PREDICTING RECIDIVISM IN THE CRIMINAL JUSTICE SYSTEM

Jackson Le
wnj9tx

Austin Shi
szb8fc

Emily Chen
qeq3cf

December 12, 2024

1 Abstract

Recidivism, the tendency of convicted individuals to reoffend, presents significant challenges within criminal justice systems. In this project, we aim to predict the likelihood of recidivism within one, two, or three years following parole, as well as examine which features contribute the most to this likelihood. Using a dataset from the Virginia criminal justice system, which includes demographic and criminal history details, we apply machine learning classification techniques to track trends and provide predictive insights. By identifying the key factors that contribute to reoffending, this analysis seeks to inform more effective decision-making processes in the criminal justice system, ultimately supporting rehabilitative efforts and reducing recidivism.

2 Introduction

Recidivism rates serve as indicators of the effectiveness of the rehabilitation and reintegration efforts of the Virginia criminal justice system. High recidivism rates often suggest gaps in support systems and may highlight the need for policy adjustments. Through this project, we will explore how accurately we can predict recidivism using demographic and criminal history data, such as gender, race, type of offense, and drug test results. Through this analysis, we hope to reveal critical insights about the factors that most strongly correlate with the risk of recidivism. By identifying patterns in these data, the project seeks to support improvements in parole management and resource allocation, enabling the justice system to better address the needs of at-risk individuals and reduce the likelihood of reoffending.

3 Method

For preliminary analysis, we imported the dataset into a Pandas Dataframe Object and labeled the object as df. Then we used the .head() and .info() command to look at the first few rows to check if it is valid. The raw data set has 18028 entities, with 53 total columns. We dropped a few features, including "ID" and "Recidivism in 1, 2, and 3 years" because it correlates with what we want to predict, which is Recidivism within 3 years.

For cleaning, we first checked out which features contained empty values and cleaned it as follows: for null Gang_Affiliated value, we opted to default it to False because we believed if they belonged to a gang, they would have had an entry. For Supervision Risk Score First, we opted to use the average value as there are many different values. For the feature Supervision_Level_First we chose to have it be labeled "Standard" as that is the equivalent of the most normal value present. Prison_Offense will be labeled as "Other" to avoid the complexity of introducing random crimes. Avg_Days_per_DrugTest was also the average value. Any DrugTest_Positive related feature will be 0 by default because we believed that if an offender did drugs, it would be noted. With that same logic, we opted to fill the jobs per year as 0 for empty space.

We opted to use Logistic Regressions, Decision Tree Classifier, and Forest Classifier first because they can be used to classify true or false, or classify the value for our Recidivism within 3 years value. Then later we attempted to do

K-means clustering to see if there is any similarity between grouping that can help determine the relationship between features.

4 Experiments

With our model, we attempted to predict Recidivism within 3 years feature.

We set up a pipeline system with imputers that dealt with the feature listed above. For Jobs_per_year we directly engineered the values to be rounded up to an integers and filled empty row with 0 so that it can be used along the numerical pipeline. For any other category pipeline or other object we simply OneHotEncoded it with a value because there are many different responses.

Once the pipeline was set up, we made a train set, test set, and validation set ready to be used for our experiments. We then transformed the dataset with our pipeline and imported all of our models to be trained. We called fit_transform() on the training and testing data, respectively.

For our Logistic Regression Model, we got an accuracy of about 72%, a precision of 74%, recall of 81%, and F1 Score of 77%. For the feature importance, the 5 features that influenced whether or not our model made the prediction on whether a former criminal would recidivate were the percentage of days employed, True for Gang Affiliation, when they are released at age 48+, when they are released at 18-22 years old, and when they have a Delinquency Report of 4 or more. We believe this model does pretty well and we were able to isolate the important features in predicting recidivism.

For the Decision Tree Model, we observed an accuracy of about 62.23%, a precision of 68.93%, a recall of 66.29%, and an F1 score of 67.59%. For the feature importance, the 5 that had the highest importance were percentage of days employeed, Average Days per Drug Test, Residence PUMA (Public Use Microdata Area which is a code for a geographical area), First Risk Score, and True for Gang Affiliation. We think that this tree model was not as accurate because some of these features did not show a clear pattern when we originally graphed it out in our notebook.

For the Random Forest Model, we achieved an accuracy of about 71.38%, a precision of 73.44%, a recall of 81.19%, and an F1 score of 77.12%. The 5 features that stood out were percentage of days employed, Average Days per Drug Test, First Supervision Score First, Residence PUMA, and Job Per Year. We believe that this is one of the better models we have because its accuracy is similar to our Logistic Regression Model, but Random Forest is known to handle more complex data.

We settled with Random Forest because it is more suited for the kind of task we were trying to carry out, and we tuned it with 5 folds and optimized with F1 score using GridSearchCV. After tuning, we found that the best parameters were 10 for max depth, log_2 for max features, 2 for min samples leaf, 100 for n estimators and 5 for min sample splits. These parameters scored the best F1 score of around 76%.

Our best model from GridSearchCV gave us 71% for accuracy, 72% for precision, 85% for recall, and 78% for F1 score. Its most important features were Percent Days Employed, First Supervision Risk Score, True for Gang Affiliation, Average Day per DrugTest, and Prior Arrest Episodes PPViolationCharges.

We also implemented means clustering to see which features are prominent while sorting. First, we graphed out the elbow method to determine the best amount of cluster to use, and we ended up using k = 4, which gave us a silhouette score of 0.618. We then got the recidivism rate which is between 57-59%. While the results are close, it still revealed a tiny bit of nuance that can be pointed out.

Cluster 1 and Cluster 2 have the highest Recidivism rate, and Cluster 0 and 3 are slightly lower in the rate. For Cluster 0, the low recidivism most likely benefits from its frequent drug check-ins with its average day between checks being the lowest, it also has the highest employment rate compared to another cluster, indicating that individuals with stable employment are less likely to commit crimes. Cluster 1, has a longer drug test period, which indicates limited oversight.

It has one of the employment and education statuses too, with those features not making the top 6 of the feature relevancy. Cluster 2 is similar in that it also has a longer drug test period, low employment and low education, indicating a potential need to improve in that area. Cluster 3 has a more frequent drug test period too; one thing worth noting is that it has the lowest THC level with it being around 0.15, which is the lowest compared to the other cluster, indicating that low drug use would make people less likely to commit a crime again.

5 Results

Link to code: https://github.com/Quole0812/ML4VA

Logistic Regression Model:

Metrics	Training Sets
Accuracy	0.722684
Precision	0.7442254
Recall	0.812324
F1 Score	0.776785

Feature Importance for Logistic Regression Model:

Feature	Importance
Percent_Days_Employed	1.762329
Gang_Affiliated_True	0.823677
Age_at_Release_18-22	0.744661
Age_at_Release_48 or older	0.716675
Delinquency_Reports$_1$	0.526803

Decision Tree Model:

Metrics	Training Sets
Accuracy	0.626178
Precision	0.691787
Recall	0.668534
F1 Score	0.679962

Feature Importance for Decision Tree Model:

Feature	Importance
Percent_Days_Employed	0.130948
Avg_Days_per_DrugTest	0.076618
Residence_PUMA	0.059841
Risk_Score_First	0.041820
Gang_Affiliated_True	0.025029

Random Forest Model:

Metrics	Training Sets
Accuracy	0.715751
Precision	0.731262
Recall	0.824463
F1 Score	0.775071

Feature Importance for Random Forest Model:

Feature	Importance
Percent_Days_Employed	0.080020
Avg_Days_per_DrugTest	0.050255
Residence_PUMA	0.042861
Risk_Score_Firs	0.041544
Jobs_Per_Year	0.020216

Random Forest Classifier:

Metrics	Training Sets
Accuracy	0.715474
Precision	0.719685
Recall	0.853408
F1 Score	0.780862

Feature Importance for Random Forest Classifier:

Feature	Importance
Percent_Days_Employed	0.154333
Risk_Score_First	0.052958
Gang_Affiliated_True	0.049428
Avg_Days_per_DrugTest	0.038190
PPViolationCharges_0	0.037956

Cluster 0 Feature Results

Feature	Mean
Avg_Days_per_DrugTest	32.71
Supervision_Risk_Score_First	6.33
Gender_M	0.87
Program_UnexcusedAbsences_0	0.73
Delinquency_Reports_0	0.64
Percent$_{D}ays_{E}mployed$	0.54

Cluster 1 Feature Results

Feature	Mean
Avg_Days_per_DrugTest	414.63
Supervision_Risk_Score_First	6.03
Gender_M	0.93
Program_UnexcusedAbsences_0	0.86
Delinquency_Reports_0	0.72
Race_BLACK	0.65

Cluster 2 Feature Results

Feature	Mean
Avg_Days_per_DrugTest	210.54
Supervision_Risk_Score_First	5.99
Gender_M	0.89
Program_UnexcusedAbsences_0	0.84
Delinquency_Reports_0	0.69
Race_BLACK	0.64

Cluster 3 Feature Results

Feature	Mean
Avg_Days_per_DrugTest	95.96
Supervision_Risk_Score_First	5.83
Gender_M	0.88
Program_UnexcusedAbsences_0	0.87
Delinquency_Reports_0	0.71
Program_Attendances_0	0.65

Cluster Recidivism Rate

Cluster number	Recividism rate
Cluster 0	0.578440
Cluster 1	0.588629
Cluster 2	0.592985
Cluster 3	0.575113

6 Conclusion

We utilized machine learning techniques to predict recidivism within 3 years of parole using a dataset from a criminal justice system, which concluded that Random Forest was the best model. Random Forest identified the key characteristics that influenced the risk of a parolee recidivating, which were Percent Days Employed, Average Days per DrugTest, First Supervision Score First, Residence PUMA, and Job Per Year. Kmeans cluster pointed out that the features highly relevant to recidivism are the Employment rate, Average day per drug test, and Education rate. With this data, Virginia policymakers can adapt the policy they propose to allocate resources more effectively in order to create a safer society. These models can support parole officers in making data-informed decisions to maximize the impacts of rehabilitation, prioritizing cases where individuals are at the highest risk of recidivism. For example, investing in employment programs for parolees or targeted interventions for gang-affiliated individuals could significantly reduce reoffending rates. These impact the well being of all Virginians because reducing the statistic of convicted offenders who reoffend gives more opportunities for them to reunite with their loved ones and restart their lives, as well as reducing crime rates overall which protects public safety. One shortcoming of this data is that there are other factors that impact recidivism rates – access to effective resources such as mental health counseling and community support – that are not easily quantifiable to be put in a dataset. Adding features like this to the dataset in the future could uncover more findings about the criminal justice system that would further benefit policymakers.

7 Contributions

Jackson Le worked on data preparation, data visualization, and model training, as well as training a random forest model. He also trained the cluster algorithm and analyzed it. Emily Chen wrote the checkpoint paper and performed model comparison analysis, Austin Shi trained the logistic regression and decision tree models, tuned hyperparameters for all models, and visualized the data.

8 References

https://data.ojp.usdoj.gov/Courts/NIJ-s-Recidivism-Challenge-Full-Dataset/ynf5-u8nk/data$_preview$
https://www.bls.gov/opub/reports/race-and-ethnicity/2022/