Tracking Machine Learning Bias Creep in Traditional and Online Lending Systems with Covariance Analysis

Conference or Workshop Item

Tracking Machine Learning Bias Creep in Traditional and Online Lending Systems with Covariance Analysis

Ángel Pavón Pérez
The Open University
Milton Keynes, United Kingdom

Miriam Fernandez
The Open University
Milton Keynes, United Kingdom

Hasan Al-Madfai
Visa Europe
London, United Kingdom

Grégoire Burel
The Open University
Milton Keynes, United Kingdom

Harith Alani
The Open University
Milton Keynes, United Kingdom

ABSTRACT

Machine Learning (ML) algorithms are embedded within online banking services, proposing decisions about consumers' credit cards, car loans, and mortgages. These algorithms are sometimes biased, resulting in unfair decisions toward certain groups. One common approach for addressing such bias is simply dropping the sensitive attributes from the training data (e.g. gender). However, sensitive attributes can indirectly be represented by other attributes in the data (e.g. maternity leave taken). This paper addresses the problem of identifying attributes that can mimic sensitive attributes by proposing a new approach based on covariance analysis. Our evaluation conducted on two different credit datasets, extracted from a traditional and an online banking institution respectively, shows how our approach: (i) effectively identifies the attributes from the data that encapsulate sensitive information and, (ii) leads to the reduction of biases in ML models, while maintaining their overall performance.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Feature selection*; • **Information systems** → *World Wide Web*.

KEYWORDS

Machine Learning, Financial Services, Bias in data, Bias identification, Bias mitigation

ACM Reference Format:
Ángel Pavón Pérez, Miriam Fernandez, Hasan Al-Madfai, Grégoire Burel, and Harith Alani. 2023. Tracking Machine Learning Bias Creep in Traditional and Online Lending Systems with Covariance Analysis. In *15th ACM Web Science Conference 2023 (WebSci '23), April 30–May 01, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3578503.3583605

1 INTRODUCTION

Automatic decision-making based on large amounts of data ingested by Machine Learning (ML) models has become increasingly present in all aspects of our daily life (online and offline). These models have become essential tools in multiple domains, from health care to the financial sector, taking decisions, for example, on whether individuals should or not get a loan based on automatically computed credit risks.

Recent research has however shown how these algorithms can be discriminative when considering sensitive characteristics such as gender, ethnicity, disability, or sexual orientation. [8, 34].

Algorithmic bias exists even when there is no discrimination intention since sometimes biases may be inherent to the data sources used to train these systems. Even when the sensitive attributes have been suppressed from the input (e.g., gender), a ML algorithm may still discriminate on the basis of such sensitive attributes because they may be indirectly represented by other information (e.g., parental leave taken).[1]

Various solutions have been recently proposed in the literature to address the issue of algorithmic bias. Works can be divided into three main categories: (i) understanding bias (i.e., how bias is created in our societies and enters our socio-technical systems), (ii) mitigating bias (i.e., how different approaches target bias in different stages of ML-decision making, focusing on data inputs, learning algorithms and model outputs), and (iii) accounting for bias (via bias-aware data collection, or explaining ML-decisions in human terms) [30, 32].

In this work, we address the problem of identifying and mitigating bias by focusing on the data used to train these ML systems. Bias in data can be manifested through sensitive attributes and their causal influences, or through under/over-representation of certain groups. Our work tackles bias by identifying the attributes in the data that indirectly represent sensitive information (e.g., gender). Once these attributes are identified we: (i) modify the data accordingly (i.e., remove such attributes to ensure that sensitive information can no longer be singled out) and (ii) create ML models using the modified data as a mitigation technique.

We apply our approach to the financial domain, particularly to lending systems [46]. These systems use ML algorithms to determine the credit risk associated with an individual and hence, whether the individual should be granted or not their requested credit. We have chosen this domain of application due to the wide range of information gathered by these systems, and the need for automatic methods to discard data that can directly or indirectly

WebSci '23, April 30–May 01, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0089-7/23/04…$15.00
https://doi.org/10.1145/3578503.3583605

[1] Algorithmic bias: from discrimination discovery to fairness-aware data mining. A Tutorial at KDD'16. https://francescobonchi.com/algorithmic_bias_tutorial.html

encapsulate biases. It is also important to highlight that financial biases can seriously affect the lives of individuals and groups [4, 15].

We represent this domain with two real datasets (see Section 4). The first one (German Credit Dataset) is an example of data extracted from traditional offline banking while the second one (Home Credit dataset) contains data extracted from an online banking institution that broadens financial inclusion for the unbanked population.

Our evaluation, conducted over these two datasets, shows how our approach: (i) reduces the propagation of data biases into the creation of ML models while maintaining their performance level and, (ii) contributes to the analysis and understanding of how societal biases are encapsulated within data.

Our contributions can be summarised as follows:

- A review of existing works that aimed to reduce bias in data, particularly for the financial sector.
- A novel approach based on the application of statistical tests as a feature selection technique based on covariance to identify attributes within the data that may indirectly represent sensitive attributes.
- An evaluation of the effect of our proposed approach within the financial domain (particularly lending systems) in terms of ML model performance and bias reduction.

The rest of the paper is structured as follows: Section 2 discusses related work. Section 3 describes our proposed approach to identify and mitigate bias. Section 4 introduces the data used in our experiments, which encapsulates the domain of application, in this case lending systems. Section 5 presents our experiments and results. Discussions and conclusions are presented in Sections 6, and 7 respectively.

2 LITERATURE REVIEW

With the creation of the Web and the rise of digital data, the biases that have been intrinsically embedded in culture and history, are now spreading faster and affecting more people. Minoritized groups are harmed by algorithms that replicate, and in some cases amplify, our existing societal biases [2].

A big part of the problem is the data used to train these algorithms. As discussed by [9], data biases that can lead to the creation of unfair and biased ML models. Some of the most popular biases found in data include selection bias, due to how the data is collected (e.g. class imbalance), and historical bias due to the inherent human biases existing within data [2, 33].

Studies have attempted to *identify* bias and *mitigate* them at different stages in the development of ML models [30, 32]. Bias identification consists of finding or measuring those factors that may cause an ML model to be biased [14, 45]. Bias mitigation, on the other hand, is focused on preventing and reducing bias. Bias mitigation is done at three different stages in the development of ML models: (i) Before training the model, known as pre-processing methods [23, 42, 44]. (ii) During the model's development, also known as in-processing methods [7, 24, 38]. And (iii) after the model's training and deployment phase, also referred to as post-processing methods [24, 26].

Works that focus on bias identification have proposed a wide range of metrics [14, 19, 43, 45]. Most of these metrics focus on identifying biases within the models' predictions (algorithmic decisions), once the ML models have already been generated, but do not consider where those biases originate, and whether the data used to train the models may be at fault. Metrics that have attempted to identify biases in data focus on the relationship between the class (the element that the ML model aims to predict, e.g., an individual's credit risk) and the sensitive attributes or information (e.g., gender, ethnicity, etc.) used to predict that class. These metrics measure the relationships between the class and sensitive attributes, from class imbalance to divergences of sensitive group distributions with regard to the class [43]. As opposed to these works, our work focuses on identifying the relationships between the sensitive attributes and other attributes within the data, as well as measuring the strength of such relationships.

Multiple works have also targeted the problem of bias mitigation. A popular method in the literature is *fairness through unawareness* [27]. This method removes the sensitive attributes from the data so that the model does not learn to discriminate. However, sensitive attributes are sometimes indirectly represented by other information within the data, and they could still learn to discriminate even when the sensitive attribute is no longer present [31, 32]. Our approach towards bias mitigation builds on this idea, but removes from the data, not only the sensitive attribute but also all the other attributes that indirectly represent it. While previous works claim that attribute removal could lead to a downgrade in model performance [7], our experiments show that this is not necessarily the case and that, by removing a selected set of attributes (identified by our approach), we can still maintain or even improve model performance.

Some studies have considered the use of correlations when mitigating bias. Kamiran [24] proposed to modify ML algorithms (in-processing approaches) by creating decision trees that seek a high correlation with the target and a low correlation with the sensitive attribute. Kamishima [25] proposed adding a regularised parameter to Logistic Regression for taking correlation into account. Other methods try to remove these relations, for example, by training a model for each possible value of the sensitive attribute [7] or directly creating a new representation of the data without information to identify if a person belongs to a protected group [47]. While these studies use correlation to mitigate model bias, they do not identify the subset of attributes on the data that might be biasing the model due to these correlations.

The most similar approach that we have found in the work presented here is from Kamiran and colleagues [23]. One of the methods they propose to mitigate bias is to remove attributes based on correlation but without stating how many or which attributes should be removed. A key aspect of our proposed approach is that it provides a confidence level that indicates which set of attributes are the ones related to the sensitive attribute and should be removed to unbias the model.

When looking at identifying and mitigating biases in the financial domain, the limitations are similar to the ones previously presented. Zhang and colleagues [48] present a comprehensive discussion of how existing bias identification and mitigation methods could be applied to the financial industry. Hassani and colleagues [21] assess whether bias exists in the data by attempting to predict the sensitive attribute (gender) using customer information, but

without considering any bias mitigation. Das [14], proposes metrics to identify bias in the data, although focusing on the relationship of the sensitive attribute to the class and not identifying other attributes that may be biasing the model.

This paper proposes a novel approach that contributes to the literature on bias identification and mitigation. Our approach differs from existing works by identifying which data attributes indirectly represent a sensitive attribute and by providing a score determined by the confidence/significance of the strength of such relation. We used the acquired knowledge to provide a step toward bias mitigation by removing the identified attributes from the data. Our results, obtained by applying the proposed approach to the financial domain using two credit datasets, show that removing the identified information does not lead to a performance drop in ML models.

3 IDENTIFYING AND MITIGATING BIAS WITH COVARIANCE ANALYSIS

As previously described, the key objective of this work is to identify how a sensitive attribute (e.g. gender or ethnicity), may be indirectly represented by other attributes with covariance analysis. For that purpose, we propose a novel approach based on the use of statistical tests and ML models. In our experiments, we will be creating two types of ML models: (i) the first type of ML models target the sensitive attribute and show how non-sensitive attributes can indirectly represent it, (ii) the second type of ML models targets the class (in our use case the credit risk) for evaluating the effects of covariance in model bias and the effects of mitigation techniques based on that covariance.

Our proposed approached can be visualised in Figure 1. This approach is divided in three main stages: (i) attribute relationship analysis, (ii) indirect attribute representation of sensitive attributes and, (iii) training and evaluating ML models.

Attributes relationship analysis. The aim is to use statistical tests for analysing covariance as a methodology to identify the attributes that indirectly represent the sensitive attribute. The main steps are the following:

- *Identify an appropriate statistical test.* To do so, several considerations should be taken into account depending on the available data. For example, attributes within the data (e.g., customer's age, employment status, etc.) can be continuous or discrete (i.e., have a finite set of values), which may vary the selected test type. Therefore, when choosing statistical tests to check if two attributes are dependent or not, we should consider the type of attributes (e.g. categorical or numerical), their distributions and if the observations are independent among others. More information is provided in Section 5.1.
- *Apply statistical test.* In this step, we apply the statistical tests previously selected for covariance analysis, to identify which other attributes in the data indirectly represent the selected sensitive attribute. It is important to highlight that a key difference of our proposed approach with respect to previous methods, is that the statistical tests considered in our approach provide us with a level of significance/confidence[2], i.e., we have an indication of how strongly each of the identified attributes represents the sensitive attribute. This is important because it allows us to establish a threshold of confidence when discarding attributes that might bias a ML model. As an output of this step, we obtain the attributes sorted according to the level of relationship with the sensitive attribute.

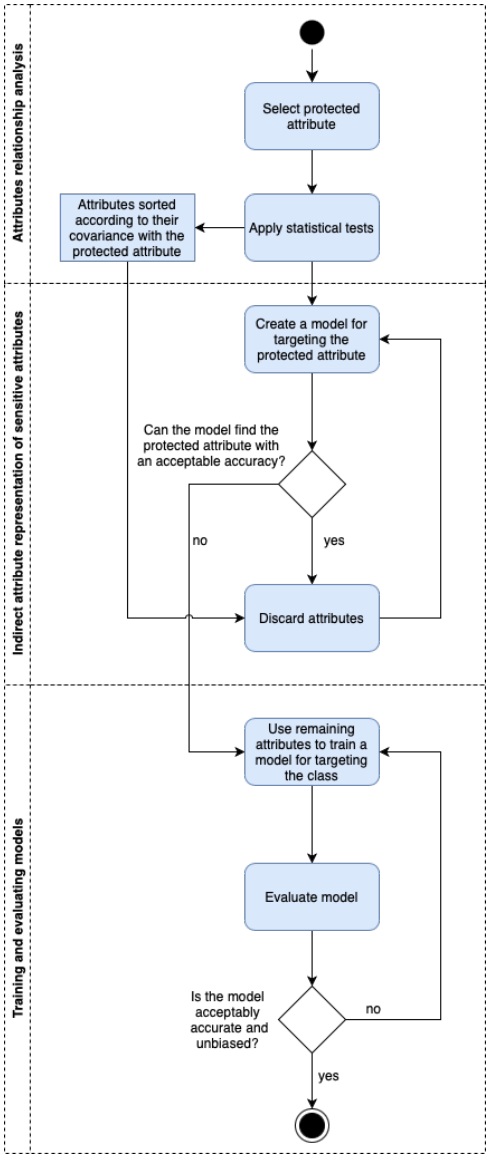

Figure 1: Experiments pipeline

Indirect representation of sensitive attributes. In this part of the pipeline we determine how many of the previously identified attributes we need to delete from the data in order to ensure that

[2]Statistical literacy guide. Confidence intervals and statistical significance. https://researchbriefings.files.parliament.uk/documents/SN04448/SN04448.pdf

the sensitive attribute can not be identified. The main steps are the following:

- *Create a ML model to predict the sensitive attribute.* In this step we create a ML model that uses the full set of attributes from the data (e.g., owning a car, employment status, maternity leave taken, etc.), except the selected sensitive attribute, to to predict the selected sensitive attribute (e.g., gender).
- *Discard attributes related with the sensitive attribute.* In this step, we progressively discard attributes from the data based on their significance (as identified in the attribute relationship analysis step) and retrain and test the previously created ML model, until we can no longer predict the sensitive attribute i.e., the accuracy of the ML model targeting the sensitive attribute must be equal or lower than the majority group ratio of the sensitive attribute (e.g. if a dataset has 60% males and 40% females, we will consider that we can not predict gender when the ML model accuracy is 60% or lower). This allows us to remove the minimum amount of information from the data while ensuring that the sensitive attribute is no longer indirectly represented up to the point that can be predicted.

Training and evaluating ML models to predict the class. In this part of the approach, we train ML models to predict the class (e.g., a person's credit risk) and analyse their performance and remaining bias levels. The main steps are the following:

- *Create a ML model with the remaining attributes.* After having removed the sensitive attribute, and the attributes related to it, it is expected that a ML model predicting the class will be less biased in comparison to a model that includes in its training the attributes that indirectly represent the sensitive one.
- *Evaluate the model on both accuracy and fairness.* The goal of this step is to assess the accuracy and fairness of the previously generated ML model. With this purpose, various performance and fairness metrics are considered. Specific details can be seen in Section 5.3.

4 FINANCIAL DOMAIN

In this section, we present the datasets that are used to model the financial domain. We describe their key characteristics and attributes, as well as the pre-processing conducted over the data.

4.1 Datasets

We use two datasets as use cases for our research, the German Credit dataset [22] and the Home Credit dataset [13]. We have chosen these datasets because their attributes contain personal and financial information, including some sensitive attributes such as gender. Additionally, both datasets have as class whether the individual has a good or bad credit risk. In Figure 2, we show a reduced example of how items in both datasets are represented.

The German Credit dataset comes from a german bank in the 70s, containing 1000 individuals with 20 personal attributes. Previous work conducted on this dataset has shown how models trained with this data have led to biases, particularly when considering gender [6, 28]. The dataset provides us with important information on the bias of the time.

Person id	*Gender*	*Accommodation type*	*Requested credit*	*[Other attributes]*	*Credit risk*
001	female	Rented house	1000$	...	Bad
002	male	Rented house	1500$	...	Good
003	male	Owned house	2000$	...	Good

Figure 2: Example of item representation

Home Credit is an online company primarily broadening financial inclusion for the unbanked population. Their models use a wide range of data to predict their clients' repayment capacities. The individuals in the home credit dataset are at risk of exclusion since the platform is oriented toward unbanked people (the majority of whom are women). The company made a dataset public in 2018 for a Kaggle competition focused on creating explainable lending models. This dataset comprises seven tables with information on Home Credit clients applying for loans. The main table contains information on the client's loan application. There are two other tables on clients' previous loan applications at other institutions and four tables with information on clients' previous loan applications at Home Credit. In total, these tables have over 200 attributes and over 300.000 individuals. The Home Credit dataset is a complex online financial dataset that allows us to analyse the bias of current online lending tools.

Despite the difference in the time when they were created and the size of these datasets, both have as main class whether individual has good or bad credit risk. They also share a range of attributes, such as the amount of credit requested, the family status e.g., single or married, or the type of accommodation e.g., owned or rented.

The individuals who apply for a loan in both datasets present important differences. For example, the percentage of females applying for a loan varies between datasets, being 31% in the German credit dataset and around 65% in the Home Credit dataset. Furthermore, it can be seen in the German credit dataset that 7.5% more males have a good credit risk than females. However, in Home Credit, the gap is much smaller, a 3% in favour of females. These datasets, therefore, give us a different perspective between traditional financial services and modern online banking.

It's important to highlight that, due to regulation and privacy concerns in the financial sector, data is generally not publicly available. To the best of our knowledge, these are the only two public datasets that include information about the credit risk of individuals and their sensitive attributes, such as gender. Some simulated datasets exist [35], but in this paper, we have focused on real data.

4.2 Preprocessing

In this section, we present the considerations that have been taken when assessing the data and the pre-processing that has been conducted accordingly.

One-Hot Encoding has been applied to both datasets to improve ML model performance. One-Hot Encoding enables us to obtain one attribute for each possible value of the categorical attributes. E.g., in the German credit dataset, as we have 50 different values for categorical values, 7 numerical values and we left class and gender unprocessed as boolean attributes for fairness and accuracy

measuring, we will have, after the encoding, 59 attributes in total. Note that this processing is only for training the ML models. When applying the statistical tests, we will use the appropriate tests with the categorical attributes (see Section 5.1).

For the Home Credit dataset, we extract new attributes from the non-main tables. To achieve that, we group the attributes according to the user identifier to extract the average of the numerical and encoded categorical attributes or count the number of previous instalments or applications. For example, if a customer has applied for 2000$, 6000$ and 7000$ loans in the past, we will extract two new attributes: the number of past loans, which will have a value of 3 and the average loan amount, which will have a value of 5000$. Thus, in total, we will have 370 attributes. Furthermore, we remove columns with more than 25% nan values to avoid adding noise to the ML model. As a result, we are left with 293 attributes. Most of the discarded attributes are from the table containing credit balance information for individuals in previous Home Credit loan applications, probably because most users do not have a credit history, given the nature of the platform. Then, we treat the remaining nan values according to the variable type. We replace them with a new class *null* for categorical variables and with the median for numerical variables. Thus, we modified 233 columns and replaced almost 5 million nan values (around 5% of the data). Finally, when training ML models with this dataset, we also apply random under-sampling regarding the class in the training set to avoid overfitting, reducing 307511 rows to 79499 in our training data.

For the German Credit dataset, we transform the attribute *status_and_sex* (family status and gender information) to only gender. We do this simplification because all females in the dataset have a generic family status that is different from males' family status, and therefore, this attribute ends up being a duplicate of gender.

Also, for both datasets, when using them for training ML models, we randomly extract 80% of data in a stratified way (according to the target) for training, leaving the remaining 20% for testing. Note that the Home Credit dataset has specific test data for kaggle that we have not used because the class (i.e., credit risk) has not been made public.

5 EXPERIMENTS

While digitalisation has provided an avenue to promote financial inclusion in areas where traditional banking services have declined [39], it has also brought attention to the importance of ensuring that no one is discriminated against on the basis of their gender in the provision of financial services. Legislation in some countries prohibits such discrimination [49], underscoring the need to address potential gender-based biases. In our experiments, we selected gender as the sensitive attribute. This attribute captures a larger number of individuals, and it is shared across both of the datasets used.

Thus, we analyse the covariance of other attributes with gender to understand how those attributes may indirectly represent gender and we use that knowledge to reduce the bias. Following our proposed approach (see Section 3), the experiments are divided into three parts: (i) The attributes relationship analysis, (ii) the key variables to find sensitive attributes and (iii) the training and evaluating of ML models.

5.1 Attribute relationship analysis

As mentioned earlier, sensitive attributes can be represented by other attributes. Relationship analysis between these attributes is key to understanding where the bias of the models may emerge from. For analysing the attribute relationships, we use statistical tests as metrics for feature selection to measure the covariance between attributes. Feature selection reduces the number of attributes used as input in a ML model's training to reduce computational costs and improve model performance for predicting the target attribute [5]. This attribute selection can follow a variety of metrics and algorithms [10], several of them based on correlation with the target [18]. With this in mind, we visualise the problem as a feature selection problem in which the target will be gender, and the selected attributes will be the ones to be discarded when building the final model. The statistical tests we used for feature selection are:

- **Chi-Square independence test [36]**: The chi-squared test of independence establishes a significance value to declare which categorical attributes are likely to be independent of the target (also categorical). That is an advantage with respect to other metrics like Mutual Information in which there is a difficulty of finding a cut-off to establish which values are considered dependent and which independent [41]. Furthermore, all Chi-Square independence test assumptions are met in our datasets (all observations are independent, cells in the contingency table are mutually exclusive and their expected value is 5 or greater in at least 80% of cells). We applied this statistical test with the categorical attributes set out in our datasets with respect to the gender attribute.
- **Mann-Whitney U test [29]**: To analyse whether numerical attributes (e.g. age) are distributed differently across the categorical target/outcome (e.g. gender) we have used the Mann-Whithey U test. The Mann-Whitney U test is a non-parametric test that can tell us whether two samples are likely to have originated from the same population or not with a set confidence level (i.e. a significant Mann-Whitney U test, indicates that there is a relationship between the categorical outcome and the numerical variable). In this way, our two samples to compare will be the values of a numerical variable for each value of our sensitive attribute (e.g. males age and females age) and, thanks to the test, know if their distributions are likely to be similar and therefore less likely to bias the model. Also, all assumptions are met in our datasets (all observations are independent and they follow an ordinal measurement scale). We applied this statistical test with the numerical attributes set out in our datasets with respect to the gender.

Note that for each dataset, the most appropriate metric or test should be used, looking at, among others, the assumptions of each statistical test [29, 36]. For example, if our numerical attributes had followed a normal distribution, we could have applied the independent t-test [12].

For the German Credit dataset, as can be seen in Figure 4, the Chi-Square test indicates that the attributes with a higher covariance with gender are: (i) housing type, (ii) since when the individual is employed, (iii) credit purpose and (iv) credit history (dependent attributes according to the Chi-Square test with a significance level

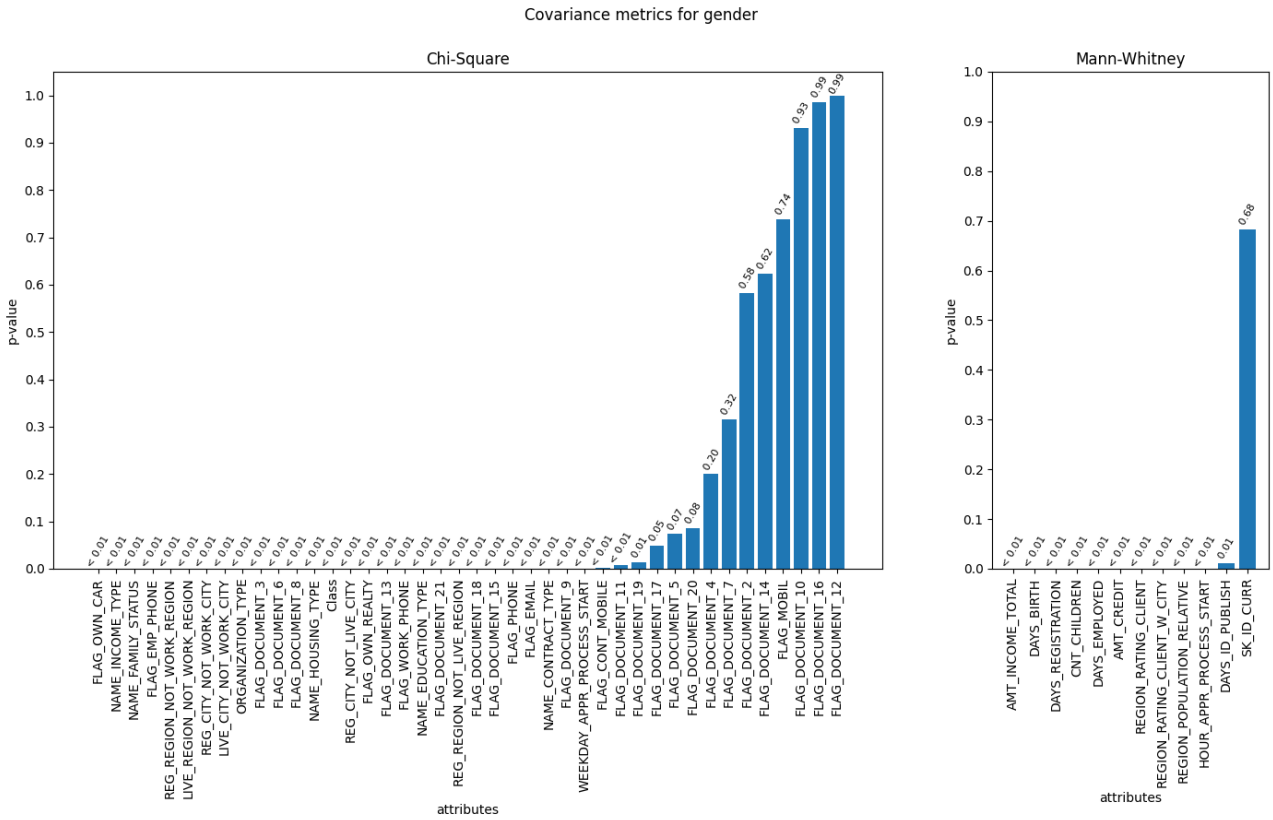

Figure 3: Chi-Square and Mann-Whitney U sample results for the Home Credit dataset

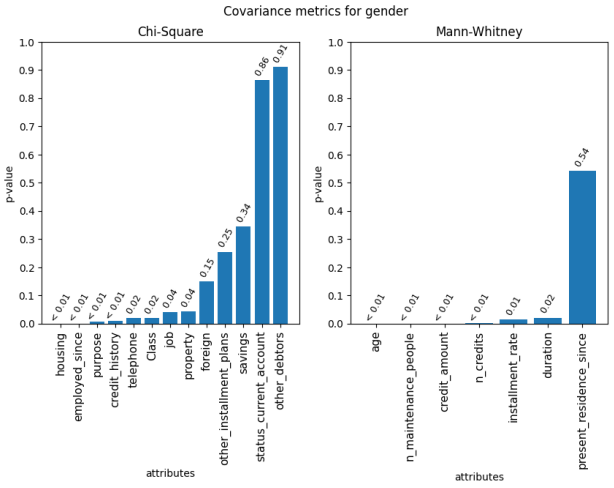

Figure 4: Chi-Square and Mann-Whitney U results for the German Credit dataset

of 1%). Also, as seen in the same figure, the Mann-Whitney U test, with a significance level of 1%, indicates that: (i) age, (ii) number of persons maintained, (iii) credit amount, and (iv) number of credits, are related with gender as the test indicates that they follow different distributions between males and females.

For the Home Credit dataset, both statistical tests show that more than half of the attributes are highly related to gender (with a significance level of 1%). As in this dataset, we have more than 200 attributes, for this paper, we show a sample of these results in Figure 3. As can be seen, some of these related attributes are if they own a car, their income type, family status, housing type, income amount or age, among others. Thus, in this dataset of modern online banking, we see how some of the relationships of traditional banking (age and housing type) are maintained while new relationships are added due to the significant increase in data. Therefore, it is vitally important to consider these relationships in modern online banking to avoid perpetuating old biases or even adding new ones.

5.2 Indirect attribute representation of sensitive attributes

This step aims to obtain a ML model targeting the sensitive attribute that explains the relationships between the attributes and the sensitive attribute, and to find the confidence level needed to discard attributes to remove the sensitive attribute indirect representation. The aim is to see to what extent there is a relationship between non-sensitive and sensitive attributes, how they relate to each other,

and how they can potentially introduce bias into a model. Thus, ML models based on decision trees were used to view the relationships of the attributes representing the gender as they are easy to interpret, visualise and show clearly the relation between attributes [37]. Accordingly, for the German Credit dataset, we used decision trees, as it got as good performance results as more complex ML models like random forest and XGBoost when targeting gender (probably due to the simplicity of the dataset) and for the purpose of this paper to show how gender can be indirectly represented by other attributes in a simple way. Meanwhile, for Home Credit, we used XGBoost models as they got the best performance results when targeting gender (compared to other ML models like logistic regression and random forest). Furthermore, we also use a forward search algorithm to create these models [1], in order to get the best possible performance in predicting gender.

The forward search that we use consists of creating a model for each attribute to predict the sensitive attribute. We keep the model's attribute that gave the best accuracy, X_1. Then, we repeat the process by creating a model for each attribute but, this time, also adding X_1 to try to improve accuracy. Again, we keep the model's attributes with the best accuracy. We keep adding attributes until a stop criterion is met. The stop criterion is when the accuracy does not improve any further when adding a new attribute to the model, i.e. when accuracy starts decreasing or is equal. In this way, the result should be an explainable ML model that uses the best attributes for getting the best performance when predicting that sensitive attribute (gender in our case).

As stated before, we trained decision trees (German Credit dataset) and XGBoost models (Home Credit dataset) using this approach to explain the relationships between attributes and gender.

An example of a decision tree trained in German Credit dataset can be seen in Figure 5. At each node of the decision tree, it can be seen the value that the attribute takes to split, the Gini index, the percentage of training samples, the probability of belonging to males or females and the majority class. In this case, with 77.5% accuracy (against 69% for the majority class), the gender of an individual can be guessed by the number of people supported (*n_maintenance_people*), whether they live in rented accommodation (*housing_3*) and whether they have not taken out another credit (*credit_history_4*). For example, we can say that if the number of people supported is more significant than one and previous credits have been taken, with the certainty that these data provide, that individual will be a male.

Similarly, we obtained an XGBoost model for the Home Credit dataset with 83.5% accuracy when targeting gender (against 65.8% for the majority class). In this case, gender can be targeted by several attributes, some of them self-explanatory, like if individuals are widows or not (*name_family_status_8*) and others less self-explanatory like if they own a car (*flag_own_car*) or if their city permanent address does not match their city work address (*reg_city_not_work_city*).

In this way, we demonstrate not only the covariance between these attributes but also show how they are related, proving that gender can indeed be indirectly represented by other attributes and helping to visualise where the problem of bias in the data lies to mitigate it later.

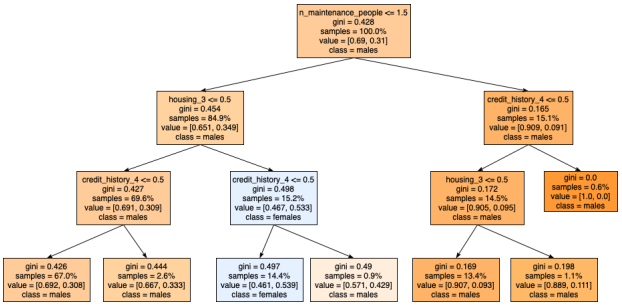

Figure 5: Decision tree example for finding gender in the German Credit dataset

Due to random components (e.g. feature split permutation), each ML model targeting gender may use the attributes in different ways and obtain different results (which will also influence the forward search as the accuracy of these models guides it). Therefore, due to these random components, each ML model will be different. That is why we have trained 100 ML models to obtain the average as, after 100 models, no significant changes in the average results are observed. The objective is to see which attributes are most likely to be selected and, thus, indicate the best attributes for predicting gender according to the ML models. Moreover, thanks to the One-Hot Encoding performed in the data processing, we can know the most used attributes and the values of these most related with gender. The results can be seen in Table 1 for the German Credit dataset and in Table 2 for the Home Credit dataset.

Table 1: Attributes that were selected more than 15% times in the German Credit dataset to predict gender

Variable name (after hot encoding)	Meaning	% of times it was selected for predicting gender
housing_3	rent	74%
housing_1	own a house	49%
n_maintenance_people	number of maintained people	40%
purpose_1	radio or tv	29%
employed_since_1	More than 7 years	27%
housing_2	for free	27%
age	client age	22%
purpose_8	repairs	19%
employed_since_5	less than 1 year	17%
purpose_9	other purposes	16%
purpose_6	business	16%
employed_since_4	unemployed	16%

As we can see, in the German Credit dataset, the housing attribute (especially for its rental value) is often the most used attribute to predict gender as well as other attributes such as the number of persons maintained. On the other hand, in the Home Credit dataset, the family status (especially for its widow value)

Table 2: Attributes that were selected more than 50% times in the Home Credit dataset to predict gender

Variable name	Meaning	% of times it was selected for predicting gender
name_family_status_8	widow	79%
name_income_type_2	state servant	70%
flag_own_car	own a car	65%
name_family_status_5	separated	64%
reg_city_not_work_city	Flag if client's permanent address does not match work address.	57%
name_income_type_4	pensioner	55%
name_family_status_3	Civil marriage	53%

is often the most used attribute to predict gender. This finding is interesting, as family status was originally embedded in gender in the German Credit dataset, as mentioned in Section 4.2. This finding implies that although the family status is a different attribute in more modern data, family status can still be an important proxy to represent gender indirectly. Furthermore, due to the increase of information in digital data, new attributes emerge that can indirectly represent gender, such as whether individuals own a car or whether their city of work is different from where they live.

Once the relationship between gender and dependent attributes has been observed, we can use a similar strategy to look for attributes that cannot predict gender in order to use them in the final ML model targeting the credit risk and prevent that model from learning to distinguish individuals by a sensitive attribute, i.e., we will discard attributes that can predict gender.

For this purpose, the gender-independent attributes are obtained using Chi-Square and Mann-Whitney U tests at 0.01 significance levels. With these remaining attributes, for the Home Credit dataset, we could not train a model to target gender that exceeded the accuracy of the majority class (65%). Because of this, for our processed Home Credit dataset, with 99% confidence, we consider that there are 282 gender-dependent attributes and 88 gender-independent attributes.

However, for the German Credit dataset, it was still possible to build a model that predicted gender with 72% accuracy (compared to 69% accuracy of the majority class). Thus, we also discarded the attributes that had a p-value lower than 0.05 in the statistical tests (i.e. we increased the significance level to 0.05), and with those remaining attributes, we could not train a ML model targeting gender that exceeded the accuracy of the majority class. Because of this, these are the attributes that we will use for creating a final ML model for predicting the class (credit risk) in the German Credit dataset: *foreign*, *other_installment_plans*, *savings*, *status_current_account*, *other_debtors* and *present_residence_since*. Note that *foreign* can also be considered a sensitive attribute. However, for this paper's purpose, we have not taken it into account.

5.3 Training and evaluating models

After analysing the covariance and seeing how it influence the final ML models targeting the class (credit risk), we trained three different ML models to make comparisons of ML models that include (a) sensitive attributes, (b) that do not include sensitive attributes and (c) that do not include other indirect representations of the sensitive attributes in order to see the influence of the covariance analysed before on performance and fairness metrics. Thus, the models are:

- **Model with all attributes**: ML model in which all available attributes have been used for training, including the sensitive attribute, gender. This ML model aims to look at the levels of bias when using the sensitive attribute.
- **Model without gender**: ML model trained with all attributes except for our sensitive attribute, gender. The objective of this ML model is to see how bias is maintained even when the sensitive attribute is removed because other attributes indirectly represent it.
- **Model without gender related attributes**: ML model trained using only the attributes considered gender-independent by the Chi-Square and Mann-Whitney U statistical tests. This ML model aims to see if the bias is reduced by removing the attributes related with the sensitive attribute to eliminate that indirect representation.

As ML models, we used XGBoost on both datasets, as these are the ML models with which we obtained the best performance when predicting credit risk. In addition, to ensure reliable results, we ran 100 different trainings for each type of ML model to obtain the mean and standard deviation of the accuracy and fairness metrics as, after 100 trainings, no significant changes were found in the average results. For fairness, we use some of the metrics explained in [45] as their definitions are in line with the kind of fairness we would expect for a model that predicts credit risk. Thus, the metrics we use are:

- **Overall accuracy equality [3]**: It is the difference between the model accuracy between the sensitive attribute groups values, also called privileged and unprivileged groups. In our case, between males and females.
- **Equal opportunity difference [20] or false-negative rate difference [11]**: The difference between the model FNR ($\frac{FN}{TP+FN}$) or sensitivity ($\frac{TP}{TP+FN}$) between the unprivileged and privileged groups.
- **False-positive error rate balance [11]**: The difference between the model FPR ($\frac{FP}{FP+TN}$) between the unprivileged and privileged groups.
- **Equalised odds [20]**: It combines the previous two metrics and is calculated as the average of the absolute difference in FPR ($\frac{FP}{FP+TN}$) and sensitivity ($\frac{TP}{TP+FN}$) between the unprivileged and privileged groups.
- **Predictive parity difference [11]**: The difference between the model precision ($\frac{TP}{TP+FP}$) between the unprivileged and privileged groups.
- **Statistical parity difference [16]**: The difference in the probability of being assigned to the positive class between the unprivileged and privileged groups.

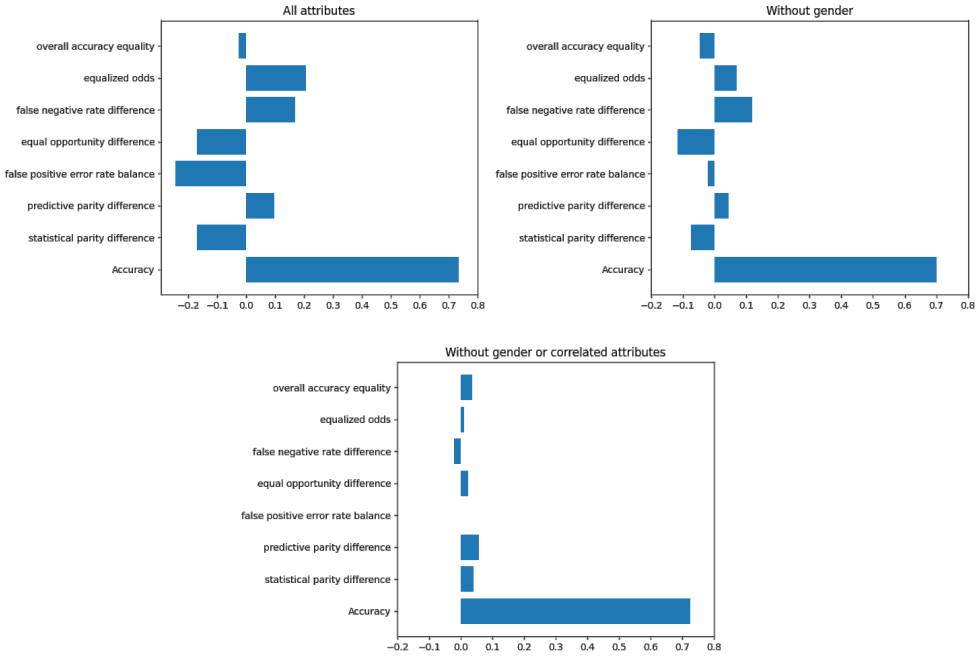

Figure 6: Fairness models comparison German Credit dataset

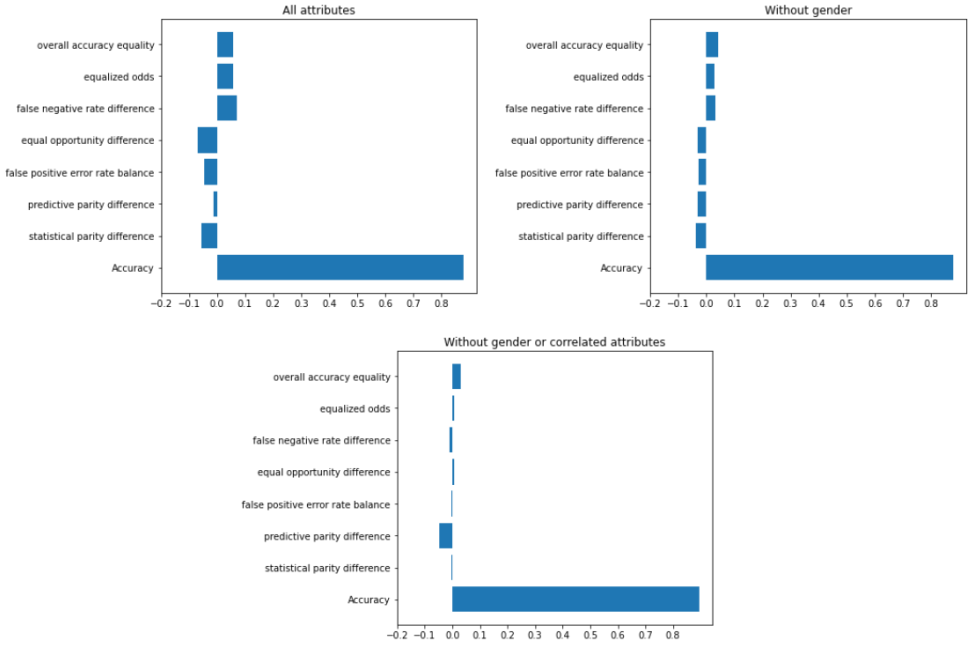

Figure 7: Fairness models comparison Home Credit dataset

Note that these metrics evaluate the fairness of a classification ML model, so they need a predicted label (the prediction of our model) and ground truth (the Class label of our data, credit risk). Also, it is important to mention that not all of these definitions of fairness can be achieved at the same time, as there are trade-offs between some of them [17].

The fairness metric results, as well as the accuracy of the models, can be seen in Figure 6 for the German Credit dataset and in Figure 7 for the Home Credit dataset.

As can be seen in Figures 6 and 7, the ML model trained with all attributes targeting credit risk (including gender and its related attributes) has significantly higher bias values in most metrics. Furthermore, removing only gender does not solve the problem (although it reduces bias in the German Credit dataset and it slightly reduces bias in the Home Credit dataset). On the other hand, removing gender and its related attributes shows a considerable decrease in most metrics. For example, the statistical parity difference improves from -0.18 to 0.03, indicating that we have gone from having 18% more males than females receiving good credit scores to a much smaller gap of 3% in favour of females. These results demonstrate how removing attributes that indirectly represent the sensitive attribute can help improve fairness results. However, we also see how some metrics, such as predictive parity (in the German Credit dataset), can have significantly higher values due to other biases in the model, like technical or algorithmic bias [2, 40].

Moreover, as mentioned in [7], removing correlated attributes is expected to lower the accuracy significantly. However, this is not necessarily always true. As shown in Figure 7, in the ML models trained on the Home Credit dataset targeting credit risk, the bias has been reduced, and at the same time, the accuracy has been increased (from 87% to 89%).

Improving accuracy and reducing bias at the same time might seem unlikely when the test data is biased, as the two goals may appear to be incompatible. To explain why this is happening in our use case let's take the Home Credit dataset as an example. Let's have as the first model, the model that uses all attributes to target credit risk with 87% accuracy and 6% statistical parity. Then, in a second model, when removing the gender-related attributes to target credit risk we have 89% accuracy and 0.04% statistical parity. The first results are worse than the second results because the first model has an accuracy of 83% with males while it has almost 90% accuracy with females. The main reason why the accuracy difference between males and females is high is that 10% of male predictions are false negatives (compared to 5% of female predictions). The bigger number of false negatives makes the first model's accuracy lower and with fewer positive predictions on males (which increases the statistical parity measure). On the other hand, the second model reduces these false negatives, making more correct predictions (which improves the accuracy) and reduces the difference in positive predictions between males and females (which reduces the statistical parity to ideal values close to 0).

Furthermore, in the German Credit dataset, as we show in Figure 8, as we eliminate the attributes most related to gender, the accuracy drops in the case of an XGBoost classifier. However, when using XGBoost with a feedforward search, the accuracy can remain stable (from 77% to 80%) until 75% or more of the attributes are eliminated. However, it is important to note that even within the stable range, the accuracy may also decrease. Nevertheless, with these experiments, we show that the impact on accuracy does not necessarily have to be so significant when using our proposed approach to remove the indirect representation of sensitive attributes that cause the appearance of bias in the ML models.

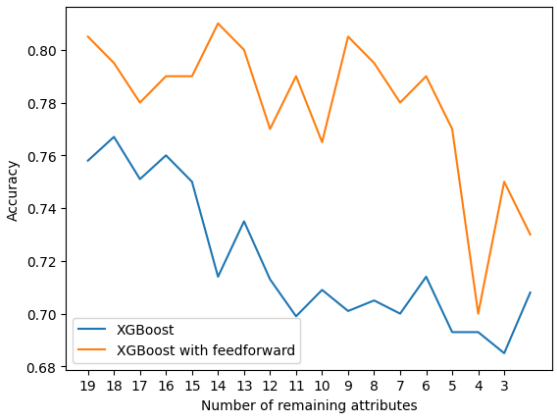

Figure 8: Accuracy deprecation when removing attributes on the German Credit dataset

6 DISCUSSION AND FUTURE WORK

In this paper we prose a novel approach to identify bias in data, particularly the attributes in the data that indirectly represent sensitive information. Data is the starting point for building ML prediction models, and early identification of biases in the data is an important step to avoid their propagation through the ML pipeline. This is relevant in multiple domains of application but particularly in the financial domain, where biases can seriously impact the lives of affected individuals.

The approach proposed in this work uses statistical tests to compute the level of covariance that the attributes in the data present with respect to a selected sensitive attribute. Statistical tests provide a level of confidence/significance helping to measure and explain up to which point these attributes can indirectly represent sensitive information. This is a particularly important aspect of our proposed approach since it not only identifies the concrete attributes within the data, but also provides their confidence levels, which are important insights that can help the developers of these systems to understand where data biases are coming from. The automatic identification of these attributes is key to multiple application domains, but particularly for the detection of credit risk, where hundreds of attributes are considered, and hence biases can be encapsulated by a wide range of information.

In this paper we have shown how gender, as selected sensitive attribute, is embedded in traditional credit scoring processes from a priori non-sensitive attributes. When analysing the two selected datasets we observed that those attributes encapsulating biases in the past are still present in modern online banking alongside a large set of new, and potentially problematic, attributes. Identifying these relationships in the data is not only relevant to avoid that data biases propagate within the construction of ML models, but our approach can also be a relevant tool to study how biases are represented in our past and present societies. We have seen in our experiments how elements like owning a car or working far from home, are still proxies to identify gender.

In terms of the domain under study, it is important to highlight that the datasets analysed in this work do not reflect the fullness of the financial sector, but a small sample of it. The German credit

dataset has a reduced set of individuals from one country. The home credit dataset, although much richer in terms of information, covers individuals who do not usually have a credit history in developing countries. While a wider range of financial services datasets should be used for a deeper analysis and understanding of the domain of application, public data is very difficult to obtain due to the sensitive nature of financial information. An important step of our future work is therefore the use of simulated data to study and reflect on how data biases could be identified through different scenarios.

In terms of bias mitigation, our approach allows us to identify exactly the attributes that need to be removed so that the sensitive attribute can not be predicted from other attributes. As we have observed from our experiments, this allows us to modify the data while preserving sufficient information to create ML models that balance performance and fairness. This effect is non-trivial since removing information from the training data generally leads to drops in the performance of ML models. However, as shown in our experiments, when removing the identified attributes the performance of the ML models for both datasets not only did not diminish but slightly improved. While this is an encouraging result more research needs to be conducted to study how performance and fairness can and should be balanced within financial systems.

It is worth mentioning that our method of bias identification could be used in datasets with more than one sensitive attribute. When it comes to mitigating bias the approach should slightly be adapted to select which attributes to eliminate. For these cases, weights to the sensitive attributes could be assigned as well as making a weighted average of the covariance metrics. The study of intersectionality, and how our approach could target multiple sensitive attributes simultaneously is part of our future work.

7 CONCLUSIONS

In this paper, we propose a novel approach to identify and mitigate bias in data by analysing the covariance between a sensitive attribute (e.g. gender) and all other attributes available within the data. Our approach specifically identifies which attributes indirectly represent a sensitive attribute with a level of confidence and removes them as a way to mitigate bias. Experiments conducted on two datasets from two different financial institutions show how our proposed approach helps to improve fairness while maintaining the overall performance of the ML models that predict individuals' credit risk.

REFERENCES

[1] Steve Austin, Richard Schwartz, and Paul Placeway. 1991. The forward-backward search algorithm. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 697–700.

[2] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[4] Kirsten MM Beyer, Yuhong Zhou, Kevin Matthews, Amin Bemanian, Purushottam W Laud, and Ann B Nattinger. 2016. New spatially continuous indices of redlining and racial bias in mortgage lending: links to survival after breast cancer diagnosis and implications for health disparities research. *Health & place* 40 (2016), 34–43.

[5] Concha Bielza and Larrañaga Pedro. 2021. *Data-driven computational neuroscience: machine learning and statistical models*. Cambridge University Press, 226–261.

[6] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*. 642–653.

[7] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data mining and knowledge discovery* 21, 2 (2010), 277–292.

[8] CDEI. 2020. *Review into bias in algorithmic decision-making*. https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making

[9] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in machine learning software: why? how? what to do?. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 429–440.

[10] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40, 1 (2014), 16–28.

[11] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[12] Harald Cramér. 1999. *Mathematical methods of statistics*. Vol. 43. Princeton university press.

[13] Home Credit. 2018. *Home Credit Default Risk*. Technical Report. https://www.kaggle.com/competitions/home-credit-default-risk/

[14] Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Muhammad Bilal Zafar. 2021. Fairness Measures for Machine Learning in Finance. *The Journal of Financial Data Science* 3, 4 (2021), 33–64.

[15] Asli Demirgüç-Kunt and Ross Levine. 2009. Finance and inequality: Theory and evidence. *Annu. Rev. Financ. Econ.* 1, 1 (2009), 287–318.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[17] Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 3662–3666.

[18] Mark Andrew Hall et al. 1999. Correlation-based feature selection for machine learning. (1999).

[19] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, et al. 2021. Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. *arXiv preprint arXiv:2109.03285* (2021).

[20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[21] Bertrand K Hassani. 2021. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics* 1, 3 (2021), 239–247.

[22] Hans Hofmann. 1994. *German Credit Dataset*. Technical Report. Hamburg University. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[23] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[24] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. IEEE, 869–874.

[25] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 35–50.

[26] Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 247–254.

[27] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[28] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. 2019. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2847–2851.

[29] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[31] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. 2021. Exacerbating algorithmic bias through fairness attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8930–8938.

[32] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[33] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.
[34] Florian Ostmann and Cosmina Dorobantu. 2021. AI in financial services. *Alan Turing Institute. doi* 10 (2021).
[35] Rohan Paris. 2021. *Credit Score Classification.* Technical Report. https://www.kaggle.com/datasets/parisrohan/credit-score-classification
[36] Robin L Plackett. 1983. Karl Pearson and the chi-squared test. *International statistical review/revue internationale de statistique* (1983), 59–72.
[37] J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning* 1, 1 (1986), 81–106.
[38] Edward Raff, Jared Sylvester, and Steven Mills. 2018. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 243–250.
[39] Ms Ratna Sahay, Mr Ulric Eriksson von Allmen, Ms Amina Lahreche, Purva Khera, Ms Sumiko Ogawa, Majid Bazarbash, and Ms Kimberly Beaton. 2020. *The promise of fintech: Financial inclusion in the post COVID-19 era.* International Monetary Fund.
[40] Sebastian Schelter and Julia Stoyanovich. 2020. Taming technical bias in machine learning pipelines. *Bulletin of the Technical Committee on Data Engineering* 43, 4 (2020).
[41] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.
[42] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 358–364.
[43] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. Lift: A scalable framework for measuring fairness in ml applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2773–2780.
[44] Sahil Verma, Michael Ernst, and Rene Just. 2021. Removing biased data to improve fairness and accuracy. *arXiv preprint arXiv:2102.03054* (2021).
[45] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.
[46] Neil Vigdor. 2019. Apple card investigated after gender discrimination complaints. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html
[47] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.
[48] Yukun Zhang and Longsheng Zhou. 2019. Fairness assessment for artificial intelligence in financial industry. *arXiv preprint arXiv:1912.07211* (2019).
[49] Frederik J Zuiderveen Borgesius. 2020. Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights* 24, 10 (2020), 1572–1593.