Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models

by

Yunhan Zheng

Bachelor of Management and Bachelor of Economics, Peking University (2018)

Submitted to the Department of Urban Studies and Planning in partial fulfillment of the requirements for the degree of

Master in City Planning
and
Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Urban Studies and Planning
May 18, 2021

Certified by .....
Jinhua Zhao
Associate Professor, Department of Urban Studies and Planning
Thesis Supervisor

Certified by .....
Shenhao Wang
Research Scientist, Department of Urban Studies and Planning
Thesis Supervisor

Accepted by .....
Ceasar McDowell
Professor of the Practice Chair, MCP Committee
Department of Urban Studies and Planning

Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models

by

Yunhan Zheng

Submitted to the Department of Urban Studies and Planning
on May 18, 2021, in partial fulfillment of the
requirements for the degree of
Master in City Planning
and
Master of Science in Transportation

Abstract

Although researchers increasingly adopt machine learning to model travel behavior, they predominantly focus on prediction accuracy, while largely ignore the ethical challenges and the adverse social impacts embedded in the machine learning algorithms. This study introduces the important missing dimension - computational fairness - to travel behavioral analysis. It highlights the accuracy-fairness tradeoff instead of the single dimensional focus on prediction accuracy in the contexts of deep neural network (DNN) and discrete choice models (DCM). The author firstly operationalizes computational fairness by *equality of opportunity*, then differentiates between the bias inherent in data and the bias introduced by modeling. The models inheriting the inherent biases can risk perpetuating the existing inequality in the data structure, and the biases in modeling can further exacerbate it. The author then demonstrates the prediction disparities in travel behavioral modeling using the National Household Travel Survey 2017. Empirically, DNN and DCM reveal consistent prediction disparities across multiple social groups, although DNN can outperform DCM in prediction disparities because of DNN's smaller misspecification error. To mitigate prediction disparities, this study introduces an absolute correlation regularization method, which is evaluated with the synthetic and the real-world data. The results demonstrate the prevalence of prediction disparity in travel behavior modeling, which can exacerbate social inequity if the prediction results without fairness adjustment are used for transportation policy making. As such, the author advocates for careful considerations of the fairness problem in travel behavior modeling, and the use of bias mitigation algorithms for fair transport decisions.

Thesis Supervisor: Jinhua Zhao
Title: Associate Professor, Department of Urban Studies and Planning

Thesis Supervisor: Shenhao Wang
Title: Research Scientist, Department of Urban Studies and Planning

Acknowledgments

It would not be possible to complete this thesis without the help and support of the kind people around me, and I would like to express my sincere gratitude to them.

Above all, I owe my deepest gratitude to Professor Jinhua Zhao, who provided me constant support, guidance and patience throughout my past three years at MIT. Thank you, Jinhua, for imparting your knowledge, expertise and innovative research ideas during our conversations, and for always reminding me to connect my research with real-world impacts and applications.

I am also grateful to Shenhao Wang, who encouraged me to dive into this research area. His unwavering enthusiasm for exploring the potential of machine learning in travel behavior analysis kept me constantly engaged with my research, and his suggestions on many technical details helped improve this thesis a lot.

My sincere thank also goes to my fellow labmates in JTL-Transit lab. Thank you, Baichuan Mo, Qingyi Wang, Xiaotong Guo and Rachel Luo, for helping me develop various research ideas and for being my supportive friends. Thank you, Joanna Moody and Hui Kong, for providing invaluable advice and assistance during our research collaborations. I also want to thank all other members in JTL-Transit lab for sharing their knowledge and for sustaining such a warm and collegial lab atmosphere.

Finally, I would like to express my heartfelt gratitude to my parents, for their continuous love, caring and support, and my boyfriend, for being with me at every step of this incredible journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent years, an increasing literature has adopted machine learning models, particularly deep neural networks (DNNs), to predict travel behavior. The common practice of machine learning is to identify the best model by fitting the training data and being evaluated on the test data, with the objectives of performance optimization and output prediction in various scenarios [70, 22]. Comparing DNN with the traditional logit models, previous studies have shown that DNN has higher predictive power and typically makes fewer mistakes in predictions compared to the multinomial logit models (MNL) [38, 55]. Specifically, DNN is powerful owing to factors such as the relaxation of linear relationships among variables [83], automatic feature learning [83] and the ability to accommodate various data structures [59, 43].

However, machine learning also poses tremendous ethical challenges. Many studies found that the machine learning models can produce much worse prediction results for disadvantaged groups such as black people, female and low-income population, leading to unfair treatment on these populations. For example, the software based on machine learning algorithms to predict future criminals is biased against blacks [5]. Research on online advertisement showed that ads (e.g. people smart ads, public records ads) for public records on a person appears disproportionately higher for black-identifying names [78]. Literature concerning text classification demonstrated gender bias in word embeddings trained on Google News, which systematically associate men with computer programmers and women with homemakers [15]. Although the fairness problem in machine learning has been revealed across a large number of contexts, thus far no study has examined the computational fairness issue in

travel behavior modeling. In fact, fairness has been an enduring topic in the transportation field [73, 63, 66]. The traditional transport fairness research either adopts a highly qualitative approach or presents quantitative metrics without being integrated with algorithms. In this computational era, it is critical to demonstrate the risks of naively adopting models without fairness concerns and showcase the integration of fairness metrics into modeling practices and policy decisions.

Motivated by these research gaps, this research investigates how to measure, evaluate, and mitigate prediction disparities of travel behavior models regarding the protected variables - race, gender, income, medical condition and region. The author takes the following three steps. First, the author introduces *equality of opportunity* to measure computational fairness in travel behavioral modeling based on the fairness theory in the machine learning research. Second, the author identifies and measures the prediction disparities in travel behavior modeling using binary logistic regression (BLR) and DNN. Third, building upon the approach by Beutel et al. [11], the author adopts an absolute correlation regularization method to mitigate the prediction biases, and evaluate the performance of the new model with bias mitigation. The second and the third steps are deployed on both synthetic datasets and the NHTS dataset. Experiments conducted on the synthetic datasets show how the prediction disparity varies with data structure, number of predictors, sample size and the degree of model misspecification. The fairness analysis is then deployed on the NHTS 2017 dataset which has wide coverage of geographic areas and populations with different characteristics, as well as the large sample size and the diversity of input features.

Prediction disparities could bias transport policy decisions unfavorably against socially disadvantaged groups, such as low-income and ethnic minority populations. For example, when the estimation of African-American communities' mobility needs for transit has a higher error rate, the transit agencies would make more mistakes in considering adding more bus routes and investing in new transit stations for the minority neighborhoods. Echoing Title VI, the DOT regulations at 49 CFR Part 21 are designed to ensure that "no person in the United States, based on race, color, or national origin, is excluded from participation in, denied the benefits of, or otherwise subjected to discrimination under any program that DOT financially assists" [79]. Therefore, people from different groups of interest (such as

race, gender, income) should be treated fairly by the models and algorithms which have been widely deployed to inform transportation project planning and policy making.

The paper is organized as follows: the next section reviews travel behavior modeling and fairness in machine learning. Section 3 introduces data and models, where I introduce the methods of measuring and mitigating prediction disparities, and also illustrate how the methods are tied to the fairness theories in machine learning. Section 4 presents the results, which include the simulation experiments on the synthetic data, the quantification of prediction disparity across various dependent variables and protected variables in the NHTS data, as well as the results of bias mitigation for both the synthetic and the NHTS data. Section 5 and 6 summarizes the key findings and discusses the future research directions.

Chapter 2

Literature Review

Discrete choice modeling has been used in numerous studies in the transportation field regarding people's travel behavior predictions. With the development of the artificial intelligence, machine learning techniques such as DNN have become increasingly popular and been adopted widely to achieve higher prediction accuracy [28, 82, 4]. Researchers adopted DNN to predict travel mode choices [85, 19], car ownership [70], traffic accidents [91], travellers' decision rules [80], driving behaviors [50], traffic flows [71, 86] and many other decisions in the transportation field. Some researchers focused on comparing the performance of various machine learning models such as DCM, DNN, SVM, random forest on travel behavior prediction [23, 84, 69, 56, 77]. However, nearly all of these studies evaluate the performance of different models in terms of accuracy, stability and interpretability.

The previous machine learning literature has presented a great body of evidence showing that a model can act discriminatorily on one population in a variety of settings including, but not limited to, criminal risk assessment [5, 24], clinical care [72, 44], online advertisement delivery [30, 78], text classifications [34, 15] and credit risk evaluation [31, 61]. Approaches have been taken to understand and address unfairness in machine-learned models, which include formalizations of fairness in machine learning [26, 41], designing fairness-enhancing algorithm [40, 53, 89] and solving the fairness concerns in real-world industries [58, 47]. In contrast to the rich literature in computer science focusing on fairness analysis, no study has investigated the fairness issue in travel behavior prediction.

On the other hand, fairness has long been a crucial concern in transportation planning,

and there has been numerous studies conducting transport equity analysis [13, 42, 62, 32]. Among the variety of equity-focused transportation studies, most of them carried out fairness discussions in terms of how the subject matter of planning such as transportation infrastructure, mobility services, opportunities and rights are distributed among the population [73, 63]. Research in this domain usually involves the cost and benefit analysis of specific groups or individuals concerning a specific transportation project [63]. The analysis of cost may include transportation cost [66] and environmental cost [76], such as the air and noise pollution produced by transportation-related construction. The benefit analysis focuses on the benefits people receive in terms of accessibility, mobility and economic vitality [13, 67]. In these studies, fairness is usually evaluated based on the distributions of costs and benefits among different demographics, neglecting the fact that the machine-learned models deployed to estimate travel demand - which is critical for costs and benefits analysis - itself can be unfair [8, 41]. Therefore, instead of focusing on substantive fairness which emphasizes resource allocation and decision making, the author takes a step back and examines bias exhibited in the model estimation results. As such, this study aims to enrich the fairness discussion by focusing on the machine learned models, investigating the fairness issue that comes up in the modeling process and analyzing the biases that arise in the modeling results.

The Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) community states the fairness principle as "Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms." [1] This principle is applied in this research and the detailed definitions of fairness are described below. The word "bias" is used to refer to the systematic favoring of one group over another by the algorithm in this research.

2.1 Fairness Notion

Table 2.1 summarizes five core fairness notions that are most widely known in the computer science literature: equality of opportunity, equality of odds, demographic disparity, fairness through unawareness and counterfactual fairness. These five fairness criteria either belong to the disparate impact analysis or the disparate treatment analysis.

Table 2.1: Different Types of Fairness Criteria

Type of Analysis	Fairness Criterion
Disparate impact analysis	Equality of opportunity
	Equality of odds
	Demographic parity
Disparate treatment Analysis	Fairness through unawareness
	Counterfactual fairness

Among a variety of fairness notions, this paper uses equality of opportunity to define fairness. Equality of opportunity is a relaxation of the fairness measure equality of odds, and equality of odds states that the protected and unprotected groups should have equal rates for true positives and false positives [68, 49]. Since in practice achieving equal rates for both measures (true positives and false positives) is usually hard, equality of opportunity is adopted instead in many cases stating that the protected and unprotected groups should have equal true positive rates [68, 49]. A similar fairness definition - "reciprocal equality of opportunity" - is also adopted in this study, which requires the protected and unprotected groups having equal true negative rates [12].

Equality of opportunity is a type of "disparate impact" analysis which evaluates fairness based on model impact (results) - specifically whether policies or practices have a disproportionately adverse impact on protected classes [8]. This fairness notion is chosen as it is inherently connected with the notion of equality of opportunity in the traditional transportation equity literature. In the traditional transport equity literature, equality of opportunity focuses on the applications and resource allocations. It asserts that the education, employment and consumer opportunities accessible to residents should be equal between different groups [63]. The violation of equality of opportunity in the travel behavior predictions can consequently affect the transportation resources different populations can get, thus will perpetuate inequality of opportunity in reality.

In addition to equality of opportunity, another widely-adopted fairness measure focusing on disparate impact is "demographic parity" [68, 49], which is achieved when the likelihood of a positive outcome is the same regardless of whether the person is in the protected group. For example, when studying gender fairness in predicting the usage of public transit, the demo-

graphic parity is achieved when the proportion of people predicted as using public transit frequently is the same between male and female. Equality of opportunity is preferred to demographic parity since the latter fails to account for discrimination which is explainable in terms of legitimate grounds [14].

Apart from disparate impact analysis, another strand of research - called "disparate treatment" analysis - evaluates fairness in terms of treatment rather than modeling result to see if the decisions are made (partly) based on the subject's sensitive attribute information [88]. This type of fairness includes "fairness through unawareness" and "counterfactual fairness". In the definition of fairness through unawareness, an algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process [45, 36]. On the contrary, counterfactual fairness deems a predictor to be fair if its output remains the same when the protected attribute is flipped to its counterfactual value [41, 60]. Disparate treatment emphasizes explicit formal classification and intentional discrimination. Therefore, in many machine learning modeling cases where there is no discriminatory intent, disparate impact doctrine is more suited to analyzing unintended biases in data mining compared with disparate treatment doctrine.

The various fairness definitions can also be categorized based on whether they are individual-focused or group-focused. Group fairness, such as equality of opportunities, requires a fair model to treat different groups equally, whereas individual fairness refers to the rule that deems a predictor fair if it produces similar outputs for similar individuals [41, 52]. Though the author uses equality of opportunities as the fairness definition in this research, the above-mentioned fairness definitions can be adopted for future studies in this area.

2.2 Sources of Bias

There are many ways that the bias can seep into the data and the modeling process and consequently lead to the discriminatory prediction results, and the source of bias can happen at the data collection phase or the modeling phase. Below the author lists the sources of bias that have been widely discussed in the machine learning community.

- At the data collection phase

    1. Representation bias

       Representation bias happens if the data for the protected groups are incorrect or nonrepresentative [6]. For instance, in a survey collection process, if the authorities heavily depend on mobile phone for data collection, they may exclude a large portion of the protected population who do not frequently use phones. Survey conducted in English may also exclude people who do not speak English [6]. Therefore, the conclusions drawn from these data will not generalize well to the protected populations.

    2. Sampling bias

       Previous research has found that even the individual records in a dataset are of high quality and can represent the subpopulation well, the statistical bias can still exist. It is often because the protected group have much fewer number of observations than the other group in the training data, and this is often known as the imbalanced data problem [8, 68]. Research showed that if a particular class is underrepresented in the sample, the results of the analysis of that sample may skew against the underrepresented class [35, 8].

    3. Feature selection bias

       Feature selection bias happens when the selected variables fail to capture enough details that account for different outcomes [8, 68]. This bias occurs often because it is cost efficient for the decision maker to rely on easily accessible information to make decisions [8].

- At the modeling phase

    1. Algorithmic bias

       Algorithmic bias is defined as the bias added by the algorithm itself and not present in the input data [7]. This bias exists inherently because the goal of the algorithm is to minimize its prediction error, and there is no motivation for the algorithm to enhance prediction fairness in the training process [57].

2. Bias attributed to the variable correlation

   The algorithm can also produce discriminatory result because of the correlations in the data. If the protected variable is highly correlated with an explanatory variable, then the algorithm will necessarily result in systematically less favorable prediction result for the protected group by using this explanatory variable as the predictor [8, 35]. This bias may occur even if the protected variable has been removed from the dataset, the variables are diverse and granular, and the data are free from latent prejudice or bias [8]. As stated by Hajian et al. (2016), this is a novel and challenging research area for the data mining community [48].

Figuring out the exact sources of bias for a specific data analytic task is difficult. As such, the author tries to gain a better understanding about the bias emerging from the modeling phase by conducting a series of synthetic experiments.

2.3 Method to address the prediction bias

The methods to mitigate the bias fall into three categories:

- Pre-processing

  Collecting more data from the protected groups to create a more balanced and diverse dataset is the most straightforward way to alleviate the prediction bias. However, this approach is often costly, and may not be feasible in many situations.

  Data augmentation is an alternative approach to increase the size and diversity of the training data. It is often applied in the computer vision and the natural language processing fields, such as cropping, darkening and flipping the input images [6].

  Another widely used technique is resampling, which is to randomly sample more data from the protected group [37]. An variant to this is the synthetic minority over-sampling technique, which is to generate "synthetic" minority class members and used them in the training process [20].

  Recent years, scholars also try to optimize the data transformation strategy for data

discrimination prevention. For instance, Flavio et al. (2017) proposed a convex optimization for learning a data transformation that can controlling discrimination, limiting distortion in individual data samples, and preserving utility [18].

- In-processing

  In-processing techniques try to modify and change the learning algorithms to remove discrimination during the model training process. This is achieved either by modifying the objective function or adding a constraint. Some popular in-processing methods include model regularizations [11, 74], adversarial training [12, 89, 65], variational fair autoencoders [64, 27], transfer learning [75] and multi-task learning [29]

- Post-processing

  Post-processing techniques take the model predictions and protected variables and calibrate the model's result to meet the fairness criteria. Post-processing algorithms are easy to apply to existing classifiers without retraining [10]. For instance, Hardt et al. (2016) solved a linear program to find probabilities with which to change output labels to optimize equalized odds [49]. Kamishima et al. (2012) gave favorable outcomes to disadvantaged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty [54].

In this research, an in-processing approach is taken, and the bias mitigation method is adapted from the work of Beutel et al. (2019) [11].

2.4 Computational Fairness in different domains

In this section, the author highlights some domain-specific computational fairness problems that have been studied in previous literature.

- Criminal risk assessment

  The machine learning bias problem in criminal risk assessment was widely popularized by Angwin et al. (2016), who studied COMPAS which is a tool for the criminal risk assessment [5]. They found that a nonrecidivating black defendant is twice as likely to be assessed as high risk as a white defendant, and a recidivating black defendant is nearly half as likely to be assessed as low risk as a white defendant. Chouldechova

(2017) showed mathematically how these differences can result in disparate impact under policy wherein a high-risk assessment leads to a stricter penalty for the defendant. Such policies may be used to inform bail, parole or sentencing decisions. They found that in both the nonrecidivating and the recidivating subgroups, black defendants are observed to receive higher sentences than the white defendants [24].

- Mortgage lending

  Machine algorithms are increasingly adopted in high-stakes decisions such as mortgage lending, insurance and employment. Hardt et al. (2016) demonstrated that if the loan company goes with the max profit strategy, the minorities may be classified poorly and so treated poorly [49]. Bartlett et al. (2019) empirically found that online lenders charge higher interest rates to African American and Latino borrowers, earning 11.5% higher profits on such loans [9]. Studies also showed that despite the direct discrimination that the mortgage lending decisions are explicitly based on the protected attributes, the indirect discrimination can also exist. The indirect discrimination refers to the situations where the treatment is based on apparently neutral non-protected attributes but still results in unjustified distinctions against individuals from the protected group [90]. For instance, if the ZIP code of an individual is used for deciding whether to grant a loan, the decision could exhibit racial discrimination since ZIP code could indicate race.

- Health care

  As AI techniques become increasingly popular in health care applications, researchers found evidence of disparate impacts of AI adoption in health care. Chen et al. (2019) found machine bias with respect to gender and insurance type in terms of ICU mortality and insurance policy for psychiatric 30-day readmission. For gender, female patients have a higher model error rate than male patients; for insurance type, public insurance patients have a much higher model error rate than private insurance patients [21].

Researchers also summarized the negative consequences associated with the machine bias. Consider a situation where the hospitals constructed a model to predict the length of stay of the inpatients in order to allocate the limited case management resources among them. If the ethnic minority neighborhoods predicted greater length

of stay, these health care resources may be directed to the richer, ethnic majority neighborhoods and away from these ethnic minority neighborhoods [72].

Chapter 3

Data and Methods

3.1 Equality of opportunity as the definition of fairness

This study measures fairness by equality of opportunity, mathematically denoted as $P(\hat{y} = 1|z = 0, y = 1) = P(\hat{y} = 1|z = 1, y = 1)$, in which y represents the binary travel behavioral outcomes, $\hat{y}$ represents the predicted values, z represents the protected variable such as race and gender. Intuitively, equality of opportunity requires the predicted travel behavior to be conditionally independent of the protected attributes given that the real outcome is positive [49]. Taking racial disparity as an example, equality of opportunity implies that the predicted travel demand is independent of the travelers being in the minority or majority groups, thus achieving a socially non-discriminatory predictive performance [81]. A related concept is referred to as reciprocal equality of opportunity, denoted as $P(\hat{y} = 1|z = 0, y = 0) = P(\hat{y} = 1|z = 1, y = 0)$, implying that the predicted travel behavioral outcome is conditionally independent of the protected attributes given that the real outcome is negative [12].

3.2 Data and Variables

In this study, the numerical experiments are conducted on two datasets: a group of synthetic datasets and the 2017 National Household Travel Survey data. For each of these two datasets, both BLR and DNN are deployed for model estimations. The experiments on the synthetic data are used to systematically show how the covariance between a protected variable and an explanatory variable may lead to disparate results, and how this prediction

disparity varies with the covariance of these two variables, the sample size, and the number of input variables. The author then tests the bias mitigation method on the synthetic data. For the NHTS data, the author examines the prediction disparity for a series of protected variables (e.g. race, gender, income, medical condition and urban-rural divide) and different dependent variables. The bias mitigation method is later tested on this real-world dataset as well.

3.3 Models and Bias Measurement

In this study, the author adopts two models BLR and DNN for model predictions, which are later evaluated by fairness metrics for different demographics.

3.3.1 Binary Logistic Regression (BLR)

As a classic travel behavior modeling method, BLR has been widely deployed to predict the probability of a certain outcome. The outcome probability is defined as follows:

$$P(y_i = 1|\boldsymbol{x_i}) = \frac{1}{1 + exp(-(\alpha + \boldsymbol{x_i}^\top \boldsymbol{\beta}))}$$

where y_i identifies the dependent variable for individual i, $\boldsymbol{x_i}$ represents the vector of all the independent variables, α is the intercept, $\boldsymbol{\beta}$ is the vector of parameters associated with attribute $\boldsymbol{x_i}$, which is estimated by the negative log likelihood loss function.

3.3.2 Deep Neural Network Modeling (DNN)

DNN usually outperforms traditional method regarding prediction accuracy, because of the non-linear transformation. The outcome probability derived from the DNN deployment can be expressed as [69, 84]:

$$P(y_i = 1|\boldsymbol{x_i}) = \frac{1}{1 + exp(-\Phi(\boldsymbol{x_i}, \boldsymbol{\beta}))}$$

where $\Phi(\boldsymbol{x_i}, \boldsymbol{\beta})$ represents a layer-by-layer transformation: $\Phi(\boldsymbol{x_i}, \boldsymbol{\beta}) = (g_m \circ ... g_2 \circ g_1)(\boldsymbol{x_i}; \boldsymbol{\beta})$, in which each $g_l(\boldsymbol{x_i}^\top \boldsymbol{\beta}) = ReLU(\boldsymbol{x_i}^\top \boldsymbol{\beta} + b_l)$ denotes one standard module in DNN which consists of linear and rectified linear unit (ReLU) transformations. The architecture of the DNNs used in this study is shown in Figure 3-1, which includes 3 hidden layers and 200

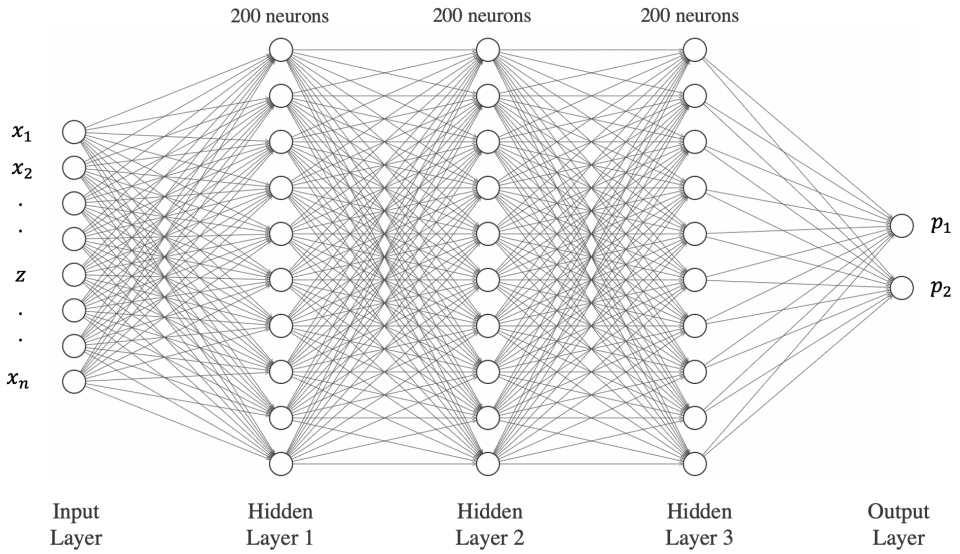


Figure 3-1: DNN architecture

neurons in each layer. Noted that the protected variable z is also included in $\boldsymbol{x}$ as an explanatory variable.

3.3.3 Bias Measurement

After applying BLR and DNN models for a specific prediction task, the author measures the fairness metrics of the prediction result. Based on the fairness definition of equality of opportunity, the unfairness occurs when the machine-learned models offer much worse quality results for some demographic groups than others [49], and the degree of unfairness is measured by the false positive rate (FPR) gap or the false negative rate (FNR) gap between two groups depending on the specific context. The two fairness metrics are calculated as:

$$False\ Positive\ Rate\ (FPR)\ Gap = \frac{FP_{z=0}}{TN_{z=0} + FP_{z=0}} - \frac{FP_{z=1}}{TN_{z=1} + FP_{z=1}} \tag{3.1}$$

$$False\ Negative\ Rate\ (FNR)\ Gap = \frac{FN_{z=0}}{TP_{z=0} + FN_{z=0}} - \frac{FN_{z=1}}{TP_{z=1} + FN_{z=1}} \tag{3.2}$$

In the above expressions, TP_z, FP_z, TN_z, and FN_z represent the number of true positives, false positives, true negatives and false negatives in class z, with $z = 0$ representing the disadvantaged group. For example, when examining the racial bias in predicting the frequent usaage of rideshare, $FN_{z=0}$ represents the number of individuals in the minority group who frequently use rideshare but are wrongly categorized as not doing so, and $FP_{z=0}$ represents the number of individuals in the minority group who do not frequently use rideshare but

are wrongly categorized as doing so. Higher false negative rate or false positive rate for the minority group intuitively suggests that the algorithm makes more mistakes on the ethnic minority group with regard to predicting whether an individual uses rideshare frequently, which might lead to significant mismatch between demand and supply of the TNC service in the minority neighborhoods. Therefore, this fairness definition captures the essential intuition in transport equity discussions. For these two fairness matrices, lower absolute value is better.

3.4 Bias Mitigation

Adapted from the work of Beutel et al. (2019) [11], this research mitigates the prediction disparity through adding a regularization term to the loss function. While Beutel et al. (2019) [11] used the correlation between the output distributions of two groups as their regularization term since the outputs in their studies are continuous scores, the author uses the correlation between the predicted probability distributions of two groups instead, which makes the regularization term differentiable in our classification tasks. This regularization loss term helps shrink the difference of prediction disparity across groups towards zero. Compared with other approaches which generally come with notable engineering concerns, this approach is lightweight, can be easily adapted to real-world system and has achieved good empirical results [11]. The loss function is specified as:

$$\min_{p} \quad (1-\lambda)L_{primary} + \lambda|Corr(p(\boldsymbol{x}), z|y=q)| \tag{3.3}$$

where

$$L_{primary} = \sum_{i=1}^{N}[-y_i log(p(\boldsymbol{x_i})) - (1-y_i)log(1-p(\boldsymbol{x_i}))] \tag{3.4}$$

$$Corr(p(\boldsymbol{x}), z|y=q) = \frac{\sum_{i\in S_q}(p(\boldsymbol{x_i}) - \overline{p(\boldsymbol{x_i})})(z_i - \overline{z_i})}{(\sqrt{\sum_{i\in S_q}(p(\boldsymbol{x_i}) - \overline{p(\boldsymbol{x_i})})^2} + \epsilon) * (\sqrt{\sum_{i\in S_q}(z_i - \overline{z_i})^2} + \epsilon)} \tag{3.5}$$

$$S_q = \{i|y_i = q\}\ ;\ \epsilon = e^{-20} \tag{3.6}$$

In the above equation, $\boldsymbol{x_i}$ is the vector representing the explanatory variables. y_i denotes the true outcome and $p(\boldsymbol{x_i})$ represents the estimated probability of the output y_i=1 using the explanatory variables. q equals 1 if the false negative bias is examined and equals 0 if the

false positive bias is examined. z_i denotes the value of the protected variable. ϵ is added to prevent the denominator from becoming zero. S_q represents the set of samples with $y_i = q$, which is used to compute the correlation loss.

$L_{primary}$ is a negative log likelihood loss function for the DNN. The correlation term is a penalty added to the model based on the distribution of the predictions. By reducing the correlation between $p(\boldsymbol{x_i})$ and z_i conditioning on $y_i = q$, it minimizes the conditional dependence between the distribution of the predicted probabilities and the group membership determined by the protected variable. λ is a parameter controlling the tradeoff between primary loss and the fairness loss. When $\lambda = 0$, no bias mitigation is employed. In this research, the author demonstrates the effectiveness of using this fairness-adjusted loss function in both BLR and DNN.

Chapter 4

Experiments

In this section, the experiment results for the synthetic datasets and the NHTS dataset are reported. For both synthetic and NHTS data, we first conduct BLR and DNN for model estimations, then examine the fairness issues on the prediction results, and lastly apply the bias mitigation methods. DNN is implemented using the TensorFlow library in Python. BLR is implemented using the scikit-learn library when examining the fairness issues while using the TensorFlow library for bias mitigation, since TensorFlow allows us to modify the loss function.

Experiments conducted in TensorFlow all use the mini-batch stochastic gradient descent method with the batch size equaling 1,000 and the step size equaling 0.0001 in each training. The author draws samples without replacement to generate the mini-batches within an epoch. After a mini-batch is generated, the algorithm calculates the prediction loss and updates the coefficients. The model which produces the lowest training loss among the 50 epochs is chosen and later performs prediction over the test data. The author ran 5 trials of 5-fold validation for each experiment.

4.1 Synthetic experiment

In the simulations, the author considers the type of bias that arises when the true predictor and the protected variable are highly correlated. In this case, the true predictor of the outcome also happens to serve as a reliable proxy for class membership in the training set [8, 92]. For example, if the usage of rideshare is positively associated with income, and

ethnic minorities in the training data tend to have lower income, the algorithm will tend to predict low usage of rideshare for the minority population, even if the true contributing factor for rideshare usage is income rather than race. This type of bias is inherent in the existing population inequality and has been less emphasized in the previous literature.[1]

4.1.1 Data generation process

In the synthetic dataset, each data point can be represented by a tuple (z_i,x_i,$\boldsymbol{k_i}$,y_i), where z represents the protected variable (e.g. race, gender), x is the explanatory variable that is correlated with z (e.g. income), $\boldsymbol{k}$ is a vector of explanatory variables which does not include x and z. y represents the binary outcome.

First, x and all the elements in $\boldsymbol{k}$ are drawn from the standard normal distributions and are independent with each other. z is generated as a binary variable that is positively correlated with x and is independent with $\boldsymbol{k}$. The label y is drawn from a binomial distribution with probability $Pr(y = 1) = \frac{1}{1+\exp(-V)}$. The systematic utility function V takes x and $\boldsymbol{k}$ as the input variables.

The author tests two scenarios with the true utility function taking a linear form and a quadratic form respectively. Let $V = \alpha + \boldsymbol{w}\phi(x, \boldsymbol{k})$. In the first scenario, $\phi(x, \boldsymbol{k})$ takes the linear transformation: $\phi(x, \boldsymbol{k}) = [x, k_1, k_2, ..., k_d]$. The weight for x is set as 1, and the weights for other explanatory variables takes {-0.5,0.5} with equal probabilities. In the second scenario, $\phi(x, \boldsymbol{k})$ takes the quadratic transformation: $\phi(x, \boldsymbol{k}) = [x, k_1, ..., k_d, x^2, k_1^2, ..., k_d^2]$. The weights for x and x^2 are set as 1 and 0.5, and the weights for other explanatory variables takes {-0.5,0.5} with equal probabilities. This data generation process makes sure that the mean value of z is 0.5, which means that there are approximately equal numbers of $z = 0$ and $z = 1$, thus giving us a balanced dataset. The detailed descriptions about the data generation process can be found in Appendix A.

[1] Besides this inherent bias, other sources of bias could exist, such as the imbalanced training data problem. This problem arises when the disadvantaged group have insufficient training data or is misrepresented, in which case the model will either fail to learn a correct statistical pattern or favor the majority group during the estimations, since the training data in the disadvantaged groups often misrepresent the true population when they are insufficient [87, 46, 72, 39, 17].

The cases where $Cov(z, x)$ is non-negative and x positively affects V are examined. The author therefore uses $z = 0$ to mimic the disadvantaged population and uses $z = 1$ to mimic the privileged population, since in the real world the privileged population (e.g. the ethnic majority group) is often positively correlated with the factor (e.g. income) that has a positive effect on the utility of an advanced mobility service (e.g. the utility of using a ride-hailing service).

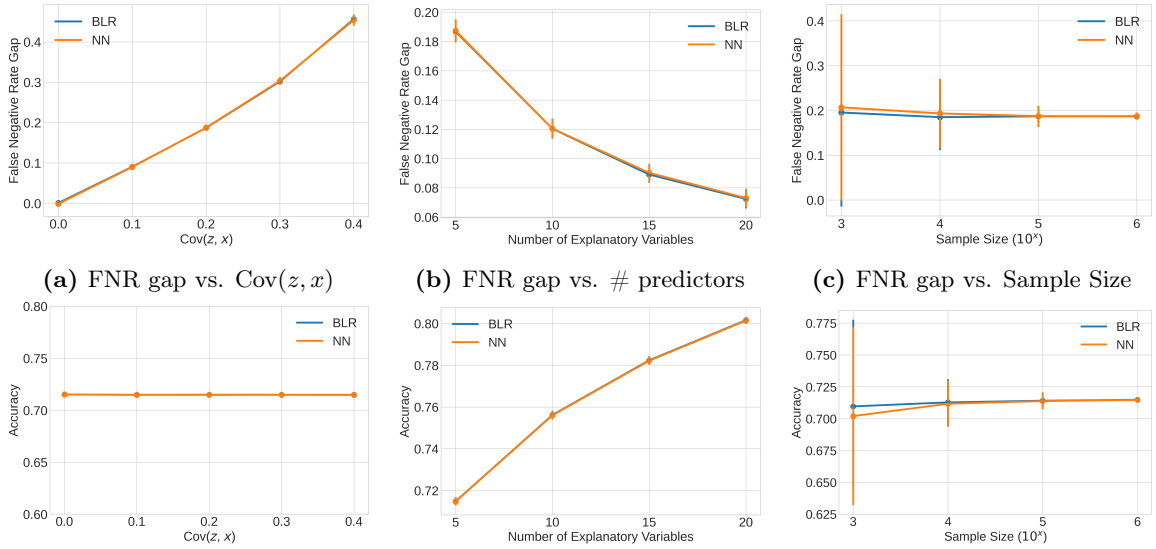
4.1.2 Fairness measurement results

In the prediction phase, the author uses z, x and $\boldsymbol{k}$ as the explanatory variables for both the logit model and the DNN model, so these two choice models are defined as $Logit(z, x, \boldsymbol{k})$ and $DNN(z, x, \boldsymbol{k})$.

First, we want to examine how fairness measure and accuracy vary regarding the correlation between the sensitive attribute z and the explanatory variable x, the sample size and the number of explanatory variables in the data generation process. Therefore, the author runs experiments along these three dimensions, and when each dimension is examined, the author sets the other two dimensions as the default values. The default values for $Cov(z, x)$, sample size and number of explanatory variables are 0.2, 10^6 and 5. For each experiment, three datasets are randomly generated based on the above data generation process.[2]

Figure 4-1 presents the results of the linear data generating process, while in Figure 4-2, the true data generating process is quadratic in variables. In both figures, the first row shows the FNR gap (see Equation 3.2) between the disadvantaged group (defined as $z = 0$) and the privileged group (defined as $z = 1$) as the measure of fairness and the second row shows the prediction accuracy. The x-axis of the first, second and third columns respectively represent the covariance between z and x, the number of explanatory variables in the data generation process and the sample size. Each figure plots both the BLR and DNN results, which are represented by the blue and orange colors. The figures plot the values averaged

[2]Occasionally, the data generation process in Scenario 2 produced datasets with highly unbalanced outcomes (e.g. when the minority class accounts for less than 30% of the total samples). In that case, the author would drop that unbalanced dataset and generate another one. The author iterated this process until all the datasets are roughly balanced (when the minority class accounts for 40%-50% of the total samples)



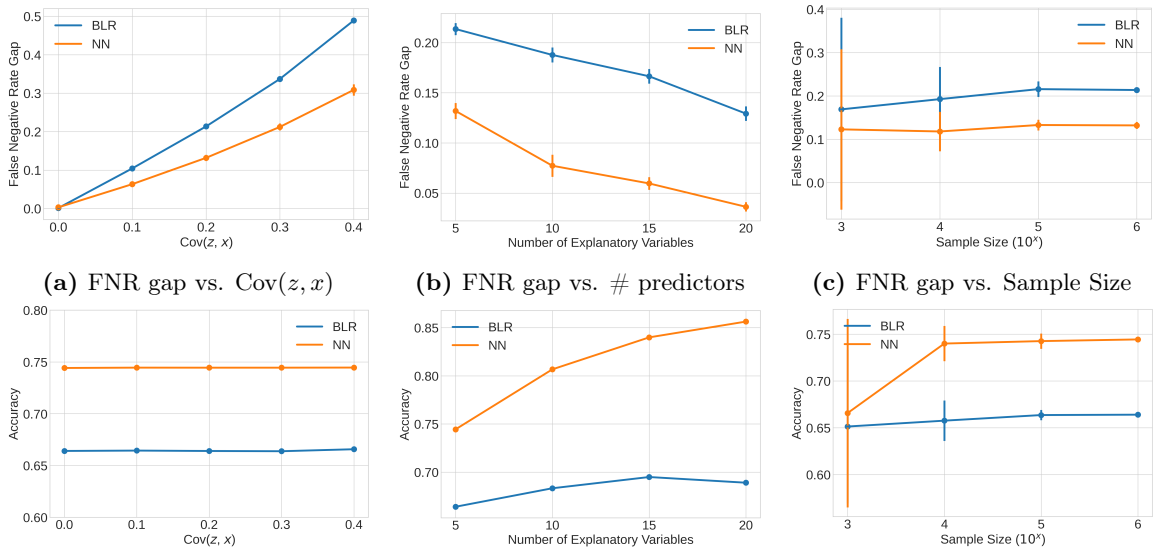
(a) FNR gap vs. Cov(z, x) **(b)** FNR gap vs. # predictors **(c)** FNR gap vs. Sample Size

(d) Accuracy vs. Cov(z, x) **(e)** Accuracy vs. # predictors **(f)** Accuracy vs. Sample Size

Figure 4-1: Fairness metric and accuracy with different parameters (BLR vs. DNN): true model taking the linear form; estimation models: $Logit(z, x, \boldsymbol{k})$ and $DNN(z, x, \boldsymbol{k})$. $Logit(z, x, \boldsymbol{k})$ adopts a linear specification, so both models contain the true model.

across 5 trials of 5-fold validation in 3 datasets for each experiment; the error bar indicates the standard deviation multiplied by 1.96, which approximates the confidence interval of the estimations.

As shown in Figure 4-1, the results of BLR and DNN mostly overlap since they both recover the true linear model. Figure 4-1a shows that the FNR gap increases with $Cov(z, x)$, indicating that as x becoming more positively correlated with z, the algorithm is more likely to falsely associate the disadvantaged population (z=0) with the negative outcomes, even if their real outcomes are actually positive. Prediction disparity is a metric relatively independent of the predictive performance, as there is no difference in prediction accuracy with different $Cov(z, x)$ (Figure 4-1d); it is also not due to the unbalanced training data problem, as the outcome variable, the protected variable and all the explanatory variable are balanced in the training data. The prediction disparity is purely inherent in the relationship among variables in the data. Figure 4-1b shows that the FNR gap decreases with the number of explanatory variables, which is probably because increasing the number of predictors dilutes the influence of x on the outcome. Figure 4-1c shows that the variances of the fairness and accuracy estimations decrease as the sample size increases.



(a) FNR gap vs. Cov(z, x) **(b)** FNR gap vs. # predictors **(c)** FNR gap vs. Sample Size

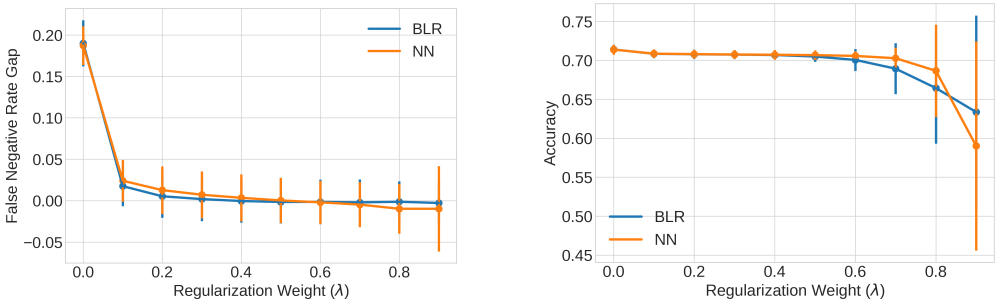
(d) Accuracy vs. Cov(z, x) **(e)** Accuracy vs. # predictors **(f)** Accuracy vs. Sample Size

Figure 4-2: Fairness metric and accuracy with different parameters (BLR vs. DNN): true model taking the quadratic form; estimation models: $Logit(z, x, \boldsymbol{k})$ and $DNN(z, x, \boldsymbol{k})$. $Logit(z, x, \boldsymbol{k})$ follows a linear model specification, so it has the misspecification error.

Figure 4-2 shows the results of prediction fairness and accuracy when the true data generation model takes a quadratic form. In this case, the BLR with linear specification has the model misspecification error while DNN does not. Figure 4-2a shows that the prediction disparity still increases with the increase of $Cov(z, x)$, and DNN is always associated with smaller FNR compared with BLR for $Cov(z, x)>0$. This result indicates that the model misspecification not only induces more prediction error, but also harms prediction fairness. Figure 4-2b shows that though increasing the number of explanatory variables can reduce the prediction disparity, the magnitude of the prediction disparity caused by the model misspecification was not significantly reduced. Figure 4-2c indicates that the fairness prediction result becomes more stable as the sample size increases.

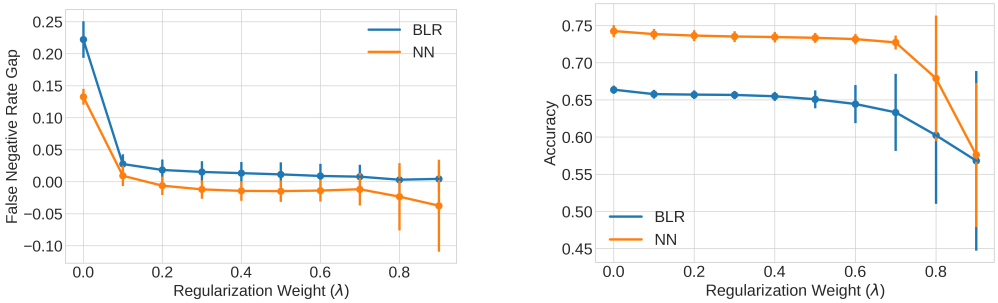
4.1.3 Bias mitigation results

To address prediction disparity, the author applies the bias mitigation method as illustrated in Section 3.4 to the synthetic datasets with $Cov(z, x) = 0.2$, sample size equaling 100,000 and number of predictors equaling 5. For each regularization weight λ, the author runs the training procedure 5 times for each of the 3 datasets with 5-fold cross-validation and reports the average results in Figure 4-3 for Scenario 1 and Figure 4-4 for Scenario 2. The error bars in the figures indicates the standard deviation multiplied by 1.96, which approximates



(a) FNR gap vs. Regularization Weight **(b)** Accuracy vs. Regularization Weight

Figure 4-3: Fairness and accuracy by bias mitigation weight (λ): true model taking the linear form (Scenario 1)



(a) FNR gap vs. Regularization Weight **(b)** Accuracy vs. Regularization Weight

Figure 4-4: Fairness and accuracy by bias mitigation weight (λ): true model taking the quadratic form (Scenario 2)

the confidence interval.

Figure 4-3a and 4-4a show that in both scenarios, applying the regularization even with a small weight (e.g. $\lambda = 0.1$) can substantially reduce the prediction disparity and this finding holds for both BLR and DNN. Given the model misspecification for BLR in Scenario 2, Figure 4-4a shows that the method is still effective in reducing the prediction bias to as small as zero.

Figure 4-3b and 4-4b report the corresponding model accuracy as the regularization weight varies. The results show that when $\lambda < 0.7$, the accuracy only slightly decreases. These results suggest that the improvement in prediction fairness can be achieved with a minimal cost of accuracy.

4.2 The NHTS dataset

4.2.1 Data and variables

The NHTS data are collected directly from a stratified random sample of U.S. households. The richness of the dataset enables us to examine fairness in predictions with varying dependent variables and protected attributes. Protected attributes are the variables we want to protect against in the model prediction process, which include race, gender, income, medical condition and urban-rural divide in this study. In terms of "race", the ethnic minority is defined as the non-white population. In terms of the variable "income", low-income households are identified based on the combination of household size and last year's household income following the 2017 Health and Human Services poverty guidelines [25]. An individual is deemed having a "medical condition" if he or she answered "yes" to the question "do you have a condition or handicap that makes it difficult to travel outside of the home?" in the survey. Regarding the protected variable "region", the question "household in urban area? Answer 'yes' or 'no'." is used to identify whether the individual is an urban or rural resident.

The dependent variables examined in this study can be categorized into two groups: the first group contains four variables indicating the "yes" or "no" answers to "usually work from home", "have the option of working from home", "agree that travel is a financial burden" and "agree that gas price affects travel"; the second group contains four variables indicating the high frequent usage of four travel modes: bike, car, bus and rideshare. These eight dependent variables are all binary variables, with "yes" taking the value 1. A detailed description of these variables can be found in Appendix B.

The distributions of the dependent variables except "travel burden" and "gas price impact" are highly skewed, and previous research has found that when the outcome class sizes are highly imbalanced, the classification algorithms tend to strongly favor the majority outcome class, resulting in very low or even no detection of the minority outcome class [16]. Therefore, for each of the six imbalanced dependent variables, the author balances the data to facilitate training by downsampling the majority class. The summary statistics of the independent and dependent variables are reported in Appendix B. The distributions of two groups of dependent variables by different protected attributes are reported in Appendix C.

4.2.2 Bias mitigation method with sample weights

As previously mentioned, one source of bias in the data is that the training data might not be representative of the overall population. Luckily, NHTS contains the sample weight[3] [3] for each individual, which can be used to largely address the representation bias. The author incorporates the sample weights in the training and evaluation phases. Weighted accuracy and weighted fairness metrics are used for model evaluation. To be specific, the weighted accuracy is calculated as:

$$Weighted\ Accuracy = \frac{\sum_i^N 1(\hat{y}_i = y_i) w_i}{\sum_i^N w_i} \tag{4.1}$$

where w_i represents the sample weight for sample i, y_i is the label and $\hat{y}_i$ is the predicted outcome. Similarly, the sample weight is applied for each sample when calculating the fairness metrics (FNR and FPR). N denotes the sample size.

Corresponding to the weighted evaluation metrics, the sample weights are also applied in the loss function during the training process. The weighted loss function is written as:

$$\min_p \quad (1-\lambda) L_{primary} + \lambda |Corr(p(\boldsymbol{x}), z | y = q)| \tag{4.2}$$

where

$$L_{primary} = \sum_{i=1}^{N} \frac{w_i L_i}{\sum_i^N w_i} \ ; \ L_i = -y_i log(p(\boldsymbol{x_i})) - (1-y_i) log(1 - p(\boldsymbol{x_i})) \tag{4.3}$$

$$Corr(p(\boldsymbol{x}), z | y = q) = \frac{\sum_{i \in S_q} w_i (p(\boldsymbol{x_i}) - m(p(\boldsymbol{x_i})))(z_i - m(z_i))}{(\sqrt{\sum_{i \in S_q} w_i (p(\boldsymbol{x_i}) - m(p(\boldsymbol{x_i})))^2} + \epsilon) * (\sqrt{\sum_{i \in S_q} w_i (z_i - m(z_i))^2} + \epsilon)} \tag{4.4}$$

$$m(p(\boldsymbol{x_i})) = \frac{\sum_{i \in S_q} w_i p(\boldsymbol{x_i})}{\sum_{i \in S_q} w_i} \ ; \ m(z_i) = \frac{\sum_{i \in S_q} w_i z_i}{\sum_{i \in S_q} w_i} \tag{4.5}$$

$$S_q = \{i | y_i = q\} \ ; \ \epsilon = e^{-20} \tag{4.6}$$

BLR is implemented through the scikit-learn library when evaluating the fairness issues and through TensorFlow when mitigating the bias. For DNN, all the experiments are imple-

[3] which is primarily calculated as the inverse of the probability of selection of the person in the given sampling stratum from the sampling frame

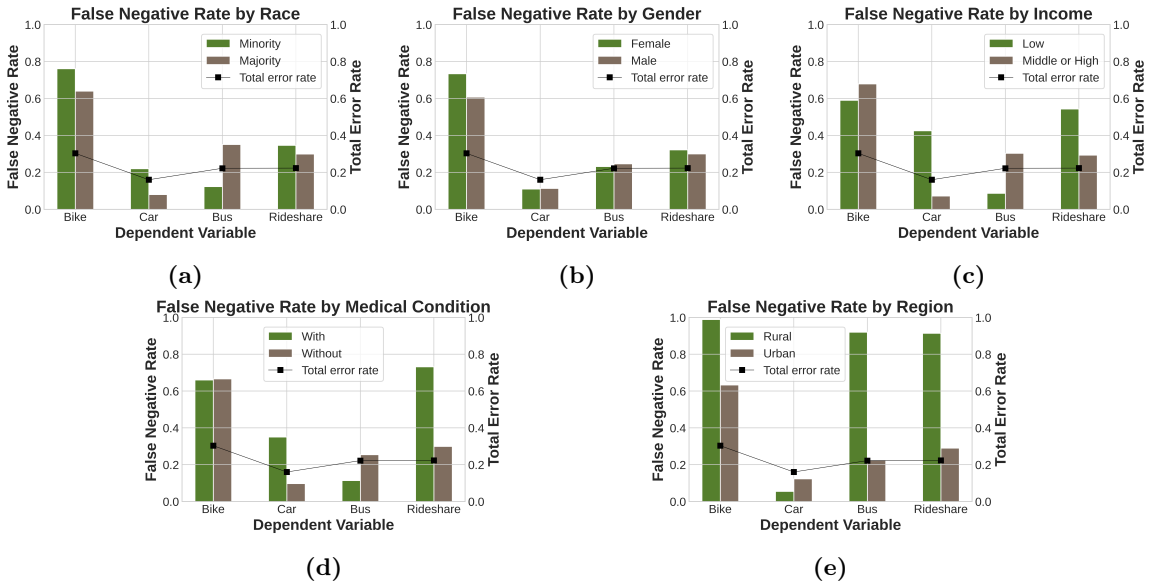


Figure 4-5: Disparity of prediction accuracy (BLR): frequent usage of bike, car, bus and rideshare

mented in TensorFlow. Each of the experiments conducted in TensorFlow uses the mini-batch gradient descent method with step size 0.0001 during training. The best model among the 5000 epochs is chosen and later performs prediction over the test data. 3 trials of 5-fold validation are conducted for each experiment.

4.2.3 Fairness issues in the adoption of BLR and DNN

The comparison of prediction accuracy with respect to various protected variables are presented by the bar charts in Figure 4-5 and 4-6 for BLR and in Figure 4-7 and 4-8 for DNN. Each bar chart depicts the prediction accuracy of two populations grouped by a specific protected variable (race/gender/income/medical condition/region) side by side. The height gap of two adjacent bars shows the prediction disparity for that protected variable. Figure 4-5 and 4-7 illustrate the prediction results of the dependent variables regarding travel mode usage by BLR and DNN. Figure 4-6 and 4-8 plot the prediction results of the dependent variables "work from home", "work from home option", "travel burden" and "gas price impact" by BLR and DNN. The dependent variables are specified on the x-axis of the bar charts.

The y-axis of the bar charts represents one of the error rates: FPR or FNR. The author examines FNR for the first group of dependent variables - since we are concerned about

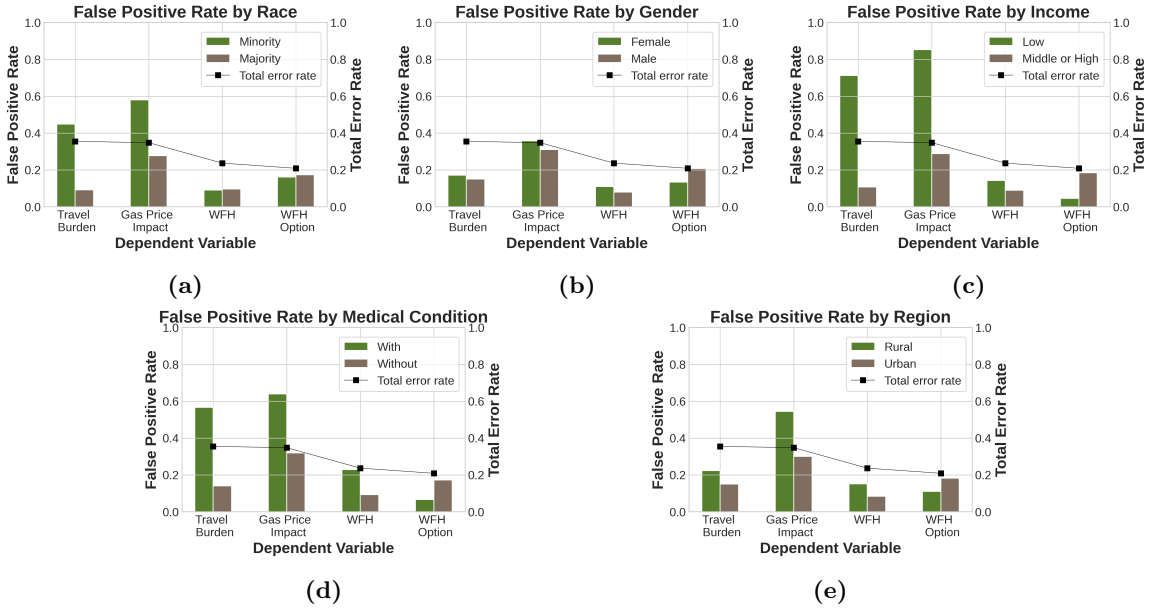


Figure 4-6: Disparity of prediction accuracy (BLR): work from home, work from home option, travel burden, gas price impact

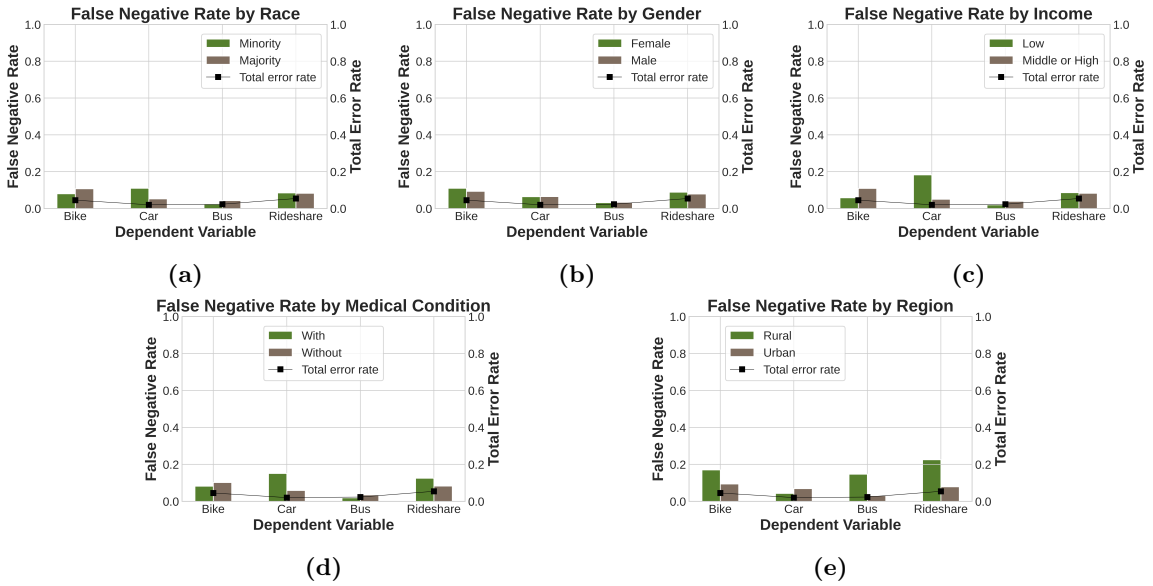


Figure 4-7: Disparity of prediction accuracy (DNN): frequent usage of bike, car, bus and rideshare

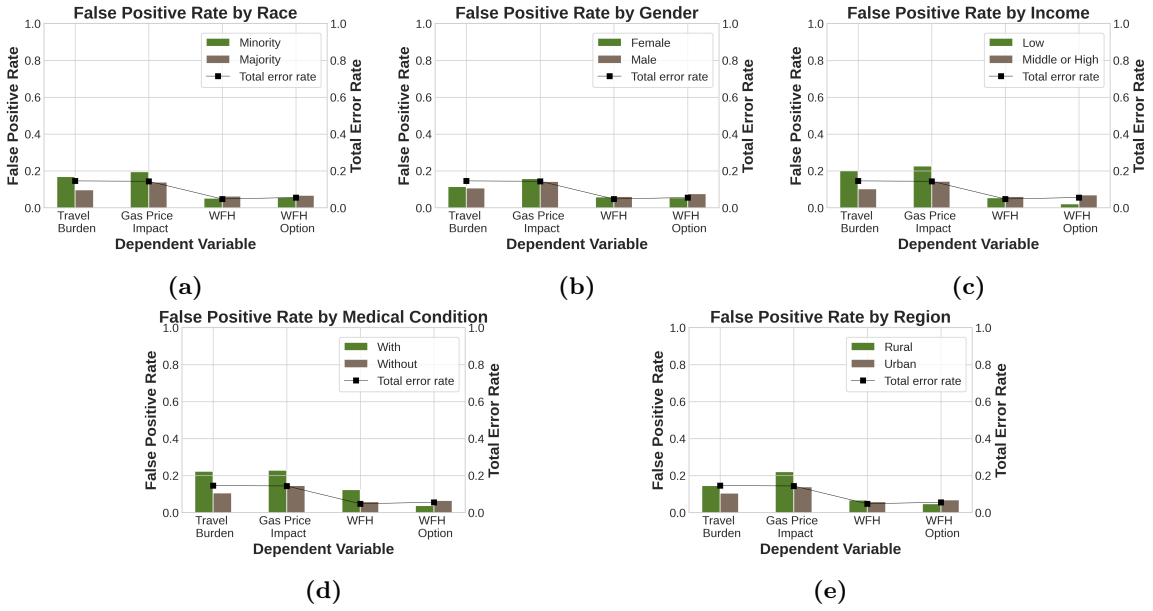


Figure 4-8: Disparity of prediction accuracy (DNN): work from home, work from home option, travel burden, gas price impact

cases where active users of a certain travel mode are not identified (Figure 4-5 and 4-7), and FPR for the second group of dependent variables (Figure 4-6 and 4-8) - since we want to focus on instances which have negative outcomes but are wrongly identified as positive (e.g. people who do not have the option of working from home but are mistakenly identified as having the option). In each bar chart, the height of a bar represents the magnitude of the class-specific FNR rate or FPR rate. The total error rate is also presented which refers to one minus the weighted accuracy for all samples.

First, we focus on the BLR results. Figure 4-5 and 4-6 show that prediction disparities widely exist with the implementation of BLR. Figure 4-5 presents the false negative bias across different populations regarding the frequent usage of different travel modes. The plots show that except for the fairly consistent prediction accuracy between male and female, the significant disparity of prediction accuracy exists for all other protected variables. Racial bias is significant regarding the prediction of frequent usage of car and bus; income bias and bias regarding the medical condition are significant regarding the predictions of car, bus and rideshare. Among all the protected attributes, the regional disparity is the largest with respect to predicting the frequent usage of bike, bus and rideshare. As a result, the proportion of rural residents that use bike, bus and rideshare frequently is underestimated

compared with that of the urban residents. If the policy makers use the modeling results to inform transportation resource allocations such as the planning of bike lanes and bus routes without considering the prediction bias, the rural area will very likely be under-served.

In terms of the dependent variables, when estimating the high frequent usage of rideshare, it is found that for all the protected variables, the disadvantaged group always has higher FNR than the other group with either BLR or DNN. This finding indicates that the communities where these disadvantaged social groups (the ethnic minority, female, low-income group, people who have medical conditions and rural residents) dominate could procure less ride sharing service if the decision makers use these data for rideshare demand estimation and do not account for the prediction disparity.

Figure 4-6 illustrates significant racial bias, income bias and health-related bias for the predictions of "travel burden" and "gas price impact". In other words, the ethnic minority population, the low-income population and people with health conditions are more likely to be predicted as "regarding travel as a financial burden" and "agreeing that price of gasoline affects travel" when the true outcome is actually negative. These biases could lead to disadvantageous consequences for the vulnerable populations. For example, given that banks are less likely to loan to someone if they perceived that person to be under financial stress, if the ethnic minorities suffer from higher FPR in travel burden prediction, they will more likely get rejected when applying for loans.

Next, the author compares the results between BLR and DNN. The prediction disparity and the total error rate of BLR are considerably larger than those of DNN in all the scenarios, showing that the prediction disparity of BLR largely comes from model misspecifications. By reducing model misspecification, the implementation of DNN can not only increase prediction accuracy, but also improve computational fairness. These findings are consistent with the simulation results as illustrated in Figure 4-2, which show the positive relationship between the magnitude of fairness gap and the degree of model misspecifications.

However, the prediction results of DNN (Figure 4-7 and Figure 4-8) also show that even if the prediction achieves very high accuracy (>94% for all dependent variables in Figure 4-7

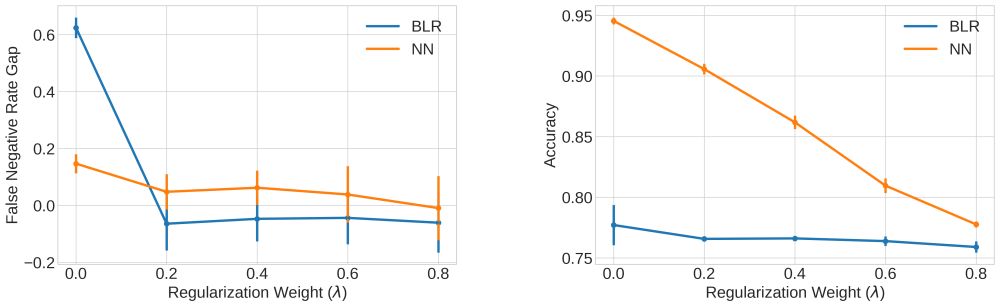
and >85% for all dependent variables in Figure 4-8), the fairness gap still exists, and for some sensitive variables the prediction disparity can be higher than 15% (e.g. the FNR gap for the frequent usage of rideshare prediction between rural and urban residents). This finding tells us that deploying complex models aiming at improving prediction accuracy cannot guarantee to eliminate prediction biases. Therefore, dedicated method should be deployed to mitigate the unintended bias other than merely attempting to improve the model prediction power.

Besides, except for the racial difference in bike frequency prediction, the signs of the fairness gaps for other protected variable and dependent variable combinations are the same between BLR and DNN. The consistency between the two models suggest that the prediction disparity may reveal the bias inherent in the existing inequality in the society which is baked in the data. For example, in the predictions of mode usage, it is found that the socially disadvantaged groups such as the ethnic minority group and people having medical conditions tend to suffer from higher FNR in the predictions of car and rideshare usage, probably because they are negatively associated with certain factors (e.g. income) that positively contribute to the usage of these modes which are more expensive compared with other modes.

4.2.4 Bias mitigation results

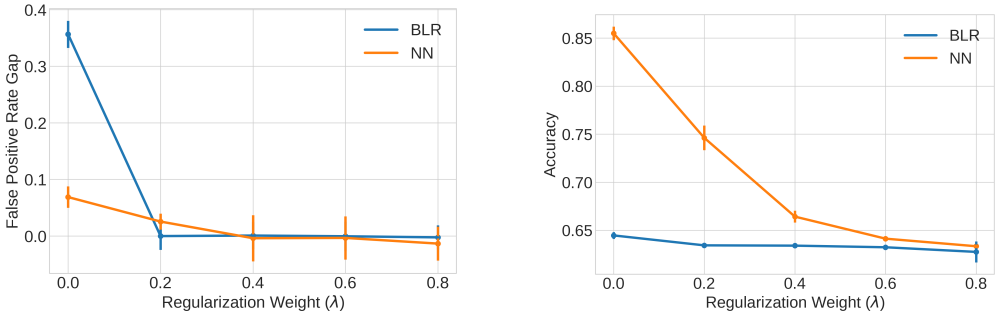
The author adopts the absolute correlation regularization method to mitigate the unintended bias for both BLR and DNN models. The method is applied to mitigate two types of prediction disparities: the FNR gap for estimating the frequent rideshare usage between rural and urban residents, and the FPR gap for the prediction of "travel burden" between the ethnic minority group and the majority group, as the initial prediction disparity in these two prediction tasks is substantial compared with those in other prediction tasks. In Figure 4-9 and 4-10, plot (a) reports the average prediction disparity and plot (b) reports the average prediction accuracy in 3 trials of 5-fold cross-validation with varied bias mitigation weight (λ). In both plots, higher weight indicates larger punishment on the prediction bias.

Both Figure 4-9 and 4-10 demonstrate the effectiveness of the bias mitigation method. The prediction disparity decreases as λ increases, and this effect is particularly significant for BLR. The blue curves in these two figures show that the fairness gap diminishes sharply



(a) FNR gap vs. Regularization Weight **(b)** Accuracy vs. Regularization Weight

Figure 4-9: Fairness and accuracy by bias mitigation weight (λ): regional bias in the prediction of frequent rideshare usage



(a) FPR gap vs. Regularization Weight **(b)** Accuracy vs. Regularization Weight

Figure 4-10: Fairness and accuracy by bias mitigation weight (λ): racial bias in the prediction of travel burden

even if only a small degree of bias mitigation weight is applied.

In Figure 4-9, it is found that increasing λ from 0 to 0.2 can reduce the FPR gap from 62.4% to -6.4% for BLR and from 14.5% to 4.8% for DNN, only at the expense of reducing the overall accuracy from 77.7% to 76.6% for BLR and from 94.5% to 90.6% for DNN, and similar effect is found in Figure 4-10. These results indicate that with the adoption of the bias mitigation method, we can reduce the fairness gap in BLR to a similar level of that in DNN. In Appendix D, the author also shows the convergence of the loss functions in training.

Chapter 5

Conclusions

This study investigates equality of opportunity as a measurement of computational fairness in travel behavior modeling with BLR and DNN. Fairness has long been a critical concern in transportation studies. However, past transportation equity studies generally evaluated equity based on the cost-benefit analysis for different populations and neglected the potential fairness problem in the machine-learned models. This research aims to enrich the transportation equity research by emphasizing the integration of fairness metric into the modeling process to improve fair transportation decision making and resource allocations.

First, this research uses the concept of equality of opportunity to define computational fairness in travel behavior modeling, which is measured by the gap of true positive rates between two groups of populations. This definition is connected to the equality of opportunity in traditional transportation equity literature, since choice probabilities in the discrete choice models are a natural metric for opportunities. The unfair prediction of choice probabilities in modeling can lead to the allocation of inadequate transportation resources to the disadvantaged neighborhood.

Then, the author conducts the simulated experiments to show two sources of prediction disparities: the bias inherent in the data structure and the bias in the modeling process. Prediction disparity increases as the correlation between the protected variables and the explanatory variable(s) increases, exhibiting the bias inherent in the data structure but not caused by modelers. The inherent bias illustrates how prediction disparities may still exist without the presence of human errors. When the true model specification cannot be cap-

tured by BLR but can be captured by DNN, DNN can produce lower prediction bias than BLR, owing to its ability to capture the complex relationships among variables. This type of bias is the algorithmic bias that can be mitigated by choosing a more fitting model, thus is a sort of human bias.

Next, the author conducts the computational fairness analysis on the NHTS dataset using BLR and DNN. The results reveal the prevalence of the prediction disparities in travel behavior modeling particularly when BLR is adopted. Though the magnitudes of unfairness are different, the signs of the fairness gap (e.g. which group has higher FNR/FPR) are the same between BLR and DNN in the vast majority of the prediction cases, probably reflecting the bias inherent in the existing population disparity. The regional disparity in the predictions of the frequent usage of bike, bus and rideshare is the largest among all the protected attributes, indicating that the resources regarding the bike, bus and rideshare services allocated by the predictions will be withheld from the rural residents compared with the urban residents. When estimating the high frequent usage of rideshare, the disadvantaged group has higher FNR than the other group for all the protected variables, indicating that these disadvantaged social groups might receive insufficient ride sharing service if the TNC companies use these data for rideshare demand estimation but do not consider the prediction disparity.

Finally, the author adopts the absolute correlation regularization method to mitigate the bias in the BLR and DNN prediction using the synthetic and the NHTS datasets. The method proves to be effective for both BLR and DNN. When the initial fairness gap is large, applying the bias mitigation method with only a small weight of regularization can considerably improve the fairness result. Though there is an accuracy-fairness trade-off in most predictions, we can achieve substantial improvement in prediction fairness with only a small reduction of accuracy by careful selection of the bias mitigation weight.

All in all, the author argues that researchers and policy makers should be aware of the normative aspect in the seemingly value-neural machine learning predictions, since the prediction disparities can lead to severely unfair treatment on the already marginalized social groups. Only after acknowledging the existence of the biases, can policy makers start to

adopt effective remedies.

Chapter 6

Discussions and Future Work

This research mainly focuses on ways to address fairness from a modeling perspective. However, improvement is also needed on the institutional side. One important way to achieve fair outcomes is to have a "human in the loop." [6] Firstly, the decision makers should ensure that the data collection and modeling process can be monitored or queried externally, and then invite representatives from different population groups, especially the disadvantage population groups, to audit the decision making process, from data collection to algorithm implementations. Second, the decision makers should also ensure that there are specific people responsible for addressing the machine bias issues.

This study demonstrates that adjusting BLR and DNN models to reduce prediction disparity may lead to lower prediction accuracy, but how to trade off accuracy and fairness essentially involves a value judgement, on which the author does not take a stance. In fact, this needs to be figured out on the basis of different normative principles, and may involve some ethical debates about what should be considered "fair" in a given context. Therefore, it is crucial to involve people who will be influenced by the data analysis results to jointly define what is a fair decision making system. In addition, mitigating the unintended prediction biases can involve more costs. Policy makers should determine whether the gain in computational fairness can outweigh the efforts of improving computational fairness, such as employing bias mitigation algorithms or collecting data that represent the socially marginalized groups. Further benefit-cost analysis is necessary to bridge the gap between the algorithmic fairness tools and the social justice needs in the real world.

On the other hand, this research also reveals the needs to incorporate the fairness analysis into future transportation studies. Several potential applications of the algorithmic fairness analysis by travel modes and their policy relevance are discussed below.

- Ridehailing service

  The ridehailing system often experiences an imbalance between the travel demand and vehicle supply. As a result, the TNC companies like Uber adopt a vehicle rebalancing strategy that proactively relocates the vehicles based on the expected demand for trips [2]. The goal of the strategy is to minimize the operational cost by reducing the empty vehicle miles travelled while providing a reasonable waiting time for the passengers [33].

  Since the feet repositioning algorithms require forecasts of future demands, the bias exhibited in the demand prediction may lead to unfair ridehailing service provision among different neighborhoods. To be specific, if the ridehailing demand for the marginal neighborhood (e.g. the neighborhood where the disadvantaged population dominates) is systematically underestimated, the vehicles allocated to these neighborhoods may not be enough to serve the demand.

  Therefore, the TNC companies should scrutinize the fairness dimension of their demand forecasting tools, and take actions to address the demand forecasting bias if the algorithmic fairness problem exists. In this way, the company can contribute to the social fairness of the system and build up their reputations, but this may require giving up a proportion of their profits.

  The transportation authority should also conduct the fairness assessment of the algorithms used by the TNC companies. Unfortunately, the authority often does not have the access to the data and algorithms. Therefore, the TNC companies should consider increasing the transparency on the algorithmic decision making process to show that the modeling result is fair and unbiased.

- Emerging transportation technology: autonomous vehicle, electric vehicle

The computational fairness analysis is especially important for estimating the demand of the new mobility services, such as autonomous vehicle and electric vehicle. These new mobility services either have not been introduced, or have not been widely implemented in the U.S. Therefore, the vehicle providers may depend on the demand prediction models to determine how many vehicles and how much corresponding service to provide. A biased demand prediction model can therefore lead to the under-provision of the new mobility infrastructure, product and service. Intuitively, the disadvantaged populations are more likely to have a higher FNR than the privileged population. This is because previous research has shown that more prosperous people are often more acceptable of the new mobility service [51], whereas the disadvantaged populations usually have lower income.

For the autonomous vehicle, using the fairness-adjusted demand estimation model will change the downstream supply-side decisions such as service allocation and vehicle routing, matching and rebalancing. The magnitude and significance of the change is yet unknown, which needs further investigation.

Other than the service provision solutions, the pricing scheme can also be affected by the fairness-adjusted model. To determine a road pricing or congestion charging strategy, the traffic demand information such as the origin-destination demand estimation is usually required. Therefore, the biased demand estimation may also affect the pricing decisions.

In terms of electric vehicles, the decision of where the electric vehicle charging stations should be allocated among different regions also take the spatial demand estimations as the input. If the demand analysis is biased against the marginal neighborhood, the service providers may underestimate the needs for electric vehicles in those neighborhoods, leading to insufficient provision of the charging stations.

- Active travel modes: public transit, walking, biking

  Incorporating fairness metric into the demand prediction for active travel modes in-

cluding public transit, walking and biking can inform equitable transportation infrastructure planning related to these modes.

For public transit, the fairness-aware estimation model should be applied to predict the demand that can serve as an input for the bus line/subway line planning, fleet size arrangement and headway optimization. For walking, the fairness-adjusted demand model should be utilized for the determination of sidewalk investment, and the fairness analysis should also be incorporated into the pedestrian safety analysis. For biking, the fairness-adjusted demand estimation is also important for the bike-sharing facility investment and the bike plane planning.

Unlike the vehicle supply optimization problem where the decision is made at real-time, the infrastructure planning for active modes is usually a long-term decision, and will have enduring impact on the transportation system. Therefore, it is especially important to assess the algorithmic fairness in the whole model development process.

Appendix A

The Synthetic Data Generation Process

The label y is generated based on the following equations:

$$U_i = V_i + \varepsilon_i = f(x_i, \boldsymbol{k_i}) + \varepsilon_i \tag{A.1}$$

$$Pr(y_i = 1) = \frac{1}{1 + \exp(-V_i)} \tag{A.2}$$

Equation A.1 denotes the true utility function specification, where ε is the extreme value distributed random term. Equation A.2 is the Sigmoid function calculating the probability of the binary outcomes. In simulations, for each individual i we draw y_i from a binomial distribution which takes value 1 with $Pr(y_i = 1)$. Noted that this is where the sampling errors may occur.

$$(a_i, x_i) \sim N\left(0, \begin{pmatrix} 1 & Cov_{ax} \\ Cov_{ax} & 1 \end{pmatrix}\right) \tag{A.3}$$

$$z_i = \begin{cases} 1, & \text{if } a_i \geq 0 \\ 0, & \text{if } a_i < 0 \end{cases} \tag{A.4}$$

$$\boldsymbol{k_i} \sim N(0, I) \tag{A.5}$$

z, x and $\boldsymbol{k}$ are derived based on Equation A.3, A.4 and A.5. a is an intermediate variable serves for the creation of z. a, x and all the elements in $\boldsymbol{k}$ are drawn from a multivariate Gaussian distribution with zero-mean and unit-variance. The above variables drawn from this distribution are all independent to each other, except that a, x are correlated with the covariance being Cov_{ax}. z takes value 1 if $a \geq 0$ and takes value 0 otherwise. Therefore, the

mean value of z is 0.5, indicating that the number of positive values and negative values of z is approximately the same. In our simulations, the value of Cov_{ax} varies across 0, 0.25, 0.5, 0.75, 1. The corresponding values of $Cov(z, x)$ are 0, 0.1, 0.2, 0.3, 0.4, each of which is empirically calculated as the value of $Cov(z, x)$ averaged across the 3 independently generated datasets used in the simulations for each Cov_{ax}.

$f(x_i, \boldsymbol{k_i})$ in Equation A.1 indicates the true systematic utility function. We consider two scenarios:

$$f(x_i, \boldsymbol{k_i}) = \alpha + x_i\beta_{x_1} + \boldsymbol{k_i}^\top \boldsymbol{\beta_{k_1}} \tag{A.6}$$

$$f(x_i, \boldsymbol{k_i}) = \alpha + x_i\beta_{x_1} + x_i^2\beta_{x_2} + \boldsymbol{k_i}^\top \boldsymbol{\beta_{k_1}} + (\boldsymbol{k_i} \odot \boldsymbol{k_i})^\top \boldsymbol{\beta_{k_2}} \tag{A.7}$$

In Scenario 1, the utility specification takes a linear form (Equation A.6). We set α to 0 and β_{x_1} to 1. Each entry of $\boldsymbol{\beta_{k_1}}$ takes {-0.5,0.5} values with equal probabilities. As such, x will have the strongest positive influence on V compared with other explanatory variables.

In Scenario 2, the utility specification takes a quadratic form (Equation A.7). We set α to -0.5, β_{x_1} to 1 and β_{x_2} to 0.5. Each entry of $\boldsymbol{\beta_{k_1}}$ and $\boldsymbol{\beta_{k_2}}$ takes {-0.5,0.5} values with equal probabilities.

This setup for both scenarios makes sure that $E(V)$ equals 0, so on average the numbers of outcomes y taking the value 1 and 0 are approximately the same.

Appendix B

Descriptive Statistics

1. *Dependent variables: travel behavior and attitude indicators*

    - Work from home: "Do you usually work from home?" (Yes=1)
    - Work from home option: "Do you have the option of working from home?" (Yes=1)
    - Travel burden: "Do you agree that travel is a financial burden?" (“Strongly agree” or “Agree”= 1)
    - Gas price impact: "Do your agree that gas price affects travel?" (“Strongly agree” or “Agree”= 1)

2. *Dependent variables: variables regarding travel mode usage*

    - Frequent usage of bike: "Do you use bike for travel daily or a few times a week?" (Yes=1)
    - Frequent usage of car: "Do you use car for travel daily or a few times a week?" (Yes=1)
    - Frequent usage of bus: "Do you use bus for travel daily or a few times a week?" (Yes=1)
    - Frequent usage of rideshare: "Do you use rideshare at least once in the past 30 days?" (Yes=1)

Variable	Mean	Std.	Min	Median	Max
Socio-demographics:					
Age	45.834	17.372	6.0	45.0	92.0
*Gender (Male=1)	0.479	0.500	0.0	0.0	1.0
*Ethnic minority (Yes=1)	0.240	0.427	0.0	0.0	1.0
*Low income household (Yes=1)	0.130	0.337	0.0	0.0	1.0
*Medical condition (Yes=1)	0.065	0.246	0.0	0.0	1.0
Driver status (Yes=1)	0.915	0.280	0.0	1.0	1.0
Education: Less than high school (Yes=1)	0.068	0.252	0.0	0.0	1.0
Education: High school graduate (Yes=1)	0.190	0.392	0.0	0.0	1.0
Education: College degree (Yes=1)	0.293	0.455	0.0	0.0	1.0
Education: Bachelor's degree (Yes=1)	0.245	0.430	0.0	0.0	1.0
Home ownership (Yes=1)	0.661	0.473	0.0	1.0	1.0
Primary activity: absent (Yes=1)	0.027	0.163	0.0	0.0	1.0
Primary activity: homemaker (Yes=1)	0.073	0.260	0.0	0.0	1.0
Primary activity: unemployed (Yes=1)	0.035	0.184	0.0	0.0	1.0
Primary activity: retired (Yes=1)	0.158	0.365	0.0	0.0	1.0
Primary activity: going to school (Yes=1)	0.067	0.250	0.0	0.0	1.0
Born in the U.S. (Yes=1)	0.861	0.346	0.0	1.0	1.0
More than one job (Yes=1)	0.075	0.263	0.0	0.0	1.0
Health level (Excellent=1, Poor=5)	2.142	0.960	1.0	2.0	5.0
Job: Sales or service (Yes=1)	0.168	0.373	0.0	0.0	1.0
Job: Clerical or administrative support (Yes=1)	0.070	0.256	0.0	0.0	1.0
Job: Manufacturing type (Yes=1)	0.087	0.282	0.0	0.0	1.0
Work for pay (Yes=1)	0.069	0.253	0.0	0.0	1.0
Level of physical activity (Never/rarely=1, Vigorous=3)	2.173	0.594	1.0	2.0	3.0
In public or private school (Yes=1)	0.024	0.154	0.0	0.0	1.0
Full-time worker (Yes=1)	0.504	0.500	0.0	1.0	1.0
Household Variables:					
Number of drivers in the HH	2.038	0.945	0.0	2.0	9.0

Note: the sample weights are used to compute the summary statistics; (*) denotes the independent variables that are also treated as the protected variables based on which we examine the prediction disparities.

Table B1: Summary statistics of the explanatory and dependent variables

Variable	Mean	Std.	Min	Median	Max
Last year's household income (K $)	84.009	64.924	5.0	62.5	250.0
Number of household members	2.915	1.435	1.0	3.0	13.0
Number of household vehicles	2.187	1.305	0.0	2.0	12.0
No child in the HH (Yes=1)	0.352	0.478	0.0	0.0	1.0
Youngest child's age $<$ 15 in the HH (Yes=1)	0.343	0.475	0.0	0.0	1.0
Youngest child's age $<$ 21 in the HH (Yes=1)	0.102	0.303	0.0	0.0	1.0
Number of workers in household	1.495	1.004	0.0	1.0	7.0
Built Environment:					
*Household in urban area (Yes=1)	0.836	0.371	0.0	1.0	1.0
Northeast Region (Yes=1)	0.178	0.382	0.0	0.0	1.0
Midwest Region (Yes=1)	0.217	0.412	0.0	0.0	1.0
Gas Price ($)	2.395	0.208	2.0	2.4	3.0
% of renter-occupied housing in the block group	31.226	22.877	2.0	20.0	97.5
Housing units/sq.m in the block group (in thousands)	3.004	5.409	0.0	1.5	30.0
Second City (Yes=1)	0.201	0.401	0.0	0.0	1.0
Suburban (Yes=1)	0.235	0.424	0.0	0.0	1.0
Small Town (Yes=1)	0.195	0.396	0.0	0.0	1.0
Number of workers/sq.m in the census tract (in thousands)	1.751	1.703	0.0	1.5	5.0
MSA population$>$ 1 million, without rail (Yes=1)	0.282	0.450	0.0	0.0	1.0
MSA population $<$ 1 million (Yes=1)	0.299	0.458	0.0	0.0	1.0
Population size of the MSA (in millions)	1.959	1.643	0.0	2.0	4.0
In an urban area (Yes=1)	0.741	0.438	0.0	1.0	1.0
In an urban cluster (Yes=1)	0.095	0.293	0.0	0.0	1.0
Area surrounded by urban areas (Yes=1)	0.000	0.022	0.0	0.0	1.0
Urban area size (in millions)	1.018	0.910	0.0	0.8	2.0
Travel Pattern and Internet Usage:					
Flexible work time (Yes=1)	0.307	0.461	0.0	0.0	1.0
Travel day began at home location (Yes=1)	0.938	0.241	0.0	1.0	1.0
Frequent internet use (Yes=1)	0.955	0.208	0.0	1.0	1.0
Dependent Variables:					
Work from home option (Yes=1)	0.184	0.388	0.0	0.0	1.0
Work from home (Yes=1)	0.116	0.321	0.0	0.0	1.0
Travel is a financial burden (Yes=1)	0.396	0.489	0.0	0.0	1.0
Gas price affects travel (Yes=1)	0.478	0.500	0.0	0.0	1.0
Frequent usage of bike (Yes=1)	0.066	0.248	0.0	0.0	1.0
Frequent usage of bus (Yes=1)	0.063	0.243	0.0	0.0	1.0
Frequent usage of car (Yes=1)	0.926	0.262	0.0	1.0	1.0
Frequent usage of rideshare (Yes=1)	0.114	0.317	0.0	0.0	1.0

Note: the sample weights are used to compute the summary statistics; (*) denotes the independent variables that are also treated as the protected variables based on which we examine the prediction disparities.

Table B2: (Cont.) Summary statistics of the explanatory and dependent variables

Appendix C

Distribution of Dependent Variables by Protected Variables

Protected Variable		Dependent Variable			
Type	Value	WFH	WFHO	TB	GPI
Race	Minority	29.73%	31.25%	46.6%	56.59%
	Majority	34.02%	33.79%	33.02%	42.98%
Gender	Male	33.52%	37.67%	34.21%	43.66%
	Female	33.19%	28.92%	35.95%	46.37%
Income	Low	37.62%	15.53%	56.81%	67.23%
	Middle or High	33.1%	34.29%	33.14%	43.07%
Medical Condition	With	49.72%	31.54%	49.2%	57.53%
	Without	32.99%	33.41%	34.1%	44.19%
Region	Urban	32.71%	34.9%	33.63%	42.71%
	Rural	35.9%	26.76%	40.67%	53.85%
	Sample Size	160110	204603	703647	703647
	Number of Positives	53400	68296	247300	317432

Note: "WFH" stands for "work from home", "WFHO" stands for "work from home option", "TB" stands for "travel burden", "GPI" stands for "gas price impact"; each percentage number indicates the proportion of positives to the total number of the corresponding dependent variable in the subset of the corresponding protected variable; since the outcome distributions of "WFH" and "WFHO" are highly skewed, for these two variables the data is balanced so that the ratio of the major outcome class to the minor outcome class instances is 2:1.

Table C1: Summary statistics of dependent variables by protected variables in the training data set

Protected Variable		Frequent Usage of			
Type	Value	Bus	Bike	Car	Rideshare
Race	Minority	55.03%	30.91%	48.35%	37.28%
	Majority	27.1%	33.83%	71.28%	32.56%
Gender	Male	33.03%	35.99%	66.65%	35.09%
	Female	33.67%	30.96%	66.6%	31.75%
Income	Low	62.76%	39.06%	29.27%	20.76%
	Middle or High	28.54%	32.82%	73.98%	34.34%
Medical Condition	With	46.37%	26.01%	38.92%	13.31%
	Without	32.21%	33.88%	69.98%	34.53%
Region	Urban	37.81%	35.8%	63.82%	37.96%
	Rural	10.79%	23.09%	78.85%	11.43%
	Sample Size	72235	130939	95323	181132
	Number of Positives	24106	43712	63504	60378

Note: since the distributions of the travel mode usage are highly skewed, for each dependent variable the data is balanced so that the ratio of the major outcome class to the minor outcome class instances is 2:1.

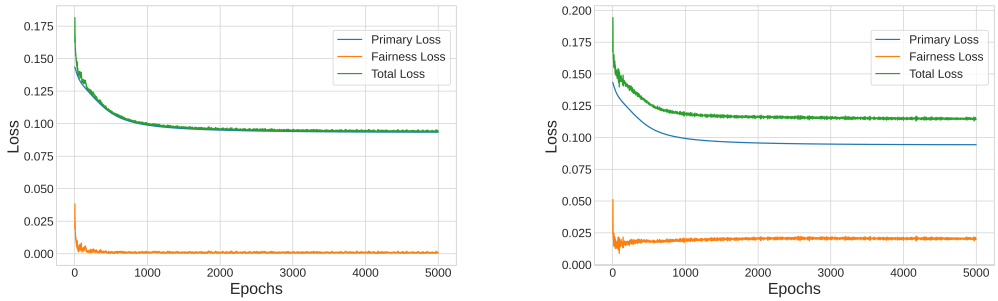
Table C2: (Cont.) Summary statistics of dependent variables by protected variables in the training data set

Appendix D

Convergence of Loss Values in the Training Process

The convergence of loss values during the training process is shown in Figure D1, where we try to mitigate the prediction disparity between rural and urban residents when predicting the frequent usage of rideshare using DNN. The bias mitigation weight is 0.8. The primary loss refers to $(1 - \lambda)$ times the cross-entropy loss which is used to increase the prediction accuracy, whereas the fairness loss refers to λ times the correlation loss which is applied to mitigate the prediction bias. The total loss is computed as the sum of the primary and the fairness loss.

From Figure D1a, we can see that in the first few epochs, the reduction of total loss is mainly driven by the reduction of fairness loss as the values of both losses drop sharply. As the fairness loss drops to nearly zero, the algorithm then primarily tries to mitigate the



(a) Training data **(b)** Testing data

Figure D1: Change of loss values in the training process. Protected variable: urban-rural divide; dependent variable: the frequent usage of rideshare; mitigation weight (λ): 0.8.

prediction loss, while the fairness loss remains very small in the following training steps. The testing data shows similar trends of loss values during the training process (Figure D1b), except that the fairness loss never drops to the near-zero level. All in all, Figure D1 illustrates how our bias mitigation method works, and how the algorithm manages to substantially reduce the fairness loss at the early stage of training.

Bibliography

[1] Principles for accountable algorithms and a social impact statement for algorithms.

[2] How uber uses data to improve their service and create the new wave of mobility, Jan 2020.

[3] Federal Highway Administration. 2017 nhts data user guide. 2019.

[4] Mahdieh Allahviranloo and Will Recker. Daily activity pattern recognition by using support vector machines with multiple classes. *Transportation Research Part B: Methodological*, 58:16–43, 2013.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[6] Yazeed Awwad, Richard Fletcher, Daniel Frey, Amit Gandhi, Maryam Najafian, and Mike Teodorescu. Exploring fairness in machine learning for international development. Technical report, CITE MIT D-Lab, 2020.

[7] Ricardo Baeza-Yates. Bias on the web, Jun 2018.

[8] Solon Barocas. Big data's disparate impact. 104:63.

[9] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the fintech era. Technical report, National Bureau of Economic Research, 2019.

[10] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[11] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459, 2019.

[12] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[13] Tierra S. Bills. Looking beyond the mean for equity analysis: Examining distributional impacts of transportation improvements. *Transport Policy*, 54:61–69, 2017.

[14] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159, 2018.

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[16] Eunshin Byon, Abhishek K Shrivastava, and Yu Ding. A classification procedure for highly imbalanced class sizes. *IIE Transactions*, 42(4):288–303, 2010.

[17] Toon Calders and Indrė Žliobaitė. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*, pages 43–57. Springer, 2013.

[18] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3995–4004, 2017.

[19] Giulio Erberto Cantarella and Stefano de Luca. Multilayer feedforward networks for transportation mode choice analysis: An analysis and a comparison with random utility models. *Transportation Research Part C: Emerging Technologies*, 13(2):121–155, 2005.

[20] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[21] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[22] Long Cheng, Xuewu Chen, Jonas De Vos, Xinjun Lai, and Frank Witlox. Applying a random forest method approach to model travel mode choice behavior. *Travel behaviour and society*, 14:1–10, 2019.

[23] Long Cheng, Xuewu Chen, Jonas De Vos, Xinjun Lai, and Frank Witlox. Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society*, 14:1–10, 2019.

[24] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[25] N Cochran. Annual update of the hhs poverty guidelines. *Department of Health and Human Services*, 2017.

[26] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[27] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pages 1436–1445. PMLR, 2019.

[28] Yu Cui, Qing He, and Alireza Khani. Travel behavior classification: An approach with social network and deep learning. *Transportation Research Record*, 2672:68–80, 2018.

[29] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[30] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

[31] Solomon Y Deku, Alper Kara, and Philip Molyneux. Access to consumer credit in the uk. *The European Journal of Finance*, 22(10):941–964, 2016.

[32] Alexa Delbosc and Graham Currie. Using lorenz curves to assess public transport equity. *Journal of Transport Geography*, 19:1252–1259, 2011.

[33] Yuntian Deng, Hao Chen, Shiping Shao, Jiacheng Tang, Jianzong Pi, and Abhishek Gupta. Multi-objective vehicle rebalancing for ridehailing system using a reinforcement learning approach. *arXiv preprint arXiv:2007.06801*, 2020.

[34] Lucas Dixon, John Li, and Jeffrey Sorensen. Measuring and mitigating unintended bias in text classification. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[35] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 2020.

[36] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[37] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[38] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

[39] Kadija Ferryman and Mikaela Pitcan. Fairness in precision medicine. *Data & Society*, 1, 2018.

[40] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 329–338. ACM Press.

[41] Pratik Gajane. On formalizing fairness in prediction with machine learning. *CoRR*, abs/1710.03184, 2017.

[42] Aaron Golub. Welfare and equity impacts of gasoline price changes under different public transportation service levels. *Journal of Public Transportation*, 13:1–21, 2010.

[43] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[44] Steven N Goodman, Sharad Goel, and Mark R Cullen. Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*, 169(12):883–884, 2018.

[45] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.

[46] Xiaowei Gu, Plamen P Angelov, and Eduardo A Soares. A self-adaptive synthetic oversampling technique for imbalanced classification. *International Journal of Intelligent Systems*, 35(6):923–943, 2020.

[47] Deniz Gunduz, Paul de Kerret, Nicholas D. Sidiropoulos, David Gesbert, Chandra R. Murthy, and Mihaela van der Schaar. Machine learning in the air. 37(10):2184–2199.

[48] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126, 2016.

[49] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, page 3315–3323, 2016.

[50] Xiuling Huang, Jie Sun, and Jian Sun. A car-following model considering asymmetric driving behavior based on long short-term memory neural networks. *Transportation research part C: emerging technologies*, 95:346–362, 2018.

[51] John Hudson, Marta Orviska, and Jan Hunady. People's attitudes to autonomous vehicles. *Transportation research part A: policy and practice*, 121:164–176, 2019.

[52] Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. *arXiv preprint arXiv:2006.11737*, 2020.

[53] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, 2011.

[54] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[55] Matthew G Karlaftis and Eleni I Vlahogianni. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399, 2011.

[56] Matthew G Karlaftis and Eleni I Vlahogianni. Statistical methods versus neural networks in transportation research:differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19:387–399, 2011.

[57] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

[58] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–16. ACM Press.

[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[60] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

[61] Michelle Seng Ah Lee. Context-conscious fairness in using machine learning to make decisions. *AI Matters*, 5(2):23–29, 2019.

[62] Orly Linovski, Dwayne Marshall Baker, and Kevin Manaugh. Equity in practice? evaluations of equity in planning for bus rapid transit. *Transportation Research Part A: Policy and Practice*, 113:75–87, 2018.

[63] Todd Litman. Evaluating transportation equity. *World Transport Policy Practice*, 8(2):50–65, 2002.

[64] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[65] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[66] Karel Martens. Substance precedes methodology: on cost–benefit analysis and equity. *Transportation*, 38:959–974, 2011.

[67] Karel Martens, Aaron Golub, and Glenn Robinson. A justice-theoretic approach to the distribution of transportation benefits: Implications for transportation planning practice in the united states. *Transportation Research Part A: Policy and Practice*, 46:684–695, 2012.

[68] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[69] Hichem Omrani. Predicting travel mode of individuals by machine learning. *Transportation Research Procedia*, pages 840–849, 2015.

[70] Miguel Paredes, Erik Hemberg, Una-May O'Reilly, and Chris Zegras. Machine learning or discrete choice models for car ownership demand estimation and prediction? In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 780–785. IEEE, 2017.

[71] Nicholas G Polson and Vadim O Sokolov. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 79:1–17, 2017.

[72] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169:866, 2018.

[73] Farideh Ramjerdi. Equity Measures and Their Performance in Transportation. *Transportation Research Record*, 1983(1):67–74, 2006.

[74] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[75] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017.

[76] Lisa Schweitzer and Abel Valenzuela. Environmental injustice and transportation: The claims and the evidence. *Journal of Planning Literature*, 18:383–398, 2004.

[77] Ch.Ravi Sekhar, Minal, and E. Madhu. Mode choice analysis using random forrest decision trees. *Transportation Research Procedia*, 17:644–652, 2016.

[78] Latanya Sweeney. Discrimination in online ad delivery. *ACM Queue*, 2013.

[79] The Department of Transportation. The department of transportation title vi program, 2019.

[80] Sander van Cranenburgh and Ahmad Alwosheel. An artificial neural network based approach to investigate travellers' decision rules. *Transportation Research Part C: Emerging Technologies*, 98:152–166, 2019.

[81] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

[82] Shenhao Wang, Baichuan Mo, and Jinhua Zhao. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies*, 112:234–251, 2020.

[83] Shenhao Wang, Qingyi Wang, Nate Bailey, and Jinhua Zhao. Deep neural networks for choice analysis: A statistical learning theory perspective. *arXiv preprint arXiv:1810.10465*, 2018.

[84] Shenhao Wang, Qingyi Wang, Nate Bailey, and Jinhua Zhao. Deep neural networks for choice analysis: A statistical learning theory perspective. *arXiv:1810.10465 [econ, q-fin]*, 2019.

[85] Shenhao Wang, Qingyi Wang, and Jinhua Zhao. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies*, 118:102701, 2020.

[86] Yuankai Wu, Huachun Tan, Lingqiao Qin, Bin Ran, and Zhuxi Jiang. A hybrid deep learning based traffic flow prediction method and its understanding. *Transportation Research Part C: Emerging Technologies*, 90:166–180, 2018.

[87] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. *arXiv preprint arXiv:1705.08804*, 2017.

[88] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[89] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM.

[90] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

[91] Zhenhua Zhang, Qing He, Jing Gao, and Ming Ni. A deep learning approach for detecting traffic accidents from social media data. *Transportation research part C: emerging technologies*, 86:580–596, 2018.

[92] Indre Zliobaite. Fairness-aware machine learning: a perspective. *arXiv preprint arXiv:1708.00754*, 2017.