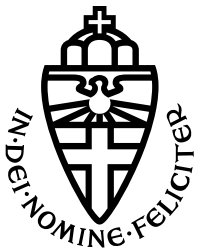
Radboud University Nijmegen



Faculty of Social Sciences

Exploring the usefulness of explainable machine learning in assessing fairness

Thesis BSc Artificial Intelligence

Author:
Daphne Smits
(s1005509)

Supervisor:
prof. Tom Heskes
Institute for Computing and
Information Sciences (iCIS)

Second reader:
Max Hinne
Donders Centre for Cognition

July 2020

Abstract

Investigating the fairness of an algorithm has become more important since such algorithms have been employed in more sensitive areas, such as credit risk assessment and criminal justice. There exists no firm consensus regarding the various existing fairness measures, which can lead to an uninformed use of any of these measures. This research aims to find a relation between the field of explainable artificial intelligence and the field of fair artificial intelligence. If such relation exists, this could evoke a more transparent and informed fairness assessment. This research focuses on the state-of-the-art explainability method SHAP and investigates the usefulness of this method in assessing fairness. This is done in three ways: (1) the relationship between SHAP and existing fairness measures is studied; (2) a possible improvement of one fairness measure using SHAP is examined; (3) a usability study is conducted to explain existing measures with SHAP. The results of this study show a promising relationship between SHAP and the field of fair artificial intelligence.

Contents

Chapter 1

Introduction

As Artificial Intelligence is rising, algorithms are used more often in a wide variety of domains. Classification algorithms, that can classify samples into different groups based on the input features, are rising as well. Applications of such algorithms are seen in a wide variety of domains, such as speech analysis, image recognition and classification of biological data. This growing use enables us to make educated decisions based on those outcomes. Because these algorithms are used in fields of great importance, concerns about the fairness of these models are emerging. It is crucial to ensure that an algorithm is just and unbiased before using it. There are several obstacles in ensuring fairness. One of them is that there are many different fairness formalizations, with no firm consensus [1]. This makes claiming fairness ambiguous, as an algorithm may be fair according to one formalization and unfair according to another. The correct formalization to use is based on the context [2], which makes a general fairness assessment, fit for all scenarios, impossible. A second problem is the fact that most algorithms are working as a 'black box'. It is extremely difficult to understand how the classification decisions are reached. If one cannot understand how a decision is made, it is also difficult to make claims about the fairness of the algorithm.

The aim of this thesis is to propose a novel method of ensuring fairness of a classification algorithm based on opening this black box. The field of explainable artificial intelligence focuses on visualizing and explaining the internal behaviour of such black box models [3]. One of the most prominent techniques is SHAP (SHapley Additive exPlanations), which assigns each input feature an importance value for a given prediction [4].

1.1 Theoretical Background

1.1.1 Fairness

The use of decision making algorithms is growing, also in sensitive areas such as crime prediction [5] and job hiring [6]. Because of the legal or ethical implications of such decisions, it is of great importance to ensure that the decision making algorithm is fair. This has led to an increase in interest in designing fair algorithms and in numerous definitions to measure fairness [7][8][2][1][9]. Despite the growing advances in this field, there is not yet a firm consensus about the proper way to measure fairness. This makes it difficult to choose the suitable definition of fairness for

practical use.

Although the topic of algorithmic fairness is rather novel, the literature on fairness in other fields such as economics and game theory is extensive [10][11]. The definitions used to define algorithmic fairness are often derived from the methods in those fields. Several of such fairness metrics will be discussed in this paper. For these definitions, the papers of Verma and Rubin [2] and Kusner et al. [12] are used as a reference. Throughout the paper, the following notations will be used.

- A: The set of protected attributes, the attributes that should not be discriminated against.
- y: The actual, to be predicted outcome.
- $\hat{y}$: The outcome as predicted by the model.
- $P(x)$: The probability of x, which can be read as the number of times x occurred, divided by the times x occurred plus the times x did not occur.

The fairness metrics that will be discussed here are statistical parity, conditional statistical parity, equal opportunity, overall accuracy equality and individual fairness.

Definition 1. Statistical Parity: *A predictor satisfies statistical parity if subjects in both the protected and unprotected group have equal probability to be assigned the favourable predicted class:* $P(\hat{y} = 1|A = 0) = P(\hat{y} = 1|A = 1)$

Definition 2. Conditional Statistical Parity: *A predictor satisfies conditional statistical parity if subjects in both the protected and unprotected group have equal probability to be assigned the favourable predicted class, when controlling for some legitimate factors L:* $P(\hat{y} = 1|A = 0, L = l) = P(\hat{y} = 1|A = 1, L = l)$

Definition 3. Equal Opportunity: *A predictor satisfies equal opportunity if subjects in both the protected and unprotected group have equal probability to be wrongly assigned the unfavourable predicted class:* $P(\hat{y} = 0|A = 0, y = 1) = P(\hat{y} = 0|A = 1, y = 1)$

Definition 4. Overall Accuracy Equality: *A predictor satisfies overall accuracy equality if subjects in both the protected and unprotected group have equal accuracy:* $P(\hat{y} = y|A = 0) = P(\hat{y} = y|A = 1)$

Definition 5. Individual Fairness: *A predictor satisfies individual fairness if similar subjects receive similar predictions. The similarity of subjects is measured by some distance metric* $d(i, j)$*: If* $d(i, j)$ *is small then* $\hat{y}_i \approx \hat{y}_j$

To measure the extent to which an algorithm satisfies the notion of individual fairness, the consistency is computed as proposed by Zemel et al. [13].

Definition 6. Consistency: *Comparison of the prediction of a sample to its k-nearest neighbors: Consistency* $= 1 - \frac{1}{Nk} \sum_{n} |\hat{y}_n - \sum_{j \in KNN(n)} \hat{y}_j|$

1.1.2 Explainability

The growing use of decision making algorithms also led to an increasing wish to explain what the algorithm is doing and why it came to a certain decision. Due to the development of deep learning, and the availability of immense datasets, decision making algorithms have a better performance than ever before. This performance comes at a cost however, as these complex models often lack transparency. Where simple models were often easy to understand and explain, these complex models are employed in a black box manner, with no information available about the decision making process. To be able to explain the decision is often as crucial as the outcome, for instance in the medical field [14]. This led to the rise of Explainable Artificial Intelligence, which focuses on visualizing and explaining these complex models [3], [4], [15], [16].

Different techniques to open the black box were developed. As stated by Amina and Barrada [15], these different techniques can differ on three levels. First, there are differences in the complexity of interpretability. In general, the more complex models are harder to interpret and explain than simpler models. This has led to two approaches. The first, and most straightforward approach, is to design intrinsic interpretable algorithms. One example of such method is the Bayesian Rule List model by Letham et al. [17]. The second approach aims to offer a post-hoc explanation of a complex, black box model. This approach enables the use of more complex methods, which are often more accurate than the simpler ones. One example is the work of Mahendran and Vedaldi [18], where such a post-hoc method for image representation is introduced. Secondly, there are differences in the scope of the explanation. Some of the methods focus on explaining the general internal workings of the model [19], these are called global explanations. Others explain each individual decision made [20], which are called local explanations. Recent work is also done on combining both local and global explanations, e.g. by Linsley et al. [21]. Lastly, there are differences regarding the models they are appropriate for. Methods can be model-specific or model-agnostic. While model-specific methods are limited to only one type of model, model-agnostic methods can be applied to any type of model.

SHAP

This thesis will focus on the explainability method SHAP (SHapley Additive exPlanations) [4]. Lundberg and Lee defined the class of additive feature attribution methods, which unifies six existing methods such as LIME [20] and deepLIFT [22]:

Definition 7. ADDITIVE FEATURE ATTRIBUTION METHODS: *Methods that belong to this class have an explanation model that is a linear function of binary variables:*

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

where g is the explanation model, $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in R$.

In this definition, z' is the simplified input, where $z'_i = 1$ means that feature i is observed and $z'_i = 0$ means feature i is missing. ϕ_i is the attribution value of feature i. Lundberg and Lee [4] state that there is only one unique solution in this

class that satisfies the following three properties: (1) local accuracy, the explanation model must match the outcome of the model we want to explain; (2) missingness, missing features must have no impact; (3) consistency, if the model is changed to depend more on a certain feature, the attribution of that feature must never decrease. The only values that satisfy all of the properties are the Shapley values from cooperative game theory. SHAP uses these Shapley values to explain specific decisions, by assigning an importance value, the SHAP value, to each feature.

SHAP explains models with the use of an explanation model, which makes it a post-hoc explanation. The SHAP values explain how to get from the base value, what would be predicted if we did not know any features, to the current output. This is done per sample, which makes it a local explanation method. However, by aggregating these local SHAP values, global SHAP values can be obtained. It can therefore also be used as a global explanation method. As computing exact SHAP values is very challenging, approximation methods are proposed. Both model-agnostic and model-specific approximation methods are presented within the SHAP framework.

SHAP thus unifies many existing methods and is the sole method that satisfies the desired properties. These are strong motivations to use SHAP as the explainability method for this thesis.

1.2 Research Questions

The aim of this project is to explore the usefulness of SHAP in assessing fairness of decision making algorithms. Therefore, the main research question addressed in this paper is:

> To what extent is the SHAP framework useful in assessing fairness of an algorithm?

To answer this question, the following three sub research questions are investigated:

1. To what extent are SHAP values related to the existing fairness measures?
2. Can SHAP values be used to enhance or improve existing fairness measures?
3. Can SHAP values be used to gain more insight in the results of existing fairness measures?

This investigation is a useful addition to the present literature. Concluding fairness of an algorithm is not yet easily done, because no firm consensus has been met about the existing fairness measures. Furthermore, the measures are mathematically incompatible [23], [24]. Where an algorithm may be fair according to one measure, it can be unfair according to another. Therefore it is of importance to make fairness measures more intuitive. Within the field of explainable artificial intelligence, the goal is to make complex algorithms transparent and understandable. Using these techniques to make the current fairness measures transparent and understandable as well, is therefore a promising idea.

Chapter 2

Methods

This chapter gives an overview of the methods used to explore the usefulness of SHAP in the field of fair artificial intelligence[1]. All of the research is done on two models. One model is based on the COMPAS Dataset [5] and the second one is based on the German Credit Dataset [25]. These models will be discussed in section 2.1. The relation will be investigated between SHAP values and the following fairness measures: (1) statistical parity; (2) conditional statistical parity; (3) equal opportunity and (4) overall accuracy equality. This will be discussed in section 2.2. The exploration regarding the enhancement of existing fairness measures is done on the measure Individual Fairness and will be discussed in section 2.3. The usefulness of explaining fairness measures with SHAP will be discussed in section 2.4.

2.1 Models

2.1.1 Recidivism model

The first model predicts the risk of recidivism and is based on the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset. This dataset contains data concerning users of the COMPAS model. The COMPAS model predicts recidivism risk scores based on 41 scales, concerning 137 variables [26]. It is thus a very complicated model, and moreover, it is a proprietary model. The model itself can thus not be accessed. The dataset that was used therefore does not contain the exact input features of the model, but several general features concerning the background of each participant, the COMPAS score and whether the participant indeed relapsed into criminal behaviour after two years of the risk assessment.

The fairness of this dataset has been disputed. The investigation done by ProPublica has concluded that it is biased against African American convicts [5]. This makes it an interesting dataset to use when investigating fairness measures, such as has been done by Corbett-Davies et al. [27] and Chouldechova [24].

For the same reason, this dataset was used to create one of the models for this thesis. First, the data needed to be preprocessed. This was already done for the research of Rudin et al. [28]. They created one dataset combining all seven tables from the original database. Furthermore, the data was made numerical, which is the appropriate format for model creation. The preprocessed data consists of 8952

[1] https://github.com/DaphneSmits/Bachelor-Thesis

samples and 33 input features, such as age at the first offense, number of felonies and number of violent charges. This data was randomly split in a train and test set, with a test size of 0.2. The training data was used to create a random forest classification model, which was then tested with the test set. The attribute describing whether or nor the participant relapsed was used as the target value, the 33 personal information attributes were used as input values. The accuracy was calculated for 100 different models, based on 100 different train/test splits. This gave an average accuracy of 0.691 with a standard deviation of 0.033. This is similar to the result of Wadsworth et al., whose recidivism model had an accuracy of 0.70 [29].

In this model, the protected attributes *gender* and *race* are taken into account. The legitimate factor to control for when using conditional statistical parity of this model is *p_arrest*, which is the number of earlier arrests. This is considered a fair and logical attribute to use for classification.

2.1.2 Credit risk model

The second model classifies people as high or low credit risk and is based on the German Credit Risk dataset [25]. This dataset contains data about 1000 loan applicants, described by 20 attributes such as credit account status, duration, credit history, credit amount and saving account. One additional attribute states the risk score that was the classification outcome. This dataset is commonly used in fairness literature as well, as it exhibits a gender-related bias [2].

Before being able to create a model using this dataset, preprocessing needs to be done. This consisted of making the categorical data numerical using a one-hot-encoding. This resulted in a dataset of 1000 samples and 38 input features. This data was randomly split in a train and test set, with a test size of 0.2. The training data was used to create a random forest classification model, which was then tested with the test set. The accuracy was calculated for 100 different models, based on 100 different train/test splits. This gave an average accuracy of 0.808 with a standard deviation of 0.030. This accuracy is slightly higher than that of Van Sang et al., which had an average accuracy of 0.734 [30].

In this model the protected attribute *gender* is taken into account. The legitimate factor to control for when using conditional statistical parity of this model is *credit_account_status*, which is the amount of money in your credit account. This is considered a fair and logical attribute to use for classification.

2.2 Relationship between SHAP and existing fairness measures

One way the SHAP values could prove to be useful, is if they could be used to replace existing measures. The intuitive nature of SHAP values could make them a better alternative than the existing measures. To find out whether the SHAP values are in some way related to the measures, the correlation between them is calculated. This is done for the fairness measures statistical parity, conditional statistical parity, equal opportunity and overall accuracy equality.

The SHAP values give a local explanation of the model, explaining the decision for each individual sample. To compare SHAP values with existing fairness mea-

sures, these local explanations need to be aggregated into global explanations. This was done according to the procedure described in appendix A of the work of Tan et al. [19]. The global SHAP values are the result of averaging the local attributions at each unique feature value. The result is a global SHAP value that states the attribution of each unique feature value. For example, for the feature value $age = 21$ in the credit risk model, the global SHAP value is 0.048. This means that an age of 21 attributes on average 0.048 to the difference between the predicted outcome and the expected output without the attribute *age*.

To be able to correlate these global SHAP values, the fairness measures need to be computed for each unique feature value as well. All four fairness measures compute the difference between two groups. These groups are made based on each unique feature value. For example, for the fairness measure related to $age = 21$, the samples are split in two groups, where one group contains all samples with $age = 21$ and one group contains all samples with $age \neq 21$. For all fairness measures, the difference is calculated between the groups. This results in a fairness measure for each unique feature value.

For each unique feature value, a global SHAP value and a fairness value for each of the four fairness measures is obtained. For all the fairness measures, the correlation between these fairness values and global SHAP values can then be computed. To get an average correlation value, and a standard deviation, this computation is repeated 100 times according to the bootstrap method. This way the statistical significance can be investigated.

2.3 Improvement of existing fairness measures using SHAP

Another way SHAP can prove to be useful, is to improve an existing fairness measure. A fairness measure that could be improved using SHAP is the notion of Individual Fairness. An algorithm satisfies this notion of fairness if similar samples have similar output. Different metrics can be used to determine the similarity between samples. The classical approach to compute Individual Fairness computes the similarity as the Euclidean distance between the features. If two samples have similar features, the samples are considered similar. However, the features are often on a different scale, which makes it difficult to compute similarity of such different features. To improve this notion of Individual Fairness, the similarity of samples can be computed using SHAP values. If two samples have similar SHAP values, the samples are considered similar. Because SHAP values all have the same unit, determining similarity may become easier and more accurate. Clustering samples based on SHAP values was already done by Lundberg and Lee [31]. They conclude improved clustering results. This is a positive indicator that clustering based on SHAP values could lead to better results in computing individual fairness.

To investigate this, individual fairness was computed once in the classical manner and once using SHAP values. The neighbors package of sklearn was used to find the k closest neighbors for all samples [32]. For this classical approach, the standard NearestNeighbors function was used, which automatically uses the Euclidean distance to compute the distances between samples. For the novel approach the same neighbors package was used to find the nearest neighbors, but now implemented

with the different distance function. The parameter k, which represents the number of neighbours that the algorithm returns, was varied from 1 to 5. This was done to verify that the found result does not depend on this parameter. If the novel approach is indeed an improvement to the classical approach, the result should agree more with known biases than the old result. This was tested using a newly created method based on counting the occurrences of positive and negative biases for certain protected groups. To determine the bias a sample experiences, first the average outcome for all of the neighbor groups is computed. To minimize the noise, only the neighbours with opposite protected attribute will be considered. All samples that have an outcome that is lower than the mean of their group are considered to receive a positive bias and all samples with a higher outcome are considered to receive a negative bias. Remember that 0 is the preferred outcome for both models. In figure 1, the pseudo-code for these computations is given.

Binomial tests are then carried out to test whether the number of positive and negative biases are significantly different for the different attributes. The expectation would be that groups the algorithm is biased against receive significantly more negative bias and groups the algorithm favours receive significantly more positive bias. It is probable that the approach that shows this effect most clearly, is the most accurate approach.

```
   input  : Dataset, protected attribute, k
   output: The positive and negative bias occurrences for the different values
           of the protected attribute
 1 value0_neg_bias ← 0;
 2 value0_pos_bias ← 0;
 3 value1_neg_bias ← 0;
 4 value1_pos_bias ← 0;
 5 neighbors ← calculate k nearest neighbors for all samples in dataset;
 6 for sample ← Dataset do
 7 |   bias ← prediction of sample - average prediction for neighbors of sample;
 8 |   if bias < 0 then
 9 |   |   if sample[protected attribute] is 1 then
10 |   |   value1_neg_bias += 1;
11 |   |   else value0_neg_bias += 1;
12 |   end
13 |   else
14 |   |   if sample[protected attribute] is 1 then
15 |   |   value1_pos_bias += 1;
16 |   |   else value0_pos_bias += 1;
17 |   end
18 end
19 return value0_pos_bias, value0_neg_bias value1_pos_bias, value1_neg_bias;
```

Algorithm 1: Counting the positive and negative bias occurrences for both values of a protected variable.

2.4 Explaining fairness measures with SHAP

The third way that SHAP could prove to be useful in the field of fair machine learning, is by explaining the results of the existing fairness measures. According to Corbett-Davies and Goel [33] the existing mathematical definitions of fairness have shortcomings. Furthermore, it can be shown that several of the known measures are mutually incompatible [23], [24]. These complications make it difficult to choose the correct fairness measure for the context in which a machine learning model is being used. Using these measures in a black-box manner is thus not advisable.

An approach to use these measures in a more informed manner is therefore desired. One way this could be achieved is by using SHAP values as explained in a blog post by Lundberg [34]. SHAP values are most often used to decompose the model output into feature attribution values. However, SHAP can also be used to decompose fairness measures into feature attribution values. Concretely, these values are computed by first computing the global SHAP values and then computing the fairness measure of each feature using these SHAP values instead of the model outcomes. This is equivalent to decomposing the outcome of the fairness measure using SHAP, as computing SHAP values and computing fairness measures are both linear operations. The two operations can therefore be interchanged and still give the same results. The obtained value of an attribute then represents how much that attribute contributed to the fairness measure. Because the SHAP values add up to the model output, the obtained fairness attribution values also add up to the overall fairness measure of the model.

This was done for the fairness measures statistical parity and conditional statistical parity. The fairness measures equal opportunity and overall accuracy equality cannot be decomposed in the same way since they compare the true outcome with the predicted outcome. SHAP values are only concerned with the model output and not with the true outcome, which makes it impossible to apply this technique to methods that rely on the true outcome.

Chapter 3

Results

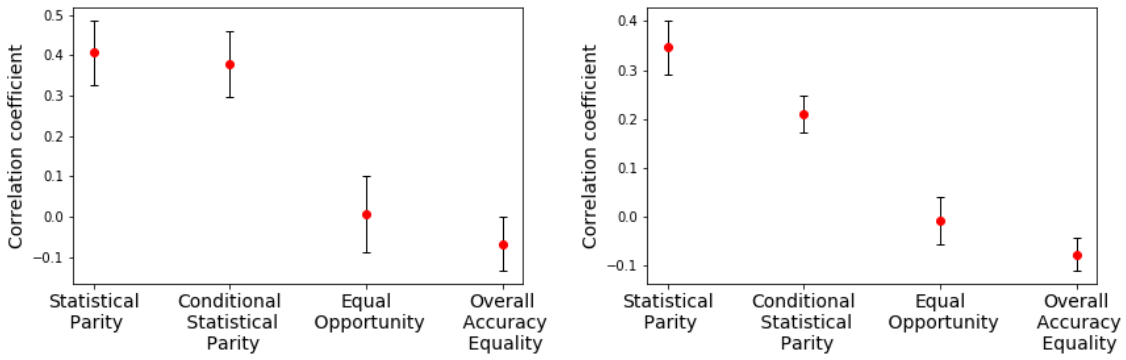
In this chapter the results of the research will be presented and discussed. In section 3.1 the correlation coefficients between the SHAP values and the various fairness measures will be discussed. The results of improving the notion of individual fairness with SHAP values will be discussed in section 3.2. Section 3.3 will show the result of the application study regarding the explanation of fairness measures using SHAP.

3.1 Relationship between SHAP and existing fairness measures

The first analysis was done to investigate the relation between SHAP values and the existing fairness measures. This was done by computing the correlation coefficients between them. In figures 3.1a and 3.1b the average correlation coefficients are plotted along with the corresponding standard deviation for each model. These values are obtained by computing the values for 100 bootstrap samples and averaging these values. From these figures, we can see that the correlation with SHAP values is relatively high for statistical parity difference and conditional statistical parity difference and relatively low for equal opportunity difference and overall accuracy equality difference.

To test if these means are significantly different from zero, the one sample t-test could be conducted. Because of the normality assumption of this statistical test, first the Shapiro-Wilk test was conducted. For all data, this resulted in a p-value higher than 0.05, which means that the null hypothesis stating that the data came from a normal distribution, could not be rejected given the chosen alpha value of 0.05. The one sample t-tests showed that the correlation between SHAP values and statistical parity difference and conditional statistical parity difference is significantly higher than 0 at the $p = 0.01$ level. For the chose alpha value of 0.05, no significance was found for the correlation between SHAP values and equal opportunity difference. The correlation between SHAP values and the overall accuracy equality difference showed to be significantly lower than zero, also at the $p=0.01$ level.

This result is partially as was expected. SHAP values do not entail information about the correct predictions, which are of importance to the equal opportunity difference and overall accuracy equality difference. This could explain the lower correlation value for those two measures. The found negative correlation for the overall accuracy equality difference is more counter intuitive. However, several studies show



(a) Credit model

(b) Recidivism model

Figure 3.1: Correlation between SHAP and various fairness measures. Figure (a) and (b) show the results of the credit model and the recidivism model respectively.

that various fairness measures are incompatible, which could explain this finding [23], [24]. The significantly high correlation coefficients for statistical parity difference and conditional statistical parity difference are an indicator of the similarity between the two concepts and SHAP values.

3.2 Improvement of existing fairness measures using SHAP

3.2.1 Credit risk model

The second analysis aimed to improve the fairness measure individual fairness using SHAP values. In order to do this, first the consistency of the individual fairness measure is computed for both approaches. Altering the parameter k did not show large differences, which is why here we will only consider the results for $k = 4$. This value was chosen as it gave the most representative result. According to the classical approach the consistency of the credit model was 0.824. When using the novel approach, using SHAP values, the outcome was 0.94. The novel approach clearly shows an improvement of consistency. This is due to the fact that the SHAP values correspond directly to the amount a feature contributes to the outcome of the model. Therefore, the features that contribute little to the outcome also contribute little to the appointment of nearest neighbours. This results in such a high consistency. But which method resembles the truth regarding this fairness measure? Table 3.1a shows the positive and negative bias occurrences for the attribute gender. According to other fairness measures and other research, the German credit dataset is biased against women [2]. The expectation would therefore also be that male participants more often receive a positive bias and female participants more often receive a negative bias.

As can be seen from table 3.1a, the classical approach, using a feature based distance function, shows this effect partially. The male participants receive significantly more often a positive bias than a negative bias at the p = 0.01 level, which corresponds with the literature. However, according to this method there exists no significant difference between the amount of positive and negative biases women

encounter. This does not correspond to the literature [2]. The novel approach shows exactly the opposite. It shows no significant differences for the male bias counts, which does not correspond to the literature. But it does show that women receive significantly more often a negative bias than a positive bias, which corresponds to literature. Despite the fact that two of the analyses show no significant differences, the observed differences are in the direction that is expected. Therefore both approaches resemble the literature reasonably well. Although the different approaches show different results for this model, no conclusions can be made regarding which approach is favoured.

3.2.2 Recidivism model

Again, the consistency of the individual fairness measure is computed for both approaches. According to the classical approach the consistency of the recidivism model was 0.816. When using the novel approach, using SHAP values, the outcome was 0.960. Again, the consistency of the model is higher according to the novel method, which is as expected.

To test which of these outcomes is more sensible, the positive and negative bias occurrences are counted. These can be seen in tables 3.1b and 3.1c, which cover the protected attributes *gender* and *race*, respectively. As can be seen from this table, the classical method, using features for computing distances between samples, shows the known bias really well. All attributes show a significant difference: male and African American offenders receive significantly more negative bias while female and non African American offenders receive significantly more positive bias. This complies exactly with the known biases from earlier work [5] [35].

When the distance between samples is computed using SHAP values, similar results can be seen. Again, the male offenders receive significantly more negative bias than positive bias, while the opposite is true for female offenders. Also from the data related to the attribute *race*, it is apparent that this method is sensible. The found significant difference between the positive and negative bias counts of African American offenders reflects the known bias against African American offenders. The data also show that non African American offenders receive significantly more positive bias. As both approaches show similar results, it cannot be concluded that one of the approaches is favoured.

3.3 Explaining existing fairness measures with SHAP

3.3.1 Statistical parity difference

The statistical parity difference has been explained using SHAP values for both models. As both models show similar results, this part of the thesis will only discuss the findings of the credit model. The results of the recidivism model can be found in the appendix. For the credit risk model, the statistical parity difference for protective attribute *gender* has a value of -0.025. This means that women have a slightly higher predicted outcome than men. Women are thus more often predicted to have a high risk score than men. The question that arises, is whether this difference can be explained legally, or if this is due to an unfair gender bias. Explaining this statistical parity difference with SHAP values results in figure 3.2. This figure

	Male			Female		
	positive bias	negative bias	p-value	positive bias	negative bias	p-value
Feature based distance function	57	12	<1e-3	14	17	0.72
SHAP based distance function	20	12	0.22	0	13	<1e-3

(a) Credit model

	Male			Female		
	positive bias	negative bias	p-value	positive bias	negative bias	p-value
Feature based distance function	269	342	0.004	96	30	<1e-3
SHAP based distance function	90	180	<1e-3	20	8	0.036

(b) Recidivism model for protective attribute gender

	African American			Not African American		
	positive bias	negative bias	p-value	positive bias	negative bias	p-value
Feature based distance function	159	292	<1e-3	297	75	<1e-3
SHAP based distance function	36	157	<1e-3	76	19	<1e-3

(c) Recidivism model for protective attribute race

Table 3.1: Positive and negative bias occurrences using the feature based distance function or the distance function using SHAP values. Table (a) shows these occurrences for the credit model on protected attribute *gender*. Table (b) shows the results of the recidivism model regarding protected attribute *gender* and table (c) concerns protected attribute *race*. The p-values are the results of binomial tests on these positive and negative bias occurrences.

shows the decomposition of the statistical parity difference among all of the input features. A negative value means that the feature has an impact on the statistical parity difference, and a positive value implies a positive impact.

From this figure it can be seen that the feature *age* has the biggest impact. One explanation could be that the data used to create the model is skewed and contains more data from women with an age that lead to a high risk score. This skewed data would then be the cause of the gender bias. Another possible explanation could be that women applying for a loan are truly more often in the sensitive age category. This would mean that the found statistical parity difference between men and women is a result of the age differences between men and women, and therefore has a legal basis. When analyzing this feature further, it appears that the women in this dataset are on average 33 years old, and men on average 37. The global SHAP values for these feature values are 0.007 and -0.030 respectively. This means that age = 37 slightly increases the outcome (thus resulting in a higher risk score) and age $\neq$ 37 slightly decreases this outcome (thus resulting in a lower risk score). This explains the impact of age on the statistical parity difference of the credit model. Further research is needed if this difference in age is true for the entire population, or only exists in this sample.

Another feature that is worth noting is the feature *f_div/sep/mar*. This feature has a value of 1 when the applicant is female (either divorced, separated or married). The dataset does not contain single female applicants, so this feature can be interpreted as the feature *female*. This feature also shows a relatively large negative impact on the statistical parity difference. This can be interpreted to mean that the statistical parity difference for gender depends partly on the gender feature. This is an indicator of unfair gender bias.

3.3.2 Conditional statistical parity difference

The conditional statistical parity difference was explained using SHAP in the same manner. Again, only the explanations of the credit model are discussed in this section. The plots of the recidivism model can be found in the appendix. The conditional statistical parity difference was computed for the credit graph while controlling for the legitimate factor *credit_account_status*. This feature entails information about the amount of money currently on your checking account. This is an important and logical feature to check when allowing a loan to an applicant, which is why it is considered a legitimate factor. The protected attribute is again *gender*. The conditional statistical parity difference has a value of 0.076. This means that men are more often predicted to have a high risk score than women, when controlling for the credit account status. In 3.3 this value is decomposed into all model features. The most prominent feature is again *f_div/sep/mar*, which now contributes positively to the fairness measure, which hints to an unfair gender bias. Note that this feature now contributes positively to the fairness measure. Apparently the gender feature has a different effect when controlling for this legitimate feature. This is an indicator of the relation between the gender bias and the legitimate factor *credit_account_status*. The attribution of feature *credit_account_status* has clearly decreased, which we would expect.

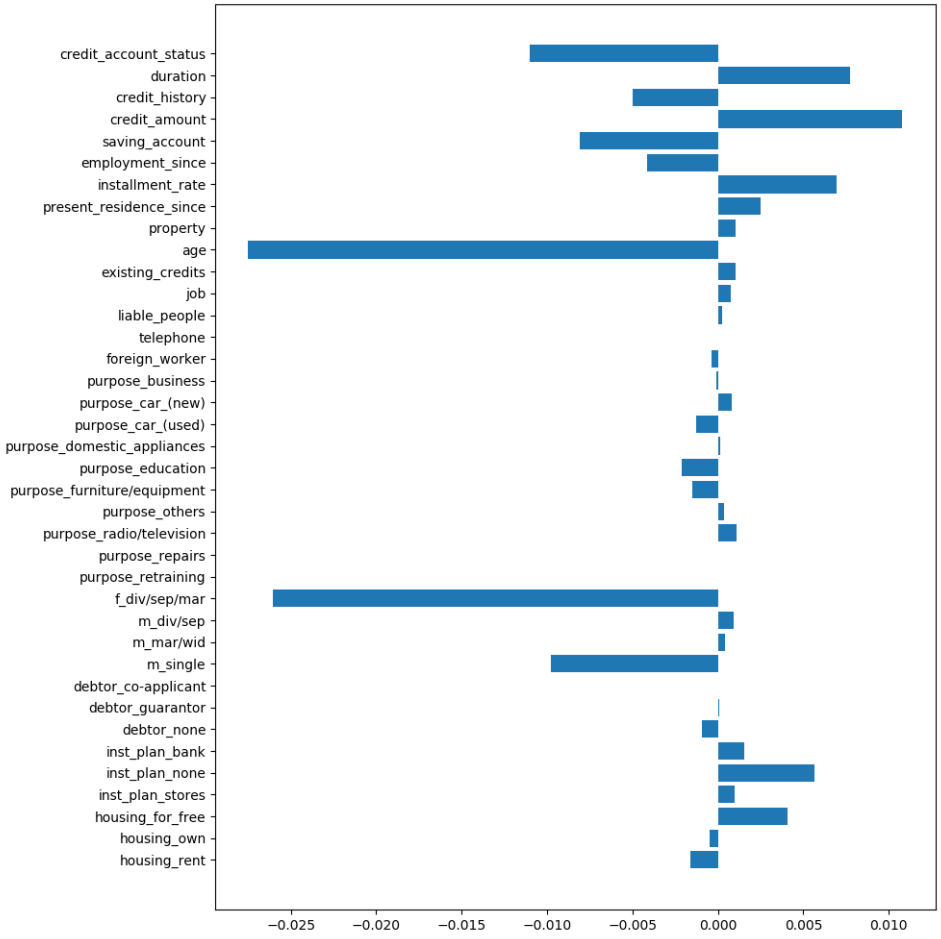


Figure 3.2: Explanation graph of the statistical parity difference of -0.025 between male and female applicants of the credit model. Attributes with large absolute values have a large impact on the statistical parity difference. The relatively large absolute value for the feature *age* could mean that the gender bias is the result of a bias on age. The relatively large absolute value for the feature *f_div/sep/mar* could mean that the model truly contains an unfair gender bias.

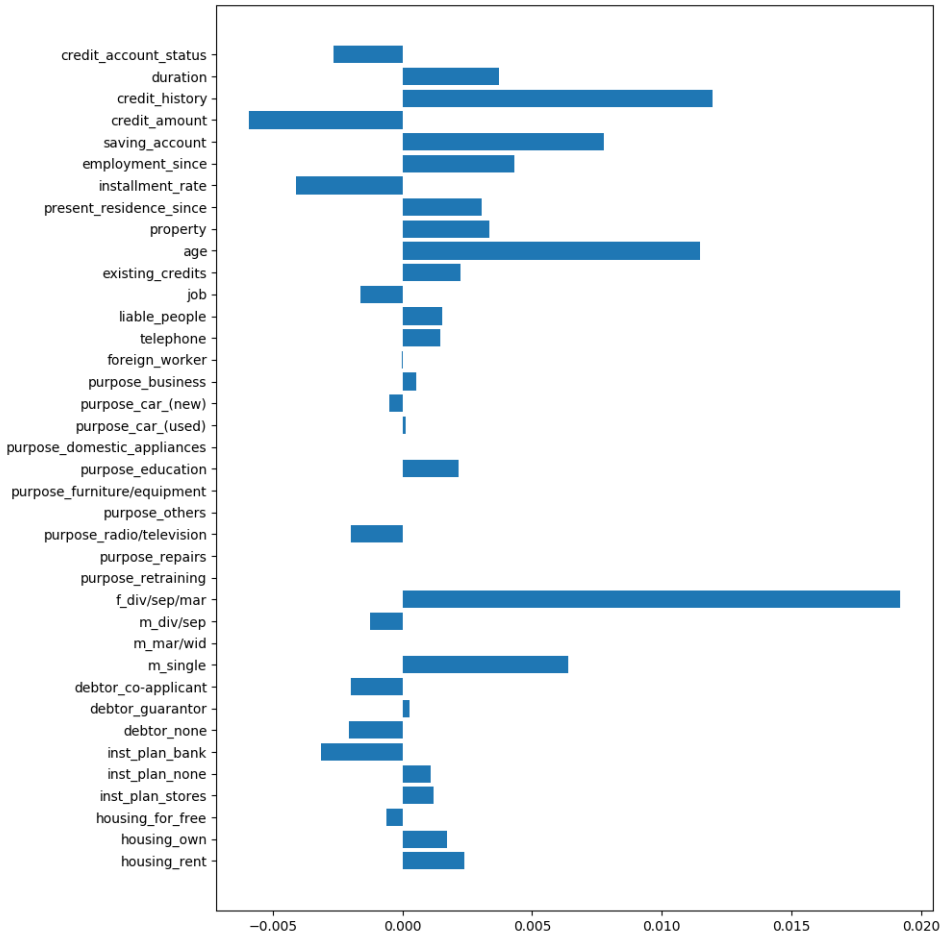


Figure 3.3: Explanation graph of the conditional statistical parity difference of 0.076 between male and female applicants of the credit model. The legitimate factor controlled for is *credit_account_status*. This attribute now has a much smaller absolute value than when not controlling for it. Additionally, the feature *f_div/sep/mar* has changed direction. It now has a positive impact on the fairness measure. This means that when controlling for the feature *credit_account_status*, the model is biased against male participants.

Chapter 4

Conclusion and discussion

The aim of this thesis was to investigate the usefulness of the explainability method SHAP in assessing fairness of a decision making algorithm. The findings of this research support the claim that SHAP is indeed useful in the field of fair artificial intelligence in three different ways. First, the study shows a relatively strong relation between SHAP and various existing fairness measures. Second, while no improvement of the fairness measure individual fairness is found, the novel approach did show equally well results as the classical approach. Third, the usability study of Lundberg's technique to explain measures of fairness with SHAP showed promising results. These three investigations together provide evidence to support this claim of usefulness.

The shown relationship between SHAP values and the fairness measures statistical parity and conditional statistical parity is a first indication of the close relation between SHAP and fair artificial intelligence. This finding can encourage future researchers to examine this relationship even further. Although the found correlation coefficients are already of significance, they do not show an unmistakably strong relationship. Future work could try to discover the reason of this disparity and find ways to improve it.

The attempt to improve the notion of individual fairness with SHAP values was not successful. When looking at the accordance of the bias with earlier research, SHAP values do not improve this notion. However, it did not show a decrease either. Similar results were obtained for the two approaches, which suggests that the novel approach is equally suitable to determine individual fairness as the classic approach. The two approaches are compared only in one manner in this thesis. A further study could aim to assess the differences between these approaches from different approaches. The results found in this study do show that it is a promising topic, which should be researched further.

The small usability study of explaining measures of fairness with SHAP values shows some promising results. This part of the study shows that SHAP values are not only useful in the field of fair artificial intelligence by substituting or altering existing measures. SHAP can also be used along with the existing measures to gain more insight in these outcomes. This application of SHAP could lead to a more substantiated use of the existing measures and a better understanding of the notion of fairness. This thesis only considers a small aspect of this use of SHAP. The only two fairness measures that are explained using SHAP are statistical parity and conditional statistical parity. Making other fairness measures suitable for this

explanation method would be an interesting topic to research further. Furthermore, the explanation method does not give one conclusive answer concerning the fairness of the model. The results show what aspects of the model are worth looking into, but only through further analysis of these aspects can one make useful conclusions about the fairness of the model. Future work could be done on creating a more exact framework to use when explaining fairness measures with SHAP.

Overall, this research has shown that SHAP can have some promising applications in the field of fair artificial intelligence. As shown in this study, SHAP shows a close relation with the existing measures and it proves to be useful when improving and explaining these measures. This is a promising result, which could lead to further research to the relation between explainable artificial intelligence and fair artificial intelligence. This in turn could lead to both a better understanding of the algorithms used in our every day lives, and an assurance of the fairness of these algorithms.

Bibliography

[1] P. Gajane and M. Pechenizkiy, "On formalizing fairness in prediction with machine learning," *arXiv preprint arXiv:1710.03184*, 2017.

[2] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7, IEEE, 2018.

[3] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[5] T. Brennan, W. Dieterich, and B. Ehret, "Evaluating the predictive validity of the COMPAS risk and needs assessment system," *Criminal Justice and Behavior*, vol. 36, no. 1, pp. 21–40, 2009.

[6] A. Lambrecht and C. Tucker, "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," *Management Science*, vol. 65, no. 7, pp. 2966–2981, 2019.

[7] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, "The case for process fairness in learning: Feature selection for fair decision making," in *NIPS Symposium on Machine Learning and the Law*, vol. 1, p. 2, 2016.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

[9] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, 2018.

[10] M. Rabin, "Incorporating fairness into game theory and economics," *The American Economic Review*, pp. 1281–1302, 1993.

[11] D. Kahneman, J. L. Knetsch, and R. H. Thaler, "Fairness and the assumptions of economics," *Journal of Business*, pp. S285–S300, 1986.

[12] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

[13] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, pp. 325–333, 2013.

[14] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial Intelligence in Medicine*, vol. 94, pp. 42–53, 2019.

[15] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[16] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[17] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[18] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5188–5196, 2015.

[19] S. Tan, R. Caruana, G. Hooker, P. Koch, and A. Gordo, "Learning global additive explanations for neural nets using model distillation," *arXiv preprint arXiv:1801.08640*, 2018.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[21] D. Linsley, D. Shiebler, S. Eberhardt, and T. Serre, "Global-and-local attention networks for visual recognition," *arXiv preprint arXiv:1805.08819*, 2018.

[22] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 3145–3153, JMLR. org, 2017.

[23] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[24] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[25] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017. `http://archive.ics.uci.edu/ml`, Accessed: 28.04.2020.

[26] Northpointe inc., "Measurement and treatment implications of COMPAS core scales," 2009. `https://www.michigan.gov/documents/corrections/Timothy_Brenne_Ph.D.__Meaning_and_Treatment_Implications_of_COMPA_Core_Scales_297495_7.pdf`, Accessed: 12.05.2020.

[27] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017.

[28] C. Rudin, C. Wang, and B. Coker, "The age of secrecy and unfairness in recidivism prediction," *arXiv preprint arXiv:1811.00731*, 2018.

[29] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: an application to recidivism prediction," *arXiv preprint arXiv:1807.00199*, 2018.

[30] H. Van Sang, N. H. Nam, and N. D. Nhan, "A novel credit scoring prediction model based on feature selection approach and parallel random forest," *Indian Journal of Science and Technology*, vol. 9, no. 20, pp. 1–6, 2016.

[31] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[34] S. Lundberg, "Explaining measures of fairness," 2020. `https://towardsdatascience.com/explaining-measures-of-fairness-f0e419d4e0d7`, Accessed: 27.03.2020.

[35] Y. Li, "Algorithmic discrimination in the US justice system: A quantitative assessment of racial and gender bias encoded in the data analytics model of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)," 2017.

Appendix A

Explanation plots of the recidivism model

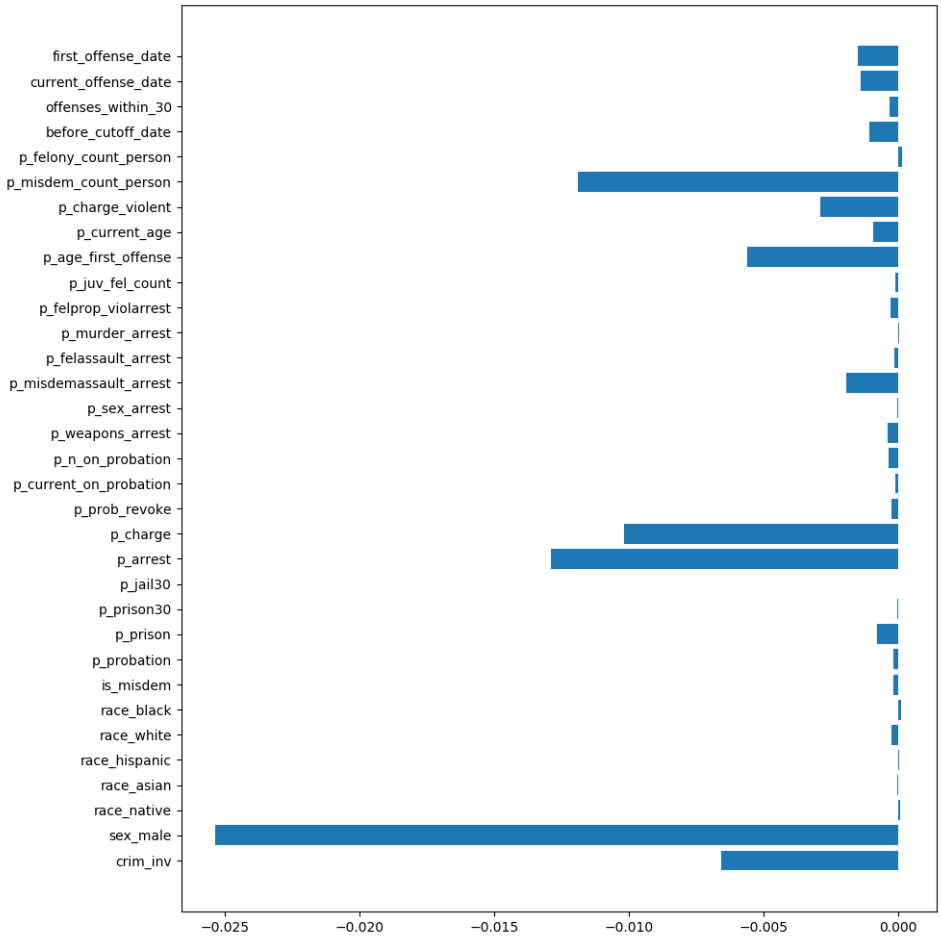


Figure A.1: Explanation graph of the statistical parity difference of -0.134 between male and female offenders. Attributes with large absolute values have a large impact on the statistical parity difference. The relatively large absolute value for the feature *sex_male* could mean that the model truly contains an unfair gender bias. The other features that show a relatively large absolute value (*p_misdem_count_person*, *p_charge*, *p_arrest*, *crim_inv*) could indicate that the bias on gender is partially a result of these legitimate features.

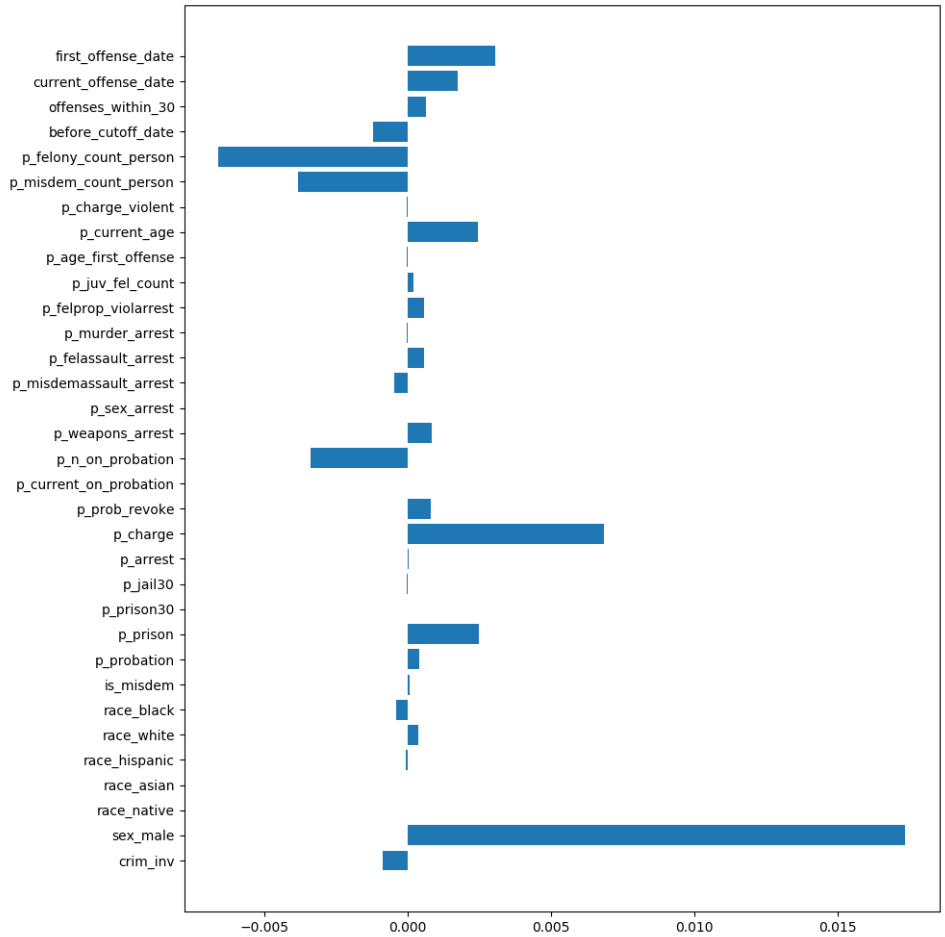


Figure A.2: Explanation graph of the statistical parity difference of 0.035 between male and female offenders, controlling for the feature *p_arrest*. This attribute now has a much smaller absolute value than when not controlling for it. Additionally, the feature *sex_male* has changed direction. It now has a positive impact on the fairness measure. This means that when controlling for the feature *p_arrest*, the model is biased against female offenders.