Recidivism: Relapse Into Prison

MGT 6203: Data Analytics for Business

Group Project Final Report

December 1st, 2023

Team #34

Andrea Ruiz Marquez, Christina Alyse Young, Praswas Shakya, Rekha Ran

Data Source: https://www.kaggle.com/datasets/uocoeeds/recidivism

Github: https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-34

Project Background: Recidivism in the US

The US has one of the largest prison populations in the world, and it spends more than $80 billion annually on the prison system. With the social costs factored in, that amount is closer to $500 billion (National Institution of Corrections). While there are programs and assistance to rehabilitate former inmates and help bring down recidivism rates, there are limited resources available to accomplish this. This paper looks at where limited resources for rehabilitation should be directed in order to reduce recidivism rates. We looked at which portions of the population of former inmates are at higher risk of failing to adapt to life outside of prison so that rehabilitation programs can target these populations more effectively. This will benefit taxpayers in the US since there will be a lower prison population to pay for, it will benefit former inmates at higher risk of recidivism since they will receive more help at adapting to life outside of prison, and it will benefit the families and communities of former inmates.

While there is not a large amount of research done on the effectiveness of machine learning models predicting recidivism, there have been a few studies that achieved high accuracy in predicting recidivism. These studies used a variety of different datasets, from a women's correctional institution in Thailand (Bulsara) to Ukrainian penitentiary institutions (Kovalchuk). Our project will provide additional insight into the effectiveness of machine learning models to predict recidivism. By using regression models on another dataset, we will be able to study if the variables and features that were found to be important in other datasets hold for other populations and provide additional evidence for or against the effectiveness of machine learning models to predict recidivism.

Description of Dataset

(https://www.kaggle.com/datasets/uocoeeds/recidivism)

We obtained the dataset used in this study from Kaggle. Some of the key columns provided are age, race, gender, education, drug use, gang affiliation, prior arrests and convictions, and whether the individual was arrested within three years of being released from prison. Each column in the dataset represents a variable related to an individual and each row in the dataset represents an individual that served a sentence in prison. The sample size was 25, 835.

Initial Hypothesis

Our initial hypothesis was that less education, lack of dependents, older age, and longer incarceration time would increase the probability of recidivism within 3 years. After looking at the relationships between these variables and recidivism, we found that we were wrong in many of our initial assumptions. Education and dependents didn't have a large effect on recidivism, and older people were actually much less likely to return to prison. While it is beyond the scope of this paper to explore why these factors had these effects on recidivism, it would be a useful area for future study.

Methodology

One of the first steps to model building was to reduce the predicting variables of the dataset by using only relevant data and getting rid of noise. Initially, the dataset contained 49 predicting variables, to avoid overfitting and make our model easier to interpret and use in practice, we started to see how we can remove independent variables without adding bias. We divided the dataset into a training set, a validation set, and a testing set. We then removed the race variable, due to the many issues with racial biases that can be caused in machine learning algorithms (Travaini). We also removed all female values in the gender variable, since the prison population in the US is over 90% male, it makes sense to focus on male recidivism, since this would have the biggest impact. The dataset we have was also systematically missing some of the variables for females, such as gang affiliation. We removed a location variable since there were many different places covered by the location variable, and creating many features by hot encoding the data would make our results convoluted. Finally, we removed a variable that measured the average number of days between a drug test.

Exploratory Data Analysis

The key influencers visual in Power BI provide insights into the factors that drive a metric, in this case, Recidivism within 3 years. It analyzes data and ranks the independent variables and designates key influencers. Power BI uses regression analysis and calculates how the dependent variable changes based on the independent variables. It is trying to find correlation between variables and contrasts the importance of these factors.

The top key influencer is gang affiliation. When Gang affiliated is True, the likelihood of recidivism within 3 years being true increased 1.42 times. The second influencer is Prior arrest episodes of parole or probation violations. When the prior arrest episodes for parole and probation violation was 5 or more, the likelihood of recidivism within 3 years being true, increased 1.34 times. Similarly, the third key influencer is Prior arrest episodes - Property charges. When prior arrest episodes - property charges were 5 or more, the likelihood of recidivism within 3 years increased by 1.28 times. The fourth key influencer is prior convictions episodes. If a person has 3 or more convictions on property charges, the likelihood of recidivism within 3 years being true increased 1.27 times. In Fig. 1 and Fig 2., we have the output of what was described above

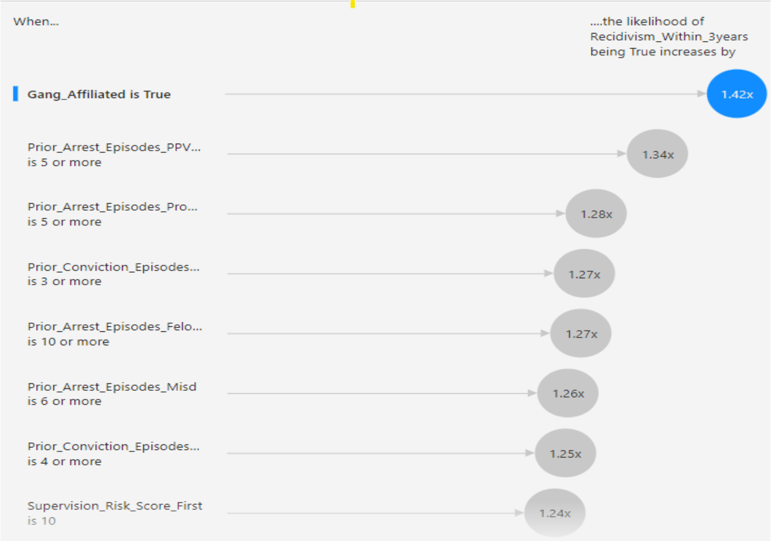


Fig. 1: Top influencer of recidivism

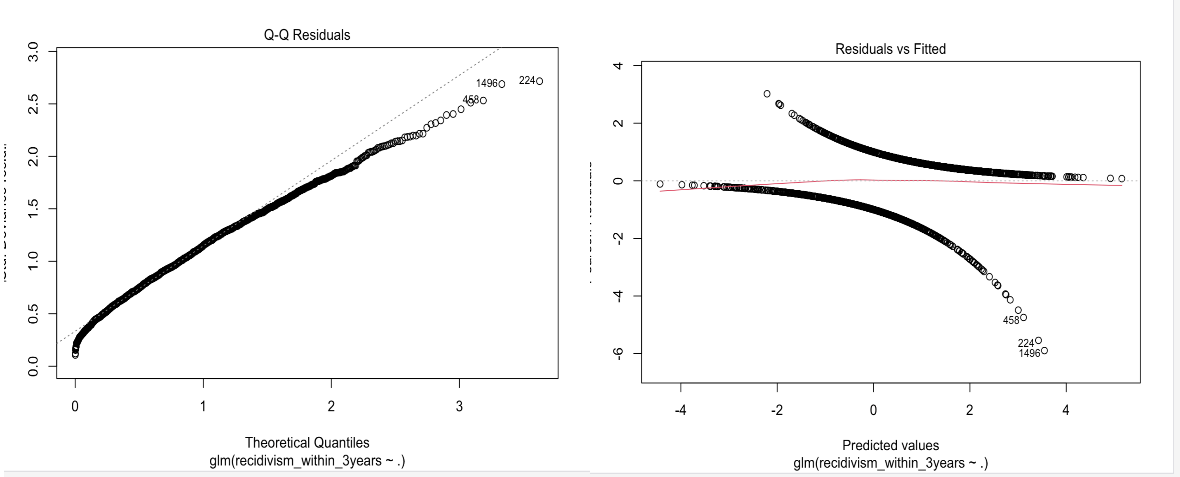
Top 10 Key Influencers	**Likelihood of recidivism being true increases by:**
Gang affiliated is True	1.42X
Prior Arrest Episodes PPV is 5 or more	1.34X
Prior Arrest Episodes Property is 5 or more	1.28X
Prior Conviction Episodes Prop is 3 or more	1.27X
Prior Arrest Episodes Felony is 10 or more	1.27X
Prior Arrest Episodes Misdemeanor is 6 or more	1.26X
Prior Conviction Episodes Misdemeanor is 4 or more	1.25X
Supervision Risk Score First is 10	1.24X
Age at release is 18-22	1.24X
Condition MH SA is true	1.22X

Fig. 2: Top 10 influencers of recidivism and likelihood ordered greatest to least.

Residuals Plots

To get better insight into a model checking graphically examining plots of the residuals.

The Q-Q Plot the data points fall within line, so it means that it follows the normal distribution. Most of the extreme data points are on the top right. The residual vs fitted plot shows if the data follows a linear pattern or not. There was no nonlinearity seen on the data.



Variance Inflation Factor (VIF)

The most common way to detect Multicollinearity is by using VIF which measures the correlation and strength of correlation between the predictor variables. Through the VIF function it is determined that most of the variables lie between 1-2 by using the following formula: (GVIF^(1/(2*Df)) that is square root of GVIF since most of the variables are categorical variables with just two levels. This indicates the absence of strong multicollinearity.

Model Evaluation

For our initial model we ran a logistic regression on all of the variables that were not removed in our initial analysis. We found that many of the features were highly statistically significant, such as age at release, whether the person had tested positive for various drugs, or if the person had been arrested many times previously. The area under the curve (AUC) for our regression was .769, and with a cutoff point of .5, the model achieved an accuracy of .72. The advantage of logistic regression is the easy interpretability of the results, and this model provides useful information about how much the factors in our dataset affect recidivism.

Figure 3 shows the statistically significant scaled coefficients from the logistic model and their confidence intervals. The coefficients are standardized in order to make it easier to compare the influence of all of the features in the model. As shown in Figure 3, the features that most decrease the chance of recidivism are an older age at release, a higher percentage of days employed, and higher attendance in rehabilitation programs. The features that most increase the probability of recidivism are high numbers of previous felony and misdemeanor arrests, delinquency reports, and being affiliated with a gang.

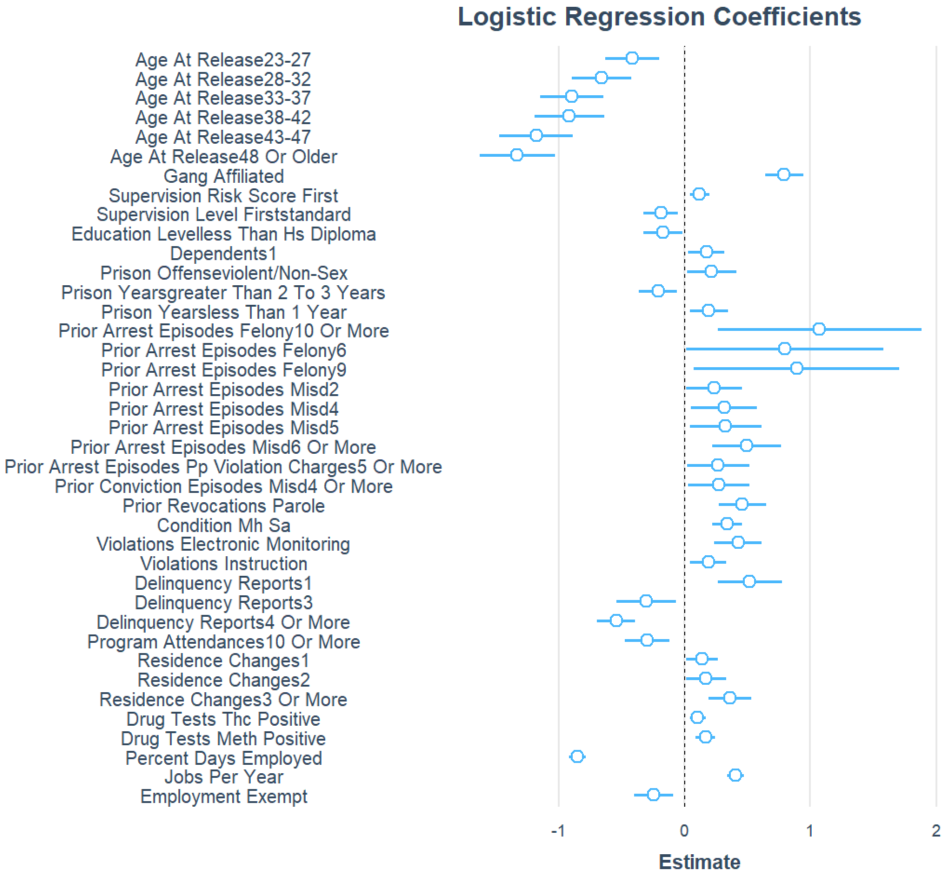


Fig. 3: Scaled coefficients of logistic model.

While logistic regression is useful for explaining how much the factors in our dataset affect recidivism, it is likely due to overfitting due to the large number of variables. We experimented with several different methods of variable selection to lessen the effects of multicollinearity and reduce overfitting in our model. First, we ran a stepwise regression, which removes and adds back in variables to lower the AIC. The variables that the stepwise model chose were similar to the key features we found while doing our initial data exploration in Power BI. Some of these features were age, gang affiliation, drug test results, and prior arrest record. The AUC of the model was .7603, still slightly lower than using all of the variables. However, since a stepwise regression is a greedy algorithm that can stop at a locally optimal result instead of the overall best result, we moved on to other methods of variable selection.

Next, we tried lasso regression, ridge regression, and elastic net regression and achieved similar AUC values for each of them: .7683 for lasso regression, .7688 for ridge regression, and

.7683 for elastic net regression. We chose elastic net regression as our final model to test our validation data set. Although our elastic net regression had a slightly lower AUC than our initial logistic regression that included all variables, elastic net can reduce overfitting and addresses the problem of multicollinearity. This means that our elastic net model was less likely to learn false patterns from the training data. Elastic net combines the feature selection of lasso with the ability of ridge to lessen the effects of multicollinearity and is a good compromise between the two. It is well-equipped to handle multicollinearity and overfitting.

Type	**AUC**
Stepwise Regression	0.7603
Logistic Regression	0.769
Lasso Regression	0.7683
Elastic Net Regression*	0.7683
Ridge Regression	0.7688

* Selected model for prediction

We ran a final elastic net regression on the test set. The purpose of this final test was to see what unbiased results we would expect our model to get on new data. Since we tested the validation data set on multiple models, it is likely that random chance played a part in our best model achieving as high accuracy on the validation data as it did. Using a new and untested data set allows us to account for this bias. With the test data set, our elastic net regression got an AUC of 0.7683. Figure 4 shows the AUC plot for the elastic net regression predicting our test set values.

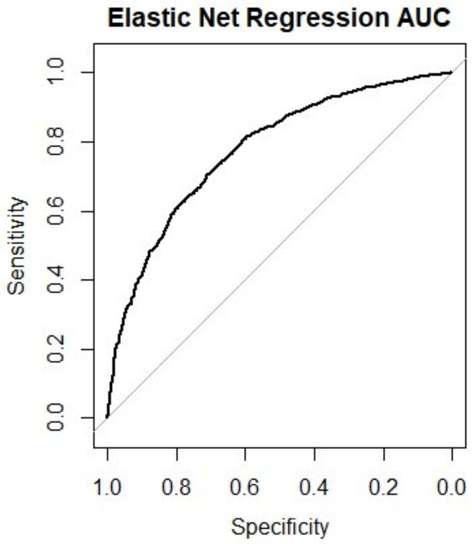


Fig. 4: Elastic Net Regression, AUC

Conclusion

In conclusion, the factors that are most likely to cause recidivism are younger age, unemployment, failure to attend rehabilitation programs, affiliation with a gang, prior arrests, and drug use. These results provide evidence of the importance of rehabilitation programs to help former prisoners integrate into society. Helping younger people who have spent time in prison find employment and a healthy community are likely to lower recidivism rates. Helping former prisoners stop drug use is also likely to lower the probability of recidivism. These results give insight into where rehabilitation programs can focus their efforts. If a rehabilitation program has limited funds and staffing and is unable to help all former prisoners in its vicinity, this study shows that the program can be most effective in reducing recidivism rates by focusing on younger, unemployed people with gang affiliation.

Even with experimenting with various algorithms, we were unable to achieve very high accuracy in predicting whether or not a specific person would return to prison. There are many factors that lead to a person returning to prison, and in this paper, we were only able to look at a limited number of those factors. To create a successful machine learning model that predicts whether a former prisoner should have access to a rehabilitation program, a model would need many features, and it would be important that it had better accuracy than current methods of predicting recidivism probability. In addition, since our dataset was limited to one specific geographic region, it is possible that our findings do not carry over to other regions. However, our model was still able to provide useful information about which factors on average can increase recidivism rates.

If given more time or resources, it would be useful to test the dataset's specificity to a particular geographic location since it is likely that our dataset does not fully represent the diversity of recidivism patterns across different regions. Expanding the demographic scope would also help us to better understand how recidivism varies among different racial and socioeconomic groups. Adding additional variables such as mental health indicators could provide additional insight into the causes of recidivism. It could also be useful to test the ability of other algorithms to predict recidivism, such as a probit model or a clustering model.

Studies that support the Conclusion

Results from a series of proportional hazard models indicate that gang membership, drug dependence, and institutional behavior are critical factors in predicting the timing of reconviction. Contrary to expectations, gun use was not related to post release involvement in the criminal justice system. Institutional misconduct may be an important marker of sustained gang membership, making institutional programming and appropriate aftercare services a priority for this group of offenders. (Huebner, Varona, Bynum)

The evidence presented herein indicates that CBT programs have proven to be the most effective in reducing prison misconduct. Moreover, these programs, including substance abuse treatment and sex offender treatment, have consistently demonstrated success in decreasing recidivism. (Duwe, Grant)

Works Cited:

Anonymous. (n.d.). *The economic burden of incarceration in the U.S (2016)*. National Institute of Corrections. https://nicic.gov/weblink/economic-burden-incarceration-us-2016

Butsara, N., Athonthitichot, P., & Jodpimai, P. (n.d.). *Predicting recidivism to drug distribution using machine learning Techniques.* Semantic Scholar. https://www.semanticscholar.org/paper/Predicting-Recidivism-to-Drug-Distribution-using-Butsara-Athonthitichot/c7b30dfcf98da78376a66772d133d0a033304e25

Kovalchuk, O., Karpinski, M., Banakh, S., Kasianchuk, M., Shevchuk, R., &amp; Zagorodna, N. (2023, March 3). Prediction machine learning models on propensity convicts to criminal recidivism. MDPI. https://www.mdpi.com/2078-2489/14/3/161

Sevigny, E. L., Johnson, T. L., & Greathouse, J. A. (2021, August 31). *Predicting Recidivism Fairly: A Machine Learning Application Using Contextual and Individual Data*. Georgia State University. https://www.ojp.gov/pdffiles1/nij/grants/305036.pdf

Travaini, G. V., Pacchioni, F., Bellumore, S., Bosia, M., & De Micco, F. (2022, August 25). *Machine Learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction*. MDPI. https://www.mdpi.com/1660-4601/19/17/10594

Huebner BETH M. , Varona SEAN P., Bynum TIMOTHY S.(2007, April 7) GANGS, *GUNS, AND DRUGS: RECIDIVISM AMONG SERIOUS, YOUNG OFFENDERS.* https://www.umsl.edu/ccj/files/pdfs/CPP_Huebner.pdf

Duwe Grant (2017, June) *The Use and Impact of Correctional Programming for Inmates on Pre- and Post-Release Outcomes*. https://www.ojp.gov/pdffiles1/nij/250476.pdf