# Looking at Factors Influencing Recidivism

Nidhi Ladda

Babson College

AQM 2000-05: Predictive Business Analytics

Professor Guinta

April 28, 2023

# Table of Contents

Abstract

The study looks into criminal justice reforms by quantifying the impact of numerous factors that influence recidivism, further advocating. It dives deep into unraveling the complex interplay of race, education, prison work programs and age on the rates of recidivism, further encouraging the development of beneficial and efficient strategies. The dataset Recid has been utilized to perform a comprehensive analysis to address our research question. Our approach involved using Logistic Regression and KNN Classification to examine the impact of education and priors along with prison work programs on returning felons respectively. Additionally, we used Linear Regression to understand the role of age on numerous variables related to criminal history and Classification and Regression Trees to investigate the aspect of race.

# Looking at Factors Influencing Recidivism

Recidivism refers to individuals reverting to criminal behavior after prior criminal history or encounters with the justice system. This concept poses significant challenges for community leaders, law enforcement agencies and policy makers around the world. The recurring criminal behavior reflects shortcomings within correctional strategies and Rehabilitation processes along with undermining public safety. Understanding the complex factors that influence returning felons is crucial for creating efficient strategies that reduce the rate of offenses committed by individuals with criminal history.

Recidivism is affected due to numerous factors; but some have a significant impact when compared to others. These include race, education, prison work programs and age. Therefore, we decided to conduct a study reflecting the impact of these factors on returning felons. A study was conducted to prove that "education opportunities in prison are key to reducing crime" with the intention to reduce the growing prison population (Wetzel, 2023). This study also highlighted that "41% of incarcerated individuals do not hold a high school diploma", further showing the role of education on individuals with criminal history (Wetzel, 2023). Additionally, research shows that inmates that worked in prisons are 14% more likely to be employed after release, further proving that they would reduce recidivism rates.

Furthermore, racial disparities have been evident in the justice system for many years and have been the topic of discussion when referring to criminal justice. "According to Nellis, African Americans are 5 times more likely to face incarceration than European Americans" (Thomas, 2016). In addition, Bureau of Justice Statistics states that sixty-three percent of African Americans released from prison have reoffended within the span of 36 months of their release. Therefore, we

believe it relevant to understand the influence of race on returning felons to address systemic inequalities.

In terms of age, studies show that the "younger an offender, the more likely they are to return to prison" (Recidivism of young adults, 2014). The ages 18-24 have the highest rate of reoffending, with 60% of the eighteen and nineteen years old returning to prison. Additionally, if recidivism was measured in terms of felons returning to corrections in general instead of the ones returning to prison, then the recidivism rate for eighteen- to nineteen-year-olds would read 90%. These statistics reflect research that shoes offending peaks in late adolescence. As a result, analyzing the impact of age on returning felons can provide insights into life-stage factors in terms of reoffending. As a result, we decided to address the research question: To what extent does race, age, education and prison work programs have an influence on recidivism.

To address the above stated research question, we choose the dataset Recid as it is the most appropriate one mainly for three reasons. Firstly, the data set includes 16 distinct variables apart from race and felons, further giving us the opportunity to analyze the effect of both a felon's lifestyle and criminal history on recidivism. Additionally, the data incorporates 1445 observations, making it the ideal data set to help draw statistically significant conclusions which might not be apparent or accurate with data of smaller sizes. We also find it interesting that all participants involved in the survey that helped extract the data were individuals with a criminal history involving jail time, further allowing us to gather more insight into returning felons.

## Linear Regression

A linear regression model is created to model the relationship between a dependent variable and one or more independent variables. This relationship is depicted by a line of best fit, the main

goal of creating the model is to find the line that best describes the relationship between the two variables.

We have used linear regression to model the relation between 'age' and 'priors', 'tserved' and 'durat'. Our goal was to see how the age of a convict impacts the prior convictions, time served and duration of the serve. The linear regression model predicts for a convict' age based on their prior convictions, the time served and duration. This model was vital in helping answer the question "To what extent does race, age, education and prison work programs have an influence on recidivism?" since it dealt with answering how age plays a role on whether a convict commits a felon again or not. The effect of age on the prior convictions tells us if age leads to recidivism.

The MAPE of 0.23 and RMSE of 8.552 are relatively low. This means that the percentage error of the model is 23% and on average, each prediction is off by +/- 8.552 years. This indicates that the model's predictions do not deviate significantly from the observed values, which suggests that it has a good predictive performance. In contrast, the benchmark MAPE is 94% and the benchmark RMSE is 28.32, hence our model performs significantly better than the benchmark or average values.

Our model equation is *"Age = 24.08 + 1.66priors - 0.03tserved + 0.058durat"*. Our B0 is 24.08 and this means that when all independent variables are 0, the age is around 24 years. This means that on average if a person is not a prior convict and has not served any time and duration, their age would be 24 years. 1.66 priors is a positive coefficient, and it tells us that for every one unit increase in the number of prior convictions, age increases by 1.66 years, given that all other variables are constant. This means that convicts with more prior convictions are likely to be older (keeping all other predictors constant), which could be because of how many convictions they have had over time, or other factors related to age and criminal history. The negative coefficient of '-

0.03' for time served suggests that when the time served increases by a unit, age decreases by 0.03 years (keeping all other variables constant). It is slightly surprising that as time served increases the age decreases, as it means that people who serve longer are younger. That could be interpreted as younger people committing worse crimes that require serving longer sentences or that the older people getting released via bail faster. Lastly, duration has a positive coefficient, which depicts that as the duration increases by a single unit, the age increases by 0.058 years. This suggests that there is a positive correlation between the two variables. This equation enabled us to understand the relation between age and the selected independent variables. By understanding this relationship, we can answer the question on how age affects the repetition of the convict repeating crime.

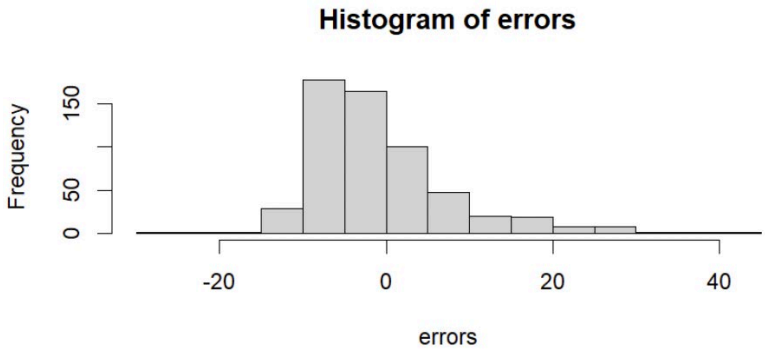


*Figure 1: Histogram of Errors*

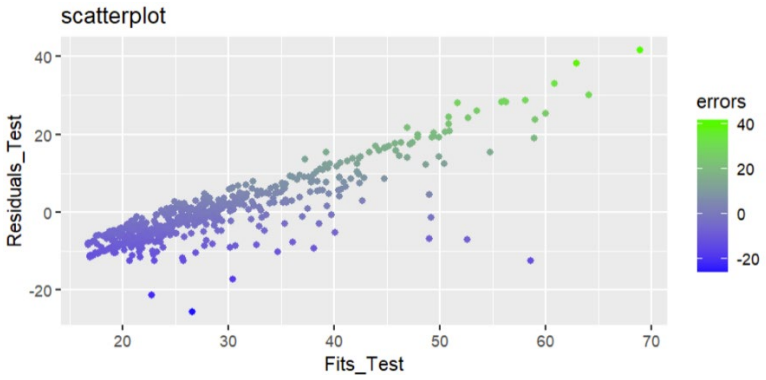


*Figure 2: Scatterplot*

All the independent variables except ‘tserved’ have *** which indicates the highest level of significance (p-value is close to 0) and that there is a strong relationship between the dependent and independent variables, which indicates that age is influenced by them. ‘tserved’ does not have a single star, further suggesting that the relationship between it and age is not as strong as the rest of the variables, and therefore as analyzed earlier that younger people serving more time doesn’t fully stand true. Furthermore, the median value is –2.803 which suggests that the model is underestimated.

```
Call:
lm(formula = age ~ priors + tserved + durat, data = training)

Residuals:
    Min      1Q  Median      3Q     Max
-32.897  -6.258  -2.802   3.156  48.027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.08360    0.89989  26.763  < 2e-16 ***
priors       1.65731    0.11183  14.820  < 2e-16 ***
tserved     -0.03121    0.01668  -1.871   0.0617 .
durat        0.05797    0.01238   4.681 3.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.367 on 863 degrees of freedom
Multiple R-squared:  0.2119,    Adjusted R-squared:  0.2091
F-statistic: 77.33 on 3 and 863 DF,  p-value: < 2.2e-16
```

*Figure 3: Console Window output*

## Logistic regression

A logistic regression model is used to predict categorical data and enables one to determine the probability of an event occurring. We chose to use this model primarily because it is strong at predicting binary outcomes (such as categorical data) and is easy to implement and interpret since it directly models the probability of outcomes. We decided to create two models with both predicting for “black” which refers to whether the individual is black or not. For the first model, we used “priors” as our predictor (which refers to whether there were any prior convictions against

the individual) while we used “educ” (which refers to the number of years of schooling) as our predictor for the second model.

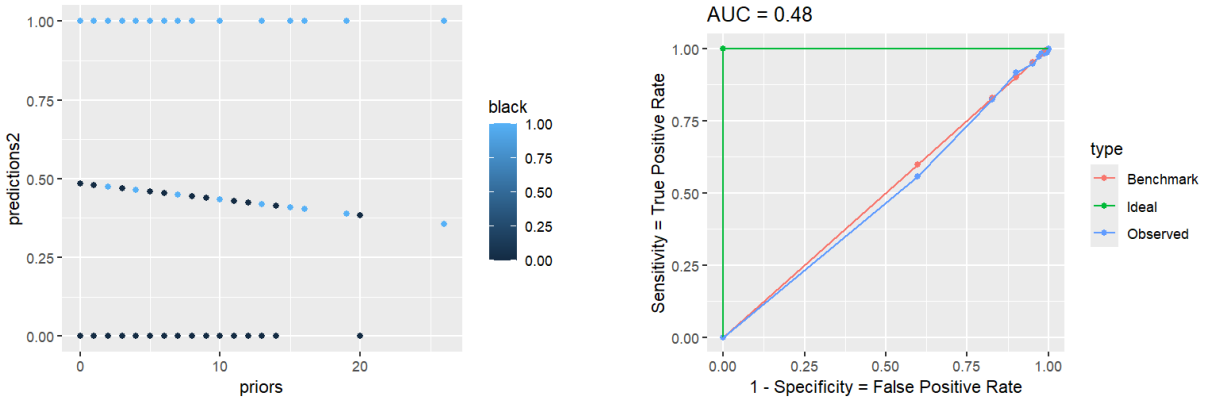


*Figure 4 and 5: ROC and Lift Chart for Model 1*

In model 1 (black based on priors), we found that our model did not form the traditional S-shape and instead formed a somewhat linear graph with a negative slope. In other words, as the number of prior convictions increased, the probability of the individual being black decreased. This is highly surprising based on our preconceived notions of the American justice system and goes against our expectations. Our equation for the model is yhat = (e^(-0.06 - 0.02x))/1 + e^(-0.06 - 0.02x). This means that our odds ratio is e^(-0.02) which is 0.98. Hence, for every 1 unit increased in the number of prior convictions, the odds of the individual being black increases by a factor of 0.98 or decreases by a factor of 0.02. Moreover, the sensitivity (which refers to the % of true predictions being accurately predicted) is 98.26%. However, the specificity (which refers to the % of false predictions being accurately predicted) is only 1.03%. This is a major problem and suggests that the model is poor. In addition, the model has an error rate of 51%, hence for every prediction there is a 51% chance that it is incorrect. Though this percentage is not too bad as around half of the predictions can be accurate, the ROC chart (on the right) highlights a significant issue. The AUC (area under the curve) is only 0.48, while graphically the observed line (in blue) is

beyond the benchmark line (in red) and is far away from the ideal line (in green). This suggests that the model is poor for prediction.

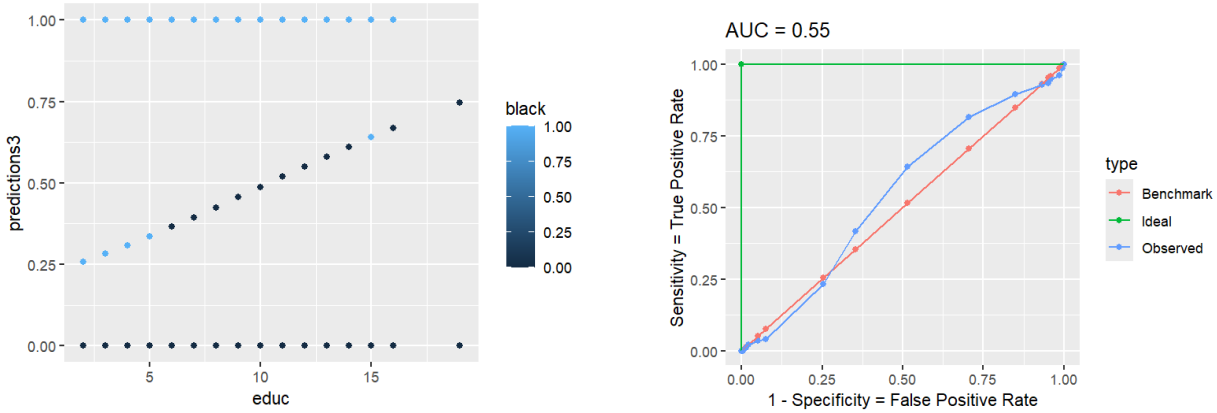


*Figure 6 and 7: ROC and Lift Chart for Model 2*

In model 2 (black based on education), we found that our model also did not form the traditional S-shape and instead formed somewhat of a linear graph with a positive slope. Hence, as the number of years of schooling increased, the probability of the individual being black increased. This too is extremely surprising as societal expectations suggest the opposite due to various stereotypes and biases: as education increases, the probability of being black decreases. Our model equation is yhat = (e^(-1.31 + 0.12x))/1 + e^(-1.31 + 0.12x). This means that our odds ratio for this model is e^(0.12) which is 1.13. Unlike the previous model, this model has a positive odds ratio which suggests a positive correlation between the dependent and independent variables. An odds ratio of 1.13 reflects that for every 1 unit increase in the number of years of schooling, the odds of the individual being black increases by a factor of 1.13. Furthermore, the sensitivity of the model is 41.81% while the specificity is 64.6%. Hence, the % of true that the model accurately predicts is 41.81%, while the % of false that the model accurately predicts is 64.6%. Though the sensitivity is on the lower end, this is more of a balance between the two key performance indicators which is a positive. Additionally, the error rate is 47% which is lower than the previous model, suggesting that for every prediction, the chances of it being accurate is higher than the

previous model. The AUC in the ROC chart is also higher at 0.55 and the observed line intersects the benchmark line and moves further to the ideal line which is significantly better than model 1.

Overall, though both models have their individual limitations, model 2 performs better than model 1 in almost every key performance indicator and as per the ROC chart.

## KNN Classification

K-Nearest Neighbors (KNN) is a powerful algorithm used in order to solve problems in machine learning. The simplicity, yet the effectiveness of the model's methodology is extremely beneficial, making it appropriate for investigating our research question. Additionally, we wanted to use values that were closest to the variable being predicted for in the same neighborhood to enhance accuracy, hence we decided to include KNN in our study.

KNN Classification was used in this study to investigate wheether or not a retuning felon has been in a prison work program. As a result, we choose our dependent variable to be "workprg", which was whether or not they were in N.C prison work program, and our independent variables to be priors, years served, length of the following period, duration of the prison stay, age in terms of years, and race. We choose these independent and dependent variables to investigate the effect of being in a prison work program on returning felons.

High quality research with respect to workforce programs suggests a positive impact on the rates of recidivism. For instance, a report authored by Steven Sprick Schuster and Ben Stickle highlighted that the "prison workforce programs reduce the likelihood of recidivism by 14.8%" (intext). The research in this report makes it clear that investments in prison-based workforce programs are worth it and generate a safer community. As a result, we decided to use a statistical study to understand the results in detail.

While coding the KNN Classification program on R, we identified the number of points in the neighborhood to be 13. This was determined using the KnnCrossval line of code which was imported from the Babson Analytics C file. We believe the size of the neighborhood (13) is appropriate since it is not too big or small to result in overfitting or underfitting. This ensures that it will not get influenced by the noise in the dataset while also avoiding randomization, further not having a negative effect on the model. Additionally, all points were not used, further not affecting the split of the data.

Aftering running the code, the final results indicated that the error rate for the model with standardization was 36.33% while the error rate without standardation was 37.54%. Moreover, the error rate benchmark was determined as 47.75%. This proves that the model with standardization is better than the model without standardization and the benchmark values since the error rate for the model with standardization is lower. Additionally, this interpretation suggests that the model allows for a good prediction of the impact of prison work programs on returning felons.

**Classification and Regression Trees**

According to a 2015 study conducted by the Drug Policy Alliance, the United States makes up less than 5% of the overall global population, however it makes up an alarming 25% of the incarcerated population, largely due to drugs. Of the total prison population, Black felons comprise 30% of those arrested for drug law violations and 40% of those incarcerated in state or federal prison for drug law violations. Additionally, mandatory minimum sentencing laws, adopted in the 1980s and 90s, have contributed greatly to the number of people of color behind bars. Furthermore, according to study published in the Journal of Prison Education & Reentry conducted by Susan Klinker Lockwood and John M. Nally from the Indiana Department of correction, Taiping Ho, who is the head of the Department of Criminal Justice and Criminology at Ball State University in

Indiana, and Katie Knutson who is a senior consultant at Public Consulting Group, found that African American ex-prisoners had a higher unemployment rate and recidivism rate than Caucasian ex-prisoners. This data highlights the disproportionate impact of drug laws and sentencing on communities of color, especially African Americans, who not only face higher arrest rates but also higher recidivism rates.

Classification and Regression Tree, CART, is a simple yet powerful analytical tool that helps determine the most significant variables in a particular data set by uncovering hidden underlying patterns within the data set. With the specific data set used in this research study, recid, which looks at recidivism data and the various characteristics of the felons who re-enter the criminal justice system, a CART model is particularly useful as it can aid prison administrators, policymakers, researchers, and other non-technical stakeholders understand the factors that most influence recidivism. Additionally, a CART model is versatile as it is able to handle both numerical and categorical data which is present in this data set with variables such as age, type of crime (drugs, alcohol, crime against person, etc.), years of schooling, etc., picking up on the non-linear relationships present between variables. Furthermore, CART models handle missing data exceptionally well and this this a crucial feature when looking at recidivism data, however in this case for the Recid data set, there were not any missing values. The presiding CART Models below highlight the underlying presence of how various recidivism factors disproportionately impact returning felons who are black.

When building out the CART model using the Recid data set, the dependent variable is Black which includes categorical data with 2 factors, Black (1) or Not Black (0). The RStudio's algorithm then classifies the data to pick out the most significant variables influencing a returning felon to be Black or Not Black. In this initial construction of the CART model, the error rate of

the model is 43.9% which is better than the error bench of 49.6%, however overfitting is present as two of the leaf nodes represent less than 3% of the data in each node (at 1% and 2%) which indicates high model complexity resulting in inaccuracy. This leads us to minimize the overfitting presented in the original model using two approaches: stopping rules and easy prune which are depicted by the following two models below.

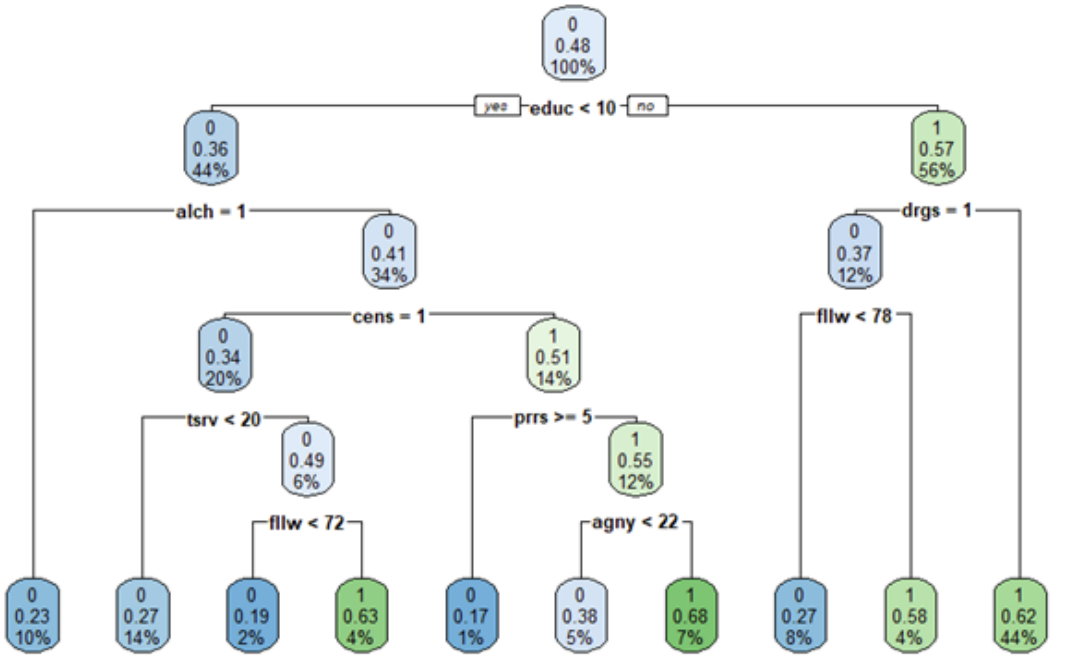


*Figure 8: Over Fitted CART Model*

In efforts to minimize the effects of over-fitting that was present in the first CART model, stopping rules were applied with a minsplit of 30 and minbucket of 25. This resulted in all leaf nodes to be greater than 3% with only 6 splits in the model compared to 9 in the overfitted model. The error rate in this model is 41.3% which is better than the benchmark and the error rate of the overfitted model. This error rate can be deemed more reliable as the model complexity is not as high as the overfitted model.

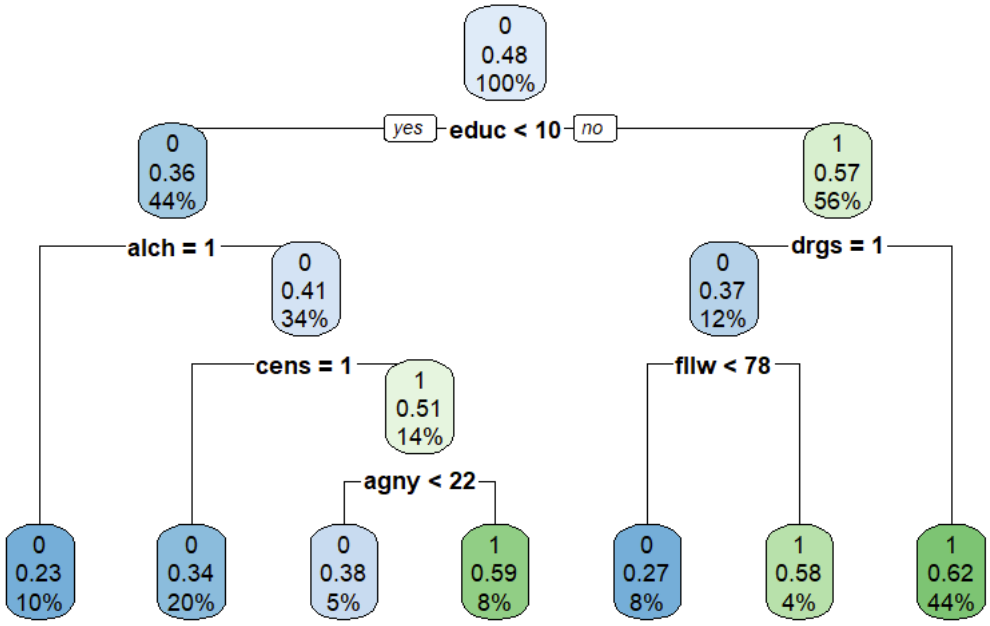


*Figure 9: Stopping Rules CART Model*

The second method used to minimize the high model complexity issue of the original over fitted model was easy prune. In this model, the error rate is 40.6% which is better than the benchmark (49.6%), the overfitted model (43.9%), and the stopping rules model (41.3%). Additionally, there are only 2 splits in the model compared to 9 which indicates underfitting making it an unreliable model.

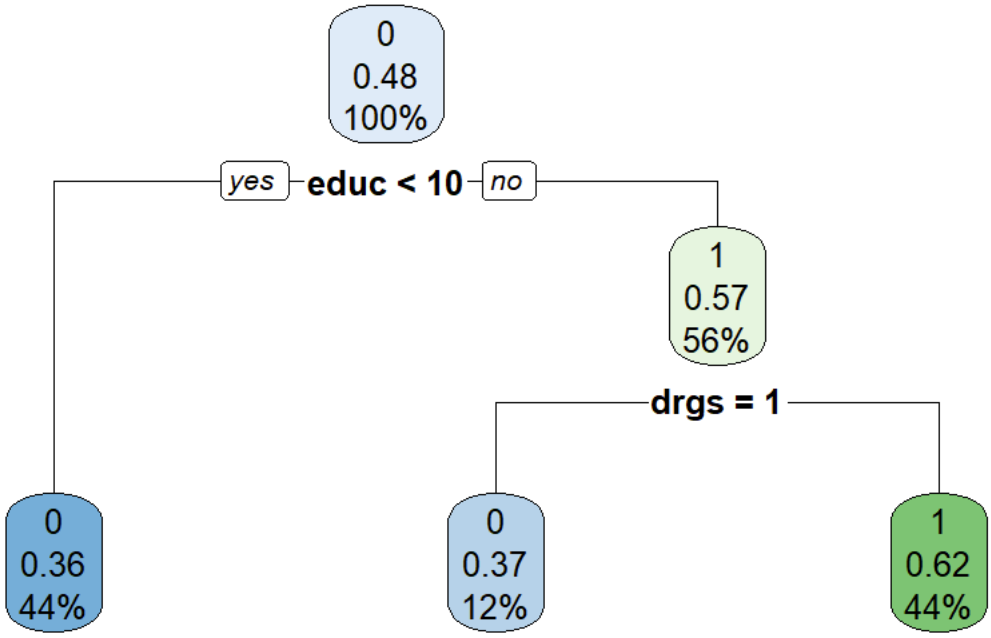


*Figure 10: Easy Prune CART Model*

After building the three models, the stopping rules model as presented in Figure 9 is the most accurate to highlight the underlying implications of race and recidivism as its error rate is most reliable and avoids under or over fitting.

The initial split is based on the educ variable. If educ is greater than or equal to 10, the model predicts a higher probability of the individual being Black (1), with a probability of 0.57, and 56% of the cases in this branch. This suggests that education level is a significant predictor in this dataset, with higher education levels being associated with black individuals. This is contradictory to common belief that state black people who enter the criminal justice system often have much lower education than Non-Black individuals. The second split in the CART model is based on the presence of drug history or not. According to the CART model built from the recid data set, returning felons with a history of drugs not present are black, as out of 44% of the data 62% are Black. When looking at the side of the split with education < 10, it can be noted that out

of 10% of the data only 23% of returning are black and have a history of alcohol abuse. The two leaf nodes with the rules: IF (Education > or equal to 10 years) AND (Drugs = 0), THEN (Black = 1), and IF (Education < 10 years) AND (Alcohol = 1), THEN (Black =0), both imply regardless of education re-offending black felons are for the majority of the time not receding because of drugs or alcohol which according to this model are two of the most significant variables. As a result, this highlights the complexity of factors influencing recidivism beyond the commonly discussed issues of substance abuse. It suggests that the relationships between race, education, and criminal behavior are not as straightforward as commonly perceived and that other underlying factors may play critical roles. The CART model's findings challenge us to look deeper into the social determinants that influence the likelihood of reoffending.

## Conclusion

In conclusion, tackling the issue of recidivism requires a deep and thoughtful approach that moves beyond easy answers and quick fixes. Our comprehensive study highlights how race, education, age, and involvement in prison work programs significantly influence the likelihood of someone returning to crime. The findings from various statistical models—including linear regression, logistic regression, KNN classification, and CART analysis—reveal not just the key factors but also the complex ways these factors interact. This complexity challenges us, as a society, to rethink our strategies and to focus on root causes rather than just symptoms. It is clear that effective solutions must not only improve educational and employment opportunities for those in the criminal justice system but also tackle the broader systemic issues that disproportionately affect certain groups. By taking a more holistic approach that addresses the specific needs and circumstances of returning felons, we can build a more just and effective system. Such an approach

will not only reduce the rates of recidivism but also enhance public safety and help create a more equitable society for all.

# References

Lockwood, K. S., Nally, J. M., Ho, T., & Knutson, K. (2015). Racial Disparities and Similarities in Post-Release Recidivism and Employment Among Ex-prisoners with a Different Level of Education. Prison Education & Re-entry, 1-16.

*Recidivism of young adults*. Office of the Inspector of Custodial Services. (2014, September 26). https://www.oics.wa.gov.au/reports/recidivism-rates-impact-treatment-programs/key-findings/recidivism-young-adults/#:~:text=There%20are%20undoubtedly%20some%20uncontrollable,heightened%20sensitivity%20to%20social%20pressures.

The drug war, mass incarceration and Race. (n.d.). https://www.unodc.org/documents/ungass2016/Contributions/Civil/DrugPolicyAlliance/DPA_Fact_Sheet_Drug_War_Mass_Incarceration_and_Race_June2015.pdf

Thomas, M. (n.d.-a). *An Exploration of Recidivism Based on Education and Race*. Walden University. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=9610&context=dissertations#:~:text=According%20to%20Nellis%20(2016)%2C,of%20Justice%20Statistics%2C%202011

Wetzel, H. (n.d.). *Research finds prison education programs reduce recidivism*. Mackinac Center. https://www.mackinac.org/pressroom/2023/research-finds-prison-education-programs-reduce-recidivism#:~:text=They%20found%20in%20their%20review,extra%20%24131%20in%20quarterly%20wages.