Bias in Adult Income Dataset

Submitted by: Group 2

Course: DATA-6550

1. Introduction

Artificial intelligence and machine learning systems are rapdily deployed in the high-stakes decision-making contexts that directly impacts people's lives, from the automated resume screening and credit card scoring to the health risk assessment and mostly criminal justice predictions. However, combining these evidences show that these frameworks are mostly perpetuate and extend existing societal biases. In 2018, Amazon stopped using an AI recruiting tool which was systematically downgrading women's resumes. In 2019, Apple Card came across a regulatory check when its algorithm was caught offering women lower credit limits as compared to men who had similar financial portfolios. These drawbacks share a similar root cause: automated machine learning systems only learn patterns from the historical data, and when that particular data reflects historical differences through bias, inequalities, the resulting models result and automate that discrimination in the decision process.

At the main point of algorithmic bias is the training data. In this project we investigate most of these challenges using a detailed analysis of the Adult Income dataset taken from the UCI Machine Learning Repository which is one of the most widely used datasets in the machine learning community for research and education.

Using a comparative analysis of four classification models which were considered for this dataset (logistic regression, decision tree, random forest, and SVM), we demonstrate bias across multiple fairness metrics, also evaluate trade-offs between accuracy and fairness, and moreover, we proposed concrete mitigation strategies at the different levels. Our major findings show that the dataset shows severe gender imbalance (67.5% male), racial imbalance (86% White), and also income disparities (males 2.7x more likely to earn >$50K), which we tested using all models learn. Demographic parity violations were ranging from 2.4% on SVM to 19.2% with Decision Tree. The main objectives of our analysis are are briefly presented below:

1. Identify and then quantify biases in the dataset, specifically gender and racial bias in income distribution

2. Evaluate demographic distributions to reveal sampling and representation biases

3. Analyze how biases affect automated machine learning model performance across multiple algorithms

4. Compare fairness metrics across different classification models and propose mitigation strategies

2. Dataset

The Adult Income dataset which is extracted from the 1994 U.S. Census Bureau by researchers named Ronny Kohavi and Barry Becker. This UCI dataset holds 48,842

records with 14 demographic and employment features (age, education, occupation, race, gender, hours worked) are used to estimate if the annual income exceeds $50,000. We further choose this dataset because it has commonly been used in many academic papers and multiple educational courses, which makes it prominent of how bias goes through the ML models. More importantly, it holds explicit protected attributes which are race and gender from a historical period of 1994. This allows us to measure how historical differences reflect in data and goes through modern algorithms. Also, the evaluation of the demographic composition of dataset reveals significant imbalances in terms of gender and racial distribution:

Gender Distribution: Males make 67.5% of the dataset with 32,975 records, while female representation is only 32.5% which makes 15,867 records. The ratio of 2:1 clearly shows that there is sampling bias that can affect the overall model generalization.

Racial Composition: The UCI Adult income dataset consists of mostly White individuals who constitute 86.0% of the overall dataset while the Black individuals represent only 9.3% of the overall dataset. The people from Asian-Pacific Islanders have 2.9% representation in the data. Also, American-Indian-Eskimos 1.0%, and Other ethnicities are 0.8%. This overall great imbalance not just limits the ability of model to learn the patterns from minority groups but it also leads to inaccurate automated model performance.

Furthermore, the distribution of among different demographic groups shows disparities not just on gender basis but on race as well which are further discussed below in by gender and by race bullets.

By Gender: Male candidates have a total of 31.2% more earnings compared to 11.4% females and this is the earning above $50K. Comparing 31.2% with 11.4% tells us that this gap is huge and it reflects the wage inequality that the females have experienced over the past decades.

By Race: Comparing racial group suggest that Asian-Pacific Islander have the highest high-income rate which is 28.3% followed by White which is 26.2% and all other races have around 12.7%. Specifically the Black has 12.6% and American-Indian-Eskimo have 12.2% of the overall dataset. These kinds of huge differences tell us how complex historical factors like immigration patterns, and educational access have underrepresented certain races for a long time.

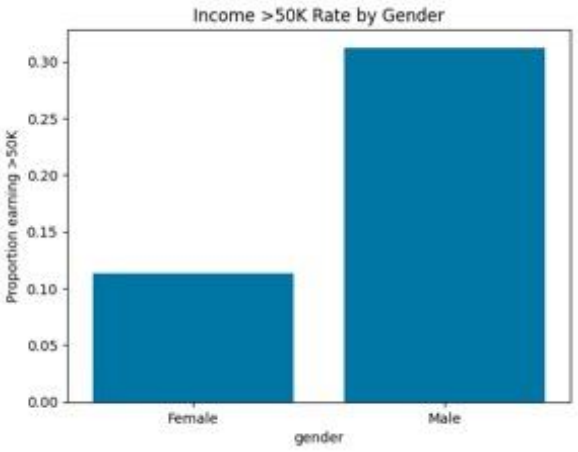

Figure 1: Income >$50K Rate by Gender

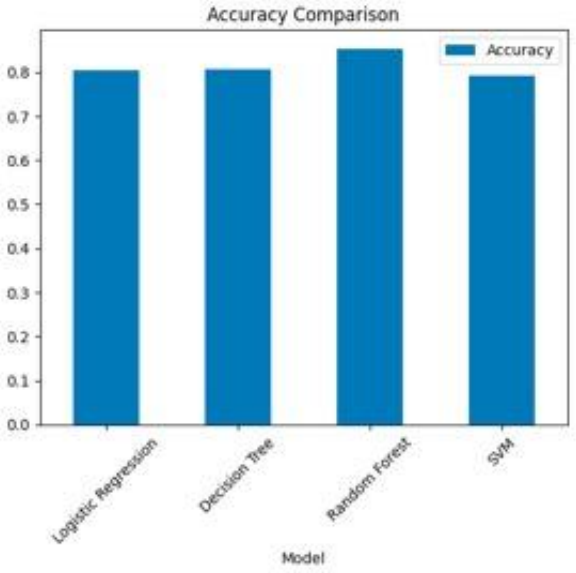

Figure 2: Income >$50K Rate by Race

3. Methodology

We started the project by doing preprocessing of the data. We observed in the dataset that the missing values were marked as '?'. We took care of these values and analysed their distribution across features. We did binary encoding for features of gender in which 0 is for female and 1 is for male and then we made the income target variable as 0 for income less than $50K and 1 income more than $50K. The dataset was well-structured although it had some missing value. Although almost 7% of the overall dataset had records which lacked occupation and work information.

We also conducted a detailed exploratory data analysis which showed us the distribution of features and made it easy for us to identify potential biases. We analyzed the features of race and gender in detail. Cross-tabulations in the features of race and gender and income clearly showed disparities that showed deeper investigation. The bar chart visualizations showed inequality in females in the low-income category.

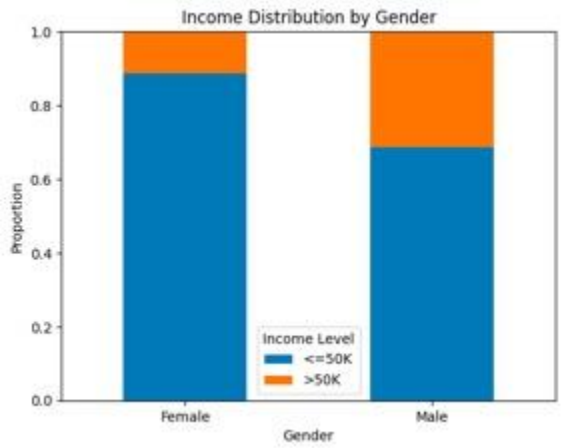

Figure 3: Income Distribution by Gender

We employed and compared four classification algorithms to analyze both predictive performance and the fairness characteristics:

Logistic Regression: A linear model is providing interpretable coefficients, which are trained with the LBFGS solver for a maximum of 1000 iterations.

Decision Tree: A non-linear model which uses recursive partitioning, and offers high interpretability through some rule-based decisions.

Random Forest: An ensemble method which combines multiple decision trees to improve the generalization and also reduce overfitting.

Support Vector Machine (SVM): A maximum-margin classifier which finds the optimal decision boundary in the feature space.

The dataset was split into 80% training and 20% testing sets using the stratified sampling to maintain class proportions in our dataset. All models were employed using scikit-learn library with the default hyperparameters to establish baseline performance.

To evaluate bias, we computed two main fairness metrics:

Demographic Parity Difference: Measures the positive prediction values between males and females. A value of 0 shows perfect demographic parity, whereas larger absolute values show greater disparate impact. This metric evaluates either the model's predictions are fairly distributed across the protected groups.

True Positive Rate (TPR) Difference: Also called the equal opportunity difference, this evaluates the difference in true positive rates among males and females. It analyzes whether people who truly earn >$50K are equally likely to be identified regardless of sex. Smaller values give fairer treatment of qualified candidates among groups.

4. Analysis and Results

Underconsidered four classification models achieved good accuracy scores, with Random Forest also known as RF gave best overall. However, only the metric of accuracy hides the significant fairness concerns:

Model	Accuracy	Demographic Parity Diff	TPR Diff
Logistic Regression	80.4%	0.119	0.181
Decision Tree	80.6%	0.192	0.060
Random Forest	**85.3%**	0.178	0.049
SVM	79.3%	**0.024**	**0.047**

Table 1: Classification models Performance Metrics

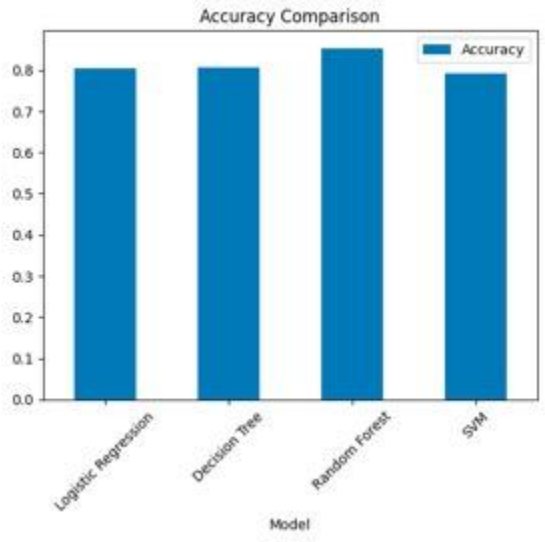

Figure 4: Model Accuracy Comparison

The metric of fairness analysis shows a critical trade-off among accuracy and fairness. Random Forest goes to the highest accuracy of 85.3% but it shows prominent demographic parity violation which is 0.178, which means that males are 17.8% more likely to be predicted as the high earners than females candidates. On the other hand, SVM achieves the overall best fairness value, with only a 2.4% change in demographic parity, but accuracy of SVM is greatly low which is only 79.3%.

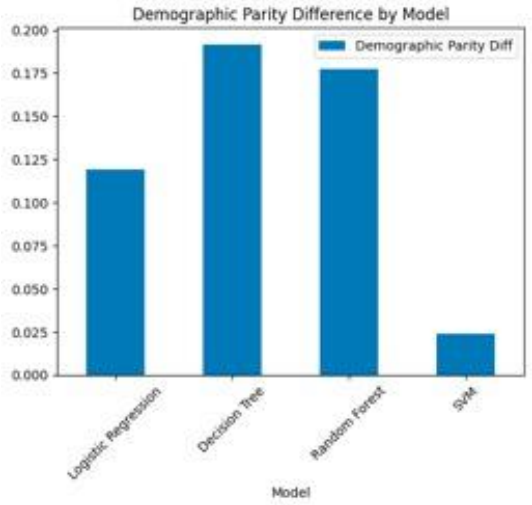

Figure 5: Demographic Parity Difference by Model

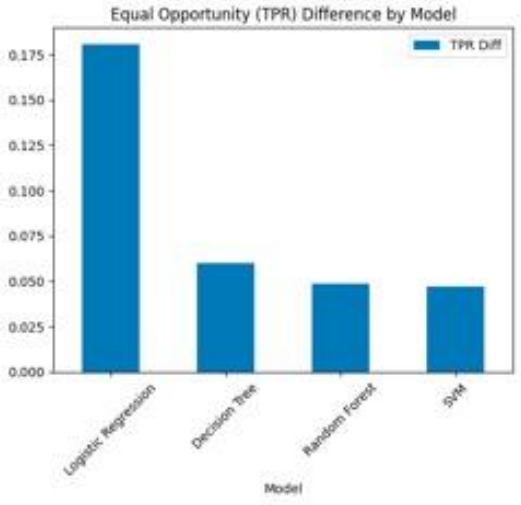

Figure 6: Equal Opportunity (TPR) Difference by Model

The decision tree reflects the most bad demographic parity (0.192), it learns and extends gender bias strongly. However, its performance is good on Equal Opportunity (TPR difference of 0.060), which shows that, among high earners or people earning more than 50K, it treats both genders similarly. This shows that different fairness metrics can demonstrate different stories about the model behavior.

If we consider the gender bias analysis of the dataset further strengthens the statement that all models somehow favor males in the income predictions. SVM which proved itself as the fairest model also exhibits measurable bias. This rose from the data distribution, in which the males are overrepresented in high-income individuals and higher educational levels. We employed four models and they learned from these patterns during the time of training and mainly what it does is it passes historical discrimination into decisions through automated systems.

If we consider the TPR metric, we can see that the TRP difference evaluation also says that the logistic regression gives the largest equal opportunity violation which is 0.181, and simply means that among individuals who really earn more than $50K, males are 18.1% highly likely to be accurately predicted than females. This type of bias is specially concerning in fields like hiring or promotion decisions. In these jobs it means that qualified female candidates in the pool are more likely to be overlooked.

The UCI Adult income dataset clearly shows the patterns of segregation by gender in terms of occupation. Men are in higher numbers in higher-paying roles like technical and managerial, on the other hand females are considered in lower-paying services and administrative roles. As we can see this distribution we can say that this difference in income reflects a structural bias hidden in the labor market data. When

automated models like machine learning train on this data, they encode these historical patterns of discrimination, potentially reflecting them into the future estimations.

Also there is a racial imbalance in the dataset as well which says that 86% are White individuals and this creates many problems. Number one, the models have inadequate data to learn and train on true patterns for minority groups, resulting in higher error rates for these groups. Second, the similarities of race and gender gives compound drawbacks; minority women experience the most underrepresentation and experience the big prediction errors. Third, the high-income values of Asian-Pacific Islander individuals at 28.3% and White at 26.2%, show higher than Black 12.6% showing historical economic inequalities that these models will learn and then use in future predictions.

5. Discussion

After carrying out this detailed analysis we can say that there are a lot of fundamental challenges in creating fair automated machine learning systems. We spotted in our analyses that there is a prominent accuracy-fairness trade-off: Random Forest had the highest accuracy of 85.3% but with a bias of 17.8% demographic parity difference. Now if we consider SVM we can see that SVM showed better fairness of 2.4% demographic parity difference at the cost of lower accuracy which is 79.3%. This requires some context-dependent decisions which are: high-stakes fields like hiring may prove accepting lower accuracy for the fairer outcomes. Although, fairness itself cannot be considered singular: The decision tree in our study showed poor demographic parity of 19.2% yet gave a nice equal opportunity showing in 6.0% TPR difference, reflecting that different fairness metrics can conflict with each other. Data scientists must choose criteria closely aligned with their task which can ensure equal opportunity for employment, and also achieve demographic parity in the resource allocation.

Deployment of classification models which are trained on this kind of data holds serious risks. Even if gender is eliminated as a main feature, education, field of work and hours in the field work as proxies, which enables systematic discrimination. The unalignment in the data collected in 1994 and the current model execution is equally problematic. Female laborers participate and their occupational distribution has changed greatly over past decades, which means that models trained on historical data risk on random patterns show outdated discrimination and violate the equal opportunity laws. The data scientists must evaluate if certain estimation tasks should be carried out through automated systems, if the data in hand is valid for the intended use, and what actual safeguards must be put in place to prevent any harm. The use of the dataset, which is Adult Income dataset, in ML research has proved as the deployment of biased models in real systems.

6. Bias Mitigation Strategies

Addressing the bias needs observations at multiple levels. Let's talk about the data level first, we can balance the dataset using resampling and reweighting minority groups like females and racial minorities, supplement 1994 census data. Then at the model level, we can use fairness constraints during training to use prominent libraries like Fairlearn or AIF360 to make sure demographic parity. Results presented in this report suggest that fairness achieved using SVM could also be achieved by random forest if we maintain appropriate constraints. Some adversarial debiasing trains models with important components that penalize estimation, which forces the primary model to give predictions that don't reveal gender or race. Optimization applies different automated decision thresholds for different groups, while maintaining the overall accuracy to improve fairness without any retraining.

We come to the **process level** afterwards, we then build the bias audits to measure the model performance among different subgroups with the help of multiple fairness metrics. Fairness definitions go along if we involve the representatives from the communities using focus groups with experiences and identify any harm that other metrics may miss out. Then finally a **hybrid approach** is also proposed which can build on the basic concept of data augmentation while using recent and balanced datasets to lower the historical and sampling bias. Also, we can apply fairness constraints during the training to target demographic parity less than 5% and TPR difference less than 5%. Moreover, we can use threshold optimization to fine-tune the fairness criteria without sacrificing excessive accuracy; and implement the fairness auditing system if metrics go bad. This multi-layered approach takes care of bias at not just the data level but also at model and process levels to maximise the chances of achieving both accurate and fair estimations while maintaining accountability.

7. Conclusion

This overall analysis of the dataset using different machine learning algorithms showed a great gender and racial bias among all different models. Males are estimated to make above $50,000 which is 2.7 times higher than females which shows pure dataset imbalances. Our overall analysis also showed a fundamental trade-off among accuracy and fairness. Random Forest although gave the highest accuracy of 85.3% with significant bias of 17.8% demographic parity gap, while SVM came up with the best fairness of 2.4% but at the lower accuracy of 79.3%.

We presented some actionable mitigation strategies which are data augmentation, fairness-constrained learning and the threshold optimization along with ongoing auditing. Moreover, just presenting some technical solutions is not sufficient. Meaningful progress needs commitment to fairness, transparency on the trade-offs, and willingness to redo the deployment phase when the fairness metric cannot be

assured. As today, AI systems greatly influence employment decisions so it is important to identify and mitigate bias as an ethical obligation. Since the dataset is from 1994, we can say that it holds historic bias. We should take the responsibility of building our systems in such a way that we can ensure an equal opportunity future for all the individuals.