Fair and Transparent Machine Learning Methods mini-project: Adversarial manipulation of local vs global explanation frameworks

Tim Mondorf CJK681

Abstract

The LIME detection framework can identify discrimination, but Slack demonstrated that it could be manipulated. I demonstrate how this is related to the fact that LIME is a local explanation framework. Global explanation frameworks, identifying statistical parity and performance parity, will not be susceptible to the same dark strategies but may instead be open to fairness gerrymandering. The outcome of an extended game between a an agent with the intent to operate an unfair model and a regulator concerned with detecting this practice can be analyzed in a generative adversarial network.

1 Using LIME to detect unfairness

In my bachelor project I discussed the relationship between Explainable Machine Learning (Explainable ML) and the general explanation frameworks known from philosophy of science (Mondorf, 2023, p. 17, [7]). One such general framework, the deductive-nomological model (Woodward, 2018, sect. 2, [3]) relies on general statements, whereas ML often expresses probabilistic and heuristic statements. Other general explanation models, such as the statistical relevance model, rely on conditional probabilities and therefore are akin to an ML model rather than a separate framework that could explain the model. Instead of relying on general models of scientific explanation, literature on Explainable ML has therefore put forward a number of explanation frameworks that are specific to ML.

One of these is LIME (Local Interpretable Model-agnostic Explanations) presented by Ribeiro (Ribeiro, 2016, p. 1, [4]). LIME analyses an ML model and a set of input features and output labels and is able to generate local explanations that quantify the relative importance of the individual features. In the oft-cited example of a loan approval model, LIME can provide an individual applicant with an explanation in the form of a ranking of the features according to their importance and sign. For example, the main reason that a loan was rejected was the level of existing debt, whereas the level of income in itself actually was a positive factor. LIME performs this local, linear quantification by generating a number of synthetic observations around each real observation, inputing them to the model and performing a linear approximating in the local area around the observation in question. Although LIME is a local model, the original article did generalize from local explanations to global model properties. In a famous example, a number of classifications of dogs versus huskies were performed, and in each example, the local explanation was the same, i.e., huskies had snow in the background and dogs did not (Ribeiro, 2016, p. 7, [4]). LIME was thus able to detect global model failures through generalization from several local explanations.

In the dog-husky example, LIME detected model failure by revealing that irrelevant variables played a role in model decisions. This suggested that LIME could also be applied to the detection of model unfairness. LIME could generate a multitude of local explanations. If these local explanations all indicated that the same protected attribute such as race was significant, it would indicate that the model was unfair.

Slack demonstrated that a smart, unfair algorithm could actually avoid detection by LIME (Slack, 2020, p. 3, [2]). The following is my own simple example that aims to demonstrate the basic traits of Ribeiro's LIME-model and of Slack's manipulation example through simpler examples without replicating their work directly.

1.1 Terminology

Instead of the term discriminatory model, used by Ribeiro and Slack, the term 'unfair model' will be used in this paper. That is because Section (5) will review literature on adversarial models, such as Goodfellow (Goodfellow, 2016, section 20.10.4, p. 691). Goodfellow uses the term ’discriminator’ to refer to the role played by the party trying to learn the underlying nature of a model (that I call GOV).

I will analyze first individual fairness and then group fairness. Both are defined by Elliott (Elliott, 2023, p. 4, [10]). Elliott defines individual fairness as the condition that similar individuals should be treated similarly. He defines group fairness as fairness relying on statistical parity conditions

The term model in this paper refers exclusively to the loan decision model operated by the actor that I call BANK. While literature also refers to LIME as a model, to avoid confusion I will refer to LIME, LIME’ and similar models as ’frameworks’. I will also refer to scientific explanation models as frameworks.

1.2 BANK, GOV and LIME’

A bank, called BANK, wants to base its loan decisions on an ML model that involves two variables x_1 and x_2 that are both attributes of the individual applicants. Assume that the general population is normally distributed in the (x_1, x_2)-space with mean $(\frac{5}{2}, 2)$ and variance 2 but with the restriction that (x_1, x_2) only take on positive values. BANK considers both variables relevant. The model used by BANK calculates credit scores according to $f(x_1, x_2) = -3x_1^3 + 5x_2^2$. It is assumed that a positive credit score is positive for the applicant and vice versa, so that the data can meaningfully be categorized according to whether credit scores are negative or positive.

A regulator, here called GOV, considers x_1 protected, i.e., it should not factor in loan decisions, whereas GOV considers x_2 legitimate. GOV will try to determine whether the model used by BANK does in fact incorporate x_1.

BANK is obligated to submit decisions and model to GOV. BANK submits two real decisions $(x_1, x_2, f(x_1, x_2))$ that are $(1, 2, 17)$ and $(6, 2, -628)$. GOV applies the LIME detection framework to identify unfairness. A simpler framework called LIME’ defined by myself, rather than the actual LIME, is used. All calculations are shown in the jupyter notebook that is attached as a zip-file.

Since the two real observations have the same x_2-values, the large difference in the credit score would already indicate unfairness based on the protected attribute x_1. This is confirmed when GOV applies LIME’ to generate two sets of synthetic observations in the vicinity of the two real observations and calculate the ensuing credit score f. Figure (1) shows the two real observations marked by stars, the synthetic observations marked by dots, the credit scores shown along the z-axis with positive scores shown in green and negative scores shown in red.

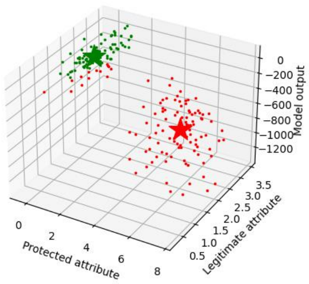
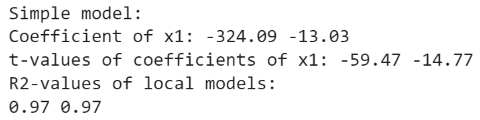


Figure 1: Credit scores simple unfair model



```
Simple model:
Coefficient of x1: -324.09 -13.03
t-values of coefficients of x1: -59.47 -14.77
R2-values of local models:
0.97 0.97
```

Figure 2: Results of local linear approximations simple unfair model

GOV then generates local explanations for the two loan decisions by for each of the two clusters performing linear approximations of the importance of x_1 and x_2. The results are shown in Figure (2). GOV focusses on the significance of the protected attribute x_1 in each of the two local decisions.

A full output of the linear regression is again shown in the notebook. It is evident from Figure (2) that x_1 plays a highly significant role in both local decisions. Numerical t-values are much higher than two and R2 of both local models is very high. GOV therefore concludes that the model is unfair, awards a fine to BANK along with a ban on basing loan decisions on the model. LIME’ has been effective in preventing unfairness.

2 Manipulating LIME’

As suggested by Slack (ibid), BANK then invents a dark strategy. BANK either has insight into LIME’ or performs reinforcement learning by submitting

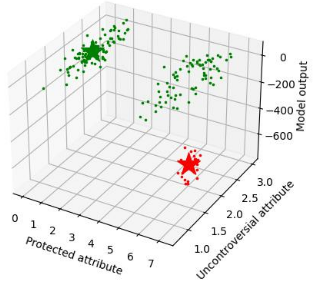
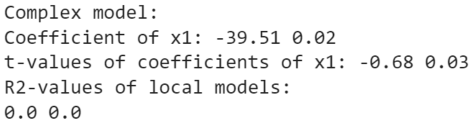


Figure 3: Credit scores complex unfair model

several models for approval. BANK is aware that LIME' generates a cluster of synthetic feature combinations around the real observations. This means that the model input can be clustered into one cluster for each real observation. Knowing that the cluster is generated by adding noise with mean zero to each real observation, BANK also knows that observations that lie close to the centroid of their cluster have a high probability of being real observations, and vice versa. BANK therefore updates the model to first learn the cluster structure of the input via a clustering algorithm such as k-means. BANK then calculates the distance to the cluster center for each observation. If the observation is among the 25 percent of observations that are nearest to their respective centroid, there is a good chance that this is a real observation, and BANK can then assign the credit score based on the underlying, unfair algorithm. All other observations are processed according to an ostentatious algorithm that only takes into account the legitimate attribute x_2: $f_{ostentatious}(x_1, x_2) = 5x_2^2$. The result is shown in Figure (3). The dark strategy preserves the unfair property of the model for the observations that are estimated to represent real applicants but now assigns positive credit scores (green dots) to several synthetic applicants with high values of x_1. The red cloud in Figure (1) is now replaced by a small red cloud close to the real observation represented by a star (real applicants who did not get a loan) and a green cloud of observations that are further from the star (synthetic applicants that did get a loan).

LIME' again calculates the linear approximations for each local decision and since each cloud now contains 75 percent of observations for which the label is generated according to the ostentatious algorithm, the coefficient of x_1 no longer becomes significant as can be seen in Figure (4). Here, the numerical t-values for the coefficient of x_1 are below two for each of the two local decisions.



```
Complex model:
Coefficient of x1: -39.51 0.02
t-values of coefficients of x1: -0.68 0.03
R2-values of local models:
0.0 0.0
```

Figure 4: Results of local linear approximations complex unfair model

BANK has succesfully manipulated LIME' into concluding that no unfairness based on x_1 occurs while being able to maintain unfair practices for real applicants. R2, i.e., overall explanatory power, in both local frameworks, is now very low which might arouse GOV's suspicion. I have thus replicated Slack's experiment demonstrating that in this simple example, using LIME' instead of LIME, the local explanation framework is susceptible to manipulation.

3 Using global explanation frameworks to detect unfairness

The research on Explainable ML evolved rapidly since the release of LIME in 2016 and even since Slack's critique in 2020. While the L in LIME stands for local, Søgaard (Søgaard, 2022, p. 3, [11]) emphasizes the difference between local and global explanation frameworks (Søgaard uses the term 'models') and posits that the distribution of model predictions is a global model property that must be analyzed based on global explanation frameworks. Bias and unfairness are related to this global property.

Slack, in his discussion of LIME, refers to the term of 'post hoc'-explanation technique. A post-hoc explanation framework considers the model a black box and, as described above, analyzes the relationship between model input and model output. The alternative is an intrinsic explanation framework that generates explanations as part of model output. Søgaard dismisses this distinction claiming that intrinsic explanation of model training is different from intrinsic explanation of model prediction. This means that the statement 'our model emphasizes existing debt in loan approvals because our training data held many defaults caused by high existing debt' is different from the statement 'your loan was rejected because of your high level of individual debt'. Søgaard instead introduces a distinction between explanations based on forward versus backward passes in neural networks.

This critique of Slack would indicate that his experiment should be reiterated, this time applying a global explanation framework.

3.1 Fairness criteria in local and global frameworks

Since the purpose of the analysis is to detect unfairness, it must first be analyzed which fairness criteria are the most appropriate to apply in such a global framework. The two real observations in the above example had the exact same value of the legitimate attribute but ended up with vastly different credit scores. This is an example of individual (un)fairness such as defined by Elliott above (Elliott, 2023, p. 4, [10]). Applying a global model requires us to rely on group fairness and statistical parity conditions.

Such statistical parity conditions are defined, e.g., by Søgaard (Søgaard, 2023, p. 17, [12]). He defines statistical parity as the requirement that $P(f > 0|x_1 < 3.5) = P(f > 0|x_1 \geq 3.5)$ where I have updated the notation to match this example. This condition translates to the requirement that the probability of getting a positive credit rating, given that the individual is below the population mean with regard to protected attribute x_1, must be equal to the probability of a positive credit score for an individual who is above the mean of x_1.

Søgaard also introduces an alternative fairness criterion called performance parity, defined as the requirement that $P(f > 0|x_1 < 3.5, x_2 > 2) = P(f > 0|x_1 \geq 3.5, x_2 > 2)$ where I have again updated the notation to match the example. Performance parity means that the probability of receiving a positive credit score, given that the applicant is below average on the protected attribute and above average on the legitimate attribute, must be equal to the probability of a positive credit score given the applicant is above average on the protected attribute as well as on the legitimate attribute. Performance parity therefore would compensate for correlation between protected and legitimate attributes and allow bias with regard to the protected attribute, as long as this bias can be explained by differences in other, legitimate attributes. For example, a protected attribute such as criminal record and a legitimate attribute such as education level may be proxies of one another. In that case, the model may seem unfair under statistical parity, because individuals with criminal records have significantly lower credit scores, whereas the model will in fact exhibit performance parity because individuals with high education levels have the same credit scores irrespective of their criminal records.

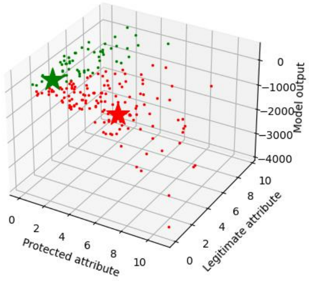


Figure 5: Loan values sampled from a global distribution

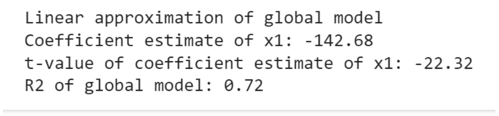
3.2 Analyzing BANK and GOV in a global framework

Having defined statistical parity and performance parity (neither of which was discussed by Slack), we can now apply a global framework to evaluate fairness and study robustness to manipulation.

As suggested by Søgaard above, an appropriate global explanation framework would be to analyze the distribution of model predictions over the entire domain of input feature variables. Instead of performing local perturbations, GOV can sample from a global input distribution and generate model output. BANK can no longer manipulate the detection framework by distinguishing between real and synthetic feature combinations, because each feature combination entered is generated from the same distribution and no longer clustered around the two real observations. The bias, in form of the strong relationship between the green dots and their position along the x_1-axis is now visually evident from Figure (5) where 200 observations are sampled and input to the model along with the original two loan applicants again shown by stars.

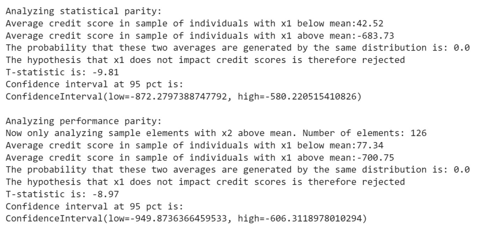
GOV may be able to detect the underlying, nonlinear function, but even a global linear approximation will detect unfairness with high explanatory power (R2 is above 70 percent) and a highly significant x_1 (t-value well above 2), as shown in Figure (6).

The global explanation framework has therefore clearly detected unfairness by concluding that the protected attribute plays a significant role in model



```
Linear approximation of global model
Coefficient estimate of x1: -142.68
t-value of coefficient estimate of x1: -22.32
R2 of global model: 0.72
```

Figure 6: Results of linear approximations of loan values sampled from a global distribution



```
Analyzing statistical parity:
Average credit score in sample of individuals with x1 below mean:42.52
Average credit score in sample of individuals with x1 above mean:-683.73
The probability that these two averages are generated by the same distribution is: 0.0
The hypothesis that x1 does not impact credit scores is therefore rejected
T-statistic is: -9.81
Confidence interval at 95 pct is:
ConfidenceInterval(low=-872.2797388747792, high=-580.220515410826)

Analyzing performance parity:
Now only analyzing sample elements with x2 above mean. Number of elements: 126
Average credit score in sample of individuals with x1 below mean:77.34
Average credit score in sample of individuals with x1 above mean:-700.75
The probability that these two averages are generated by the same distribution is: 0.0
The hypothesis that x1 does not impact credit scores is therefore rejected
T-statistic is: -8.97
Confidence interval at 95 pct is:
ConfidenceInterval(low=-949.8736366459533, high=-606.3118978010294)
```

Figure 7: Analyzing statistical parity and performance parity of credit scores sampled from a global distribution

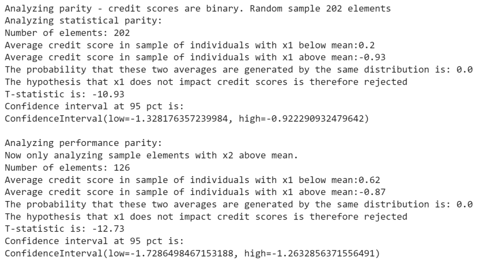
predictions.

In addition, statistical parity and performance parity are now calculated as described above and shown in Figure (7).

The full calculations are shown in the notebook. To analyze statistical parity, a t-test is performed between the credit scores of sample individuals with x_1 values below mean and credit scores of sample individuals with x_1 values above mean. It is clear from the very low p-value that it is improbable that the credit scores of the two groups are sampled from the same distribution. Therefore, the model is not fair for individuals with high x_1-values.

Figure (7) also demonstrates that the model does not exhibit performance parity. A t-test is performed of the credit scores of individuals with x_1-values below mean and x_2-values above mean against the credit scores of individuals with x_1-values above mean and x_2-values above mean. As can be seen from the very low p-value, it is improbable that the two are generated by the same distribution. x_1 and x_2 are not proxies of one another. In the population of individuals who all things being equal would have high credit scores due to their value of the legitimate attribute x_2, unfair treatment according to the protected attribute x_1 persists.

The global analysis applied here is less well suited for local explanations. To explain why the individual at (6,2) had his loan rejected, it can be noted that he has an x_1-value higher than the population average and that the model is highly biased against x_1, but the linear approximation of a non-linear model cannot quantify the exact contribution of (x_1, x_2) around (6,2). For instance, if the person asked 'which x_2-value should I try to attain to compensate for my high x_1-value?', the global framework would not be able to help, but would refer to a local approximation such as LIME. Global and local explanation techniques serve different purposes.



```
Analyzing parity - credit scores are binary. Random sample 202 elements
Analyzing statistical parity:
Number of elements: 202
Average credit score in sample of individuals with x1 below mean:0.2
Average credit score in sample of individuals with x1 above mean:-0.93
The probability that these two averages are generated by the same distribution is: 0.0
The hypothesis that x1 does not impact credit scores is therefore rejected
T-statistic is: -10.93
Confidence interval at 95 pct is:
ConfidenceInterval(low=-1.328176357239984, high=-0.922290932479642)

Analyzing performance parity:
Now only analyzing sample elements with x2 above mean.
Number of elements: 126
Average credit score in sample of individuals with x1 below mean:0.62
Average credit score in sample of individuals with x1 above mean:-0.87
The probability that these two averages are generated by the same distribution is: 0.0
The hypothesis that x1 does not impact credit scores is therefore rejected
T-statistic is: -12.73
Confidence interval at 95 pct is:
ConfidenceInterval(low=-1.7286498467153188, high=-1.2632856371556491)
```

Figure 8: Statistical parity and performance parity binary labels random samples

4 Manipulating global explanation frameworks

The global approach has therefore outsmarted the unfair algorithm that was used to manipulate the LIME'-framework. BANK must therefore attempt another tactic if he wants to continue to base model decisions on x_1 without regulatory sanction. 'Fairness gerrymandering', as described by Elliott (Elliott, 2023, p. 12, [10]) and Kearns (Kearns, 2018, p. 1, [5]), exploits structure in data in order to hide unfairness in models. The original example in Kearns (example 1.1., ibid) involves binary labels and two protected attributes. I will perform a version of fairness gerrymandering that builds on the above example, mantaining one protected attribute and replacing real-valued credit scores by their sign to have binary labels.

The above credit scores are transformed to the sign of the original scores and can thus be interpreted as whether a loan was awarded or not. Again, a sample is taken and the two fairness metrics clearly demonstrate that the protected attribute is a significant factor in awarding a loan, as shown in Figure (8).

As before, the sample consists of the two original observations as well as 200 randomly selected observations from a population of 10.000. Assume that BANK is aware that fairness is always measured based on his actual client base plus 200 randomly selected observations. Assume that BANK is able to calculate the credit scores of the entire population. BANK can then target a demographic that has above-mean values of the protected at-

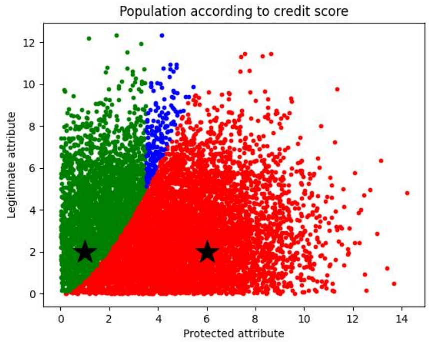
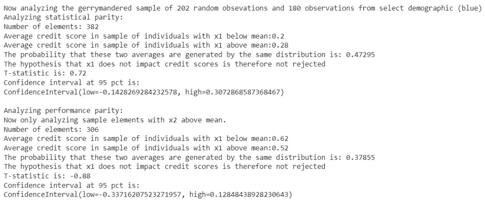


Figure 9: Positive credit scores (green), minority overachievers with positive credit scores (blue), negative credit scores (red)



```
Now analyzing the gerrymandered sample of 202 random obsevations and 180 observations from select demographic (blue)
Analyzing statistical parity:
Number of elements: 382
Average credit score in sample of individuals with x1 below mean:0.2
Average credit score in sample of individuals with x1 above mean:0.28
The probability that these two averages are generated by the same distribution is: 0.47295
The hypothesis that x1 does not impact credit scores is therefore not rejected
T-statistic is: 0.72
Confidence interval at 95 pct is:
ConfidenceInterval(low=-0.1428269284232578, high=0.3072868587368467)

Analyzing performance parity:
Now only analyzing sample elements with x2 above mean.
Number of elements: 306
Average credit score in sample of individuals with x1 below mean:0.62
Average credit score in sample of individuals with x1 above mean:0.52
The probability that these two averages are generated by the same distribution is: 0.37855
The hypothesis that x1 does not impact credit scores is therefore not rejected
T-statistic is: -0.88
Confidence interval at 95 pct is:
ConfidenceInterval(low=-0.33716207523271957, high=0.12848438928230643)
```

Figure 10: Statistical parity and performance parity binary labels gerrymandered example

tribute but values of the legitimate attribute that are so high that the credit score will still be positive according to the unfair model. Such a demographic can be interpreted as 'minority overachievers'. This demographic is shown in Figure (9) as the blue slice, between the green labels of other individuals with positive credit scores and the red labels of individuals with negative credit scores. The two original observations are shown as black stars.

The targeting of minority overachievers can be done via select marketing or corporate communication praising the importance of diversity. BANK now submits the model and existing client base to GOV. GOV then evaluates a dataset that consists of the two original observations, the 106 minority overachievers and 200 randomly selected individuals. It can be seen from Figure (10) that this sample now exhibits both statistical parity and performance parity.

The inclusion of a select demographic with high values of the protected attribute as well as high values of the legitimate attribute has obscured the unfair nature of the model.

If the experiment is replicated (not shown) with actual rather than binary credit scores, the unfair nature of the model is in fact detected. The same may happen if GOV adjusts the size of the random sample as a function of the size of the dataset that is provided by BANK. An alternative form of gerrymandering (not shown) would be if BANK would be willing to change the unfair model in a way that disregards the value of the protected attribute, but only for individuals with high values of the legitimate attribute and again pad the client base with a large number of minority overachievers.

5 An adversarial learning framework

Rather than pursue the above specific examples, a more general approach is taken. ML literature provides such a set-up in the form of adversarial frameworks (often called adversarial models or Generative Adversarial Networks (GAN)). Goodfellow (Goodfellow, 2016, chap. 20.10.4., p. 690, [1]) and Rocca (Rocca, 2019, p. 1, [9]) describe GAN. A simple example of GAN is where a model A generates synthetic pictures of dogs that cannot be distinguished from pictures of real dogs. A generates a synthetic picture, then randomly decides whether to present this synthetic picture or a real picture. An adversarial model B then tries to determine whether the picture is synthetic or real. A learns from the process of being rewarded (B mistook a synthetic picture for a real picture). B learns from the detection process (B detected a synthetic picture as such).

5.1 Analyzing BANK and GOV in an adversarial framework

Apply this approach to BANK and GOV. As outlined by Slack (ibid), GOV can demand all real model decisions as well as model decisions for any synthetic feature combinations. GOV will try to learn the exact model in use. GOV's loss function is calculated as the error of the learned model on a separate test data set. The game is zero-sum, meaning that BANK's loss function is the negative of the loss function of GOV, so the better BANK can dissimulate the true model the higher the reward. Both models will then be trained simultaneously, the output of one being the input of the other. The example could be brought even closer to Goodfellow's example by allowing BANK to generate several manipulated models, potentially expressing the same underlying unfairness, combined with a dataset of fair models, sampling equally from both and asking GOV to distinguish unfair from fair models. If both BANK and GOV apply neural networks, there is according to Goodfellow (ibid) no mathematical guarantee of convergence. However, most actual experiments, such as the above mentioned model to generate artificial pictures of dogs, do converge at an equilibrium where the artificial images cannot be distinguished from real images. It would depend on an actual experiment whether in an adversarial framework, BANK would be able to learn to dissimulate unfair practices to perfection. An example of distinguishing credit scores in the form of one real value would be simpler than distinguishing images.

Goodfellow makes one very relevant observation, i.e., that GAN's can 'fit probability distributions that assign zero probability to the training points' (Goodfellow, 2016, section 20.10.4., p. 691, [1]). This means that an artificially generated output will never include actual training data. Therefore, potentially, a GAN can be trained on highly unfair input that is never in itself reproduced. If BANK can train its side of the adversarial model on labelled training data generated by the actual, unfair model, the question would be whether unfairness would be maintained in the output or dissimulated enough to avoid detection by GOV.

5.2 Comparison to Heo's framework

Heo (Heo, 2017, p. 1, [8]) first establishes a set of models that for each prediction identify the role played by each individual input feature in arriving at that prediction. For image classification models, this is done using saliency maps, and for other quantitative data sets, this is done by returning a relevance score for each feature. Where this has previously been called 'intrinsic explanations', Søgaard (Søgaard, 2022, p. 3, [11]) would call it a forward-pass explanation because it relates the input to the neural network with the output.

I will now try to apply Heo's thinking to the examples in this paper. Again, this is done in my own, simplified way that is inspired by the sources without attempting a full replication.

Where BANK previously had the loan approval model $(x_1, x_2, f(x_1, x_2))$ it would now have to operate $(x_1, x_2, f(x_1, x_2), f.x_1 - score, f.x_2 - score)$. This can be seen from the 'Adult income example', Figure 1 (left panel) in Heo (Heo, 2017, Figure 1, p.2, [8]). In other words, where the model previously returned a decision in the form of a credit score, it will now return a decision with three attributes. $f(x_1, x_2)$ is the credit score, $f.x_1 - score$ is the relevance of x_1 in the decision and $f.x_2 - score$ is similar. The two relevance scores can be thought of as summing to unity.

BANK knows that GOV will bark at a high $f.x_1 - score$. BANK therefore creates his own two-model setup. First, the original, unfair model produces a set of training, validation and testing data $x_1, x_2, f(x_1, x_2)$. Using this training data, BANK trains a neural network that will serve as the new loan approval model, since it will take x_1, x_2 as input and return an estimated credit score $\hat{f}(x_1, x_2)$. It will also return the two relevance scores. BANK now trains its neural network with a loss function that, as any ML model, penalizes errors relative to the labelled training data, but which also penalizes high $f.x_1 - score$-values. Where the above examples relied on linear regression, neural networks are better at emulating non-linear and multicriteria structure in data according to Igel (Igel, 2023, p. 11, [6]). Assuming BANK can train the neural network well, it can serve as a model that both performs the unfair decisions that BANK ultimately wants while presenting inconspicuous relevance scores to GOV.

Heo succesfully demonstrates that it is possible to fit a model that both has the desired outcome (i.e., unfair loan decisions) while dissimulating the importance of protected attributes. His work is remarkable because it demonstrates that several fooling techniques work in practice across very diverse datasets. Race and sex have high relevance scores in Heo's 'real' model but low relevance scores in the 'fool'-model. As for the other model attributes, some relevance scores are unchanged in Heo's example while others ('marital status', 'relationship status' and marginally 'WorkClass') increase from the real model to the fooling model. Heo does not explicitly discuss whether this can also be considered fairness gerrymandering, i.e., BANK exploits correlation between protected and legitimate attributes. If 'WorkClass' was completely uncorrelated with 'race', it is not clear whether the fooling-strategy of Heo would work.

Heo does not discuss local versus global explanation frameworks but it is clear that his is a local approach, attributing specific decisions to specific feature values. A global explanation framework such as applied above would input a large and even more diverse dataset into the model and analyze all decisions and their relevance scores. This might shed another light on the success of the fooling-strategies.

It is, however, clear that the potential for manipulating fairness characteristics of ML models persist, not least in neural network-based models since these can exploit more complex structure in data than the simpler models.

References

[1] Ian Goodfellow Yoshua Bengio Aaron Courville. 2016. Deep learning. MIT Press.

[2] David Slack et al. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. AIES 20.

[3] James Woordward et al. 2018. Scientific explanation. Stanford Encyclopedia of Philosophy.

[4] Marco Ribeiro et al. 2016. Why should i trust you? explaining the predictions of any classifier. arXiv.

[5] Michael Kearns et al. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. arXiv.

[6] Christian Igel. 2023. Deep learning. Lecture slides, Advanced Deep Learning, DIKU.

[7] Tim Mondorf. 2023. Does machine learning need a theory of everything? Bachelor project, BSc in Machine Learning and Data Science, DIKU.

[8] Juyeon Heo Sunghwan Joo Taesup Moon. 2019. Fooling neural network interpretations via adversarial model manipulation. Neurips 2019, Vancouver.

[9] Joseph Rocca. 2019. Understanding generative adversarial networks (gans). Towards Data Science.

[10] Desmond Elliott Karolina Stanczak. 2023. Statistical notions of fairness. Lecture slides in Fair and Transparent Machine Learning Methods.

[11] Anders Søgaard. 2022. Explainable natural language processing. Springer.

[12] Anders Søgaard. 2023. Adl-2023-fairness. Lecture slides, Advanced Deep Learning, DIKU 2023.

A Notebook

The notebook with all calculations is submitted as a zip-file.