Exploring Fairness in an Adaptive Learning Platform

Introduction

About DreamBox Math

DreamBox Math from Discovery Education is an adaptive K-8 math program, rated ESSA-Strong, and serves 5 million students across all 50 states, as well as in Canada and Mexico. Designed by educators, DreamBox Math aims to introduce, reinforce, and formatively assess foundational math concepts, supplementing classroom curricula. The program supports teachers and students by identifying and addressing gaps in student knowledge, thereby strengthening foundational skills and ensuring students are prepared for classroom learning.

DreamBox Math lessons are crafted to be engaging and enjoyable for students, emphasizing real-world applications and models. Students are empowered to choose their next lesson from a selection curated by DreamBox's proprietary adaptive engine. This adaptive engine is a core component of DreamBox, meeting students at their current level, recognizing each child's unique strengths and areas for development, and providing a personalized learning experience tailored to their individual needs.

DreamBox and AI

With the rapid advancements in Generative AI (GenAI) and increased public awareness around tools like ChatGPT, companies face significant pressure to integrate these technologies into their core products. However, at DreamBox, we remain focused on achieving optimal and equitable learning outcomes for all students. Guided by our core mission, we have adopted a conscientious approach as we explore moving beyond "AI 1.0" and begin integrating generative pre-trained transformers into our products.

A key concern in deploying these advanced technologies is the potential for inherent biases. Large language models (LLMs) can inadvertently perpetuate and amplify existing biases in their training data, manifesting as gender, racial, or cultural stereotypes. Biased

AI systems could further widen pre-existing educational disparities, making it crucial to ensure that AI-generated recommendations are equitable and unbiased.

Addressing these concerns aligns with national guidelines for AI in education, such as those outlined by the U.S. Department of Education's Office of Educational Technology (Artificial Intelligence and the Future of Teaching and Learning), which emphasize fairness, accountability, and transparency in AI systems. These guidelines advocate for rigorous testing and validation to ensure AI tools are inclusive and free of bias. By committing to these principles, our efforts to develop predictive and generative models for DreamBox Math aim not only to enhance learning outcomes but also to promote fairness and inclusivity. This ethical approach to AI deployment underscores the broader significance of our research, striving to create a more equitable and effective educational landscape for all students.

Research Overview

The purpose of this technical document is to aid the advancement of transparent and equitable AI-systems by sharing publicly some of our research into factors which influence student learning outcomes in DreamBox Math, including but not limited to the impact of student ethnicity.

Methodology

Data Collection and Preparation

For this study, we collected usage data from a single U.S. public school district accumulating 1.24M play events from 7,707 students spanning grades 3 through 6. A play event is defined as an individual instance of a lesson being played by a student at a specific time. This data was collected from October 2022 to June 2023, during the 2022-23 school year.

For the purposes of this research study and others, the district under investigation provided student ethnicity data. Each record in the dataset represents a play event, detailing the associated student's behavioral characteristics with product usage, various play-level metrics, and classroom features at that point in time. We note that there were several crucial diversity dimensions (such as free-lunch, ELL, household income, independent learning plans, etc…) that were unavailable in this study.

The dataset includes metrics such as product usage, class size, grade level, lesson difficulty, student pass rate, and ethnicity. These metrics were chosen to evaluate their influence on student success and to ensure a comprehensive analysis of the factors affecting learning outcomes in DreamBox Math.

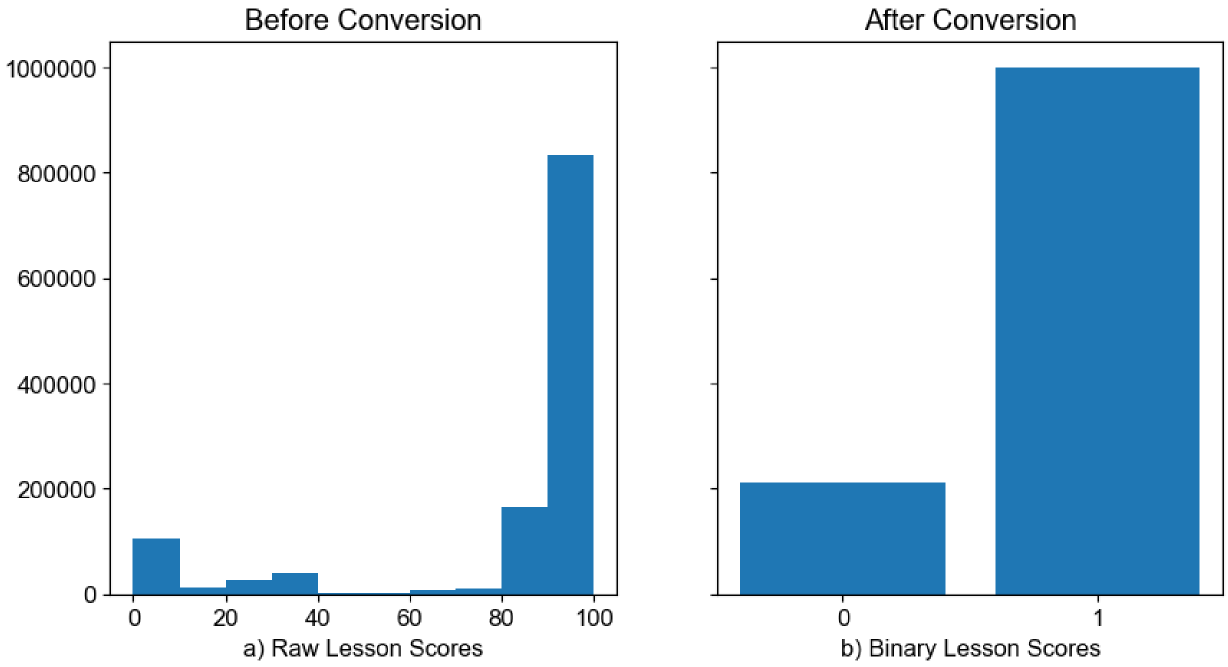


Fig 1. Target variable overview. a) Raw lesson scores from 1.24M play events spanning grades 3 through 6. b) The raw lesson scores are converted into a binary variable (lesson_score) used in this analysis. For the purposes of this analysis, raw lesson scores at or above 80 are considered a pass (lesson_score = 1).

The target variable in our dataset is the lesson score, which reflects the student's performance in a given lesson at a particular time. Originally a continuous variable spanning a range of 0 to 100, we converted the lesson score into a binary variable: 0 if the lesson score < 80, and 1 otherwise.

A common problem in developing predictive models is dealing with large class imbalances in the target variable. In this case, 83% of the play events included instances where the student passed. To handle this class imbalance of the raw data set, we have randomly under-sampled the pass play events (lesson_score = 1) resulting in a data set with approximately 244,954 play events from 7,296 students.

Below, we provide a detailed **data dictionary** describing the individual features (predictive factors) that might impact learning outcomes (target variable)

Features

Play-level features

- *student_passrate_90*: The ratio of successfully completed lesson play attempts to the total number of lesson play attempts by the student, calculated from 90 days prior to the current play attempt's timestamp up to the current play attempt. This metric only includes complete play attempts, excluding those left unfinished.
- *usage_frequency*: The average number of lessons played per week by the student from the start of the school year to the current lesson play event.
- *usage_consistency*: The standard deviation of the average number of lessons played per week by the student from the start of the school year to the current lesson play event.
- *usage_depth*: The total number of lesson plays by the student from the start of the school year to the current lesson play event.
- *usage_retention*: The total number of weeks in which the student engaged in lesson plays from the start of the school year to the current lesson play event.
- *play_seconds_of_duration*: The total time, in seconds, spent by the student to complete a particular lesson play event.
- *days_since_last_play*: The number of days since the student last engaged in a lesson play event.
- *random_num*: A randomly generated value ranging between 0 and 1. We include this for analysis of feature importance of the predictive model.
- *lesson_passrate_90*: This is the ratio of successfully completed attempts to the total number of attempts for a specific lesson, measured from 90 days prior to the current play attempt's timestamp up to the current play attempt.
- *season:* The season at which the play event occurred. One-hot-encoded into three variables:

  Fall: August to November, **Winter**: December to February, **Spring**: March to May

Student-level features

- *ethnicity*: The ethnic group to which the student belongs. This data was provided by the school district and is mutually exclusive categorical variable for each student. This data was one-hot-encoded into the five ethnicity groups provided by the district: Asian, Black, Hispanic, Indigenous, and White.
- *class_grade_numeric*: The grade level of the class that the student was part of during a play event. This categorical variable is mutually exclusive and was one-hot-encoded before feeding into the predictive model.
- *number_of_classes*: The total number of classes the student was enrolled in throughout the entire school year (July 2022 – June 2023).

Class-level features

- *students_per_class*: The number of students enrolled in the class that the student was part of during a lesson play event.

Target variable

- *lesson_score*: Originally a continuous variable ranging from 0 to 100, reflecting the performance by the student for a given play event. For the analysis, this target variable has been binarized it indicate pass/fail.

A note on leakage

To avoid leakage (a situation where the data within the feature variables contain information that would only be available after the outcome event), we ensured that all aggregated usage metrics were calculated before the play event. For example, for usage_frequency, the metric is calculated for each play event for that student from the start of the year up until, but excluding that play event. student_passrate_90 (the ratio of lessons passed to lessons played by a student) and lesson_passrate_90 (the ratio of passed attempts to total attempts for a specific lesson) were calculated using a trailing 90-day window from each lesson play event. For example, for a play event on March 3rd 2023, we calculate the student’s average pass rate over the past 90 days and encode that pass rate as the student_passrate for that specific play event.

Correlation Analysis

We begin the analysis by examining correlations between the individual features and the target variable (lesson score). A positive correlation is observed between lesson score and metrics such as student_passrate_90 (correlation coefficient = 0.43) and lesson_passrate_90 (correlation coefficient = 0.50), indicating that higher rates of student and lesson passes are linked to elevated lesson scores. Conversely, weaker correlations are observed with features like usage_consistency (correlation coefficient = -0.01) and usage_retention (correlation coefficient = 0.02), suggesting a more nuanced relationship between these metrics and lesson outcomes.

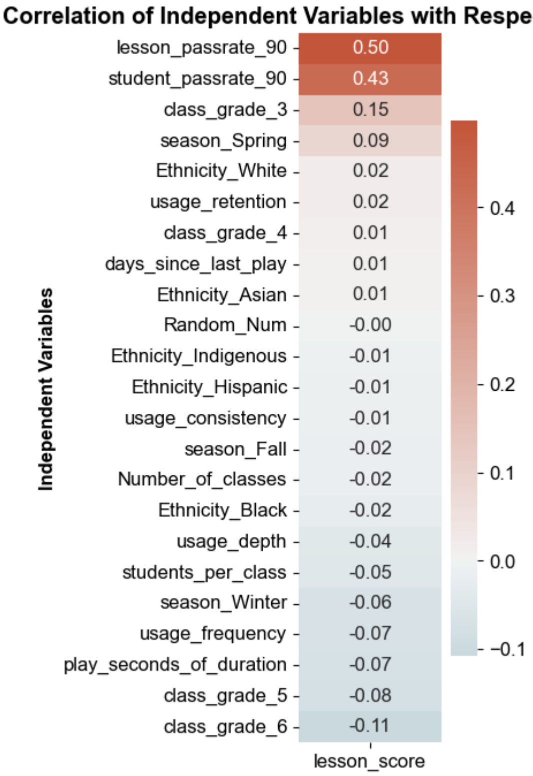


Fig 2. Correlation of all features over Lesson Score

Based on the correlation analysis between each feature and lesson_score, features can be categorized into distinct tiers based on their influence on predicting lesson scores through linear relationships.

In Tier 1 we have features like lesson_passrate_90 and student_passrate_90 which exhibit strong positive correlations with lesson_score, indicating a significant contribution to predicting lesson outcomes. Moderately correlated features, such as class_grade_3 and season_Spring, fall into Tier 2, contributing positively but to a lesser extent. Tier 3 comprises features with weak positive correlations, like usage_retention and Ethnicity_White, which have minor but discernible impacts on predicting lesson scores. Conversely, features in Tier 4 display weak negative correlations or no significant correlation with lesson_score, suggesting minimal influence or even counter productivity in a linear relationship model. Understanding the tiered influence of these features is pivotal for constructing effective predictive models for lesson scores, ensuring informed decision-making in educational settings.

Next, we turn our attention to examining correlations between features. For the first part of this predictive modeling problem, we're interested in trying to understand how features relate to each other and how they get stack ranked as "important" in making predictions. Most models cannot provide informative feature rankings when the features are correlated with each other. Below we discuss the inter-correlation between features among themselves -

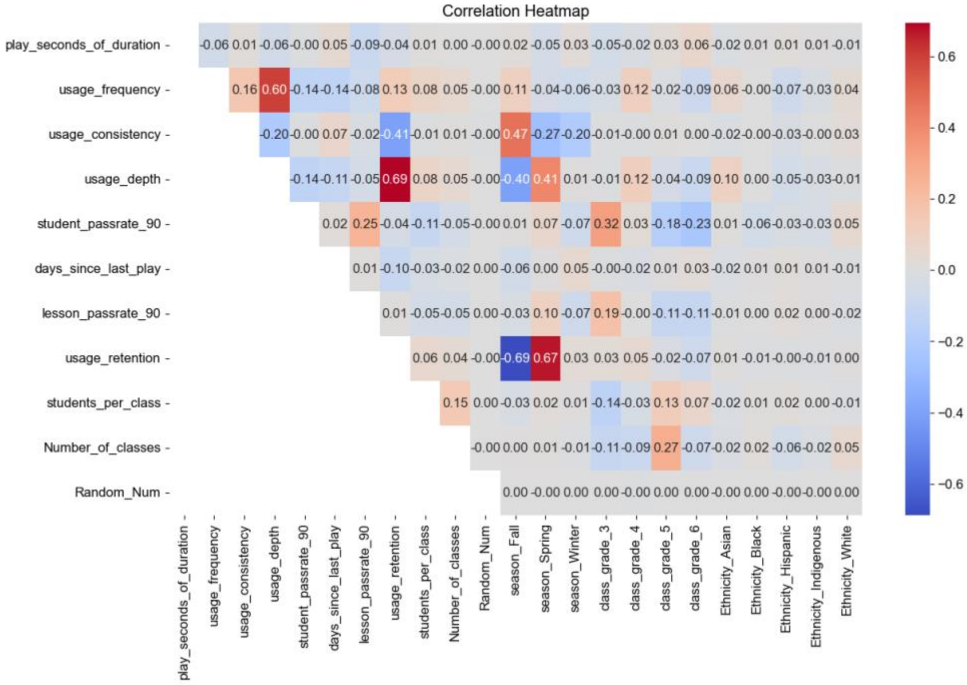


Fig 3. Correlation Heatmap among all features

From above figure we identify Inter-feature relationships with moderate correlations within the dataset. Notably, usage_frequency exhibits a positive association with usage_depth, with a correlation coefficient of 0.61, indicating a potential link between the frequency and depth of product usage. Conversely, a negative correlation is observed between usage_consistency and usage_retention, with a correlation coefficient of -0.40, suggesting a trade-off between usage consistency and retention over time.

Furthermore, *student_passrate_90* shows a correlation with lesson_passrate_90, implying that as students' passrate improves, they are more adept at handling more challenging lessons. Additionally, metrics such as usage_depth, usage_frequency, usage_consistency, and usage_retention are closely interrelated, with increases in one metric often corresponding to increases in others.

Moreover, there is a positive correlation between class_grade_3 and both *student_passrate_90* and lesson_passrate_90. However, this correlation gradually decreases as class grade level increases, indicating that students at higher grade levels may find it slightly more challenging to pass lessons due to the increased complexity of the curriculum across higher grades.

Analysis of Distributions: Towards Predictivity

The above correlation analysis highlights the complex relationships between predictor variables and the target variable, lesson_score. We'd like to get a bit better understanding beyond a simple correlation of the relationship between each variable and it's potential value as a predictive feature. Fig 4. shows plots of the normalized distributions (kernal density estimations) for a subset of potentially informative features. Each distribution is plotted separately for the pass/fail target events, potentially highlighting areas of key distinction between student learning outcomes.

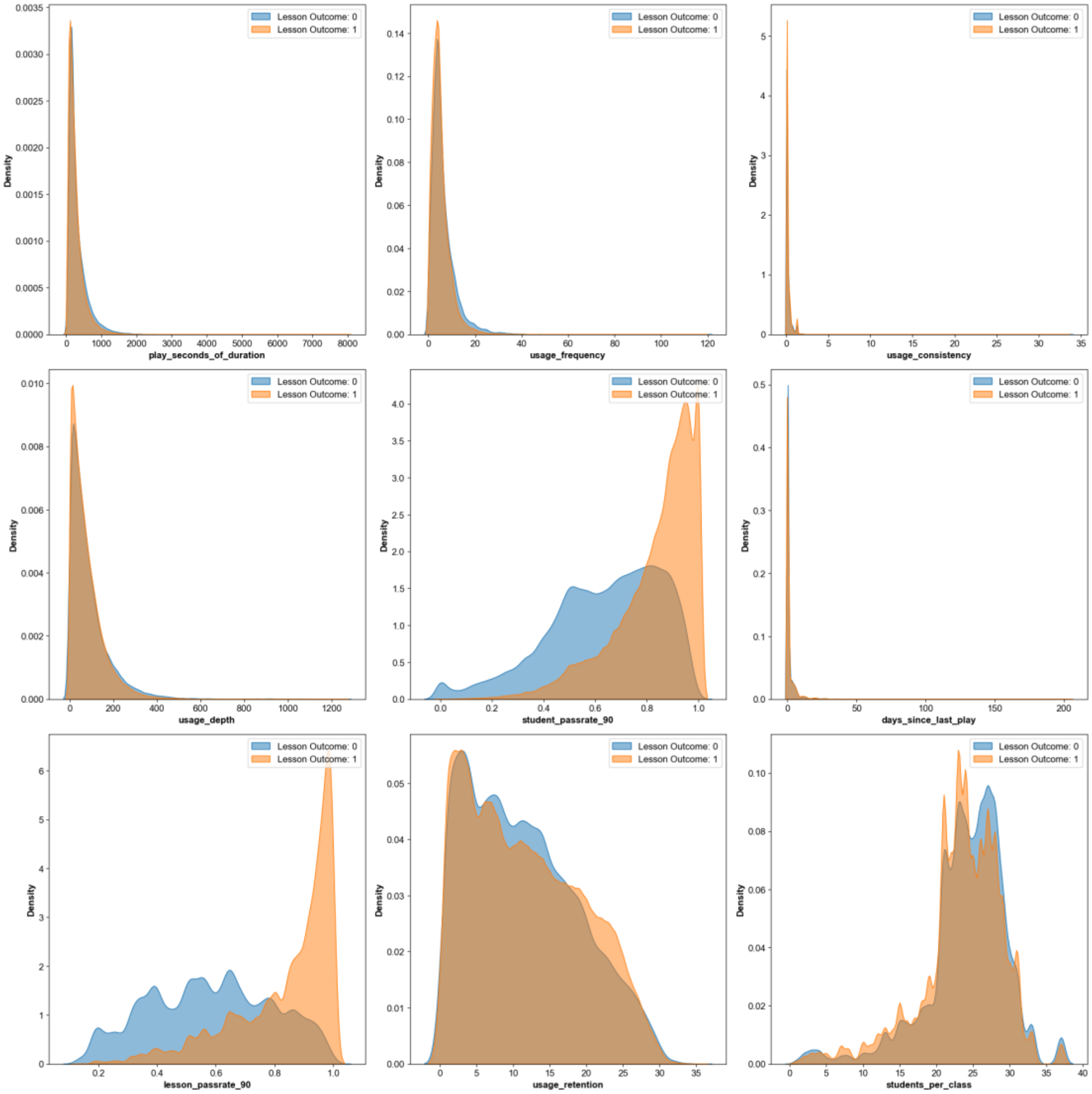


Fig 4. Split plot of Continous Features with respect to Lesson Outcome (Pass: 1, Fail: 0)

Fig. 4 shows that student_passrate_90 and lesson_passrate_90 exhibit distinct separation between lesson outcomes, indicating their strong predictive potential. In agreement with the correlation analysis, the data indicates that lesson_passrate_90 and student_passrate_90 are likely the two most predictive factors in this problem. This suggests that lessons which have historically been challenging to students are likely a leading indicator for subsequent student success. Similarly, the sizable splitting for student_passrate_90 seen in Figure 4e suggests that past student performance on lessons is another strong indicator of future success. On the other hand, and somewhat surprisingly, features such as days_since_last_play and usage_consistency show significant overlap in distributions between pass/fail, suggesting limited linear predictive power.

However, the distributions for usage_frequency, play_seconds_of_duration, and students_per_class reveal subtle distinctions that might be exploitable by more sophisticated non-linear models.

The long tail distribution of feature 'play_seconds_of_duration' indicates that students often pause a lesson after starting it, which may cause them to struggle with the remaining questions when they return. This can result in a failure to complete the lesson. Other factors, such as interruptions during the lesson, can also hinder student progress, leading to higher failure rates.

Model Selection

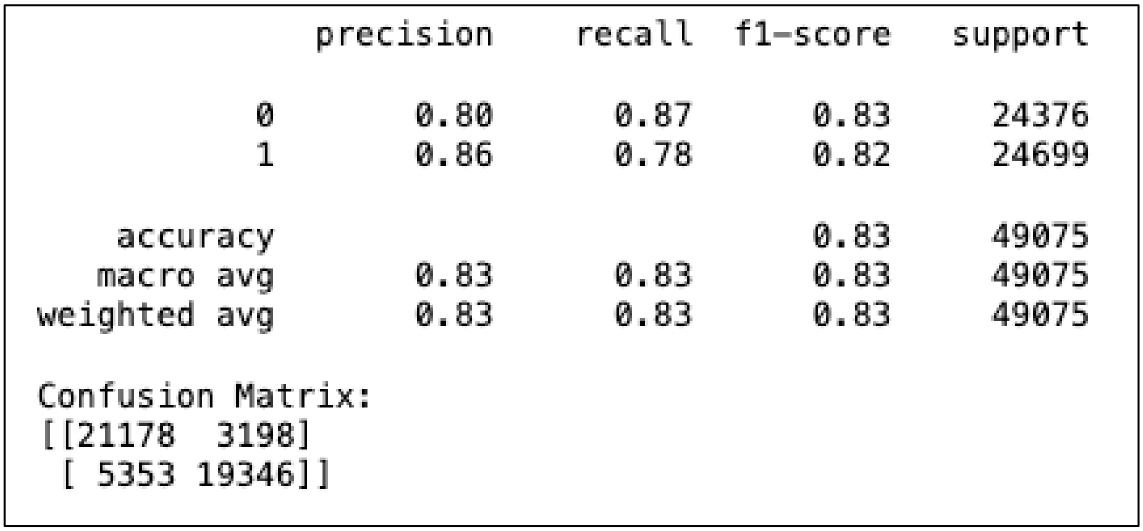
The above analysis highlights the significance of non-linear and complex relationships within the dataset. These interactions, often missed in linear evaluations, can provide substantial predictive power. We have structured the data as a moment-in-time binary classification problem and chosen to use a Random Forest Classifier (RFC) for the subsequent analysis. RFCs are particularly well-suited for this task due to their ability to handle mixed data types and capture non-linear relationships, making them ideal for predicting learning outcomes. By constructing multiple decision trees and combining their predictions, RFCs effectively model intricate decision boundaries within continuous variables. The feature importance metrics help identify influential variables, refining feature selection and enhancing model interpretability. Additionally, RFCs can accommodate heterogeneous data types without extensive preprocessing, further underscoring their utility in creating accurate and interpretable predictive models. Most importantly, Random Forest models are highly interpretable, as it is straightforward to trace how individual features contribute to the final prediction.

The data is randomly split 80/20 into a training set and an evaluation set. The RFC is then applied to the training set, and we evaluate the trained model’s performance on the evaluation set.

Results

Model Evaluation

We first examine the overall performance of the trained RFC model on the evaluation data set by looking at the confusion matrix and classification report. The confusion matrix provides a breakdown of predicted versus actual outcomes. For lesson_score = 0, the classifier correctly predicts 21,172 instances as failures (true negatives), but misclassifies 3,204 instances as successes (false positives). Similarly, for class 1, the classifier accurately predicts 19,318 instances as successes (true positives), but misses 5,381 instances, incorrectly classifying them as failures (false negatives).



```
              precision    recall  f1-score   support

           0       0.80      0.87      0.83     24376
           1       0.86      0.78      0.82     24699

    accuracy                           0.83     49075
   macro avg       0.83      0.83      0.83     49075
weighted avg       0.83      0.83      0.83     49075

Confusion Matrix:
[[21178  3198]
 [ 5353 19346]]
```

Table 1. Classification report for the trained Random Forest Classifier applied to the evaluation data set.

Next, we examine the model's accuracy, precision, and recall. The trained RFC exhibits an accuracy of 83%, indicating that it correctly predicts the outcome for 83% of the instances in the evaluation dataset. Upon closer examination of the precision and recall metrics, we observe that for lesson_score = 0 (indicating a lesson was not passed), the classifier achieves a precision of 80% and a recall of 87%. This indicates that out of all instances predicted as lesson_score = 0, 80% are indeed true negatives, while the classifier successfully identifies 87% of the actual negative instances. Conversely, for lesson_score = 1 (indicating passing a lesson), the precision is 86%, indicating that out of all instances predicted as lesson_score = 1, 86% are true positives. However, the recall for lesson_score = 1 is slightly lower at 78%, implying that the classifier misses approximately 22% of the actual positive instances.

We note here that for the purposes of this analysis the overall accuracy score of the model is not a quantity of interest. We are not attempting to build the "most predictive" model, rather we are interested in model interpretability and obtaining a relative stack ranking of leading indicators. For that reason, we have included a completely random number as a feature in the data set against which we can compare other features. This inclusion inherently reduces the accuracy of a trained model. Regardless, we can extract a few insights from the classifier performance. 1. The model accuracy could be improved further, likely indicating that there are other demographic or behavioral features missing from the training data which might be informative. 2. The fit model fairly balances the tradeoff between precision and recall, with both macro values at 83%. When precision and recall are balanced, the predictive features identified in a feature importance analysis are likely to contribute to fairly between accurately detecting relevant success instances (true positives) and minimizing incorrect success predictions (false positives). This means that the features deemed important are consistently valuable across different aspects of the classification task. If the model had instead been trained and optimized to, for example, maximize precision, then this favoritism towards certain kinds of predicted success would be similarly reflected in the feature importance analysis.

Evaluation via Feature Importance

We examined feature importance using three techniques: model feature importance (Gini impurity), permutation importance, and the **SH**apley **A**dditive ex**P**lanations (SHAP) method. The RFC's model feature importance measures the average decrease in Gini impurity caused by a feature across all trees in the forest, indicating its contribution to the model. Permutation importance assesses the importance of a feature by randomly shuffling its values and measuring the decrease in model performance, thus showing how crucial the feature is for accurate predictions. The SHAP method, based on cooperative game theory, explains the model output by indicating the contribution of each feature to individual predictions relative to the average prediction. The results of all three methods were generally consistent. In the following discussion, we focus our attention on the results from the SHAP feature importance.

Fig 5. shows the SHAP feature importances for the evaluation data set. The SHAP values represent the impact of each feature on the model output and can be used to identify which features contribute the most information to the model. The features are stack ranked from top to bottom according to their overall impact in influencing predictions (feature importance). The individual dots represent individual predictions, and the color of each dot represents the feature value (red: high; blue: low). The x-axis of a SHAP visualization represents the SHAP value, which indicates the impact of each feature on the model's output for a particular prediction. Positive SHAP values indicate that the feature pushes the prediction towards a higher value, while negative SHAP values indicate that the feature pushes the prediction towards a lower value. It is important to note that SHAP values are not constrained to a binary interpretation of success (1) or failure (0), but rather they show how much each feature contributes to the deviation from the average prediction.

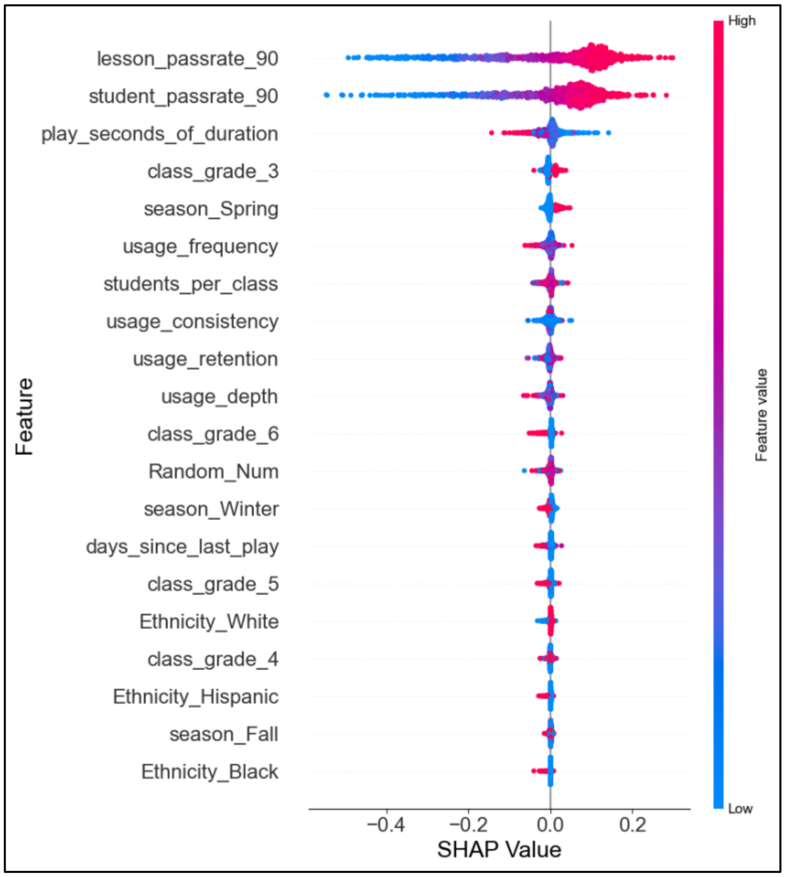


Fig 5. SHAP Feature Importance

Based of above SHAP plot, we can infer following primary influences in determining lesson outcome for each student. A low ***lesson_passrate_90*** value pushes the prediction strongly towards failure (lesson_score = 0) and a high ***student_passrate_90*** value pushes the prediction towards success (lesson_score = 1), but with a slightly weaker contribution to the overall prediction than *lesson_passrate_90*.

We note that the correlation analysis, analysis of the raw distributions, and feature importance have each indicated that these two variables are the leading indicators for predicting whether or not a student will pass a lesson at any given moment in time. We interpret this result as follows. At any given moment in time, the two most predictive factors for whether or not that student will pass the lesson presented to them is determined by their past history of success with prior DreamBox Math lessons and the passrate of the lesson presented to them. While other factors do contribute to success above random chance, they contribute to the prediction with weaker weights than those two primary factors.

Next, we would like to briefly discuss a few other notable results from the SHAP analysis. First, the first usage-specific metric that bubbles to the top is play_seconds_of_duration, indicating that the time that a student spends on the lesson is highly predictive. The next highest usage feature is usage_frequency, which suggests that students who have been using DreamBox math lessons more often throughout the school year do perform better on the lessons. **Lastly, the model has weighted all of the ethnicity features to contribute to predictions of lesson success at a level at or below random chance.** We will discuss this point more later.

Feature Reduction

Given the dominance of the two primary features, and their conceptual limitation in helping us answer the issue of equitable outcomes in student success (e.g. students who have demonstrated success (over the trailing 90 days) continue to show success, and hard lessons tend to be challenging for students), we continue our analysis by removing these two features and retraining the RFC on the reduced data set. This provides a more nuanced understanding of the influence of product usage factors and the impact of ethnicity. The resulting model had an accuracy of 66%, macro precision of 66%, and macro recall of 68%, indicating a worsened performance in the absence of those two key predictive features. Similar to the preceding analysis on feature importance, we found that the SHAP, permutation, and Gini approaches to feature importance generally showed similar results. We next discuss the feature importance results obtained using the SHAP method (see Fig. 6).

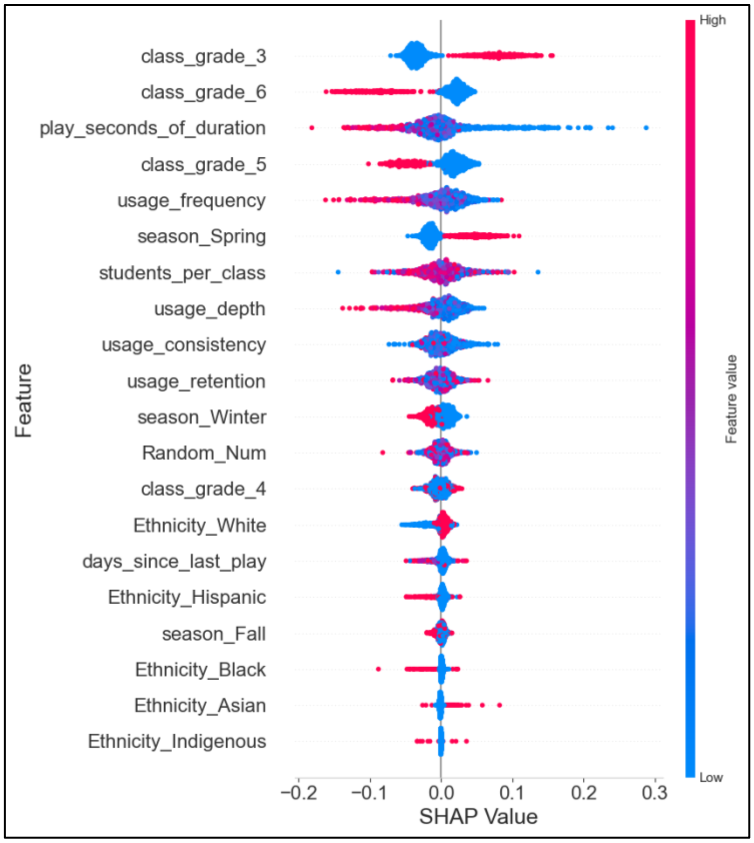


Fig. 6. SHAP Plot Following Feature Reduction: Model analysis with Lesson Difficulty and Student Proficiency removed.

Fig. 6 shows that a number of features contribute to the model's prediction with comparable strength. No single feature or subset of features completely determines the model's outputs. We notice immediately that several grade level features bubble to the top, particularly grade 3, 5, and 6. These grades also showed moderate correlations with the lesson pass rate feature (see Fig. 3), suggesting that the dominance of these variables are most likely stemming from the nature of the typical pass rates of these lessons in these grades. We note here that "more difficult" or "easier" lessons by grade level is not an absolute statement of difficulty but rather that the model has noticed and parsed out typical performance variabilities in the lessons presented *on average* in these grade levels. Since DreamBox Math provides adaptive learning pathways, students in these grades still progress and show learning growth even with slight variations in overall pass rate by grade.

This analysis also shows that the top in-product usage feature is the play_seconds_of_duration, or the time that students actually take to complete the lesson. The negative influence of play_seconds_of_duration highlights the impact of frequent interruptions and pauses during lessons on student performance, which often leads to higher failure rates. We also find that top trailing usage measure that predicts success is usage_frequency. This result indicates that the average number of lessons played per week is a strong determiner of student success in DreamBox Math. We find the remaining usage-based trailing indicators (usage_consistency, usage_retention, usage_depth, and days_since_last_play). It was surprising to us to learn that the days since last play and usage consistency were not strong factors. We also note here that usage_depth is correlated with usage_frequency as these are conceptually similar measures. Thus, the model has selected usage_frequency as the better of the two definitions for model predictions. A model oriented around accuracy would likely benefit from collapsing these two features into one variable using a dimensional reduction technique such as PCA.

This analysis also shows that the algorithm again finds that student ethnicity contributed to prediction of student success at a scale below random chance. We can fairly conclude at this point that ethnicity is not a strong factor determining these moment-in-time predictions for whether or not an individual student will pass a given lesson in DreamBox Math. Although usage patterns for this school district generally mirror those found across the whole of student usage within DreamBox Math, we stress that this conclusion is strictly only valid for the data and school district under investigation in this study.

Comments on Seasonality

Seasonality is a critical variable underlying usage patterns in Education. Back-to-school, winter break, Mother’s day, 100th day of school, assessment season are all typical markers of changes in usage patterns throughout EdTech products, particularly in grades 3-6. Below, we plot in Fig 7a, the weekly average pass rate for the full data set used for this analysis.

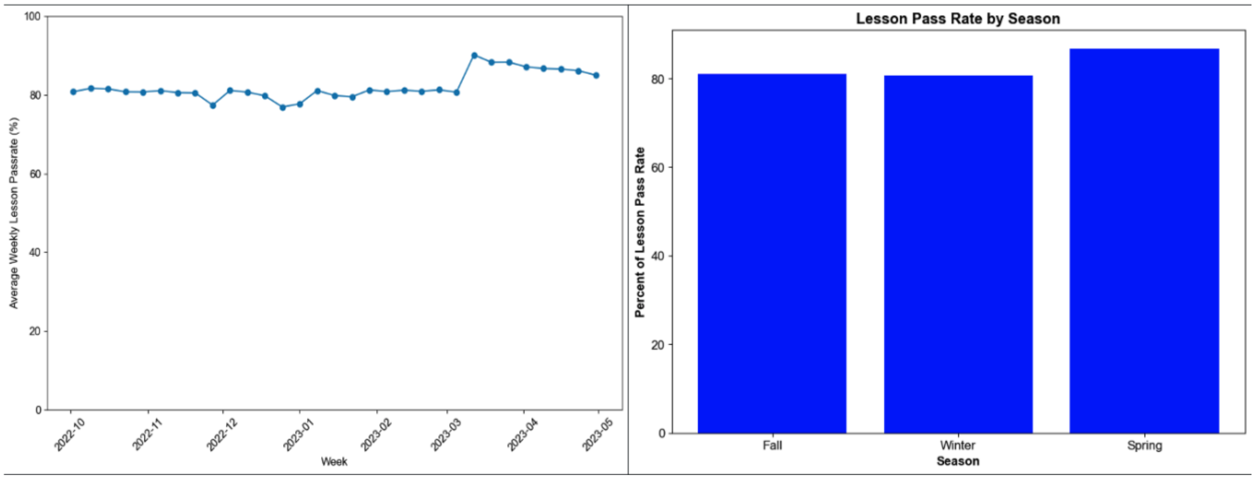


Fig 7a. Average Weekly Pass Rate & 7b. Seasonality Analysis

Figure 7a presents the average weekly lesson pass rates from October 2022 to May 2023. The data remains relatively stable initially, with a minor dip observed in early December, likely due to school breaks and holidays. This is followed by a sharp recovery in January 2023 as students resume their regular schedules. Subsequent weeks demonstrate a general upward trend in pass rates, with a slight decline beginning in mid-March, suggesting dynamic fluctuations in student performance potentially linked to academic calendar events.

Figure 7b illustrates the distribution of lesson pass rates across the Fall, Winter, and Spring seasons. The data indicates a consistent pattern across all seasons, with approximately 80% of lessons resulting in a pass, reflecting stable performance throughout the academic year. Notably, there is a marked improvement in pass rates from Winter to Spring, possibly due to factors such as increased familiarity with the curriculum, improved teaching methods, and heightened student motivation as the academic year progresses.

Disparate Impact: An Alternate Measure of Student Equity

Introduction to Disparate Impact

In the above analysis, we evaluated the performance of a machine learning model to gain insights into the role that various factors (including student ethnicity) play in determining the moment-in-time likelihood that an individual student will pass a given DreamBox Math lesson. This is certainly one way to explore the problem, but acknowledge that such an approach is akin to looking through at a problem through a single lens. We wish to develop multiple measures to understand our effectiveness in providing equitable outcomes for student learning.

Towards that goal, we now introduce the concept of “disparate impact.” Disparate impact is a metric designed to assess fairness by providing a measure of favorable outcomes across diverse demographic groups. The ideal for equitable outcomes is found when the disparate impact ratio, a value ranging from 0 to 1, is equal to 1 across all groups.

$$\boxed{DI = \frac{P(\hat{Y} = 1 | A = \text{Group\_A})}{P(\hat{Y} = 1 | A = \text{Group\_B})}}$$

Fig 8. Formula to calculate Disparate Impact (DI)

To calculate Disparate Impact, we assess the pass rates of DreamBox Math lessons among two ethnic groups, Group A and Group B. We calculate the pass rate for each group by finding the ratio of passed lessons to total students as shown in Fig 8 where, A is the Student sub-group and P(Y=1) is the Probability of positive outcome for the sub-group. We then compare these ratios to identify any disparities. Finally, we evaluate the extent of these disparities among different groups using these values, where a value close to 1 indicates fair outcomes. Significant deviations from 1 suggest biases that require further investigation to ensure equitable educational opportunities.

Benchmarking Disparate Impact in DreamBox Math

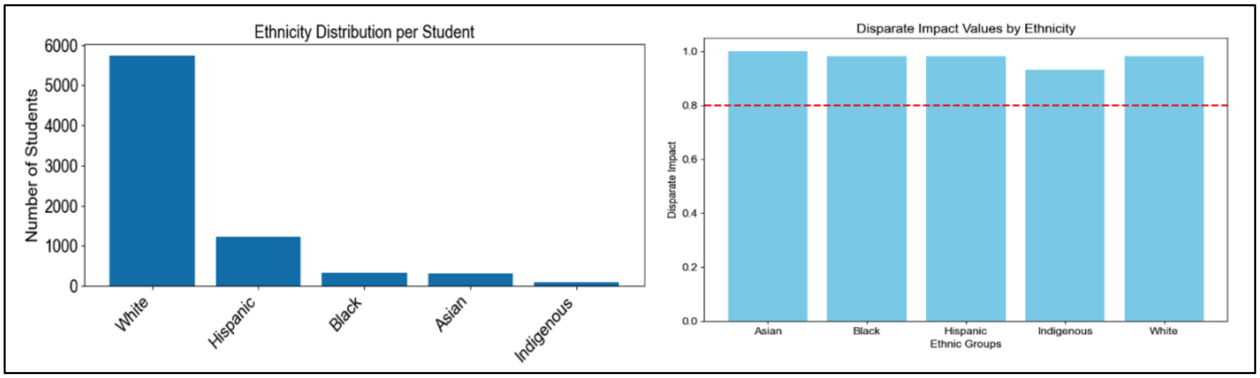


Fig 9a. Breakdown of Students by Demography & 9b.
Disparate impact factor for each Demographic group

An analysis of disparate impact in DreamBox Math. a) Bar chart showing the breakdown of students by each demographic group. In the data provided, 74% of students are White, 16% are Hispanic, 4% are black and asian each and 1% Indigenous. The Disparate Impact analysis assesses the ratio of favorable outcomes across various ethnic groups within the context of student success in DreamBox Math. The ratios obtained for each group are as follows: Asian (1.00), Black (0.98), Hispanic (0.98), Indigenous (0.93), and White (0.98). **The analysis reveals a consistent proportion of positive outcomes across all ethnic groups, indicating a relatively uniform distribution of success rates among students.** These findings, in conjunction with the above analysis of feature importance, suggest a minimal level of bias or unfairness in the outcomes across different ethnic groups within the DreamBox Math environment for the specific School District. We again stress that this analysis is conducted on a usage data from a single school district, and are not definitively generalizable to the entire student population. Further exploration and analysis on a broader scale would be necessary to validate these observations and ensure equitable educational opportunities for all students using DreamBox Math.

The 80% Rule

This rule states that a selection rate for any racial, ethnic, or gender group that is less than 80% of the rate for the group with the highest rate signals potential disparate impact (unfairness). The 80% rule, serves as a crucial benchmark for assessing disparate impact, particularly in employment practices, but its principles extend to other domains such as education and algorithmic fairness. In our case, the 80% rule falls, coincidentally, at 0.8 as the highest disparate impact score is 1.0. All demographics in this study displayed a disparate impact at or above 0.93, significantly above the 80% rule, demonstrating that engagement with DreamBox Math lessons have shown a fair balance of outcomes amongst the sub groups.

Applying the 80% rule in the context of AI in education is essential to ensure fair treatment across diverse student populations. As AI becomes increasingly integrated into educational tools, adherence to this guideline helps identify and mitigate biases, fostering equitable outcomes. This aligns with national guidelines on ethical AI deployment, such as those outlined by the U.S. Department of Education’s Office of Educational Technology. Their report, "Artificial Intelligence and the Future of Teaching and Learning," underscores the importance of transparency, inclusivity, and fairness in educational AI applications, reinforcing the necessity of rigorous bias evaluation methods like the 80% rule to promote equitable educational opportunities for all students. For further details on the 80% rule and its application, refer to the EEOC Uniform Guidelines (EEOC).

Conclusions and Outlook

In this study, we explored factors which might predict student learning outcomes (passing or failing a particular lesson) in DreamBox Math at any moment-in-time during the 2022/23 school year. We partnered with a U.S. public school district who provided student ethnicity data. While we did not have access to other important student characteristics (e.g. free lunch status), we were able to compare the importance of student ethnicity against other factors, such as usage and lesson difficulty in determining the likelihood of a student passing or failing a DreamBox Math Lesson. We prepared the data as a binary classification problem, carefully structured the features to reflect conceptual usage patterns, mitigated leakage, and applied a random forest classifier to the data. In evaluating the model, we found that the model (and variations of different models) never weighted student ethnicity above the level of random chance when making predictions about student learning outcomes. We also introduced a new measure, Disparate Impact, which has more commonly been used to analyze fairness practices in the housing market, as an alternative benchmark for interpreting fairness in AI-based learning platforms.

As part of our partnership with the Bill & Melinda Gates Foundation, we set out to proactively participate in contributing to responsible AI in Education and EdTech products. Our first step towards those goals has been to carefully evaluate our existing adaptive engine as it makes moment-in-time adaptive lesson recommendations to students. Since its release to students over a decade ago, we have been monitoring this system internally, and have continued to evaluate the effectiveness of the adaptive system as we have updated and improved the content and engine over time. As the world moves further and further towards GenAI systems which are increasingly black box and trained on questionable and unknowable data sources, it is more important than ever to provide as transparent a view as possible into the nature of the systems which are actively teaching children at immense scale (> 5M students).