Recidivism Forecasting: Comparing Logistic Regression, LASSO, and Random Forest Models

Debang Ou
University of Connecticut

December 7, 2024

Abstract

Recidivism, the tendency of previously incarcerated individuals to reoffend, lead to significant challenges to public safety and rehabilitation efforts. This study compares three predictive models—Logistic Regression, LASSO, and Random Forest—in their ability to forecast recidivism using a dataset from the Georgia Department of Community Supervision. The primary objective is to assess these models in terms of accuracy and fairness(FPR and FNR), and to determine whether more advanced machine learning models offer improvements over simpler techniques without exacerbating biases.

1 Introduction

Recidivism, the tendency of individuals who have served time to reoffend, has been a critical issue in the criminal justice system. High recidivism rates pose a challenge to public safety and impact rehabilitation efforts to help individuals reintegrate into society. Research shows that nearly two-thirds of people released from prison in the United States are rearrested within three years (Durose et al., 2014), highlighting the urgent need for improved predictive models to guide the implementation of interventions. By accurately identifying individuals at higher risk of recidivism, predictive models can enhance decision-making on probation, parole, and rehabilitation, ultimately reducing recidivism rates (Berk, 2009).

Although traditional statistical models such as logistic regression have been widely used to predict recidivism, these models often fall short in capturing the complex nonlinear relationships between risk factors and outcomes (James et al., 2013). Additionally, recent research has raised questions about the fairness of recidivism prediction models because these tools may inadvertently reinforce biases based on race, socioeconomic status, and other factors (Angwin et al., 2016; Chouldechova, 2017). Machine learning techniques (such as LASSO and random forests) offer new possibilities that can improve prediction performance through feature selection and nonlinear interactions (Hastie et al., 2015). However, it is currently unclear how these advanced models compare to simpler models in terms of predictive accuracy and fairness. Existing research often focuses on the predictive accuracy of models, but few

comprehensive analyses balance accuracy and fairness in recidivism prediction (Berk et al., 2018).

This study aims to fill this gap by systematically comparing the predictive capabilities of three models (logistic regression, LASSO, and random forests) on a large recidivism data set. Specifically, we explore the following questions: (1) How accurate and fair are these models in predicting recidivism? (2) Can advanced models like LASSO and Random Forest significantly improve prediction performance without exacerbating bias?

The remainder of this article is structured as follows. We first provide a detailed description of the data used in our study, focusing on key features relevant to the prediction of recidivism. Next, the research methodology is described, including the specific metrics used to evaluate the accuracy and fairness of each model. In the results section, we present and analyze the performance of each model. Finally, we delve into the implications of the results for policy development, limitations of the study in the Discussion section.

To provide a structured approach to evaluating model performance, we articulate the following hypotheses:

- **Hypothesis 1**: The Random Forest model will achieve higher predictive accuracy compared to Logistic Regression and LASSO due to its capability to model complex non-linear relationships.
- **Hypothesis 2**: LASSO will effectively reduce overfitting by feature selection, providing a balance between model simplicity and predictive power, but may not achieve the same level of accuracy as Random Forest.

2 Data

This study uses data from the Georgia Prison Recidivism Prediction Dataset, provided by the National Institute of Justice (NIJ). The dataset includes records for approximately 26,000 individuals who were released on parole under the supervision of the Georgia Department of Community Supervision (GDCS) between January 1, 2013, and December 31, 2015.

2.1 Data Content

Key attributes in the dataset include:

- **Demographic Information**: Basic demographic information includes `Gender`, `Race`, and `Age_at_Release`. The `Residence_PUMA` column indicates the geographical area of residence, and `Gang_Affiliated` reflects any recorded gang affiliation.
- **Supervision and Risk Scores**: `Supervision_Risk_Score_First` and `Supervision_Level_First` provide details on each individual's risk assessment and supervision level at the time of release.
- **Education and Family Structure**: `Education_Level` describes the highest education level attained, and `Dependents` captures the family responsibility, both of which are potential factors influencing recidivism risk.

- **Criminal History**: Numerous fields track prior arrests and convictions, broken down by type:
  - Felony offenses (`Prior_Arrest_Episodes_Felony`, `Prior_Conviction_Episodes_Felony`)
  - Misdemeanor, violent, drug-related, and property-related offenses
  - Probation and parole violations, domestic violence, and gun charges
- **Supervision Conditions and Violations**: Supervision violation data includes variables such as `Violations_ElectronicMonitoring`, `Violations_FailToReport`, and `Violations_MoveWithoutPermission`. Additional fields capture conditions for participation in mental health, substance abuse, and cognitive education programs (`Condition_MH_SA`, `Condition_Cog_Ed`).
- **Program Engagement and Employment**: Engagement in assigned programs is recorded in `Program_Attendances` and `Program_UnexcusedAbsences`, while employment data, such as `Percent_Days_Employed` and `Jobs_Per_Year`, reflects economic stability during supervision.
- **Drug Testing Results**: Positive results for THC, cocaine, methamphetamine, and other substances are recorded in columns such as `DrugTests_THC_Positive`, `DrugTests_Cocaine_Positive`, and `DrugTests_Meth_Positive`.
- **Recidivism Outcomes**: The variable `Recidivism_Within_3years` indicates any recidivism within three years, while year-specific indicators (`Recidivism_Arrest_Year1`, `Recidivism_Arrest_Year2`, `Recidivism_Arrest_Year3`) track the timing of reoffenses.

2.2 Data Sources

This dataset is collaboratively provided by GDCS and the Georgia Bureau of Investigation (GBI). Data from GDCS includes demographic details, information on incarceration and parole cases, prior community supervision history, probation and parole conditions (as set by the Board of Pardons and Paroles), and records of supervision activities (such as violation records, drug testing, program participation, employment, residential moves, and accumulated reports of violations for breaking parole conditions). The GBI's data comes from the Georgia Crime Information Center (GCIC) statewide criminal history database. The GCIC database records an individual's arrest and conviction history prior to incarceration, detailing each arrest incident with the most serious charge. Certain offenses, like domestic violence or firearms violations, include all relevant charges. Additionally, GCIC data provides a measure of recidivism, capturing any new felony or misdemeanor arrests within three years of starting parole supervision.

2.3 Data Preprocessing

To prepare the data for modeling, the following preprocessing steps were conducted to enhance data quality, ensure compatibility with chosen models, and prevent data leakage:

- **Removed Non-Predictive Columns:** Columns such as `ID`, `Recidivism_Arrest_Year1`, `Recidivism_Arrest_Year2`, `Recidivism_Arrest_Year3`, and `Training_Sample` were removed. These columns were either identifiers or specific to training, and therefore irrelevant to prediction.

- **Categorical Conversion:** Columns representing categorical information, such as `Gender`, `Race`, `Age_at_Release`, `Residence_PUMA`, `Gang_Affiliated`, `Supervision_Level_First`, `Education_Level`, `Dependents`, and `Prison_Offense`, were converted to categorical types to prepare them for encoding.

- **Handling Missing Values:** Missing values in numerical columns were filled with the mean value, while categorical columns were filled with the mode (most frequent value), preserving the distribution of the data.

- **Encoding Categorical Variables:** Categorical variables were one-hot encoded to enable their use in models requiring numerical input.

- **Feature Scaling:** All numerical features were standardized using z-score normalization, a necessary step for Logistic Regression and LASSO, which are sensitive to scale.

- **Adding Interaction Terms:** For the Logistic Regression model, interaction terms were included to capture combined effects between variable pairs, thus enhancing the model's ability to capture complex relationships.

These preprocessing steps ensure that each feature is optimized for use in the modeling process, reducing noise and improving interpretability.

3 Methods

To compare the effectiveness of different modeling approaches, we implemented three models: Logistic Regression with interaction terms, LASSO, and Random Forest. Each model has distinct strengths in capturing linear and non-linear relationships, as well as handling feature selection.

3.1 Model Selection

- **Logistic Regression with Interaction Terms:** This model allows us to examine relationships between predictor variables and recidivism through a linear model with interaction terms. Interaction terms were added to capture potential combined effects between predictors, as they can often reveal deeper insights into factors influencing recidivism.

- **LASSO:** LASSO, or Least Absolute Shrinkage and Selection Operator, applies regularization to feature selection, reducing model complexity and enhancing interpretability by penalizing coefficients of less significant features.

- **Random Forest:** Random Forest, an ensemble model, combines multiple decision trees and aggregates their outputs. This model effectively handles non-linear relationships and is less susceptible to overfitting due to its voting-based approach across multiple decision trees.

3.2 Assumptions of the Models

Each predictive model has underlying assumptions that impact its suitability and performance:

- **Logistic Regression**: Assumes a linear relationship between predictors and the log-odds of the outcome. It also assumes no multicollinearity among predictors and requires a sufficiently large sample size for stable estimates.
- **LASSO**: Similar to logistic regression, LASSO assumes linearity between predictors and the outcome.
- **Random Forest**: Unlike the other two models, Random Forest, as an advanced machine learning algorithm, does not require linear relationships. However, it assumes sufficient variability in predictor features to form meaningful splits, and is generally robust to overfitting given enough trees are used.

3.3 Evaluation Metrics

To assess model performance, we used the following metrics:

- **Mean Accuracy and ROC AUC:** These metrics provide an overview of model accuracy and discriminatory ability. ROC AUC, specifically, measures the models' ability to distinguish between recidivists and non-recidivists.
- **False Positive Rate (FPR) and False Negative Rate (FNR):** In a criminal justice context, FPR indicates the rate at which non-recidivists are incorrectly classified as recidivists, while FNR measures the rate at which recidivists are misclassified as non-recidivists. Both rates are crucial in evaluating fairness and the potential social impact of model predictions.

3.4 Model Performance and Interpretation

Table 1 summarizes the performance of each model based on the outlined evaluation metrics.

Interpretation The results demonstrate distinct strengths and limitations across the models:

- **Logistic Regression with Interactions:** With a mean accuracy of 0.610 and ROC AUC of 0.642, Logistic Regression shows relatively low predictive performance, likely due to its linear nature and limited capacity to capture complex non-linear interactions between risk factors. The FPR and FNR rates indicate higher misclassification of recidivists and non-recidivists, posing concerns for real-world application.

Table 1: Model Performance Comparison

Model	Mean Accuracy	Mean ROC AUC	FPR	FNR
Logistic Regression with Interactions	0.610	0.642	0.432	0.358
LASSO	0.709	0.690	0.443	0.176
Random Forest	0.730	0.793	0.414	0.161

- **LASSO:** LASSO achieves a higher mean accuracy of 0.709 and ROC AUC of 0.690, demonstrating improved performance over Logistic Regression by leveraging feature selection. However, its FPR remains similar to Logistic Regression's, while its FNR is largely reduced, indicating improved identification of recidivists.
- **Random Forest:** With the highest accuracy (0.730) and ROC AUC (0.793), Random Forest outperforms the other models in predictive performance. Its relatively low FPR and FNR show a balanced performance, reducing both types of classification errors, making it the most reliable model in terms of accuracy and fairness.

These results suggest that Random Forest may be the most effective model for recidivism prediction, offering both accuracy and reduced bias. Also, the 2 hypotheses mentioned in section one were validated.

4 Discussion and Conclusion

This study contributes to the field of recidivism forecasting by comparing three modeling approaches—Logistic Regression with interaction terms, LASSO, and Random Forest—using a large and detailed dataset from the Georgia Department of Community Supervision. The findings indicate that the Random Forest model outperforms the other models in terms of predictive accuracy and discriminatory power, offering a more nuanced capture of non-linear relationships in recidivism risk factors. These results underscore the potential of machine learning models to improve decision-making within the criminal justice system, particularly by enabling a more accurate identification of individuals at high risk of reoffending.

4.1 Research Questions and Key Findings

The research questions posed in the introduction were addressed as follows:

- **How accurate and fair are Logistic Regression, LASSO, and Random Forest models in predicting recidivism?** Our results show that Random Forest provides the highest accuracy and ROC AUC, followed by LASSO, with Logistic Regression showing lower performance. Random Forest also demonstrated a balance in reducing both False Positive and False Negative rates, making it the most promising model in terms of both accuracy and fairness.
- **Can advanced models like LASSO and Random Forest improve prediction performance without exacerbating bias?** The study suggests that machine learning models like Random Forest, while highly accurate, require careful consideration of

fairness metrics. Though Random Forest showed strong predictive power, its use in practical applications must include measures to monitor and mitigate potential biases, especially those arising from systemic factors within the dataset.

4.2 Limitations of the Study

While the findings provide valuable insights, several limitations should be acknowledged:

- **Lack of Fine-Tuning for Models**: The models used in this study were not fine-tuned extensively due to computational limitations. Hyperparameter optimization, particularly for complex models like Random Forest, could potentially enhance performance further and may reveal different patterns in predictive accuracy or fairness.
- **Limited Cross-Validation Splits**: We applied a limited number of splits in Monte Carlo cross-validation, again due to constraints in computational resources. A higher number of splits would provide a more robust evaluation of model generalizability and performance, reducing variability in the metrics reported.
- **Limited discussion regarding fairness**: We evaluated fairness by comparing FPR and FNR across different models instead of comparing FPR and FNR across subgroups. This comparison provides a basic understanding of fairness by revealing differences in misclassification rates, but a thorough fairness evaluation requires analyzing these metrics across subgroups and including other fairness measures like demographic parity or calibration, which are infeasible in this study.

These limitations suggest that future studies should focus on optimizing model parameters and increasing the number of cross-validation splits to obtain more refined and reliable results.

4.3 Future Research Directions

Building on this study, future research could investigate additional machine learning algorithms, such as gradient boosting or hybrid models, which may further enhance predictive accuracy while managing model interpretability and fairness. There is also a need for fairness-aware machine learning methods that incorporate fairness as a central criterion alongside accuracy. Such methods could help ensure that predictive models do not inadvertently reinforce systemic biases, especially in the criminal justice system. Furthermore, analyzing the impact of individual features—such as employment stability or gang affiliation—on recidivism risk could inform the design of targeted intervention programs aimed at reducing reoffense rates.

4.4 Conclusion

In summary, this study highlights the potential of machine learning models, particularly Random Forest, in forecasting recidivism. The comparative analysis shows that Random Forest achieves superior predictive performance while maintaining relatively balanced error

rates. However, the use of machine learning in recidivism prediction must be approached with a strong emphasis on fairness and ethical considerations. By addressing limitations in model fine-tuning and cross-validation, and by exploring fairness-aware algorithms, future research can further advance the field of recidivism forecasting, eventually contributing to a more equitable and effective criminal justice system.

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine bias.

Berk, R., H. Heidari, S. Jabbari, M. Joseph, and M. Kearns (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research 47*(1), 3–44.

Berk, R. A. (2009). The role of race in forecasts of violent crime. *Race and Social Problems 1*(4), 231–242.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data 5*(2), 153–163.

Durose, M. R., A. D. Cooper, and H. N. Snyder (2014). *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*. Bureau of Justice Statistics Special Report.

Hastie, T., R. Tibshirani, and J. Friedman (2015). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.