INTERACTIVE MITIGATION OF BIASES IN MACHINE LEARNING MODELS

by

Kelly Van Busum

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Computer and Information Science at IUPUI

Indianapolis, Indiana

August 2024

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Shiaofen Fang, Chair
Department of Computer and Information Science

Dr. Snehasis Mukhopadhyay
Department of Computer and Information Science

Dr. Yuni Xia
Department of Computer and Information Science

Dr. Mihran Tuceryan
Department of Computer and Information Science

Approved by:
Dr. Mihran Tuceryan

For Jamie and Ellie,

so that the walls in our house will sing again

ACKNOWLEDGMENTS

Completing this dissertation has been a dream many years in the making, and I want to express gratitude for everyone who walked with me along the way. Without your encouragement, guidance, and support, reaching this goal would not have been possible.

To Dr. Shiaofen Fang, thank you for encouraging me to finish my Ph.D., serving as my thesis advisor, and for mentoring me both as a colleague and as a student. Your patience, flexibility, expertise, and feedback have helped me grow. Thank you for shaping me as a scholar and guiding me in my career.

To Dr. Snehasis Mukhopadhyay, Dr. Yuni Xia, and Dr. Mihran Tuceryan, thank you for serving on my committee and patiently answering my questions. I appreciate the time you spent helping me develop my research skills.

To Jamie and Ellie Bell, you were my motivation through the late nights. Thank you for your patience, support, and love. I am better because of you.

To my mom, Dr. Brenda Ragle, thank you for showing me it could be done, despite the climb.

To Kristin Van Busum, thank you for being my cheerleader, and for showing me how to live a life where you dream big.

To JCG, thank you for reminding me of my strength, and for sending food on the hard nights.

To Dax Lowery, Shauna Moore, Lori Hart, Dr. Rebecca Upton, Linzi Jones, and the rest of my tribe, thank you for your unwavering encouragement, unconditional love, and support.

To Anna Vadella, thank you for your work on the user interface.

To the Butler CSSE Department, especially Dr. Ankur Gupta, Dr. Jon Sorenson, and Dr. Ryan Rybarczyk, thank you for encouraging me, mentoring me, and helping to protect my time during my first year at Butler.

To Dr. John Gersting and Dr. Judy Gersting, thank you for your wisdom, time, and patience.

To Dr. Gavriil Tsechpenakis, thank you for your bits of wisdom, like telling me to solve the small problems to solve the big problems, and teaching me to think like a researcher.

To the IUPUI CSCI Department, for supporting my career and encouraging me as I worked on this degree.

I'd also like to thank the following people at IUPUI for their help in providing the dataset used in this study, for answering questions about the dataset format, and for explaining policies related to admissions: Jane Williams, Joe Thompson, Steve Graunke, Matt Moody, Norma Fewell, and Lori Hart.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Bias and fairness issues in artificial intelligence algorithms are major concerns as people do not want to use AI software they cannot trust. This work uses college admissions data as a case study to develop methodology to define and detect bias, and then introduces a new method for interactive bias mitigation.

Admissions data spanning six years was used to create machine learning-based predictive models to determine whether a given student would be directly admitted into the School of Science under various scenarios at a large urban research university. During this time, submission of standardized test scores as part of a student's application became optional which led to interesting questions about the impact of standardized test scores on admission decisions. We developed and analyzed predictive models to understand which variables are important in admissions decisions, and how the decision to exclude test scores affects the demographics of the students who are admitted.

Then, using a variety of bias and fairness metrics, we analyzed these predictive models to detect biases the models may carry with respect to three variables chosen to represent sensitive populations: gender, race, and whether a student was the first in his/her family to attend college. We found that high accuracy rates can mask underlying algorithmic bias towards these sensitive groups.

Finally, we describe our method for bias mitigation which uses a combination of machine learning and user interaction. Because bias is intrinsically a subjective and context-dependent matter, it requires human input and feedback. Our approach allows the user to iteratively and incrementally adjust bias and fairness metrics to change the training dataset for an AI model to make the model more fair. This interactive bias mitigation approach was then used to successfully decrease the biases in three AI models in the context of undergraduate student admissions.

CHAPTER 1. INTRODUCTION

Artificial Intelligence (AI) has started to play an increasingly important role in almost every part of society. It is used in health care recommendations, hiring and promotion decisions, the criminal justice system, customer service, education, and general everyday life. There are certainly many benefits to the use of AI – decisions can often be made more rapidly, more objectively, more consistently, and lead to a greater understanding of data. However, the use of AI also has the potential to create harm. AI can reflect human bias, potentially making biased decisions faster, and many times these decisions disproportionally hurt marginalized groups. Even worse, in its attempt to optimize decision making, AI can introduce new biases, above and beyond existing human biases.

Defining and detecting bias in AI is a difficult problem because bias is a human-defined concept, and any attempt to capture it purely by objective mathematical functions may, at best, be incomplete. Different people may have different notions of bias, and attempts must be made to address these diverse human viewpoints to achieve community acceptance and trust in machine learning solutions. Furthermore, the concept of fairness is broader than bias, and is subjective and context dependent. Decisions can be biased but considered fair, or unbiased but considered unfair. Even if there is agreement that decisions are unfair, there can be disagreement about why and what to do to fix it, because it is often difficult to achieve a balance among different fairness and bias metrics as they are sometimes interdependent and even contradictory. Therefore, there is a strong need for human input as part of an AI system to achieve desired trade-offs and compromises.

Complicating things, AI is often considered a "black box," where data goes in and recommendations come out, but the innerworkings of the AI are too mathematically complex to be visualized or understood, especially by people without a technical background in AI development. But, since bias is context-dependent, AI systems need human input from people with domain-specific knowledge, who many times are not the people with technical expertise. As AI becomes ubiquitous, it is therefore critical to develop techniques to allow non-expert stakeholders the ability to select context-specific bias metrics, and to adjust these bias metrics incrementally and iteratively. As these adjustments are made, it becomes possible to observe how

changing one metric affects the other metrics allowing stakeholders to agree on acceptable trade-offs, the motivation for this research.

This thesis describes a project where an existing university admissions dataset was analyzed as a case study. Students who attend large urban universities, such as the one where this data was collected, have some unique challenges. They may be more likely than students who attend other types of universities to face the competing demands of work, family, and school leading to poorer academic experiences. Attending a university in a city with a high cost-of-living and the continuing effects of COVID-19 are both factors that can increase the financial demands on students, also leading to increased stress and poorer academic outcomes. Although urban universities tend to be more diverse, minority students still report struggling with a sense of belonging, which can negatively affect academic performance.

In an attempt to level the playing field for disadvantaged students, many universities are experimenting with an admissions policy that no longer requires students to submit standardized test scores. These test-optional policies were designed to address the concern that standardized test scores are biased metrics for predicting student success and to increase equity in admissions procedures, but more research needs to be done to fully understand how these policies might change the demographics of the students admitted to the university and other impacts. Such fundamental changes in admission policies can significantly affect our higher education system for different populations, so to best support all students it is important to have a strong understanding of the implications of such policy changes. Analyzing the behavior and features of AI-based predictive models built from existing datasets such as admissions data can provide a powerful opportunity to better understand the impact of policy changes such as this in the U.S. higher education system. Identifying the important factors in admissions decisions and how test-optional policies might change admitted student demographics is one goal of this project [1].

When AI models are trained using existing datasets, the models can introduce new biases unintentionally increasing inequity. It is critical to have techniques for carefully examining predictive models for evidence of bias as AI becomes more widely used in higher education. Defining and detecting bias in AI models built to predict admission decisions is another goal of this project [2]

If evidence of bias in predictive models is found and deemed to be unfair, it is important to have a procedure for effectively mitigating the bias. Because bias and fairness statistics

interact and can conflict, humans must be part of the bias mitigation process so that the decisions made by the models reflect the values of human decision-makers. This motivates the third goal of this project, the development of an interactive approach for bias mitigation [3].

1.1 Overview of Project

Admissions data from the School of Science at a large urban research university was used to create and analyze machine learning-based AI models. These models predict whether a student would be directly admitted into the School of Science, or not, under a variety of scenarios. The dataset spans six years, and over this time, the admissions policy of the university changed from requiring students to submit standardized test scores as part of their application, to making test scores optional.

This project focuses on answering the following research questions:

1. Which variables are important in admissions decisions? How does excluding test scores affect who is admitted?

2. Is there bias in the machine learning model toward sensitive groups? How can this bias be defined and detected?

3. If evidence of bias is found, how do we mitigate this bias, given that fairness is subjective and context-dependent?

We begin by constructing several AI predictive models, including one that focuses on students admitted when test scores were required (the "Test-Required Cohort") and one that focuses on students admitted after submitting test scores became optional (the "Test-Optional Cohort"). The predictive models contain a variety of demographic variables, and three variables were chosen to represent sensitive populations (the "sensitive variables"): Gender, Race, and First-Generation, whether a student was the first in his/her family to attend college. We then conducted an exploratory analysis to better understand the factors that are most important in admissions decisions, and how the change from a test-required to a test-optional policy affected the demographics of the students who were admitted.

Next, we define some bias metrics and use these to carefully evaluate the AI models for presence of potential biases with respect to performance relative to these three sensitive variables. The result of this analysis provides some evidence that AI algorithms can be harmful when used as part of the admission decision-making process if bias is not effectively mitigated.

To mitigate this bias, we present a human-in-the-loop method that uses a second machine learning model. A user can evaluate bias and fairness statistics on the results given by an "Admissions Model", a machine learning model which predicts admissions decisions. If the model is determined to be biased in a way that is unfair, the user can specify adjustments to be made to these statistics which would make the model more fair. A second machine learning model, the "Bias Mitigation Model," takes these user-adjusted statistics, and predicts the adjustments needed to the training set to be used for constructing an adjusted Admissions Model. This adjusted Admissions Model is then built and reevaluated based on the updated bias and fairness statistics. This interactive adjustment process continues until the user is satisfied with the results. Finally, we demonstrate and evaluate this bias mitigation method using three scenarios.

1.2 Contributions

This work makes the following contributions:

1. College admissions data was used to generate AI predictive models, evaluated on their overall accuracy and effectiveness. These models were used to better understand the primary factors in admissions decisions and their variations within different cohorts.

2. The models were highly accurate overall, but a thorough analysis uncovers evidence of bias with respect to sensitive populations. This serves as a warning and contributes to an understanding of how bias can be defined and detected.

3. Fairness metrics were included in the analysis of the models, illustrating some limitations with the use of these metrics.

4. A novel approach to bias mitigation is introduced and evaluated which uses a combination of machine learning and user interaction. This work demonstrates how carefully designed interactive adjustments to the training sets used in the construction of an AI predictive model can effectively mitigate bias.

5. Recognizing that the concept of fairness is subjective, the bias mitigation method presented here allows the user to interactively adjust various bias and fairness metrics used to construct the AI model to create a model that is fair within the context in which it is being used. To our knowledge, data-driven human-in-the-loop techniques have not been used for algorithmic bias mitigation in AI models.

1.3 Organization

This thesis is organized as follows. In CHAPTER 2, we provide an overview of the literature related to this work and its impact. CHAPTER 3 describes the admissions dataset used in this project, the results of an exploratory analysis, and the effects of a test-optional policy on admissions demographics. CHAPTER 4 defines the bias metrics and methodology we used to analyze the admissions dataset, presents evidence of bias, and introduces fairness metrics. In CHAPTER 5 we describe our interactive bias mitigation method. In CHAPTER 6, we demonstrate our bias mitigation method using three scenarios. In CHAPTER 7, we provide some concluding remarks.

6.5 User Interface

A user interface was created to allow the user to view bias and fairness metrics and make interactive adjustments to the Training dataset. The next few pages provide screenshots of this interface using Scenario 1, Race as described in Section 6.2. Note that the same user adjustments can lead to different results because of the randomness inherent in choosing data from the Reserves dataset to adjust the Training dataset.

To begin, the user must select a sensitive variable, and in this scenario, Race was chosen, as shown in Figure 6.19.

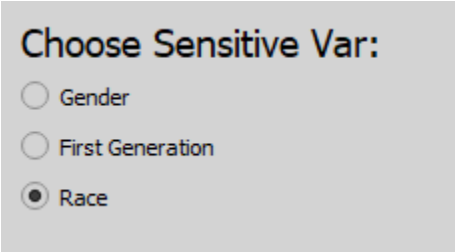

Figure 6.19. User First Chooses a Sensitive Variable

The interface displays the bias and fairness statistics for the Baseline dataset at Time=0, its initial state, as shown in Figure 6.20.

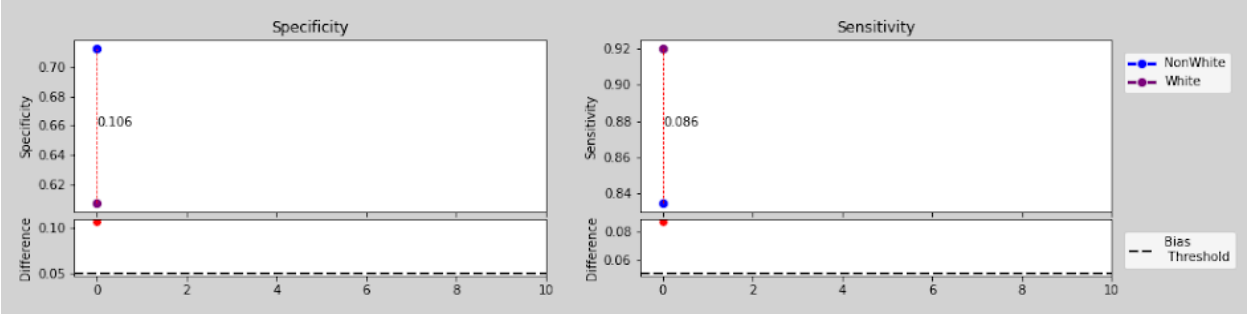

Figure 6.20. Evidence of Bias

The user sees that there is Specificity bias and Sensitivity bias present in the model and wants to decrease each of these by .01 in Time=1. The user enters -1 for all other metrics to indicate that the algorithm should estimate these. The estimation is done by finding the closest datapoint (using Euclidian distance) to the user's desired changes in the Training dataset for the Bias Mitigation Model and using the values of those metrics. This portion of the interface is shown below in Figure 6.21.

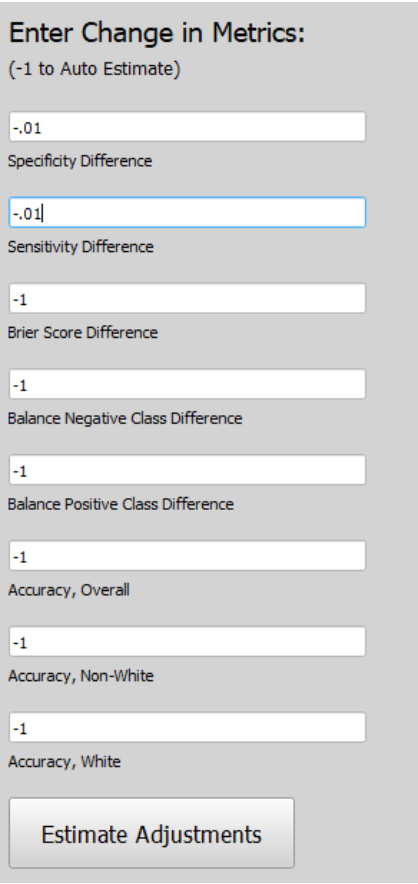

Figure 6.21. User Enters Desired Changes in Metrics

The user clicks "Estimate Adjustments" which sends the user's desired values for the bias and fairness statistics as input to the Bias Mitigation Model.

The Bias Mitigation Model returns the Estimated Adjustments to the Training Set, which are displayed for the user to review, as shown in Figure 6.22. The adjustments are the predicted changes (in percent) needed for each demographic group in the Training dataset to change the bias and fairness metrics as the user indicated. The user should review these changes to see if they are reasonable, and if they are not, should choose different modifications to the bias and fairness statistics. Recall that the Bias Mitigation Model was trained on changes from 0-40% for each demographic group, so adjustment values below 0 or much above 40 may indicate decreased accuracy. Also, note that data is never removed from the Training dataset, so when a variable's adjustment value is below 0, the Training dataset remains unmodified with respect to that variable.

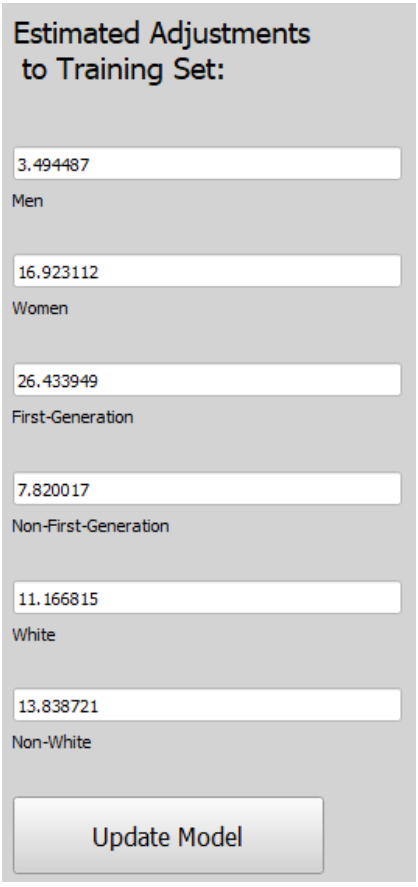

Figure 6.22. User Reviews Training Set Adjustments

When the user is satisfied with the results, he can click "Update Model" which then adjusts the Training set, retrains the Admissions Model, and graphs the resulting bias and

fairness statistics. The user can then adjust the metrics again, based on these new results. This iterative process is shown in Figure 6.23, Figure 6.24, Figure 6.25, Figure 6.26, Figure 6.27, Figure 6.28, Figure 6.29, and Figure 6.30.

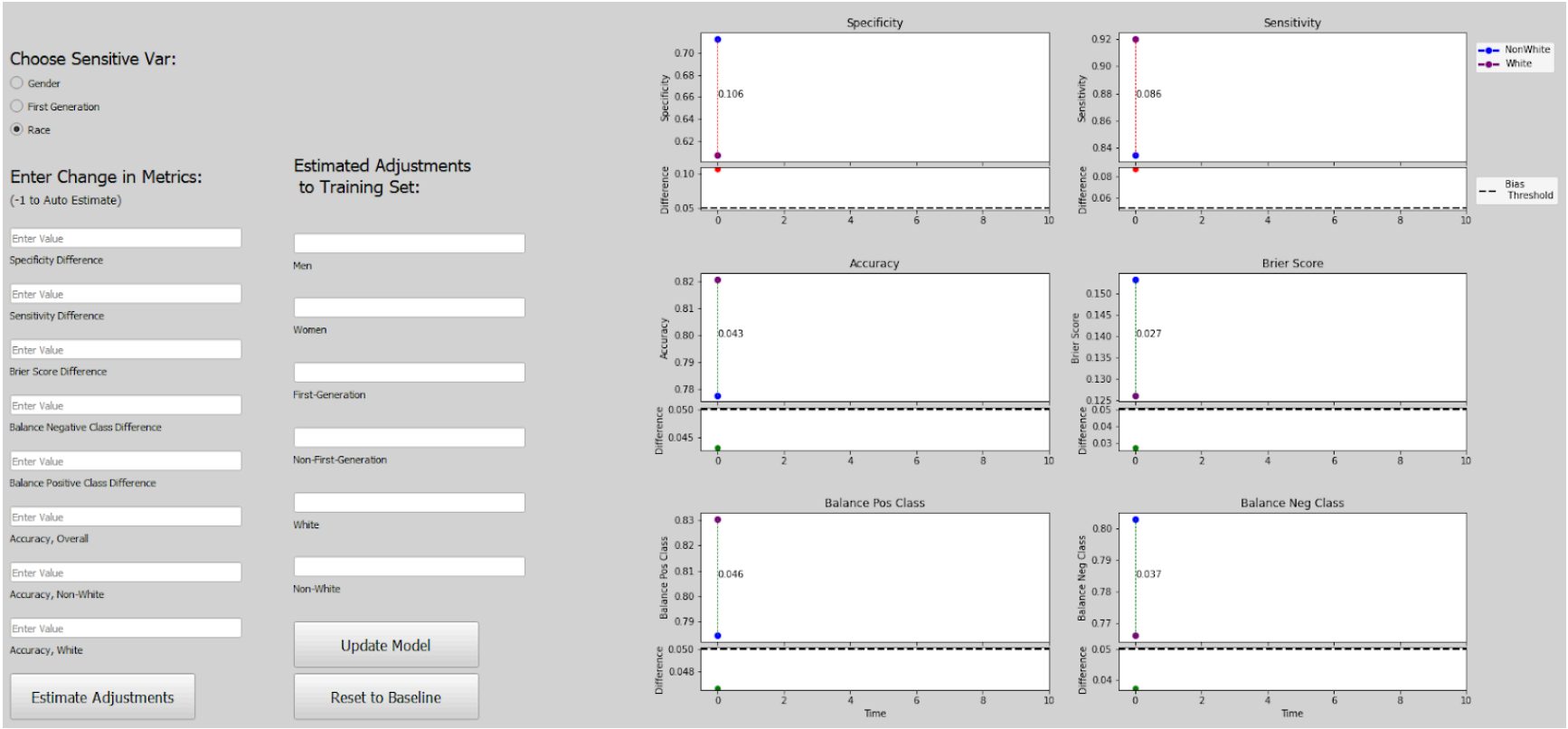

Figure 6.23. User Interface, Time=0, Bias in Baseline Dataset Displayed

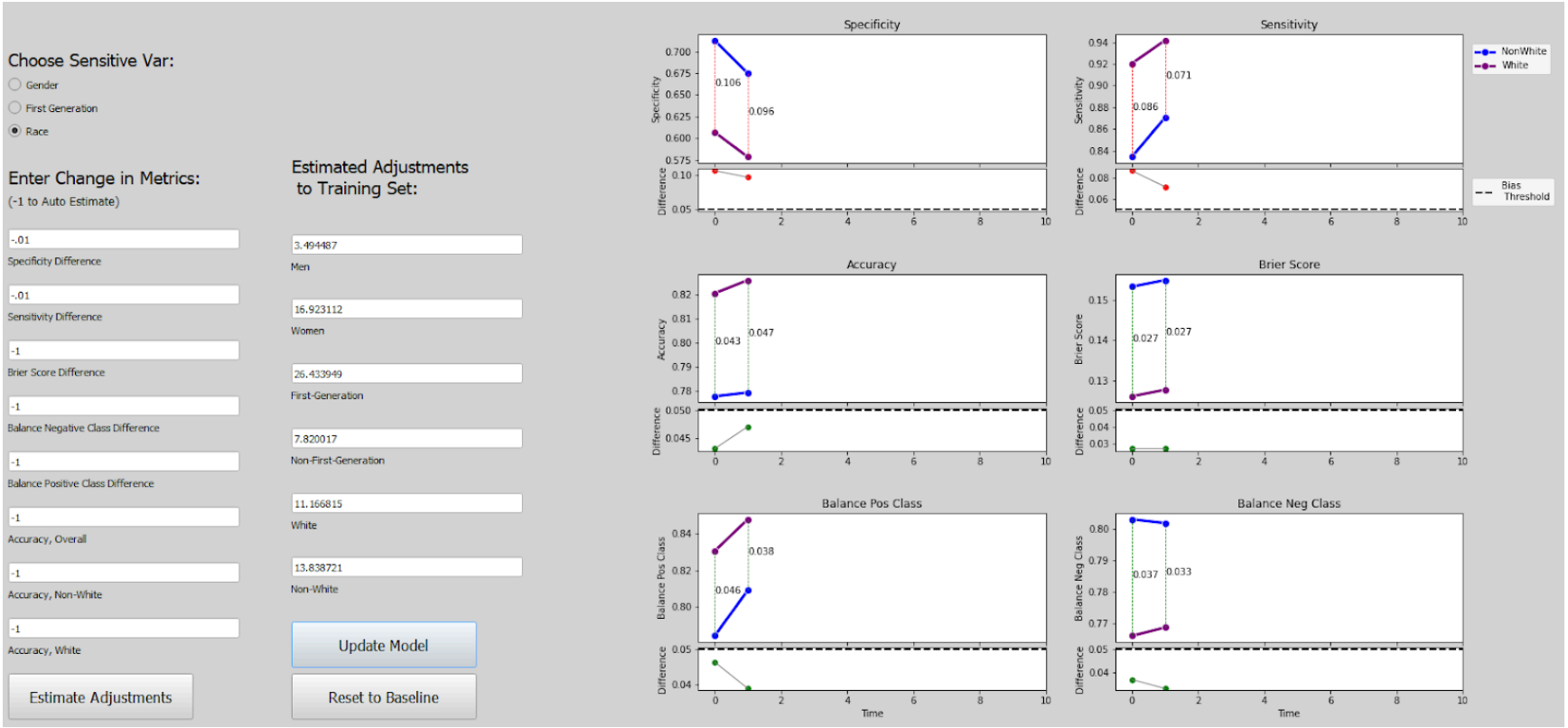

Figure 6.24. User Interface, Time=1

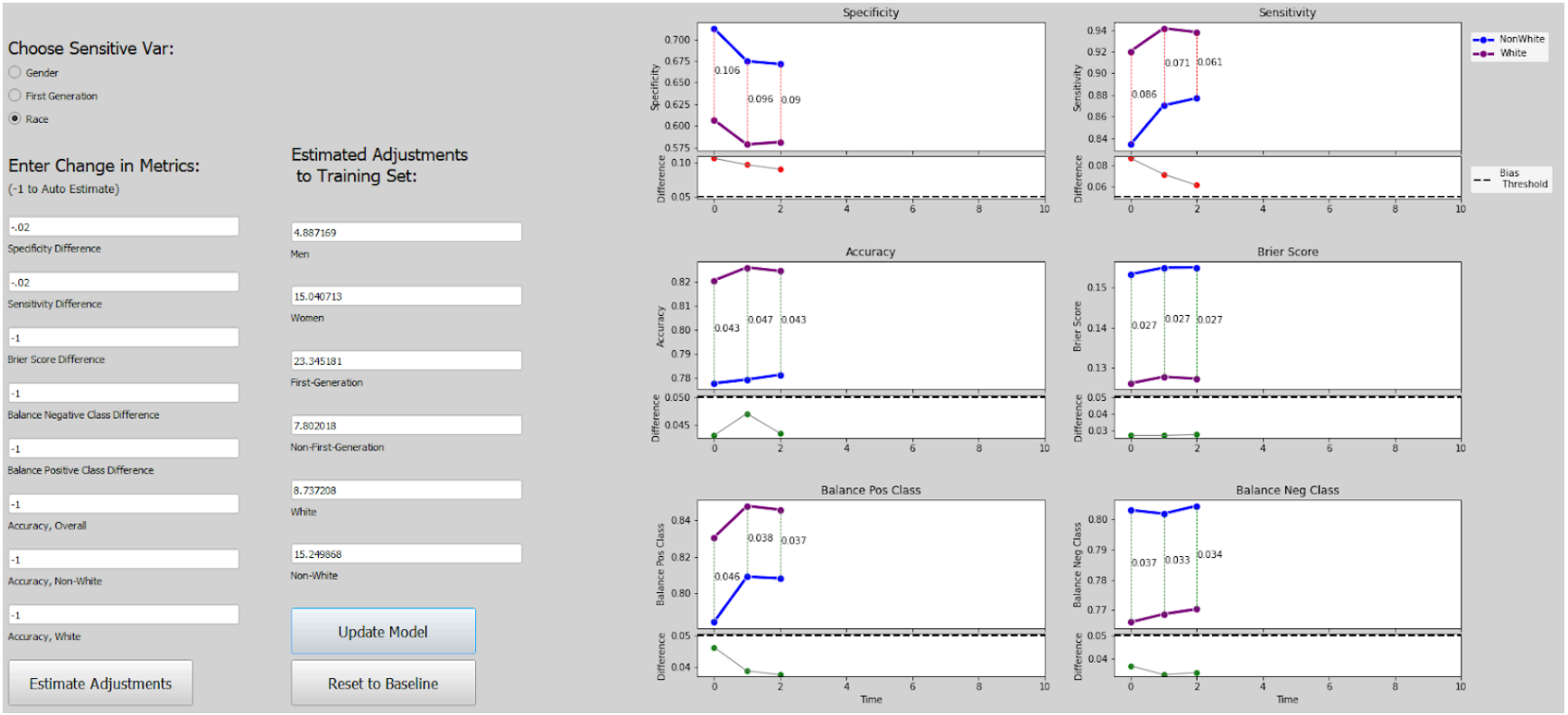

Figure 6.25. User Interface, Time=2

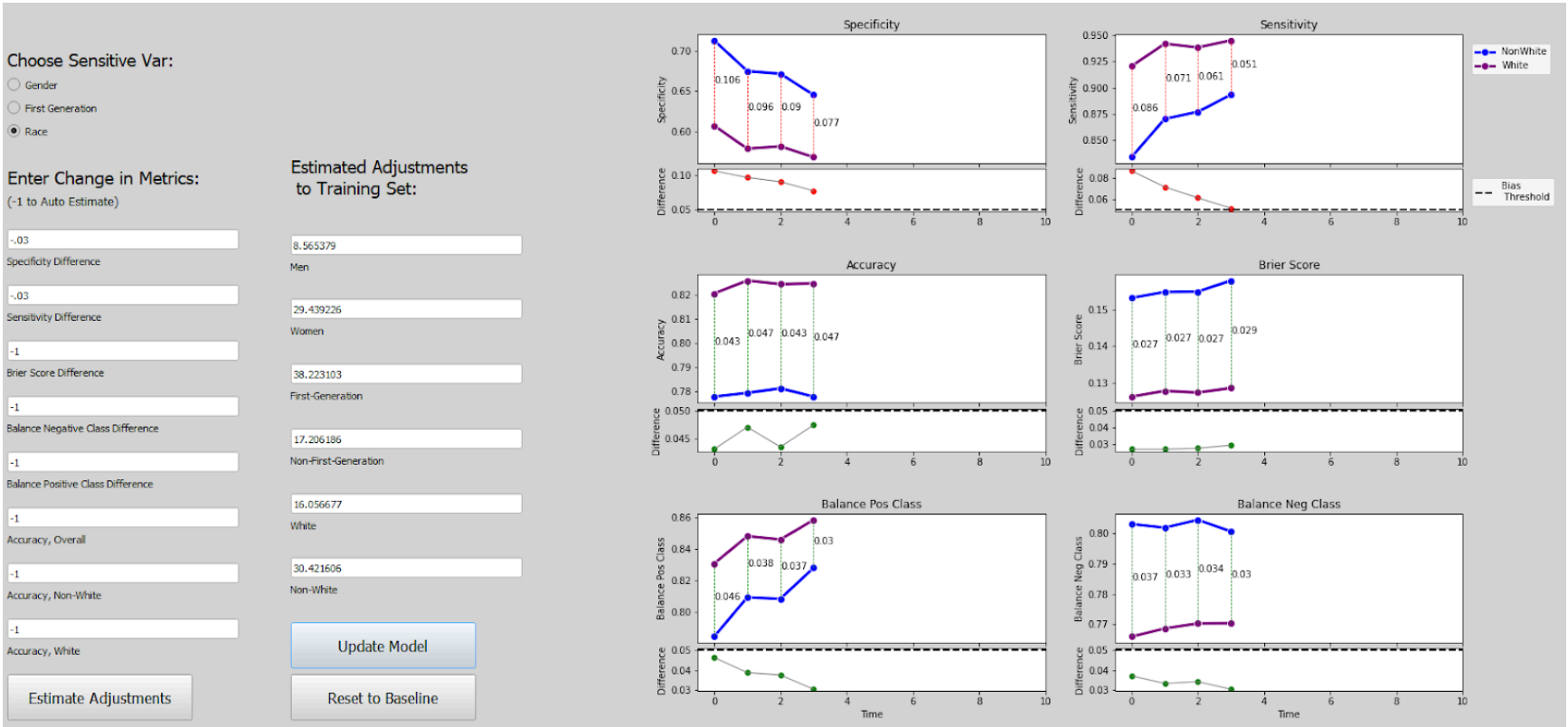

Figure 6.26. User Interface, Time=3

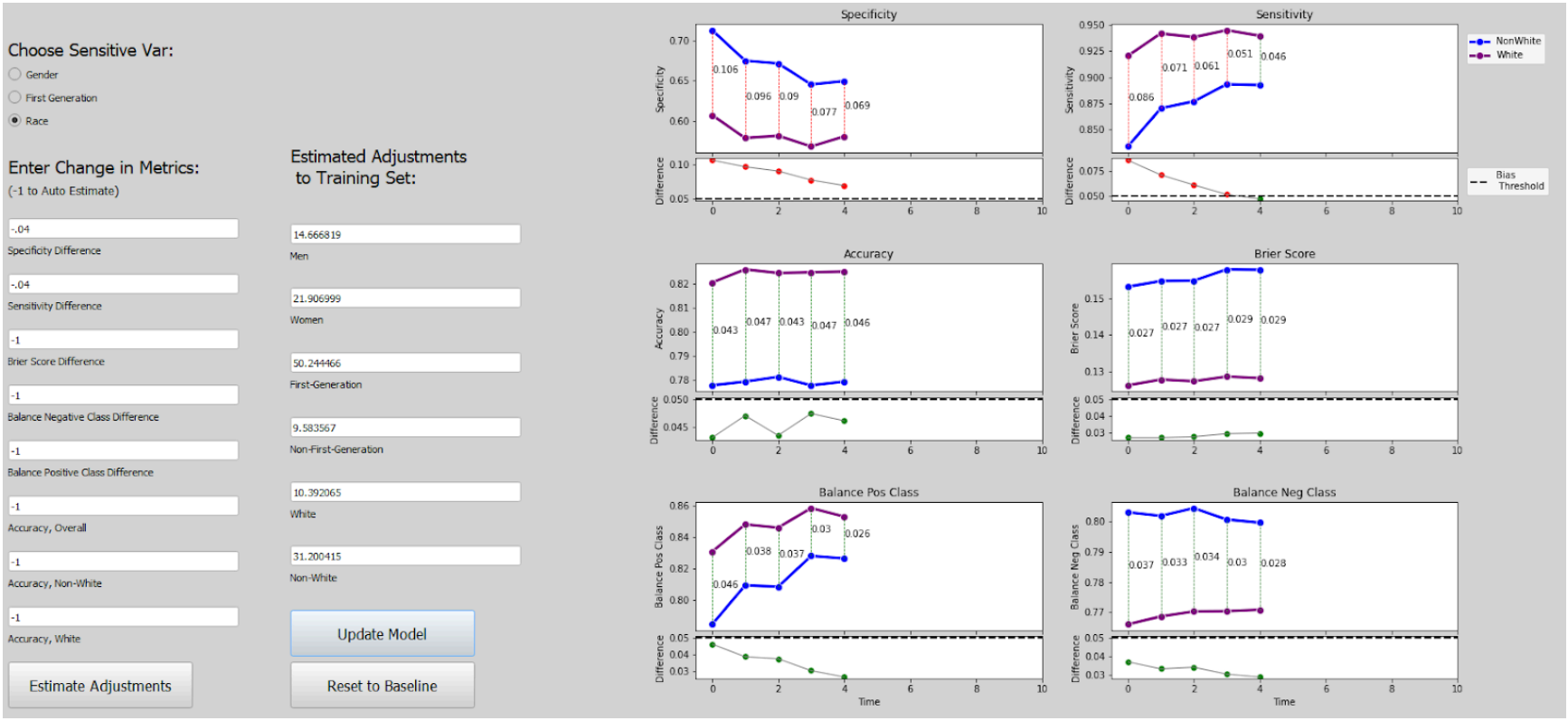

Figure 6.27. User Interface, Time=4

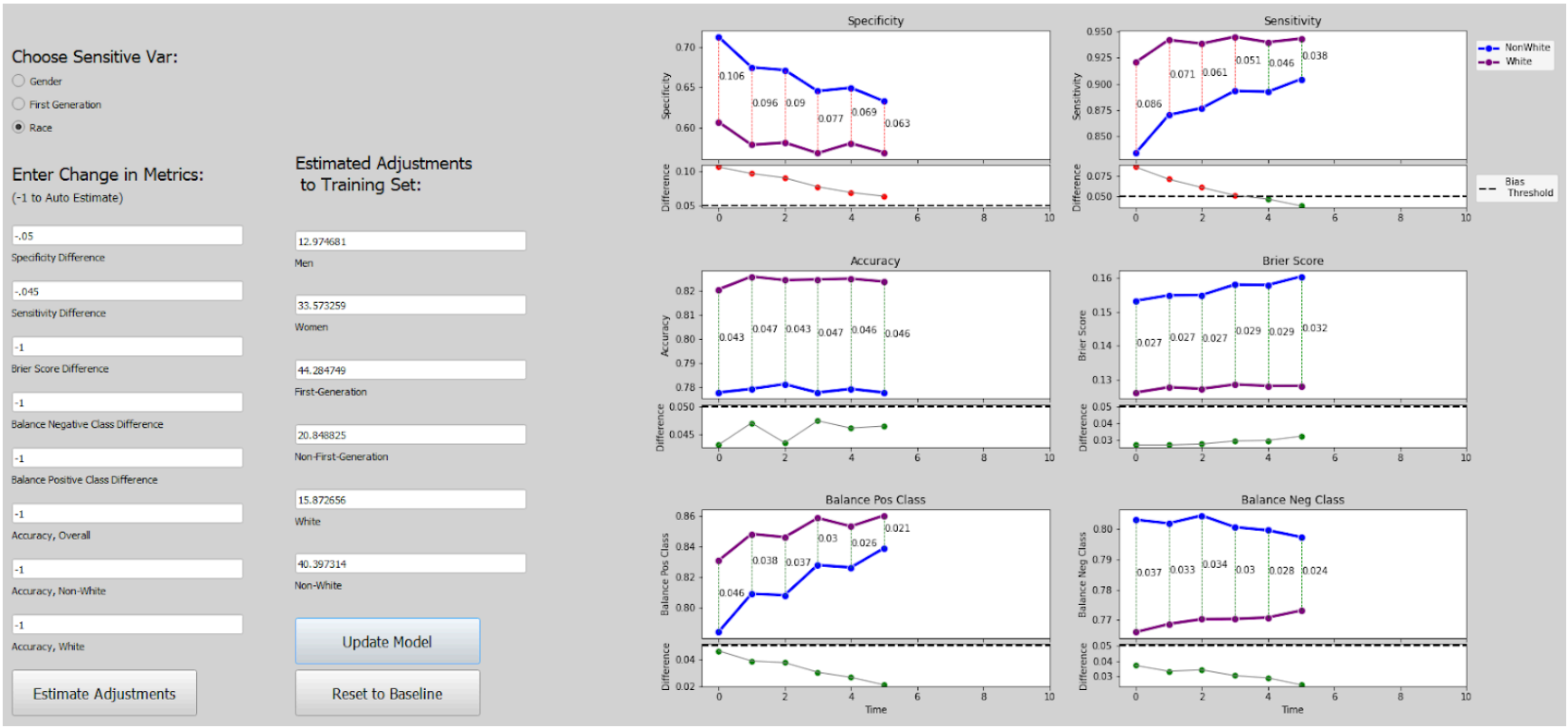

Figure 6.28. User Interface, Time=5

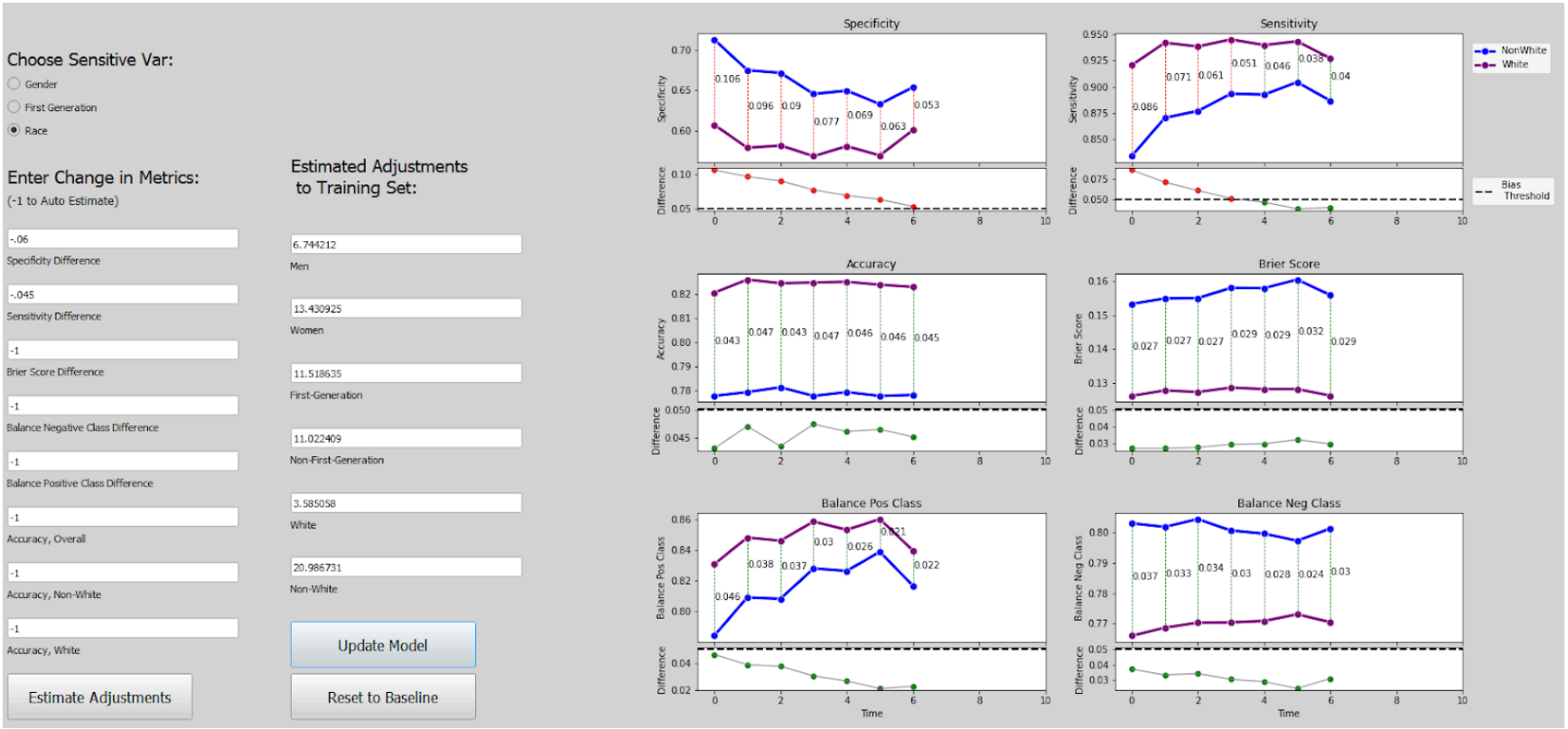

Figure 6.29. User Interface, Time=6

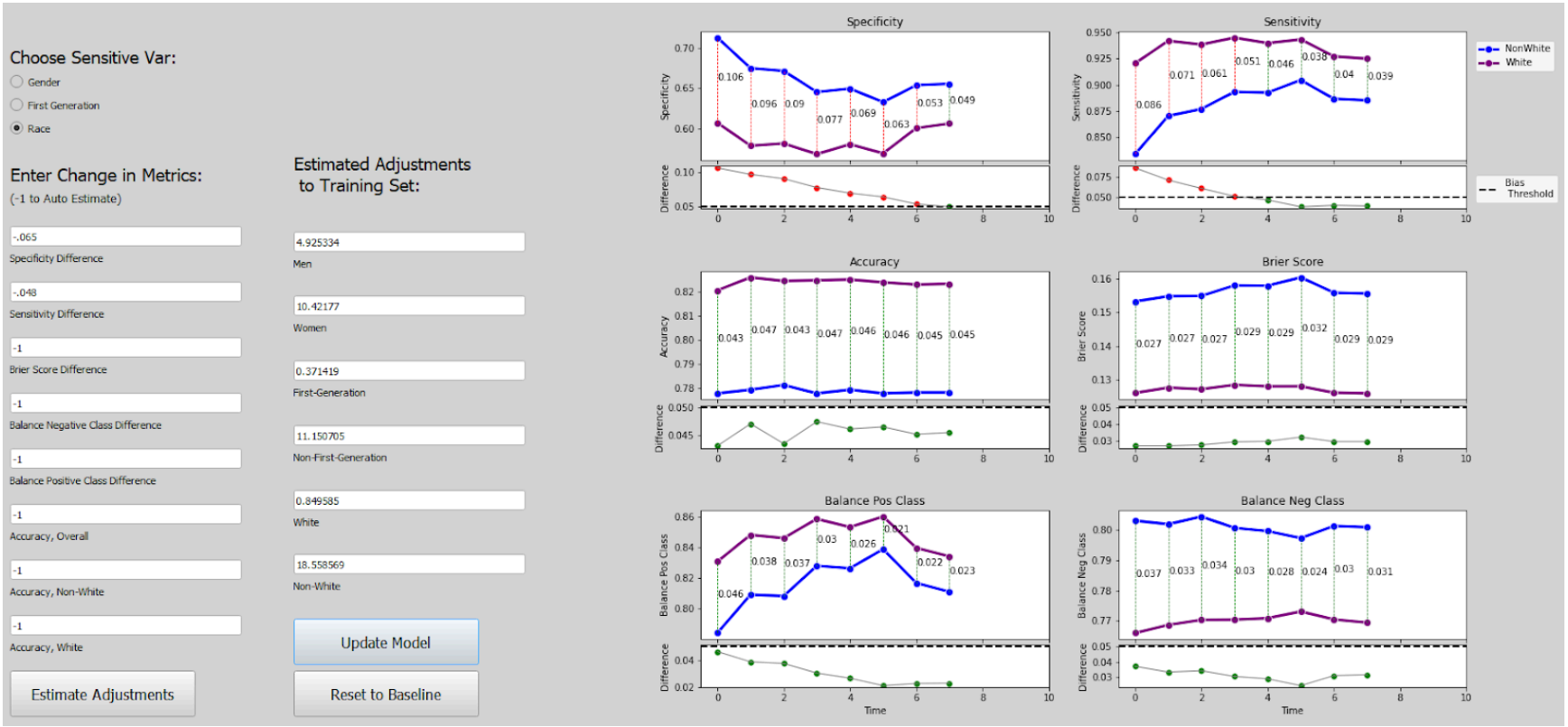

Figure 6.30. User Interface, Time=7