# *Age of Unfairness*: Implementation details

Beau Coker, Caroline Wang

February 3, 2019

## Data processing [1]

Our data includes the same raw data collected by ProPublica, which includes COMPAS scores for all individuals who were scored in 2013 and 2014, obtained from the Broward County Sheriff's Office. There are 18,610 individuals, but we follow ProPublica in examining only the 11,757 of these records which were assessed at the pretrial stage. We also used public criminal records from the Broward County Clerk's Office to obtain the events/documents and disposition for each case, which we used in our analysis to infer probation events.

In their analysis [**?**, **?**], ProPublica processed the raw data, which includes charge, arrest, prison, and jail information, into features aggregated by person, like the number of priors or whether or not a new charge occurred within two years. We too process the raw data into features, partly to ensure the quality of the features and partly to create new features as defined by the components of the COMPAS subscales (see Tables **??**-**??**). Note that while ProPublica publishes the code for their analysis and the raw data, they do not publish the code for processing the raw data. Thus we did that from scratch.

The *screening date* is the date on which the COMPAS score was calculated.

- Our features correspond to an individual on a particular screening date. If a person has multiple screening dates, we compute the features for each screening date, such that the set of events for calculating features for earlier screening dates is included in the set of events for later screening dates.
- On occasion, an individual will have multiple COMPAS scores calculated on the same date. There appears to be no information distinguishing these scores other than their identification number. We take the scores with the larger identification number.
- Any charge with degree "(0)" seems to be a very minor offense, so we exclude these charges. All other charge degrees are included, meaning charge degrees other than felonies and misdemeanors are included.

[1] Reproduced from paper

- Some components of the Violence Subscale require classifying the type of each offense (*e.g.*, whether or not it is a weapons offense). We infer this from the statute number, most of which correspond to statute numbers from the Florida state crime code.

- The raw data includes arrest data as well as charge data. Because the arrest data does not include the statute, which is necessary for the Violence Subscale, we use the charge data and not the arrest data throughout the analysis. While the COMPAS subscales appear to be based on arrest data, we believe the charge data should provide similar results.

- For each person on each COMPAS screening date, we identify the offense — which we call the *current offense* — that we believe triggered the COMPAS screening. The *current offense date* is the date of the most recent charge that occurred on or before the COMPAS screening date. Any charge that occurred on the current offense date is part of the current offense. In some cases, there is no prior charge that occurred near the COMPAS screening date, suggesting charges may be missing from the dataset. For this reason we consider charges that occurred within 30 days of the screening date for computing the current offense. If there are no charges in this range, we say the current offense is missing. For any part of our analysis that requires criminal history, we exclude observations with missing current offenses. All components of the COMPAS subscales that we compute are based on data that occurred prior to (not including) the current offense date, which is consistent with how the COMPAS score is calculated according to [**?**].

- The events/documents data includes a number of events (*e.g.*, "File Affidavit Of Defense" or "File Order Dismissing Appeal") related to each case, and thus to each person. To determine how many prior offenses occurred while on probation, or if the current offense occurred while on probation, we define a list of event descriptions indicating that an individual was taken on or off probation. Unfortunately, there appear to be missing events, as individuals often have consecutive "On" or consecutive "Off" events (e.g., two "On" events in a row, without an "Off" in between). In these cases, or if the first event is an "Off" event or the last event is an "On" event, we define two thresholds, $t_{\text{on}}$ and $t_{\text{off}}$. If an offense occurred within $t_{\text{on}}$ days after an "On" event or $t_{\text{off}}$ days before an "Off" event, we count the offense as occurring while on probation. We set $t_{\text{on}}$ to 365 and $t_{\text{off}}$ to 30. On the other hand, the "number of times on probation" feature is just the count of "On" events and the "number of times the probation was revoked" feature is just the count of "File order of Revocation of Probation" event descriptions (*i.e.*, there is no logic for inferring missing probation events for these two features).

- Age is defined as the age in years, rounded down to the nearest integer, on the COMPAS screening date.

- Recidivism is defined as any charge that occurred within two years of the COMPAS screening date. For any part of our analysis that requires recidivism, we use only observations for which we have two years of subsequent data.
- A juvenile charge is defined as an offense that occurred prior to the defendant's 18th birthday.

## Machine learning implementation

Here we discuss the implementation of the various machine learning methods used in this paper. To predict the COMPAS general and violent raw score remainders (Tables **??**, **??**, and **??**), we use a linear regression (base `R`), random forests (`randomForest` package), Extreme Gradient Boosting (`xgboost` package), and SVM (`e1071` package). To clarify, we predict the COMPAS raw scores (not the decile scores, since these are computed by comparing the raw scores to a normalization group) after subtracting the age polynomials ($f_{\text{age}}$ for the general raw score and $f_{\text{viol age}}$ for the violent raw score). For XGBoost and SVM we select hyperparameters by performing 5-fold cross validation on a grid of hyperparameters and then re-train the method on the set of hyperparameters with the smallest cross validation error. For random forest we use the default selection of hyperparameters. For the COMPAS general raw score remainder, we use the available Criminal Involvement Subscale features (Table **??**), while for the COMPAS violent raw score remainder, we use the available History of Violence Subscale and History of Noncompliance Subscale features listed in tables Tables **??** and **??**, respectively. For both types of COMPAS raw scores, we also use the age at first offense. Race and age at screening date may or may not be included as features, as indicated when the results are discussed. To predict general and violent two-year recidivism (Tables **??** and **??**), we use the same methods, features, and cross validation technique as used to predict the raw COMPAS score remainders, except we adapt each method for classification instead of regression (for linear regression, we substitute logistic regression) and we include the current offense in the features.