Fair Data Projection for Mitigating Bias in Law School Bar Passage Rate Predictions

Julie Hinge (juhi@itu.dk)
IT University of Copenhagen, July 29, 2024

1 Introduction

Machine learning is increasingly used in critical decision-making processes such as application filtering, crime prediction, and medical diagnostics [1]. An estimation made by *Intelligent* in 2023 screened 400 education professionals and found that an estimated 50% of higher education admissions offices are using AI to screen applicants as an efficiency tool , and that this number was expected to rise to 80% in 2024 [2]. While it can be considered true that these methods can save companies or institutions a great number of resources, the practise of using AI to evaluate real humans can do more societal harm as these AI systems are often encoded with structural biases [3]. Furthermore, modern AI systems primarily look for correlations in large datasets without understanding the reasons behind them, which can perpetuate and even amplify existing biases. These correlations, while useful for identifying trends, do not inherently explain the underlying causes [4].

Bias in machine learning typically refers to any systematic error made during either data collection, feature engineering, training, or deployment [5]. However, in this report, the definition of bias will have a closer relation to how we know it in everyday life which, refers to prejudice toward or against one person or group based on their characteristics.

To mitigate bias, it is essential for evaluation tools that make use of machine learning to incorporate fairness into their algorithms. Fairness, in this context, means the absence of any prejudice towards an individual or group. This paper explores how frameworks for evaluating the bar passing rates of law students may be biased against marginalized groups and examines whether this bias can be reduced by integrating fairness into machine learning models.

2 Background

2.1 How can machine learning models be biased?

In machine learning, predictions are based on features, which vary depending on the model and use case. For instance, when predicting outcomes such as whether law school students will pass the bar exam, data about their lives and characteristics are often utilized. In fairness analysis, these sensitive attributes could include characteristics such as race, gender, ethnicity, religion, or socioeconomic background, and we refer to them as **protected variables**. Bias in machine learning can originate from:

1. **Historical Bias:** Prejudices reflected in historical prejudices against sensitive groups, such as the segregation of African-Americans or the perception of women as unfit for certain jobs. These biases are reflected in the training data, and when models are trained on this data, they amplify these biases in decision processes. An example of this is Amazon's sexist hiring algorithm [6].

2. **Feedback Loops:** Model outcomes influence future data, potentially reinforcing biases. For instance, social media platforms use user interactions to recommend content, which then influences further user interactions. This can lead to an echo chamber, underrepresenting diverse perspectives [7].

3. **Unbalanced Samples:** Skewed training data leads to poor performance on minority classes. If the minority class represents a protected group, the model's biased predictions perpetuate and amplify societal biases, leading to unfair treatment.

4. **Proxy Variables:** Proxy variables are features closely related to protected features. Using these can lead to unfair models, as they effectively allow the model to use protected variables for predictions.

While these are not the only sources of bias, they are common issues that are often overlooked when developing machine learning models that impact real people's lives.

This study will employ a dataset called **Law School Admissions Bar Passage**. which has been constructed by the Law School Admissions Council (LSAC) and collects data from 27.000 law students through law school, graduation, and sittings for bar exams from 1991 through 1997. The dataset was originally used in a study by Linda Wightman in 1998 to investigate if anecdotes that passing rates were low among colored students were correct [8]. These anecdotes stated that the use of law school resources to support minority special admission programs (programs to increase enrollment of students of color) were not justifiable. According to the critics, the great majority of special admission students were either flunking out of law school or unable to pass the bar examination.

Now imagine if law schools used this data to train machine learning models for predicting bar exam success, identifying at-risk students, and assessing the effectiveness of admission policies. Here, historical biases present in the data could influence the outcomes, making the machine learning model biased as well. The dataset may reflect earlier inequities in educational opportunities, socio-economic disadvantages, and institutional biases. Additionally, minority students from the 1990s could have faced gaps in preparation due to disparities in primary and secondary education, access to LSAT preparation resources, and support systems crucial for success [8].

2.2 How can fairness be achieved?

To address fairness, we must first define what it means to be fair. In machine learning literature, there are several definitions of fairness, often categorized into group fairness, subgroup fairness, and individual fairness. [9].

This paper uses the methodology proposed by [10] to learn a fair lower-rank representation of the data independent of protected features. Any model trained on this representation should then be considered as a fair model. This method is called fair PCA and by adopting this approach, I aim to ensure that the trained models do not exhibit bias toward any specific group, thereby aligning with the broader goals of fairness in machine learning

This paper will use two protected groups: Race and Gender. However, as the original study focused purely on race, I thought it would be interesting to also focus mainly on race to investigate if machine learning models could exhibit biased behaviors towards people of color, specifically Black people.

3 Exploratory Data Analysis

As aforementioned, this paper will employ the dataset **Law School Admissions Bar Passage**, tracking 27,000 law students from 1991 through 1997 [8].

The data needed quite a lot of cleaning as there were many similar columns or incomplete columns. A description of the columns that were dropped and why can be found in Appendix 3. After discarding irrelevant and incomplete columns, the columns which remained were can be found in table 1.

Variable	Description
decile1, decile1b, decile3	Describes the law school ranking by decile of each candidate in year 1 (semester 1, semester 2) and year 3. I'm not sure why the remaining years are not included though.
lsat	The LSAT score of each candidate (although it has been altered somehow as it's not within the normal range, but it still correlated with the target (passed/not passed) so therefore I decided to keep it).
grad	This is a binary column describing whether a student graduated or not.
fulltime	If the student is a full-time student.
fam_inc	Family income by quintile.
tier	What tier law school did the student attend by quintile.
race1	The race of the student.
sex	The gender of the student.
pass_bar	This is the target variable. Did the student pass the bar.

Table 1: Description of Dataset Variables

Out of all students, 94.8 % passed the bar and 5.2% failed. The percentages of student who passed the bar within the race groups can be seen in Fig 1 and a corresponding figure for the gender group can be seen in Appendix 5.

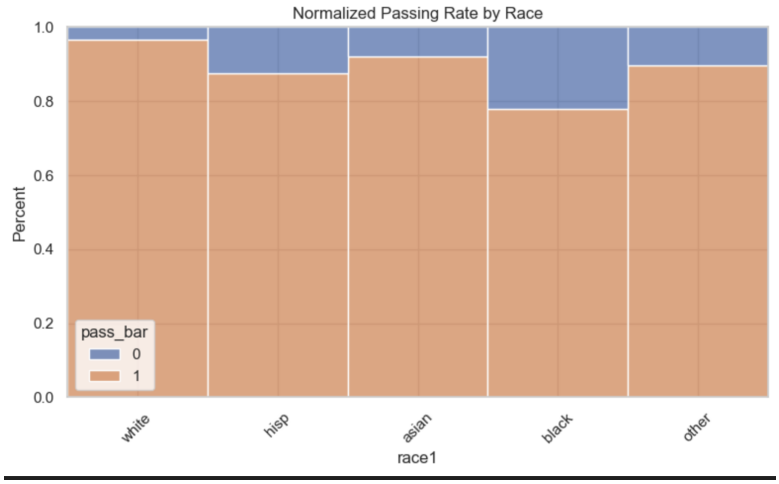

Figure 1: Passing rate by Race

In Fig 1, we do see that significantly more Black people failed the bar compared to other races.

4 Methodology

4.1 Baseline model

A baseline model using a Random Forest classifier has been chosen to predict whether a student will pass the bar or not. The advantages of using Random Forest are that it is a robust ensemble model capable of handling non-linearity between features, while still providing interpretability through feature importance measures[11]. The raw dataset is unbalanced, meaning that far more students pass the bar than fail it. To address this, I have set **class_weight='balanced'** in the Random Forest model so it adjusts the loss function by giving more importance to minority class samples and less importance to majority class samples during training. The model does this by internally adjusting the weights assigned to each class during training, specifically by computing class weights as $w_j = \frac{n}{k \cdot n_j}$ where: n is the total number of samples, k is the number of classes, n_j is the number of samples in class j.

4.1.1 Standard PCA

Another model will be trained on the reprojected version of the data using standard Principal Component Analysis (PCA). PCA transforms the data into a lower-dimensional space by projecting it onto the directions (principal components) that maximize variance. This approach is included to evaluate the performance of a more straightforward dimensionality reduction technique compared to more advanced methods.

4.1.2 Metrics

It was decided to use False Negative Rate (FNR) as a benchmark for measuring the fairness of the employed methods. This is due to the heavy impact that false negatives can have in the context predicting students' passing rates. A false negative occurs when a student who will pass the bar has been predicted as failing it which can lead to significant negative consequences for the individual. These missed opportunities can disproportionately affect certain demographic groups, leading to systemic inequalities. By focusing on FNR, I aim to ensure that all demographic groups have an equal chance of being correctly identified as passing the bar. Lowering the disparity in FNR between groups is crucial because it directly addresses the fairness and equity of the decision-making process. While other metrics like False Positive Rate (FPR) or accuracy could also be considered, FNR is particularly relevant in this scenario due to its direct impact on the students' opportunities.

4.2 Fair PCA

This analysis uses Fair PCA per Section 3.4 of [10] for multiple demographic groups. fair PCA extends the standard PCA algorithm to incorporate fairness by mitigating biases towards protected features. Given some data $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n is the number of samples and d is the number of features, we want to find a projection matrix $\mathbf{U} \in \mathbb{R}^{d \times k}$ that maximizes the variance of the projected data (just like standard PCA) while mitigating bias towards the protected features (extension of standard PCA). This can be described as the following optimization problem:

$$\arg \max_{\mathbf{U} \in \mathbb{R}^{d \times k}, \mathbf{U}^T\mathbf{U} = \mathbf{I}_{k \times k}} \text{trace}(\mathbf{U}^T\mathbf{X}^T\mathbf{X}\mathbf{U}) \quad \text{subject to} \quad \mathbf{Z}^T\mathbf{X}\mathbf{U} = 0,$$

where $\mathbf{Z} \in \mathbb{R}^{n \times p}$ represents the protected features, with each column of $\mathbf{Z}$ corresponding to a different demographic group. The constraint $\mathbf{Z}^T\mathbf{X}\mathbf{U} = 0$ makes sure that the principal components are orthogonal to the protected features, which is responsible for reducing bias [10].

To implement this, we first compute the null space of $\mathbf{Z}^T\mathbf{X}$, which we denote as $\mathbf{R} \in \mathbb{R}^{d \times (d-m+1)}$, where m is the number of protected groups.

Next, we find the top k eigenvectors of the matrix $\mathbf{R}^T\mathbf{X}^T\mathbf{X}\mathbf{R}$ which we define as $\mathbf{\Lambda}$. This is the matrix consisting of the top k eigenvectors, which are selected based on the largest eigenvalues. The matrix $\mathbf{\Lambda}$ defines the optimal projection within the subspace spanned by $\mathbf{R}$.

Finally, we construct $\mathbf{U}$ as $\mathbf{U} = \mathbf{R}\mathbf{\Lambda}$. This projection matrix $\mathbf{U}$ maximizes the variance of the projected data while ensuring that the fairness constraint $\mathbf{Z}^T\mathbf{X}\mathbf{U} = 0$ is satisfied.

4.3 SHAP

To see which features contributed most to the outcome, one can use shapley values. This method can be used to see if sensitive attributes like race or gender have a larger predictive power and if they contribute positively or negatively in a machine learning model, which they ideally shouldn't. This method computes an importance value for each feature based on how the model performs with that feature and without it. These two model predictions are compared by subtracting the model score without the feature involved and with the feature (the model score is simply the metric chosen for model evaluation, in my case accuracy score).

The method is based on the Shapley values, which equation is as follows [12]:

$$\phi(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(v(S \cup \{i\}) - v(S)\right)$$

Here,

- $v(s)$ denotes the model's prediction for a specific subset of features.
- N is the set of all features.
- $S \subseteq N \setminus \{i\}$ denotes the set of features excluding feature i.
- $|S|$ is the features without feature i.
- $\sum_{S \subseteq N \setminus \{i\}}$ denotes the model's prediction when the feature i is included in the subset S. i.e., it's the model's output when both the features in S and the feature i are considered.

Using the SHAP library, I train a Random forest classifier using the training data and then obtain Shapley scores using the testing data. These values can be used to plot a butterfly plot showing the ranked absolute feature importance.

4.4 Counterfactuals

As another method of evaluating whether the fair PCA transformation improves fairness, I propose a counterfactual analysis using the top *n* observations that had nearly equal probabilities of being predicted as passing the bar exam.

The process is as follows:

1. **Identification of Ambiguous Observations**: From the original (non-transformed) dataset, identify the top *n* observations where the Random Forest model predicted probabilities close to 0.5 for both classes. These observations represent cases with the highest classification ambiguity.

2. **Apply fair PCA Transformation**: Transform the dataset using fair PCA to obtain reprojected data aimed at reducing bias while preserving relevant information.

3. **Predict Using Reprojected Data**: Use the Random Forest model to predict class probabilities for these top n observations in the reprojected dataset.

4. **Compare Predictions**: Assess if any candidates were incorrectly classified as failing in the baseline model, compared to the model using the fair PCA reprojected data.

5 Results

5.1 Baseline Model, PCA projected and Fair PCA projected results

Table 1 shows the comparison of accuracy scores using standard data, PCA-projected data, and fair PCA-projected data. The baseline model achieved 82% accuracy and a 17% false negative rate (FNR). Standard PCA reached 85% accuracy and a 13% FNR, while fair PCA scored 79% accuracy and a 19% FNR.

Method	Accuracy	FNR
Standard Data	0.82	0.17
Standard PCA	0.85	0.13
Fair PCA	0.79	0.19

Table 2: Comparison of accuracy and False Negative Rate (FNR) for different methods.

The results indicate a trade-off between accuracy and False Negative Rate (FNR) when applying fair PCA compared to the baseline method. The baseline method achieves a slightly higher accuracy and slightly lower FNR. This implies that while fair PCA may address fairness concerns, it does so at the cost of reduced accuracy and an increased rate of false negatives. This trade-off highlights the challenge of balancing fairness with model performance.

5.1.1 Groups

The comparison of the FNR for all methods for each of the subgroups within the protected feature Race, can be seen in fig 2.

In the baseline model, FNR is highest for Black and Hispanic students, second highest for "other," third for Asian, and lowest for White students. These results clearly show the need for fairness being implemented in the model. After implementing the fair PCA algorithm, the FNR has been lowered for the groups that were the most affected by it in the baseline model (i.e. Black, Hispanice, and Other), while it has been slighly increased for the Asian and White group. Overall though, it has balanced the FNR across all groups. The PCA projected data model shows that the FNR has significantly increased

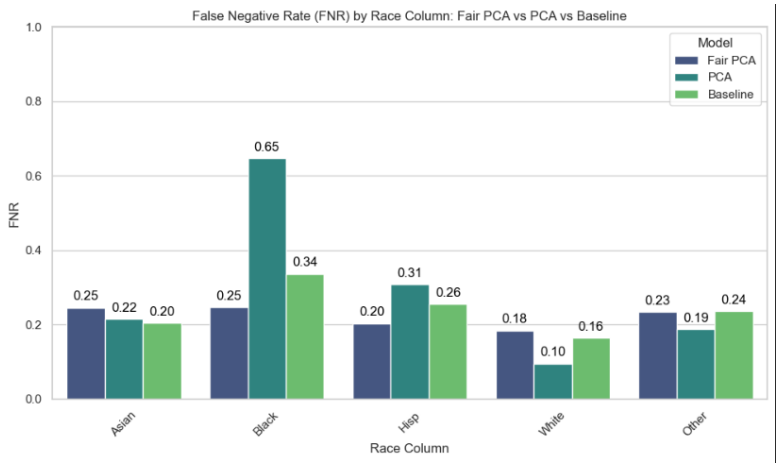

Figure 2: False Negative Rate by Race: FairPCA vs. baseline

for the Black and Hispanic subgroup and has decreased for the White subgroup.

It is also worth mentioning that the accuracy was highest for White students (83%) and lowest for Black students (67%) in the baseline model. After implementing fair PCA the accuracy decreased for all groups. A corresponding plot for the accuracy scores within each group can be found in Appendix 6, as well as corresponding plots for the gender group (Appendix 7, Appendix 8).

5.2 Correlation matrices

To ensure that the projection of the data onto the null space created by the projected features were correct, I evaluated the correlation between the projected data and the protected features. To visualize the impact of debiasing, I compared the correlation matrices from the standard PCA algorithm provided by Scikit-learn with those from my fair PCA implementation, as shown in Fig 3.

(a) Correlation between principal components and protected groups using standard PCA

(b) Correlation between principal components and protected groups using fair PCA

Figure 3: The correlation between the principal components and the protected features, for the PCA and Fair PCA

5.3 SHAP Feature Importance

The SHAP butterfly plot illustrating feature importance on the model is shown in Fig 4. We see that most features have an intuitive impact on the model. For instance, it is beneficial for passing the bar

to be in a high decile, have a high LSAT score, and undergraduate GPA, and attend a law school from a higher tier. These columns also have the most significant impact on the model's outcome.

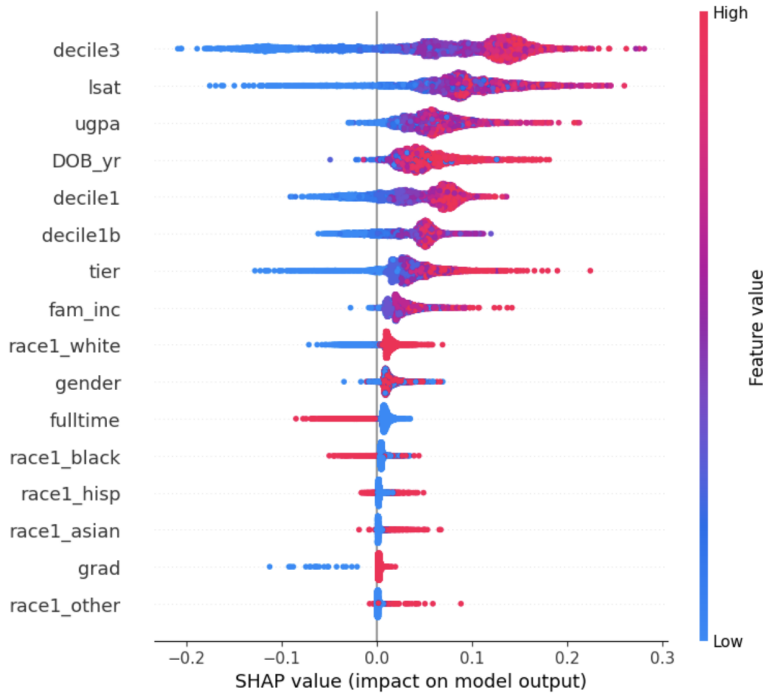

Figure 4: SHAP butterfly plot illustrating feature importance

To clearly illustrate how belonging to a specific race influences the overall prediction, it was decided not to group the individual races together. The SHAP butterfly plot in Fig 4 shows that not identifying as Hispanic, Asian, or Other has no impact on the model, whereas being a member of these races slightly positively affects the outcome. On the other hand, being a White student positively impacts the prediction, while being a Black student has a mixed but predominantly negative impact.

Regarding the gender column, the impact is mixed, with slightly more female observations contributing positively to the outcome. These results are consistent with the initial evaluations using the baseline Random Forest model.

5.4 Counterfactuals

To assess whether the fair PCA transformation results in different predictions for the most balanced observations, I compared the predictions for the top 10 observations (arbitrary number) with nearly equal probabilities from the original dataset against those from the fair PCA transformed dataset.

Among these top 10 observations, 9 were initially predicted as not passing the bar (class 0) and with the fair PCA data, they were predicted as passing the bar (class 1). Notably, half of these observations predicted as passing after the transformation were from non-white racial groups. It is also important to mention that only 3 of these 9 observations had an actual outcome of passing the bar.

6 Discussion and Conclusion

This paper aimed to incorporate fairness in the prediction of bar passage rates among law students, focusing on mitigating biases related to race and gender. The baseline model, which did not account for fairness, revealed imbalance in prediction accuracy and false negative rates across different racial groups showing the highest false negative rates for Black and Hispanic students, indicating a bias against these groups. These results show the need to incorporatate fairness meassures into predictive models before applying them to real-life situations.
By implementing Fair PCA, I observed an improvement in balancing the false negative rates across racial groups. While the overall accuracy of the fair PCA model was lower compared to the baseline and standard PCA models, the fair PCA method successfully decreased bias toward racial groups. This trade-off between accuracy and fairness is a common challenge in machine learning, but the results demonstrate that it is possible to create more equitable models by sacrificing a degree of accuracy.
SHAP analysis confirms that sensitive attributes like race and gender impact model predictions, emphasizing the need for fairness constraints in model training to reduce bias.

From a philosophical view, the reliance that companies and institutions are starting to have on big data and machine learning demonstrates a departure from the GOFAI (Good Old Fashioned Artificial Intelligence) paradigm, which was more based on rule-based algorithms. Modern AI's approach instead prioritises pattern recognition over explicit reasoning, and although big data and deep learning often provide quite reliable insights, they do not inherently offer explanations for why correlations exist. This shift shows how important it is to have some human oversight in interpreting AI results, including of fairness methods and metrics, and ensuring that these models do not perpetuate or amplify existing biases[4].

In conclusion, this study demonstrates that integrating fairness into machine learning models is crucial for addressing biases in educational predictions. While there is a trade-off between accuracy and fairness, the benefits of creating equitable models that do not disproportionately disadvantage any group, outweigh the decrease in accuracy. Future research could be conducted to explore other fairness-enhancing techniques to ensure that predictive models serve all demographic groups fairly.

References

[1] Kanisha Karunakaran. The role of machine learning in automating decision-making processes. https://www.latentview.com/blog/the-role-of-machine-learning-in-automating-decision-making-processes/, May 2023.

[2] Cole Claybourn. Is ai affecting college admissions? https://www.usnews.com/education/best-colleges/articles/is-ai-affecting-college-admissions, 2023. US news.

[3] UNESCO. Artificial intelligence: examples of ethical dilemmas. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases, April 2023.

[4] Pawel Grabarczyk. Philosophy and big data. Slide Presentation, 2024. Algorithmic fairness, Pawel Grabarczyk, Spring semester, Lecture 13.

[5] Mary Reagan. Understanding bias and fairness in ai systems. https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3, March 2021.

[6] Maude Lavanchy. Amazon's sexist hiring algorithm could still be better than a human. https://www.imd.org/research-knowledge/digital/articles/amazons-sexist-hiring-algorithm-could-still-be-better-than-a-human/: :text=Amazon

[7] Mike Loukides. The biggest problem with social media has nothing to do with free speech. https://qz.com/1714598/information-feedback-loops-make-social-media-more-dangerous, 2019. Quartz.

[8] Linda F. Wightman. *LSAC National Longitudinal Bar Passage Study*. LSAC research report series. Law School Admission Council, 1998.

[9] Zeyu Tang and Kun Zhang. Attainability and optimality: The equalized odds fairness revisited, 2022. URL https://arxiv.org/abs/2202.11853.

[10] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair pca for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5250–5270. PMLR, 2023.

[11] Christoph Molnar. Interpretable machine learning. In *Interpretable Machine Learning*, chapter 5.4. Christoph Molnar, 2024.

[12] Aditya Bhattacharya. *Applied machine learning explainability techniques: Best practices for making ML algorithms interpretable in the real-world applications using lime, shap and others*. Packt Publishing, 2022.

7 Apendix

Column Name	Description
dnn_bar_pass_prediction	Models prediction of whether or not the student will pass
Cluster	Unable to find a description and unclear representation
index6040, indxgrp, indexgrp2	Unclear what these columns represented
gpa, zgpa, zfygpa	Completely correlated with UGPA (undergraduate GPA); redundant
bar1, bar1_yr, bar2, bar2_yr, bar_passed	Provides nuanced information about student passing; redundant
race, race2, asian, black, hisp	Redundant; race1 is sufficient as others have missing values
Dropout	Column consisting of zeroes only; redundant
parttime	Correlated perfectly with "full-time"; dropped
age	Contains many negative values, which are not plausible
male, sex	Redundant as gender column contains this information

Table 3: Descriptions of columns and the reasons for their exclusion from the dataset.

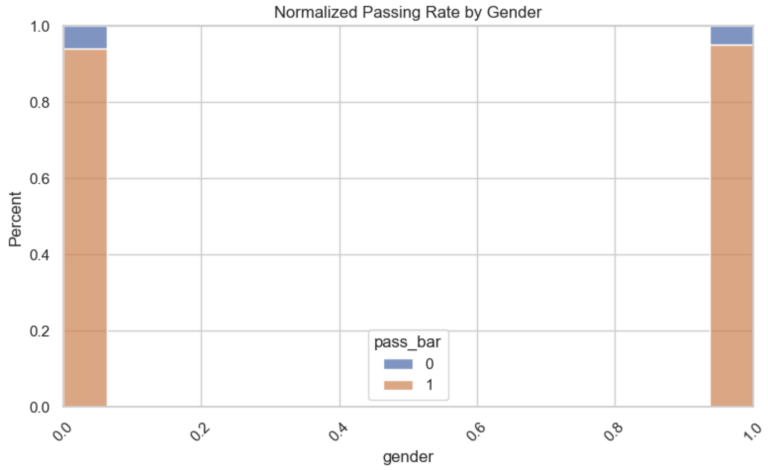

Figure 5: Passing rate by Gender

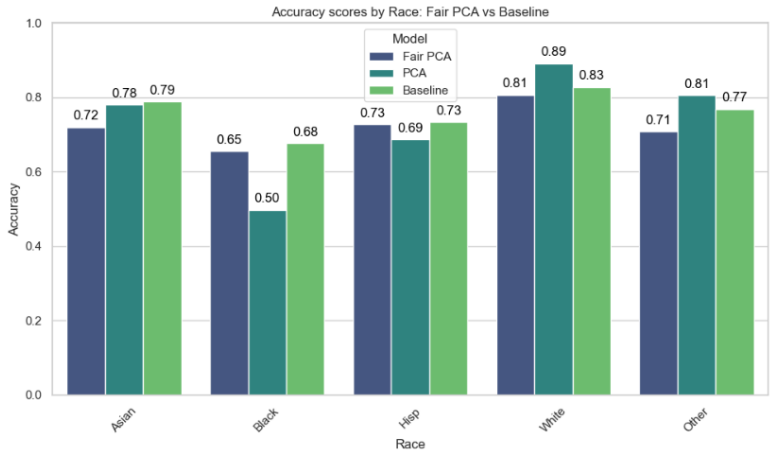

Figure 6: Accuracy by Race for baseline model, PCA projected data, and fair PCA projected data

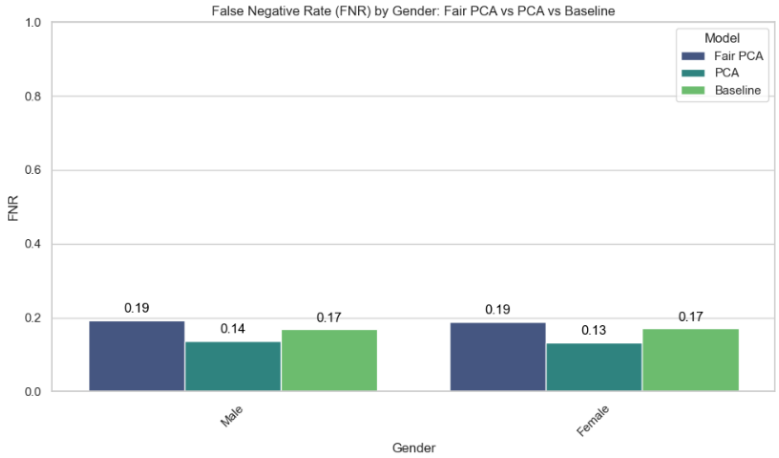

Figure 7: FNR by Gender for baseline model, PCA projected data, and fairPCA projected data

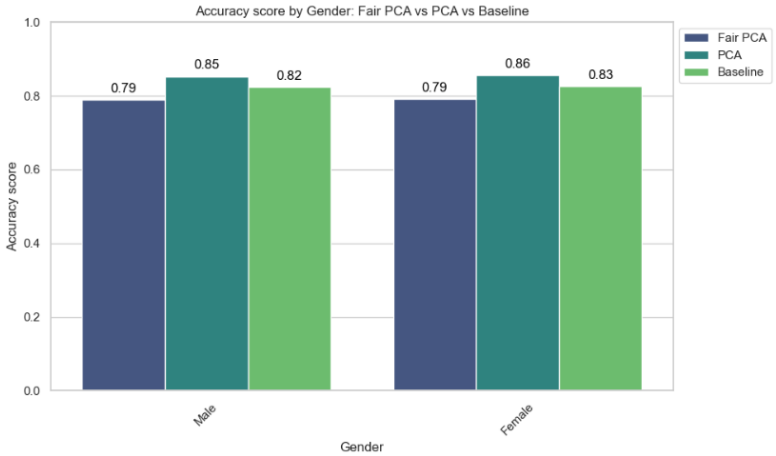

Figure 8: Accuracy by Gender for baseline model, PCA projected data, and fair PCA projected data