KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Beyond the Bars: Predicting Recidivism Using Employment and Risk Data

Julyana Ayache– May 2025

Dr. Junjun Huo

INTRODUCTION

- Recidivism, or the tendency for formerly incarcerated individuals to be re-arrested, poses a significant challenge for the criminal justice system in the United States.
- The dataset was obtained from the NIJ Recidivism Challenge from Data.gov and includes over 25,000 formerly incarcerated individuals tracked for 3 years post-release.
- This project aimed to identify key personal, criminal, and behavioral factors that predict recidivism within three years to support more effective intervention strategies. The analysis included variables such as gender (1), race (2), age at release (3), gang affiliation (4), education level (5), prison offense and years served (6), prior arrests and convictions across various crime types (7–13), parole and probation violations (14), mental health and substance abuse conditions (15), supervision risk score (16), employment history (17), drug test results (18), program attendance (19), and housing stability indicators (20).

METHODS

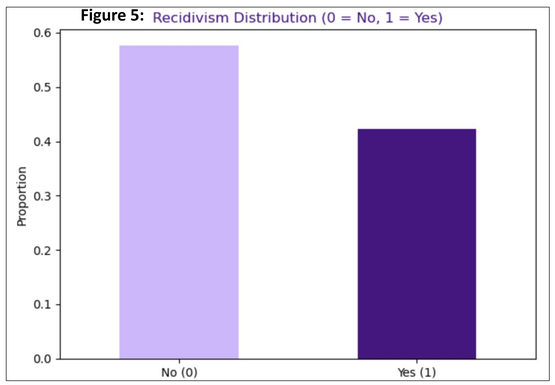
- Feature Selection (Figure 1): ANOVA F-tests were used to select statistically significant predictors of recidivism, keeping only those with p-values < 0.05.
- Correlation (Figure 2): A correlation bar chart assessed the direction and strength of each predictor's relationship with recidivism.
- Plots: Histograms were created to visualize how Percent_Days_Employed and Supervision_Risk_Score_First varied by recidivism status.
- Logistic Regression: Used as a baseline model to evaluate the relationship between employment, supervision, and recidivism.
- Random Forest: A random forest classifier was used to capture nonlinear relationships, achieving an AUC of 0.71 on the test set.
- Model Comparison: K-Nearest Neighbors and Support Vector Machines were tested but underperformed compared to Random Forest.
- Evaluation: Accuracy, AUC, and classification reports were used to compare model outcomes.
- Feature Importance: A Random Forest feature importance chart highlighted Avg_Days_per_DrugTest and Percent_Days_Employed as the most influential variables.

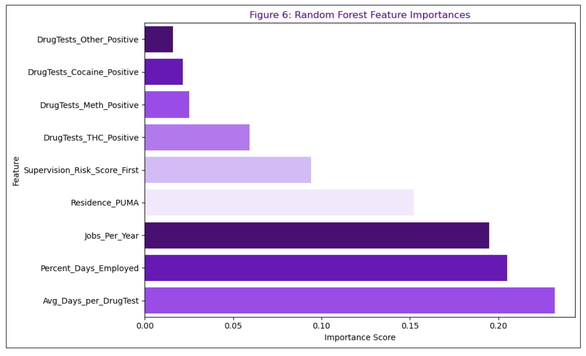
RESULTS

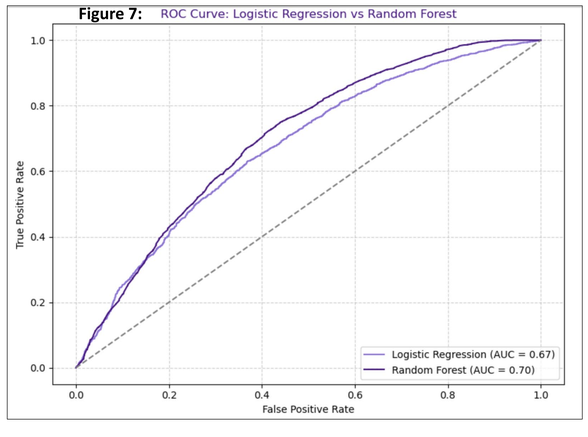
- Random Forest Model (Figure 5)
  - AUC = 0.71, indicating strong classification performance
  - Accuracy = 67–70%, outperforming logistic regression and KNN
- Top Predictors (Figure 6):
  - Avg_Days_per_DrugTest – Most important; frequent drug testing linked to higher recidivism risk
  - Percent_Days_Employed – Fewer days employed strongly associated with reoffending
  - Jobs_Per_Year – High job turnover indicated higher recidivism
  - Residence_PUMA – Geographic region influenced recidivism patterns
  - Supervision_Risk_Score_First – Higher risk scores increased reoffense likelihood
  - DrugTests_THC_Positive – Substance use was a moderate predictor
- Recidivism Distribution: The class distribution is imbalanced, with more individuals not reoffending than reoffending. Accuracy alone may be misleading — AUC, precision, and recall were included for deeper analysis.

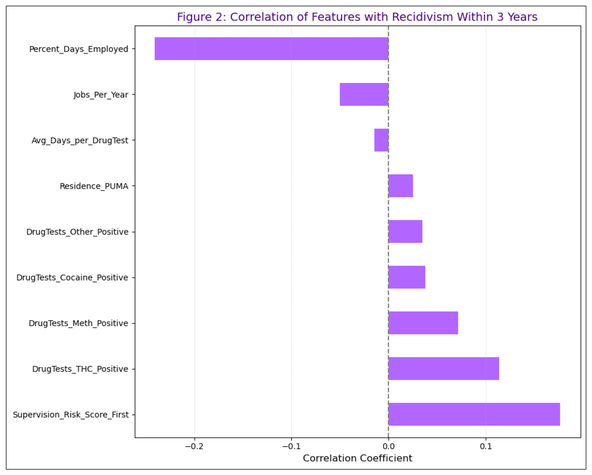
Figure 1: F-values and P-values (Most to Least Significant)

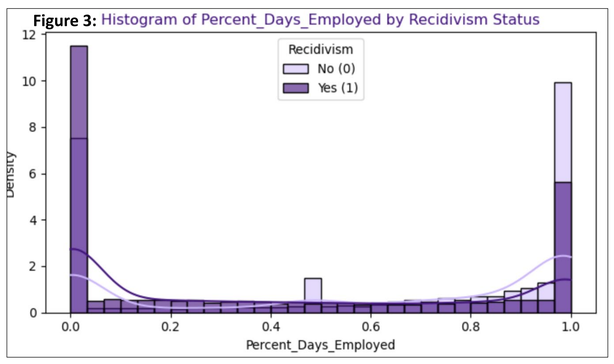
Feature	F Value	P Value
Percent_Days_Employed	1587.67	0.00000
Supervision_Risk_Score_First	834.65	0.00000
DrugTests_THC_Positive	339.84	0.00000
DrugTests_Meth_Positive	132.43	0.00000
Jobs_Per_Year	64.86	0.00000
DrugTests_Cocaine_Positive	37.77	0.00000
DrugTests_Other_Positive	31.36	0.00000
Residence_PUMA	16.42	0.00005
Avg_Days_per_DrugTest	5.30	0.02133

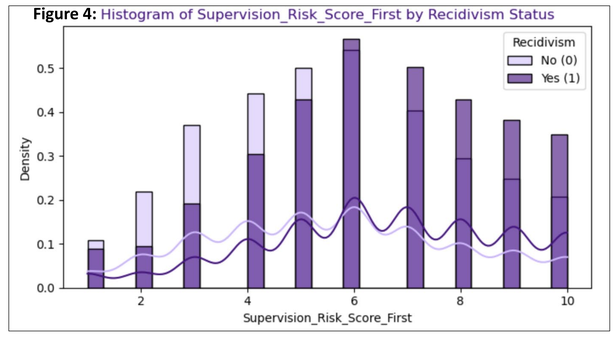












Figures 3 & 4: These histograms visualize how key predictors differ between individuals who recidivated (Yes) and those who did not (No). Recidivism is more common among individuals with higher supervision risk scores and lower employment percentages. These visualizations support the model's findings and illustrate why these features were among the most important in predicting recidivism.

Figure 8:

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.646	0.664	0.782	0.718	0.675
Random Forest	0.669	0.698	0.752	0.724	0.701

DISCUSSION

- Support Employment Stability: Stable employment was the strongest protective factor; job training and placement should be prioritized in reentry programs.
- Strengthen Supervision Assessment: High supervision scores predicted recidivism; agencies should enhance monitoring and allocate resources accordingly.
- Address Drug Use Patterns: Positive and frequent drug tests correlated with recidivism; expanding access to treatment may reduce reoffending.
- Target High-Risk Regions: Geographic factors mattered; localized programs should be developed to meet community-specific needs.
- Leverage Predictive Models: Random Forest helped identify key predictors; early intervention strategies can benefit from such tools if applied ethically.
- Additional Insights
  - Model Limitations: Results are data-driven but may not capture social or systemic influences like bias in supervision or policing.
  - Ethical Use of AI: Predictive tools must be monitored for fairness to avoid reinforcing inequality in criminal justice decisions.
  - Policy Implications: Agencies can use insights to tailor supervision intensity and reentry support based on individual profiles.
  - Data Constraints: Results are based on available variables — future work should include social factors, community resources, and mental health access.

CODE

Data Cleaning
```
df.fillna(df.mean(numeric_only=True), inplace=True)
```

LogisticRegression
```
log_model = LogisticRegression(C=1, max_iter=1000)
log_model.fit(X_train, y_train)
```

RandomForestClassifier
```
rf_model = RandomForestClassifier(n_estimators=100)
rf_model.fit(X_train, y_train)
```

ROC/AUC Score
```
roc_auc_score(y_test, model.predict_proba(X_test)[:, 1])
```

Evaluation
```
print(classification_report(y_test, model.predict(X_test)))
```

Linkedin