Ethical and Socially-Aware Data Labels

Elena Beretta[1,3][0000−0002−6090−2765], Antonio Vetrò[1,2][0000−0003−2027−3308], Bruno Lepri[3][0000−0003−1275−2333], and Juan Carlos De Martin[1][0000−0002−7867−1926]

[1] Nexa Center for Internet & Society, DAUIN, Politecnico di Torino, Italy
{elena.beretta,antonio.vetro,demartin}@polito.it
[2] Future Urban Legacy Lab, Politecnico di Torino, Italy
[3] Fondazione Bruno Kessler, Italy
lepri@fbk.eu

Abstract. Many software systems today make use of large amount of personal data to make recommendations or decisions that affect our daily lives. These software systems generally operate without guarantees of non-discriminatory practices, as instead often required to human decision-makers, and therefore attracting increasing scrutiny. Our research is focused on the specific problem of biased software-based decisions caused from biased input data. In this regard, we propose a data labeling framework based on the identification of measurable data characteristics that could lead to downstream discriminating effects. We test the proposed framework on a real dataset, which allowed us to detect risks of discrimination for the case of population groups.

Keywords: Data Ethics, automated decisions, data quality

1 Introduction

The availability of large-scale data, often regarding human behavior, is profoundly changing the world in which we live. The automated flow and analysis of this type of data offers an unprecedented opportunity for actors in both public and private sectors to observe human behaviors for a large variety of purposes: to provide insights to policy-makers; to build personalized services like automated recommendations on online purchases; to optimize business value chains; to automate decisions; etc. However, the way data are collected, tested and analyzed poses a number of risks and questions related to the context of use [3]. Many researchers, in fact, identified a number of ethical and legal issues where the application of software automated techniques in decision-making processes has led to intended and unintended negative consequences, and especially disproportionate adverse outcomes for disadvantaged groups [1, 12]. Recent scandals such as the one involving Cambridge Analytica and Facebook[4] or the study conducted by ProPublica of the COMPAS Recidivism Algorithm[5], are two well-known examples of the relevance of these issues for our societies. Recent research efforts have

[4] https://bit.ly/2Hoa2q7
[5] See https://bit.ly/1XMKh5R

focused on the data collection and data exploitation issues of software systems (e.g., in the field of machine learning [7] or, more in general, in software-related conferences [11]). We place the problem in the context of software engineering practice proposing a data labeling framework, the Ethical and Socially-Aware Data Labels (EASAL), to identify data input properties that could lead to downstream potential risks of discrimination towards specific population groups. The beta version of the framework, presented here, relies on three building blocks, each one supported by previously published evidence drawn from different disciplines. We describe our data labeling in Section 2, and we show and discuss the results from testing EASAL on a real dataset in Section 3 . We conclude by summarizing our contribution and providing indications for future research work in Section 4.

2 Ethically and Socially-Aware data Labeling (EASAL)

Many software systems today rely on statistical techniques and prediction models, fed by large amount of available data. Such data is used for training algorithms whose scope is to recognize patterns and find relationships in data. A problem that characterizes automatic approaches that rely solely on data and algorithms is that they miss the human capability to perform important tasks, among which the context-aware interpretation of the results, the elaboration of explanations and cause-effect relationships, the recognition of biases (and possibly their correction). Regarding the latter, which is the focus of our work, we report a statement made by the mathematician Cathy O'Neal in her book "Weapons of Math Destruction" [13]:

> if the admission models to American universities had been trained on the basis of data from the 1960s, we would probably now have very few women enrolled, because the models would have been trained to recognize successful white males.

This observation entails an important fact: not only data processes such collection and analysis have ethical consequences, but also input data properties are connected to important ethical issues. In fact, some characteristics of the collected data involve ethical issues, and those problems propagate throughout all subsequent phases of the data life-cycle in software systems, until affecting the output, i.e. the decisions or recommendations made by the software. Our hypothesis is that certain data characteristics may lead to discriminatory decisions and therefore it is important to identify them and show the potential risks.

Moved by these motivations, we defined the Ethically And Socially-Aware Labeling (EASAL) framework, which is a way of labeling datasets using measures of certain input data characteristics (e.g., uneven distribution in gender balance, co-linearity of attributes, etc.) that represent a risks of discrimination if used in decision making (or decision support) systems. We believe that this information will be useful to software engineers to be more aware of the risks of discriminations and to use the dataset in an more ethically and socially-aware

manner. In addition, it could be used by third parties to certify such risks on a given dataset.

To the best of our knowledge, labelling approaches for ethical purposes are being investigated in two other ongoing research initiatives. The first one is a collaboration between the Berkman Klein Center at Harvard University and the MIT Media Lab, which led to "The Dataset Nutrition Label Project"[6]. The project aims to avoid that incomplete, misunderstood or problematic data have an adverse impact on artificial intelligence algorithms. The second research is conducted by Gebru *et al.* [6], who propose "Datasheets for Datasets": with respect to our proposal, this approach is towards more discursive technical sheets to encourage better communication between creators and users of a dataset. These approaches are not mutually exclusive, instead they can be seen as mutually reinforcing. Herein we describe the building blocks of EASAL. Herein we describe the building blocks of EASAL.

2.1 Disproportionate datasets

Most of today software-automated decisions are based on the analysis of historical data. This is very often done with machine learning models. It has been proven that problems of fairness and discrimination inevitably arise, mainly due to disproportionate datasets [13]. Disproportionate datasets lead to disproportionate results, generating problems of representativity when the data are sampled - thus leading to an underestimation or an overestimation of the groups - and of imbalance when the dataset used has not been generated using random probabilistic sampling methods. Many of the datasets used today have not been generated using these methods, but are rather selected through non probabilistic methods, which do not provide to each unit of the population the same opportunity to be part of the sample; this means that some groups or individuals are more likely to be chosen, others less. For this reason, it is essential to keep this aspect under control in non-probabilistic samples. In general, solutions relating to demographic or statistical parity are useful in cases where there is no deliberate and legitimate intention to differentiate a group considered protected, which would otherwise be penalized [4]. It should therefore be borne in mind that the solutions vary according to both the nature and use of the data. Take as an example a type of analysis that includes in its attributes individual income. If the choice to include in the sample only individuals with a high income is voluntary, no representativity problems arise, since the choice of a given group is based on the purposes of the analysis. However, if the probability of being included in the sample is lower as the income is lower, then the sample income will on average be higher in the overall income of the population.

2.2 Correlations and collinearity

In statistics two variables x_1 and x_2 are called collinear variables when one is the linear transformation of the other and therefore there is a high correlation.

[6] See `https://datanutrition.media.mit.edu`

In general there are always relationships between variables that involve a certain degree of linear dependence, but it is essential to keep this aspect under control to avoid negative effects: in fact, in case of collinearity, small variations in the data may correspond to significant variations in the estimated values. Since the analysis of collinearity reveals the presence of redundant connections between variables, it is useful in those areas more sensitive to the risk of discrimination. To prevent this effect some researchers adopt a naïve approach that precludes the use of sensitive attributes such as gender, race, religion and family information, but in some cases may not be effective. The use of geographic attributes, for example, is shown to be unsuitable when the use of protected data is to be foreclosed, because it easily leads to tracing protected attributes, such as race [12]. Hardt [7] points out that the condition of non-collinearity requires that the predictor ($\hat{Y}$) and the protected attribute (A) are independent conditional on Y: e.g., if *income* has to be predicted, it must be independent of *gender*. Another common error is *"mistake correlation with causation"* [8]; cause-effect ratios are often confused with correlations when features are used as proxies to explain variables to be predicted. For example, the IQ test is a test that measures logical-cognitive abilities, but if used as a proxy to select the smarter students for admission to a university course, it would almost certainly reveal itself as an imperfect proxy, since intelligence is a too broad concept to be measured by a number only. As a consequence, although there is a correlation between the test value of the IQ and the predicted variable, it is not sufficient to explain it. As Friedler *et al.* [5] remark, "*determining which features should be considered is a part of the determination of how the decision should be made*". In light of the problems mentioned in the previous paragraphs, we expect EASAL synthetically summarizes the analysis of collinearity and correlation between protected attributes in order to avoid possible discriminatory results.

2.3 Data quality

In computer science, "garbage in, garbage out" (GIGO) is a popular sentence to identify where "flawed, or nonsense input data produces nonsense output" [7]. The GIGO principle implies that the quality of the software is affected by the quality of the underlying data. As a consequence, computer generated recommendations or decisions are affected by poor input data quality. For these reasons, we include data quality as third building block of EASAL. The ISO/IEC standard 25012 [9] defines 15 data quality characteristics, operationalized by 63 metrics defined in the ISO/IEC 25024 [10]. Recent research efforts (e.g., [2] [15] [14]) showed that a measurement approach is effective in revealing data quality problems, especially for the inherent quality dimensions, that are also more effective for our purposes, because they are not affected by the context of use (e.g., hardware and software environment, computer-human interface). We propose the ISO/IEC 25012 and 25024 standards models as a reference for quantitatively assessing the quality of data input and the consequential confidence of the decision made out

[7] See `https://en.wikipedia.org/wiki/Garbage\_in,\_garbage\_out`

of that data. In particular, we refer to the inherent quality dimensions: accuracy, completeness, consistency, credibility, currentness [8].

3 Testing EASAL on real datasets

We tested the EASAL on Credit Card Default dataset, that *contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005*[9]. The field of creditworthiness often appears in the literature alongside issues related to ethical decisions. Recently, some studies have shown that access to credit for black people is modulated by certain attributes such as race, rather than by information about the payer's status [10]. The dataset that we use does not contain the protected attribute *race*, but contains other personal information that can be used in a discriminatory way if applied to assess creditworthiness, such as gender and level of education.

Disproportion Figure 1 reports an example of how *disproportion plots* should look like. The figures shows that 60% of individuals are women, 46.7% of individuals have attended university, the age group most represented is that of 25 to 40 years, the proportion of married individuals is the same for single individuals. We conclude that the potential risk identified by the label is with respect to age, since older people are less represented, and we do not have information on the real proportions in the whole population and at the time of dataset creation.

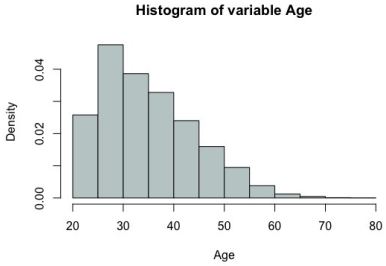

Fig. 1. Frequency of variable *age*

Correlation and collinearity We perform the analysis for each protected attribute in relation to default payment. Figure 2 contains an example of mosaic plot [11], the levels of Pearson residuals for each cell are indicated by the colours; blue indicates cases in which there are more observations than those that the null model would present - meaning in case of independence -; the case in which the observations are lower than the expected number is indicated with red. The performed tests show that the association between the protected attributes and the *default payment* variable (1 = yes, 0 = no) is present in all groups for the *gender*

[8] For the definitions of inherent quality measures see [10]

[9] `https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset`

[10] See `https://bit.ly/2NyNVPx`

[11] *A mosaic plot is an area proportional visualization of a (possibly higher-dimensional) table of expected frequencies. From:*`https://bit.ly/2LvQIbk`

and *education* variables, while it is only found in the *default payment group = yes* for the marital status variable. The residual levels allow the identification of the groups most responsible for the dependency: the *education* variable for *default payment = yes*, and the male group for *default payment = yes*. More in-depth analyses show a trend in contrast to the previous one, in which the group *Female* is the most responsible for addiction. This may indeed indicate a possible problem of overestimation in the prediction phase.

Inherent data quality The nature of dataset allows to test only two of the five inherent quality dimensions, *accuracy* and *completeness*; of these, five metrics are selected from the ISOIEC 25024, namely: *Acc-I-4: Risk of data set inaccuracy*, *Com-I-1: Record completeness*, *Com-I-2: Attribute completeness*, *Com-I-4: Data values completeness*, *Com-I-5: Empty records in a data file* [12]. The test provides the extreme positive value of the index, therefore it is not necessary to report the values obtained.

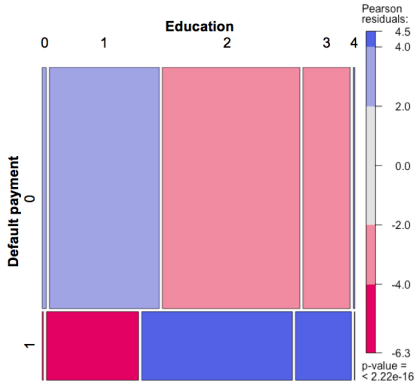

Fig. 2. Conditional mosaic plot for conditional independence of variable *education* *Legend: 0 = na; 1 = graduate school; 2 = university; 3 = high school; 4 = others*

4 Conclusions

We presented a theoretical framework for labeling input data used in decision making software and for identifying risks of discrimination towards specific population groups. The Ethically and Socially Aware Labels (EASAL) are composed of three building blocks: measures for assessing disproportion; measures for assessing correlation and collinearity involving protected attributes; measures for assessing data quality. The building blocks have been identified on the base of literature studies and authors experience. We intend to address our future work along the following directions: test and specification of the use of correlation and collinearity metrics for different types of statistical variables; graphical design of an intuitive label that could help software engineers in quickly understanding the discriminatory risks of using a dataset; automation of label creation and source code freely available to allow replication studies and improvements. We also invite the software engineering community to contribute to this initial work by improving the building blocks measures, by identifying new building blocks and by applying EASAL for benchmarking purposes. This would facilitate an

[12] Details on how to compute each metric can be retrieved from [10]

increase of awareness of software practitioners regarding the ethical implications of the data-driven systems that they design, build, and to which are probably subject, at least in some scenarios.

References

1. Barocas, S., Selbst, A.D.: Big data's disparate impact. California Law Review **104**(3), 671–732 (2016)
2. Corrales, D.C., Corrales, J.C., Ledezma, A.: How to address the data quality issues in regression models: A guided process for data cleaning. Symmetry **10**(4) (2018). https://doi.org/10.3390/sym10040099, `https://bit.ly/2xOLVzN`
3. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S.J., O'Brien, D., Shieber, S., Waldo, J., Weinberger, D., Wood, A.: Accountability of ai under the law: The role of explanation. Berkman Center Research Publication Forthcoming. Harvard Public Law Working Paper **18**(07) (2017)
4. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemeln, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226. ACM (2012)
5. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236 (2016)
6. Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daume, H.I., Crawford, K.: Datasheets for datasets. arXiv:1803.09010r (2018)
7. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems (2016)
8. Hosni, H., Vulpiani, A.: Forecasting in light of big data. Philosophy & Technology pp. 1–13 (2017)
9. ISO-IEC: ISO/IEC 25012:2008 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model. Standard, International Organization for Standardization, Geneva, CH (december 2008)
10. ISO-IEC: ISO/IEC 25024:2015 – Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality. Standard, International Organization for Standardization, Geneva, CH (october 2015)
11. Karim, N.S.A., Ammar, F.A., Aziz, R.: Ethical software: Integrating code of ethics into software development life cycle. In: 2017 International Conference on Computer and Applications (ICCA). pp. 290–298 (Sept 2017). https://doi.org/10.1109/COMAPP.2017.8079763
12. Lepri, B., Staiano, J., Sangokoya, D., Letouz, E., Oliver, N.: The Tyranny of Data? The bright and dark sides of data-driven decision-making for social good. Springer, Cham (2017)
13. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York, NY (2016)
14. Torchiano, M., Vetrò, A., Iuliano, F.: Preserving the benefits of open government data by measuring and improving their quality: An empirical study. In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). vol. 1, pp. 144–153 (July 2017). https://doi.org/10.1109/COMPSAC.2017.192
15. Vetrò, A., Canova, L., Torchiano, M., Minotas, C.O., Iemma, R., Morando, F.: Open data quality measurement framework: Definition and application to open government data. Government Information Quarterly **33**(2), 325 – 337 (2016). https://doi.org/https://doi.org/10.1016/j.giq.2016.02.001