Understanding Predictive Models of Student Success with a Multiverse Analysis

Yunxuan Tang
University of Michigan
yunxuant@umich.edu

Emma Harvey
Cornell University
evh29@cornell.edu

Chengyuan Yao
Columbia University
cy2706@tc.columbia.edu

Renzhe Yu
Columbia University
renzheyu@tc.columbia.edu

Rene F. Kizilcec
Cornell University
kizilcec@cornell.edu

Christopher Brooks
University of Michigan
brooksch@umich.edu

ABSTRACT

Predictive models of student success can provide timely information to inform interventions in K-12 and higher education. However, the design and implementation of these predictive models require various stakeholders to make decisions about the prediction target, data sources, processing, training, models, and deployment strategies. These choices are often poorly documented in the scholarly literature, even when code is openly available, limiting our ability to generalize and translate research findings to other institutions or contexts. More importantly, it obfuscates the potential trade-offs of decisions that are made with respect to prediction performance and other objectives, such as group fairness criteria. To address these challenges, we advocate for a multiverse approach in student success modeling and demonstrate the approach using a case study. In the multiverse framework, each plausible choice made to refine the problem space results in separate analyses being completed (each being referred to as a "universe"), with the final result being the collection of all universes explored. We demonstrate the mechanics and merits of this approach by building a first-year retention model for higher education. We interpret the findings of this analysis, specifically considering both model goodness-of-fit and fairness by group, demonstrating the value of the multiverse technique in engaging education-specific stakeholders—from administrative supervisors to model developers—in making predictive models that are robust, reproducible, and equitable.

Keywords

Multiverse Analysis, Machine Learning, Educational Technology, Predicting Student Success

1. INTRODUCTION

Although cutting-edge machine learning (ML) models have been developed for student success prediction with increasing performance in different decision-making scenarios [5, 44, 29], the complexity and lack of transparency in the technical pipelines hinder their replicability and generalizability across different contexts [35]. Stakeholders involved in the process of model development and deployment encounter numerous choices from prediction target selection, feature inclusion, handling of missing values, model selection, and more, which can significantly influence the predictive performance and also group fairness of the final model [27, 19]. In most studies, researchers make numerous decisions regarding data processing and model settings but only present their findings based on the final chosen configuration. This practice raises concerns regarding the transparency and robustness of study findings [39, 20, 10, 45]. Additionally, it prompts questions about which decisions or choices impact the results significantly [34, 39, 11].

With a systematic way to address these decisions macroscopically, researchers can better understand the effect of decisions on model performance and fairness and reflect the intrinsic correlations (inherent relationships among various factors that affect a model's performance and fairness) of these choices to confidently develop powerful and equitable ML models [37]. In this work, we apply the multiverse analysis, a method that systematically explores a range of plausible analytical decisions, to predictive models of student's re-enrollment [39]. This approach enables us to quantify the impact of various administrative, data processing, and model tuning decisions on both model performance and fairness metrics. By generating and evaluating thousands of model specifications, we provide insights into the trade-offs between performance and fairness associated with different decision pathways. Our main contributions are:

1. Demonstrating the application of the multiverse analysis technique in educational ML models through a case study
2. Evaluating the effect of decisions on the performance of models through multiverse analysis
3. Assessing the effect of decisions on model fairness using multiverse analysis

This approach enables a comprehensive understanding of how various decisions by distinct stakeholders influence prediction outcomes, facilitating the development of more robust, equitable, and reproducible ML models in education.

Yunxuan Tang, Emma Harvey, Chengyuan Yao, Renzhe Yu, Rene Kizilcec, and Christopher Brooks. Understanding Predictive Models of Student Success with a Multiverse Analysis. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 518–525. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870276

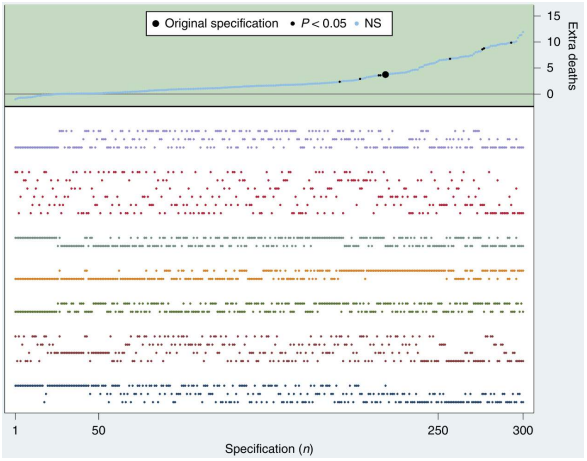

Figure 1: Specification Curve Analysis of Different Multiverses (dots) for Number of Extra Deaths by Hurricanes vs. Unique Specification of Decisions, from Simonsohn et al. [37]

2. MULTIVERSE ANALYSIS TECHNIQUE

2.1 Overview

Researchers often encounter a multitude of choices in data processing, experimental design, metric choices, and analysis methods and as a result, there are many decisions to make for a transparent and robust research [39, 20, 45, 10]. This issue of choice has led to the rise of the *multiverse analysis* technique. In multiverse analysis, researchers make every plausible combination of dataset decision, data processing, model specification, etc. Each combination of decisions represents a possible *universe* of outcomes. The researchers can then analyze results aggregated across these unique *universes* of research [8], increasing the transparency of the research and by identifying key choices that result in the conclusions [39].

The most well-known application of the multiverse analysis comes from Simonsohn et al. [37] who were following up on a report by Jung and colleagues [21] on the relationship between the names of hurricanes and their level of casualty. In the original work by Jung and colleagues, six decades of death rates from US hurricanes show that those hurricanes that were feminine-named caused significantly more deaths than masculine-named hurricanes, an effect they attributed to the public not perceiving feminine-named storms as dangerous as those which were masculine-named. Simonsohn et al. [37], however, reconsidered the research through the lens of over 1,700 unique analytical decisions including those dealing with "which storms to choose" and "the type of regression models" [37]. From this set of choices, they selected 300 specifications – individual groups of parameter value that were plausible given the research question at hand – and plotted the number of extra deaths versus each specification, where each dot on the curve depicted the estimated additional fatality of a hurricane with a feminine name rather than a masculine name (figure 1). The curve showed that the vast majority of specifications had an estimated additional deaths larger than 0, providing additional evidence that hurricanes with feminine names are deadlier [37]. This work demonstrated that, while there are reasonable choices a researcher might make that could show masculine-named hurricanes result in more casualties, there was more evidence that supported the conclusions of Jung et al. when the cases were considered more systematically.

Researchers have also applied multiverse analysis in various other fields including psycholinguistic research, vocabulary research, and corpus research [26]. For instance, in psycholinguistic research, Maie et al. [26] created 162 distinct universes by varying processing steps to reveal that evidence for implicit knowledge development was sensitive to the data-processing choices. In vocabulary research [26], they applied multiverse analysis to generate 54 universes based on decisions such as "how to define outliers" and "how to treat pretest scores in modeling", and their findings suggested that different combinations of choices could lead to varying conclusions about the effectiveness of repeated exposure on vocabulary learning. In corpus research (study of language through large collections of authentic text), Maie et al. [26] applied multiverse analysis to the dataset used originally in Eguchi et al. [12]'s study, by exploring different analytical choices, such as factor extraction methods and criteria for determining the number of factors. The variations of choices led to variations in results, but together they more robustly demonstrated that lexical sophistication factors explain a significant portion of variance in oral proficiency scores, as claimed in the original study.

Despite the fact that multiverse analysis is a well-established technique to enhance analytic rigor, and the need to better understand how research choices impact predictive models in education, we could not find any prior research that has applied multiverse analysis in educational data mining, learning analytics, or in the learning sciences.

2.2 Procedure

As researchers have to make choices about parameters at different levels, multiverse analysis generally consists of (1) a data multiverse and (2) a model multiverse [39]. The data multiverse classifies the universes based on data processing parameters. Researchers decide what kind of data to include and what methods to use to filter the dataset. For instance, Steegen et al. studied a data multiverse created by decisions like "assessment of relationship status" (by assigning an integer value) and "exclusion of women based on cycle length" to create a total of 210 combinations of universes for analysis [39]. The model multiverse investigates different modeling assumptions or modeling methods to arrive at analytic results [39]. An example of model multiverse analysis comes from Patel et al. [31], who examined 417 variables' associations with all-cause mortality, demonstrating the choice of predictors and co-variates can significantly impact the associations.

While the multiverse analysis has predominantly been applied for data analysis, some researchers have also extended it for data-collection purposes. Harder [16] used an adaptation of multiverse analysis to analyze 19 studies on shooting decisions with varied data-collection methods such as participant sample size. The multiverse analysis demonstrated how different data-collection choices can influence the eventual research outcome, highlighting the robustness of the finding.

Multiverse analysis can be extended to the process of hyper-

parameter optimization in ML. Common hyper-parameter optimization employs techniques such as grid search and Bayesian optimization to explore the optimal configuration for an ML model systematically [6]. Multiverse analysis in hyper-parameter optimization relies on the ideas of these common techniques with variations of parameter values, and it is similar to the model multiverse discussed above. In this case, the model specifications are tied to the parameter values for the ML models rather than the traditional statistical modeling. Beyond hyper-parameter optimization, integrating multiverse analysis with ML models generally involves a mixture of data multiverse (data processing) and model multiverse (hyper-parameter choices and evaluation metrics). This comprehensive approach allows researchers to understand how combinations of decisions at different levels can affect the model's performance and robustness. For example, Wayland et al. [46] contributed to the PRESTO framework which utilizes multiverse analysis to map latent representations in ML models.

2.3 Distinction from Grid Search

Traditional grid search is a prevalent method for hyper-parameter optimization in machine learning. It systematically evaluates a defined set of parameter combinations to identify the configuration producing the best performance metric, often treating the model as a black box without considering the broader modeling decisions [6]. In contrast, multiverse analysis extends beyond hyper-parameter tuning by systematically exploring a wide array of plausible choices of different stakeholders, including the configuration of the dataset, data preprocessing steps, model architectures, and evaluation metrics. This comprehensive approach assesses the robustness and variability of results across different reasonable analytical decisions at different levels, and hence improves transparency and reproducibility in machine learning research [4].

3. THE EDUCATIONAL CONTEXT

Educational data mining includes a broad range of tasks such as predicting student performance [13, 23], identifying at-risk students [25, 47], analyzing student engagement patterns [14, 24], and personalizing the learning experience [17]. Incorporating multiverse analysis into these studies has the potential to deepen our understanding of how different analytical decisions impact findings, leading to more robust and generalizable conclusions.

Predicting student success is a fundamental challenge in K-12 and higher education [32], with practical value for individual classrooms all the way to full school districts [1]; a large amount of educational data mining research has been centered on predictive modeling and learning analytics. For example, learning analytics tools often use real-time student engagement and performance metrics from learning management systems to help individual faculty identify which students may be struggling at a given time, thereby informing in-time outreach and tailored instructional changes to help every student succeed [36, 1]. At a programmatic level administrators use similar approaches to forecast enrollment, degree completion, and performance, enabling institutions to develop strategies that enhance retention and graduation rates, particularly among underrepresented groups [42].

In the context of educational ML models, various stakeholders are involved in decision-making processes that significantly impact data handling and model development. At the highest level, there are administrators who oversee the dataset and research directions, making critical decisions such as determining the range of data to include, deciding whether to include sensitive characteristics like race and sex, as well as assessing whether to include data collected during the COVID-19 pandemic. These decisions often involve legal and ethical considerations directly related to the persons of interest, namely the students. The National Center for Education Statistics (NCES) requires educators to be aware of regulations and practices regarding data collection and reporting [28]. The second-level stakeholders are data scientists and analysts who typically operationalize data cleaning based on the decisions of the high-level administrators. Their work includes considerations such as how missing values are handled and determining the proportion of datasets used for training and validation. This role is crucial in preparing data for subsequent modeling and analysis [28], and these stakeholders often have a wide variety of techniques available in the changing technology landscape. The third-level stakeholders are machine learning engineers who generally share similar responsibilities to data scientists but often at a more technical or infrastructure level. Here we will classify data scientists as individuals who conduct data processing tasks, including encoding, scaling, and sampling, while machine learning engineers focus on building and tuning machine learning models. As we briefly discussed in the previous section, machine learning engineers have to choose an optimal model and corresponding hyper-parameters to build a model.

Each stakeholder's decisions play a vital role in ensuring a powerful and ethical ML model. With this span of choices at different levels, it paves the way for the application of multiverse analysis.

4. METHODS

As the name implies, multiverse analysis focuses on experimenting with multiple universes. Hall et al. [15] explained the basic elements in multiverse analysis, where a **universe** is one of the analyses conducted in the multiverse analysis report. A **parameter** is a characteristic that varies across the multiverse, and a **parameter value** is a value that the parameter can take. Each combination of parameter values defines a unique universe. Hall et al. [15] provided a simple example: a paper proposes three methods to handle outliers, (1) no exclusion, (2) excluding data 2 Standard Deviations (SDs) from the mean, and (3) excluding data 3 SDs from the mean. Here, the method to handle outliers is the parameter, and each of the three methods is a parameter value for this parameter. Each of the parameter values (potentially combined with parameter values for other parameters) defines a particular universe. Each of the analysis reports also has an **outcome**, which could be the p-value in the previous example [15]. The outcomes are analyzed in an aggregate manner.

4.1 Data Overview

The dataset used in this study contains de-identified information about enrolled students at a public research university in the United States with over 30,000 students for approximately three decades. Because of the enormous amount

of raw data (over 400k entries), we processed them by selecting a fifteen-year time span that covers students enrolled between Fall 2007 and Fall 2022, which results in a dataset with roughly 100k entries.

4.2 Prediction Target

For this work, we chose a single target goal: predicting if a first-year student would re-enroll at the institution in the subsequent fall semester. This target had a binary value of yes/no (represented as 1/0) without any missing values. Whether a student re-enrolls next fall is calculated by checking if the dataset contains enrollment information for the student in that particular semester. Roughly 98% of students in our dataset re-enrolled.

We then processed the dataset with features that can represent a student and indicate a student's potential college success such as GPA and test scores [40, 7, 38]. There are seventeen features in total, and some notable features we calculated include the student's cumulative GPA at the institution before next fall, whether a student is a first-generation college student, the declared major of the student, the number of credits taken in the first year, and the aggregated/converted SAT scores.

4.3 Parameterizing the Multiverse

We applied multiverse analysis in three progressive stages, each one aligning with our different groups of stakeholders: (1) administrative decisions, (2) data processing decisions, and (3) ML model tuning decisions. The choices selected for administrative decisions are relevant because of legal and ethical issues [9], while the choices for data processing and ML models are commonly faced by researchers [33]—they are arbitrarily picked for demonstration in this work. In total, we examine 11 decisions for the re-enrollment target (Table 1), yielding 6,912 choice combinations for our multiverse analysis.

Administrative decisions include (i) whether to include transfer students, (ii) whether to include students enrolled during the COVID period (winter 2020 term to winter 2022 term), (iii) whether to include the student's sex information in the features for training, and (iv) whether to include the student's race information in the features for training.

Data processing decisions include (i) how to handle missing values, either by dropping them or using a simple imputation of the most frequent values, (ii) the size of the training data for the train-test split, with 70% or 80% being selected as examples in this case study, (iii) whether to use one-hot encoding or ordinal encoding of non-ratio data, (iv) whether a standard scaling method is used, and (v) whether the balance sample through use of over-sampling techniques (in this study, SMOTE), under-sampling techniques (NearMiss), or no sampling (full training dataset only).

Machine learning model tuning decisions simply include (i) the choice of classifier employed, and we reduce this to the set of random forest, gradient boosting classifier, or logistic regression, and (ii) one parameter for each of the classifieries, such as the n_estimators of 50, 100, 150 for random forest, the learning_rate of 0.01, 0.1, 1 for gradient boosting classifier, and a C value of 0.01, 0.1, 1 for logistic regression.

Table 1: Overview of the Parameter Space with Stakeholders

Stakeholders	Parameter	Choices
Administrators	Include Transfer Data	True vs. False
Administrators	Include COVID Data	True vs. False
Administrators	Include Sex	True vs. False
Administrators	Include Race	True vs. False
Data Scientists	Handle NaN	Drop vs. Impute
Data Scientists	Train Size	70% vs. 80%
Data Scientists	Encoder	One Hot Encoder vs. Ordinal Encoder
Data Scientists	Scaler	Standard Scaler vs. None
Data Scientists	Sampler	SMOTE; NearMiss; None
ML Engineers	Classifier	Random Forest; Gradient Boosting; Logistic Regression
ML Engineers	Hyper-parameters	N_estimators - *Random Forest*: 50 vs. 100 vs. 150; Learning Rate - *Gradient Boosting*: 0.01 vs. 0.1 vs. 1; C - *Logistic Regression*: 0.01 vs. 0.1 vs. 1

4.4 Training & Testing

With the parameter space defined above, the training of the multiverse fits a scikit-learn pipeline based on the specification of the parameters, the trained model then evaluates the test dataset's AUC score (the higher the better) as well as the Equalized Odds Difference (EOD) (the lower the fairer—smaller difference) concerning sex and race using the Fairlearn package [41]. The specification evaluations are aggregated and analyzed in the next section.

4.5 Specification Curve Analysis

After aggregating and sorting the test set results of different combinations, we plot specification curves that explicitly display each parameter's effect. In multiverse analysis, the specification curves are sorted along the x-axis according to the metric of interest (e.g., AUC), with each specification number representing a unique parameter combination (for a total of 6,912). The y-axis is the variable of interest (AUC score or Equalized Odds Difference in case of fairness), and the different colored curves separate the assignment of one parameter value of interest. For instance, the specification curve in figure 2 looks at the goodness-of-fit of models predicting first-year re-enrollment, specifically comparing the data scientist's choice of whether to drop or impute missing values. Across the 3,500 specifications, we see there is a general improvement of the model which was trained on dropped data, as shown by the higher orange line, suggesting that imputation leads to worse quality models for this task. The specification curves help researchers instantly see the effect of changing one parameter across the breadth of other possible choices.

5. RESULTS

5.1 Effect of Sampler

One noticeable discovery of our multiverse analysis is the effect of employing samplers on the model's performance and fairness. The goodness of fit is significantly better when the model employs an over-sampling technique or does not adopt any sampler. As demonstrated in Figure 3, AUC scores

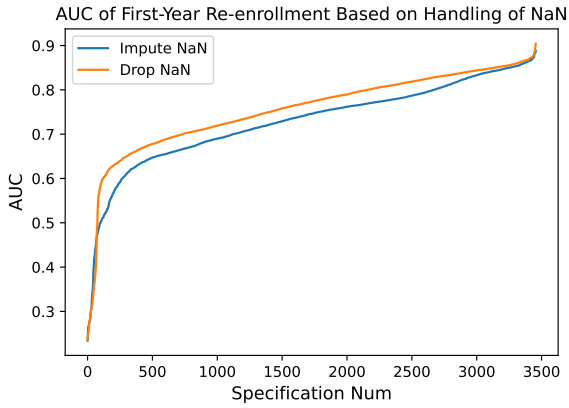

Figure 2: AUC Based on Handling of NaN

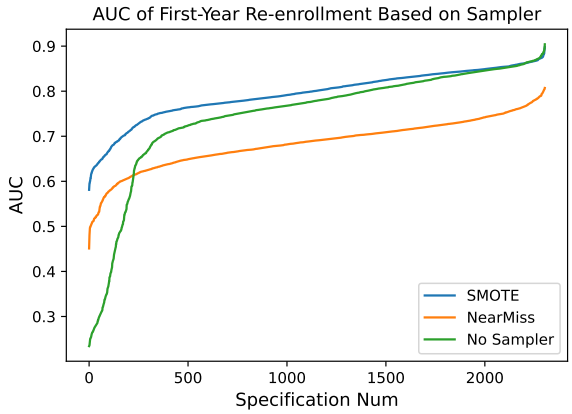

Figure 3: AUC Based on Sampler

for the majority of specifications that use SMOTE or no sampler are about 0.2-0.25 higher than those of NearMiss. This performance gap is likely accounted for by the fact that the dataset contains 98% students who re-enrolled, and an under-sampling process using NearMiss would balance the dataset by drastically decreasing the size of re-enrolled students for training. This extremely small training dataset results in a worse AUC score. While over-sampling through SMOTE creates a balanced dataset by fabricating more values of students not re-enrolled, leading to a slightly higher AUC score than no sampling at all.

The equalized odds difference (EOD) across sexes of the re-enrollment target follows a similar trend as the AUC curves. As shown in Figure 4, NearMiss has a notably higher EOD value for nearly all specifications than SMOTE or no sampler. The specification without any sampler has the lowest equalized odds difference, indicating the fairest model among the three. This indicates that the model with SMOTE sacrifices some fairness in exchange for a slightly better performance in AUC, which is achieved by performing better in the majority group (males).

On the other hand, the equalized odds difference across different racial groups of the re-enrollment target indicates a different trend. In Figure 5, the equalized odds difference of most NearMiss specifications is about 0.3-0.6 lower than

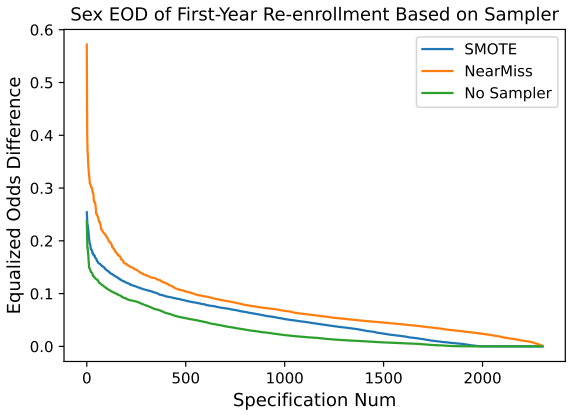

Figure 4: EOD across Sexes Based on Sampler

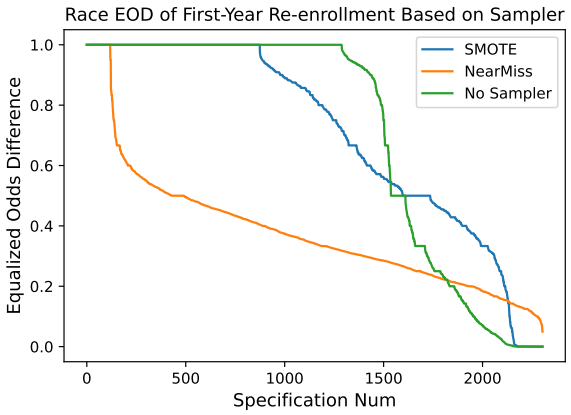

Figure 5: EOD across Races Based on Sampler

those of SMOTE or no sampler. Based on the equalized odds difference here, it seems to suggest that NearMiss is "fairer" concerning different race groups. However, incorporating the analysis from the AUC graph 3 reveals that the low equalized odds difference among races for NearMiss is likely the result of predicting poorly among all the races. Through these aggregate plots, the multiverse analysis demonstrates the interesting effect of samplers/sampling methods on the model performance and fairness.

5.2 Inclusion of Race

The decision of whether to include race as a feature for model training and testing is another valuable discovery from multiverse analysis. The merits of including protected attributes, such as race, have been debated in the research literature. Some argue that the inclusion of such traits improves model performance and addresses systemic inequities [3, 22, 47], while some argue that the inclusion of them is generally a privacy concern, and likely results in bias against minority groups [30, 18, 2]. In our study, the exclusion of races was only narrowly the preferred choice in the multiverse analysis.

Specifically, the AUC curves for including race as a feature and not including race as a feature overlap with each other as displayed in Figure 6. Since the two curves overlap for

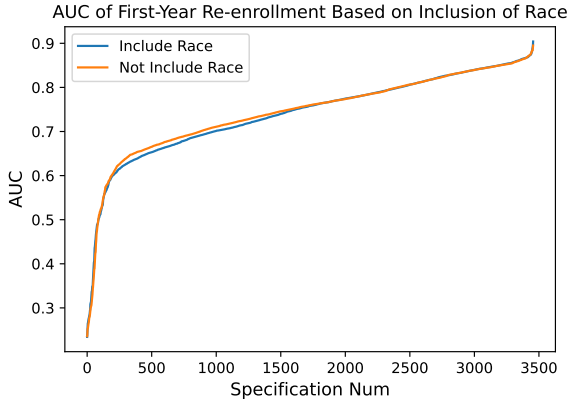

Figure 6: AUC Based on Inclusion of Race

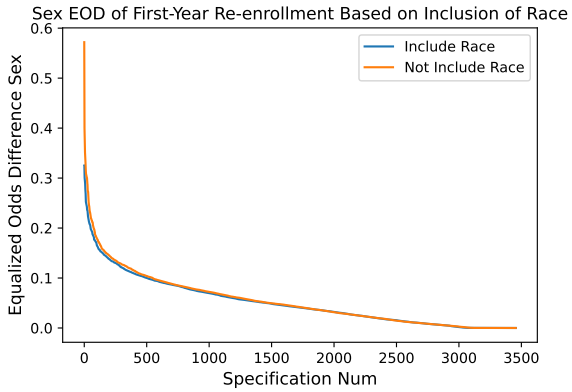

Figure 7: EOD across Sexes Based on Inclusion of Race

most of the specifications, it suggests the inclusion/exclusion of race as a training feature does not impact the model's AUC score in this dataset; this finding is consistent with Yu et al. [47] observations. A similar pattern is observed in the curves for the EOD among sexes as shown in Figure 7, demonstrating that the effect of inclusion/exclusion of race as a training feature on the EOD among sexes is minimal.

On the other hand, the EOD across different racial groups suggests that the exclusion of races can result in slightly fairer models, as illustrated in Figure 8. Without including race as a feature, the equalized odds difference across races is about 0.1 lower for some of the specifications, suggesting fairer models with respect to race. This result aligns with the argument that excluding race from training data prevents the model from learning and propagating existing societal prejudices, resulting in a fairer model with respect to race [43].

Since the AUC and EOD across sexes are about the same without race as a training feature, and the EOD across races is slightly improved when excluding race as a feature, it appears that the exclusion of race does not affect model performance and slightly improves the fairness with respect to race—at least demonstrated in the specific context of this case study.

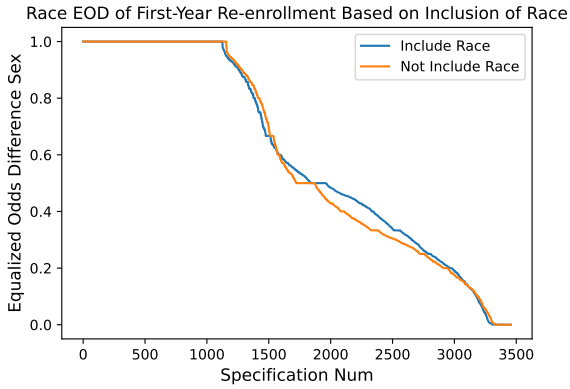

Figure 8: EOD across Races Based on Inclusion of Race

6. CONCLUSION

Multiverse analysis offers a powerful lens for evaluating and improving the robustness, reproducibility, and transparency of predictive models in education. In our case study on predicting first-year re-enrollment, we analyzed thousands of plausible *universes* of decisions that led to several notable findings. We saw that sampling strategies—especially over-sampling and under-sampling—can have significant trade-offs: while over-sampling through SMOTE often yielded gains in AUC score, it also increased disparity across certain subgroups. Similarly, excluding sensitive features such as race did not degrade predictive performance, but it reduced performance disparities by race. These findings emphasize that certain modeling decisions can inadvertently favor one metric, such as goodness-of-fit, at the cost of another, such as group fairness.

By systematically examining each decision that the administrators, data scientists, and ML engineers might make—ranging from whether to include protected features, to how to handle missing data and tune hyper-parameters—our multiverse framework showcased explicitly the interplay among analytic decisions, model performance, and equity implications. Beyond just picking "the best" model on a single metric, this approach highlights the need to weigh contextual and ethical considerations alongside technical performance.

More broadly, our study demonstrates how multiverse analysis can help education stakeholders see and discuss the inherent trade-offs embedded in ML model design. We anticipate that future work will extend this approach to a wider range of educational prediction tasks in K-12, higher education, or lifelong learning with larger or broader datasets to reveal similarly nuanced decision pathways. Ultimately, the multiverse framework equips researchers with a systematic method for constructing learning analytics pipelines that are robust, transparent, and reproducible, thereby better supporting effective interventions and policies aimed at improving student success.

7. ACKNOWLEDGMENT

We gratefully acknowledge the funding support from the Learning Engineering Virtual Institute (LEVI) and the Michigan Institute for Data & AI in Society (MIDAS) - Propelling Original Data Science (PODS) program, funded in part by Microsoft (MIDASPODS2418).

8. REFERENCES

[1] E. Alyahyan and D. Düştegör. Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1):1–21, 2020.

[2] R. S. Baker, L. Esbenshade, J. Vitale, and S. Karumbaiah. Using demographic data as predictor variables: a questionable choice. *Journal of Educational Data Mining*, 15(2):22–52, 2023.

[3] I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, P.-C. Kuo, M. P. Lungren, L. Palmer, B. J. Price, S. Purkayastha, A. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H. Trivedi, R. Wang, Z. Zaiman, H. Zhang, and J. W. Gichoya. Reading race: Ai recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*, 2021.

[4] S. J. Bell, O. P. Kampman, J. Dodge, and N. D. Lawrence. Modeling the machine learning multiverse. *arXiv preprint arXiv:2206.05985*, 2022.

[5] M. Beseiso. Enhancing student success prediction: A comparative analysis of machine learning techniques. *TechTrends*, 69:372–384, January 2025.

[6] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2):1–43, 2023.

[7] N. W. Burton and L. Ramist. Predicting success in college: Sat® studies of classes graduating since 1980. Technical report, College Entrance Examination Board, 2001.

[8] G. G. Cantone and V. Tomaselli. Theory and methods of the multiverse: an application for panel-based models. *Quality & Quantity*, 58:1447–1480, 2024.

[9] G. N. Carlizzi and G. Quattrone. The algorithmic public decision, between explainability, administrative law and artificial intelligence. In D. Marino and M. Monaca, editors, *Artificial Intelligence and Economics: the Key to the Future*, pages 153–168. Springer, Cham, Switzerland, 2022.

[10] O. S. Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):943–950, August 2015.

[11] S. Demir and E. K. Sahin. The effectiveness of data pre-processing methods on the performance of machine learning techniques using rf, svr, cubist and sgb: a study on undrained shear strength prediction. *Stochastic Environmental Research and Risk Assessment*, 38:3273–3290, 2024.

[12] M. Eguchi and K. Kyle. Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, 104(2):381–400, 2020.

[13] S. Ghosh, A. Mondal, and D. De. Student performance analysis and prediction in classroom learning: A regressive approach. *Education and Information Technologies*, 26(1):205–240, 2021.

[14] Y. Guo, C. Gunay, S. Tangirala, D. Kerven, W. Jin, J. C. Savage, and S. Lee. Identifying critical lms features for predicting at-risk students. *arXiv preprint arXiv:2204.13700*, 2022.

[15] B. D. Hall, Y. Liu, Y. Jansen, P. Dragicevic, F. Chevalier, and M. Kay. A survey of tasks and visualizations in multiverse analysis reports. *Computer Graphics Forum*, 41(1):402–426, 2022.

[16] J. A. Harder. The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, 15(5):1158–1177, 2020.

[17] M. Injadat, A. Moubayed, A. Bou Nassif, and A. Shami. Systematic ensemble model selection approach for educational data mining. *arXiv preprint arXiv:2005.06647*, 2020.

[18] C. Intahchomphoo and O. E. Gundersen. Artificial intelligence and race: a systematic review. *Legal Information Management*, 20(2):74–84, 2020.

[19] H. Jeong, H. Wang, and F. P. Calmon. Fairness without imputation: A decision tree approach for fair prediction with missing values. *arXiv preprint arXiv:2109.10431*, 2021. Accessed: 2025-02-10.

[20] L. K. John, G. Loewenstein, and D. Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532, 2012.

[21] K. Jung, S. Shavitt, M. Viswanathan, and J. M. Hilbe. Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 111(24):8782–8787, 2014.

[22] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. In W. Holmes and K. Porayska-Pomsta, editors, *Ethics in Artificial Intelligence in Education*, pages 174–202. Taylor & Francis, London, United Kingdom, 2022.

[23] R. F. Kizilcec, M. Pérez-Sanagustín, and J. J. Maldonado. Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & education*, 104:18–33, 2017.

[24] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 170–179, 2013.

[25] H. Li, W. Ding, and Z. Liu. Identifying at-risk k-12 students in multimodal online environments: A machine learning approach. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, pages 431–437. International Educational Data Mining Society, 2020.

[26] R. Maie, M. Eguchi, and T. Uchihara. Arbitrary choices, arbitrary results: Three cases of multiverse analysis in l2 research. *Research Methods in Applied Linguistics*, 3(2):1–20, 2024.

[27] F. Martínez-Plumed, C. Ferri, D. Nieves, and J. Hernández-Orallo. Fairness and missing values. *arXiv preprint arXiv:1905.12728*, 2019. Accessed: 2025-02-10.

[28] National Forum on Education Statistics. Forum guide to data ethics, 2010.

[29] F. A. Orji and J. Vassileva. Machine learning approach for predicting students' academic

performance and study strategies based on their motivation. *arXiv preprint arXiv:2210.08186*, 2022. Accessed: 2025-02-10.

[30] T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, R. M. Peixoto, G. A. S. Guimarães, G. O. R. Cruz, M. M. Araujo, L. L. Santos, M. A. S. Cruz, E. L. S. Oliveira, I. Winkler, and E. G. S. Nascimento. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1):1–31, 2023.

[31] C. J. Patel, B. Burford, and J. P. A. Ioannidis. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9):1046–1058, 2015.

[32] J. C. Perdomo, T. Britton, M. Hardt, and R. Abebe. Difficult lessons on social prediction from wisconsin public schools. *arXiv preprint arXiv:2304.06205*, 2023.

[33] P. Probst, A.-L. Boulesteix, and B. Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53):1–32, 2019.

[34] S. Samuel, F. Löffler, and B. König-Ries. Towards explaining the effects of data preprocessing on machine learning. In *2019 15th International Conference on eScience (eScience)*, pages 340–341, 2019.

[35] S. Samuel, F. Löffler, and B. König-Ries. Machine learning pipelines: Provenance, reproducibility and fair data principles. In B. Glavic, V. Braganholo, and D. Koop, editors, *Provenance and Annotation of Data and Processes: 8th and 9th International Provenance and Annotation Workshop, IPAW 2020 + IPAW 2021, Virtual Event, July 19–22, 2021, Proceedings*, volume 12839 of *Lecture Notes in Computer Science*, pages 226–230. Springer, 2021.

[36] L. Shepard, G. Rehrey, and D. Groth. Faculty engagement with learning analytics: Advancing a student success culture in higher education. In M. Shah, S. Kift, and L. Thomas, editors, *Student Retention and Success in Higher Education*, pages 89–107. Palgrave Macmillan, Cham, 2021.

[37] U. Simonsohn, J. P. Simmons, and L. D. Nelson. Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*, pages 1–29, 2019.

[38] A. Singh. Investigating the predictors of first-time student retention at uw. Technical report, University of Wyoming, Office of Institutional Analysis, 2019.

[39] S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

[40] T. L. Strayhorn. Factors influencing the academic achievement of first-generation college students. *NASPA Journal*, 43(4):82–111, 2006.

[41] F. D. Team. *Common fairness metrics*, 2025. Accessed: 2025-02-10.

[42] M. N. Tedeschi, T. M. Hose, E. K. Mehlman, S. Franklin, and T. E. Wong. Improving models for student retention and graduation using markov chains. *PLOS ONE*, 18(6):1–14, 2023.

[43] I. Valentim, N. Lourenço, and N. Antunes. The impact of data preparation on the fairness of software systems. *arXiv preprint arXiv:1910.02321*, 2019.

[44] A. Villar and C. R. V. de Andrade. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, 4(2):1–24, 2024.

[45] E.-J. Wagenmakers, R. Wetzels, D. Borsboom, and H. L. J. van der Maas. Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 100(3):426–432, 2011.

[46] J. Wayland, C. Coupette, and B. Rieck. Mapping the multiverse of latent representations. *arXiv preprint arXiv:2402.01514*, 2024.

[47] R. Yu, H. Lee, and R. F. Kizilcec. Should college dropout prediction models include protected attributes? In *Proceedings of the eighth ACM conference on learning@ scale*, pages 91–100, 2021.

APPENDIX

Code for this paper is accessible at: `https://github.com/educational-technology-collective/multiverse-code`