CIS 6930: Privacy & Machine Learning

Project Proposal: Can Explainable AI techniques explain Unfairness?

Kiana Alikhademi
kalikhademi@ufl.edu

Emma Drobina
edrobina@ufl.edu

Brianna Richardson
richardsonb@ufl.edu

September 27, 2019

1 Introduction

Recently, applications of Machine Learning (ML) and Deep Learning (DL) have hit an all-time high. Within nearly every domain of technology, machine learning's predictive power is being used to replace automated processes. However, often these models are assessed based on accuracy alone, with no checks and no assurances that these tools are utilizing appropriate parameters to make their decisions. Furthermore, often these models are so complex that it's difficult to determine how they make their decisions. With recent failings of machine learning models [1] [2], there has been a push towards holding machine learning engineers accountable via tools like Black-box interpreters and explainable interfaces, which work to explain the results of complicated ML models. While such tools can point out why ML algorithms made the decision, very little research has been done on whether major issues like fairness can be illustrated via Black-box interpreters and what metrics should be used to deduce fairness from the results.

We intend to utilize data sets which previous research dictates are highly biased. Classifiers will be constructed to produce the highest possible accuracy for each respective data set. The resulting models will be run through Black-box interpreters and explainability interfaces to determine if such tools identify issues of fairness. Our work aims to create a fairness rubric in which ML models can be evaluated via explainability and interpretability tests.

2 Background and Related Work

2.1 Fairness

Issues of fairness are prevalent in Machine Learning. As machine learning systems usually learn from the data provided by humans, there is a strong likelihood that biases that exist in the data will be reflected in the models. One popular example is the COMPAS dataset, which has higher false-positive rates for black people than similar white individuals [3]. Mitigating the bias and unfairness in machine learning is a necessity, as it has the potential to impact everyone. To talk about fairness, we first need to define it. There are multiple fairness definitions in the literature. Chouldechova and Roth [4] stated that unfairness in machine learning is the result of the following factors:

- Biased data
- Ignoring the different distributions of classes within a data
- The need to explore for more data before making the decision

In the algorithmic fairness domain, experts defined two notions of fairness: statistical and individual [4]. In the statistical notion, some protected demographic groups such as racial groups are considered. The parity of some statistical measure across all of these groups is the key for preserving the statistical

fairness [4]. However, according to Chouldechova and Roth [4], there are some drawbacks to the statistical notion of fairness. First, any two different statistical measures (e.g., false negative and false positive rate) contradict each other and cannot be optimized simultaneously. Also, learning subjects concerning this notion could be computationally hard [4].

In the individual idea of fairness, each pair of individual will be compared according to specific sets of criteria [4]. Although this notion seems more natural to the users, there is a need to predefined detailed assumptions to measure the similarity of instances and providing similar results. There are not holistic guidelines to measure fairness in the machine learning domain. Therefore, this project is aimed to address this gap.

2.2 Explainable AI

ML algorithms are capable of facilitating the prediction process by synthesizing the data to extract the underlying relations and generating the relevant results for the target data. In many of these domains, the complex nature of these algorithms discourage professionals from utilizing it's predictive power. For example, research has been done to predict patients' risk of suicide by considering the diverse set of factors such as substance abuse, family history, outpatient, and inpatient visits [5, 6, 7]. Although most of the ML algorithms predict with high accuracy, most of the experts such as clinicians or police officers are hesitant to use them. The inability of ML algorithms to explain themselves is one of the main factors in the underlying mistrust.

Holzinger et al. [8] discussed that the success of AI/ML models in any domains heavily relies on their ability to explain the results to a human in a simple and yet understandable way. Explainability of ML models, based on the multiple sources of data, would help experts to track the decisions made by the machines and improve the trust and transparency. Therefore, Explainable Artificial Intelligence (XAI) has recently become a popular research topic. One of the popular XAI systems is Local Interpretable Model-Agnostic Explanations known as LIME[9]. LIME is learning the prediction model locally to make the explanation [9]. PreJu [10] is another XAI technique which is a java applet to generate human-centered explanations based on simple classifiers.

3 Proposed Approach & Plan

We propose to study the generalizability of several Explainable AI (XAI) tools that generate human-readable explanations. The output of multiple models built using a variety of datasets and classifiers will be tested concerning the fairness heuristics. We will use the results of these experiments to help define a series of heuristics for evaluating fairness and explainability.

3.1 Explainable AI Toolkits

Within our project, both LIME and PreJu will be used to explain the results of different classifiers. Based on the literature, we identified these as currently popular tools.

3.2 Datasets and classifiers

We plan to test these toolkits on several types of datasets to ensure generalizability of our results. We intend to use both ProPublica's released COMPAS dataset [11] as well as traffic stop data from Stanford's Open Policing Project. Specifically, we plan to use data from Nashville, TN, since it has a large number of records (3,092,351). Nashville has also released data for all possible features, making this one of the least sparse datasets available through the Open Policing Project. Importantly, these datasets have been studied with a critical eye for bias [3] [12], so we can refer to outside auditors in addition to our analysis.

Furthermore, effective classifiers for both COMPAS and the Open Policing Project have been identified across several papers [11], [13], [14]. Based on the results of these papers, we plan to use logistic regression, state vector machines (SVM), multinomial naive Bayes, and neural networks.

We also intend to complete a similar analysis of hate-speech identification datasets. Specifically, we will use Founta et al. 's hate speech twitter collection [15] and Blodgett et al. 's African-American English on Twitter collection [16]. [15]'s dataset consists of 91, 951 tweets which have been coded as abusive, hateful, spam, or normal. These labels will be used to train a logistic regression model using the bag-of-words features. [16]'s dataset consists of 59.2 million tweets labeled by the race, which will be used as a test set. These collections were chosen because previous research [17] reflects apparent racial disparities when machine learning is used to identify hate speech.

The last collection of data is centered around health. Specifically, we will be using datasets collected by Agniel et al [18]. Work done in [18] depict that the analysis of Electronic Health Record (EHR) can be used to detect inconsistencies and biases within patient treatmeent and survival rate. We intend to use encoding and logistic regression to be able to predict 3-year survival rate for patients in this dataset.

3.3 Research Plan

We will begin by drafting our initial rubric for unfairness. This will be based off our review of the literature, such as Florida's [19] work on developing a framework for ethical AI in society and Albarghouthi's [20] research into programmatically applying checks for fairness.

We plan to clean and process the data from the datasets we identified in section 3.2. A subset of this processed data will be used to train our datasets. We will then apply the LIME and PreJu tools to our models and review the results. Based on the outcome of this stage, we will revise our initial heuristics. Then, we will retrain our classifiers on our full dataset, re-apply LIME and PreJu, and use our final heuristic test to evaluate the results.

Additionally, we plan to incorporate the results of this project into an ongoing research project on predictive policing and fairness in machine learning headed by Dr. Juan Gilbert and Dr. Duncan Purves.

4 Timeline

Table 1: Project Timeline

Sept 26	Complete Initial Proposal. Identify classifiers, Data sets, AI techniques.
Oct 11	Draft the rubric for fairness. Complete Data Processing; Complete Classifier construction.
Oct 25	Complete run on Explainability and Interpretability (EI) tests with classifiers and partial data. Make appropriate changes to the rubric.
Nov 20	Complete final run of EI tests with classifiers and complete data. Complete Rating of classifiers with respect to the results from the EI tests. Prepare these results for Mid-semester Project Feedback.
Nov 21	Submit Mid-semester Project Feedback.
Dec 6	Complete Data Analysis. Draft the final report. Draft presentation.
Dec 11	Submit final project report.

References

[1] A. Hern, "Google's solution to accidental algorithmic racism: ban gorillas," 2018.

[2] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," 2019.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica, May*, vol. 23, p. 2016, 2016.

[4] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.

[5] Y. Barak-Corren, V. M. Castro, S. Javitt, A. G. Hoffnagle, Y. Dai, R. H. Perlis, M. K. Nock, J. W. Smoller, and B. Y. Reis, "Predicting suicidal behavior from longitudinal electronic health records," *American journal of psychiatry*, vol. 174, no. 2, pp. 154–162, 2016.

[6] G. E. Simon, E. Johnson, J. M. Lawrence, R. C. Rossom, B. Ahmedani, F. L. Lynch, A. Beck, B. Waitzfelder, R. Ziebell, R. B. Penfold, *et al.*, "Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records," *American Journal of Psychiatry*, vol. 175, no. 10, pp. 951–960, 2018.

[7] C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, "Predicting risk of suicide attempts over time through machine learning," *Clinical Psychological Science*, vol. 5, no. 3, pp. 457–469, 2017.

[8] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?," *arXiv preprint arXiv:1712.09923*, 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, ACM, 2016.

[10] O. Biran and K. R. McKeown, "Human-centric justification of machine learning predictions.," in *IJCAI*, pp. 1461–1467, 2017.

[11] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the compas recidivism algorithm," *ProPublica (5 2016)*, vol. 9, 2016.

[12] A. Chohlas-Wood, S. Goel, A. Shoemaker, and R. Shroff, "An analysis of the metropolitan nashville police department's traffic stop practices," tech. rep., Technical report, Stanford Computational Policy Lab, 2018.

[13] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018.

[14] M. Saviano and S. Tieu, "When to stop-and-frisk," 2017.

[15] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Proceedings of the Twelth International AAAI Conference on Web and Social Media*, pp. 491–500, AAAI, 2018.

[16] S. L. Blodgett, L. Green, and B. O'Connor, "Demographic dialectal variation in social media: A case study of african-american english," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, Association of Computational Linguistics, 2016.

[17] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," in *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35, Association of Computational Linguistics, 2019.

[18] D. Agniel, I. S. Kohane, and G. M. Weber, "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study," *BMJ*, vol. 361, 2018.

[19] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, *et al.*, "Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.

[20] A. Albarghouthi and S. Vinitsky, "Fairness-aware programming," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 211–219, ACM, 2019.