Performance, Fairness, and Explainability in Automated Mortgage Loan Application Decision-Making

Simon M. Breum smbr@itu.dk
Daniel V. Egdal daeg@itu.dk
Victor G. Mortensen vmor@itu.dk
IT University of Copenhagen
May 19, 2023

1 Abstract

Using 2017 HMDA data from Tennessee, we train a decision-making system for predicting mortgage loan application approval. We use data exploration techniques to highlight potential biases in the data, and use FairPCA to de-bias the data. Using equal FNRs as a fairness metric, we show that FairPCA can be used to de-bias the data, leading to a more fair model. However, doing so reduces model performance and obscures model explanations. We explore the performance vs. fairness trade-off version of FairPCA in an attempt to find a model with good performance and fairness, however, we find that there is no point where performance increases before fairness decreases. Furthermore, we discuss the impact of this method on model explanations, in particular how it obscures the explanations. Lastly, we discuss issues and concerns regarding the deployment of automated decision-making systems - particularly the danger of blindly accepting their decisions as truths without fully understanding the reasoning behind their decisions. This discussion is accompanied by discussion of the ethical implications of using such systems.

2 Introduction and Previous Works

Deciding which applicants should be granted a mortgage loan is an important decision for lenders, as giving loans to borrowers who will not be able to pay back loans may result in great monetary losses. This task can be viewed as a classification task of loan applicants into "good" and "bad" applicants. From the perspective of the lender, the goal is to accept as many "good" applicants as possible, while rejecting the "bad" ones. Machine learning techniques can be deployed with the purpose of automating the process of evaluating mortgage applications, and systems have already been developed to estimate credit risk [1].

While potentially beneficial for the bank, there have been multiple examples highlighting the threat automated decision-making systems may pose to minority groups [2, 3]. In a mortgage loan application setting, a concern might be that the machine learning algorithm learns to separate "good" applicants from "bad" ones by using information about so-called *protected attributes*, e.g., it uses information about sex or race to make decisions.

To address this, there has been a recent increase in literature focusing on the creation of techniques for making machine learning systems less unfair [4]. One example is FairPCA, which aims at transforming data into a low-rank representation where protected features are uncorrelated with the transformed data [5]. Currently, many different notions of fairness exist, and it has been shown that creating algorithms that uphold all notions of fairness simultaneously is impossible [6]. Furthermore, it is often the case that introducing fairness as a constraint to a machine learning system will lead to lower performance [7].

Another challenge in creating machine learning decision-making systems in these settings are the need for explainable models. The European General Data Protection Regulation grants citizens the right to contest decisions made by fully-automated systems, as well as the right to transparency regarding the logic involved in the system [8]. However, this becomes challenging when data transformation techniques are used.

Using a dataset [9] consisting of mortgage loan application data, we build a logistic regression classifier to predict whether or not an applicant had their application approved or denied. Under the assumption that loans were granted to "good" applicants and not to "bad" ones, such an algorithm could be used to automatically make decisions about who should get a loan or not. This paper explores the idea of creating such a system that is both well-performing, fair, and explainable. In particular, we adapt the technique described by Kleindessner et al. [5, section 3.1] on the trade-off between predictive performance and fairness, varying λ in an attempt to find a model that achieves both good performance and fairness metrics. Furthermore, we examine the explainability of our model after applying FairPCA to get an understanding of how the technique affects model explainability.
Lastly, we discuss ethical implications of using machine learning for automated decision-making.

3 Data

3.1 Pre-processing

We work with data published under the Home Mortgage Disclosure Act (HMDA) [9]. The data consists of mortgage applications made to multiple financial institutions (modified to protect applicant and borrower privacy). We use the 2017 data from Tennessee.
Our data initially consists of 78 features, consisting of information about the financial institution in which the loan was applied for, information about the loan itself, as well as demographic data for the loan applicant and co-applicant. In order to simplify our approach, we keep only a subset of the original features. Appendix A shows the list of features we kept, as well as their explanations. With the exception of our target feature, we dummy-encode all categorical features, i.e., we one-hot encode the features such that a feature with K categories is encoded as K-1 one-hot vectors. Initially, the data also contained information regarding the sex and race of co-applicants. To simplify our approach, we chose to re-code this information into a categorical feature with the value 1 if someone applied with a co-applicant, 0 if they applied without, and "not applicable" for the rest. We then dummy-encoded this feature in the same way as the other features.
To simplify our prediction task, we re-code the `action_taken` column, such that "Loan Originated" and "Application approved but not accepted" are coded as 1, indicating that the applicant was granted the opportunity of accepting the loan, and likewise "Application denied by financial institution" is coded as 0. Rows without any of these are dropped from the dataset, as they have no clear indication of whether a loan was offered or denied. Our prediction task is thus whether an applicant had their application approved or denied.
We split our data into a random 75%-25% train-test split. In our training data, we observe that 79.8% of all applicants had their applications approved. To improve model performance, we downsample the applicants who had their loans approved, such that there are equally many approved and denied applications in our training data. In the end, our training data consists of 57,384 rows and 17 columns.

3.2 Potential biases

There have been multiple examples of automatic decision-making systems that discriminate between minority groups [2, 3]. A common concern of current machine learning techniques is the notion of "garbage in, garbage out"; that no machine learning model is going to be better than the data it is trained on. Therefore, it is important to look for potential biases in the data before training a model.
For our task, we suspect that our approach is susceptible to evaluation bias [10] because we only have information regarding whether an applicant had their loan approved, and not whether they were able to pay it back. For this reason, we cannot evaluate how good our model is at detecting "good" and "bad" applicants - only how much its decisions overlap with those of the financial institutions. Thus, the performance evaluation of our model may be inherently biased, since the labels are decisions

performed by the financial institutions. Additionally, our model may be subject to what Mehrabi et al. [10] refer to as *omitted variable bias*. Since our model does not have access to other relevant information such as employment status, credit score, or previous debt, it makes it more difficult for the model to correctly predict whether someone should be granted a mortgage loan or not.

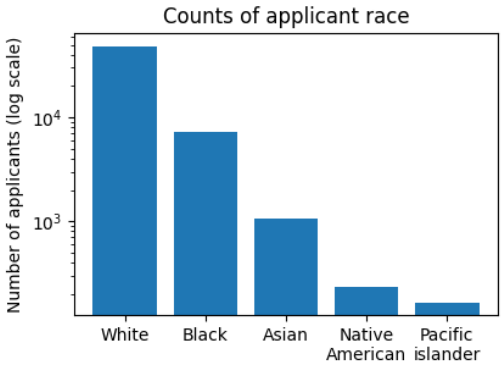


Figure 1: Distribution of Race for training data. Note that the y-axis is in log-scale.

Another bias that might be present in our data is *representation bias* [10], since we observe very large differences in the number of applicants from different demographic groups (see Figure 1). This representation bias means that our model may have a harder time generalizing well to some groups, since the model only observes a small number of examples for these. Additionally, histograms of minority population percentage by racial group (see Figure 2a) indicate differences between different racial groups. For example, we observe no white applicants living in areas where more than 60% of the population are minority groups, whereas this range is higher for Black/African American applicants, indicating the presence of a historical bias in our data (since historically, groups of the same race tend to live closer to one another).

Another concern regarding this historical bias is that places with a larger share of minorities tend to have lower incomes. Figure 2b shows income distributions for different racial groups. We generally observe that native Americans and Pacific islanders tend to have smaller incomes than the other groups (possibly due to the representational bias already discussed). Furthermore, Figure 3 shows the percentage of applicants within each racial group who had their mortgage loan application accepted. Here, we see a bias against certain minority groups, as we clearly notice that Black/African American applicants, as well as Native American and Pacific islanders, have their applicants accepted at a much lower rate compared to White or Asian applicants. This further supports our hypothesis that race should be treated as a protected group to reduce bias.

We conducted a similar analysis for male and female. In general, we found that distributions between sexes are quite similar, especially compared to the distribution differences between racial groups. However, we noticed that that a larger fraction of the females in the data were Black/African American compared to that of males, and that more applications from males were accepted (see appendix B). This indicates that there is a potential for our model to learn an undesirable gender bias. For this reason, we choose to also debias the data with respect to sex.

One of the simplest approaches to debiasing a machine learning system is to remove information about protected features, i.e., not training the model on sex or race. However, this is often not enough due to the presence of *proxy features*. This means that the model may implicitly train on protected attributes through its correlations with other features. Using `dython.nominal.associations`, we show these correlations between attributes in our training data (see Figure 4). We see that our data has correlations between protected and non-protected features. For example, we observe a Correlation Ratio of 0.53 between the applicant's race and the minority population percentage of the applicant's tract. This supports our hypothesis that a historical bias is present in our data. It also means that we do not expect to remove a potential racial bias in our model by simply removing race as a feature.

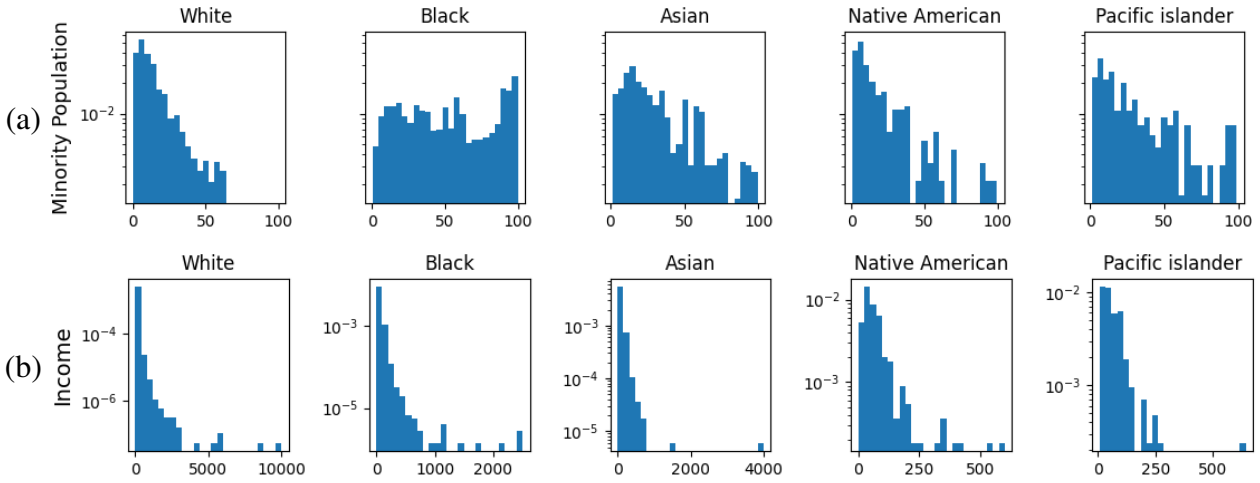


Figure 2: Histograms by racial group showing a) Minority population percentage distribution. b) Income (in thousands) distribution. Note that axes are not aligned and that y-axes are in log-scale.

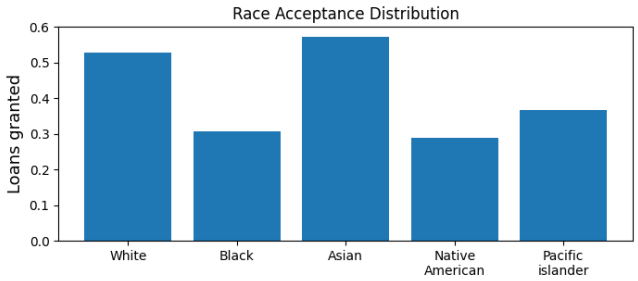


Figure 3: Percentage of applicants of a certain racial group who had their loan application accepted.

4 Methods

We train a simple logistic regression model as a baseline for predicting whether someone should be granted a loan or not. While better machine learning architectures exist for prediction tasks, we opt for logistic regression since it will be easier to explain compared to something like a deep neural network, due to each feature having only a single weight. We perform a grid-search using 5-fold-cross validation for tuning the regularization type (L1 or L2) and regularization strength. Due to our unbalanced data, we evaluate the performance of our models using Matthew's Correlation Coefficient (MCC) since this metric takes all aspects of the confusion matrix into account, as opposed to accuracy or F1-score. We find multiple models that achieve the same MCC (to four decimal places). Of these options, we choose the one using L1-regularization with the highest regularization strength (inverse regularization strength of 0.1), since this will be more likely to force weights to 0, making our model easier to explain.

The correlations discussed in section 3.2 motivate our use of FairPCA as a way of debiasing our data. FairPCA is similar to PCA in that it attempts to find a lower-dimensional, linear projection of the data, however, FairPCA differs from standard PCA in that it attempts to find the best projection such that the projected data is no longer correlated with protected demographic information [5]. We use sex and race as the information we wish to de-correlate against.

To evaluate the fairness of our models, we compare the False Negative Rates (FNRs) within different protected groups. FNR measures the fraction of people who should have gotten a loan, but were predicted as *denied*. Ideally, all groups should have the same FNR. We focus on minimizing the absolute difference in FNR between protected groups.

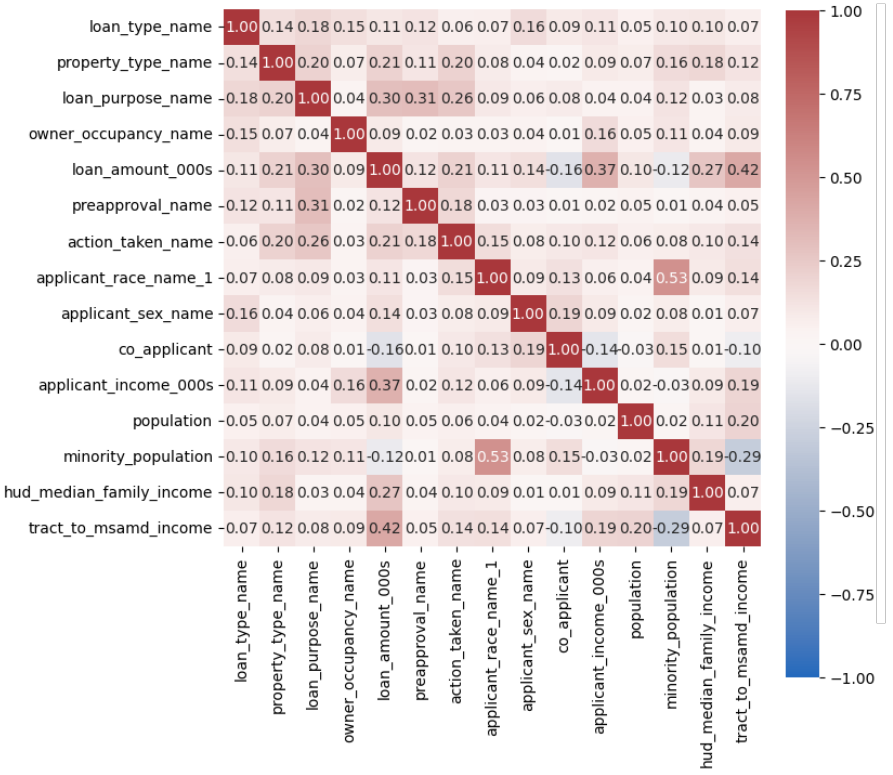


Figure 4: Correlations between features in the down-sampled training data, using the `associations` function from the `dython.nominal` module. The function uses the appropriate correlation metrics depending on the types of the compared features.

We choose to focus on FNR since being falsely denied a loan will have strong negative consequences for the applicant. If certain demographic groups are falsely denied a loan more frequently than others, then said group will unjustly be at a disadvantage in the real world as they will be unable to purchase a house. We focus on FNR over other fairness measures, e.g. equalized odds, since the consequence from the perspective of applicant is larger if they are unjustly denied a mortgage loan, compared to being given a mortgage loan they shouldn't have gotten (false positive rate). We focus on promoting the individuals' fairness by using FNR, whereas using FPR could be the focus for a bank, as false positives are what causes them to lose money.

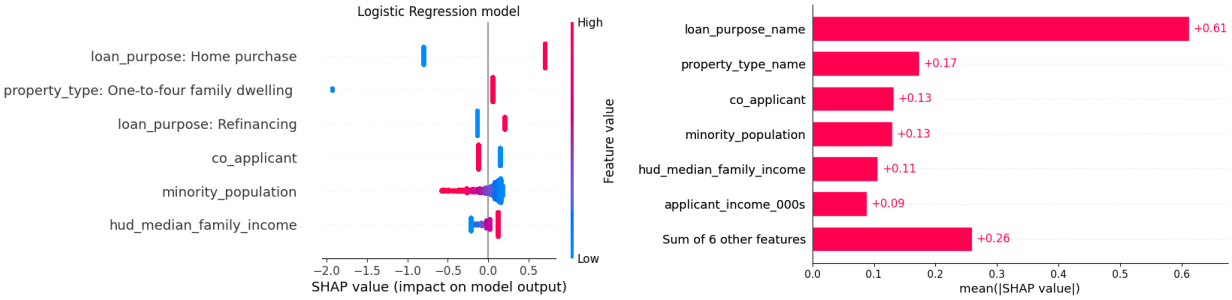
When increasing the fairness of a model, one often observes a decrease in performance [7]. Kleindessner et al. [5] suggest a modification of FairPCA that allows for a trade of lower fairness for higher performance by concatenating the FairPCA representation of a datapoint with a scaled version of the standard PCA representation of the datapoint, i.e., the new representation becomes $(U_{fair}^T x; \lambda \cdot U_{st}^T x)$, where $U_{fair}^T x$ is the fair representation of a datapoint, x, $U_{st}^T x$ is the standard PCA representation of x, and λ is the scaling parameter used to trade fairness for performance. In order for this method to work, λ should be small; $0 \leq \lambda \ll 1$. We try 33 logarithmically spaced lambda values in the range $[10^{-4}; 0.1]$ to examine if a good trade-off between performance and fairness is possible for our data.

Lastly, because the applicants should have the possibility of contesting the decision-making system, we need a way of examining our model logic to understand how it makes its predictions. For this purpose, we use SHAP-values [11] to compute feature importance. These can be used to explain to applicants why their application was rejected by the model and what features were the most influential. Additionally, it can also be used to examine how the previously mentioned biases impact our model.

5 Results

Our baseline logistic regression model achieves an MCC of 0.31 - much better than most-frequent or random guessing. However, looking at Figure 6a which shows the FNR across different racial groups for our baseline model, we clearly see that the model is unfair. The largest absolute difference in FNR between racial groups is 0.27. We observe a much lower FNR for Asians compared to all other racial groups - in fact, the FNR for Asians is less than half of the FNR for Black/African Americans,

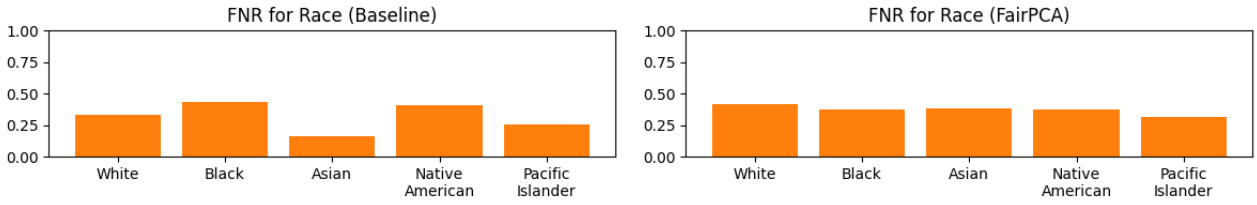
indicating that the historical bias has slipped into our model, despite it not being trained on race. Note that the FNR for Pacific islanders might be this low due to how little they are represented in our data. For sex, the difference in FNR is 0.05 (see Appendix C (a)). The model is thus more fair across sexes compared to racial groups, which makes sense given the generally similar feature distributions for male and female. However, we still observe a slightly higher FNR for females, indicating that there may be room for improvement in terms of model fairness with respect to sex.
Using the `shap` Python library, we compute SHAP-values for our model predictions and plot them (see Figure 5). We see that the minority_population feature is within the top-5 most important features used by the model. Additionally, if a higher fraction of the population of an applicant's tract are of a minority group, then this contributes negatively towards having their application accepted. This further supports our suggestion of using FairPCA to de-correlate our data with the protected features.



(a) (b)

Figure 5: a) SHAP summary plot. SHAP-values are computed for our baseline model's predictions on the test set. b) Feature importance estimated as mean absolute SHAP-value computed for each feature. SHAP-values for categorical variables were aggregated before feature importance was computed.

We compute the first $m - N_{protected\_groups} = 10$ fair PCs (principal components), where $m = 17$ is the number of non-protected features in our data and $N_{protected\_groups} = 7$ is the number of protected groups (two for sex, five for race). We then train a model on our data projected into FairPCA space. This yields a model with an MCC of 0.14 on the test set, which is much worse than our baseline performance. However, looking at Figure 6b, we see much smaller differences in FNRs across racial groups, indicating that our model has become more fair. The largest absolute difference in FNR has been lowered to 0.09, which is much lower compared to the baseline. For male and female, the absolute difference in FNR is 0.02 (see Appendix C (b)), so FairPCA was able to make the model more fair with respect to both protected attributes. Unfortunately, better fairness comes at the cost of higher FNRs for almost all racial groups, which contributes to the decreased performance.



(a) (b)

Figure 6: Test prediction FNRs for different racial groups for a) our baseline model and b) a model trained on data in FairPCA space.

While the predictions for this model are clearly more fair, performance decreases by quite a bit. For this reason, we implement the performance vs. fairness trade-off for FairPCA [5, section 3.1] to

examine the possibility of finding a model that simultaneously maintains decent fairness *and* performance. Figure 7 shows how our model performance and fairness change for different values of λ. Here, we measure fairness as the largest absolute difference in FNR across racial groups. We find that the method works as described in [5]; as we increase λ, our performance starts to increase, but at the cost of fairness. There does not seem to be a λ for which performance increases without lowering the fairness of the model. However, the method could be used to create a model that trades some performance for more fairness. This may however have consequences for model explainability (see section 6). Appendix D shows how fairness for sex changes for different values of λ. Here, we note that we are capable of slightly increasing fairness while also slightly increasing fairness with respect to sexes. However, this still leads to a decrease in fairness with respect to race (as shown in Figure 7).

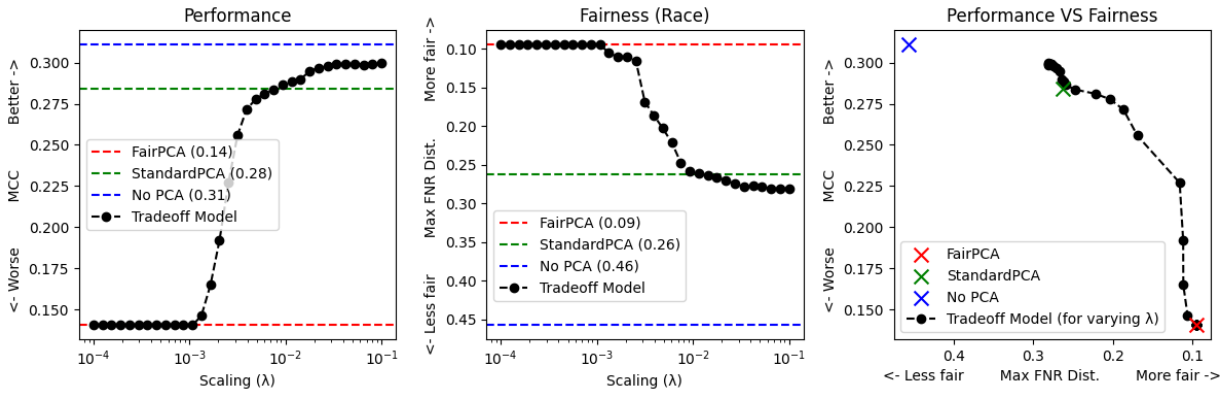


Figure 7: Trade-off of performance vs. fairness for including (a scaled version of) standard PCA as training data. Fairness is measured as the largest absolute difference in False Negative Rate between different racial groups.

While FairPCA clearly improves the fairness of our predictions, Kleindessner et al. [5] do not discuss the implications that the method may have on explainability. One of the drawbacks of using data projection to de-bias a model is that the model becomes much harder to interpret. To estimate feature importance, we first use SHAP-values to estimate the importance of each FairPC as their mean absolute SHAP-value. We let s be the vector containing the importance score of each FairPC, then scale the projection matrix, U_{fair}, column-wise based on s. Since the importance scores are strictly positive, we multiply s with $sign(w)$, where w is the coefficient weights of our logistic regression model trained on the data projected into FairPCA space, in order to account for whether the component contributes positively or negatively towards a positive label. In short, we compute $U_{fair} * (s * sign(w))$, which is an $n\_components \times m$ matrix. We then sum along the rows of this matrix to get an estimate of how each of our original features impact the prediction of our model (see Figure 8). Notice that minority_population now has much less importance than before (see figure 5), suggesting that FairPCA was indeed capable of de-biasing our data. We also observe that VA and FSA/RHS loans are more likely to be accepted, which makes sense given that applicants applying for these types of loan have an organization who helps by providing guarantees for loans.

To verify that our approach can be used to explain our model, we used the estimated feature importance to make counterfactual examples, i.e., changing the applicant's features to alter the outcome of the model prediction. We select two applicants from our data; one who was accepted, and one who was not (see applicant features in appendix E). For each person, we focus only on attributes they can influence themselves. For the person who was originally denied a loan, the most important feature they can alter is the co_applicant feature. We find that if they were to apply with a co-applicant, then they would be granted the loan. It makes sense that having two people who commit to paying back a loan should increase the chances of the loan being successfully paid back, and thus increase the chance of getting accepted. Thus, this seems like a reasonable explanation that could be provided from the bank.

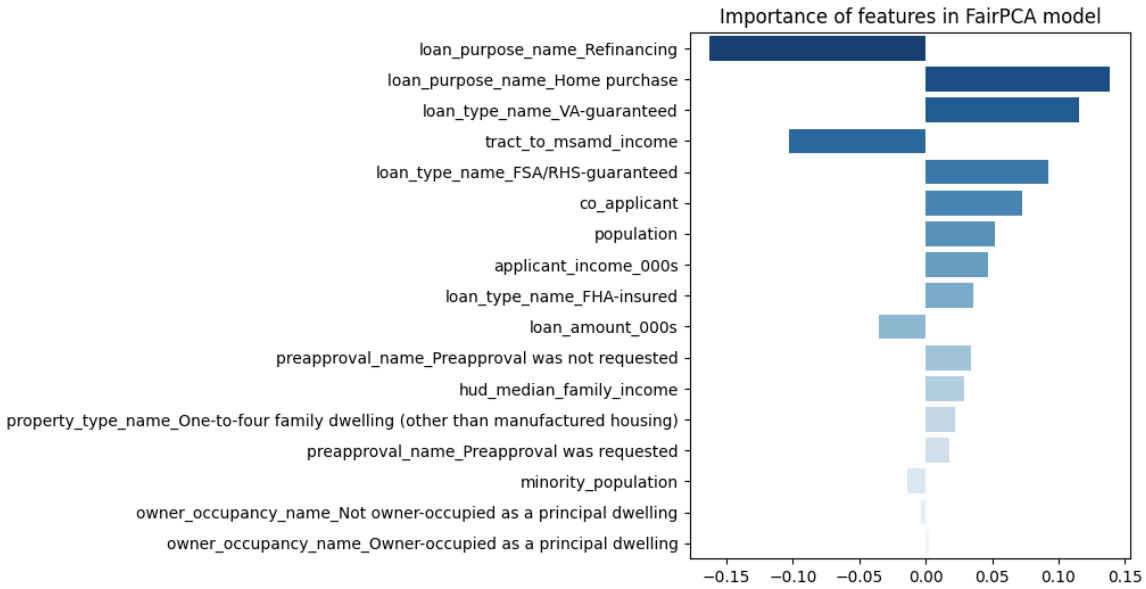


Figure 8: Estimated feature importance of original features for model trained on data projected into FairPCA space.

For the person whose application was accepted, we change their loan type from a VA-guaranteed loan to a conventional loan. This was enough for the person to no longer have their loan accepted. This gives an explanation as to how someone whose loan was guaranteed by the Veteran's Association may have been denied under other circumstances. Thus, it seems that our estimated feature importance can be used to explain the overall decision-making of our model, although the specific process of estimating the feature importance is more tricky to understand.

6 Discussion

6.1 FairPCA

Our results show that the methodology of [5] does seem to reduce biases in the data, leading to a more fair decision-making system. However, we also find that this comes at the cost of performance, which may be undesirable to the institution deploying a decision-making system similar to this one.
We reiterate that since our target labels are based on decisions taken by the financial institutions, rather than if the applicant was able to pay back their loan or not, our way of measuring model performance may be biased due to the bias against certain racial groups observed in the data. If our training labels themselves are biased, then it will clearly be difficult to create a model that is fair and well-performing in the way we measure performance. A better way to measure performance may be to predict how likely it is that a person will default on their mortgage loan, but since we do not have such data this is impossible for us to do. It should also be mentioned that it remains unclear how robust our model is with respect to how it generalizes to applicants from other states or countries. Since we only train the model on data from Tennessee, we should not expect the model to be transferable to other states or countries, since applicants outside Tennessee may have widely different feature distributions than what the model is trained on. This would require further analysis which we leave to future works.

To balance performance and fairness, we implemented the *performance vs. fairness trade-off* method described in [5, section 3.1], and saw that it could indeed be used to trade some amount of fairness for higher performance. However, this comes with an additional cost in terms of reduced model explainability. Examining feature importance in a projected space (such as the one created by FairPCA) is challenging in itself, but using the trade-off method, each datapoint is represented as a combination of

two different projections (the FairPCA and standard PCA projection), which makes it very difficult to reason about what information the model uses, and how. Although it may be possible to create such explanations, it is beyond the scope of this project, and we simply note that while data augmentation techniques can be used to de-bias a dataset, it may conflict with other desirable properties such as performance and explainability.

6.2 The Intricacies of Explainability in Modern Machine Learning

While data scientists and researchers may have an understanding of how decision-making algorithms work and what they base their decisions on, model explanations such as the ones provided in this paper may not be enough for someone to fully grasp how the system works if they are not familiar with machine learning (ML) and artificial intelligence (AI). In particular, there may be a discrepancy in how a layperson understands the inner workings of ML compared to how current ML techniques actually work. For example, when computers were first introduced to science, they were used to enhance the mechanistic approaches of science, i.e., to find explanations around cause and effect. This was mostly done using algorithms and simulations, where each step of a process could easily be explained, and we imagined a future where these computers would eventually grow to become more intelligent than humans. This "Good Old Fashioned AI" (GOFAI) may be how ordinary people understand AI - as programs with super-human intelligence whose decisions should not be questioned. In reality, modern AI relies mostly on the discovery of correlations in large amounts of data. This presents a new paradigm of science which no longer relies on theory, but solely on data [12]. This makes it dangerous to view modern AI as GOFAI, since correlations between variables in a dataset may be random or have little causal effect. This essentially means that our ML algorithms will not be better than the data on which they are trained - something researchers colloquially refer to as "garbage in, garbage out". This also means that biases present in our training data will be learned and potentially amplified by our machine learning model, as we saw when we trained our baseline model.

This highlights some of the dangers of blindly accepting automated decision systems. Viewing the outputs of these systems as indisputably true answers does not make sense, since the systems only learn whatever correlations may be present in your data. They have no understanding for whether or not it makes sense to base their decisions on these correlations. There is no mechanistic explanation to be found within them, as was the case for the GOFAI algorithms.
This raises the question if we should even be using automated decision-making systems for decisions that have such large impacts on the lives of individuals. Clearly, modern AI provides benefits for governments and large companies, since they can relatively easily build such automated systems due to the abundance of data they can collect. These systems are also easier to build than GOFAI systems, since one no longer has to come up with 'rules' for rule-based decision-making - this is simply learned by the model based on the data. Once an automated system has been built, the efficiency with which decisions can be made increases, which is the main benefit of deploying these systems [13]. But do these benefits outweigh the potential harm these systems may bring onto others?
A potential solution to this concern is to have a 'human in the loop' and use the decision-making systems as an advisor, rather than as the sole decision maker. This would use the capabilities of the decision-making system, but let a human be the one with the 'final say' regarding the decision. Unfortunately, there have been cases where such a setup was employed, but where its decisions were rarely challenged [13], which further highlights how laypeople may put too much trust towards the predictive performance of the decision-making system.
Thus, while we can clearly improve the fairness of the decision-making systems using methods such as the ones deployed in this paper, this does not address the concern that laypeople do not truly understand modern AI systems. Additionally, the decrease in model performance may not be acceptable to the companies deploying these systems as they have to stay competitive in their respective market.

Whether or not automated decision-systems should be used is thus an ethical question that from the utilitarian perspective translates to "does the benefits for the company deploying the decision-making system outweigh the potential harm it may cause loan applicants?" The answer to this question is probably no, given how many people would be affected by these automated systems, but we could also imagine a world in which human decision-making is inherently biased, and a decision-making system that is equally biased would not make the situation any worse - ultimately, it may not be easy to dispute a decision taken by a human either. In such a scenario, utilitarianism would not necessarily view automated systems as unethical, since they would not reduce the overall happiness or increase suffering. However, one may argue that viewing the problem from a utilitarian point of view does not truly assess the justness of deploying these systems [14], and it is often the case that modern ML approaches amplify the biases found in their training data. For these reasons, one needs to consider carefully when and where to deploy automated decision-making systems, and particularly how it should be done. In particular, it is important that decisions can truly be contested, regardless of whether they were done by a human, an AI, or a combination of both.

7 Conclusion

In this paper, we used 2017 HMDA data from Tennessee and built a logistic regression model to predict whether a mortgage loan applicants should be granted a loan.
Early data exploration revealed several biases in the training data; certain minority groups were very underrepresented, a historical bias of Black/African American applicants living in neighborhoods with a larger minority population, and a differing acceptance rate across different racial groups. These biases led to the hypothesis that a classifier trained to predict approval of loan applications would be biased against certain minority groups.
Our baseline logistic regression model achieved an MCC of 0.31. When evaluating the fairness of the model, we observed large differences in False Negative Rate (FNR) across different racial groups. For example, the FNR was much lower for Asian and much higher for Black/African American applicants. We used SHAP-values to compute feature importance for our baseline model, which showed that `minority_population` was one of the most important features used by the model.
Since we observed a large correlation between applicant race and `minority_population`, we implemented FairPCA [5] to de-bias our data. Training the model on the de-biased data achieved an MCC of 0.14, which was much worse than the baseline. However, the difference in FNRs across races was lower, meaning that the model is more fair according to our notion of fairness. Unfortunately, this came at the cost of increasing FNR for most of the racial groups, and also made model explanation much more complex - although we did outline a method for estimating feature importance in the original feature space, which we successfully used to make counterfactual examples.
In an attempt to create a model that achieved a better trade-off between performance and fairness, we implemented the performance vs. fairness trade-off for FairPCA [5, section 3.1]. Using the method, we were able to trade some fairness for better performance, but we did not observe a point where we could increase performance without hurting fairness. Furthermore, we discussed concerns regarding the explainability of the models trained using this method.
Finally, we highlighted broader concerns regarding automated decision-making systems. In particular, whether people truly understand how the automated decisions are made, and how this may complicate the process of contesting their decisions - a right given to citizens of the European Union via the European General Data Protection Regulation (GDPR) [8]. Furthermore, we raise ethical questions about whether automated systems should even be used for such important decisions, in particular, from the utilitarian point of view.

References

[1] Bhekisipho Twala. Multiple classifier application to credit risk assessment. *Expert systems with applications*, 37(4):3326–3336, 2010.

[2] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the COMPAS recidivism algorithm. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[3] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. 2015. URL https://arxiv.org/pdf/1408.6491.pdf.

[4] Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1), November 2022. doi: 10.1007/s44176-022-00006-z. URL https://doi.org/10.1007/s44176-022-00006-z.

[5] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair PCA for fair representation learning, 2023. URL https://arxiv.org/abs/2302.13319.

[6] Suvodeep Majumder, Joymallya Chakraborty, Gina R Bai, Kathryn T Stolee, and Tim Menzies. Fair enough: Searching for sufficient measures of fairness. *arXiv preprint arXiv:2110.13029*, 2021.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[8] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL https://data.europa.eu/eli/reg/2016/679/oj.

[9] Consumer Financial Protection Bureau. Download HMDA data. 2017. URL https://www.consumerfinance.gov/data-research/hmda/historic-data/?geo=tn&records=all-records&field_descriptions=labels.

[10] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

[11] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

[12] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 2014. doi: 10.1177/2053951714528481. URL https://doi.org/10.1177/2053951714528481.

[13] Maciej Kuziemski and Gianluca Misuraca. Ai governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44(6), 2020. ISSN 0308-5961. doi: https://doi.org/10.1016/j.telpol.2020.101976. URL https://www.sciencedirect.com/science/article/pii/S0308596120300689.

[14] John Rawls. Justice as fairness. *Philosophical Review*, 67:164–194, 1958. URL http://fs2.american.edu/dfagel/www/Philosophers/Rawls/JusticeAsFairness.pdf.

Appendix

A Features after pre-processing

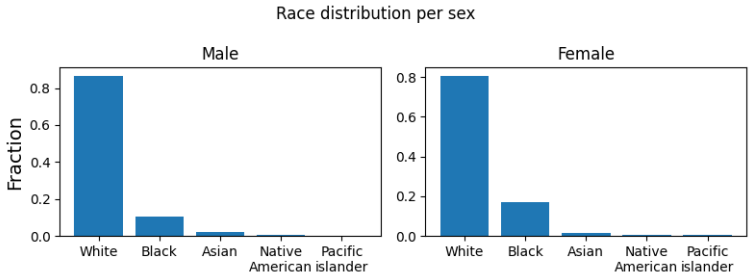
One-hot-encoded features are described together. Note that some feature names in the notebook are suffixed with '`_name`'. This is because we have used the string columns, rather than the numerical columns, which has no effect on the outcome of the analysis.

Feature Name	**Description**	**Values**
loan_amount_000s	The requested loan in thousands	[1; 3500]
action_taken	Whether or not the applicant's loan was accepted. This is our target variable	0, 1
co_applicant	Whether the main applicant has a co-applicant	0, 1
applicant_income_ 000s	The applicant's gross annual income in thousands	[1; 9999]
population	The total population in the tract	[0; 24746]
minority_population	Percentage of the population in the tract that consists of minorities	[0; 100]
hud_median_ family_income	The median family income in the MSA/MD	[47900; 67500]
tract_to_ msamd_income	The percentage difference in income from the applicant's tract to their MSA/MD	[0; 325.58]
loan_type	Describes the loan type	Conventional *, FHA-insured, VA-guaranteed, FSA/RHS-guaranteed
property_type	The housing type of the property	Not manufactured housing Manufactured housing *
loan_purpose	What the loan is used for	Home purchase, Refinancing, Home improvement *
owner_occupancy	Describes if the owner lives in the property	Owner-occupied as a principal dwelling, Not owner-occupied as a principal dwelling, Not applicable *
preapproval	Describes if a preapproval loan was requested	Not applicable *, Preapproval was not requested Preapproval was requested
applicant_race_1	The race of the applicant. Mixed races have been dropped.	White, Black or African American, Asian, Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native *
applicant_sex	Sex of applicant	Male, Female *

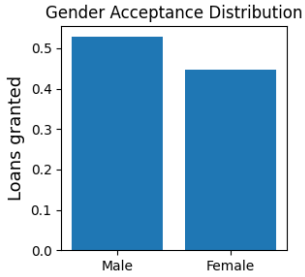
* Reference/baseline value for dummy encoded features, i.e., the value whose column we drop.

B Sex Distributions

a) Distribution of racial groups across sexes.

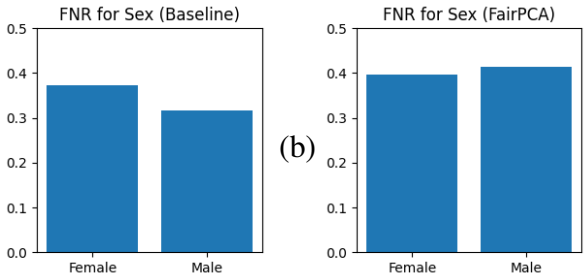
(a) 

b) The fraction applicants within each sex who had their application accepted.

(b) 

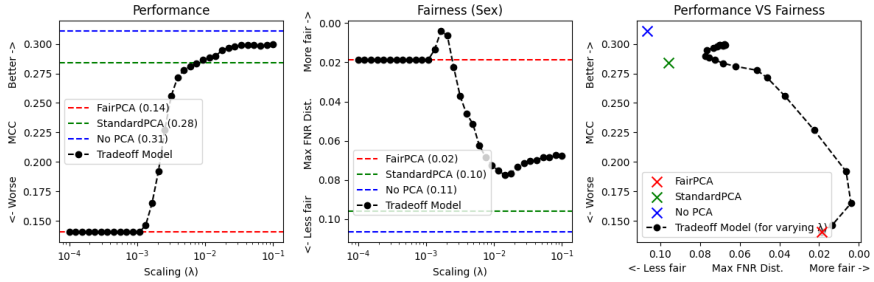
C Sex FNR

Test prediction FNRs for sexes for a) our baseline model and b) a model trained on data in FairPCA space.

(a) (b) 

D Sex Tradeoff

Trade-off of performance vs. fairness for including (a scaled version of) standard PCA as training data. Fairness is measured as the absolute difference in FNR between sexes.



E Counterfactuals

Counterfactuals for two samples in the data, before and after changing features. Using the information from Figure 8 it is simple to determine and show the changes needed to change an approved application to be denied, or vice versa.

Counterfactual	**1**	**2**
ID	259345	149356
loan_amount_000s	43	163
co_applicant	$0 \rightarrow 1$	0
applicant_income_000s	64	71
population	2662	5977
minority_population	64.46	51.03
hud_median_family_income	59500	55800
tract_to_msamd_income	46.93	111.72
loan_type_FHA-insured	0	0
loan_type_FSA/RHS-guaranteed	0	0
loan_type_VA-guaranteed	0	$1 \rightarrow 0$
property_type_name_One-to-four family dwelling	1	1
loan_purpose_Home purchase	0	1
loan_purpose_Refinancing	0	0
owner_occupancy_Not owner-occupied as a principal dwelling	0	0
owner_occupancy_Owner-occupied as a principal dwelling	1	1
preapproval_Preapproval was not requested	0	1
preapproval_Preapproval was requested	0	0
applicant_race_1_Asian	0	0
applicant_race_1_Black or African American	1	0
applicant_race_1_Native Hawaiian or Other ...	0	0
applicant_race_1_White	0	1
applicant_sex_Male	1	1
% probability of being accepted (needs $\geq$50%)	$46.96 \rightarrow 52.30$	$60.95 \rightarrow 49.58$

An arrow ($\rightarrow$) indicates the change for the counterfactual.