Analyzing Parole Violations

ST635 Group 1: Ethan Clabaugh, Jaime Fastino,
Cayty Fitzgerald, Nicole Romangsuriat, Jessica Smith

Agenda

01 Dataset

02 Research Questions

03 Methodology

04 Findings

05 Model Comparisons

06 Conclusion

Dataset

❑ Kaggle <- US 2004 National Corrections Reporting Program

  ❑ Nationwide census of parole releases in 2004

❑ Only paroles serving <= 6 months

❑ Only parolees whose max sentence <= 18 months

❑ Response: parole violation (yes/no)

❑ Predictors: 5 categorical, 3 quantitative

❑ 675 data values

Research Questions

- ❑ Which individual predictor has the greatest impact on parole violations?
- ❑ Which combination of variables are most predictive of a parole violation?
- ❑ Can we create a model that accurately predicts parole violations?

Methodology

Understanding & Cleaning Data

- Exploratory Data Analysis (no. of values, missing values, variable types)
- Summary Statistics
- Mean Data

Logistic Regression

- ➢ Fit model with all predictors
- ➢ Compare p-values
- ➢ Revise the model

Ridge & LASSO

- ➢ Divide data into training/testing sets
- ➢ Fit two logistic regression models
- ➢ Observed lambda values

PCA

- Find predictors, group with R
- Choose PCs (Scree Plot, 80-90% Rule)
- Fit linear regression

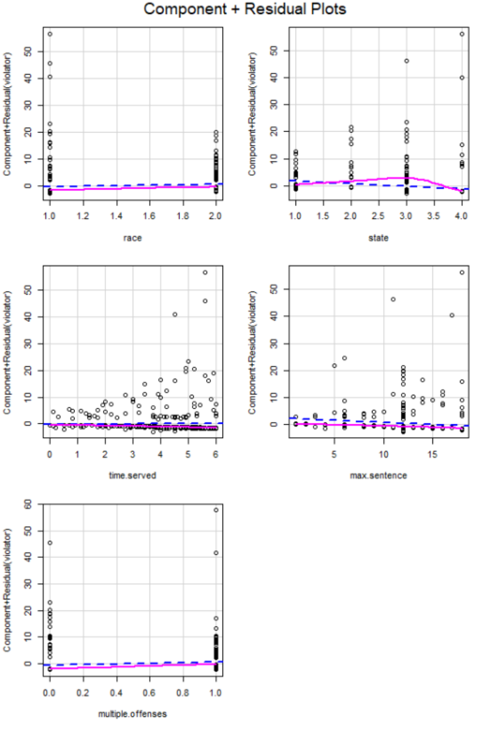
Classification Tree

- Split into training/testing data
- Build full tree with CP = 0
- Prune tree - prevent overfitting
- Look at xerror in CP tables, pick smallest values

Model Comparisons

- Model Utility (AUC)
- Goodness of Fit & Model Assumptions
- Accuracy (Confusion Matrix & R^2)

Findings



Logistic Regression

- ❖ AUC: 0.767 > 0.5 -> good utility
- ❖ Likelihood Ratio: near 0 -> goodness-of-fit
- ❖ Straight residuals -> linearity
- ❖ VIF close to 1 -> no multicollinearity
- ❖ 88.3% confusion matrix accuracy
- ❖ 0.99 sensitivity, 0.06 specificity

Ridge & LASSO

- AUC 0.6909 and 0.6959 > 0.5 -> good utility
- ^ Similar -> No severe overfitting/multicollinearity
- ^^ Lower than logistic regression -> penalization not useful
- Lambda values close to 0 -> optimization regulation has little impact

PCA

- ➢ PC1 had highest sd value
- ➢ PC1 had highest proportion of variance
- ➢ 1st 5 PCs had cumulative proportion of 81%

- ➢ Adjusted R^2: 0.7267 -> accurate

Classification Tree

- Pruned Tree
  - AUC: 0.799 -> good utility
  - Confusion Matrix accuracy: 87.68% -> accurate model
- Full Tree
  - AUC: 0.758, accuracy: 85.22%
- <u>Pruned tree has the optimal model</u>

Model Comparison

	Logistic Regression	Ridge & LASSO	PCA	Decision Tree
AUC Values	0.696	0.767	-	0.799
Accuracy (CM, Adjusted R^2)	88.3%	88.8%	0.727	87.7%

Answers to RQs

Best Overall Model:
Classification Tree

PCA - group variables
Other 3 - binary models

3 models can accurately
predict violations

Most Predictive Variable Combo:
age, multiple offenses, race, state

Most Impactful Predictor:
multiple offenses

Mean of 4 predictors -> 0.71
probability of not violating

Thank you!

Questions are welcome.