Predicting Recidivism in the United States Criminal Justice System

by Alex Free

Abstract

We investigated how artificial intelligence creates bias in the US criminal justice system. Recidivism prediction algorithms used in courtrooms have been shown to produce unnecessarily harsh risk assessment scores against Black individuals, leading them to have more strict bail guidelines. This contributes to a greater narrative against Black Americans. In order to explore how different algorithms might be biased in predicting recidivism, we used 3 different classification models to predict whether individuals in the Virginia Felony Database dataset would recidivate. We tested three different (optimized) machine learning algorithms — Logistic Regression, Random Forest, and k-nearest-neighbors and compared the resulting confusion matrices to determine which was the most effective. The final Logistic Regression model had a final accuracy of 70.44%; the Random Forest algorithm had the best at 75.64% and K-nearest-neighbors had 72.24%. All algorithms had a greater false positive (false arrest prediction) for white individuals, likely due to the dataset containing many examples of white criminals. In our model white defendants were discriminated against because of data bias, which demonstrates how algorithms can treat different racial groups differently depending on the data that goes in. Given historical bias, we can see how this same pattern could occur in the opposite direction favoring white individuals more frequently in other datasets.

Introduction

There are real-life effects of machine bias in courtrooms. American lives are completely changed based on criminal sentencing, which can mean the difference between rehabilitation and recidivism. This issue is urgent and finding a solution to ensure equality is imperative. Our approach was to use machine learning to classify defendants using categorical data (race, gender, etc.) to output a label (guilty/not guilty) in order to best emulate recidivism prediction algorithms. This was a supervised learning problem.

Background

The legacy of slavery has created a gap in cumulative wealth between white and Black Americans. White Americans have had centuries longer to build up a “nest egg” due to economic privilege, thereby indirectly profiting off of slavery even after generations. Black Americans have suffered from lack of financial support (such as not being offered or paying a premium on loans due to poor credit history) and simply inadequate wages. Healthcare is another area where Blacks are discriminated against: from poor grocery store placement and selection in neighborhoods predominantly of color to environmental racism, America has made access to basic needs more difficult for Black people. Redlining prevents neighborhoods of color from prospering economically because it keeps valuable investments out of less-established, “higher-risk” areas. This is particularly an issue because these neighborhoods ache for funding the most.

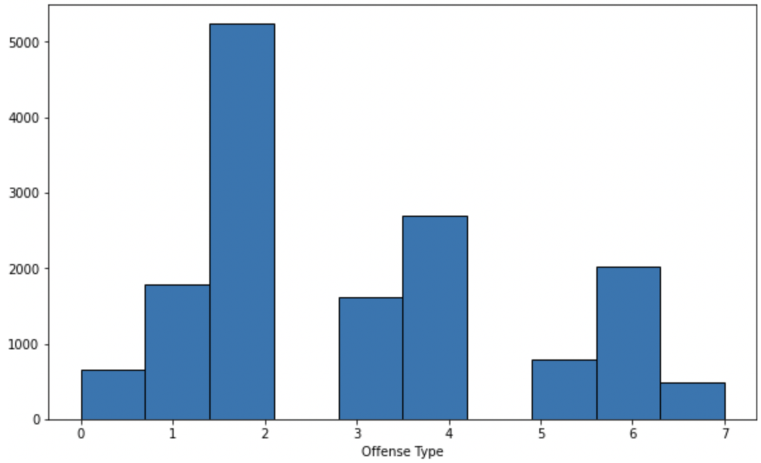
The disparity between white and Black American children is a key cog in the inescapable “school-to-prison-pipeline”: faulty legislature creates a lack of access to proper schooling for children of color, meaning insufficient opportunities to access higher career outcomes, and the cycle of poverty continues. Youth of color are forced into a life of crime to meet basic needs, but are punished even without a choice in the matter. These people often face manipulative plea deals that prey off of their inability to find adequate legal representation. These oppressive factors are vital components of slavery’s legacy and highlight how America’s racist history continues to plague it today, signaling an urgent need for change.

Propublica investigated one proprietary recidivism risk assessment algorithm, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). The article concluded COMPAS algorithm incorrectly disadvantages Black defendants in its recidivism calculations, while incorrectly marking white defendants as innocent at a higher rate. Blacks are predicted as more risky than reality, but whites are predicted the opposite, even when controlling for certain factors. The gap is greater for violent recidivism predictions. From this, Larson et al. (2016) caution that while risk assessment scores can serve as a baseline for judges, one must be very mindful of the bias and rely on human judgment.

The University of Mumbai published an article proposing a new way of using machine learning to predict recidivism. The study tested the same three algorithms used in our research. The Random Forest algorithm was the most accurate with ~88% accuracy rate. One limitation of the study’s dataset was that it focused on underage males, which does not allow the model to control for gender or age. Also, the article does not distinguish between accuracy and fairness. Even though accuracy rates may be similar across races, one race may be more frequently predicted to have positive outcomes and one may more frequently be predicted to have negative ones.

Dataset

We used the “Racial Disparities in Virginia Felony Court Cases, 2007-2015” dataset from the Rand Corporation from this project. We also considered using the Propublica “COMPAS Recidivism Risk Score Data” dataset, but this data was already analyzed in an article by Propublica so we decided to use the Rand Corporation dataset. The dataset we chose contained 12 features and 15287 samples. Our goal was to closely resemble the COMPAS algorithm as possible, so we included features that related to defendant data but excluded court outcome data to avoid giving the machine learning algorithm data directly related to the output. The final features we used were gender, sentence minimums for the accused crime, race, year, crime type, county, and a few other features relating to the type of crime the defendant is accused of.



Offense Type Graph (# of defendants per crime-code)
Key (respectively): burglary, fraud, larceny, motor vehicle, narcotics, other, violent, weapon

Methodology/Models

We tested three machine learning models in this project — logistic regression, random forest, and k-nearest neighbors. Logistic regression uses statistical modeling for classification. Random Forest algorithms use decision trees. K-nearest-neighbors algorithms classify based on the nearest “neighbors” (examples).

We first split the dataset into train/val/test (72%/8%/20%) and removed any features relating to the outcome of the case. We then input the data into models using the SciKit-Learn Python package and optimized hyperparameters (with our validation set) and analyzed results, comparing the confusion matrix and overall accuracy score.

Classification Using Logistic Regression

For the logistic regression method, we attempted to optimize the max_iter (iterations) hyperparameter. We graphed the error from a max_iter value of 250 to 2000 (in increments of 250), but noticed no change across max iterations. We also analyzed the model’s error by solver type, and the default “lbfgs” showed the least error.

Classification Using Random Forests

For the random forest model, we analyzed the max_depth hyperparameter between 1 and 16. The optimal depth was 11 — higher depths had similar error rates but took longer. We also tested the n_estimators hyperparameter, which was optimal in a range from 70 to 100. The model was most successful with n_estimators=100 (default).

Classification Using K-nearest Neighbors

For K-nearest-neighbors, we graphed the k_neighbors hyperparameter (the amount of neighbors) in between 1 and 300. The optimal value for k_neighbors was 15.

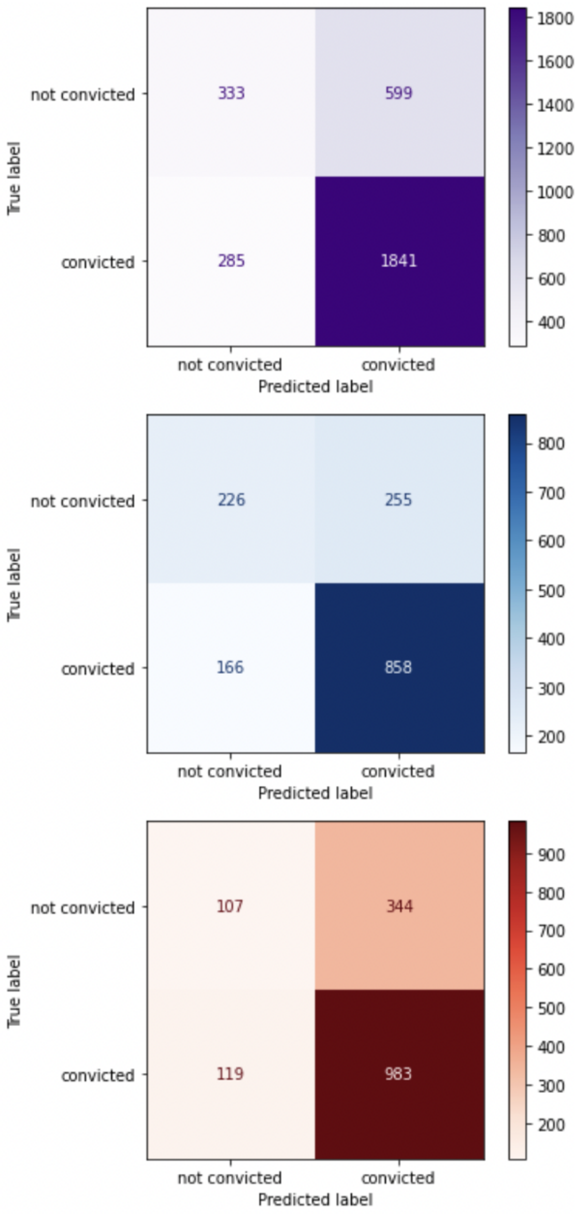
Results and Discussion

The logistic regression model proved to have the worst overall accuracy rate of all of the methods with 70.44%. The confusion matrices we generated by race showed statistics as follows: 16.27% False Positive rate for Black individuals, 11.43% False Negative rate for Black individuals, 25.98% False Positive rate for white individuals, 5.5% False Negative rate for white individuals.

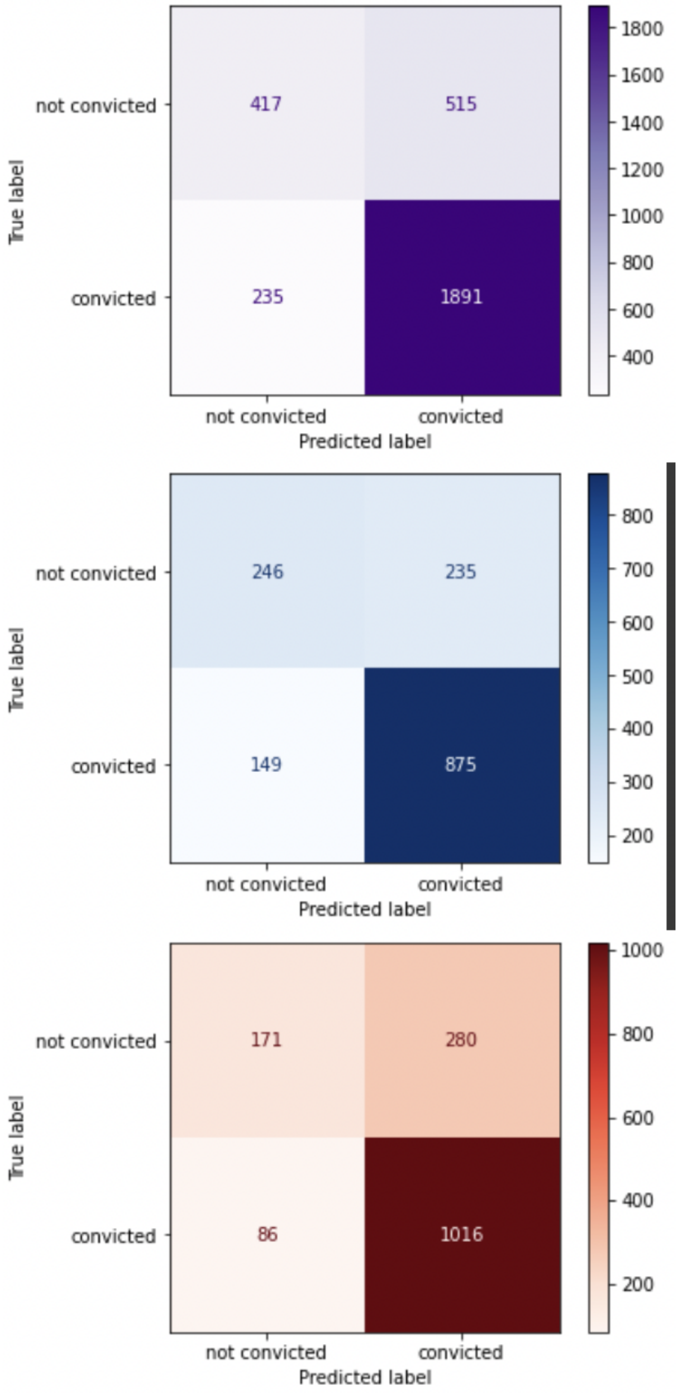
The overall accuracy rate with tuned hyperparameters in the random forest model was 75.64%. The confusion matrices we generated by race showed statistics as follows: 15.8% False Positive rate for Black individuals, 7.9% False Negative rate for Black individuals, 20.61% False Positive rate for white individuals, 4.38% False Negative rate for white individuals.

The optimal overall accuracy rate with tuned hyperparameters was 72.24%. The confusion matrices we generated by race showed statistics as follows: 18.5% False Positive rate for Black individuals, 8.98% False Negative rate for Black individuals, 21.31% False Positive rate for white individuals, 6.72% False Negative rate for white individuals.

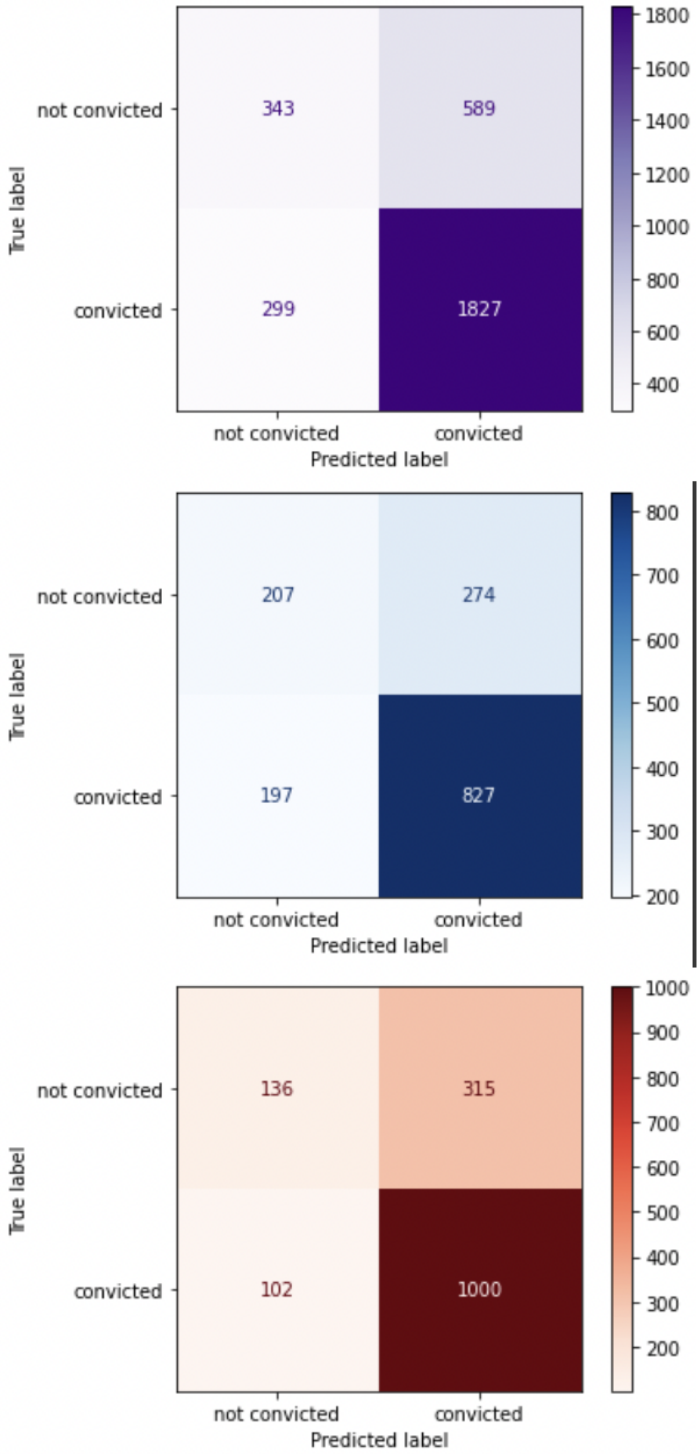
The primary source of error would likely be the dataset used: it displays many examples of white defendants who ended up convicted, so the algorithm learns from that and incarcerates white defendants at a higher rate. This phenomenon is similar to the bias against Black individuals in the justice system that the ProPublica article identified but with data leading to opposite outcomes. Criminal justice data is scarce and frequently tainted by racism, so recidivism prediction models are operating off of faulty input data that makes the conclusion biased as well. This Virginia dataset is unrepresentative of data that is largely used across the US: in practice, data will likely look more like the ProPublica study’s dataset than in this one. Across the country, Black Americans are incarcerated at significantly higher rates than white Americans, so this dataset is an outlier.



Logistic Regression Algorithm Confusion Matrices (from top to bottom: combined, Black defendants, white defendants)



Random Forest Algorithm Confusion Matrices (from top to bottom: combined, Black defendants, white defendants)



K-nearest-neighbors Algorithm Confusion Matrices (from top to bottom: combined, Black defendants, white defendants)

Conclusion

The Random Forest model was a pretty effective classifier when measured by overall accuracy, although even its accuracy might be considered low in such a high stakes decision making environment, but all three tested models were prone to error from bias in the data. It is not possible to create a perfect algorithm, and numbers can't reflect feelings nor meaning very well. Models can only be as accurate as their input data: there is always a compromise between fairness and accuracy. The problem arises when input data (training data for machine learning) is biased because of past racism/bias with human officers, then the algorithm will react accordingly and cannot be fair. Artificial intelligence may have a role to play in the criminal justice system in the future, but it is currently plagued by the effects of some of the glaring inequities in our society.

References

Algorithmic Risk Assessments and the Double-Edged Sword of Youth. https://openscholarship.wustl.edu/cgi/viewcontent.cgi?article=6353&context=law_lawreview.

"Classification of Criminal Recidivism Using Machine Learning Techniques." *Researchgate*, https://www.researchgate.net/publication/342436748_Classification_of_Criminal_Recidivism_Using_Machine_Learning_Techniques.

Compas Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Compas Validation 1-26-11 - Florida State University. https://criminology.fsu.edu/sites/g/files/upcbnu3076/files/2021-03/Validation-of-the-COMPAS-Risk-Assessment-Classification-Instrument.pdf.

Hao, Karen. "Can You Make AI Fairer than a Judge? Play Our Courtroom Algorithm Game." *MIT Technology Review*, MIT Technology Review, 10 Jan. 2022, https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

Jeff Larson, Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, 23 May 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Johndrow, James E., and Kristian Lum. "An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction." *Project Euclid*, Institute of Mathematical Statistics, https://projecteuclid.org/journals/annals-of-applied-statistics/volume-13/issue-1/An-algorithm-for-removing-sensitive-information--Application-to-race/10.1214/18-AOAS1201.full.

Julia Angwin, Jeff Larson. “Machine Bias.” *ProPublica*, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.