Evaluating Fairness in Predictive Policing Using Domain Knowledge

Ava Downey, Sheikh Rabiul Islam, Md Kamruzzman Sarker
University of Hartford
avdowney@hartford.edu, shislam@hartford.edu, sarker@hartford.edu

Abstract

As an increasing number of Artificial Intelligence (AI) systems are ingrained in our day-to-day lives, it is crucial that they are fair and trustworthy. Unfortunately, this is often not the case for predictive policing systems, where there is evidence of bias towards age as well as race and sex leading to many people being mistakenly labeled as likely to be involved in a crime. In a system that already is under criticism for its unjust treatment of minority groups, it is crucial to find ways to mitigate this negative trend. In this work, we explored and evaluated the infusion of domain knowledge in the predictive policing system to minimize the prevailing fairness issues. The experimental results demonstrate an increase in fairness across all of the metrics for all of the protected classes bringing more trust into the predictive policing system by reducing the unfair policing of people.

Introduction

A predictive policing system predicts either who may be involved in a crime, or where a crime may take place. The information gathered through this system can be used for crime control and forecasting, though the data used in these systems are often not disclosed and may contain inaccuracy and bias (Richardson et al., 2019).

Data with missing values, is incorrect, or is a bad representation of what the dataset is supposed to encompass, is called “dirty data”. The term can also be expanded to include data that has been gathered through corrupt, biased, and unlawful practices. The risk assessment score, founded on this dirty data, is remarkably unreliable in predicting if someone will recommit a violent crime. While not explicitly in practice now, up until the 1970's race (Angwin et al., 2016), nationality, and skin color were used in determining the risk assessment of a person. Now, metrics such as poverty, joblessness, and social marginalization are used in predicting risk, although it still comes across issues with race. This assessment is only slightly more accurate than flipping a coin and is used to make decisions over a person's freedom and criminal sentencing in many states (Angwin et al., 2016). While there have been several documented instances of predictive policing systems used in government investigations, consent decrees, or other documentation of corrupt, racially-biased, or otherwise illegal police practices, there has been little to no effort observed of police departments or predictive system vendors to attempt to mitigate the problem (Richardson et al., 2019).

The consequences of this are not insignificant. Dirty data in the criminal justice system negatively impacts the areas that end up being overpoliced because newly observed data reflects this increased police presence disproportionally. Ignorance and bigotry in the police force lead to long-term consequences for already marginalized and vulnerable communities. This makes it harder for them to obtain jobs, housing, and get help at publicly run shelters. It also perpetuates negative stereotypes about these communities which can lead to improper reports of suspicious activity which further feeds into the dirty data (Richardson et al., 2019). Finding a way to mitigate bias through predictive policing can have many advantages for communities across America, especially those of people of color. This might be done by first detecting the existence of bias and discrimination in the data, then infusing domain knowledge into the data to mitigate bias. Domain knowledge is the abstract understanding of a specific topic. In the case of this problem, domain knowledge would be helpful towards mitigating potential detected bias and fairness issues (Goolsby et al., 2022) in the predictive policing system.

The goal of this research paper is to explore ways to increase public safety in a more fair way. If this transition to a more “data-driven practice” is done correctly, then it will help to decrease bias issues in policing, including bias pertaining to race. Making predictive policing more equitable is a task easier said than done. By understanding and implementing domain knowledge regarding the topic, these algorithms will be able to give a more fair and unbiased risk score to individuals, therefore mitigating racial bias in policing.

Background

It is not a new finding that predictive policing algorithms have fairness concerns, or as Ninareh Mehrabi et. al broadly defines, the "absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making" (Mehrabi et al., 2021). California has banned the use of predictive policing in an effort to eliminate racism in policing because of how it is biased toward people of color. These algorithms reinforce racist practices because they are based on previous police data where there was significant racial bias (Asher-Schapiro, 2020). Since these practices are still being followed, dirty data leads to even more dirty data resulting from a constant feedback loop that dissuades the problem from being fixed.

Bias, such as racial bias in predictive policing algorithms, is not intentionally implemented into algorithms but is the result of the data it has been given. For example, ads for STEM jobs are supposed to be gender-neutral, but young women were deemed to be a more valuable and costly subgroup due to the data it was trained on. Young women were assumed to be more expensive to show advertisements to so they did not see the ads which then caused them not to apply for the jobs (Mehrabi et al., 2021). This introduction of bias into advertising data early on had profound effects on the gender disparity seen in STEM today. This same trend can be seen in how people of color have been treated by police throughout the history of the United States, and in the disparity between white people and people of color who are incarcerated today. This past bias has been ingrained into today's algorithms because of the bias in the data they are being trained on.

It is also very important that the data used is a fair representation of everyone or everything it is predicting in order to avoid discrimination. Mehrabi et al defines discrimination as a source of unfairness from either intentional or unintentional prejudice or stereotyping from humans during the data collection process (Mehrabi et al., 2021). The sample of data being used should encompass an accurate subsection of the population the data is being used to represent. If this is not taken into account, it can have dangerous outcomes that could have been avoided. For example, many medical datasets are based on people of European descent, meaning that minority groups are not equally represented in medical studies. Discrimination, like bias, can also lead to unfairness (Mehrabi et al., 2021).

Arrest records are just one way that data is being collected to predict someone's likelihood to commit a crime. Due to recent advances in technology, everyone is surveilled through their routine activity in day-to-day life. If someone is to participate in society, such as using a bank, sending an email, receiving medical care, or conducting an internet search they are being surveilled. Big data surveillance allows the police to gather data, and identify suspicious patterns, locations, activities, and individuals before a criminal incident has occurred. The fourth amendment of the United States Constitution prohibits unreasonable search and seizure without probable cause, but big data surveillance allows data to be collected without explicitly violating this right. Big data has the potential to reduce bias, increase efficiency, and improve prediction accuracy when used in surveillance practices. Data-driven decision-making, founded on this new source of data, has just become more prevalent in the practices of law enforcement in recent decades. The switch to big data analytics both amplifies and transforms the surveillance practices used by the police force previously. For example, an officer's assessment of the risk for a person to re-offend is now supported by risk scores. Based on the way that big data may be implemented into policing practices and calculations of risk scores, discriminatory practices may become negligible (Brayne, 2017). However, this is a difficult goal to reach.

The data we have currently is difficult to separate into "good" and "bad", because of the number of different types of manipulation in the data (Richardson et al., 2019). There are many algorithms developed to try to tackle this issue. These algorithms are able to mitigate some bias that people might introduce from burnout or other environmental factors but also introduces new bias based on unfairness in the data it is trained from.

Bias in artificial intelligence can be addressed through three different stages. The first stage is pre-processing where discriminatory data can be transformed or removed. This enables the dataset to be in a form suitable to be worked on in the next stage. In-processing is the second stage and is where algorithms attempt to remove bias through the training process. This is where knowledge is learned about the data using different algorithms. The third stage is post-processing, where knowledge extracted from the previous step can be further processed, simplified, and documented against previous knowledge (Bruha and Famili, 2022).

Casual models remove attributes such as gender or race when designing certain systems and policies because they are considered sensitive attributes during the in-processing stage. This helps to solve fairness-related concerns. In order to tackle this issue and develop more fair machine learning algorithms, tools have been developed to warn developers about bias in their data (Mehrabi et al., 2021).

One tool is AIF360 (AI Fairness 360, 2021), an open-source tool created by IBM which is used to find and mitigate bias and fairness concerns in algorithms and data, such as those of predictive policing. The tool provides researchers a platform to experiment with bias detection and mitigation, contribute new algorithms, and contribute new datasets to analyze bias. A Dataset consists of training da-

ta, validation data, testing data, and associated protected attributes. The training section of a dataset is the data that the models are trained on and is the largest subset of the data. The validation section of a dataset is used to tune hyper-parameters and for model selection. This helps to create the best model. The testing subsection of a dataset is used to evaluate the performance of the final model based on the training and validation data (Myrianthous, 2021). Any dataset preprocessed using AIF360 is randomly split into 50% training, 20% validation, and 30% test partitions. This allows the data to be thoroughly tested for accuracy based on a prediction. There are three paths when making a fair prediction that each corresponds to a bias mitigation algorithm implemented in AIF360. They are fair pre-processing, fair in-processing, and fair post-processing. They improve fairness metrics by modifying training data, learning algorithms, or predictions. As the pipelines run, there are several stages where bias can be assessed using the fairness metrics of the tool (Bellamy et al., 2019). By using tools such as AIF360, the bias, and fairness of predictive policing algorithms can be tested and improved.

Despite many existing works on predictive policing, external knowledge (i.e., domain knowledge) infusion (Islam et al., 2019, 2020a, 2020b; Goolsby et al., 2022) still has unlimited potential to improve fairness in predictive policing. Our work is an attempt to fill the gap.

Data

The data used by Chicago Police for their predictive policing models was made publicly available in 2020 and is one of the only of its kind (Chicago Data Portal, 2017). This dataset is called *Strategic Subject List (SSL) - Historical* (Chicago Data Portal, 2017) and it is used by the city of Chicago, Illinois to predict the likelihood of residents being involved in a shooting. This dataset is a de-identified listing of arrest data for 398,684 people from August 1, 2012, to July 31, 2016, used by the Chicago Police Department's Strategic Subject Algorithm, or their predictive policing algorithm. The SSL score is the main focus of this dataset and is based on several variables which include

1. The number of times an individual has been a victim of a shooting
2. The age of an individual at their most recent arrest
3. The number of times an individual was the victim of aggravated battery or assault
4. The number of violent offense arrests previously of an individual
5. The number of narcotics arrests previously of an individual
6. The number of unlawful use of a weapon arrests of an individual
7. An individual's recent trend in criminal activity
8. The gang affiliation of an individual

Knowing this information, the dataset can then be preprocessed into a form where algorithms can be applied to predict the SSL, and AIF360 can be applied to determine the bias of the data (Posadas, 2017).

Experiments and Results

We used Python and Python-based packages for our experiment. The Python package for AIF360 was used to quantitatively measure bias among the different algorithms and data. The dataset was split using a 75/25 training and test split and many predictive models were run on the data including logistic regression, random forest, support vector machine (SVM), and artificial neural networks (ANN). This is done using several Python packages such as scikit-learn and TensorFlow.

Data Preprocessing

The raw dataset includes 48 columns and 398,684 rows. Many of these columns are not necessary for the scope of the issue being looked into and are removed. This leaves only 13 columns in the dataset, those columns being the SSL score, its predictors, and the protected classes being looked into. The predictors previously calculated by the city of Chicago are described in Table 1. The protected classes include age with people under 30 years old as the disadvantaged group, sex with females as the disadvantaged group, and race with non-white people as the disadvantaged group.

Predictor	Description
PREDICTOR RAT AGE AT LATEST ARREST	The age of an individual at their latest arrest
PREDICTOR RAT VICTIM SHOOTING INCIDENTS	The number of times an individual has been a victim of a shooting
PREDICTOR RAT VICTIM BATTERY OR ASSAULT	The number of times an individual has been a victim of aggravated assault and/or aggravated battery
PREDICTOR RAT ARRESTS VIOLENT OFFENSES	The number of times an individual has been arrested for a violent offense
PREDICTOR RAT GANG AFFILIATION	If the individual is involved in a gang or not
PREDICTOR RAT NARCOTIC ARRESTS	The number of times an individual has been arrested for a narcotics offense
PREDICTOR RAT TREND IN CRIMINAL ACTIVITY	The trend of an individual's recent criminal activity
PREDICTOR RAT UUW ARRESTS	The number of times an individual has been arrested for Unlawful Use of Weapons

Table 1: Descriptions of Predictor Columns in the *Strategic Subject List - Historical* dataset (Chicago Data Portal, 2017)

The empty rows are also dropped, and the remaining ones are scaled to be either 0 or 1. The SSL score was split into high risk (score above 250) and low risk (score at or below 250), with the high-risk scores being scaled to 1, and low risk 0. The high-risk SSL score was determined to be above 250 by the city of Chicago and means that a person will be more surveilled than their low-risk counterparts due to how to predictive policing algorithm classifies them (Posadas, 2017). Race, sex, and age were also scaled with the disadvantaged group(s) being scaled to a 0 and the advantaged groups being scaled to a 1. Lastly, the predictor columns were one-hot encoded in order to be used in model creation. This leaves the preprocessed dataset with a shape of 13 columns and 227,070 rows.

Domain knowledge was then able to be integrated into the dataset. This was done using the census tract in which each of the deidentified people (rows) in the dataset reside. The Python packages, geopandas, and census were used to read data from the 2014 census. From the census data being read, the average rate of education, employment, and poverty were calculated and introduced to the dataset. The goal of incorporating this domain knowledge into the predictive policing algorithms used today is to minimize bias and fairness concerns while maintaining or improving the accuracy of assessment previously achieved through these algorithms. Education, employment, and poverty were chosen due to the correlation between them and crime (Quednau, 2021). This leaves the dataset with a shape of 16 columns and 227,070 rows.

Model Creation

The SSL score is the defining metric of the dataset and is the value to be predicted based on the other features of the dataset. Figures 1, 2, and 3 all show how the dataset splits the different demographics into high and low-risk SSL scores. The data is scaled by density to account for disproportional amounts of data between the protected and unprotected classes with the high-risk (blue) bars adding up to 1.000 or 100% and the low-risk (orange) bars adding up to 1.000 or 100% to show how the data favors or disfavors a group. Looking at Figure 1, it can be seen that there is a bias of 5.4% between high and low-risk SSL scores among White-Hispanic people in the dataset, while White people have a bias of 4.4% in favor of them having a low SSL Score. Figure 2 highlights this lack of fairness as well with males being 4% more likely to be low risk than females who are 4% more likely to be high risk. However, the most significant disparity is among age as seen in Figure 3. There is a clear bias toward people over the age of 40 to have a low SSL score. While it is unlikely that every person under 20 years old in this dataset would be involved in a shooting, 0% of people in the low-risk SSL Score are less than 20 years old, and only 0.1% of people ages 20-30 years old make up the low-risk SSL Score.

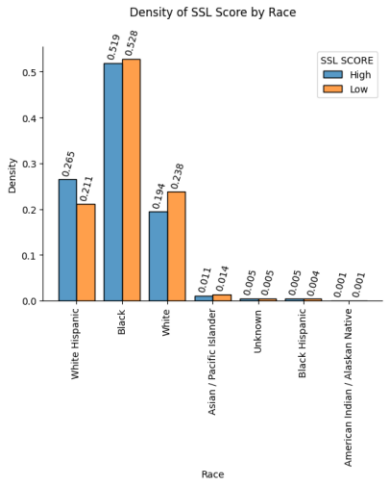

Figure 1: Density of SSL Score by Race

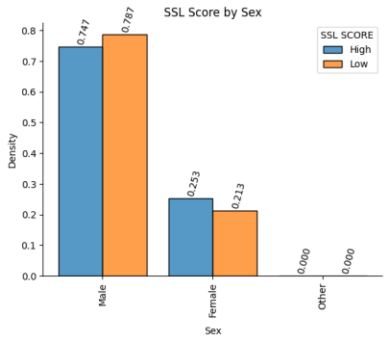

Figure 2: Density of SSL Score by Sex

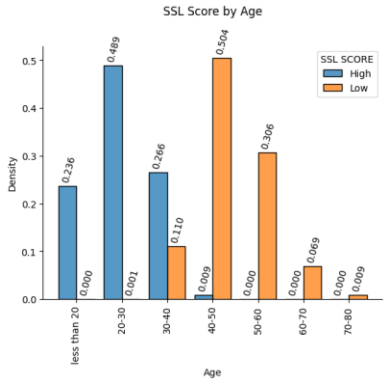

Figure 3: Density of SSL Score by Age

These graphs can be backed up through several fairness metrics, though it is important to determine the best model for the problem in order to get the most accurate fairness assessment. Since the city of Chicago has not

disclosed the algorithm that they used to determine the exact SSL score of an individual, several algorithms were tested to get the closest to accurately predicting the SSL score possible from the base data. Though many of the algorithms had similar results, Random Forest was found to be one of the best models for the task. The Random Forest algorithm was then run with the Domain Knowledge to determine how it affects the classification metrics. Table 2 shows the accuracy, precision, recall, and F1 scores of the Random Forest algorithm with the base data as well as the base data with poverty rate, education rate, and employment rate.

Classification Metric	Without Domain Knowledge	With Domain Knowledge
Accuracy	0.9689	0.964
Precision	0.964	0.9662
Recall	0.9941	0.9847
F1 Score	0.9788	0.9753

Table 2: Classification metrics for Random Forest with and without Domain Knowledge

Incorporating domain knowledge into the prediction process only minimally affects the classification metrics of the models. The accuracy is only affected by -0.49% between the model without domain knowledge and the model with domain knowledge. The recall and F1 scores of the model are also only minimally affected with a -0.94% difference with recall and a -0.35% difference for the F1 score. Precision was positively affected by a change of 0.22% between the model without domain knowledge and the model with it.

Fairness Metrics Calculation

Since incorporating domain knowledge into the data only minimally affects the accuracy of the model, the fairness of the random forest models both with and without domain knowledge can be evaluated. This is calculated using the IBM AIF360 Python package on a binary version of the dataset. The protected attribute (protected class) is selected, and the privileged and unprivileged values are specified in the code. The target variable is also specified, which in the scope of this thesis is the 'SSL Score'. Classification algorithms are then run using these conditions, which for this project included logistic regression, random forest, linear SVM, RBF SVM, and an ANN. The datasets are then reweighed by AIF360 and the fairness metrics are computed for each of the algorithms (Goolsby et al., 2022). A sample of 75,000 rows was selected to compute the fairness metrics for the original model as well as the two best models with Domain Knowledge shown in Table 3.

	Original Data			w/ Poverty-Employment Data			w/ All Census Data		
	Age	Race	Sex	Age	Race	Sex	Age	Race	Sex
Statistical Parity Difference	0.5567	0.0648	0.0399	0.5536	0.0618	0.0397	0.5534	0.0615	0.0371
Disparate Impact	2.2566	1.0954	1.0546	2.2411	1.0892	1.0517	2.2402	1.0887	1.0506
Average Odds Difference	0.2308	0.0132	0.0077	0.2276	0.0091	0.0055	0.2273	0.0090	0.0047
Equal Opportunity Difference	0.0631	0.0039	0.0014	0.0589	0.0023	-0.0003	0.0579	0.0017	-0.0011
Theil Index	0.0224	0.0288	0.0234	0.0218	0.0206	0.0226	0.0215	0.0205	0.0222

Table 3: Fairness metrics for Random Forest Algorithm with and without Domain Knowledge

The fairness metrics computed include statistical parity difference, disparate impact, average odds difference, equal opportunity difference, and Theil index. These metrics are briefly defined below (Goolsby et al., 2022):

- The *statistical parity difference* is fair when it is equal to 0 and represents the ratio between the number of favorable outcomes for the unprivileged group to the number of favorable outcomes for the privileged group.
- The *disparate impact* is fair when it is equal to 1 and represents the ratio of a favorable outcome for the unprivileged group over the favorable outcome of the privileged group.
- The *average odds difference* is fair when it is equal to 0 and represents the average difference of false and true positive rate between the unprivileged and privileged group.
- The *equal opportunity difference* is fair when it is equal to 0 and represents the difference of true positive rates between the privileged and unprivileged groups.
- The *Theil index* is also fair when it is equal to 0 and is a measure of the inequality in benefit allocation for individuals.

Looking into the data in Table 3, it can be seen that adding domain knowledge on poverty and employment into the model increases the fairness across all of the metrics for all of the protected classes. Figures 4, 5, 6, 7, and 8 depict the comparisons of the fairness metrics between the original data, and the data with the domain knowledge with the best outcomes for mitigating bias. The most drastic improvement was seen in the Theil index for race. The original data had a Theil index score of 0.0288, and after adding domain knowledge on poverty and employment it dropped to 0.0206, and with domain knowledge on poverty, employment, and education it dropped further to 0.0205. While the domain knowledge does not make the

predictive policing algorithm perfectly fair, it is a step in the right direction toward achieving a system with no bias or fairness concerns.

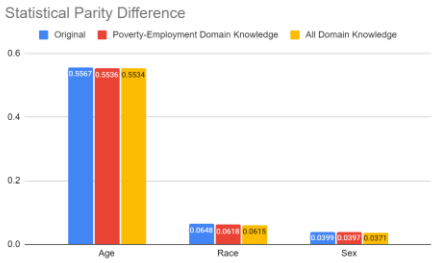

Figure 4: Statistical Parity Difference for Random Forest with and without Domain Knowledge

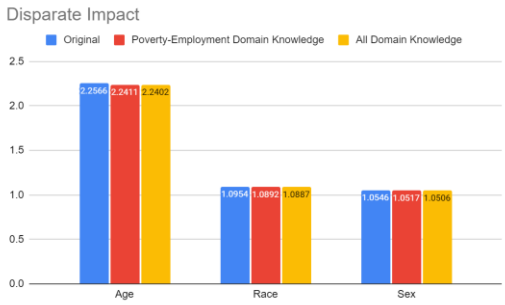

Figure 5: Disparate Impact for Random Forest with and without Domain Knowledge

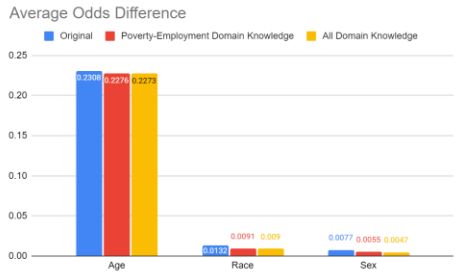

Figure 6: Average Odds Difference for Random Forest with and without Domain Knowledge

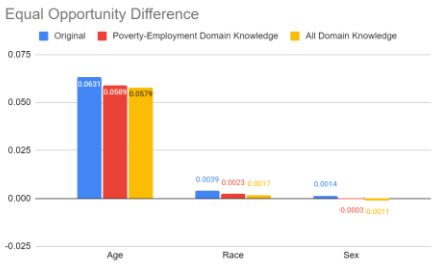

Figure 7: Equal Opportunity Difference for Random Forest with and without Domain Knowledge

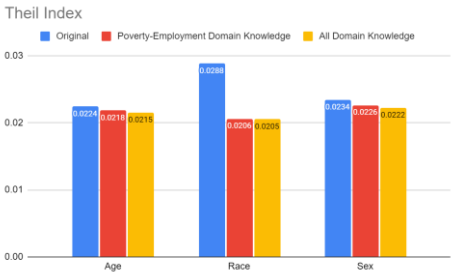

Figure 8: Theil Index for Random Forest with and without Domain Knowledge

Conclusion

We explored useful domain knowledge and infused that into a predictive policing system to mitigate fairness-related disparities among different protected groups. Looking specifically at data provided by the city of Chicago, domain knowledge on poverty rate, education rate, and employment rate has been integrated into their predictive policing algorithms to increase the fairness in predictions. Looking into poverty, education, and employment help to even out the implicit bias on age, sex, and race by adding more features that correlate to crime. Other possible pieces of domain knowledge might have similar or better effects on improving fairness such as income or earnings. Going forward, looking into these new variables or other similar variables may be able to reduce bias even more than seen in this work.

Acknowledgments

This work was supported in full or in part by a University of Hartford Greenberg Junior Faculty Research Grant. This support does not necessarily imply endorsement by the University of Hartford of project conclusions.

References

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). *Machine bias*. ProPublica. Retrieved May 20, 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Asher-Schapiro, A. (2020, June 24). *California city bans predictive policing in U.S. first*. Reuters. Retrieved September 22, 2022, from https://www.reuters.com/article/us-usa-police-tech-trfn/california-city-bans-predictive- policing-in-u-s-first-idUSKBN23V2XC

Bellamy Rachel, K. E., Kuntal, D., Michael, H., Hoffman Samuel, C., Stephanie, H., Kalapriya, K., ... & Zhang, Y. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4), 5.

Brayne, S. (2017). Big data surveillance: The case of policing. *American sociological review*, *82*(5), 977-1008.

Bruha, I., & Famili, A. (F. (n.d.). *Postprocessing in machine learning and Data Mining*. Retrieved October 24, 2022, from https://kdd.org/exploration_files/KDD2000PostWkshp.pdf

Chicago, C. of. (2017, May 1). *Strategic subject list - historical: City of chicago: Data Portal*. Chicago Data Portal. Retrieved October 6, 2022, from https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np

Goolsby, T., Islam, S. R., & Russell, I. (2022). Advancing Fairness in Public Funding Using Domain Knowledge. *AAAI 2022 Spring Symposium.*

Islam, S. R., Eberle, W., Ghafoor, S. K., Siraj, A., & Rogers, M. (2020a). Domain knowledge aided explainable artificial intelligence for intrusion detection and response. *AAAI-MAKE 2020*.

Islam, S. R., Eberle, W., & Ghafoor, S. K. (2020b). Towards quantification of explainability in explainable artificial intelligence methods. *FLAIRS-33*.

Islam, S. R., Eberle, W., Bundy, S., & Ghafoor, S. K. (2019). Infusing domain knowledge in ai-based" black box" models for better explainability with application in bankruptcy prediction. *KDD 2019, Workshop: Anomaly Detection in Finance*

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Myrianthous, G. (2021, August 3). Training vs testing vs Validation Sets. Retrieved October 17, 2022, from https://towardsdatascience.com/training-vs-testing-vs-validation-sets-a44bed52a0e1

Posadas, B. (2017, June 26). *How strategic is Chicago's "Strategic subjects list"? upturn investigates.* Medium. Retrieved October 6, 2022, from https://medium.com/equal-future/how-strategic-is-chicagos-strategic-subjects-list-upturn-investigates-9e5b4b235a7c

Quednau, J. (2021). How are violent crime rates in US cities affected by poverty?. *The Park Place Economist*, *28*(1), 8.

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online*, *94*, 15.

Strategic subject list - historical: City of chicago: Data Portal. Chicago Data Portal. (2017, May 1). Retrieved October 6, 2022, from https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np

United States. Routledge. AI Fairness 360. Retrieved February 18, 2023, from https://aif360.mybluemix.net/

.