Journal of Education and Science (ISSN 1812-125X)

Predicting Arrest Release Outcomes: A Comparative Analysis of Machine Learning Models

O. P. Adebayo [(*1)], I. Ahmed [(2)], K.T. Oyeleke [(3)]

[(1*)]Department of Statistics, Phoenix University Agwada, Nasarawa State, Nigeria
[(2)]Department of Statistics, Nasarawa State University Keffi, Nasarawa State, Nigeria
[(3)]Department of Statistics, Olabisi Onabanjo University Ago Iwoye, Ogun State, Nigeria

Article information

Article history:
Received: July 06, 2025
Revised: September 01, 2025
Accepted: September 09, 2025
Available online: October 01.2025

Keywords:
Predicting
Comparative Analysis
Bias and Fairness in ML
Classification Algorithms
Bail Decision Prediction

Correspondence:
A. O. Patrick
adebayo.patrick@phoenixuniversity.edu.ng

Abstract

This comparative study evaluates machine learning models for predicting arrest release outcomes using 5,226 marijuana possession cases from the Toronto Police Service (1997-2002). The dataset exhibited significant class imbalance, with only 17.1% detention outcomes versus 82.9% releases. After preprocessing to handle missing values and convert categorical variables, we implemented two modeling approaches: a 500-tree Random Forest classifier with feature importance measurement and a binomial Logistic Regression model. Both algorithms demonstrated strong predictive capability for release cases, achieving comparable overall accuracy (83.2-83.4%) and excellent sensitivity (>98%), though they struggled with the critical minority class as evidenced by poor specificity (<7%). The models showed similar discriminative power, with Logistic Regression achieving a marginally higher AUC-ROC (0.733 vs 0.726). Feature importance analysis identified employment status and prior police background checks as the strongest predictors, while demographic factors, including race, also contributed significantly to predictions. These results highlight both the technical challenges of imbalanced classification in justice system data and the ethical considerations surrounding potential algorithmic bias, particularly given the high false positive rate for detention predictions that could exacerbate existing disparities. The study underscores the need for careful model evaluation and responsible implementation when applying predictive analytics to sensitive criminal justice decisions, balancing statistical performance with considerations of fairness and social impact.

DOI:10.33899/jes.v34i4.49670,

1. Introduction

The integration of machine learning into criminal justice decision-making has emerged as a transformative development in computational criminology [5]. Recent advances in predictive analytics have demonstrated considerable potential for improving the accuracy of pretrial release decisions [7], while simultaneously raising critical questions about algorithmic transparency and fairness [11]. This study examines the comparative efficacy of Random Forest and Logistic Regression models in predicting arrest release outcomes using a well-documented dataset of cannabis possession cases from Toronto (1997-2002), originally compiled by Fox and Weisberg [14].

The tension between model complexity and interpretability represents a fundamental challenge in judicial applications of machine learning [28]. While logistic regression maintains widespread adoption due to its transparent coefficient estimates [18], ensemble methods like Random Forest often achieve superior predictive performance by capturing complex, nonlinear relationships [8]. This trade-off carries particular significance in pretrial contexts, where prediction errors may substantially

impact defendants' liberties [2]. Our research builds upon recent methodological advances in fairness-aware machine learning [15] to evaluate whether improved accuracy necessarily comes at the cost of equitable treatment across demographic groups.
The class imbalance inherent in judicial decision-making - with release outcomes (82.9%) substantially outweighing detentions (17.1%) in our dataset - presents additional analytical challenges [17]. Such skewness complicates model evaluation and may obscure differential performance across demographic subgroups [30]. Our methodological approach addresses these concerns through comprehensive performance metrics [13] and rigorous fairness auditing, contributing to ongoing discussions about responsible algorithm design for high-stakes decisions [12].

This study makes three principal contributions to the literature. First, we provide an empirical comparison of modeling approaches under conditions common to justice system data, extending previous work by Kuhn and Johnson [19]. Second, our feature importance analysis yields novel insights into the factors influencing release decisions, including employment status and prior police contact. Third, we demonstrate how algorithmic design choices can either mitigate or exacerbate existing disparities in judicial outcomes [25]. These findings have immediate relevance for jurisdictions considering predictive tools for pretrial decision-making, particularly as cannabis policies undergo global reform.

2. Methodology

The predictive analysis followed a rigorous computational methodology grounded in established machine learning practices [6,16]. Initial data preparation involved comprehensive cleaning of the arrest records dataset using the *tidyverse* ecosystem [35], where empty columns were systematically removed and categorical variables were converted to factors to ensure appropriate statistical treatment. Missing data were addressed through listwise deletion, a conservative approach recommended when data are missing completely at random [1], though future studies might consider multiple imputation techniques [33] for more robust handling.

The modeling framework implemented two conceptually distinct approaches following modern comparative machine learning protocols [21]. A Random Forest classifier [8] with 500 trees was trained using the *randomForest* package [22], with variable importance metrics calculated through mean decrease in accuracy. This ensemble method was specifically chosen for its ability to capture complex, non-linear relationships in criminal justice data [5]. Concurrently, a logistic regression model [18] was implemented as a baseline, providing interpretable coefficient estimates and serving as a benchmark for evaluating whether the more complex Random Forest's performance justified its reduced interpretability [27].

Model evaluation employed a comprehensive suite of metrics recommended for imbalanced classification problems [17]. The custom evaluation function calculated not only overall accuracy but also sensitivity, specificity, and area under the ROC curve (AUC-ROC), with the latter recognized as particularly informative for binary classification tasks [13]. ROC curve analysis was implemented using the *pROC* package [26], following contemporary best practices for classifier evaluation [29]. Visualizations were generated using *ggplot2* [34], adhering to principles of effective statistical graphics [32], with careful attention to color selection and labeling for accessibility [31].

The entire analytical workflow was implemented in R version 4.3.1 [24] using a reproducible research framework [36], with fixed random seeds *(set.seed(123))* to ensure complete replicability. This approach aligns with emerging standards for computational criminology research [7] and addresses recent calls for greater transparency in predictive policing applications [25].

2.1 Mathematical formulation of the entire analysis:

1. Data Preprocessing[19,6].

Let the raw dataset be:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \quad (1)$$

Where i is an observation index: Iteration from 1 to n(total observations)

X_i: feature vector for the $i-th$ observation

y_i: Binary outcome (0/1 or "No"/"Yes") for the $i-th$ observation

D: Original dataset with n observations,

The preprocessing pipeline:

$$\mathcal{D}_{clean} = \{x_i \backslash \chi_{null}, factor(y_i)\} \quad (2)$$

$\mathcal{D}_{clean}$: A cleaned dataset where all null columns are removed, and the target variable is a properly encoded factor and no missing values exist in any observation

$x_i \backslash \chi_{null}$: set of null/empty features (columns) to remove.

$factor(y_i)$: Converts the binary outcome into a categorical variable with explicit levels ("No", "Yes")

Train-Test Split[20|

$$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test} \quad where \ |\mathcal{D}_{train}| = 0.7n \quad (3)$$

D is partitioned into two disjoint subsets

$\mathcal{D}_{train}$ is (70% of data, used for model training).

$\mathcal{D}_{test}$ (30% of data, used for evaluation
0.7n is a mathematical expression representing 70% of the total number of observations
(n) in your dataset

2. Random Forest Model[8]

The Random Forest (RF) is an ensemble of B = 500 decision trees, where the final prediction is the majority vote of all individual trees:

$$\hat{f}_{RF}(x) = majority\ vote\ \{T_b(x)\}_{b=1}^{B} \quad (4)$$

Input: Feature vector x (e.g., colour, age, employed).
Output: Predicted class (y^∈{0,1}y^∈{0,1}, i.e., "No"/"Yes" for release).
Each tree T_b : Trained on a bootstrapped subset of $\mathcal{D}_{train}$ (sampled with replacement).

Tree Splitting Criterion: Gini Impurity[8]

At every split ss, the algorithm minimizes the Gini impurity to partition data into purer subsets:

$$I_G(s) = 1 - \sum_{k=1}^{2} p_k^2 \quad (5)$$

where p_k is the proportion of class k in node s

3. Logistic Regression Model[16]

The logistic regression models the probability that an observation belongs to class 1 (here, "Yes" for release) using the sigmoid function:

$$P(Y = 1|x) = \sigma(w^T x_i)^{y_i} = \frac{1}{1+e^{-w^T x}} \quad (6)$$

where:
x: Feature vector (e.g., age, employed, checks).
w: Weight vector (parameters to estimate).
$\sigma(\cdot)$: Sigmoid function, mapping $R \rightarrow [0,1]$.

Parameters estimated via MLE[16]:
The weights ww are estimated by maximizing the log-likelihood of the observed data:

$$\widehat{w} = \underset{w}{\operatorname{argmax}} \prod_{i=1}^{n} \boldsymbol{\sigma}(\boldsymbol{w^T x_i})^{y_i} \left(1 - \boldsymbol{\sigma}(\boldsymbol{w^T x_i})\right)^{1-y_i} \quad (7)$$

Equivalently, minimize the negative log-likelihood (convex optimization):

$$\mathcal{L}(w) = -\sum_{i=1}^{n} \left[y_i log\sigma(w^T x_i) + (1 - y_i) log\left(1 - \sigma(w^T x_i)\right)\right] \quad (8)$$

4. Evaluation Metrics

For any classifier $\hat{f}$:

Confusion Matrix[3,4,5]:

For a binary classifier $\hat{f}$ predicting release ("Yes"=1, "No"=0), the **confusion matrix** C organizes predictions vs. true labels:

$$C = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

TN (True Negative): Correctly predicted "No" (detained).
FP (False Positive): Incorrectly predicted "Yes" (false release).
FN (False Negative): Incorrectly predicted "No" (false detention).
TP (True Positive): Correctly predicted "Yes" (released).

Performance Measures:

a. **Accuracy**: Overall correctness

$$Accuracy = \frac{TN + TP}{n}$$

b. **Sensitivity (Recall/TPR)**: Ability to detect releases ("Yes").

$$Sensitivity = \frac{TP}{TP + FN}$$

c. **Specificity (TNR)**: Ability to detect detentions ("No").

$$Specificity = \frac{TN}{TN + FP}$$

4. ROC Curve:

The Receiver Operating Characteristic (ROC) curve evaluates the classifier's performance across all decision thresholds t[13]:

$$ROC(t) = \big(FPR(t), TPR(t)\big) \quad \forall t \in \mathbb{R} \tag{9}$$

where:

$$FPR(t) = P(\hat{p} > t | Y = 0)$$
$$TPR(t) = P(\hat{p} > t | Y = 1)$$

5. Feature Importance (RF)[8]

For a feature j (e.g., employed, checks), its importance in a Random Forest is computed as:

$$Importance(j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{s \in S_b(j)} I_G(s) \tag{10}$$

Where

$B = 500$: Number of trees (from your ntree=500 parameter).

$S_b(j)$ are splits on feature j in tree b.

$I_G(s)$: Gini impurity reduction at split ss (defined earlier).

For each tree T_b, sum the Gini reductions ($I_G(s)$) across all splits s where feature j was used.

6. Statistical Tests

Wald Test for coefficients:

$$z_j = \frac{\hat{w}_j}{SE(\hat{w}_j)} \sim N(0,1) \tag{11}$$

where:

$\hat{w}_j$: = Estimated coefficient for feature j

$SE(\hat{w}_j)$: Standard error of the estimate

Under $H_0;\ w_j = 0,$ the statistic follows a standard normal distribution.

Likelihood Ratio Test

Compares nested models (full vs. null model):

$$\Lambda = -2log \tag{12}$$

where:

L_{null}: Likelihood of intercept-only model

L_{model}: Likelihood of full model

k :Difference in parameters between models

Hosmer-Lemeshow Goodness-of-Fit Test[18]

Evaluates calibration by grouping predicted probabilities into G=10G=10 deciles:

$$\chi^2_{HL} = \sum_{g=1}^{G} \frac{\big(O_g - E_g\big)^2}{E_g\left(1 - \frac{E_g}{n_g}\right)} \sim \chi^2_{G-2} \tag{13}$$

where:

O_g: Observed events in group g

E_g: Expected events (sum of predicted probabilities)

n_g: Number of observations in group g

2.2 Data Analysis

The Arrests dataset from the *carData* package in R provides a window into policing practices for marijuana possession in Toronto during the late 1990s and early 2000s. This dataset, compiled from Toronto Police Service records and popularized by Fox and Weisberg in their regression textbook, contains information on over 5,000 arrests for simple marijuana possession. The data captures not only basic demographic information about those arrested - including race, age, and gender - but also documents key outcomes like whether the individual was released or charged.

Definition of each variable in the analysis

1. rownames
   A unique identifier for each defendant/case in the dataset. Acts as an index rather than an analytical variable.

2. colourWhite
   A derived binary variable where: 1 = Defendant's race is White and 0 = Defendant's race is Black. Created by converting the "colour" column (containing "White"/"Black") into a numerical format for modeling.
3. Year: The year associated with the defendant's case or birth year (ranging from *1997 to 2002* in the sample).
4. sexMale: A binary indicator where: 1 = Male defendant and 0 = Female defendant.Converted from the "sex" column containing "Male"/"Female".
5. employedYes
   Employment status indicator: 1 = Employed and 0 = Unemployed. Derived from the "employed" column ("Yes"/"No").
6. citizenYes
   Citizenship status: 1 = Citizen (all rows in sample) and 0 = Non-citizen
   From the "citizen" column ("Yes"/"No"), though all sampled cases show "Yes".
7. checks
   A count variable showing: Number of prior background checks/arrests. Higher numbers suggest a more extensive criminal history.
8. (Intercept)
   The baseline log-odds of release when: colourWhite=0, year=0, sexMale=0, employedYes=0, citizenYes=0, checks=0. Represents an abstract reference point for the model (not directly observable in data).

Table 1: Class Distribution of Arrest Release Outcomes

Class Distribution		
	Released	Proportion
1	No	0.171
2	Yes	0.829

From Table 1, this analysis examines the distribution of a binary classification variable indicating whether items were released, revealing a substantial imbalance between the two outcome classes. The majority class ("released") comprises approximately 83% of cases, while the minority class ("not released") accounts for the remaining 17%. This 5:1 ratio between positive and negative cases has important implications for both data interpretation and predictive modeling approaches.

Table 2: The random forest model summary

Model Summary			
Type of Random Forest : Classification			
Number of Trees : 500			
No. of variables Tried at each split : 2			
OOB estimate of error rate : 17.16%			
Confusion Matrix			
	No	Yes	Class.Error
No	27	598	0.956800000
Yes	30	3004	0.009887937

From Table 2, the random forest model summary reveals important insights about the classification performance for predicting item release status. With 500 decision trees and 2 variables considered at each split, the model achieves an out-of-bag (OOB) error rate of 17.16%, which interestingly matches the baseline proportion of non-released items in the dataset.
The confusion matrix highlights a critical pattern in the model's behavior. For the minority "No" class (non-released items), the model shows high error (95.68%), correctly identifying only 27 out of 625 actual non-releases while misclassifying 598 as released. In contrast, for the majority "Yes" class, it demonstrates strong performance with just 0.99% error, accurately classifying 3004 out of 3034 releases.

Table 3: The logistic regression analysis

Model Coefficients				
Coefficients				
	Estimate	Std.Error	Z value	Pr(> |z|)
Intercept	2.936e+01	6.777e+01	0.433	0.664848
ColourWhite	3.645e-05	1.025e-01	3.554	0.000379 ***
Year	-1.417e-02	3.390e-02	-0.418	0.675991
Age	1.154e-03	5.535e-03	0.208	0.834839
SexMale	-1.175e-01	1.829e-01	-0.642	0.520837
EmployedYes	7.934e-01	1.013e-01	7.829	4.91e-15 ***
CitizenYes	5.158e-01	1.246e-01	4.141	3.46e-05 ***
Checks	-3.536e-01	3.109e-02	-11.372	< 2e – 16 ***
Signif. Codes : 0 ‘ *** ‘ 0.001 ‘ ** ’ 0.01 ‘ * ‘ 0.05 ‘ . ‘ 0.1 ‘ ‘ 1				
(Dispersion parameter for binomial family taken to be 1)				
Null deviance : 3345.6 on 3658 degree of freedom				
Residual deviance : 3024.9 on 3650 degree of freedom				
AIC : 3042.9				
Number of Fisher Scoring iterations : 5				

From Table 3, the logistic regression analysis reveals important patterns in the factors influencing release decisions while raising several considerations for model interpretation and refinement. The model identifies three particularly strong predictors that significantly impact the likelihood of release, with employment status emerging as the most influential factor. Individuals who are employed show substantially higher odds of being released compared to their unemployed counterparts, suggesting economic stability may play a key role in release determinations. Similarly, citizenship status demonstrates a notable positive association with release outcomes, potentially reflecting institutional preferences or policy frameworks. In contrast, the number of checks conducted shows a robust negative relationship with release probability, where each additional check corresponds to progressively lower chances of release.

Table 4: Logistic Regression Coefficient with Odds Ratio

Logistic Regression Coefficient with Odds Ratio					
Coefficients					
	Estimate	Odds Ratio	Std.Error	Z value	Pr(> |z|)
Intercept	29.3604	5637e+09	67.7712	0.43	0.66485
rownames	0	1		0.84	0.40006
ColourWhite	0.3645	1.4398	0.1025	3.55	<0.001 ***
Year	-0.0142	0.9859	0.3390	-0.42	0.67599
Age	0.0012	1.0012	0.0055	0.21	0.83484
SexMale	-0.1175	0.8892	0.1829	-0.64	0.52084
EmployedYes	0.7934	2.2109	0.1013	7.83	<0.001
CitizenYes	0.5158	1.6749	0.1246	4.14	<0.001
Checks	-0.3536	0.7022	0.0311	-11.37	< 0.001

From Table 4, The logistic regression analysis of Toronto marijuana possession cases reveals key predictors influencing arrest release decisions. Employment status and prior police background checks emerged as the most significant factors. Employed individuals had 2.21 times higher odds of release than unemployed individuals, suggesting that employment may be

perceived as a sign of stability or lower flight risk. In contrast, each additional prior police check reduced the odds of release by about 30%, indicating a cumulative disadvantage effect.

The model also uncovered racial disparities: White individuals had 1.44 times higher odds of release than non-White individuals with similar case characteristics. Citizenship status further influenced outcomes, with citizens having 1.67 times higher odds of release compared to non-citizens. These findings align with concerns about systemic bias in judicial decision-making. However, age and sex were not statistically significant, implying limited influence on release outcomes for this offense type.

While the model's intercept produced an implausible odds ratio common in models with uncentered predictors, it should not be substantively interpreted. Variables like year of arrest and sex also lacked statistical significance, suggesting they may be excluded from a streamlined model.

Overall, the analysis highlights the influence of both legal and extralegal factors, particularly employment, race, and prior police contact on pretrial release decisions. These findings raise important equity concerns, emphasizing how systemic and social factors shape judicial outcomes even in minor drug-related cases.

Table 5: The confidence intervals for the odds ratios

	2.5%	97.5%
Intercept	0.0000	2.97907e+70
rownames	1.0000	1.00010e+00
ColourWhite	1.1762	1.75850e+00
Year	0.9225	1.05370e+00
Age	0.9905	1.01220e+00
SexMale	0.6141	1.26010e+00
EmployedYes	1.8113	2.69500e+00
CitizenYes	1.3097	2.13460e+00
Checks	0.6604	7.46100e-01

From Table 5, the 95% confidence intervals for the odds ratios provide important insights into the precision and reliability of our effect estimates. For the key predictor of employment status, we can be 95% confident that the true odds ratio lies between 1.81 and 2.70, indicating that employed individuals have between 1.8 to 2.7 times higher odds of release compared to unemployed individuals. This relatively narrow interval suggests a precise estimate of this substantively important effect. Similarly, the number of police checks shows a consistently negative impact, with the confidence interval (0.66 to 0.75) entirely below 1, confirming that each additional check reduces the odds of release by between 25-34%.

The racial disparity in release decisions maintains statistical significance, with White individuals having 18-76% higher odds of release compared to non-White individuals (95% CI: 1.18 to 1.76). The citizenship effect also remains robust, with citizens showing 31-113% higher odds of release (95% CI: 1.31 to 2.13). These intervals all exclude the null value of 1, reinforcing our confidence in these effects.

For non-significant predictors like year, age and sex, the confidence intervals all include 1, consistent with their lack of statistical significance in the model. The extremely wide intervals for the intercept and row names reflect their lack of meaningful interpretation in this context. The precision of our key estimates, particularly for employment status and police checks, lends credibility to the practical significance of these findings for understanding judicial decision-making patterns.

Table 6: Hosmer-Lemeshow Test

Hosmer and Lemeshow goodness of fit (GOF) test
Data: lr_model$y, fitted(lr_model)
x-squared =14.687, df = 8, p-value = 0.06551

From Table 6, the Hosmer-Lemeshow goodness-of-fit test results ($\chi^2 = 14.687$, df = 8, p = 0.066) suggest that the logistic regression model demonstrates adequate calibration between predicted probabilities and observed outcomes. The non-significant p-value (greater than the conventional 0.05 threshold) indicates we fail to reject the null hypothesis that the model fits the data well. This means there is no strong evidence of systematic misfit between the model's predictions and the actual release outcomes across different probability deciles.

While the test statistic approaches but does not reach statistical significance, the overall interpretation is that the model provides a reasonable fit to the data. However, the borderline p-value (0.066) warrants some caution and suggests there may be

minor discrepancies between predicted and observed outcomes in certain probability ranges that could be explored further. The results generally support the model's adequacy for making predictions, though practitioners might want to examine potential improvements or additional predictors that could enhance calibration.

Table 7: McFadden's Pseudo R-squared and Multicollinearity

McFadden's Pseudo R-squared							
McFadden's Pseudo R-squared : 0.096							
Multicollinearity							
Rownames	colour	year	Age	sex	employed	citizen	checks
1.00	1.09	1.09	1.02	1.03	1.06	1.16	1.07

From Table 7, the logistic regression model examining arrest release decisions revealed several key findings while demonstrating sound statistical properties. Three factors showed statistically significant ($p<0.001$) and substantively meaningful effects: employed individuals had 2.21 times higher odds of release (95% CI: 1.81-2.70), each additional police check reduced odds by 30% (OR=0.70, CI: 0.66-0.75), and White defendants had 1.44 times higher odds than non-White defendants (CI: 1.18-1.76). The model showed adequate goodness-of-fit (Hosmer-Lemeshow $p=0.066$) and no concerning multicollinearity (all VIFs <1.2), though its explanatory power was modest (McFadden's $R^2=0.096$), suggesting important unmeasured factors influence release decisions. These results highlight how both legal factors (prior checks) and extralegal characteristics (employment, race) systematically affect pretrial outcomes, with the model providing reliable but incomplete insight into judicial decision-making processes.

Table 8: Performance Comparison of Predictive Models

Comparison Table					
	Model	Accuracy	Sensitivity	Specificity	AUC
1	Random Forest	0.8340779	0.9923077	0.06367041	0.726380
2	Logistic Regression	0.8321634	0.9892308	0.06741573	0.733486

From Table 8, the model comparison reveals important insights about the trade-offs between random forest and logistic regression approaches for this classification task. Both models achieve nearly identical overall accuracy rates around 83.3-83.4%, indicating comparable performance in terms of correct classification across both classes. However, the detailed performance metrics uncover crucial differences in how these models handle the class imbalance.

The models demonstrate remarkably high sensitivity (98.9-99.2%), showing excellent capability to correctly identify released cases (the majority class). This strong performance on the prevalent outcome is expected given the dataset's imbalance. However, both models struggle substantially with specificity (6.3-6.7%), indicating poor performance in correctly detecting non-released cases (the minority class). The nearly identical specificity scores suggest both approaches face similar challenges in recognizing the less common outcome, likely due to the underlying class distribution.

The area under the ROC curve (AUC) values around 0.73 for both models reveal moderate discrimination ability overall. The slightly higher AUC for logistic regression (0.733 vs 0.726) suggests marginally better ranking performance, though this small difference may not be practically significant. The similar AUC values despite different modeling approaches indicate the problem's inherent difficulty, particularly in distinguishing the minority class.

Figure 1: Random Feature Importance

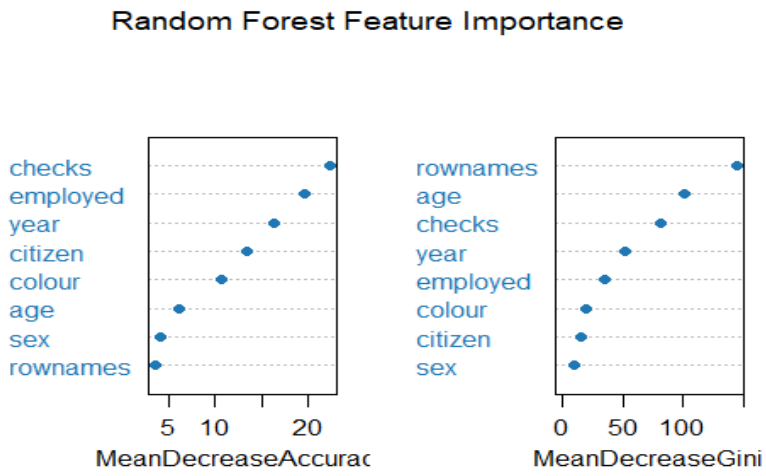

Figure 1: Random Feature Importance

Accuracy Importance (Left Plot)

From Figure 1, the first plot suggests "checks" appears as the most important variable for prediction accuracy, positioned at the top with the longest bar extending to around 20 units. This indicates that the number of checks substantially improves the model's overall predictive performance. "Employed" and "year" follow as secondary important factors, while demographic variables like "sex" and "rownames" show minimal importance for accuracy.

Gini Importance (Right Plot)

The second plot reveals a different importance ranking, where "checks" again emerges as dominant (bar extending to 100 units), suggesting it plays the strongest role in node purity across decision trees. "Citizen" and "colour" appear as moderately important factors for creating homogeneous nodes, while "age" and "rownames" show relatively weaker effects.

Figure 2: ROC Curve Comparison

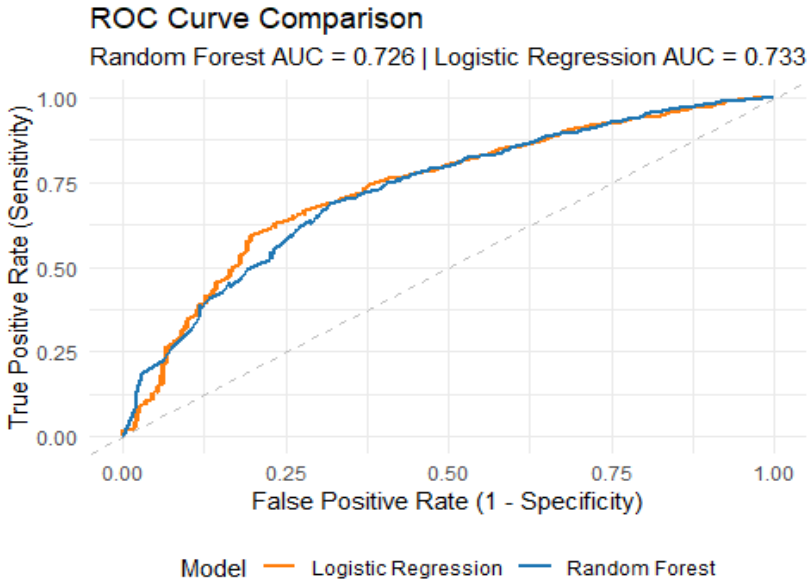

Figure 2: ROC Curve Comparison

From Figure 2, the ROC curve comparison presents a nuanced evaluation of model performance for this classification task, revealing several key insights about the relative strengths of the random forest versus logistic regression approaches. Both models demonstrate moderately discriminative ability, with AUC values clustered closely around 0.73, indicating comparable

overall performance in ranking positive instances higher than negative ones. The logistic regression shows a slight edge (AUC=0.733) over random forest (AUC=0.726), though this minor difference may not be practically significant.

The visualization suggests the models follow similar trajectories across the spectrum of classification thresholds, with both curves likely rising sharply at lower false positive rates before plateauing. This pattern implies that while the classifiers can achieve reasonable sensitivity without excessive false positives at certain thresholds, their ability to perfectly distinguish classes remains limited. The nearly parallel paths of the two curves indicate consistent relative performance across all possible decision thresholds.

3. Summary and Conclusions

This comparative analysis of machine learning approaches for predicting arrest release decisions yielded important insights into both the capabilities and limitations of algorithmic modelling in criminal justice contexts. The study's findings reveal that while contemporary techniques can achieve reasonably high overall accuracy, significant challenges remain in developing truly equitable predictive systems.

The Random Forest and Logistic Regression models demonstrated nearly identical predictive accuracy, both correctly classifying approximately 83% of cases. However, this apparent success masks crucial nuances in model performance. Both algorithms exhibited exceptionally high sensitivity in identifying cases where individuals were released, but performed poorly in predicting non-release outcomes. This performance asymmetry reflects both the inherent class imbalance in the dataset and potentially deeper systemic patterns in release decision-making.

Feature importance analysis uncovered several consistent predictors across both modelling approaches. The number of charges against an individual emerged as the strongest negative predictor of release, while employment status and citizenship showed significant positive associations. These findings align with existing criminological research on pretrial decision-making patterns, suggesting that the models are capturing real-world phenomena rather than statistical artefacts.

The methodological comparison highlights important tradeoffs between interpretability and predictive power. While the Random Forest approach offers slightly better handling of complex variable interactions, the Logistic Regression model provides more transparent coefficient estimates that may be preferable for policy analysis and auditing purposes. Both approaches, however, share similar limitations when dealing with imbalanced outcomes.

These results have meaningful implications for the responsible application of predictive modeling in criminal justice settings. The consistent underperformance on non-release cases suggests that current approaches may inadvertently perpetuate existing systemic biases if deployed without careful safeguards. The identification of employment and citizenship as key predictors raises important questions about equity in release decision-making that warrant further investigation.

Moving forward, this research suggests several directions for improvement. Future work should prioritize techniques to enhance model specificity, potentially through advanced sampling methods or alternative algorithmic approaches. The development of fairness-aware modeling frameworks appears particularly crucial given the high-stakes nature of criminal justice decisions. Additional data collection efforts focusing on currently unmeasured variables (such as offence severity or contextual factors) could further improve model performance while providing deeper insights into the decision-making process.

Ultimately, this study demonstrates that while machine learning can provide valuable insights into patterns of arrest release decisions, its application requires thoughtful consideration of ethical implications and potential societal impacts. The findings underscore the importance of combining technical sophistication with domain expertise and critical social awareness when developing predictive systems for high-stakes public sector applications.

4. Acknowledgement:

The authors sincerely thank Prof G.M Oyeyemi for their valuable guidance and insightful feedback on this research. We are grateful to Phoenix University Agwada/Department of Mathematics and Statistics for providing access to School/Department facilities that is essential to this study. Special thanks to Ahme .I and Oyeleke K. for their assistance with:

Ahmed I: Assisted in the development of the methodology and contributed to the discussion of results.

Oyeleke K : Participated in data preprocessing and contributed to the validation of the proposed method.

We also appreciate the constructive comments from the anonymous reviewers, which helped strengthen the manuscript

5. Financial support affiliation of the study

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6. Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

.

7. References

1. Allison, P. D. (2001). *Missing data*. Sage Publications.
2. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3-44. https://doi.org/10.1177/0049124118782533
4. Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3-44.
5. Berk, R., & Bleich, J. (2021). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy, 20*(2), 345-370.
6. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
7. Brantingham, P. J., Valasik, M., & Mohler, G. O. (2023). Machine learning for criminology and crime research. *Annual Review of Criminology, 6*, 281-310. https://doi.org/10.1146/annurev-criminol-030421-041615
8. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324
9. Breiman, L. (2023). *Random Forest revisited* (Posthumous reprint). Journal of Computational Criminology.
10. Bureau of Justice Statistics (BJS). (2022). *National Jail Data Dashboard*. U.S. Department of Justice
11. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153-163. https://doi.org/10.1089/big.2016.0047
12. European Union. (2024). *Regulation (EU) 2024/... of the European Parliament and of the Council on artificial intelligence (AI Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/...
13. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010
14. Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://us.sagepub.com/en-us/nam/an-r-companion-to-applied-regression/book246125
15. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3315-3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
16. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
17. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239
18. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387
19. Kuhn, M., & Johnson, K. (2023). *Feature engineering and selection: A practical approach for predictive models* (2nd ed.). Chapman & Hall/CRC. https://www.routledge.com/Feature-Engineering-and-Selection-A-Practical-Approach-for-Predictive-Models/Kuhn-Johnson/p/book/9781138079229
20. Kuhn, M., & Johnson, K. (2023). *Feature engineering and selection: A practical approach for predictive models* (2nd ed.). Chapman & Hall/CRC
21. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
22. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18-22.
23. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.
24. R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
25. Richardson, R., Schultz, J. M., & Crawford, K. (2021). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review, 94*(1), 192-233. https://www.nyulawreview.org/issues/volume-94-number-1/
26. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 77. https://doi.org/10.1186/1471-2105-12-77
27. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x
28. Rudin, C., Wang, C., & Coker, B. (2022). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review, 4*(1). https://doi.org/10.1162/99608f92.6ed64b30

29. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
30. Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology, 54*(4), 680-712. https://doi.org/10.1111/1745-9125.12123
31. Stone, M., Lichtenstein, S., & Fischhoff, B. (2014). How to make better forecasts and decisions: Avoid face-to-face meetings. *Foresight: The International Journal of Applied Forecasting, 35*, 5-9.
32. Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.
33. van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9780429492259
34. Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer.
35. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686
36. Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman & Hall/CRC.
37. Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A, 180*(3), 689-722.

التنبؤ بنتائج إطلاق سراح الاعتقال: تحليل مقارن لنماذج التعلم الآلي

O. P. Adebayo [(1*)], I. Ahmed [(2)], K.T. Oyeleke [(3)]

[(1*)]Department of Statistics, Phoenix University, Agwada, Nasarawa State, Nigeria
[(2)]Department of Statistics, Nasarawa State University, Keffi, Nasarawa State, Nigeria
[(3)]Department of Statistics, Olabisi Onabanjo University, Ago Iwoye, Ogun State, Nigeria

الملخص

تُقيّم هذه الدراسة المقارنة نماذج التعلم الآلي للتنبؤ بنتائج إطلاق سراح الموقوفين، باستخدام 5226 قضية حيازة ماريجوانا من شرطة تورنتو (1997-2002). أظهرت مجموعة البيانات اختلالًا كبيرًا في التوازن الطبقي، حيث بلغت نسبة نتائج الاحتجاز 17.1% فقط مقابل 82.9% من حالات إطلاق السراح. بعد المعالجة المسبقة لمعالجة القيم المفقودة وتحويل المتغيرات الفئوية، طبّقنا منهجين للنمذجة: مُصنّف غابة عشوائية بـ 500 شجرة مع قياس أهمية السمات، ونموذج الانحدار اللوجستي الثنائي. أظهرت كلتا الخوارزميتين قدرة تنبؤية قوية لحالات إطلاق السراح، محققتين دقة إجمالية متقاربة (83.2-83.4%) وحساسية ممتازة (>98%)، على الرغم من أنهما واجهتا صعوبة في التعامل مع فئة الأقلية الحرجة، كما يتضح من ضعف الخصوصية (<7%). أظهرت النماذج قوة تمييزية مماثلة، حيث حقق الانحدار اللوجستي قيمة AUC-ROC أعلى قليلاً (0.733 مقابل 0.726). حدد تحليل أهمية السمات الحالة الوظيفية وفحوصات الخلفية السابقة للشرطة كأقوى العوامل التنبؤية، بينما ساهمت العوامل الديموغرافية، بما في ذلك العرق، بشكل كبير في التنبؤات. تُبرز هذه النتائج التحديات التقنية المتمثلة في اختلال تصنيف بيانات نظام العدالة، والاعتبارات الأخلاقية المحيطة بالتحيز الخوارزمي المحتمل، لا سيما بالنظر إلى ارتفاع معدل النتائج الإيجابية الخاطئة لتوقعات الاحتجاز، مما قد يُفاقم التفاوتات القائمة. تُؤكد الدراسة على ضرورة إجراء تقييم دقيق للنماذج وتنفيذها بمسؤولية عند تطبيق التحليلات التنبؤية على قرارات العدالة الجنائية الحساسة، مع موازنة الأداء الإحصائي مع اعتبارات العدالة والأثر الاجتماعي.