Framework for Trustworthy AI in the Health Sector

Mykhailo Danilevskyi - D22126578@myTUDublin.ie
School of Enterprise Computing and Digital Transformation,
Faculty of Computing, Digital and Data, TU Dublin, Ireland
Supervisors: Dr. Fernando Perez Tellez, Dr. Davide Buscaldi

Introduction

The European Commission defines that Trustworthy AI should be lawful, ethical and robust. The ethical component and its technical methods are the main focus of the research. According to this, the initial research goal is to create a methodology for evaluating datasets for ML modeling using ethical principles in the healthcare domain. Ethical risk assessment will help to ensure compliance with principles such as privacy, fairness, safety and transparency which are especially important for the Health Care sector. At the same time, risks must be evaluated with respect to AI model performance and possible scenarios of risk mitigation. Ethical risk mitigation techniques involve data modification, elimination of private information from datasets that directly influence AI modelling. Therefore ethical risk mitigation techniques should be carefully selected depending on domain and context. In this research work, we present an analysis of these techniques.

Privacy

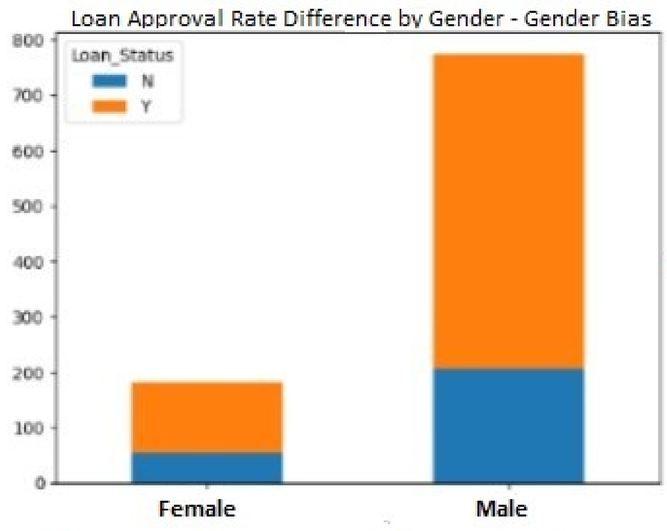
Privacy principle declares that data is protected and used with owner consent. The aim of the section is to research methods for private data detection and de-identification with minimum information loss.
The following methods have been researched for the best suitability for healthcare data: Regex, Conditional Random Fields, Machine Learning (LR, SVM, Decision Trees), Deep learning (CNN, RNN, LSTM, BERT)

Fairness

Compliance with fairness principle ensures that model decisions are not discriminating and equally correct for any group of AI users by gender, age, race, etc. For healthcare AI powered solutions, fair and unbiased decisions are critical for people's health. For example, younger people have higher underdiagnosis AI decision rate than older people, because they are naturally healthier and this is reflected in AI train data. In order to avoid such cases, it is required to measure and mitigate bias in train data. The following detection methods and techniques are being used in experiments with healthcare data: Data pre-processing methods - Re-weighing [1], Optimized pre-processing [2], Learning fair representations [3], Disparate impact remover [4]; Data in-processing methods - Adversarial debiasing [5], Prejudice remover [6]; Data post-processing - Equalized odds post-processing [7], Reject option classification [8]. Tools: AIF360, Google What-if, Fairkit, Fairlearn.

Case Study: Improve Fairness using Disparate Impact Remover Algorithm

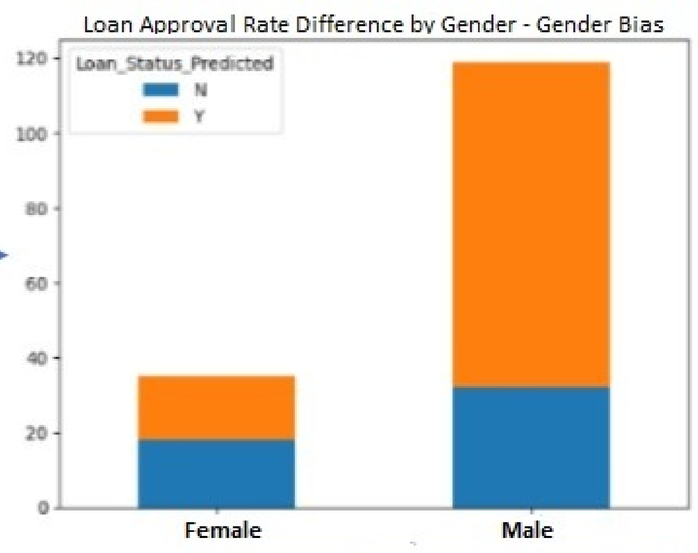
The initial experiment was conducted on the dataset that contains information about the loan applicants*. Gender was selected as applicants protected attribute which means that in the ideal world loan approval rate is equal for applicant regardless of gender. Fairness was estimated with Disparate Impact metric. If the value of the metric is greater or equal to 0.80, then it is considered that there is no discrimination in loan approvals between genders [9]. During the experiment, Disparate Impact Remover algorithm was applied to improve fairness of the loan approval logistic regression model. Disparate Impact Remover algorithm transforms numeric data to minimize differences between genders in loan amount, terms etc. The algorithm was selected randomly for initial exploration of the fairness problem. The algorithm improved Disparate Impact metric from 0.66 to 0.71. The improved result is still lower than normal 0.80, which means that further analysis of other methods is required.
* - experiments with healthcare related data will be conducted in the next couple of months.



Gender	N	Approval Rate
Female	624	68%
Male	145	74%

DI = 0,92
(0,68 / 0,74)

Accuracy: 0.82
Precision: 0.88
Recall: 0.86



DI = 0,71
(0,51 / 0,72)

Gender	N (testset)	Approval Rate
Female	35	51%
Male	119	72%

(without DIR Transformation)

Gender	N (testset)	Approval Rate
Female	35	48%
Male	119	73%

DI = 0,66
(0,48 / 0,73)

Disparate Impact (DI) >= 0,80 considered as normal

Conclusions and Future Work

Creation of a framework of trustworthy AI in the healthcare sector is critically important for ensuring privacy, bias-free and fair decision by AI-powered healthcare systems. As a future work, the following subgoals are planned to achieve: 1) Identify peculiarities of health datasets; 2) Apply methods and techniques for detecting and de-identification of private information. Define de-identification methods that best work with healthcare data taking into consideration possible loss of meaningful data; 3) Apply methods and techniques for detection and mitigation of bias in datasets. Define the best methods for healthcare.

References

[1] Kamiran, Faisal & Calders, Toon. (2011). Data Pre-Processing Techniques for Classification without Discrimination. Knowledge and Information Systems. 33.10.1007/s10115-011-0463-8. [2] Calmon, Flavio & Wei, Dennis & Natesan Ramamurthy, Karthikeyan & Varshney, Kush. (2017). Optimized Data Pre-Processing for Discrimination Prevention. [3] Zemel, Richard & Wu, Y. & Swersky, K. & Pitassi, T. & Dwork, C.. (2013). Learning fair representations. 30th International Conference on Machine Learning, ICML2013.1362-1370. [4] Friedler, Sorelle & Scheidegger, Carlos & Venkatasubramanian, Suresh. (2014). Certifying and Removing Disparate Impact. 10.1145/2783258.2783311. [5] Zhang, Brian & Lemoine, Blake & Mitchell, Margaret. (2018). Mitigating Unwanted Biases with Adversarial Learning. 335-340.10.1145/3278721.3278779. [6] Kamishima, Toshihiro & Akaho, Shotaro & Asoh, Hideki & Sakuma, Jun. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. [7] Hardt, Moritz and Price, Eric & Srebro, Nathan. (2016). Equality of Opportunity in Supervised Learning.[8] Kamiran, Faisal & Karim, Asim & Zhang, Xiangliang. (2012). Decision Theory for Discrimination-Aware Classification. Proceedings - IEEE International Conference on Data Mining, ICDM.924-929.10.1109/ICDM.2012.45. [9] Dan Biddle (2006). Adverse Impact And Test Validation: A Practitioner's Guide to Valid And Defensible Employment Testing. Aldershot, Hants, England: Gower Technical Press. pp.2–5.