CIS 6930: Privacy & Machine Learning
Final Project Report: **Can Explainable AI Explain Unfairness?**

Kiana Alikhademi
(Point of Contact)
kalikhademi@ufl.edu

Emma Drobina
edrobina@ufl.edu

Brianna Richardson
richardsonb@ufl.edu

December 12, 2019

1 Introduction

Within nearly every domain of technology, the predictive power of machine learning (ML) is being used to replace automated processes, and its prevalence is only growing. However, ML models are often assessed based on accuracy alone, with no checks and no assurances that these tools are utilizing appropriate parameters to make their decisions. Furthermore, these models can be so complex that it is impossible for a human to determine how decisions were made independently. With recent failings of machine learning models [11] [21], 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 how ML algorithms made a specific decision, very little research has been done on whether major issues like fairness can be illustrated via black-box explainers and what metrics should be used to deduce fairness from the results.

We defined a rubric evaluating XAI tools in terms of their use in evaluating fairness and applied this rubric to two XAI tools and two inherently explainable ML models. We created a total of twelve models (four models each for three datasets), analyzed our models using the XAI tools we identified, and completed our rubric based on the results of our analysis. So that our project could discuss fairness in a meaningful way, we selected data sets to train our models that previous research indicated were likely to be biased.

1.1 Major Contributions

- Developed holistic fairness rubrics concerning the access and capabilities of XAI.
- Examined the state-of-the-art fairness tools with respects to our comprehensive rubrics.
- Outlined the functional gaps within this area.

2 Background

2.1 Fairness

Issues of fairness are prevalent in ML. 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 famous example of this phenomenon is the COMPAS dataset, which has higher false-positive rates for black people than similar white individuals [2]. Mitigating the bias and unfairness in machine learning is a necessity as ML systems become implemented in practice, as unfair systems will lead to user distrust. Furthermore, ML developers have an ethical obligation not to develop systems that hurt people, to follow existing non-discrimination laws. However, before we can determine if a system is fair, we first need to define fairness. Fairness is a perhaps surprisingly complex topic that is defined in several conflicting ways in the literature.

Experts in algorithmic fairness have defined two notions of fairness: statistical and individual [7]. Under the statistical definition of fairness, minority groups as a whole should be treated the same as majority groups as a whole. The parity of some statistical measure across all of these groups is the key for preserving the statistical fairness [7]. However, according to Chouldechova and Roth [7], there are some drawbacks to the statistical defintion 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 [7].

Under the individual definition of fairness, each pair of individuals is compared according to specific sets of criteria [7]. Essentially, "similar individuals should be treated similarly". This notion can be more intuitive for humans to understand, but detailed assumptions about how to measure the similarity of instances and provide similar results must be made.

2.2 Datasets

Three separate data sets were chosen from literature based on controversial issues of unfairness that were identified after further study. This data served as case studies from which we generated our rubric and evaluated XAI tools. The data sets and their respective fairness issues are detailed below.

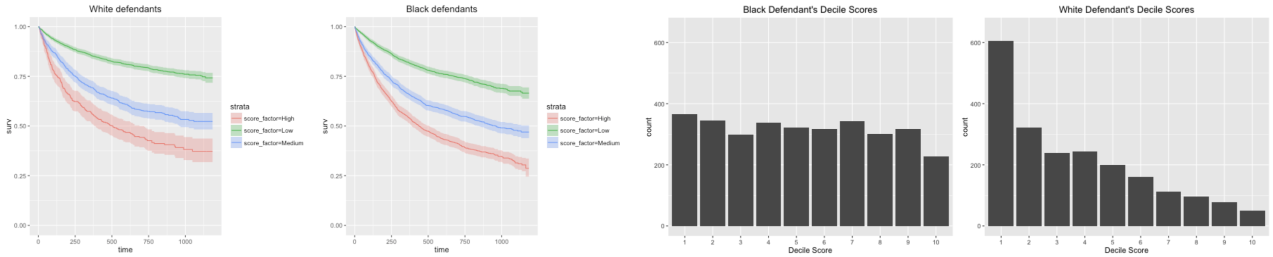
2.2.1 COMPAS

We plan to test XAI tools on several types of datasets to ensure the generalizability of our results and rubrics. We intend to use ProPublica's released COMPAS dataset [13]. Importantly, this dataset has been studied with a critical eye for bias [2] [6], so we can refer to outside auditors in addition to our analysis. Furthermore, effective classifiers for COMPAS has been identified across several papers [13], [9], [18]. COMPAS data consists of 6167 records and eight features. Some of the main biases found in the ProPublica's study has shown in the following:

- Black defendants were often predicted to be at a higher risk of recidivism in comparison to their white counterparts as shown in Figure 1a.

- Due to the racial bias in the system and data, White defendants were often mistakenly marked to be less risky than black ones as shown in Figure 1b[13].

- The racial bias led to a 77 percent higher chance that a black defendant is classified with a higher risk of violent crimes in comparison to white defendants.



(a) Racial bias distribution across different score levels for white and black defendants

(b) Decile score distribution comparison between the white and black defendants

Figure 1: Visualization of biases and issues in COMPAS

2.2.2 Health

The second collection of data is centered around health. Specifically, we used datasets collected by Agniel et al [1] with laboratory results and corresponding facility information for 272 unique tests across 228,592 patients. Work done in [1] depicts that the analysis of Electronic Health Record (EHR) can be used to detect inconsistencies and biases within patient treatment and survival rate. As can be seen in 2, the original work depicted the drop in survival rates for patients who had tests done early in the morning, on the weekend, or consecutively close together. Researchers urged from these results that more attention be placed on healthcare infrastructure and procedure that might influence these trends. Therefore, this data set was a prime example of biased events that are reflected in the data.

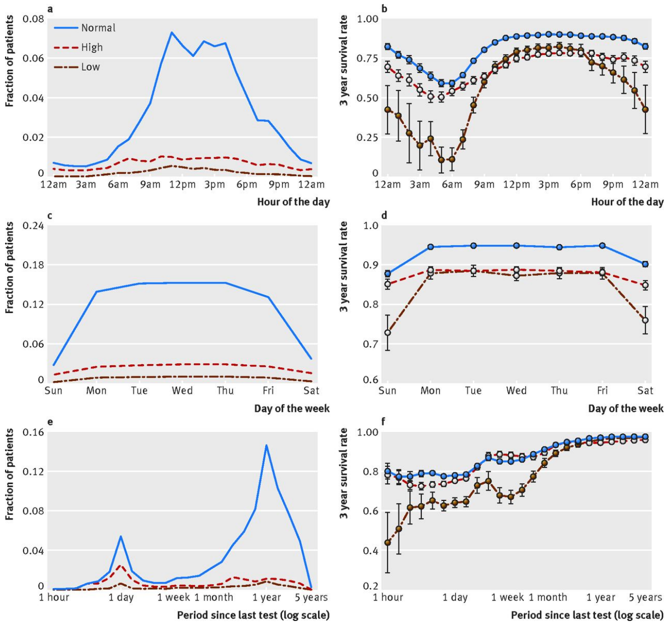
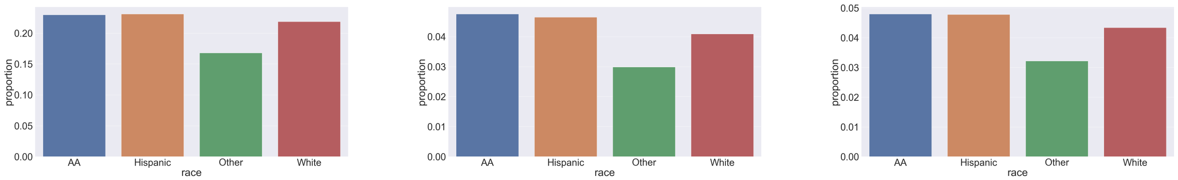


Figure 2: Results of previous work where survival rates could be best predicted based on hospital procedure in comparison to actual lab results.

2.2.3 Twitter

We also completed a similar analysis of hate-speech identification datasets. Specifically, we used Founta et al.'s hate speech twitter collection [10] and Blodgett et al. 's African-American English on Twitter collection [5]. [10]'s dataset consists of 91,951 tweets which have been coded as abusive, hateful, spam, or normal. [5]'s dataset consists of 59.2 million tweets labeled by the race, which was used as a test set for our models. These collections were chosen because previous research [8] identified racial disparities in the identification of abusive language on social media. We sought to replicate this study.



(a) Random forest classifier model.

(b) Keras LSTM.

(c) Logistic regression model.

Figure 3: Proportion of tweets marked as derogatory by race for each model.

2.3 Explainable Artificial Intelligence

Although many ML algorithms predict with high accuracy, many experts with no ML background, 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. Therefore, many researchers have focused on designing tools called Explainable AI (XAI) tool to overcome the uncertainty in the black-box models. As we discussed previously, one of the popular XAI tools is Local Interpretable Model-Agnostic Explanations, known as LIME [17]. LIME learns the prediction model locally to make the explanation [17] for any classifier. LIME is a local agnostic model that can be used across different models for providing more understandable explanations. Therefore, we decided to evaluate the explanations generated by LIME on our datasets concerning fairness metrics. Another tool that has gained popularity in the area of fairness and explainability is the AI Fairness 360 toolkit (AIF360) developed by Bellamy et al. [4] at IBM. The package includes a comprehensive set of fairness metrics for datasets and models, explanations for these metrics, and algorithms to mitigate bias in datasets and models. AIF360 can be used in any step of the machine learning process from pre-processing to post-processing (ad-hoc analysis).

3 XAI Rubric

As machine learning becomes more prevalent, XAI tools will provide necessary auditing abilities, both for internal QA testers and governmental and regulatory bodies [16]. Among these auditing, tasks will be reviews of fairness, both to prevent discrimination in the legal sense and to prevent other kinds of bias distorting a model. However, there is little analysis in the literature of how suitable various XAI tools are for reviewing fairness.

To address this need, we developed a rubric to evaluate the usefulness of XAI tools in analyzing fairness. This rubric is intended to help developers of XAI tools understand user needs; to help ML developers as they review their models for accuracy and fairness; to help lay users scrutinize the results of ML models [20]. By critically reviewing these XAI tools, we hope to motivate XAI development. We want XAI to be widely used so that it can increase trust and ensure fair and ethical decision-making [14].

We distinguished four major areas of need: 1) identifying biased data and biases that are reflected in the data; 2) ensuring pre-processing procedures were appropriate; 3) reviewing the selection and optimization of the ML model under observation; 4) ensuring that the results outputted by the XAI tool are useful and presented in a clear and comprehensible manner.

	Random Forest Feature Importance	**LIME**	**Fairness 360**	**Ad-hoc Explainability**
Model used	Random Forest	Deep learning		Logistic Regression
Issues with Biased Data				
Imbalanced data	0	0	1	0
Influential variable identification	1	1	0	1
Preprocessing issues	0	0	1	0
Sensitive attributes	0	0	1	0
Issues involved in Machine Learning Models				
Model-Specific influences	0	0	0	0
Accuracy equity	0	1	1	0
Issues involved with XAI results				
Target audience	0	0	0	0
Presentation of explanations	1	1	0	1

Table 1: Final rubric with XAI tools being evaluated based on whether or not they included these fairness considerations.

As can be seen in Table 1, fairness issues found in literature were categorized into 8 general categories. Within each category, we evaluated the respective XAI tools ability to identify and warn the user that the respective fairness issue is compromised. A detailed breakdown of the rubric can be found in the appendix.

Once models were built, as described in the subsequent sections, we evaluated their performance with respect to the rubric. Table 1 shows the scores given to each XAI tool, where 1 depicts the existence of features to assist with the respective fairness issue.

4 Methodology & Results

For each of our three datasets, we created 3 separate models that could be used to evaluate the different XAI (as is shown in Table 1). Logistic regression, random forest, and neural network models were built for each data set. Details on the preprocessing and model creation can be found in the Appendix. These models were then used in conjunction with the XAI tools identified in the rubric to assess the quality of the tool with respect to fairness.

4.1 COMPAS

4.1.1 Logistic Regression

The logistic regression got an accuracy of 72% in predicting if the defendants will re-offend. Figure 4 is a graph of the coefficients of the logistic regression, displaying the magnitude and direction of influence each feature will have on the target variable.

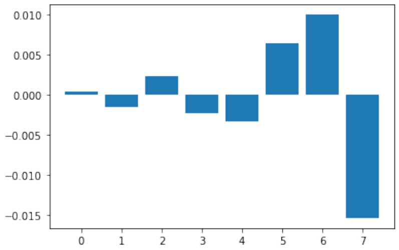
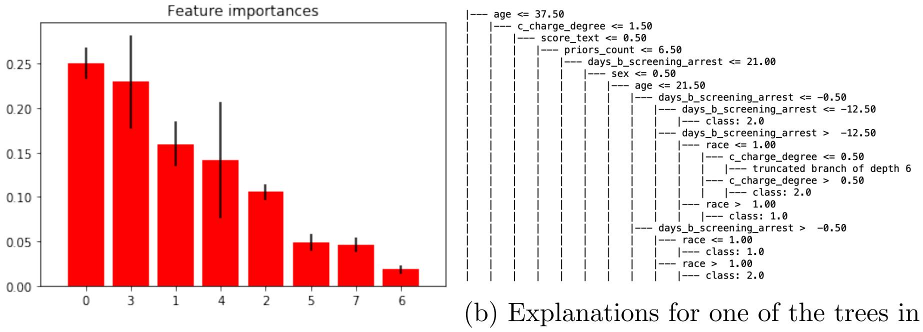


Figure 4: Explanations about the coefficients using Logistic Regression Model

4.1.2 Random Forest

Our random forest for COMPAS had an the accuracy of 70%. The random forests model was explained using the in-built properties of random forests. We selected individual trees from the forest and used scikit-learn's native function to export a decision tree as a series of text rules to analyze them.



```
|--- age <= 37.50
|   |--- c_charge_degree <= 1.50
|   |   |--- score_text <= 0.50
|   |   |   |--- priors_count <= 6.50
|   |   |   |   |--- days_b_screening_arrest <= 21.00
|   |   |   |   |   |--- sex <= 0.50
|   |   |   |   |   |   |--- age <= 21.50
|   |   |   |   |   |   |   |--- days_b_screening_arrest <= -0.50
|   |   |   |   |   |   |   |   |--- days_b_screening_arrest <= -12.50
|   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |--- days_b_screening_arrest >  -12.50
|   |   |   |   |   |   |   |   |   |--- race <= 1.00
|   |   |   |   |   |   |   |   |   |   |--- c_charge_degree <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 6
|   |   |   |   |   |   |   |   |   |   |--- c_charge_degree >  0.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |   |--- race >  1.00
|   |   |   |   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |   |--- days_b_screening_arrest >  -0.50
|   |   |   |   |   |   |   |   |--- race <= 1.00
|   |   |   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |   |   |--- race >  1.00
|   |   |   |   |   |   |   |   |   |--- class: 2.0
```

(a) Random Forest Feature Importance

(b) Explanations for one of the trees in the COMPAS random forest model.

Figure 5

4.1.3 Sequential Deep Neural Network

As COMPAS data is not a sequence or an image, we could not use convolutional or recurrent neural networks. Therefore, we used a deep neural network. The final accuracy of our model is 73%.

We fed our deep neural network into LIME. The results for one record can be seen in Figure 6. We can see the features in order of their impact on the prediction.

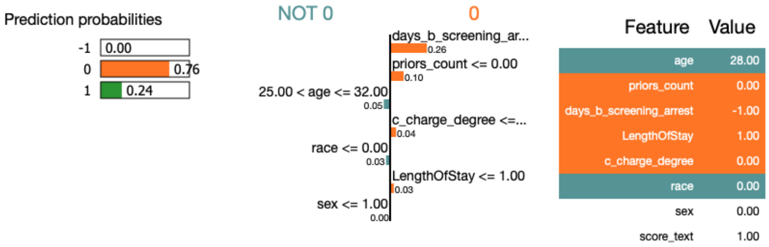


Figure 6: Deep Neural Network Explanations.

Based on the results of the AIF360 tool, the primary difference in mean outcomes between unprivileged and privileged groups is −0.145954. We transformed our data into an AIF360 dataset and retrained our COMPAS deep neural network, then used it to generate 8, a graph of the balanced accuracy vs. the disparate impact and classification threshold.

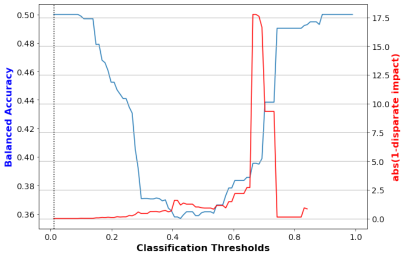
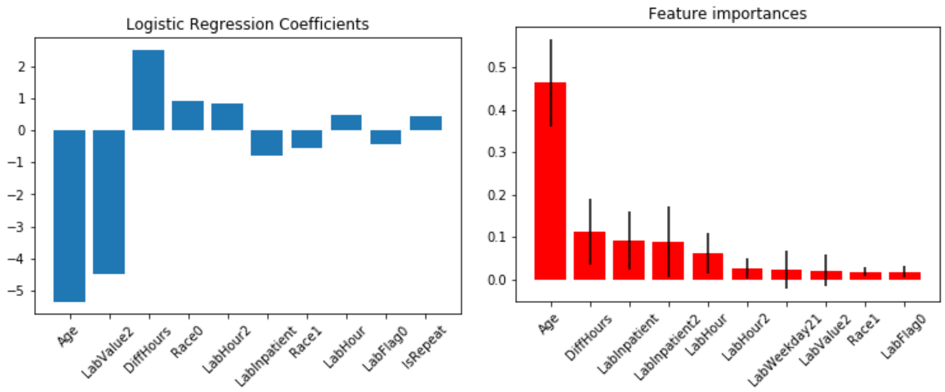


Figure 7: Deep Neural Network Explanations.

Figure 8: AI Fairness 360 accuracy versus the disparate impact for deep neural network

4.2 Health

4.2.1 Logistic Regression



(a) Logistic Regression Coefficients. (b) Random Forest Feature Importance

Figure 9: Explainable output for Healthcare Data

Logistic regression was used to evaluate Ad-hoc explainability. A regressor was built with 94.76% accuracy of predicting three-year survival rate. Coefficients from this model can be seen on the lefthand side of Figure 9. Based on these results, a user can determine that based on the model, age, lab value, and hours between consecutive tests were the most impactful variables. While this is valuable, there is no additional explanation or justification. A user might not catch to the fact that LabHour is as significant as it is.

4.2.2 Random Forests

A Random Forest model was also built with an accuracy of 90%. Similarly to the other datasets above, feature importance from the resulting model could also be used to extract which variables were most important for the model. These results can be seen on the righthand side of Figure 9. Interestingly, the top parameters are similar between Logistic Regression and Random Forest and the differences are not explained.

4.2.3 Neural Networks

A deep neural network was built with Keras; it contained two hidden layers with 10 and 4 nodes each. It's accuracy was 90%, similar to how Random Forest performed. As is, the model revealed no information as to how decisions were made, so we employed two XAI tools: LIME and Fairness 360. Similar to other datasets, LIME presented information for singular samples. While these results are revealing, most users would be interested in a more global response. Furthermore, AIF360 was used, but we were not able to generate unique differences between individuals who had labs done in the night vs in the day, preventing this model from being able to assist in identifying the major issue of fairness in this data set.

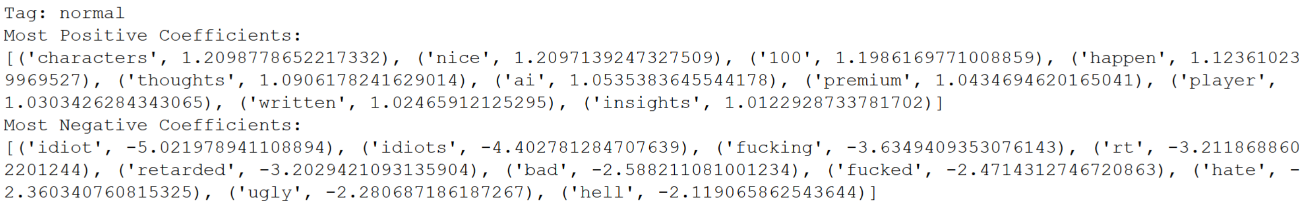
4.3 Twitter

The classifiers we used to analyze the Twitter data were random forests, logistic regression, and LSTM neural networks. All of our models were trained on Founta et al. 's hate speech twitter collection [10], which labeled tweets by derogatory/normal/spam. We tested on Blodgett et al.'s African-American English on Twitter collection, which included tweets labeled by race of poster, and asked our model to predict whether their class [5]. Additional evaluation was necessary to determine whether the results of our model on the test data were biased.

4.3.1 Logistic Regression

Our model's final accuracy was 0.777. It was then tested on Blodgett et al.'s African-American English on Twitter collection [5]. It found a low proportion of derogatory tweets - about 3%-5%.

We analyzed the logistic regression model using its in-built features. By combining the feature names from the vectorizer and the coefficients of the logistic regression classifier, we can see how much the inclusion of a given word weights a tweet's predicted class towards or away from another class. This easily lets users see how to adjust a tweet's wording to classify it differently. The explanations generated by logistic regression do not directly answer questions of fairness, but by comparing the words most heavily correlated with a given race and then analyzing how those words are weighted towards different classes can reveal fairness issues.



```
Tag: normal
Most Positive Coefficients:
[('characters', 1.2098778652217332), ('nice', 1.2097139247327509), ('100', 1.1986169771008859), ('happen', 1.12361023
9969527), ('thoughts', 1.0906178241629014), ('ai', 1.0535383645544178), ('premium', 1.0434694620165041), ('player',
1.0303426284343065), ('written', 1.02465912125295), ('insights', 1.0122928733781702)]
Most Negative Coefficients:
[('idiot', -5.021978941108894), ('idiots', -4.402781284707639), ('fucking', -3.6349409353076143), ('rt', -3.211868860
2201244), ('retarded', -3.2029421093135904), ('bad', -2.588211081001234), ('fucked', -2.4714312746720863), ('hate', -
2.360340760815325), ('ugly', -2.280687186187267), ('hell', -2.119065862543644)]
```

Figure 10: Words that had the greatest impact on the inclusion or exclusion of a tweet in the class ”normal” (i.e. whose appearance was most closely correlated with that class in the training data.

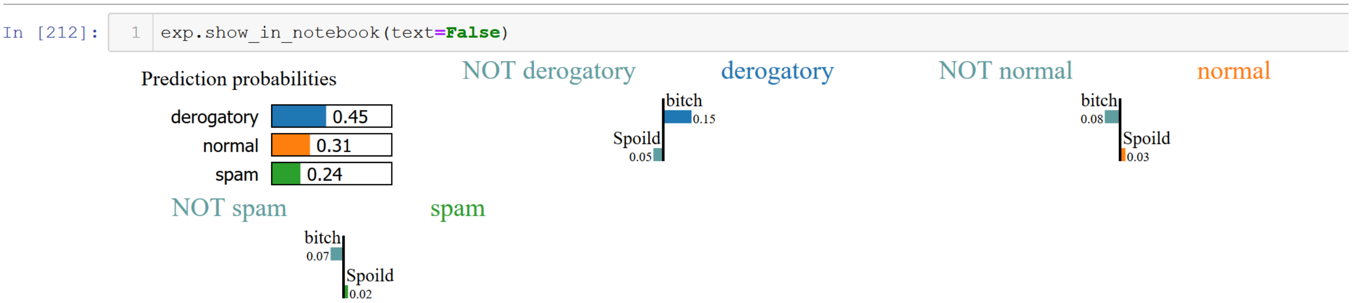
4.3.2 Random Forests

The final accuracy of our random forests model was 0.715, and this model was saved in *model_RF3.pkl*. When it was tested on the African American English on Twitter dataset, it predicted the highest proportion of derogatory tweets to all tweets out of all our models. Our training dataset contained approximately 13.3% derogatory tweets, but our random forests model found 16%-23% of the tweets to be derogatory. Because our text data did not have clear feature names, it was difficult to directly interpret the rules of the tree by exporting it to text.

4.3.3 LSTM

After 50 epochs, our LSTM model had a validation accuracy of 0.74. When it was tested on Blodgett’s African American English dataset, it found a similar proportion of derogatory tweets as the logistic regression model did, about 3%-5%. This suggests that logistic regression and LSTM have somewhat comparable results.

Unlike logistic regression and random forests, however, the LSTM required us to use outside XAI tools to explain it. We began by using LIME to evaluate individual samples.



```
In [212]:  1 exp.show_in_notebook(text=False)
```

Figure 11: Example LIME output for a tweet.

Secondly, we used AIF360 to evaluate our model for bias. Since AIF360 requires that we select privileged groups from our dataset, we classified our test data using the LSTM model outlined about. The predicted classes were combined with the original test data and evaluated in regards to the whether tweets labeled ”African American” were more likely to be labeled ”derogatory” than other races. We found a difference in mean outcomes between tweets by African Americans and tweets by posters of other races of only 0.8%, showing a lack of bias in the predictions of our LSTM model.

5 Discussion

The results of our evaluations were used to complete the rubric shown in Section 3. This section discusses our evaluation of each XAI tool in more detail. Each subsection for an XAI tool considers the results of our analysis for all three datasets.

5.1 Logistic Regression

The logistic regression model is considered to be a simple version of a neural network that has some ad-hoc explainability. Viewing the coefficients of the regression shows the influence of the features on the model's classification decision. These coefficients can be viewed locally for a single sample or globally for the feature values that have the greatest effect on classification. Logistic regression is easy to implement and its coefficient are easy for a human understand and interpret. However, they require additional processing to understand fairness. Furthermore, logistic regression does not allow developers to highlight sensitive attributes; additionally, since it is just a model, it does not address issues with preprocessing or model choice.

5.2 Random Forests

Random forests are bagged decision tree models that split on a subset of features on each split. Random forests are popular due to their versatility, parallelization, high training speed, compatibility with high dimensionality, ability to handle unbalanced data, and low bias and variance. However, it also has multiple drawbacks, such as a low level of interpretability, high memory usage, and relying on parameter tuning to avoid overfitting.

Feature importance is one of the ways that random forest can present some explanations. In this plot, the features are sorted from the most to the least influential. Features with importance less than the mean value will be considered as the ones with negative influence while other ones have positive impacts on the target variable. We can also generate the decision rules for individual decision trees in our random forest. These rules are human-readable in theory, but with large number of features and/or complex decision trees they quickly become impractical to actually interpret by hand. Furthermore, the decision rules need additional evaluation to determine how they correspond to fairness.

5.3 LIME

Explainability is most often discussed in the context of deep neural networks with numerous hidden layers and sophisticated architecture. Therefore, we also decided to create neural network models based on our data and use XAI tools to generate explanations.

LIME looks into specific predictions made by any classifier and provides explanations of how much each feature impacted the final decision. Like logistic regression, LIME gave detailed information on what details of a sample lead to the model's decision about what class it is. It additionally presents the information in a more user-friendly way that allows users to easily visualize the magnitude of each word's influence on the model's decision and whether that influence was positive or negative. LIME, however, does not give a global perspective on what the most notable features for the whole model are. This leads to LIME's greatest weakness as an XAI tool for fairness: it requires significant post-processing to find patterns across explanations for samples and compare these patterns across sensitive. Furthermore, LIME does not allow users to denote specific sensitive features that the model should pay special attention to in its explanations.

5.4 AIF360

In the AIF360 tool, users are able to specify sensitive attributes and privileged vs. unprivileged groups. AIF360 then considers the fairness and parity between user-determined unprivileged and privileged groups for each sensitive attribute. Interestingly, fairness and parity are outlined across subgroups to provide valuable insight for users. Additionally, it can handle the imbalanced data and bias appropriately. Its main drawback is its limited ability to handle multiclass classifiers and categorical data, as well as the need to perform additional processing to perform more sophisticated analysis regarding the fairness.

6 Conclusion & Future Work

Based on our rubric and our experiences evaluating our models, we have several recommendations for designers of future XAI tools.

1. LIME's level detail for individual samples and AIF360's global perspective on bias in the data are both extremely valuable. Future XAI tools should work to incorporate flexibility so that users can choose whether they want to view individual or global-level explanations.

2. Allowing users to select features that they want to focus their evaluations on is a key feature for explainability. This has a clear purpose for evaluating fairness in regards to protected class, as well as in non-fairness-related explanations where certain features should be weighed more or less heavily in class selection than others.

3. Consider the target audience and purpose of your XAI tool. As ML permeates more areas of our society, multiple groups of people will need to evaluate ML models. These groups will have different goals and different levels of prior knowledge of ML. Consider the case of a ML developer who wants to perform QA tests of their model versus a government auditor who wants to ensure that a model is conforming to legal standards - these two individuals will be interested in different levels of detail presented in different ways. XAI tools with specific purposes as well as target audiences outside of ML developers will be increasingly valuable in the coming years.

We are interested in continuing our work by incorporating more XAI tools into our evaluations, such as Google's What-If Tool and IBM's AI Explainability 360 (an expansion of AI Fairness 360 for deep neural networks). Furthermore, we want to revise our rubric and our scoring mechanisms so that we can provide more detailed feedback. There is a great deal of nuance in the specific functions of XAI tools that may not be covered in our current 0/1 scoring model; some tools may be more flexible or powerful than others, even if they technically fulfill the same functions.

Our evaluations reveal that while current XAI tools provide important functions for data and model analyis, they are still lacking when it comes to analyzing fairness. This is a critical gap in XAI research, given several notable scandals in recent years in regards to bias in ML [11] [21]. ML developers and outside auditors and critics alike want to know if their models are fair, and XAI tools should address this need.

7 Links

All the codes and figures will be accessible using the following link (`https://github.com/kalikhademi/MLinprivacy`) on GitHub.

References

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

[2] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica, May 23* (2016), 2016.

[3] Bansal, G. Explanatory dialogs: Towards actionable, interactive explanations. In *2018 Proceedings of the conference on Artificial Intelligence, Ethics, & Society* (12 2018), ACM, pp. 356–357.

[4] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[5] Blodgett, S. L., Green, L., and O'Connor, B. 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* (2016), Association of Computational Linguistics, pp. 1119–1130.

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

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

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

[9] Dressel, J., and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science advances 4*, 1 (2018), eaao5580.

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

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

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

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

[14] Lipton, Z. C. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[15] Mittelstadt, B., Russell, C., and Wachter, S. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency* (2019), ACM, pp. 279–288.

[16] Ras, G., van Gerven, M., and Haselager, P. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pp. 19–36.

[17] Ribeiro, M. T., Singh, S., and Guestrin, C. 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* (2016), ACM, pp. 1135–1144.

[18] Saviano, M., and Tieu, S. When to stop-and-frisk.

[19] Veale, M., and Binns, R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society 4*, 2 (2017), 2053951717743530.

[20] Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, p. 601.

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

Appendices

A Detailed Breakdown of Rubric

Issues with Biased Data.

- Imbalanced Data. [19]
  - Does the XAI tool identify subpopulations that are undersampled?
  - Does XAI recognize the samples which had less influence on the final decision?
  - Does XAI tool detect that data is imbalanced or not?
- Influential Variable Identification.
  - Does the XAI tool identify variables that are most influential on the final system decision for a sample? [12]
  - Does the XAI tool identify variables that are most influential on the final system decision across all samples?
  - Does the XAI tool show for each variable, the influence it had on the final system decision?
  - Does the XAI tool show whether the influence it had on the final system decision was positive or negative for each variable?
  - Does the researcher have the ability to use a metric of their choice to evaluate how variables influence the final model?
- Pre-processing issues [19].
  - Does the XAI tool identify misuse of categorical variables in the machine learning model? (Loss of information)
- Sensitive attributes.
  - Does the researcher have the ability to highlight sensitive attributes in the XAI tool?
  - Does the XAI evaluate the model with respect to the sensitive attributes?

Issues with Machine Learning Models.

- Model-Specific influences [19].
  - Does the XAI tool identify/predict issues that might arise from the choice of model?
  - Does the XAI tool identify/predict issues that might arise from the selection of parameters?
  - Does XAI tool identify the issues raising from feeding the encoded data into the machine learning algorithm?
  - Does the XAI tool identify statistical changes made in the model creation that could potentially change the meaning of the results?
- Accuracy Equity [19]
  - Does the XAI tool identify inconsistencies with accuracy?
  - Does the XAI tool identify distribution of desired prediction across all sub-groups?

Issues with XAI Results.

- Target Audience [15]
  - Does the XAI tool identify a primary target audience for its explanations?
  - Does the XAI tool identify a primary function for its explanations?
- Presentation of explanations [15]
  - Can the results of the XAI tool be understood without additional processing?
  - Does the XAI tool tell the user how the input would have to be changed in order to be classified differently? [3]

B Detailed Methodology

B.1 COMPAS

B.1.1 Preprocessing

The COMPAS dataset was used to create all the machine learning models. Logistic regression and random forest techniques were implemented to show the results of the method with some level of explainability. Due to the black-box nature of deep learning methods, we have used the XAI tools mentioned in the Background section to provide more explanations for results. In the preprocessing steps, the categorical variables have been encoded using the label encoder of Scikit Learn's library. As ProPublica's study showed that there is a correlation between the recidivism score and the length of stay for defendants, we have computed the feature-length of stay and used it in our analysis. The dataset has been split into the train, and the test dataset consists of 9405 and 2352 records, respectively. The following list of features have been used in our analysis:

- age
- priors_count
- days_b_screening_arrest
- Length Of Stay in jail
- c_charge_degree_ids
- race_ids
- sex_ids

B.1.2 Neural Network Construction

Therefore, we have defined a simple deep neural network using Keras API in the TensorFlow backend. The network has consisted of two hidden layers with six neurons in each of them and three neurons in the output layer. Relu activation function has been used in both hidden layers, while the output layer has been computed using the softmax activation function. To optimize the learning across our deep neural network, we have used Adam optimization while computing the loss based on the "categorical_crossentropy" method.