mean, missingness flag
GPA percentile for students in the same year	mean, missingness flag
GPA percentile for students in the same year and school	mean, missingness flag
GPA percentile for students in the same year and district	mean, missingness flag
Features from other provided tables	**Imputation**
interventions per year since grade 6 of each category	0, missingness flag
absences per year since grade 6 of each category	0, missingness flag

Table 2: Feature List

max_iters	[100, 500, 1000]
penalty	["elasticnet"]
l1_ratio	[0, 0.5, 1]
C	[0.001, 0.01, 0.1, 1]

Table 3: Logistic Regression Initial Grid

max_iters	[100,200, 300]
penalty	["elasticnet"]
l1_ratio	[0, 0.1, 0.2, 0.3, 0.4]
C	[0.01, 0.05, 0.1, 0.2, 0.3]

Table 4: Logistic Regression Updated Grid

Decision Tree:

max_depth	[3, 5, 10, 50, None]
min_samples_split	[2, 5, 10]
min_samples_life	[1, 2, 5, 10]

Table 5: Decision Tree Grid

Gradient-Boosted Decision Tree:

learning_rate	[0.0001]
n_estimators	[100, 500, 5000, 10000]
max_depth	[3, 30]

Table 6: Gradient-Boosted Decision Tree Grid

Random Forests:

n_estimators	[100, 500, 1000, 5000, 10000]
min_samples_leaf	[0.01, 0.1, 0.3, 1]
max_depth	[3, 10, 50, 100, None]
max_features	["sqrt"]

Table 7: Random Forests Initial Grid

n_estimators	[2500, 5000,7500,10000,12500]
min_samples_leaf	[0.2, 0.3, 0.4]
max_depth	[75, 100, 125]
max_features	[None (all), "log2", "sqrt"]

Table 8: Random Forests Updated Grid

Kernelized SVM:

kernel	['linear', 'poly', 'rbf', 'sigmoid']
C	[0.0001, 0.01, 0.1, 1]

Table 9: Kernelized SVM Initial Grid

kernel	['linear', 'poly']
C	[0.05, 0.1, 0.3, 0.5, 0.7, 1]

Table 10: Kernelized SVM Updated Grid

Multi-Layered Perceptrons: To make neural networks effectively converge and avoid overfitting, we used the technique of early stopping, and stopped training if the performance on the validation set did not increase for a long time.

Architectures	[[128, 64, 32], [128, 16], [128, 32], [64, 32, 16], [32, 16]]
Optimizers	[Adam, SGD]
L2 regularization	[0.001, 0.01, 0.1, 1]
Batch Size	[32, 64, 128, 256]
Learning Rate	[0.001, 0.01, 0.1]

Table 11: Multi-Layered Perceptron Grid. Each list for architectures corresponds to the number of neurons in each hidden layer.

A.4. Validation and Training Sets

We collected data for the following training and validation data splits. We show all validation sets we used in Figure 6.

A.5. Temporal Graph of Primary Evaluation Metric

Recall that the primary evaluation metric is precision at top 5% of highest risk students. We show the graph of precision at 5% with time in Figure 7.

A.6. Criteria Used For Top Models

As discussed in Section 7, we do not have any one criteria to select the "top" model. The "top" model is defined as per the needs of the policymaker, and our pipeline is flexible enough to support multiple different methods/criteria to pick the best model. In this report, we presented results of default three criteria that we chose for policymakers which are fully explained in Sections 6.5 and 6.6. These criteria evaluate models on precision at top 5%, fairness among ethnicities (by calculating the ratio of false negative rate in blacks to whites), fairness among school districts (by calculating the ratio of median false negative rate in school districts that collect little data to the ones that collect sufficient data). As said before, the exact details and formulae are presented in Sections 6.5 and 6.6.

A.7. Sample Results

In this section, we present our full results of the top 1-2 models in all three categories. To keep the presentation

	Train Set					Validation Set				
Train-Validation Pair ID	Start date for rows	End date for rows	Interval/Δ between dates for each entity across rows	Start date for labels	End date for labels	Start date for rows	End date for rows	Interval/Δ between dates for rows	Start date for labels	End date for labels
1	2007-05-31	2007-05-31	0 days	2007-05-31	2009-05-31	2009-05-31	2009-05-31	0 days	2009-05-31	2011-05-31
2	2007-05-31	2008-05-31	0 days	2007-05-31	2010-05-31	2010-05-31	2010-05-31	0 days	2010-05-31	2012-05-31
3	2007-05-31	2009-05-31	0 days	2007-05-31	2011-05-31	2011-05-31	2011-05-31	0 days	2011-05-31	2013-05-31
4	2007-05-31	2010-05-31	0 days	2007-05-31	2012-05-31	2012-05-31	2012-05-31	0 days	2012-05-31	2014-05-31
5	2007-05-31	2011-05-31	0 days	2007-05-31	2013-05-31	2013-05-31	2013-05-31	0 days	2013-05-31	2015-05-31
6	2007-05-31	2012-05-31	0 days	2007-05-31	2014-05-31	2014-05-31	2014-05-31	0 days	2014-05-31	2016-05-31

Figure 6: All Training-validation Pairs Used

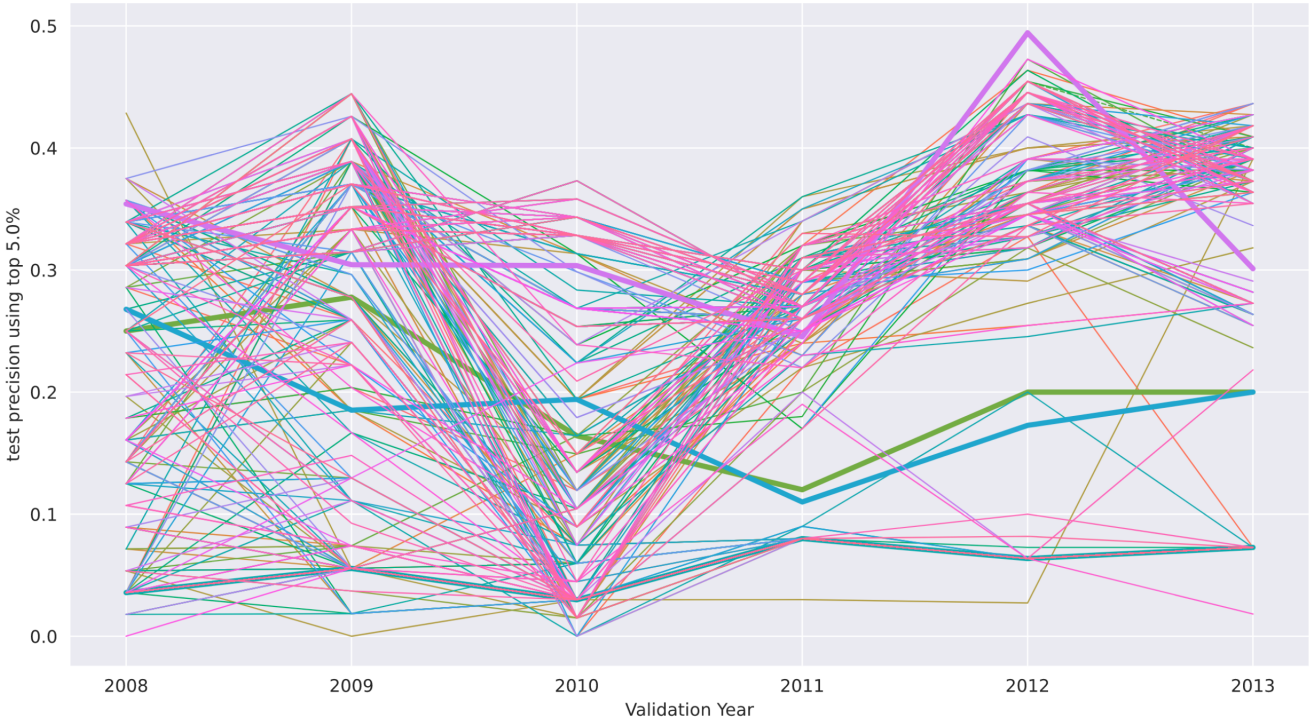


Figure 7: The graph of precision at 5% vs validation year. Each line corresponds to a different model, and all baselines are shown in thicker lines. So, the thick purple line at the very top of year 2012 is GPA baseline, the next green thick line is the absenteeism baseline, the next blue thick line is the discipline baseline, and finally the last blue thick line (overlaid by a pink line) is the random model which corresponds to the prior rate.

of results succinct and focused, we present results of cross-tabs, feature importances, and PRK curves as evaluated on the most recent validation set (of year 2013) because our models have most training data for the last validation set, and it also would be of most interest to policymakers due to its recency. All models results are shared below, and the model type and hyperparameter settings are shared in Table 12. Since we use over 350 features, we have not included

Criteria	Model Type	Hyperparameters
Most Stable Precision	Random Forest	n_estimators: 2500, max_depth: 100, max_features: log2, min_samples_leaf: 0.3
Second Most Stable Precision	Logistic Regression	penalty: elasticnet, l1_ratio: 0.3, inverse regularization strength: 0.01, max_iters: 100
Most Stable Fairness Over Districts	Neural Network	Architecture: [64, 32, 16], Batch size: 128, Optimizer: SGD, Learning Rate: 0.01, Weight Decay: 0.01
Second Most Stable Fairness Over Districts	Random Forest	n_estimators: 12500, max_depth: 75, max_features: None (all), min_samples_leaf: 0.3
Most Stable Fairness Over Ethnicities	SVM	Kernel: Poly, Inverse Regularization Strength (C): 0.05

Table 12: Hyperparameters of the models whose results are presented in Section A.7

feature importance results for all features in this document to save space. However, they are available for at https://tinyurl.com/y9x3xm7f.

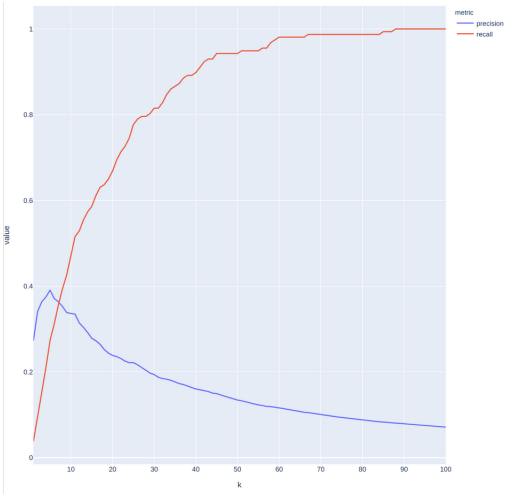
Most Stable Precision This is the model whose results are shown in Figure 5a already. So, we here only show the remaining graphs needed like the PRK curve, and cross tabs in Figure 8.

Second Most Stable Precision All results of this model are shown in 9.

Most Stable Fairness Over Districts This is the model whose results are shown in Figure 5c already. So, we here only show the remaining graphs needed like the PRK curve, and cross tabs in Figure 10.

Second Most Stable Fairness Over Districts All results of this model are shown in 11.

Most Stable Fairness Over School Ethnicities This is the model whose results are shown in Figure 5b already. So, we here only show the remaining graphs needed like the PRK curve, and cross tabs in Figure 12.

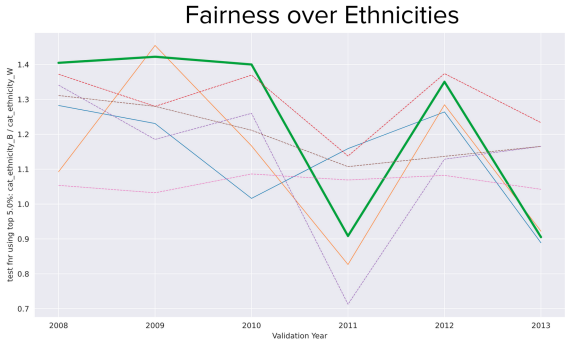


Feature	Mean Value for students with predicted risk of late graduation	Mean value for students without a risk
Current GPA	1.38	2.87
Last Year's Discipline Incidents	5.48	0.93
GPA Percentile Among All Students	0.88	0.46
GPA Percentile Among Districts	0.87	0.46
GPA Percentile Among Schools	0.87	0.46
Total Discipline Incidents	13.17	2.12
Current Year Discipline Incidents	4.92	0.7
Previous Year GPA	1.4	2.79
Number of Absences Per Year	15.1	8.04

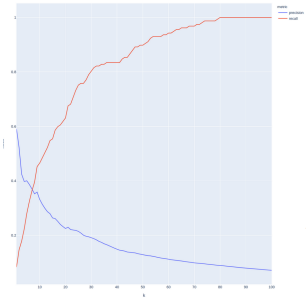
Figure 8: PRK curves and Cross Tabs of the model with most stable precision. The initial bump in PRK curve is just due to randomness, and we have checked that this is not the case for other validation sets (and we had provided evidence for the same in Project Update 6). Other plots for model are shown in Figure 5a.

Cross Tabs

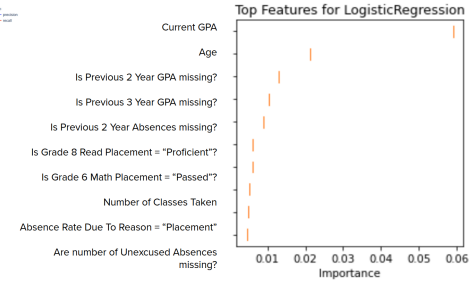
Feature	Mean Value for students with predicted risk of late graduation	Mean value for students without a risk
Current GPA	1.73	2.86
Age	16.44	15.55
Number of Unexcused Absences	2.14	0.81
Last Year's Discipline Incidents	4.57	0.98
GPA Percentile Among All Students	0.79	0.47
GPA Percentile Among Districts	0.78	0.47
GPA Percentile Among Schools	0.78	0.47
Number of Classes	31.82	38.39
Total Discipline Incidents	11.11	2.23



PRk Curve



Top Features



Fairness over Districts

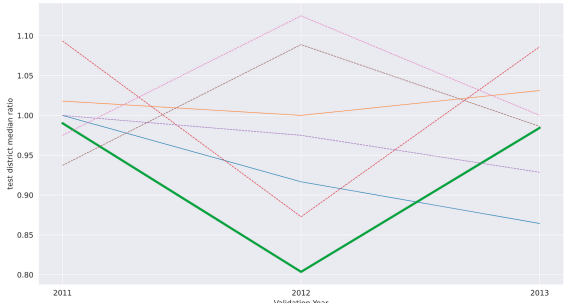
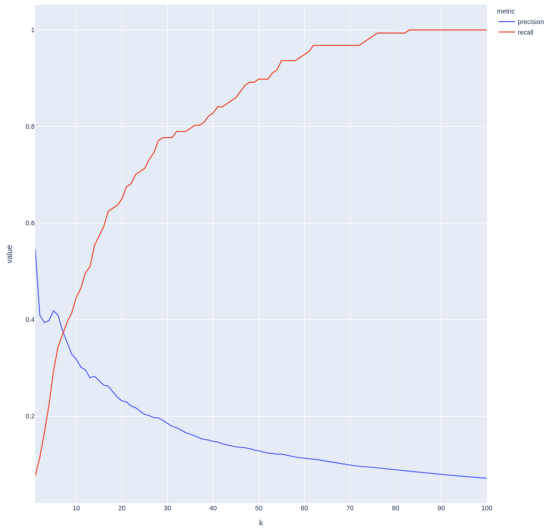


Figure 9: Cross Tabs, PRk curves, Feature Importances, and bias/fairness audit plots of the model with second most stable precision.

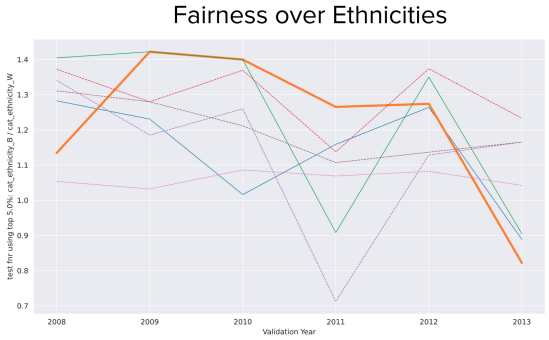


Feature	Mean Value for students with predicted risk of late graduation	Mean value for students without a risk
Last Year's Discipline Incidents	6.52	0.87
Current GPA	1.67	2.86
Total Discipline Incidents	14.2	2.07
Number of Unexcused Absences	2.13	0.81
GPA Percentile Among All Students	0.81	0.46
Number of Classes	30.41	38.46
GPA Percentile Among Schools	0.81	0.46
GPA Percentile Among Districts	0.81	0.46
Age	16.24	15.57

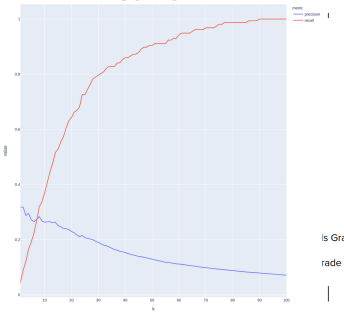
Figure 10: PRK curves and Cross Tabs of the model with most stable fairness over districts. Other plots for model are shown in Figure 5c.

Cross Tabs

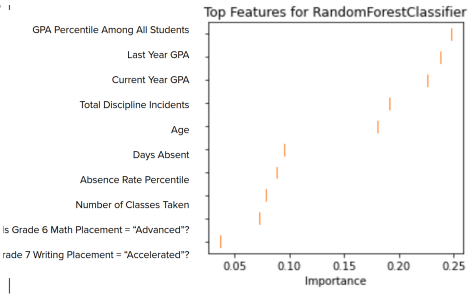
Feature	Mean Value for students with predicted risk of late graduation	Mean value for students without a risk
Current GPA	1.80	2.86
GPA Percentile Among All Students	0.79	0.47
GPA Percentile Among Schools	0.78	0.47
GPA Percentile Among Districts	0.77	0.47
Previous 1 Year Discipline Incidents	3.47	1.03
Last Year GPA	1.71	2.77
Economic Disadvantagement	0.82	0.58
Absence Rate Percentile	0.70	0.48
Number of Days Absent	14.33	8.12



PRk Curve



Top Features



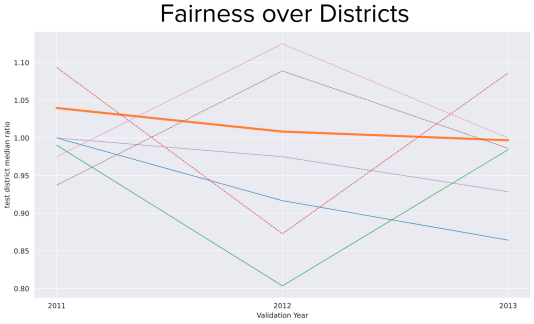
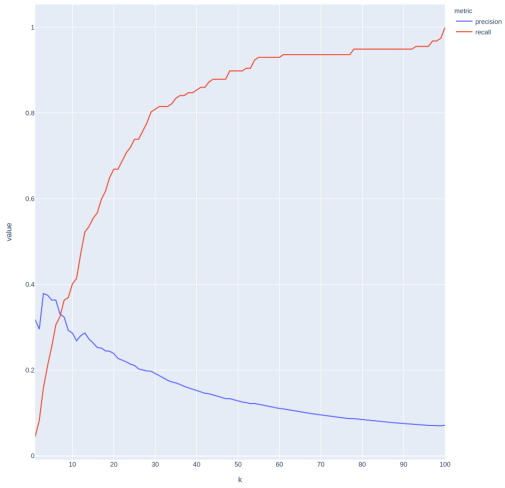


Figure 11: Cross Tabs, PRk curves, Feature Importances, and bias/fairness audit plots of the model with second most stable fairness over districts. Other plots for model are shown in Figure 5b.



Feature	Mean Value for students with predicted risk of late graduation	Mean value for students without a risk
Current GPA	1.44	2.87
Last Year's Discipline Incidents	5.78	0.91
GPA Percentile Among All Students	0.87	0.47
GPA Percentile Among Schools	0.86	0.46
GPA Percentile Among Districts	0.86	0.46
Number of Absences Per Year	17.73	7.91
Days Absent	19.24	7.86
Total Discipline Incidents	12.63	2.15
Number of absentees in last year	15.76	7.89

Figure 12: PRK curves and Cross Tabs of the model with most stable fairness over ethnicities. The initial bump in PRK curve is just due to randomness since the criteria used to select this model had nothing to do with precision.