MS&E 226: Fundamentals of Data Science
Project Part 2: Inference and Causality
Predicting Recidivism

Ben Rosenfeld, Ness Arikan

1. Prediction on the holdout set

□ We applied our best model from Part 1 (neural network, AUC = 0.70) to the held-out test set. For comparison, we fit a logistic regression model achieving test log loss of 0.3781, accuracy of 0.8568, and AUC of 0.7029. Performance is consistent with Part 1 (validation log loss 0.6948, AUC 0.70). The logistic regression performs similarly in AUC, suggesting linear relationships are sufficient. The log loss difference (0.38 vs 0.69) may reflect the neural network's tendency toward extreme probability estimates versus logistic regression's better calibration.

2. Inference

2.1 Statistical Significance of Coefficients

□ (a) We fit logistic regression on training data (n = 63,739). Using $\alpha = 0.05$, we found 14 statistically significant coefficients. Statistical significance means that under the null hypothesis (coefficient = 0), we would observe such an extreme coefficient less than 5% of the time by chance, suggesting non-zero associations unlikely due to random variation.

Most practically meaningful for the CJA: **pending_misd** (coef = 0.955, $p < 10^{-248}$) is the strongest predictor; **pending_nonvfo** (0.633, $p < 10^{-79}$) and **pending_vfo** (0.608, $p < 10^{-38}$) are also strong; **prior_misd_cnt** (0.102, $p < 10^{-79}$) increases log-odds by 0.102 per additional conviction; **age** (-0.202, $p < 10^{-48}$) decreases risk; **male** (0.224, $p < 10^{-14}$) increases risk.

2.2 Significance on Test Data

□ (b) On test data (n = 16,016), we found 10 significant coefficients. Four variables lost significance: **black** ($p = 0.877$ vs $p < 10^{-5}$), **nyc** ($p = 0.332$ vs $p < 10^{-7}$), **nmr_requested** ($p = 0.358$ vs $p = 0.002$), and **prior_nonvfo_cnt** ($p = 0.275$ vs $p = 0.001$). Differences reflect: (1) sampling variability, (2) overfitting—spurious training patterns, (3) reduced power from smaller test size. Strongest predictors (pending charges, prior misdemeanor counts, age) remained significant, suggesting robustness. Loss of demographic significance (black, nyc) suggests less stable associations.

2.3 Bootstrap Confidence Intervals

□ (c) We computed bootstrap CIs using 1,000 samples. Bootstrap and standard CIs were very similar (mean width difference = 0.0099, median = 0.0031). For **age**, bootstrap CI was [-0.230, -0.174] vs standard [-0.229, -0.175]. Similarity suggests standard assumptions are reasonable. Bootstrap advantages: no distributional assumptions, intuitive resampling interpretation, more defensible to non-statisticians. For the CJA, we would report bootstrap CIs because they are more robust, intuitive, and defensible. Their similarity to standard intervals strengthens confidence.

2.4 Multiple Hypothesis Testing

□ (d) We tested 19 hypotheses. Without correction, 13 were significant at $\alpha = 0.05$. After Bonferroni ($\alpha_{corrected} = 0.0026$) and Benjamini-Hochberg, all 13 remained significant, reflecting extremely small p-values ($< 10^{-30}$ for key predictors). For stakeholder presentation, we would **not** apply correction because: (1) conclusions unchanged, (2) CJA needs all predictive factors, (3) Bonferroni is conservative, (4) prediction/risk assessment prioritizes understanding all factors over strict Type I error control. We would communicate that multiple hypotheses were tested and some associations (black, nyc) were not robust across splits.

2.5 Post-Selection Inference

□ (e) Our process suffers from post-selection inference: (1) **Model selection**: Comparing Lasso, Ridge, Neural Network and selecting best inflates p-values. (2) **Feature engineering**: Preprocessing decisions (charge_score creation, missingness indicators, scaling, interactions) were data-informed, making inference conditional. (3) **Variable selection**: Examining correlations and excluding variables (e.g., judge names) conditions inference. (4) **Hyperparameter tuning**: Grid search conditions results. Standard p-values don't account for selection. Proper accounting requires sample splitting or post-selection techniques. However, given large sample size and relationships remaining significant after multiple testing correction, key findings appear robust.

3. Stakeholder Guidance

3.1 Informing Decisions

□ (a) Our results inform CJA pretrial release and supervision decisions. **Statistical significance**: Strong associations (pending charges, prior convictions, age, gender) with p-values $< 10^{-30}$ indicate non-spurious findings. **Practical significance**: **Pending charges** are strongest—pending misdemeanor increases rearrest odds 2.6x (exp(0.955) = 2.60). **Age** is protective—one SD increase (approximately 12 years) decreases odds 18% (exp(-0.202) = 0.82). **Prior misdemeanors** accumulate risk—each additional increases odds 11% (exp(0.102) = 1.11). **Gender** matters—males have 25% higher odds (exp(0.224) = 1.25). CJA can: (1) develop risk tiers, (2) allocate supervision efficiently, (3) provide targeted support.

3.2 Causal vs Correlational Relationships

□ (b) The **pending charges**-rearrest relationship is critical for causality. If causal, pending charges directly increase rearrest risk (stress, legal complications, criminal network embedding). If correlational, they mark underlying risk factors (criminal history, socioeconomic status, neighborhood). The distinction matters: if **causal**, CJA might adjust release decisions or support types for pending charges. If **correlational**, pending charges remain useful risk markers, but CJA should focus interventions on underlying causal factors.

3.3 Evaluating Causal Interpretation

□ (c) For pending charges-rearrest, threats to causal inference: **Confounding**: (1) Criminal history—unmeasured aspects (gang affiliation, drug addiction) may confound despite controlling for prior counts. (2) Socioeconomic factors—poverty, unemployment, housing instability (not in dataset) may independently increase risk. (3) Neighborhood characteristics—high-crime areas may increase both pending charges and rearrest risk. (4) Legal representation quality—poor representation may increase pending charges and worsen outcomes. **Selection effects**: Analysis includes only released defendants, creating selection bias if release factors correlate with both pending charges and rearrest. **Measurement issues**: Pending charges measured at current arrest; timing/accuracy varies; some may resolve before rearrest window. **Conclusion**: Relationship is **primarily correlational**. While pending charges may create direct stress/complications, strong confounding by criminal history, socioeconomic factors, and neighborhood characteristics likely explains much association. Pending charges are likely markers for deeper risk factors, not direct causes. However, they remain highly useful for CJA risk assessment.

3.4 Strengthening Causal Claims

□ (d) To strengthen causal claims: (1) **Natural experiment/IV**: Policy changes affecting pending charge assignment, or IV using judge leniency/prosecutor workload. (2) **Richer covariates**: Socioeconomic variables (income, employment, education, housing), neighborhood characteristics (crime rates, poverty, police presence), mental health, substance use, family/social support. (3) **Longitudinal data**: Following defendants over time to see how pending charge status changes relate to rearrest, controlling for time-invariant characteristics. (4) **RCT**: Randomizing supervision interventions, though ethically/practically challenging. **Feasibility**: Most feasible are richer covariate data and natural experiments (policy changes, judge assignment variation). RCT is difficult due to ethical concerns. CJA could partner with researchers using judge assignment as natural experiment, as judges vary in release tendencies, providing quasi-random variation affecting pending charge accumulation.

4. AI use

□ (a) We used: Cursor AI (Auto agent) for R code generation, preprocessing, analysis, and report writing; ChatGPT/GPT-4 for code structure and debugging.
□ (b) **AI was helpful for**: (1) Code generation/debugging—quickly generating bootstrap, multiple testing, and preprocessing code. (2) Statistical methodology—explanations of bootstrap, multiple testing, post-selection inference. (3) Report structure—organizing sections comprehensively. **AI was less helpful for**: (1) Domain interpretation—understanding criminal justice implications required our context knowledge. (2) Causal reasoning—evaluating causality required careful consideration of confounding, selection, measurement that AI could assist but not fully reason through. (3) Stakeholder communication—translating findings for CJA required understanding their specific needs.

5. Code upload

□ Code is available at: `https://github.com/rosenfeldb/226`

Figures

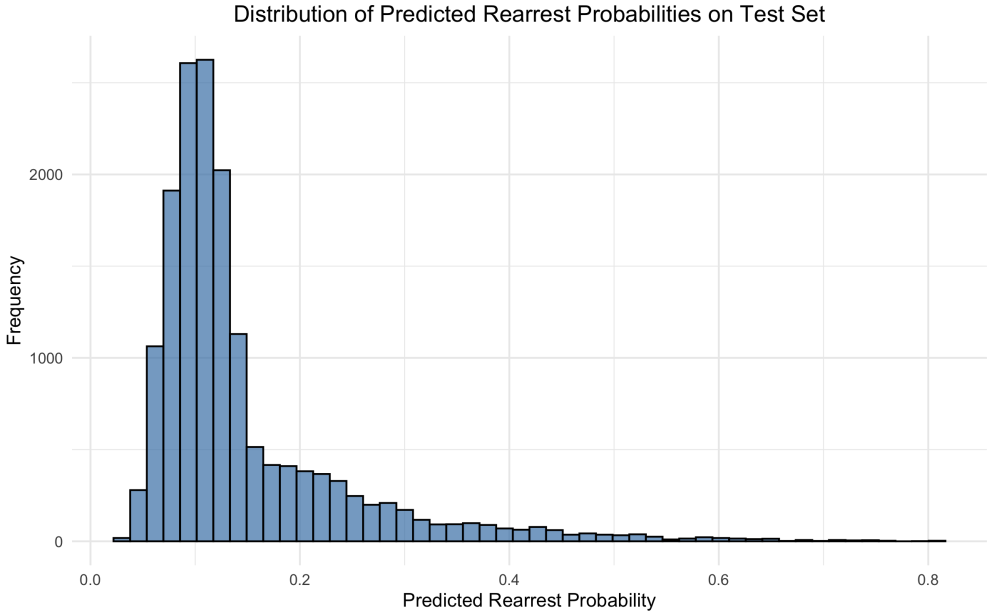

Figure 1: Distribution of predicted rearrest probabilities on test set

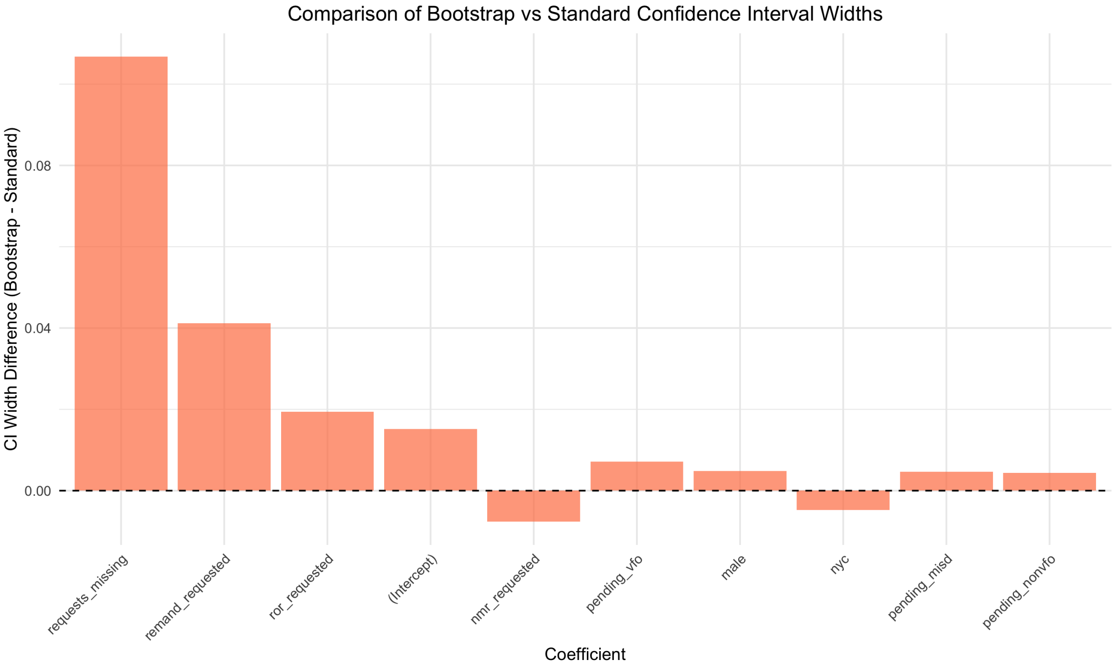

Figure 2: Comparison of bootstrap vs standard confidence interval widths

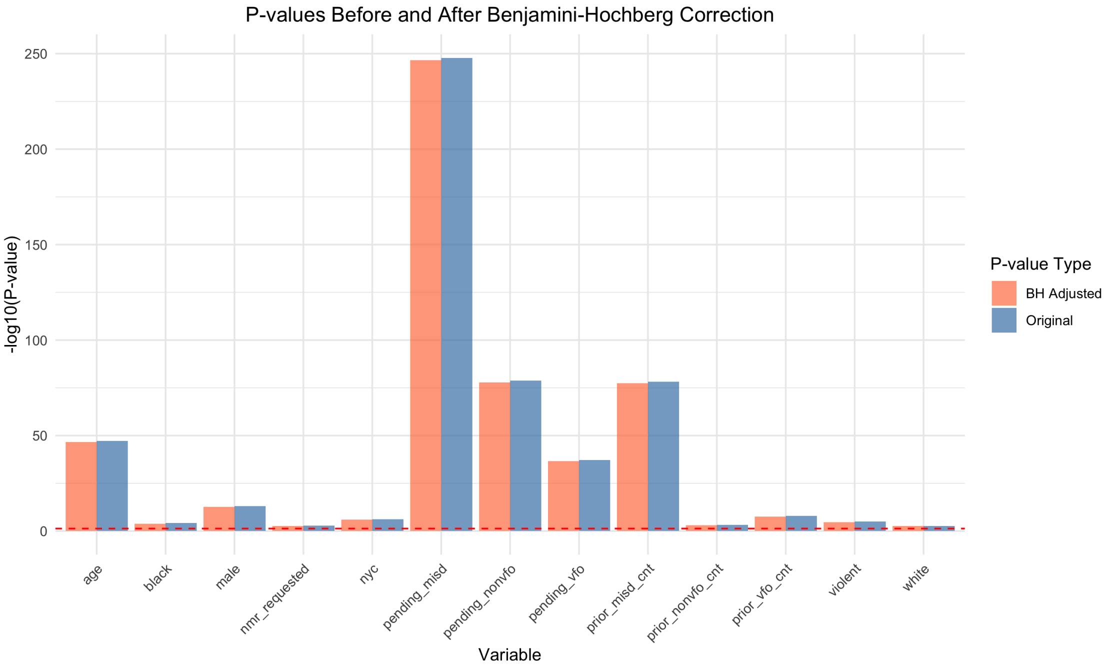

Figure 3: P-values before and after Benjamini-Hochberg correction