Resampling technique for Bias Mitigation

Giovanni De Lucia, Alessandro Ferri
Prof. Letizia Tanca
Supervisors: Chiara Criscuolo, Tommaso Dolci

May 22, 2023

Contents

1 Introduction

Thanks to the fast evolution of machine learning, data-driven decision-making is quickly growing in popularity, thus raising the obvious question: how trustworthy is the decision made by a system? The system will learn what is represented by the data, but what happens if the data fail to represent a subpopulation? After all, data are just a small sample drawn from a population, it is rather easy to leave out some minorities when sampling the data. This phenomenon is called *Representation Bias*, and it happens when the training data underrepresents some part of a target population. We will focus in particular on the concept of *fairness* of the data, which is crucial to avoid discrimination based on sensitive attributes (gender, race etc.) during the training of the machine learning model.

The purpose of this project is to apply a particular technique to achieve fairness on a biased dataset: the *Resampling technique*, which performs oversampling or undersampling of certain classes to have a more fair dataset to learn on.

2 Project Objectives

Our work can be divided in three steps:

- Reading papers to better understand the data fairness topic
- Searching for libraries that implement bias-mitigation techniques
- Applying one of the techniques (Resampling) to a biased dataset

2.1 Papers

The first task of this project was understanding the main topic: fairness and bias in the data. The papers we read are the following:

- Managing bias and unfairness in data for decision support[1]:
  This paper covers the creation of fairness metrics, fairness identification and mitigation methods, and biases in crowdsourcing activities, it also argues about a new data-centered approach to overcome the limitations of algorithmic-centered methods.
- A Survey on Bias and Fairness in Machine Learning[2]:
  The authors of this paper investigated many real-world applications that have presented bias in some ways, and listed some sources of biases that can affect Artificial Intelligence (AI) applications.
- Representation Bias in Data: A Survey on Identification and Resolution Techniques[3]:
  This paper focuses on categorizing various techniques for bias identification and mitigation in structured and unstructured data.

- Data preprocessing techniques for classification without discrimination[4]: This last paper presents various techniques for bias mitigation, including the pseudo-code to implement such techniques in practice. The DALEX library that has been used for this project uses this paper as a reference for many of its implemented techniques.

These papers provided us with a better understanding of the impact that unfair datasets might have in critical applications, together with some methods to achieve fairness.

2.2 Libraries research

The second task we were given was to search for some open-source libraries that offered bias-mitigation techniques different from the ones that had already been tested by the Professor and her team. Although we found some interesting repositories, most of them involved techniques for NLP (Natural Language Processing) (WEFE) or techniques for bias measurement but not mitigation (FairLens).

After discussing with our Supervisors about the repositories we found, they proposed to search for an implementation of the *Resampling Technique*. We found a library called DALEX, which implements Fairness measurement and mitigation techniques, including Resampling, and, as already mentioned above, it follows the last paper listed in the previous section.

2.3 Resampling technique

The Python code related to this section can be found in the following repository: https://github.com/gio-del/TIS-Project

The last step was to apply the Resampling technique to a biased dataset: the technique works by performing oversampling (adding samples) or undersampling (removing samples) on the different classes of the model in order to have a more balanced distribution of the population. The dataset (Kaggle Diabetes Dataset) has been used to train a machine learning model to classify a patient as diabetic or non-diabetic based on some metrics (age, glucose level, BMI, insuline level etc..). Since we did not have a strong knowledge in Machine Learning or Applied Statistics, we had to do some research in order to understand how to perform the experiment; luckily, the DALEX repository contained some examples that we studied to create our own code.

After importing and splitting the dataset between training set and test set, we applied three different Machine Learning Algorithms:

- Logistic Regression
- Random Forest
- Decision Tree

We have evaluated the fairness of each of the three models and obtained the following results:

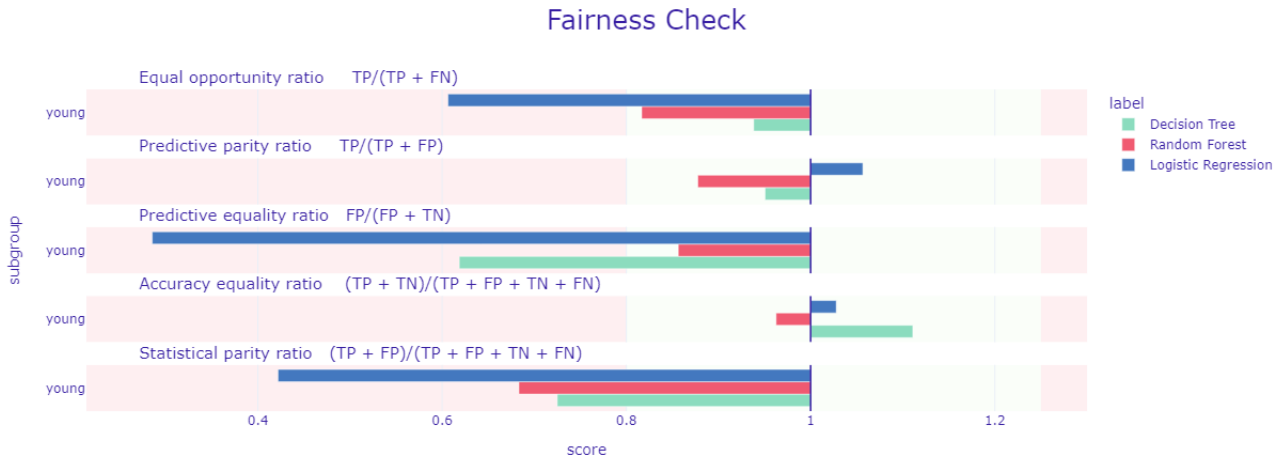

Figure 1: Fairness based on common metrics

As we can see, all the models have bias in some metrics, thus making our dataset not fair (an older patient has a higher chance of being classified as diabetic). In particular:

1. Logistic Regression
   The logistic regression model exhibited the highest bias among the three models. It demonstrated a higher likelihood of classifying older patients as diabetic, indicating a potential age bias.

2. Random Forest
   The random forest model displayed relatively lower bias compared to logistic regression. It achieved a more balanced classification of patients across different age groups, suggesting a better overall fairness performance.

3. Decision Tree
   The decision tree model fell between logistic regression and random forest in terms of bias.

The differences in bias among the models could be attributed to the characteristics of the dataset itself. It is possible that the dataset may have inherent complexities and relationships that certain models, such as random forest and decision tree, were better able to capture compared to logistic regression.

2.3.1 Resampling Technique: Application

In order to mitigate data bias, we used the resampling technique implemented by the DALEX library. Specifically, we utilized the resample function from the dalex.fairness.mitigation module.
This function allowed us to address bias by resampling the data based on the protected attribute(s) and the target variable. By specifying the desired type of resampling, such as 'uniform', we were able to control the resampling process. Additionally, we had the option to provide probabilities associated with each instance. By leveraging this resampling technique, we aimed to improve the fairness of our dataset and subsequently enhance the performance of our models.

It is important to note that the resample function relies on randomness, which can lead to different results with each run. This variability arises from the random selection of instances during the resampling process. Therefore, multiple runs or additional analyses are recommended to assess the stability and effectiveness of the resampling technique in mitigating data bias.

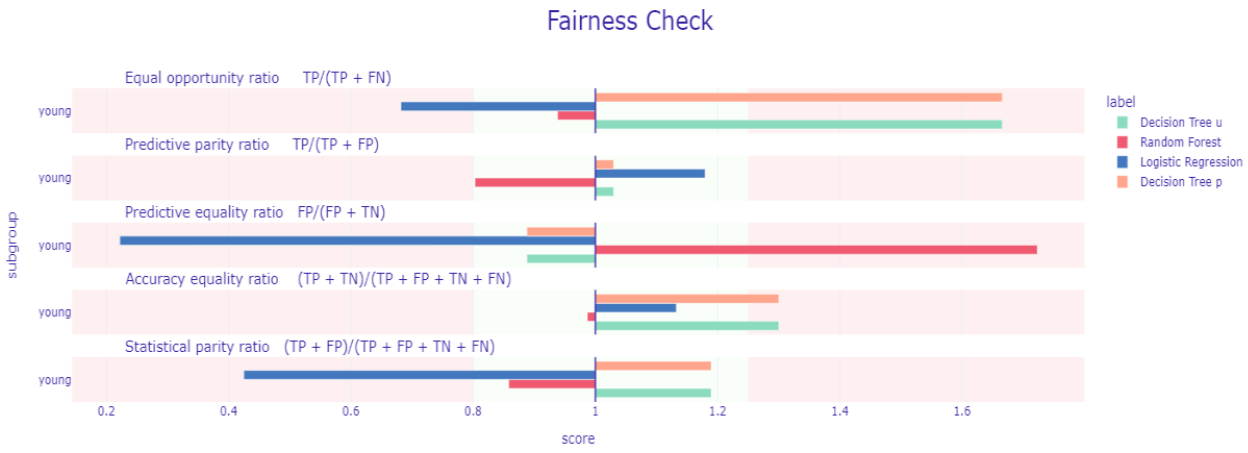

Figure 2: Fairness after applying the Resampling technique

After applying the resampling technique, the models' performance rankings remained consistent: Random Forest performed the best, followed by Decision Tree, and Logistic Regression.

For the Decision Tree model, the resampling technique improved the False Positive Rate (FPR) and Statistical Parity (STP), but led to slight decreases in Accuracy (ACC) and True Positive Rate (TPR).

The Random Forest model showed an increase in FPR but improved the STP and TPR.

The Logistic Regression model exhibited minimal changes in its parameters after resampling.

The consistent pattern of Random Forest outperforming Decision Tree and Logistic Regression even after the application of the mitigation technique could be due to inherent differences in the algorithms' capabilities to handle bias and capture complex relationships in the data.

3 Conclusions

The resampling technique had varying effects on the models, with improvements seen in some metrics for Decision Tree and Random Forest, while Logistic Regression remained relatively stable. Further analysis is required to understand these changes and evaluate the overall impact of resampling in mitigating bias.

References

[1] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5):739–768, 2021.

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

[3] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 2023.

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