Interpreting Recidivism Prediction Models

Alasia Buschkopf
University of Missouri-St. Louis

May 8, 2024

Abstract

Introduction America's prison and jail population is over 2 million and rising. There are many approaches to try and solve the mass incarceration problem in America, many of which involve the concept of recidivism. Recidivism refers to the likelihood of a defendant to re-offend after release. Models to predict recidivism are currently in use across the country by judges, parole boards, and other agencies, sometimes as a part of the decisions regarding a defendant's sentencing. In the case of black box models, the lack of transparency is upsetting and leaves the model at risk for unintentional bias. To make the use of these models equitable for all, it is imperative that the public have an understanding of how they work.
Methods This project analyzed a neural network trained on a dataset containing 54 data fields on 26,000 defendants from the state of Georgia. The dataset had a baseline accuracy of 57.7% and the model(s) achieved accuracy ranging from 70% to 74%. While this is not the highest accuracy, it is still higher than the accuracy of some models in use today. The low accuracy may affect the interpretability, however.
Results The main model was analyzed using the LIME library. The top features contributing to the predictions were consistently percent days employed, gender, age at release, and gang affiliation. To try to further explain the model, it was trained on datasets split on top features. The model was trained on only male and only female data. The explanations for only male data lined up with the general model, but there was more nuance in the only female data model explanations. The possibility of different features contributing to accurate vs inaccurate prediction instances was also explored, though nothing significant was noted.
Discussion It is difficult to determine if the model is equitable based on just this analysis alone. Further exploration on the characteristics of the subsets of accurate vs inaccurate predictions, further splitting the data and training models for more in depth explanations, and determining if other features may be represented by proxy in the features in the explanations would all be helpful for further consideration.

Contents

List of Tables

List of Figures

1 Introduction

America has a mass incarceration problem, with more than 2 million people in prisons or jails [1], and it is getting worse. There are many approaches being taken by justice organizations and non-profits to attempt to reduce the prison population. Many of these efforts involve the concept of recidivism. Recidivism "refers to a person's relapse into criminal behavior, often after the person receives sanctions or undergoes intervention for a previous crime." Recidivism rates are used in many ways, such as measuring the success of intervention programs. Many levels of the justice system use recidivism prediction models to determine the risk of a person committing an additional criminal act within 3 years after release. These predictions are used by parole boards, judges, and government agencies nationwide when making life altering decisions for defendants. Many of these recidivism prediction models are not transparent, raising concerns for how they work and what bias may exist within them. ProPublica's "analysis of Northpointe's tool, called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions), found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk." [2] Without understanding how a model is reaching a prediction, it is difficult to trust. Interpreting these models can help to identify bias that may exist within them and provide more equitable outcomes for defendants. Additionally, understanding how the models work can help the judicial system with rehabilitation and intervention programs for parolees with high risk factors as determined by these models.

2 Methods

2.1 Dataset and Models Analyzed

Many tools exist to help predict the risk of a defendant re-offending. These tools range in complexity from simple assessments, to complex algorithms, to commercial models with no transparency. Many of the models currently in use are proprietary or difficult to gain access to the model itself. In order to do many interpretation methods, access to the model is desirable. For this project, a neural network model was the primary model analyzed.

Some prediction models utilize data from a questionnaire given to defendants when they are initially booked in jail, while some utilize data based specifically on the defendants' criminal history and general demographic information. The dataset used for this project was the NIJ's Recidivism Challenge Dataset. [3] This is a tabular dataset containing 54 data fields on nearly 26 thousand defendants. The data is from 2021 was provided by the Georgia Crime Information Center. Data fields include gender, race, employment history, gang affiliations, age at release, length of prison term, number of children, education level, and types of prior convictions.

The dataset was preprocessed to convert any non-numerical columns to numerical values and one hot encoded. The data was randomly shuffled and then split into 80% training and 20% validation.

2.2 Model Accuracy

The dataset contains 25,835 people and 14,904 of those are classified as True, meaning they did re-offend within 3 years of release. Therefore, this dataset has a baseline accuracy of 57.7%. The models analyzed in this project ranged in accuracy from 70% to 74%. While this is not a high accuracy rate, it is still higher than baseline accuracy and higher than the accuracy found in some models currently in use today. However, it is important to keep in mind the low accuracy when considering any interpretation methods. If the model is not performing well, the explanations will be less meaningful.

3 Interpretation and Results

3.1 Local Interpretable Model-Agnostic Explanations (LIME)

LIME is an interpretability method that works by "training local surrogate models to explain individual predictions." [4] When using LIME, an explainer is set up to try and explain which features contributed most to the particular prediction being analyzed. Perturbations surrounding that instance are then randomly generated and fed into the black box model to determine how the model will classify them. Using these randomly generated new data points and their associated predictions, a linear model is developed. Using this linear model, LIME can determine how, locally, the model seems to be making decisions. It will then generate an explanation, in that it will determine the features which had the highest weight according to the simpler, local model. Because this is a local model, there may be accuracy issues with LIME interpretations if the instance that is being explained is too near to an inflection point in the model. A LIME explainer was run using the initial Neural Network model on randomly selected instances of the validatation dataset. LIME Notebook

The first explanation instance shows the features with the biggest impact on the model's prediction, according to a local surrogate model. The model predicted an 84% chance that the defendant would re-offend within three years (True). The actual label for this defendant was also True. The features which had the strongest impact on the prediction that the criminal would commit another crime were 0% days employed, gender being male, their age at release being younger than 48, and their gang affiliation. (see **Figure 1** and **Figure 2**).

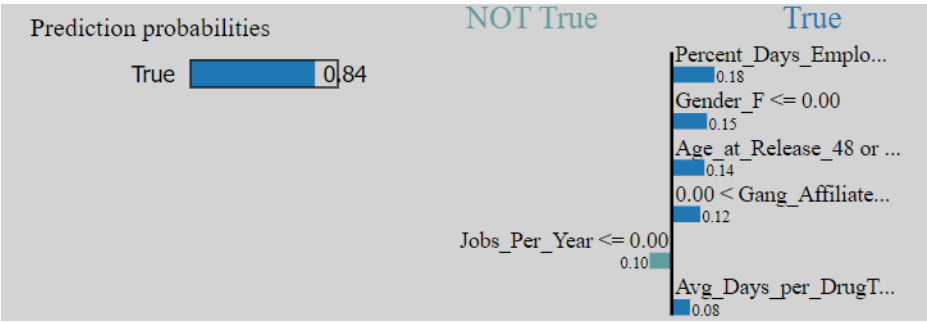


Figure 1: LIME Explanation Instance 1 - Actual Label: True

Feature	Value
Percent_Days_Employed	0.00
Gender_F	0.00
Age_at_Release_48 or older	0.00
Gang_Affiliated	1.00
Jobs_Per_Year	0.00
Avg_Days_per_DrugTest	236.00

Figure 2: LIME Explanation Instance 1 Values - Actual Label: True

In a second instance, the model predicted a 59% chance that the defendant would re-offend within three years (True). The actual label for this defendant was False. The features which had the strongest impact on the prediction that the criminal would commit another crime were gender being male and their age at release being younger than 48. However, both 100% days employment and no gang affiliation also were very important features towards classifying this defendant as False. (see **Figure 3** and **Figure 4**).

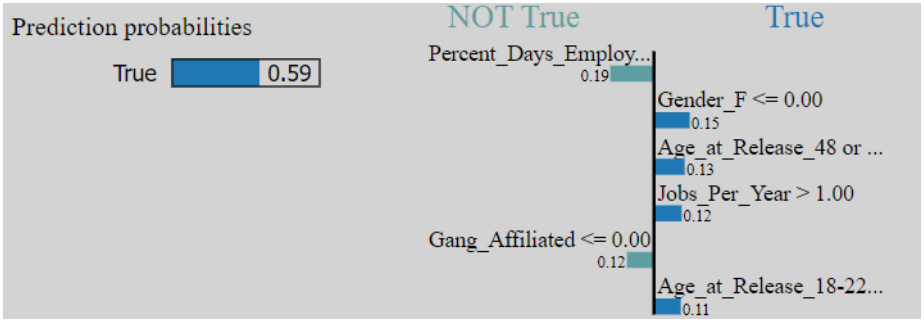


Figure 3: LIME Explanation Instance 2 - Actual Label: False

Feature	Value
Percent_Days_Employed	1.00
Gender_F	0.00
Age_at_Release_48 or older	0.00
Jobs_Per_Year	3.00
Gang_Affiliated	0.00
Age_at_Release_18-22	1.00

Figure 4: LIME Explanation Instance 2 Values - Actual Label: False

In a third instance, the model predicted a 37% chance that the defendant would re-offend within three years (False). The actual label for this defendant was False. The features which had the strongest impact on the prediction that the criminal would not commit another crime were gender being female, 73% days employed, and not being in the age group of 18-22. The two main features that caused the model to increase the probability of recidivism were age not being over 48 years and a gang affiliation. (see **Figure 5** and **Figure 6**).

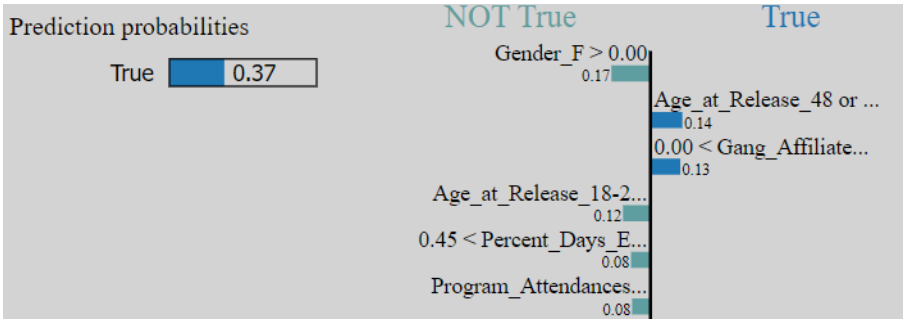


Figure 5: LIME Explanation Instance 3 - Actual Label: False

Feature	Value
Gender_F	1.00
Age_at_Release_48 or older	0.00
Gang_Affiliated	1.00
Age_at_Release_18-22	0.00
Percent_Days_Employed	0.73
Program_Attendances	10.00

Figure 6: LIME Explanation Instance 3 Values - Actual Label: False

Multiple other instances were examined with similar results. This led to the conclusion that gender, percent days employed, age, and gang affiliation were among the top four features used by this model when predicting recidivism.

3.2 LIME on Models Trained With Feature Separated Datasets

Some sources have noted that splitting the datasets and building separate models for male and female defendants resulted in higher results. Considering this, modified models were trained by splitting the data into subsets based on the top features determined by the previous LIME explorations . Once these models were trained, LIME was then applied to random instances of the validation dataset on each model to see how the prediction's explanation varies when the top features are split out.

When the model was trained on only female datapoints, LIME explanations seemed to prioritize features that didn't even appear in the top 4 of the mixed gender trained model. When trained on only male datapoints, the explanations were similar to that in the mixed gender trained model, albeit without the gender feature.

A sample instance of the female data only model predicted a 42% chance that the defendant would re-offend within three years (False). The actual label for this defendant was False. The feature which had the strongest impact on the prediction that the criminal would not commit another crime was their age being 48 or over. Interestingly, despite the false classification, the top three features actually point towards a true classification. Those three features that caused the model to increase the probability of recidivism were program attendance being zero, prison years not being greater than 2-3, and 0% days employed. The varying top features was common

among other female only model instances. (see **Figure 7** and **Figure 8**).
Female Data Only Notebook

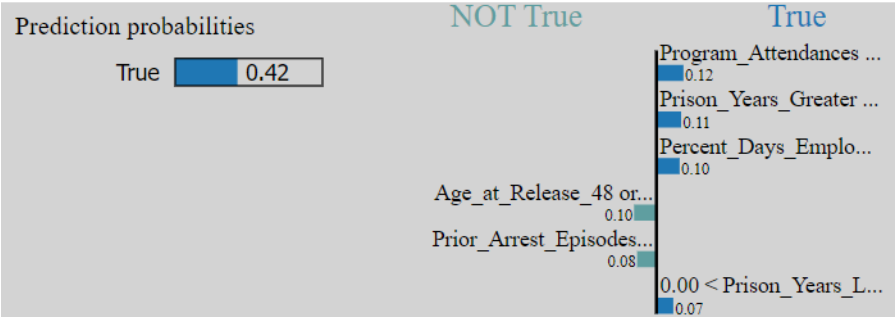


Figure 7: LIME Explanation - Female Only Dataset Instance 4 - Actual Label: False

Feature	Value
Program_Attendances	0.00
Prison_Years_Greater than 2 to 3 years	0.00
Percent_Days_Employed	0.00
Age_at_Release_48 or older	1.00
Prior_Arrest_Episodes_Misd	0.00
Prison_Years_Less than 1 year	1.00

Figure 8: LIME Explanation - Female Only Dataset Instance 4 Values - Actual Label: False

A sample instance of the male data only model predicted a 75% chance that the defendant would re-offend within three years (True). The actual label for this defendant was True. The top features for this prediction were very similar to the mixed gender model, 0% days employed, gang affiliation, and age. This held true for the other instances analyzed for the male only model. (see **Figure 9** and **Figure 10**).
Male Data Only Notebook

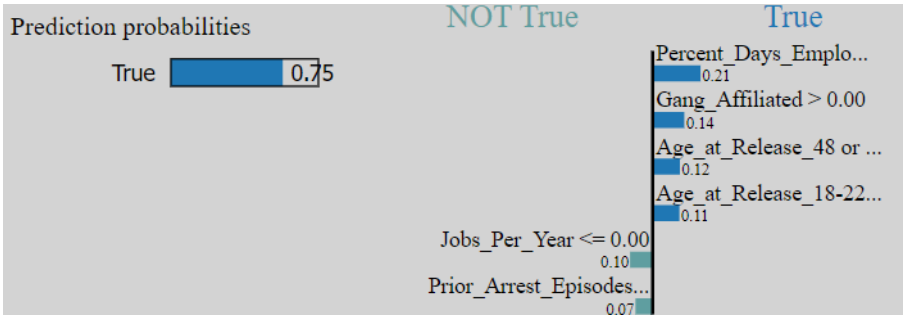
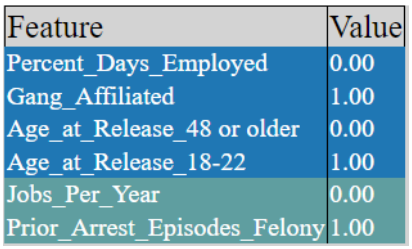


Figure 9: LIME Explanation - Male Only Dataset Instance 5 - Actual Label: True

A similar analysis was performed on a model trained on datasets separated based on gang affiliation. The gang affiliated trained model had similar explanations as the general model and



Feature	Value
Percent_Days_Employed	0.00
Gang_Affiliated	1.00
Age_at_Release_48 or older	0.00
Age_at_Release_18-22	1.00
Jobs_Per_Year	0.00
Prior_Arrest_Episodes_Felony	1.00

Figure 10: LIME Explanation - Male Only Dataset Instance 5 Values - Actual Label: True

the male only model, while the non-gang affiliated trained model had more variance, with features such as prior felonies and delinquency reports appearing in the top four features.

Gang Affiliated Data Only Notebook Not Gang Affiliated Data Only Notebook

3.3 LIME Explanation Summary on Predicted Vs Actual Value Subsets

One concern with a model that is predicting recidivism and being used to make life altering decisions regarding parole and sentencing for defendants, is the population of people who do not re-offend but were predicted to re-offend by the model. In order to explore this subset, the data instances in the validation set were separated into four subsets: 1 - Predicted True, Actual True; 2 - Predicted False, Actual False; 3 - Predicted True, Actual False; 4 - Predicted False, Actual True. Due to computing time limitations, 200 instances were randomly selected from each subset for this process, for a total of 800 instances.
Considering each subset individually, LIME was applied in a loop to the relevant instances and the top four features retained in a list. After explanations for all instances in each subset were identified, the top contributing factors to each subsets' predictions were tallied.
It was of note that the subsets of accurate predictions (Predicted True, Actual True and Predicted False, Actual False) had different top factors when compared to those instances with inaccurate predictions (Predicted True, Actual False and Predicted False, Actual True) (see **Table 1**, **Table 2**, **Table 3**, and **Table 4**). Unfortunately, this analysis did not provide much insight as the top features did not vary significantly between these subsets. It would still be an interesting area to further explore with more time. Predicted Vs Actual Explanations Notebook

Table 1: A table showing the most frequent top 4 features in a sample of 200 predicted true, actual true instances

Predicted True, Actual True	Most Frequent	Second Most Frequent
Feature 1	Percent Days Employed	Gender
Feature 2	Gender	Age at Release
Feature 3	Age at Release	Gang Affiliation
Feature 4	Gang Affiliation	Percent Days Employed

Table 2: A table showing the most frequent top 4 features in a sample of 200 predicted false, actual false instances

Predicted False, Actual False	Most Frequent	Second Most Frequent
Feature 1	Percent Days Employed	Gender
Feature 2	Gender	Age at Release
Feature 3	Age at Release	Gang Affiliation
Feature 4	Gang Affiliation	Percent Days Employed

Table 3: A table showing the most frequent top 4 features in a sample of 200 predicted true, actual false instances

Predicted True, Actual False	Most Frequent	Second Most Frequent
Feature 1	Percent Days Employed	Gender
Feature 2	Gender	Age at Release
Feature 3	Age at Release	Gang Affiliation
Feature 4	Gang Affiliation	Percent Days Employed

3.4 Partial Dependency Plots

After using LIME to narrow down the most important features used by this model, partial dependency plots (PDP) of the top two features (38 - Percent Days Employed and 41 - Gender) were generated. The PDP for Percent Days Employed shows the most variance. This implies that this feature has a significant impact on predictions in relation to the other features. (see **Figure 11**)

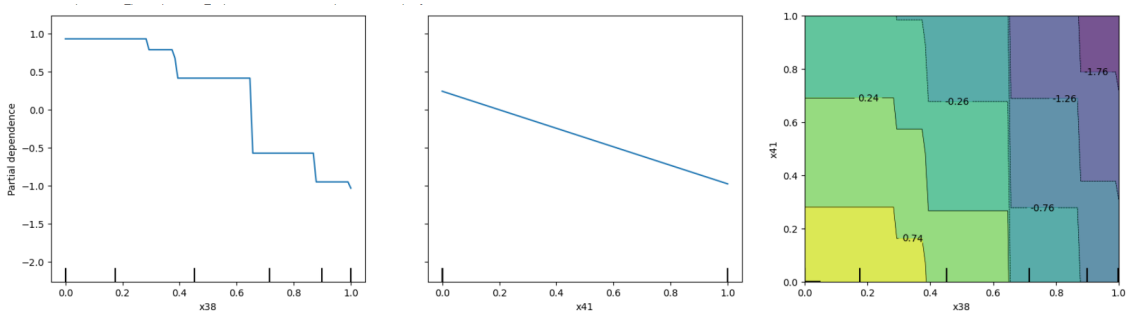


Figure 11: PDPs of Feature 38 (Percent Days Employed) and Feature 41 (Gender)

4 Conclusions

This project attempted to interpret a black box neural network used to predict recidivism likelihood. Using both LIME and PDPs, it was shown that the most important features

Table 4: A table showing the most frequent top 4 features in a sample of 200 predicted false, actual true instances

Predicted False, Actual True	Most Frequent	Second Most Frequent
Feature 1	Gender	Percent Days Employed
Feature 2	Gender	Age at Release
Feature 3	Age at Release	Gang Affiliation
Feature 4	Gang Affiliation	Percent Days Employed

considered by this model when making a prediction are percent days employed, gender, gang affiliation, and age at release.

By separating out the dataset into groups based on top features, other patterns in how a model reaches a decision were noticed. Models trained on datasets containing males only and gang affiliated only show similar prediction explanations as the general model. However, models trained on datasets containing females only and non gang affiliated only had more nuance in their explanations, with features such as delinquency reports, prior felonies, and others appearing in their top features.

The possibility of different features explaining accurate vs inaccurate predictions was also considered by tabulating the top features for samples from each subset of predicted vs actual instances. While there were no significant differences found in this experiment, it would be an interesting area for further analysis, including comparing the populations in each of these subsets.

Based on the analysis of this project, it appears that employment, age, gender, and gang affiliation are the main features explaining how this recidivism prediction model is working. Whether or not these features would be considered biased or inequitable features to base parole and sentencing decisions on is it's own topic. It is also important to consider the other demographic features that may be represented by proxy in these features. Next steps might include analyzing the demographics of each subset of predicted vs actual, performing feature importance analysis, and splitting data out into more subsets to train models to review interpretations.

References

[1] ACLU. https://www.aclu.org/issues/smart-justice/mass-incarceration. Accessed: 2024-05-06.

[2] Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin . How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, 2016.

[3] Jational Institute of Justice - Recidivism Challenge Full Dataset. https://data.ojp.usdoj.gov/Courts/NIJ-s-Recidivism-Challenge-Full-Dataset/ynf5-u8nk/about_data. Accessed: 2024-05-06.

[4] Christoph Molnar. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. Independently Published, 2023.