Assessing the Fairness of Mortgage Lending Decisions:

Leveraging Explainability to Quantify the Trade-Off between Fairness and Performance in Algorithmic Decision-Making

Hauke Schwarz

Dissertation written under the supervision of Professor **Ana Guedes**

Dissertation submitted in partial fulfillment of requirements for the **MSc in Business Analytics**, at the Universidade Católica Portuguesa, May 31st, 2024.

Abstract

Title: Assessing the Fairness of Mortgage Lending Decisions:
Leveraging Explainability to Quantify the Trade-Off between Fairness and Performance in Algorithmic Decision-Making
Author: Hauke Schwarz

The increase in the application of algorithmic decision-making has given rise to concerns about the fairness of machine-made decisions. Specifically in areas that heavily impact individuals like healthcare, recidivism, or finance, a thorough understanding of how decisions have been made is crucial from a regulatory and moral standpoint.

However, improving algorithms in terms of their predictive performance often comes at the cost of increased complexity and thereby reduced understandability. This is especially true for algorithms without a direct interpretation method like neural networks, so-called "black box" models.

This thesis aimed to examine whether 2022 Home Mortgage Disclosure Act mortgage lending data can be used to train a neural network to predict whether a mortgage will be granted based on demographic data, specifically focusing on geography and sensitive attributes such as applicant race. Both fairness and explainability requirements were incorporated into the process, all with the aim of keeping the model's predictive performance high.

Iteratively adapting the initial neural network with different fairness-focused algorithms as well as trying to uncover its inner workings using explainability algorithms showed mixed results. While some results were promising with regards to the scope of this thesis, none of the iterations applied managed to significantly improve fairness and/or predictive performance of the proposed model. This is most likely the effect of underlying discriminatory factors underlying in the data, that cannot directly be mitigated by controlling for the race of mortgage applicants, leaving an important focus for future research.

Keywords: Interpretability, Explainability, Fairness in Algorithmic Decision-Making

Abstract Portuguese

Título: Avaliando a Justiça das Decisões de Empréstimos Hipotecários:
Utilizando Explicabilidade para Quantificar o Compromisso entre Justiça e Desempenho em Tomadas de Decisão Algorítmicas
Autor: Hauke Schwarz

O aumento na aplicação de tomada de decisão algorítmica tem gerado preocupações sobre a justiça das decisões automatizadas. Em áreas com impacto significativo sobre os indivíduos, como saúde, reincidência e finanças, é crucial entender minuciosamente os processos de tomada de decisão, tanto do ponto de vista regulatório quanto moral.

No entanto, melhorar o desempenho preditivo dos algoritmos muitas vezes aumenta a sua complexidade e reduz a sua compreensibilidade. Isto é particularmente verdadeiro para modelos de “caixa-negra", como as redes neurais, que carecem de interpretabilidade direta.

Esta tese examina se os dados de concessão de hipotecas de 2022, que têm como base o Home Mortgage Disclosure Act, podem ser usados para treinar uma rede neural para prever a aprovação de hipotecas com base em dados demográficos, com foco em geografia e atributos sensíveis, como a raça do solicitante.

A adaptação iterativa do modelo de redes neurais inicial, utilizando diferentes algoritmos focados na justiça e a aplicação de várias técnicas de explicabilidade, produziu resultados mistos. Embora alguns resultados tenham sido promissores no âmbito desta tese, nenhum melhorou significativamente a justiça ou o desempenho preditivo do modelo proposto.

Tal deve-se provavelmente a fatores discriminatórios subjacentes nos dados que não podem ser mitigados apenas controlando a raça dos solicitantes, destacando uma área importante para pesquisas futuras.

Palavras-Chave: Interpretabilidade, Explicabilidade, Justiça na tomada de decisões algorítmicas

Contents

List of Figures

List of Tables

List of Formulas

List of Abbreviations

Abbreviation	Definition
adam	Adaptive Moment Estimation
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AIF360	AI Fairness 360
AUC	Area Under the Curve
AUS	Automated Underwriting System
Cal. Eq. Odds	Calibrated Equalized Odds
Corr. Rem.	Correlation Removal
csv	Comma-Separated Values
dtype	Data Type
EDA	Exploratory Data Analysis
e.g.	exempli gratia (for example)
ERS	Economic Research Service
FHA	Federal Housing Administration
F&I	Fairness and Interpretations
FIPS	Federal Information Processing Standards
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FSA	Farm Service Agency
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HMDA	Home Mortgage Disclosure Act
HOEPA	Home Ownership and Equity Protection Act
i.e.	id est (that is)
ibid.	ibidem (in the same place)
IFF	Interpretability for Fairness

knn	k-Nearest Neighbors
KPI	Key Performance Indicator
LA	Louisiana
lei	Legal Entity Identifier
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
MAR	Missing at Random
Max	Maximum
MCAR	Missing Completely at Random
Min	Minimum
ML	Machine Learning
MNAR	Missing Not at Random
MoE	Mixture of Experts
MT	Montana
Perc.	Percentage
ReLU	Rectified Linear Unit
RHS	Rural Housing Service
ROC	Receiver Operating Characteristic
SD	South Dakota
SHAP	SHapley Additive exPlanations
Std.	Standard Deviation
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
TX	Texas
USD	United States Dollar ($)
USDA	United States Department of Agriculture
UT	Utah
VA	Veterans Affairs
xAI	Explainable Artificial Intelligence
xLSTM	Extended Long Short-Term Memory

1 Introduction

The Home Mortgage Disclosure Act (*Home Mortgage Disclosure Act (HMDA) Data* 2022) is a dataset of publicly available U.S. mortgage data. Its nature of consisting of demographic information including such potentially prone to discrimination like gender, race or geography of mortgage applicants alongside information on mortgage approval or disapproval have made the HMDA an often-used source for fairness research ever since its initial release[1].

This thesis aims to make a novel contribution to this field not only by its combined analysis of fairness and explainability (as will be discussed below and in **chapter 3**), but also by the data being used. The analyses in this thesis are based on the 2022 HMDA (Home Mortgage Disclosure Act) dataset, which is a more recent data source than the academic literature analyzed relied upon. Moreover, the dataset has been enriched with additional data sources based on geographical information, which will be discussed in **chapter 4.1.2**. This enrichment does not only serve the purpose of an improved explainability, but also aims to tackle the issue of *indirect discrimination* (i.e. discrimination through non-sensitive attributes like zip codes being correlated with sensitive attributes like race), which has been discussed in academic literature as a major issue in the field of fairness (Mehrabi et al., 2021). This results in the **Research Question** that this thesis investigates:

"Can underlying unfairness in mortgage decision-making be detected, explained, and iteratively mitigated without sacrificing predictive performance?"

As the Research Question states, the second important object of analysis in this thesis next to fairness will be the aspect of explainability (which will be introduced in more detail in **chapter 2.2**). Only by understanding which factors drive model decisions, concrete measures to mitigate potential unfairness can be taken. Therefore, fairness and explainability are considered as two somewhat intertwined concepts in this work.

[1] See e.g. So et al., 2022, M. S. A. Lee and Floridi, 2021, or Singh et al., 2022.

2 Literature Review

2.1 Fairness in Algorithmic Decision Making

Due to the increasing use of decision-making algorithms for varying applications (Mehrabi et al., 2021), the topic of *Fairness* has become a heavily researched area in the field of Machine Learning and AI. While fairness concerns do not take an important role in all kinds of algorithmic decision making (some algorithms simply do not have grave enough implications to make fairness a concern, an example being buying recommendation algorithms (Marcinkevičs and Vogt, 2023)), they need to be considered in applications like hiring processes or criminal justice systems, where the decisions could heavily impact individuals (Barocas and Selbst, 2016). This is the case in the research question of this work, which focuses on mortgage lending and takes into consideration demographic attributes like gender or race, making it crucial to focus on fairness concerns.

2.1.1 Overview of Fairness in Algorithmic Decision Making

The increasing use of automated decision-making via Artificial Intelligence has shed light on how difficult it is to ensure that these algorithms act in a way humans would perceive as fair. Recent examples like an automated Amazon hiring algorithm systematically discriminating against women (Chang, 2023) prove that even the use of highly sophisticated algorithms cannot guarantee fair results when they rely on biased training data.

The intensive research of fairness in algorithmic decision making sparked by issues like this results in a wide range of definitions of fairness as well as methods to assess it (Corbett-Davies et al., 2023). Most of these are of *mathematical* or quantitative nature, however, there also is a somewhat separate approach of assessing fairness *individually* in a qualitative way (Chouldechova and Roth, 2018). It should be noted that while there are different reasons for unfairness to arise (see ibid. for a brief overview), bias in the underlying data (through discrimination, erroneous inputs or imbalances in the distributions) is the main reason for unfairness to arise (Choraś et al., 2020)[1].

A crucial element in the fairness discussion is the concept of *protected* or *sensitive attributes*. These describe attributes that are legally protected from being the basis of potentially discrimina-

[1] The reasons for bias in the underlying data are a highly interesting area of research, but out of the scope of this thesis. See Mehrabi et al., 2021 or Jui and Rivas, 2024 for an extensive list and discussion for types and reasons for biased data

tory decisions (Datta et al., 2017). Protected attributes usually contain demographic information like age, race or gender (identity) (Teodorescu et al., 2020). There are legal definitions of relevant protected attributes available specifically for the case of mortgage lending, among them race, sex, or religion (Chen et al., 2019)[2]. Even though recent studies have suggested that racial bias might not be as prevalent in both human-made and algorithmic decisions as it has been assumed (Bhutta, Hizmo, et al., 2022), it is still apparent that bias-related fairness concerns are a major issue in algorithmic decision making (Mehrabi et al., 2021). Moreover, race is just one of the protected attributes present in the dataset analyzed in this thesis, meaning that other features present in the data can still be a source of unfairness.

There are a multitude of approaches to fair ML (Machine Learning) algorithms being discussed in academic literature (compare e.g. ibid.). The main approaches are *Fair Modeling*, i.e. the development of fair ML algorithms, and *Fairness Assessment*, i.e. the evaluation of the fairness of a given model

Fair Modeling

Fair Modeling is the process of developing fair ML algorithms. This can be done in three different stages: *Pre-processing*, *In-processing* and *Post-processing* (Pessach and Shmueli, 2020).

Pre-Processing focuses on modifying the training data in a way that removes potential biases in order to have the model learn from unbiased data (Bellamy et al., 2019). Several methods have been proposed in academic literature, among them *massaging*, *reweighing* and *sampling* (Kamiran and Calders, 2012), editing features and labels of the underlying dataset following fairness criteria based on a probabilistic framework (Calmon et al., 2017), or attempts to remove disparate impact by feature value adjustment (Feldman et al., 2015).

In-Processing aims to address unfairness during model training (d'Alessandro et al., 2017). In practice, this is usually achieved by imposing constraints or adding regularization parameters to the models that explicitly aim to affect fairness criteria (Pessach and Shmueli, 2020). Practical examples from the academic literature include the *prejudice remover regularizer* introduced by Kamishima et al. (Kamishima et al., 2012), which is a regularizer that aims to enforce independence of predictions from protected attributes, or a method of imposing both accuracy and fairness constraints to tackle disparate impact proposed by Zafar et al. (Zafar et al., 2017).

Post-Processing can be applied after the model training is finished. It does however rely on the availability of previously unseen holdout data (d'Alessandro et al., 2017). Nevertheless, it is the only viable option if neither the underlying data nor the model itself can be modified (Bellamy et al., 2019). Increased fairness through post-processing can e.g. be achieved by selecting different thresholds for different demographic groups or learning different classifiers for each demographic group (Pessach and Shmueli, 2020). A practical implementation of this

[2]Note that the attributes defined in the paper by Chen et al. apply specifically to laws of the United States of America

approach is the work of Hardt et al. (Hardt et al., 2016), where a post-processing algorithm is proposed that aims to adjust the model's predictions to achieve equalized odds (i.e. equal True Positive Rates and False Positive Rates for all demographic groups) while trying to improve accuracy at the same time, relying on complete information on labels, predictions and protected attributes.

The increasing importance of fairness considerations in Machine Learning has also led to an increased availability of tools and libraries to implement fair ML algorithms. Two of the most prominent libraries are *Aequitas*[3] and *AI Fairness 360 (AIF360)*[4]. Both of these libraries offer a range of features from the fair modeling and fairness assessment steps mentioned above.

A concrete approach on how to implement fair ML algorithms in organizations and monitor their performance is proposed by Teodorescu et al. (Teodorescu et al., 2020): The proposed framework has three phases: *Design*, *Development* and *Post-hoc Model Assessment*. Based on the presence and relevance of protected attributes, the authors propose different decision steps to rule out potential unfairness in the algorithm implementation. Although being out of the scope of this thesis, this framework is a promising approach to implement fair ML algorithms in practice.

Fairness Assessment

Fairness Assessment refers to the post-hoc evaluation of the fairness of model results. The scope of fairness can roughly be divided into the following three categories (Mehrabi et al., 2021):

- **Individual fairness**: Individual fairness refers to the fairness of the model's predictions for individual instances. Examples are the concept of *Fairness through unawareness* (Kusner et al., 2017), *Fairness through awareness* (Dwork et al., 2012) or *Fairness through counterfactuals* (Kusner et al., 2017). Counterfactuals (see also **chapter 2.2.2**) are a method to assess individual fairness by analyzing how the input features would need to be changed in order to change the model's prediction (Wachter et al., 2017).
- **Group fairness**: Group fairness refers to the fairness of the model's predictions for different groups.
- **Subgroup fairness**: Subgroup fairness combines the aforementioned concepts by attempting to apply group fairness measures to subsets of the underlying dataset (Kearns et al., 2019)

[3] https://github.com/dssg/aequitas
[4] https://github.com/Trusted-AI/AIF360

2.1.2 Applications of Fairness in Mortgage Lending

Specifically in the Banking and Finance sector, algorithmic decision-making is rapidly being adopted due to the vast amount of data available and the potential upsides in terms of efficiency (Sargeant, 2022). Common usage examples include risk management, loan approvals, or the assessment of creditworthiness. While the actual level of automatization in these areas varies strongly based on factors like the bank or the amount in question, it is often hard to determine whether and to what extent decisions have been made by a human or by a machine, as in many areas, banks do not need to disclose details on that topic (Kelp and Schneider, 2023). In Europe, this is in part addressed by the General Data Protection Regulation (GDPR), which, in certain cases, requires disclosure of what factors influenced a decision (*Regulation - 2016/679 - EN - gdpr - EUR-Lex* 2024). But just as in other sectors that are partly or completely reliant on algorithms to facilitate decision-making, potential discrimination is an issue. In the USA, it is assumed that People of Color are 40-80% more likely to be denied a loan than their White counterparts, based on 2019 application data (Martinez and Kirchner, 2021).

Several studies and papers have focused on specifically analyzing the fairness of algorithmic decision making in the field of mortgage lending. Nam and Yun (Nam and Yun, 2022) follow an approach very similar to this thesis when they use Machine Learning (specifically GANs) to infer decision rules underlying the 2018 HMDA data set in order to compare the decisions made by algorithms to man-made decisions. Using the 2020 HMDA data set, So et al. (So et al., 2022) raise the point that indirect discrimination is present in historic loan application data and propose *reparative algorithms* to tackle the issue of indirect discrimination in mortgage lending. Their point is backed by other studies, such as the ones by Rugh et al. (Rugh et al., 2015) and Faber (Faber, 2013), who prove that indirect discrimination is present in mortgage lending data and that it is a major issue in the field of fairness in mortgage lending. Research concerning both fairness and explainability has been conducted by Sharma et al. (Sharma et al., 2022), who propose a Mixture of Experts (MoE) model aiming to combine these two aspects as well as the ability to detect drift over time. While this research is very promising and in some parts overlaps with the scope of this thesis, its main focus is the time component, which is not part of the scope of this thesis.

One of the most frequented sources for fairness research specifically in the area of mortgage lending is the publicly available HMDA (Home Mortgage Disclosure Act) Data Browser[5], a database on mortgage applications, originally created in 1975 to tackle discrimination against low-income borrowers (Bogen et al., 2020). Examples of the HMDA data being use to assess fairness are the paper by Ghoba and Colaner (Ghoba and Colaner, 2021), who use counterfactuals to tackle the issue of fairness in a 2019 version of the HDMA Dataset, or Singh et al. (Singh

[5] https://ffiec.cfpb.gov/data-browser/data/2022?category=states

et al., 2022), who use HMDA data from 2018 to 2020, filtered to pre-defined states in order to develop the *DualFair* loan classifier algorithm. While Singh et al. only excluded features that had a missingness >25%, Ghoba and Colaner narrowed the features used down further, only including the following:

- *interest rate*
- *applicant sex*
- *income*
- *applicant race*
- *state code*
- *loan type*
- *debt to income ratio*
- *loan to value ratio*
- *lien status*

It is suggested by previous works that protected attributes do in fact play a statistically significant role in the decision-making process of mortgage lending. Exemplarily, Lindsey-Taliefero and Kelly analyzed the role of race, age, and gender on the probability of mortgage lending specifically to research mortgages with the 2019 HMDA dataset and came to the conclusion that all of these factors had a measurable impact on the probability of a mortgage being granted (Lindsey-Taliefero and Kelly, 2021). Their results are backed by the study of Cyree and Winters, who analyzed HMDA data from 2007 to 2016 and found that every subgroup in their data was statistically discriminated when compared to White males that applied for a mortgage together with at least one co-applicant (Cyree and Winters, 2023).

Finding benchmarks for model performance directly comparable to the scope of this work in the academic literature is hampered by the recency of the used data and the individual approach to filtering states (see **chapter 3.1.1**). Papers with a similar approach were able to achieve a ROC AUC of *0.768* on a regression task aiming to predict mortgage amounts (Ghoba and Colaner, 2021) with somewhat comparable data, or reported an accuracy of 91% using a deep neural network for classification (Hodges. et al., 2024), albeit with a different timeframe and a higher amount of used features.

2.2 Explainability and Interpretability in Machine Learning

Given the constantly increasing research into AI interpretability and explainability, there is surprisingly little consent on how to precisely define these concepts and how to distinguish them (Linardatos et al., 2021). While both terms are used interchangeably in a multitude of publications, several studies have tried to distinguish both concepts in terms of their scope, giving rise to terms like 'xAI' (Explainable Artificial Intelligence) (Gunning and Aha, 2019) and occasionally also introducing related concepts like *Understandability*, *Comprehensibility* (Guidotti et al., 2018), or *Intelligibility* (Caruana et al., 2015).

One of the most adapted definitions for *Interpretability* has been made by Doshi-Velez and Kim, who define it as the "ability to explain or to present in understandable terms to a human" (Doshi-Velez and B. Kim, 2017). However, this definition does not only heavily intersect with common definitions of explainability, it also appears to be rather general and not easily applicable in a scientific context. Among other unclear definitions, this has led Lipton to state that in the scientific discussion, the term interpretability is "ill-defined", leading to many papers in this research area only exhibiting a "quasi-scientific character" (Lipton, 2018), while other authors deemed interpretability to be a "broad, poorly defined concept" (Murdoch et al., 2019). Usually, the concept of interpretability is focused on the ability to logically comprehend the inner workings of AI algorithms, i.e. the user being able to predict outputs from inputs (B. Kim et al., 2016).

Explainability or *xAI* (which will be used interchangeably with the term interpretability in this work), even though being subject to a similarly wide range of definitions, is usually defined to be more concerned with explaining the rationale behind decisions made to generate trust instead of precisely dissecting the inner workings mathematically (Gunning and Aha, 2019). Compared to interpretable AI, which has been discussed in academic literature for a comparably longer time-frame, explainability is a newer concept, which however seems to gather momentum in academic interest very fast, most likely due to the more and more widespread adoption of not inherently explainable Deep Learning Models (Barredo Arrieta et al., 2020).

2.2.1 Inherently Interpretable Models vs. Black Box Models

Machine Learning algorithms vary in their degree of interpretability. There usually is a trade-off between their *predictive accuracy* (i.e. how well they perform on prediction tasks) and their *descriptive accuracy* (i.e. how well they can be understood by humans) (Murdoch et al., 2019), although more recent studies are challenging this assumption, compare e.g. Cooper et al., 2024. While some models, like linear regressions, are inherently interpretable, others, like Neural

Networks, are considered to be 'black boxes' (Guidotti et al., 2018) or 'opaque models' (Burrell, 2016), as their inner workings are not intuitively understandable for humans. However, there is increased demand for models that have a high predictive accuracy while still being explainable due to, among others, legal requirements like the GDPR (*Regulation - 2016/679 - EN - gdpr - EUR-Lex* 2024) and ethical considerations (Guidotti et al., 2018), leading to increased demand for explainability algorithms.

2.2.2 Explainability Algorithms

In the academic literature, explainability methods are usually categorized based on their properties (see **figure 2.1**), such as:

- **Model-specific vs. Model-agnostic**: *Model-specific* methods are designed to explain the outcomes of a specific model, while *model-agnostic* methods are designed to be applicable to (nearly) any model by describing how individual features influence the model outcome on average (Molnar, 2023).
- **Local vs. Global**: Model-agnostic methods can be either *local* or *global* (ibid.). Local methods are designed to explain individual predictions, while global methods should explain the model's behavior as a whole (Saleem et al., 2022).
- A special case of model-agnostic local explanations are **Counterfactual Explanations**, which aim to explain predictions by analyzing how the input features would need to be changed in order to change the model's prediction (Wachter et al., 2017).
- **Post-hoc vs. Ante-hoc**: *Post-hoc* methods are applied after the model outcome has been generated, but can only explain individual outcomes without making the model workings transparent (Lipton, 2018), while *ante-hoc* methods are usually model-specific algorithms that aim to make all the steps taken by a model transparent. (Saleem et al., 2022).

The two most widely adapted local, model-agnostic, post-hoc methods are *SHAP* (SHapley Additive exPlanations) (Lundberg and S.-I. Lee, 2017) and *LIME* (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016).

SHAP is based upon the use of Shapley values used in coalitional game theory and aims to weight the importance of each feature to the result by game-theoretically distributing the value of the final prediction among all features being included (Molnar, 2023). The key features of SHAP are (based on ibid.)

- **Additivity**: All feature contributions can be summed up in a linear way, benefitting understandability of the explanations.

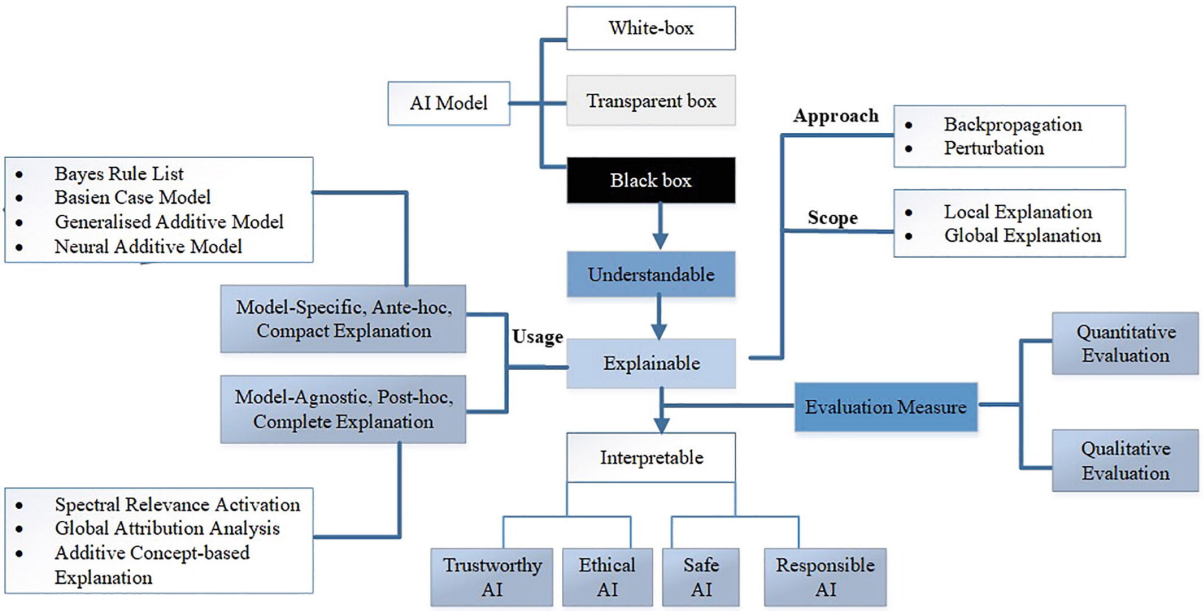


Figure 2.1: Explainability Overview, from Saleem et al., 2022 - The concept of *explainability* must be understood in the context of similar and related concepts like *understandability*.

- **Local Accuracy**: SHAP values are locally accurate, as predictions for a given input can be calculated as a result of the expected model output and the specific score for the local input.
- **Missingness**: As missingness in features is attributed zero values, missingness does not skew outcomes.
- **Consistency**: SHAP values do not change upon model change, but only when the actual feature contributions change.

LIME is based upon trying to approximate the model's decision boundary locally by perturbing the input data and observing the change in the model's output (Molnar, 2023).

A way of globally assessing model explainability is the use of **(Global) Surrogate Models** (as opposed to e.g. the LIME algorithm, which can be considered a local surrogate model). These aim to make black-box algorithm decisions explainable by implementing white-box models to learn the decision criteria that the black-box model has established by being trained on the black-box predictions (compare e.g. Karim et al., 2023). Instead of making the predictions themselves explainable, these models aim to making the underlying decision criteria understandable (Molnar, 2023).

2.2.3 Explainability and Fairness

Explainability and fairness within the field of Machine Learning do not inherently share the same scope. On the one hand, as an example, explainability is a rather objective concept, focusing on describing a model's inner workings neutrally. Comparing that to fairness it becomes apparent that the latter is a more subjective concept, as it is dependent on the context and man-made definitions, which are often rooted in certain (e.g. political) interpretations of fairness (Padmanabhan et al., 2021). Yet, both concepts are somewhat dependent as assessing model fairness requires an understanding of the factors influencing a prediction (Zhou et al., 2022). Subsequently, approaches to connect both concepts have been proposed in academic literature. However, these concepts are only stated on theoretical level without practical implications in this paper and are therefore not directly applicable for practitioners.

Another promising approach to merge the concepts of fairness and explainability is the identification of root-causes for a certain model behavior by identifying *influential subsets* within the underlying datasets as proposed by Pradhan (Pradhan et al., 2022)[6]. By combining explainability criteria ('Why is a certain subgroup influential on the model's outcome?') with fairness criteria ('In how far can intervention in the training data regarding these subsets improve fairness?'), the authors manage to implement an efficient algorithm combining both concepts.

Just as it is the case for the aspect of fairness (as discussed in **chapter 2.1.2**), the field of mortgage lending has been specifically targeted by previous research on the aspect of explainability. Maass et al. contribute to the discussion by proposing the use of *conceptual models* to capture the relationships in data, exemplarily using the HMDA 2020 dataset to predict mortgage approvals (Maass et al., 2022). By using a total of 52 features clustered into different *concepts* (i.e. abstracted groups of features), they determine the most important concepts relative to one another for the model's decision-making process.

In a comparative review of different explainability methods, Anderson evaluates the accuracy of explanations of different ML models on the 2019 HMDA dataset (Anderson, 2023). While all of the algorithms reviewed *(most points lost, Shapley values, Shapley additive explanations, and the method of integrated gradients)* were able to provide explanations for the model's decisions, the accuracy of the explanations varied greatly between the different methods, with SHAP being the overall best-performing approach.

A very recent contribution to a research question from a similar field as this thesis has been made by Shiam et al. (Al Shiam et al., 2024), who leverage explainable AI to assess the risk

[6] https://github.com/romilapradhan/gopher

of credit default (as opposed to mortgage approval focused in this work), and apply SHAP to explain predictions of a default classifier

This thesis will add value to the discussion of fairness and explainability in Machine Learning by providing a novel approach to the analysis of both concepts in the field of mortgage lending: By using explainability algorithms and enrichment data, the analysis of fairness will be supported and iteratively adjusted with the aim of finding the optimal balance of both concepts.

3 Data and Methodology

3.1 Data

3.1.1 HMDA Data

The main dataset was retrieved from the HMDA Data Browser[1] (see **chapter 2.1.2**). The data selection process is loosely based on the paper by Ghoba and Colaner (Ghoba and Colaner, 2021, see **chapter 2.1.2**), but adjusted to the scope of this thesis (see below). All data were generated in **2022**, which supports the claim of this thesis to provide data that are newer than the studies available and potentially already include effects from the COVID pandemic. **All financial institutions** were included, but to narrow down the research focus to racial fairness specifically, the approach of Ghoba and Colaner was followed by only including the two predominant racial groups (**Non-Hispanic White** and **Black or African Americans**) in the analysis. In terms of geographical filtering, inspiration was drawn from the paper by Singh et al. (Singh et al., 2022, see **chapter 2.1.2**), but instead of including different-sized states, a fairness-based approach was taken: U.S.News publishes a ranking of US states by their equality, by measure of distributions by race, gender and other fairness-related aspects and aggregate them to overall equality levels[2]. To use a subset of the HMDA data that is likely to include fairness issues, the five states considered least equal in the year 2023 were included in the data, being **South Dakota, Louisiana, Utah, Texas,** and **Montana**. The resulting dataframe has *867.401 rows and 99 columns*.

The features included for the analysis are close to the ones proposed by Ghoba and Colaner to maintain a degree of comparability, but adjusted to the scope of this thesis by including a different target variable as well as different geographical features: *action_taken, county_code, interest_rate, applicant_sex, applicant_race-1, loan_type, debt_to_income_ratio, loan_to_value_ratio,* and *lien_status*[3]. Missing values were present in *interest_rate (320.335), loan_to_value_ratio (257.408), debt_to_income_ratio (229.442),* and *county_code (9.053)*. Furthermore, three columns contained the string *Exempt* as a value *(interest_rate: 20.591 occurrences; loan_to_value_ratio, and debt_to_income_ratio: 20.533 occurrences each)*, preventing them from initially being cast as numerical types. In order to clean and reshape the data in a easily analyzable format, several steps were taken that are described in the following and summarized in **table 3.1**.

[1] https://ffiec.cfpb.gov/data-browser/data/2022?category=states

[2] 2023 ranking:https://www.usnews.com/news/best-states/rankings/opportunity/equality?sort=rank-desc, methodology: https://www.usnews.com/news/best-states/articles/methodology

[3] https://ffiec.cfpb.gov/documentation/publications/loan-level-datasets/lar-data-fields

	Transformation Step	Reasoning
Rows with missing county code	Dropping	Insignificant amount, imputation not feasible, presumably MCAR
All variables	Typecasting	Required for further analysis
loan_granted	Creation	Target variable for the classification task
Exempt	Recoding	Recoding to zero for proper typecasting
debt_to_income_ratio	Binning	Reduction of total categories
loan_to_value_ratio	Outlier Removal	Few outliers skewed the initial distribution
interest_rate and **loan_to_value_ratio**	Imputation	Mean Imputation to tackle missingness

Table 3.1: Transformation Steps in the HMDA Mortgage Data - Several transformation steps have been applied to the HMDA mortgage data in order to make it easily analyzable.

All rows including missing values for the *county_code* were **dropped**, as their relative amount was insignificant, imputation was not logically possible for this variable, and missingness completely at random could be assumed from visual inspection using the missingno package[4] (see **Figure 3.1**). All features were **cast** to their appropriate types *(county_code: string; applicant_race-1, applicant_sex, lien_status, loan_type: category)*. Furthermore, a **new target variable** *(loan_granted)* was created from *action_taken* (which was dropped), assuming 1 for granted loans and 0 for denials instead of including different reasons for (dis-)approval.

All *Exempt* strings were **recoded** as zero, allowing proper typecasting (float) for the *interest_rate* and *loan_to_value_ratio* variables. For *_to_income_ratio*, further **binning** was introduced, creating the new categories *36%-41%, 41%-45%*, and *46%-49%*. As later analysis suggested that missingness in the *debt_to_income_ratio* was a feature that heavily impacted predictions, the missingness was encoded as a separate category. After assessing the distributions of *interest_rate* and *loan_to_value_ratio*, extreme outlier values *(loan_to_value_ratio > 250)* were **dropped**.

To tackle missingness in *interest_rate* and *loan_to_value_ratio*, **imputation** was utilized. As the preferred way of imputing, the KNNImputer from the scikit-learn package[5], proved to be too

[4]https://github.com/ResidentMario/missingno
[5]https://scikit-learn.org/stable/

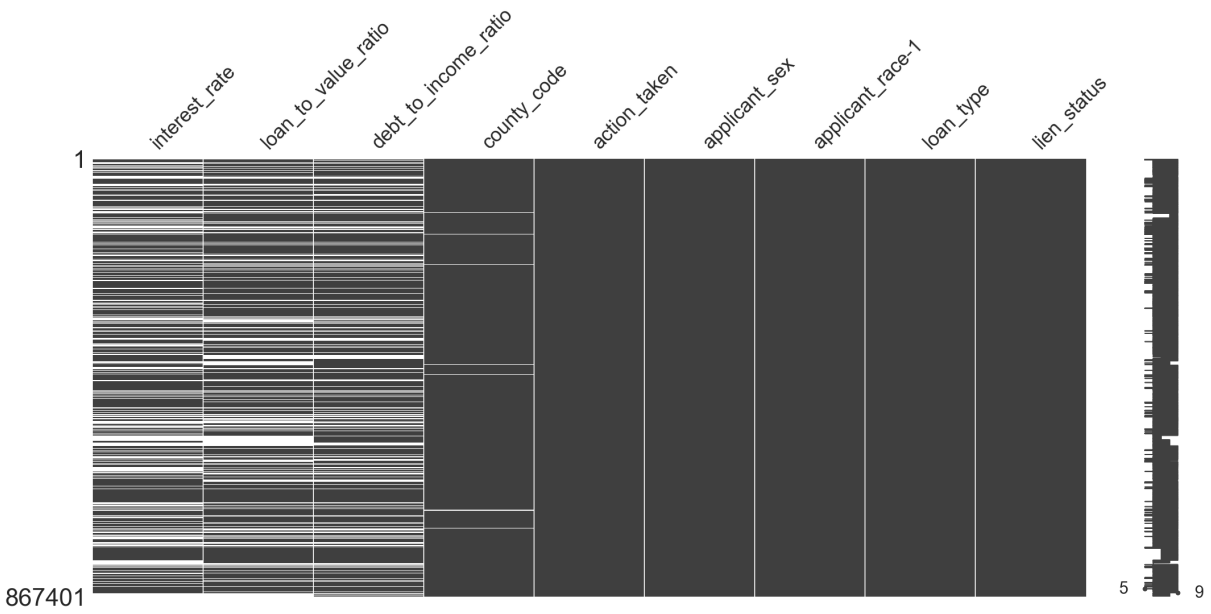


Figure 3.1: Visual Inspection of Missingness in the missingno Package - Visual Inspection of Missingness shows that some datapoints seem to be missing at random in *interest_rate*, *loan_to_value_ratio*, and *debt_to_income_ratio*, while the other features expose no or very little missingness.

computationally expensive to be efficiently used, the IterativeImputer from the same package was used as an alternative. This multivariate imputing technique is supposed to infer feature values from other features[6]. It is however set up to default to the mean as the imputation value if no satisfying solution is found.

In order to prepare the HMDA data for the classification task, the numerical variables *(interest_rate, loan_to_value_ratio)* were *standardized* using Z-score standardization:

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

The target variable *loan_granted* was **encoded** as a binary variable. Furthermore, all categorical variables were **one-hot encoded**, dropping the first column to avoid multicollinearity. The data was **split** into *train, test*, and *validation* sets with an 80/20 split each using the *train_test_split* function from sklearn. These steps add to the list of transformation steps that were applied to the HMDA mortgage data, which are summarized in **table 3.2**.

[6]https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

	Transformation Step	Reasoning
Rows with missing county code	Dropping	Insignificant amount, imputation not feasible, presumably MCAR
All variables	Typecasting	Required for further analysis
loan_granted	Creation	Target variable for the classification task
Exempt	Recoding	Recoding to zero for proper typecasting
debt_to_income_ratio	Binning	Reduction of total categories
loan_to_value_ratio	Outlier Removal	Few outliers skewed the initial distribution
interest_rate and **loan_to_value_ratio**	Imputation	Mean Imputation to tackle missingness
interest_rate and **loan_to_value_ratio**	Standardization	Z-Score Standardization for better model performance
Categorical Variables	One-Hot Encoding	Required for model training
Data	Splitting	Train, Test, and Validation sets for model training and testing

Table 3.2: Transformation Steps in the HMDA Mortgage Data (including preparation for model training and testing) - Several transformation steps have been applied to the HMDA mortgage data in order to make it easily analyzable.

3.1.2 Enrichment Data

The enrichment data was obtained from the *USDA ERS page*[7]. All data used are structured, tabular data that are publicly available. Privacy concerns are not relevant, as the data is anonymized and aggregated. All available reports were downloaded, specifically the following datasets:

- **Poverty** (2021 latest)
- **Population** (2022 latest)

[7] https://www.ers.usda.gov/data-products/county-level-data-sets/

- **Unemployment, and Median Household Income** (annual average 2022 unemployment and 2021 median income latest)
- **Education** (2017–21, 5-year average latest).

In order to clean and reshape the data in a easily analyzable format, several steps were taken that are described in the following and summarized in **table 3.3**.

	Transformation Step	Reasoning
Geography	Filtering	Only include the five states in question *("SD", "LA", "UT", "TX", "MT")*
Year (where applicable)	Filtering	Only include the newest datapoints
Features	Reducing	Only include the most relevant features *(Poverty, Population, Unemployment and Median Household Income, Education)*
Data Structure	Pivoting	Use the attributes as feature names
Indexing	Indexing	Use the FIPS code and the name of the respective county as index
Renaming	Renaming	Rename the feature columns for clarity
Merging	Merging	Merge all datasets into a single dataframe

Table 3.3: Transformation Steps in the Enrichment Data - Several transformation steps have been applied to the geographical enrichment data in order to make it easily analyzable.

All datasets were *filtered* to only contain county-level data for the five states in question *("SD", "LA", "UT", "TX", "MT")*, not the aggregated values for the full states, and, in case multiple years of analysis were available, to only include the newest datapoints. Where there were more features available than would be useful for the analysis, the datasets were *reduced* to only include the most relevant features, being:

- **Poverty:** Only the percentage of the population living in poverty (PCTPOVALL_2021) was included
- **Population:** Only the total population (POP_ESTIMATE_2022) was included
- **Unemployment, and Median Household Income:** The unemployment rate (Unemployment_rate_2022) and the median household income (Median_Household_Income_2021) were included

- **Education:** All relative values, i.e. percentages of adults with their corresponding highest degrees were included.

All datasets were *pivoted* in order to use the attributes as feature names and *indexed* by the FIPS code and the name of the respective county. After basic *checks for completeness*, the feature columns were *renamed* for clarity. Finally, all datasets were *merged* into a single dataframe, which was then *exported* as a pickle file for further use in the analysis.

3.2 Methodology

The methodology used in this research is summarized in the flowchart in **figure 3.2**. It shows the three different main tasks that were focused on in this work:

1. A Neural Network model aiming to predict mortgage approval was trained and tested on the HMDA mortgage dataset, serving as a **Benchmark** (see **chapter 3.2.1**). Its performance was evaluated using a set of predefined metrics, *metrics #1*, and its fairness was assessed using a set of fairness metrics, *metrics #2* (see **chapter 3.2.1**)[8].
2. Insights into the decision criteria of the benchmark model were gained through the application of three different **Explainability** algorithms: SHAP, LIME, and a Global Surrogate Model. Furthermore, geographical *Enrichment Data* were utilized to further the understanding of which criteria may influence mortgage approval. For details on that step see **chapter 3.2.2**.
3. Finally, three different **Fairness Adjustments** algorithms were applied to the model or the data preparation process iteratively, each aiming to improve either *metrics #1* or *metrics #2*. These algorithms are described in **chapter 3.2.3**.

3.2.1 Mortgage Classifier (Benchmark)

Initially, a model aiming to predict mortgage approval as a classification task was set up. The model chosen comprised a comparably simple sequential neural network, implemented with the keras package[9]. Model details are provided in **table 3.4**.

To increase the efficiency of the training process and to prevent overfitting, *callbacks* for early stopping (with a patience of 5 iterations) and best model selection, both based on the validation loss, have been implemented. *Adam* was chosen as the optimizer, due to the nature of the classification task, loss evaluation was based on *binary crossentropy*, and *accuracy* was selected

[8] All code produced for this thesis can be found at https://github.com/HaukeSchwarz/Thesis/tree/main/notebooks
[9] https://keras.io/about/

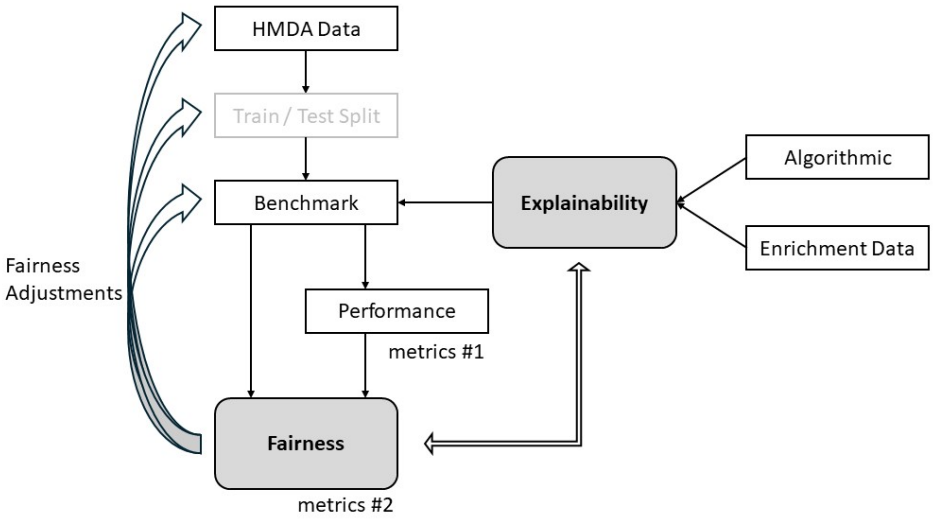


Figure 3.2: Methodology - The procedure chosen for this thesis follows an iterative approach: An initial model for classification of mortgage (dis)approval is trained and tested, serving as a *Benchmark*. For this benchmark, fairness and performance are assessed based on a predefined set of metrics, *metrics #1* for model performance and *metrics #2* for model fairness. Additionally, different *Explainability* techniques will be leveraged to gain insight into the decision criteria underlying the predictions of the benchmark model. Subsequently, three different *Fairness Adjustments* are made to the model and the data preparation process iteratively, each aiming to improve one or more of the metrics. The results of each iteration are once again evaluated using the pre-defined metric sets.

as the target metric. Training took place with a batch size of 48 and for a maximum of 30 epochs.

Performance Assessment (metrics #1)

The model originally put out numerical probabilities, which were then converted to binary values, using a threshold of 0.5. Using these original probabilities, **ROC AUC** curves could be plotted and the ROC AUC score could be calculated, which was used as one of the assessment measures. Additionally, a common set of classification metrics, being **accuracy**, **precision**, **recall**, and **F1 score** (see **equations 3.2** to **3.5**), are calculated alongside the **confusion matrix**. This set of performance metrics is referred to as *metrics #1* in the methodology (see **figure 3.2**).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

Layer (type)	**Output Shape**	**Param #**
dense	(None, 32)	640
dense_1	(None, 64)	2112
dropout	(None, 64)	0
dense_2	(None, 128)	8320
dropout_1	(None, 128)	0
dense_3	(None, 64)	8256
dropout_2	(None, 64)	0
dense_4	(None, 1)	65
Total params		19,393
Trainable params		19,393
Non-trainable params		0

Table 3.4: Summary of the Neural Network - The neural network consisted of 5 dense layers (with 32, 64, 128, 64, and 1 neuron per layer), with 3 dropout layers (with dropout rates of 0.1, 0.25, and 0.1 respectively) in between. L2 regularization (0.001) is utilized in each dense layer. The total number of parameters is 19,393.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.4}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3.5}$$

Fairness Assessment (metrics #2)

The aequitas package provides a multitude of fairness metrics to provide model fairness. In order to work with a uniform framework in this thesis, four easy to understand yet highly relevant metrics are chosen: The disparities in **False Positive Rate** (FPR) and **False Negative Rate** (FNR), as well as the **True Positive Rate** (TPR) and **True Negative Rate** (TNR) are calculated for the different groups. Analyzing these will inform about how much more likely the model is to make one of the four predictions for Black or African American Americans compared to White applicants. From the original paper (Saleiro et al., 2018), the formulas for the calculation of the disparities can be inferred (see **equations 3.6** to **3.9**):

$$FPR_{g_{\text{disp}}} = \frac{FPR_{a_i}}{FPR_{a_r}} = \frac{\Pr(\hat{Y} = 1|Y = 0, A = a_i)}{\Pr(\hat{Y} = 1|Y = 0, A = a_r)} \tag{3.6}$$

$$FNR_{g_{\text{disp}}} = \frac{FNR_{a_i}}{FNR_{a_r}} = \frac{\Pr(\hat{Y} = 0|Y = 1, A = a_i)}{\Pr(\hat{Y} = 0|Y = 1, A = a_r)} \tag{3.7}$$

$$TPR_{g_{\text{disp}}} = \frac{TPR_{a_i}}{TPR_{a_r}} = \frac{\Pr(\hat{Y} = 1|Y = 1, A = a_i)}{\Pr(\hat{Y} = 1|Y = 1, A = a_r)} \tag{3.8}$$

$$TNR_{g_{\text{disp}}} = \frac{TNR_{a_i}}{TNR_{a_r}} = \frac{\Pr(\hat{Y} = 0|Y = 0, A = a_i)}{\Pr(\hat{Y} = 0|Y = 0, A = a_r)} \tag{3.9}$$

Following the original paper, the optimal value for the disparities is **1**, as this would indicate parity between the groups. This means, that increased fairness is indicated by values closer to **1** when comparing model results to each other. Analyzing these disparities result in the first set of fairness metrics assessed, denoted as *metrics #2* in the methodology (see **figure 3.2**).

The second part of *metrics #2* is a more specified analysis of the model performance for the different groups. If the model performed vastly different for the races in question, this would indicate a potential bias in the model. Therefore, race-specific performance metrics were analyzed. More precisely, the outcomes were masked to only include either **White** or **Black or African American** applicants respectively, and the following metrics were calculated for each group:

- **Accuracy**: The percentage of correct predictions made by the model.
- **Precision**: The percentage of correct positive predictions made by the model.
- **Recall**: The percentage of actual positive cases that were correctly predicted by the model.
- **F1 Score**: The harmonic mean of precision and recall.
- **AUC** (where applicable): The area under the ROC curve.

3.2.2 Explainability

As laid out in the research question for this thesis (see **chapter 1**), an important part of assessing the models fairness is the ability to actually explain its decisions. A total of three different explainability algorithms (see **chapter 2.2.2** for theoretical background) were utilized for two reasons: Firstly, to provide a comprehensive overview of the models' decision-making process,

and secondly, to understand potential differences in the results of the algorithms, in case any may arise.

- The **SHAP** algorithm was applied to a set of individual predictions, aiming to understand feature importance and the impact of individual features on the model's decision (Lundberg and S.-I. Lee, 2017).
- The **LIME** algorithm was used to challenge the predictions made by the SHAP algorithms to assess whether both provide comparable explanations (Ribeiro et al., 2016).
- A **Global Surrogate Model** was trained on the model's prediction to assess whether the local explanations made by LIME and SHAP are also reflected on a global level (compare e.g. Molnar, 2023).

The combination of these three algorithms aims to benefit not only model understanding, but also the analysis of potential differences in the predictions.

A separate approach to uncover factors that might influence the predictions of the model and therefore improve the ability to explain it is the use of geographical data (see **chapter 4.1.2**). As the dataset contains information on the geographical location of the applicants, it was possible to analyze the distribution of granted loans across different regions by joining the mortgage data to economic variables of the corresponding regions using FIPS identifiers. While this procedure does not build upon the usage of an explainability algorithm per se, it does support the understanding of the model's decision-making process (thereby addressing both explainability and fairness concerns) by providing additional context that might not be directly grasped from the data itself.

In order to operationalize this approach, the following steps were taken:

- The data was aggregated on *county_code* level.
- Aggregated KPIs were created:
  - **Sum of Applications**: The total number of applications per county, mainly important to filter out counties with a low number of applications.
  - **Percentage of Grants**: The overall likelihood of a positive decision per county.
  - **Percentage White Applicants**: The percentage of White applicants per county.
- These KPIs were then related to the predictions made by the model by creating the **Perc. Pred. Grants** feature, which is the percentage of granted loans per county predicted by the model.
- Based on this set of information, several scatterplots relating these factors were created to analyze the distribution of granted loans across different regions, incorporating information on potentially discriminating factors.

3.2.3 Fairness Adjustments

Based on the results of the performance and fairness assessments for the initial model run (see **chapter 3.2.3**), adjustments to the model and the data preparation process were made in multiple iterations, aiming to improve at least one of these aspects in each run.

As a first iteration, **Reweighing** was applied. This *pre-processing* procedure was initially developed by Calders et al. (Calders et al., 2009) and is implemented in the **AIF360** package[10]. It works by adding a weight to each sample in the training data with the aim of balancing the weights of the different groups without actually adjusting any values. The practical application of this technique includes the following steps: Initially, the weights of the samples need to be calculated. This can be achieved using the *Reweighing* class of the *aif360.algorithms.preprocessing* module. Being supplied with the training dataset and information on the privileged group (in this case, the *White* applicants) and the unprivileged group (*Black and African American* applicants), the algorithm calculates the weights for each sample. In order to ensure comparability of the results, an exact copy of the neural network described in **figure 3.4** is created and compiled. During fitting however, the weights calculated by the reweighing algorithm are passed to the *sample_weight* parameter of the model, causing keras to apply them during model fitting. The results of this iteration can be found in **chapter 3.2.3**.

The second iteration consisted of the application of the **Correlation Remover**, another pre-processing technique that was proposed by Weerts et al. in their paper on the Fairlearn Python package (Weerts et al., 2023). It aims to remove any correlation between the sensitive attribute and the features of the dataset, while changes in non-sensitive features are kept as low as possible. Practically implementing this technique requires a definition of the sensitive attribute s (as opposed to non-sensitive attributes z), which will then be used to fulfill the constraint of the optimization problem as displayed in **equation 3.10**. Once again, the neural network needs to be retrained on the dataset after the application of the Correlation Remover. It must however be noted, that, as opposed to the reweighing technique, the underlying data are actually altered by this technique, meaning that a model fitted on correlation-removed data will only perform well on correlation-removed validation and testing data. The results of this iteration can be found in **chapter 3.2.3**.

$$\begin{aligned} \min_{z_1,\ldots,z_n} \quad & \sum_{i=1}^{n} ||z_i - x_i||^2 \\ \text{subject to} \quad & \frac{1}{n}\sum_{i=1}^{n} z_i (s_i - \bar{s})^T = 0 \end{aligned} \tag{3.10}$$

[10] https://github.com/Trusted-AI/AIF360

As a third iteration, a post-processing technique was applied: The **Calibrated Equalized Odds Postprocessing** algorithm, initially proposed by Pleiss et al. (Pleiss et al., 2017). It aims to satisfy the *Equalized Odds* criterion (see **chapter 2.1.1**) while keeping the results *calibrated*, i.e. making sure that the predictions probabilities are interpretable as levels of confidence. This algorithm has been implemented in the *AIF360* package [11] and can be applied to the model's predictions after the model has been trained by transforming the data with regards to the selected fairness constraint. The results of this iteration can be found in **chapter 3.2.3**.

[11] More information on practical applications can be found under `https://github.com/Trusted-AI/AIF360/blob/main/examples/demo_calibrated_eqodds_postprocessing.ipynb`

4 Results

4.1 Exploratory Data Analysis

4.1.1 HMDA Data

After all preparation steps detailed in **chapter 3.1.1**, the dataset contained *851,936* observations and *9* features (including the target variable, *loan_granted*).

EDA of the Target Variable

The target variable, *loan_granted*, is binary (*Granted* or *Not Granted*) and slightly imbalanced, with *57.9%* of all loans being granted. There were no missing values. The distribution of the target variable can be seen in **figure 4.1**.

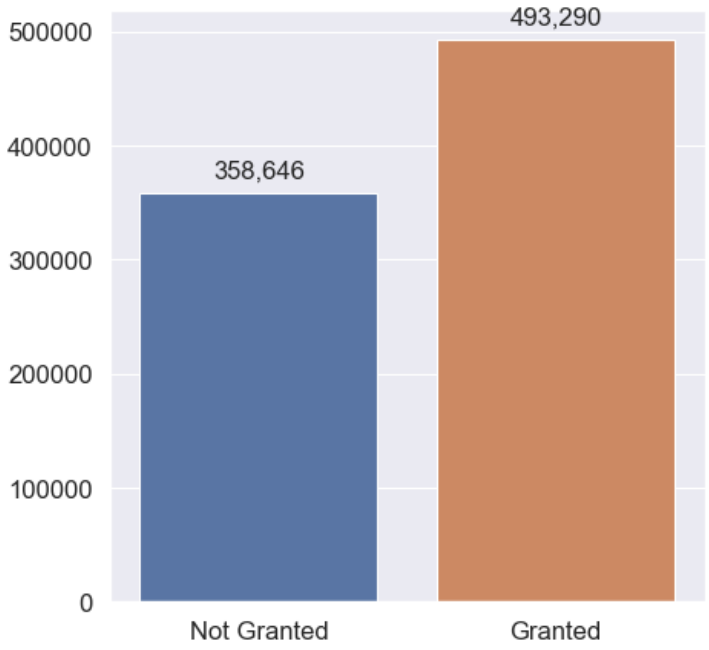


Figure 4.1: HMDA: Distribution of the Target Variable - The target variable is slightly imbalanced, with 57.9% (493,290) of all loans being granted.

Analyzing the correlation of the target variable with the categorical and numerical features showed that the *debt_to_income_ratio* feature has the highest correlation (moderate negative correlation of **-0.61**) with the target variable, while the other features show only weak correlations. The correlation of the target variable with the numerical and categorical features was calculated by pairwise Pearson correlation, with the categorical features being encoded as

integers before the calculation. The results can be seen in **table 4.1** and **figure 4.2**, respectively.

Feature	Correlation Coefficient
Race	0.11
Loan Type	0.025
Lien Status	-0.0051
Sex	-0.034
Debt to Income Ratio	-0.61

Table 4.1: HMDA: Correlation of the Target Variable with Categorical Features - While there are only weak correlations between *granted loans* and *applicant race, lien status, loan type*, and *applicant sex*, there is a moderate negative correlation of **-0.61** between *granted loans* and *debt to income ratio*.

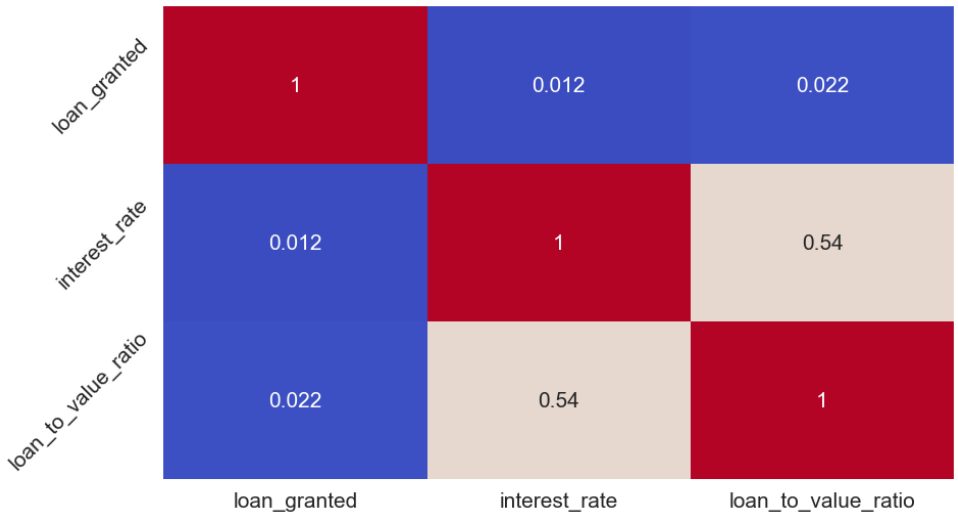


Figure 4.2: HMDA: Correlation Between the Numerical Features - While both numerical features are only weakly correlated with *loan_granted*, there is a moderate positive correlation of **0.54** inbetween them.

EDA of the Features

The processed dataset contained two *numerical* features: **interest_rate** and **loan_to_value_ratio**. Their distributions can be seen in **figure 4.3**.

Exploratory data analysis of the categorical variables showed that the majority of loan applicants are **White** (65%) and **Male** (85%), resulting in 57% of all applicants being both White and Male. Most loans applied for are **Conventional** (82%) and **First Lien** (86%). Aside from missingness in the *debt_to_income_ratio* feature, which alone accounts for 26% of all values, the data in this feature are roughly normally distributed, with the mode being 14% of values

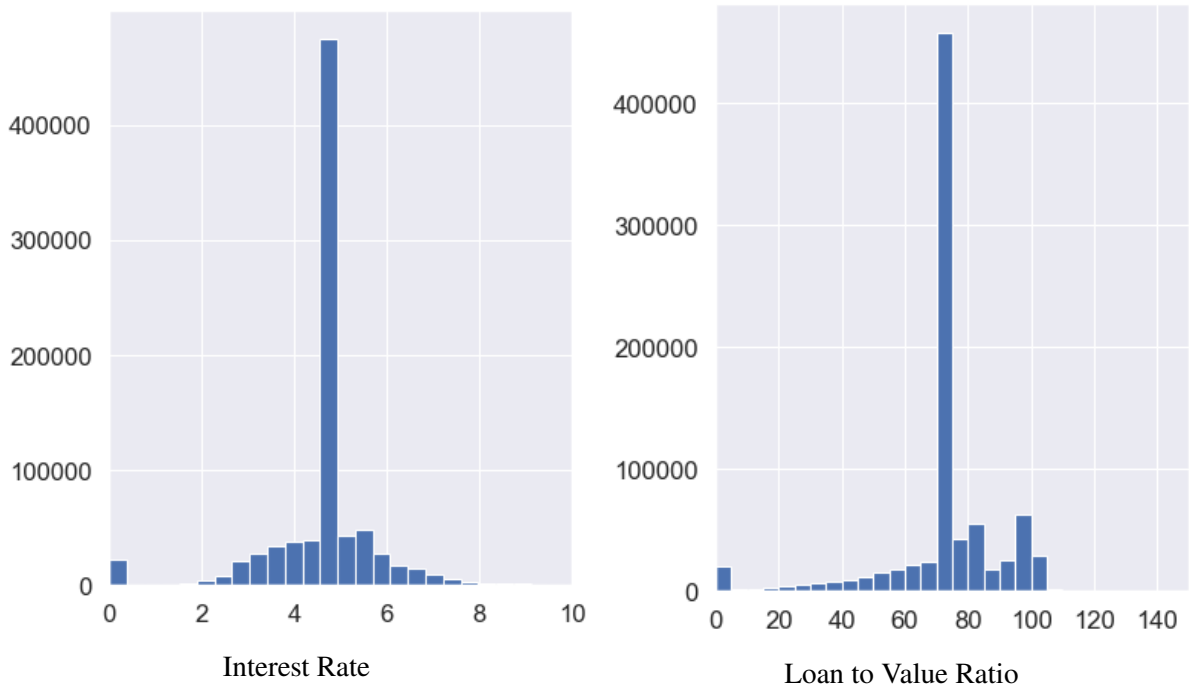


Figure 4.3: HMDA: Histograms of the Numerical Features - The results of the KNNImputer applying mean (*4.56%* for the interest rate and *71.24%* for the Loan to Value Ratio) values for the loan to value ratio values for all missing values show clearly here by the high amount of values at the mean.

in the **36%-41%** range. The distributions of the categorical variables can be seen in **figure 4.4**.

EDA of Fairness Aspects

Following the scope of this thesis specified in **chapter 1**, a special focus needs to be put on fairness, specifically equality with regards to the protected attribute(s). This potential unfairness in the underlying data can be identified from assessing the distribution of the target variable across different groups. **Figure 4.5** shows the amount of (not) granted loans per race and by sex, the probabilities of being granted a loan across these groups can be found in **table 4.2**.

Even though the focus of the analysis is on the *applicant_race-1* attribute, *applicant_sex* has been included as a second discriminating factor, as it also constitutes a protected attribute. Inspection of the results depicted here did however imply that the issue of racial equality is more pronounced than that of inequality between the sexes. A chi-squared test of independence proved that assumption of underlying inequality between races in the data, as the p-value is <0.01 and therefore H0 (equality in granted loans) could be rejected at any significance level. Using the aforementioned **AIF360** package to assess the mean difference of granted loans between the races in the underlying data amounted to a *14.9%* difference.

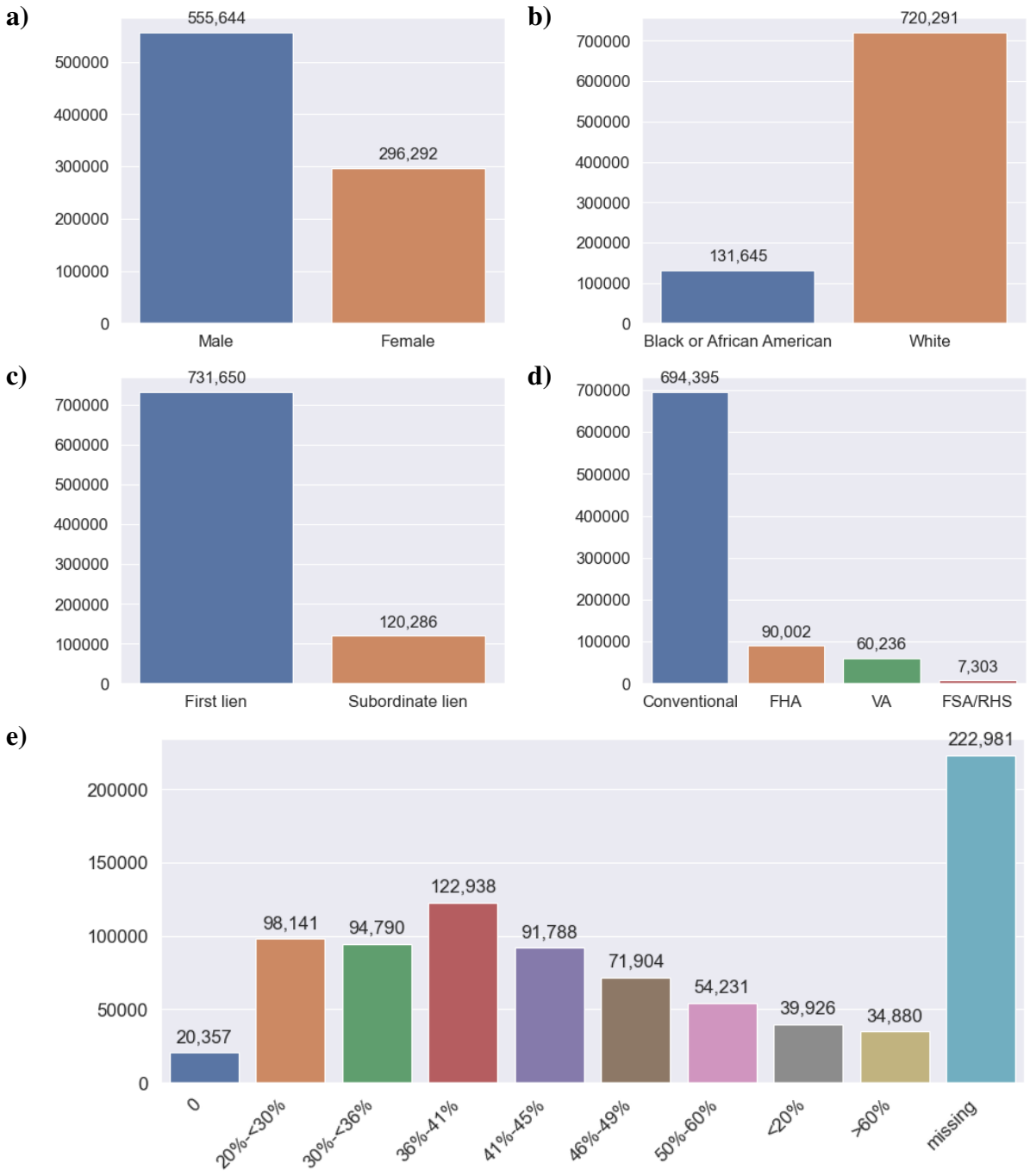


Figure 4.4: HMDA: Distributions of the Categorical Features - The majority of loan applicants are *Male* and *White* (see **a)** and **b)**). Most loans applied for are *Conventional* and *First Lien* (see **c)** and **d)**). The *debt_to_income_ratio* feature is (apart from the missing values) roughly normally distributed, with the mode being 14% of values in the **36%-41%** range (see **e)**.

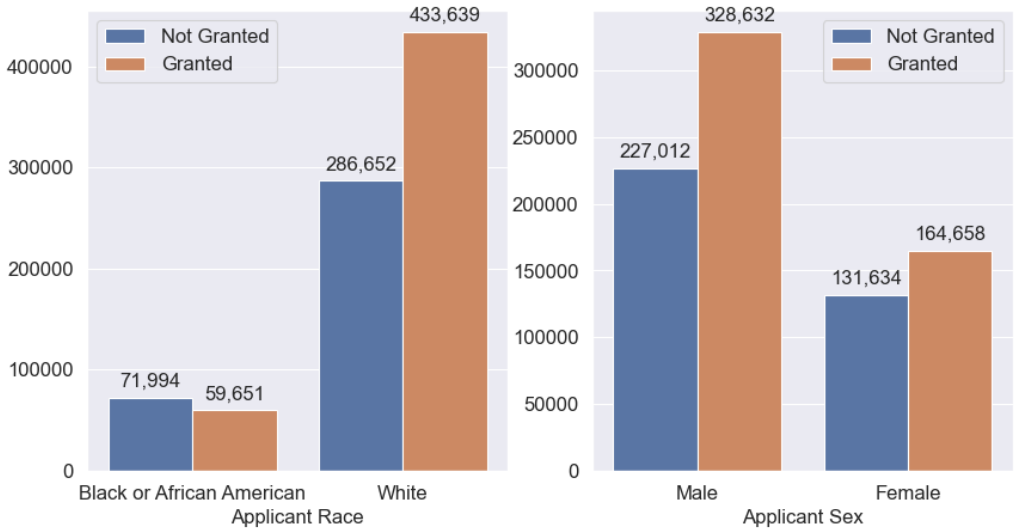


Figure 4.5: Loan Grant by Protected Attribute - Discrimination by race is apparent in the data, as the amount of granted loans differs significantly.

Applicant Race	**Applicant Sex**	**Loan Granted (%)**
Black or African American	Male	46.4
	Female	44.1
White	Male	60.9
	Female	58.8

Table 4.2: Loan Granting Statistics by Applicant Race and Sex - Analyzing the differences in percentages of loans being granted among race and sex of applicants shows differences in the underlying data: Regardless of their gender, Black or African American applicants are less likely to be granted a loan than White applicants.

4.1.2 Enrichment Data

The cleaned enrichment data (see **chapter 4.1.2**) are *complete* in a way that there are values available for all features and for all of the 469 counties. Basic summary statistics can be found in **table 4.3**. Across all 469 counties that are included in this analysis, a high school degree is the most common highest degree, with the mean of population percentages with that degree being **32.9%** overall, followed by a college degree with **30.9%**. A high discrepancy in the educational standard between the different observational units becomes apparent when considering the presence of counties where **81.6%** of the population have less than a high school degree, while in others, **55.2%** have a bachelors degree or higher. In terms of population, the counties in this analysis are heterogenous as well, with the mean population being **85.36 thousand** with standard deviation of **318.70 thousand** and the most inhabited county having nearly 100,000 times as many inhabitants as the least inhabited. The socio-economic factors expose a wide range of values as well, with the poverty rate ranging from **3.9%** to **43.5%**, the median household income from **25.65 thousand** to **124.35 thousand**, and the unemployment rate from **0.6%** to **11.0%**.

	Count	Mean	Std	Min	Max
College Degree	469.00	30.9%	6.03	0.00%	76.92%
Bachelor Degree or Higher	469.00	21.7%	8.42	0.00%	55.17%
High School Degree	469.00	32.9%	6.60	12.93%	51.17%
Less than High School Degree	469.00	14.5%	8.18	0.60%	81.55%
Population (thousands)	469.00	85.36	318.70	0.05	4,780.91
Poverty Rate	469.00	15.9%	6.23	3.90%	43.50%
Median Household Income (thousands)	469.00	56.6	13.94	25.65	124.35
Unemployment Rate	469.00	3.6%	1.21	0.60%	11.00%

Table 4.3: Summary Statistics of the Enrichment Data - The summary statistics show that the data is complete and has a wide range of values.

Expectedly, there is a high correlation between some of the features (see **figure 4.6**). While a bachelors degree or higher tends to be associated with a higher median household income (positive correlation of **0.62**), a high school degree or less than a high school degree as a highest degree is associated with a higher poverty rate (positive correlation of **0.33** and **0.54** respectively). A similar relation of socio-economic factors with education can be observed in the correlation of the unemployment rate with the highest degree, with a negative correlation of **-0.42** for holders of a bachelors degree or higher, compared to a positive correlation of **0.53** for less than a high school degree, implying that on average, observational units with less than a high school degree a more likely to be unemployed than those with a bachelors degree or

higher.

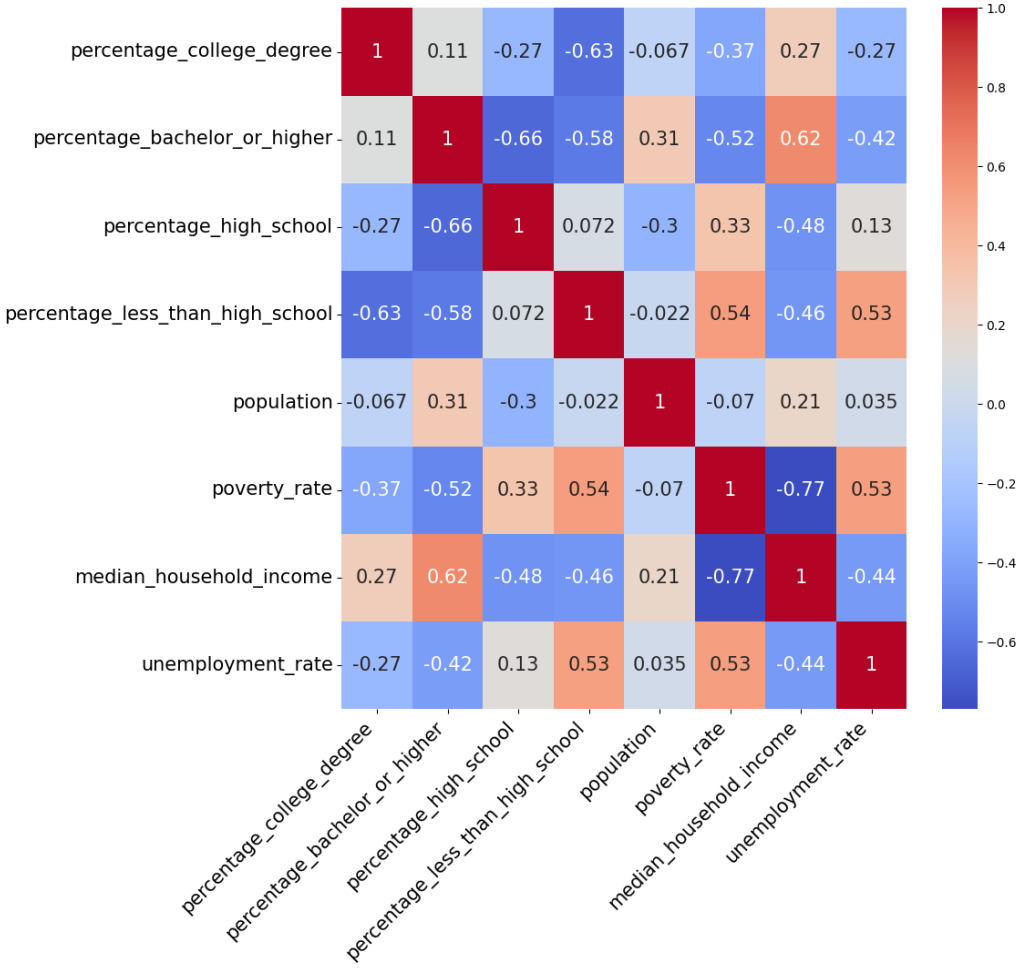


Figure 4.6: Correlation Between the Enrichment Features - Some of the enrichment features expectedly show a high correlation, an example being *median_household_income* and *poverty_rate*.

As stated before, analyzing the geographical information provided in the enrichment dataset (see **chapter 3.2.2**) was expected to provide additional insights into the fairness of the model. A sign of potential discrimination in the data could be the correlation between race and potentially favorable outcomes, such as a higher percentage of granted loans or a lower poverty rate. **Figure 4.7** shows a scatterplot that relates the percentage of White applicants to the percentage of granted loans per county. It indicates that a higher percentage of white applicants per county does not only seem to be correlated with a higher percentage of these loans actually being granted, but also with a lower poverty rate on average.

It should be noted that the distributions within the enrichment data themselves are skewed by nature, as there are way higher numbers of White applicants in the data set and only few counties have a substantial number of predicted mortgage grants (see **figure 4.8**).

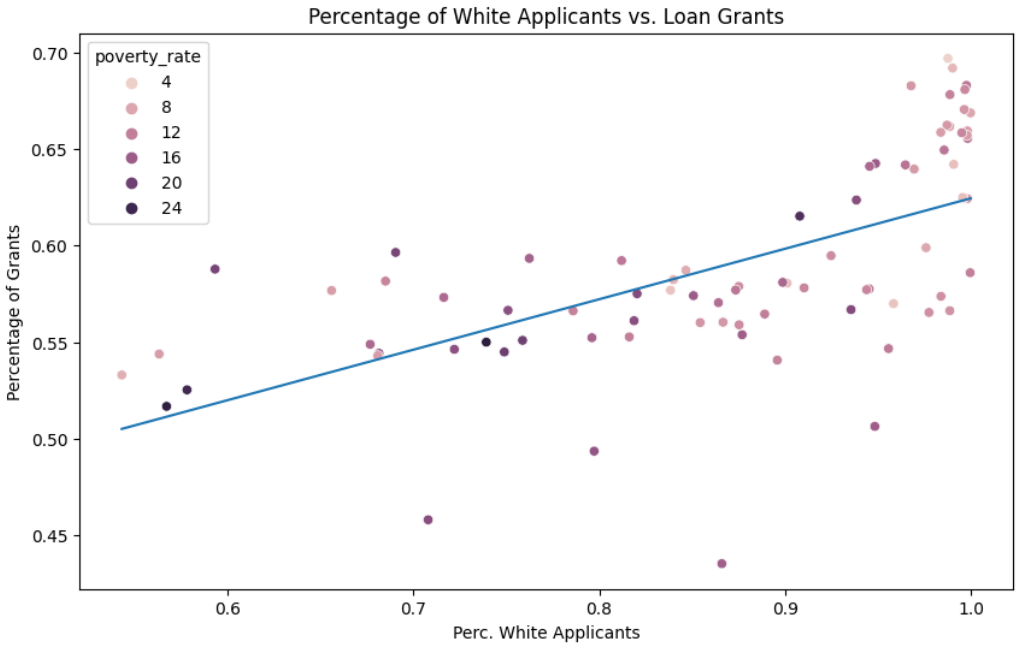


Figure 4.7: Relationship between Applicant Race, Poverty Rate and Loan Grants - Counties with predominantly White applicants do not only tend to have a lower poverty rate on average, but also see a higher percentage of loans being granted on average.

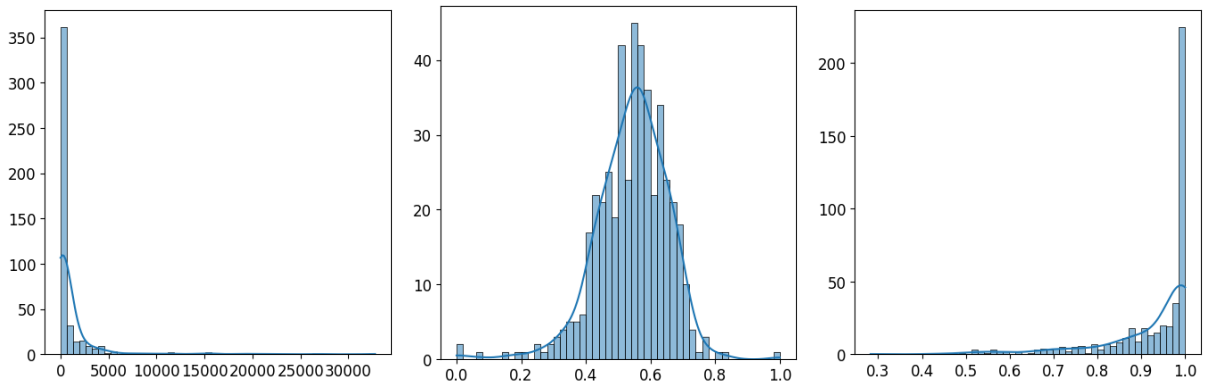


Figure 4.8: Enrichment Data EDA - Analyzing the enrichment data shows that while the percentage of predictions per county appears to be normally distributed, the distributions of the percentage of White applicants and the percentage of predicted loans per county are skewed.

4.2 Results

4.2.1 Mortgage Classifier (Benchmark)

In order to assess whether a predictive algorithm would pick up on and reproduce bias in the data, an initial classification model (as described in **chapter 3.2.1** and detailed in **table 3.4**) was trained on the HMDA dataset (see **chapter 3.1.1**) with the goal of predicting whether a mortgage would be granted or not for a given applicant. The results of this model were assessed in terms of performance and fairness.

Performance Assessment

When fitting the neural network to the training data, the *training accuracy* of the model improved rapidly initially, leveling off after a few epochs. The *validation accuracy* started at a high level and constantly improved by small increments, suggesting that both the model learning process as well as the ability to generalize to previously unseen data were successful. The training process was stopped by an early_stopping callback. The training results of the best epoch were:

- *Training Accuracy*: 0.90
- *Validation Accuracy*: 0.90
- *Training Loss*: 0.28
- *Validation Loss*: 0.28

The history of the training process is depicted in subplot a) of **figure 4.9**.

The model was then evaluated on the test dataset, which was not seen by the model during training. The results of the performance evaluation (i.e. *metrics #1*) are shown in **table 4.4**. The model achieved an *accuracy* of 0.90, a *precision* of 0.88, a *recall* of 0.97, and an *F1-score* of 0.92. As stated in **chapter 3.2.1**, the original model output were probabilities between 0 and 1. These could be used to calculate ROC AUC and plot the corresponding ROC curve, which can be seen in subplot b) of **figure 4.9**. The *ROC-AUC* score was 0.94, indicating a high level of model performance. Converting the probabilities into predictions with a threshold of 0.5 fulfilled the classification requirement. The *confusion matrix* is depicted in subplot c) of **figure 4.9**. The model managed to achieve a high number of true positives and true negatives, while the number of false negatives was low. However, the number of false positives was nearly 8% of all predictions.

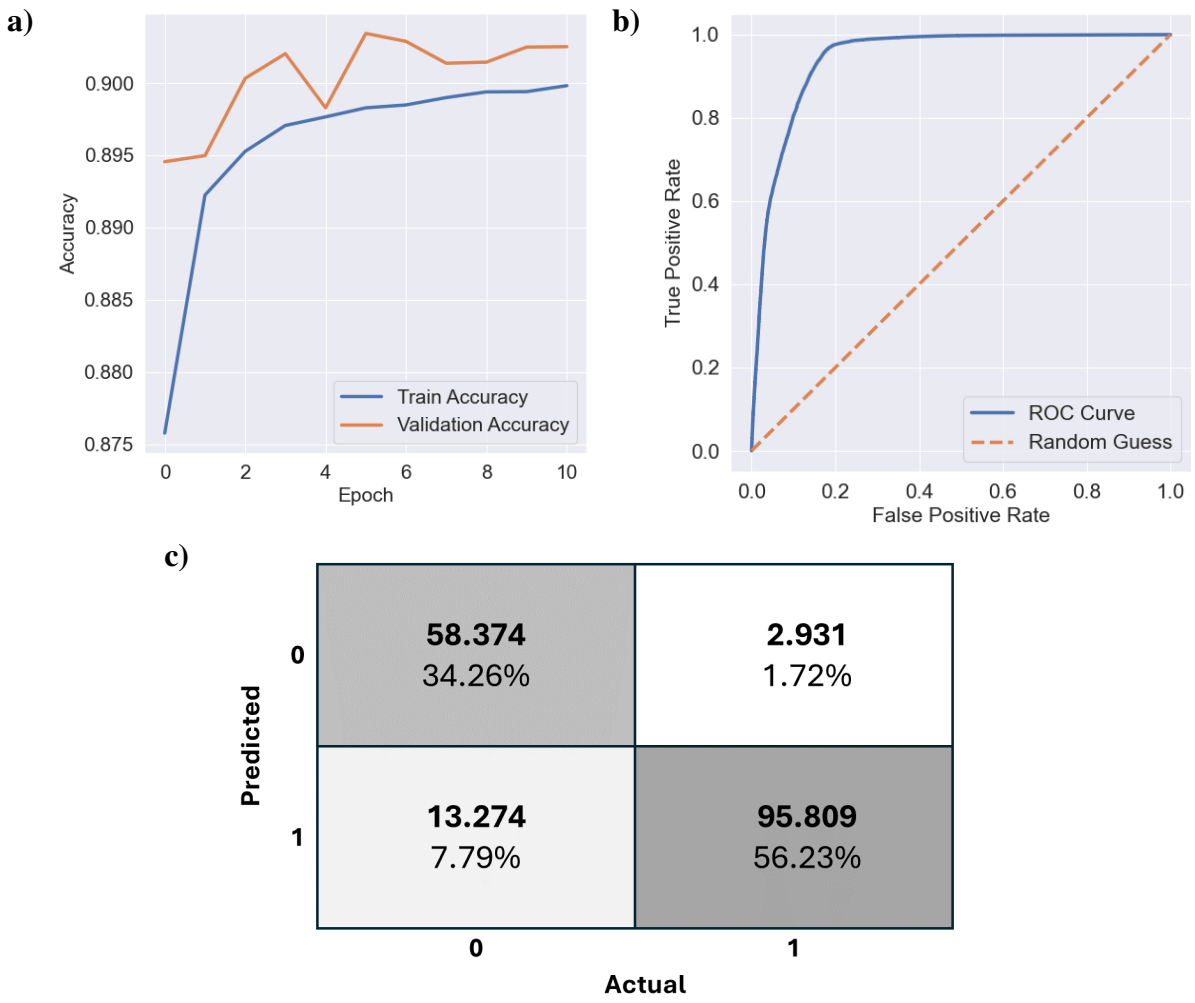


Figure 4.9: Training History, ROC curve, and Confusion Matrix of the Mortgage Classifier Model - **a)** The training history of the initial mortgage classifier model, showing the training and validation accuracy and loss over the course of the training process. The training accuracy improved constantly until the early_stopping callback. The validation accuracy constantly improved, suggesting a successful learning process. **b)** The ROC curve is significantly above the diagonal baseline, indicating high predictive performance. **c)** The confusion matrix of the mortgage classifier model on the test dataset. The model achieved a high number of true positives and true negatives. The number of false negatives was low, however, false positives made up nearly 8% of all predictions.

Metric	Value
accuracy	0.90
precision	0.88
recall	0.97
f1	0.92

Table 4.4: Metrics #1: Initial Model - The mortgage classifier model was evaluated on the test dataset, achieving an accuracy of 0.90, a precision of 0.88, a recall of 0.97, and an F1-score of 0.92.

Fairness Assessment

Following the research question (see **chapter 1**), the *detection of unfairness* in the predictions is an explicit goal of this thesis. To address this, the fairness assessment outlined in **chapter 3.2.1** was applied to the predictions. The results of the fairness assessment (i.e. *metrics #2*) are shown in **table 4.5**. The model performed slightly better for *White* applicants than for *Black* applicants. The *accuracy* for *White* applicants was 0.91, while it was 0.88 for *Black* applicants. The *precision* for *White* applicants was 0.89, while it was 0.82 for *Black* applicants. The *recall* for *White* applicants was 0.97, while it was 0.96 for *Black* applicants. The *F1-score* for *White* applicants was 0.93, while it was 0.88 for *Black* applicants. The *AUC* for *White* applicants was 0.94, while it was 0.95 for *Black* applicants. In terms of disparities (where the optimal value is **1**), the model performed comparably well in all disciplines except the *fnr_disparity*, which was 1.42. This indicates that the model was more likely to predict a false negative for *Black* applicants than for *White* applicants.

4.2.2 Explainability

As stated in **chapter 3.2.2**, three different approaches to explainability were utilized not only to support the analysis of fairness by providing insights into the model's decision-making process, but also to provide a better understanding of the model's behavior: *SHAP*, *LIME*, and a *Global Surrogate Model*.

SHAP

As stated in **chapter 2.2.2**, the SHAP algorithm tries to game-theoretically distribute the value of the final prediction among the individual features considered. **Figure 4.10** shows the SHAP beeswarm plot, which displays the distribution of the SHAP values for each feature in the dataset. The color indicates the feature value (red = higher; blue = lower), while the x-axis shows the SHAP value (left of center = negative; right of center = positive). The y-axis shows the feature

Metric	Value
Accuracy White	0.91
Precision White	0.89
Recall White	0.97
F1 Score White	0.93
AUC White	0.94
Accuracy Black	0.88
Precision Black	0.82
Recall Black	0.96
F1 Score Black	0.88
AUC Black	0.95
tpr_disparity	0.99
fpr_disparity	0.96
tnr_disparity	1.01
fnr_disparity	1.42

Table 4.5: Metrics #2: Initial Model - The benchmark model showed a slightly better performance for *White* applicants than for *Black* applicants. The disparities were comparably low, except for the *fnr_disparity*, which was 1.42.

name. The plot includes 150 values from the SHAP values of the test dataset. It showed that the most influential features according to SHAP were *debt_to_income_ratio_missing, interest_rate, loan_to_value_ratio*, and *debt_to_income_ratio_>60%*. While it was apparent how missingness in the debt to income ratio (missingness is negative, availability is positive) and values >60% in the debt to income ratio (higher is negative, lower is positive) affected the model decision, interest rates and the loan-to-value-ratios as the numerical variables were less intuitive to interpret. Although most medium to high interest rates seemed to be related with slight negative impacts, there also was a cluster of higher values for these variables corresponding with positive prediction influences. As all other variables were categorical, their interpretations were straightforward. Albeit their absolute impact in terms of SHAP values was limited, it is noteworthy that the *protected variables* of race and sex were in fact picked up on by the model. In every case where a decision was negatively influenced by the *race* of the applicant, the applicant in question was Black or African American (as can be inferred from all values left of the center of the x-axis for *applicant_race-1_White* being colored blue for this feature, meaning a lower value, which in turn means Black or African American ethnicity). A similar picture was observed for the *sex* of the applicant: In most cases where the model picked up on the sex being an influential factor on the model decision, the applicant in question was *female* and vice versa.

The *expected value* of the SHAP values (i.e. the baseline before consideration of any feature importance) was *0.57*. This corresponds to the imbalance in the original HMDA data (see **chapter 4.1**). **Figure 4.11** shows the SHAP values for four selected applicants. The force plots displayed show how the individual values of the features influence the model's decision according to SHAP. Each individual prediction results from the aforementioned *expected value* and the sum of inferred importances of the features (exemplarily, in the first plot displayed in **figure 4.11**, the feature importances amount to roughly positive 0.4, leading to a total value of 0.97 and therefore a positive prediction, i.e. a granted mortgage). It shows that SHAP attributes a high importance to (missingness of) the Debt to income ratio. Due to the values being scaled, a value of *-0.59* corresponds to *debt_to_income_ratio_missing == False* and a value of *1.68* corresponds to *debt_to_income_ratio_missing == True*. Therefore, SHAP considers missingness in this variable as negative and availability as positive. Considering the last applicant displayed (*Black or African American Male, Debt to Income Ratio available*), it does however show that even availabilty of the Debt to Income Ratio does not guarantee a positive model decision. According to SHAP, none of the decisions displayed here (and in the whole set of predictions in general) were significantly informed by any protected attribute. However, confounding factors might still be present, as the model might have learned to discriminate based on other features that are correlated with the protected attributes.

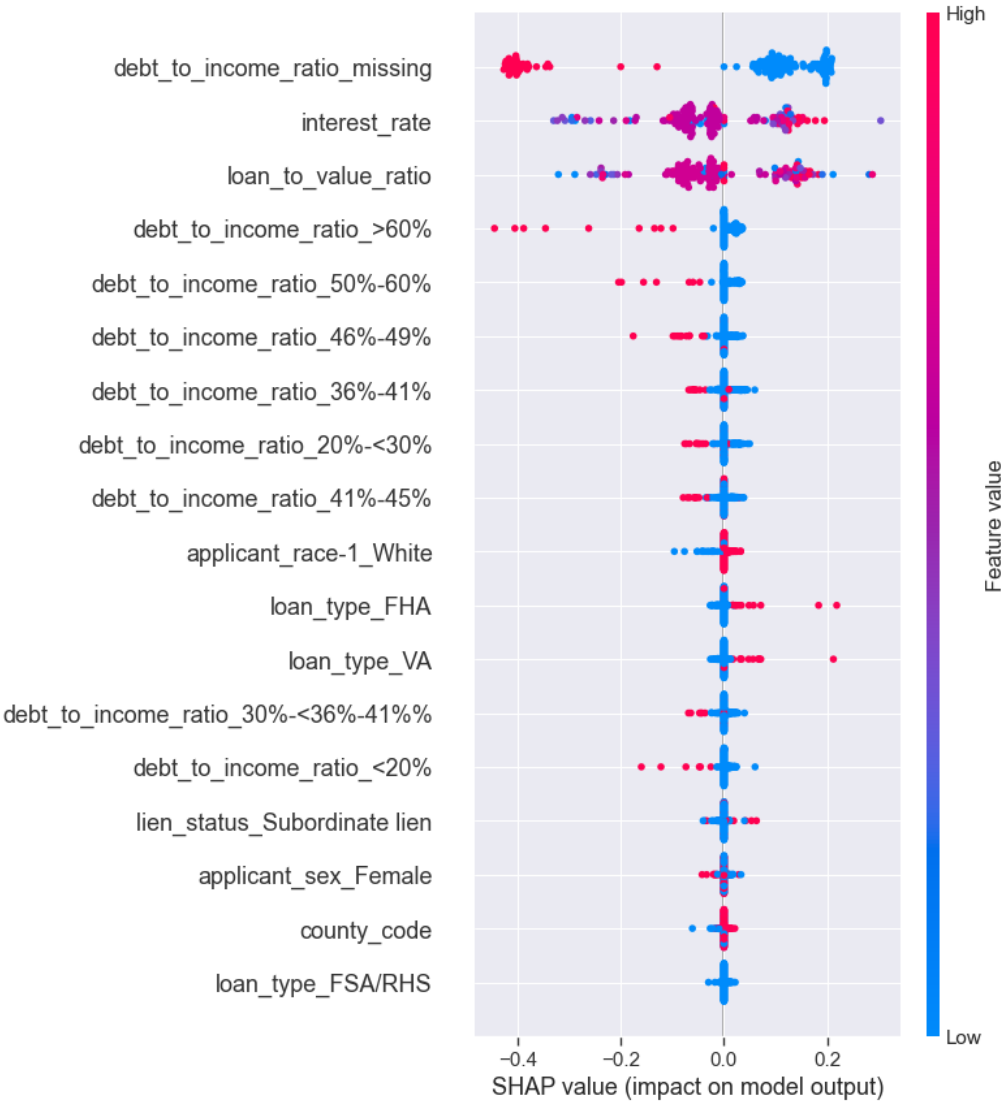
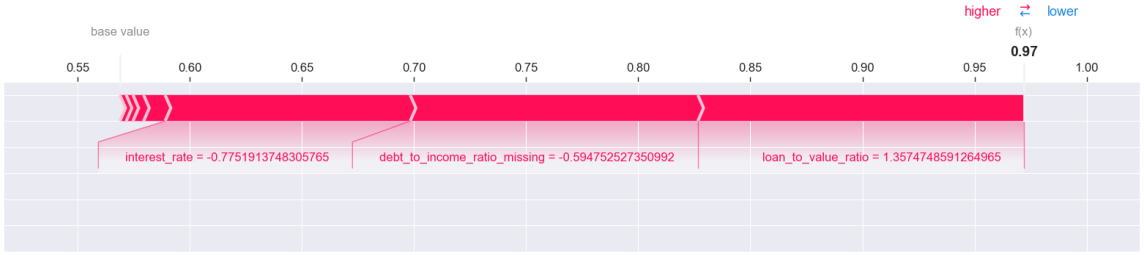
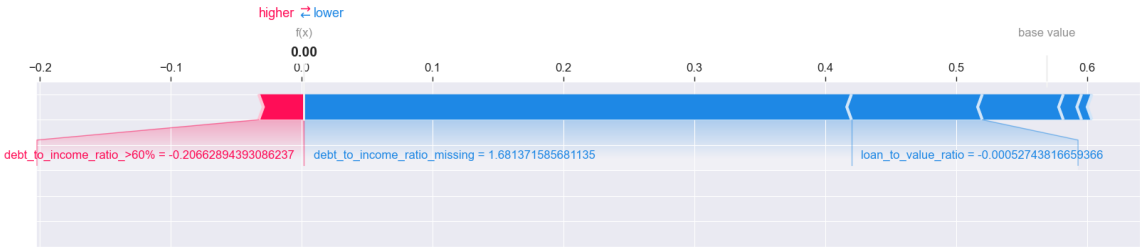


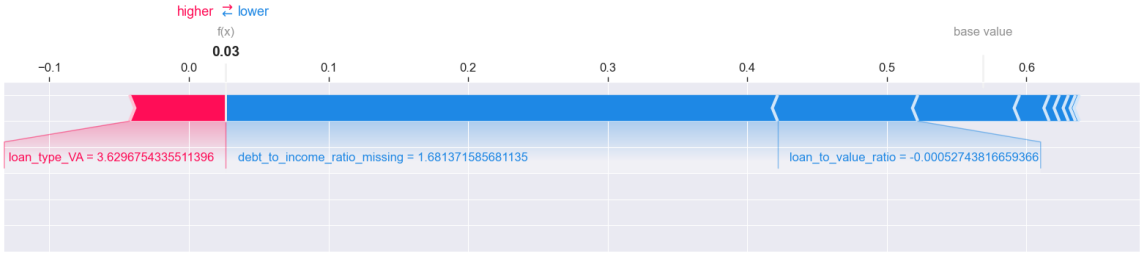
Figure 4.10: SHAP beeswarm plot - The SHAP beeswarm plot shows the distribution of the SHAP values for each feature in the dataset. The color indicates the feature value, while the x-axis shows the SHAP value. The y-axis shows the feature name.



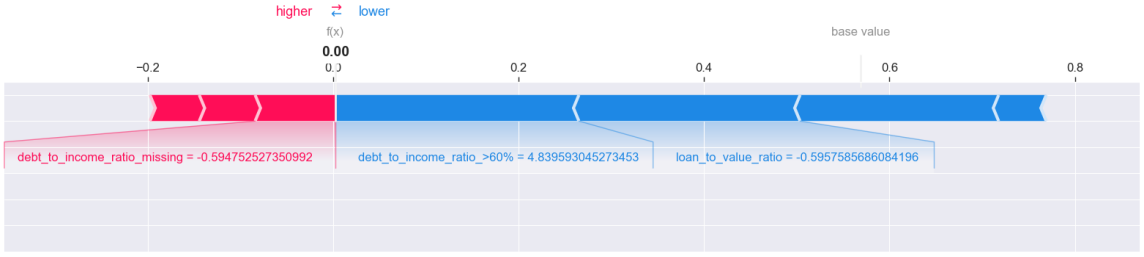
White Male, Debt to Income Ratio available



White Male, Debt to Income Ratio missing



Black or African American Male, Debt to Income Ratio missing



Black or African American Male, Debt to Income Ratio available

Figure 4.11: Selected SHAP Individual Analyses - Comparing four selected Male applicants with different characteristics shows that, in general, SHAP attributes a high importance to (missingness of) the Debt to Income Ratio. However, it is not the sole decision criterion, as the last applicant displayed shows.

LIME

The LIME algorithm, in contrast to SHAP, tries to explain the model's decision on a local level by approximating the model's behavior around a single prediction (see **chapter 2.2.2**). **Figure 4.12** shows the LIME individual feature importance plot for a selected applicant, specifically the same applicant that is denoted as *White Male, Debt to Income Ratio available* in **figure 4.11**. The x-axis shows the feature importance, while the y-axis shows the feature name. It showed that LIME attributes a high importance to the *loan type* and the *debt to income ratio*. Once again, the *protected attributes* had very little absolute influence on the model's decision according to LIME.

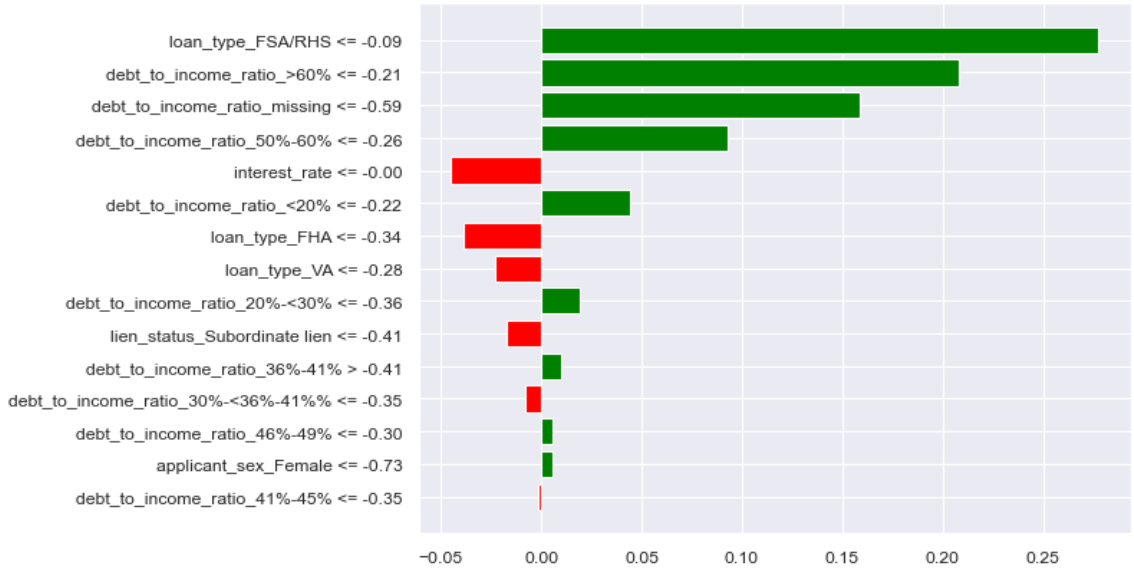


Figure 4.12: LIME Individual Feature Importance - The LIME individual feature importance plot shows the direction and the impact of the features on the model's decision for a selected applicant. The x-axis shows the feature importance, while the y-axis shows the feature name.

While the overall explanations on which features are influential are similar for both SHAP and LIME, the actual impact of the features varies significantly. While this is not a direct threat to the quality of the results of this thesis, it is a reminder that explainability algorithms need to be analyzed carefully. This ties with the findings of Krishna et al. (Krishna et al., 2022), who emphasize the importance of understanding the underlying assumptions of explainability algorithms and the need for a more comprehensive evaluation of their results.

Global Surrogate Model

To validate the results of the local explanations, a *Global Surrogate Model* was used. **figure 4.13** shows the results of the global surrogate model. Specifically, the five most important features according to the global surrogate model are compared to the SHAP and LIME explanations in terms of their relative performance. It showed that all three explanation algorithms mainly agree

on the three most important features in the data (*debt_to_income_ratio_missing, interest_rate*, and *debt_to_income_ratio_>60%*), although LIME attributes a different order of importance to them compared to SHAP and the global surrogate model.

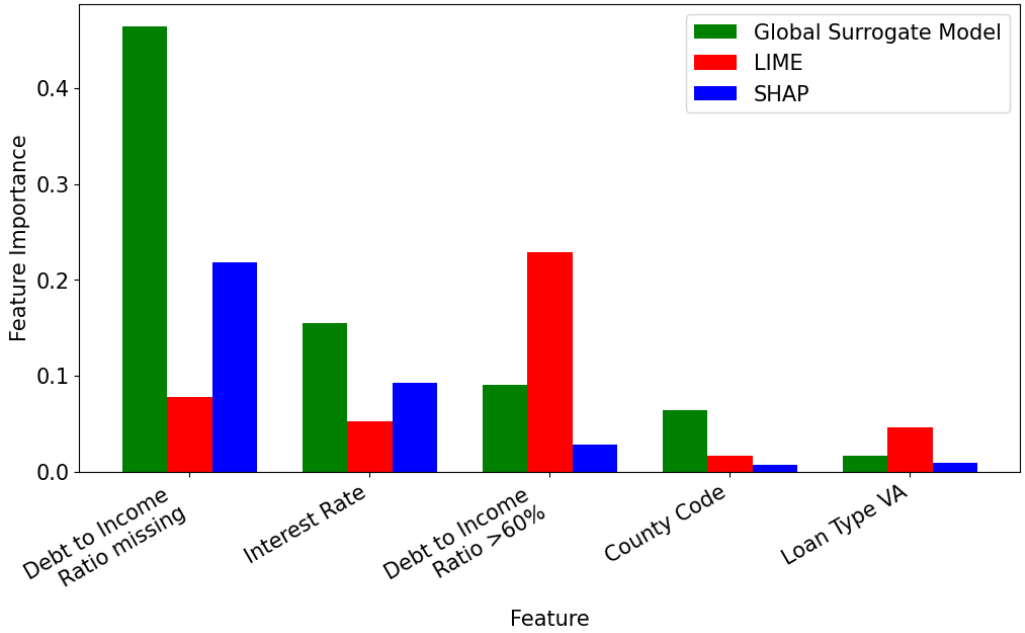


Figure 4.13: Global Surrogate Model compared to SHAP and LIME - Analyzing the 5 most important features according to the global surrogate model implies that the overall trends of SHAP and LIME are close to the global explanations. Although LIME attributes a higher importance to the *Debt to Income Ratio* being >60%, than to missingness, the three most important features are similar in all three models.

Summary

While the SHAP and LIME algorithms agreed on the most important features, the actual impact of these features varied significantly. The global surrogate model confirmed the results of the local explanations, showing that the most important features were the *debt to income ratio* and the *interest rate*. It is noteworthy that missingness in the *debt_to_income_ratio* cannot be considered to be at random as it is highly predictive of the model's decision. While individual values put out by SHAP and LIME might differ due to the different approaches, the overall trends of the explanations were similar, suggesting that the model's decision-making process could be explained with some level of confidence.

4.2.3 Fairness Adjustments

While the performance of the benchmark mortgage classifier detailed in **chapter 4.2.1** was satisfactory, the scope of this thesis (see **chapter 1**) included taking an iterative approach to improve fairness without sacrificing predictive performance, as outlined in **chapter 3.2.3**. To this end, the following fairness adjustments were applied to the model: *Reweighing*, *Correlation Remover* and *Calibrated Equalized Odds*. The aim was to reach improvement in at least one of the two metric sets, compared to the benchmark performance depicted in **table 4.4** (*metrics #1*) and **table 4.5** (*metrics #2*).

Reweighing

Using the *reweighing* pre-processing algorithm to assign specific weights to the training data without actually adjusting the underlying data (for details, refer to **chapter 3.2.3**), yielded positive results in terms of model performance. While the predictive performance of the model (as measured by *metrics #1*) could be kept on the same level that was reached by the benchmark model (see **table 4.6**), the fairness in terms of the disparities in predictions between *White* and *Black or African American* applicants could be reduced in comparison (see **table 4.7**)

Metric	Value
accuracy	0.90
precision	0.88
recall	0.97
f1	0.92

Table 4.6: Metrics #1: Reweighing - The performance of the *reweighed* model was on par with the benchmark (compare **table 4.4**).

Correlation Remover

As described in **chapter 3.2.3**, the *correlation remover* algorithm aims to minimize any correlation of selected protected attributes with any other variables within the data. While the performance of the model (see **table 4.8**) with the accordingly data pre-processed was comparable to or in parts slightly better than the benchmark model (compare **table 4.4**), it did not perform well in terms of fairness (see **table 4.9**). Subgroup performance could not be optimized compared to the benchmark and the disparities were higher, indicating less equality in the model predictions.

Metric	Value
Accuracy White	0.91
Precision White	0.89
Recall White	0.97
F1 Score White	0.93
AUC White	0.94
Accuracy Black	0.88
Precision Black	0.81
Recall Black	0.97
F1 Score Black	0.88
AUC Black	0.95
tpr_disparity	0.99
fpr_disparity	1.00
tnr_disparity	1.00
fnr_disparity	1.20

Table 4.7: Metrics #2: Reweighing - While the subgroup performances of the *reweighed* model were comparable to those of the benchmark model (compare **table 4.5**), the disparities were nearly nullified with the exception of the *fnr disparity*, which still is on a good level with 1.20.

Metric	Value
accuracy	0.91
precision	0.88
recall	0.97
f1	0.92

Table 4.8: Metrics #1: Correlation Remover - The performance of the *correlation remover* model was very similar to the benchmark (compare **table 4.4**).

Calibrated Equalized Odds

The only postprocessing algorithm within the fairness adjustments, the *calibrated equalized odds* algorithm (see **chapter 3.2.3** for details), aims to adjust the model's predictions to ensure equalized odds between the protected attributes while keeping the results calibrated. **Tables 4.10** and **4.11** show that the performance of this algorithm is not optimal within the context of this thesis. Both *Metrics #1* and *Metrics #2* indicate a lower performance of this algorithm compared to the benchmark (compare **table 4.4**).

Metric	Value
Accuracy White	0.91
Precision White	0.89
Recall White	0.97
F1 Score White	0.93
AUC White	NA
Accuracy Black	0.89
Precision Black	0.84
Recall Black	0.93
F1 Score Black	0.88
AUC Black	NA
tpr_disparity	0.96
fpr_disparity	0.80
tnr_disparity	1.05
fnr_disparity	2.67

Table 4.9: Metrics #2: Correlation Remover - While the subgroup performances of the *correlation remover* model were comparable to those of the benchmark model (compare **table 4.5**), the disparities had higher absolute distances from 1, indicating a lower level of fairness.

Metric	Value
accuracy	0.73
precision	0.69
recall	0.97
f1	0.81

Table 4.10: Metrics #1: Calibrated Equalized Odds - Compared to the benchmark model (see **table 4.5**), the model performance of the *calibrated equalized odds* is significantly worse. Besides the *recall*, which was kept on the same level, all performance metrics are well below those of the benchmark model.

Summary

Considering the overall *performance* of the different approaches, all models except the *calibrated equalized odds* algorithms performed on a similar, good level (see **table 4.12**). The *correlation removal* algorithm managed to slightly outperform the benchmark model in terms of *accuracy*, *precision*, and *f1-score*, while the *calibrated equalized odds* algorithm managed to slightly outperform the benchmark model in terms of *recall*.

Metric	Value
Accuracy White	0.72
Precision White	0.68
Recall White	0.98
F1 Score White	0.81
AUC White	NA
Accuracy Black	0.81
Precision Black	0.73
Recall Black	0.93
F1 Score Black	0.82
AUC Black	NA
tpr_disparity	0.95
fpr_disparity	0.43
tnr_disparity	2.23
fnr_disparity	3.25

Table 4.11: Metrics #2: Calibrated Equalized Odds - Similar to the total performance of the model (see **table 4.10**), the performance of the *calibrated equalized odds* model in terms of fairness was significantly worse than that of the benchmark model (see **table 4.5**) for both subgroups, as were the disparities.

	Initial Model	Reweighing	Calibrated Equalized Odds	Correlation Removal
accuracy	0.90	0.90	0.73	**0.91**
precision	0.88	0.88	0.69	**0.88**
recall	0.97	0.97	**0.97**	0.97
f1	0.92	0.92	0.81	**0.92**
roc_auc	**0.94**	0.94	NA	NA

Table 4.12: Metrics #1: Fairness Adjustments Summary - Results of *metrics #1* for all applied algorithms. The values are rounded, the highest scores are marked **bold**.

In terms of *fairness*, no significant optimizations could be achieved. The *reweighing* technique managed to improve the overall fairness of the model, but the other techniques did not manage to reach the same level of fairness. **Figure 4.14** shows that the differences in loan grants between *White* and *Black or African American* applicants was not substantially reduced by any of the iterations, with the calibrated equalized odds algorithm even increasing the difference.

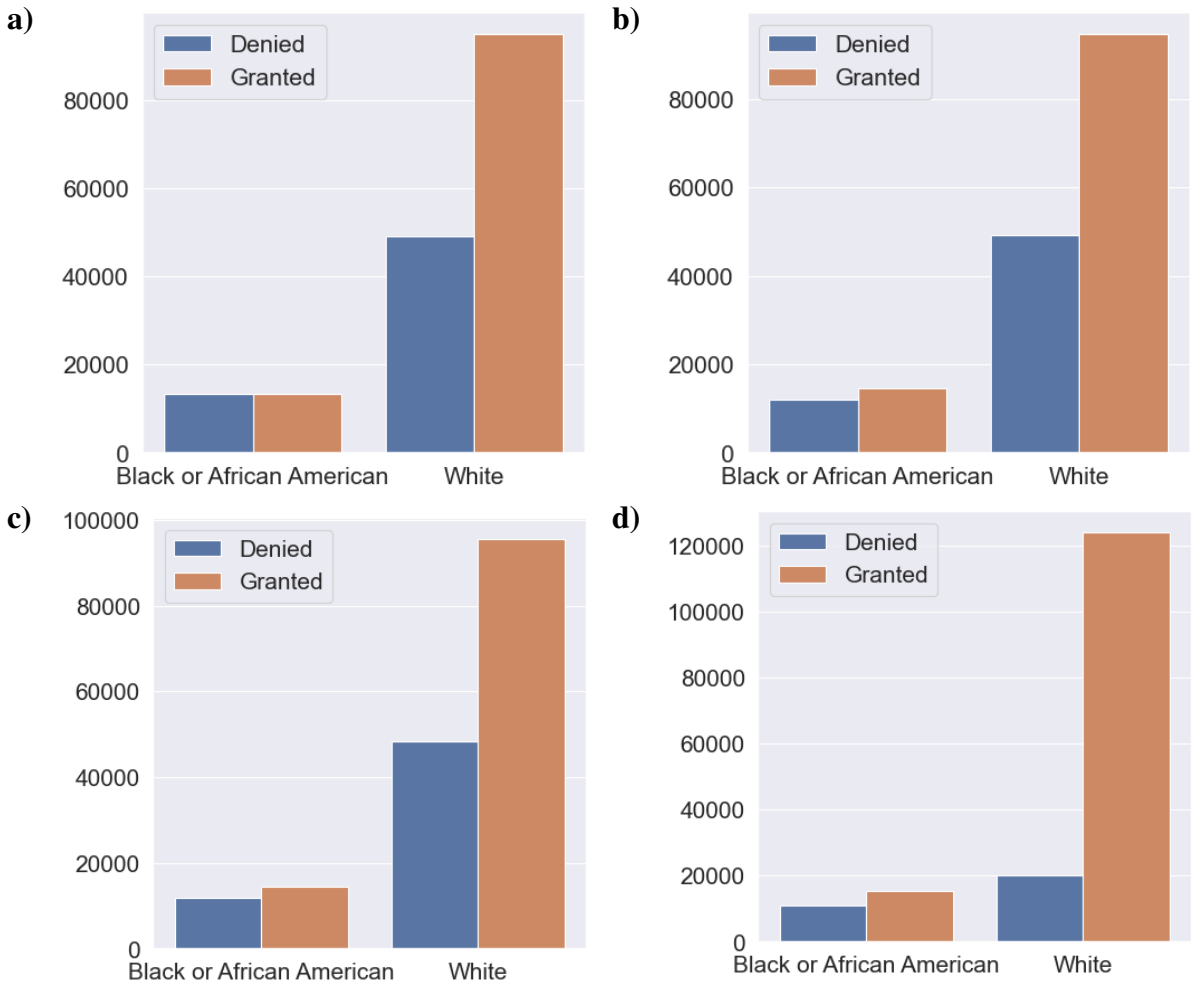


Figure 4.14: Differences in Positive Predictions per Model - Only the *Calibrated Equalized Odds* model **d)**, which exhibits a higher amount of overall predicted grants, changed the ratio of granted mortgages among the races significantly compared to the benchmark model **a)**. The *reweighted* model **b)** and the *correlation removal* model **c)** did not significantly change the ratio.

Figure 4.15 shows the results of the fairness adjustments in terms of the *disparities* of the model. The disparities were calculated for the *true positive rate*, the *false positive rate*, the *true negative rate*, and the *false negative rate*. The disparities were calculated for the *White* and *Black or African American* applicants. As they are relative terms, only the relation of the disparities to the other group is shown. The disparities were comparably low for the *true positive rate*, while all the other disparities show at least one outlier. The *reweighed* model was the only one that produces satisfactory results in terms of fairness as measured by disparities across all four KPIs, with the benchmark model coming in second.

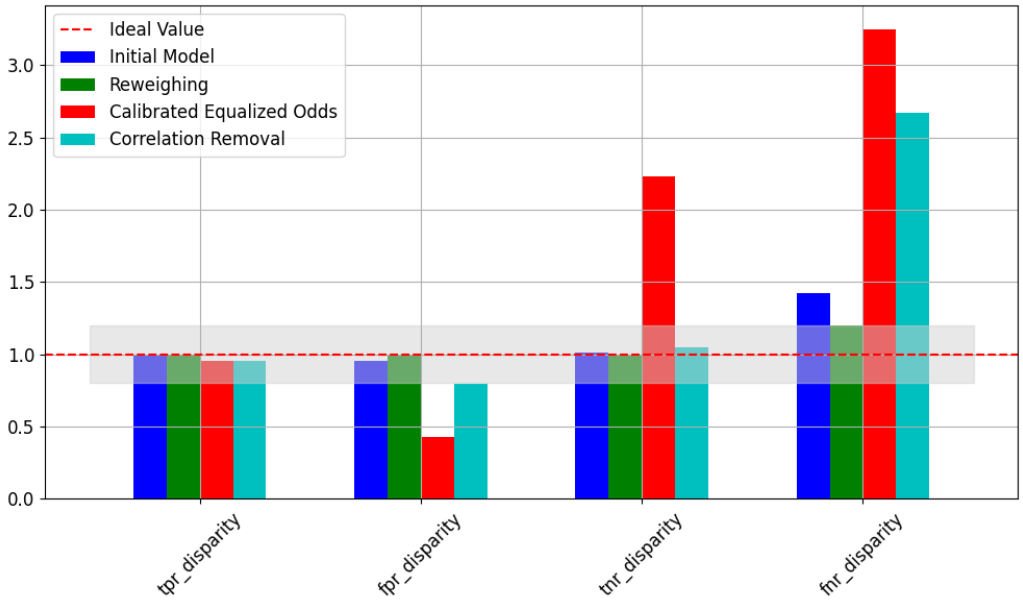


Figure 4.15: Fairness Adjustments Results - The results of the fairness adjustments in terms of the disparities of the model. The disparities were calculated for the *true positive rate*, the *false positive rate*, the *true negative rate*, and the *false negative rate*. They were calculated for the *White* and *Black or African American applicants*. Values closer to 1 are better, the gray area represents a good level of fairness.

Adding up all fairness metrics for all iterations (see **table 4.13**) resulted in a mixed picture. While the *reweighing* algorithm managed to improve the fairness of the outcomes in terms of disparities (as can also be inferred from **figure 4.15**), the predictive power of the model for subgroups did not show a single optimal model.

	Initial Model	Reweighing	Calibrated Equalized Odds	Correlation Removal
Accuracy White	0.91	0.91	0.72	**0.91**
Precision White	**0.89**	0.89	0.68	0.89
Recall White	0.97	0.97	**0.98**	0.98
F1 Score White	0.93	0.93	0.81	**0.93**
AUC White	**0.94**	0.94	NA	NA
Accuracy Black	0.88	0.88	0.81	**0.89**
Precision Black	0.82	0.81	0.73	**0.84**
Recall Black	0.96	**0.97**	0.93	0.93
F1 Score Black	0.88	**0.88**	0.82	0.88
AUC Black	**0.95**	0.95	NA	NA
tpr_disparity	0.99	**0.99**	0.95	0.96
fpr_disparity	0.96	**1.00**	0.43	0.80
tnr_disparity	1.01	**1.00**	2.23	1.05
fnr_disparity	1.42	**1.20**	3.25	2.67

Table 4.13: Metrics #2: Fairness Adjustments Summary - Results of *metrics #2* for all applied algorithms. The values are rounded, the highest scores (respectively those closest to 1 for the disparity calculations) are marked **bold**.

5 Conclusion

5.1 Discussion, Interpretation, and Limitations

The **research question** of this thesis included investigating the fairness of a machine learning model in mortgage lending and exploring the potential of explainability algorithms to detect and mitigate unfairness in the models:

"Can underlying unfairness in mortgage decision-making be detected, explained, and iteratively mitigated without sacrificing predictive performance?"

The findings of this thesis suggest that the underlying unfairness in mortgage decision-making can be detected and partly explained using explainability algorithms. The analyses conducted supported the theory that the mortgage decisions according to the 2022 HMDA dataset favored **White** applicants (who also accounted for the majority of applications), leading to a neural network model trained on it being biased towards favoring **White** applicants over **Black or African American** applicants. However, the iterative mitigation of this unfairness proved hard to achieve with the chosen algorithms. The fairness adjustments conducted in **chapter 3.2.3** partly yielded improvements in fairness, but the models still favored **White** applicants over **Black or African American** applicants, even when the protected attribute itself was controlled for. This assumption of a "deeper" inequality present in the data, resulting from **Black or African American** applicants historically being associated with worse attributes in terms of worthiness for a mortgage was supported by the analysis of the geographical enrichment data in **chapter 4.1.2**, showing that the regions with a higher proportion of **Black or African American** applicants were correlated with lower chances of having a mortgage granted.

When comparing the findings of this thesis to the research conducted by Martinez and Kirchner (Martinez and Kirchner, 2021), one common factor becomes apparent. Even though the differences in granted mortgages do not vary as extreme between the races analyzed (Martinez and Kircher report **Black or African American** applicants to have a 40% - 80% higher risk of having their applications denied), the factors influencing the decision-making process are similar. Just as the findings of the explainability algorithms displayed in **chapter 3.2.2** suggest, the *debt to income ratio* appeared to be a major factor in the decision-making process, with **Black or African American** applicants having a higher risk of being denied a mortgage due to this factor. While a direct comparison of the results is not possible due to the different datasets used and the different filters applied, the similarities in the factors influencing the decision-making process suggest that the underlying unfairness in mortgage decision-making is a systemic issue that is not easily mitigated.

Taking a historical perspective, the findings of this thesis are not surprising. The systemic discrimination against **Black or African American** applicants in mortgage lending has been well-documented in the past. Different studies on less recent versions of the HMDA data, such as those by Cherian (Cherian, 2014 for the years 1992 to 2003), the US Federal Reserve (Bhutta and Canner, 2013 for 2012), or Cyree and Winters (Cyree and Winters, 2023 for the years 2007 to 2016) have shown that **Black or African American** applicants are systematically discriminated against in mortgage lending. The findings of this thesis suggest that this discrimination is still present in the data used for the analysis, even though the data was collected in 2022, which supports the assumption that a big part of that discrimination is not "direct" but rather results from skin color being related to other decision factors (see **chapter 4.1.2**). It can be assumed that a more granular control for these factors might have provided a more detailed explanation of the unfairness present in the data. This is where studies like the one by Delis and Papadopoulos come in. They suggest that, even though discrimination cannot be ruled out for mortgage lending decisions, a tighter control of the omitted variables might show that race alone is not the deciding factor in the decision-making process (Delis and Papadopoulos, 2019).

Both the performance (see **table 4.4**) and the subgroup performance (see **table 4.5**) of the initial model surpass the performance of the algorithms promoted by Ghoba and Colaner (Ghoba and Colaner, 2021), which were used as a basis for the variable selection in this thesis. However, direct comparisons are impaired by the fact that no detailed methodology was provided in their paper. Additionally, no feature importance has been assessed in this study, so a direct comparison of the results is not possible. Apart from the aforementioned study by Martinez and Kirchner, few papers report on feature importance in comparable setups. One of them is a pre-print by Alves et al., that suggests high feature importance of *derived_loan_product_type, intro_rate_period* (both of which have not been included in this thesis), as well as *race* and *gender* of the applicants, which have been included in this thesis but did not show high feature importance (Alves et al., 2020). As stated before, it cannot be ruled out that this is the case because of omitted variables impacting the feature values of the included variables instead of reflecting actual feature importance.

The results of this thesis tie in with some of the findings of the study by Zou and Khern-am-Nui (Zou and Khern-am-nuai, 2023), who conducted a similar analysis on the 2019 HMDA dataset. They did not only come to the conclusion that the HMDA data do contain discrimination against **Black or African American** applicants, but also that the discrimination is not only present in the data but also amplified in the models trained on it. The findings of this thesis suggest that the same is true for the 2022 HMDA dataset. However, Zou and Khern-am-Nui found a more effective way to mitigate the unfairness in the data by using an *exponentiated gradient* algorithm, which was not applied in this thesis. This might be a promising approach for future research to improve the fairness of the models applied.

Limitations

Limitations of the research conducted in this thesis can be found in the **data** used, the **models** applied, and the **fairness adjustments** conducted.

Although the **data** selection process was thorough and the data was cleaned and preprocessed following strict criteria, only focusing the analysis on five American states naturally limits the generalizability of the results for other geographies within America and internationally. In terms of the features used it must be noted that with missingness in the *debt_to_income_ratio* appearing to be the most influential factor in the decision-making process, alternative approaches to handling missing data might have yielded different results. Using more of the 99 features available in the HMDA dataset was not feasible from the standpoint of computational efficiency in this thesis but might have also yielded different results, as the assumed unfairness in the data might be more deeply rooted in other features not analyzed in this thesis. Neither for the HMDA data, nor for the geographical enrichment data, causation can be assumed from the results presented in this thesis. The correlations found in the data might be due to other factors not analyzed in this thesis.

While the **model** choice of the neural network detailed in **table 3.4** can be considered appropriate because of the widespread application of such models in practice, its generally good performance and it being a typical example of a *black-box* algorithm, other models might have yielded different results. Specifically, it has not been tested whether a less complex might have been able to achieve comparable results without the trade-off in inherent explainability.

The **fairness adjustments** conducted in this thesis were limited to the three iterations detailed in **chapter 3.2.3**. While these adjustments were chosen based on their theoretical soundness, other fairness adjustments might have been appropriate for this task. For example, no *in-processing* algorithm was applied to keep the model used identical and therefore comparable across the iterations. This might have contributed to none of the fairness adjustments being able to fully mitigate the unfairness present in the data. For the *calibrated equalized odds* model, it must be noted that the authors themselves showed that enforcing calibration and equalized odds at the same time might impact performance in certain settings (Pleiss et al., 2017), which can be assumed to have been the case in this thesis as well, given the comparably bad results of this adjustment.

Revisiting the research questions of this thesis (see **chapter 1**), it must be concluded that the initial aim of this work could only be achieved in parts. While the exploratory data analysis, coupled with the analysis of the geographical enrichment data was useful to *detect* and partly *explain* unfairness within the data, the *mitigation* of this unfairness proved hard to achieve.

5.2 Recommendations and Conclusion

Recommendations and Outlook

These results yield important implications for the practical implementation of machine learning models, both in mortgage lending and generally. Not only must it be taken into account that automated decision-making in critical areas such as mortgage lending might learn from and amplify existing inequalities, but also that these inequalities might not be easily uncovered, even when attempting to explain the decision-making process of the models with common algorithmic techniques. This implies that before the widespread application of machine learning models in mortgage lending (or, in fact, any are where the decision-making might have critical impact on individuals), a thorough analysis of the data used and the models applied must be conducted to ensure that the models do not amplify existing inequalities.

At the same time, the development of Artificial Intelligence progresses in a staggering and still increasing pace. By the time of writing this thesis, several multimodal Large Language Models (LLMs) are available to the broad public (examples being GPT-4o by OpenAI[1], Google Gemini[2], or the Microsoft Copilot[3]), while other promising model architectures like the Extended Long Short-Term Memory (xLSTM) are already challenging the transformer architecture (Beck et al., 2024). These models are not only more powerful in terms of predictive performance but also in terms of their complexity and therefore their *black-box* nature. In times where such AI models are becoming an integral part of everyone's daily life to an extent that companies like OpenAI are already planning for the introduction of Artificial General Intelligence (AGI) (Altman, 2023), while Google assumes that users will form 'deep relationships' with AI products (Harris, 2024), it becomes very apparent that the need for explainability and fairness in AI models is dire due to their sheer impact. It must be assumed that Artificial Intelligence will be applied to even more sensitive areas in the future (examples being healthcare, military, or the justice system). Consequences of unfairness in the models can be even more severe than in mortgage lending in these cases. While some researchers are already tackling this issue by integrating explainability aspects into their work (a recent practical example being the CIMPLE project to tackle information manipulation[4]), additional research and practical application is required not only in the field of explainability but specifically also in its intersection with fairness considerations.

This is where this thesis aims to contribute to the field. By incorporating both explainability and fairness considerations into the model evaluation process, it aims not only to raise awareness that

[1] https://openai.com/index/hello-gpt-4o/
[2] https://deepmind.google/technologies/gemini/
[3] https://copilot.microsoft.com/
[4] https://www.chistera.eu/projects/cimple

both concepts are important both in their own right but also in their intersection, while trying to provide a practical example of how these concepts can be applied in a real-world setting. While the methodology of this thesis was focused specifically on unfairness in mortgage lending and does not provide a one-size-fits-all solution, the approach might be generalized to other areas after considering how to specifically adjust it to the given problem:

- Is the problem one of binary classification or of a different nature?
- Are there subgroups that can be expected to be treated unfairly?
- Must the explainability algorithms be adjusted to the nature of the available data?
- Are the used fairness metrics feasible for the given problem?

For the data in this thesis, the procedure was feasible, because the data was binary, the subgroups were clearly defined and contained one (**Black or African American**) that could be expected to be systematically discriminated against, the explainability algorithms were applicable to the data, and the fairness metrics were chosen specifically for subgroup fairness. However, as already stated in the **limitations** considerations (see **section 5.1**), the attempted fairness adjustments might not have been the most appropriate for the given problem.

While this thesis used a combination of different approaches to the topic of fairness, one way to truly improve the fairness of the models might be to include an even broader scope in terms of research focuses. As Kim et al. (S. D. Kim et al., 2023) suggest in their review of fairness in credit, true multidisciplinary research is needed to tackle the issue of fairness in credit scoring. But even within the narrower scope of this thesis, further research might be conducted to improve the fairness of the models applied. For example, the application of different fairness adjustments, the use of different models, or the inclusion of more features in the analysis might yield different results. Other opportunities for future research close to the scope of this thesis might include the application of a different geographical scope, different target variables (e.g. the interest rate of the mortgage), or a stronger focus on causal inference to uncover the underlying reasons for the unfairness present in the data. In a broader scope, a generalization of a methodology similar to the one applied in this thesis to other areas of machine learning might be a promising approach to improve the fairness of models in general.

Concluding Summary

By combining the concepts of explainability and fairness in the context of mortgage lending, this thesis aimed to provide a practical example of how these concepts and specifically their combination can add value to the evaluation of machine learning models. The combination of three different explainability algorithms enriched with a geographical perspective provided insights into the decision-making process of the initially used neural network, while a set of performance- and fairness-focused metrics helped to evaluate the model from different perspectives. The iterative fairness adjustments conducted in the final chapter of this thesis aimed to mitigate the unfairness present in the data, but only partly succeeded in doing so, leaving room for further research in this area.

Bibliography

Al Shiam, Sarder Abdulla et al. (2024). “Credit Risk Prediction Using Explainable AI”. In: *Journal of Business and Management Studies* 2.6, pp. 61–66.

Altman, Sam (02/2023). “Planning for AGI and beyond”. In: *OpenAI*.

Alves, Guilherme et al. (12/2020). “FixOut: an ensemble approach to fairer models”. working paper or preprint.

Anderson, A.A. (2023). “Testing machine learning explanation methods”. In: *Neural Computing and Applications* 35, pp. 18073–18084.

Barocas, Solon and Andrew D. Selbst (2016). “Big Data’s Disparate Impact”. In: *California Law Review* 104.3, pp. 671–732.

Barredo Arrieta, Alejandro et al. (06/2020). “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58, pp. 82–115.

Beck, Maximilian et al. (2024). *xLSTM: Extended Long Short-Term Memory*.

Bellamy, Rachel et al. (09/2019). “AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias”. In: *IBM Journal of Research and Development* PP, pp. 1–1.

Bhutta, Neil and Glenn Canner (01/2013). “Mortgage Market Conditions and Borrower Outcomes: Evidence from the 2012 HMDA Data and Matched HMDA–Credit Record Data”. In: *Federal Reserve Bulletin* 99.

Bhutta, Neil, Aurel Hizmo, and Daniel Ringo (10/2022). “How Much Does Racial Bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions”. en. In.

Bogen, Miranda, Aaron Rieke, and Shazeda Ahmed (2020). “Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, pp. 492–500.

Burrell, Jenna (06/2016). “How the machine ‘thinks’: Understanding opacity in machine learning algorithms”. In: *Big Data & Society* 3.1.

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy (12/2009). “Building Classifiers with Independency Constraints”. In: *2009 IEEE International Conference on Data Mining Workshops*. ISSN: 2375-9259, pp. 13–18.

Calmon, Flavio et al. (2017). “Optimized Pre-Processing for Discrimination Prevention”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.

Caruana, Rich et al. (08/2015). “Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’15. New York, NY, USA: Association for Computing Machinery, pp. 1721–1730.

Chang, Xinyu (09/2023). “Gender Bias in Hiring: An Analysis of the Impact of Amazon’s Recruiting Algorithm”. In: *Advances in Economics, Management and Political Sciences* 23, pp. 134–140.

Chen, Jiahao et al. (2019). “Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, pp. 339–348.

Cherian, Madhavi (2014). “Race in the Mortgage Market: An Empirical Investigation Using HMDA Data”. In: *Race, Gender & Class* 21.1/2, pp. 48–63.

Choraś, Michał et al. (2020). “Machine Learning – The Results Are Not the only Thing that Matters! What About Security, Explainability and Fairness?” In: *Computational Science – ICCS 2020*. Ed. by Valeria V. Krzhizhanovskaya et al. Cham: Springer International Publishing, pp. 615–628.

Chouldechova, Alexandra and Aaron Roth (10/2018). *The Frontiers of Fairness in Machine Learning*. Tech. rep. arXiv:1810.08810 [cs, stat] type: article. arXiv.

Cooper, A. et al. (02/2024). “Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification”. In.

Corbett-Davies, Sam et al. (08/2023). *The Measure and Mismeasure of Fairness*. Tech. rep. 24. arXiv:1808.00023 [cs] type: article. arXiv, pp. 1–117.

Cyree, K.B. and D.B. Winters (2023). “Investigating bank lending discrimination in the US using CRA-rated banks’ HMDA loan data.” In: *Public Choice* 197, pp. 371–395.

d’Alessandro, Brian, Cathy O’Neil, and Tom LaGatta (06/2017). “Conscientious Classification: A Data Scientist’s Guide to Discrimination-Aware Classification”. In: *Big Data* 5, pp. 120–134.

Datta, Anupam et al. (07/2017). *Proxy Non-Discrimination in Data-Driven Systems*. Tech. rep. arXiv:1707.08120 [cs] type: article. arXiv.

Delis, M.D. and P. Papadopoulos (2019). "Mortgage Lending Discrimination Across the U.S.: New Methodology and New Evidence". In: *Journal of Financial Services Research* 56, pp. 341–368.

Doshi-Velez, Finale and Been Kim (03/2017). *Towards A Rigorous Science of Interpretable Machine Learning*. Tech. rep. arXiv:1702.08608 [cs, stat] type: article. arXiv.

Dwork, Cynthia et al. (01/2012). "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. New York, NY, USA: Association for Computing Machinery, pp. 214–226.

Faber, Jacob W. (04/2013). "Racial Dynamics of Subprime Mortgage Lending at the Peak". In: *Housing Policy Debate* 23.2, pp. 328–349.

Feldman, Michael et al. (2015). "Certifying and Removing Disparate Impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Sydney, NSW, Australia: Association for Computing Machinery, pp. 259–268.

Ghoba, Sama and Nathan Colaner (12/2021). "Counterfactual Fairness in Mortgage Lending via Matching and Randomization". In: *35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia*.

Guidotti, Riccardo et al. (08/2018). "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys* 51.5, 93:1–93:42.

Gunning, David and David W. Aha (2019). "DARPA's Explainable Artificial Intelligence Program". en. In: *AI Magazine* 40.2, pp. 44–58.

Hardt, Moritz, Eric Price, and Nathan Srebro (10/2016). *Equality of Opportunity in Supervised Learning*. Tech. rep. arXiv:1610.02413 [cs] type: article. arXiv.

Harris, Jamie (05/2024). "AI DO! People will get into 'deep relationships' with AI bots as the technology becomes more powerful, Google boss predicts". In: *The Sun*.

Hodges., Hope, Carolyn Garrity., and James Pope. (2024). "Deep Learning, Feature Selection and Model Bias with Home Mortgage Loan Classification". In: *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods - ICPRAM*. INSTICC. SciTePress, pp. 248–255.

Home Mortgage Disclosure Act (HMDA) Data (09/2022). en.

Jui, T.D. and P. Rivas (01/2024). "Fairness issues, current approaches, and challenges in machine learning models". In: *International Journal of Machine Learning and Cybernetics*.

Kamiran, Faisal and Toon Calders (10/2012). "Data preprocessing techniques for classification without discrimination". en. In: *Knowledge and Information Systems* 33.1, pp. 1–33.

Kamishima, Toshihiro et al. (2012). "Fairness-Aware Classifier with Prejudice Remover Regularizer". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 35–50.

Karim, Rezaul et al. (10/2023). "Interpreting Black-box Machine Learning Models for High Dimensional Datasets". In: pp. 1–10.

Kearns, Michael et al. (01/2019). "An Empirical Study of Rich Subgroup Fairness for Machine Learning". In: pp. 100–109.

Kelp, Dr Torsten and Dr Martina Schneider (07/2023). "When an algorithm decides on the loan". In: *BaFin Expert Articles*.

Kim, Been, Rajiv Khanna, and Oluwasanmi O Koyejo (2016). "Examples are not enough, learn to criticize! Criticism for Interpretability". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc.

Kim, Savina D., Galina Andreeva, and Michael Rovatsos (05/2023). "The 40-year journey for fairness in credit: A systematic review and future research directions". In.

Krishna, Satyapriya et al. (02/2022). "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective". In.

Kusner, Matt et al. (03/2017). "Counterfactual Fairness". In: *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*.

Lee, Michelle Seng Ah and Luciano Floridi (03/2021). "Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs". en. In: *Minds and Machines* 31.1, pp. 165–191.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (01/2021). "Explainable AI: A Review of Machine Learning Interpretability Methods". en. In: *Entropy* 23.1, p. 18.

Lindsey-Taliefero, D. and L. Kelly (2021). "Reverse Mortgage Lending Disparities and the Economically Vulnerable". In: *Int Adv Econ Res* 27, pp. 159–169.

Lipton, Zachary C. (06/2018). "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57.

Lundberg, Scott and Su-In Lee (12/2017). "A Unified Approach to Interpreting Model Predictions". In.

Maass, Wolfgang et al. (2022). "AI Explainability: Embedding Conceptual Models". In: *ICIS 2022 Proceedings*. Vol. 12.

Marcinkevičs, Ričards and Julia E. Vogt (2023). "Interpretable and explainable machine learning: A methods-centric overview with concrete examples". en. In: *WIREs Data Mining and Knowledge Discovery* 13.3, e1493.

Martinez, Emmanuel and Lauren Kirchner (08/2021). "The Secret Bias Hidden in Mortgage-Approval Algorithms". In: *The Markup*.

Mehrabi, Ninareh et al. (07/2021). "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6, 115:1–115:35.

Molnar, Christoph (2023). *Interpretable Machine Learning*.

Murdoch, W. James et al. (10/2019). "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44.

Nam, Tong-yob and Hayong Yun (10/2022). *Algorithms and Fairness in Lending Markets: When Do Humans and Machines Disagree?* en. Tech. rep. 4261484. Rochester, NY.

Padmanabhan, Deepak, Sanil V., and Joemon M. Jose (2021). "On Fairness and Interpretability". In: *IJCAI 2021 Workshop on AI for Social Good*.

Pessach, Dana and Erez Shmueli (01/2020). *Algorithmic Fairness*. Tech. rep. arXiv:2001.09784 [cs, stat] type: article. arXiv.

Pleiss, Geoff et al. (2017). "On Fairness and Calibration". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.

Pradhan, Romila et al. (06/2022). "Interpretable Data-Based Explanations for Fairness Debugging". In: *Proceedings of the 2022 International Conference on Management of Data*. SIGMOD '22. New York, NY, USA: Association for Computing Machinery, pp. 247–261.

Regulation - 2016/679 - EN - gdpr - EUR-Lex (2024). en.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""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*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 1135–1144.

Rugh, Jacob S., Len Albright, and Douglas S. Massey (05/2015). "Race, Space, and Cumulative Disadvantage: A Case Study of the Subprime Lending Collapse". In: *Social Problems* 62.2, pp. 186–218.

Saleem, Rabia et al. (2022). "Explaining deep neural networks: A survey on the global interpretation methods". In: *Neurocomputing* 513, pp. 165–180.

Saleiro, Pedro et al. (2018). "Aequitas: A Bias and Fairness Audit Toolkit". In: *arXiv preprint arXiv:1811.05577*.

Sargeant, Holli (11/2022). "Algorithmic decision-making in financial services: economic and normative outcomes in consumer credit". In: *AI and Ethics* 3.

Sharma, Shubham, Jette Henderson, and Joydeep Ghosh (10/2022). *FEAMOE: Fair, Explainable and Adaptive Mixture of Experts*. Tech. rep. arXiv:2210.04995 [cs] type: article. arXiv.

Singh, Arashdeep et al. (03/2022). "Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair". en. In: *Machine Learning and Knowledge Extraction* 4.1, pp. 240–253.

So, Wonyoung et al. (2022). "Beyond Fairness: Reparative Algorithms to Address Historical Injustices of Housing Discrimination in the US". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, pp. 988–1004.

Teodorescu, Mike Horia et al. (08/2020). "A Framework for Fairer Machine Learning in Organizations". In: *Academy of Management Proceedings* 2020.1, p. 16889.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2017). "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31, pp. 841–887.

Weerts, Hilde et al. (2023). "Fairlearn: Assessing and Improving Fairness of AI Systems". In: *J. Mach. Learn. Res.* 24, 257:1–257:8.

Zafar, Muhammad Bilal et al. (04/2017). "Fairness Constraints: Mechanisms for Fair Classification". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, pp. 962–970.

Zhou, Jianlong, Fang Chen, and Andreas Holzinger (2022). "Towards Explainability for AI Fairness". In: *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Ed. by Andreas Holzinger et al. Cham: Springer International Publishing, pp. 375–386.

Zou, L. and W. Khern-am-nuai (2023). "AI and housing discrimination: the case of mortgage applications". In: *AI Ethics* 3, pp. 1271–1281.