Carnegie Mellon University

Prioritizing Mental Health Outreach to Reduce Recidivism

Machine Learning for Public Policy Lab

14 Dec 2022

Jameson Carter
Sebastian Dodt
Sarah Nance

Table of Contents

Executive Summary **3**

Background and Introduction **3**

Related Work **4**

Problem formulation **5**
- Cohort Definition 5
- Target Variable Definition 5
- Goals 6

Data Description **7**

Solution **7**
- Methods 7
- Tools 8
- Analysis 8
- Model Types 8
- Hyperparameters 9
- Features 9

Evaluation **9**
- Results 9

Discussion **13**

Deployment **15**

Policy Recommendations **16**

Limitations **16**
- Analysis Caveats 16
- Data Caveats 17
- Future work 17

Bibliography **18**

Table of figures **20**

Appendix **21**
- Triage configurations file 21
- Temporal Validation Splits 22
- Temporal performance graph for all final models 24
- Criteria Used to Select Top Models 25
- PRK Curves, Feature Importance, Crosstabs of Top 5 Models and Baseline 26
- Feature importance tables 48

Executive Summary

Jails have increasingly operated as care providers for individuals with mental illness who may be at risk of falling into a cycle of illness-induced re-incarceration. Johnson County officials are concerned about this pattern manifesting in their jail system; our project aims to support the County's desire to provide proactive, preventative outreach to individuals trapped in that cycle. For this paper, we developed a machine learning-based solution that could feasibly be used to identify, at the beginning of a month, which 100 ex-inmates with mental health issues are most likely to return to the Johnson County jail within the subsequent year. We then map these implications of using this model onto the Johnson County use-case, describe how to validate the model, and make policy recommendations. Throughout the paper, we define and discuss essential tradeoffs between the equity, effectiveness, and efficiency of the model.

To develop this system, we used data from several services based in Johnson County, jail bookings and court case data as well as mental health and medical data. The data were de-identified to protect the identity of the individual. Because our data spans multiple decades, we used temporal blocked cross-validation, a training and validation method that trains models on past data and validates on future data. Using this method, we identified the model which might generalize the best to a real-world deployment setting where the future is unknown.

We trained more than 2,000 combinations of models and hyperparameters on 26 time-splits. To validate each model, we used precision among the 100 highest-scored individuals to assess its efficiency. Our best performing model, a random forest bagging model, was able to achieve an average of 61.7% precision. This represents an improvement of roughly 10 percentage points over a simple non-ML baseline. Secondly, we assess the recall of people of color and women to assess the equity goal of our model. Our best performing model has a better recall for people of color than white individuals; however, it performs worse for women than for men.

For future work on this project, we recommended a field trial that would be able to validate the assumptions of our model and test the effectiveness of the outreach that it is guiding. If the model performs well in the deployment phase, we suggest its implementation by mental health services in Johnson County and similar interventions in other counties.

Background and Introduction

Johnson County, Kansas has expressed concern that some of their residents are caught in a cycle of poor mental health and incarceration. These individuals have existing mental health problems, but the county jail is not capable of treating or handling the mental illness properly. After they are released, their illness still exists, and the cycle continues. In 2016, the jail implemented a short mental health survey for all being booked into the county jail and in the next year found that 27% of inmates exhibited signs of severe mental illness (Sullivan, 2018). Further, in examining the cycle of mental health and recidivism in Johnson County, individuals who have been to jail before July 2017 have a 21% chance of returning to jail if they have a mental health need whereas it's only a 13% chance if they don't have one.

Various studies have also explained the harmful relationship between mental health and incarceration. There is a causal relationship between longer jail time and more severe mental health needs (Gendreau et al., 1999). Additionally, increased jail time leads to an increased chance of committing a crime after release. Going to jail in general for any reason has negative ramifications for individuals. If someone has been convicted of a crime, they will likely struggle to find work

upon release, and if they do it is unlikely to be well-paid. As such, preventing any ex-convict's return to jail is incredibly important for their well-being.

The current system in Johnson County is largely reactive; they evaluate mental health and refer people to services while they are already in or on the way to jail. Instead, if the County was able to intervene before a re-arrest, they could connect individuals with treatment for their mental illness and perhaps insulate them from returning to jail altogether. In doing this we aim to be *efficient* in maximizing the number of individuals we provide outreach to who actually need help, *effective* in reducing recidivism rates among community members with both mental health needs and past criminal records, and *equitable* in distributing outreach attempts among people with different racial and gender identities. In achieving these goals, we could break the cycle of mental-illness induced re-incarceration and improve the lives of those who are trapped in it.

Related Work

The linkages between mental illness and re-incarceration are well-studied considering that local jails now provide more mental health treatment than state or private facilities (Morgan et al., 2012). Many researchers focus on interventions occurring in jails themselves. For example, one study found that inmates who receive mental health/substance abuse treatment in jail have larger gaps between arrests (Peters & Matthews, 2002). Another found that women in a mental health/trauma treatment program were less likely to be re-incarcerated (Zlotnick et al., 2009).

Other approaches consider how post-release services may prevent re-incarceration. A descriptive study of a behavioral health program in Pennsylvania found that participants were less likely to be re-incarcerated (Zubritsky et al., 2018). The authors argue that the host of services provided were designed to ease the transition back into the community. They suggest that by connecting inmates to community organizations which provide post-release services *before* release, these individuals engage in said services, which are beneficial to individuals suffering from mental illness (Durose et al., 2014).

However, some researchers argue that insufficient services provided during incarceration may not alleviate symptoms (Reingle Gonzalez & Connell, 2014). This criticism rightly points out that these studies, for the most part, do not identify a causal relationship between service utilization and outcomes. Additionally, jails differ widely across the country and results in one system may not generalize to another. These studies generally focus on participants who opted into programs, rather than targeting specific individuals for interventions, as we are here.

Though there appear to be few deployed machine learning (ML) systems targeting specific ex-inmates with mental illness, there are many which attempt to prevent recidivism more generally (Travaini et al., 2022). The Minnesota Department of Corrections has utilized a machine learning model to predict whether ex-inmates are at high risk of committing felonies, nonviolent crimes, sexual crimes, or violent crimes (Minnesota Department of Corrections, 2016). These individuals are prioritized for more intensive supervised release (Guckenburg et al., 2019). The model has been validated to outperform generic risk modeling scores (Duwe, 2021) and saves state funds (Duwe & Rocque, 2017), and the treatment was found to reduce re-offense among individuals with high risk of committing violent crimes (Duwe & McNeeley, 2021).

Our work uniquely builds on this landscape of ML systems which have been successful in predicting re-incarceration elsewhere, while targeting a population that has mostly been explored in descriptive studies. This new ML system has the potential to specifically identify individuals with mental illness who are most likely to be re-incarcerated. In following validation work, we

would be able to disentangle some of the causal links between services and outcome that have been assumed, but often not proven, within the literature. This could result in a system that is specially calibrated to maximize the impact of Johnson County's mental health services on individuals who are at the highest risk of re-incarceration.

Problem formulation

Cohort Definition

The core goal of our project is to identify those individuals trapped in a cycle of mental health issues that causes their return to prison. However, determining the causal reason for any arrest, particularly in relation to mental health, seems impossible given the tools and data at our disposal. Thus, we chose the proxy of predicting the return to jail *among* those ex-inmates that had shown indicators of mental health problems in the past five years. We determined that an individual had a mental health need if they were:

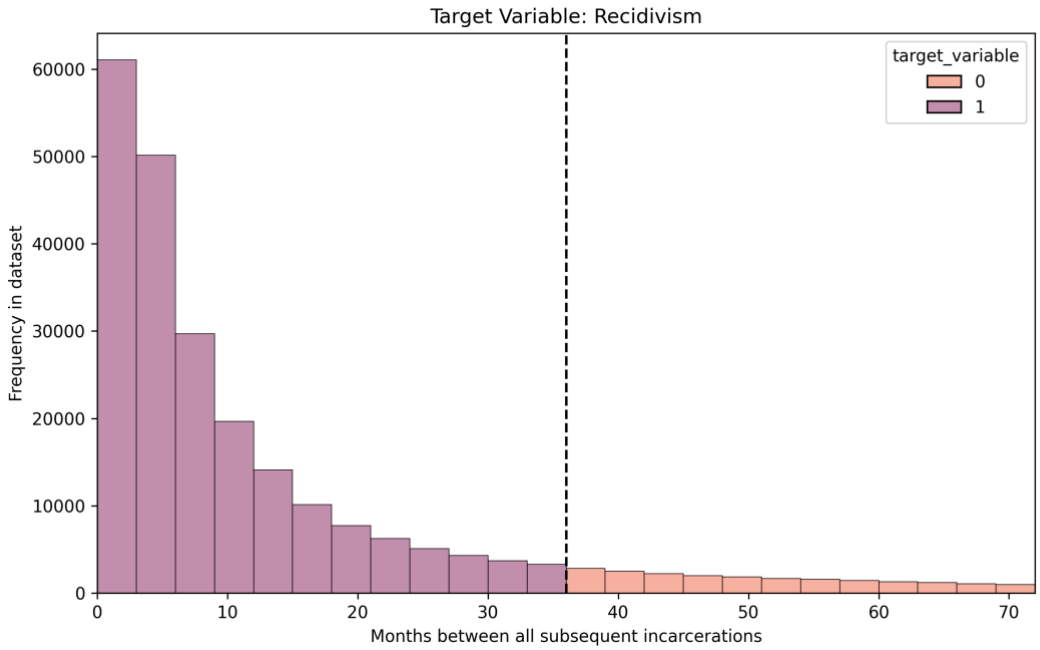
- Referred to mental services after completing a Brief Jail Mental Health Screening upon starting a jail sentence in Johnson County;
- Received mental health services by the Johnson County Mental Health Services other than intake appointments or minor notes to the individual's file;
- Flagged as mentally ill or suffering from substance abuse during a pre-trial assessment in Johnson County;
- Diagnosed with a mental illness by Johnson County Mental Health Services; or
- Had an emergency services encounter with Johnson County first responders (MED-ACT) that was flagged as related to mental health problems.

If an individual met one or more of the above conditions, we assumed that they had a mental health condition that could put them at increased risk of returning to jail up to five years after the mental health flag was noted. Lastly, we defined that an individual was in a "cycle" of jail sentences when they had a jail booking less than three years after their last booking. Thus, our cohort consisted of everyone who had demonstrated a mental health need in the last five years *and* was released from jail in the last three years. Furthermore, everyone in our cohort had to be currently free and a resident of Johnson County. In any given year between 2010 and 2019, these criteria applied to 2,500 to 6,000 people per year – with an increasing trend.

Target Variable Definition

Having defined the cohort to include everyone released from jail in the last three years, and filtered for those with mental health problems, we chose our target variable (or label) to be whether an individual returned to jail in the year after the respective cutoff date. In other words, our label was one if an individual who had been released in the three years before a given date was booked in jail in the year following. It was zero if the individual did not return to jail within said year. In the dataset provided to us by the Johnson County Justice Information Management Systems (JIMS), most individuals were booked only once (53%) in their lifetime. Whenever these individuals appear in our cohort, their target variable is zero. However, the remaining individuals were booked two or more times. The time span in-between their release and subsequent rebooking date is shown in Figure 1, as well as our cutoff time at 3 years.

Figure 1: Histogram of difference between date of release and date of subsequent booking



Goals

We designed our model in a way that it could be run at the beginning of every month to identify the individuals most in need of pre-emptive mental health outreach. In summary, our problem statement therefore was: At the beginning of each month, among everyone who has been released from jail in the last 3 years, is currently free, and demonstrated a mental health need, through assessments and medical incidents, or received mental health services/diagnoses over the past 5 years, we want to identify the 100 individuals at highest risk of getting booked back into jail within a year to proactively offer mental health services and prevent their return.

Through this formulation, we set a goal of maximizing the *efficiency* of the outreach that we are informing. To enforce this goal, we set the key metric that we used to check our model's performance as the precision it achieved among the 100 individuals that it predicted to be at the highest risk. This is to say that by maximizing precision, we are maximizing the proportion of all outreach calls which are made to people who will *actually* be re-incarcerated in the next year.

There are two additional goals that we are pursuing. First, we want to optimize *equity* which we defined as ensuring that our outreach is targeting an equitable share of inmates from protected groups, such as people of color and women. We are measuring this in two steps. Initially, when setting up the cohort – that is a subset of the Johnson County population as well as the county's jail population – we are testing whether the population that our model is choosing from is a fair representation of the base population. Subsequently, after our model has identified the individuals to reach out to, we aim to equalize recall among groups, such that protected groups do not have lower recall than majority groups.

Second, we want to maximize *effectiveness* of our outreach. Ideally, we would select individuals who would *actually* benefit from whatever outreach services Johnson County would utilize in this setting. We are not able to accurately estimate the effectiveness of our program in

reducing recidivism among our cohort while in the modelling phase. However, we are providing a blueprint of how we could test it in our description of the deployment phase on 15.

Data Description

Johnson County has provided us with all the data we used for this project. This includes information on the county jail, mental health center, services provided by the department of health, ambulance runs, and police arrests. In particular, the county jail data includes information on inmate jail bookings, demographics, charges, mental health needs, and the overall court case from the early 2000s through 2019. In this dataset there are roughly 180,000 inmates across 558,000 cases. Also, the Johnson County mental health center provides information on individuals who have interacted with them including their diagnoses, mental health outcome and treatment, demographics, indicators of socioeconomic status and the services they have received. This dataset covers 109,000 patients consistently from the 1990s through 2019. These are the two main sources of data for the project, but as previously mentioned are supplemented by additional information from Johnson County. All the data was joined across systems on individual IDs created by Johnson County data administrators

Our initial data explorations mainly focused on if an individual had recidivated at any point in time and various features we thought could be related to recidivism or mental health. In particular we looked at mental health indicators such as diagnoses, answers to the pre-trial assessment, suicide-homicide risk, and a referral for mental health services. Based on our results, several of these features formed the basis for our mental health criteria for the cohort. We also learned that just because an individual is booked into jail again does not inherently mean that they have committed a new crime. We used the booking type field to identify new charges, an instrumental element of defining our cohort and what recidivism means for us. Further we investigated age, race, and marital status. These explorations led to important features such as age at first booking because research shows that being booked into jail at a younger age increases the likelihood of reoffending (United States Sentencing Commission, 2017).

Solution

Methods

We are working with temporal data that on the one hand shows structural differences across time, and on the other, is correlated within a certain time step. Because we want to approximate our model's true error once it is deployed on future data, we cannot perform simple k-fold cross-validation to select our best model. Instead, we are hoping to best simulate the model performance if it was deployed in a real-world setting where all of the data that we have comes from the past and all of the labels that we predict on are in the future. To achieve this, we used a method called blocked cross-validation by Roberts et al. (2017) that focuses testing on spatially distant records and thus solves the issues of temporal auto-correlation.

Following the authors' best practices, we divided our data into 26 time blocks where each block ends at a certain cutoff date. The data before the cutoff date is aggregated and used as features for our model which are described in more detail on page 9 below. The year following the cutoff date is used to aggregate the labels, i.e., to determine whether an individual returned to jail. After training the data on the resulting training set, we tested the model on a training set, made up of unseen examples that came temporally *after* the last training set. In particular, we use the end date

of the last training labels as the new cutoff date. We aggregate across the time span before to produce our test feature values and use the year after the cutoff date as our test labels.

We configured this validation scheme to reflect the use-case which caseworkers would be faced with if this system were utilized by Johnson County. For example, we used individual dates to collect examples in the validation set. This is because caseworkers could feasibly run the model once at the beginning of the month and act on individuals identified by the model on that single day. We want to know how the model would perform at predicting re-incarceration for individuals identified in this way. Additionally, we chose 26 time splits because this reflects a scenario where the model is re-trained 3 times a year, over label data covering only those who were re-incarcerated over the past 10 years. This provided our team with the ability to train many features and models for individuals who interacted with the system more recently.

Tools

We defined our cohort, features, and labels in one pipeline using the Python library "Triage" developed at the University of Chicago's Center for Data Science and Public Policy and maintained and enhanced at Carnegie Mellon University. We used models from the scikit-learn package in Python as well as Aequitas developed by the University of Chicago's Center for Data Science and Public Policy. Our code can be found in this GitHub repository: https://github.com/dssg/mlpolicylab_fall22_mcrt1

Analysis

We defined the core metric of our model to be its precision across the 100 individuals that it ranked as having the highest likelihood of returning to jail among our cohort of ex-inmates with mental health problems. The precision is different for each model, each combination of its hyperparameters, and each of the time splits that it is trained and tested on. To decide on a model to use going forward, we calculated the performance of the models chosen by minimizing the regret from having chosen one particular model as the difference in precision from the best model of the respective time split. Specifically, we chose the model which most frequently exhibited a regret within 5% of the best model in each split. This model also happened to perform relatively well as a matter of average performance across all splits.

Precision is integral to reach our efficiency goal, both for the overall cohort as well as for the subsets of protected demographics. For instance, though we prioritize recall as our main equity concern, we also hoped to have a model that was at least as precise for the Black ex-inmate population as it was for the white population. We evaluated both the disparity in precision and also conducted a short evaluation of how our cohort selection affected our equity concerns. First, we checked whether our cohort was a good representation of the inmate population or whether there appeared to be biases in the flagging of mental health issues. Second, we checked whether among all individuals in our cohort, we were selecting an equitable share of individuals from all groups to receive outreach. To do this, we calculated the share of positives identified by our model among all those that were a positive (i.e., returned to jail).

Model Types

Given the classification task at hand, the main models we used were triage's scaled logistic regression, and sklearn's decision trees, random forest, and adaboost. Each of these models can work well for classification and are commonly used. We tested a naïve bayes model and quickly removed it due to consistent poor performance. Additionally, due to timing constraints, we did not

include gradient boosting although that would have been another good option. We also included dummy classifiers and strong baselines in our model space to compare the performance of complex machine learning models with easy methods that could be done in Excel. These baselines included rankings by the most recent release date, the most frequent bookings (all time and in the last five years), and most frequent services received from the JCMHC.

Hyperparameters

For each model type we also created a grid of hyperparameters to tune and create the most optimal model. For scaled logistic regression we included both l1 and l2 penalties and varied the C-value which indicates how much the model regularizes with smaller values indicating stronger regularization and larger values indicating none. For decision trees we focused on the maximum depth of each tree, the minimum number of samples required for a split, and the minimum number of samples for each leaf node. Random forest also utilized the max depth and min sample split hyperparameters in addition to the number of estimators (trees). Lastly, the adaboost model hyperparameters used the number of estimators and the learning rate.

Features

We created 2,635 different features to use in the models focusing primarily on demographics, jail booking information, assessments of mental health, calls to the mental health center, arrests, and charge details. Identifying ideas for features was based on past exploratory data analysis, research on the problem, and advice from our mentors. Following triage methodology, each feature is a specific variable aggregated in a certain manner over a certain time frame. Most of them have been aggregated over all time, 1 year, and 5 years. For specific instances such as demographics, we only aggregated over all time because these are unlikely to change often if at all. Most features had multiple aggregation options including a total for the past time frame (using count or sum depending on the data type) and the most recent value (max). In instances where it made sense, we did an average over the time frames as well. We were also often interested in the days since an incidence of something (ex: days since last booking, days since most recent release, etc.) and as a result took the minimum of the current as of date – the day of the incidence. Our triage configuration file contains the full list of all of our features, as well as the time frames and aggregations for each and is linked in the appendix below.

Evaluation

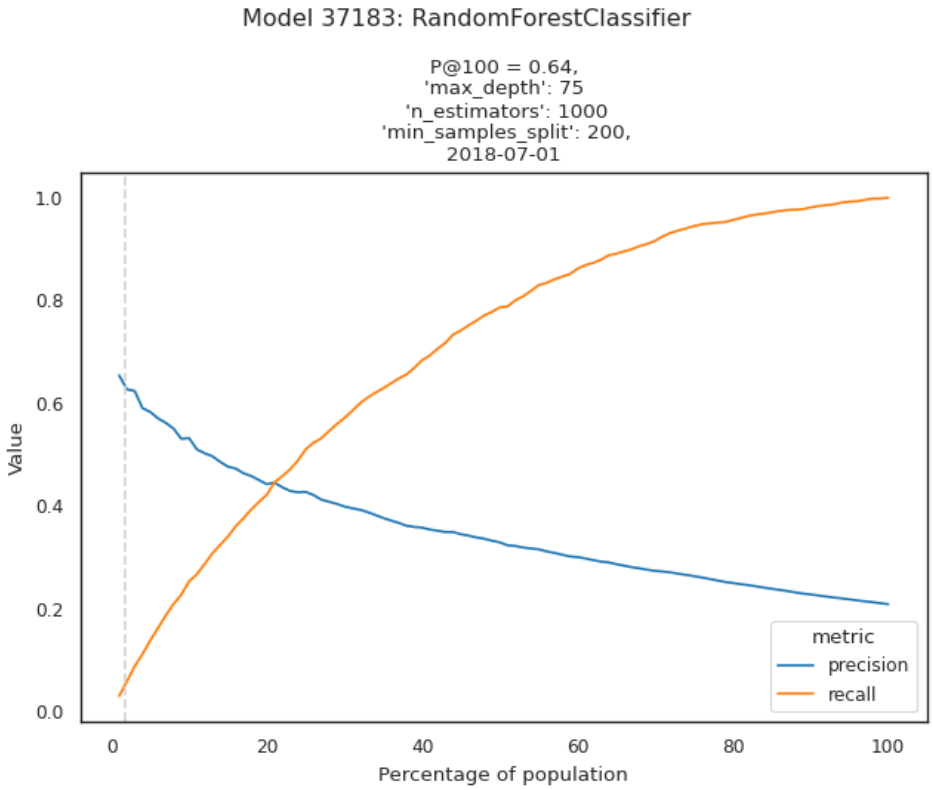
Results

Overall, the best model correctly selected individuals who return to jail ~61.7% of the time, on average across all test splits. The baseline correctly selected these individuals from the cohort ~51% of the time. This result displays that this model could help the County efficiently conduct outreach for individuals with mental illness needs who may return to jail. Instead of identifying roughly 50 such people per month using a non-ML solution, they could identify 62. This satisfies the efficiency goal stated at the beginning of this paper, though throughout this section we will explore how our other goals manifest in the results.

The degree of efficiency in this system is not static and rather depends on how many individuals the county hopes to intervene on. The County could reach more people in total if allocated more resources to target a larger percentage of the population, but the model would

identify people correctly less often. Figure 2 displays this tradeoff in the most recent validation set for the best model. The dashed line shows that currently, the people correctly chosen by our model would represent roughly 5.6% of the total population of those who are re-incarcerated. If the County wanted to reach 50% of the total population of those who are re-incarcerated, the model's precision would decline to roughly 40%. This implies that the county could serve half of the relevant population by intervening with roughly 1,500 people.

Figure 2: Increasing the number of people targeted decreases ability to precisely who will recidivate, but touches a larger number of people 'in need'



Additionally, these gains in efficiency may come at the expense of our effectiveness and equity goals. Although effectiveness is more difficult to judge until a field trial can be conducted which maps the intervention to its impacts, equity is a bit easier to examine preemptively. This section goes on to ask the following questions:

1. Who is being selected by our model vs. the baseline? How do these results reflect our concerns for equity via recall?
2. What are the important features driving results?

Who is Being Selected by the Model?

It seems like our model is selecting individuals who might be ensnared in the cycle of re-incarceration which justified this project. Table 1 shows that individuals selected by the model are admitted to the JCMHC more often, are booked into jail more often, are more likely to solicit substance abuse services from the JCMHC, and are more likely to have a history of homelessness.

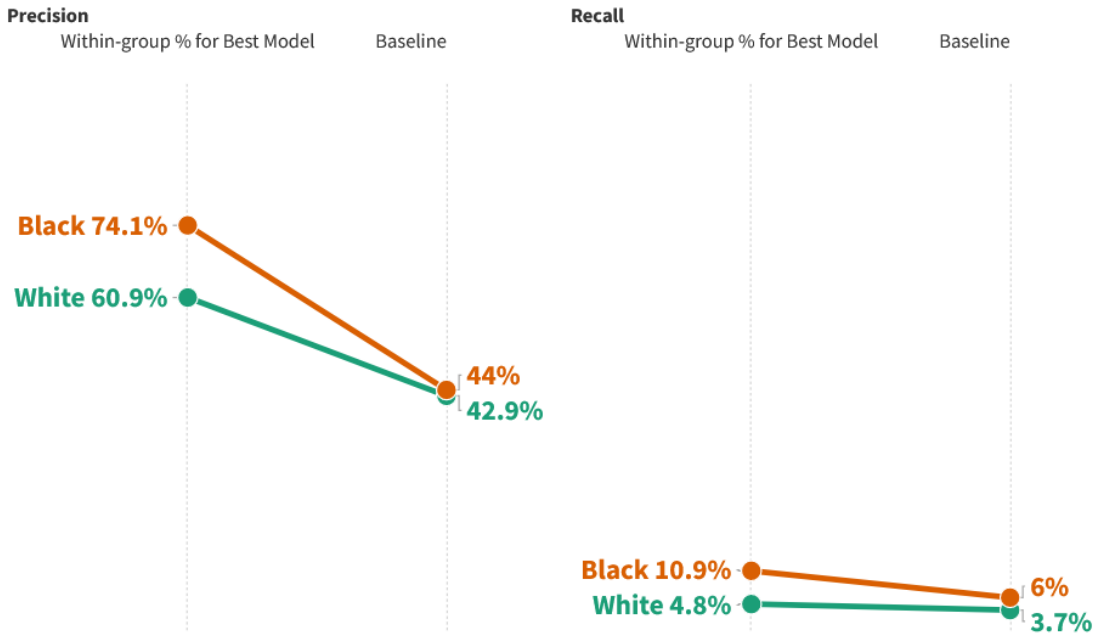
Additionally, they have made more crisis calls to the hotline and were younger when they were first booked. Although a field trial will be necessary to see whether serving this population of individuals would satisfy our effectiveness goal, it is at least promising to see that the model is selecting groups of people who could feasibly be trapped in the cycle.

Table 1: Individuals selected by the model differ from those not selected by the model

Metric, across all test splits	**Selected By Model**	**Not Selected by Model**
% Substance abuse history	64.8%	43.3%
% Homelessness history	9.3%	1.4%
Average admits to JCMHC	18	7
Average bookings	22	7
Average age at first booking	22	28
Average hotline crisis calls	10	3

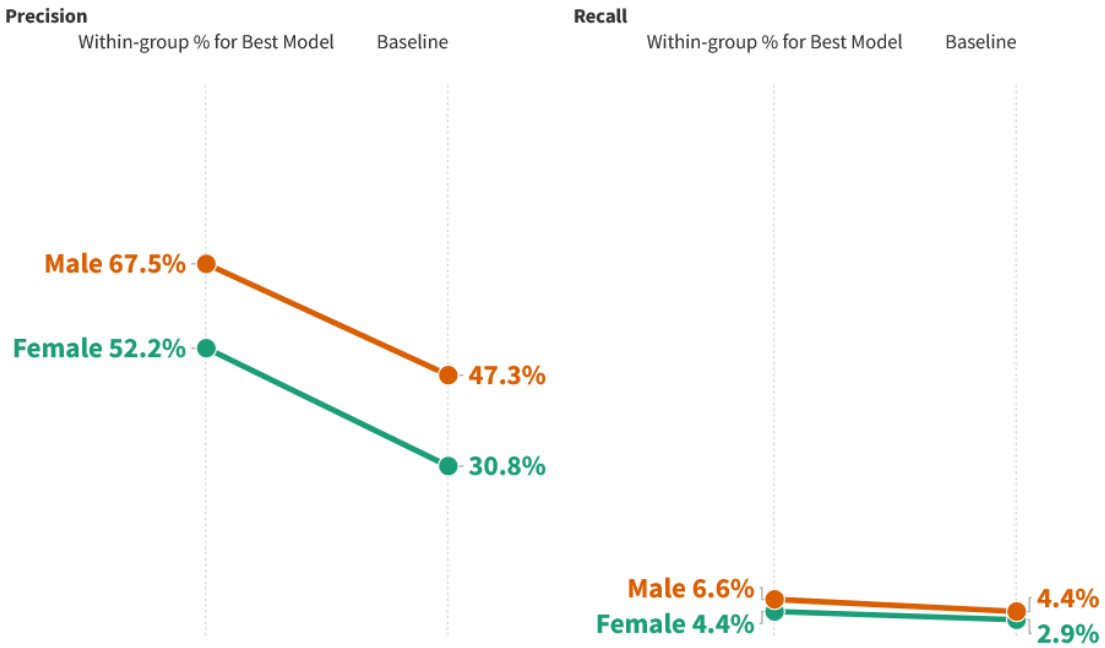
Regarding our equity goals, the model performed disparately across Black and White individuals. Figure 3's left panel shows that our model was more likely to be correct for Black individuals, and further that this disparity was increased as compared to the baseline (which had almost no disparity). Additionally, the right panel shows that the model had more favorable recall for Black vs. White individuals among the cohort. Our model also performed disparately across men and women. Figure 4 displays that there is a lasting disparity for both the baseline and the best model in both precision and recall. Men are more likely to be correctly selected, and a larger proportion of all re-incarcerated men are selected by the model.

Figure 3: Best model is more precise and has higher recall for Black individuals than White, increases disparity of precision and recall over the baseline



Note: Scale is from 0-100%. Calculated over the last test split.

Figure 4: Best model is more precise and has higher recall for male individuals than female, with similar disparities between precision and recall when comparing to the baseline

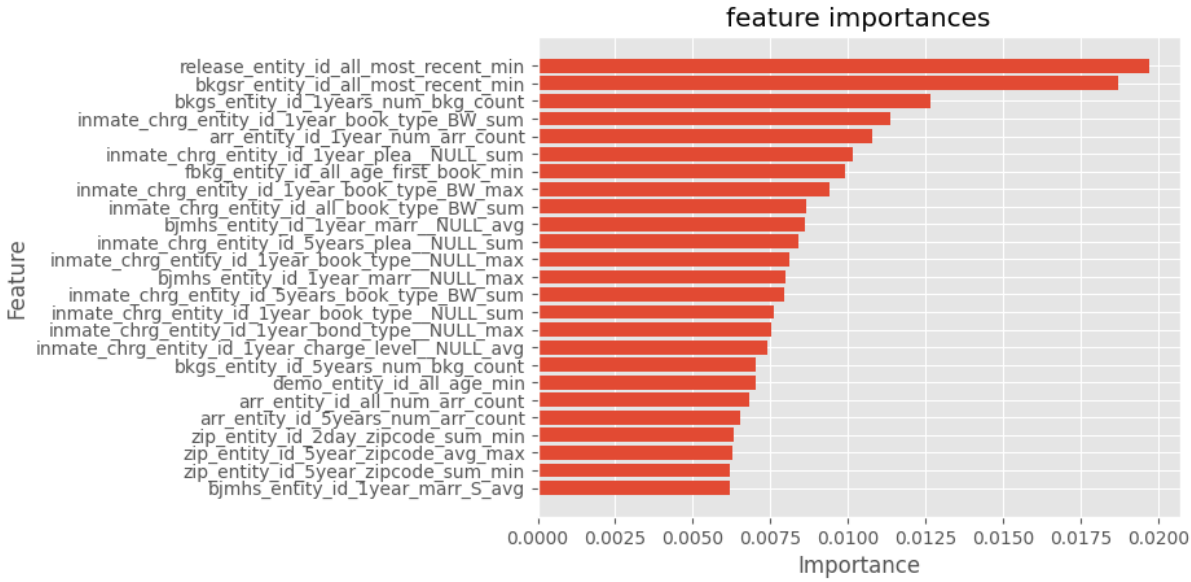


Note: Scale is from 0-100%. Calculated over the last test split.

What Features are Driving these Results?

The chart below shows that the top 25 most important variables track who was released most recently, who was booked most recently, the number of bookings over the past year, the number of bench warrants over the past year, and the number of arrests. Additionally, the age of a member of the cohort at first booking (and at a given time) is important. A series of important variables tracking booking types, bond types, and plea types are null, potentially indicating some sort of systematic reason for these codings which warrant further investigation. Our explorations of re-incarceration rates at the zip code level turned out to be important for the model, meaning that community levels of re-incarceration are associated with a person's chances of becoming re-incarcerated themselves.

Figure 5: History with criminal justice system, age at first booking, and community levels of re-incarceration drove model results



Discussion

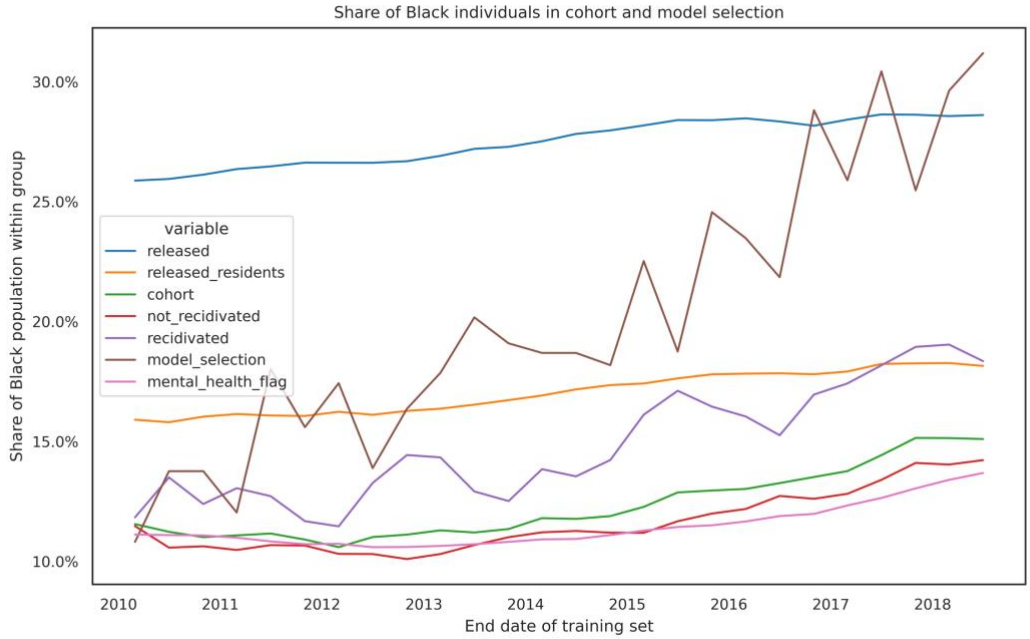
Our model appears to perform well across all of our key metrics. The precision of the results indicates an efficiency that we had hoped our model would achieve in guiding the outreach of the mental health services. An average precision of 61.7% is acceptable in relation to the base rate of around 20% -- that is, the fraction for which the cohort had positive labels. Our best baseline is performing strong with around 51% precision, yet our best model does appear to outperform it significantly. The results further show that many features we felt might be relevant were useful for the model: age at first release, history with the criminal justice system, JCMHC admissions, and substance abuse service utilization all seemed relevant.

In terms of our second goal, equity, measured in terms of the representation of groups in the final model selection being proportional to the overall cohort our model also appears to do relatively well. However, here the results are somewhat blurrier and further on-the-ground testing is needed. Looking at race, the recall performance of our model is encouraging as it appears that

the best performing model selected a larger share of Black individuals that are going to return to jail than other ethnicities. However, this does not show the full picture because our cohort is not unbiased. The overall share of inmates released in the last three years that are Black appears to be stay relatively consistently between 26% and 29% across the time splits (see top, blue line in Figure 6). When we filter by residents, the share of Black individuals drops to 16% to 17% (orange line). This could mean, on the one hand, that incarcerated Black individuals are less likely to be from Johnson County – which is a predominantly white area.

As of 2021, only 5.9% of Johnson County residents are Black (US Census Bureau, 2021). The cohort of all residents released in the last three years is further narrowed by our filter by mental health problems in the past five years. It appears that Black ex-inmates are far less likely to have a recorded mental health problem in our data (10% to 13%, pink line). As national studies have shown that people of color are more likely to suffer from a mental health issue (Vance, 2019a), this may be a result of misdiagnosing or of inaccessible mental health services. Thus, our cohort is made up of around 11 to 15 percent Black individuals (green line) which is much lower than the total jail population in Johnson County and the subset of ex-inmates that are residents. Nevertheless, our model seems to select Black individuals at a high rate (brown line) with an increasing trend over time. Indeed, in the last time split, the share of Black individuals among those selected was higher than 30%, i.e., higher than the share of Black individuals in jail. This may be due to the higher recidivism rate in this subset of the population (purple versus red line).

Figure 6: Share of Black individuals in various cohorts



There is also a clear gender disparity in our best model. However, the recall for both men and women is quite low. The poor recall is possibly due to the imbalance between men and women in our dataset. Of all unique individuals who have ever been booked into the Johnson County jail, 71.94% are men. Our model also predicts more accurately for men than women which could be the results of our cohort as men are more likely to go prison.

Deployment

Before deploying our model, it should be tested in a field trial. This helps us to validate our assumptions and our performance across our three goals.

The most crucial assumption of our model is that we define mental health problems as having one of several indicators from different data sources. We should verify whether this assumption holds through a field test. Through said test, we could validate more reliable sources, eliminate less reliable sources, and possibly add new sources. We could do this by approaching a subset of our cohort that includes individuals flagged through different programs, services, and forms, and investigate whether these individuals are truly suffering from mental health conditions. If a significant amount is not suffering from mental health issues, we may want to trace how they have been flagged. If there's a source for flagging a disproportional share of false positives, we would eliminate the source. If performance is bad across all sources, we would have to revisit our definition. On the other hand, we may also want to approach individuals that were not added into our cohort because they did not exhibit any mental health criteria. We could check whether these individuals are truly not mentally ill, and for those who are, find solutions how their illness may be recognized better and incorporated into future models.

To test the performance across our three goals, efficiency, equity, and effectiveness, we will use a randomized control trial (RCT) that works on several levels. For our RCT to work, our model has to be deployed almost as it would be in the real setting. Date pipelines between the services will have to be created so that the model can be run with up-to-date data at the beginning of each month. Whether this is possible will be a first crucial test for our models' validity. After our model has identified the top 100 individuals that it deems as having the highest risk of re-incarceration in the Johnson County jail, we will attempt to locate the individuals. Whether the individuals can be found in phone books, service records, or stored addresses is a second test for our process. We then divide the 100 individuals into two random groups, a control group in which we do not perform any outreach and a treatment group.

To test our efficiency goal, we check whether the precision of our model is as accurate in real life as it had been in our lab. To calculate this, we take the percentage of those returning to jail in the year after identification among all those who were in our control group. We will also do cross-tabulations to ensure its precision does not show discrepancies across different demographics. Secondly, to test our equity goal, we calculate the recall across different groups. For this we are taking the percentage of those who are returning to jail in our control group among all those in our cohort who returned to jail, and group by each demographic group (race and gender). Lastly, we check our effectiveness goal by calculating the effect of our outreach on recidivism rates. If the percentage of those in our control group who are returning to jail is statistically significantly higher than in our treatment group, we can conclude that (a) our outreach is effective and (b) our model is selecting individuals for whom the outreach can be effective. We should consider running the trial for a year and identify a control and treatment group every month. Note that this requires careful consideration of duplicate individuals that are flagged in multiple months, a consideration that is not present in the model currently.

Policy Recommendations

In addition to supporting the above field trial, we recommend reviewing the programs offered by the Johnson County Mental Health Center in order to fully understand each program's purpose and optimize the type of outreach that could be suggested for each individual. This review should also include an assessment of the accessibility of each program. If members of communities lack awareness of services or express concerns about their usefulness, their concerns should be addressed as quickly as possible given the center's constraints. In doing so this will ensure that individuals with mental health needs are recommended to the services that will best benefit them during the field trial and potential future deployment of this program.

For example, our results show that individuals selected for mental health outreach use substance abuse services more. As a result, this review of programs should emphasize those at the JCMHC which help people with substance use disorders. More generally, it would probably be wise to increase funding for these programs: we recommend hiring more therapists, doctors, and nurses so more people can be treated and focusing on outreach in Black communities as they are less likely to be receiving substance abuse treatment (Grooms and Ortega, 2022). Similarly, we feel that the County should review and identify which programs are being used successfully by individuals who become implicated in the system at a young age.

Lastly, we would also like to encourage each entity involved in this project (JCMHC, Johnson County Police, etc.) to employ better data practices as the data was sometimes difficult or unusable. For example, as mentioned previously, many of our most influential features were actually null values for a specific variable. Knowing exactly what these null values mean would be incredibly useful for our analysis, and if a null value can have multiple meanings, setting data practices to separate them is a good idea.

Limitations

Before any real-world deployment, it is crucial for stakeholders to understand the caveats of using our model to inform their decision making.

Analysis Caveats

First, as emphasized on our problem formulation, we are predicting re-incarceration *among* mentally ill individuals, not re-incarceration *caused* by mental health illnesses. The patterns found between our features and labels may not represent causal relationships but correlations that have held over time.

Second, without real-world testing, the effectiveness of our proposed program is largely unknown. On the one hand, this is due to a potential difference in model performance in the lab and in its deployment – despite our best efforts to estimate the true error created by our model. On the other hand, our model is set to find those most likely to return to jail who may be or may not be the same individuals for whom the outreach is most effective in preventing incarceration.

Third, we made assumptions about mental health indicators and the "cycle of incarceration" that individuals are in. The temporal parameters set (i.e., 5 years for mental health need, and 3 years for re-incarceration) were chosen to focus our attention on those with the highest need, and that the need may be most accurate for those who demonstrated it more recently. Nevertheless, these parameters should be scrutinized and tested in the deployment phase.

Data Caveats

As we have shown, data is not the only source of bias, but it is a likely one. Real-world biases were part of our modelling process, particularly as we are dealing with two areas where real-world biases are prominent: the criminal justice system and mental health data. In the former, minorities have historically been overrepresented due to over-policing and structural societal inequalities. In the latter, studies have highlighted a set of structural biases that lead people of color to be less likely to take up mental health support and be less likely to be diagnosed with a mental health illness (Vance, 2019b). These inequalities were present in our dataset as Black individuals were almost six times as likely to be in jail compared to white individuals in Johnson County in 2018, while having a 45% lower chance of having been flagged as mentally ill. As discussed on page 8, this can create a bias issue that has to be addressed.

This leads to a second critical limitation. There is no conclusive indicator for mental health problems. Mental health flags are scattered across our data sources and were pulled in from various pre-sentence screenings as well as mental health center and first responder calls and encounters. One flag suffices, but all flags from all data sources are different. Some are more trustworthy, for example diagnoses from trained psychologists at the Johnson County Mental Health Center. Others come from non-medically trained jail employees or first responders, and some even from self-reported answers to questionnaires. We chose this diverse set of sources to reduce bias in our cohort resulting from filtering narrowly by only those formally diagnosed – a medical service that may be more accurate though less accessible for marginalized populations.

Lastly, we are lacking critical data by using only Johnson County jail data to define our labels. If an individual goes to prison – not jail – within Johnson County or any penitentiary institution outside of Johnson County after their discharge, we are not able to use that information. Because jail sentences are generally for less severe crimes than prison sentences, our model may be less accurate at identifying individuals who are going to commit severe crimes.

Future work

In order to increase the precision and reliability of our model, we should integrate data from outside Johnson County into our pipeline to track individuals' future sentences. This was not possible in the current iteration of the project as it would have violated our commitment to anonymization of individuals' sensitive data.

Additionally, we could further improve our model by adding other outside data to our feature space. Census data may provide some interesting socio-economic insights into our models. This may include but is not limited to zip code and tract-level data on the individuals' current residence. As individuals move and neighborhoods develop, this could help us capture geographic variables that are correlated with increased risk of arrests, either because an area seeing high crime rates, or is facing over policing through law enforcement, or both.

Lastly, we would like to implement a field trial that tests our assumptions and performance across our three goals. This field trial may require significant resources, yet it is integral to validating and improving our process. In the deployment section on page 15, we describe a year-long field trial that tests our assumptions and performance indicators through a randomized control trial. We recommend that this field trial is to be implemented swiftly to ensure that the individuals affected by the cycle of incarceration and mental health problems are receiving this needed help as soon as possible.

Bibliography

Durose, M. R., Cooper, A. D., & Snyder, H. N. (2014). Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010. *U.S. Department of Justice Office of Justice Programs Bureau of Justice Statistics*, 1–31.

Duwe, G. (2021). Evaluating Bias, Shrinkage and the Home-Field Advantage: Results from a Revalidation of the MnSTARR 2.0. *Minnesota Department of Corrections*.

Duwe, G., & McNeeley, S. (2021). The Effects of Intensive Postrelease Correctional Supervision on Recidivism: A Natural Experiment. *Criminal Justice Policy Review*, *32*(7), 740–763. https://doi.org/10.1177/0887403421998430

Duwe, G., & Rocque, M. (2017). Effects of Automating Recidivism Risk Assessment on Reliability, Predictive Validity, and Return on Investment (ROI): Recidivism Risk Assessment. *Criminology & Public Policy*, *16*(1), 235–269. https://doi.org/10.1111/1745-9133.12270

Gendreau, P., Goggin, C., & Cullen, F. T. (1999). *The effects of prison sentences on recidivism, 1999-3*. Solicitor General.

Guckenburg, S., Persson, H., Dodge, C., & Petrosino, A. (2019). Evaluation of the Minnesota Statewide Initiative to Reduce Recidivism (MNSIRR): A summary. *WestEd Justice & Prevention Research Center*.

Minnesota Department of Corrections. (2016). Minnesota Screening Tool Assessing Recidivism Risk 2.0 (MnSTARR 2.0). *Minnesota Department of Corrections*.

Morgan, R. D., Flora, D. B., Kroner, D. G., Mills, J. F., Varghese, F., & Steffan, J. S. (2012). Treating offenders with mental illness: A research synthesis. *Law and Human Behavior*, *36*(1), 37–50. https://doi.org/10.1037/h0093964

Peters, R., & Matthews, C. (2002). Jail Treatment for Drug Abusers. In C. G. Leukefeld, F. M. Tims, & D. Farabee (Eds.), *Treatment of drug offenders: Policies and issues*. Springer Pub.

Reingle Gonzalez, J. M., & Connell, N. M. (2014). Mental Health of Prisoners: Identifying Barriers to Mental Health Treatment and Medication Continuity. *American Journal of Public Health*, *104*(12), 2328–2333. https://doi.org/10.2105/AJPH.2014.302043

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Sullivan, R. (2018). Pratical Applications of the Brief Mental Health Screen: Johnson County, Kansas. *Samhsa's Gains Center*. https://www.prainc.com/bjmhs-johnson-co-ks/

Travaini, G. V., Pacchioni, F., Bellumore, S., Bosia, M., & De Micco, F. (2022). Machine Learning and Criminal Justice: A Systematic Review of Advanced Methodology for Recidivism Risk Prediction. *International Journal of Environmental Research and Public Health*, *19*(17), 10594. https://doi.org/10.3390/ijerph191710594

US Census Bureau. (2021). Quick Facts, Johnson County Kansas. *Census.Gov*. https://www.census.gov/quickfacts/johnsoncountykansas

Vance, T. A. (2019a). Addressing Mental Health in the Black Community. *Columbia Psychiatry*. https://www.columbiapsychiatry.org/news/addressing-mental-health-black-community

Vance, T. A. (2019b). Addressing Mental Health in the Black Community. *Columbia Psychiatry*. https://www.columbiapsychiatry.org/news/addressing-mental-health-black-community

Zlotnick, C., Johnson, J., & Najavits, L. M. (2009). Randomized Controlled Pilot Study of Cognitive-Behavioral Therapy in a Sample of Incarcerated Women With Substance Use Disorder and PTSD. *Behavior Therapy*, *40*(4), 325–336. https://doi.org/10.1016/j.beth.2008.09.004

Zubritsky, C., Wald, H., Jaquette, N., & Balestra, A. (2018). Breaking the Cycle of Recidivism: From In-Jail Behavioral Health Services to Community Support. *Journal of Criminology and Forensic Studies*, *1*(2), 1–6.

Table of figures

Figure 1: Histogram of difference between date of release and date of subsequent booking 6
Figure 2: Increasing the number of people targeted decreases ability to precisely who will recidivate, but touches a larger number of people 'in need' 10
Figure 3: Best model is more precise and has higher recall for Black individuals than White, increases disparity of precision and recall over the baseline 12
Figure 4: Best model is more precise and has higher recall for male individuals than female, with similar disparities between precision and recall when comparing to the baseline 12
Figure 5: History with criminal justice system, age at first booking, and community levels of re-incarceration drove model results 13
Figure 6: Share of Black individuals in various cohorts 14
Figure 7: Temporal Graph Tracking Precision @ 100 in the Validation Set Over Time 24
Figure 8: Model Group 2101 PRK 26
Figure 9: Model Group 2101 Feature Importance 27
Figure 10: Model Group 2120 PRK 28
Figure 11: Model Group 2120 Feature Importance 28
Figure 12: Model Group 2120 Bias 29
Figure 13: Model Group 2126 PRK 32
Figure 14: Model Group 2126 Feature Importance 32
Figure 15: Model Group 2126 Bias 33
Figure 16: Model Group 2106 PRK 36
Figure 17: Model Group 2106 Feature Importance 36
Figure 18: Model Group 2106 Bias 37
Figure 19: Model Group 2123 PRK 40
Figure 20: Model Group 2123 Feature Importance 40
Figure 21: Model Group 2123 Bias 41
Figure 22: Model Group 2107 PRK 44
Figure 23: Model Group 2107 Feature Importance 44
Figure 24: Model Group 2107 Bias 45

Table 1: Individuals selected by the model differ from those not selected by the model 11
Table 2: Temporal Training and Validation Sets Used in this Pipeline 22
Table 3: Model Group 2101 Crosstabs of last time split 27
Table 4: Model Group 2120 Crosstabs 31
Table 5: Model Group 2126 Crosstabs 35
Table 6: Model Group 2106 Crosstabs 39
Table 7: Model Group 2123 Crosstabs 43
Table 8: Model Group 2107 Crosstabs 47

Appendix

Triage configurations file

Link to our final triage configuration file:
https://github.com/dssg/mlpolicylab_fall22_mcrt1/blob/main/model/triage_config/triage_config.yaml

Our cohort is defined in a separate file and fed into our triage config through the file path:
https://github.com/dssg/mlpolicylab_fall22_mcrt1/blob/main/model/cohort_definitions/triage_cohort_def_new.sql

Our above cohort file is commented to aid in viewing our exact definition, which is as follows: a person is in the cohort on a given 'as-of date' if they are released from jail over the last 3 years from that date, they are currently out of jail, are residents of Johnson County, and have exhibited a mental health need over the past 5 years from that date. We consider a person to be re-incarcerated if they are re-booked into the jail over the next year from that same as of date, for a new booking type listed as 'WARR', 'SANCT', 'CITY', 'VIEW', or 'BW,' per documentation listing these booking types as relating to a new charge levied by the criminal justice system.

Temporal Validation Splits

Table 2 displays descriptions of the 26 training and validation pairs used to tune and test the model over time. For example, for the first row listed here we collected examples of members of our cohort spanning the course of five years, from July 2012 to July 2017, and checked whether those individuals were re-incarcerated. These checks for whether individuals were re-incarcerated are reflected in the label start and end date columns. For the members of the cohort viewed in July of 2012, for example, we check to see whether they are re-incarcerated at any point before July of 2013. We follow this same process for members viewed in July of 2017, looking at whether they are re-incarcerated up at any point before July of 2018.

To reflect the use case which caseworkers would be faced with if this system were utilized by Johnson County, we used individual dates to collect examples in the validation set. This is because caseworkers would run the model once at the beginning of the month and act on individuals identified by the model on that single day. We want to know how the model would perform at predicting re-incarceration for individuals identified in this way. That is why the start date and end date have the same date values as do the values for label start and end dates.

For each model, we ran the model across every one of these splits. This was computationally intensive: we eventually stopped collecting examples of re-incarceration in March of 2009. At this point, the value of the start date column for the train set stays fixed and the overall number of examples collected for those splits is reduced.

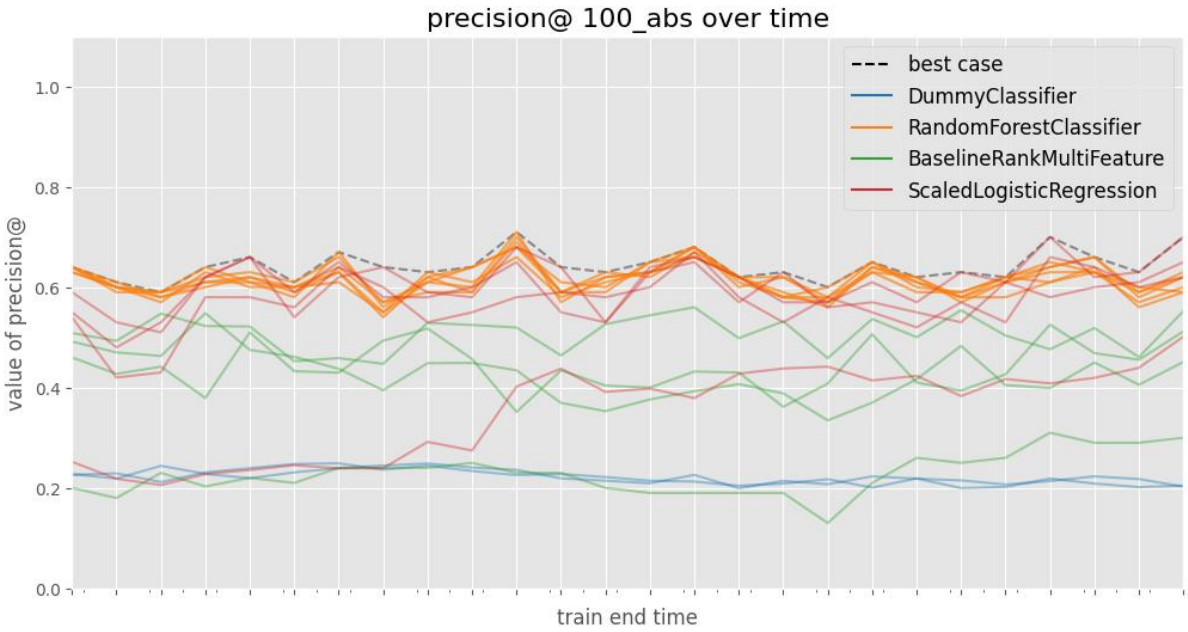
Table 2: Temporal Training and Validation Sets Used in this Pipeline

Train Set				Validation Set			
Start Date	**End Date**	**Label Start Date**	**Label End Date**	**Start Date**	**End Date**	**Label Start Date**	**Label End Date**
01-07-12	01-07-17	01-07-13	01-07-18	01-07-18	01-07-18	01-07-19	01-07-19
01-03-12	01-03-17	01-03-13	01-03-18	01-03-18	01-03-18	01-03-19	01-03-19
01-11-11	01-11-16	01-11-12	01-11-17	01-11-17	01-11-17	01-11-18	01-11-18
01-07-11	01-07-16	01-07-12	01-07-17	01-07-17	01-07-17	01-07-18	01-07-18
01-03-11	01-03-16	01-03-12	01-03-17	01-03-17	01-03-17	01-03-18	01-03-18
01-11-10	01-11-15	01-11-11	01-11-16	01-11-16	01-11-16	01-11-17	01-11-17
01-07-10	01-07-15	01-07-11	01-07-16	01-07-16	01-07-16	01-07-17	01-07-17
01-03-10	01-03-15	01-03-11	01-03-16	01-03-16	01-03-16	01-03-17	01-03-17

01-11-09	01-11-14	01-11-10	01-11-15	01-11-15	01-11-15	01-11-16	01-11-16
01-07-09	01-07-14	01-07-10	01-07-15	01-07-15	01-07-15	01-07-16	01-07-16
01-03-09	01-03-14	01-03-10	01-03-15	01-03-15	01-03-15	01-03-16	01-03-16
01-03-09	01-11-13	01-03-10	01-11-14	01-11-14	01-11-14	01-11-15	01-11-15
01-03-09	01-07-13	01-03-10	01-07-14	01-07-14	01-07-14	01-07-15	01-07-15
01-03-09	01-03-13	01-03-10	01-03-14	01-03-14	01-03-14	01-03-15	01-03-15
01-03-09	01-11-12	01-03-10	01-11-13	01-11-13	01-11-13	01-11-14	01-11-14
01-03-09	01-07-12	01-03-10	01-07-13	01-07-13	01-07-13	01-07-14	01-07-14
01-03-09	01-03-12	01-03-10	01-03-13	01-03-13	01-03-13	01-03-14	01-03-14
01-03-09	01-11-11	01-03-10	01-11-12	01-11-12	01-11-12	01-11-13	01-11-13
01-03-09	01-07-11	01-03-10	01-07-12	01-07-12	01-07-12	01-07-13	01-07-13
01-03-09	01-03-11	01-03-10	01-03-12	01-03-12	01-03-12	01-03-13	01-03-13
01-03-09	01-11-10	01-03-10	01-11-11	01-11-11	01-11-11	01-11-12	01-11-12
01-03-09	01-07-10	01-03-10	01-07-11	01-07-11	01-07-11	01-07-12	01-07-12
01-03-09	01-03-10	01-03-10	01-03-11	01-03-11	01-03-11	01-03-12	01-03-12
01-03-09	01-11-09	01-03-10	01-11-10	01-11-10	01-11-10	01-11-11	01-11-11
01-03-09	01-07-09	01-03-10	01-07-10	01-07-10	01-07-10	01-07-11	01-07-11
01-03-09	01-03-09	01-03-10	01-03-10	01-03-10	01-03-10	01-03-11	01-03-11

Temporal performance graph for all final models

Figure 7: Temporal Graph Tracking Precision @ 100 in the Validation Set Over Time



Criteria Used to Select Top Models

As shown in Figure 7**,** the splits mentioned in Table 2 performed variably for each model examined in our pipeline. Therefore, we had to make some decisions about how we selected a best model. To facilitate this, we first filtered out models which were ever more than 10% away from the best possible model in a given time split and removed models which every went below 50% precision since we knew that our best baseline was performing at slightly above 50% precision @100. This winnowed the potential best models down to 13 potential candidate models, consisting of 12 random forests and 1 logistic regression.

To determine the best final model, we consulted a series of decisions to reflect our priorities. Namely, we wanted a model which had low regret, i.e., was most frequently within a small distance of the best performing model. Secondarily, we wanted a model which performed well in the most recent test set. We felt these two concepts should be considered simultaneously because a model could theoretically frequently exhibit low regret over the 26 splits before decaying in performance in recent data. While it's uncertain whether that sort of trend exists in the data, we operate under the heuristic that recent performance matters for generalizing to present data.

We therefore tested whether models were within 5% of the best model across all splits and got the following list:

1. Model Group 2120: Random Forest with 75 max depth, 1000 estimators, 200 minimum sample split and 61.7% precision over all splits.
2. Model Group 2126: Random Forest with 300 max depth, 1000 estimators, 200 minimum sample split and 61.7% precision over all splits.
3. Model Group 2106: Random Forest with 75 max depth, 1000 estimators, 100 minimum sample split and 61% precision across all splits.
4. Model Group 2123: Random Forest with 300 max depth, 3000 estimators, 70 minimum sample split and 61.4% precision across all splits.
5. Model Group 2107: Random Forest with 75 max depth, 3000 estimators, 100 minimum sample split and 61.2% precision over all splits.

At this point, we tested to see whether the models performed well in the recent test set, and whether they performed well according to average precision metrics where recent test sets are more heavily weighted. We found that the top model, model 2107, was the fourth best model for performance @100 on the most recent split and performed somewhat well in some of the scenarios where more recent test split performance was weighted heavily. Therefore, we stuck with model 2107 as our best model.

PRK Curves, Feature Importance, Crosstabs of Top 5 Models and Baseline

The following figures show Precision-Recall @K curves displaying how the tradeoff between precision and recall performs on the most recent test split. Additionally, we show the feature importance for the top 25 features in each of these models (there were too many to show in each graph, so we chose the top 25 for this section, starting on page 48 there are full tables with all feature importances for each model group). Finally, we display crosstabs of these models and the best baseline.

Best Performing Baseline, Model Group 2101 Results

Note that the baseline's performance over bias metrics is reported in ensuing sections' model results and is explored in the results section as well. For this reason, we did not repeat measures of the baseline's mathematical biases in this section.

Figure 8: Model Group 2101 PRK

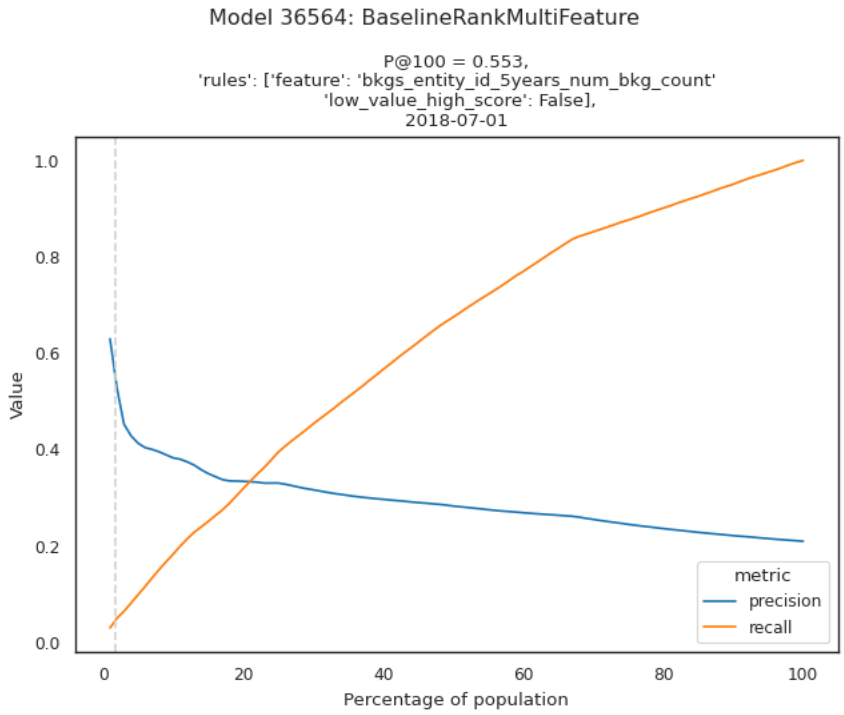


Figure 9: Model Group 2101 Feature Importance

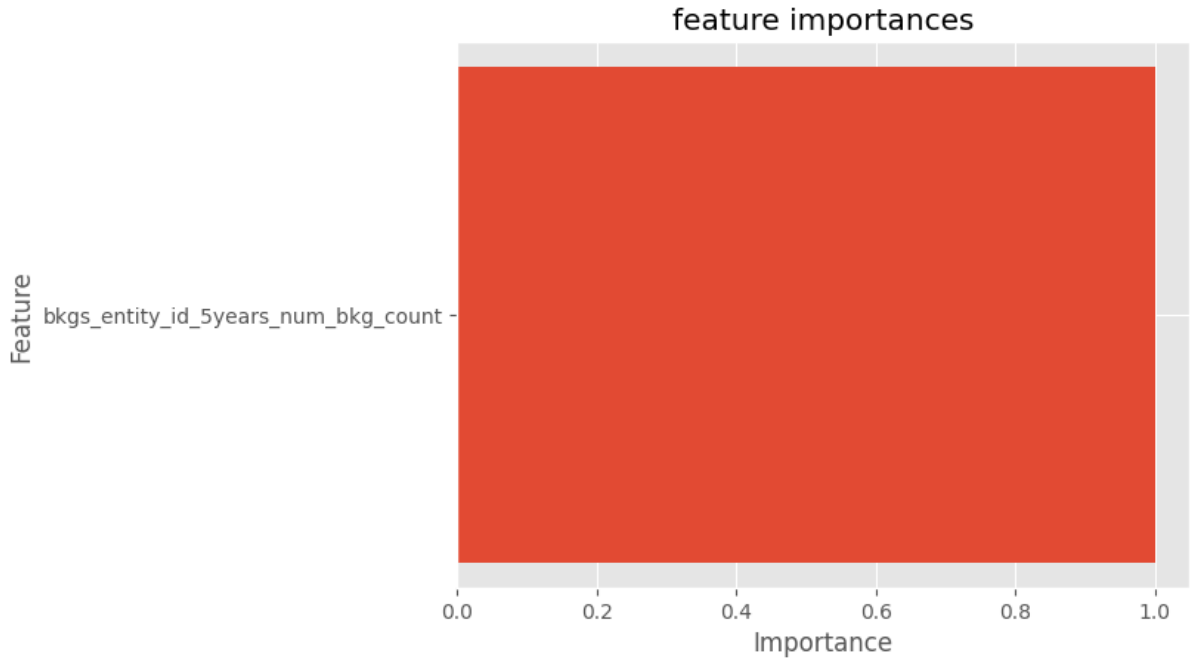


Table 3: Model Group 2101 Crosstabs of last time split

Variable Name	Mean of selected individual	Mean of the non-selected individuals
bkgs_entity_id_5years_num_bkg_count	13.68	3.11670354
inmate_chrg_entity_id_5years_plea__NULL_sum	17.68	3.36615044
hmls_entity_id_5years_hmls_sum	1.27	1.00331858
inmate_chrg_entity_id_5years_charge_level__NULL_sum	6.37	0.94008112
inmate_chrg_entity_id_5years_book_type_BW_sum	10.04	1.67146018
hmls_entity_id_all_hmls_sum	1.39	1.00497788
inmate_chrg_entity_id_5years_bond_type_CA-SU_sum	7.13	1.44542773
inmate_chrg_entity_id_all_bond_type_CA-SU_sum	7.87	1.60195428
inmate_chrg_entity_id_5years_book_type_VIEW_sum	4.03	1.18879056
inmate_chrg_entity_id_5years_bond_type_NB_sum	6.14	1.25294985

Figure 10: Model Group 2120 PRK

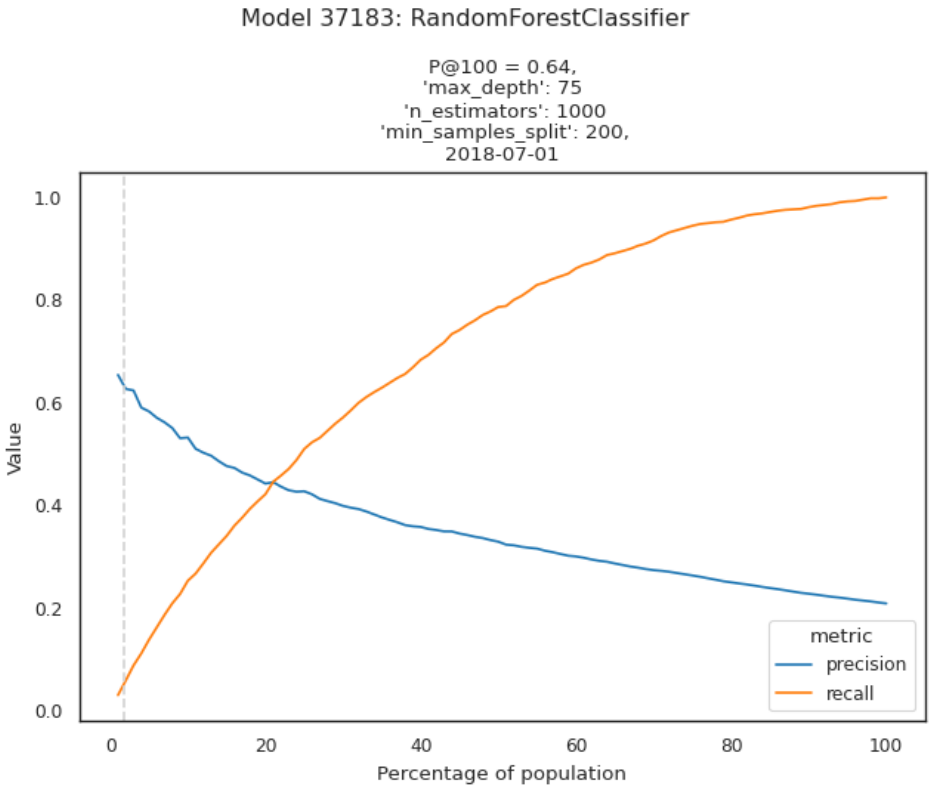


Figure 11: Model Group 2120 Feature Importance

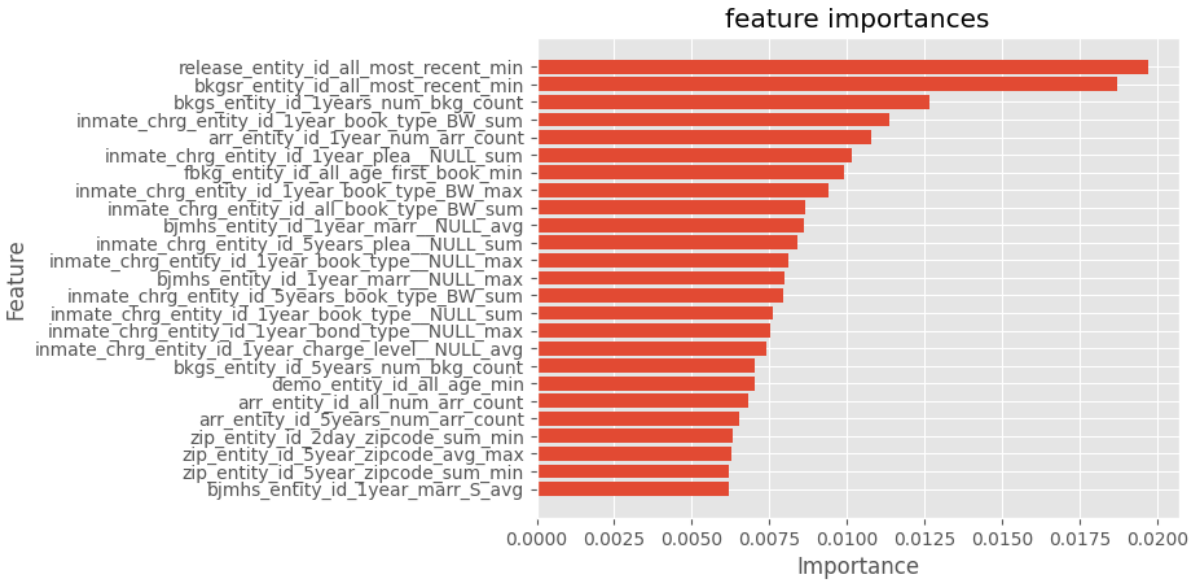
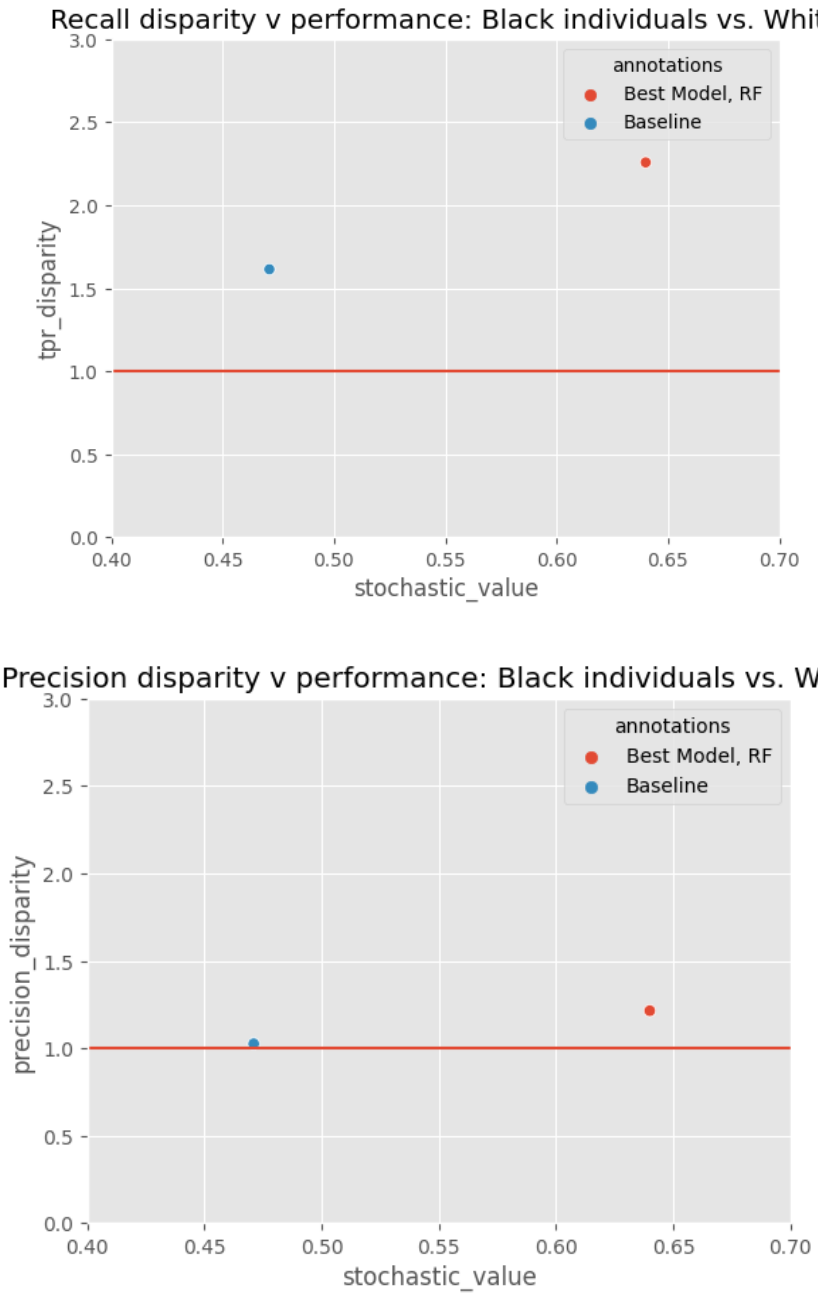
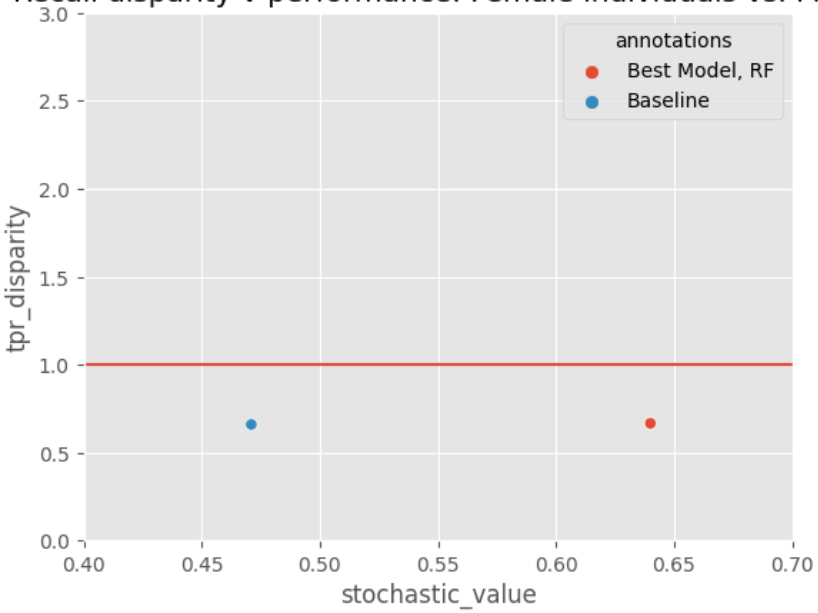


Figure 12: Model Group 2120 Bias



Recall disparity v performance: Female individuals vs. Male



Precision disparity v performance: Female individuals vs. Male

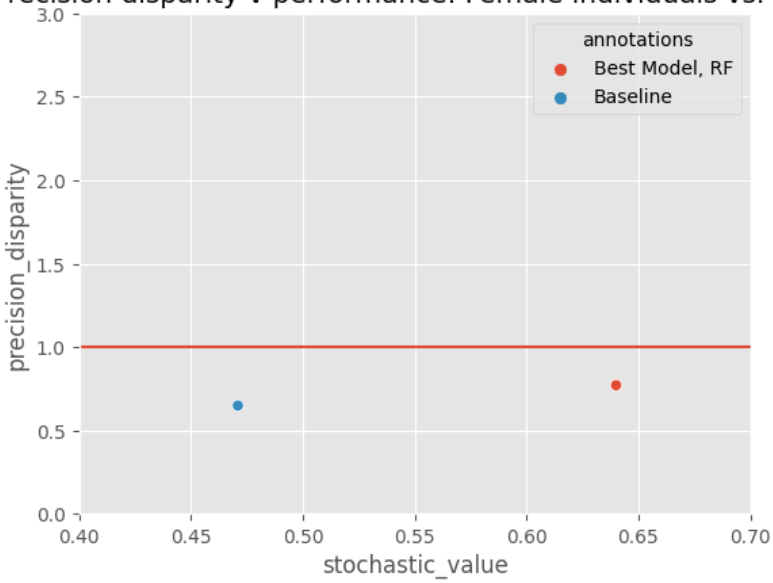


Table 4: Model Group 2120 Crosstabs

Variable Name	Mean of selected individual	Mean of the non-selected individuals
inmate_chrg_entity_id_1year_book_type_BW_sum	3.58	0.42
inmate_chrg_entity_id_1year_plea__NULL_sum	4.91	1.34
inmate_chrg_entity_id_1year_charge_level_√ú_sum	0.74	0.04
arr_entity_id_1year_num_arr_count	2.67	0.35
inmate_chrg_entity_id_5years_book_type_BW_sum	9.57	1.68
inmate_chrg_entity_id_5years_charge_level__NULL_sum	5.74	0.95
inmate_chrg_entity_id_5years_charge_type_ON_sum	4.57	0.69
inmate_chrg_entity_id_1year_charge_type_ON_sum	1.47	0.15
inmate_chrg_entity_id_5years_bond_type_PR_sum	3.49	0.63
inmate_chrg_entity_id_5years_plea__NULL_sum	13.51	3.44

Figure 13: Model Group 2126 PRK

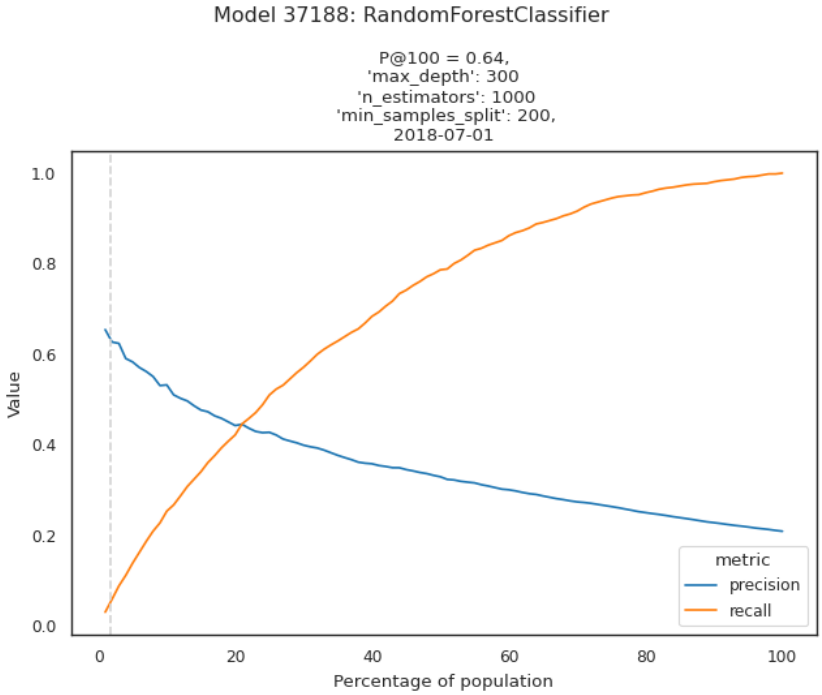


Figure 14: Model Group 2126 Feature Importance

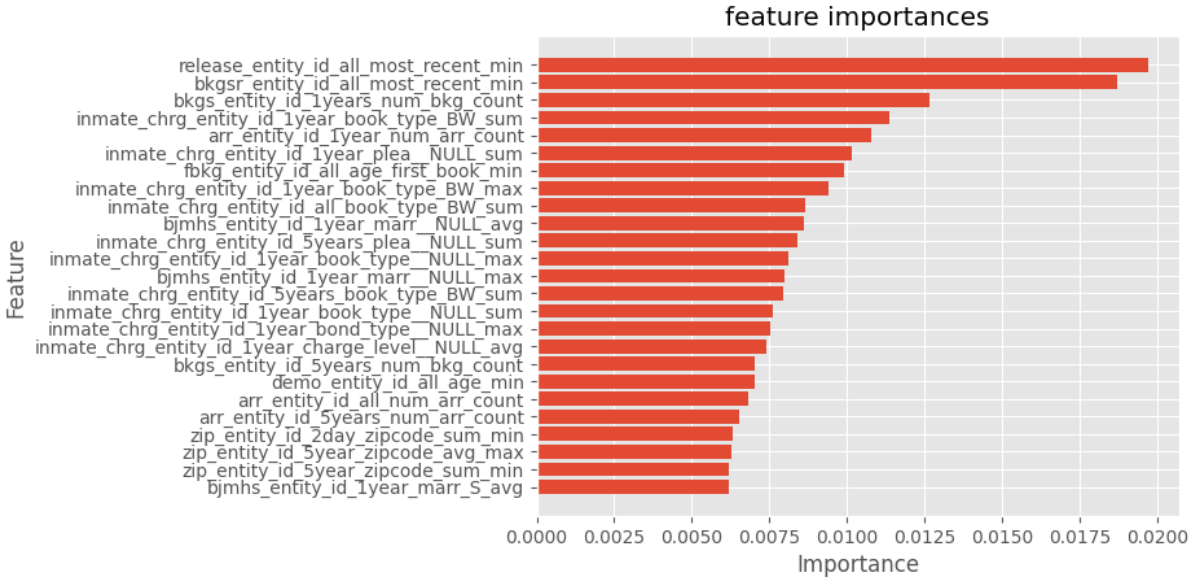
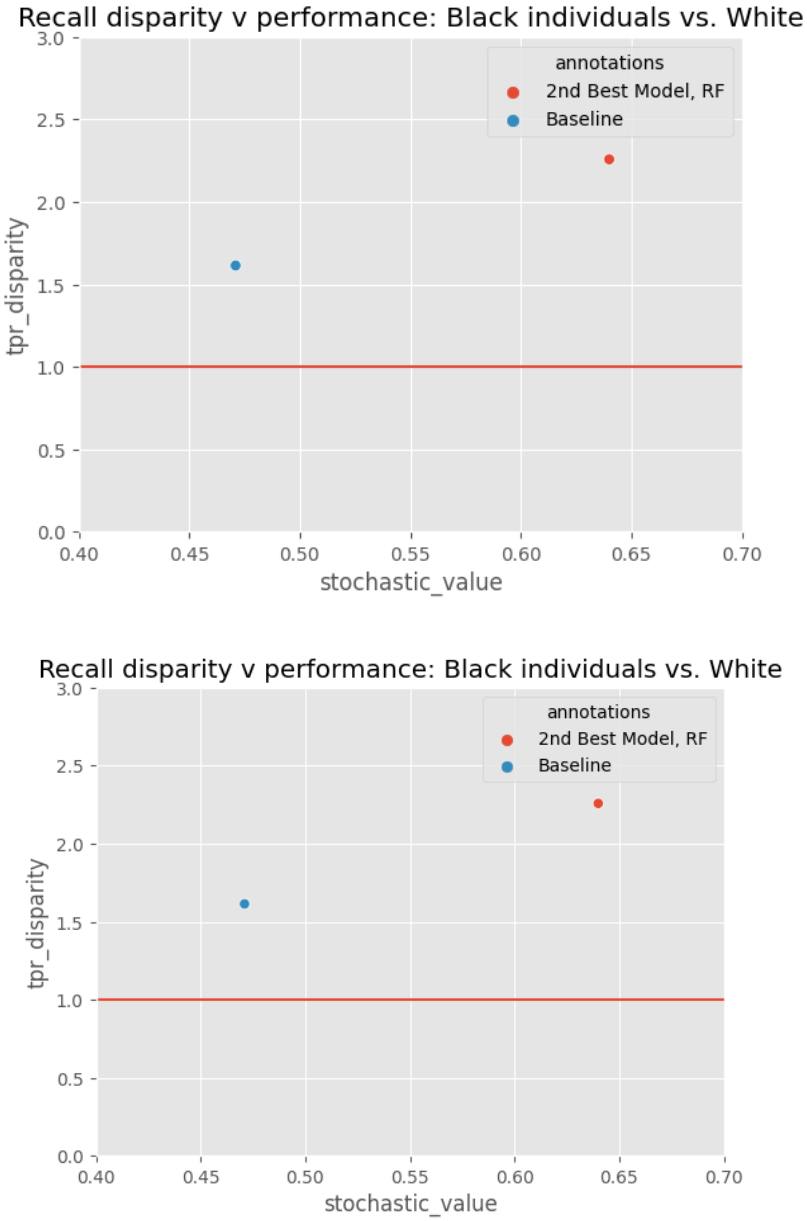
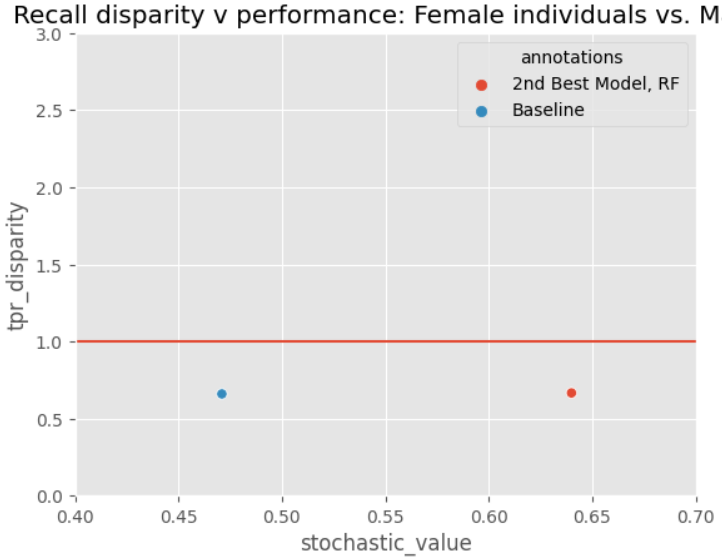


Figure 15: Model Group 2126 Bias



Recall disparity v performance: Female individuals vs. Male



Precision disparity v performance: Female individuals vs. Male

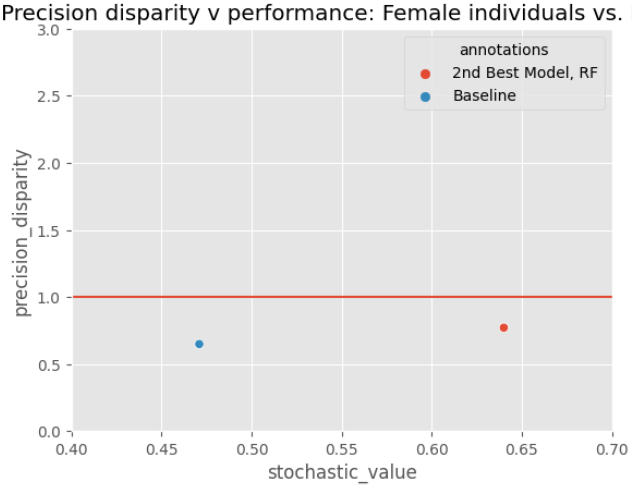


Table 5: Model Group 2126 Crosstabs

Variable Name	Mean of selected individual	Mean of the non-selected individuals
inmate_chrg_entity_id_1year_book_type_BW_sum	3.58	0.42
inmate_chrg_entity_id_1year_plea__NULL_sum	4.91	1.34
inmate_chrg_entity_id_1year_charge_level_√ú_sum	0.74	0.04
arr_entity_id_1year_num_arr_count	2.67	0.35
inmate_chrg_entity_id_5years_book_type_BW_sum	9.57	1.68
inmate_chrg_entity_id_5years_charge_level__NULL_sum	5.74	0.95
inmate_chrg_entity_id_5years_charge_type_ON_sum	4.57	0.69
inmate_chrg_entity_id_1year_charge_type_ON_sum	1.47	0.15
inmate_chrg_entity_id_5years_bond_type_PR_sum	3.49	0.63
inmate_chrg_entity_id_5years_plea__NULL_sum	13.51	3.44

Figure 16: Model Group 2106 PRK

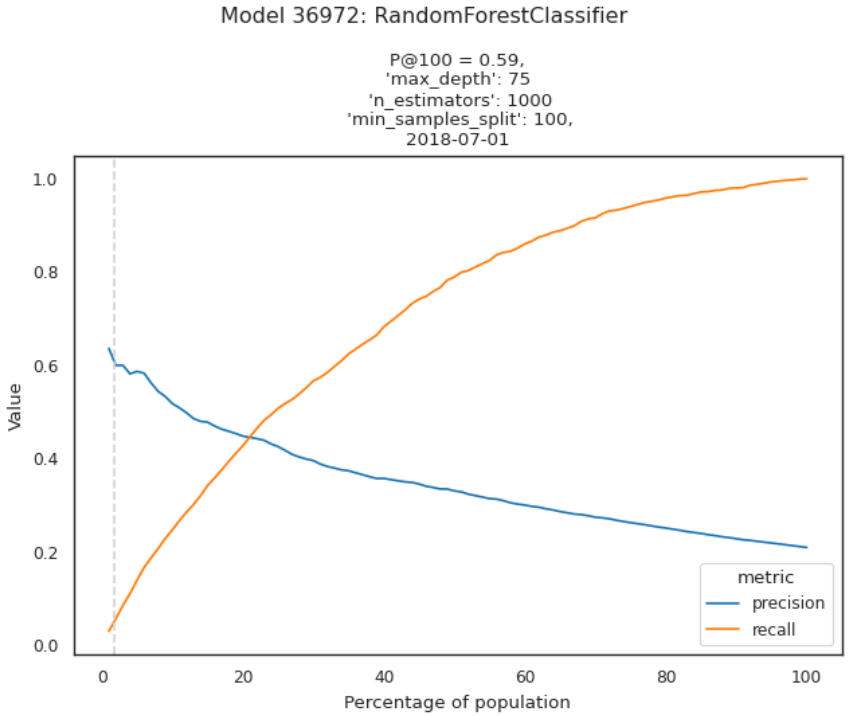


Figure 17: Model Group 2106 Feature Importance

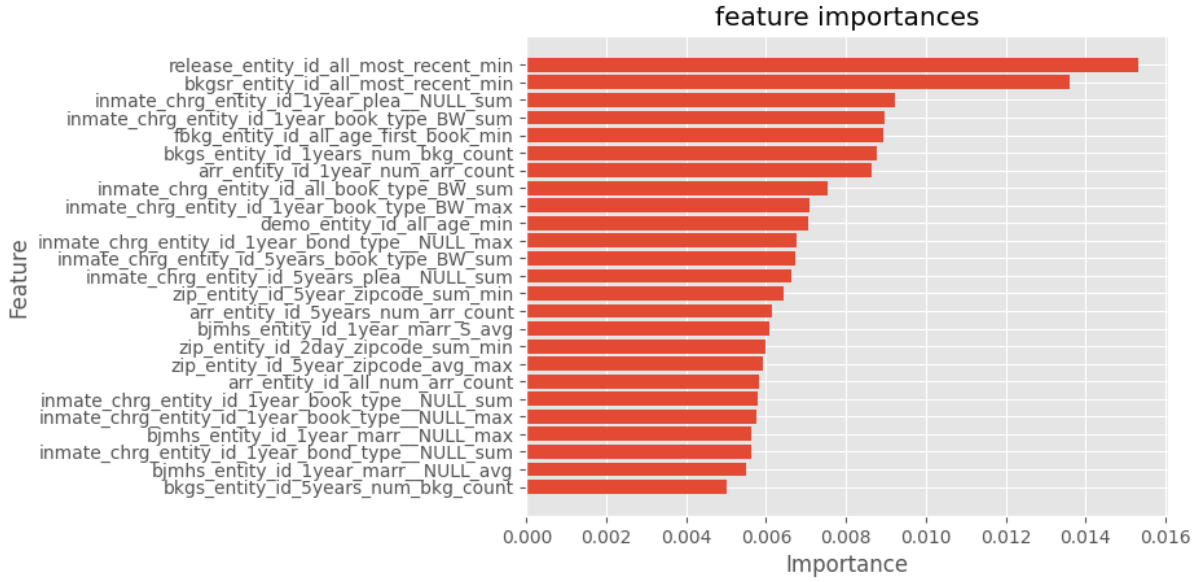
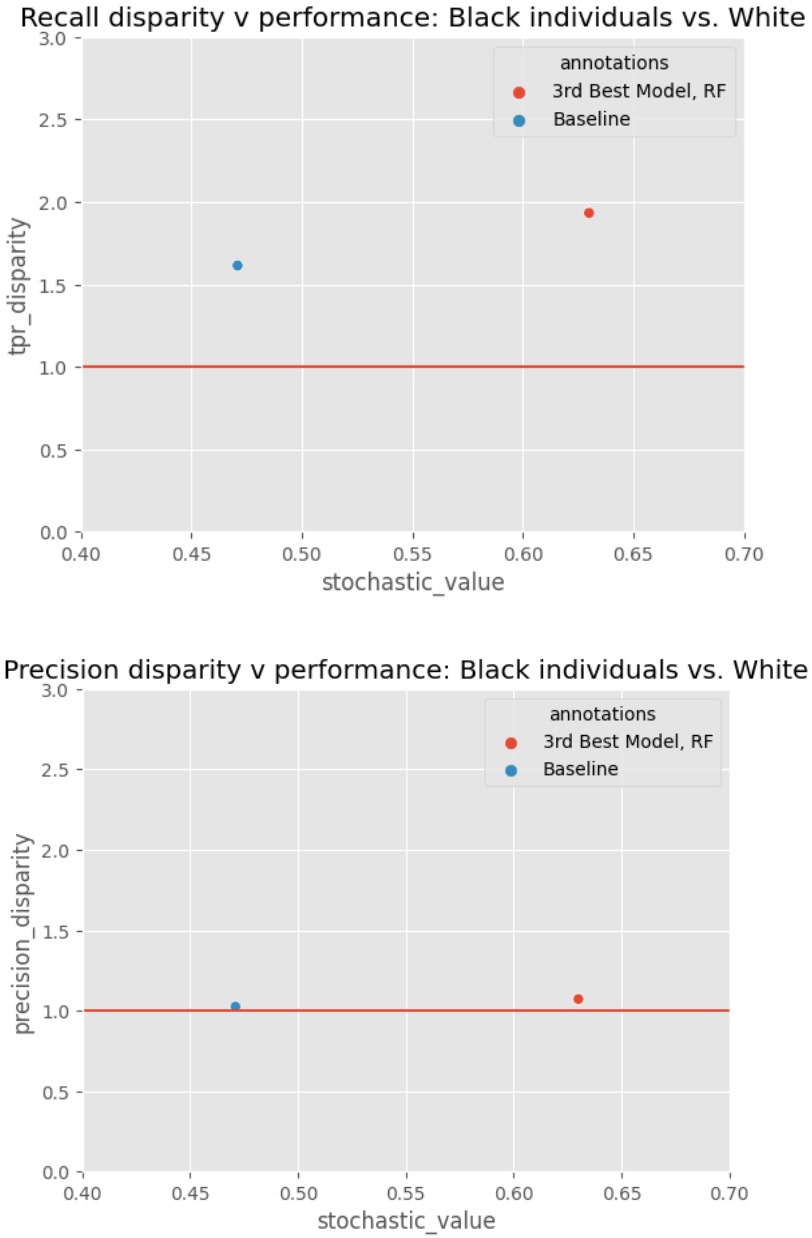
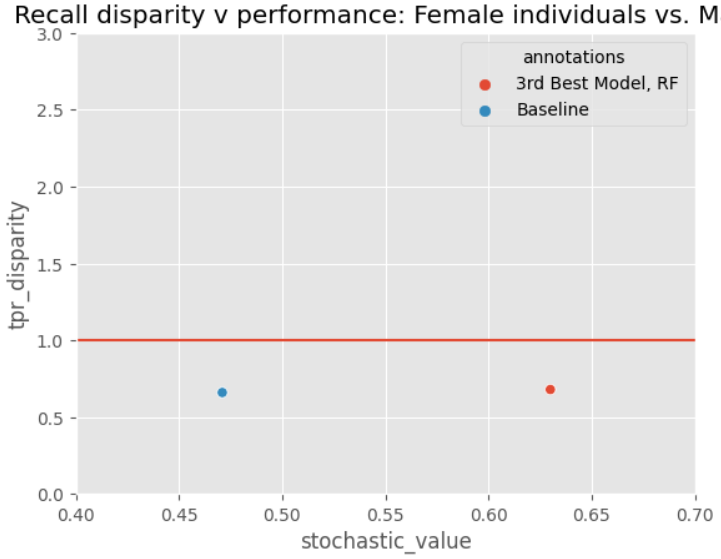


Figure 18: Model Group 2106 Bias



Recall disparity v performance: Female individuals vs. Male



Precision disparity v performance: Female individuals vs. Male

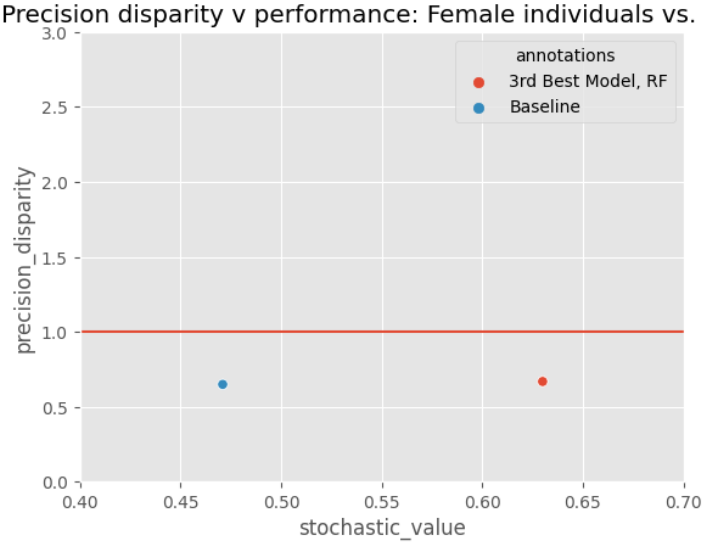


Table 6: Model Group 2106 Crosstabs

Variable Name	Mean of selected individual	Mean of the non-selected individuals
inmate_chrg_entity_id_1year_book_type_BW_sum	3.62	0.42
inmate_chrg_entity_id_1year_plea__NULL_sum	4.99	1.34
inmate_chrg_entity_id_5years_book_type_BW_sum	9.5	1.68
inmate_chrg_entity_id_5years_bond_type_PR_sum	3.55	0.63
inmate_chrg_entity_id_1year_charge_level_√ú_sum	0.7	0.04
inmate_chrg_entity_id_5years_plea__NULL_sum	13.63	3.44
arr_entity_id_1year_num_arr_count	2.47	0.35
inmate_chrg_entity_id_5years_charge_level__NULL_sum	5.55	0.96
inmate_chrg_entity_id_5years_charge_type_ON_sum	4.42	0.69
inmate_chrg_entity_id_1year_charge_type_ON_sum	1.4	0.15

Figure 19: Model Group 2123 PRK

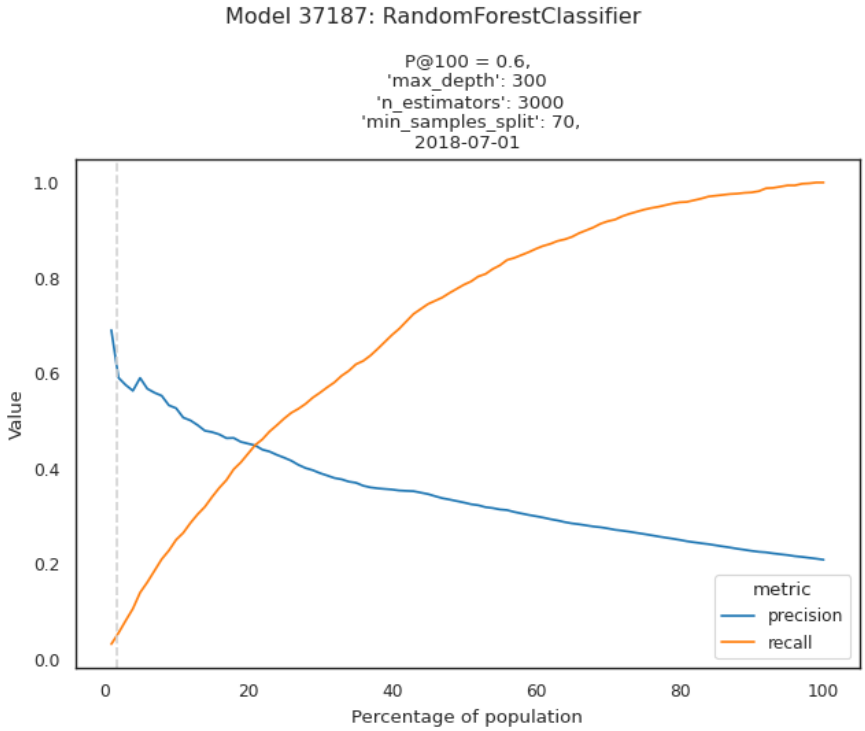


Figure 20: Model Group 2123 Feature Importance

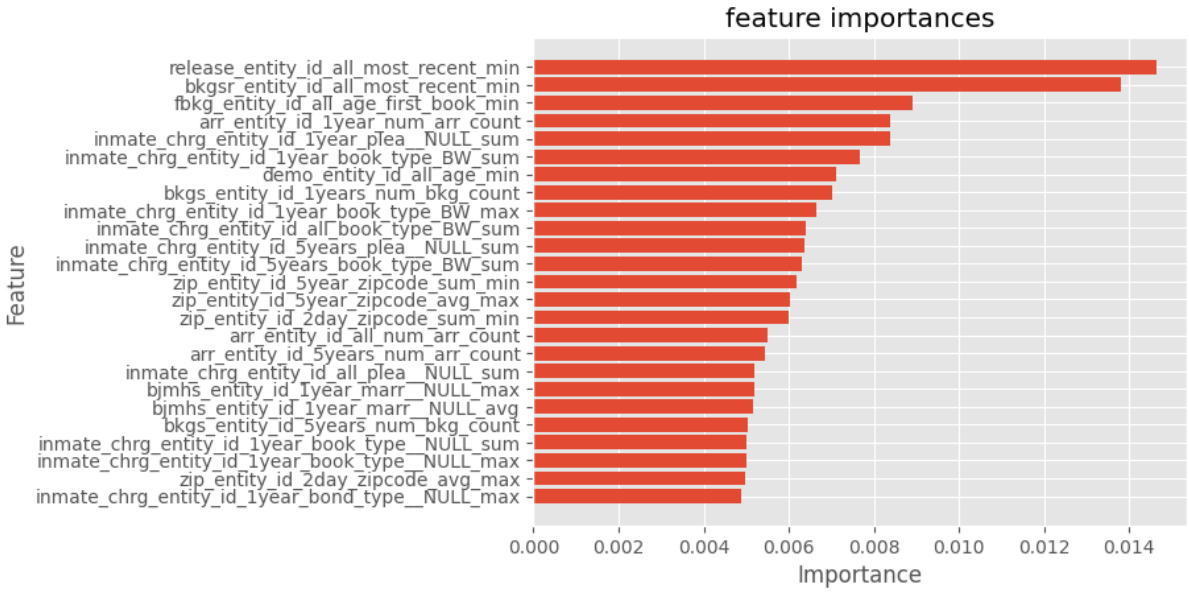
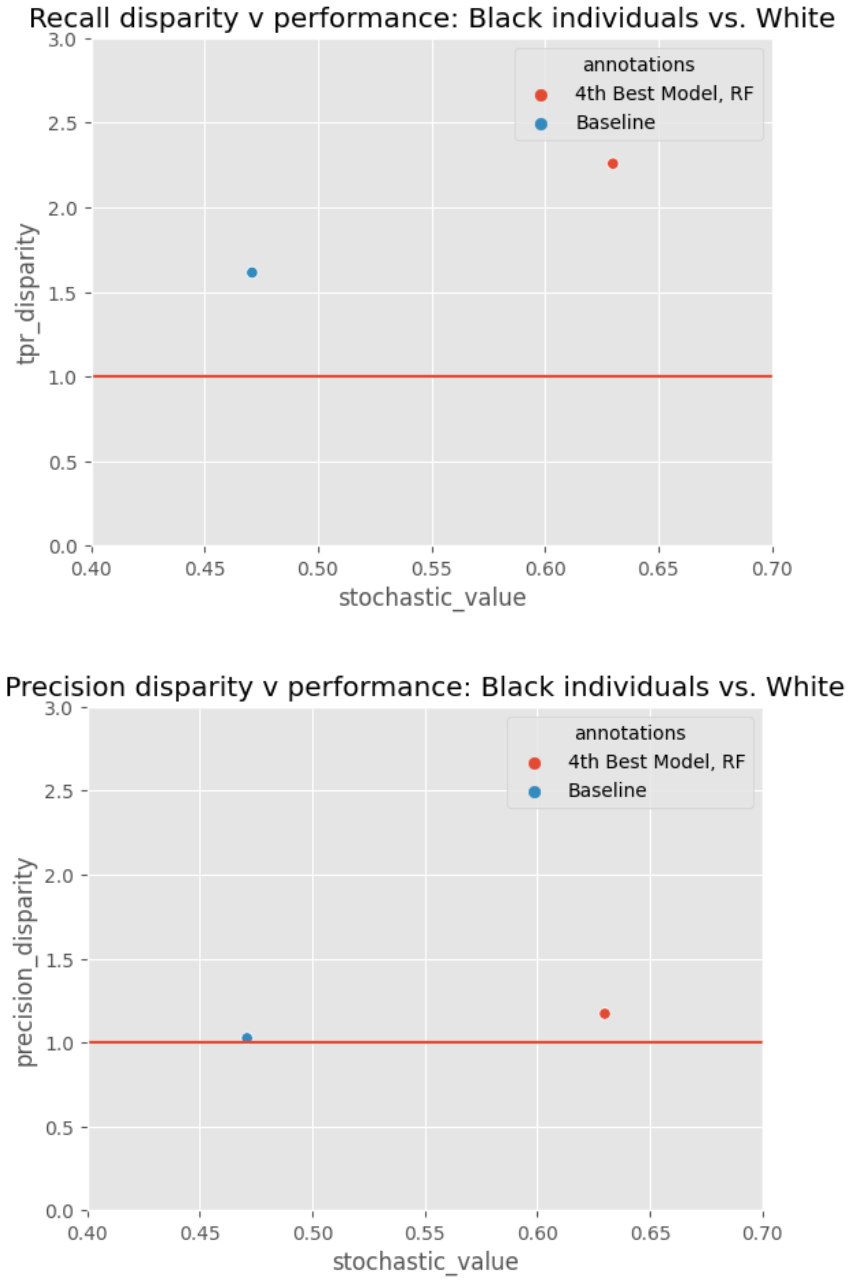
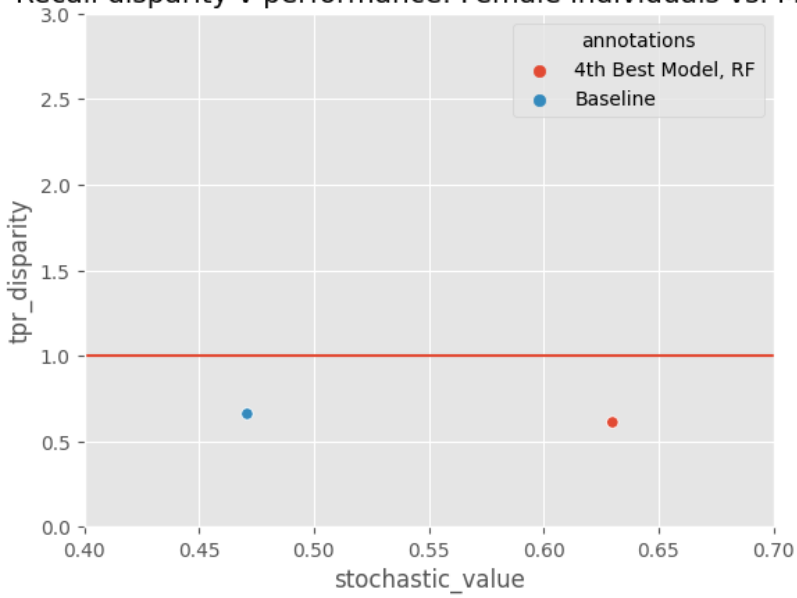


Figure 21: Model Group 2123 Bias



Recall disparity v performance: Female individuals vs. Male



Precision disparity v performance: Female individuals vs. Male

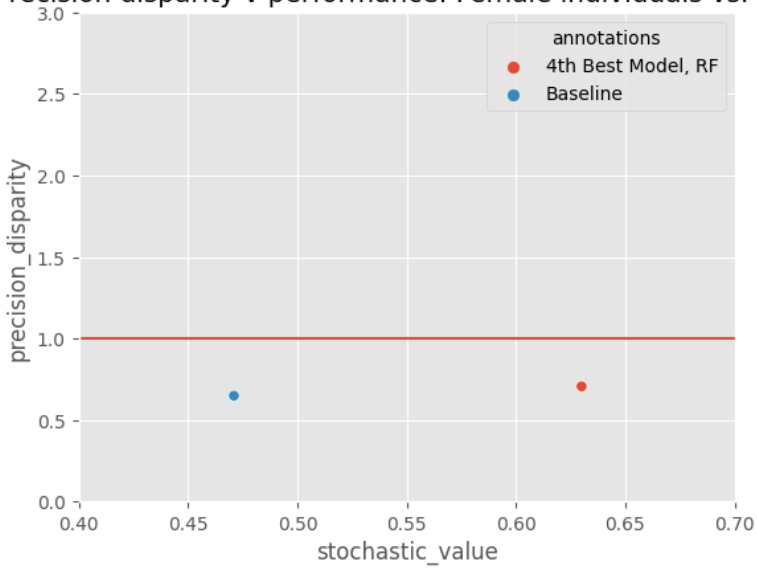


Table 7: Model Group 2123 Crosstabs

Variable Name	Mean of selected individual	Mean of the non-selected individuals
inmate_chrg_entity_id_1year_book_type_BW_sum	3.46	0.42
inmate_chrg_entity_id_5years_book_type_BW_sum	9.69	1.68
inmate_chrg_entity_id_1year_plea__NULL_sum	4.72	1.34
inmate_chrg_entity_id_5years_plea__NULL_sum	13.81	3.44
inmate_chrg_entity_id_5years_charge_level__NULL_sum	5.57	0.95
inmate_chrg_entity_id_5years_bond_type_PR_sum	3.46	0.63
inmate_chrg_entity_id_5years_charge_type_ON_sum	4.46	0.69
arr_entity_id_1year_num_arr_count	2.43	0.35
inmate_chrg_entity_id_all_book_type_BW_sum	12.18	2.37
inmate_chrg_entity_id_1year_charge_level_√ú_sum	0.65	0.04

Figure 22: Model Group 2107 PRK

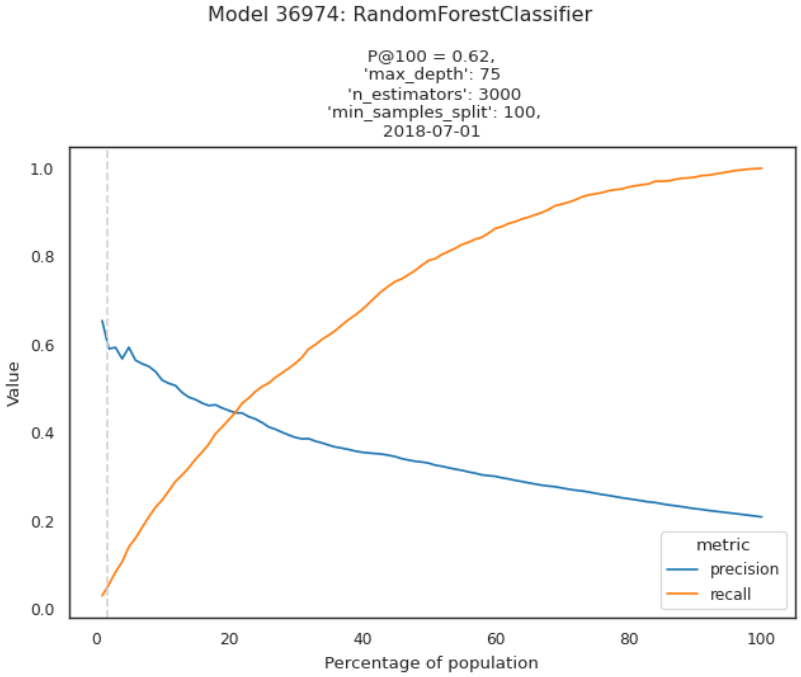


Figure 23: Model Group 2107 Feature Importance

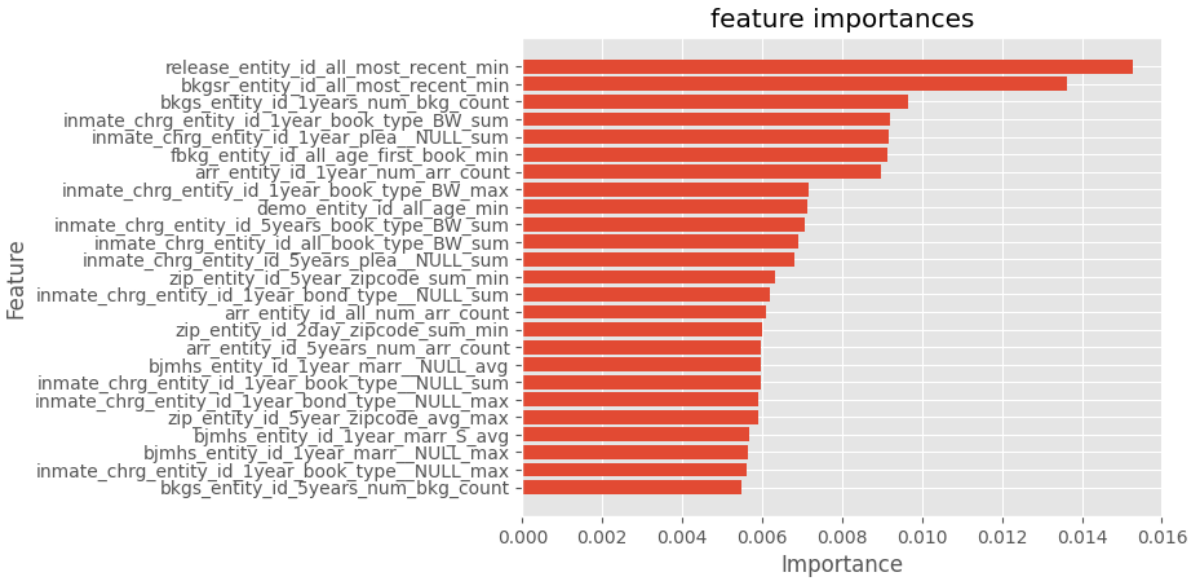
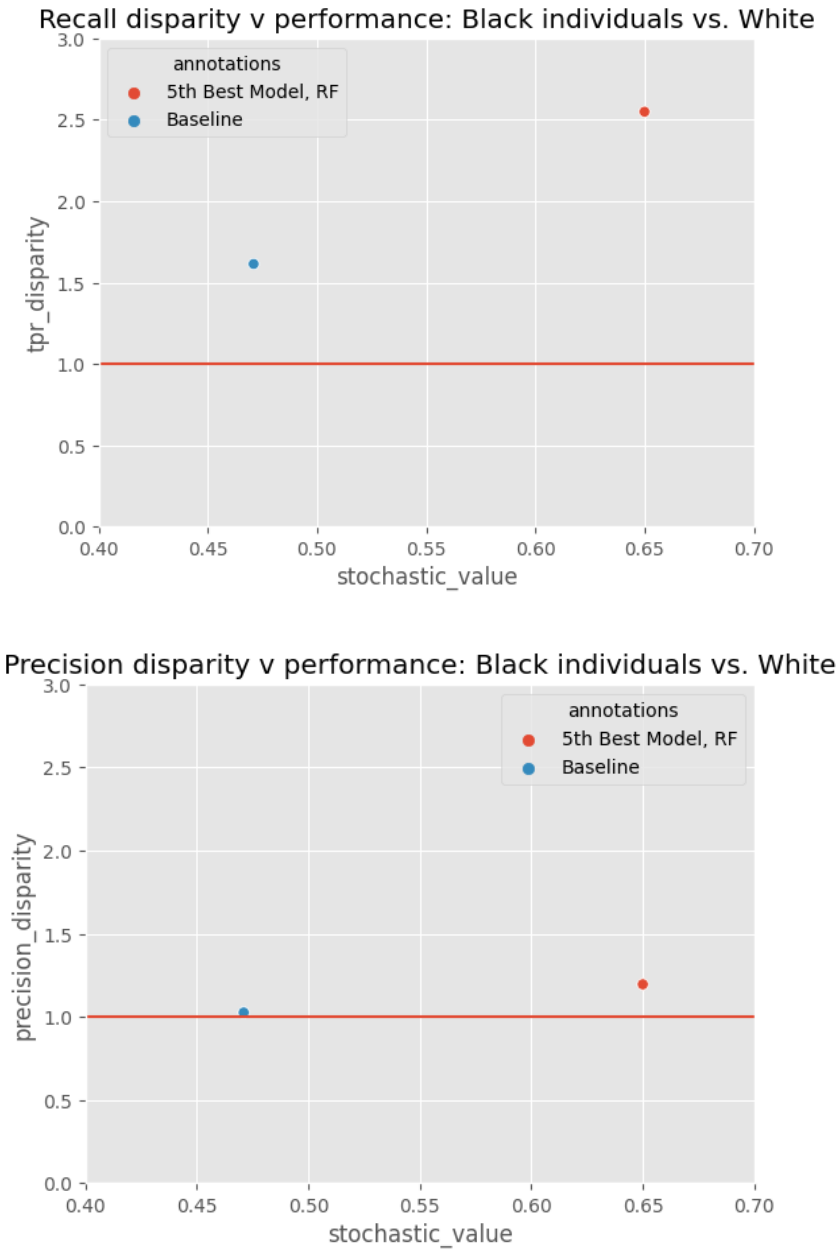
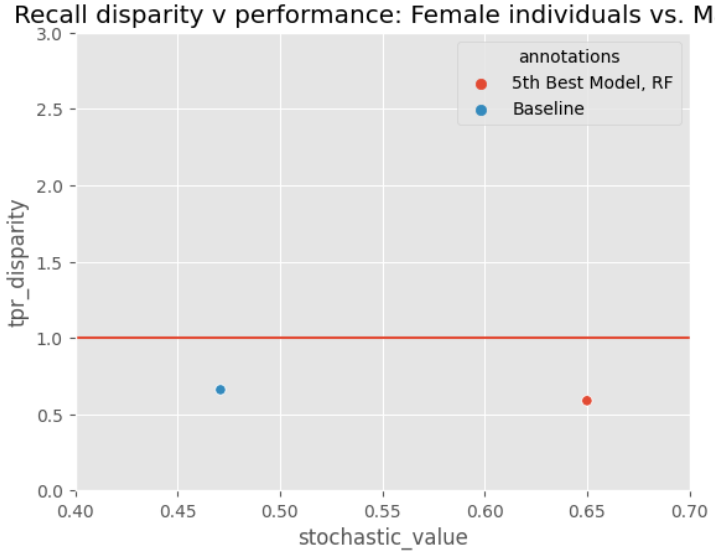


Figure 24: Model Group 2107 Bias



Recall disparity v performance: Female individuals vs. Male



Precision disparity v performance: Female individuals vs. Male

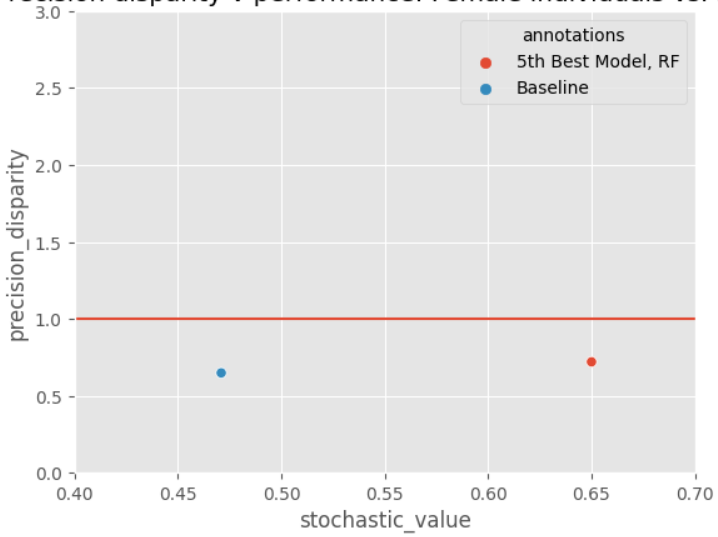


Table 8: Model Group 2107 Crosstabs

Variable Name	Mean of selected individual	Mean of the non-selected individuals
inmate_chrg_entity_id_1year_book_type_BW_sum	3.66	0.42
inmate_chrg_entity_id_1year_plea__NULL_sum	4.98	1.34
inmate_chrg_entity_id_5years_book_type_BW_sum	9.7	1.68
arr_entity_id_1year_num_arr_count	2.54	0.35
inmate_chrg_entity_id_1year_charge_level_√ú_sum	0.7	0.04
inmate_chrg_entity_id_5years_plea__NULL_sum	13.71	3.44
inmate_chrg_entity_id_5years_charge_type_ON_sum	4.54	0.69
inmate_chrg_entity_id_5years_charge_level__NULL_sum	5.62	0.95
inmate_chrg_entity_id_5years_bond_type_PR_sum	3.47	0.63
inmate_chrg_entity_id_1year_bond_type_PR_sum	1.39	0.16