Pretrial Risk Tool for the Pennsylvania Commission on Sentencing

Fall 2025

Presented By:
Amy Kang, Selina Pang, Yash Sivramkrishnan, Karry Li, & Casey Mei
Carnegie Mellon University, Heinz College of Public Policy & Information Systems in Partnership with the Pennsylvania Commission on Sentencing

Faculty Advisor:
Daniel Nagin
Teresa and H. John Heinz III University Professor of Public Policy and Statistics

With Support From:
Donovan Lappe & Jaxon Zaorski
Law students at Duquesne University

Table of Contents

For Jupyter notebooks and additional documentation, please refer to this GitHub page.

Terminology

Pretrial recidivism: An individual commits pretrial recidivism if they commit a new crime after being released from custody or from a previous arrest but *before* their initial case has been resolved. This definition differs from standard recidivism, which typically refers to reoffending after a conviction and sentence. For our purposes, a charge is considered a crime for *non-summary* offenses.

Pretrial release: Refers to the conditions of release from custody to which defendants must adhere during the time period between the filing of charges by law enforcement and court adjudication. After arrest charges are filed, courts will decide whether a defendant can be released pending trial and other proceedings. Conditions of pretrial release can include: release on personal recognizance, payment of cash bail, securing surety or property bonds, requirement to submit to electronic monitoring, and pretrial supervision (Bureau of Justice Statistics).

L1 Penalty: The L1 penalty adds the absolute value of each coefficient to a model's loss function, which pushes some coefficients exactly to zero during optimization. This creates a built-in feature selection effect, helping the model ignore irrelevant predictors and improving interpretability, especially when many features are noisy or weakly informative.

L2 Penalty: The L2 penalty adds the squared magnitude of each coefficient to a model's loss function, shrinking coefficients toward zero without eliminating them entirely. This reduces variance, stabilizes estimates, and prevents overly large weights, making it particularly effective when predictors are correlated or when the dataset risks overfitting.

Unweighted Logistic Regression (L2 Penalty): Unweighted logistic regression estimates class probabilities without adjusting for class imbalance, letting majority-class patterns dominate the optimization process. The L2 penalty shrinks coefficients toward zero by penalizing their squared magnitude, which stabilizes estimates, reduces variance, and mitigates overfitting while preserving all predictors in the model.

Balanced Logistic Regression (L1 Penalty – Lasso): Balanced logistic regression incorporates class weights so that errors on the minority class are penalized more heavily, improving learning in imbalanced datasets. When paired with an L1 penalty, the model performs both classification and feature selection by pushing some coefficients exactly to zero, making the resulting model more interpretable and often more robust.

Balanced Logistic Regression (L2 Penalty – Ridge): This approach applies class weighting to counteract imbalanced outcomes while using an L2 penalty to control coefficient growth. The ridge penalty compresses coefficients without eliminating them, distributing influence across

correlated predictors and improving model stability and generalization in settings where minority-class information is sparse.

Elastic Net Regression: Elastic net regression combines L1 and L2 penalties to balance feature selection with coefficient shrinkage, making it effective when predictors are numerous or highly correlated. By adjusting the mixing parameter between penalties, the model can simultaneously retain groups of correlated variables while discarding irrelevant ones, producing a flexible compromise between lasso and ridge behavior.

Accuracy: Accuracy measures the proportion of all predictions that the model classifies correctly. Although intuitive, it becomes misleading under class imbalance because high accuracy can be achieved by simply predicting the majority class.

Precision: Precision quantifies the proportion of positive predictions that are actually correct. It is especially informative in contexts where the cost of false positives is high, as it focuses only on the correctness of positive classifications rather than overall model performance.

Recall: Recall measures the proportion of actual positive cases that the model successfully identifies. It is crucial when missing positive cases are costly because it captures the model's ability to minimize false negatives.

F1-Score: The F1-score represents the mean of precision and recall, creating a single metric that balances both concerns. It is especially useful when classes are imbalanced or when neither precision nor recall alone adequately reflects the model's performance priorities.

Receiver Operating Characteristic – Area Under the Curve (ROC-AUC): ROC-AUC summarizes how well the model distinguishes between classes across all possible classification thresholds. A higher AUC indicates stronger separability, reflecting the model's ability to achieve high true positive rates while maintaining low false positive rates.

False Positive Rate (FPR): The false positive rate captures the proportion of negative cases incorrectly classified as positive. It reflects the model's tendency to raise unwarranted alarms and is an essential counterbalance to metrics like recall, particularly in domains where false positives carry meaningful consequences.

False Negative Rate (FNR): The false negative rate measures the proportion of positive cases that the model fails to detect. High FNR indicates blind spots in model sensitivity and is problematic in settings where overlooking positive instances leads to significant risk or harm.

Executive Summary

This project built a statewide pretrial recidivism dataset by merging PSP, MDJ, and CPMC charge-level records into 431,920 pretrial-eligible cases, engineering individualized pretrial windows, and constructing baseline predictive models. We cleaned inconsistent timestamps, reconciled case transfers, standardized date fields, and created features for demographics, criminal history, and offense characteristics. Several models were tested, including unweighted L2 logistic regression, class-weighted L1 and L2, and Elastic Net. **The class-weighted L1 logistic regression emerged as the preferred model after threshold tuning (t = 0.45), balancing operational accuracy with interpretability (Accuracy ~0.57; Precision ~0.31; Recall ~0.67; AUC ~0.66).** Key predictors included prior failures to appear, prior arrests, age, and supervision indicators. Constraints included reliance on a single heavy-compute VM and unavoidable noise in administrative timestamp data. Recommended next steps include automating the data pipeline, expanding feature engineering, and adding Responsible AI evaluations such as group fairness metrics and SHAP-based explainability.

Introduction

1.1: Project Justification & Objective

The Pennsylvania Commission on Sentencing (PCS) is a state legislative agency that plays a central role in promoting fairness, consistency, and transparency across the Commonwealth's criminal justice system via development and oversight of sentencing, resentencing, and parole guidelines. As the General Assembly considers expanding PCS's authority to include bail and pretrial release guidelines through House Bill 1454, the need for reliable, evidence-based tools becomes more urgent. Enhancing pretrial recidivism risk assessment tools is a critical component of this work, as it helps better identify the likelihood that a defendant will commit new offenses or fail to appear for court while awaiting trial. When grounded in rigorous data analysis, these assessments can reduce subjectivity in judicial decision-making.

Typically following an arrest, defendants charged with criminal behavior are brought before a magisterial district judge for an initial appearance to determine pretrial conditions (see terminology on pretrial release). However, there is a lack of uniformity and potential for bias in Pennsylvania's current pretrial system. Practices for setting bail and pretrial conditions – despite having the qualitative data-based Pretrial Risk Score (PRS) assessment – vary significantly across the state's 67 counties, creating inconsistency in decision-making.

To address these gaps, **our project objective is to develop a pretrial risk assessment tool using machine learning techniques to predict pretrial recidivism for Pennsylvania magisterial judges, as well as identify next steps for the project to expand**. What this project will not cover includes: 1) no tool on predicting failure to appear in court; 2) no model deployment strategy or polished user interface; and 3) no policy recommendations for the PCS.

1.2: Literature Review Findings

Current empirical research reveals which factors most reliably predict pretrial outcomes, providing the evidence base needed for a more consistent and transparent statewide assessment framework. The literature overwhelmingly supports the finding that static criminal history factors and the characteristics of the current charge dominate predictive power in recidivism risk assessments. Meta-analyses of pretrial risk assessments confirm that variables such as prior convictions, prior failures to appear, and prior felonies are the strongest predictors of pretrial failure and new criminal activity (VanNostrand & Lowenkamp, 2013).

Conversely, dynamic indicators of social stability, such as residence and employment status, are found to be less predictive, adding limited or marginal signal to models dominated by criminal history (Gendreau et al., 1996; Desmarais et al., 2013). While age is a strong predictor – with clear drops in rearrest rates as age increases (United States Sentencing Commission, 2017) – and gender shows conditional effects (with recidivism risk between males and females becoming

increasingly similar as convictions accumulate), these demographic factors offer modest additional predictive value compared to the core criminal record. Furthermore, studies indicate that pretrial detention itself may have a criminogenic effect, increasing the likelihood of future criminal behavior among those detained (Heaton et al., 2017).

Considerations for model design favor explainability and fairness, particularly regarding sensitive variables. Although race has been shown to strongly predict arrest outcomes, most of this racial difference in risk scores is attributable to existing criminal history factors, meaning that race effects operate primarily through these embedded variables (Skeem & Lowenkamp, 2016). Therefore, race is not used as an independent predictor in the resulting model. The reliance on parsimonious models like logistic regression for tools such as the Virginia Pretrial Risk Assessment Instrument and the Colorado Pretrial Assessment Tool demonstrates a favoring of explainability over the complexity of alternative machine learning approaches, which is critical for implementation in judicial settings (Clark, 2017).

Methodology

2.1: Data Sources

Our analysis integrates four primary datasets provided by PCS, covering all criminal filings and state police arrest records from 2015–2018. Across all sources, records are reported at the charge level, meaning that each row represents a single charged offense, rather than a unique case or individual. Thus, cases with multiple charges appear in multiple rows, and individuals with multiple cases or arrests may appear repeatedly across files.

All four datasets share a consistent state identification number (SID), which serves as a universal person identifier. The same SID appearing in the four files refers to the same individual, enabling comprehensive linkage of pretrial cases with arrest histories. The datasets include:

1. "CMU AOPC MDJ Filings 2015–2018" (MDJ): Charge-level filings from the **Magisterial District Courts**, including offense information, filing dates, disposition dates, and case identifiers. These records capture the earliest stages of criminal case processing.

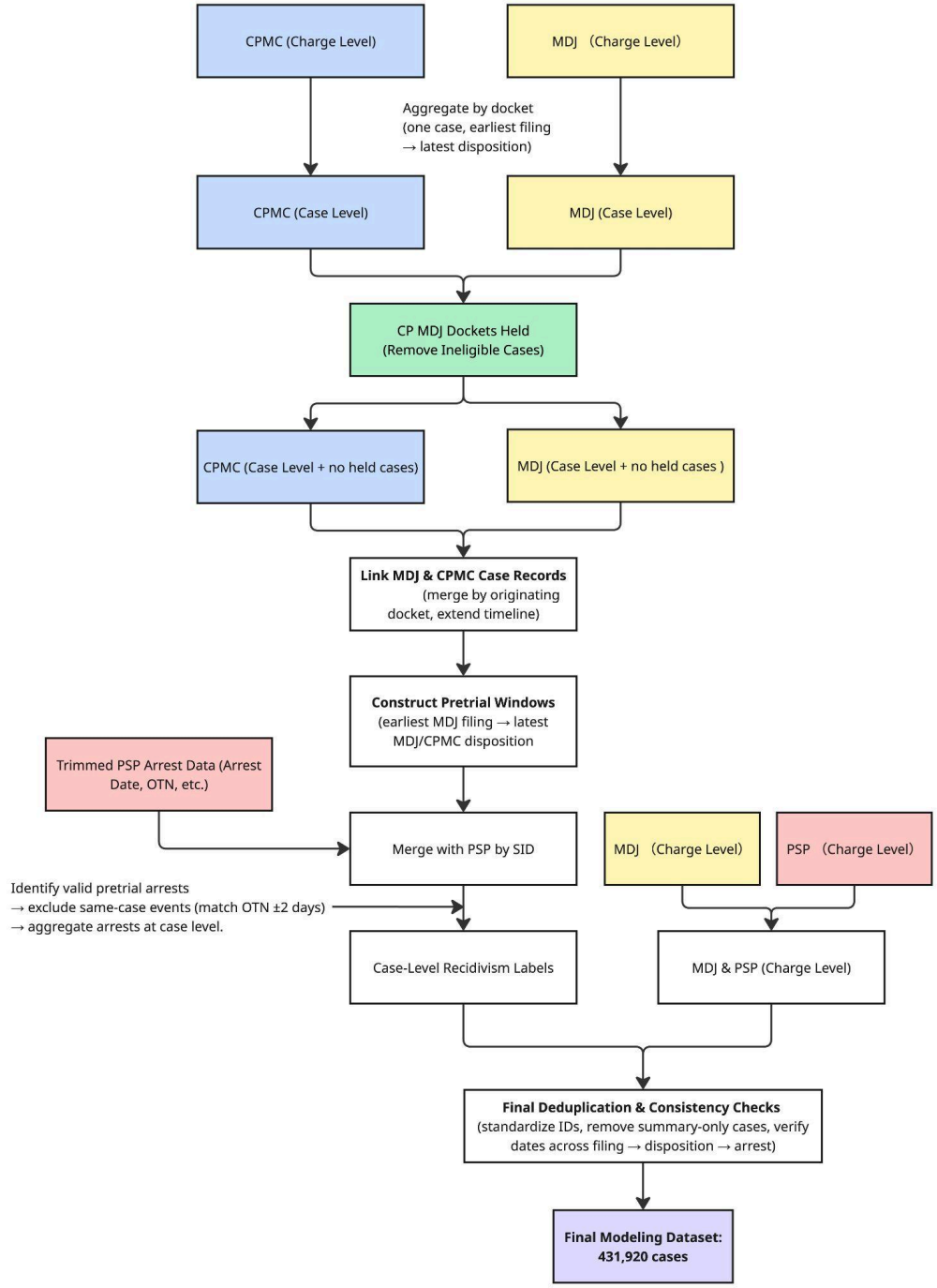
2. "CMU AOPC CP_MC Filings 2015-2018" (CPMC): Charge-level filings from the **Courts of Common Pleas** (CP) and Municipal Court (MC). This dataset contains:

- Cases transferred upward from MDJ, linked through originating docket numbers.
- Cases that originate directly in CP or MC without prior MDJ involvement. These records supply longer-term disposition dates and outcomes for cases that progress beyond the MDJ level.

3. "CMU PSP Data" (PSP): Arrest-level records from the **Pennsylvania State Police**. This data contain arrest dates and charge details for all individuals appearing in court filings and are used to identify pretrial recidivism.

4. "CP MDJ Dockets Held": A consolidated list identifying dockets in which defendants were held pretrial, either at the MDJ level or after transfer to CP/MC. Because detained individuals are not at risk of committing new pretrial offenses, these cases were excluded from modeling.

2.2: Data Cleaning & Integration



Because all datasets are delivered at the charge level, we first reconstructed complete case-level records and aligned them across MDJ, CPMC, and PSP data sources. We consolidated multiple rows per case, resolved duplicate information, and defined individualized pretrial “at-risk” windows for identifying recidivism.

2.2.1: Aggregating Charge-Level MDJ Filings into Case-Level Records

The MDJ dataset contains one row per charge, meaning a single docket can appear across multiple entries. To create case-level units of analysis, we aggregated all rows sharing the same docket number and completed the following:

- The earliest filing date was assigned as the case's start date.
- The latest disposition date across all charges was used as the MDJ case end date.
- Charge-level fields (e.g., statute, section, grade) were consolidated.
- Admin attributes (e.g., case status, county) were selected from the first non-null value.

This step converted hundreds of thousands of charge-level rows into a clean set of MDJ case-level records, each tied to a unique individual through SID.

2.2.2: Removing Held Cases

Cases in which defendants were held in custody at any point prior to final disposition are not eligible for pretrial recidivism because individuals who remain detained cannot commit new offenses in the community. Using the CP–MDJ held-dockets file, we removed all cases flagged as held either at the MDJ level or after transfer to CP/MC.

The held-dockets file includes both the MDJ 'docketnumber' and the CPMC 'originatingdocketnumber', and any case appearing in either field was excluded. This ensured that only defendants who were released pretrial and at risk of reoffending are in the dataset.

2.2.3: Linking MDJ & CPMC Filings

To capture complete case trajectories, we merged MDJ case-level records with CPMC filings. The CPMC dataset includes both cases transferred upward from MDJ and their extended court outcomes. For transferred cases, we linked MDJ dockets to CPMC records using originating docket numbers and retained the latest disposition date from either system – producing a unified case timeline reflecting the full duration of each case, including transfers beyond the MDJ stage.

2.2.4: Constructing Pretrial Windows

After integrating disposition information from MDJ and CPMC, we defined an individualized pretrial window for each case:

- pretrial_start = earliest MDJ filing date
- pretrial_end = latest available disposition date from MDJ or CPMC

This interval represents the period when a defendant is released pretrial and at risk of reoffending. These case-specific windows form the foundation for labeling pretrial recidivism.

2.2.5: Merging PSP Arrest Data & Identifying Pretrial Arrest Events
After constructing case-level records and defining individualized pretrial windows, we merged PSP arrest data onto the MDJ–CPMC case table using the shared SID identifier. We first worked with a trimmed PSP dataset containing only arrest dates, offense tracking numbers (OTNs), and offense severity indicators, filtered to arrests occurring within the overall pretrial date range.

The merged long-format table linked each case to all arrest events associated with the same individual. To isolate valid pretrial arrest events, we removed arrests that were part of the same case, identified through matching OTNs or arrest dates occurring within ±2 days of the filing date. Remaining arrests were flagged as pretrial recidivism if they occurred strictly between the case's pretrial_start and pretrial_end dates. These arrest-level records were then aggregated back to the case level to generate final recidivism labels.

The resulting case-level labels were later merged with the full MDJ–CPMC dataset to restore all demographic fields, charge information, and engineered features needed for modeling.

2.2.6: Building the MDJ–PSP Charge-Level Dataset
To consolidate MDJ and PSP records at the charge level, we performed an inner join on both the SID and OTN fields, which were consistently present across the two datasets. This merge integrated offender-level demographic attributes from MDJ with arrest-specific information from PSP, producing a unified charge-level table. Duplicate columns generated by the join were removed, and data types were standardized to ensure consistency across variables. Case timelines were updated for instances in which arrests in PSP corresponded to cases initially recorded in MDJ, ensuring each charge-level record reflected a complete and temporally accurate representation of the associated case trajectory.

Following this integration, case-level recidivism labels generated from pretrial arrest events were merged into the full MDJ–PSP charge-level dataset using SID and docket number. This step attached the pretrial recidivism indicator to each case while preserving all demographic attributes, charge information, PSP-derived offense flags, and engineered features needed for downstream modeling. The resulting dataset provided a comprehensive foundation for both charge-level and case-level analyses, ensuring that recidivism outcomes were linked to the richest set of features available.

2.2.7: Final Deduplication, Consistency Checks, & Dataset Assembly
Before modeling, we standardized identifiers, removed cases involving summary-only charges, and verified chronological consistency among filing, disposition, and arrest dates. After the full integration and cleaning pipeline, the final dataset contained 431,920 pretrial-eligible cases, each with complete case-level information and PSP-derived recidivism labels.

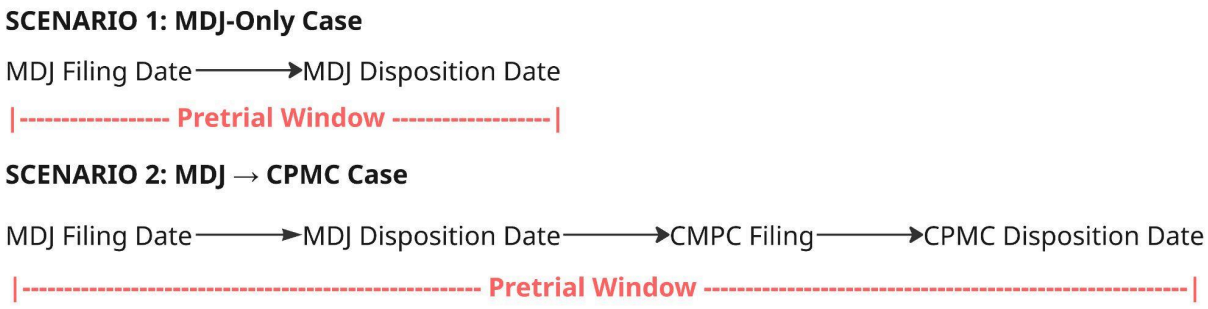
2.3: Recidivism Definition & Construction

Pretrial recidivism is defined as any new arrest that occurs after a case is filed in MDJ court and before the case reaches its final disposition in MDJ or CP/MC. Because our analysis focuses exclusively on the pretrial period, the labeling process relies on correctly defining the initial case event, constructing individualized at-risk windows, and identifying subsequent arrests that represent new criminal activity distinct from the original case.

2.3.1: Pretrial Recidivism Window

The pretrial recidivism window is defined as the period between:

- pretrial_start: the earliest MDJ filing date for the case, and
- pretrial_end: the final disposition date, taken from MDJ or CP/MC depending on the case's trajectory.



This window operationalizes the time during which a defendant is not detained and able to commit new offenses. Cases where defendants were held pretrial were excluded earlier in the pipeline and do not receive recidivism labels.

2.3.2: Identifying Subsequent Convictions (New Arrest Events)

To identify pretrial recidivism, we merged PSP arrest records onto the case-level dataset using SID identifiers and evaluated all arrest events occurring during each case's pretrial window.

A PSP arrest was counted as a new pretrial offense only if:

- The arrest occurred strictly between pretrial_start and pretrial_end;
- It did not correspond to the same case (to avoid double-counting original charges);
- It had a valid arrest timestamp and identifiable charge information.

To exclude arrests associated with the same case, we removed PSP entries that:

- matched the case's OTN, or
- occurred within ±2 days of the MDJ filing date (capturing same-incident arrests processed at different times).

Any remaining arrest event within the window was treated as a recidivism event, and case-level labels were generated by aggregating arrest activity across all PSP rows linked to the same case.

2.3.3 Crime-Specific Recidivism Logic

While the primary outcome is a binary indicator of any pretrial recidivism, we also generated crime-type-specific labels based on the offense severity recorded in PSP, including felony recidivism, misdemeanor recidivism, and an "other-category" recidivism.

These subtype labels were created by evaluating the offense grade associated with each qualifying PSP arrest during the pretrial window. They provide additional flexibility for exploratory analysis and subgroup modeling but were not used as primary modeling targets.

Analysis & Results

3.1: Exploratory Data Analysis

The dataset includes 431,920 pretrial-eligible records representing 381,376 unique cases from 2015 to 2018. Summary-only cases were excluded to focus on substantive misdemeanor and felony charges. Defendants come from multiple Pennsylvania counties, with case volume concentrated in urban and suburban jurisdictions.

3.1.1: Data Quality Insights

The dataset underwent extensive cleaning and preprocessing to ensure reliability for modeling. Columns with excessive missingness, duplicate identifiers, or minimal predictive value were removed. Numeric variables were standardized, categorical features were one-hot encoded, and missing values were imputed where appropriate. Additional features, such as county-level urban/rural classification, were engineered to support analysis and fairness assessments. After these transformations, the MDJ–PSP dataset was reduced from 2,189,279 rows and 44 columns to a final structure of 431,920 rows and 37 columns, preserving all information relevant to pretrial recidivism modeling.

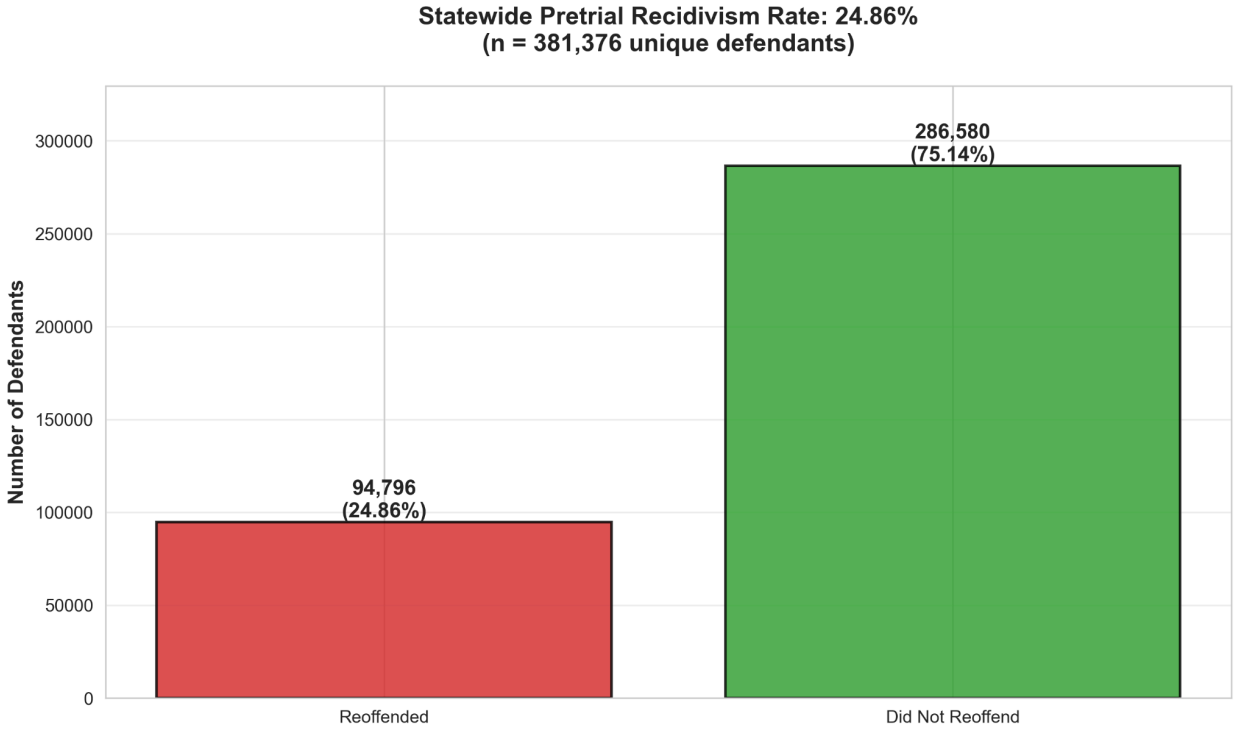
3.1.2: Dataset Statistics

Offender ages at the time of offense ranged, with a mean of 33.23 years (median: 30.89; standard deviation: 10.95). The 25–34 age group comprised the largest cohort at 36.89% (159,320 cases), followed by those under 25 (26.35%), ages 35–44 (20.71%), and 45 and older (16.05%).

Case volume was concentrated in all Pennsylvania counties: Allegheny County (19.03%), Montgomery (5.13%), York (4.88%), Delaware (4.69%), and Dauphin (4.68%), and the rest of the counties, with the top ten counties collectively representing 62.21% of cases.

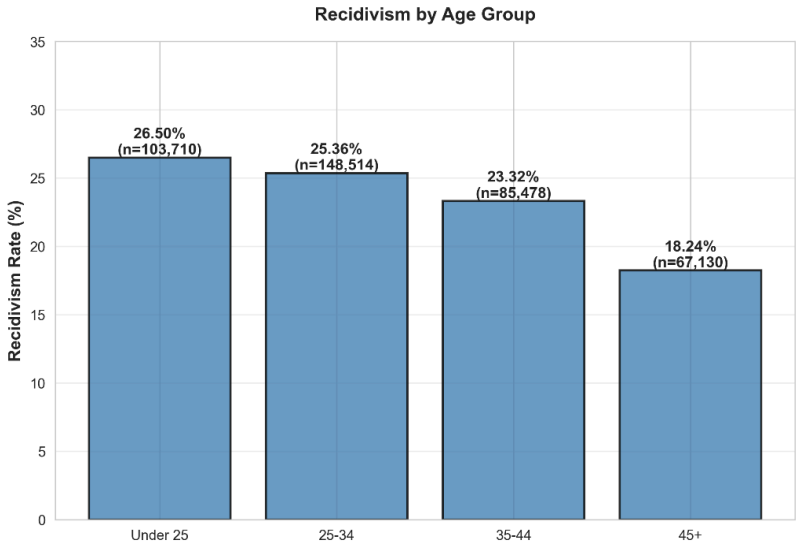
Drug-related offenses made up 34.18% of charges, followed by property crimes (23.37%) and violent offenses (19.09%). Charge severity averaged 2.80 on a 0–7 scale, with maximum severity per case averaging 3.51. A substantial portion of defendants had prior criminal involvement: 36.53% had any prior offense, and 21.72% had documented prior pretrial recidivism.

3.1.3: Overall Recidivism Rates



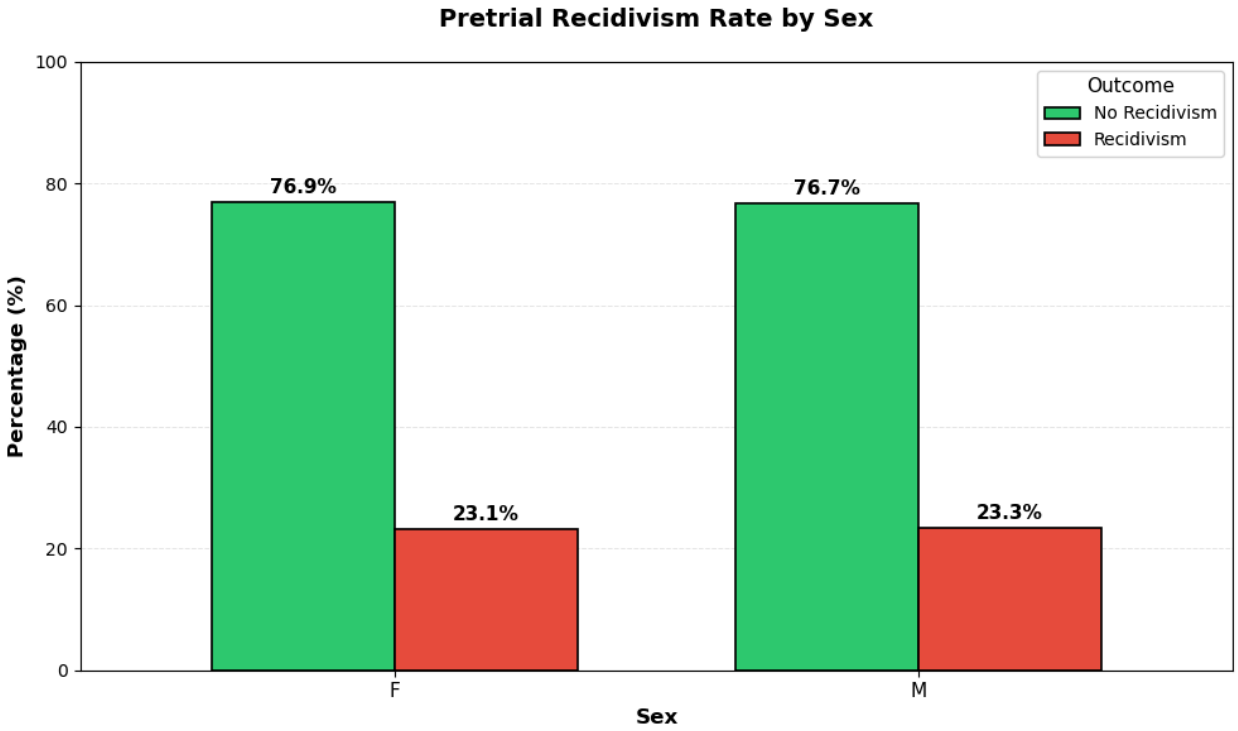
The statewide pretrial recidivism rate was 24.86%, with 94,796 of 381,376 unique defendants identified as having reoffended during the observation period. This indicates that approximately one in four defendants released pretrial went on to commit a subsequent offense, while 75.14% (286,580 defendants) remained offense-free during the monitoring period.

3.1.4: Recidivism by Demographics



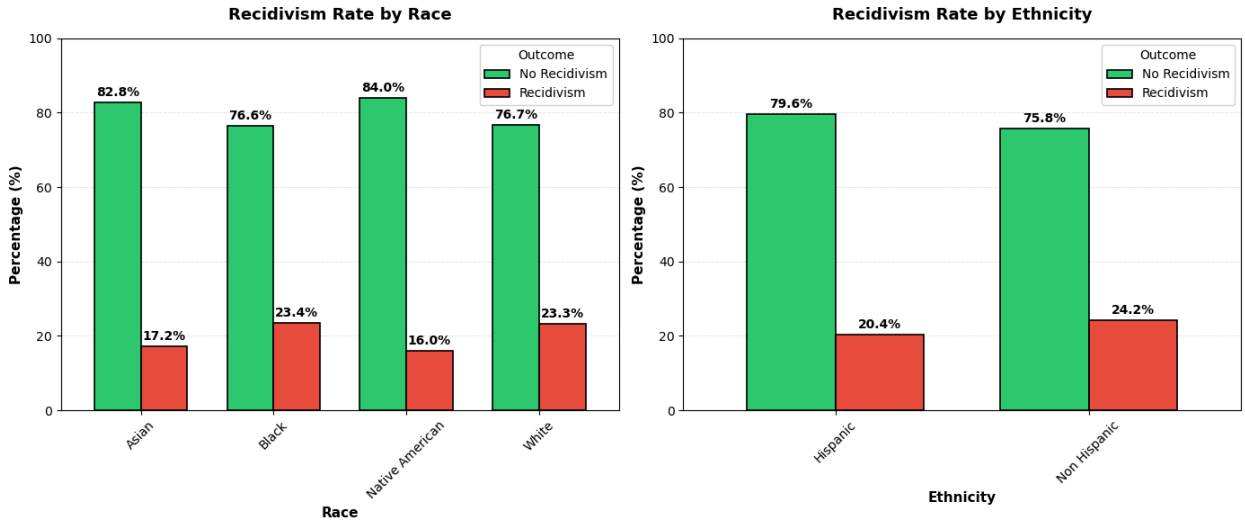
Age emerged as a notable demographic predictor of recidivism, though with a modest correlation. The correlation between age at offense and pretrial recidivism was -0.0604, indicating that younger defendants had slightly elevated recidivism risk compared to older defendants. Mean age among reoffenders was 32.03 years compared to 33.60 years among non-reoffenders, a difference of approximately 1.6 years.

Recidivism rates by age group revealed important stratification. Defendants aged 25-34 exhibited the highest recidivism rate at 25.36% (37,660 of 148,514 defendants), while those aged 35-44 showed a lower rate of 23.32% (19,937 of 85,478 defendants). The pattern of declining recidivism with age aligns with criminological research suggesting that criminal propensity and behavioral risk decrease across the lifespan.



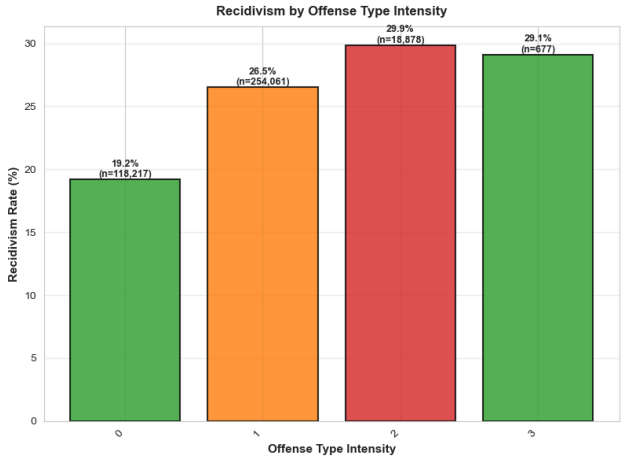
Pretrial recidivism rates show minimal variation across sex categories, with males and females exhibiting nearly identical recidivism rates of 23.33% and 23.11%, respectively. Overall, the observed differences between males and females suggest that sex is not a strong predictor of pretrial recidivism, indicating that other factors beyond demographic characteristics are more influential in determining recidivism outcomes.

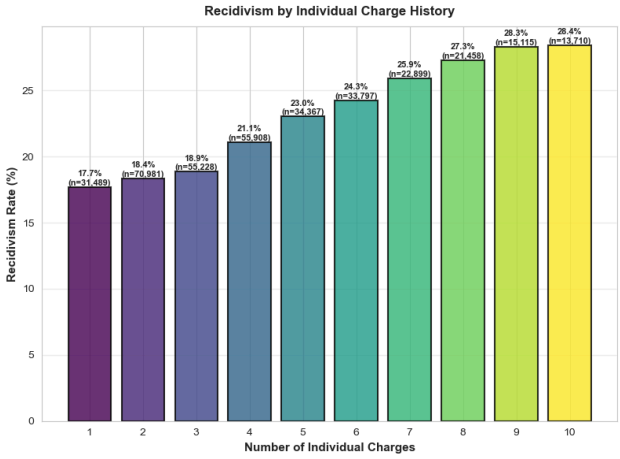
3.1.5: Race & Ethnicity

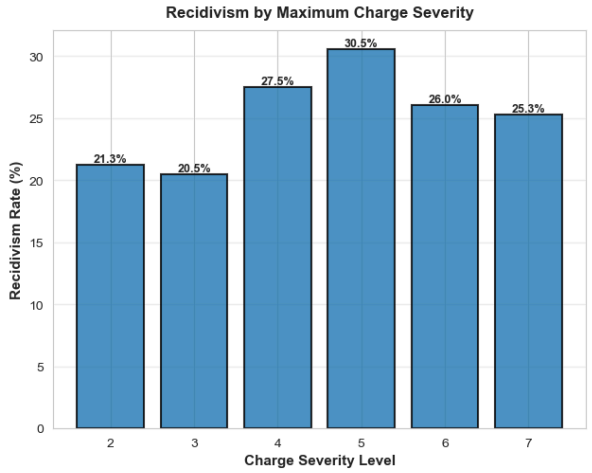


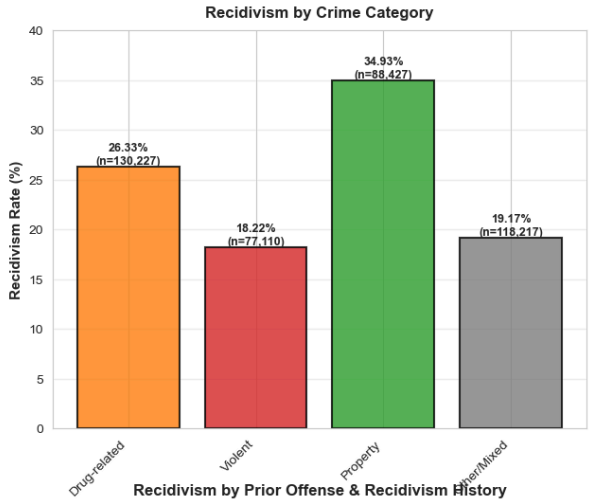
Our analysis reveals significant disparities in pretrial recidivism rates across both racial and ethnic groups. The overall pretrial recidivism rate is 23.28%, but this masks considerable variation. Race shows a 7.39 percentage point disparity, with Black defendants (23.44%) and White defendants (23.28%) exhibiting the highest recidivism rates, while Native American/Indian defendants (16.05%) show the lowest. Ethnicity demonstrates a smaller but notable 3.83 percentage point gap, with Non-Hispanic defendants recidivating at 24.23% compared to Hispanic defendants at 20.40%. Chi-square tests confirm these associations are statistically significant ($p < 0.0001$ for both race and ethnicity), indicating that demographic characteristics are meaningfully associated with pretrial recidivism outcomes. To mitigate potential algorithmic bias and ensure fairness in our predictive model, we excluded race and ethnicity as features in our final model, despite their statistical significance. This decision prioritizes equitable treatment across demographic groups and prevents the perpetuation of systemic disparities in criminal justice risk assessments.

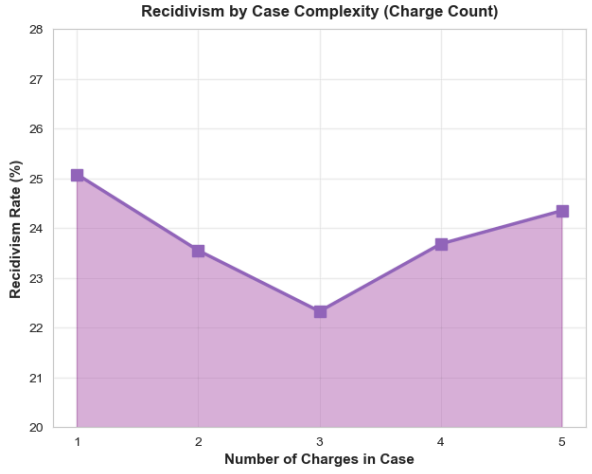
3.1.6: Recidivism by Offense Characteristics

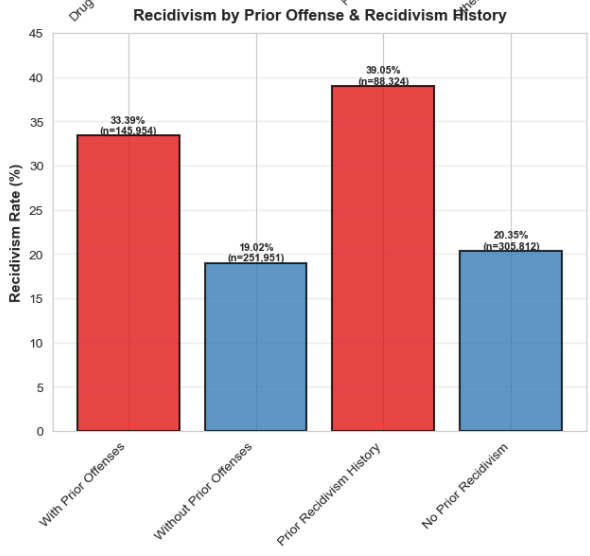












The analysis of recidivism patterns across various offense characteristics reveals that pretrial recidivism is not uniformly distributed across defendant populations. The overall pretrial recidivism rate is 24.86%, but this figure masks significant variation based on offense type, offense history, criminal background, and case complexity. Understanding these variations is critical for risk assessment, detention decisions, and case management strategies.

Offense Type Intensity

The relationship between offense type intensity and recidivism rates demonstrates a clear escalating pattern. Defendants with the lowest intensity offenses (Intensity 0) exhibit a recidivism rate of 19.2% across 118,217 cases. This baseline rate increases to 26.5% for Intensity 1 offenses (254,061 defendants), a substantial jump of 7.3 percentage points. The escalation continues with Intensity 2 offenses showing a recidivism rate of 29.9% (18,878 defendants), and Intensity 3 offenses reaching 29.1% (677 defendants).

This upward trajectory reflects the behavioral complexity captured by offense type intensity. Rather than measuring severity alone, this metric counts how many offense categories (drug, violent, property) apply to each case. Defendants engaging across multiple offense categories demonstrate broader behavioral patterns and greater versatility in their criminal activity—both factors strongly linked to increased recidivism risk. Those with Intensity 0 offenses fall within a single category, while those at Intensity 3 have demonstrated involvement across all three offense types, suggesting more entrenched and multifaceted criminal engagement. However, it is important to note that this measure is limited to three offense categories and does not account for other offense types that may exist in the broader criminal spectrum, potentially understating the true behavioral versatility of some defendants.

Notably, the recidivism rate stabilizes between Intensity 2 and 3, suggesting that beyond a certain threshold, additional categorical involvement may not substantially increase recidivism risk. This plateau could indicate that once a defendant has demonstrated involvement across multiple offense types, the presence of a third category adds limited additional predictive value. The large sample sizes for Intensity 0 and 1 offenses provide robust evidence for their risk profiles, while the smaller sample for Intensity 3 (n=677) suggests these defendants represent a specialized subset with the most versatile criminal patterns across the measured categories.

Individual Charge History

Individual criminal charge history shows a pronounced and consistent positive correlation with pretrial recidivism risk. Defendants with only one prior individual charge in their history exhibit a recidivism rate of 17.7%, while those with ten prior charges face a recidivism rate of 28.4%. This represents a 10.7 percentage point increase and a 60.6% relative increase in risk, making individual charge history a powerful predictor of recidivism.

The relationship is nearly linear across the range examined, with each additional charge in a defendant's history associated with incrementally higher recidivism risk. Defendants with 1-3 individual charges maintain recidivism rates below 19%, while those with 4-5 charges experience rates of 21-23%. By the time defendants reach 6-7 charges, rates climb to 24-26%, and those with 8-10 charges exceed 27%. This graduated pattern suggests that cumulative criminal experience significantly elevates recidivism propensity. The large sample sizes across all charge levels (ranging from 13,710 to 70,981 defendants) ensure statistical reliability of these estimates. This variable appears particularly valuable for risk stratification, as it provides clear differentiation across the defendant population.

Maximum Charge Severity

Charge severity shows a strong nonlinear relationship with pretrial recidivism. Severity levels 0, 2, and 3 demonstrate relatively moderate recidivism rates of approximately 20-21%. However, a substantial increase occurs at Severity Level 4, where the recidivism rate jumps to 27.5% (55,087 defendants). This trend continues with Severity Level 5 showing the highest rate at 30.5% (67,372 defendants), representing a 9.2 percentage point increase from Severity Level 3.

The peak at Severity 5 is notable given that this level contains the largest sample (67,372 defendants), ensuring robust estimates. Severity Levels 6 and 7 show somewhat lower rates of 26.0% and 25.3% respectively, suggesting a plateau or slight regression at the highest severity tiers. This may reflect that the most serious charges carry substantial pretrial restrictions that reduce opportunity for recidivism, or that defendants facing the most severe charges may be detained pretrial. The dramatic escalation from Severity 3 to 4 (a 7-point jump) indicates that this threshold represents a critical boundary in terms of offense seriousness and associated recidivism risk. Overall, charge severity is a meaningful predictor, with the relationship particularly pronounced at intermediate-to-high severity levels.

Crime Category

Crime category reveals substantial disparities in recidivism risk across offense types. Property crimes show the highest recidivism rate at 34.9% (88,427 defendants), significantly exceeding the overall average of 24.9%. This elevated rate suggests that individuals charged with property offenses are substantially more likely to commit additional crimes while awaiting trial, perhaps reflecting their economic motivations or persistent offending patterns.

Drug-related offenses demonstrate a recidivism rate of 26.3% (130,227 defendants), which is close to the overall average but slightly elevated. This substantial defendant population (the second largest category) represents a consistent moderate-risk group. Violent crimes show a notably lower recidivism rate at 18.2% (77,110 defendants), the lowest among categorized offenses. This may reflect that violent crime defendants face more serious consequences, higher bail amounts, or more intensive supervision. The "Other/Mixed" category, containing 118,217

defendants not classified as drug, violent, or property crimes, shows a recidivism rate of 19.2%, similar to the violent crime rate.

The 16.7 percentage point gap between property crimes (34.9%) and violent crimes (18.2%) is striking and suggests that offense type is a crucial factor in understanding recidivism risk. Property crime offenders appear to represent the highest-risk category, which has important implications for risk assessment frameworks and pretrial release decisions.

Case Complexity (Charges per Case)

Case complexity, measured by the number of charges in a single case, reveals an interesting and counterintuitive pattern compared to individual charge history. Defendants charged with a single offense in their current case have the highest recidivism rate at 25.1% (56,397 defendants). This rate decreases for cases with two charges (23.5%, 108,673 defendants) and three charges (22.3%, 76,004 defendants), representing a decline of approximately 2.8 percentage points. However, the trend reverses slightly with cases of four and five charges showing rates of 23.7% and 24.3% respectively.

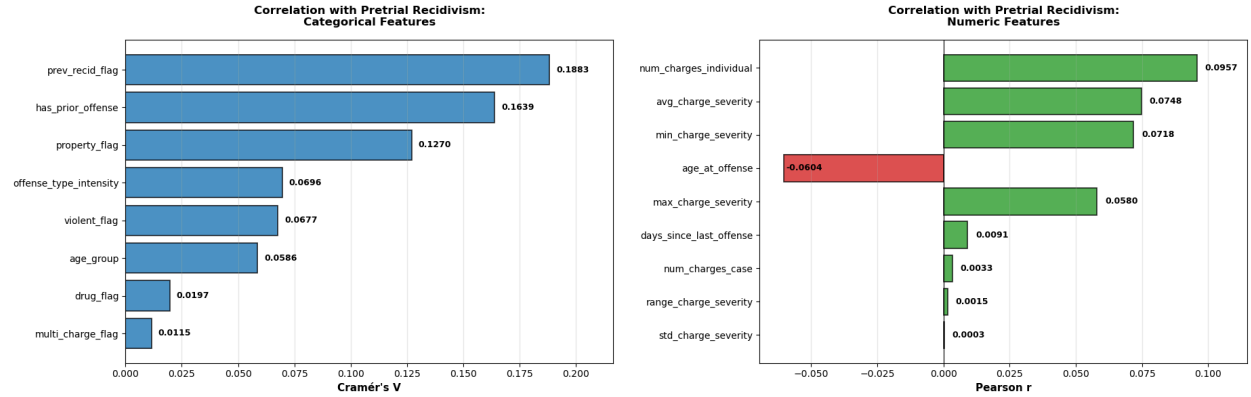
This pattern suggests that while individuals with extensive personal charge histories pose elevated risks, those facing multiple charges in a single incident may actually present lower pretrial recidivism risk, at least for cases with 2-3 charges. This could reflect that defendants involved in complex, multi-charge incidents may face more serious dispositions, higher bail amounts, or closer supervision. Alternatively, it may indicate that single-incident cases are more likely to involve single-crime specialists (such as standalone property crimes), whereas multi-charge cases may involve individuals with different risk profiles. The relatively flat variation (ranging from 22.3% to 25.1%) across case complexity levels suggests this factor alone is less predictive than individual criminal history, and that other contextual factors likely moderate the relationship between case complexity and recidivism.

Criminal History

Criminal history emerges as perhaps the most powerful predictor of pretrial recidivism among all factors examined. Defendants with prior offenses demonstrate a recidivism rate of 33.4% (145,954 defendants), compared to 19.0% for those without prior offenses (251,951 defendants). This represents a 14.4 percentage point absolute difference, or a 75.5% relative increase in risk. The relationship is even more pronounced when examining prior recidivism specifically: defendants with prior recidivism history show a rate of 39.1% (88,324 defendants), versus 20.4% for those without such history (305,812 defendants). This 18.7 percentage point gap and 92% relative increase demonstrates that past recidivism is an exceptionally strong predictor of future recidivism.

The clear stratification of defendants based on criminal history suggests that prior behavior is among the most reliable indicators of pretrial risk. The consistency of these large effect sizes across both prior offense and prior recidivism measures indicates that defendants with established patterns of criminal activity—particularly those with documented recidivism—warrant careful consideration in pretrial decision-making. The large sample sizes in all comparison groups (ranging from 88,324 to 305,812 defendants) ensure these estimates are highly reliable and not subject to sampling variation.

Comparative Summary and Risk Stratification



Across all dimensions examined, several clear patterns emerge:

Strongest Predictors: Prior recidivism flag (Cramér's V = 0.1883) and prior offense history (Cramér's V = 0.1639) are the strongest predictors of pretrial recidivism. These categorical features substantially outperform all numeric features and demonstrate that historical criminal behavior is the most reliable indicator of future recidivism risk. The magnitude of these correlations suggests they should be primary considerations in any risk assessment framework.

Moderate-Strong Predictors: Property crime involvement (Cramér's V = 0.1270) and offense type intensity (Cramér's V = 0.0696) represent secondary categorical predictors that meaningfully differentiate recidivism risk. Additionally, among numeric features, the number of individual charges (Pearson r = 0.0957) and average charge severity (Pearson r = 0.0748) show moderate predictive value, indicating that offense multiplicity and severity matter in risk assessment.

Moderate Predictors: Violent crime flag (Cramér's V = 0.0677) and age group (Cramér's V = 0.0586) show modest associations with recidivism. Among numeric features, minimum charge severity (Pearson r = 0.0718) and maximum charge severity (Pearson r = 0.0580) display similar moderate correlations, though their limited variation suggests they may be less predictive in isolation.

Notable Finding - Age Effect: Age at offense shows a negative correlation (Pearson r = -0.0604), indicating that older individuals at the time of offense have slightly lower pretrial recidivism risk, the only inverse relationship observed in the analysis.

Weak Predictors: Drug involvement flag (Cramér's V = 0.0197), multi-charge flag (Cramér's V = 0.0115), days since last offense (Pearson r = 0.0091), and various charge severity metrics (range and standard deviation) show negligible correlations with pretrial recidivism and should be de-emphasized in risk assessment models.

Critical Insight - Feature Type Disparity: Categorical features, particularly those capturing criminal history, demonstrate substantially stronger predictive power than numeric features. The strongest categorical predictor (0.1883) is nearly twice as strong as the strongest numeric predictor (0.0957), suggesting that binary indicators of criminal history and offense type provide more discriminative information than quantitative charge metrics.

Implications for Risk Assessment: A comprehensive risk framework should prioritize categorical features—especially prior recidivism and criminal history—as the foundation of pretrial risk assessment. While numeric features offer supplementary information, their modest correlations indicate they are less critical for differentiating risk levels. The multifactorial nature of recidivism risk requires integration of both historical indicators and offense characteristics, but the analysis clearly demonstrates that categorical factors related to prior behavior should receive primary weight in predictive models.

3.2: Feature Engineering

With the final dataset, literature review findings, and our exploratory data analysis, we created features related to demographics, criminal history, sentencing characteristics, and offense-level variables. See Appendix A for each variable definition and Appendix B for correlation metrics with our target pretrial recidivism variable.

Numeric features, including age, days elapsed, and charge severity scores, were retained in their original form to preserve the underlying information. Categorical features such as age_group and all binary flags (offense type indicators, supervision violations, and favorable dispositions) were encoded as numeric values. Missing values in days_since_last_offense were handled using forward fill logic, treating the first offense for each individual as NaN where applicable.

Text-based keyword matching for offense types and disposition outcomes was performed using case-insensitive searching to ensure robust categorization across variations in case documentation. County-level aggregations were computed on the training set to maintain consistency in the recidivism rate assignments. Standardization and scaling were applied to numeric features where appropriate for model compatibility.

3.3: Predictive Modeling Results

This section outlines the performance of our predictive models to estimate the likelihood that a defendant will recidivate during the pretrial period. Initial baseline models were constructed before applying more sophisticated approaches, such as regularized logistic regression. Performance was evaluated using accuracy, precision, recall, F1-score, and ROC–AUC, with particular emphasis on recall and false negative rate (FNR) due to their policy relevance in identifying high-risk individuals.

3.3.1: Baseline Models

To establish a performance benchmark for predicting pretrial recidivism, several baseline logistic regression models were trained and evaluated. These models provide essential reference points for understanding how different modeling choices like class weighting, regularization penalties, and scaling influence predictive ability in an imbalanced classification setting.

Dummy Model

To establish a clear baseline for comparison, we first implemented a dummy classifier that always predicts the most common outcome in the dataset. In our data, the majority class is Non-Recidivism, so the dummy model predicts every individual as a non-recidivist regardless of their features. As a result, the model achieves a relatively high accuracy, but this performance is misleading:

Accuracy	0.77
Precision	0.00
Recall	0.00
F1 Score	0.00
AUC	0.50
FPR	0.00
FNR	1.00

Because the model never predicts recidivism, it fails to identify any true recidivists, leading to a false negative rate of 100%. At the same time, the false positive rate is zero since no non-recidivists are incorrectly flagged.This dummy model highlights an important limitation of relying on accuracy alone in highly imbalanced datasets. While accuracy appears strong, the model provides no practical or policy-relevant value for pretrial risk assessment. It serves solely

as a benchmark to ensure that more advanced models meaningfully improve upon trivial majority-class predictions.

Unweighted Logistic Regression (L2 Penalty)

We applied a standard L2‑regularized logistic regression with no class weighting. As expected given the highly imbalanced dataset, the model predicted almost exclusively the majority class ("no recidivism"), missing all true positive cases.

Accuracy	0.767
Precision	0.0000
Recall	0.0000
F1 Score	0.0000
AUC	0.5609
FPR	0.0000
FNR	1.0000

Although accuracy appears high, the model fails to identify any individuals who reoffend pretrial. This confirms that unweighted models are unsuitable in the presence of strong class imbalance and motivates the need for class-weight adjustments.

Balanced Logistic Regression (L1 Penalty – Lasso)

Introducing class weighting and L1 regularization substantially improved the model's ability to detect recidivism. L1 regularization (Lasso) works by shrinking some coefficients all the way to zero, effectively performing feature selection. This can help the model focus on the most informative predictors, especially in high-dimensional or noisy settings. Although overall accuracy decreased, the model achieved meaningful gains in recall and F1, indicating better identification of true reoffenders.

Accuracy	0.6425
Precision	0.3366
Recall	0.5500
F1 Score	0.4176
AUC	0.6566

FPR	0.3293
FNR	0.4500

This model correctly identifies over half of all recidivists, which is an enormous improvement over the unweighted baseline. The trade-off, however, is a higher false positive rate, which highlights policy tensions between public safety and fairness.

Balanced Logistic Regression (L2 Penalty – Ridge)

Applying L2 regularization (Ridge) while retaining class weighting yielded performance very similar to the balanced L1 model. Unlike L1, L2 shrinks coefficients smoothly toward zero without eliminating them, which helps stabilize the model and prevent overfitting. The comparable performance suggests that both penalties help mitigate class imbalance and model complexity, though neither dramatically outperformed the other in this context.

Accuracy	0.6435
Precision	0.3372
Recall	0.5488
F1 Score	0.4178
AUC	0.6566
FPR	0.3277
FNR	0.4512

The similarity in performance indicates that both L1 and L2 penalties are viable at the baseline stage, though L1 remains preferable due to its interpretability and sparsity.

Elastic Net Logistic Regression

Because Elastic Net logistic regression is computationally expensive on the full dataset, training was conducted on a randomized 100,000-row subsample. This allowed efficient hyperparameter tuning across L1/L2 penalty mixtures and regularization strengths.

Accuracy	0.518
Precision	0.2681
Recall	0.6178
F1 Score	0.3740
AUC	0.5636
FPR	0.5124
FNR	0.3822

This model achieved the highest recall among baseline models but at the cost of extremely high false positive rates. While valuable for exploratory purposes, the Elastic Net model's aggressiveness makes it impractical for operational use in a pretrial setting.

The baseline outcomes indicate that unweighted models do not work well, class balanced logistic regression gives a good and easily understandable initial performance, and although Elastic Net is sensitive, it is not properly calibrated for the policy and fairness constraints related to pretrial decision making. The results from these experiments were instrumental in making the decision to use the L1 regularized logistic regression model as the basis for the final prediction system and to subsequently refine and tune it.

3.4: Model Selection and Fine-Tuning

After establishing baselines, we evaluated more advanced model configurations and conducted threshold optimization to identify operating points that balance false positives (FPR) and false negatives (FNR). This section describes how we selected **Ridge Logistic Regression(L2 penalty)** as the final model and fine-tuned it for deployment.

3.4.1 Model Selection

Although L1, L2, and Elastic Net regularization strategies were tested, Ridge Logistic Regression (L2 penalty) emerged as the preferred final model for several reasons:

1. **Stability and Reduced Overfitting:** L2 regularization discourages large coefficients, spreading influence across correlated predictors and reducing variance, which is critical in a dataset with many encoded categorical variables.
2. **Interpretability:** Ridge yields smooth, nonzero coefficients across features, supporting transparent judicial decision support and policy analysis.

3. **Fairness and Calibration:** The model maintains consistent performance across validation partitions and supports threshold tuning to balance FPR/FNR.
4. **Operational Efficiency:** Ridge logistic regression is lightweight, scalable, and easily monitored over time compared to more complex models such as Random Forest and XGBoost.

Thus, Ridge Logistic Regression offers the best combination of performance, transparency, and deployability.

3.4.2 Model Fine-Tuning: Hyperparameters and Training Configuration

To optimize model performance, we tuned several core components of the logistic framework:

- **Regularization strength (C):** controlling the complexity of the model and preventing overfitting
- **Penalty specification:** focusing on L2 for stability and fairness considerations
- **Class weights:** using the "balanced" option ensured proportional attention to the minority class
- **Classification thresholds:** rather than relying on the default 0.50 threshold, we evaluated a wide range of thresholds (0.1–1.0) to study the trade-offs between false positives and false negatives

The final Ridge Logistic Regression model was trained using:

- `max_iter = 1000` — ensures convergence
- `class_weight = "balanced"` — offsets class imbalance
- `solver = "lbfgs"` — stable and efficient for large datasets
- `penalty = "l2"` — Ridge regularization

These settings produced a model that was both interpretable and resilient, while offering the flexibility to tune the threshold to fit policy priorities.

3.4.3 Threshold Analysis

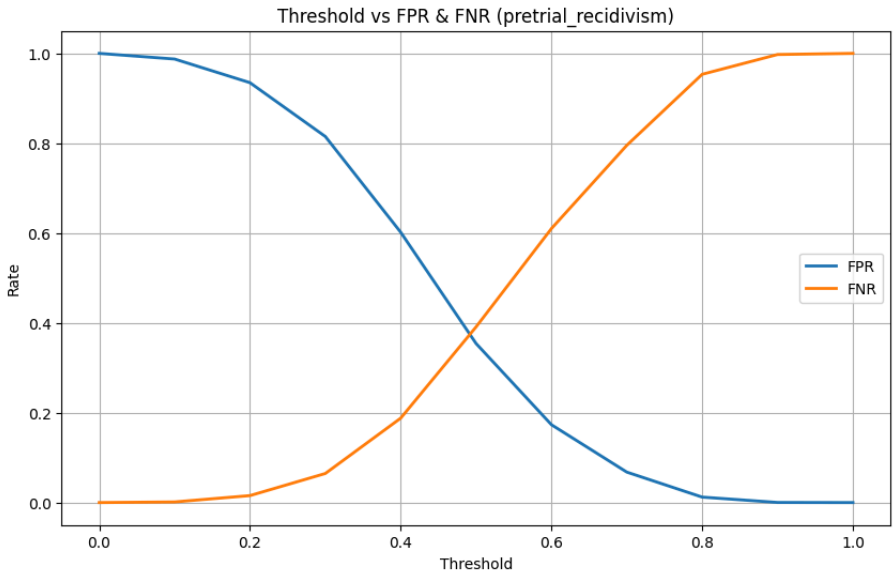
Threshold tuning proved to be central to understanding model performance. Across all three prediction tasks, pretrial, felony, and misdemeanor recidivism, the same general pattern emerged. **Low thresholds** (0.1–0.2) minimized FNR, meaning the model rarely missed true recidivists, but produced unacceptably high FPR. **High thresholds** (0.8–1.0) minimized FPR but caused FNR to

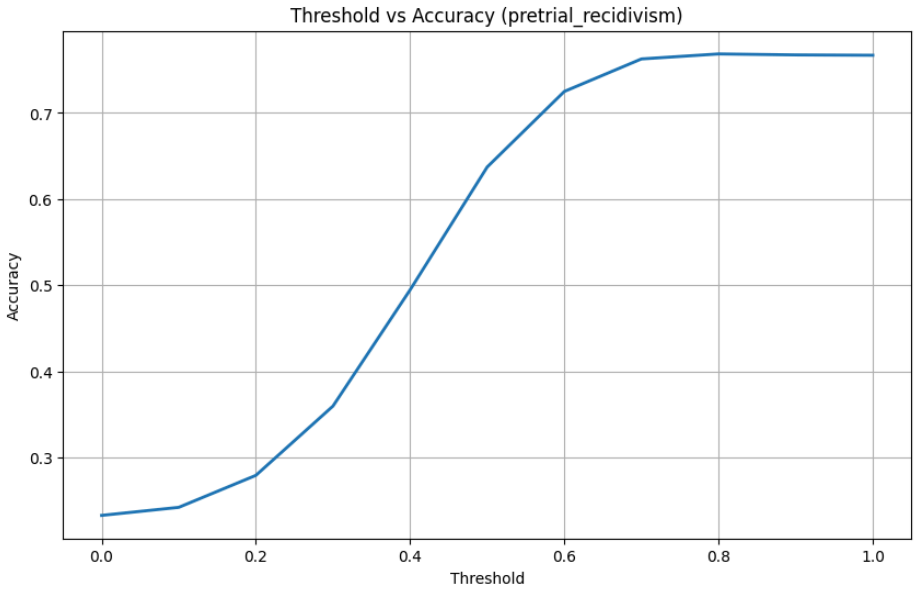
rise dramatically, eventually reaching 1.0 at threshold = 1.0, where all recidivists were misclassified. **Mid-range thresholds** delivered the most reasonable balance.

Pretrial Recidivism

For pretrial recidivism, the threshold range of **0.4–0.5** consistently produced the most stable trade-off. In this region, the model maintained a manageable FPR while keeping FNR substantially below the dummy baseline, which fails entirely to identify recidivists. Accuracy also increased sharply and plateaued, indicating that additional threshold increases delivered diminishing returns. These findings demonstrate that the Ridge model meaningfully distinguishes between recidivists and non-recidivists, unlike the dummy classifier that simply predicts the majority class.

```
====================================
TRAINING MODEL FOR TARGET = pretrial_recidivism
====================================

Top Thresholds:
    threshold  accuracy       FPR       FNR
10        1.0  0.766963  0.000000  1.000000
9         0.9  0.767364  0.000241  0.997483
8         0.8  0.768491  0.012155  0.953438
7         0.7  0.762672  0.067780  0.795337
6         0.6  0.724934  0.173315  0.609948
5         0.5  0.637279  0.354116  0.391045
4         0.4  0.494660  0.601791  0.187906
3         0.3  0.359881  0.814953  0.064711
2         0.2  0.279465  0.934755  0.015499
1         0.1  0.242360  0.987422  0.001391
```





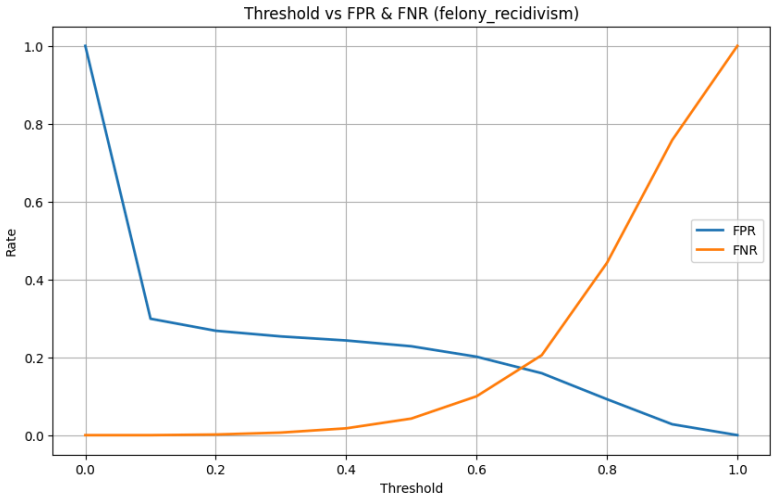
A threshold of **0.5** emerged as the most appropriate general-purpose cutoff. It offered the first meaningful improvement over the baseline classifier by producing balanced, consistent predictions across the diverse population of cases.

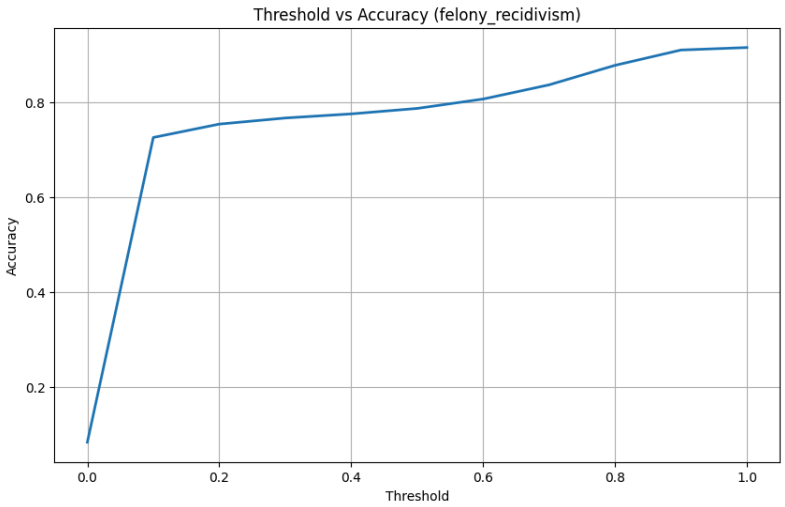
Felony Recidivism

Felony recidivism exhibited an even clearer separation between the extremes. Very low thresholds nearly eliminated false negatives but inflated false positives. Thresholds around **0.3–0.4** yielded a particularly effective balance, maintaining FNR close to zero while reducing

FPR to a reasonable range. Accuracy rose quickly and then stabilized, mirroring the balanced error tradeoff.

```
====================================
TRAINING MODEL FOR TARGET = felony_recidivism
====================================

Top Thresholds:
    threshold  accuracy       FPR       FNR
10        1.0  0.915586  0.000000  1.000000
9         0.9  0.910385  0.028001  0.757908
8         0.8  0.878110  0.092314  0.442677
7         0.7  0.836991  0.159089  0.205522
6         0.6  0.807109  0.201470  0.099835
5         0.5  0.787368  0.228308  0.042604
4         0.4  0.775838  0.243210  0.017553
3         0.3  0.767195  0.253662  0.006583
2         0.2  0.754337  0.268160  0.001646
1         0.1  0.726168  0.299078  0.000000
```



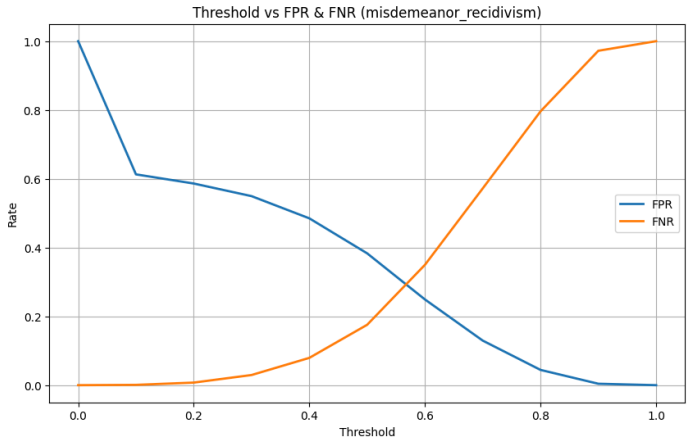


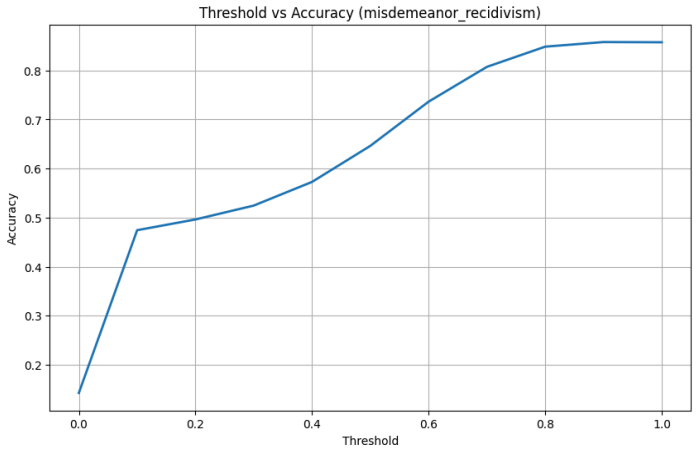
A threshold of **0.4** provided the best balance. At this level, the model maintained strong detection of true felony recidivists while avoiding the excessive over-classification associated with lower thresholds. This makes 0.4 the first threshold that is both accurate and operationally reasonable for high-severity cases.

Misdemeanor Recidivism

For misdemeanor recidivism, mid-range thresholds **between 0.5 and 0.6** provided the most viable equilibrium. At these values, the false negative rate remained manageable while the false positive rate decreased significantly. Beyond this range, FNR surged rapidly, making high thresholds impractical for real-world use. As with the other models, accuracy increased smoothly across the mid-range and plateaued, reinforcing these threshold choices.

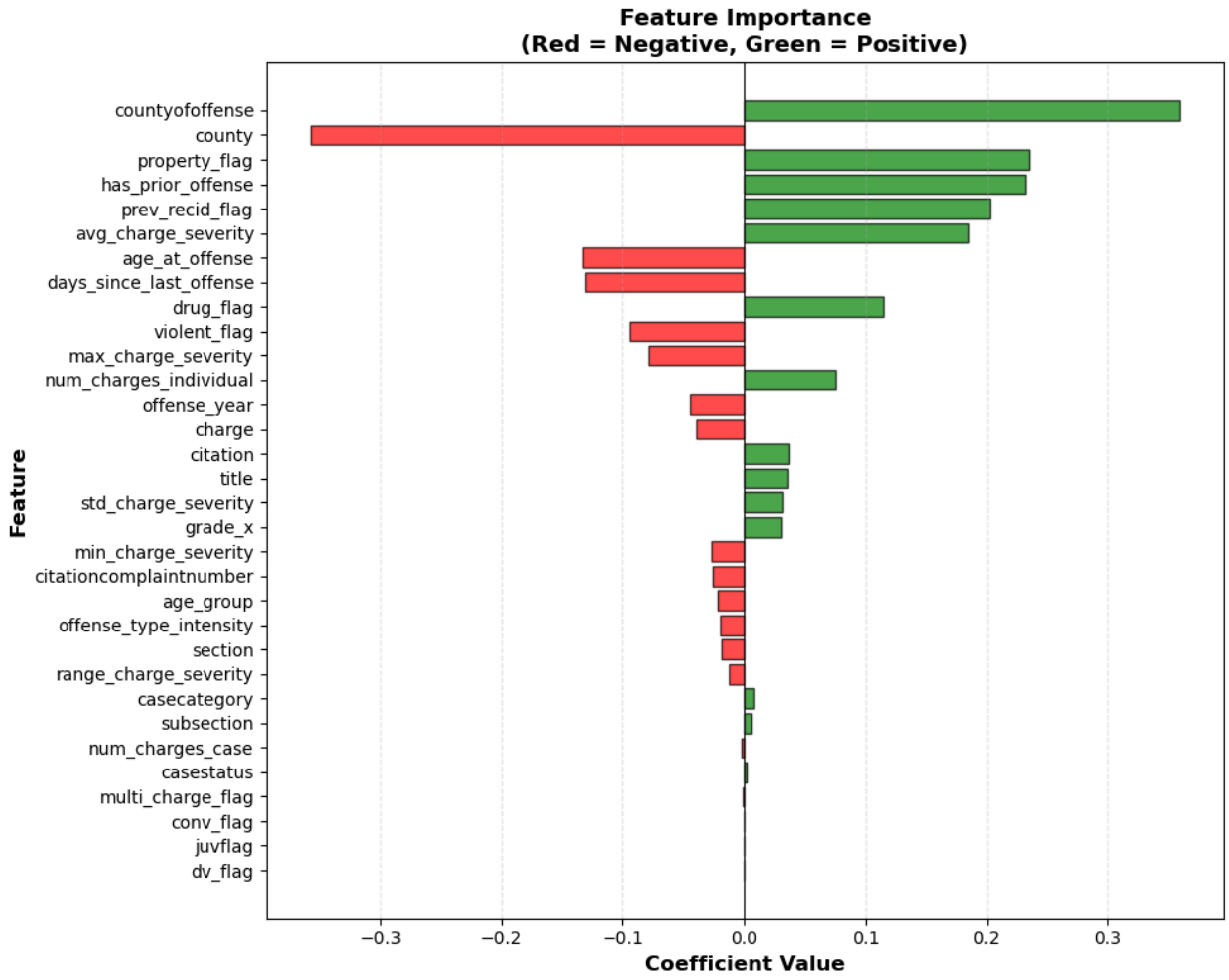
```
====================================
TRAINING MODEL FOR TARGET = misdemeanor_recidivism
====================================

Top Thresholds:
    threshold  accuracy       FPR       FNR
10        1.0  0.857628  0.000000  1.000000
9         0.9  0.858060  0.004121  0.972138
8         0.8  0.848645  0.044345  0.795967
7         0.7  0.807464  0.129544  0.571986
6         0.6  0.736494  0.249172  0.349848
5         0.5  0.646339  0.383234  0.175520
4         0.4  0.572776  0.484972  0.079358
3         0.3  0.524511  0.549528  0.029488
2         0.2  0.496203  0.586189  0.007480
1         0.1  0.474316  0.612807  0.000867
```





A threshold of **0.6** performed best. Misdemeanor patterns are more diffuse, and higher thresholds help stabilize predictions. The 0.6 threshold yielded the most reliable and interpretable results without compromising the model's ability to distinguish between reoffenders and non-reoffenders.

3.5: Feature Importance



The feature importance of our model reveals key predictors of pretrial recidivism, with the most influential factors being offense and defendant history characteristics. County-level geographic factors show the strongest effects, with countyofoffense (coefficient: +0.360) and county (coefficient: -0.358) being nearly opposite in their association with recidivism risk—suggesting that where the offense occurred versus where the defendant resides may capture important differences in court processing, local crime patterns, or case characteristics. Prior criminal history emerges as another critical predictor: defendants with a has_prior_offense flag (coefficient: +0.232) or positive prev_recid_flag (coefficient: +0.202) show substantially elevated recidivism risk, which aligns with criminological research linking prior behavior to future outcomes. Notably, several demographic and charge-level features also contribute meaningfully: property_flag (+0.238) and drug_flag (+0.117) increase recidivism likelihood, while age_at_offense (-0.134) and days_since_last_offense (-0.132) reduce it, indicating that younger age at offense and shorter time since the last offense are risk factors. Conversely, higher offense severity metrics like max_charge_severity (-0.078) and violent_flag (-0.092) actually

show negative associations with recidivism—a counterintuitive finding that may reflect selection bias (more serious cases receiving stricter supervision or detention) rather than true causal protection.

3.6: Error Tradeoffs

A central consideration in evaluating pretrial and post disposition risk models is the tradeoff between false positives and false negatives, as each type of error carries distinct operational and ethical implications. The false positive rate (FPR) reflects individuals incorrectly classified as high risk who would not recidivate, while the false negative rate (FNR) captures individuals classified as low risk who do in fact recidivate, posing potential public safety concerns.

Across all three outcomes, pretrial, felony, and misdemeanor recidivism, the threshold sweep plots reveal a consistent and expected tradeoff pattern. At low thresholds, FPR is extremely high while FNR is near zero, indicating that the model flags nearly everyone as high risk. As the threshold increases, FPR steadily declines, but this improvement is accompanied by a sharp rise in FNR, meaning the model increasingly fails to identify true recidivists. These opposing trends highlight that neither error metric can be minimized simultaneously, making threshold selection a value driven decision rather than a purely technical one.

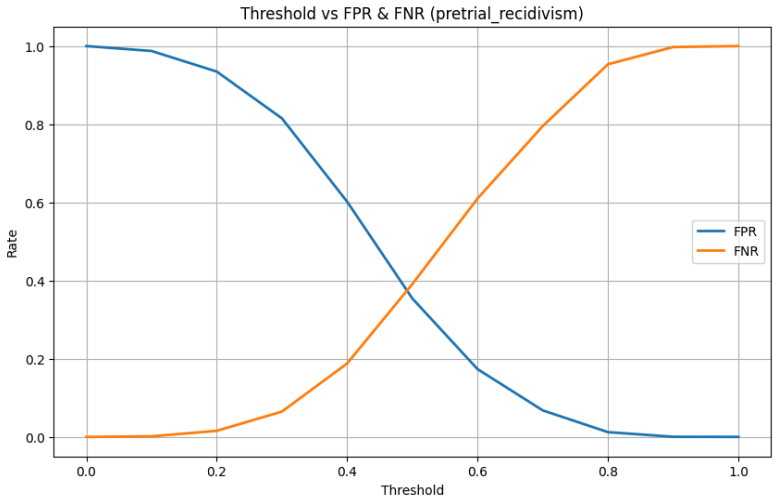
For pretrial recidivism, the crossover between FPR and FNR occurs near the mid range thresholds. Accuracy rises rapidly between thresholds 0.3 and 0.6 and then plateaus beyond roughly 0.7, indicating diminishing returns from stricter classification. A threshold around t equal to 0.5 represents a balanced operating point. At this threshold, FPR has already declined substantially from its extreme low threshold values, while FNR has not yet risen to levels that would severely compromise sensitivity. Moving to higher thresholds further reduces FPR but does so at the cost of sharply increasing FNR, meaning many true pretrial recidivists are missed despite only marginal gains in accuracy.

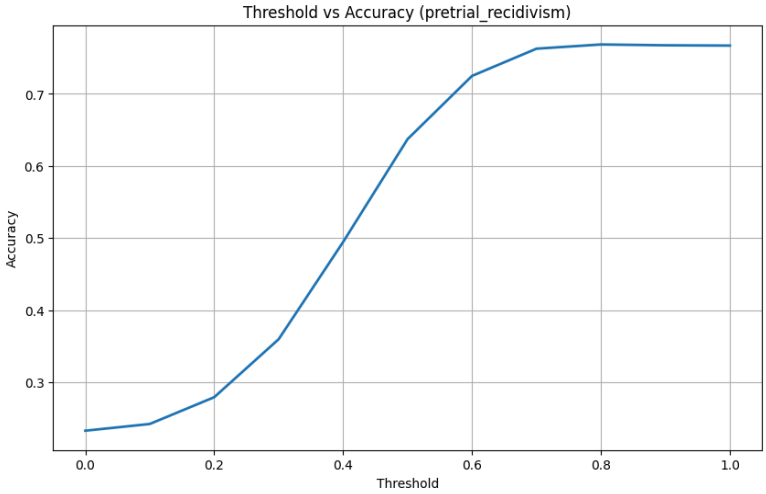
For felony recidivism, accuracy increases monotonically with the threshold and reaches its highest values at very high thresholds. However, the FNR rises steeply after approximately t equal to 0.7 and approaches 1.0 at the upper end, indicating that nearly all felony recidivists are missed. While FPR becomes very low in this region, the model effectively stops identifying high risk individuals. This demonstrates that maximizing accuracy alone leads to an operationally weak model for felony recidivism, as the reduction in false positives comes at the expense of public safety sensitivity.

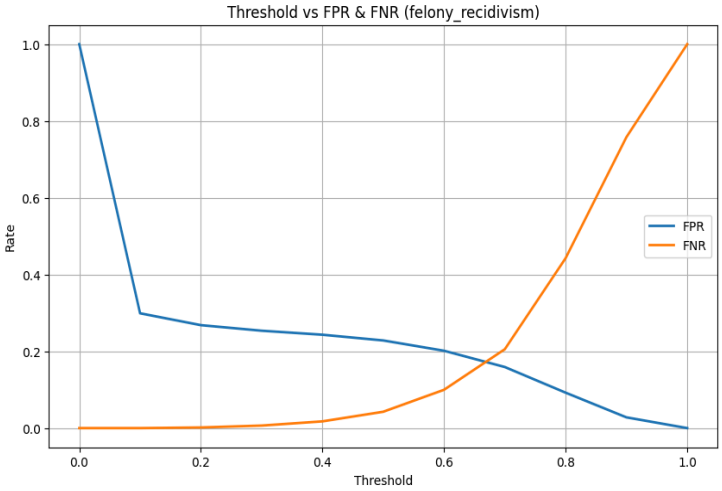
A similar pattern is observed for misdemeanor recidivism. Accuracy improves steadily with higher thresholds and plateaus near the top end, while FPR declines and FNR increases sharply beyond mid range thresholds. Thresholds above approximately 0.7 result in very low FPR but extremely high FNR, again indicating that most true recidivists are classified as low risk. As with the other outcomes, accuracy gains in this region are marginal compared to the substantial loss in recall.

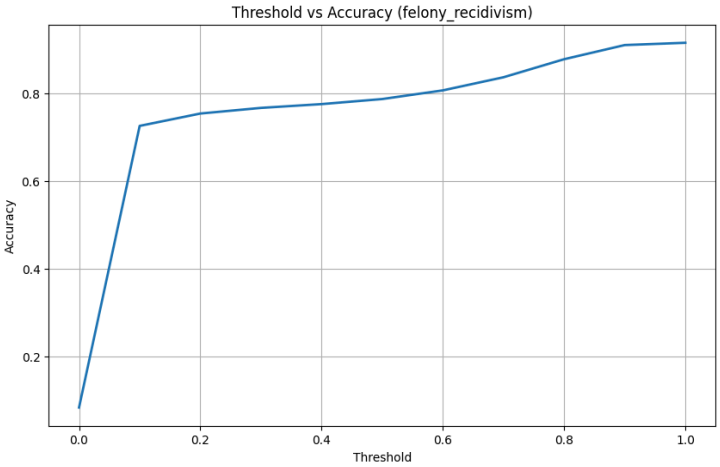
Overall, the plots demonstrate that moderate thresholds in the 0.45 to 0.60 range provide the most defensible balance across all three targets. In this region, accuracy is close to its maximum while both FPR and FNR remain within non extreme ranges. Selecting higher thresholds yields minimal accuracy improvements but produces disproportionate increases in false negatives, undermining the model's usefulness as a risk screening tool.

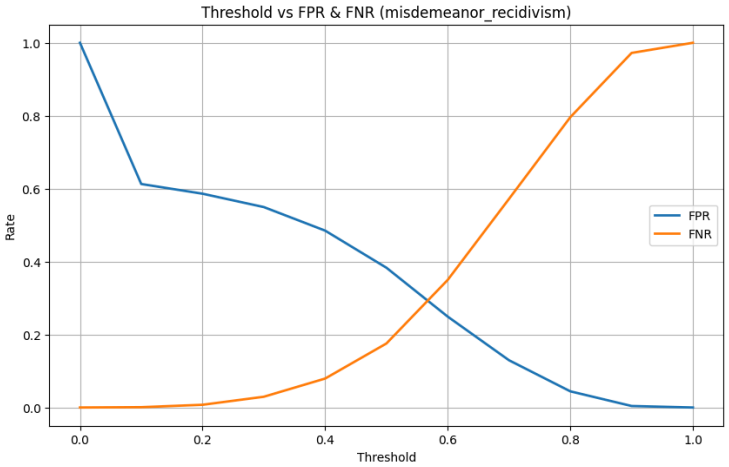
In summary, the threshold analysis confirms that aggressive thresholding prioritizes accuracy and false positive reduction at the cost of missing true recidivists, while overly permissive thresholds overwhelm the system with false positives. The mid range thresholds observed in the plots strike a practical and ethical equilibrium, supporting the model's role as a decision support tool that balances fairness, public safety, and operational feasibility rather than serving as a rigid decision rule.

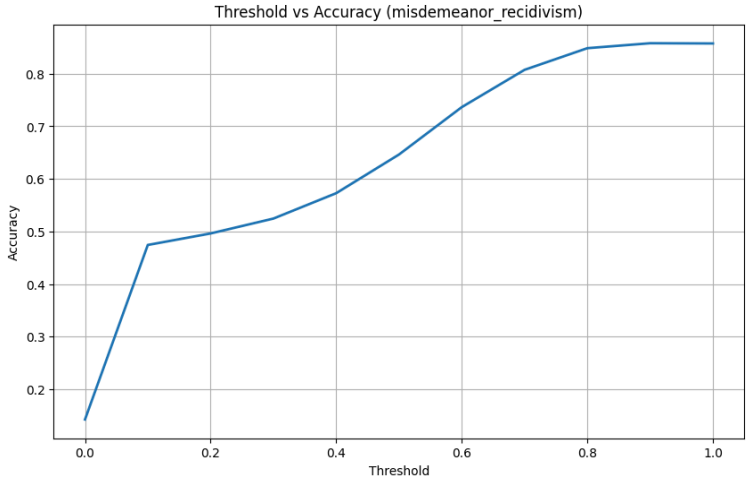





Threshold vs FPR & FNR (felony_recidivism)
Rate
Threshold
FPR
FNR
0.0
0.2
0.4
0.6
0.8
1.0


Threshold vs Accuracy (felony_recidivism)
Accuracy
Threshold
0.0
0.2
0.4
0.6
0.8
1.0


Threshold vs FPR & FNR (misdemeanor_recidivism)
Rate
Threshold
FPR
FNR
0.0
0.2
0.4
0.6
0.8
1.0


Threshold vs Accuracy (misdemeanor_recidivism)
Accuracy
Threshold
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

4.1: Challenges

A central challenge throughout the project was managing the size and complexity of the PSP, MDJ, and CPMC datasets. Each source contained millions of charge-level observations, which quickly exceeded the capabilities of local machines for tasks like loading, merging, and running basic exploratory checks. To move forward, we had to coordinate with computing services to secure a more powerful virtual machine, configure the environment for large-scale data handling, and transfer all intermediate files. Even with improved infrastructure, processing required careful sequencing, memory-efficient operations, and repeated verification to ensure that merges across sources behaved consistently and preserved all relevant information.

A second challenge involved operationalizing the pretrial window in a way that was consistent and analytically meaningful. Because each dataset reported events on different timelines and in different formats, constructing accurate time deltas required detailed inspection of date fields, correction of errors and irregularities, and iterative logic checks. We had to validate that the ordering of events aligned with procedural expectations and that the calculated windows captured true pretrial activity rather than structural artifacts in the data, with assistance from the Duquesne law students. This work demanded substantial time and methodological precision, especially given the downstream implications for modeling and interpretation.

4.2: Limitations

Our challenges introduced clear limitations. Reliance on a single virtual machine meant that our processing environment was centralized and sometimes slow to iterate. Any change in logic required re-running large portions of the pipeline, and interactive exploration was constrained by hardware limits and long processing times. This dependence on heavy computational resources reduces portability and makes it more difficult for future teams to reproduce work without similar access. It also increases the likelihood of bottlenecks if multiple collaborators need to run large jobs simultaneously.

More broadly, our analytic outputs remain influenced by the structural quality of the underlying datasets. Inconsistencies in date formats, incomplete timestamps, and variation in how agencies document events create noise that is difficult to fully resolve through cleaning alone. Although we applied validation checks, some uncertainty remains in the final pretrial window calculations, which can affect downstream descriptive statistics or modeling. Additionally, because our work focused primarily on data construction, we were not able to fully explore issues such as fairness, representativeness, or systemic bias across the merged dataset, which limits the interpretability and policy relevance of the initial findings.

4.3: Next Steps

The next cohort would benefit from **investing early in a more formalized, automated data pipeline that can reliably ingest, clean, and merge the PSP, MDJ, and CPMC datasets without extensive manual intervention**. Containerized environments, workflow managers, and standardized schema-mapping tools would improve reproducibility and help avoid redundant processing work. Documenting each preprocessing decision, along with automated quality checks, would also support long-term continuity so that future teams can focus more of their time on analysis rather than infrastructure.

Another priority is to **deepen the analytical scope by integrating responsible AI evaluation methods into the project**, such as LIME and SHAP explainability analyses. Once the dataset is fully stabilized, students should assess fairness, representation, and potential biases across demographic groups, urban v. rural regions, decision points, and case outcomes. Tools such as group-level error metrics, disparity ratios, and counterfactual fairness checks could meaningfully extend the project and provide policy-relevant insights. Additional opportunities include **expanding feature engineering for future modeling efforts**, **conducting robustness checks on the pretrial window logic**, and **exploring alternative temporal definitions to assess the sensitivity of findings**. Strengthening these components will help the project evolve from dataset construction toward a more comprehensive, policy-informed analysis.

Appendix

Appendix A: Definitions for Feature Variables

Demographic Features

age_at_offense calculates the offender's age in years at the time of the offense by computing the difference between the offense date and date of birth in days, then converting to years using the standard 365.25-day year.

age_group categorizes offenders into four age brackets (Under 25, 25-34, 35-44, 45+) based on their calculated age at offense, providing a categorical representation of age for stratified analysis.

multi_county_flag is a binary indicator showing whether a defendant's cases span multiple counties, highlighting cross-jurisdictional or mobile offending behavior that may indicate broader criminal activity.

Criminal History Features

num_charges_individual counts the cumulative number of charges for each offender across all cases in their criminal history, capturing total charge count at the individual level to reflect overall criminal history severity.

days_since_last_offense calculates the number of days between consecutive offenses for each individual by grouping records by offender ID and computing day differences between successive offense dates within each group (NaN if first offense).

prev_recid_flag is a binary variable (0/1) indicating if a defendant has prior pretrial recidivism history, serving as a strong predictor of repeat-offense risk and behavioral persistence.

offense_during_same_year_flag marks defendants who commit multiple offenses within the same calendar year, identifying short-term reoffending or impulsive behavioral patterns within a condensed time frame.

has_prior_offense is a binary indicator (0/1) denoting whether a defendant has any documented prior criminal offense in their record. This feature captures long-term criminal involvement beyond pretrial behavior specifically and serves as a foundational measure of an individual’s historical offending trajectory.

Sentencing Characteristic Features

supervision_violation_flag creates a binary indicator that flags whether an offender had a supervision violation by checking if either the suspension flag (susp_flag) or the leave of absence flag (laflag) equals 1, with the feature equaling 1 if either violation type is present and 0 otherwise.

waived_or_dismissed_flag creates a binary indicator identifying cases with favorable dispositions by searching the case disposition text for the keywords "Waived," "Dismissed," or "Withdrawn" (case-insensitive). The feature equals 1 if any of these terms appear in the disposition and 0 otherwise.

case_duration_days measures the total number of days from case filing to disposition, capturing how court processing time relates to case complexity and potential pretrial behavior patterns.

county_recid_rate represents the average pretrial recidivism rate within each county, providing a contextual measure of local reoffense tendencies and judicial policy differences.

severity_dismiss_interaction is an interaction term combining charge severity and dismissal status, highlighting how dismissal outcomes interact with offense seriousness and revealing whether defendants with more severe charges benefit from dismissal or maintain elevated reoffense risk.

Offense-Level Features

drug_flag creates a binary indicator for drug-related offenses by searching for drug-related keywords across multiple charge description fields. It flags cases where the offense title, statute section, or charge description contains terms such as "drug," "narcotic," "controlled substance," "marijuana," "cocaine," "heroin," "methamphetamine," or the specific statute code "780-113." The feature equals 1 if any drug-related term is found and 0 otherwise.

violent_flag creates a binary indicator for violent offenses by searching for violence-related keywords and statute codes across multiple charge description fields. It flags cases where the offense title, statute section, or charge description contains terms such as "assault," "aggravated," "robbery," "rape," "murder," "homicide," "kidnap," "terroristic," or specific statute codes (2701-2708). The feature equals 1 if any violent crime indicator is found and 0 otherwise.

property_flag creates a binary indicator for property crimes by searching for theft-related keywords and statute codes across multiple charge description fields. It flags cases where the offense title, statute section, or charge description contains terms such as "theft," "burglary," "trespass," "shoplifting," "receiving stolen," or specific statute codes (3921-3929). The feature equals 1 if any property crime indicator is found and 0 otherwise.

num_charges_case counts the number of charges per case, reflecting the complexity and breadth of criminal conduct within a single case.

charge_severity maps each individual charge grade to a numerical severity score using the continuous scale: F1→7, F2→6, F/F3→5, M1→4, M2→3, M/M3→2, S→1, U/empty→0. This creates a standardized numeric representation of charge severity for individual charges.

max_charge_severity extracts the most serious charge grade from cases with multiple charges and maps it to a severity score (0-7), capturing the severity of the most serious offense in the case, which often has the strongest influence on case outcomes and recidivism risk.

min_charge_severity extracts the least serious charge grade and maps it to a severity score, providing insight into the full range of offense behavior by identifying the least serious charge in multi-charge cases.

avg_charge_severity computes the average severity score across all charges in a case, providing a balanced overall measure of case seriousness when multiple charges of varying severity are present.

std_charge_severity calculates the standard deviation of severity scores across charges. High values indicate diverse charge types (e.g., mixing felonies with misdemeanors), while low values indicate similar charges, reflecting the diversity of criminal behavior.

range_charge_severity computes the difference between the maximum and minimum severity scores in a case, indicating the spread between the most and least serious charges and reflecting the diversity of criminal behavior in a single case.

offense_type_intensity counts how many offense categories (drug, violent, property) apply to the case. Higher values indicate broader offending patterns and greater behavioral versatility, both of which are linked to increased recidivism likelihood.

multi_charge_flag is a binary indicator (0/1) showing whether a case contains multiple charges, identifying defendants with more complex or varied offense behavior that may be associated with greater criminal versatility or elevated recidivism risk.

Appendix B: Correlation Metrics for Feature Variables

The following table presents correlation metrics quantifying the relationship between each feature variable and pretrial recidivism. Pearson's r is used for numeric variables, while Cramér's V is used for categorical variables. Both metrics range from 0 to 1 (or -1 to 1 for Pearson's r), where higher absolute values indicate stronger associations with recidivism.

Feature	**Type**	**Metric**	**Correlation Value**
prev_recid_flag	Categorical	Cramér's V	0.1883
has_prior_offense	Categorical	Cramér's V	0.1639
property_flag	Categorical	Cramér's V	0.1270
num_charges_individual	Numeric	Pearson r	0.0957
avg_charge_severity	Numeric	Pearson r	0.0748
min_charge_severity	Numeric	Pearson r	0.0718
offense_type_intensity	Categorical	Cramér's V	0.0696
violent_flag	Categorical	Cramér's V	0.0677
age_at_offense	Numeric	Pearson r	-0.0604
age_group	Categorical	Cramér's V	0.0586
max_charge_severity	Numeric	Pearson r	0.0580
drug_flag	Categorical	Cramér's V	0.0197

multi_charge_flag	Categorical	Cramér's V	0.0115
days_since_last_offense	Numeric	Pearson r	0.0091
num_charges_case	Numeric	Pearson r	0.0033
range_charge_severity	Numeric	Pearson r	0.0015
std_charge_severity	Numeric	Pearson r	0.0003

References

Bureau of Justice Statistics. (n.d.). Pretrial release. Retrieved December 2, 2025, from https://bjs.ojp.gov/topics/courts/pretrial-release

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2013). Risk assessment instruments validated and implemented in correctional settings in the United States. CSG Justice Center. https://csgjusticecenter.org/wp-content/uploads/2020/02/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! Criminology, 34(4), 575-608. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-9125.1996.tb01220.x

Heaton, P., Mayson, S., & Stevenson, M. (2017). The downstream consequences of misdemeanor pretrial detention. Stanford Law Review, 69(3), 711-786. https://www.stanfordlawreview.org/print/article/the-downstream-consequences-of-misdemeanor-pretrial-detention/

Pretrial Justice Institute. (n.d.). Overview of research findings on pretrial risk assessment and pretrial supervision. U.S. Pretrial Justice Institute. https://www.sog.unc.edu/sites/default/files/course_materials/Clark_Overviewofresearchfindingsonpretrialriskassessmentandsupervision.pdf

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. Criminology, 54(4), 674-716. https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-9125.12123

United States Sentencing Commission. (2017). The effects of aging on recidivism among federal offenders. U.S. Sentencing Commission. https://www.ussc.gov/research/research-reports/effects-aging-recidivism-among-federal-offenders

VanNostrand, M., & Lowenkamp, C. T. (2013). Assessing pretrial risk without a defendant interview. Laura & John Arnold Foundation. https://static.prisonpolicy.org/scans/ljaf/LJAF_Report_no-interview_FNL.pdf