Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva

Alexander Stevens
Research centre for information systems engineering (LIRIS)
KU Leuven
Leuven, Belgium
0000-0001-6140-8788

Peter Deruyck
Research centre for information systems engineering (LIRIS)
KU Leuven
Leuven, Belgium
peter.deruyck@student.kuleuven.be

Ziboud Van Veldhoven
Research centre for information systems engineering (LIRIS)
KU Leuven
Leuven, Belgium
0000-0001-6013-7437

Jan Vanthienen
Research centre for information systems engineering (LIRIS)
KU Leuven
Leuven, Belgium
0000-0002-3867-7055

***Abstract*—Artificial Intelligence is finding its way to ever more applications. Nonetheless, it is increasingly required that decision-making procedures must be explainable and fair. As many applications are based on black-box models, there is a strong need for more explainable AI algorithms. For this reason, our paper explores the practical implications and effectiveness of four bias mitigation algorithms (learning fair representations, reweighing, Equality of Odds, and Reject Option based Classification) based on a standard XGBoost classifier to build an explainable and fair prediction model on a real-world loan dataset. The models were evaluated based on their performance, fairness, and explainability. Potential biases, i.e. fairness, were detected with the use of NLP techniques and evaluated with the AIF360 metrics, whereas the explainability of the model was tested by post-hoc explanations (SHAP method). The best results were obtained by the reweighing algorithm that improved the fairness while maintaining a high model performance and explainability.**

Keywords— explainable artificial intelligence, XAI, machine learning, fairness

I. Introduction

Artificial Intelligence (AI) is one of the main drivers of digital transformation in recent years. However, apart from the range of opportunities in terms of model optimisation and innovation, there must be some caution when implementing AI algorithms. Many AI applications generate outcomes without justified explanations of how the result is obtained. This lack of explainability in machine learning is closely related to the adjacent lack of fairness. For example, an algorithm can be unfair due to human biases in its input data. Although several studies have already examined methods to address fairness and explainable AI (XAI) on a theoretical level, less attention has been devoted to the effectiveness of these methods on real-world datasets.

The overall objective of this study is to provide insights into the performance of fair and explainable algorithms on a real-world dataset. The investigated data originates from Kiva.org, a non-profit microfinance institution that developed a platform to lend money to low-income entrepreneurs. The platform consists of three major players. Firstly, the borrowers who make loan requests. Secondly, the field partners who pre-disburse the loan to the borrower, while relying on the funding made through the Kiva platform. Third, the lenders who fund the loan requests on the Kiva platform and who are eventually paid back by the borrower through the field partner. We evaluate several algorithms for a fair and explainable recommendation system that predicts whether a loan request on Kiva.org will be completely funded or not. This system can improve the allocation of funds and therefore reduce the number of expired loans. Furthermore, the predictions can indicate the risk of a loan request which could be used by kiva to determine the priority of the loan request. The key research questions are as follows: (1) Which biases are incorporated in the Kiva dataset? (2) How can we accurately predict whether a loan will be fully funded by using a fair and explainable recommendation system? (3) How does the fair and explainable recommendation system perform compared to a traditional recommendation system without a fairness adjustment?

The remainder of this paper is structured as follows. In Section II, an overview is given about explainability and fairness in machine learning. Section III explains the methodology used to construct five recommendation models. In section IV, we discuss the performance, explainability, and fairness of each model. The different models, whereby the baseline model served as the benchmark, are compared in Section V. Finally, we give a summary of the key findings in section VI.

II. Explainability and Fairness

A. The Black Box Model Discrimination

The current generation of AI applications often relies on black-box models such as deep-learning networks. These models can obtain high performance but are often unable to explain their outcomes to end-users and therefore do not guarantee fairness-aware decisions. As the use cases of AI increase, so does the need for models with more transparency and fairness [1]–[3]. For example, a recent study found that black defendants were significantly more incorrectly labelled as recidivism risks compared to white defendants, despite both groups being predicted with similar accuracy [1].

In the case of a loan funding process, AI applications must be explainable to effectively utilize and trust the outcomes. For instance, gender or race should not be a defining feature to determine if a loan will be funded or not. Previous studies already demonstrated that AI substantially inherits biases exhibited by humans in text documents [4], [5], e.g. in the loan description. As a result, XAI is needed to advance data-driven decision-making processes [6].

978-1-7281-2547-3/20/$31.00 ©2020 European Union

B. Model Explainability

In the domain of text classification, several studies proposed the usage of NLP techniques to determine the most important words in a document to detect biases and gain insights [5], [7]. As such, the explainability of the classification process is exemplified by tracing the classification decision back to individual words. One such technique is the use of term frequency-inverse document frequency (TF-IDF) scores that make a distinction between relevant and irrelevant words. To overcome the inability of the TF-IDF scores to quantify the similarity between two words [8], a combination with the Word2vec model is recommended [9]. One of the main learning algorithms of a Word2Vec model, the continuous bag-of-words (CBOW), tries to predict a target word based on its context [10]. Thus, words without a semantic difference will be taken together before calculating the TF-IDF scores.

Another way of enhancing the explainability of a model is by providing post-hoc explanations that describe the relationships of the inputs with the output to understand how the model works [1], [11]. This can also be done without interpreting the internal working of the model [1], [3], [11]. The SHapley Additive exPlanations (SHAP) method [2] is based on the concept of SHAP values, calculated for each instance-feature combination whereby the calculation is based on coalitional game theory [12]. The goal of the SHAP values is to explain the difference between the prediction of an instance and the average predicted probability by determining which feature values had the highest marginal contribution [2], [12].

C. AIF360 Fairness Metrics

The concept of fairness in machine learning is often divided into two categories: individual fairness and group fairness [13], [14]. Individual fairness denotes the level of difference in predictions for similar instances, whereas group fairness indicates the equal treatment of different groups. The work of Bellamy et al. [15] introduces a new open-source Python toolkit for algorithmic fairness called AI Fairness 360 (AIF360). The AI360 library provides metrics, shown in Table 1, to measure both individual and group fairness. For individual fairness, the consistency metric is used that compares the prediction of a certain individual with the predictions of its k-nearest neighbours [16].

TABLE I. FAIRNESS METRICS

Metric	Fairness Condition	Fairness Level
Statistical Parity	-0.10 ≤ fair ≤ 0.10	Group
Equal Opportunity	-0.10 ≤ fair ≤ 0.10	Group
Average Absolute Odds	-0.10 ≤ fair ≤ 0.10	Group
Disparate Impact	0.80 ≤ fair ≤ 1.20	Group
Theil Index	0 ≤ fair ≤ 0.25	Individual and Group
Consistency	0.90 ≤ fair ≤ 1.10	Individual

For group fairness, four metrics are commonly applied. Firstly, the Equal Opportunity Difference metric measures the difference in true positive rates (TPR) between an unprivileged group and a privileged group. Secondly, the difference between false-positive rates (FPR) and TPR between an unprivileged and a privileged group can be measured with the Average Absolute Odds Difference [17]. Thirdly, the Disparate Impact compares the proportion of individuals who receive a positive output for two groups: an unprivileged group and a privileged group [18]. Fourthly, the Statistical Parity Difference (SPD) calculates the difference in the probability of favourable outcomes between the unprivileged group and the privileged group [16]. The library also contains the Theil Index, a subclass of the generalised entropy index [14], which is a metric that captures both group and individual fairness.

D. Bias Mitigation Algorithms

A first strand of the literature suggests the use of fair pre-processing algorithms. Fair pre-processing modifies the input data to reduce the bias against the unprivileged group which was discovered with the fairness metrics. Afterwards, an explainable classifier can be trained on this modified data resulting in a transparent and fair model. Several algorithms are already developed for fair pre-processing [16], [18]–[20].

The Learning Fair Representations (LFR) is a pre-processing algorithm that transforms the input variables to a latent representation existing out of K clusters [16]. These clusters ensure individual and group fairness for the protected attribute. Afterwards, this latent representation is mapped to the response variable. In this way, the algorithm constructs a fair representation of the data by modifying the input variables and the response variable. The LFR algorithm minimises an objective function consisting of three components. The first component L_x measures the loss of information by mapping the input variables to a latent representation. The second component L_y quantifies the error made by mapping the latent representation to the response variable. The third component L_z ensures that each cluster complies with the notion of SPD. Each of these components has a corresponding weight which can be adjusted. As such, these hyperparameters A_x, A_y and A_z make trade-offs between each component as can be seen in the following formula [16]:

$$L = A_x.L_x + A_y.L_y + A_z.L_z \tag{1}$$

Compared to the LFR technique, less complex pre-processing techniques exist which resample or reweigh the dataset instead of using latent representations [20]. For example, the reweighing method assigns weights to the training set to mitigate bias [20]. This implies that certain instances from a privileged group, which are more likely to have a favourable outcome, will get a lower weight while instances from an unprivileged group will get a higher weight. The calculation of these weights is based on two probabilities. Firstly, the expected probability P_{exp} is calculated based on the situation where the protected attribute S with groups b_k does not affect the class outcome, and vice versa [20]:

$$P(S = b_k \cap class = +) = \frac{|S=b_k|}{|D|} \times \frac{|class=+|}{|D|} \tag{2}$$

978-1-7281-2547-3/20/$31.00 ©2020 European Union

Secondly, the observed probability P_{obs} of a dataset D measures the probability of being a member of the group b_k and having a favourable outcome [20]:

$$P(S = b_k \cap class = +) = \frac{|S=b_k \cap class=+|}{|D|} \quad (3)$$

Instances from a certain group b_k will get a higher weight when the expected probability is higher than the observed probability because this group perceives a bias.

The second strand of literature focuses on fair in-processing algorithms whereby a constraint or a regularisation term is added during the optimisation at training time [21]–[23]. This constraint must guarantee the group fairness between a privileged and an unprivileged group, but an even stronger restriction would amount to an additional individual fairness constraint. Although these algorithms can achieve both high model performance and a high degree of fairness, they are less suitable for explaining and interpreting the results of the model.

The third strand of literature consists of the algorithms which use fair post-processing [17], [24]. Unlike the prior discussed studies, post-processing algorithms do not require access to the training process of a classifier. These algorithms are applied after a classifier has been trained on a training set. Next, the predictions made by the classifier are transformed by the fair post-processing algorithms to ensure fairness. What distinguishes post-processing from in-processing is that it can be used with any model. This ensures that the internal operation of an explainable model is still usable, whereas in-processing algorithms require an adjustment of the optimisation process.

The Equality of Odds (EQO) method [17] keeps the original training process but adds a post-learning step. This step tries to build an equalized odds or equal opportunity predictor from a possibly discriminatory learned binary predictor which was obtained by the existing training pipeline. The concept of equalized odds is related to the average absolute odds difference whereby it enforces both equalised TPR and FPR across an unprivileged group and a privileged group for a derived predictor $\tilde{Y}$ [17]. The algorithm uses a linear program to find the trade-off between FPR and TPR that optimises the expected loss between fair predictions and the true responses while satisfying equalized odds.

A similar post-processing fairness technique called Reject Option based Classification (ROBC) [24] gives favourable outcomes to the unprivileged group and unfavourable outcomes to the privileged group in a confidence band around the decision boundary. Kamiran, Karim, and Zhang [24] defined a critical region that contains all the instances labelled as rejected if their posterior probabilities are close to 0.5. For a classification problem, the critical region can be defined for all instances $x \in X$ for which $\max[P(Y = 1 \mid x), 1 - P(Y = 1 \mid x)] \leq \theta$ $(where\ 0.5 \leq \theta \leq 1)$ with label $Y \in \{0,1\}$ [24].

Generally, the loss associated with misclassified instances is equal for both positive and negative labels. This is modified by the ROBC method for the instances belonging to this critical region. The method assigns a higher cost to unprivileged groups who receive a negative label compared to those receiving a positive label. Similarly, the privileged groups who receive a positive label will get a higher cost compared to those receiving a negative label. This ensures less discrimination while still achieving a high level of accuracy [24].

III. Methodology

A. Kiva Dataset

The Kiva dataset consists of 614,010 loan instances each having 34 features. The target variable *status,* which explains whether the loan request was completely funded or not*,* has two categories: funded and expired. This results in a highly imbalanced dataset with 579,041 funded loans and 34,969 expired loans.

The independent features give information about the loan requests, the borrowers, and the field partners. TABLE II. shows a partial list of the most important features in the dataset, together with a short description. The feature *description_ENG* is a string variable which represents the texts published on the Kiva site. These descriptions consist of two to fifteen sentences in which a borrower describes some general information about themself along with the purpose of the loan. The feature *borrower_genders* gives for each loan a list with the genders of each borrower. For example, an instance with three female borrowers would contain the list 'female,female,female'. The feature *distribution_model* indicates if a loan was administered by a field partner or not. Loans without a field partner are called direct loans and are made through a digital payment system. The focus of the analysis in this paper lies on the loans distributed through a field partner.

TABLE II. Original Input Features

Variable name	Variable description	Type
Loan_amount	Total dollar amount of a loan request	Numeric
Funded_amount	Total dollar amount of funds made by lenders	Numeric
Repayment_interval	Four categories: irregular, monthly, bullet and weekly	Nominal
Activity	Loan activity type	Nominal
Sector	Sector of loan activity as shown to lenders	Nominal
Country	Country of borrower	Nominal
World region	World region of borrower	Nominal
Borrower_genders	Gender of borrower(s)	Nominal
Use	One sentence explaining loan use	String
Description_ENG	Borrower description on site	String
Distribution_model	Field partner or direct	Nominal
Partner_id	Unique ID for field partners	Nominal
Field partner name	Name of field partner	Nominal

978-1-7281-2547-3/20/$31.00 ©2020 European Union

B. Recommendation System Development

The methodology applied in this paper consists of four steps. In the first step, pre-processing, we transform several original features to enhance the analysis. The feature *gender_reclassified* was created by transforming the original feature *borrower_genders*. To investigate a gender bias, the feature *gender reclassified* needs to have a single-gender label for each loan. For group loans, the gender of the group leader was used. The feature *loan_amount_bin* was created based on a quantile binning method by cutting the feature *loan_amount* into roughly equal groups using vigintiles. This was done because the LFR algorithm cannot handle continuous features on a large scale.

In the second step, we construct an NLP model to determine whether the lenders were prejudiced when they allocate their funds. Additionally, the results of the NLP model are used for the AIF360 algorithms, as explained in step three. We vectorize the loan descriptions based on the TF-IDF method before building a text classifier which reveals the most important words to classify a loan based on the model coefficients. In addition, a principal component analysis (PCA) was used to reduce the high-dimensional TF-IDF vectors to two components to find out the maximum amount of variance, and therefore the quality of the NLP model [25]. PCA was chosen over other dimensionality reduction algorithms such as t-SNE and LDA as it preserves the distance and maximizes the explained variance [26]. The TF-IDF vectors were built on a balanced sample of 20,000 loans that was constructed by randomly sampling 10,000 funded loans and 10,000 expired loans.

Before vectorizing this balanced sample, several pre-processing steps were required. First, we removed special characters such as punctuations, backslashes, or HTML code. Next, lemmatization was used to derive the common base of a word. Subsequently, several irrelevant but frequently used words in the descriptions were removed based on a predefined list obtained from the NLTK library in python. The gender-related words were removed from this list to reveal a possible gender bias. The last step used the CBOW learning algorithm of the word2vec model to combine similar words [27] which have at least 1,000 occurrences in the dataset. Finally, the balanced sample was split into a training set to train a text classifier using logistic regression and a test set to assess the performance. The training set was created by randomly sampling 80% of the data, while the other 20% of the data was kept for testing. The performance of the NLP model was evaluated on the test set of 4,000 loans based on the accuracy, precision, recall and the Area Under the Curve (AUC).

In the third step, we constructed the baseline model, based on XGBoost, and the AIF360 algorithms which were used to mitigate the biases from the baseline model. Because the XGBoost model cannot handle categorical values, each variable had to be converted with one-hot encoding. Next, the build-in feature selection of the XGBoost model allowed us to distinguish irrelevant features from the most relevant ones. This list of relevant features was further reduced by looking at the accuracy change when performing feature removal. This step was preferred over an exclusive feature selection method because the incorporation of domain knowledge made it possible to choose more interpretable features. To compare the performance of the fair and explainable recommendation algorithms to a baseline model without fairness adjustments, an XGBoost classifier was used to construct a baseline model based on the feature selection results. This original baseline model was not adjusted to mitigate any biases because it serves as a benchmark to evaluate the four fairness adjusted models.

Each AIF360 algorithm requires the specification of a protected feature with a privileged group and an unprivileged group. Based on the results of the NLP model, this protected feature was defined by the feature *gender_reclassified_female*. The loans labelled as female were specified as the privileged group, while the unprivileged group was defined by the loans labelled as male. Each algorithm was constructed based on the general pipeline developed by Bellamy et al. [15]. The fair pre-processing algorithms, reweighing and LFR, were applied to modify both the training and test sets to mitigate the biases. In addition, the EOQ and the ROBC algorithm served as the fair post-processing algorithms which adjust the predictions made by the baseline model to ensure fairness. In summary, step three produces five models: a baseline model based on XGBoost, LFR, reweighing, EQO, and ROBC.

Downsampling was used to overcome the class imbalance problem which resulted in a balanced sample of 69,938 loans [28]. Afterwards, stratified 10-fold cross-validation was used to shuffle and split the balanced sample into ten folds which preserve the percentage of samples for each class [29]. For each fold, the test set is evaluated based on accuracy, precision, recall, and F1-score. By averaging these evaluation metrics across the 10 folds, we provide an accurate performance overview of the five models.

In the fourth step, the explainability and fairness of each model obtained from step three were investigated. For the baseline model and the fair pre-processing algorithms, the explainability was assessed by using feature importance plots based on the SHAP method. The SHAP method could not be used for the fair post-processing algorithms (EQO & ROBC), as these construct their own rules to make fair predictions. As such, the effect of the EQO algorithm was observed by looking at the confusion matrices of the two genders both before and after the application of the algorithm. For the ROBC algorithm, the transformation from biased predictions to fair predictions was clarified. In addition, the fairness of each model was assessed based on the fairness conditions defined in TABLE I.

IV. Results

A. Bias Detection Based on NLP

Before fitting a text classifier using logistic regression, the TF-IDF vectors were inspected based on Fig. 1. Here, the PCA decomposition of the obtained TF-IDF vectors is shown. The PCA indicates both some key distinctions and an overlap between the two categories. The overlap reveals that some words appeared in both categories decreasing the accuracy of the text classifier. The NLP model resulted in an accuracy of 75.5%, a precision of 76%, and a recall of 76% whereby all three metrics indicate a well-performing model. Moreover, the ROC curve was created with a corresponding AUC value of 83%. As such, the NLP model has high discriminative power between the descriptions of funded loans and expired loans

978-1-7281-2547-3/20/$31.00 ©2020 European Union

For the funded loans, the top five most important unigrams and/or bigrams are: she, philippines, enable she, widow, and fee. On the other hand, the top five of the expired loans consist of: he, store, merchandise, work, and english. External analysis, out of the scope of this research, showed that the female Filipino borrowers represented more than 20% of the loans, which can influence the top-funded loans. For the word 'he', a possible explanation remains unclear, therefore these results imply a significant difference in male and female loans being funded, which is not desired.

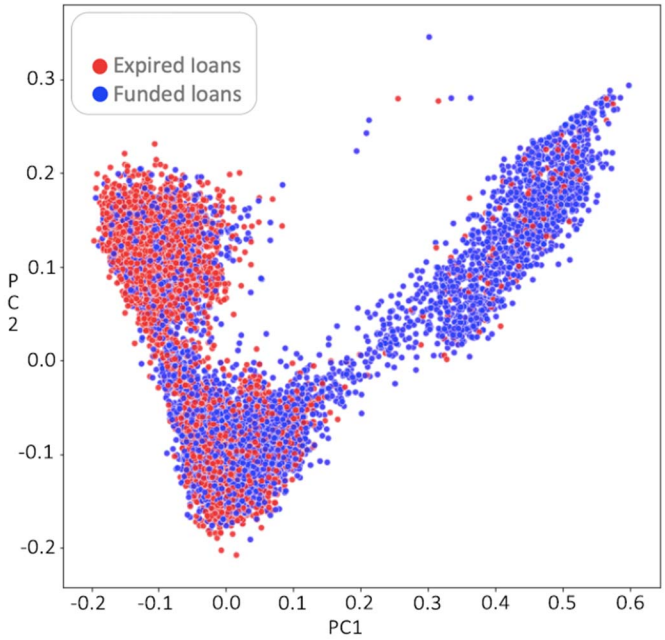

Fig. 1. PCA decomposition of top 500 features

B. *Baseline Model*

For each model, the classification results are depicted in TABLE III. , and the AIF360 fairness metrics are shown in TABLE IV. The results that satisfy the fairness conditions are marked with *, while the highest score is boldfaced. The standard deviation for each performance metric is added in brackets in TABLE III. The baseline model obtained an accuracy of 0.7501, a recall of 0.6370, a precision of 0.8208, and an F1 score of 0.7190. The recall-precision comparison indicates that more funded loans are misclassified compared to expired loans. In addition, the high precision reveals that many predicted funded loans are truly funded. Looking at TABLE IV. four metrics do not comply with the fairness conditions. Only the Theil index and the consistency metric indicate a fair model, meaning that the baseline model has high individual but low group fairness.

TABLE III. PERFORMANCE OF EACH MODEL

	Baseline	**LFR**	**Reweighing**	**EQO**	**ROBC**
Accuracy	**0.7501** (±0.0046)	0.6605 (±0.0064)	0.7411 (±0.0054)	0.6058 (±0.0072)	0.6915 (±0.0069)
Recall	0.6370 (±0.0094)	0.5838 (±0.0097)	0.6358 (±0.0063)	0.3226 (±0.0203)	**0.6575** (±0.0699)
Precision	**0.8208** (±0.0056)	0.6896 (±0.0093)	0.8054 (±0.0082)	0.7442 (±0.0091)	0.7111 (±0.0414)
F1-score	**0.7190** (±0.0067)	0.6322 (±0.0070)	0.7106 (±0.0060)	0.4497 (±0.0198)	0.6791 (±0.0237)

TABLE IV. AIF METRICS RESULTS FOR EACH MODEL

	Baseline	**LFR**	**Reweighing**	**EQO**	**ROBC**
Statistical Parity	-0.4431	-0.0945*	-0.1188	**-0.0484***	-0.0492*
Equal Opportunity	-0.3977	**0.0051***	-0.1728	0.0741*	0.0917*
Average Absolute Odds	0.3221	**0.0029***	0.1212	0.0387*	0.0934*
Disparate Impact	0.1924	0.8249*	0.7221	0.7880	**0.8634***
Theil Index	**0.2260***	0.0024*	0.2296*	0.4417	0.2923
Consistency	0.9361*	**0.9960***	0.9257*	0.7771	0.9198*

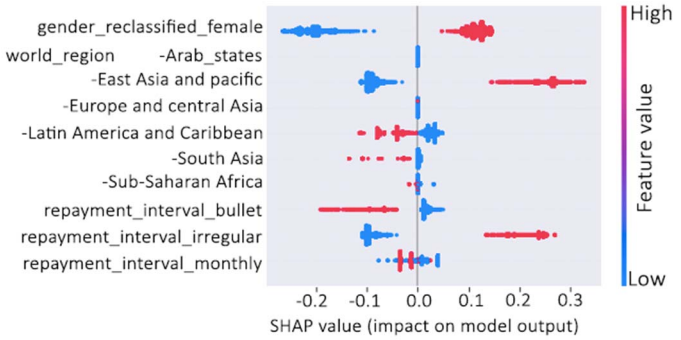

Fig. 2. Baseline model summary plot

The equal opportunity of the baseline model shows a negative result of -0.3977 which means that more funded loans of female borrowers are correctly classified as funded compared to the loans of male borrowers. By comparing the result of the equal opportunity metric with the average absolute odds difference, it can be seen that the FPR difference is 0.2465. These results indicate that the baseline model does not obtain an equal performance in predicting the loan outcomes of both female and male borrowers. The SPD and disparate impact metrics indicate that loan requests from female borrowers are more likely to be predicted as completely funded

To explain the predictions of the baseline model, a global feature importance plot was used based on the SHAP method which revealed that the features *repayment_interval*, *world_region*, *loan_amount,* and *gender_reclassified_female* are the most important to predict a loan outcome. The summary plot, shown in Fig. 2 indicates that the feature value female leads to a higher probability of having a funded loan according to the SHAP values. These findings confirm the results obtained from the AIF360 fairness metrics. Two more feature values that lead to a higher probability of obtaining a completely funded loan can be distinguished *repayment_interval_irregular* and *world_region_East_Asia and the pacific*.

978-1-7281-2547-3/20/$31.00 ©2020 European Union

C. Learning Fair Representations Model

The results of the bias mitigation algorithms will be compared with the baseline model to assess the relative changes in fairness, explainability, and performance. The default settings of the hyperparameters for the LFR model, $A_x = 0.01, A_y = 1, A_z = 50$ and $K = 5$ resulted in an accuracy of 70% and two out of six fairness metrics still implied a gender bias. Moreover, the SPD and disparate impact metrics were lower compared to those of the original baseline model. The fairness could be improved by adjusting the hyperparameters to $A_x = 1, A_y = 0.21, A_z = 1$ and $K = 5$. As shown in TABLE IV. , this completely mitigates the gender bias because now both the individual and the group fairness metrics satisfy the fairness conditions. Although the gender bias was removed, the accuracy and the precision dropped to 0.6605 and 0.6896 respectively. The recall remained stable at 0.6358 and the F1 score dropped to 0.6322. Due to the lack of compatible AIF360 code to perform an exhaustive grid search along with the required computing time, these results cannot be seen as the optimal settings as the findings are based on a limited combination of parameter values.

To better understand the internal working of the LFR model, a summary plot is shown in Fig. 3 whereby the dummy variables were not aggregated because of the latent representations which transformed the original binary variables to continuous variables. Nevertheless, the plot can still rank the features according to their importance. Looking at this ranking, one sees that it does not contain the feature *gender_reclassified_female*. As a result, this feature is not considerably used in the XGBoost classifier to predict the outcome of a loan which confirms the findings of the previous section that showed a very fair model. In the LFR model, the dummy variables *sector_name_Food*, *loan_amount_bin_35-40Q*, *loan_amount_bin_10-15Q,* and *world_region_East Asia and the Pacific* have the highest importance to classify a loan. These variables are very different compared to the baseline model, which possibly explains the decline of accuracy. Fig. 3 also depicts the distorting effect where the continuous variables are indicated by purple dots. As such, these variables make the model less appropriate to explain the prediction of a loan outcome.

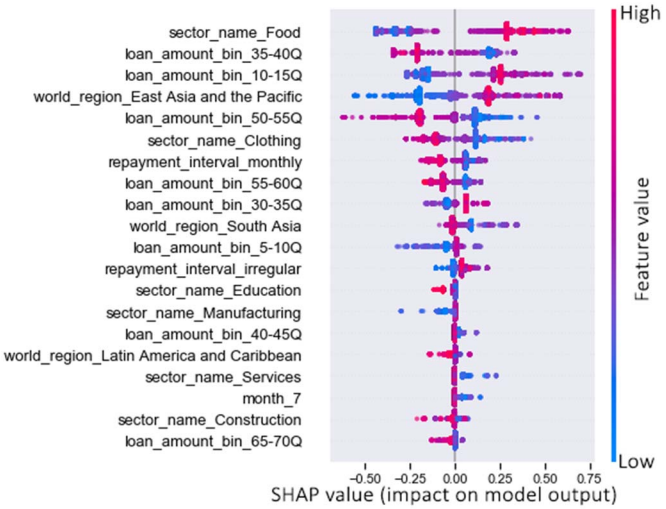

Fig. 3. LFR summary plot

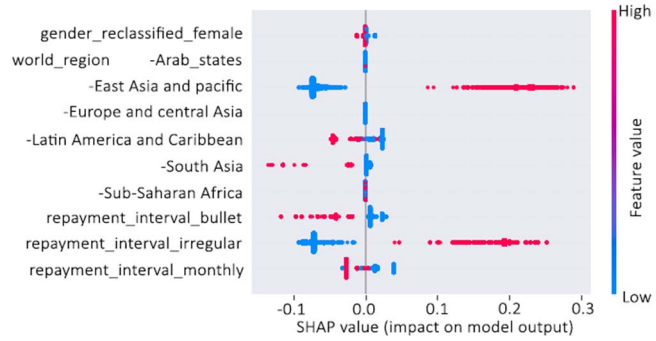

Fig. 4. Reweighing model summary plot for the features gender reclassified female, world region, and repayment interval

D. Reweighing Model

By comparing the reweighing model with the baseline model, it is clear that the individual fairness remained the same, whereas the group fairness improved. nevertheless, four group fairness metrics still do not satisfy the fairness conditions. As shown in TABLE III. , the performance metrics of the reweighing model show an accuracy of 0.7427, a recall of 0.6346, a precision of 0.8076, and an F1 score of 0.7107. Fig. 4 that the feature *gender_reclassified_female* has a negligible impact on predicting a loan outcome. Moreover, the reweighing model classifies based on similar feature values compared to the baseline model. A loan is still more likely to be predicted as completely funded when the loan request has an irregular repayment interval and originates from *world_region_East Asia and the Pacific*.

E. Equality of Odds Model

For the EQO model, three out of six metrics, cf. table III, suggest a gender bias but compared to the baseline model the group fairness has considerably improved. In contrast, individual fairness has declined which implies that loan requests with similar characteristics are treated differently by the model. Moreover, the fairness adjustments come with a performance cost whereby the accuracy decreases to 0.6058, and the F1 score drops to 0.4497. Compared to the baseline model, the recall and the precision dropped to 0.3226 and 0.7442 respectively.

Before the bias mitigation, both the TPR and the FPR were quite different. After the application of the EQO algorithm, these two rates are equalised based on a linear program. To obtain equalized odds, the algorithm has significantly reduced the TPR of women. This means that some loans that have a rather high probability to be funded are predicted as expired only because the borrower is a woman. Intuitively, the algorithm classified each group as poorly as the disadvantaged group.

F. Reject Option Classification Model

For the ROCB, the Theil index suggests an unfair model while the group fairness improved compared to the baseline model. The consistency metric shows a slight decrease in individual fairness and therefore also explaining the result of the Theil index. The results also show a limited trade-off between fairness and accuracy. The accuracy after the application of the algorithm stabilises to 0.6915. As such, the improvement in group fairness only implies a small reduction of accuracy.

978-1-7281-2547-3/20/$31.00 ©2020 European Union

However, the F1 score drops to 0.6791 compared to the baseline model with a recall of 0.6575 and a precision of 0.7111. For the instances lying in the critical region, another decision rule is added by the ROBC algorithm which ensures that the unprivileged male group receives the funded outcome, while the privileged female group receives the expired outcome. As the algorithm makes each loan outcome fair based on the baseline predictions, it is not possible to provide either ante-hoc or post-hoc explanations because the predictions of the ROBC algorithm cannot be linked with the input features.

V. Discussion

Previous studies have emphasized the need for more explainability and fairness in machine learning [1], [13]. The purpose of this paper was to study the concepts of explainability and fairness in machine learning based on a real-world dataset. Our analyses answered three research questions which led to several key findings. Firstly, the NLP model revealed that female borrowers are more likely to be funded. This finding was quantified by AIF360 fairness metrics which showed a high degree of individual fairness but a low degree of group fairness for the baseline model. As a result, we found that the baseline model systematically favoured the loan requests from female borrowers leading to unfair recommendations.

Secondly, four bias mitigation algorithms were compared (LFR, reweighing, EQO, and ROBC) to investigate the performance of fair and explainable recommendation systems. The results show that only the reweighing model could achieve a high level of fairness, explainability, and performance. The low accuracy cost of the reweighing model is explained by the fact that it mainly focuses on group fairness by using a limited number of data modifications. The ROBC model had similar fairness and performance compared to the reweighing model. However, it could not provide post-hoc explanations of the predictions leading to a reduced explainability.

One interesting finding is that the LFR model had a high degree of fairness but a low degree of explainability and model performance. This suggests that the fairness introduced by the latent representations causes a decline in model interpretability and model performance. Moreover, the latent representations ensured individual fairness by modifying each loan instance causing a significant decline of accuracy. The EQO model had the worst model performance and individual fairness among all the models. This model could also not provide explainability by using post-hoc explanations for the predictions.

Thirdly, the four models were compared with the baseline model based on explainability, fairness, and performance to answer the third research question. The results indicate that the reweighing model could obtain a similar explainability and performance while improving fairness. The post-hoc explanations provided by the SHAP method showed that both the baseline model and the reweighing model use similar feature values to predict a loan outcome. The reweighing model also obtained similar results as the baseline model on both the accuracy and F1 score. For both models, the precision was higher than the recall, signifying that the models produced less false positives compared to false negatives. In addition, the reweighing model improved the group fairness, whereas the individual fairness metrics remained at a high level. For Kiva, the reweighing model gives the lenders and the field partners a more accurate view of the loan outcomes. As a result, there is a more efficient allocation of funds and a reduction of the amount of money that flows back to the lenders due to expired loans. This gives the lenders more impact and a higher sense of satisfaction. For the field partners, the recommendations are an indication of the risk of a loan request that could be used to determine the priority of a loan request.

These findings will be of interest to companies engaging in data mining applications where fairness and explainability must be accounted for. Even though the LFR algorithm and both post-processing algorithms are very valuable in theory, they are less capable of incorporating explainability. Extending these methods to improve not only fairness and explainability but also performance would be a fruitful area for future work. Furthermore, more research needs to be conducted about the trade-offs that these models make.

In our research, the bias mitigations techniques were limited to fair pre-processing and fair post-processing algorithms. A possible extension could be to develop a fair in-processing algorithm based on an explainable classifier. Another limitation is that our findings are based on one baseline model. Future research is needed to test the bias mitigation algorithms with other explainable classifiers to measure the effect of the classifier. Third, our study assumed that the true negatives and the false positives have the same misclassification costs. For the Kiva platform, one may argue that false negatives have a higher cost. Therefore, it would also be possible to assign a higher misclassification cost when a funded loan is misclassified as an expired loan, which could increase the recall.

VI. Conclusion

We combined and compared four bias mitigation techniques with the explainable classifier XGBoost based on the performance, explainability, and fairness. The study indicates that the reweighing algorithm could mitigate the gender bias while ensuring model explainability through post-hoc SHAP explanations.

References

[1] Z. C. Lipton, "The mythos of model interpretability," *ACM Queue*, vol. 61, no. 3, pp. 31–57, 2018.

[2] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, pp. 4766–4775, May 2017.

[3] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *arXiv Prepr. arXiv1708.08296*, Aug. 2017.

[4] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science (80-.).*, vol. 356, no. 6334, pp. 183–186, 2017.

[5] L. Arras, F. Horn, G. Montavon, K. R. Müller, and W. Samek, "'What is relevant in a text document?': An interpretable machine learning approach," *PLoS One*,

978-1-7281-2547-3/20/$31.00 ©2020 European Union

vol. 12, no. 8, pp. 1–23, 2017.

[6] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making,” *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[7] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, “Explainable Prediction of Medical Codes from Clinical Text,” *arXiv Prepr. arXiv1802.05695*, 2018.

[8] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018.

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations ofwords and phrases and their compositionality,” *Adv. Neural Inf. Process. Syst.*, pp. 3111–3119, 2013.

[10] X. Rong, “word2vec Parameter Learning Explained,” *arXiv Prepr. arXiv1411.2738*, pp. 1–21, 2014.

[11] A. Barredo Arrieta *et al.*, “Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[12] L. S. Shapley, “A value for n-person games,” *Contrib. to Theory Games*, vol. 2, no. 28, pp. 307–317, 1988.

[13] A. Chouldechova and A. Roth, “The Frontiers of Fairness in Machine Learning,” *arXiv Prepr. arXiv1810.08810*, pp. 1–13, 2018.

[14] T. Speicher *et al.*, “A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 2239–2248.

[15] R. K. E. Bellamy *et al.*, “AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias,” *arXiv e-prints arXiv:1810.01943*, p. 20, 2018.

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

[17] M. Hardt, E. Price, and N. Srebro, “Equality of Opportunity in Supervised Learning,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[18] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing disparate impact,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 259–268.

[19] F. P. Calmon, D. Wei, B. Vinzamuri, K. R. Varshney, and K. N. Ramamurthy, “Optimized Data Pre-Processing for Discrimination Prevention,” in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[20] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012.

[21] M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, “Rawlsian Fairness for Machine Learning,” *arXiv Prepr. arXiv1610.09559*, vol. 1, no. 2, pp. 1–26, 2016.

[22] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, “Fairness-Aware Classifier with Prejudice Remover Regularizer,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012, pp. 35–50.

[23] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, “Fairness constraints: Mechanisms for fair classification,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017, pp. 962–970.

[24] F. Kamiran, A. Karim, and X. Zhang, “Decision theory for discrimination-aware classification,” in *IEEE 12th International Conference on Data Mining*, 2012, pp. 924–929.

[25] C. Ding and X. He, “K -means Clustering via Principal Component Analysis,” *Proc. Twenty-First Int. Conf. Mach. Learn.*, p. 29, 2004.

[26] L. van der Maaten and G. Hinton, “Multiobjective evolutionary algorithms to identify highly autocorrelated areas: The case of spatial distribution in financially compromised farms,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–3605, 2008.

[27] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, “Comparative study of word embedding methods in topic segmentation,” in *Procedia computer science*, 2017, vol. 112, pp. 340–349.

[28] N. V. Chawla, N. Japkowicz, and A. Kotcz, “Editorial: Special Issue on Learning from Imbalanced Data Sets,” *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[29] N. A. Diamantidis, D. Karlis, and E. A. Giakoumakis, “Unsupervised stratification of cross-validation for accuracy estimation,” *Artif. Intell.*, vol. 116, no. 1–2, pp. 1–16, 2000.

978-1-7281-2547-3/20/$31.00 ©2020 European Union