Enhancing Accuracy and Explainability of Recidivism Prediction Models

Tammy Babad, Soon Ae Chun
City University of New York - College of Staten Island, NY 10314, USA
tammybabad516@gmail.com, Soon.Chun@csi.cuny.edu

Abstract

Predicting recidivism is a challenging task, but it helps support courts in their decision-making process. Automated prediction models suffer from low accuracy and are associated with criticism for biased and unexplainable decision making. In this poster, we present different machine learning models with just a few selected features that achieve accuracies as good as models that use larger sets of features. In addition, we investigate the influencing features that contribute to recidivism prediction, which can enhance the explainability of the learned models.

Introduction

Recidivism risk assessment by algorithm is being used throughout the criminal justice system. Systems like COMPAS (Correctional Offender Management Profiling System for Alternative Sanctions), PSA (Public Safety Assessment) and LSI-R (Level of Service Inventory Revised), are influencing decisions on parole, bonds, criminal sentencing, rehabilitations, etc. Problematically, these systems also suffered from inherent bias (Angwin et al. 2016). Recidivism risk prediction algorithms can range from simple linear combination models to complicated neural networks (Brennan et al. 2008; Larson et al. 2016; Barenstein, 2019). These models are often criticized for a lack of transparency, interpretability, and fairness. In this poster, we use blackbox ML models with very few features for predicting recidivism. Through four different experiments, we examine if the predicted results can be explained with the influencing features.

Related Work

Much of the literature on AI and recidivism focuses on what it means to be fair or transparent. There are those that argue proprietary models, like COMPAS, are not transparent and therefore unfair (Rudin et al. 2020). Others argue that given the nature of the problem, it is impossible to meet certain measures of fairness (Chouldechova 2017). Still others suggest getting rid of machine learning models in favor of AI-Human hybrid models or solely human based decision makers, since the accuracy of these methods tends to be similar to their machine learning counterparts (Tan et al. 2018; Dressel et al. 2018). Most saliently a 2022 study pinpointed several models with high accuracies that might replace tools like COMPAS (Wang et al. 2022).

Data Sets

For this project, we used datasets from ProPublica and the National Institute of Justice (NIJ). The ProPublica dataset consists of 13,419 data points, each one representing an arrested citizen. Although this dataset was first analyzed by ProPublica, it was gathered by the Broward County Police Department. Each data point includes demographic information about the subject, their criminal history, the COMPAS score, and whether the subject did recidivate within the span of three years. The dataset primarily contains males under the age of 35. The NIJ dataset is also skewed towards men, however, the age range is more balanced.

The NIJ dataset has roughly 26,000 data points and includes more nuanced information regarding priors, employment, relocation, dependents, and other relevant features. The dataset was gathered by the Georgia Department of Community Supervision (GDCS) and the Georgia Bureau of Investigation (GBI). This data is publicly available on the NIJ website for use in statistical and research environments. Similar to the ProPublica dataset, this dataset measures recidivism within a three-year period.

Based on the two data sources, we created four experimental datasets. Table 1 shows these datasets with their respective features, sources, and number of records. Dataset 4 combines the NIJ dataset with the ProPublica dataset. The integrated dataset required the unifying of features into the same format. For instance, we calculated jail time by converting two features from strings to date time objects, subtracting them and converting them into categorical values. Additionally, we combined multiple features for priors in the NIJ dataset into one priors_count feature to match with the Broward county dataset.

Recidivism Prediction Models

We developed ML models to predict recidivism. We included the following models in our experiments: Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Linear Support Vector Machine (SVC), and Shallow Neural Networks (NN). These models allow us to pinpoint influencing features in the decision-making process. For instance, the DT model generates a decision tree, allowing one to trace how the model makes it's decisions. RF models allow one to see the top features influencing the decision-making process. In addition to RF, models commonly used to predict recidivism in the literature include SVCs and LRs. We also experimented with NNs because simple NNs can often outperform many other varieties of ML. We split each dataset into 75/25% for training and testing.

Results & Discussions

We compare the performance of different models on the four experimental datasets using the overall accuracy of each model, summarized in Table 2.

As per the United States Sentencing Commission, time spent incarcerated tends to have a large impact on recidivism (Cotter, 2022). This is why experiment 2 includes prison years in addition to the three features from experiment 1. With this added feature, the accuracy of most of our models decreased, except for the NN, which increased dramatically. RF also increased slightly to .697.

We also compared the performance of each model by measuring the accuracy in AUC. Figure 1 shows the AUC for the models from experiment 1 using dataset 1.

In this figure, the ROCs for each model are plotted and their AUCs are named. The blue line, representing the RF classifier, exhibits the highest AUC score. The closer an AUC score is to 1, the better the model performed. In general, the more closely a ROC curve hugs the top left corner, the better the fit of the model is. In this case one can see that the blue curve bows most heavily towards the upper left corner. It is closely followed by the red curve (SVC). Interestingly, while in all other cases the accuracy was higher than the AUC, in the case of the neural networks, the AUC was much higher. In Experiment two it was .78 using a shallow NN.

Influencing Features and Explainability

We analyzed model decisions and their influencing features in predicting recidivism. The RF model generates a ranked list of important features as shown in table 3. In experiment 2, we observed that, when all hyper perimeters are left untuned, the number of years spent in prison usurped the number of priors as the most important feature in the RF model. Conversely, when the hyper parameters *max_depth* and *min_samples_leaf* are tuned in a particular way, *priors* becomes the most important feature, followed by *age* and then *prison days*. When *amount of time spent incarcerated* is the most important feature, *number of priors* becomes about a fifth as important as it was before the *amount of time incarcerated* was added to the equation.

Our findings show that in all of our tested models - all of which performed on par with, or slightly below, what other authors have gotten on this dataset (Dressel, 2018; Rudin, 2020; Tan, 2018; Wang, 2022) - priors and age are primarily important features. Adding prison years as in Experiment 2 puts that at the forefront of the decision making in the ProPublica dataset, but it is significantly less important in the NIJ and joint datasets where priors are still the most important features by far.

The relationship between age's importance and the addition of the incarceration time and hyperparameter constraints is interesting, and something I would like to explore further, with the use of influence measures.

The DT model achieved the fourth highest accuracy (however not AUC) and provides a visual decision tree. This tree, however, is difficult to read since it is incredibly wide. By restricting the tree to a readable size, with slight suffering in accuracy (less than .012), one can identify decision making features and rules to provide transparency in recidivism prediction.

The Decision Tree depicted in Figure 2, is that of experiment 1. Here the features were Priors count, age, and gender. Similar to the RF analysis seen in table 3, priors count seems to be of the highest significance in this model, as it appears at the top of the decision tree. Once more, it is followed by age. While the ways in which the tree reaches a case of recidivism are varied, there are only two paths that lead to no recidivism. In both cases the subjects have more than 4 priors. On one path they are younger than 31.5. The second path taken to no recidivism, is when the subject is older than 31.5 and has more than 16 priors and has more than 11 priors.

Conclusion & Future Studies

In this paper, we created four different datasets for training recidivism classification models which achieved comparable results to other existing risk scoring systems while using a small number of features. In addition, we used features importance tables and decision tree paths to explain the model predictions. We plan to implement a tool that can generate the explanation for the model's decision making by leveraging the feature importances and rules we discovered by running these experiments. Future work includes deep learning models and more in-depth analyses on fairness, age and other economic/social and environmental features to enhance accuracy and create a more holistic approach in predicting recidivism.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. 2016. Machine bias. In *Ethics of data and analytics* 254-264.

Barenstein, M. 2019. ProPublica's COMPAS data revisited. At *arXiv preprint.* Available at: arXiv:1906.04711

Brennan, T., Dieterich, W. and Ehret, B. 2008. Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment system. In *Criminal Justice and Behavior 36(1),* 21–40.

Chouldechova A. 2017 Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data 5(2)* 153–163. Available at: https://doi.org/10.1089/big.2016.0047

Cotter, R. 2022 Length of incarceration and recidivism. In *United States Sentencing Commission*. Available at: https://www.ussc.gov/research/research-reports/length-incarceration-and-recidivism-2022

Dressel, J., & Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. In *Science advances, 4(1)*.

Larson, J., Mattu,S., Kirchner, L., & Angwin, J. 2016. How We Analyzed the COMPAS Recidivism Algorithm. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Rudin C., Wang C., Coker B. 2020. The Age of Secrecy and Unfairness in Recidivism Prediction. In *Harvard Data Science Review.* Available at: https://doi.org/10.1162/99608f92.6ed64b30

Tan S., Adebayo J., Inkpen K., Kamar E. 2018. Investigating Human + Machine Complementarity for Recidivism Predictions. Available at: https://doi.org/10.48550/arXiv.1808.09123

Travaini G., Pacchioni F., Bellumore S., Mosia M., De Micco F. 2022. Machine Learning and Criminal Justice: A Systematic Review of Advanced Methodology for Recidivism Risk Prediction. In *International Journal of Environmental Research and Public Health 19(17):10594*.

Wang, C. Han B., Patel B., Rudin C. 2022. In pursuit of interpretable, fair and Accurate Machine Learning for criminal recidivism prediction. In *Journal of Quantitative Criminology.*

Appendix

Table 1 Experimental Dataset Information

	Features	**Sources**	**Number Records**
Dataset 1	age, gender, # of priors	ProPublica	13,419
Dataset 2	age, gender, # of priors, # of years in prison	ProPublica	13,419
Dataset 3	age, gender, education level, number of dependents, # of years in prison, delinquency report, number of residency changes, 13 kinds of priors.	NIJ	25,835
Dataset 4	age, priors, gender, time spent in jail, # of priors	ProPublica & NIJ	39,254

Table 2 Accuracy of Models in Four Experimental Datasets

Accuracy	**DT**	**Log Reg**	**RF**	**SVC**	**NN**
Dataset 1	**0.703**	0.693	0.696	0.695	0.702
Dataset 2	0.697	0.690	0.697	0.693	**0.713**
Dataset3	0.678	0.699	**0.705**	0.701	0.692
Dataset 4	0.675	0.667	**0.676**	0.671	0.672

Table 3 Feature Importances in Four Experiments in RF

	Dataset #1	Dataset #2	Dataset #3	Dataset #4
Rank #1	Priors Count (.58)	Prison Days (.57)	Percent Days Employed (.12)	Priors Count (.69)
Rank #2	Age (.37)	Age (.25)	Prior Arrest Episodes Felony (.06)	Age 48+ (.07)
Rank #3	Male (.02)	Priors Count (.16)	Prior Arrest Episodes PPViolation Charges (.05)	Age 18-22 (.04)

Figure 1 AUC Plot for Experiment #1

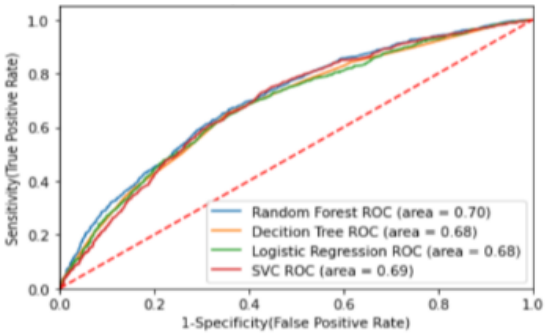

Figure 2 Decision Tree Path