Quantitative Input Influence

A study on influence of inputs on automated recidivism decisions

Karina Huang, Lipika Ramaswamy
AC221: Critical Thinking in Data Science
Spring 2019

For further details, see our Github (https://github.com/kareenaaahuang/Algorithm-Transparency).

Out[6]: Click here to toggle on/off the raw code.

I. Introduction

Responsible data science is necessary because of its impact on decision making. Consider the fairness of automated decision making in hiring and sentencing procedures. While fairness of algorithms remains debatable depending on how the term is defined, lack of interpretability in algorithms complicates the debate; if it remains unclear what contributes to a decision, it will be difficult to compartmentalize what makes the decision fair or unfair. Therefore, it is necessary that an algorithm be presented as clearly as possible. Granting algorithmic transparency helps facilitate discussions on the fairness of model design and consequences.

One approach to quantifying the interpretability of a model is the Quantitative Input Influence (QII) framework introduced by Datta et al. in their paper, *Algorithmic Transparency via Quantitative Input Influence*. These are causal measures that explain the influence of an input or a set of inputs on the decision made by machine learning algorithms. Transparency queries that use QII measures can be used to explain decisions about the classification outcome for an individual or a group. Specifically, for a quantity of infuence, Q, and an input feature, i, the QII of i on Q is the difference in Q when i is changed via intervention. In other words, we can replace features with random values from the population, and examine the distribution over outcomes for changes.

In this study, we examined the QII framework using the COMPAS dataset (https://github.com/propublica/compas-analysis). ProPublica has previously published a study (https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm) presenting the scoring system as biased against African American individuals. Our objectives were to evaluate the validity of such claims by:

- examining the average importance of individual attributes on individual outcomes in modeling the COMPAS decile score categories (low, medium, high)
- comparing the QII outcomes on group disparity, with respect to the probability of being predicted in a decile score category
- investigating marginal influence of attributes on individuals in the dataset through transparency reports

II. Data

ProPublica included 3 datasets on their Github page. Due to a lack of documentation regarding how the data was cleaned (we found discrepancies in the their reported methods and the actual dataset), we examined each of the dataset and chose to use the dataset in `compas-scores-two-years.csv` . Below we report the preliminary characteristics of the dataset.

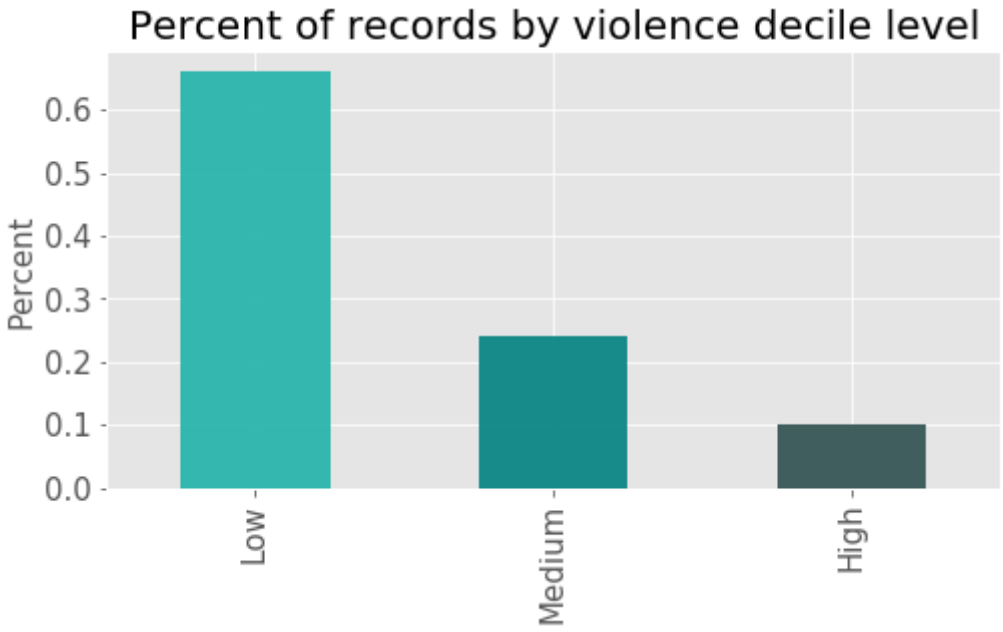
```
The size of this dataset is (7214, 15) ; Note that the number of unique
individuals is  7214
Here is a snapshot of the first five rows of the dataset:
```

Out[7]:

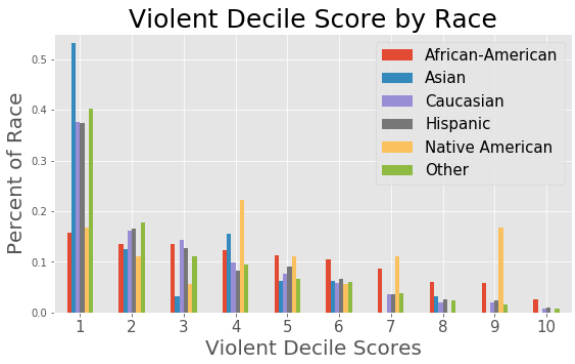
	id	sex	age	race	juv_fel_count	juv_misd_count	juv_other_count	priors_count	decile_s
0	1	Male	69	Other	0	0	0	0	
1	3	Male	34	African-American	0	0	0	0	
2	4	Male	24	African-American	0	0	1	4	
3	5	Male	23	African-American	0	1	0	1	
4	6	Male	43	Other	0	0	0	2	

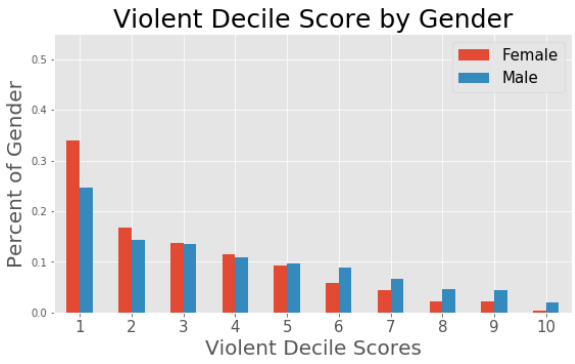
For each individual, there are decile scores and categories. We choose to explore the violent scores and categories. Note that scores of 1 to 4 correspond to 'Low', 5 to 7 correspond to 'Medium' and 8 to 10 correspond to 'High'.

We can see that the majority of records in the dataset correspond to a COMPAS score in the 'Low' category (66%), followed by 24% in the 'Medium' category, and 10% in the 'High' category.



The visualizations below reveal the distribution of decile scores by race and sex. For all African-Americans in the dataset, the percent assigned to each COMPAS score is more uniform than is the case for most other groups. Of note is also the Native American group, of which a large percentage is assigned to medium to high violent decile scores. In comparison, roughly 40% of those in the Asian, Caucasian, Hispanic and Other groups are assigned to the lowest violent decile score.
Similarly for gender, more females are assigned lower violent decile scores than males.





III. Methodology

Models

Our modelling goal was to reconstruct COMPAS assignments of individuals to violence decile scores given that it is a blackbox algorithm. Using ProPublica's data, we model the violence decile categorization (low, medium, high) of an individual using their other attributes, including age, juvenile misdemeanor count, priors count, sex and race. Note that we use the predictors that are very similar to the ones specified by Northpointe in the COMPAS user guide (http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf). We used violence decile categorization instead of numeric scores visualized above to closely mirror ProPublica's approach. Given the classification task at hand, we chose to employ Logistic Regression, a Support Vector Machine and a Decision Tree for this reconstruction. We don't expect to have perfect reconstruction of the COMPAS algorithm, but aim to have a reasonable proxy.

Quantitative Input Influence

Given a black-box algorithm, A, we first define a quantity of interest, Q, which represents a property of the behavior of the algorithm for a given input distribution. A operates on inputs or features, i, and every i has a set of possible states it can take on, and $\mathbf{x} \in X$ is any vector representing a row of the dataset that is drawn from the true *underlying* distribution represented by the random variable X. In the particular problem explored here, an example of an algorithm is Logistic Regression and an example of a feature is `race`, which can take on the following states: African-American, Caucasian, Hispanic, Asian, Native American and Other.

Formally, for a quantity of interest $Q_A(\cdot)$ and an input i, the Quantitative Input Influence of i on $Q_A(\cdot)$ is defined as:

$$\iota^{Q_A} = Q_A(X) - Q_A(X_{-i}U_i),$$

where the random variable $X_{-i}U_i$ represents the random variable with input i replaced with a random sample, and represents the *intervened* distribution.

Unary QII

1. QII for individual outcomes:

   One use of QII is to provide individuals with information on a particular classification outcome. In order to quantify the use of an input for individual outcomes, we define the quantity of interest to be the classification outcome, c, for a particular individual, $\mathbf{x}$, given by $Q^{\mathbf{x}}_{ind}(\cdot) = \mathbb{E}(c(\cdot) = 1 | X = \mathbf{x})$. We can define the QII as:

   $$\iota^{\mathbf{x}}_{ind} = \mathbb{E}(c(X) = 1 | X = \mathbf{x}) - \mathbb{E}(c(X_{-i}U_i) = 1 | X = \mathbf{x})$$

   In our specific application, we are interested in the probability that a given feature is pivotal to the classification of an individual. We define this mathematically as:

   $$\sum_{\mathbf{x} \in X} Pr(X = \mathbf{x}) \cdot \mathbb{E}(i \text{ is pivotal for } c(X) | X = \mathbf{x}) = \mathbb{E}(i \text{ is pivotal for } c(X))$$

2. QII for group disparity:

   QII on group disparity denotes the association between classification outcomes and membership in a group. In this case, the quantity of interest is the absolute difference of postive classification outcomes between a given group and all others.

   $$Q^{\mathcal{Y}}_{disp}(\cdot) = |\mathbb{E}(c(\cdot) = 1 | X \in \mathcal{Y}) - \mathbb{E}(c(\cdot) = 1 | X \notin \mathcal{Y})|$$

   We can thus define the QII as:

   $$\iota^{\mathcal{Y}}_{disp}(i) = Q^{\mathcal{Y}}_{disp}(X) - Q^{\mathcal{Y}}_{disp}(X_{-i}U_i)$$

Set and Marginal QII

In the overwhelming majority of real datasets, it is impossible to look at the *individual* influence of one feature on an outcome, as it is likely correlated with other features that were included in the decision making process. Thus, if we intervene on only one feature, changes to the outcome will be less likely, as the features correlated with it will still capture some of the feature in question's impact. So if we intervene on sets of features including a given feature, we may get a better understanding of the influence of this feature by looking at its marginal influence.

For a quantity of interest, Q, and an input i, the QII of input i over a set $S \subseteq N$ on Q is defined to be

$$\iota^{Q}(i, S) = Q(X_{-S} U_S) - Q(X_{-S \cup \{i\}} U_{S \cup \{i\}})$$

It is clear from the equation that the marginal QII of i is the additional value in transparency achieved by including i in a subset of features.

The marginal contribution of i can vary depending on which set S is considered, so we report the aggregate marginal contribution of i to S, where S is sampled from some distribution over subsets of features. We use the Shapley value as a measure of this aggregate marginal contribution. This is what we present in transparency reports for an individual.

IV. Evaluation and Results

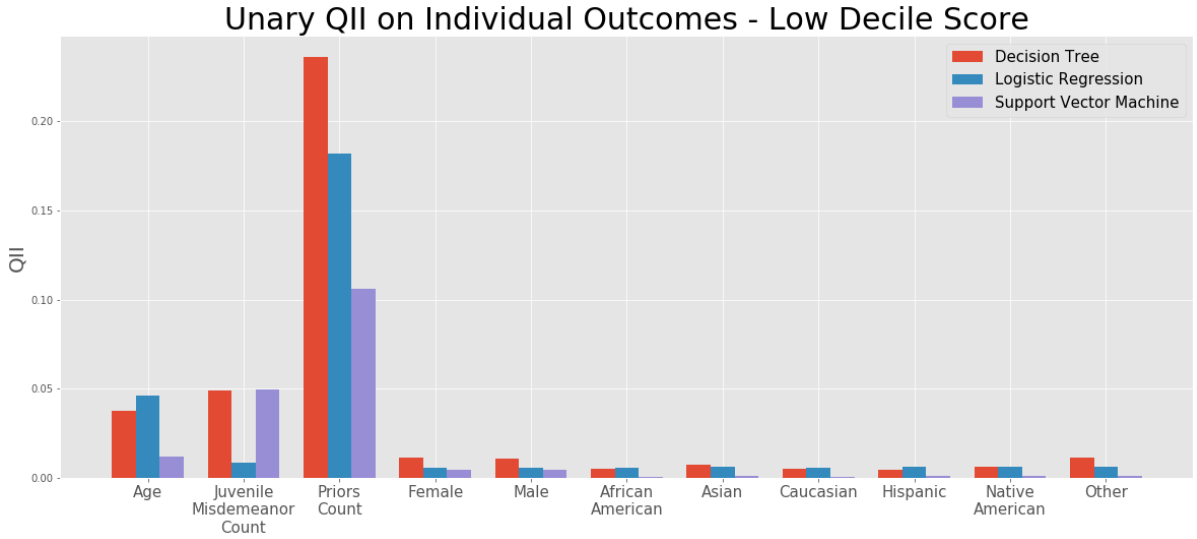
Classification Accuracy

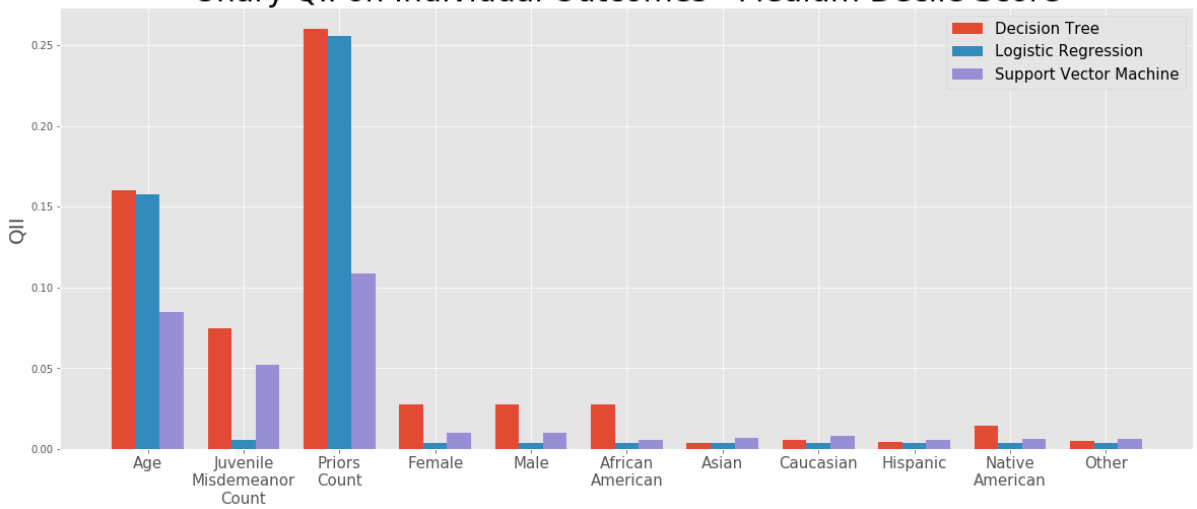
Overall the Support Vector Machine generated the best out-of-sample prediction accuracy and F1-score; a high f1 score indicates low false positives and false negatives.

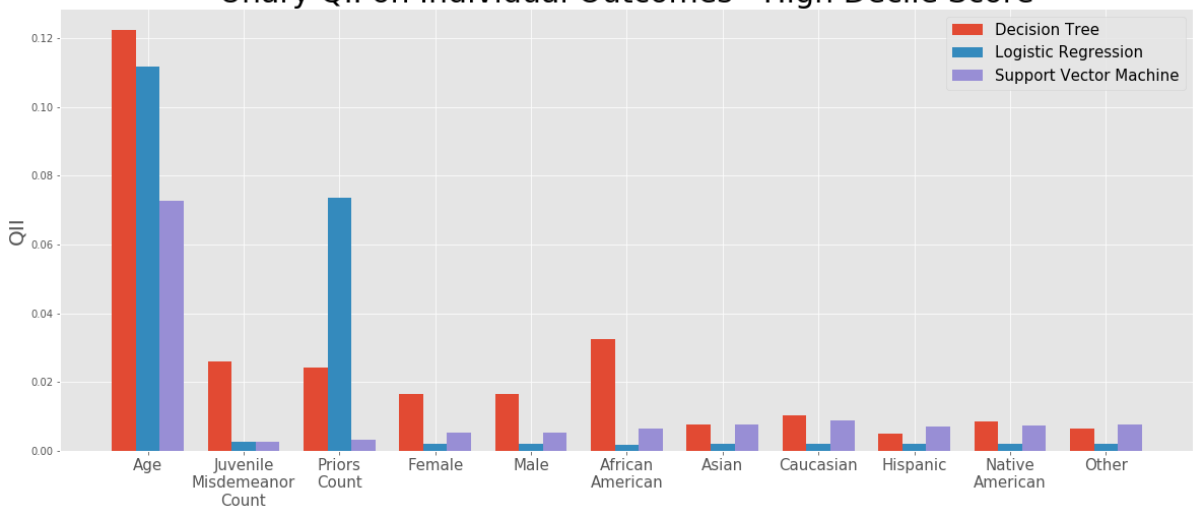
```
Decision Tree:
Test Prediction Accuracy:  0.7172557172557172
Test F1 Score:  0.7282290644902742
------------------------------------------------------------------------
--------------------------------------
Logistic Regression:
Test Prediction Accuracy:  0.7331947331947332
Test F1 Score:  0.7905109954134039
------------------------------------------------------------------------
--------------------------------------
Support Vector Machine:
Test Prediction Accuracy:  0.762993762993763
Test F1 Score:  0.7969124345782171
------------------------------------------------------------------------
--------------------------------------
```

Unary QII on Individual Outcomes

The classifiers rendered overall the same feature importance ranking. The plots below display the difference in feature importance by classifiers and decile score catogories. The y axis indicates the probability of being classified into the respective decile score category between using the actual data and some randomly permuted data for all individuals. The larger this QII value, the more the respective classifier and the decision relies on the given feature. For low and medium decile scores, *priors count* was the most significant factor of classification, followed by *age* and *juvenile misdemeanor count*. Interestingly, *age* was the most significant determinant in categorization of high violent decile scores. Note that the feature importances vary depending on the classifier, however, the magnitude is arguably minor.


Unary QII on Individual Outcomes - Low Decile Score
Decision Tree
Logistic Regression
Support Vector Machine
QII
0.00
0.05
0.10
0.15
0.20
Age
Juvenile Misdemeanor Count
Priors Count
Female
Male
African American
Asian
Caucasian
Hispanic
Native American
Other


Unary QII on Individual Outcomes - Medium Decile Score
Decision Tree
Logistic Regression
Support Vector Machine
QII
0.00
0.05
0.10
0.15
0.20
0.25
Age
Juvenile Misdemeanor Count
Priors Count
Female
Male
African American
Asian
Caucasian
Hispanic
Native American
Other


Unary QII on Individual Outcomes - High Decile Score
Decision Tree
Logistic Regression
Support Vector Machine
QII
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Age
Juvenile Misdemeanor Count
Priors Count
Female
Male
African American
Asian
Caucasian
Hispanic
Native American
Other

Unary QII on Group Disparity

There are two grouping factors in the current dataset: 1) sex, and 2) race. Our function allows for comparison between any given reference group to the rest. For demonstration purposes, we report below results of 1) group disparity between male and female, and 2) group disparty between Caucasian and other. As for unary QII on individual outcomes, the y-axis indicates the discrepancy between being classified a certain decile score group given the actual vs. random attributes. For group disparity, we held all other variables constant and randomly sampled the group assignment for each individual in the dataset. Therefore, a positive QII indicates a more consistent disparity in probability of being assigned a certain decile score group due to the randomized demographics information.

Male vs. Female

For the low decile score group, the plots indicate that there is some discrepancy in being assigned this decile score category between male and female groups. This discrepancy is attributed to age by all three classifiers. Interestingly, priors count did not seem to matter for the decision tree classifier.

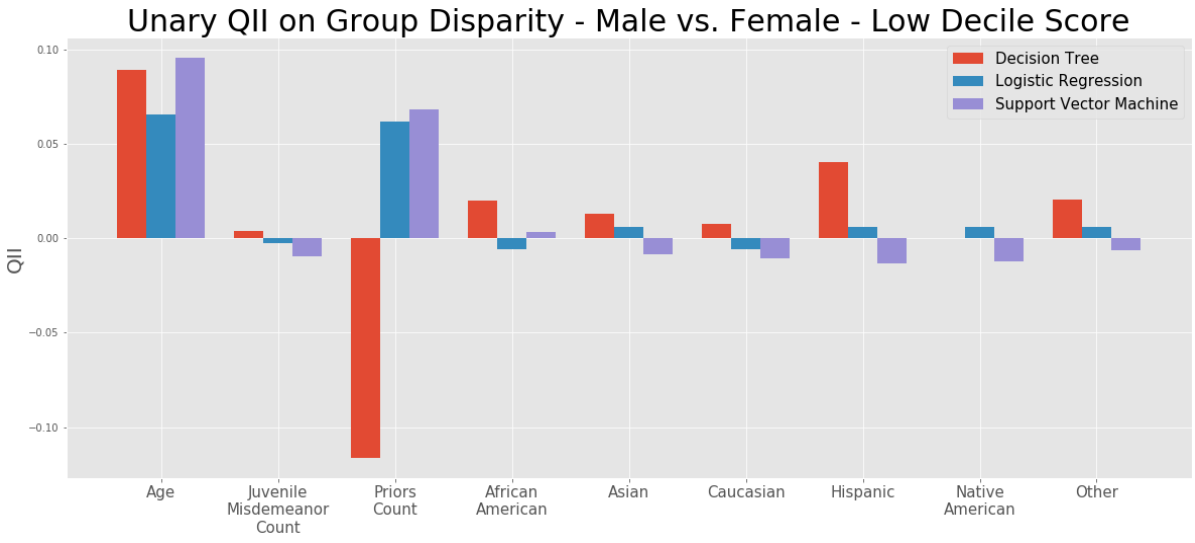
For the medium decile score group, the plots indicate that age, priors count, and juvenile misdemeanor counts contributed to the discrepancy of assignment due to gender.

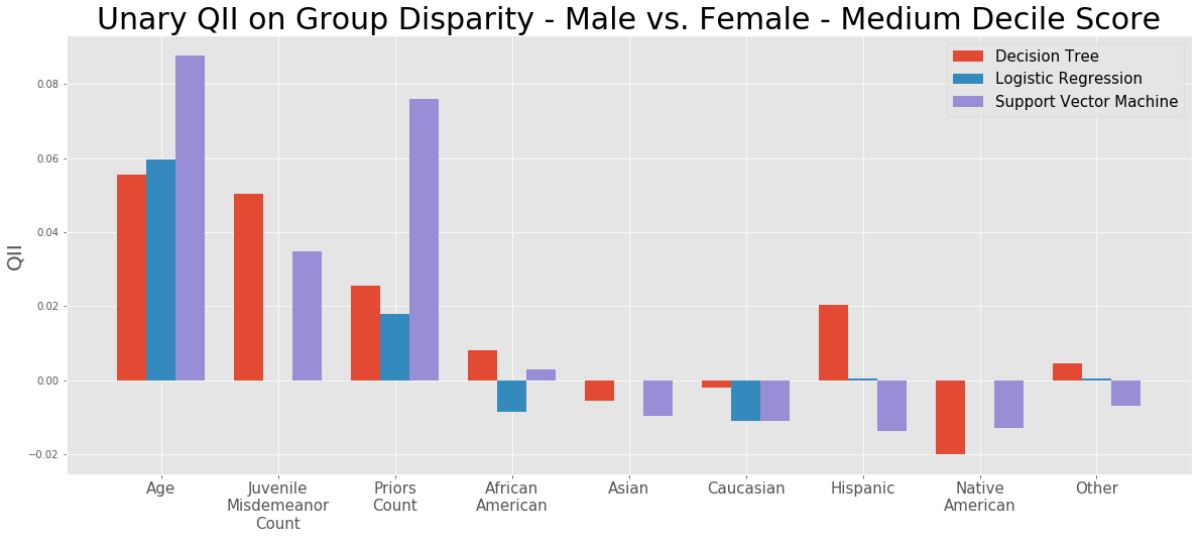
Interestigly, for the high decile score group, juvenile misdemeanor count and priors count did not seem to be discriminating inputs for discrepancy between assignments for male and female.

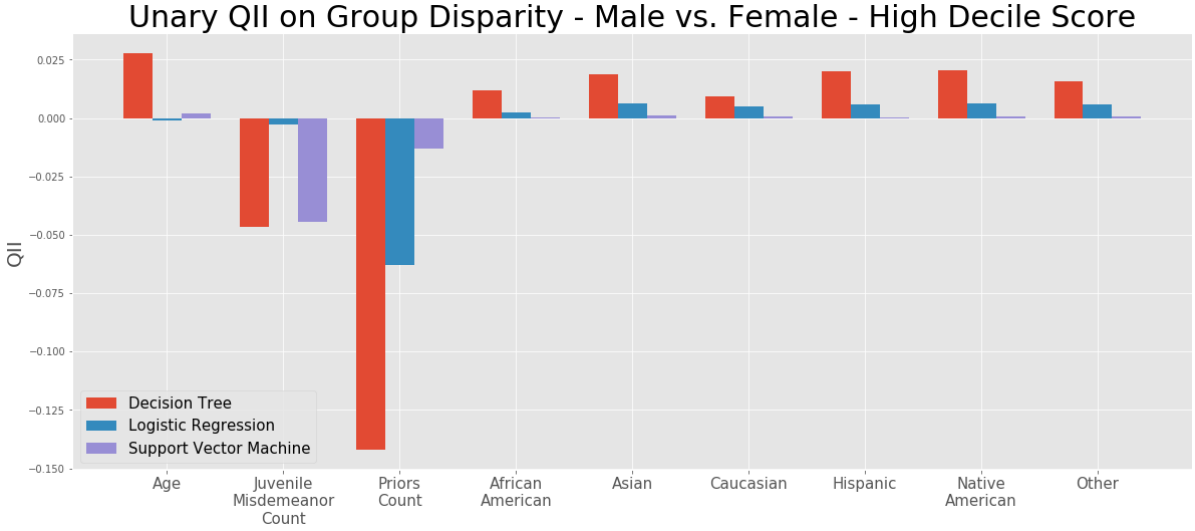
While the discrepancy is arguably minor, race appeared to have some effect towards gender group disparity. Most notably, for all decile score categories, the decision tree classifier always considered race to be important.

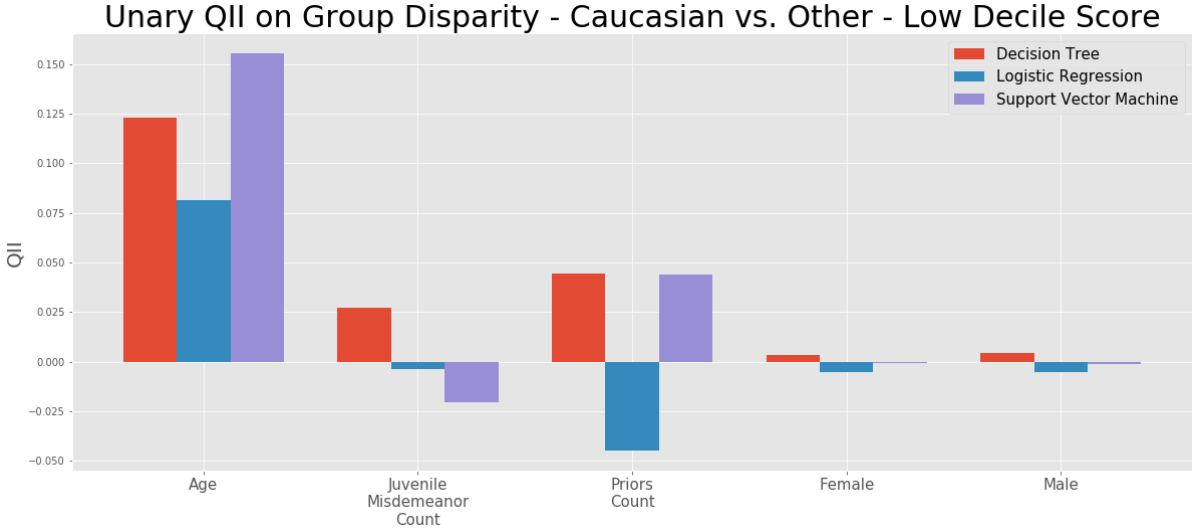
Caucasian vs. Other

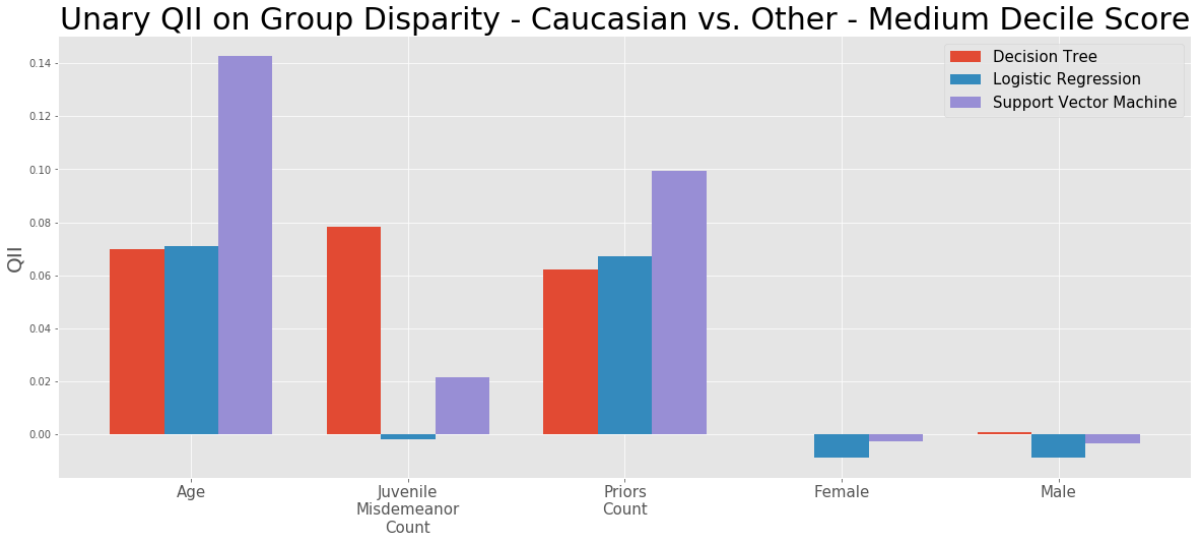
For all decile score categories and classifiers, age appeared to be an important contributor to discrepancy between assignments for Caucasian and non-Caucasian individuals. Priors count seemed to matter for low and medium decile score categories, but interestingly, not for high decile score category.

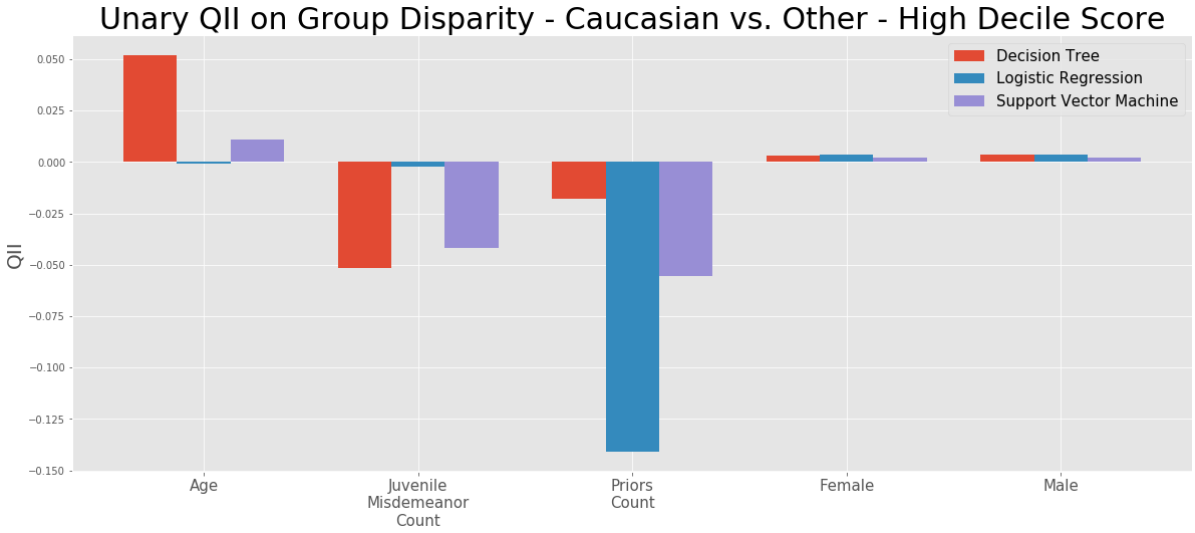



Unary QII on Group Disparity - Male vs. Female - Medium Decile Score
Decision Tree
Logistic Regression
Support Vector Machine
QII
0.08
0.06
0.04
0.02
0.00
−0.02
Age
Juvenile Misdemeanor Count
Priors Count
African American
Asian
Caucasian
Hispanic
Native American
Other


Unary QII on Group Disparity - Male vs. Female - High Decile Score
Decision Tree
Logistic Regression
Support Vector Machine
QII
0.025
0.000
−0.025
−0.050
−0.075
−0.100
−0.125
−0.150
Age
Juvenile Misdemeanor Count
Priors Count
African American
Asian
Caucasian
Hispanic
Native American
Other


Unary QII on Group Disparity - Caucasian vs. Other - Low Decile Score
Decision Tree
Logistic Regression
Support Vector Machine
QII
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
Age
Juvenile Misdemeanor Count
Priors Count
Female
Male





Transparency Report - Marginal QII

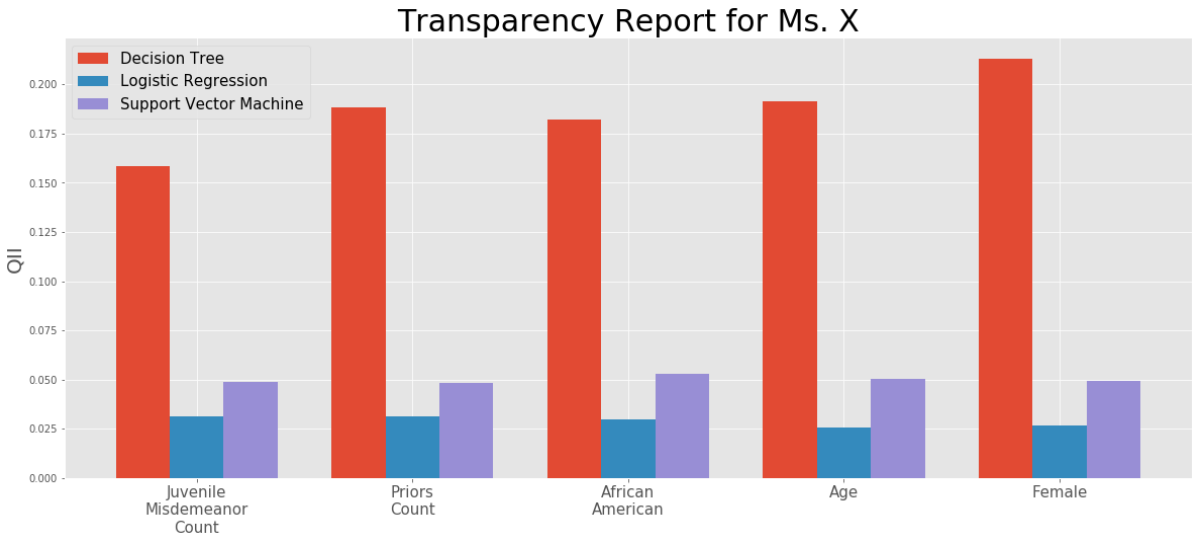
Transparency reports can provide individuals, such as Ms.X presented below, with detailed information on their classification outcomes. In the figure, we observe that the decision tree classifier appeared to have weighted each feature more heavily than the other classifiers. Essentially, this method can be applied to all individuals in the dataset in investigating featuer importance. Note that the classifiers do not necesarily yield the same classification outcomes for a given individual.

The transparency report for Ms. X show that her gender and age contributed most significantly to the algorithm predicting her decile score group to be 'Low'.

Information on Ms.X:

Out[105]:

	Ms.X
id	241
age	26
juv_fel_count	0
juv_misd_count	0
juv_other_count	0
priors_count	2
decile_score	4
score_text	Low
is_recid	1
v_decile_score	4
v_score_text	Low
is_violent_recid	0
two_year_recid	1
sex_Female	1
sex_Male	0
race_African-American	1
race_Asian	0
race_Caucasian	0
race_Hispanic	0
race_Native American	0
race_Other	0
pred	Low



V. Conclusion

In conclusion, the above methods provide some insight into the decisions that are made by blackbox algorithms. Specifically, we did not observe significant evidence that corroborated ProPublica's rhetoric of discriminatory behaviour of the COMPAS algorithm, with the caveat that our stand-in models are far from perfect. It would certainly benefit this study to include more data, and more features that are used in the original algorithm.

The QII framework is a very intuitive route for examining feature importance in automated decision making. We were able to employ this framework in our study to observe some interesting trends across race and gender. It is also worth noting that the three models we explore don't always perform consistently, which only reinforces the fact algorithms are all blackboxes in their own way.

Future work

Improve classifier performance

This problem has been tackled using classifiers that have been minimally tuned. We recognize that the best classification accuracy achieved by any of our models on out-of-sample predictions is 76.3% (SVM), which clearly falls short of perfectly predicting COMPAS scores. Had our goal been to predict COMPAS scores accurately, we would have balanced classes using some synthetic oversampling technique. Further, hyperparameter tuning could also be done to try for better model performance.

Differentially private releases

Given their construction, transparency reports can reveal sensitive information about the all data used to train the decision aiding classifier. We recommend making either the classifiers themselves differentially private (see Abadi et al. (https://arxiv.org/pdf/1607.00133.pdf) or making the release of transparency reports differentially private by addition of noise. Note that these procedures are computationally expensive for large data, so we leave it as an extension to our QII work.