CITY,
UNIVERSITY OF LONDON

MASTER'S THESIS

Intersectional biases in loan denial prediction. An interpretability with SHAP values.

Author:
Antonio Jose
LOPEZ ROLDAN

Supervisor:
Dr. Oleksandr
GALKIN

A thesis submitted in fulfillment of the requirements for the degree of Master's Degree in Data Science

in the

Department Of Computer Science

October 1, 2024

Declaration of Authorship

I, Antonio Jose
LOPEZ ROLDAN, declare that this thesis titled, "Intersectional biases in loan denial prediction. An interpretability with SHAP values." and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Antonio Jose Lopez Roldan

Date: 01/10/2024

CITY,
UNIVERSITY OF LONDON

Abstract

Department Of Computer Science

Master's Degree in Data Science

Intersectional biases in loan denial prediction. An interpretability with SHAP values.

by Antonio Jose
LOPEZ ROLDAN

This study evaluates the performance of Logistic Regression and LightGBM models in predicting loan denial using the Home Mortgage Disclosure Act dataset for the year 2017 in New York City, focusing on fairness and bias mitigation across various intersectional demographic groups. The project applies re-weighting (from AIF360) for bias reduction and explores their limitations in addressing disparities in model results. Using SHAP values for interpretability, the analysis reveals the influence of financial and demographic factors on predictions, highlighting the presence of demographic bias for underrepresented groups. Additionally, the study demonstrates how dataset size impacts fairness and model performance...

Acknowledgements

To my project advisor, for being the lighthouse when I needed guidance. . .

Contents

List of Figures

List of Tables

Dedicated to
my parents, who are the pillars of who I have become as a person. . .

Chapter 1

Introduction

1.1 Introduction

Credit risk analysis (assessment) is the process of evaluating the probability that a borrower will default on their financial obligations. It involves assessing a counterparty's ability and willingness to fulfill their credit commitments, considering factors such as their payment history, current financial situation, and future capacity to repay. The purpose of this analysis is to balance the potential benefits of lending against the risk of non-payment (probability of default), ultimately aiming to minimize financial losses while maintaining profitable outcomes.

Credit risk management can be traced back to ancient times, with some evidence in Hammurabi's Code from 1750-1755 BC in Mesopotamia. This suggests that credit risk management has existed for centuries and has evolved across different cultures and eras. Today, the methods and tools used for credit risk management continue to vary based on culture, region, and levels of economic development. For example, the United States, Europe, India, and China, have specific regulations, which differ from one another. However, most relevant international banks are under the umbrella of an international regulatory framework that we will see later.

On the other hand, credit risk assessment employs many types of approaches.

- Judgmental, which is the oldest and the simplest relying on the expertise of a professional in the field rather than theory, or empirical data. It is the most vulnerable to human bias.
- Data-driven empirical models, which are based on historical data about loans accepted, rejected, defaults, and paid. This approach can be traced back to the late 1960s. (Tounsi, Hassouni, and Anoun, 2017). Bias identified by models can be exacerbated.
- Financial models, which are based on theory.

This piece of work will be centred on the probability of default, which falls under the scope of the data-driven empirical model.

In recent years, credit risk management, and in particular data-driven approaches, have experienced an extraordinary transformation. The integration of artificial intelligence, big data, mobile applications and evolving regulatory frameworks has revolutionized the field as financial institutions have widened the variety of assessments

applied, such as behavioural scoring, collection scoring, recovery score, etc. (Tounsi, Hassouni, and Anoun, 2017).

Despite these advancements, AI-driven credit risk models can inadvertently lead to discrimination based on race, sex, age, marital status, and other characteristics. In this way, as models learn from data they can perpetuate or exacerbate any presence of societal biases. (Garcia, Garcia, and Rigobon, 2023) covers in depth many scenarios where biases are present. For example, individuals from historically underserved communities may receive unfavourable credit assessments due to limited financial documentation, rather than their true creditworthiness. (Hurley and Adebayo, 2016)

This scenario not only underscores the critical need to address inequalities in credit risk assessment but also ignores the impact of unfairly excluding individuals from access to credit on companies' income.

Last but not least, multiple authors stress that most studies only focus on one sensitive attribute at a time, such an approach is limited and doesn't reflect reality since in real-world scenarios multiple sensitive attributes come into play when assessing subjects. (Garcia, Garcia, and Rigobon, 2023) mention some of these authors as well as make a "call for more research considering multiple attributes at a time". Among those authors worth citing the innovative approach of (Singh et al., 2022), Dualfair, to analyze multiple sensitive attributes simultaneously.

1.2 Research Questions

Under this context, this work will be centered on "unfair" denials that may arise when assessing the probability of default (PD) on individuals. To assess the outcomes of the PD we will follow the path done by other pieces of work such as (Garcia, Garcia, and Rigobon, 2023) and (Das, Stanton, and Wallace, 2023). We will use the Receiver Operating Characteristics (ROC) and the Log Loss (Negative Log Likelihood) to evaluate the models; these models are, Linear Regression simple, Linear Regression advanced, LightGBM simple, LightGBM advanced. In addition, we will use 3 different dataset sizes; 100%. 50% and 25% of the total dataset after cleaning and preprocessing it:

- Models' performance in terms of making predictions using the metrics ROC-AUC and Log Loss. For the three datasets and the 4 different models, with and without bias mitigation:
  - Which model seems to be more robust to the different datasets?
  - Which five intersectional groups have the best model performance?
  - Which five intersectional groups have the worst performance?
- Disparate impact. For the three datasets and the 4 different models, with and without bias mitigation:
  - Overall, which model seems to have the biggest disparate impact, and which the least? is the disparate impact affected by the different batches?
  - Which five intersectional groups have the best disparate impact?

  - Which five intersectional groups have the best disparate impact?
- Explainability. Next we would use SHAP values to explain the decisions of our best model:
  - In general, which variables influenced the most on the outcomes?
  - For the three intersectional groups with the most favourable disparate impact. Which are the variables that influenced the most on these outcomes?
  - For the three intersectional groups with the least favourable disparate impact. Which are the variables that influenced the most on these outcomes?
- Trade-off. We will evaluate the Trade-off between model performance and bias mitigation (re-weighting) for the advanced models on the largest batch:
  - Does re-weighting improve model fairness?
  - Is there any effect on the models' performance?

1.3 Expected outcomes

- By using different batches we expect to assess models' robustness and identify under which batch size our models perform the best.
- In this way, we also expect to see how the dataset size affects the disparate impact across the intersectional groups.
- If there are any biases, we expect to add some clarifications on the potential reasons for them across the most prominent intersectional groups.
- We expect to narrow down which intersectional groups are affected the most by biases.
- We expect to assess the effectiveness of re-weighting as bias mitigation technique.
- We expect to discover the impact of bias mitigation on model performance.

1.4 Beneficiaries

Overall, the results could have a positive impact on different economic agents:

- Bias visibility on minorities. (Garcia, Garcia, and Rigobon, 2023) conduct an exhaustive analysis of the biases present in society. In terms of gender, they identify higher bias against women, and in terms of race higher bias against black people, hence with our study, we will potentially uncover that one of the most punished group will probably be black women, pointing out unfair discrimination is the first step to abolish it.
- Through their local branches, financial institutions would see an increased their customer base without increasing the risk assumed apart from improving their

public image and reducing the possibilities of penalties due to discriminatory practices regulated by law.

- On a broader scale, the economy as a whole. Money is the lifeblood of economic systems, and expanding access to financial services for previously marginalized groups could ripple through society with varying degrees of impact. However, the magnitude of this effect would depend on factors such as the proportion of the population affected or the cultural landscape among others. A more inclusive financial system could act as a catalyst, igniting latent entrepreneurial spirits and fuelling consumer spending. Moreover, reducing financial disparities may contribute to a more balanced and prosperous economic landscape. (Park and Mercado, 2015); (Chinnakum, 2023)
- Our exhaustive screening on the Home Mortgage Disclosure Act (HMDA) dataset with the application of four models and three different datsets will potentially add extra light on systemic inequalities in housing finance supporting housing advocacy groups on their cause by providing evidence for their work.

1.5 Structure of the report

The remainder of this thesis is structured as follows:

- Chapter 2, literature review.
- Chapter 3, methodology used.
- Chapter 4, results obtained.
- Chapter 5, analysis and discussion.
- Chapter 6, conclusions.

Chapter 2

Context

2.1 Current State of Credit risk assessment

As mentioned credit risk assessment and economic development, among other factors, go hand in hand, therefore the regulation framework, evolution, et cetera have been different for the last years across the different regions from around the globe.

This paper is mainly centred on the case of Europe, and EEUU where back in the 1930s and 1940s credit scoring began with statistical classification methods becoming commercialized in the 1950s. The introduction of the FICO score by Fair Isaac Corporation in 1956 standardized credit scoring (Das, Stanton, and Wallace, 2023). Credit scoring models estimate the probability of default (PD), this probability of default represents the most important component of a credit risk model which in turn estimates the expected financial loss that a credit institution suffers if a borrower defaults to pay back a loan (Bussmann et al., 2021). Generally, before the 1980s most lending decisions were based on a judgemental approach which was potentially subjected to personal bias (Hurley and Adebayo, 2016). During the 1980s and early 1990s, some attempts were made to incorporate some analytical abilities into purely judgmental systems. These attempts mostly focused on the expert systems (ESs) technology, which was emerging during that period. ESs provided structured reasoning capabilities, often combines with database management, and can be considered as the first attempt towards more sophisticated artificial intelligence approaches that have attracted considerable interest recently (Doumpos et al., 2019).

Artificial intelligence (AI) can be defined as the development of computer systems to perform tasks that ordinarily require human intelligence, in this sense, AI and machine learning (ML) act as an "extension" of the human expert in the field to elevate their capabilities. Interest in ML in particular has become increasingly popular, due to, a combination of more digitized data, faster computers, and better algorithms to analyze data, but not only that, the increasing availability of resources like cloud platforms (e.g. Amazon Web Services (AWS)) offer scalable and cost-effective computing. This allows businesses and individuals to experiment and deploy ML models without heavy upfront investments in infrastructure (Das, Stanton, and Wallace, 2023). Additionally, recent studies underscore the transformative impact of these advancements. According to (Suhadolnik, Ueyama, and Da Silva, 2023), financial institutions and regulators increasingly rely on large-scale data analysis, particularly machine learning, for making credit decisions.

(Noriega, Rivera, and Herrera, 2023) further emphasizes the importance of AI and ML in assessing credit risk by analyzing large volumes of information (Big data). They identify key challenges such as the need for explanatory artificial intelligence and addressing data imbalance issues. Addressing these challenges is critical for improving model performance and ensuring fair credit decisions.

Moreover, (Bussmann et al., 2021) highlights the economic significance of the retail credit market, where ML-based scoring models play a crucial role in loan approvals. At this point it is worth remembering that by 2010, over 90 per cent of lenders used FICO scores for lending decisions and the potential impact on society (Hurley and Adebayo, 2016). Continuing with (Bussmann et al., 2021) they argue that the ability of machine learning models to capture non-linear relationships within financial data makes them particularly well-suited for credit risk assessment. The integration of these technologies enables the discovery of complex patterns and interactions, providing a more comprehensive view of a borrower's creditworthiness.

On the other hand, alternative credit-scoring techniques while potentially reaching customers excluded by traditional methods, are not the panacea. Some models employ unconventional data types with a lack of clear connections to financial behaviour. Although these tools may benefit certain consumers, they also pose significant risks since by integrating thousands of data points collected without consumer knowledge, they create serious transparency problems and consequently, consumers find it difficult to call into question unfair decisions or understand how to improve their credit standing (Hurley and Adebayo, 2016).

2.2 Advantages, Disadvantages, Challenges

At this level, we acknowledge that the use of machine learning in credit scoring presents both benefits and drawbacks, as well as hurdles to overcome. We mention some of them below:

2.2.1 Advantages

- Artificial intelligence and machine learning can analyze data in real-time, providing up-to-date risk assessments. This is particularly useful for dynamic financial environments where borrower profiles can change rapidly (Ray and Luz, 2024).
- AI and ML models can continuously learn and adapt to new data, allowing for more dynamic and up-to-date risk assessments compared to traditional static models (Leo, Sharma, and Maddulety, 2019).
- Machine learning models can handle large volumes of data. This is particularly relevant for financial institutions dealing with large numbers of transactions and customer data (Hurley and Adebayo, 2016).
- ML models can be notably effective at detecting patterns of potential fraudulent activity, helping to reduce losses due to fraud (Ray, 2022).

2.2.2 Disadvantages

- Transparency issues. The intricate processes involved in feature selection and model construction make it difficult for consumers to understand how their credit scores are determined (Hurley and Adebayo, 2016).
- The reliance on large data sets increases the risk of incorporating not genuine correlations or biased data into the model. Even when models are statistically accurate, they may perpetuate existing social inequalities by reinforcing stereotypes or penalizing individuals based on group characteristics, leading to outcomes that are not just based on personal credit behaviour (Hurley and Adebayo, 2016).
- Sensitive characteristics such as race or gender may not be directly used as inputs, but the model could still learn to associate these characteristics with creditworthiness through proxy variables like zip codes or consumption patterns. This can lead to discriminatory outcomes, where people are unfairly judged based on generalized data profiles rather than their individual merits (Hurley and Adebayo, 2016).

2.2.3 Challenges

- According to (Hurley and Adebayo, 2016) ensuring fairness requires minimizing outcome disparities across groups. The authors also note that bias is interconnected with causality and privacy, suggesting that a lack of privacy can potentially reveal or activate biased algorithms.
- Many authors note the fact that advanced ML models are prone to overfitting (Doumpos et al., 2019); whereas (Das, Stanton, and Wallace, 2023) highlight the need for skilled model fitting and regularization techniques to improve performance.
- There's also concern that some tools may be designed to target vulnerable individuals for high-cost loan products rather than predict creditworthiness (Hurley and Adebayo, 2016).
- Simple statistical learning models, such as linear and logistic regression models, provide high interpretability but possibly limited predictive accuracy. Complex machine learning models, such as neural networks and sofisticated tree-based models, provide high predictive accuracy at the expense of limited interpretability (Bussmann et al., 2021).

2.3 Societal Impact of Credit Assessment

The impact of credit scores and reports on society is huge. Employers often use them in hiring and promotion decisions, while landlords utilize them to screen potential tenants. In addition, in many countries, credit access is crucial to achieving major life goals like buying homes, cars, or getting access to higher education (Hurley and Adebayo, 2016).

This environment turns people into vulnerable actors in the economic system. Many people become economically insolvent due to situations that they cannot predict and make them unable to repay their debts. Loss of work, the emergence of an extraordinary economic need, a severe economic crisis at the regional or country level, the irregular and drastic political changes that are common in, mainly emerging markets, can significantly affect the ability of a borrower to meet existing debt obligations (Doumpos et al., 2019).

2.4 Theoretical Framework

2.4.1 Bias, fairness, disparate impact, disparate treatment

We are aware that the definition of fairness and discrimination is quite complex and a current challenge to solve in the application of AI in credit risk assessment. Many papers have dealt with the fairness definition and there is a clear consensus, "there is no consensus on what is fair". As an example (Fazelpour and Danks, 2021), highlight the lack of consensus on what constitutes fairness in the context of algorithmic design and deployment. They also emphasize that pluralism can even worsen algorithms, they stress the difficulty of achieving a universally accepted notion of fairness in algorithmic systems.

In addition (Verma and Rubin, 2018), state that "the answer to this question depends on the notion of fairness one wants to adopt. We believe more work is needed to clarify which definitions are appropriate to each particular situation." The paper also notes that there are more than twenty different notions of fairness, further emphasizing the lack of agreement on a single definition.

Besides, it is crucial to distinguish between fairness and bias. Following with (Fazelpour and Danks, 2021); Bias is fundamentally about systematic deviation in the algorithm's output, performance, or impact relative to a specific norm or standard. This deviation can be statistical, moral, or social, depending on the chosen benchmark. For instance, an algorithm may exhibit statistical bias if its predictions deviate from the training data, or moral bias if its decisions unfairly favour one group over another based on sensitive attributes like race or gender. Fairness, on the other hand, is a more complex and multifaceted concept. It involves ensuring that the algorithm's outcomes align with ethical and social norms avoiding discrimination.

Additionally, (Chouldechova, 2017) add a crucial point when emphasising that fairness, including the concept of disparate impact (used in this paper), is fundamentally rooted in social and ethical considerations rather than purely statistical ones, Alexandra also note that even a predictive tool meeting specific fairness criteria could still result in unequal outcomes depending on its application context and implementation methods.

It is also important to differentiate disparate impact from disparate treatment, as highlighted by (Feldman et al., 2015). Disparate impact occurs when a model's predictions or decisions result in significantly different outcomes for different groups, despite seeming neutral. While disparate impact refers to unintentional discrimination

through model's outputs, disparate treatment involves direct, intentional discrimination. Finally, it's worth highlighting that algorithmic discrimination is the manifestation of bias.

2.4.2 The Choice of ML Model & Explainability Tool

Given the need for interpretability and explainability, we evaluated the following tools: Partial Dependence Plots (PDP), Local Interpretable Model-Agnostic Explanations (LIME), Accumulated Local Effects (ALE), and SHapley. We found SHapley to be the most advantageous for the following reasons: : (1) SHAP is based on Shapley values from cooperative game theory, ensuring fair attribution and consistency. (2) SHAP offers both global and local explanations that align with each other (3) SHAP inherently considers feature iterations.

Regarding model selection, we initially considered logistic regression and decision trees as baseline models, with Neural Networks (NN), Support Vector Machines (SVM), XGBoost, and LightGBM as potential candidates for our advanced model. However, we conclude LightGBM to be the best option against the rest of the models. According to (Hlongwane, Ramaboa, and Mongwe, 2024) LightGBM offers faster and more efficient performance, especially with large datasets, compared to XGBoost. It provides better interpretability than Neural Networks (Lundberg et al., 2019) and, overall is more advantageous than Support Vector Machines (Lextrait, 2023), although this last point might vary depending on the dataset and other factors. Logistic regression will serve as our baseline model as is one of the most established models in credit risk (Lessmanna et al., 2013).

Additionally, the findings of many papers corroborate LightGBM as the best option. The empirical analysis conducted by (Suhadolnik, Ueyama, and Da Silva, 2023) demonstrates that ensemble models, like XGBoost, outperform traditional algorithms like logistic regression in credit classification tasks. (Ponsam et al., 2021) go further concluding that LightGBM outperforms XGBoost in terms of both training and testing speeds, reaching the conclusion that LightGBM is the better choice between the two of them. This highlights the superiority of modern ML techniques over the considered more traditional methods.

Furthermore, the systematic review undertaken by (Noriega, Rivera, and Herrera, 2023) of machine learning applications in credit risk prediction identifies ensemble methods, such as LightGBM, as the most researched and effective for this purpose.

2.4.3 Understanding Interpretability & Explainability

(Arrieta et al., 2020) define interpretability as an intrinsic characteristic of a model. It refers to how easily a person can understand the model's inner workings and the reasoning behind its decisions, without needing additional explanations or tools. In this context, according to the literature reviewed LightGBM demonstrates superior interpretability among the evaluated models (SVM, NN, XGBoost) in the context of credit risk assessment.

Conversely, explainability is not inherent to the model itself, but rather the ability to provide clarifications or details about how the model works.

2.5 Legal and Regulatory Frameworks

We aim to create our work according to the Basel framework. The Basel Framework is a set of rules created by the Basel Committee on Banking Supervision (BCBS). The BCBS is the main global group that sets standards for regulating banks. It operates under the Bank for International Settlements (BIS) but is separate from it. All BCBS members have agreed to fully implement these rules for international banks in their countries. Some BCBS members are China, Russia, Saudi Arabia, United Kingdom and Argentina.

However, the regulatory framework becomes narrower as we get closer to a specific region. For instance, in Europe we have the European Banking Authority (EBA) in charge of providing instructions about the requirements for the creation of machine learning for IRB (Internal Rating-Based) models, in contrast to the US where four bodies are involved in the regulation of machine learning for IRB models; the Federal Reserve Board (FRB), the Office of the Comptroller of the Currency (OCC), the Federal Deposit Insurance Corporation (FDIC), the Consumer Financial Protection Bureau (CFPB).

Our work is designed with a global scope in mind, aiming to align with international regulatory expectations and best practices in risk management. We aim to create a work that adheres to the (BCBS) standards (Bank for International Settlements (BIS), 2022). As such, we will briefly enumerate the minimum requirements summarized by (Doumpos et al., 2019):

- Meaningful differentiation of risk: a credit system should focus on accurately distinguishing different levels of credit risk. Borrowers in the same grade should be treated differently based on the specific characteristics of their transactions. Moreover, it is crucial to ensure that credit exposures are well-distributed across all grades to avoid excessive concentration in any one grade.
- Continuous evaluation of borrowers: credit portfolios require a comprehensive annual review of all borrowers and their associated loans. This process should integrate any recent developments or changes in the borrower's situation and financial progress. This ongoing assessment is crucial for maintaining an up-to-date and reliable credit risk management framework.
- Oversight of the system's operations: Ongoing oversight of credit scoring/rating systems is essential for maintaining accuracy and performance. Regular stress testing and a strong feedback loop among stakeholders are crucial. These practices ensure system integrity and the reliability of risk evaluations.
- Correct selection of risk assessment attributes: Credit model developers should employ relevant risk factors to accurately assess borrower creditworthiness. The analysis must be predictive, utilizing current borrower data and external conditions to forecast future performance. Selected attributes should meaningfully reflect actual risk levels.

- Collecting a substantial database: Developing and evaluating credit rating systems requires a robust, representative database. This should encompass historical borrower data, including past scores, ratings, default probability estimates, credit migrations, and payment records. Such comprehensive data ensures the system's accuracy and real-world applicability in risk assessment.

Chapter 3

Methods

3.1 Data

The dataset utilized in this study is the Home Mortgage Disclosure Act(HMDA) for New York, year 2017, and is publicly available from the Consumer Finance Protection Bureau website. There are some HMDA datasets ready to use on Kaggle, nonetheless, they are indeed overused. Besides, we were looking for a dataset large enough so the data size is not a constrain, moreover, as our study is directly linked to race, due to its multiculturalism and diversity, New York is one of the naturally less unbalanced HMDA datasets available since it is home for almost every nationality in the world and the proportion of white people, which in the United States is the majority, has less relevance in comparison with other locations.

The original dataset is full of information with 78 columns and 446,902 records offering a wide overview of the mortgage landscape. Key features include loan type, purpose, and amount; applicant race, sex, and income; as well as loan outcomes (originated, approved but not accepted, or denied). Those characteristics make it notably suitable for creating and evaluating credit risk models aligning with our objective of assessing the intersectional bias against profitability.

However, due to the high demand on computing resources, to deal with the dataset easily we split it into smaller pieces and then iterated through each chunk. Although we later used cloud computing services, such as Amazon Web Services (AWS) and Google Colab.

Since this study is focused on intersectional bias (ethnicity-race-sex) related to sole applicants, the first step was to extract data related to the cases where only one person applied, in other words, without co-applicants. Consequently, the original dataset became more manageable and was reduced to 242.296 records. In addition, we created the columns "loan to income ratio" (after preprocessing our dataset) and "ethnicity-race-sex" and then excluded some other columns to lighten the dataset for our exploratory data analysis.

3.2 Exploratory data analysis

From the initial analysis of the new column created (ethnicity-race-sex), we discovered that males are the majority applicants for each ethnicity-race except for black

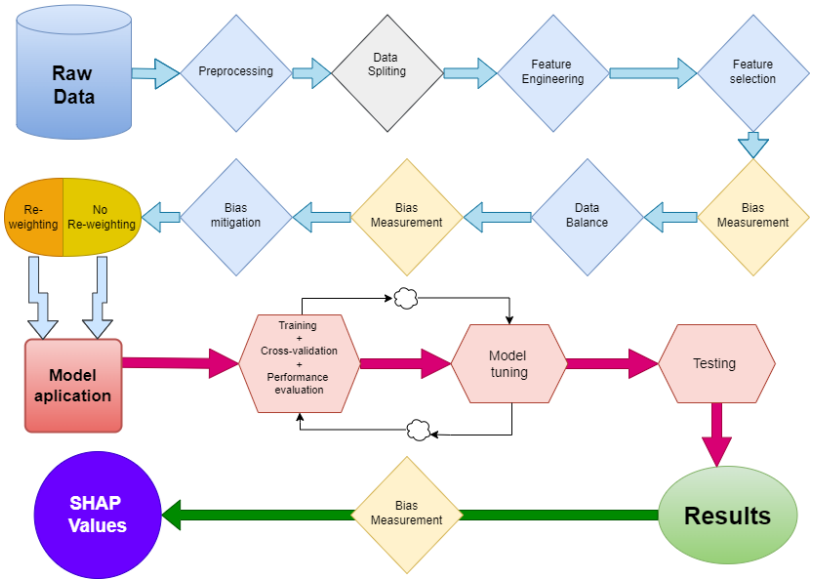

FIGURE 3.1: Workflow

individuals. We calculated the ratio of females against men (female/male) and overall females apply less than males. The results add some interesting facts:

- Hispanic or Latino females have a fraction of 0.55:1, they apply almost half of the time than men. This makes sense due to a cultural connotation (traditional gender roles).
- On the opposite, Black women surpass applications to men by 1.19:1 times. In other words, when men make 5 applications women make 6 applications.

Strikingly, black females' applications exceeded males. We further investigated the case and found out that a considerable amount of data linked to race was missing whereas data for sex was not missing, which might explain why black women's applications are notably higher to black males. The key point here is that the missing race data creates an incomplete picture, making it seem like Black women applications are higher than black males, when in reality, the missing data might hide some of the Black male applicants.

While doing our exploratory data analysis (EDA) many questions were raised. Questions that might add context to interpret and understand the final model's outcomes:

1. Is there any relationship between missing values from geographical-related columns and the approval/denial rate?:

   The data shows that white men are the most affected (negatively) when information linked to potential downsides is provided and the most beneficiated when is not provided. it can denote "you are a white male trustful until you demonstrate the opposite by showing where you live"

2. What is the ratio of "denials" to "approvals" for each record type from the column "ethnicity_race_sex"?:

- Highest ratios: Hispanic Native Hawaiian or Pacific Islander males (1.47:1) and females (1.24:1) meaning that these individuals have a higher ratio of denied to approved loan applications compared to other communities.
- Lowest ratios: Non-Hispanic Asian females have the lowest ratio at 0.23:1, closely followed by non-Hispanic white males at 0.25:1. These groups have the highest chances of getting approved their mortgage application.

3. Impact of missing information.
   - In regards of denials, White males and white females are the groups less affected when information is missed.
   - Black or African American (male and female), as well as, American Indian or Alaska Native, Males stand out as the most punished groups when some pieces of information are not provided.
4. Other pieces of relevant information extracted from our analysis:
   - The number of applications from men was 68% higher than from women.
   - Average loan amount requested by men was $276.97 (in thousands) whereas by women $212.03 (in thousands)
   - "Not hispanic or latino asian males" hold on average the highest income, circa 145 thousand ($), while "Indian or alaska native females" are the lowest with circa 74 thousand ($).
   - Minorities are also overcharged the most when loans are approved, and white females/males are the least punished with overcharging represented by the ("rate spread") column.

Finally, we show the proportions of applications approved - denied per group in the figure 3.2 and for broader context figure 3.3 displays the percentage of applications made by each group relative to the total number of applications.

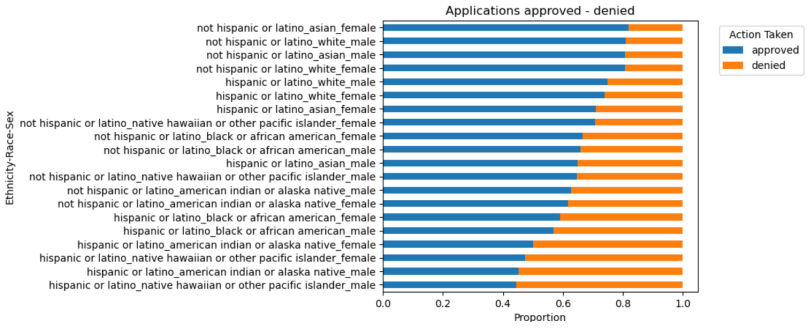

FIGURE 3.2: Proportion of applications approved - denied

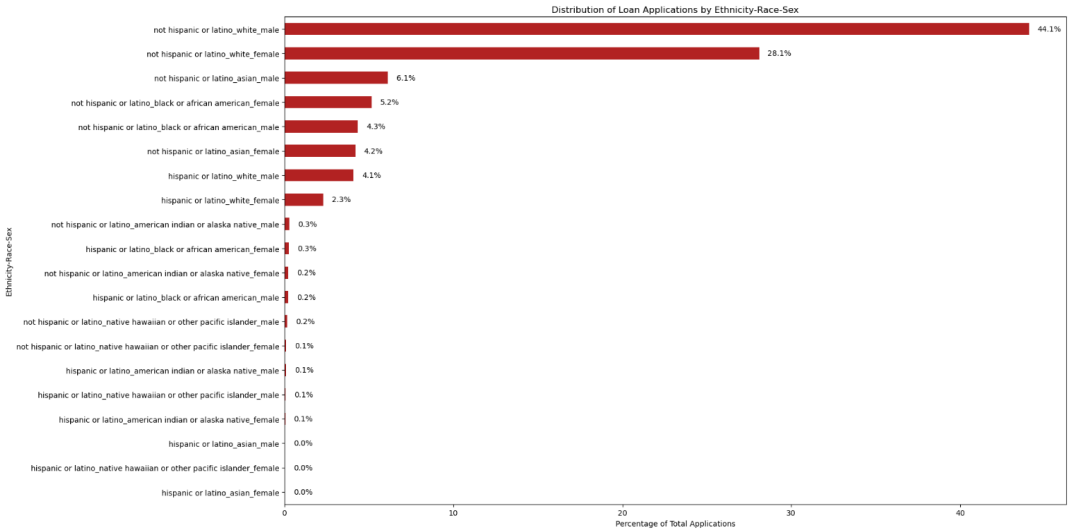

FIGURE 3.3: Total applications in (%) per group

3.3 Preprocessing and feature engineering

During the following steps, we would clean and preprocess our data to improve its quality as input. Some of the most notable tasks done were:

- Since we considered the previously missing data linked to census columns to be meaningful due to the proportion represented in certain groups. Our approach is based on many other works such as (Nasima, 2021), (Emmanuel et al., 2021), (Brownlee, 2020). To try to capture this information:
    1. We would create new columns that mirror the missing values for the original columns. We keep the original columns as well.
    2. We would apply median imputation to fill out the missing values from the "original columns". For the imputation, we will consider each group from the column "ethnicity_race_sex" as benchmarking.
- Removed outliers by trimming data below the 1st and above the 99th percentile.
- We evaluated skewness to decide whether to apply log transformation, Z-score standardization, or both to our data.
- Apart from cross-validation, we also applied other techniques to avoid data leakage, for example:
    1. We applied scaling and one-hot encoding separately to training and test datasets.
    2. We would first, fit preprocessing steps on training data only.
    3. Then, we would use the parameters learned from the training data to transform both, the training and test sets.

- To apply one hot encoding we used "OneHotEncoder" instead of other methods such as "get_dummies". "OneHotEncoder" prevents data leakage and handles unseen data adequately as (Géron, 2022) mentions in his work.
- Created a new column combining the three different columns of ethnicity, sex, and race. Initially, ethnicity was omitted however, ethnicity is mostly about being Hispanic/Latino which from our literature review, we concluded that the ethnicity factor would potentially have an impact on credit risk assessment. We also created the loan-to-income ratio column, which replaces the applicant income and loan amount columns on our feature input list.
- Defined "action taken" variable as our binary target variable. It contains two classes, denied or approved.
- Unbalanced target variable. We applied Synthetic Minority Oversampling Technique (SMOTE) to increase the number of cases and balance our dataset:

Approved (0)	81489 (Records)
Denied (1)	21213 (Records)

TABLE 3.1: Before smote

Approved (0)	81489 (Records)
Denied (1)	81489 (Records)

TABLE 3.2: After smote

3.4 Model and feature selection

- Feature selection:
  - Checked for correlations among our features vs target variable with Logarithmic coefficients.
  - Not checking for multicollinearity would be big error that could ruin our work. We checked it with variance inflation factor (VIF) and excluded any feature with a VIF greater than 5 units. The work of (Chan et al., 2022) was especially insightful for this purpose.
  - Manual experiments. Following our previous readings and learnings we also intuitively selected variables to check the baseline model's performance (Simple Logistic Regression).

  The goal of this study is to predict the probability of default (PD). To evaluate our models, we focused on two key metrics: Receiver Operating Characteristic (ROC-AUC), used in other pieces of work to evaluate the PD such as in (Doumpos et al., 2019) as well as Logarithmic Loss. Feature selection was guided by the model's performance on these metrics. Table 3.3 provides a detailed description of each selected feature note that the data type is "float64" as all of them are ready to use as input.
- Bias Measurement. To measure potential biases, we employed the disparate impact metric. We assessed disparate impact for each intersectional group individually, as well as across the overall model outputs. Our approach was

TABLE 3.3: Feature Description

Feature	Description	Data Type
applicant_income_000s	The applicant's income in thousands of dollars.	float64
hud_median_family_income	The Department of Housing and Urban Development (HUD) median family income for the area.	float64
hud_median_family_income_missing	Indicator if the HUD median family income data is missing.	float64
loan_amount_000s	The amount of the loan in thousands of dollars.	float64
loan_to_income_ratio	The ratio of the loan amount to the applicant's income.	float64
action_taken_binary	Binary indicator for whether the loan was approved (0) or denied (1).	float64
minority_population	The proportion of minority population in a census tract.	float64
minority_population_missing	Indicator if the minority population data is missing.	float64
tract_to_msamd_income	The ratio of tract median family income to the median family income of the metropolitan area.	float64
tract_to_msamd_income_missing	Flag if the tract-to-MSA (metropolitan statistical area)/MD (metropolitan division) income data is null.	float64
loan_type_name_Conventional	Loan type is conventional (not insured by a government program).	float64
loan_type_name_FHA-insured	Loan is insured by the Federal Housing Administration (FHA).	float64
loan_type_name_FSA/RHS-guaranteed	Loan is guaranteed by the Farm Service Agency (FSA) or Rural Housing Service (RHS).	float64
loan_type_name_VA-guaranteed	Loan is guaranteed by the Department of Veterans Affairs (VA).	float64
loan_purpose_name_Home improvement	Loan purpose is for home improvement.	float64
loan_purpose_name_Home purchase	Loan purpose is for home purchase.	float64
loan_purpose_name_Refinancing	Loan purpose is for refinancing an existing loan.	float64
property_type_name_Manufactured housing	Property type is a manufactured housing unit.	float64
property_type_name_Multifamily dwelling	Property type is a multifamily dwelling (e.g., flats).	float64
property_type_name_One-to-four family dwelling	Property type is a one-to-four family dwelling (excluding manufactured housing).	float64
lien_status_name_Not secured by a lien	Loan is not secured by a lien.	float64
lien_status_name_Secured by a first lien	Loan is secured by a first lien.	float64
lien_status_name_Secured by a subordinate lien	Loan is secured by a subordinate lien.	float64
ethnicity_race_sex_hispanic or latino_american indian or alaska native_female	Hispanic or Latino ethnicity, American Indian or Alaska Native race, and female gender.	float64
ethnicity_race_sex_hispanic or latino_american indian or alaska native_male	Hispanic or Latino ethnicity, American Indian or Alaska Native race, and male gender.	float64
ethnicity_race_sex_hispanic or latino_asian_female	Hispanic or Latino ethnicity, Asian race, and female gender.	float64
ethnicity_race_sex_hispanic or latino_asian_male	Hispanic or Latino ethnicity, Asian race, and male gender.	float64
ethnicity_race_sex_hispanic or latino_black or african american_female	Hispanic or Latino ethnicity, Black or African American race, and female gender.	float64
ethnicity_race_sex_hispanic or latino_black or african american_male	Hispanic or Latino ethnicity, Black or African American race, and male gender.	float64
ethnicity_race_sex_hispanic or latino_native hawaiian or other pacific islander_female	Hispanic or Latino ethnicity, Native Hawaiian or Other Pacific Islander race, and female gender.	float64
ethnicity_race_sex_hispanic or latino_native hawaiian or other pacific islander_male	hispanic or latino ethnicity, Native Hawaiian or Other Pacific Islander race, and male gender.	float64
ethnicity_race_sex_hispanic or latino_white_female	Hispanic or Latino ethnicity, White race, and female gender.	float64
ethnicity_race_sex_hispanic or latino_white_male	Hispanic or Latino ethnicity, White race, and male gender.	float64
ethnicity_race_sex_not hispanic or latino_american indian or alaska native_female	Not Hispanic or Latino ethnicity, American Indian or Alaska Native race, and female gender.	float64
ethnicity_race_sex_not hispanic or latino_american indian or alaska native_male	Not Hispanic or Latino ethnicity, American Indian or Alaska Native race, and male gender.	float64
ethnicity_race_sex_not hispanic or latino_asian_female	Not Hispanic or Latino ethnicity, Asian race, and female gender.	float64
ethnicity_race_sex_not hispanic or latino_asian_male	Not Hispanic or Latino ethnicity, Asian race, and male gender.	float64
ethnicity_race_sex_not hispanic or latino_black or african american_female	Not Hispanic or Latino ethnicity, Black or African American race, and female gender.	float64
ethnicity_race_sex_not hispanic or latino_black or african american_male	Not Hispanic or Latino ethnicity, Black or African American race, and male gender.	float64
ethnicity_race_sex_not hispanic or latino_native hawaiian or other pacific islander_female	Not Hispanic or Latino ethnicity, Native Hawaiian or Other Pacific Islander race, and female gender.	float64
ethnicity_race_sex_not hispanic or latino_native hawaiian or other pacific islander_male	Not Hispanic or Latino ethnicity, Native Hawaiian or Other Pacific Islander race, and male gender.	float64
ethnicity_race_sex_not hispanic or latino_white_female	Not Hispanic or Latino ethnicity, White race, and female gender.	float64
ethnicity_race_sex_not hispanic or latino_white_male	Not Hispanic or Latino ethnicity, White race, and male gender.	float64

guided by key works in the field, particularly (Bellamy et al., 2018), who developed an extensible toolkit for bias measurement. It is important to clarify beforehand some concepts:

- Disparate impact (DI): Measures the ratio of favorable outcomes between the privileged group (in our case "not hispanic white male") and the unprivileged groups.
  1. A value of 1 means both groups have equal outcomes, hence there isn't any bias.
  2. Values less than 1, means that the output is less favourable for unprivileged groups.
  3. If values are greater than 1, means that the outputs are more favourable for the declared "unprivileged" group.
- Statistical Parity Difference (SDP) tells us the likelihood that the unprivileged group will receive a favourable outcome compared to the privileged group:
  1. If SPD = 0, there is no disparity.
  2. If SPD < 0, unprivileged group has a lower proportion of favourable outcomes.
  3. if SPD > 0, unprivileged group has higher proportions.
- Mean Difference (MD) measures the average difference in predictions made by the model between groups:
  1. MD = 0 means both groups have the same average prediction (no bias).
  2. MD < 0 the unprivileged group has lower average predictions.
  3. MD > 0 the unprivileged group has higher average predictions.

- Bias mitigation. We used re-weighting. It is a technique used to mitigate bias in machine learning by adjusting the importance or influence of different data points in the training set. We mainly used the same reference from above (Bellamy et al., 2018) when trying to mitigate biases by re-weighting.
- Model selection
  - Logistic Regression was chosen as a baseline model due to its stability, simplicity, predictive capabilities and proven reliable starting point in other studies such as (Feldman et al., 2015). Its inherent interpretability allows for tracing the individual contribution of each feature to the outputs. This transparency is particularly valuable in the predictions made by the model where understanding the decision-making process is crucial for compliance and for explaining model decisions within the legal framework.

Moreover, logistic regression is well-suited for binary classification tasks, such as predicting loan defaults (Hurlin, Pérignon, and Saurin, 2022), and it has been a commonly used and established model in the industry (Lessmanna et al., 2013). Therefore its reputation will contribute to both evaluating the LightGBM model's application and contextualizing its performance against a classic, industry-standard approach (Brotcke, 2022). The parameters used on the architecture of the models and a brief description of them are shown in the table 3.5.

– LightGBM is a tree-based learning algorithm. Was selected as the primary model for this study due to its demonstrated ability to handle large datasets and its potential for high predictive accuracy. It has shown promising results in various classification tasks, including credit scoring, often outperforming traditional methods.

  Its capacity to capture complex relationships within the data makes it a valuable tool for identifying subtle patterns that may contribute to credit risk assessment. The interpretability can be a challenge however SHAP values can effectively help in providing insights about the model's decisions. (Ponsam et al., 2021)

• SHapley Additive exPlanations (SHAP) is a model-agnostic approach. It's a method used to explain the predictions of machine learning models by breaking down each prediction into the contributions of individual features. In the context of LightGBM, SHAP helps trace the impact of each feature on the model's output, allowing for a clearer understanding of how predictions are made.

Parameter	Meaning & function
n_estimators	Number of boosting rounds (maximum number of trees in the model)
learning_rate	Shrinks the contribution of each tree by the learning rate
num_leaves	Maximum number of leaves in one tree (controls model complexity)
max_depth	Maximum depth of the tree (-1 means no limit)
min_child_samples	Minimum number of samples a leaf node must have
subsample	Fraction of data used for building each tree (used for bagging)
colsample_bytree	Fraction of features (columns) used when building each tree
reg_alpha	L1 regularization term (controls feature sparsity)
reg_lambda	L2 regularization term (controls feature weights)

TABLE 3.4: Significance of the parameters used on LightGBM

Parameter	Meaning & function
C	Inverse of regularization strength. Smaller values specify stronger regularization (similar to alpha in Ridge/Lasso)
penalty	Type of regularization to apply: • 'l1': Lasso (absolute value regularization) • 'l2': Ridge (squared value regularization) • 'elasticnet': Combination of L1 and L2 regularization
solver	Algorithm to use for optimization: • 'saga': Supports L1, L2, and elasticnet penalties • 'lbfgs': Limited-memory Broyden–Fletcher–Goldfarb–Shanno (supports only L2)
max_iter	Maximum number of iterations the solver can take to converge
l1_ratio	Ratio between L1 and L2 penalties for elasticnet regularization (only used when penalty is elasticnet)

TABLE 3.5: Significance of the parameters used on Logistic Regression model

Architecture justification

The reasons behind our final advanced models' architecture lay in, first, try-error approach. After building the simplest models, we built up a more complex model by adding parameters together with RandomizedSearchCV or GridSearchCV (depending on the case) for hyperparameter tuning. This helped to optimize the models based on our dataset with the most complex configuration, getting a mix of valid outputs and errors until we ended up with the final models:

- In our LightGBM architecture each parameter selected has a reason for existence, here are some examples:

  1. n_estimators: In your case, this value balances prediction accuracy with model complexity. Low values would tend to underfitting, high values would lead to overfitting.
  2. learning_rate: Used in other models such as Deep Reinforcement Learning. Low values ensure that the model coverage slowly and learns gradually, which is important when predicting loan denial.
  3. num_leaves: controls the complexity of individual trees.
  4. max_depth: it gives control over how deep the tree grows. Leaving the model unleashed would lead to overfitting.

  In general, the chosen parameters helped us to have control over generalization, overfitting as well as allowing our model to remain interpretable with SHAP values setting parameters like learning_rate and max_depth. (Ke et al., 2017).

- As we did with LightGBM, for logistic regression we show some examples and their justification of use:

  1. C parameter: controls the strength of regularization. Small values represent strong regularization. Since we have many features, this helps to prevent overfitting.
  2. The penalty controls which type of regularization is applied. "l1" is useful with demographic fractures for example as enforce sparsity. "l2" Shrink coefficient to generalize. "elsticnet" is a combination of both.
  3. Solver, configures the optimization algorithm ('saga'/ 'lbfgs')

  In general, our parameter choice for Logistic Regression intends to guarantee fairness, controlling overfitting and optimizing performance.

Chapter 4

Results

4.1 Model's performance. ROC-AUC, Logistic Loss

In this section, we first display the results obtained for four different models across the different dataset sizes (100%, 50%, 25%) before and after we try to mitigate bias: Logistic Regression Simple, Logistic Regression Advanced, LightGBM Simple, and LightGBM Advanced. Later we dive into a deeper detail with tables showing the model's results for each intersectional group in order to attain a deeper understanding of models' behaviour and biases.

The two metrics used to evaluate our models in their ability to make predictions are:

- ROC-AUC: A measure of the model's ability to distinguish between the two classes (approved - denied), where the closer the values are to 1 (range from 0 to 1) the better the performance.
- Log Loss: compares the predicted probabilities with the actual outcomes. Lower values mean that the predictions made by the model are closer to the true labels and indicate better performance. (Range from 0 to, in theory, infinite).

General models' performance prior and after bias mitigation

From the observation of 4.1 and 4.2 we appreciate:

- The most relevant information is that every model that used the full dataset size, produce the same results before and after bias mitigation. This fact is noticeable for the logistic regression model, both simple and advanced architecture seem to show similar performance, which is somehow surprising since theoretically, the advanced model should have benefited from the advantage of hyperparameter tunning ("RandomizedSearchCV" and "GridSearch") and model tuning. Also, a similar performance is registered for each model across the different batches.
- Besides that, we appreciate exceptional variations such as LightGBM simple's ROC-AUC at 50% dataset size decreased from 0.7100 to 0.6660 after bias mitigation.

- The logistic regression model is also less affected by the different dataset sizes. It exactly produces the same results in each set except for the advanced architecture, 25% of the batch. In general produces slightly better results in ROC-AUC in comparison with lightGBM.
- On the other hand, lightGBM generally seems to perform slightly better in terms of log loss results.

Group	ROC-AUC			Log Loss		
	100%	50%	25%	100%	50%	25%
LR simple	0.7022	0.7019	0.7019	0.6286	0.6296	0.6296
LR advanced	0.7023	0.7019	0.7095	0.6287	0.6296	0.6275
LightGBM simple	0.7152	0.7100	0.7120	0.5541	0.5469	0.5355
LightGBM advanced	0.7008	0.6833	0.6819	0.5332	0.5424	0.5435

TABLE 4.1: Models' performance prior bias mitigation across the different dataset sizes

Group	ROC-AUC			Log Loss		
	100%	50%	25%	100%	50%	25%
LR simple	0.7022	0.7019	0.7022	0.6286	0.6296	0.6286
LR advanced	0.7023	0.7097	0.7095	0.6287	0.6237	0.6275
LightGBM simple	0.7152	0.6660	0.7120	0.5541	0.6275	0.5355
LightGBM advanced	0.7008	0.6824	0.6819	0.5332	0.5423	0.5435

TABLE 4.2: Model's performance with bias mitigation across the different dataset sizes

Models' performance prior and after bias mitigation. Intersectional groups.

In the previous section, we evaluated the models' performance and observed that the application of bias mitigation had an extremely mild effect on the results (overall). Next, we will go further and examine models' behaviour across the intersectional group.

Despite the increased granularity of the analysis, which led us to anticipate greater variation in the results, the observed outcomes appear to align with those of from our previous tables, exhibiting no substantial deviations. The mentioned results are gathered in the tables A.1, A.2, A.3, A.4, A.5, A.6, A.7, A.8, and we can appreciate:

- In most cases, the ROC-AUC values remain relatively stable after bias mitigation. The differences are generally minor.
- Said so, without bias mitigation, the model's performance is generally less stable for smaller and underrepresented groups (but still the differences are minimal).
- Whereas more represented groups in the dataset (e.g., White, Male, and Female) experience a relatively consistent performance regardless of application or not of bias mitigation.
- Both LightGBM (simple and advanced) and Logistic Regression (simple and advanced) show minimal differences in ROC-AUC with or without bias mitigation. We could mention for example "Hispanic Native Hawaiian Female" in the case of LightGBM model.

- The overfitting observed in some small datasets (like hispanic asian female with perfect ROC-AUC (=1)) and underfitting in other groups (like hispanic asian male (=0)) demonstrate that the model may not generalize well when there are fewer samples or imbalanced data. There are some missing data for certain groups in the 25% batches (Hawaiian female), also the same intersectional group reach a value of 1 at ROC-AUC, indicative of overfitting.
- The differences in performance across intersectional groups suggest potential biases. The model seems to perform consistently better for "not hispanic white female" and "not hispanic white male", while groups like "hispanic asian male" and "hispanic american indian female" experience higher variability and lower performance with smaller dataset sizes.
- Interestingly, some groups (e.g., Hispanic Black males, lightgbm simple with and without bias mitigation) show improved performance metrics with the 50% dataset compared to both 100% and 25% datasets. This counterintuitive result suggests that more data doesn't always lead to better model performance for all subgroups, highlighting the complexity of intersectional fairness in machine learning models.
- Performance varies significantly across different demographic groups for all models. Some groups show consistent performance across models, while others vary greatly.
  - The model shows remarkable stability in performance for some groups, especially non-Hispanic White and Asian females and males. On the other hand, Hispanic intersectional groups (especially Hispanic Asian and Hispanic Native Hawaiian) show the highest variability across dataset sizes.
  - For some intersectional groups gender is a key factor. Within some ethnic groups (e.g., Hispanic American Indian, Hispanic Asian), there are notable differences in how male and female subgroups are affected by dataset size changes in some models.
- Performance metrics often change as the data percentage decreases from 100% to 25%, but not always consistently.

Regardless bias

Regardless bias mitigation is applied or not, this is, considering both scenarios and the full dataset, there are huge differences in model performance across the different intersectional groups. We noticed interesting facts: The following intersectional groups appear repeatedly as the groups with the best model predictions:

1. "hispanic american native female" (ROC-AUC: 0.8667, Log Loss: 0.5035)
2. "hispanic black male" (ROC-AUC: 0.8613, Log Loss: 0.5507)
3. "hispanic american native male" (ROC-AUC: 0.8589, Log Loss: 0.5369)
4. "hispanic asian female" (ROC-AUC: 0.8333, Log Loss: 0.4942)
5. "hispanic native hawaiian female" (ROC-AUC: 0.8000, Log Loss: 0.6544)

For further detail we also present the five groups with frequent low performance.

1. "hispanic asian male" (ROC-AUC: 0.5062, Log Loss: 0.7317)
2. "not hispanic american native male" (ROC-AUC: 0.6469, Log Loss: 0.6398)
3. "not hispanic asian female" (ROC-AUC: 0.6725, Log Loss: 0.5627)
4. "hispanic native hawaiian male" (ROC-AUC: 0.6923, Log Loss: 0.6674)
5. "not hispanic white male" (ROC-AUC: 0.6984, Log Loss: 0.5335)

Finally, we will show the list of the most affected intersectional groups by the application of bias mitigation:

1. Not Hispanic Black Male
    - Change in ROC-AUC: From 0.6841 (before) to 0.6841 (after)
    - Change in Log Loss: From 0.7727 (before) to 0.7772 (after).
2. Hispanic Black Male
    - Change in ROC-AUC: From 0.8282 (before) to 0.8228 (after)
    - Change in Log Loss: From 0.7024 (before) to 0.5874 (after).
3. Hispanic Native Hawaiian Male
    - Change in ROC-AUC: From 0.6356 (before) to 0.6518 (after)
    - Change in Log Loss: From 0.6870 (before) to 0.6871 (after).
4. Hispanic White Male
    - Change in ROC-AUC: From 0.6992 (before) to 0.7020 (after)
    - Change in Log Loss: From 0.7059 (before) to 0.6104 (after).
5. Hispanic American Indian Male
    - Change in ROC-AUC: From 0.8804 (before) to 0.8589 (after)
    - Change in Log Loss: From 0.7541 (before) to 0.6834 (after).

List with the least affected:

1. Hispanic Asian Female
    - Change in ROC-AUC: From 0.7667 (before) to 0.8000 (after)
    - Change in Log Loss: From 0.5663 (before) to 0.5010 (after).
2. Not Hispanic White Female
    - Change in ROC-AUC: From 0.6974 (before) to 0.6992 (after)
    - Change in Log Loss: From 0.5983 (before) to 0.4978 (after).
3. Not Hispanic White Male
    - Change in ROC-AUC: From 0.6791 (before) to 0.6889 (after)
    - Change in Log Loss: From 0.6100 (before) to 0.5089 (after).
4. Hispanic Black Female
    - Change in ROC-AUC: From 0.7467 (before) to 0.6932 (after)
    - Change in Log Loss: From 0.7521 (before) to 0.7674 (after).
5. Hispanic Native Hawaiian Female
    - Change in ROC-AUC: From 0.8600 (before) to 0.8600 (after)
    - Change in Log Loss: From 0.8298 (before) to 0.5357 (after).

4.2 Bias measurement. Disparate impact

We will assess disparate impact across different models and batches, beginning with the tables of comparison 4.3. Generally speaking, across all models, the application of bias mitigation did not lead to any noticeable improvement in fairness. The values of disparate impact(DI), statistical parity difference (SPD), and mean difference (MD) remain largely unchanged, suggesting that the bias mitigation applied was not effective in significantly reducing bias for either Logistic Regression or LightGBM models. As the variations between models are minimal, no meaningful conclusions can be drawn from these minor differences.

Model	Without Bias Mitigation			With Bias Mitigation		
	DI	SPD	MD	DI	SPD	MD
Logistic Regression Simple	0.9176	-0.0511	-0.0511	0.9176	-0.0511	-0.0511
Logistic Regression Advanced	0.9197	-0.0498	-0.0498	0.9197	-0.0498	-0.0498
LightGBM Simple	0.8924	-0.0823	-0.0823	0.8924	-0.0823	-0.0823
LightGBM Advanced	0.9055	-0.0762	-0.0762	0.9055	-0.0762	-0.0762

TABLE 4.3: Comparison of Disparate Impact (DI), Statistical Parity Difference(SPD), Mean Difference (MD). With and without Bias Mitigation. 100% batch

From the screening across the comparative tables 4.5 and 4.4:

- The model with the most unfavourable disparate impact is Logistic Regression simple, for the intersectional group "Hispanic american indian male", with a DI of 0.0393, before and after bias mitigation.
- The model with the most favourable DI is LightGBM, for the intersectional group "not hispanic asian female", with a DI of 1.2434 before bias mitigation and 1.2434 after bias mitigation.

When considering the analysis of all comparative tables, including different models and different dataset sizes, focusing on disparate impact we found:

- Although in most cases, Disparate Impact tends to increase when the batch size decreases, indicating that the fairness of the models gets affected badly. An example is the intersectional group.
- In certain cases the behaviour is the other way around and the DI increases as does the dataset size such as in the intersectional group.

In addition, the biggest variation of disparate impact across the different batches, considering bias mitigation cases, was registered in the intersectional group "hispanic native female", Linear Regression advanced model with a DI of 0 for the 50% batch to 1.7377 for the 100% batch, a difference of 1.7377.

On the contrary, the smallest variation across the different batches is for the intersectional group "not hispanic black female" with a maximum difference of only 0.0018 across the different dataset sizes.

Finally, we will list the 5 intersectional groups with the largest DI (fairest), after bias mitigation, in the full dataset size:

1. "not hispanic asian female" (Logistic Regression simple): DI = 1.2434
2. "not hispanic asian male" (Logistic Regression advanced): DI = 1.1852
3. "not hispanic white female" (LightGBM simple): DI = 1.0161
4. "hispanic asian male" (Logistic Regression simple): DI = 1.0149
5. "not hispanic native Hawaiian female" (LightGBM advanced): DI = 0.8891

As well as the 5 groups with the smallest DI (unfairest):

1. "hispanic american native male" (Logistic Regression simple): DI = 0.0393.
2. "hispanic native hawaiian female" (Logistic Regression simple): DI = 0.1075.
3. "hispanic native hawaiian male" (Logistic Regression simple): DI = 0.1511.
4. "hispanic black male" (Logistic Regression advanced): DI = 0.2873.
5. "hispanic black female" (Logistic Regression simple): DI = 0.403

Group	Model	DInoW	SPDnoW	DI	SPD
not_hispanic_black_female	LR Simple	0.4993	-0.3106	0.4993	-0.3106
	LR Advanced	0.5026	-0.3083	0.5026	-0.3083
	LightGBM Simple	0.5360	-0.3549	0.5360	-0.3549
	LightGBM Advanced	0.6223	-0.3047	0.6223	-0.3047
not_hispanic_black_male	LR Simple	0.4827	-0.3209	0.4827	-0.3209
	LR Advanced	0.4892	-0.3166	0.4892	-0.3166
	LightGBM Simple	0.5566	-0.3392	0.5566	-0.3392
	LightGBM Advanced	0.6380	-0.2921	0.6380	-0.2921
not_hispanic_native_hawaiian_female	LR Simple	0.8410	-0.0987	0.8410	-0.0987
	LR Advanced	0.7016	-0.1850	0.7016	-0.1850
	LightGBM Simple	0.8525	-0.1128	0.8525	-0.1128
	LightGBM Advanced	0.8891	-0.0895	0.8891	-0.0895
not_hispanic_native_hawaiian_male	LR Simple	0.7069	-0.1818	0.7069	-0.1818
	LR Advanced	0.7077	-0.1811	0.7077	-0.1811
	LightGBM Simple	0.6651	-0.2562	0.6651	-0.2562
	LightGBM Advanced	0.7393	-0.2104	0.7393	-0.2104
not_hispanic_white_female	LR Simple	1.0061	0.0038	1.0061	0.0038
	LR Advanced	1.0063	0.0039	1.0063	0.0039
	LightGBM Simple	1.0161	0.0123	1.0161	0.0123
	LightGBM Advanced	1.0071	0.0057	1.0071	0.0057
not_hispanic_white_male	LR Simple	0.8133	-0.1158	0.8133	-0.1158
	LR Advanced	0.8107	-0.1173	0.8107	-0.1173
	LightGBM Simple	0.8106	-0.1449	0.8106	-0.1449
	LightGBM Advanced	0.8271	-0.1395	0.8271	-0.1395

TABLE 4.4: Table 2 of Comparison of Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW) with bias mitigation, and without, Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW). 100% batch.

Group	Model	DInoW	SPDnoW	DI	SPD
hispanic_american_native_female	LR Simple	0.4446	-0.3445	0.4446	-0.3445
	LR Advanced	0.5008	-0.3094	0.5008	-0.3094
	LightGBM Simple	0.4508	-0.4202	0.4508	-0.4202
	LightGBM Advanced	0.5556	-0.3586	0.5556	-0.3586
hispanic_american_native_male	LR Simple	0.0393	-0.5960	0.0393	-0.5960
	LR Advanced	0.1574	-0.5222	0.1574	-0.5222
	LightGBM Simple	0.5101	-0.3748	0.5101	-0.3748
	LightGBM Advanced	0.4232	-0.4654	0.4232	-0.4654
hispanic_asian_female	LR Simple	0.7585	-0.1498	0.7585	-0.1498
	LR Advanced	0.6644	-0.2080	0.6644	-0.2080
	LightGBM Simple	0.6920	-0.2356	0.6920	-0.2356
	LightGBM Advanced	0.8020	-0.1598	0.8020	-0.1598
hispanic_asian_male	LR Simple	1.0149	0.0092	1.0149	0.0092
	LR Advanced	0.8367	-0.1012	0.8367	-0.1012
	LightGBM Simple	0.9199	-0.0613	0.9199	-0.0613
	LightGBM Advanced	0.8263	-0.1402	0.8263	-0.1402
hispanic_black_female	LR Simple	0.4030	-0.3704	0.4030	-0.3704
	LR Advanced	0.5110	-0.3031	0.5110	-0.3031
	LightGBM Simple	0.6863	-0.2400	0.6863	-0.2400
	LightGBM Advanced	0.6610	-0.2735	0.6610	-0.2735
hispanic_black_male	LR Simple	0.2873	-0.4422	0.2873	-0.4422
	LR Advanced	0.3834	-0.3821	0.3834	-0.3821
	LightGBM Simple	0.5565	-0.3393	0.5565	-0.3393
	LightGBM Advanced	0.5154	-0.3910	0.5154	-0.3910
hispanic_native_hawaiian_female	LR Simple	0.1075	-0.5537	0.1075	-0.5537
	LR Advanced	0.6454	-0.2197	0.6454	-0.2197
	LightGBM Simple	0.6972	-0.2317	0.6972	-0.2317
	LightGBM Advanced	0.6610	-0.2735	0.6610	-0.2735
hispanic_native_hawaiian_male	LR Simple	0.1511	-0.5267	0.1511	-0.5267
	LR Advanced	0.4034	-0.3697	0.4034	-0.3697
	LightGBM Simple	0.5310	-0.3588	0.5310	-0.3588
	LightGBM Advanced	0.2711	-0.5881	0.2711	-0.5881
hispanic_white_female	LR Simple	0.8293	-0.1059	0.8293	-0.1059
	LR Advanced	0.8319	-0.1042	0.8319	-0.1042
	LightGBM Simple	0.7933	-0.1582	0.7933	-0.1582
	LightGBM Advanced	0.8151	-0.1492	0.8151	-0.1492
hispanic_white_male	LR Simple	0.8133	-0.1158	0.8133	-0.1158
	LR Advanced	0.8107	-0.1173	0.8107	-0.1173
	LightGBM Simple	0.8106	-0.1449	0.8106	-0.1449
	LightGBM Advanced	0.8271	-0.1395	0.8271	-0.1395
not_hispanic_american_native_female	LR Simple	0.4925	-0.3149	0.4925	-0.3149
	LR Advanced	0.4482	-0.3420	0.4482	-0.3420
	LightGBM Simple	0.5689	-0.3298	0.5689	-0.3298
	LightGBM Advanced	0.7230	-0.2235	0.7230	-0.2235
not_hispanic_american_native_male	LR Simple	0.5874	-0.2560	0.5874	-0.2560
	LR Advanced	0.6154	-0.2384	0.6154	-0.2384
	LightGBM Simple	0.8308	-0.1294	0.8308	-0.1294
	LightGBM Advanced	0.8718	-0.1035	0.8718	-0.1035
not_hispanic_asian_female	LR Simple	1.2434	0.1510	1.2434	0.1510
	LR Advanced	1.2429	0.1506	1.2429	0.1506
	LightGBM Simple	0.9744	-0.0196	0.9744	-0.0196
	LightGBM Advanced	0.9711	-0.0233	0.9711	-0.0233
not_hispanic_asian_male	LR Simple	1.1845	0.1145	1.1845	0.1145
	LR Advanced	1.1852	0.1148	1.1852	0.1148
	LightGBM Simple	0.9671	-0.0252	0.9671	-0.0252
	LightGBM Advanced	0.9759	-0.0195	0.9759	-0.0195

TABLE 4.5: Table 1 of Comparison of Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW) with bias mitigation, and without, Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW). 100% batch.

4.3 Explainability. SHAP values

Before we proceed, as we will do later with the other types of SHAP graphics, we will remember some helpful points to understand the waterfall plots which we would use to provide a more granular presentation of our results:

- f(x): Represents the final model prediction (PB of loan denial) for the intersectional group we are analyzing. For example, if f(x) = 0.7, means the model predicts a 70% of PB of loan denial.
- E[f(x)]: Expected value. Represents the average prediction. For instance, if E[f(x)] = 0.3 means that the model predicts, on average, a 30% PB of loan denial across the whole dataset.
- Blue-red bars:
  - In our case, blue bars (-) decrease the probability of loan denial.
  - Red bars (+) increase the probability of loan denial.

4.3.1 Model evaluation: LightGBM advanced after bias mitigation

SHAP absolute coefficients

The absolute SHAP values show how much each feature contributed in total, to the model prediction. It is important to point out here that, features that might had a high impact on rare occasions but were meaningless most of the time, could seem relevant. The graph from figure: 4.1, display:

- Most important features: "hud median family income", "loan purpose name Home purchase", "loan to income ratio, minority population", "tract to msamd income".
- Observe that some of the features that appear to have some relevance on the model decision refer to specific intersectional groups (e.g., "ethnicity race sex not hispanic" or "latino white female"), although these are generally less influential than income or loan-related features.

SHAP bar plot

On the other hand, The SHAP bar plot, figure: 4.2, represents the average contribution, and highlights features that consistently impact predictions across many instances, hence those features with high impact on rare occasions will appear less important:

- As in the first plot, "hud median family income" and "loan purpose name Home purchase" are the most important features, but here has a different meaning, it reflects their consistent influence across the dataset, rather than just their total contribution.
- Demographic features (like ethnicity, race, and sex combinations) still appear, but with lower average impact, indicating they have smaller but steady contributions to model predictions. Let's pay particular attention to "not hispanic white, female and male".

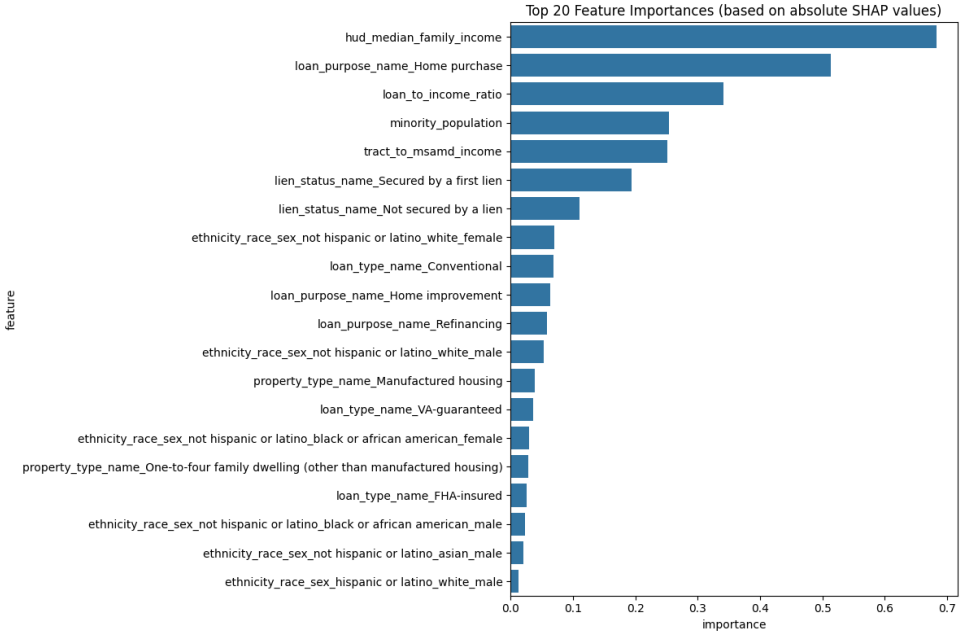

FIGURE 4.1: LightGBM absolute SHAP values

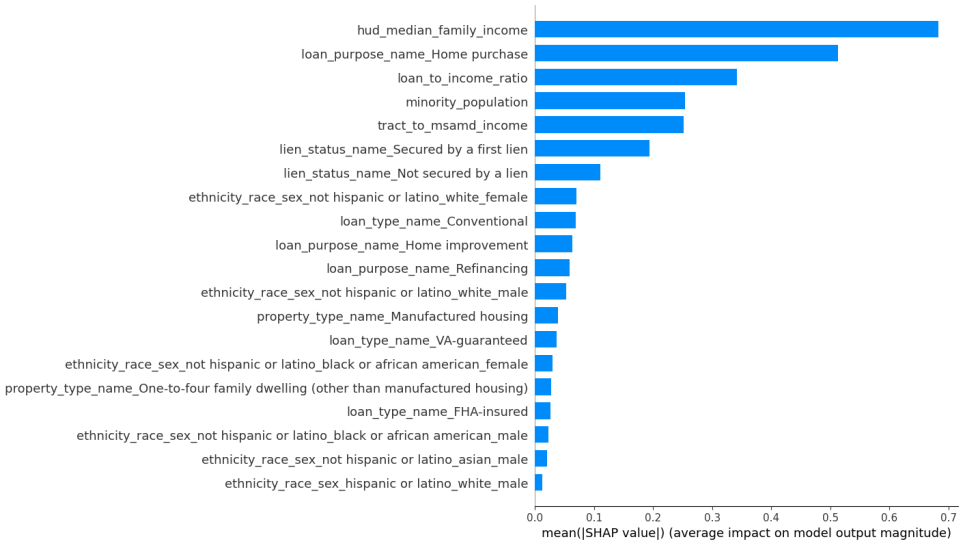

FIGURE 4.2: LightGBM SHAP bar plot

SHAP summary plot

SHAP summary plot, figure: 4.3,visualize the individual contribution to the model decision represented by each one of the dots we can appreciate. The spread of the dots shows the distribution of SHAP values for that feature. The width of the beeswarm shows how spread the SHAP values are for each feature across different predictions. Finally, two keays components to interpret the results:

- The magnitude of the impact (SHAP value) tells you how much a feature contributes to the prediction for an instance.
- The feature value (the instance value) shows whether the feature is high or low for that instance, which often influences whether the SHAP value is positive or negative.

From the chart we perceive:

- "hud median family income": Has the highest overall impact on predictions (together with ("loan to income ratio"). Higher incomes (red) generally increase SHAP values to contribute to a higher probability of loan denial, while lower incomes (blue) have the opposite effect. The widespread indicates varied impact across different instances.
- "loan purpose name Home purchase": Home purchases (red) tend to have a positive impact, and "Non-home" purchases (blue) a negative.
- "loan to income ratio": Higher ratios (red), negative impact. This suggests a higher risk. Lower ratios (blue) positive impact (lower risk)

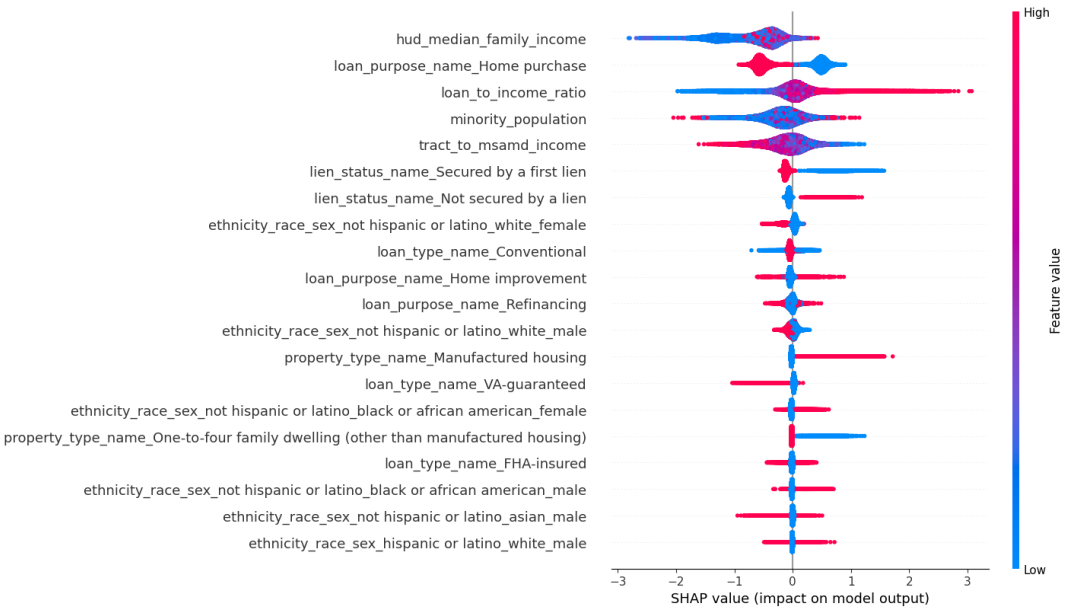

FIGURE 4.3: LightGBM advanced summary plot

4.3.2 Intersectional group evaluation

We will evaluate the 5 intersectional groups with the largest disparate impact (fairest) and the 5 intersectional groups with the lowest disparate impact (unfairest). For the intersectional

groups with the three fairest and unfairest disparate impact, we will show both, the corresponding summary plot and waterfall plot. Pay attention to certain intersectional groups such "not hispanic white, female and male" as they are present in the SHAP summary plots.

Three fairest disparate impact

Summary plot

From the evaluations of the following summary plots: 4.4, 4.5, 4.6 we witness:

- Factors such as income, loan purpose, and secured loans heavily influence the favourable outcomes for these intersectional groups.

Waterfall plot

From the analysis of the waterfall values displayed on figures 4.7, 4.8, 4.9:

- Features related to wealth seem to dominate the outcomes. Loan to income ratio, loan purpose(home purchase), median family income, and secured loan status consistently contribute to decrease the probability of loan denial.

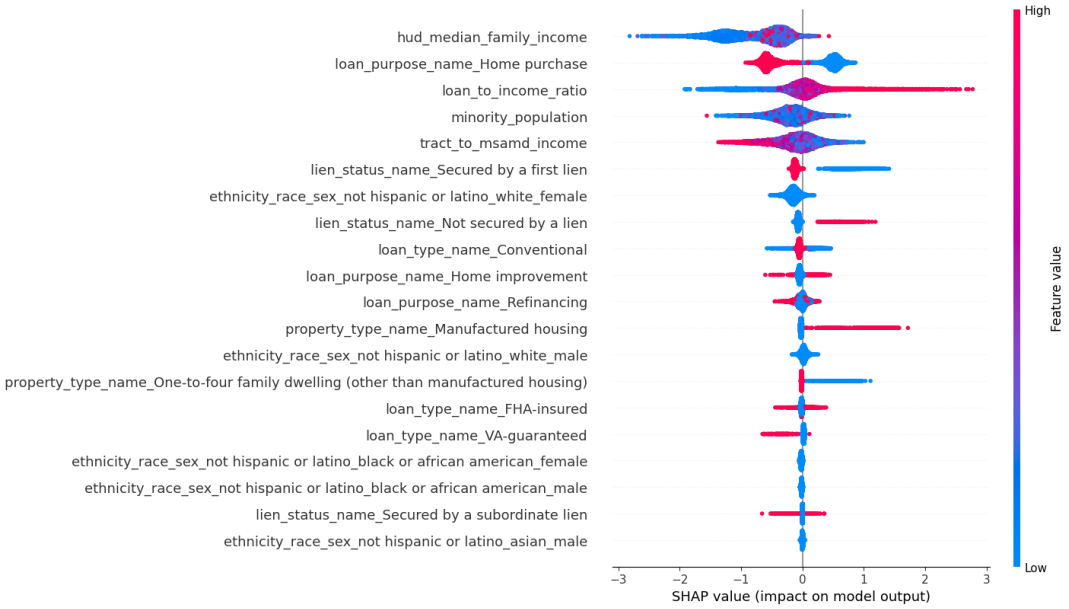

FIGURE 4.4: Fairest summary plot. Not hispanic white female

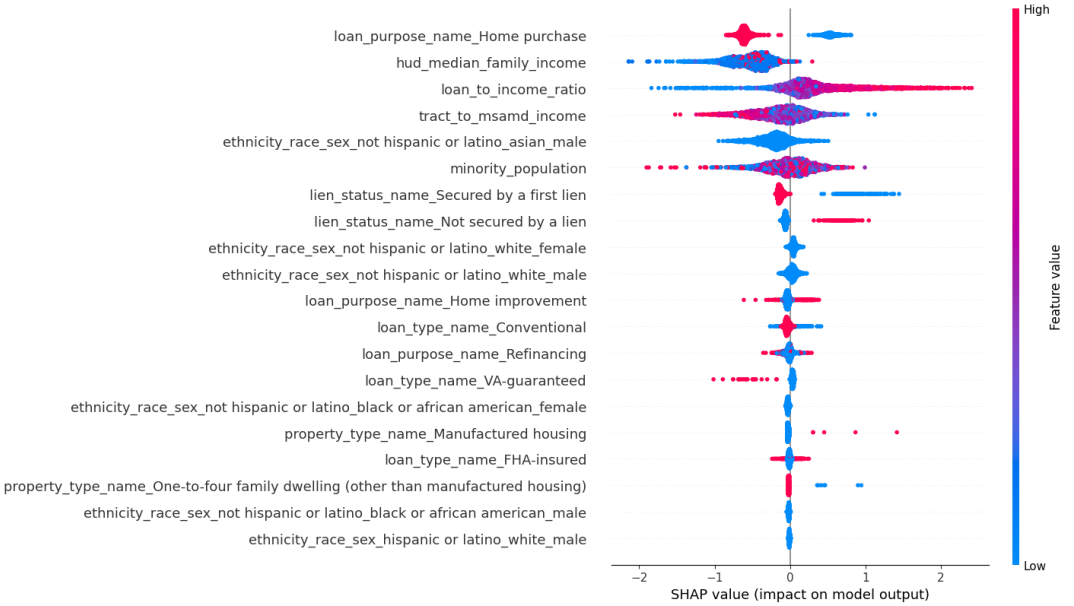

FIGURE 4.5: Fairest summary plot. Not hispanic asian male

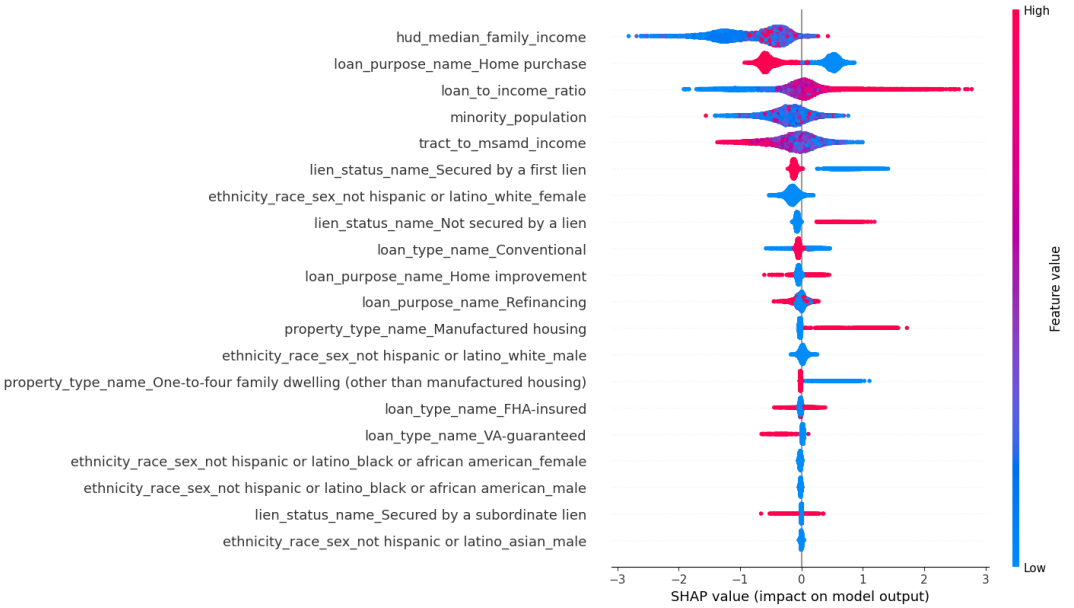

FIGURE 4.6: Fairest summary plot. Not hispanic asian female

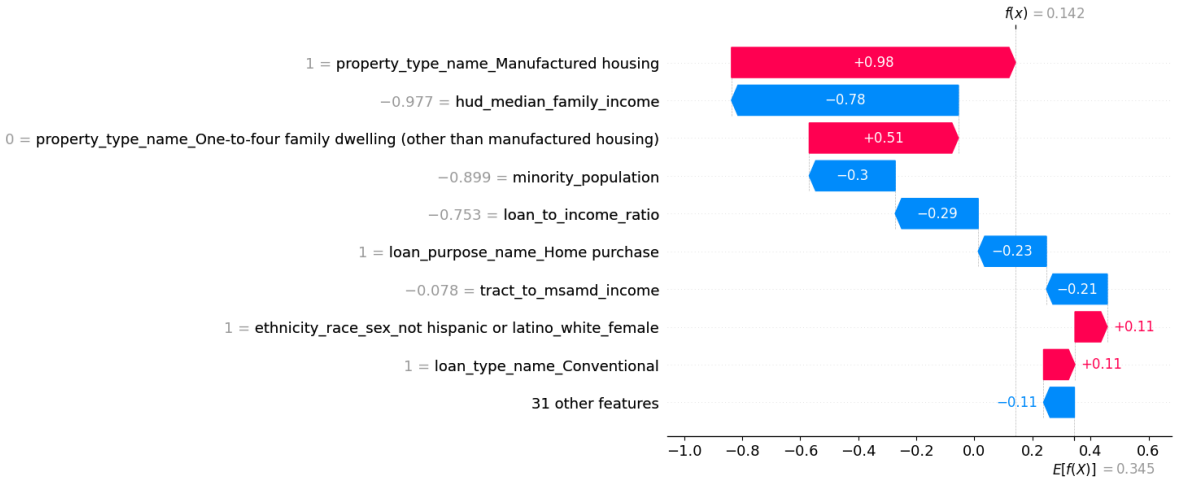

FIGURE 4.7: Fairest waterfall plot. Not hispanic white female

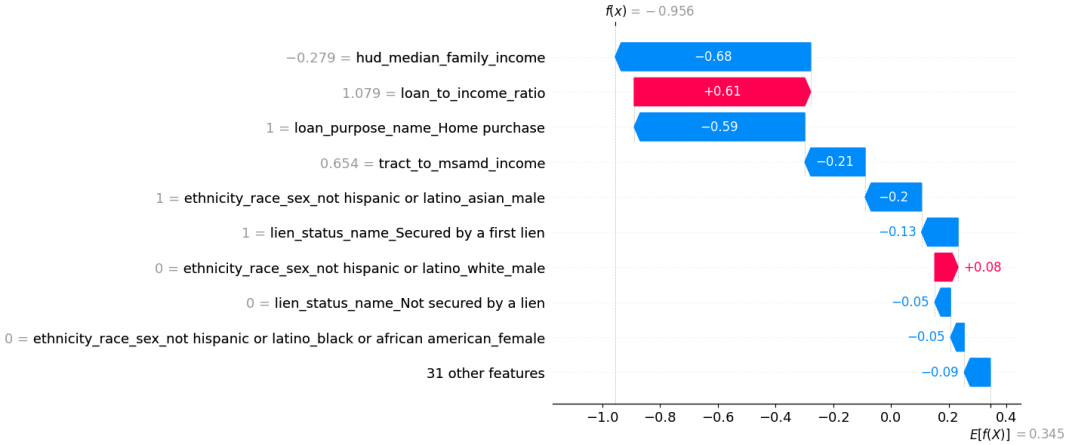

FIGURE 4.8: Fairest waterfall Plot. Not hispanic asian male

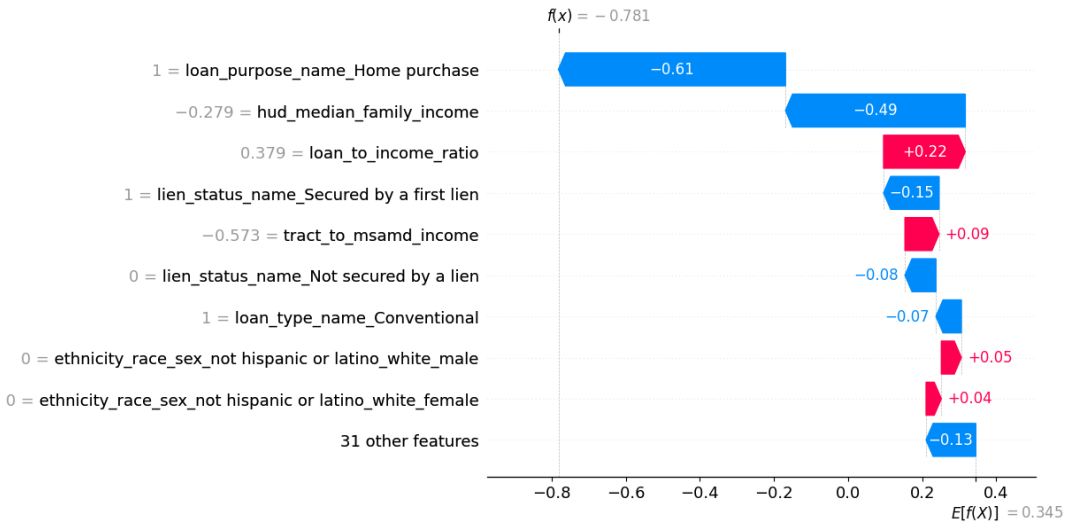

FIGURE 4.9: Fairest waterfall plot. Not hispanic asian female

Final review of both chart types:

- In both, wealth-related features play the most influential role in predicting loan denial. The more favourable are these the less the likelihood of loan denial.
- A key variable to mention is loan to income ratio which has a strong impact across the three different groups assessed.
- Belonging or not to a particular intersectional group affects the probability of loan denial.

Three unfairest disparate impact

Summary plot For the three intersectional groups with the unfairest disparate impact, the plots 4.10, 4.11, 4.12, these are our findings:

- Financial factors like those already mentioned are the key drivers of the model outcomes. Loan-income ratio, secured loans, median family income, and loan purpose carry a heavy weight on the loan denial predictions.
- Being part of some intersectional groups noticeably contributes to the model predictions. For example, "Hispanic or Latino black or African American males", or " Hispanic or Latino American Indian or Alaska Native males".
- property type (manufactured housing) and " one-to-four family dwellings" have an impact, particularly "manufactured housing" which is linked to a higher likelihood of loan denial.

Waterfall plot The three waterfall plots 4.13, 4.14, **??** show:

- Certain intersectional groups strongly increase the probability of loan denial such as "hispanic or latino black or African american male" and "hispanic or latino american indian or Alaska native male", even though their financial situation seems to be adequate, these groups, tend to face a higher risk of loan denial.
- The already mentioned variables, "high family income, "loan to income ratio", or home purchase loans remain as important features on model predictions.
- The model treats the "minority population" feature inconsistently, which can be a reflection of potential biases.

4.3.3 Final observations

From our perspective, these are the most outstanding facts:

- For the favourable groups, the model appears to prioritize financial health, treating demographic characteristics (ethnicity, race, sex) as secondary considerations. The fairer outcomes seem to be based on economic factors.
- For the unfavourable groups, demographic features are key drivers of loan denials, despite favourable wealth-related variables.

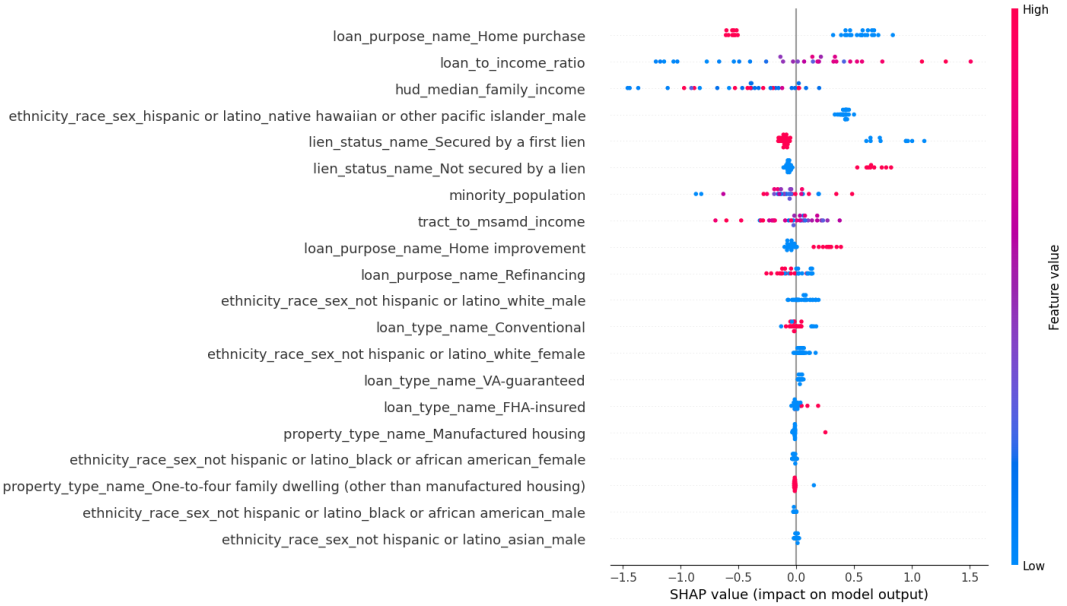

FIGURE 4.10: Unfairest summary plot. Hispanic hawaiian male

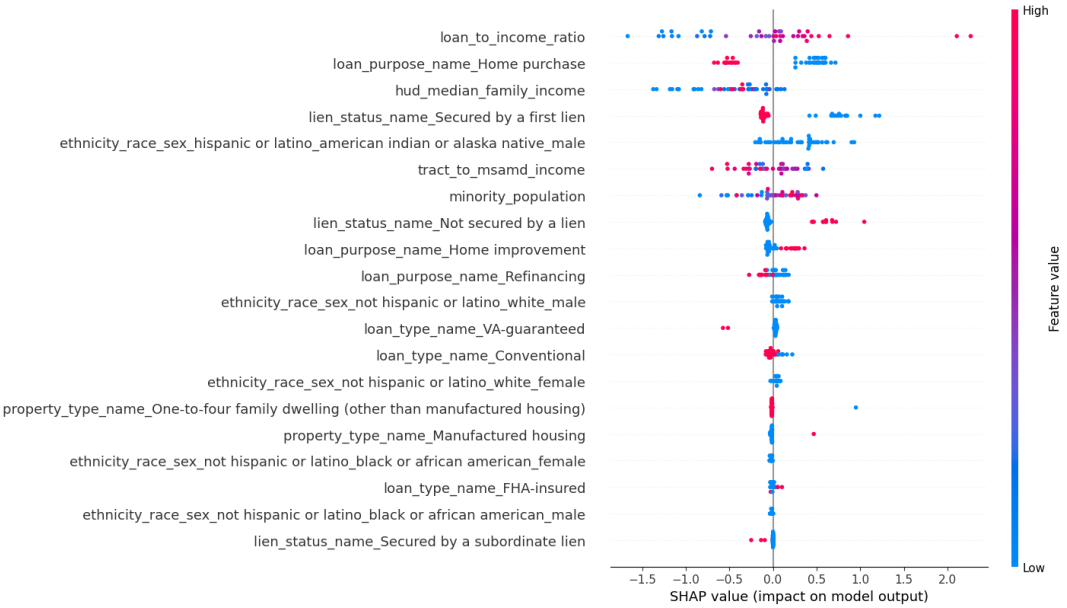

FIGURE 4.11: Unfairest summary plot. Hispanic native male

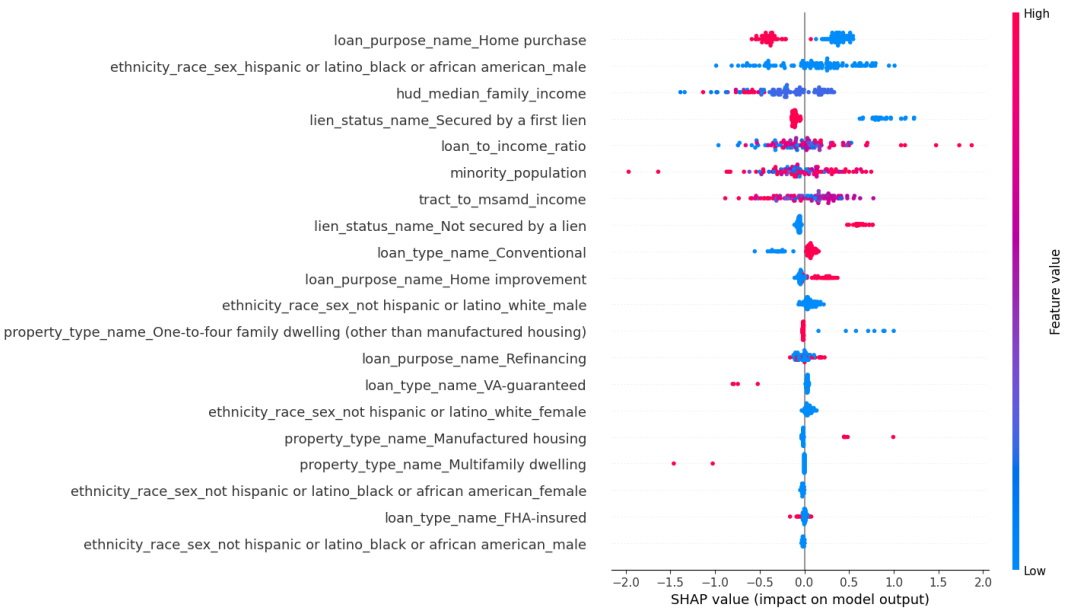

FIGURE 4.12: Unfairest summary plot. Hispanic black male

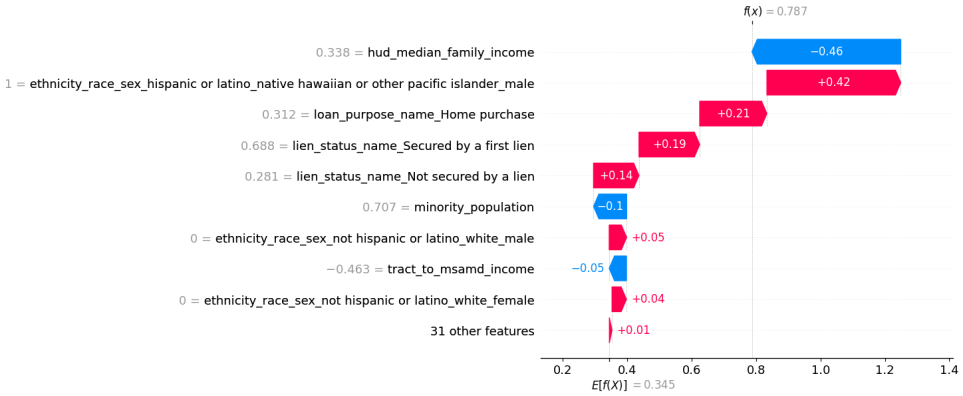

FIGURE 4.13: Unfairest waterfall plot. Hispanic hawaiian male

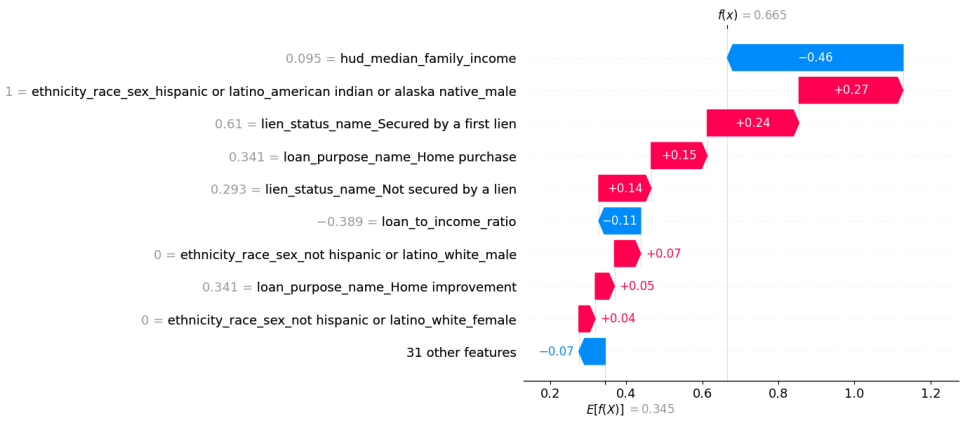

FIGURE 4.14: Unfairest waterfall plot. Hispanic American native male

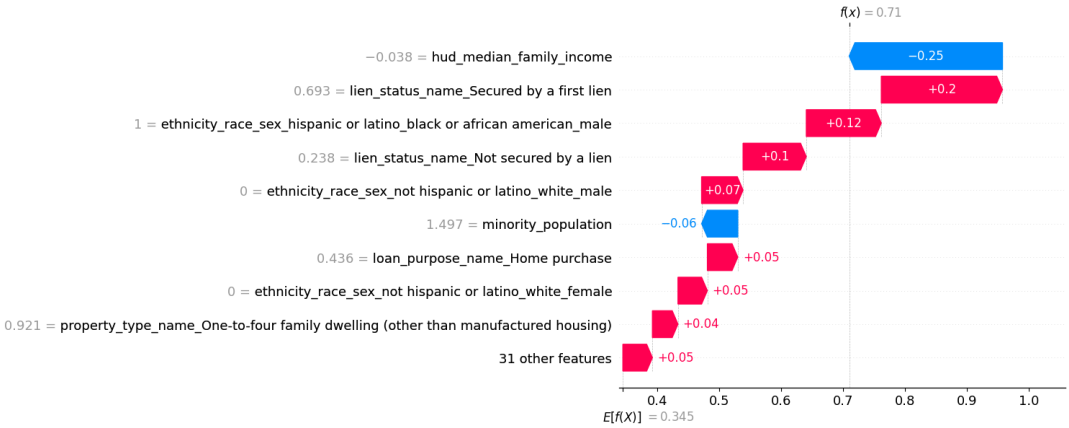

FIGURE 4.15: Unfairest waterfall plot. Hispanic black male

Chapter 5

Discussion

Our primary objective was to assess the performance of Logistic Regression and LightGBM models in predicting loan denial, as well as disparate impact variability across various intersectional groups. Additionally, we aimed to understand the effect of bias mitigation techniques on these models and how they perform across different dataset sizes (100%, 50%, and 25%).

5.1 Model performance. ROC-AUC and Log Loss

Generally speaking, the stability shown by every model on each batch is out of expectations. Let's remember, that for both, logistic regression and lightGBM, we are analysing 4 models, 3 batches, and a total of 24 models' outputs. To appreciate the meaningless of the variations we present the following table:

Model	**ROC-AUC Range**	**Log Loss Range**
Logistic Regression	0.7019 - 0.7095	0.6237 - 0.6296
LightGBM	0.6660 - 0.7152	0.5332 - 0.6665

TABLE 5.1: Metrics range for Logistic Regression and LightGBM

Now, questions arise... Why? What is behind this robust performance? Furthermore, why does bias mitigation (re-weighting) not affect model outputs?

While it is challenging to assert, evidence suggests that the lack of significant difference between models seems to point towards the still fresh reliability of logistic regression in our study, and it looks like sophistication in architecture is not always the answer to improving results in the prediction of loan denied/approved. Yet another question surge, What about bias mitigation? The previous statement could have passed as valid if there were some differences between prior and post, application of re-weighting. However, this is not the case, leading us to the next appreciation, which is the fact that both models are equally insensitive to re-weighting since both produce extremely close results before and after the bias mitigation. Then, if both models with different working ways yield similar results and behaviour across the different dataset sizes, perhaps the root of the cause lies in the data input.

5.1.1 Analysis of results

The data itself

Looking at the data itself, the report provided by (*Controlling machine-learning algorithms and their biases — mckinsey.com* n.d.) points out the fact that re-weighting might have little or no effect on bias mitigation on certain scenarios. The text explains that systemic bias embedded in the dataset could not be handled adequately by the aforementioned technique. For example, socio-economic factors that have a strong influence on loan denial as we witnessed earlier on our SHAP waterfall values could be responsible for the ineffectiveness of re-weighting.

The preprocessing followed

In our examination of the stages involved in preparing the input data for our models, we acknowledge that these steps might have affected the potential biases:

- The application of Log transformation and scaling standardised features would have levelled the playing field for both models.
- SMOTE is used to balance class distributions by creating synthetic examples for underrepresented groups. This process already addresses some degree of class imbalance, indeed, we measured biases before and after the application of smote and there is a difference. As a result, is reasonable that applying additional bias mitigation techniques might not show much improvement.
- Correcting for skewness helps to normalize the distribution of features. This process reduces the impact of outliers and ensures that the model treats highly skewed features more evenly. Therefore since the skewness correction was applied uniformly across groups, it could have normalized the treatment of intersectional groups as well. Hence, re-weighting would find fewer disparities to address, leading to similar results before and after mitigation.

5.1.2 Answers to our initial questions for models' performance

Which model seems to be more robust to the different datasets?

The model that seems to be more robust is logistic regression, even though both models present stable performance.

Even though the model seems to produce stable results before and after bias application, when looking at the results for the different intersectional groups, across the different batches the difference is immense for example, "hispanic asian male" obtained a ROC-AUC of 0.5062 whereas "hispanic american native female" obtained a ROC-AUC of 0.8667.

Some intersectional groups, particularly the ones that include Hispanic and Native Hawaiian/Pacific Islander, show significant variations in model performance across different dataset sizes. This suggests that the model's reliability and fairness may be inconsistent for these groups depending on the amount of data used.

Which five intersectional groups have the best model performance?

As we saw previously the following groups present the best performance overall:

1. "hispanic american native female"
2. "hispanic black male"
3. "hispanic american native male"
4. "hispanic asian female"
5. "hispanic native hawaiian female"

Which five intersectional groups have the worst model performance

The following groups receive the less favourable outcomes.

1. "hispanic asian male"
2. "not hispanic american native male"
3. "not hispanic asian female"
4. "hispanic native hawaiian male"
5. "not hispanic white male"

Regarding the previous lists, it is interesting to note that the model produces one of the best outcomes for an intersectional group considered as a "minority group" even when they represent an insignificant proportion of the total applications. When we look back at figure 3.3, we see for example that "hispanic hawaiian female" category, with around only 0.1% on the loan applications, appears on the top list.

5.2 Disparate Impact

In Chapter 4, we observed that the application of bias mitigation did not result in any significant change in disparate impact across the models. This is consistent with our earlier observation that the re-weighting technique applied had minimal effect on the outcomes. At this stage, we confirm that both models, logistic regression and LightGBM produced very close results before and after bias mitigation in both, disparate impact and models' metrics (ROC-AUC, Log Loss). Showing such similar results before and after bias mitigation is an additional argument to support the ineffectiveness of re-weighting to diminish biases in our model outputs.

5.2.1 Answers to the initial questions for disparate impact

which model seems to have the most unfavourable disparate impact?

The model with the most unfavourable disparate impact seems to be logistic regression simple, 25% batch, in particular for "hispanic american native male" group with a disparate impact as low as 0.0393. Focusing on this result, the disparate impact is in line with our previous expectations from our literature review which already pointed out the discrimination of machine learning models towards minorities due to the lack of data for these communities in comparison with the "privileged" group, "not hispanic white male, female". Overall, this fact, the scarcity of data for the fair assessment of minorities seems to be more noticed when the data set is reduced.

Which model seems to have the most favourable disparate impact?

The model with the most favourable disparate impact is linear regression simple, dataset of 100% for the intersectional group "not hispanic asian female", with a disparate impact of 1.2434 before bias mitigation and 1.2434 after bias mitigation. However, why does this intersectional group have a favourable DI, if it is not one of the closest to the "privileged" group in terms of loan applications? it is interesting that the simple logistic regression model performed favourably for this group. This raises the question of whether the model's simplicity might have played a role in its ability to maintain fairness or perhaps has been the preprocessing done in the early stages that benefitted this model in particular. Seeing that each of the stages was mainly tailored to logistic regression given that lightGBM is a more flexible model.

Is the disparate impact affected by the different dataset sizes?

Our analysis reveals that disparate impact is influenced by dataset size, with the smallest dataset (25%) resulting in a higher variability and poorer fairness across the intersectional groups. As the dataset increased to 50%, the models demonstrated more stability in fairness outcomes, it seems that particularly in LightGBM variations. This stability is evidenced by the reduced fluctuation in the disparity impact values when the dataset is larger. It is clear that a larger dataset provided more representative samples of minority groups, and improved their assessment, this fact was reflected in our results when we saw in some cases a disparity impact of 0.03... and a ROC-AUC of 0 or even 1 (overfitting), the impact of data set size in fairness varied by model and group.

In addition, although the differences in disparate impact between Logistic Regression and LightGBM are not always large, LightGBM generally shows more stable fairness outcomes across different intersectional groups, particularly as dataset size increase

Which five intersectional groups have the most favourable disparate impact?

This question answered earlier revealed that the group with the most favourable disparate impact after bias mitigation "not hispanic asian female" is not the group next to the privileged group in terms of loan applications (see figure 3.3, which is interesting, given that a disparate impact higher than 1 means that the model would likely produce more favourable outcomes for these groups than for the reference group (privileged group; "not hispanic white male"). It is possible that models produced better results for minority groups due to larger datasets.

Which five intersectional groups have the most unfavourable disparate impact (full dataset)?

This question was already answered in the past chapter 4. The lowest disparate impact value was 0.0393 for 'hispanic american indian male' in the logistic regression simple model, indicating that this group was disproportionately less likely to receive favourable outcomes compared to the reference group. This pattern persisted across both models, logistic regression and lightGBM, suggesting that biases might be rooted in the dataset influencing the models' performance. However, this intersectional group has one of the lowest application proportions, therefore the bad results in disparate impact could be due to insufficient data available for these groups.

Finally, it's worth mentioning that we observed changes in both, before and after the application of SMOTE, as well as before and after the application of bias mitigation. However, the models produced practically the same results with or without bias mitigation.

5.3 Explainability with SHAP values

5.3.1 Most influencing variables

In general, which variables influenced the most on the outcomes?

To discuss the SHAP values we are using the results for the lightGBM advanced model.

In our quest to understand, which features most influenced the model's predictions, we conducted a SHAP analysis. The SHAP plots offer three different perspectives: absolute SHAP values, average SHAP values, and a SHAP summary plot. These plots allow us to assess both the magnitude and direction of feature importance. Across all three perspectives, several key features consistently emerged as the most influential

In summary, the analysis shows that the top three features—"hud median family income", "loan purpose name Home purchase", and "loan to income ratio"—consistently ranked as the most influential across all SHAP plots for the model lightGBM advanced. This suggests that the model heavily relied on financial stability indicators and the purpose of the loan when making decisions providing some light on the fairness and transparency of the loan approval process.

5.3.2 Most favoured intersectional groups

For the three intersectional groups with the most favourable disparate impact. Which are the variables that influenced the most on these outcomes?

Across the three intersectional groups with the most favourable disparate impact The most influential features were consistently "hud median family income", "loan purpose name Home purchase", and "loan to income ratio", reflecting the model's reliance on financial stability and loan characteristics in making predictions. However, the impact of these features varied by group, with the loan-to-income ratio contributing positively for some groups and negatively for others, depending on their specific circumstances.

Additionally, the model's sensitivity to geographic and demographic factors, such as minority population and tract income, suggests that broader socioeconomic conditions significantly influenced predictions, raising concerns about geographic bias. Moreover, the presence of other intersectional groups playing a decisive role in the model predictions supports these concerns about potential biases.

In the observation of f(x) values, we noticed "not hispanic white female" applicants had a moderate probability of loan denial, reflected in a positive f(x) value of 0.142, indicating that while the risk was not high, certain features (like loan-to-income ratio and geographic factors) contributed to a slight increase in the denial risk. On the other hand, "not hispanic asian female" group had a much lower risk of denial f(x) value of -0.78, suggesting that the model predicted a stronger likelihood of loan approval for this group.

The fact that these three groups had the most favourable disparate impact suggests that the model was generally fairer to them compared to other groups, offering higher chances of

loan approval. However, the specific feature impacts reveal that fairness does not guarantee uniformity each group experiences the influence of features differently.

Demographic impact raises important questions about fairness, are these demographic features being used as proxies for other variables (e.g., socioeconomic status) or are they introducing unintentional biases into the model? For example, certain demographic groups may face systemic disadvantages in lending decisions, and the model may reflect these patterns.

Another interesting observation was how badly was punished for the intersectional "not hispanic white female" when the "property type was manufacturing house".

5.3.3 Most unfavoured intersectional groups

For the three intersectional groups with the least favourable disparate impact. Which are the variables that influenced the most on these outcomes?

In our analysis of the three intersectional groups with the least favorable disparate impact "hispanic native hawaiian male", "hispanic native american indian or alaska native male", and "hispanic black male", we found key information influencing on the likelihood of loan denial:

The most noticeable event is that demographic variables played a disproportionately large role in increasing the likelihood of loan denial. This was evident for "hispanic black male" where its intersectional group added +12 points towards the likelihood of the loan being denied and the "loan income ratio" wasn't even present among the most important features.

We found a strong influence of manufactured housing on denial prediction, which reflects broader economic and social disparities, as this type of housing is often associated with lower-income applicants and historically underserved communities. Furthermore, the combination of property type and demographic factors could further exacerbate the disparities for minority groups who are indeed more likely to seek loans for manufactured housing.

Besides that, this finding may also reflect geographic biases as manufactured housing is more prevalent in certain regions with lower median incomes and higher minority populations.

5.3.4 Summary

- The model leans heavily on financial factors for loan denial decisions, but demographic traits still have some influence, even for groups with favourable outcomes. This is where the fairness concern lies: ideally, demographic factors should have no role if the model were entirely fair.
- While financial factors such as loan-to-income ratio and income are critical drivers across all groups, the model penalizes these factors more heavily for the least favourable groups. For the most favourable groups, these financial indicators reduce the likelihood of denial, demonstrating a disparity in how financial risk is assessed based on group membership.
- Demographic features have a stronger impact on the least favourable groups, often contributing more to denial risk than financial factors. In contrast, demographic characteristics have a neutral or mitigating effect on the most favourable groups, suggesting that the model treats certain groups as inherently riskier.

- Manufactured housing consistently increases denial risk for the least favourable groups, while geographic factors such as tract to msamd income show inconsistent effects. These findings highlight the potential for geographic bias and the model's reliance on economic and regional factors to influence decisions.

Chapter 6

Evaluation, Reflections and Conclusions:

6.1 Evaluation and project achievements

6.1.1 The origin

The primary objective of this research was to assess the probability of loan denial through various machine learning models, with a particular focus on understanding how these models performed across different intersectional groups and identifying potential biases in their predictions. The key objectives were:

- Evaluation of model performance in terms of predictability, using the metrics ROC-AUC and Log Loss.
- Widely measure disparate impact across the different models and dataset batches.
- Mitigation of biases with the application of re-weighting
- Explainability of model decisions through the use of SHAP values.
- Assessment of the trade-off between bias mitigation and model performance.

The ultimate goal was to provide a comprehensive analysis of how model performance and fairness intersect in predicting loan denials, with the intention of uncovering systemic biases and identifying which groups are most affected by loan denial outcomes, investigating potentially biased decisions.

6.1.2 Evaluation of methodology

Model performance

Initially, we built a simple logistic regression model which would serve as a baseline model. The expectation was that by applying grid search and parameter tuning we would improve the model performance, nevertheless, as we were adding parameters and building the more sophisticated version the results obtained for the 100% of the data didn't vary significantly. The same phenomena would be seen when we built the different versions of lightGBM, it was even more surprising for this model as according to previous papers such as (Suhadolnik, Ueyama, and Da Silva, 2023) we reviewed we had higher expectations on lightGBM. The results were that in terms of performance, the variations between logistic regression and

lightGBM were minimal. There were only variations in models' performance in smaller datasets and those outcomes are for the worst.

This event could be explained by many factors. According to the previous citations in Chapter 5, we believe that the fact that the same preprocess data was used in both models has something to be with that. We implemented an exhaustive preprocessing focused mainly on logistic regression requirements. Theoretically, this wouldn't have affected to the performance of the more versatile model, lightGBM. However, the stability ensured through the preprocessing, may have masked some of the complexities that a model like LightGBM is designed to capture and take advantage of.

Model performance were so stable that not even the application of bias mitigation changed their results when assessing the full dataset size.

Finally, the predictions made by our lightGBM models were far from what we expected from other pieces of work such as that achieved by (Ponsam et al., 2021) with an AUC score of 0.90 for lightGBM.

Disparate impact and bias mitigation

The application of bias mitigation (re-weighting) was a failure. For the models (100% dataset) we obtained exactly the same results with and without bias mitigation. Hence, while did not affect negatively to the model outputs in terms of ROC-AUC and log loss, it also did not mitigate the disparate impact faced by certain intersectional groups. A revision of the previous steps given before the application of biased mitigation should be done.

In this aspect, we already proved that a potential reason for the inefficient application of bias mitigation can be due to the use of SMOTE. The use of SMOTE combined with bias mitigation might overlap (Popoola and Sheppard, 2024).

In addition, the data preprocessing could have introduced existing biases into the model.

SHAP values

Despite logistic regression is naturally interpretable through its coefficients, we utilized SHAP values on both models to:

- Create a comparative interpretability framework across both models, logistic regression and lightGBM. This facilitated the direct comparison of feature contributions in each model although we only considered lightGBM in our final evaluation.
- Evaluate how specific features affect specific intersectional groups.
- It also was key in evaluating the disparate impact by understanding how the contributions made by the features were different across the intersectional groups.

Dataset

Considering the starting point smartly would be a key point in favour since the beginning of the project. The dataset used in this work was from New York City, the reason was mainly its large population together with its cultural diversity hence these factors would help us since the beginning in our goals of assessing biases and having enough data for assessing the different intersectional groups. The results highly likely wouldn't have been the same if, for example, we had chosen Charleston, in West Virginia (US) where the Hispanic population is less than 1.1% whereas in New York City is about 19%, source: DataUSA.io

6.2 Reflections

6.2.1 Challenges

Massive dataset

The initial dataset was massive with over 400k rows. As mentioned I had to chunk it and iterate through to reduce the volume by deleting the most obvious insignificant variables. Later, the dataset still was big enough for delaying eternally, so we opted to run our models on Google collab, and also made some attempts on Amazon AWS.

Average SHAP values plot cut off

The use of SHAP was full of setbacks, the most remarkable was that the average SHAP value plot, for some reason was cut off in some models. We debugged the issues for days without finding any solution, indeed some of these kinds of charts in our reports are still cut off. However, in our case, since we are assessing fairness at an intersectional level, the waterfall plot together with the the absolute coefficients plot used as a guide were enough to reach sound conclusions.

Mirror on missing data

There was missing data for relevant columns. These missing data were perfectly correlated with "census-related" columns. The solution we found was to create new columns to mirror the missing data. The null cells in the original column were filled out with different imputation strategies.

6.2.2 Assumptions and limitations

Subestimated intersecional groups

When we started this project, we assumed we would obtain similar model performance to other pieces of work however we obtained poor results in the end. We subestimated the influence of the assessment on intersectional groups. Let's remind we merged three columns into one, "ethnicity-race-sex".

Bias identification vs mitigation

The challenge of bias in predictive models extends beyond mere identification. While recognizing the reasons behind model decisions is crucial for spotting biases and underlying issues, this awareness alone doesn't rectify the problem. The core dilemma lies in how these models often reflect real-world inequalities that can't be easily changed through algorithmic adjustments.

6.2.3 Interpretations and reflections

Dataset

During our dataset selection, we appreciated how the economic cycles enormously affect to the number of loan applications and there is a high probability that also would affect the outcomes. for the US we found that prior to the 2008 crisis, in 2007 the nationwide number of applications was over one million and then decreased drastically to a minimum of 2014

389,279 (check this inf) according to after 2014 from 2015 to 2017 applications were between 400k and 500k Then during covid time in 2020 and 2021 we also detected a massive disparity with respect to other years. We selected data before COVID-19 as data size/applications varied drastically (circa 30% up) from 2019 to 2020 and 2021 even in posterior years are different pattern

SHAP values

From the SHAP values, we also observed how the different variables played different roles (and impacts), depending on the intersectional group.

The SHAP analysis consistently reveals that financial factors such as loan-to-income ratio, secured loans, median family income, and loan purpose are the primary drivers of loan denial predictions across all groups. However, these factors alone do not fully explain the disparities seen in some intersectional groups. In particular, the demographic characteristics of Hispanic or Latino black or African American males and Hispanic or Latino American Indian or Alaska Native males noticeably contribute to their higher likelihood of denial, even when their financial profiles appear adequate.

This observation leads us to a critical point: the intersection between financial and demographic features suggests that the model may be implicitly assigning higher risks to these groups, reflecting historical lending biases. The combination of demographic features and high loan-to-income ratios or home purchase loans creates a compounded effect, amplifying the likelihood of loan denial beyond what financial indicators alone would suggest. This intersectional impact highlights the complex nature of potential biases within the model and the need for a more nuanced understanding of how different factors interact in loan denial predictions.

Law vs companies

From our research into the regulatory framework, we conclude that the relationship between law and credit risk companies resembles a game of "cat and mouse" where credit risk companies using the latest technology consistently stay several steps ahead. Where the mouse (AI companies) always seems to outsmart the cat (regulatory bodies). This dynamic highlights the ongoing challenge of creating effective, proactive legislation in a rapidly evolving technological landscape.

Economic impact

Our previous assumption about the economic impact of the results was potentially quite limited. Before our analysis, we didn't anticipate the actual distribution of minorities affected by discrimination. While there are many intersectional groups affected, their relevance to the overall picture is relatively insignificant compared to the groups in the upper 4% of total applications.

Relationship between low applications, poor performance, higher biases

A clear pattern emerges after analyzing the intersectional groups with the most unfavourable and favourable disparate impact. Centring our attention on the most unfavourable groups such as "hispanic native american indian or Alaska native male", consistently demonstrated both high disparate impact (indicating bias) and poor model performance (as reflected in lower ROC-AUC and/or higher Log Loss). These groups are also underrepresented in the

dataset, which likely contributed to the model's inability to accurately predict loan denial outcomes for them. The opposite effect is seen in the most favourable intersectional groups.

6.3 LESSONS LEARNED

Bias

A key takeaway from this experience is the difficulty to mitigate biases in scenarios close to the real world, even more in our particular case where we narrowed the communities into intersectional groups. Even after applying one of the most innovative methods (Bellamy et al., 2018) in bias mitigation, the model continued showing exactly the same disparate impact values underscoring the need for more advanced fairness strategies. Peha

Understanding complex data

Data must be questioned in depth. During our preprocessing, we found many inconsistencies that when we investigated them turned out to be meaningful. One of the most impactful in our final results was the case of "manufactured housing".

We found that the dataset contains 261 missing entries in "census-related" columns, which represents an insignificant proportion (approximately 0.001) of the total records. However, upon closer examination, an interesting pattern emerges. All these missing values are associated with "denial" cases, and their distribution across different demographic groups is noteworthy.

Surprisingly, the intersectional group affected the most by these missing values is "not hispanic White, men and women". This is particularly significant because this group already has the lowest denial rate. Consequently, if we were to remove these missing entries from the analysis, it would further accentuate the disparity with the rest of the intersectional groups.

Standard practices

We did what we had to do yet we didn't obtain the expected results. In our data preprocessing, we cleaned the data properly, excluding outliers, addressing null values etc. Later, created the binary column, split the data addressing potential data leakages, and applied feature engineering (Log transformation / Scaling / one-hotencoding) yet our models performed poorly.

"Best standard practices on data preprocessing don't guarantee better results".

It would be done differently

- One of the key learnings from this project is that bias mitigation through re-weighting did not produced the expected results. If we were to start again, we would invest more time beforehand exploring a broader range of bias mitigation techniques, such as adversarial debiasing or fair representation learning, rather than solely relying on re-weighting.
- Creating smaller intersectional datasets, perhaps considering first 2 columns instead of 3 at once and experimenting with different batch sizes early on might have exposed fairness issues sooner, allowing for better handling of biases across the groups.

- A key finding was that logistic regression and LightGBM performed similarly, likely due to the exhaustive preprocessing. If we started again, we would experiment more with feature engineering, potentially exploring new features or even just applying raw data from the columns as input on the models. would have created a very simple model and would have evaluated every change on the features list used.

6.4 FUTURE WORK

Fixing bias mitigation

First of all, any future work should focus on recreating a scenario where the bias mitigation could have any meaningful effect on the models' results so we could complete adequately the unachieved goal of this work of assessing the trade-off between models' predictability and fairness. In general, a similar work where the effect of the bias mitigation would be assessed on the model performance could be the object of interest. Providing a broader report across the differences between the different dataset sizes would be like an X-ray of the existence of hidden biases. Perhaps using a less preprocessed data input. Concerning this, the recommendation, according to the experience from this work would be to focus on one single model first when executing the whole process, then implement it to the others.

Profit vs intersectional biases

Like (Kozodoi, Jacob, and Lessmann, 2022) mention many papers ignore the profitability part when addressing fairness. This author provides a method to include the profitability side against fairness which in combination with intersectional biases has the potential to become a unique piece of work in a barely explored "niche".

Economic variables vs intersectional biases

Taking into account or linking any economic variable to metrics like the model predictability would be quite interesting leveraging the different datasets available from the HMDA or any other similar to compare fairness, bias mitigation, and model performance with the different economic cycles. There are data for periods related to economic crises like the one in 2008, or periods of adversity like the Covid years, indeed, when we were selecting our dataset, we witnessed huge differences in regard to the number of loan applications (dataset size in both, Mb and by number of records).

6.5 Final conclusions

This study set out to explore the performance of Logistic Regression and LightGBM models in predicting loan denial, focusing on fairness, bias, and explainability across various intersectional groups. A critical objective was to assess the impact of bias mitigation techniques, specifically re-weighting, and how dataset sizes influenced model performance and fairness.

The results demonstrated that, overall, both, Logistic Regression and LightGBM performed consistently across the different datasets, with minimal variation in ROC-AUC and Log Loss. However, variations emerged when analyzing performance across different intersectional groups. For instance, the "hispanic qsian male" group consistently showed one of the worst performances, while the "hispanic american native female" group exhibited one of the best.

Despite applying bias mitigation techniques, the models showed limited improvements in fairness as measured by disparate impact, with minimal differences before and after re-weighting. These findings suggest that the systemic biases embedded in the data, particularly geographic factors, were not effectively addressed by re-weighting alone. Preprocessing steps, such as SMOTE and log transformation, further contributed to balancing the data, potentially reducing the impact of re-weighting.

The SHAP analysis revealed the most influential features in loan denial predictions, with financial indicators such as "HUD median family income," "loan purpose (home purchase)," and "loan-to-income ratio" playing the most significant roles. However, demographic features, still had some influence on the model's decisions, raising concerns about fairness, especially for certain minority groups.

In conclusion, this project highlights both the strengths and limitations of using machine learning models like Logistic Regression and LightGBM for predicting loan denial, particularly in addressing fairness. Bias mitigation techniques, such as re-weighting, were shown to be limited in reducing disparities among intersectional groups, and the SHAP analysis revealed the influence of financial factors alongside potential demographic bias. The project contributes by demonstrating the limitations of re-weighting in credit risk assessments, introducing intersectional analysis, and showing how dataset size affects fairness, offering practical insights for improving model equity

Appendix A

Model performance. ROC-AUC, Log Loss

Group	ROC-AUC			Log Loss		
	100%	50%	25%	100%	50%	25%
hispanic_american_native_female	0.8286	0.8571	0.7000	0.6291	0.6303	0.8994
hispanic_american_native_male	0.8804	0.7551	0.9000	0.7541	0.7248	0.7372
hispanic_asian_female	0.7667	0.7143	1.0000	0.5663	0.5806	0.3707
hispanic_asian_male	0.4444	0.2222	0.0000	0.7370	1.0852	1.0285
hispanic_black_female	0.7467	0.7791	0.6405	0.7521	0.6727	0.7414
hispanic_black_male	0.8228	0.9196	0.8750	0.7024	0.5874	0.7726
hispanic_native_hawaiian_female	0.8600	1.0000	N/A	0.8298	0.6502	N/A
hispanic_native_hawaiian_male	0.6356	0.7344	0.8571	0.6870	0.8649	0.4374
hispanic_white_female	0.7379	0.7333	0.7045	0.6893	0.6835	0.6688
hispanic_white_male	0.6992	0.7280	0.7055	0.7059	0.6602	0.6415
not_hispanic_american_native_female	0.8121	0.8512	0.7851	0.6038	0.6146	0.7581
not_hispanic_american_native_male	0.6316	0.5682	0.6254	0.7955	0.7303	0.6937
not_hispanic_asian_female	0.6768	0.6880	0.6637	0.5903	0.5921	0.5664
not_hispanic_asian_male	0.7064	0.6740	0.7138	0.5914	0.6042	0.5968
not_hispanic_black_female	0.6843	0.6878	0.6506	0.7907	0.7728	0.7819
not_hispanic_black_male	0.6841	0.6669	0.6818	0.7727	0.8258	0.7882
not_hispanic_native_hawaiian_female	0.7806	0.5882	0.8000	0.6575	0.7103	0.4974
not_hispanic_native_hawaiian_male	0.7878	0.7067	0.6531	0.6104	0.7013	0.6184
not_hispanic_white_female	0.6974	0.7035	0.7166	0.5983	0.6001	0.5835
not_hispanic_white_male	0.6791	0.6765	0.6845	0.6100	0.6123	0.6242

TABLE A.1: Model performance. Logistic Regression Simple without bias mitigation

Group	ROC-AUC			Log Loss		
	100%	50%	25%	100%	50%	25%
hispanic_american_native_female	0.8429	0.8571	0.7000	0.6072	0.6046	0.7352
hispanic_american_native_male	0.8804	0.7551	0.9000	0.6834	0.7201	0.6605
hispanic_asian_female	0.7667	0.7143	1.0000	0.6861	0.5795	0.3673
hispanic_asian_male	0.4444	0.2222	0.0000	0.7829	1.0539	0.8936
hispanic_black_female	0.7464	0.7791	0.6364	0.7210	0.6771	0.7491
hispanic_black_male	0.8240	0.9196	0.8750	0.6727	0.5879	0.7103
hispanic_native_hawaiian_female	0.8600	1.0000	N/A	0.6301	0.6366	N/A
hispanic_native_hawaiian_male	0.6356	0.7344	0.8571	0.6414	0.8581	0.4691
hispanic_white_female	0.7374	0.7334	0.7038	0.6916	0.6839	0.6705
hispanic_white_male	0.6994	0.7280	0.7046	0.7077	0.6605	0.6425
not_hispanic_american_native_female	0.8121	0.8512	0.7807	0.6272	0.6096	0.7298
not_hispanic_american_native_male	0.6313	0.5694	0.6321	0.7706	0.7287	0.6978
not_hispanic_asian_female	0.6770	0.6881	0.6638	0.5920	0.5923	0.5684
not_hispanic_asian_male	0.7063	0.6740	0.7138	0.5925	0.6044	0.5972
not_hispanic_black_female	0.6847	0.6878	0.6514	0.7891	0.7724	0.7804
not_hispanic_black_male	0.6846	0.6669	0.6835	0.7709	0.8256	0.7855
not_hispanic_native_hawaiian_female	0.7806	0.5882	0.8000	0.7255	0.7260	0.6286
not_hispanic_native_hawaiian_male	0.7878	0.7067	0.6531	0.6090	0.6960	0.6227
not_hispanic_white_female	0.6974	0.7035	0.7165	0.5986	0.6001	0.5838
not_hispanic_white_male	0.6792	0.6765	0.6842	0.6102	0.6123	0.6244

TABLE A.2: Model performance. Logistic Regression advanced without bias mitigation

Group	ROC-AUC			Log Loss		
	100%	50%	25%	100%	50%	25%
hispanic_american_native_female	0.8667	0.8857	0.8000	0.5035	0.5638	0.5756
hispanic_american_native_male	0.8589	0.7755	0.8333	0.5369	0.5619	0.5305
hispanic_asian_female	0.8333	0.7143	1.0000	0.4942	0.5204	0.1909
hispanic_asian_male	0.5062	0.0000	0.0000	0.7317	0.8548	0.7625
hispanic_black_female	0.7437	0.7844	0.6529	0.6773	0.6230	0.6960
hispanic_black_male	0.8613	0.9263	0.8750	0.5507	0.5070	0.7103
hispanic_native_hawaiian_female	0.8000	1.0000	1.0000	0.6544	0.3365	0.0000
hispanic_native_hawaiian_male	0.6923	0.6719	0.8571	0.6674	0.7548	0.4691
hispanic_white_female	0.7450	0.7570	0.7038	0.6077	0.5712	0.6705
hispanic_white_male	0.7196	0.7379	0.7046	0.6170	0.5742	0.6425
not_hispanic_american_native_female	0.7893	0.8348	0.7807	0.5437	0.5104	0.7298
not_hispanic_american_native_male	0.6469	0.5720	0.6321	0.6398	0.7409	0.6978
not_hispanic_asian_female	0.6725	0.6632	0.6638	0.5627	0.5595	0.5684
not_hispanic_asian_male	0.7076	0.6700	0.7138	0.5548	0.5627	0.5972
not_hispanic_black_female	0.6988	0.6936	0.6514	0.7116	0.6793	0.7804
not_hispanic_black_male	0.6989	0.6805	0.6835	0.6854	0.7083	0.7855
not_hispanic_native_hawaiian_female	0.7139	0.6471	0.8000	0.5550	0.6465	0.6286
not_hispanic_native_hawaiian_male	0.7047	0.5800	0.6531	0.6297	0.7240	0.6227
not_hispanic_white_female	0.7134	0.7216	0.7165	0.5186	0.5111	0.5838
not_hispanic_white_male	0.6984	0.6873	0.6842	0.5335	0.5290	0.6244

TABLE A.3: Model performance. LightGBM Simple without bias mitigation

Group	ROC-AUC			Log Loss		
	100%	**50%**	**25%**	**100%**	**50%**	**25%**
hispanic_american_native_female	0.8190	0.7714	0.8000	0.5684	0.7888	0.4604
hispanic_american_native_male	0.8038	0.6888	0.7000	0.6024	0.6781	0.6414
hispanic_asian_female	0.8000	0.7143	1.0000	0.5010	0.5356	0.1729
hispanic_asian_male	0.4444	0.1111	0.0000	0.7215	0.6833	0.5997
hispanic_black_female	0.6932	0.7050	0.6694	0.7674	0.6409	0.6943
hispanic_black_male	0.8360	0.9219	0.8229	0.5922	0.4406	0.7486
hispanic_native_hawaiian_female	0.8600	1.0000	N/A	0.5357	0.3295	N/A
hispanic_native_hawaiian_male	0.6518	0.6250	0.8571	0.7241	0.7614	0.6397
hispanic_white_female	0.6906	0.7380	0.6438	0.6104	0.5637	0.7061
hispanic_white_male	0.7020	0.6596	0.7104	0.6046	0.6126	0.5533
not_hispanic_american_native_female	0.7763	0.7731	0.8904	0.5611	0.5351	0.4700
not_hispanic_american_native_male	0.6322	0.6326	0.6254	0.6571	0.7201	0.6979
not_hispanic_asian_female	0.6509	0.6522	0.5796	0.5311	0.5432	0.5736
not_hispanic_asian_male	0.6914	0.6523	0.6501	0.5279	0.5440	0.5529
not_hispanic_black_female	0.6788	0.6697	0.6385	0.6956	0.6900	0.7171
not_hispanic_black_male	0.6841	0.6744	0.6411	0.6689	0.7141	0.7446
not_hispanic_native_hawaiian_female	0.6028	0.5098	0.9000	0.5880	0.5717	0.5491
not_hispanic_native_hawaiian_male	0.6663	0.6533	0.7551	0.7676	0.6906	0.7228
not_hispanic_white_female	0.6992	0.6894	0.6924	0.4978	0.5049	0.4962
not_hispanic_white_male	0.6889	0.6622	0.6736	0.5089	0.5221	0.5160

TABLE A.4: Model performance. LightGBM advanced without bias mitigation

Group	ROC-AUC			Log Loss		
	100%	**50%**	**25%**	**100%**	**50%**	**25%**
hispanic_american_native_female	0.8286	0.8571	0.8286	0.6291	0.6303	0.6291
hispanic_american_native_male	0.8804	0.7551	0.8804	0.7541	0.7248	0.7548
hispanic_asian_female	0.7667	0.7143	0.7667	0.5663	0.5806	0.5677
hispanic_asian_male	0.4444	0.2222	0.4444	0.7370	1.0852	0.7375
hispanic_black_female	0.7467	0.7791	0.7467	0.7521	0.6727	0.7520
hispanic_black_male	0.8228	0.9196	0.8228	0.7024	0.5874	0.7020
hispanic_native_hawaiian_female	0.8600	1.0000	0.8600	0.8298	0.6502	0.8284
hispanic_native_hawaiian_male	0.6356	0.7344	0.6356	0.6870	0.8649	0.6871
hispanic_white_female	0.7379	0.7333	0.7379	0.6893	0.6835	0.6895
hispanic_white_male	0.6992	0.7280	0.6992	0.7059	0.6602	0.7057
not_hispanic_american_native_female	0.8121	0.8512	0.8121	0.6038	0.6146	0.6042
not_hispanic_american_native_male	0.6316	0.5682	0.6316	0.7955	0.7303	0.7959
not_hispanic_asian_female	0.6768	0.6880	0.6768	0.5903	0.5921	0.5902
not_hispanic_asian_male	0.7064	0.6740	0.7064	0.5914	0.6042	0.5914
not_hispanic_black_female	0.6843	0.6878	0.6843	0.7907	0.7728	0.7908
not_hispanic_black_male	0.6841	0.6669	0.6841	0.7727	0.8258	0.7727
not_hispanic_native_hawaiian_female	0.7806	0.5882	0.7806	0.6575	0.7103	0.6572
not_hispanic_native_hawaiian_male	0.7878	0.7067	0.7878	0.6104	0.7013	0.6104
not_hispanic_white_female	0.6974	0.7035	0.6974	0.5983	0.6001	0.5983
not_hispanic_white_male	0.6791	0.6765	0.6791	0.6100	0.6123	0.6100

TABLE A.5: Model performance. Logistic Regression simple with bias mitigation

Group	ROC-AUC			Log Loss		
	100%	50%	25%	100%	50%	25%
hispanic_american_native_female	0.8429	0.7143	0.7000	0.6072	0.7558	0.7352
hispanic_american_native_male	0.8804	0.6778	0.9000	0.6834	0.9121	0.6605
hispanic_asian_female	0.7667	0.0000	1.0000	0.6861	0.9318	0.3673
hispanic_asian_male	0.4444	0.4375	0.0000	0.7829	0.6362	0.8936
hispanic_black_female	0.7464	0.7026	0.6364	0.7210	0.7407	0.7491
hispanic_black_male	0.8240	0.7259	0.8750	0.6727	0.8473	0.7103
hispanic_native_hawaiian_female	0.8600	0.6875	-	0.6301	0.7843	-
hispanic_native_hawaiian_male	0.6356	0.8571	0.8571	0.6414	0.7467	0.4691
hispanic_white_female	0.7374	0.7338	0.7038	0.6916	0.6763	0.6705
hispanic_white_male	0.6994	0.7018	0.7046	0.7077	0.6839	0.6425
not_hispanic_american_native_female	0.8121	0.8245	0.7807	0.6272	0.6728	0.7298
not_hispanic_american_native_male	0.6313	0.6467	0.6321	0.7706	0.7355	0.6978
not_hispanic_asian_female	0.6770	0.7145	0.6638	0.5920	0.5749	0.5684
not_hispanic_asian_male	0.7063	0.7274	0.7138	0.5925	0.5783	0.5972
not_hispanic_black_female	0.6847	0.6983	0.6514	0.7891	0.7589	0.7804
not_hispanic_black_male	0.6846	0.7044	0.6835	0.7709	0.7546	0.7855
not_hispanic_native_hawaiian_female	0.7806	0.8503	0.8000	0.7255	0.4692	0.6286
not_hispanic_native_hawaiian_male	0.7878	0.6748	0.6531	0.6090	0.7488	0.6227
not_hispanic_white_female	0.6974	0.7130	0.7165	0.5986	0.5931	0.5838
not_hispanic_white_male	0.6792	0.6726	0.6842	0.6102	0.6113	0.6244

TABLE A.6: Model performance. Logistic Regression advanced with bias mitigation

Group	ROC-AUC			Log Loss		
	100%	50%	25%	100%	50%	25%
hispanic_american_native_female	0.8667	0.7714	0.8000	0.5035	0.6268	0.5756
hispanic_american_native_male	0.8589	0.7398	0.8333	0.5369	0.5880	0.5305
hispanic_asian_female	0.8333	0.6667	1.0000	0.4942	0.4876	0.1909
hispanic_asian_male	0.5062	0.2222	0.0000	0.7317	0.7497	0.7625
hispanic_black_female	0.7437	0.7044	0.6529	0.6773	0.6924	0.6960
hispanic_black_male	0.8613	0.8750	0.8958	0.5507	0.5613	0.5719
hispanic_native_hawaiian_female	0.8000	0.8333	N/A	0.6544	0.6121	N/A
hispanic_native_hawaiian_male	0.6923	0.5625	0.5714	0.6674	0.7893	0.6491
hispanic_white_female	0.7450	0.6718	0.7044	0.6077	0.6823	0.5948
hispanic_white_male	0.7196	0.6632	0.7378	0.6170	0.6596	0.5512
not_hispanic_american_native_female	0.7893	0.7858	0.8202	0.5437	0.7165	0.5181
not_hispanic_american_native_male	0.6469	0.5379	0.6355	0.6398	0.7436	0.6156
not_hispanic_asian_female	0.6725	0.6382	0.6401	0.5627	0.5916	0.5228
not_hispanic_asian_male	0.7076	0.5929	0.6972	0.5548	0.6212	0.5404
not_hispanic_black_female	0.6988	0.6690	0.6655	0.7116	0.7415	0.6816
not_hispanic_black_male	0.6989	0.6486	0.6859	0.6854	0.7797	0.6695
not_hispanic_native_hawaiian_female	0.7139	0.8039	1.0000	0.5550	0.6927	0.3370
not_hispanic_native_hawaiian_male	0.7047	0.5467	0.6939	0.6297	0.7031	0.6482
not_hispanic_white_female	0.7134	0.6654	0.7213	0.5186	0.6020	0.4922
not_hispanic_white_male	0.6984	0.6398	0.6938	0.5335	0.6128	0.5260

TABLE A.7: Model performance. LightGBM simple with bias mitigation

Group	ROC-AUC			Log Loss		
	100%	**50%**	**25%**	**100%**	**50%**	**25%**
hispanic_american_native_female	0.8190	0.7714	0.8000	0.5684	0.6268	0.4604
hispanic_american_native_male	0.8038	0.7398	0.7000	0.6024	0.5880	0.6414
hispanic_asian_female	0.8000	0.6667	1.0000	0.5010	0.4876	0.1729
hispanic_asian_male	0.4444	0.2222	0.0000	0.7215	0.7497	0.5997
hispanic_black_female	0.6932	0.7044	0.6694	0.7674	0.6924	0.6943
hispanic_black_male	0.8360	0.8750	0.8229	0.5922	0.5613	0.7486
hispanic_native_hawaiian_female	0.8600	0.8333	N/A	0.5357	0.6121	N/A
hispanic_native_hawaiian_male	0.6518	0.5625	0.8571	0.7241	0.7893	0.6397
hispanic_white_female	0.6906	0.6718	0.6438	0.6104	0.6823	0.7061
hispanic_white_male	0.7020	0.6632	0.7104	0.6046	0.6596	0.5533
not_hispanic_american_native_female	0.7763	0.7858	0.8904	0.5611	0.7165	0.4700
not_hispanic_american_native_male	0.6322	0.5379	0.6254	0.6571	0.7436	0.6979
not_hispanic_asian_female	0.6509	0.6382	0.5796	0.5311	0.5916	0.5736
not_hispanic_asian_male	0.6914	0.5929	0.6501	0.5279	0.6212	0.5529
not_hispanic_black_female	0.6788	0.6690	0.6385	0.6956	0.7415	0.7171
not_hispanic_black_male	0.6841	0.6486	0.6411	0.6689	0.7797	0.7446
not_hispanic_native_hawaiian_female	0.6028	0.8039	0.9000	0.5880	0.6927	0.5491
not_hispanic_native_hawaiian_male	0.6663	0.5467	0.7551	0.7676	0.7031	0.7228
not_hispanic_white_female	0.6992	0.6654	0.6924	0.4978	0.6020	0.4962
not_hispanic_white_male	0.6889	0.6398	0.6736	0.5089	0.6128	0.5160

TABLE A.8: Model performance. LightGBM Advanced with bias mitigation

Appendix B

Disparate impact, all dataset sizes

Group	Model	DInoW	SPDnoW	DI)	SPD
hispanic_american_indian_female	LR Simple	0.4446	-0.3445	0.4446	-0.3445
	LR Advanced	0.5008	-0.3094	0.5008	-0.3094
	LightGBM Simple	0.4508	-0.4202	0.4508	-0.4202
	LightGBM Advanced	0.5556	-0.3586	0.5556	-0.3586
hispanic_american_indian_male	LR Simple	0.0393	-0.5960	0.0393	-0.5960
	LR Advanced	0.1574	-0.5222	0.1574	-0.5222
	LightGBM Simple	0.5101	-0.3748	0.5101	-0.3748
	LightGBM Advanced	0.4232	-0.4654	0.4232	-0.4654
hispanic_asian_female	LR Simple	0.7585	-0.1498	0.7585	-0.1498
	LR Advanced	0.6644	-0.2080	0.6644	-0.2080
	LightGBM Simple	0.6920	-0.2356	0.6920	-0.2356
	LightGBM Advanced	0.8020	-0.1598	0.8020	-0.1598
hispanic_asian_male	LR Simple	1.0149	0.0092	1.0149	0.0092
	LR Advanced	0.8367	-0.1012	0.8367	-0.1012
	LightGBM Simple	0.9199	-0.0613	0.9199	-0.0613
	LightGBM Advanced	0.8263	-0.1402	0.8263	-0.1402
hispanic_black_female	LR Simple	0.4030	-0.3704	0.4030	-0.3704
	LR Advanced	0.5110	-0.3031	0.5110	-0.3031
	LightGBM Simple	0.6863	-0.2400	0.6863	-0.2400
	LightGBM Advanced	0.6610	-0.2735	0.6610	-0.2735
hispanic_black_male	LR Simple	0.2873	-0.4422	0.2873	-0.4422
	LR Advanced	0.3834	-0.3821	0.3834	-0.3821
	LightGBM Simple	0.5565	-0.3393	0.5565	-0.3393
	LightGBM Advanced	0.5154	-0.3910	0.5154	-0.3910
hispanic_native_hawaiian_female	LR Simple	0.1075	-0.5537	0.1075	-0.5537
	LR Advanced	0.6454	-0.2197	0.6454	-0.2197
	LightGBM Simple	0.6972	-0.2317	0.6972	-0.2317
	LightGBM Advanced	0.6610	-0.2735	0.6610	-0.2735
hispanic_native_hawaiian_male	LR Simple	0.1511	-0.5267	0.1511	-0.5267
	LR Advanced	0.4034	-0.3697	0.4034	-0.3697
	LightGBM Simple	0.5310	-0.3588	0.5310	-0.3588
	LightGBM Advanced	0.2711	-0.5881	0.2711	-0.5881
hispanic_white_female	LR Simple	0.8293	-0.1059	0.8293	-0.1059
	LR Advanced	0.8319	-0.1042	0.8319	-0.1042
	LightGBM Simple	0.7933	-0.1582	0.7933	-0.1582
	LightGBM Advanced	0.8151	-0.1492	0.8151	-0.1492
hispanic_white_male	LR Simple	0.8133	-0.1158	0.8133	-0.1158
	LR Advanced	0.8107	-0.1173	0.8107	-0.1173
	LightGBM Simple	0.8106	-0.1449	0.8106	-0.1449
	LightGBM Advanced	0.8271	-0.1395	0.8271	-0.1395
not_hispanic_american_indian_female	LR Simple	0.4925	-0.3149	0.4925	-0.3149
	LR Advanced	0.4482	-0.3420	0.4482	-0.3420
	LightGBM Simple	0.5689	-0.3298	0.5689	-0.3298
	LightGBM Advanced	0.7230	-0.2235	0.7230	-0.2235
not_hispanic_american_indian_male	LR Simple	0.5874	-0.2560	0.5874	-0.2560
	LR Advanced	0.6154	-0.2384	0.6154	-0.2384
	LightGBM Simple	0.8308	-0.1294	0.8308	-0.1294
	LightGBM Advanced	0.8718	-0.1035	0.8718	-0.1035
not_hispanic_asian_female	LR Simple	1.2434	0.1510	1.2434	0.1510
	LR Advanced	1.2429	0.1506	1.2429	0.1506
	LightGBM Simple	0.9744	-0.0196	0.9744	-0.0196
	LightGBM Advanced	0.9711	-0.0233	0.9711	-0.0233
not_hispanic_asian_male	LR Simple	1.1845	0.1145	1.1845	0.1145
	LR Advanced	1.1852	0.1148	1.1852	0.1148
	LightGBM Simple	0.9671	-0.0252	0.9671	-0.0252
	LightGBM Advanced	0.9759	-0.0195	0.9759	-0.0195

TABLE B.1: Table 1 of Comparison of Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW) with bias mitigation, and without, Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW). 100% batch.

Group	Model	DInoW	SPDnoW	DI)	SPD
not_hispanic_black_female	LR Simple	0.4993	-0.3106	0.4993	-0.3106
	LR Advanced	0.5026	-0.3083	0.5026	-0.3083
	LightGBM Simple	0.5360	-0.3549	0.5360	-0.3549
	LightGBM Advanced	0.6223	-0.3047	0.6223	-0.3047
not_hispanic_black_male	LR Simple	0.4827	-0.3209	0.4827	-0.3209
	LR Advanced	0.4892	-0.3166	0.4892	-0.3166
	LightGBM Simple	0.5566	-0.3392	0.5566	-0.3392
	LightGBM Advanced	0.6380	-0.2921	0.6380	-0.2921
not_hispanic_native_hawaiian_female	LR Simple	0.8410	-0.0987	0.8410	-0.0987
	LR Advanced	0.7016	-0.1850	0.7016	-0.1850
	LightGBM Simple	0.8525	-0.1128	0.8525	-0.1128
	LightGBM Advanced	0.8891	-0.0895	0.8891	-0.0895
not_hispanic_native_hawaiian_male	LR Simple	0.7069	-0.1818	0.7069	-0.1818
	LR Advanced	0.7077	-0.1811	0.7077	-0.1811
	LightGBM Simple	0.6651	-0.2562	0.6651	-0.2562
	LightGBM Advanced	0.7393	-0.2104	0.7393	-0.2104
not_hispanic_white_female	LR Simple	1.0061	0.0038	1.0061	0.0038
	LR Advanced	1.0063	0.0039	1.0063	0.0039
	LightGBM Simple	1.0161	0.0123	1.0161	0.0123
	LightGBM Advanced	1.0071	0.0057	1.0071	0.0057
not_hispanic_white_male	LR Simple	0.8133	-0.1158	0.8133	-0.1158
	LR Advanced	0.8107	-0.1173	0.8107	-0.1173
	LightGBM Simple	0.8106	-0.1449	0.8106	-0.1449
	LightGBM Advanced	0.8271	-0.1395	0.8271	-0.1395

TABLE B.2: Table 2 of Comparison of Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW) with bias mitigation, and without, Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW). 100% batch.

Group	Batch	Model	DInoW	SPDnoW	DIW	SPDW
hispanic_white_female	25%	LR Simple	0.8291	-0.1060	0.8291	-0.1060
		LR Advanced	0.8460	-0.0886	0.8460	-0.0886
		LightGBM Simple	0.7929	-0.1570	0.7929	-0.1570
		LightGBM Advanced	0.8294	-0.1362	0.8294	-0.1362
	50%	LR Simple	0.8661	-0.0827	0.8664	-0.0825
		LR Advanced	0.8701	-0.0792	0.8701	-0.0792
		LightGBM Simple	0.8265	-0.1360	0.7021	-0.2261
		LightGBM Advanced	0.8186	-0.1462	0.8325	-0.1360
hispanic_white_male	25%	LR Simple	0.8139	-0.1155	0.8139	-0.1155
		LR Advanced	0.8133	-0.1158	0.8133	-0.1158
		LightGBM Simple	0.8139	-0.1155	0.8139	-0.1155
		LightGBM Advanced	0.8271	-0.1395	0.8271	-0.1395
	50%	LR Simple	0.8525	-0.0911	0.8491	-0.0931
		LR Advanced	0.8511	-0.0919	0.8722	-0.0779
		LightGBM Simple	0.8628	-0.1075	0.8760	-0.0941
		LightGBM Advanced	0.8918	-0.0879	0.8918	-0.0879
hispanic_american_native_female	25%	LR Simple	0.4445	-0.3447	0.4445	-0.3447
		LR Advanced	0.7447	-0.1469	0.7447	-0.1469
		LightGBM Simple	0.5655	-0.3293	0.5655	-0.3293
		LightGBM Advanced	0.8946	-0.0842	0.8946	-0.0842
	50%	LR Simple	0.4050	-0.3673	0.4050	-0.3673
		LR Advanced	0.5401	-0.2839	0.3514	-0.3955
		LightGBM Simple	0.6589	-0.2589	0.6589	-0.2589
		LightGBM Advanced	0.3078	-0.5623	0.3078	-0.5623
hispanic or latino_american_native_male	25%	LR Simple	0.0393	-0.5962	0.0393	-0.5962
		LR Advanced	0.3159	-0.3937	0.3159	-0.3937
		LightGBM Simple	0.4798	-0.3942	0.4798	-0.3942
		LightGBM Advanced	0.6831	-0.2530	0.6831	-0.2530
	50%	LR Simple	0.1736	-0.5102	0.1736	-0.5102
		LR Advanced	0.0000	-0.6098	0.0000	-0.6098
		LightGBM Simple	0.6118	-0.2946	0.6118	-0.2946
		LightGBM Advanced	0.6156	-0.3123	0.6156	-0.3123
hispanic_asian_female	25%	LR Simple	0.7583	-0.1500	0.7583	-0.1500
		LR Advanced	0.8688	-0.0755	0.8688	-0.0755
		LightGBM Simple	0.6598	-0.2579	0.6598	-0.2579
		LightGBM Advanced	0.6262	-0.2985	0.6262	-0.2985
	50%	LR Simple	1.2959	0.1827	1.2959	0.1827
		LR Advanced	0.6560	-0.2098	0.6560	-0.2098
		LightGBM Simple	1.1860	0.1411	1.1860	0.1411
		LightGBM Advanced	0.8618	-0.1123	0.8618	-0.1123
hispanic_asian_male	25%	LR Simple	1.0146	0.0091	1.0146	0.0091
		LR Advanced	0.9930	-0.0040	0.9930	-0.0040
		LightGBM Simple	0.7540	-0.1864	0.7540	-0.1864
		LightGBM Advanced	1.2524	0.2015	1.2524	0.2015
	50%	LR Simple	0.3240	-0.4173	0.3240	-0.4173
		LR Advanced	0.4861	-0.3172	0.9840	-0.0098
		LightGBM Simple	0.7658	-0.1835	0.9224	-0.0589
		LightGBM Advanced	0.8618	-0.1123	0.8618	-0.1123

TABLE B.3: Intersectional groups. Table 1 of Comparison of Disparate Impact (DInoW) and Statistical Parity Difference(SPDnoW) with bias mitigation, and without, Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW). Batch: 50%, 25%.

Group	Batch	Model	DInoW	SPDnoW	DIW	SPDW
hispanic_black_female	25%	LR Simple	0.4029	-0.3706	0.4029	-0.3706
		LR Advanced	0.7372	-0.1512	0.7372	-0.1512
		LightGBM Simple	0.8397	-0.1215	0.8397	-0.1215
		LightGBM Advanced	0.8349	-0.1318	0.8349	-0.1318
	50%	LR Simple	0.5400	-0.2840	0.5400	-0.2840
		LR Advanced	0.5090	-0.2994	0.5238	-0.2904
		LightGBM Simple	0.6704	-0.2501	0.7389	-0.2046
		LightGBM Advanced	0.7775	-0.1807	0.6689	-0.2690
hispanic_black_male	25%	LR Simple	0.2872	-0.4423	0.2872	-0.4423
		LR Advanced	0.3475	-0.3755	0.3475	-0.3755
		LightGBM Simple	0.4618	-0.4079	0.4618	-0.4079
		LightGBM Advanced	0.5010	-0.3985	0.5010	-0.3985
	50%	LR Simple	0.6259	-0.2310	0.6259	-0.2310
		LR Advanced	0.2050	-0.4848	0.2050	-0.4848
		LightGBM Simple	0.6289	-0.2816	0.5801	-0.3290
		LightGBM Advanced	0.6435	-0.2895	0.6203	-0.3061
hispanic_native_hawaiian_female	25%	LR Simple	0.1074	-0.5539	0.1074	-0.5539
		LR Advanced	1.7377	0.4245	1.7377	0.4245
		LightGBM Simple	1.3195	0.2421	1.3195	0.2421
		LightGBM Advanced	1.2524	0.2015	1.2524	0.2015
	50%	LR Simple	0.0000	-0.6173	0.0000	-0.6173
		LR Advanced	0.0000	-0.6098	0.0000	-0.6098
		LightGBM Simple	0.3765	-0.4732	0.5317	-0.3775
		LightGBM Advanced	0.7035	-0.2408	0.5386	-0.3748
hispanic_native_hawaiian_male	25%	LR Simple	0.1511	-0.5268	0.1511	-0.5268
		LR Advanced	0.2172	-0.4505	0.2172	-0.4505
		LightGBM Simple	0.3299	-0.5079	0.3299	-0.5079
		LightGBM Advanced	0.6262	-0.2985	0.6262	-0.2985
	50%	LR Simple	0.0000	-0.6173	0.0000	-0.6173
		LR Advanced	0.0000	-0.6098	0.0000	-0.6098
		LightGBM Simple	0.3191	-0.5335	0.0824	-0.6964
		LightGBM Advanced	0.5386	-0.3748	0.9428	-0.0461

TABLE B.4: Intersectional groups. Table 2 of Comparison of Disparate Impact (DInoW) and Statistical Parity Difference(SPDnoW) with bias mitigation, and without, Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW). Batch: 50%, 25%.

Group	Batch	Model	DInoW	SPDnoW	DIW	SPDW
not_hispanic_black_female	25%	LR Simple	0.4992	-0.3108	0.4992	-0.3108
		LR Advanced	0.6056	-0.2270	0.6056	-0.2270
		LightGBM Simple	0.6398	-0.2730	0.6398	-0.2730
		LightGBM Advanced	0.7190	-0.2244	0.7190	-0.2244
	50%	LR Simple	0.5001	-0.3087	0.4975	-0.3102
		LR Advanced	0.5004	-0.3084	0.5238	-0.2904
		LightGBM Simple	0.5840	-0.3259	0.6704	-0.2501
		LightGBM Advanced	0.6653	-0.2698	0.6689	-0.2690
not_hispanic_black_male	25%	LR Simple	0.4826	-0.3211	0.4826	-0.3211
		LR Advanced	0.4478	-0.3178	0.4478	-0.3178
		LightGBM Simple	0.6395	-0.2732	0.6395	-0.2732
		LightGBM Advanced	0.6992	-0.2402	0.6992	-0.2402
	50%	LR Simple	0.3776	-0.3844	0.3690	-0.3895
		LR Advanced	0.3708	-0.3883	0.5217	-0.2917
		LightGBM Simple	0.5528	-0.3504	0.6289	-0.2816
		LightGBM Advanced	0.6289	-0.2991	0.6360	-0.2957
not_hispanic_native_american_female	25%	LR Simple	0.4924	-0.3150	0.4924	-0.3150
		LR Advanced	0.4484	-0.3174	0.4484	-0.3174
		LightGBM Simple	0.6810	-0.2417	0.6810	-0.2417
		LightGBM Advanced	0.6868	-0.2501	0.6868	-0.2501
	50%	LR Simple	0.7085	-0.1800	0.7087	-0.1798
		LR Advanced	0.7088	-0.1797	0.5727	-0.2606
		LightGBM Simple	0.7977	-0.1585	0.4118	-0.4464
		LightGBM Advanced	0.8464	-0.1248	0.8464	-0.1248
not_hispanic_native_american_male	25%	LR Simple	0.5872	-0.2562	0.5872	-0.2562
		LR Advanced	0.8206	-0.1033	0.8206	-0.1033
		LightGBM Simple	0.9530	-0.0356	0.9530	-0.0356
		LightGBM Advanced	0.8349	-0.1318	0.8349	-0.1318
	50%	LR Simple	0.5114	-0.3017	0.5115	-0.3015
		LR Advanced	0.5401	-0.2839	0.6427	-0.2179
		LightGBM Simple	0.7389	-0.2046	0.3237	-0.5133
		LightGBM Advanced	0.8855	-0.0930	0.8855	-0.0930
not_hispanic_asian_female	25%	LR Simple	1.2431	0.1508	1.2431	0.1508
		LR Advanced	1.3739	0.2152	1.3739	0.2152
		LightGBM Simple	1.0638	0.0484	1.0638	0.0484
		LightGBM Advanced	1.0181	0.0145	1.0181	0.0145
	50%	LR Simple	1.2357	0.1455	1.2327	0.1436
		LR Advanced	1.2706	0.1650	1.2706	0.1650
		LightGBM Simple	0.9589	-0.0322	1.0688	0.0522
		LightGBM Advanced	0.9539	-0.0375	0.9539	-0.0375
not_hispanic_asian_male	25%	LR Simple	1.1842	0.1143	1.1842	0.1143
		LR Advanced	1.2588	0.1489	1.2588	0.1489
		LightGBM Simple	1.0320	0.0242	1.0320	0.0242
		LightGBM Advanced	0.9614	-0.0308	0.9614	-0.0308
	50%	LR Simple	1.1909	0.1179	1.1924	0.1188
		LR Advanced	1.2171	0.1324	1.2171	0.1324
		LightGBM Simple	0.9690	-0.0243	1.1021	0.0775
		LightGBM Advanced	0.9604	-0.0322	0.9604	-0.0322

TABLE B.5: Intersectional groups. Table 3 of Comparison of Disparate Impact (DInoW) and Statistical Parity Difference(SPDnoW) with bias mitigation, and without, Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW). Batch: 50%, 25%.

Group	Batch	Model	DInoW	SPDnoW	DIW	SPDW
not_hispanic_native_hawaiian_female	25%	LR Simple	0.8408	-0.0988	0.8408	-0.0988
		LR Advanced	0.5792	-0.2421	0.5792	-0.2421
		LightGBM Simple	1.0996	0.0755	1.0996	0.0755
		LightGBM Advanced	0.7306	-0.2151	0.7306	-0.2151
	50%	LR Simple	0.8907	-0.0675	0.8909	-0.0673
		LR Advanced	0.8101	-0.1172	1.1714	0.1045
		LightGBM Simple	0.8934	-0.0835	0.8565	-0.1089
		LightGBM Advanced	0.9849	-0.0123	0.9849	-0.0123
not_hispanic_native_hawaiian_male	25%	LR Simple	0.7068	-0.1820	0.7068	-0.1820
		LR Advanced	0.8688	-0.0755	0.8688	-0.0755
		LightGBM Simple	0.9425	-0.0436	0.9425	-0.0436
		LightGBM Advanced	0.8946	-0.0842	0.8946	-0.0842
	50%	LR Simple	0.6481	-0.2172	0.6480	-0.2173
		LR Advanced	0.5762	-0.2584	0.5762	-0.2584
		LightGBM Simple	0.8168	-0.1435	0.3690	-0.4789
		LightGBM Advanced	0.9357	-0.0523	0.9357	-0.0523
not_hispanic_white_female	25%	LR Simple	1.0060	0.0037	1.0060	0.0037
		LR Advanced	1.0761	0.0438	1.0761	0.0438
		LightGBM Simple	1.0625	0.0474	1.0625	0.0474
		LightGBM Advanced	1.0296	0.0236	1.0296	0.0236
	50%	LR Simple	0.9824	-0.0109	0.9824	-0.0109
		LR Advanced	1.0094	0.0058	1.0094	0.0058
		LightGBM Simple	1.0206	0.0156	1.0206	0.0156
		LightGBM Advanced	1.0107	0.0087	1.0107	0.0087
not_hispanic_white_male	25%	LR Simple	0.8139	-0.1155	0.8139	-0.1155
		LR Advanced	0.8107	-0.1173	0.8107	-0.1173
		LightGBM Simple	0.8106	-0.1449	0.8106	-0.1449
		LightGBM Advanced	0.8271	-0.1395	0.8271	-0.1395
	50%	LR Simple	0.8491	-0.0931	0.8722	-0.0779
		LR Advanced	0.8511	-0.0919	0.8722	-0.0779
		LightGBM Simple	0.8760	-0.0941	0.8889	-0.0895
		LightGBM Advanced	0.8918	-0.0879	0.8918	-0.0879

TABLE B.6: Intersectional groups. Table 4 of Comparison of Disparate Impact (DInoW) and Statistical Parity Difference(SPDnoW) with bias mitigation, and without, Disparate Impact (DInoW), Statistical Parity Difference(SPDnoW). Batch: 50%, 25%.

References

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

Bank for International Settlements (BIS) (2022). *IRB approach: minimum requirements to use IRB approach*. Accessed: 28 July 2024. URL: https://www.bis.org/basel_framework/standard/CRE.

Bellamy, Rachel K. E. et al. (2018). *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. arXiv: 1810.01943 [cs.AI]. URL: https://arxiv.org/abs/1810.01943.

Brotcke, Liming (2022). "Time to assess bias in machine learning models for credit decisions". In: *Journal of Risk and Financial Management* 15.4, p. 165.

Brownlee, Jason (2020). *Add binary flags for missing values for machine learning*. URL: https://machinelearningmastery.com/binary-flags-for-missing-values-for-machine-learning/.

Bussmann, Niklas et al. (2021). "Explainable machine learning in credit risk management". In: *Computational Economics* 57.1, pp. 203–216.

Chan, Jireh Yi-Le et al. (2022). "Mitigating the multicollinearity problem and its machine learning approach: a review". In: *Mathematics* 10.8, p. 1283.

Chinnakum, Warattaya (2023). "Impacts of financial inclusion on poverty and income inequality in developing Asia". In: *The Singapore Economic Review* 68.04, pp. 1375–1391.

Chouldechova, Alexandra (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2, pp. 153–163.

Controlling machine-learning algorithms and their biases — mckinsey.com (n.d.). https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/controlling-machine-learning-algorithms-and-their-biases. [Accessed 01-10-2024].

Das, Sanjiv, Richard Stanton, and Nancy Wallace (2023). "Algorithmic fairness". In: *Annual Review of Financial Economics* 15.1, pp. 565–593.

Doumpos, Michalis et al. (2019). "Analytical techniques in the assessment of credit risk". In: *EURO advanced tutorials on operational research. Cham: Springer International Publishing*.

Emmanuel, Tlamelo et al. (2021). "A survey on missing data in machine learning". In: *Journal of Big data* 8, pp. 1–37.

Fazelpour, Sina and David Danks (2021). "Algorithmic bias: Senses, sources, solutions". In: *Philosophy Compass* 16.8, e12760.

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*, pp. 259–268.

Garcia, Ana Cristina Bicharra, Marcio Gomes Pinto Garcia, and Roberto Rigobon (2023). "Algorithmic discrimination in the credit domain: what do we know about it?" In: *AI & SOCIETY*, pp. 1–40.

Géron, Aurélien (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."

Hlongwane, Rivalani, Kutlwano KKM Ramaboa, and Wilson Mongwe (2024). "Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data". In: *Plos one* 19.5, e0303566.

Hurley, Mikella and Julius Adebayo (2016). "Credit scoring in the era of big data". In: *Yale JL & Tech.* 18, p. 148.

Hurlin, Christophe, Christophe Pérignon, and Sébastien Saurin (2022). "The fairness of credit scoring models". In: *arXiv preprint arXiv:2205.10200*.

Ke, Guolin et al. (2017). "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30.

Kozodoi, Nikita, Johannes Jacob, and Stefan Lessmann (2022). "Fairness in credit scoring: Assessment, implementation and profit implications". In: *European Journal of Operational Research* 297.3, pp. 1083–1094.

Leo, Martin, Suneel Sharma, and Koilakuntla Maddulety (2019). "Machine learning in banking risk management: A literature review". In: *Risks* 7.1, p. 29.

Lessmanna, Stefan et al. (2013). "Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update". In: *Credit Research Centre, Conference Archive*.

Lextrait, Bastien (2023). "Scaling up SMEs' credit scoring scope with LightGBM". In: *Applied Economics* 55.9, pp. 925–943.

Lundberg, Scott M et al. (2019). "Explainable AI for trees: From local explanations to global understanding". In: *arXiv preprint arXiv:1905.04610*.

Nasima (2021). *Effective strategies for handling missing values in data analysis*. Accessed: 20 August 2024. URL: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/.

Noriega, Jomark Pablo, Luis Antonio Rivera, and José Alfredo Herrera (2023). "Machine Learning for Credit Risk Prediction: A Systematic Literature Review". In: *Data* 8.11, p. 169.

Park, Cyn-Young and Rogelio Mercado (2015). "Financial inclusion, poverty, and income inequality in developing Asia". In: *Asian Development Bank Economics Working Paper Series* 426.

Ponsam, J Godwin et al. (2021). "Credit Risk Analysis using LightGBM and a comparative study of popular algorithms". In: *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*. IEEE, pp. 634–641.

Popoola, Gideon and John Sheppard (2024). "Investigating and Mitigating the Performance–Fairness Tradeoff via Protected-Category Sampling". In: *Electronics* 13.15, p. 3024.

Ray, Eben and Ayuns Luz (Mar. 2024). "AI-powered credit scoring and risk assessment models". In.

Ray, Samrat (2022). "Fraud detection in e-Commerce using machine learning". In: *BOHR International Journal of Advances in Management Research* 1.1.

Singh, Arashdeep et al. (2022). "Developing a novel fair-loan classifier through a multi-sensitive debiasing pipeline: Dualfair". In: *Machine Learning and Knowledge Extraction* 4.1, pp. 240–253.

Suhadolnik, Nicolas, Jo Ueyama, and Sergio Da Silva (2023). "Machine learning for enhanced credit risk assessment: An empirical approach". In: *Journal of Risk and Financial Management* 16.12, p. 496.

Tounsi, Youssef, Larbi Hassouni, and Houda Anoun (2017). "Credit scoring in the age of big data–A state-of-the-art". In: *International Journal of Computer Science and Information Security (IJCSIS)* 15.7, pp. 134–145.

Verma, Sahil and Julia Rubin (2018). "Fairness definitions explained". In: *Proceedings of the international workshop on software fairness*, pp. 1–7.