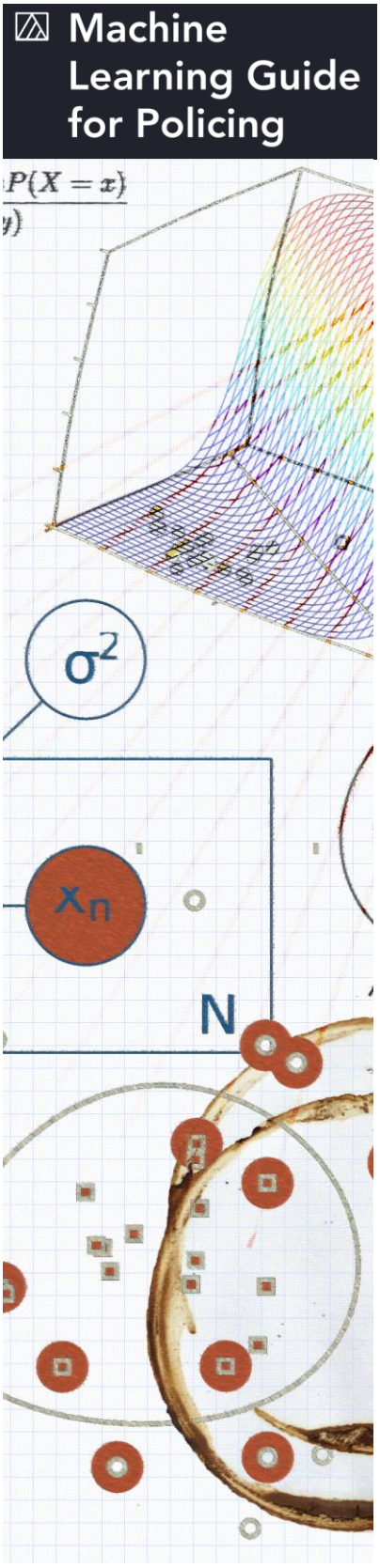


RUDI

The worked example

Tamara Polajnar, Hazel Sayer, Ruth Spence
STAR 2023/2024

Contents

Introduction

In this document, we are outlining an example of a machine learning model and filling out parts of the RUDI framework as a guideline. The full framework guide is available online as a webpage or as PDF. Here we do not provide as much detail as final documentation of a finished product might, but just a sketch with a very basic solution, so there will be sections we are unable to replicate.

We'd like to point out that in this case the bulk of the code is concerned with data preparation and analysis, as well as visualisation of test results and analysis of model fairness. This is likely to be the case with many similar projects, most of the effort will be in accurate data representation and task understanding rather than in programming the model. Model evaluation and maintenance are also an ongoing concern if the project proceeds.

We are working in the context and with knowledge of policing in England and Wales therefore there may be errors in interpreting our chosen publicly available dataset (COMPAS) which comes from Florida, USA and was created for a different purpose. The COMPAS dataset comes from a freedom of information request by Pro Publica and contains data from an algorithm that was used to predict the probability of recidivism for the purpose of aiding judges in sentencing decisions.

In England and Wales there have been several models that predict future risk of suspects, for example HART used by Durham Constabulary or using prediction to forecast domestic abuse (DA) homicides. In essence, all of these models are doing the same thing: predicting potential future harm based on past histories. There is a drive to automate risk prediction for serial and violent DA offenders and therefore many forces are considering producing algorithms like these.

We use the COMPAS dataset to mimic one of these tasks and are disregarding the original purpose of the data. To do this a leap of imagination is needed as the underlying data will embody different statistical properties than data from England and Wales with regards to label distributions and racial biases.

It is important to note that just because it is possible to train a model to predict potential for future harm with some degree of accuracy it may not be ethical to implement such a model, even with the best attempts to introduce fairness during training and mitigate underlying biases in data.

The **most important** aspect of fairness is how a model is implemented, monitored, and used by people who understand the context. All models will have incomplete understanding of the problem and will be trained on incomplete data. Therefore, they will embody biases of all the previous training data or models that have been used to make them. We must hold these limitations in mind when using a model to aid decision making that affects human lives.

Background

When building a machine learning model to replicate a **predictive labelling task** (supervised learning), we often have a dataset consisting of input data and a label we would like to assign to that data. We are training a model to replicate the task of automatically labelling the dataset.

The way the target labels are defined, and the original data sampled can increase or decrease the difficulty of classification. In a well-defined task, the more balanced the classes are within the dataset the easier the problem is.

Although, it is unknown how the COMPAS data was sampled and what the relationship is with the original incident report data, it contains individuals who were scored by the COMPAS system and there might be pre-screening for that too.

The COMPAS problem, as it was framed at the point of sentencing is easier than analysis of the raw suspect data, as the DA crime prediction was framed, because the class of interest makes a higher proportion of the dataset. That is people who are included in the COMPAS dataset have a higher likelihood of committing a serious crime (in the future component of the dataset, as we set out the experiment later on) than the ones that are contained in the incident report datasets from England and Wales because they have been sampled at the point of conviction and sentencing.

The types of crimes and crime labelling is different in this data and idiosyncratic to Florida. Therefore, we considered felony charges as an indication of high risk, whereas in English and Welsh data we would have the use of the higher-grained Cambridge Crime Harm Index. Disparate crimes are considered felonies including drug offences, car jackings, murder, sexual assaults, etc. For each case identifier there appear to be several charges of various levels.

The COMPAS dataset is most often used for fairness modelling so there are many instances of its use. More information on the dataset can be found in the following resources:

- Original article by ProPublica entitled Machine Bias
- ProPublica Methodology article and GitHub link
- Fairness in ML tutorial that covers a lot of important points related to working with COMPAS as well as a few notes on the dataset itself at about the 1hr mark
- Fair prediction with disparate impact: A study of bias in recidivism prediction instruments an article describing why calibration and fairness in error rates are incompatible. Calibration ensures that when the algorithm returns the value of 0.8 about 8 out of 10 people with that value in the test data were correctly labelled as relapsing within the next two years. Obviously, we don't know the true accuracy of such model after it has been implemented as it would have influenced the future outcomes through sentencing. Due to the population statistics of the dataset (available in the Development section) any errors by the calibrated model would lead to disparate outcomes for the different populations.
- COMPAS data datasheet from the Datasheet Repository for Criminal Justice Datasets

Prior to Modelling

Introducing modelling is a commitment in terms of time and resources. To successfully develop and implement algorithmic modelling in policing, diverse stakeholders must work together to define why and how modelling will work in practice, as well as have the specific expertise needed to develop and test large data models. We recommend police develop and build their own models in-house for two reasons: firstly, the model will be specific to the aims and data of each force and secondly, any model needs to be maintained over time.

Conceptualisation

We recommend police consider what they want to achieve through data modelling. Although guidance on how to conceptualise problems suitable for modelling is outside the scope of this framework, before beginning data modelling forces should be able to answer the following:

Conceptualisation template

What is the problem to be addressed?	We want to reduce the number of felonies committed in Florida
What is the proposed solution?	A machine learning algorithm that can predict if someone who has been arrested will commit a felony over the next 12 months so that they can be offered an intervention
What is the overall aim of the initiative and why is this important?	The aim is to find suitable candidates (those who are predicted to commit a felony within 12 months of their most recent arrest) for our intervention. This is important because it will reduce the likelihood of new felonies being committed
Who/what does the initiative target and why?	It targets suspects who have been arrested and who the algorithm predicts will commit a felony in the 12 months following their arrest
What are the key definitions being used and how are they operationalised? (e.g., high harm, recidivism risk)	Arrestee – anyone who has been arrested of any crime, using their most recent arrest as a trigger to be included in the algorithm. They don't have to have been convicted or charged with any crime, just arrested. Suspects are either considered likely to commit a felony in a 12-month period (high risk) or not (low risk) Felony – uses the definition from Florida Felony Charges: Know the Different Levels of Offenses (criminaldefenseattorneytpa.com)
What is the mechanism (e.g. professional judgement, structured professional judgement using a tool, static algorithm,	A machine learning based algorithm is being used due to the large quantity of arrest data

machine-learning based algorithm) by which cases of interest will be identified and why?	
Will identified cases go through further assessment? If so, what does this look like?	Cases identified by the algorithm will be sent to the intervention team to be assessed as to whether they believe they are a suitable candidate based on their own criteria
What kind of action will be taken? How is this justified and is there resource available for this?	Suspects who are deemed suitable candidates will be offered the intervention
How does all of the above fit within your forces legal and ethical framework?	The intervention is voluntary and will target predictors of antisocial behaviour
What evidence base are you using to justify all of the above steps? (briefly state underlying theory & hypotheses)	For this analysis example, we did not draw on any evidence base or theory

Note: *We suggest you do not proceed any further until you can complete the above table*

RUDI

RUDI is a framework police forces can follow to build and implement data modelling. It is specific to complex machine learning models which can drive police action and have potential public impact, such as models that prioritise individuals or predict future criminal behaviour or locations. It covers four main stages:

Rationale: Documenting the process, making decisions explicit and justifying actions during unification, development and implementation.

Unification: Merge data sources together for modelling and ensure validity and reliability of data.

Development: Build and test models, evaluating for bias, performance and limitations and choose preferred model.

Implementation: How the model feeds into current practice and how it will be maintained over time.

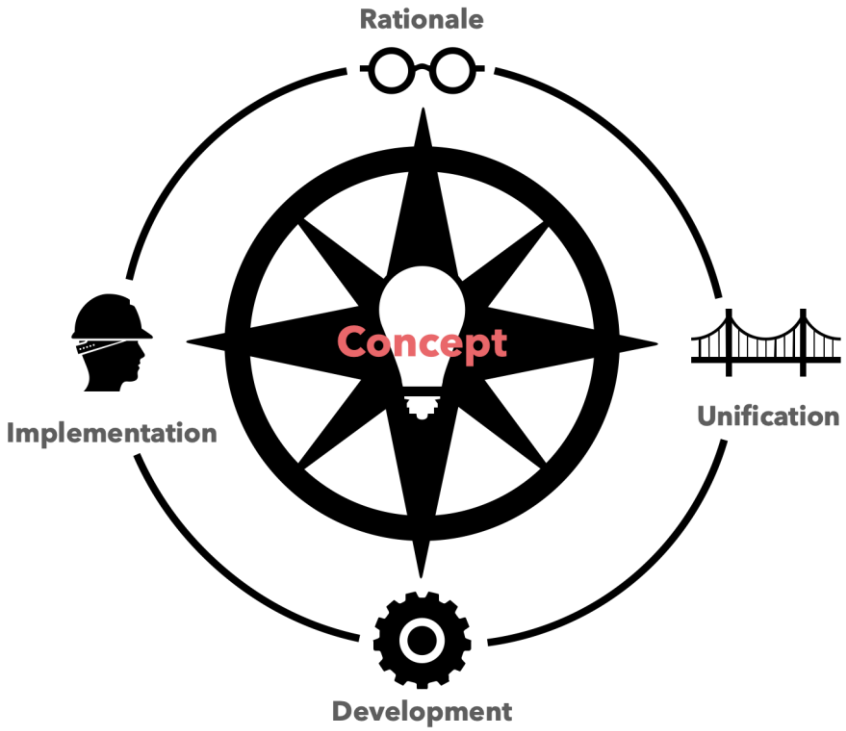


Figure 1: RUDI process

Rationale

Responsible team members: The whole team

The RUDI rationale section outlines the reasoning behind the project as guided by the business template and encourages formation of a multidisciplinary team to ensure that the project development is ethically sound and fits into organisational structure. Refer to the full RUDI guidelines for details.

BUSINESS CASE TEMPLATE

Team	
Senior Responsible Officer	
Data Engineer(s)	
Data Scientist(s)	
Domain Expert(s)	
Validator(s)	
Governance Expert(s)	
Plan for if someone leaves post	There are several data engineers and data scientists who can be debriefed should someone leave. The model and data cards, plus meeting minutes and these templates can be used to bring domain experts or a new SRO up to speed.
Model	

Outline the problem to be addressed and the overall aim of the model, including if relevant, who the model is being used to target, and why	There are many felonies committed in Florida every year, the model aims to identify those who will commit a felony within 12 months of being arrested. This is to identify suitable candidates for an intervention that reduces serious offending.
Why is algorithmic modelling rather than other options (e.g., professional judgement) best suited to solving this problem?	Due to the large amount of data, algorithmic modelling can help identify new cases which may be relevant for our intervention. This will be in parallel to professional judgment as officers can also refer to the intervention team.
Alignment with force priorities	The force has committed to reducing the worst crimes
Alignment with national strategic priorities	Addresses the national strategy of reducing violent crime
Briefly state underlying theory & hypotheses	For this example, we did not draw on any underlying theory or have a hypothesis about the link between arrests and felonies as all individuals in the dataset had an arrest
Desired outcome(s) i.e., how will you know the model is working?	The number of felonies in Florida will decline
Possible undesired outcome(s) (e.g., either directly or through misuse)	The intervention will have too many cases to deal with and there will not be enough space to take suitable cases The intervention team will become over-reliant on the model when making decisions about suitability Model could be used to justify more punitive methods rather than being referred to the intervention. The intervention could stop being offered on a voluntary basis – given this is suspects rather than convicted people. Model bias could propagate stereotypes
Model design (e.g., classification, ranking) & rationale	We have chosen to use ranking because there may be issues with capacity (numbers of people the intervention can accept) therefore we want to focus on the most likely people who will commit a felony
Data features needed & rationale	Arrest data – this is the trigger Current and past charge data – criminal history, including current and past charges

	are thought to increase the risk of future felonies Prison and jail histories – we expect those with prison histories to be more likely to commit a future felony, but if they are sent to prison after their arrest it may explain some false positives Ideally previous intervention data would also be used if available. This is so we can see the effect of previous interventions – again it might explain some false positives i.e., reduces risk of people who may otherwise commit a felony
Data analysis plan & rationale (e.g., how bias will be assessed, how the model will be evaluated and analysis of errors)	Bias will be assessed through fairness rebalancing using demographic parity because ethnicity is expected to affect model results. Model will be evaluated using a combination precision, recall and ROC AUC because it's a classification model which is then used to rank people
Inclusion and exclusion criteria for cases being included in the model & rationale	Anyone being arrested in Florida will be included. This is because we are focusing on Floridian crime and arrests are the trigger Excluded – anyone not being arrested or being arrested outside the state of Florida. We are not interested in arrests occurring in other states or those who do not reach the threshold for arrest
Plan for storing & sharing output	Output will be stored in the suspect database. Other than the data science team, the output will only be shared with the intervention team who will be given training on the model and its output
Who will use the output & will training on using the output be provided, if yes what? and if no, why not?	The data science team will run the model and produce a report for each meeting of the intervention team. The intervention team will be given training on the model and output. Training will include model limitations and bias, model drift and decision-making bias
How will the model's outputs be incorporated into officer decision-making? Are processes in place to catch cases where the model is inaccurate?	The outputs will be used along with multiple other sources of information, including case files and HCR-20 ratings, when the intervention team is assessing potential candidates for the intervention. Their professional judgment will allow them to over-ride the model and OICs can independently bring cases they believe are suitable candidates.

If relevant, what is the intervention for the cases the model identifies? Are there situations where identified cases will not be considered for an intervention?	Suspects identified as likely to predict a felony will be referred to the intervention team for further assessment. The intervention focuses on how to mediate predictors of antisocial behaviours. The intervention team will decide on a case-by-case basis if arrestees highlighted by the algorithm should not be given the intervention based on professional judgment.
What are the implications of an error (both false positives and false negatives) e.g., ethical, legal, reputational	False positive – the suspect will be asked to attend a voluntary intervention which may be beneficial. May take time & space on the intervention that could have been given to a more suitable candidate. False negative – a suspect may go on to commit a felony offence. This could cause reputational risk as they will have been missed. Also, ethical risk as it may involve harm to victims.
What is the plan for ongoing model evaluation (e.g., thresholds and inputs/outputs that trigger model retraining)	When intervention team and algorithm-flagged cases disagree on more than 7/10 cases per session over a 6-month period. Additionally, if data starts to skew from projected model performance and general population statistics more than 10% (the dataset was already skewed, this is merely for illustrative purposes).
Can iterative changes be made to the model as needed? (e.g., to account for feedback loops due to the effects of interventions)	No, we will retrain the model as needed rather than make iterative changes because of the low computational overhead involved
Costs & Resourcing	
What costs/resources will be needed for set up & piloting?	Costs are needed for computational power, resources include personnel - data science team and intervention team time
What costs/resources will be needed for model maintenance?	Data science team are needed to maintain the model and intervention team need to work with them to ensure model continues to function effectively. As dataset grows computational power needed may increase.
Changes & Trade-offs (to be completed during project lifecycle)	
Any changes to model design & rationale	
Any changes to analysis plan & rationale	

Any trade-offs & rationale e.g., false positive/negative rates, accuracy vs. explicability	

Note: *this is not exhaustive and should be added to in line with the force's own concerns. Responses will change as the project progresses*

Unification

Responsible team members: Data Engineer and Data Scientist to lead

In this case we're working with ProPublica COMPAS data. The data comes in SQL database format with 7 tables: casearrest, prisonhistory, jailhistory, compas, summary, charge, people. There are multiple errors in the tables with empty columns and inconsistencies in, for example, name fields. One table will have separate first and last name fields, while another will have the two names concatenated in a single field. There are issues with some of the date fields as well. Some are null, while the date charge filed field is the same as the date offence committed field in the charges table. This is unfortunate because there might be implicit information about the seriousness of the offence in the time between the offence being committed and the charges being filed.

The tables are already pre-aligned by person_id and name, so there is nothing needed for further unification on this dataset other than future joins to form different types of datasets for different studies.

We have included one html page per table generated by ydata-profiling Python package, which allows a quick overview of the table structure and values.

Details of the original compas tables:

- Casearrest
- Prisonhistory
- Jailhistory
- Compas
- Charge
- People
- Summary - empty

Development

Responsible team members: Data Scientist to lead

This section is intended as indicative of the steps involved in analysis not an exhaustive list of what analysis to conduct

PROJECT PREPARATION

It is important to take the time in the beginning of the project to set up data and code versioning. As you explore the solution space including various iterations of features and models, it will be important to keep track of the steps that led to the best performing solutions.

This is just a small example, and we chose not to include extra packages for experiment tracking at this time. Although, we did devise dataset and task naming schemes to keep track of the versioning manually.

DATA PREPARATION (PREPROCESSING)

Importing: Write code that gathers dataset features from the available unified data and verifies data integrity based on any assumptions that are made on formats, values, or quality.

- The data comes in SQL database format with 7 tables: casearrest, prisonhistory, jailhistory, compas, summary, charge, people.
- We discard the compas table and all its contents and any compas-specific columns. We join the other tables per person on person_id taking care to double check that the only correlating feature we have, name, also matches.

Cleaning and feature manipulation: Remove or correct entries so the data is valid and reliable. This includes removing cases that are inappropriate, as outlined in rationale, as well as assessing missing data. You may also end up scaling or combining features, so you need to pay special attention to how scaling factors translate to unseen examples.

- There were columns in the database without content or with repeated content and we discarded those

Labelling: Accurate data labelling improves model predictions. Labels should be consistently applied based on the aim of the model using pre-determined criteria, so the labels reflect this as closely as possible (e.g., high risk cases are labelled as such if they meet certain criterion, such as having a particular harm score or a certain number of convictions).

- We take a subset of columns that might be relevant for this project including age, race, sex, juvenile offence counts, charge types (determined from a series of codes), charge descriptions, jail and prison histories.
- We consider each time someone was arrested as a possible trigger for the model. The time of the arrest is then considered the "time now" and any crimes occurring before that time constitute past arrest history. This means that the same person can occur several times in the training data.
- The label in this case will be the number of felonies committed in 1 year following "time now"
    - To calculate the label accurately we must save any crimes that occur in the final year of the dataset for creating labels, so we automatically skip those. We also skip any crimes that are followed with a prison term within the next year as that is considered a significant intervention that would limit the ability to commit a felony. Ideally, we would want to consider all types of interventions and their effects.
    - This leads to 38774 data points of which 28109 are negative and 10665 are positive. This still a high ratio of positive examples because the people sampled in this dataset have already been referred to courts.

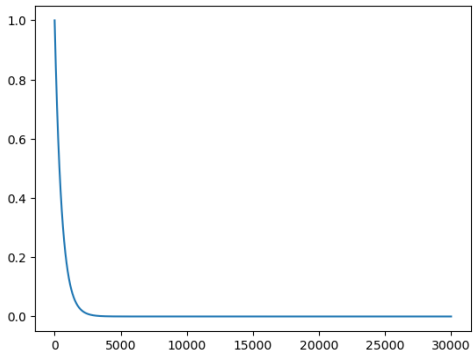
- We looked at using next two years to generate but that leads to a higher ratio of positive to negative labels (11809 positive to 16681 negative), making it an easier problem to solve, but also less comparable with the setting where the data is sampled from general arrest data and not from people who have already been referred to courts. This also leads to fewer data points because we need to exclude all crimes within two years of the dataset finishing time.
- Charges labelled M* are misdemeanours, F* are felonies. Top 30 charges in the dataset are

(M1)_Possess Cannabis/20 Grams Or Less	5405
(M1)_Battery	5202
(M2)_Driving License Suspended	4962
(F3)_Possession of Cocaine	4659
(M1)_Resist/Obstruct W/O Violence	4565
(0)_Fail Wear Safety Belt/Operator	4357
(F3)_Grand Theft in the 3rd Degree	4314
(0)_Pers/Inj/Prot/Ins Require	4107
(M2)_Operating W/O Valid License	3713
(M1)_Possess Drug Paraphernalia	3025
(0)_License Suspended W/O Knowledge	2197
(0)_Expired Tag/Infraction	2026
(M2)_Petit Theft	1862
(0)_Disobey/Ran Stop Sign	1776
(0)_Oper Veh Unsafe/Improp Equip	1775
(0)_Evasion/Fail To Pay Toll	1697
(0)_Speed Posted Municip/State Rd-Driver	1671
(M2)_Susp Drivers Lic 1st Offense	1594
(0)_Unlawful Speed (Requires Speeds)	1498
(M2)_Unlaw LicTag/Sticker Attach	1443
(0)_License Not Carried Exhibited	1421
(0)_Red Light Camera Violation	1367
(0)_Side Wind/Rest Sunscreen	1320
(M2)_Fail Register Vehicle	1224
(0)_Speed Posted Municip/State Rd	1193
(F3)_Driving While License Revoked	1171
(F2)_Burglary Unoccupied Dwelling	1156
(M1)_Driving Under The Influence	1130
(0)_Fail To Stop Steady Red Signal	1081
(0)_Careless Driving	1060

- For the purposes of this example, we do a minimum of processing to create a feature set including:
  - person_id, age, race, sex, j_felonies, j_misdemeanor, j_other, felonies_case, misdemeanors_case, other_cases, felonies_charge, misdemeanors_charge, other_charges, pred_label, , total_felonies_charges, total_misdemeanors_charges, total_other_cases, 1_felonies_case, 1_misdemeanors_case, 1_other_cases, 1_total_prison, 1_total_jail, 1_time_norm
    - Where j_* is the count of junior infractions grouped into felonies, misdemeanours, and other directly from the person table.
    - Total_* is the count of the number of felony, misdemeanour, and other charges that we have in the whole criminal history. When you group by person and then by time or case number it becomes clear that each time someone is arrested, they are charged with several offenses of different levels.

- *_case variables refer to values in the current case (the one we're considering "time now"
- *_charge variables refer to the charges table lines that correspond to the "time_now" cases
- n_* where n is some number between 1 and number of total past cases you want to consider
    - Time norm is a decayed value of the number of days from the past charges to "time now" calculated in python using the expression 0.5^(diff_days/365.)

Function value vs diff_days



- In this example, we consider 5 past cases only
- We do a small amount of word processing to separate out unique tokens from the charge descriptions and choose words that occur more than 5 times across the dataset, removing some high frequency connectives and symbols. We make two sets of token vectors, ones describing the current charges and the ones describing all of the past charges. Ideally you would want to normalise word vectors separately row-wise, so that you get the influence of each word per person rather than across people, but we leave raw counts as are in this case.

  Words from the charge descriptions are chosen using the following formula:
    - exclude these words: ['an','of', 'to', 'without', 'while', 'with', 'on', 'none', 'at', 'by', 'as', 'the', 'or', 'in', 'over', 'and', 'nd', 'too', 'one', 'two', 'when', 'for']
    - exclude any words that occur fewer than 5 times and have a single character

Details of the processed dataset can be seen in the HTML report.

Assess and mitigate bias: Try to discover and control the biases before using the data to train models

There are visible biases in the data with regards to race distribution. Below is a table captured on USA Facts

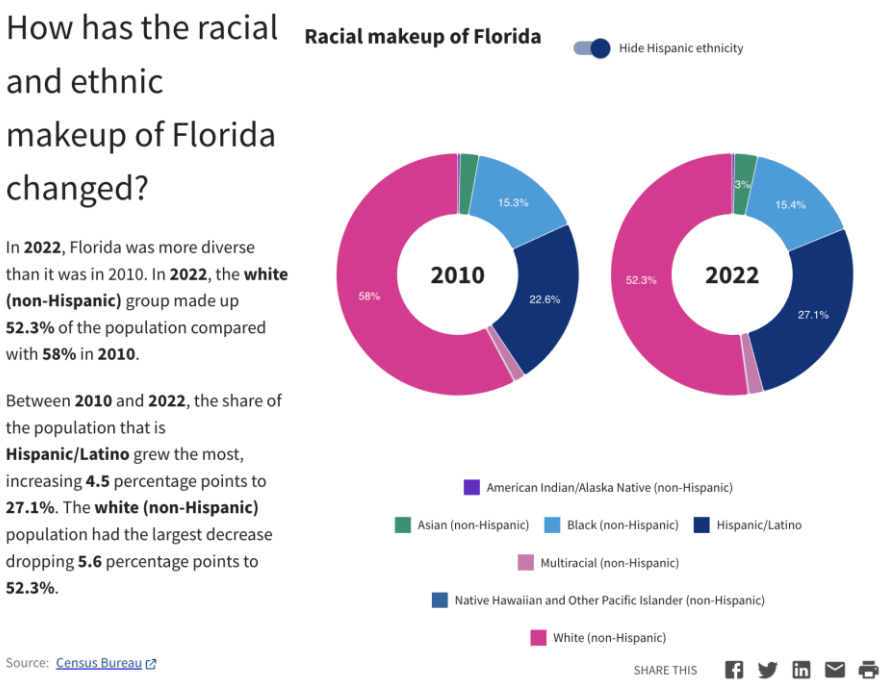


Figure 1

We can see that the black population is about 15% of the overall population of Florida, where the COMPAS dataset originates. In the people table of the dataset we see the following distribution where almost 50% of the suspects are reported as African American.

Value	Count	Frequency (%)
african-american	5813	49.3%
caucasian	4085	34.6%
hispanic	1100	9.3%
other	661	5.6%
asian	58	0.5%
native	40	0.3%

Figure 2

Therefore, the amount of societal bias encoded in this dataset is difficult to unpack. We can observe some more trends.

In the following graphs we're checking the correlation between Filing Agency as location proxy and race.

By normalising row-wise (Figure 3) we see that certain locations like Broward Sherrif Office/Lauderdale Lakes or grand jury predominantly recommend charges for African-American suspects; County Court predominantly deals with Hispanic suspects; while Fire Marshalls and US Marshals seem to deal with Caucasian suspects.

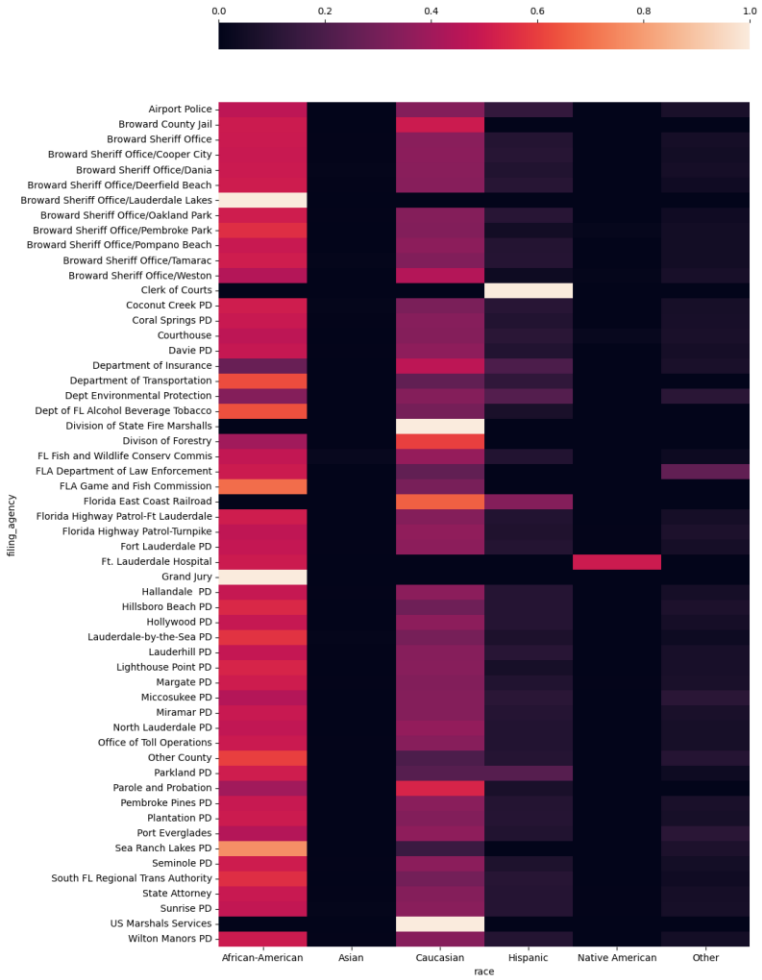


Figure 3

By normalising column-wise, we're looking at distribution of locations for each race. So, we can see that Broward Sheriff Office seems to cover a large population centre with most of the groups represented there but slightly more of the Asian community is found there than in other places. Ft Lauderdale PD also seems to have a large population catch but interacts with a larger percentage of the Native American population than any of the other offices.

Figure 4

We can look at the similar breakdown of charge vs race. By normalising across rows we can see that almost on every charge type except CT (Contempt of Court), F5 (Sex Battery Deft 18+/Vict 11- or Armed Sexual Battery), or X (None or Reckless Driving), has a higher percentage of Arican American suspects.

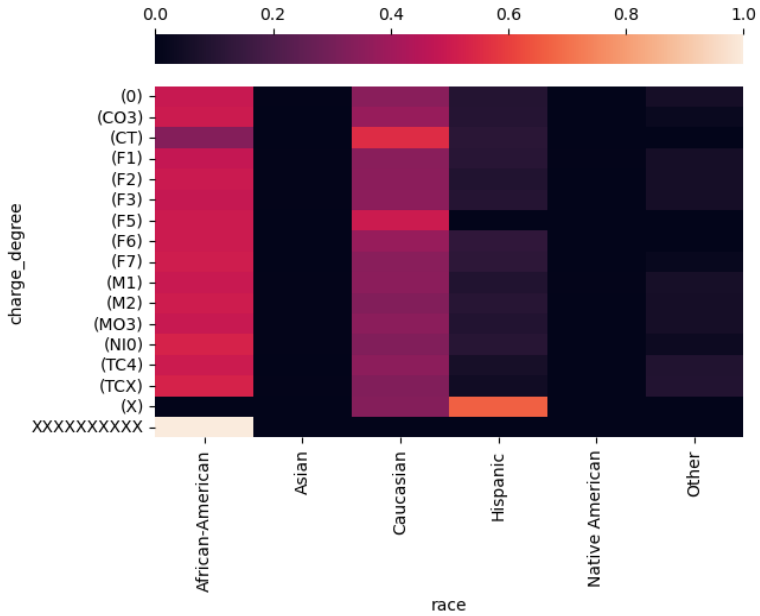


Figure 5

By normalising across the columns, we can see what charges each race is more likely to incur.

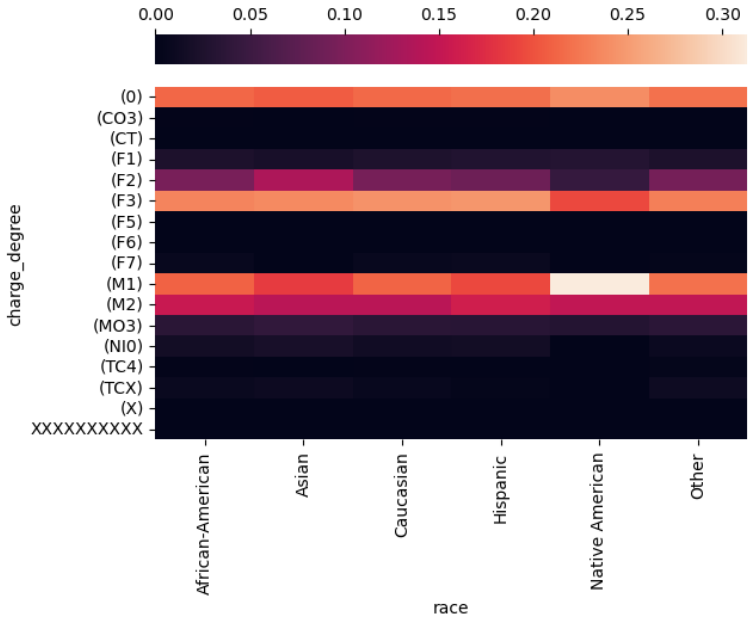


Figure 6

We can also observe trends by gender. By normalising across the rows, we can see what that men are more likely to commit most crimes other than CT - contempt of court, F5 sexual battery, and F6 murder in the first degree.

An interesting side note here is that in the ML Fairness tutorial they note that contempt of court in US is not a simple issue, as people often have multiple court hearings and can be charged with failure to report for missing any one of them. For many people this means conflicts with work and care arrangements, which would disproportionately affect women.

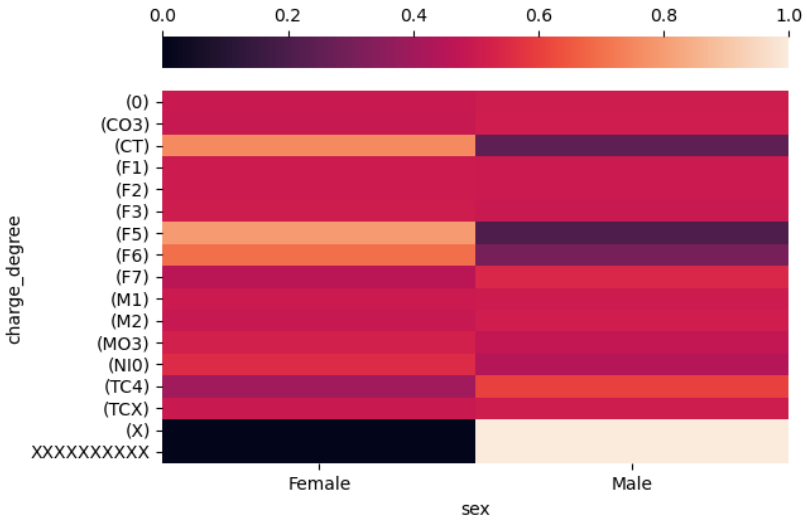


Figure 7

At this point you should engage the multidisciplinary team in analysis of the data trends, what it means for your organisation and whether it makes sense to proceed with the modelling. And if you choose to go on you need to discuss how you could mitigate the clear trends and what can be done to ensure that bias is not propagated through the model or model use.

We chose to disregard the referring agency as an input field but to keep the sex field. You should also do a more thorough correlation analysis between the protected features of interest and the other inputs.

Splitting Data

We split the data into train, validation, and test by taking the list of unique person IDs and randomly splitting them into 20% for test and 80% for training. From the training data we reserve a quarter for validation, effectively leaving us 60% of the individuals for training, 20% for validation, and 20% for testing. We then choose all the data points associated with each person and put them into the corresponding partitions. In this way we ensure that we do not have examples of the same past incidents in the training and test data. It is important we do that because we use each individual's past case data to generate several data points.

Training Data

Groups	Number of Data Points	Positive Count	Negative Count
african-american	14235	4389	9846
caucasian	6873	1633	5240
hispanic	1438	320	1118
other	856	148	708
asian	54	6	48
native american	54	20	34
total	23510	6516	16994

Validation Data

Groups	Number of Data Points	Positive Count	Negative Count
african-american	4627	1456	3171
caucasian	2206	456	1750
hispanic	515	99	416
other	336	93	243
asian	40	11	29
native american	35	9	26
total	7759	2124	5635

Test Data

Groups	Number of Data Points	Positive Count	Negative Count
african-american	4539	1435	3104
caucasian	2086	439	1647

hispanic	551	111	440
other	277	30	247
native american	31	7	24
asian	21	3	18
total	7505	2025	5480

MODELLING

Train the models on the training data and conduct error analysis on the validation data:

- Any trade-offs between accuracy and interpretability should be clearly thought through and documented
    - Evaluate how good each model is for achieving the aim based on pre-set criteria
    - Choose the best model and investigate error rates, sources of error, assess for bias and limitations within the model
    - Consider different misclassification costs and consider the asymmetry of misclassification costs (e.g., in risk assessment, incarceration versus wasted police time)
    - If you uncover patterns in the sources of error or unfairness, go back through feature engineering, data splitting, and modelling to improve the process
- Do final checks on the held-out test data
    - Check for bias on the chosen model, document the trade-offs
    - Conduct interpretability/explicability tests on the chosen model, record any known patterns

Model training

As this is just an example, we made a quick model using the AutoML tool AutoGluon. We're working on an Apple laptop with an M2 chip so there were some issues with some python packages (lightgbm), and the models take about 6 minutes to train.

```
predictor = TabularPredictor(label=current_label,
                             eval_metric='f1_macro',
                             path=outpath).fit(train.drop(columns=['race']),
                                               tuning_data = val.drop(columns=['race']),
                                               keep_only_best=True,
                                               excluded_model_types=['GBM'],
                                               num_gpus=0)
```

This is a quick way to create a robust baseline model. AutoGluon also comes with a fairness package which we can use to examine performance with regards to a protected characteristic.

This gives a final model with the following components:

```
                 model  score_test  score_val  pred_time_test  pred_time_val  \
0       NeuralNetFastAI    0.599639   0.586153        0.199461       0.215972
1  WeightedEnsemble_L2     0.586165   0.593347        0.650118       0.711113
2        NeuralNetTorch    0.585511   0.590249        0.449339       0.488049

     fit_time  pred_time_test_marginal  pred_time_val_marginal  \
```

```
0   37.273778                0.199461                0.215972
1  268.246343                0.001318                0.007092
2  228.778178                0.449339                0.488049

   fit_time_marginal  stack_level  can_infer  fit_order
0          37.273778            1       True          1
1           2.194387            2       True          3
2         228.778178            1       True          2
```

We also fit a model using the race column information. The idea is that "fairness through unawareness" doesn't work but the question often crops up whether we are discarding useful information by removing some protected characteristics. At first glance it leads to an improvement of scores and a different type of ensemble.

```
                model  score_test  score_val  pred_time_test  pred_time_val
0  WeightedEnsemble_L2    0.620699   0.609052        1.046038       0.749058
1       NeuralNetTorch    0.608124   0.587176        0.450192       0.448455
2     NeuralNetFastAI    0.600225   0.592538        0.495920       0.217587
3              XGBoost    0.553044   0.563452        0.098033       0.075340

     fit_time  pred_time_test_marginal  pred_time_val_marginal
0  318.372494                 0.001893                0.007676
1  242.028368                 0.450192                0.448455
2   39.444080                 0.495920                0.217587
3   34.177085                 0.098033                0.075340

   fit_time_marginal  stack_level  can_infer  fit_order
0           2.722961            2       True          4
1         242.028368            1       True          3
2          39.444080            1       True          1
3          34.177085            1       True          2
```

Model fairness

On the class of interest (committed a felony in the next year) the improvement is actually only 0.02 in the F1 score. We use the fairness module to decompose the performance of the original models and the performance of the models re-tuned for demographic parity. Note that we also used the fairness module to enforce multiple fairness constraints, but this just led to a recall of 1 on all racial groups and a potentially useless model.

```
# Modify predictor to enforce fairness over the training data with respect to groups given by the column 'race'
fpredictor = FairPredictor(predictor,train,'race')
# Maximize accuracy while enforcing that the demographic parity (the difference in positive decision rates between different races is at most 0.02)
fpredictor.fit(gm.f1,gm.demographic_parity,0.02)
```

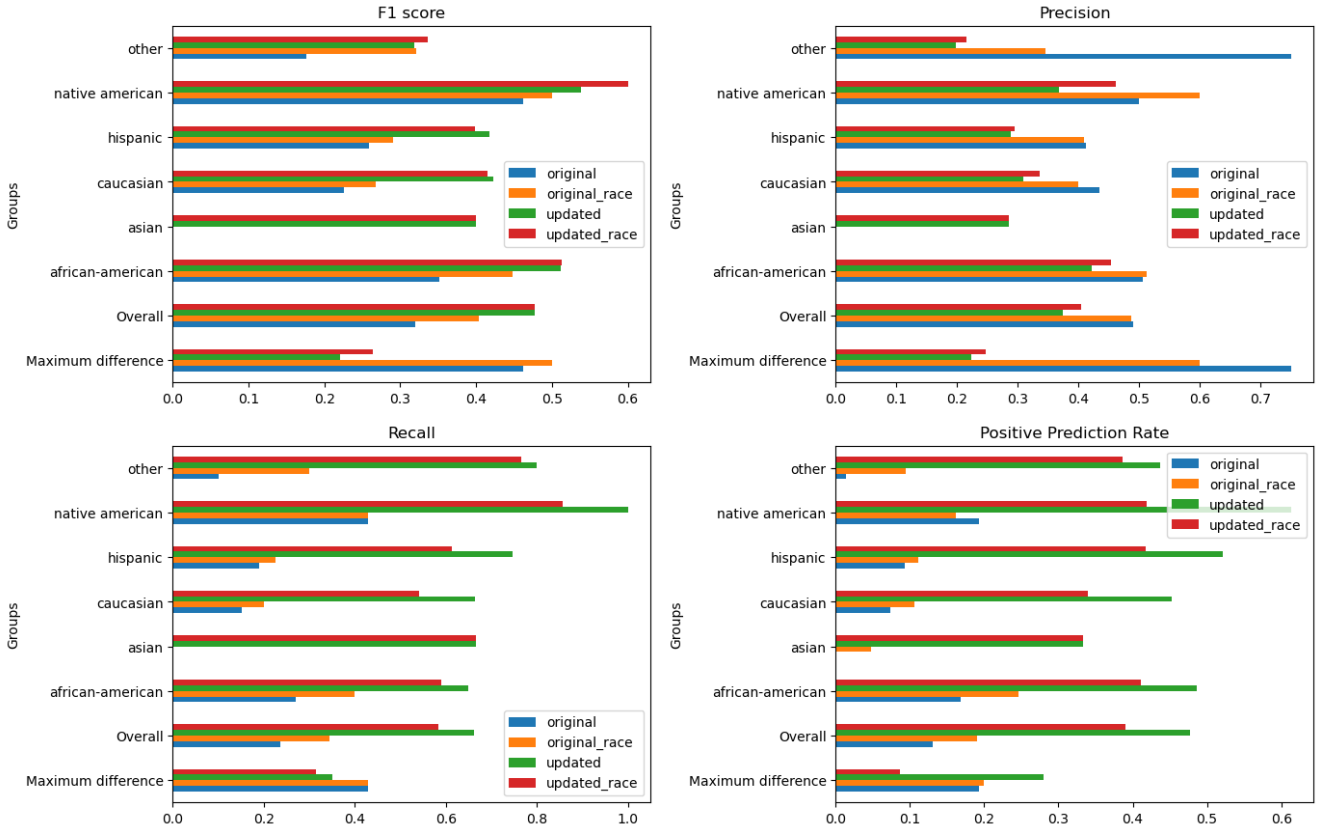
Demographic measures for the different models are shown below. Due to the large differences in population due to the inclusion of the lower represented classes some of the measures are higher than they are if you concentrate on the disparity between African American and Caucasian populations.

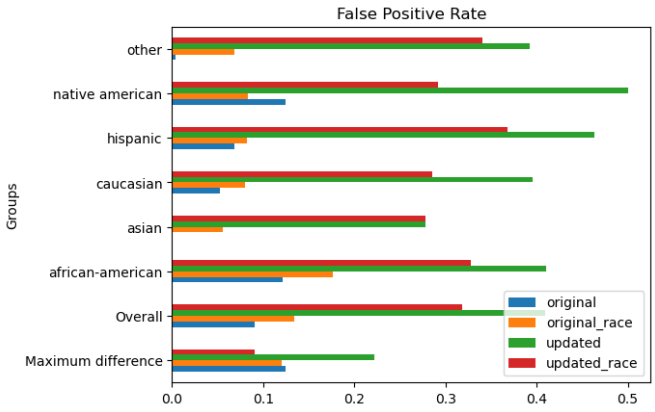
original	
Class Imbalance	2.078457e-01
Demographic Parity	2.000441e-01
Disparate Impact	0.000000e+00
Maximal Group Difference in Accuracy	2.222010e-01

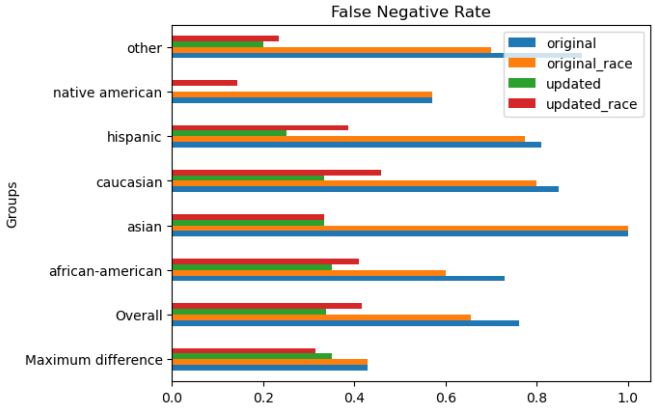
Maximal Group Difference in Recall	4.285714e-01
Maximal Group Difference in Conditional Acceptance Rate	2.999999e+06
Maximal Group Difference in Acceptance Rate	7.999998e-01
Maximal Group Difference in Specificity	1.488402e-01
Maximal Group Difference in Conditional Rejectance Rate	1.051390e-01
Maximal Group Difference in Rejection Rate	1.767885e-01
Treatment Equality	7.499998e-01
Generalized Entropy	1.683715e-01

	updated
Class Imbalance	2.078457e-01
Demographic Parity	1.616397e-01
Disparate Impact	7.362709e-01
Maximal Group Difference in Accuracy	6.558473e-02
Maximal Group Difference in Recall	3.226479e-01
Maximal Group Difference in Conditional Acceptance Rate	3.748243e-01
Maximal Group Difference in Acceptance Rate	2.164524e-01
Maximal Group Difference in Specificity	9.514172e-02
Maximal Group Difference in Conditional Rejectance Rate	5.922901e-01
Maximal Group Difference in Rejection Rate	2.099772e-01
Treatment Equality	1.200000e+07
Generalized Entropy	1.137748e-01

Below are visualisations of some of the performance metrics, keep in mind that statistics for the underrepresented classes (ones other than African American or Caucasian) will be subject to higher error.



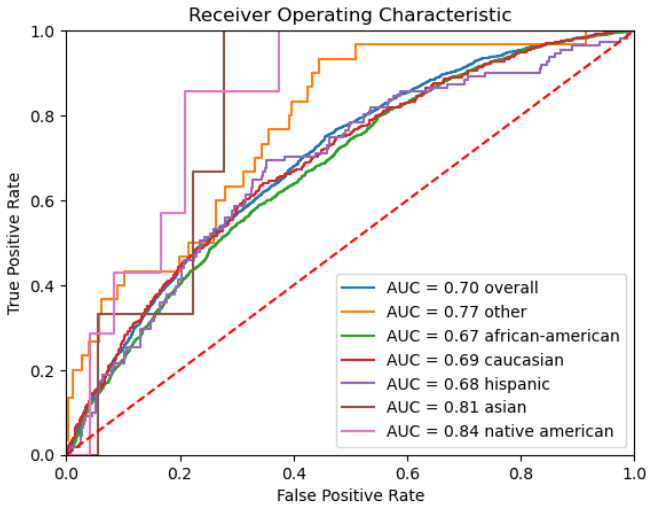




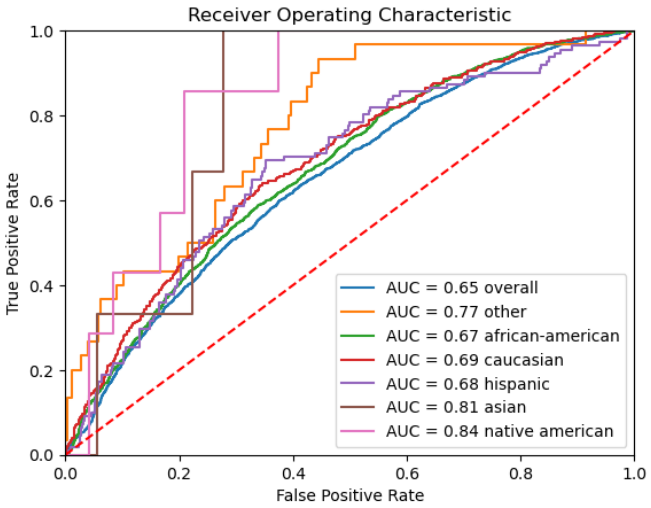
We can see the effect of using race in training data is that it leads to a lower false positive rate fairness adjusted, but that fairness measures increase overall false positive rate since they increase recall.

In unadjusted models the false negative rate is higher in models using race, but lower after adjustments in general.

original

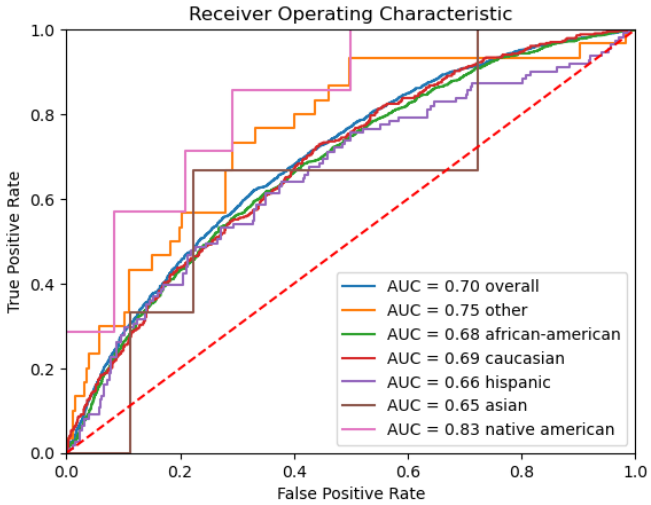


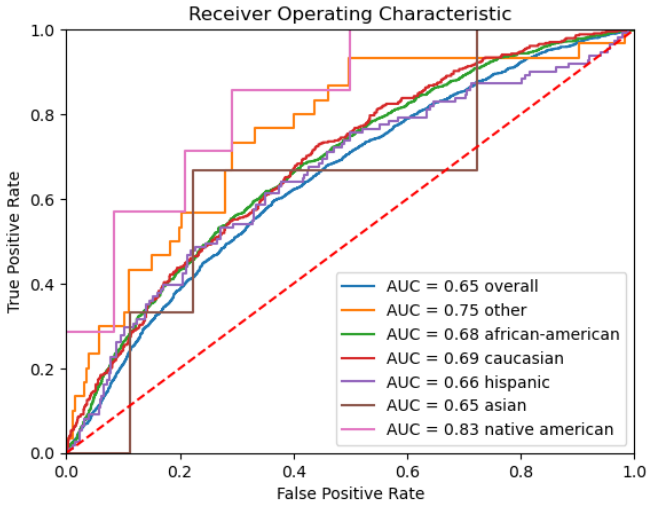
updated



original race

updated race





There is no difference in AUCs between the original and fair models and within racial categories, but overall AUC is higher in the original model. This indicates that the fairness measures are changing the order between groups by changing the ranking values.

Model explainability

We are unable to examine the feature importance of the fairness updated model, because the package does not currently have this capability, but we did look at the most predictive features in the original model. From the top 30 features we can see that the overall number of offences and the recency of those offences are key. Words indicating violent offences such as battery and drugs offences such as possession and cocaine also feature.

importance	stddev	p_value	n	p99_high	p99_low		
total_felonies_charges	0.022763	0.003953	0.000105	5	0.030902	0.014624	
total_past_incidents	0.020789	0.003690	0.000114	5	0.028386	0.013192	
total_misdemeanors_charges	0.017170	0.002246	0.000034	5	0.021794	0.012546	
past_possession	0.007608	0.002738	0.001708	5	0.013246	0.001970	
5_felonies_case	0.007510	0.001540	0.000201	5	0.010682	0.004339	
word_battery	0.007206	0.001864	0.000492	5	0.011043	0.003368	
past_felony	0.004983	0.001529	0.000941	5	0.008131	0.001836	
misdemeanors_case	0.004766	0.001639	0.001442	5	0.008140	0.001392	
2_misdemeanors_case	0.004309	0.003603	0.027775	5	0.011728	-0.003109	
word_license	0.004153	0.001462	0.001571	5	0.007163	0.001144	
total_other_cases	0.004146	0.000851	0.000201	5	0.005898	0.002395	
1_misdemeanors_case	0.003864	0.003668	0.039025	5	0.011416	-0.003688	
5_time_norm	0.003830	0.002261	0.009650	5	0.008484	-0.000825	
past_obstruct	0.003694	0.000734	0.000177	5	0.005205	0.002184	
word_trespass	0.003591	0.001404	0.002311	5	0.006481	0.000700	
3_other_cases	0.003567	0.001050	0.000806	5	0.005729	0.001404	
felonies_charge	0.003555	0.004030	0.059924	5	0.011853	-0.004743	
1_time_norm	0.003394	0.001949	0.008812	5	0.007406	-0.000618	
past_cocaine	0.003391	0.003125	0.036126	5	0.009826	-0.003043	
past_belt	0.003278	0.000829	0.000452	5	0.004985	0.001570	
2_time_norm	0.003102	0.001409	0.003954	5	0.006003	0.000201	
word_petit	0.002776	0.000809	0.000777	5	0.004443	0.001109	
4_total_prison	0.002726	0.001148	0.003028	5	0.005091	0.000361	
word_revoked	0.002716	0.000927	0.001406	5	0.004625	0.000806	

4_time_norm	0.002545	0.001809	0.017315	5	0.006269	-0.001179
past_operating	0.002444	0.001237	0.005764	5	0.004991	-0.000103
3_felonies_case	0.002143	0.001698	0.023846	5	0.005639	-0.001353
1_total_prison	0.002131	0.002636	0.072444	5	0.007559	-0.003296
past_dwelling	0.002125	0.000912	0.003237	5	0.004003	0.000247
past_petit	0.002119	0.001890	0.033168	5	0.006011	-0.001774

Model choice

The decisions of which data to include should be institutional. The deep learning ideal is more information is better, but if quality is suspect or information is found to be biasing you should definitely exclude it. If you find that perturbation of a column like race, leads to a difference in the outcome, you should exclude it. In UK data, race is self-reported and usually incomplete, and therefore to test your model you might have to subsample with that column[1]. You might also want to consider if there is a bias in the reporting of the column e.g. some groups of people may choose to avoid reporting their race to avoid bias therefore making a gap in the data a bias itself.

We are interested in the ranking task so that we can send top N candidates to our intervention clinic each week. To simulate this, we did a little experiment to calculate average precision of the top 10 retrieved suspects. Say that in the intervening time interval between clinics we get an average of 100 new cases, and we are interested in sending 10 people for referral because that is what the clinic has the capacity to process. We randomly sampled our test set 100 times (with replacement) and each time we took 100 examples out. We ranked the examples according to the predictor score and averaged that score over the number of repeated samplings.

Model	Average Score	African American	Caucasian	Hispanic
Without race	0.48	0.39	0.07	0.01
Without race + fairness	0.43	0.22	0.14	0.06
With race	0.53	0.45	0.06	0.01
With race + fairness	0.43	0.20	0.14	0.06

In this case we are looking this as **a ranking task**, the original model without fairness on average recommends 4.8 people who went on to commit a felony; however, it predominantly recommends African Americans. The model without fairness picks up 4.3/10 people correctly, but also recommends more equally across different members of the protected category we're interested in. We're assuming we want to choose the most equitable model.

If we are looking at a **classification task** we might choose one of the models with higher F-score or we might design our own metric that does the correct operation cost-benefit analysis that takes into account the cost of missing someone who committed a highly dangerous and damaging offence, the cost of various interventions, and other relevant factors. It is important that the choice of metric reflects the real costs to victims, society, and police. This may include considerations about how much data output from a model can be realistically processed and reviewed.

[1] If officer reported ethnicity is available you may consider using a combination of both.

We may note that the performance of these models looks very low but is higher than statistics of chance. Where we have randomly sampled 1s and 0s we get the performance of

	precision	recall	f1-score	support
0	0.50	0.73	0.60	3758
1	0.51	0.27	0.36	3747
accuracy			0.50	7505
macro avg	0.51	0.50	0.48	7505
weighted avg	0.51	0.50	0.48	7505

While the performance of the chosen model is

	precision	recall	f1-score	support
0	0.59	0.83	0.69	3920
1	0.66	0.37	0.48	3585
accuracy			0.61	7505
macro avg	0.63	0.60	0.58	7505
weighted avg	0.62	0.61	0.59	7505

Likewise, this is much higher than some of the performances reported on real UK police data, e.g. F=0.30 presented in the paper "Using machine learning to forecast domestic homicide via police data and super learning" as calculated from the precision and recall values in the table below:

Model	Recall	Precision	Specificity	AUC
Super learner	**0.7764**	**0.1861**	**0.6443**	**0.7104**

Documentation

- Enables other people to understand the dataset and the process through which the model was developed:
    - Data cards: supply a quick look at the properties of the dataset
    - Model cards: document processes so models can be compared and evaluated

Template Data Card

Metadata	
Senior Reporting Officer	
File name	compas_v1_all_1y-label.pqt compas_v1_all_1y-label_val.pqt compas_v1_all_1y-label_test.pqt compas_v1_all_1y-label_train.pqt

File format	Pandas data frame as a parquet file
Short summary of dataset purpose and content, and keywords	COMPAS data transformed for the purposes of training a model that predicts whether someone will commit a felony within the next year
Dataset size	38774 training examples with 860 features each
Version history	v1
Data governance plan	
% of missing values for relevant variables, nature of missingness, causes (if known) and how missing data has been handled	398808/33345640 = 0.012 or 1.2% null values mainly due to some data points not having up to 5 previous crimes. Because we were using an ensamble learning pipeline the null values were left as NaN and the data preprocessing pipeline processed the data according to the best practices for each model
Data Sources	
List sources of data and why each was included	COMPAS data : https://criminaljustice-datasheets.github.io/criminal%20charges%20&%20court%20records/compas/ https://github.com/propublica/compas-analysis?tab=readme-ov-file
Variables	• Person data: person_id, age (at the time of current arrest), race, sex, j_felonies, j_misdemeanor, j_other, • Current arrest information: felonies_case, misdemeanors_case, other_cases, • Charges stemming from current arrest: felonies_charge, misdemeanors_charge, other_charges, word_* (selected from the charge text of the current charges) • Predictive label: pred_label (whether someone commited a felony in the year after this current arrest) • Summary of previous charges: total_felonies_charges, total_misdemeanors_charges, total_other_cases,

	• Five previous cases each having the following columns: 1_felonies_case, 1_misdemeanors_case, 1_other_cases, 1_total_prison, 1_total_jail, 1_time_norm Total prison and jail times are calculated as the total prison time after this charge and before the next charge, • Past_*: selected words from the charge descriptions of the aggregate of 5 previous arrests. These are represented as bag of words style vectors.
Description of each variable (column) in the dataset	• j_* is the count of junior infractions grouped into felonies, misdemeanours, and other directly from the person table. • Total_* is the count of the number of felony, misdemeanour, and other charges that we have in the whole criminal history. When you group by person and then by time or case number it becomes clear that each time someone is arrested, they are charged with several offenses of different levels. • *_case variables refer to values in the current case (the one we're considering "time now" • *_charge variables refer to the charges table lines that correspond to the "time_now" cases • n_* where n is some number between 1 and number of total past cases you want to consider • Time norm is a decayed value of the number of days from the past charges to "time now" calculated in python using the expression 0.5^(diff_days/365) • In this example, we consider 5 past cases only • We do a small amount of word processing to separate-out unique tokens from the charge descriptions and choose words that occur more than 5 times across the dataset, removing some high frequency connectives and symbols. We make two sets of token vectors, ones describing the current charges and

	the ones describing all of the past charges. Ideally you would want to normalise word vectors separately row-wise, so that you get the influence of each word per person rather than across people, but we leave raw counts as are in this case.
Data Preparation	
How was the data sampled from the source data and how was it split into training/validation/test sets?	• We consider each time someone was arrested as a possible trigger for the model. The time of the arrest is then considered the "time now" and any crimes occurring before that time constitute past arrest history. This means that the same person can occur several times in the training data. • We split the data into train, validation, and test by taking the list of unique person IDs and randomly splitting them into 20% for test and 80% for training. From the training data we reserve a quarter for validation, effectively leaving us 60% of the individuals for training, 20% for validation, and 20% for testing. We then choose all the data points associated with each person and put them into the corresponding partitions. In this way we ensure that we do not have examples of the same past incidents in the training and test data. It is important we do that because we use each individual's past case data to generate several data points.
Describe what data cleaning took place e.g., cases removed, discretisation, coding, changes etc.	• the time_norm variable was constructed from the number of days between • Total prison and jail times for each arrest are calculated by adding time after the current arrest and before the next arrest • Words from the charge descriptions are chosen using the following formula: exclude these words: `['an','of', 'to', 'without', 'while', 'with', 'on', 'none', 'at', 'by', 'as', 'the', 'or', 'in', 'over', 'and', 'nd', 'too', 'one', 'two', 'when', 'for']` exclude any words that occur fewer

	than 5 times and have a single character
Describe data labels and what criteria were used for each?	There are two labels 1 if someone committed a felony within a year of an arrest 0 if they did not Any cases that were followed by a prison term were excluded
How was missingness assessed?	Missing values were left in as NaNs
Evaluation Data	
What data was used to train and test the model? Why?	We used COMPAS data to train and test this illustrative example.

TEMPLATE MODEL CARD

Model Details	
Senior Responsible Officer	
File name	autogluon-model-v1-norace / models are in python pickle files inside the directory
Date & Model Version	Data version: v1 Model version: v1
Type of model (e.g., random forest, neural net)	Weighted ensemble model of Neueural Net Torch Neural Net FastAI Cat Boost and Random Forest Trained by AutoGluon
Intended Use (can be taken from business case)	
Aim	The model aims to identify those who will commit a felony within 12 months of being arrested. This is to identify suitable candidates for an intervention that reduces serious offending.
Who will use the output	The output will be used by the intervention team to help them assess suitable candidates for the intervention
Appropriate & Inappropriate case examples	Appropriate – suspects arrested in Florida Inappropriate – suspects either not arrested or arrested outside of Florida
Results	

What groups are in the data? Why have these specific groups been chosen and how were they defined?	The protected groups in the data are gender and race based. In this data there are many more men than women and many more African Americans than other reported races: Caucasian, Hispanic, Asian, Native American. There is also a higher proportion of African Americans more than listed in population statistics of Florida for the period the dataset covers.
Are there any known groups missing from the data? Why might this be?	Unknown
Unitary results (i.e., by group), including uncertainty values	Precision Recall roc_auc Overall 0.367342 0.689877 0.681215 african-american 0.416452 0.677352 0.674806 asian 0.272727 1.000000 0.833333 caucasian 0.301980 0.694761 0.692764 hispanic 0.279605 0.765766 0.682453 native american 0.368421 1.000000 0.827381 other 0.200000 0.833333 0.765992
Intersectional results (i.e., where groups are combined), including uncertainty values	original updated Accuracy 0.722585 0.595736 F1 score 0.348561 0.479410 MCC 0.199398 0.222686 Prec. 0.475662 0.367342 Recall 0.275062 0.689877 roc_auc 0.696832 0.681215
Error Analysis	
What groups are in the data for which model performance might vary? (e.g., ethnicity, location, crime type)	Ethnicity, location (disregarded for the purposes of this study), crime type, gender
Have you reported differences for all relevant features? If not, why	In this study we considered only ethnicity for illustrative purposes
What has been done to mitigate any bias?	Fairness rebalancing was employed using demographic parity as a measure
Metrics	
What are the model's performance measures, including why these were selected as appropriate (e.g., false	The final choice of model was through a simulation of the operational environment where we expect to recommend 10 people for an intervention out of projected 100

positive/negatives, distribution differences across groups)	cases per time period between interventions. The model choice correlates with the analysis of models with traditional measures.
What are the decision thresholds (if relevant) including what they are and why they were chosen	We use top 10 people returned as a ranked list using model score as the ranking value
How were uncertainty and variability calculated? (e.g., variance, confidence intervals)	Not reported in this example but the designed test could be used to estimate standard deviation in the estimated results.
Risks	
What risks might be present in model use (e.g., recipients, likelihood and magnitude of possible harm)	As we are working to evaluate preventative measures, the risk is that someone does not get access to a possibly effective intervention that could reduce their criminal behaviour before they commit another crime.
What risk mitigation strategies were used during development?	We will monitor model performance and intervention effectiveness. We will assess the performance of the model and intervention pairing and increase or decrease the program as needed.
Maintenance	
Any known or expected changes relating to the population or how data were collected which may contribute to data shifts over time?	The model will be retrained in the future with the information about the intervention included.
What metrics will trigger an update and at what thresholds*? By whom and how will changes be communicated?	After each intervention clinic the data science team will record the expert opinions on algorithm ranking and use these as proxy scores. They will also keep track of the race, gender, location statistics report if there is a skew in the data that significantly differs from projected model performance and general population statistics. At each intervention officers will also bring cases and therefore we can compare model chosen cases with officer cases and monitor the performance of both methods, giving us a comparison of manual vs automated methods. In six monthly intervals there will be a review with the multidisciplinary team. At the year and a half mark there will be 6 months of intervention data which can be examined in line with the "felony after a year label".

Note: *thresholds are not relevant to time series or spatio-temporal models*

Implementation

Responsible team members: The whole team

Implementation decisions

Who will access the model results?	The data science team will run the model and produce a report for each meeting of the intervention team.
What training is provided to staff accessing model results? What provision is in place to ensure relevant staff get this training?	In this case access will be tightly controlled by the data science team. The intervention team will be given training on the output and model limitations. This will be online training available to all new starts and a refresher course taken once a year.
At what point in the investigative process will staff access the model results?	The staff will be given results of the model at an intervention triage session, without numbers but just as a list of 10 possible intervention candidates, which have to be reviewed manually.
How will model results be presented?	As a ranked list of candidates for the intervention
How will model limitations be presented?	The intervention panel will be aware that some of the candidates may not have been considered as potential felons, but it is up to them to determine who is suitable for the intervention.
When should predictions be acted upon? What should staff do?	Each person should be considered individually at the intervention triage session
Are there occasions where predictions should not be acted on? If so, when?	This depends on the intervention triage committee
How do staff document why they have either agreed/disagreed with the model's predictions? Can they document contradictory evidence?	The staff should note whether they think this individual is in danger of committing a violent crime in the near future, and whether they are recommending them for the intervention. It is essential that this data is collected as it will be incorporated in model evaluation and further training.
What additional resources will be needed to act on the model's predictions?	The intervention triage committee is essential to this process. The model needs to be reconsidered and if the committee and the intervention cease.
Are there processes in place for when the model is clearly making incorrect or biased decisions?	The model will be retrained, and the committee will continue to go with officer-led recommendations.

Model Maintenance

What metrics will be used to track model performance and why have these been chosen?	After each intervention clinic the data science team will record the expert opinions on algorithm ranking and use these as proxy scores. They will also keep track of the race, gender, location statistics report if there is a skew in the data that significantly differs from projected model performance and general population statistics. At each intervention officers will also bring cases and therefore we can compare model chosen cases with officer cases and monitor the performance of both methods, giving us a comparison of manual vs automated methods.
What thresholds for each metric will trigger an alarm and why have these been chosen?	At the six-monthly meetings the multidisciplinary team will evaluate the estimated model performance. If the triage committee is not finding suitable candidates or the model is underperforming compared to the officer cohort, it will be reevaluated. If the model statistics with regards to the protected classes deviate outside the acceptable range.(specify some range possibly related to the standard deviation of the simulation test
How many thresholds need to be breached for the model's performance to be reassessed? Why?	If there are distributional issues flagged at 6 monthly meetings. We don't have explicit thresholds set for this example.
How will the model's performance be reassessed?	The data science team will integrate statistics from the triage clinic and will present them on a tracking dashboard, which will also show underlying statistics for the cohort including protected characteristics. This will allow the multidisciplinary team to engage easily with the monitoring. In six monthly intervals there will be a review with the multidisciplinary team. At the year and a half mark there will be 6 months of intervention data which can be examined in line with the "felony after a year label".
What will happen if the model's performance needs to be reassessed? (e.g., stop using the model entirely? warn users the output may be faulty?)	Back off to officer-made suggestions, allow officers to suggest more people.
What additional resources will be needed to maintain the model?	This model requires minimal computational power but will require updated data, and continued time commitments from all members of the

	interdisciplinary team, as well as the intervention triage committee.
What is the process in place for assessing the impact on different community groups?	There are members of community groups on the multidisciplinary committee, and they have suggested the following measures...

Note: *thresholds are excluded for time series or spatio-temporal models*