Adapting Multiverse Analysis for Prediction: A Decision-Maker's Dashboard

Fifi Ding [1], Galit Shmueli [1], Travis Greene [2], Sofie Goethals [3]

[1]Institute of Service Science, National Tsing Hua University
[2]Department of Digitalization, Copenhagen Business School
[3]Department of Engineering Management, University of Antwerp
fifi.ding@iss.nthu.edu.tw, galit.shmueli@iss.nthu.edu.tw, trgr.digi@cbs.dk, sofie.goethals@uantwerpen.be

Abstract

Algorithmic decision-making systems increasingly guide consequential judgments in domains such as criminal justice, credit scoring, and healthcare. Yet, their legitimacy is undermined by predictive inconsistency. When the same person receives different risk scores depending on which data and modeling pipeline is used, even when all paths are reasonable, individuals with highly variable predictions might face unfair treatment. While prior research conceptualized this multiplicity issue and proposed multiverse analysis (specification curves) to expose variability, existing tools remain inaccessible to non-technical stakeholders who rely on algorithmic predictions in high-stakes decisions. This study introduces an interactive multiverse visualization dashboard that translates complex pipeline variability into interpretable insights for decision-makers, using regression trees in the backend. Unlike conventional explainability tools that focus on how one model treats different individuals, our framework explores how one individual can yield multiple algorithmic outcomes across plausible analytical pipelines. The dashboard integrates three innovations: (1) interactive specification curves adapted for predictive outcomes, (2) a regression tree-driven engine to help identify which modeling decisions most influence prediction variability, and (3) profile comparison that visual counterfactual exploration of feature variations. Through these features, we provide an evidence-based support tool to inform decision-makers in various high-stake domains.

Code — https://github.com/fifi-ding/basic_multiverse

Introduction

Algorithmic decision-making (ADM) systems shape important decisions across society, determining who receives bail, who qualifies for credit, and who gains access to healthcare coverage. Operating through machine learning algorithms (Brynjolfsson and Mitchell 2017; Newell and Marabelli 2015), these systems have become pervasive in high-stakes domains including criminal justice (Monahan and Skeem 2016; Garrett and Monahan 2019), credit scoring (Hand and Henley 1997; Authority 2019) , and health insurance (Cumming et al. 2002). These tools generate risk scores or probabilities indicating the likelihood an individual will experience a specific outcome such as loan default, criminal recidivism, or healthcare utilization, which directly inform consequential decisions about support or punishment (Berk, Berk, and Drougas 2019; Siddiqi 2012). Yet despite widespread deployment, their legitimacy is questionable: identical individuals can receive dramatically different risk assessments depending on which modeling pipeline choices system designers make.

Greene et al. (2022) conceptualize this phenomenon as *predictive inconsistency*, referring to situations where identical individuals receive different predicted scores because of conceptually justified yet technically distinct pipeline modeling choices (see Figure 1). Such choices include using different sources of data, measuring variables in different ways, and processing data through different techniques. In practice, these decisions manifest in numerous ways: practitioners routinely determine which risk factors to include (DeMichele et al. 2020; Duncan 2011; Metz et al. 2019), what predictive models to implement (Abdou and Pointon 2011; Ash et al. 2000), and how prediction quality should be evaluated (Bhatore, Mohan, and Reddy 2020). While these are relatively visible modeling decisions, they represent only part of the challenge. Less transparent are the processing choices that often go unreported, including data bias introduced during collection (Barocas and Selbst 2016; Caton and Haas 2024), subtle variations in questionnaire wording (Duncan 2011), and inconsistencies in defining key terms (Thomas 2000; Garrett and Monahan 2019). Taken together, both visible and hidden modeling choices contribute to predictive inconsistency and lead to the *Rashomon effect*, where different models may exhibit similar minimum error rates despite arriving at various outcomes (Breiman 2001).

This often unreported variation creates parallel versions of the same individual with divergent predicted outcomes (Wynand, De Ven, and Ellis 2000), challenging the core requirements of procedural consistency and fairness in high-stakes decision-making contexts (Lee et al. 2019). One way to understand this is to imagine an individual existing within multiple algorithmic *universes* simultaneously, where different modeling pipeline approaches applied to that same individual's profile produce significantly varying risk outcomes. These inconsistencies mirror a broader challenge in scientific research: the problem of analytical flexibility and

Workshop on Navigating Model Uncertainty and the Rashomon Effect: From Theory and Tools to Applications and Impact (AAAI 2026)

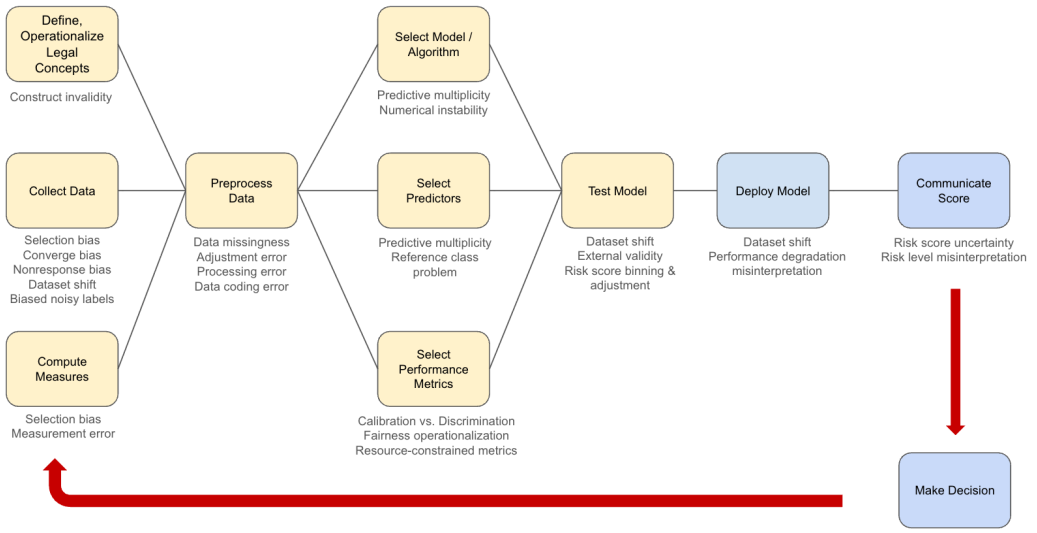

Figure 1: Sources of predictive inconsistency. Yellow nodes show flexible data-processing stages; blue nodes show post-deployment stages with less design control. (Adapted from Greene et al. (2022))

its threat to reproducibility. Simmons, Nelson, and Simonsohn (2011) termed *researcher degrees of freedom* as the flexibility researchers possess in data collection, analysis, and reporting, which can unintentionally skew results toward desired outcomes. Subsequent studies (Nosek, Spies, and Motyl 2012; Gellman and Lokem 2014) have shown how such flexibility, often motivated by disciplinary norms rather than misconduct, leads to what is known as the *garden of forking paths*. This refers to a multitude of plausible yet unacknowledged analytical pipelines that lead to inflated false-positive findings and undermine the reliability of both scientific and algorithmic inference. In response to the reproducibility crisis, specification curve analysis (SCA), as developed by Simonsohn, Simmons, and Nelson (2020), visualizes the robustness of research findings across multiple analytical choices. SCA consists of two aligned panels (Figure 2): The top panel is the *specification curve*, where each point represents the outcome of interest (e.g., effect size or p-value) from a different model specification. The bottom panel is the *specification grid*, which details the analytical decisions (e.g., model types, sample restrictions) underlying each point. The panels are aligned on the x-axis, allowing readers to trace any result on the curve down to the exact combination of choices in the grid. The colors show the statistical outcome: red is statistically significant and gray is statistically insignificant. Specifications, also known as universes, are generated by systematically testing all combinations of feasible analytical options (e.g. 3 outcome measures $\times$ 3 treatment types $\times$ 2 models $\times$ 4 offense types = 72 specifications), revealing the sensitivity of the findings to methodological variation.

Greene et al. (2022) suggested adapting SCA for prediction rather than statistical inference by focusing on predicted scores for a single individual. However, they did not provide practical tools for decision-makers to navigate predictive inconsistencies. While their work makes visible the typically hidden modeling pipeline choices that affect individual outcomes, significant gaps remain in translating multiverse analysis into actionable insights for non-technical stakeholders in real-world settings. Judges determining bail, loan officers evaluating credit applications, or insurers making coverage decisions, are typically presented with simplified outputs: a numerical probability score and a categorical recommendation. They remain unaware of the numerous modeling pipeline decisions made throughout the process, lack insight into which choices most influence outcomes, and have no way to evaluate whether alternative pipeline approaches might yield different conclusions for the individuals whose lives hang in the balance.

This study addresses these limitations by developing an interactive multiverse visualization dashboard that translates complex analytical variability into interpretable insights for decision-makers without technical expertise. In contrast to conventional approaches that analyze statistical relationships or existing explainability tools –such as SHAP– that elucidate how a single model produces different predictions across individuals, our framework shifts focus to individual-level predictions, exploring how one individual could have multiple algorithmic versions of themselves across reasonable modeling pipeline choices. The dashboard allows comparative analysis between a focal profile (the actual individual) and counterfactual profiles (the same individual with modified characteristics such as gender or race), revealing not only prediction variability but also interactions between processing pipelines and demographic features. Through three innovations: interactive specification curves adapted for probability outcomes, regression trees that identify key pipeline decisions that drive prediction differences, and pro-

file counterfactual manipulation and comparison capabilities, we make the forking paths of algorithmic decision-making both visible and comprehensible to stakeholders who must act on these predictions.

Related Work

Rashomon Effect

The Rashomon effect refers to situations where multiple models achieve similarly high predictive performance (Breiman 2001). Building on this idea, Semenova, Rudin, and Parr (2019) introduced the concept of Rashomon sets: the collection of models within a given function class that perform approximately as well as the best model. Tools such as TimberTrek (Wang et al. 2022) have been developed to help users examine these sets and select tree model variants that align with their interpretive or analytical preferences. However, these tools primarily focus on model-level differences and do not account for variability introduced in earlier data-processing stages, such as missing-value handling or class imbalance, which can also contribute to predictive inconsistency.

Rather than applying trees solely for variable selection, we use regression trees to analyze the entire modeling pipeline. Within Greene et al. (2022)'s framework, Rashomon sets can be seen as a subset of our dashboard's exploration space, focusing mainly on variations in model type and configuration.

Multiverse Visualization Tools

Multiverse analysis emerged in response to the reproducibility crisis, encouraging researchers to examine all plausible modeling decisions to ensure robustness. Liu et al. (2020) developed Boba Visualizer, which simplifies this workflow by letting users define analysis code once and explore each feature's effect on the coefficient estimations interactively, but their work remains as statistical analysis rather than a predictive one. Simson, Pfisterer, and Kern (2024) applied multiverse fairness to public health insurance data revealing how certain processing methods influence fairness. However, in his interactive tool, we are unable to tell how each decision impacted the fairness metrics without reviewing each universe individually.

Both the Rashomon effect and multiverse analysis have advanced our understanding of model variability, yet they remain limited in scope. Rashomon sets primarily capture variation across model architectures and parameters, overlooking data preprocessing choices that can equally affect predictions. Similarly, current multiverse analysis tools emphasize statistical robustness or fairness assessment but lack integration with predictive modeling workflows and do not provide a unified view of how preprocessing, modeling, and evaluation decisions interact. Our proposed dashboard addresses these shortcomings by bridging the two perspectives: enabling systematic exploration of model, variable, and preprocessing combinations within a single predictive framework. This innovation provides not only transparency into how modeling choices shape outcomes but also actionable insight for selecting models aligned with interpretability goals.

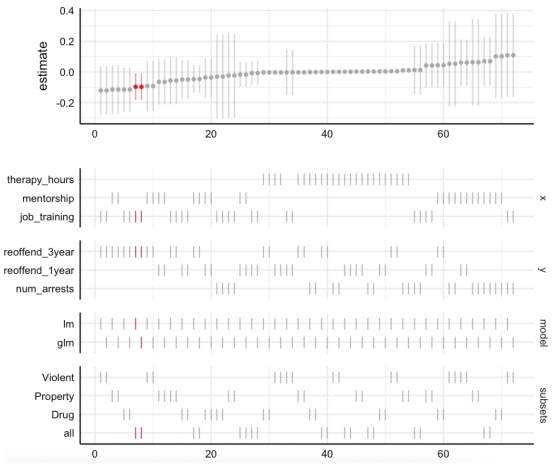

Figure 2: Specification Curve Analysis Demonstration

Decision-Making Setup

We consider three datasets of individuals who returned to jail ("recidivators") across different time periods and contexts. This allows us to compare how data practices and modeling choices influence predictions. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset (Angwin et al. 2016) is widely used in fairness research for its controversial algorithmic risk scores, containing demographic and criminal history data for over 7,000 defendants. The North Carolina (NC) prisoners dataset (Schmidt and Witte 2012) reflects an earlier data collection era, emphasizing survival analysis of 4,618 released prisoners with simpler race and offense categorizations, offering historical contrast in methodological norms. The Georgia State (GA) prisoners dataset (2013-2015), released by the National Institute of Justice's Recidivism Forecasting Challenge, includes over 25,000 parolees with rich documentation and modern predictive modeling approaches. Together, these datasets span nearly four decades of criminal justice data, allowing systematic exploration of how definitional, measurement, and methodological choices shape recidivism prediction outcomes. To ensure consistent analysis across datasets, we designed data processing and modeling pipelines that standardize variable definitions, handle missing or inconsistent measures, and automate multiverse generation. Each pipeline integrates stages such as data cleaning, feature engineering, recidivism definition, model specification, and performance evaluation. This allows flexible combinations of modeling decisions to form distinct "universes". This structure enables systematic comparison of how variations in predictors, preprocessing, and modeling methods influence outcomes within and across datasets with differing levels of completeness and preprocessing.

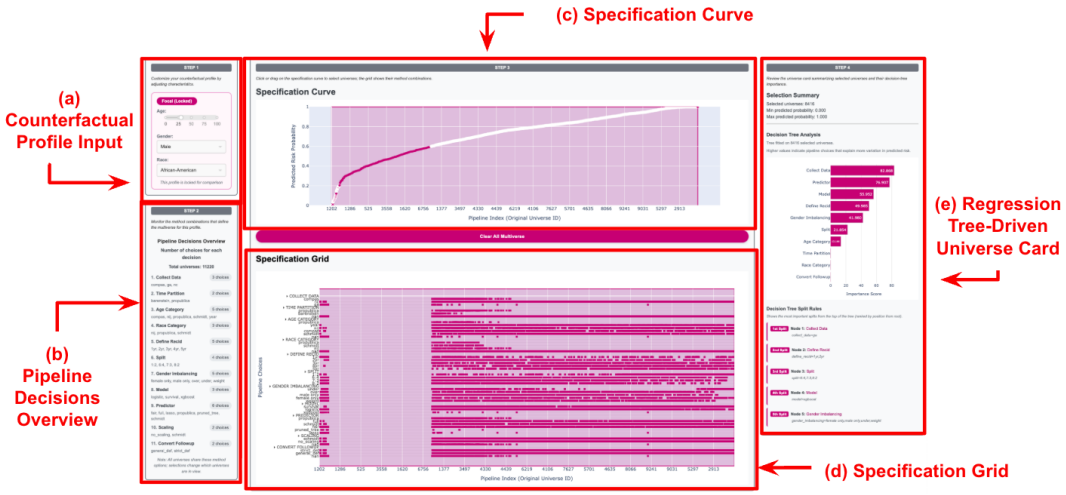

Figure 3: Multiverse Dashboard Components

Predictive Inconsistency: Dashboard Design

To demonstrate how decision-makers can engage with our multiverse dashboard, consider the scenario of a judge, who must decide whether to grant bail to an individual: a 30-year-old Black male. The judge wants to understand how different pipeline modeling choices such as variable selection, resampling strategy, and model type affect the individual's predicted risk of recidivism. Using universes built from three datasets (COMPAS, NC, and GA), the judge turns to the dashboard to explore prediction variability and identify the methodological conditions under which the individual might be assessed as high or low risk.

Focal Profile and Pipeline Decisions Overview

Figure 3(a) shows the focal profile of the individual the judge is examining. The focal profile lists out the demographic characteristics of the individual. For this 30-year-old Black male, he is recorded as a 30 year-old African American male in the dashboard. The judge proceeds to check the Pipeline Decisions Overview (Figure 3(b)), giving him/her a high-level snapshot of how the pipelines were constructed. The judge can immediately see the total number of universes produced, along with the pipeline modeling decisions that define them such as resampling strategies, feature selection methods, and model types. Each decision point lists how many alternative options were tested, revealing the full range of methodological variability behind the results.

Specification Curve Analysis

The judge then reviews the specification curve (Figure 3(c)). Specification curves visualize the distribution of predicted probabilities across all universes, where the y-axis represents predicted probabilities (ranging from 0 to 1), and the x-axis lists universes sorted in ascending order by these predicted probabilities. Below the specification curve, the specification grid (Figure 3(d)) follows the same order, ensuring that each grid column corresponds directly to a point on the curve. Along the y-axis, the grid lists the available pipeline options such as whether the sample includes only females or males, whether demographic predictors are excluded, and whether survival models are used. Each column represents a unique combination of these options, forming one universe within the multiverse. For example, a column labeled over indicates that male participants were oversampled in that universe.

The judge can interactively select specific probability ranges on the curves to filter the universes displayed in the grid. This filtering feature allows the judge to focus on subsets of universes that produce similar predictions and examine the modeling pipeline decisions that lead to those outcomes. For example, if the judge wants to find the pipeline decisions that lead to lower predicted recidivism probabilities compared with higher probabilities, s/he can select two probability ranges (e.g. 0-0.2 and 0.6-1) to see the processing choices that may have led to this difference. The judge here notices from the grid that most low recidivism predictions are due to shorter recidivism definitions (one year follow-up period). However, visual inspection alone may not reveal the key differences between the two profiles' specifications. To address this challenge, we make use of a transparent machine learning tool by running a regression tree with the predicted probabilities as the outcome, and the multitude of universe pipeline modeling decisions as predictors. The result is displayed in a regression tree-driven universe card that identifies the most influential pipeline decisions and options distinguishing the selected probability regions.

Universe Cards

The universe cards use regression tree analysis to identify which pipeline modeling decisions have the greatest impact on outcome variability. The cards present this information through two components: (1) bar charts that display the relative importance of each modeling decision (derived from impurity reduction), and (2) detailed breakdowns of the spe-

Figure 4: Implementation of counterfactual profiles using GA

cific options that create splits in the regression tree. [1] Together, these visualizations help users understand the key factors driving differences in predictions across analytical approaches.

Continuing from the previous scenario, the judge consults the universe card to explore why prediction results differ. From the bar charts (see Figure 3(e)), the judge observes that data collection, predictors, and model selection most affected the predictions. To investigate further, the judge examines the tree split rules underneath the bar chart, which also reveal that data collection affects predictions the most.

Incorporating counterfactual profiles

During dashboard development, we created several prototypes that varied based on the type of profile being analyzed and the number of datasets displayed. One extension of the basic design is a version that incorporates a counterfactual profile (see Figure 4), allowing users to examine how predictions change when a specific feature, such as race, is altered. For example, we compare a 30-year-old African American individual with a counterfactual version of the same person labeled as Caucasian (White). In addition to adding the counterfactual specification curve and grid, this prototype introduces another regression tree that analyzes the selected region of universes across both the focal and counterfactual profiles, enabling cross-profile decision analysis. For this prototype, we used the GA dataset exclusively. As shown in Figure 4, the specification curve for the African American profile becomes markedly steeper than the Caucasian curve in the 0.8-1 range. To understand this divergence, we inspected the specification grid and the corresponding universe cards and found that differences in splitting methods and recidivism definitions were key contributors, especially the absence of shorter follow-up periods in the African American profile. This highlights how even small variations in the profile, despite sharing the exact same pipelines, can affect predictions.

Discussion

The system's interactive visualizations and traceability features reveal patterns of instability across universes, improving transparency in high-stakes settings. Our study contributes three innovations: (1) interactive specification curves for predictive outcomes, (2) a regression tree–driven engine to identify influential pipeline modeling choices, and (3) profile comparison tools for visual counterfactual exploration.

Despite these advances, development surfaced challenges, including balancing transparency with usability, integrating heterogeneous datasets with missing inputs and unequal sizes, and constructing coherent individual profiles amid dataset inconsistencies. In this section, we outline these challenges and their implications for deploying predictive multiverse analysis in real-world applications.

Information overload and transparency

The addition of counterfactual features rapidly increases complexity and crowding of the visualization. Research shows that people can only hold three to nine chunks of visual information in short-term memory and excessive complexity can hinder comprehension (Few 2006). We do not

[1] The regression tree is implemented using R's `rpart()` function, where inputs are all the pipeline decisions in the multiverse and the outcome is their respective predicted probabilities.

Table 1: Sample data for Figure 3

Collect Data	Age Category	Define Recid	Split	Gender Imbalancing	Model	Predictor	Recidivism Prob
compas	propubica	2yr	6:4	under	logistic	propubica	0.675148496
compas	propubica	2yr	6:4	under	logistic	propubica	0.589255221
compas	year	2yr	6:4	under	logistic	propubica	0.688810423
compas	year	2yr	6:4	under	logistic	propubica	0.603105608
compas	nij	2yr	6:4	under	logistic	propubica	0.684402419
compas	nij	2yr	6:4	under	logistic	propubica	0.589500402

yet know whether decision-makers prefer richer counterfactual information or the simplicity of a single focal profile. Our dashboard attempts to provide as much relevant information as possible, but the current design introduces several layers of visual complexity. The counterfactual profile, method overview, and specification grid show what and how data are used and processed. The specification curves visualize how pipeline choices influence predictions, while the universe cards offer interpretable summaries of the specification grid. These components collectively risk overwhelming users. As Wilke (2019) notes, a visualization may be memorable yet still confusing if it overloads cognitive capacity.

Hence, we must assess whether increased transparency of pipeline modeling decisions actually promotes more informed judgments in practice. Prior work has shown that clarity does not necessarily improve fairness. Lee et al. (2019) found that increased transparency did not lead to fairer decisions, which raises the possibility that our dashboard may exhibit similar limitations. Some components in the dashboard contain overlapping information. The specification grid, universe cards, and method overview all communicate details about pipeline processing. One option is to consolidate the grid and method overview into a single panel by making the grid's y-axis more intuitive, such as by color-coding parameter categories and values. Another approach is removing the specification grid entirely. Displaying all universes simultaneously in the grid may also overwhelm viewers, who tend to focus on the relationship between curves and a few universe characteristics rather than carefully examining the full set of specifications. Relying solely on universe cards would reduce cognitive load, but at the cost of transparency. This creates an inherent tension between simplifying the interface and preserving full visibility of pipeline decisions, suggesting a need for adaptive interfaces that present information most relevant to the decision-maker while allowing deeper inspection on demand.

An alternative solution is to remove the specification grid, and instead dedicate more space to enlarging the universe card. Users often interpret scrollable content as less important (Few 2006), so larger cards may increase visibility and highlight regression tree insights without scrolling. User studies can determine whether the dashboard supports fairer decision-making, rather than simply improving visual clarity.

We also acknowledge that the current labels for pipeline modeling choices in the method overview may be insufficient for technical stakeholders. For example, terms like over do not specify the exact processing methodology used. In future designs, we plan to improve clarity by allowing hover-over explanations or providing a terminology dictionary.

Generalizability across datasets

As described earlier, we designed different prototypes depending on whether the dashboard includes a counterfactual profile and how many datasets need to be visualized. Figure 3 shows that data collection has the strongest impact on predicted probabilities. The dataset underlying this dashboard was created by three multiverse analyses (COMPAS, NC, and GA) each generated from universe designs tailored to the variables available in that dataset. After generating these multiverse predictions independently, we combined them into a single unified dataset for visualization.

In this combined dataset, each column represents a processing step (e.g., data collection, data splitting, gender rebalancing, predictor selection, and more), while each row corresponds to a single universe (see Table 1). The data collection column contains three possible values (COMPAS, NC, and GA) indicating which dataset was used. Some procedures, such as gender imbalancing, are shared across all datasets; therefore, each dataset includes universes oversampling females, undersampling males, weighting cases to balance male and female representation, female-only samples, and male-only samples. Other procedures apply only to specific datasets. For example, time partitioning is available only for COMPAS, so its corresponding entries for NC and GA are marked as N/A.

Multiverse analysis is effective when conducted on COMPAS, NC, and GA separately, but it becomes significantly more complex when we attempt to integrate the datasets into a single multiverse. We initially tried to standardize the universe designs across all three datasets, but even seemingly simple variables revealed deep incompatibilities. Take prior offenses, for example: GA provides a detailed breakdown of different types of prior offenses, whereas COMPAS and NC report only a total count. There is no reliable way to determine whether aggregating GA's detailed categories produces a measure comparable to the single-number priors used in COMPAS and NC.

Race categorization introduces an additional complication. COMPAS reports race using multiple granular categories (Caucasian, African American, Hispanic, Asian, Native American, and Others). NC, in contrast, collapses race into a binary indicator in which WHITE distinguishes Black individuals from everyone else (i.e., Caucasian, Hispanic, Asian, and Other). GA is even more restrictive: its authors

explicitly state that they only provide data for Black and White individuals. These inconsistencies make it difficult to meaningfully align universe designs across datasets and highlight the inherent risks of attempting cross-dataset standardization in multiverse workflows.

Profile Creation

These inconsistencies become even more problematic when using the system's predictions for a new profile, as would occur in practice. Imagine a new case: a 30-year-old Hispanic male charged with a misdemeanor and reported to have five prior offenses. This person cannot be represented in GA because GA does not include Hispanic as a racial category, nor can we map his total number of priors to GA's detailed prior-type structure. In real-world systems, we often do not know how data scientists handle these gaps. They might substitute a person with a similar criminal history but a different race, or they might impute missing information using the mean or some other aggregation. We experimented with a similar strategy to approximate missing information. First, we identified variables that were consistently present and semantically comparable across datasets such as age, gender, and race when race was measured in a compatible way. We then constructed an individual's profile using only these consistent variables. For variables that were not comparable across datasets, we searched within each dataset's test set to find individuals who matched the target person on the consistent variables and used their values to fill in the remaining fields needed for prediction. Under Greene et al. (2022) framework, profile creation itself becomes a new source of predictive inconsistency. The same person may be represented with different values depending on how variables are defined, how incompatible attributes are reconstructed, or how the system adapts once deployed. All of these factors contribute to additional layers of variation in the final predictions.

Self-consistency over accuracy

Multiverse analysis usually provides transparency in statistical inference rather than supporting predictive modeling. This difference matters because statistical and predictive analyses have different goals and constraints. In statistical analysis, researchers examine effect sizes and statistical significance (e.g. p-values) using complete datasets where outcomes are already known. Predictive modeling operates differently, where we attempt to predict unknown outcomes using only available predictor information (input variables, or features). Since predictions involve inherent uncertainty, we cannot verify their accuracy until future outcomes occur. This uncertainty allows for more pipeline variability. Researchers make numerous decisions such as which features to include, how to preprocess data, which algorithms to use, and how to tune parameters. Each researcher may pursue different strategies they believe will optimize prediction performance (e.g. accuracy), leading to *researcher degrees of freedom*. By creating many simulated universes, we can systematically vary factors that may influence the outcome. Exploring different combinations of these factors across universes allows us to observe which configurations produce better performance. However, multiverse analysis also reveals the worst outcome for the individual across these universes. Their best and worst outcomes, whether expressed as scores, probabilities, or labels, can differ dramatically. This variation illustrates why traditional model-level accuracy metrics are often uninformative in this context. A model might achieve 99% accuracy overall, yet the particular target individual might be in the 1%. From the individual's perspective, the high accuracy offers no comfort: their prediction is simply wrong. Conversely, a best universe might yield a model with only 20% overall accuracy, yet correctly predict the individual's outcome. For this reason, our multiverse does not emphasize conventional accuracy metrics. Instead, we focus on self-consistency measures that better capture the multiplicity of an individual's outcomes by examining how much their predictions vary across universes. This variance provides a more meaningful assessment of stability and reliability at the individual level.

Adaptability to other high stakes domains

While beyond the scope of this paper, we expect that applying multiverse analysis to high-stakes domains like credit scoring and healthcare will reveal how predictive variability is shaped not only by technical choices but also by regulatory, ethical, and institutional constraints. In credit scoring, design decisions about variables, proxies, and fairness constraints can reinforce inequities despite appearing legally neutral, raising questions about which universes are permissible under antidiscrimination laws. Healthcare introduces a different complexity: risk itself is multidimensional and contested, with universes varying in how they handle missing clinical data, population stratification, and competing clinical and financial priorities. Such considerations imply that extending predictive multiverse analysis requires not only technical adjustments, but also careful design choices and domain knowledge to ensure the results remain interpretable and relevant. It demands domain-specific governance, ethical consideration, and stakeholder input to determine which decisions during the process are meaningful, acceptable, and fair. Applying our dashboard to other high-stakes domains is therefore an important future step.

Conclusion

By generating multiple universe pipelines and visualizing their outcomes through an interactive dashboard, our study reveals that even seemingly minor pipeline modeling decisions can alter predictions for the same individual. Beyond methodological contributions, our study proposes a visualization framework to make these dependencies interpretable to non-technical stakeholders. We also identified tensions between transparency and fairness that require exploration through user studies and collaborative design with real-world stakeholders. In doing so, our work lays the groundwork for systems that not only reveal how predictions are produced, but also support more thoughtful and holistic decision-making across high-stakes environments.

References

Abdou, H. A.; and Pointon, J. 2011. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, 18(2-3): 59–88.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23.

Ash, A. S.; Ellis, R. P.; Pope, G. C.; Ayanian, J. Z.; Bates, D. W.; Burstin, H.; Iezzoni, L. I.; MacKay, E.; and Yu, W. 2000. Using diagnoses to describe populations and predict costs. *Health care financing review*, 21(3): 7.

Authority, F. C. 2019. Machine learning in UK financial services (2019).

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104: 671.

Berk, R.; Berk, D.; and Drougas, D. 2019. *Machine learning risk assessments in criminal justice settings*. Springer.

Bhatore, S.; Mohan, L.; and Reddy, Y. R. 2020. Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, 4(1): 111–138.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3): 199–231.

Brynjolfsson, E.; and Mitchell, T. 2017. What can machine learning do? Workforce implications. *Science*, 358(6370): 1530–1534.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7).

Cumming, R. B.; Knutson, D.; Cameron, B. A.; and Derrick, B. 2002. A comparative analysis of claims-based methods of health risk assessment for commercial populations. *Final report to the Society of Actuaries*.

DeMichele, M.; Baumgartner, P.; Wenger, M.; Barrick, K.; and Comfort, M. 2020. Public safety assessment: Predictive utility and differential prediction by race in Kentucky. *Criminology & Public Policy*, 19(2): 409–431.

Duncan, I. G. 2011. *Healthcare risk adjustment and predictive modeling*. Actex Publications.

Few, S. 2006. *Information dashboard design*. O'reilly Sebastopol, CA.

Garrett, B.; and Monahan, J. 2019. Assessing risk: The use of risk assessment in sentencing. *Judicature*, 103: 42.

Gellman, A.; and Lokem, E. 2014. The statistical crisis in science data-dependent analysis-a 'garden of forking paths'-explains why many statistically significant comparisons don't hold up. *Am. Sci*, 102(460): 460.

Greene, T.; Shmueli, G.; Fell, J.; Lin, C.-F.; and Liu, H.-W. 2022. Forks over knives: Predictive inconsistency in criminal justice algorithmic risk assessment tools. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement_2): S692–S723.

Hand, D. J.; and Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the royal statistical society: series a (statistics in society)*, 160(3): 523–541.

Lee, M. K.; Jain, A.; Cha, H. J.; Ojha, S.; and Kusbit, D. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–26.

Liu, Y.; Kale, A.; Althoff, T.; and Heer, J. 2020. Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, 27(2): 1753–1763.

Metz, A.; Monahan, J.; Garrett, B.; and Siebert, L. 2019. Risk and resources: A qualitative perspective on low-level sentencing in Virginia. *Journal of Community Psychology*, 47(6): 1476–1492.

Monahan, J.; and Skeem, J. L. 2016. Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12(1): 489–513.

Newell, S.; and Marabelli, M. 2015. Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The journal of strategic information systems*, 24(1): 3–14.

Nosek, B. A.; Spies, J. R.; and Motyl, M. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6): 615–631.

Schmidt, P.; and Witte, A. D. 2012. *Predicting recidivism using survival models*. Springer Science & Business Media.

Semenova, L.; Rudin, C.; and Parr, R. 2019. On the Existence of Simpler Machine Learning Models. *arXiv preprint arXiv:1908.01755*.

Siddiqi, N. 2012. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons.

Simmons, J. P.; Nelson, L. D.; and Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11): 1359–1366.

Simonsohn, U.; Simmons, J. P.; and Nelson, L. D. 2020. Specification curve analysis. *Nature human behaviour*, 4(11): 1208–1214.

Simson, J.; Pfisterer, F.; and Kern, C. 2024. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1305–1320.

Thomas, L. C. 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2): 149–172.

Wang, Z. J.; Zhong, C.; Xin, R.; Takagi, T.; Chen, Z.; Chau, D. H.; Rudin, C.; and Seltzer, M. 2022. TimberTrek: exploring and curating sparse decision trees with interactive visualization. In *2022 IEEE visualization and visual analytics (VIS)*, 60–64. IEEE.

Wilke, C. O. 2019. *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media.

Wynand, P.; De Ven, V.; and Ellis, R. P. 2000. Risk adjustment in competitive health plan markets. In *Handbook of health economics*, volume 1, 755–845. Elsevier.