Chunyu Chen
Zongjie Yin

Data Analysis for Crime Recidivation

Background

As one may know, in the US judicial system, criminals are sentenced based on the Compas Score given by the Compas Algorithm, which sometimes produces biased or unreasonable scores that can lead to wrong decisions. The impact of a criminal miscarriage extends beyond just the immediate harm caused to the victim. It can have far-reaching consequences on public safety and the overall happiness of society. In some cases, criminal acts can result in serious casualties, making it a matter of great concern.

Purpose

Our project aims to assist the government in identifying inmates who are at a high risk of reoffending. This would help the government to optimize the resources to track down criminals that are more likely recidivate while ensuring the accuracy of our recidivism predictions. To achieve this, we will utilize the information provided in the dataset to predict whether an individual criminal is likely to recidivate within the next 3 years.

Data

Our project utilizes a dataset sourced from the Kaggle website at https://www.kaggle.com/datasets/uocoeeds/recidivism. The dataset consists of 54 columns with 25,835 rows, recording intrinsic data such as gender and race, as well as crime activity data, including prior crime type, prison offense type, risk score, and whether each criminal will recidivate within the next 3 years.

Data Cleanning

Firstly, we need to check whether there are any duplicated values in our data set. Fortunately, we did not find any duplicate record IDs, so we retained all the data for further analysis.

Apart from checking for duplicated data, we also need to identify any outliers in our data set. To begin with, we extracted all the numerical variables and used the 3-sigma rule (which is effective under the assumption of normal distribution) along with histograms of each column to identify outliers. We found some suspicious outliers, such as in the Avg_Days_per_Drug column, where the maximum value is 1088, and over 3 percent of the data had values greater than 364. However, after careful consideration, we decided to keep the data as it is reasonable for someone who takes the drug test yearly, bi-yearly, or even once every 3 years.

Another column that we paid close attention to was Jobs_per_year, which had a maximum value of 8, and roughly 2 percent of the data had values greater than 3. Again, we decided to retain most of the data as it is not uncommon for a person to change jobs 3 times in a year. However, we decided to delete the individuals who had over 3 jobs per year.

After handling the outliers, we discovered that there were many missing values in the dataset. Usually, there are several ways to deal with them: 1) delete the column if it is not informative; 2) delete the rows with missing values if they won't affect the results compared to the number of data points; 3) fill the missing values using the median, mean, or other relative methods.

We first filtered out the columns with over 50% missing values, but fortunately, we did not have any such columns in our dataset. However, we found that some columns still had a certain number of missing values, such as Gang_Affiliated, Supervision_Risk_Score_First, Percent_Day_Employed, Prison_Offense, etc.

Then, in order to fill in the missing values for Gang_Affiliated, we decided to use the mode value as 80% of the values were "True" and 20% were "False". For Supervision_Risk_Score_First, we filled in the median value as the values for Supervision_Risk_Score_First were evenly distributed between 1 and 10. Interestingly, for the Supervision_Level_First column, we utilized the mode value based on the Supervision_Risk_Score_First value, as there were multiple values of Supervision_Level_First given the same Supervision_Risk_Score_First value. Finally, for the columns related to drug test results, we took the advantage of the mode value to fill in the missing values as most of the value of the drug test results were 0.

Data Visualization

Before moving to the Data Visualization section, some questions we might want to answer:

a) Whether the **gender** difference will affect recidivation?

b) Whether the **prior crime type** will affect recidivation?

c) Whether the **risk score** will affect recidivation?

d) Whether the **education level** will affect recidivation?

Correlation Heatmap

We first wanted to see a bigger picture of those numerical variables. To achieve this goal, we made the correlation heatmap to investigate the correlation between covariates.

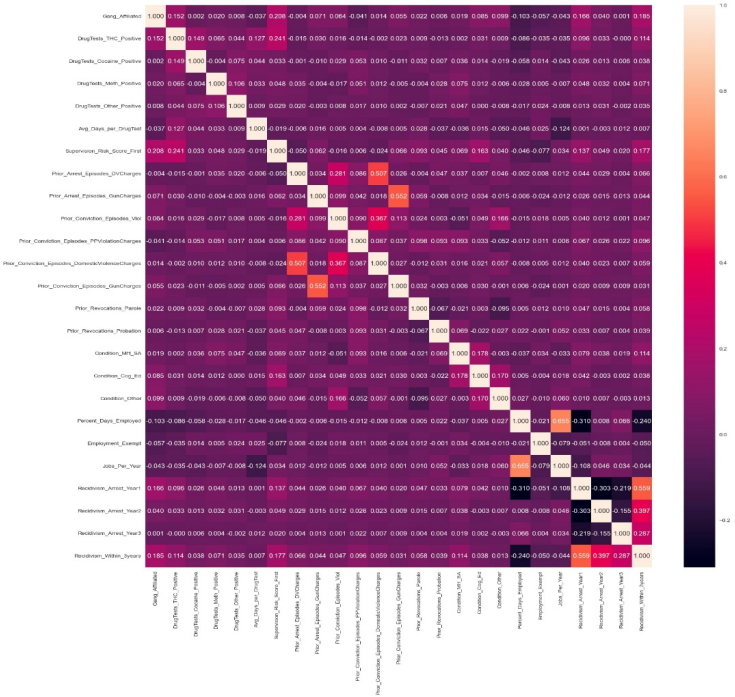


Figure 1: Correlation Heatmap

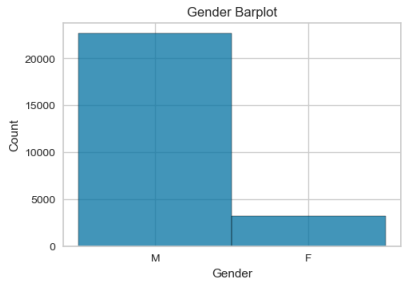
From the above correlation matrix, we could see that the Gang_Affiliated (0.185), DrugTest_THC_Positive (0.114), Supervision_Risk_Score_First (0.177), Condition_MH_SA (0.114), Percent_Days_Employed (-0.240) had some correlation with recidivism_within_3years, which we would investigate it later.

Bar plot

Gender

We first examined the gender distribution of the criminals using a bar plot and then explored the relationship between gender and recidivism using another bar plot. The bar plot on the left revealed that approximately 85% of the criminals in the dataset were male.

When we analyzed the bar plot on the right, we observed a trend suggesting that male criminals had a higher probability of recidivating within the next three years compared to female criminals. This finding suggests that gender may be a factor that influences recidivism rates. However, it is important to note that other variables, such as race, may also impact recidivism rates and should be considered in the analysis.



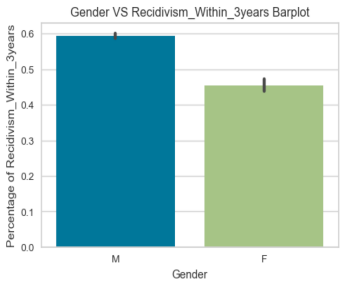
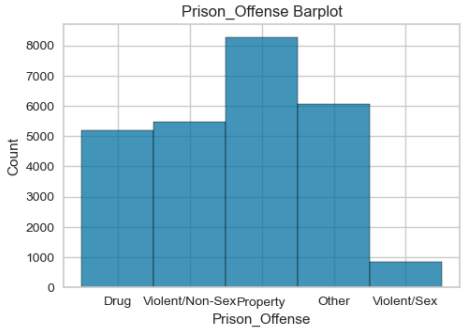


Figure 2: Gender Bar Plot *Figure 3: Gender VS Recidivism Bar Plot*

Prison Offense

We also examined the distribution of Prison Offense types among the criminals using a bar plot and then explored the trend between Prison Offense type and recidivism using another bar plot. The bar plot on the left showed that Property crime was the most frequent offense type, while violent/sex crimes were the least frequent. Upon analyzing the bar plot on the right, we observed that criminals affiliated with drug offenses or violent offenses without a sexual component had similar recidivism rates. Additionally, criminals affiliated with property crimes had the highest recidivism rate.



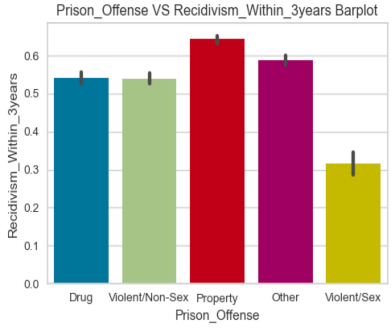
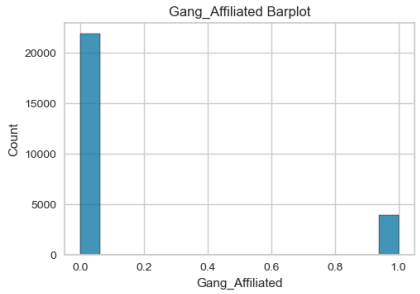


Figure 4: Prison Offense Bar Plot *Figure 5: Prison Offense VS Recidivism Bar Plot*

Gang Affiliated

From the bar plot on the left-hand side below, it can be observed that the majority of crimes were not related to gang activity. However, it is worth noting that gang-affiliated criminals exhibited a higher recidivism rate compared to non-gang-affiliated criminals.



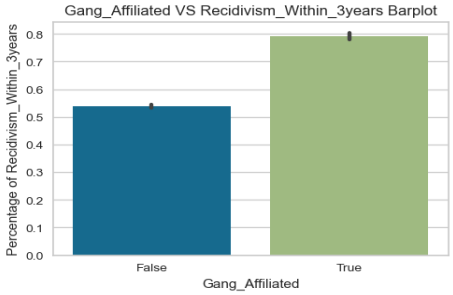
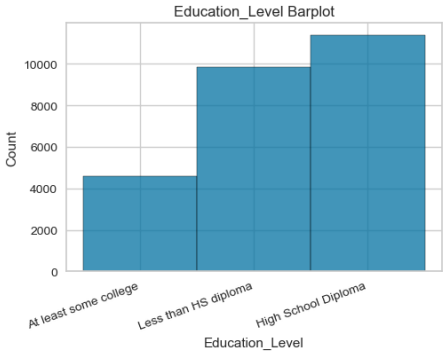


Figure 6: Gang Affiliated Bar Plot *Figure 7: Gang Affiliated VS Recidivism Bar Plot*

Education Level

The bar plot on the left-hand side shows that the majority of criminals did not have an education level higher than a high school diploma. Conversely, the plot on the right-hand side revealed that criminals with at least some college education had a lower recidivism rate.



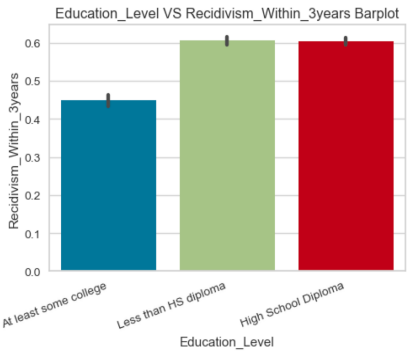
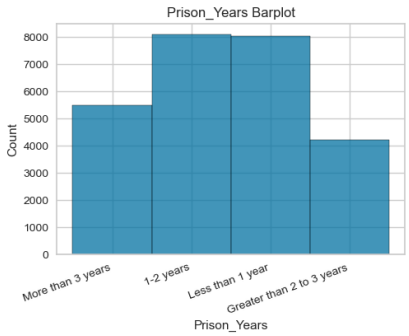


Figure 8: Education Level Bar Plot *Figure 9: Education Level VS Recidivism Bar Plot*

Prison Years

The bar plot on the left-hand side shows that most prison sentences were less than two years. On the other hand, the plot on the right-hand side revealed that criminals with more than three years of imprisonment had a lower recidivism rate.



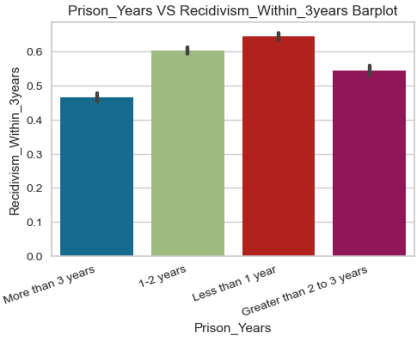
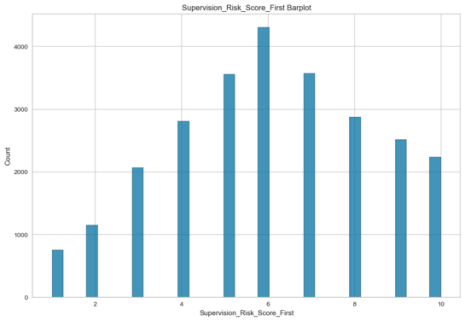


Figure 10: Prison Years Bar Plot *Figure 11: Prison Years VS Recidivism Bar Plot*

Supervision Risk Score First

The bar plot on the left-hand side shows that both high and low Supervision risk scores tended to have lower counts. Regarding the recidivism rate, the Supervision risk score initially showed an increasing trend from 2.0 to 10.0, although the score of 1.0 did not follow this trend.



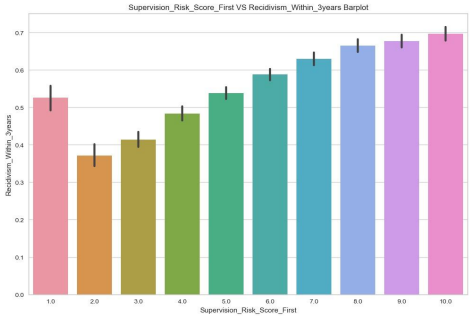
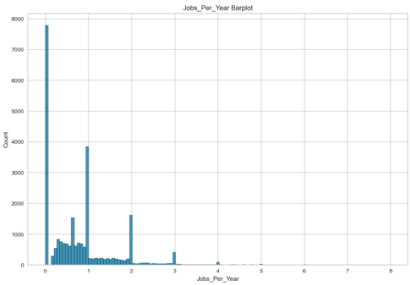


Figure 12: Risk Score Bar Plot *Figure 13: Risk Score VS Recidivism Bar Plot*

Percent Days Employed

The bar plot showed that most criminals had less than three jobs per year. On the right-hand side plot, it was observed that criminals who recidivated within three years tended to have a lower percentage of days employed.



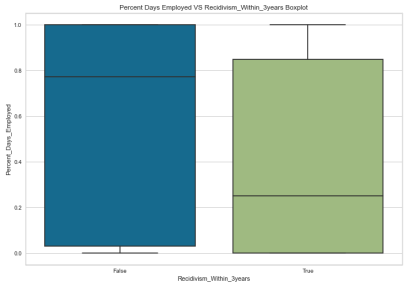


Figure 14: Jobs Per Year Bar Plot *Figure 15: Jobs Per Year VS Recidivism Bar Plot*

Cluster map

We also created a cluster map to compare criminals who recidivated in the next three years with those who did not. Horizontally, it can be observed that the percentage of days employed and the number of jobs per year exhibited similar patterns, as did the results of drug tests. However, when examining the clusters vertically, we found no distinguishable or clear clustering across different groups.

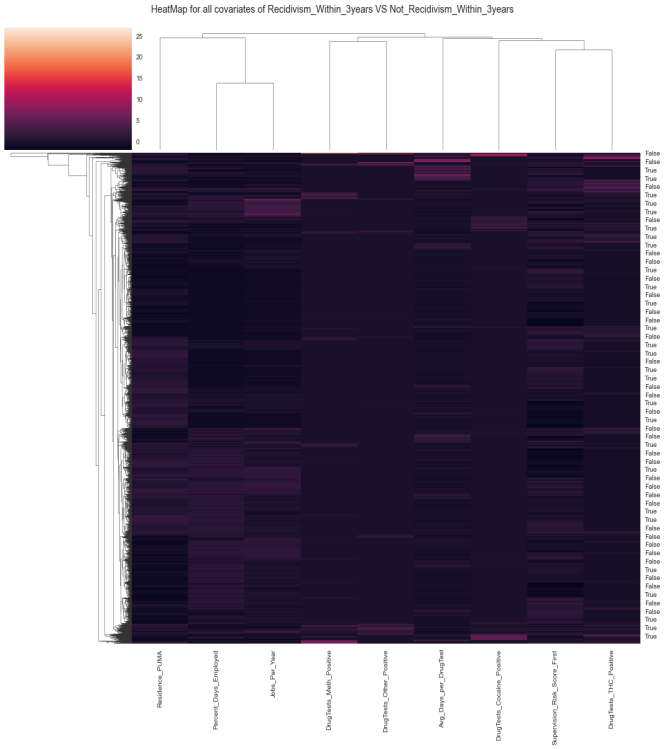


Figure 16: Cluster Map

PCA and K-means

In this analysis, our focus was to identify any distinct patterns that differentiate whether a criminal will recidivate within the next three years. To address this question, we employed standardization, PCA (Principal Component Analysis), and K-means techniques.

Firstly, we standardized all numerical variables to ensure consistency in measurements and facilitate distance calculations in the K-means algorithm. Next, we utilized PCA as a dimension reduction technique to obtain the two most significant orthogonal variables, derived from linear combinations of the 'important' variables. By performing PCA, we were able to visualize the data points on the x- and y-axes, aiming to identify underlying patterns among the criminals.

Then we employed the K-means clustering method. To determine the optimal number of clusters (k) to be generated in the K-means process, we employed the Elbow method. The Distortion Score Elbow graph indicated an optimal value of k as 3, with a distortion score of 29587.470 (representing the sum of

squared Euclidean distances from the respective cluster centroids). However, the clustering results were not considered good based on this analysis.

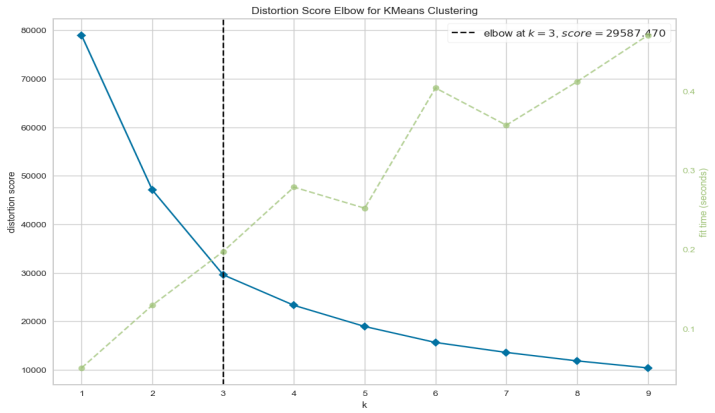


Figure 17: Elbow Method with Distortion Score Graph

Subsequently, we generated the K-means clustering graph with k = 3, which exhibited a cone-like shape. Additionally, we plotted the original dataset, differentiating between criminals who recidivated and those who did not. In the original dataset graph, the data points appeared to be randomly distributed. By comparing the results of the original dataset graph and the K-means clustering algorithm, we unfortunately observed that there was no distinct clustering pattern that aligned well with our dataset.

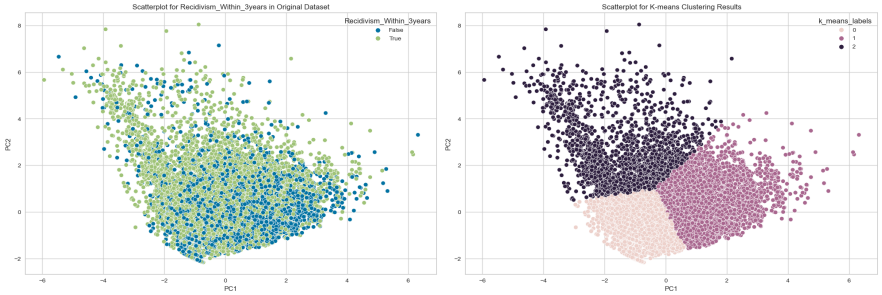


Figure 18: Original Dataset Recidivism Scatterplot *Figure 19: K-means Clustering Results Scatterplot*

Finally, we created a Cluster Characteristics graph to examine the distinctive features of each cluster resulting from the K-means analysis. Our observations revealed that cluster 2 exhibited high values in variables related to drug test results. On the other hand, cluster 1 demonstrated higher values in variables such as jobs per year and percent days employed. In contrast, cluster 0 showed low levels across most variables, except for average days per drug test, supervision risk score, and residence PUMA.

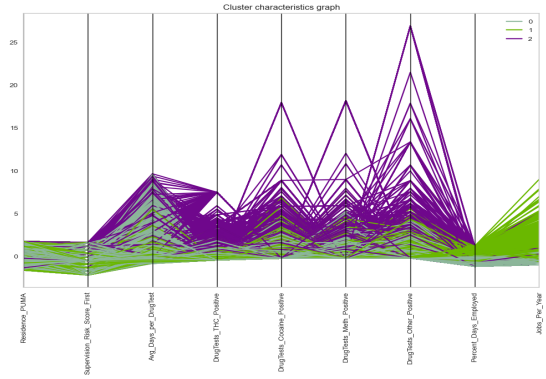


Figure 20: Cluster Characteristics Graph

Statistical Methods and Analysis

Model Set Up

In the machine learning section of our project, we conducted a comparison of results across different models. To initiate this process, we divided our dataset into an 80% training set and a 20% test set. Our primary metric for evaluating the performance of each model was recall (TP/(TP+FN)). This choice of metric was driven by our goal to predict whether a criminal would recidivate within the next three years, ensuring that we correctly identified those who would recidivate.

To ensure accurate results and facilitate meaningful model comparisons, we ran each model 100 times, generating a histogram of recall values. The models we employed for this analysis included Logistic Regression, Random Forest, and XGBoost. For the Random Forest and XGBoost tree-ensemble models, we further utilized grid search and cross-validation techniques to identify the best hyperparameter set, aiming to enhance model stability and robustness.

Data Preprocessing

In the data preprocessing stage, we implemented standardization for all numerical variables. This step aimed to address the issue of weighting imbalance that may occur due to different measurement scales across columns, ensuring fair treatment during model training. Additionally, we employed one-hot encoding to transform all categorical variables into matrix form. This transformation allowed the model to gain a better understanding of the meaning and significance of each value within the categorical columns.

Logistic Regression Model

Given the logistic regression model's suitability for estimating binary target variables, we selected it as the initial model to apply to the crime recidivism dataset. To mitigate the risk of overfitting, we leveraged ridge (L2) regularization.

In a single run of the logistic regression model, the recall for label 1 was approximately 0.79 (Figure 21). Additionally, the area under the ROC curve was 0.78 (Figure 23), and the area under the Precision-Recall Curve was 0.82 (Figure 24), indicating favorable performance. Furthermore, the confusion matrix (Figure 22) demonstrated the highest counts in True Positives and the lowest counts in False Negatives.

Results from running the model 100 times, as shown in the histogram, yielded a median recall rate of approximately 0.802 (Figure 25).

	precision	recall	f1-score	support
0	0.68	0.61	0.65	1450
1	0.74	0.79	0.77	1995
accuracy			0.72	3445
macro avg	0.71	0.70	0.71	3445
weighted avg	0.72	0.72	0.72	3445

recall = TP / (TP+FN) = 0.793984962406015

Figure 21: Results Matrix

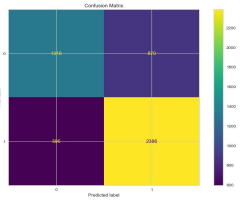


Figure 22: Confusion Matrix

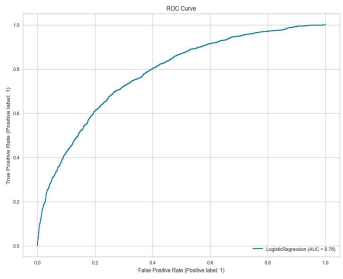


Figure 23: ROC Curve

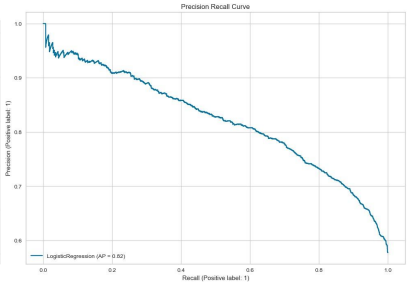


Figure 24: Precision Recall Curve

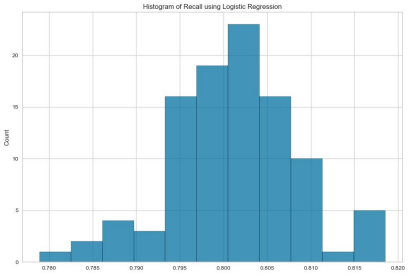


Figure 25: Recall Histogram

Random Forest

In this phase, we employed the Random Forest algorithm, which is a tree-ensemble algorithm, for our prediction task. We chose Random Forest for several reasons. Firstly, it has the ability to mitigate overfitting due to the utilization of multiple trees. Secondly, it is a flexible algorithm that can be applied to both regression and classification problems. Additionally, it operates in a parallel running mechanism, allowing for efficient processing. Lastly, Random Forest is known for its stability, thanks to the

aggregation of predictions from a large number of trees.

Initially, we used the same set of covariates as in the logistic regression model. To ensure stability and robustness, we employed the cross-validation technique. Moreover, we conducted hyperparameter tuning by testing different combinations of parameters such as the number of estimators, maximum depth, maximum features, minimum sample leaf, and minimum sample split. Our objective was to achieve the best result with the highest recall.

After experimentation, we obtained a recall of 0.847 by setting the hyperparameters as follows: number of trees: 500, minimum sample split: 5, minimum sample leaf: 2, and maximum depth of tree: 50.

.

```
Best hyperparameters: {'n_estimators': 500, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'auto', 'm
ax_depth': 50}
Recall: 0.8473666554847367
```

In a single run of the Random Forest model, the recall for label 1 was approximately 0.847 (Figure 26). Additionally, the area under the ROC curve was 0.79 (Figure 27), and the area under the Precision-Recall Curve was 0.82 (Figure 28), indicating significant improvements compared to the logistic regression model. Moreover, the confusion matrix (Figure 26) exhibited the highest counts in True Positives and the lowest counts in False Negatives.

Results from running the Random Forest model 100 times, as shown in the histogram, provided a median recall rate of approximately 0.847 (Figure 29). These results further validate the performance improvement achieved by the Random Forest model compared to the logistic regression model.

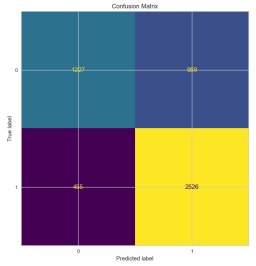


Figure 26: Confusion Matrix

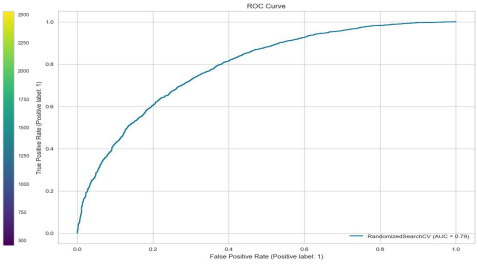


Figure 27: ROC Curve

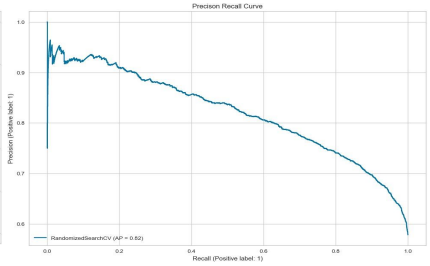


Figure 28: Precision Recall Curve

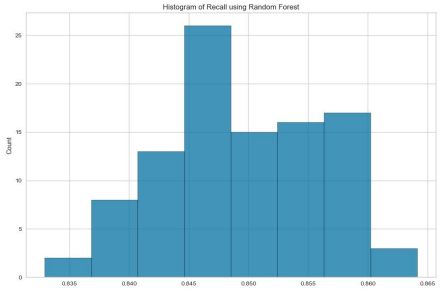


Figure 29: Recall Histogram

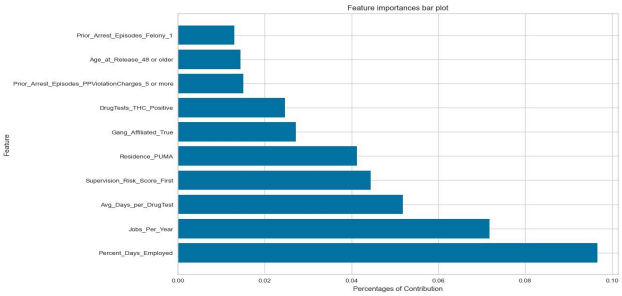


Figure 30: Features Importance Bar Plot

Upon examining the feature importance graph (Figure 30), we observed that the variable 'percent days employed' exhibited the highest level of contribution to our model. It played a crucial role in predicting recidivation rates. Following that, variables such as 'jobs per year,' 'supervision risk score,' and 'residence PUMA' also demonstrated significant predictive power.

The feature importance analysis provided valuable insights into the variables that strongly influenced the model's predictions, with 'percent days employed' being the most influential feature, followed by 'jobs per year,' 'supervision risk score,' and 'residence PUMA'.

XGBoost

We also explored another tree-ensemble algorithm, XGBoost, for our machine learning task. We were motivated to try this method for several reasons: XGBoost provides a customized regularization term, which helps prevent overfitting; The loss function in XGBoost can be tailored to specific requirements; XGBoost requires less feature engineering compared to other models.

We initially utilized the same set of covariates as in the Random Forest model. To ensure model

stability and robustness, we employed cross-validation techniques. Moreover, we conducted a thorough exploration of various hyperparameters, including the number of estimators, maximum depth, and learning rate, in order to obtain the best result with the highest recall.

Eventually, we achieved a recall of 0.849 by setting the following hyperparameters: number of trees: 300, maximum depth: 3, and learning rate: 0.1.

```
Best hyperparameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 300}
Recall: 0.849379402884938
```

In a single run of the XGBoost model, the recall for label 1 was approximately 0.849 (Figure 31). Additionally, the area under the ROC curve was 0.84 (Figure 32), and the area under the Precision-Recall Curve was 0.87 (Figure 33), demonstrating superior performance compared to the Random Forest model. Moreover, the confusion matrix (Figure 31) exhibited the highest counts in True Positives and the lowest counts in False Negatives.

Results from running the XGBoost model 100 times, as shown in the histogram, provided a median recall rate of approximately 0.835 (Figure 34). Although slightly lower than the Random Forest model, this result still reflects the strong predictive capability of XGBoost in capturing recidivation patterns.

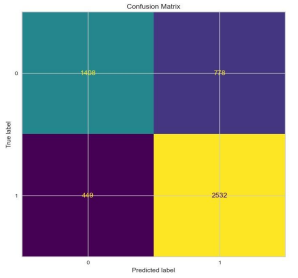


Figure 31: Confusion Matrix

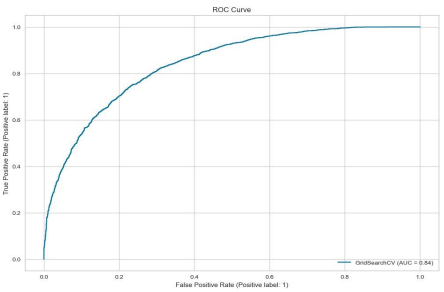


Figure 32: ROC Curve

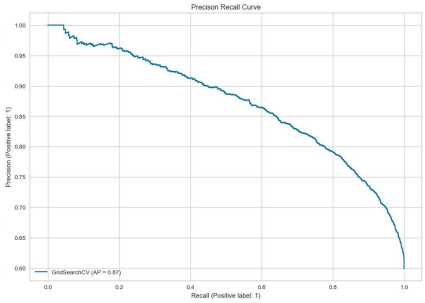


Figure 33: Precision Recall Curve

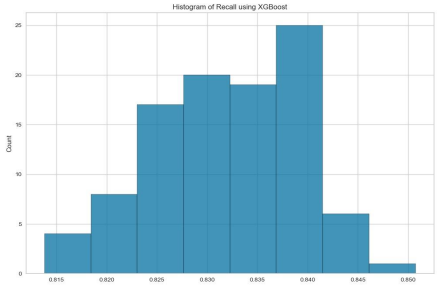


Figure 34: Recall Histogram

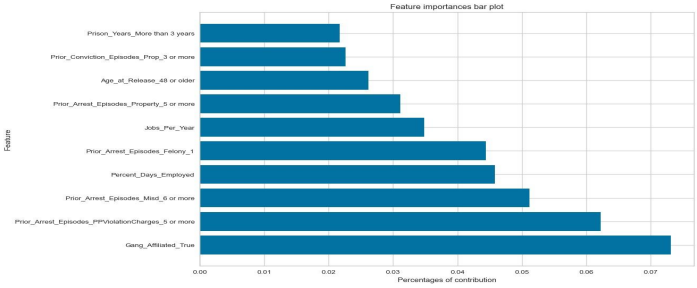


Figure 35: Features Importance Bar Plot

Upon reviewing the feature importance graph (Figure 35), we observed that several variables exhibited predictive power for the recidivation rate. Specifically, 'gang affiliated,' 'prior arrest PPViolationCharges,' 'prior arrest misdemeanor,' and 'percent days employed' emerged as influential features.These variables played a significant role in determining the model's predictions and contributed to its ability to capture patterns related to recidivation rates.

Conclusion

After conducting a comprehensive analysis, our findings revealed that several factors have an impact on the recidivation rate. These factors include gender, percent days employed, gang affiliation, prison years, prison offense type, prior crime type, risk score, and education level. The improved recall achieved by Random Forest and XGBoost models, compared to Logistic Regression, suggests the presence of interactions between these covariates.

In the Random Forest model, variables such as percent days employed, jobs per year, supervision risk score, and residence PUMA demonstrated some predictive power for the recidivation rate. Similarly, in the XGBoost model, gang affiliation, prior arrest PPViolationCharges, prior arrest misdemeanor, and percent days employed were identified as variables with predictive power for recidivation.

However, despite these findings, we did not discover a distinct pattern that could reliably differentiate criminal recidivism. Further analysis and exploration may be required to uncover additional factors or relationships that contribute to the recidivation rate.

Model	Logistic Regression	Random Forest	XGBoost
Recall	0.802	0.847	0.835

Table 1: Recall for different models

Among the three models, our team suggests using the Random Forest model to predict whether or not a criminal will recidivate within three years. By utilizing this model, the crime system can allocate its limited resources more efficiently to track down the criminals who are more likely to recidivate in the future.

Limitations

Despite our findings, there were several limitations to our analysis. Firstly, the data quality was not optimal. On one hand, we had a limited variety of data, such as only having data for the African American and White races. On the other hand, there were missing values, particularly in the columns related to drug test results. Although we used the mode value to fill in these missing values, it may have affected the accuracy of our model predictions, including the recall results.

Additionally, we faced hardware restrictions that prevented us from running the Random Forest and XGBoost models multiple times with different hyperparameter sets. This limitation impacted the stability and robustness of our models and their overall performance.

Furthermore, the Random Forest and XGBoost models lack interpretability as they are considered black box algorithms. This means that understanding the underlying reasoning behind their predictions can be challenging.

Lastly, in future work, incorporating knowledge from neural networks could potentially improve our analysis and enhance the predictive capabilities of our models.

Reference

Uocoeeds. (2022, June 23). *Recidivism*. Kaggle.https://www.kaggle.com/datasets/uocoeeds/recidivism

Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2016, May 23). *How we analyzed the compas recidivism algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing