Improving the Ethics of Prediction with Interpretable Models[*]

Mohammed Almanaa
Department of Civil Engineering

Reid Bixler
Department of Computer Science

Thomas Lux
Department of Computer Science

Stefan Nagy
Department of Computer Science

Sirui Yao
Department of Computer Science

ABSTRACT

(All members) Large volumes of data allow for modern application of statistical and mathematical models to practical social issues. Many applications of predictive models like criminal activity heatmapping, recidivism estimation, and child safety scoring rely on data that is incomplete, incorrect, or biased. Many sensitive social and historical issues can unintentionally be incorporated into predictions causing ethical mistreatment. This work attempts to address data bias by using models that produce interpretable results. The improvements offered by these models are twofold: (1) bias can be identified either statistically or by human users on a per-prediction basis; (2) data can be cleaned for bias on a per-prediction basis. Modeling techniques similar to those presented in this work could not only strengthen the ethical application of data science, but also make the process of cleaning and validating data manageable in the long term.

KEYWORDS

Ethics, Predictive Models, Multilayer Perceptron, Decision Tree, Local

ACM Reference Format:
Mohammed Almanaa, Reid Bixler, Thomas Lux, Stefan Nagy, and Sirui Yao. 2018. Improving the Ethics of Prediction with Interpretable Models. In *Proceedings of Virginia Tech (GRAD5134).* ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION (MA)

With the growth of new technologies and machine learning algorithms, smart cities and urban areas are adapting advanced devices to control and monitor many aspects of our life and thus provide better service for our communities and the society as a whole. With the explosion of the massive amounts of collected data (referred to as big data), predictive models have boomed and become an interest for many fields such as healthcare, employment, finance, government [43]. The predictive models were motivated by the critical need in our daily life to anticipate the future events to prepare and make an action in advance. Succeeding in doing so would stop or mitigate the negative consequences of future events. The predictive models play a role in our daily life, helping us answer various questions, ranging from easy to complicated:

- Do I need to take my umbrella with me when leaving home?
- How long will take me to get home/work?
- How likely this applicant would succeed in the college and thus deserves to be accepted?
- How likely a criminal defendant would not commit a crime again and thus should be released?

The significant increase of the collected data has pushed researchers and agencies in developing very accurate predictive models but not paying much attention to the ethical aspects of the models. Non-Ethical models (i.e. biased models) could lead to treat people differently based on race, color, sex, or religion which is prohibited by the U.S. federal laws. That happens when human subjects are involved in the model such as criminal risk assessment tool.

Ignoring the bias in the predictive models may create serious discriminatory consequences. Many researchers have shown clear signs of unequal and racial treatment of individuals [1, 45, 46]. For example, Angwin et al. analyzed a well-known criminal risk assessment: COMPAS and concluded the outcomes of the model are racially biased [1]. They showed that "Black defendants were twice as likely as white defendants to be misclassified as a higher risk of violent recidivism, and white recidivists were misclassified as low risk 63.2 percent more often than black defendants" . The U.S. government realized the enormous potential for negative impact of the biased predictive models and thus the White House published several reports, highlighting the potential bias that could affect adversely individuals or groups [44]. Research efforts have been conducting actively in proposing ethical frameworks and potential solutions to protect principles of ethics and privacy [11, 28, 48].

Not-ethical models do not necessarily mean they contain personal characteristics as predictors, yet they could impose racial bias, for instance, if they have other predictors correlated with race. This makes developing unbiased models harder given the researchers have to dig deep into the results, making sure the bias is not existed, and thus there is a need to use models that produce interpretable results. That could make it possible for a researcher to detect the bias before making prediction.

In this paper, we attempt to address data bias by using models that produce interpretable results. The improvements offered by these models are twofold: (1) bias can be identified either statistically or by human users on a per-prediction basis; (2) data can be cleaned for bias on a per-prediction basis. Modeling techniques similar to those presented in this work could not only strengthen the ethical application of data science, but also make the process of cleaning and validating data manageable in the long term.

[*] Authors are listed in alphabetical order by last name. This work was a final project for the Ethics and Professionalism in Data Science Graduate class at Virginia Polytechnic Institute and State University.

GRAD5134, Spring, 2018, Blacksburg, Virginia

1.1 Ethical Violation: Recovering Protected Information

In an effort to prevent biased and discriminatory decisions, policy makers and practitioners have classified discriminatory attributes such as personal attributes as protected information and thus cannot be used as predictors in the model. However, researchers have shown this approach is inadequate as the discrimination and biases are still existed (i.e. indirect bias), and more importantly, protected information can be easily recovered [11, 39]. Pedreschi et al. have shown that discrimination rules could be inferred using protected German credit dataset. This was done by linking the outcomes of the model (i.e. discrimination rules) with other background knowledge. For instance, knowing there is a majority of black people living in this neighborhood means any resultant classification rules are linked with black people. Dedeo showed the existing correlation between protected and non-protected variables made it possible for unveiling the protected information [11]. They proposed one approach to overcome this issue by de-correlating category and outcome using the Information Theory.

1.2 Global Predictive Models

Classic machine learning and data science techniques applied today often rely on solving a very specific problem. They create a global predictive model with the aim of capturing trends that exist across thousands of examples. In general, these global models are constructed given data matrix $X \in \mathbb{R}^{n \times d}$, a truth function $f : \mathbb{R}^d \to \mathbb{R}$, and *labels* $f(x^{(i)})$ for row vectors $x^{(i)} \in X$, $1 \le i \le n$. These models find the solution to

$$\min_P \| \hat{f}_P(X) - f(X) \|,$$

where $\hat{f}_P : \mathbb{R}^d \to \mathbb{R}$ is the parametric approximation, $f(X)$ is used to denote the vector with components $f(X)_i = f(x^{(i)})$, and $\| \cdot \|$ is an appropriate measure. The labels may be real numbers, like probability of recidivism estimates, or categories such as "safe" or "not safe" for an at-risk child.

The difficulty with these models is that the minimization search which identifies the parameters for the model is performed over *all* data. Whenever it is time to explain a prediction, the answer is often "all data was used to capture this trend". The models of this form that will be applied are a multilayer perceptron (MLP) and a decision tree (DT).

1.2.1 *Multilayer Perceptron.*

The neural network is a well studied and widely used method for both regression and classification tasks [19]. When using the rectified linear unit (ReLU) activation function [10] and training with the BFGS minimization technique [36], the model built by a multilayer perceptron uses layers $l : \mathbb{R}^i \to \mathbb{R}^j$ defined by

$$l(u) = \left(u^t W_l\right)_+,$$

where W_l is the i by j weight matrix for layer l. In this form, the multilayer perceptron produces a piecewise linear model of the input data. The computational complexity of training a multilayer perceptron is $O(ndm)$, where m is determined by the sizes of the layers of the network and the stopping criterion of the BFGS minimization used for finding weights. This paper uses the scikit-learn MLP regressor [38], a single hidden layer with 100 nodes, ReLU activation, and BFGS for training.

1.2.2 *Decision Tree.*

The decision tree is used because of the relatively straightforward interpretation of the prediction process. Model construction is out of the scope of this description, but is a well-studied process [40]. A prediction at a point $z \in \mathbb{R}^d$ for a decision tree constructed over a real vector space is made by traversing nested axis-aligned conditionals of the form

$$\hat{f}(z) = \hat{f}(z | z_{k^{(1)}} \ge v^{(1)}, \dots)$$

This paper uses the scikit-learn Decision Tree regressor [38], no maximum depth or number of nodes, and the Gini impurity measure of information gain.

1.3 Local Predictive Models

The construction of approximation functions $\hat{f}_P : \mathbb{R}^d \to \mathbb{R}$ as described for global models can instead be approached on a per-prediction basis. A model is henceforth referred to as *local* when any prediction made at a point $z \in \mathbb{R}^d$ is only a function of a set of points $L \subset X$, where membership in L is determined by a distance metric. The advantage of using a local model is a more compact description of *how* a prediction is made that is derived from a manageable subset of all known data. *Local* models become particularly useful when predictions regard ethically sensitive issues and need to be rigorously evaluated for bias. The source data for any prediction can be checked on the spot for fairness of representation against any number of protected attributes.

The following sections describe the three local approximation techniques that will be used to predict recidivism likelihood in this work.

1.3.1 *Nearest Neighbor.*

This algorithm will be used as a baseline for comparison because it is the most mathematically simple *local* model in this study. A prediction is made for Nearest Neighbor at point $z \in \mathbb{R}^d$ by

$$\hat{f}(z) = f\left(argmin_{x^{(i)} \in X} \|z - x^{(i)}\|_2\right).$$

This approximation technique can be applied in a wide range of applications, however it must be noted that the approximation surface it produces is not C^0 (continuous in value).

1.3.2 *Delaunay Triangulation.*

The Delaunay method of interpolation is a well studied geometric technique for producing an interpolant [33]. The Delaunay triangulation of a set of data points into simplices is such that the sphere defined by the vertices of each simplex contains no data points in the sphere's interior. For a d-simplex S with vertices $v^{(0)}$, $v^{(1)}$, …, $v^{(d)}$, $x \in S$, and data values $f(v^{(i)})$, $i = 0, \dots, d$, x is a unique convex combination of the vertices,

$$x = \sum_{i=0}^{d} w_i v^{(i)}, \quad \sum_{i=0}^{d} w_i = 1, \quad w_i \ge 0, \quad i = 0, \dots, d,$$

and the Delaunay interpolant to f at x is given by

$$p(x) = \sum_{i=0}^{d} w_i f(v^{(i)}).$$

The computational complexity of the Delaunay triangulation (for the implementation used here) is $OO(n^{1+\frac{1}{d}} d^3 + nd^4)$ per prediction, which should reasonably scale to $d < 100$. A newly released Fortran implementation of this polynomial-time Delaunay interpolation technique is used here [7].

1.3.3 Voronoi Mesh. The final of the three meshes utilizes 2-norm distances to define boundaries rather than max norm distances. A well-studied technique for classification and approximation is the nearest neighbor algorithm [9]. Nearest neighbor inherently utilizes the convex region $v^{x^{(i)}}$ (Voronoi cell [12]) consisting of all points closer to $x^{(i)}$ than any other point $x^{(j)}$. The Voronoi mesh smooths the nearest neighbor approximation by utilizing the Voronoi cells to define support via a generic basis function $V : \mathbb{R}^d \to \mathbb{R}_+$ given by

$$V^{x^{(i)}}(y) = \left(1 - \frac{\left\|y - x^{(i)}\right\|_2}{2\, d(y \mid x^{(i)})}\right)_+ ,$$

where $x^{(i)}$ is the center of the Voronoi cell, $y \in \mathbb{R}^d$ is an interpolation point, and $d(y \mid x^{(i)})$ is the distance between $x^{(i)}$ and the boundary of the Voronoi cell $v^{x^{(i)}}$ in the direction $y - x^{(i)}$. $V^{x^{(i)}}\big(x^{(j)}\big) = \delta_{ij}$ and $V^{x^{(i)}}$ has local support. While $V^{x^{(i)}}(x^{(i)}) = 1$, the 2 in the denominator causes all basis functions to go linearly to 0 at the boundary of the twice-expanded Voronoi cell. Note that this basis function is C^0 because the boundaries of the Voronoi cell are C^0. In the case that there is no boundary along the vector w, the basis function value is always 1.

While the cost of computing the exact Voronoi cells for any given set of points grows exponentially [13], the calculation of d is linear with respect to the number of control points and dimensions. Given any center $x^{(i)} \in \mathbb{R}^d$, set of control points $C \subseteq X$, and interpolation point $y \in \mathbb{R}^d$, $d\big(y \mid x^{(i)}\big)$ is the solution to

$$\max_{c \in C \setminus \{x^{(i)}\}} \frac{\left\|y - x^{(i)}\right\|_2}{2} \; \frac{y \cdot \big(c - x^{(i)}\big) - x^{(i)} \cdot \big(c - x^{(i)}\big)}{c \cdot \big(c - x^{(i)}\big) - x^{(i)} \cdot \big(c - x^{(i)}\big)}.$$

2 RELATED WORK (SN)

Algorithmic bias has long been a subject of research on machine learning [2, 16, 30]. Mitchell [35] initially defined machine learning bias as “any basis for choosing one generalization over another, other than strict consistency with the instances”. Mooney [37] expanded this definition with the following assertions: (1) that every model bears some inherent bias, and (2) that detecting a model’s bias requires comparison against others. In recent years, the broad adoption of machine learning has made algorithmic bias a factor in real-world data discrimation [1, 45, 46].

Several prior works have explored the problem of measuring algorithmic bias [25, 42, 56]. Calders & Verwer [5] formalized discrimination as, given an input characteristic, the unequal distribution of outputs for different groups. Calders-Verwer (CV) scoring is frequently [4, 14, 21, 22, 24, 26, 27, 57] used to measure *group discrimation* [53, 54]. For example, if a loan classifier produces dissimilar outcomes for both sexes (e.g. $P(Y = loan|S = male) > P(Y = loan|S = female)$), then its CV score is measured as the difference in outcomes between those groups (e.g. $P(Y = loan|S = male) - P(Y = loan|S = female)$. Many works have since expanded on CV scoring for preventing discrimination [14, 17, 22, 23, 52, 57]. Four general approaches exist: (1) *Suppression* – removing attributes most correlating with discrimination-sensitive attributes; (2) *Dataset “massaging”* – altering labels of some objects to mitigate unwanted classifier outcomes; (3) *Reweighting* – assigning data carefully chosen weights to lessen the degree of discrimination; (4) *Sampling* – under- and over-sampling certain groups to compensate.

A recent focus of machine learning and data science research has been on improving model interpretability [20, 34, 55]. More specific goals include transparency for assessing model ethicacy [15, 18], augmenting informativeness [29, 50, 51], and inferring causality [3, 47, 49]. An obstacle to model interpretability is the lack of transparency of black-box classifiers such as deep neural networks [6, 8, 32]. Post-hoc interpretability [34] (e.g. visualizations or explanations) represents a promising alternative, however, formalizing “model-agnostic” methodologies remains an ongoing research problem [41].

3 DATA (SY)

We are using the 3-Year Recidivism for Offenders Released from Prison dataset, publicly available on the webpage of U.S. GovernmentâĂŹs open data (data.gov). This dataset reports whether an offender is re-admitted to prison or not within three years from being released from prison. With this dataset, we could either perform a classification task (predict whether or not a released offender would recidivate), or a regression task (predict the probability of recidivism), for the latter, we collapse all identical instances to one and approximate the probability of recidivism with the frequency of positive instances.

We first computed simple statistics on the dataset so that we are aware of the imbalance and discrepancies in the dataset that could lead to bias. This dataset provides 21, 646 instances, Among which 14, 619 instances are negative and 7, 027 instances are positive. We also computed the population across ethinicity groups and compared it against the actual population breakdown by race in Iowa in 2016 (See figure 2), we noticed that there exist obvious gaps especially for White non-hispanic (dataset: 67.4%, actual population: 88.7%) and Black or African American (Dataset: 23.6%, actual population: 2.9%).

Next we preprocessed the features. Each instance in the dataset is described with 16 features and a label indicating whether this prisoner recidivates. First of all, 6 of these features are used to describe the new crime if recidivisms occurs and are therefore unavailable for the negative instances, so we exclude these features; Second, we found that two features can be easily computed or inferred from other predictors, specifically, "Recidivism Reporting Year" is always three years later than "Fiscal Year Released", and "Convicting Offense Type" can be uniquely identified with "Convicting Offense Subtype", so we removed these two features. Third, "Race - Ethnicity" is reserved as protected attribute and is later used

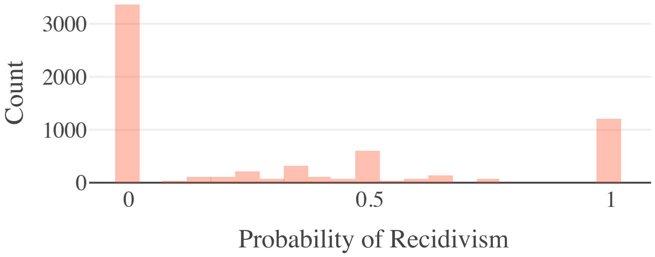


Figure 1: A 20-bin histogram of the values for recidivism likelihood in the regression task composed of 6,632 samples.

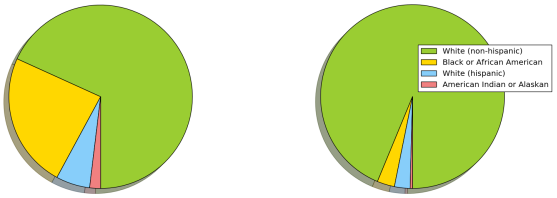


Figure 2: The distribution across ethnicity groups of the dataset (left) and the actual population in Iowa (right)

to evaluate biases in predictions. Therefore, we have 7 features left and we removed the instances that have empty entries for the selected features. Specifically, the 7 features are:

- Fiscal Year Released
- Sex
- Age At Release
- Convicting Offense Classification
- Convicting Offense Subtype
- Release Type
- Main Supervising District

Except "Fiscal Year Released", all attributes are categorical. We show some sample values of these categorical features in table 1. We discard any categories with less than 100 samples in order to reduce the resulting dimension and prevent predictions from being made with statistically insignificant amounts of supporting data. We encode categorical features with c unique categories into $\mathbb{R}^{c-1}$ by mapping each category onto one of the vertices of a regular simplex centered at 0 where all vertices v satisfy $\|v\|_2 = 1$. We also consider representing "Age At Release" categories using their average ages, to reduce the number of dimensions in input space.

Finally, in our experiments, predictions are evaluated against the labels, race & ethnicity information are used to evaluate model bias. With all the processing, we have 21,018 instances in a 50 dimensional space for classification; for regression, since all identical instances are represented with one, the number of instances drop to 6,632 and the distribution of the response values can be seen in figure 1.

4 RESULTS (TL)

NearestNeighbor

	GN	GR
TN	47.6%	19.8%
TR	20.1%	12.5%

Accuracy: 60.1%

20.6%	White - NH
19.8%	Black - NH
17.2%	White - H
16.3%	AI or NA
20.7%	Asian or PI

DecisionTreeClassifier

	GN	GR
TN	55.8%	11.6%
TR	23.6%	9.0%

Accuracy: 64.8%

11.2%	White - NH
12.1%	Black - NH
8.7%	White - H
7.9%	AI or NA
12.7%	Asian or PI

MLPClassifier

	GN	GR
TN	59.7%	7.8%
TR	24.5%	8.1%

Accuracy: 67.7%

7.9%	White - NH
8.0%	Black - NH
5.2%	White - H
3.6%	AI or NA
9.3%	Asian or PI

Figure 3: Performance of three of the algorithms on the classification problem. Left column represents the confusion matrices with rows "true no recidivism" (TN), "true recidivism" (TR), and columns "guessed no recidivism" (GN), "guessed recidivism" (GR). The right column represents the false positive rate for predicting recidivism broken down by race. The race category abbreviations are: not Hispanic (NH), Hispanic (H), American Indian (AI), Native Alaskan (NA), and Pacific Islander (PI).

The nearest neighbor (baseline), neural network, and decision tree are applied to both the classification and regression tasks. The Delaunay and Voronoi mesh interpolants are only applied to the regression task. First we analyze the classification results on the raw data. In order to best estimate the real-world performance of these algorithms, *k-fold* cross validation as described in [31] with $k = 10$ is used. All algorithms are given the same ten folds of randomized training and testing data in order to maintain comparative fairness. Note that in this scheme there will be exactly one prediction made for each data point, meaning all analysis of results is done with the same sized data as described in Section 3.

In figure 3 we see the initial results on the classification task of predicting whether or not an admitted prisoner will recidivate upon release. The overall top performer by accuracy is the multilayer perceptron, but it also has the highest false negative prediction rate and the largest discrepancy in false positives by race ("positive" refers to recidivism). Generally it would be expected that any racial crime correlations present would be largest for the race which makes up the largest percentage of data. Interestingly that trend is not observed with the racial outcomes. Larger false positive rates were attributed occasionally to the least represented races (AI or

FEATURE NAME	VALUES
Sex	F, M
Age At Release	Under 25, 25-34, 35-44, 45-54, 55 and Older
Convicting Offense Classification	Aggravated Misdemeanor, Serious Misdemeanor, Sexual Predator Community Supervision, etc
Convicting Offense Subtype	Murder, Alcohol, Weapons, Drug Possession, Assault, Traffic, Burglary, Forgery/Fraud, Animals, Theft, etc
Release Type	Special Sentence, Parole Granted, Discharged - Expiration of Sentence, Released to Special Sentence, etc
Main Supervising District	1JD, 2JD, 3JD, 4JD, 5JD, etc

Table 1: Sample values of categorical features selected

NA, Asian or PI) and also the most represented race (White - NH). It should be noted that the *explanation* for predictions made by the decision tree and multilayer perceptron must be provided in the context of all data. For ethical applications, it may be concerning that the racial representation of the prison data does not match the distribution of the population of the state of Iowa (presuming there is the potential for racial discrimination in historical criminal sentencing).

Now we consider the outputs of the global and local algorithms on the regression task of predicting recidivism likelihood. figure 4 displays the outcomes for four of the five algorithms. The column arrangement of the models is intentional, noting that nearest neighbor and regression tree perform similarly while the Voronoi mesh and neural network also perform similarly. The prediction outcomes are promising, demonstrating that 50% of recidivism likelihood predictions have less than a .15 absolute error. The largest discrepancy in false positive rate by race is observed to be the neural network (same as for classification).

The most notable result that can be observed in figure 4 is that the Voronoi Mesh (VM) algorithm that makes predictions based only on a local support of roughly $2d$ (100) points from data competes with the global fitting Multilayer Perceptron based on all data (20K) points. The benefit of using the VM to make predictions in ethically sensitive applications is that *every* prediction has a manageable set of source data that can be used to describe how a prediction is produced. Addressing the two points mentioned in Section 1, statistical tests can be run on these source data points to reduce prediction bias in desirable ways. Along with simplified statistical testing, the opportunity can be taken during any prediction to validate the source data and remove entries that are evidently outliers. Finally, this type of prediction opens up the possibility of a legal right to a representative sample from any predictive models used in official legal proceedings (such as predicting recidivism).

Results for Delaunay were the most theoretically promising, because of the guaranteed compactness of support (exactly $d+1$ source points for any prediction). Unfortunately, after two weeks of computation the Delaunay code was not able to compute predictions for the data. This suggests either a bug in the recently released Fortran code, or an unexpected geometric degeneracy in the data being used to make predictions. Regardless of the cause of slowdown, Delaunay results could not be collected for this experimentation and that is left to future attempts.

5 DISCUSSION (RB)

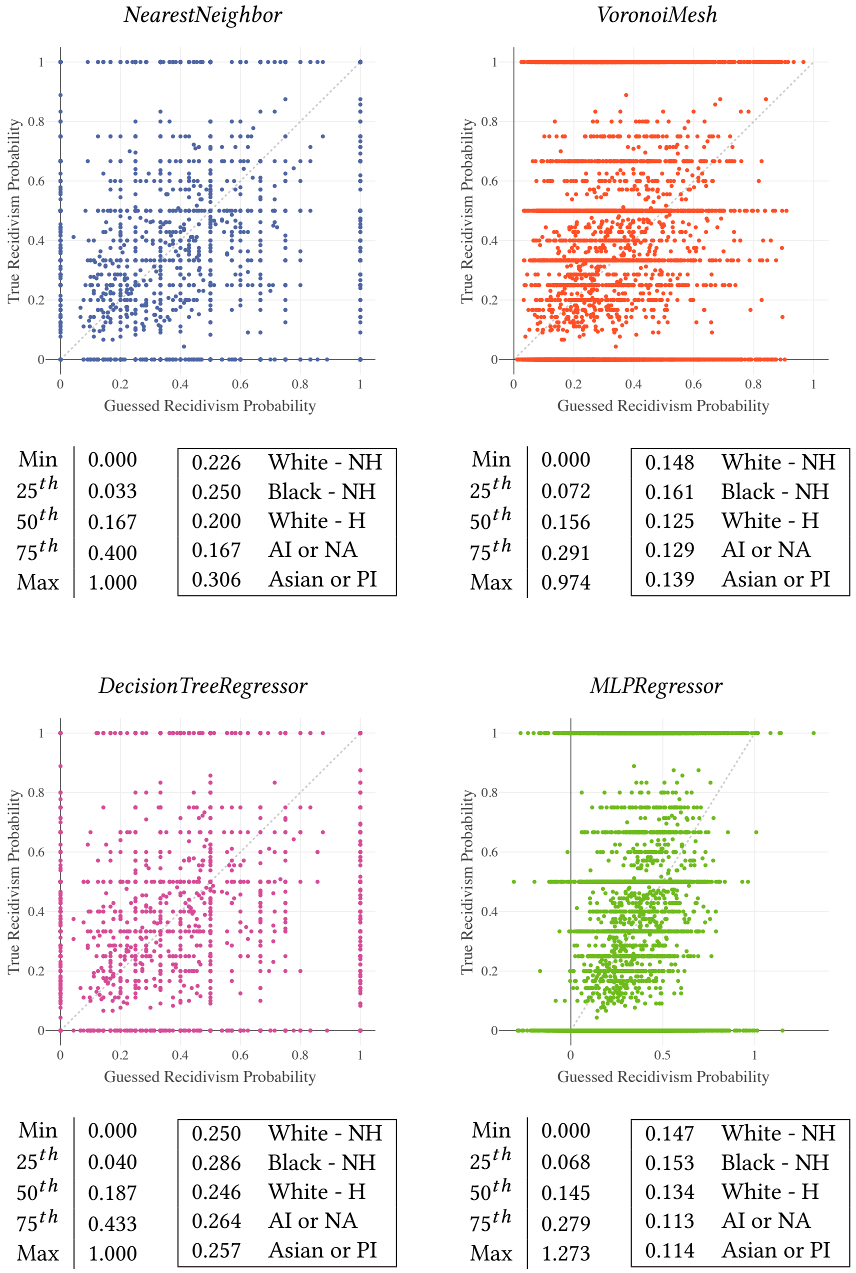
For our experiments, we ran both condensed regression and full classification on all 3 of our models, with variation in the definition of Age either as a category or as a number. From these results, it is noted that interpretting age as a category or as a number will produce negligible differences in the results on the models. The condensed regression results were fairly similar between all 3 classifiers, with the largest number of guesses on recidivism being only 3% more than the minimum. However, we ended up creating a 4th model for the Voronoi Mesh for condensed regression which happened to compete well with the Multilayered Perceptron Regressor. Based off of this, we have instead decided to focus on all the models of full classification, but only focus on the Voronoi Mesh and Multilayered Perceptron Regressors in condensed regression.

5.1 Full Classification

As can be seen from our results in figure 3, all of our models are at least 60% correct in classifying the correct recidivism of individuals, with the Multilayered Perceptron resulting in the best classification rate at approximately 67%. Notably, this model guessed nearly **half** as many recidivism candidates as the Nearest Neighbor algorithm, yet still achieved a better overall accuracy. This can likely be attributed to the fact that the *true* recidivism for this dataset was actually only 33%, meaning that guessing non-recidivism will be more likely to be correct.

In terms of interpretability, it's possible to see that for all 3 models that, regardless of the true accuracy, all of the models had a relatively even amount of bias when generating a false positive. One must also note that the Asian or Pacific Islander category was also the smallest population size by far, meaning that having the highest percentage of bias from false positives is expected for all 3 models. From these results, the Nearest Neighbor, Decision Tree, and Mulitlayered Perceptron have a false positive standard deviation of 2.03, 2.11, 2.32 from their mean respectively. This is well within a tolerable bias range considering that the more guesses on recidivism will result in a higher false positive rate overall. With this in consideration, the Nearest Neighbor, Decision Tree, and Multilayered Perceptron models ended up guessing recidivism 19.8%, 11.6%, and 7.8% of the time, which follows our assumptions on the biases.

We can conclude from our initial set of tests that all of our models have been relatively debiased while also having a reasonable accuracy when guessing recidivism or non-recidivism. These findings aside, the best model that results in the least number of false



NearestNeighbor

Min	0.000
25th	0.033
50th	0.167
75th	0.400
Max	1.000

0.226	White - NH
0.250	Black - NH
0.200	White - H
0.167	AI or NA
0.306	Asian or PI

VoronoiMesh

Min	0.000
25th	0.072
50th	0.156
75th	0.291
Max	0.974

0.148	White - NH
0.161	Black - NH
0.125	White - H
0.129	AI or NA
0.139	Asian or PI

DecisionTreeRegressor

Min	0.000
25th	0.040
50th	0.187
75th	0.433
Max	1.000

0.250	White - NH
0.286	Black - NH
0.246	White - H
0.264	AI or NA
0.257	Asian or PI

MLPRegressor

Min	0.000
25th	0.068
50th	0.145
75th	0.279
Max	1.273

0.147	White - NH
0.153	Black - NH
0.134	White - H
0.113	AI or NA
0.114	Asian or PI

Figure 4: These four plots show the true recidivism probability versus guessed recidivism probability for each of the regression techniques (except for Delaunay, explained in Section 4). The top two regression algorithms make predictions based only on local data while the bottom two algorithms are global fitting techniques. The left vertical table beneath each figure displays the percentiles of absolute errors when predicting the probability of recidivism with that algorithm. The right table beneath each figure shows the median error in recidivism likelihood for those predictions which were over-estimated (false positives) broken down by race. The race abbreviations are the same as in figure 3.

positives as well as the highest accuracy is our Multilayered Perceptron model with a 67.7% accuracy and average false positive rate of 6.8%. Interestingly, this also coincides with the trend that less guesses on recidivism tended to result in better accuracy, with the Multilayerd Perceptron guessing recidivism half as much as the Nearest Neighbors Classifier.

5.2 Condensed Regression

In this section we are just focusing on the Voronoi Mesh and Multilayer Perceptron Regressors' results, since the other 2 models didn't produce enough positive results. From the results seen in figure 4, the Voronoi Mesh has been able to produce results that compete quite well with that of the Multilayered Perceptron. This shows us that the local predictive models are able to produce just as good results as the global predictive models. In the case of the Nearest Neighbor and Decision Tree Regressors, not enough positive results in accuracy nor false positives meant that those models would not be as useful in classification of bias.

One odd thing of note is that the Multilayered Perceptron Regressor tended to produce weird prediction by nature (which were outside of the usual range [0,1], which in it of itself is undesirable in our results. This is to be expected with this type of Regressor, but our hope would be that such things would not happen in most datasets as it could produce more exact results. Regardless of these oddities, the Multilayered Perceptron Regressor tended to produce quite minimal median error in recidivism likelihood for those predictions that were false positives (at least compared to that of the 2 less than satisfactory models). The Voronoi Mesh has similar results in its minimal median error as well as doesn't produce odd predictions outside of the usual range.

5.3 Takeaways

From these results, we can assume that it should be relatively possible to use these classifiers to ensure data is clean and relatively debiased. A very positive result shows that both Global Predictive **and** Local Predictive Models can be beneficial in locating bias within datasets and can guess well on recidivism rates without much bias. In the case when the false positive standard deviation is much larger than 2, we can assume then that there is a bias towards one or more races. We can therefore utilize these algorithms to ensure that a given dataset has not been ethically comprimised due to unintentional (or perhaps even intentional) biases. Similarly, we could use these classifiers as a means to figure out which features in a dataset are the cause of significant bias in the results by generating false positive rates for different combinations of features. This is on track with our expected results in producing an improvement to identify as well as clean bias either statistically or by human users on a per-prediction bias.

6 CONCLUSION (ALL MEMBERS)

This paper presents a comparative analysis of popular global predictive modeling algorithms and less popular local prediction techniques. Results demonstrate that algorithms which rely only on local support are capable of producing predictions of comparable (and sometimes superior) accuracy to those popular techniques while also maintaining an enhanced level of interpretability. The potential for operating under more concise legal definitions and meaningful statistical analyses further supports the implementation of explainable prediction methodologies. This recidivism case study demonstrates that the use of more explainable models could not only strengthen the ethical application of data science, but also make the process of cleaning and validating data manageable in the long term.

7 ACKNOWLEDGEMENTS & CONTRIBUTIONS

The data for this project was collected from the `data.gov` project called "3-year-recidivism-for-offenders-released-from-prison" provided by the State of Iowa.

All of the code for this project is available at `https://github.com/tchlux/GRAD_5134_Project`

Component	Primary Contributer
Abstract and Conclusion	All Members
Introduction	Mohammed Almanaa
Related Work	Stefan Nagy
Data	Sirui Yao
Results + Code	Thomas Lux
Discussion	Reid Bixler

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. [n. d.]. Machine Bias: there‘s software used across the country to predict future criminals. And it‘s biased against blacks. ProPublica 2016.

[2] Ricardo Baeza-Yates. 2016. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*. ACM, 1–1.

[3] Lionel C Briand, VR Brasili, and Christopher J Hetmanski. 1993. Developing interpretable models with optimized set reduction for identifying high-risk software components. *IEEE Transactions on Software Engineering* 19, 11 (1993), 1028–1044.

[4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*. IEEE, 13–18.

[5] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (01 Sep 2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[6] Jorge Casillas, Oscar Cordón, Francisco Herrera Triguero, and Luis Magdalena. 2013. *Interpretability issues in fuzzy modeling*. Vol. 128. Springer.

[7] Tyler H Chang, Layne T Watson, Thomas CH Lux, Bo Li, Li Xu, Ali R Butt, Kirk W Cameron, and Yili Hong. 2018. A polynomial time algorithm for multivariate interpolation in arbitrary dimension via the Delaunay triangulation. In *Proceedings of the ACMSE 2018 Conference*. ACM, 12.

[8] Paulo Cortez and Mark J Embrechts. 2013. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences* 225 (2013), 1–17.

[9] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27.

[10] George E. Dahl, Tara N. Sainath, and Geoffrey E. Hinton. 2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*. IEEE, 8609–8613.

[11] Simon DeDeo. 2014. Wrong side of the tracks: Big Data and Protected Categories. *arXiv preprint arXiv:1412.4643* (2014).

[12] G Lejeune Dirichlet. 1850. Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. *Journal für die Reine und Angewandte Mathematik* 40 (1850), 209–227.

[13] Mathieu Dutour Sikirić, Achill Schürmann, and Frank Vallentin. 2009. Complexity and algorithms for computing Voronoi cells of lattices. *Math. Comp.* 78, 267 (2009), 1713–1731.

[14] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting*

on Foundations of Software Engineering. ACM, 498–510.

[15] Bryce Goodman and SR Flaxman. 2017. European Union regulations on algorithmic decision-making and a. (2017).

[16] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[17] Sara Hajian, Josep Domingo-Ferrer, and Antoni Martinez-Balleste. 2011. Discrimination prevention in data mining for intrusion and crime detection. In *Computational Intelligence in Cyber Security (CICS), 2011 IEEE Symposium on*. IEEE, 47–54.

[18] Bernease Herman, Gundula Proksch, Rachel Berney, Hillary Dawkins, Jacob Kovacs, Yahui Ma, Jacob Rich, and Amanda Tan. 2017. Data science for urban equity: Making gentrification an accessible topic for data scientists, policymakers, and the community. *arXiv preprint arXiv:1710.02447* (2017).

[19] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks* 2, 5 (1989), 359–366.

[20] Chia-Feng Juang and Chi-You Chen. 2013. Data-driven interval type-2 neural fuzzy system with high learning accuracy and improved model interpretability. *IEEE transactions on cybernetics* 43, 6 (2013), 1781–1795.

[21] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*. IEEE, 1–6.

[22] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 869–874.

[23] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2013. Techniques for discrimination-free predictive models. In *Discrimination and privacy in the information society*. Springer, 223–239.

[24] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 924–929.

[25] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems* 35, 3 (2013), 613–644.

[26] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[27] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 643–650.

[28] Nancy E Kass. 2001. An ethics framework for public health. *American journal of public health* 91, 11 (2001), 1776–1782.

[29] Been Kim. 2015. *Interactive and interpretable machine learning models for human machine collaboration*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[30] Keith Kirkpatrick. 2016. Battling algorithmic bias: how do we ensure algorithms treat us fairly? *Commun. ACM* 59, 10 (2016), 16–17.

[31] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Montreal, Canada, 1137–1145.

[32] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5686–5697.

[33] Der-Tsai Lee and Bruce J Schachter. 1980. Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer & Information Sciences* 9, 3 (1980), 219–242.

[34] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[35] Tom M Mitchell. 1980. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ. New Jersey.

[36] Martin Fodslette Møller. 1993. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks* 6, 4 (1993), 525–533.

[37] Raymond J Mooney. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. *arXiv preprint cmp-lg/9612001* (1996).

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[39] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 560–568.

[40] J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning* 1, 1 (1986), 81–106.

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. ACM, 1135–1144.

[42] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

[43] Eric Siegel. 2016. *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons Incorporated.

[44] Megan Smith, D Patil, and Cecilia Muñoz. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. *White House Report, Executive Office of the President* (2016).

[45] Gregory D Squires. 2003. Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs* 25, 4 (2003), 391–410.

[46] Michael A Stoll, Steven Raphael, and Harry J Holzer. 2004. Black job applicants and the hiring officer's race. *ILR Review* 57, 2 (2004), 267–287.

[47] Umar Syed and Golan Yona. 2009. Enzyme function prediction with interpretable models. In *Computational Systems Biology*. Springer, 373–420.

[48] Effy Vayena, Urs Gasser, Alexandra B Wood, David O'Brien, and Micah Altman. 2016. Elements of a new ethical framework for big data research. (2016).

[49] Hui-Xin Wang, Laura Fratiglioni, Giovanni B Frisoni, Matti Viitanen, and Bengt Winblad. 1999. Smoking and the occurence of Alzheimer's disease: Cross-sectional and longitudinal data in a population-based study. *American journal of epidemiology* 149, 7 (1999), 640–644.

[50] Tianfu Wu, Xilai Li, Xi Song, Wei Sun, Liang Dong, and Bo Li. 2017. Interpretable R-CNN. *arXiv preprint arXiv:1711.05226* (2017).

[51] Liping Yang, Alan M MacEachren, Prasenjit Mitra, and Teresa Onorati. 2018. Visually-Enabled Active Deep Learning for (Geo) Text and Image Classification: A Review. *ISPRS International Journal of Geo-Information* 7, 2 (2018), 65.

[52] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259* (2017).

[53] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2015. Learning fair classifiers. *arXiv preprint arXiv:1507.05259* (2015).

[54] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[55] Shang-Ming Zhou and John Q Gan. 2008. Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling. *Fuzzy Sets and Systems* 159, 23 (2008), 3091–3131.

[56] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (2015).

[57] Indre Žliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling conditional discrimination. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 992–1001.