CSE 574 Market Submission Model
Group 7: Nikhil Arora, Jacquelyn Dufresne, Mridula Rao

Proposal

Our proposed system for replacing COMPAS is a RBF kernel support vector regressor that uses the postprocessing method equal opportunity, where accuracy is used as the secondary optimization criteria. The total cost of the system's choices to society is $129,335,008 and the overall accuracy of the system is 0.6980956804458894.

Motivation

Correctional Offender Management Profiling for Alternative Sanctions tool, commonly known as COMPAS, is a widely used algorithm that assists judges in making consequential decisions for defendants such as assigning sentences or determining bail/parole eligibility by outputting a risk score that indicates how likely someone will recidivate. ProPublica, a non-profit newsroom that uses investigative journalism to incite reform, reviewed the risk scores assigned to more than 7,000 individuals arrested in Broward County, Florida and evaluated whether an individual committed another crime within the next two years.[1] This benchmark was strategically chosen, as it is the same benchmark used by Northpointe, the creators of the algorithm.[2] While evaluation of the algorithm identified correctly predicted recidivism at similar rates for both Caucasian and African American defendants, racial disparities were found when the algorithm made mistakes.[3] African American defendants were falsely flagged as future criminals at almost twice the rate of Caucasian defendants (44.9% for African Americans vs. 23.5% for Caucasians), and Caucasian defendants were mislabeled as low risk more often than African American defendants (47.7% for Caucasians vs. 28% for African Americans).[4] Northpointe responded by claiming that ProPublica did not take into account base rates of recidivism for African Americans and Caucasians, resulting in false assertions.[5] ProPublica refuted Northpointe's response, pointing out that after running a logistic regression and adjusting for recidivism, criminal history, age, and gender across races, they found that black defendants were 45% more likely to get a higher score.[6] Although race is not a factor in the COMPAS algorithm, it is possible that characteristics shared by a certain demographic may be more prevalent in a certain racial group, resulting in this disparity. In light of this analysis, we have developed a new system to replace COMPAS. As volunteers of a non-profit non-governmental organization, our mission is to provide each and every individual fair treatment in the U.S. criminal justice system. We recognize that the stakeholders in this situation are the defendants, the defendant's communities, the local governments, and state governments, and therefore have developed our system with the needs of these stakeholders in mind.

Solution

While it is difficult to define fairness, there are justifications that make certain metrics more desirable for fair machine learning applications to the U.S. criminal justice system. Our proposed replacement for COMPAS provides a notion of fairness among different racial groups by implementing the postprocessing method equal opportunity. This solution determines thresholds for each racial group such that all groups have equal true positive rates within a tolerance value epsilon of .01 and then chooses a threshold solution that satisfies this requirement with the highest accuracy on the data. Our solution solves the main issue outlined by ProPublica, ensuring that African Americans do not experience significantly higher rates of being falsely predicted to recidivate and Caucasians do not experience significantly higher rates of being incorrectly predicted to not recidivate. Our training data reduced the disparity of the false positive rate between African Americans and Caucasians to a mere 1% (23.2% for African Americans vs. 24.4% for Caucasians) and had adequate results on the test data (28.2% for African Americans vs. 34.5% for Caucasians). It is important to note that the values chosen for the parameters gamma and C in our model provide the highest accuracy on the test data without creating a racial significant racial disparity for false positive rates. A slightly higher accuracy could have been achieved with the consequence of an increased gap between African American and Caucasian false positive rates, and a lower accuracy could have been achieved that minimized the false positive gap between African American and Caucasians, but would have resulted in negative effects on the accuracy of the other racial groups. While our test data shows that Caucasians have a slightly higher false positive rate, this is a significant improvement from the racial disparities found in the COMPAS model. Our training data also reduced the disparity of the false negative rate between African Americans and Caucasians (20.2% for African Americans vs. 20.8.5% for Caucasians) and had similarly great results on the test data (28.4% for African Americans vs. 29.4% for Caucasians). It also is important to note that the racial group labeled Other produced significantly lower false positive rates and significantly higher true negative rates for the training and test data. This does not affect our models notion of fairness, as it does not focus on unfair treatment of one particular racial group.

Analysis of the impracticality of other post processing models will further justify that our solution achieves the most relevant application of fairness for the problem at hand. Applying maximum accuracy and single threshold solutions as post processing methods would have no considerations to fairness across racial groups, as a solution with a maximum accuracy operates independently

[1] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, accessed May 6, 2020, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] Ibid.
[3] Ibid.
[4] Ibid.
[5] William Dieterich, Christina Mendoz and Tim Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and predictive parity," Northpointe Inc., July 8, 2016, accessed May 6, 2020, https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html
[6] Ibid.

of any additional notions of fairness and can often worsen racial disparities in the data, and a single threshold solution fails to make considerations towards statistical measurements of fairness or underlying biases within the data. Predictive parity is also not a viable solution for several reasons. Predictive parity does not address measurements such as the false positive and false negative rates, which are the main drivers of unfairness in the COMPAS algorithm. Some researchers note that applying predictive parity can produce results that move away from fairness by further creating racial disparities.[7] Most importantly, Northpointe previously revealed that they had already met predictive parity requirements with the flawed COMPAS algorithm.[8] When applying predictive parity to our test data, the results are still desirable, however there is future risk that predictive parity could result in racial disparities with false positive and false negative rates on new data, making it an inadequate application of fairness. Some experts also advise against the use of demographic parity to enforce fairness when certain attributes are present. In particular, demographic parity is not suited for situations where base recidivism rates differ, which is present with the COMPAS algorithm and dataset.[9] Additionally, some experts argue that demographic parity can correctly accept individuals from one demographic on the intended basis but random individuals in another to meet the requirement of equal predictive positive rates.[10] The potential consequence of predicting improbable individuals to recividite to meet a predictive threshold required by demographic parity produces a level of uncertainty that cannot be tolerated when making life altering decisions for individuals in our society. It is also important to note that when applying demographic parity to our test data, the costs incurred were roughly $2,000,000 higher than the other postprocessing models, adding to the unsuitability of this method. The intolerable consequences of the postprocessing methods maximum accuracy, single threshold, predictive parity, and demographic parity and the desirable outcomes of our model make equal opportunity the best post processing choice for fairness within this situation.

The choice of our machine learning model also contributes to the quality of our solution. With our RBF kernel support vector regressor, all five postprocessing methods produce an accuracy between .6897 and .7004 and a cost between $128,523,748 and $131,895,580 for test data. Minimal changes in accuracy and cost between the postprocessing methods allowed our team to exclusively focus on choosing the method that was most appropriate for solving the problem analyzed by ProPublica. Minimal differences between these methods is not seen in models we tested previously, such as the linear support vector regressor, where costs could vary up to $6,000,000 for the test data, with equal opportunity having some of the highest societal costs. It is clear that our model provides functionality that will keep our society safe by providing an accuracy of almost 70% while also considering the societal costs incurred and bringing justice to the African American community, as they have experienced devastating racial bias through the COMPAS model.

Impact

The solution presented has proven to relieve the disparity of African American defendants being wrongfully predicted to recidivate at significantly higher rates, and eliminate the injustice of Caucasians being wrongfully predicted not to recidivate at higher rates. To understand the importance of relieving African Americans from high false positive rates that may result in higher incarceration rates, it is imperative to realize the implications that incarceration can have on individuals. In general, collateral consequences such as limited employment, applications for social programs such as public housing, and the right to vote are often experienced when reentering society after imprisonment.[11] For example, as of 2018, 87% of employers conduct background checks, which may explain that 60% of formerly incarcerated individuals remain unemployed after one year of reentry to society.[12] Additionally, family members of those incarcerated are often considered “hidden victims” due to the lack of acknowledgement of the challenges and difficulties they may experience.[13] Examples of these effects include financial challenges faced by the family due to a loss of income, and a profound effect on children’s futures, as children with incarcerated parents are six times on average to be incarcerated themselves.[14] While these are only a few of the implications on inmates and their families, it is clear that something must be done in a society where African Americans are incarcerated in state prisons at a rate of almost 5.1 times whites. It is also important that we again mention that we did find a solution that had a 2% higher accuracy and a roughly $4,000,000 decrease in societal costs. While this solution may appear to be better, the test data revealed a significant racial disparity in falsely flagging Caucasians as future criminals at a rate 16% higher than African Americans. Achieving fairness for African American defendants should not come at the expense of another racial group. To achieve fairness across all groups, this minor penalty must be incurred. Our solution will no longer allow COMPAS to subject African American defendants and their families to unjust incarcerations that can bring these devastating implications while also remaining fair to the other racial groups.

---

[7] Julia Angwin and Jeff Larson, “Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say,” ProPublica, December 30, 2016, accessed May 6, 2020, https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

[8] William Dieterich, Christina Mendoz and Tim Brennan, “COMPAS Risk Scales: Demonstrating Accuracy Equity and predictive parity.”

[9] Songül Tolan, Marius Miron and Emilia Gómez and Carlos Castillo, “Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia,” *ICAIL '19: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (2019): 83-92, https://chato.cl/papers/miron_tolan_gomez_castillo_2019_machine_learning_risk_assessment_savry.pdf

[10] Moritz Hardt, Eric Price and Nathan Srebro, “Equality of Opportunity in Supervised Learning,” *30th Conference on Neural Information Processing Systems* (2016), http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

[11] American Bar Association, “Collateral Consequences of Criminal Convictions Judicial Bench Book,” The National Inventory of Collateral Consequences of Criminal Convictions, March 2018, https://www.ncjrs.gov/pdffiles1/nij/grants/251583.pdf

[12] Ibid.

[13] Eric Martin, “Hidden Consequences: The Impact of Incarceration on Dependent Children,” National Institute of Justice, March 1, 2017, accessed May 6, 2020, https://nij.ojp.gov/topics/articles/hidden-consequences-impact-incarceration-dependent-children

[14] Ibid.