Illinois DOC Sentencing Machine Learning Analysis

Juan Gonzalez, Aaron Graff, Ilijah Pearson, Hope Winsor

2025-12-17

Abstract

Incarceration and criminal justice are two large issues currently affecting the United States. Although there has been work to mend some of the past injustices that have occurred on the basis of race, gender, sexual identity, and more, bias and discrimination undoubtedly remain apart of our society. In particular, one issue that typically exemplifies these unresolved issues is the prison system. This report takes the approach of analyzing prison specific data, from the Illinois Department of Corrections, using machine learning to predict charge category, as well as sentencing length. Predicting offense category was much more successful for offense categories with a large sample, potentially indicating more variability in predictors for lower-level offenses. For sentence length, random forest modeling performed the best, with the relationship between prison sentence and the predictor variables likely not being linear. Demographic data, such as race, sex, and weight were less important predictors for sentencing length, indicating an encouraging sign that sentence length is based primarily on criminal history and the specific crime committed, rather than being biased by physical traits.

Background

We chose to analyze a dataset recording people incarcerated in Illinois because we were curious to understand the relationships underlying an individual's prison sentence. Our data comes from kaggle, it contains three csv files; sentencing, person, and marks. Across these datasets we have a unique id for each prisoner assigned by the Illinois Department of Corrections. Because this data was scraped from the Illinois Department of Corections' website, this data required a substantial amount of cleaning. The variables in these datasets were initially separated by semicolons and the data failed to parse into designated columns because there were also semicolons used within the information scraped from the website. A parser was used to replace all within-text semicolons with commas to resolve this issue. The sentencing and marks files had different rows for each of an individual's charges and marks respectively. Thus these files were modified to group together all of the individual's rows so that each row had a unique ID and the three datasets could be merged together on ID. Within the sentencing csv, individuals recieved a sentence that was either numerical, indicating the length of time they were sentenced to serve, or one of the following categories: life, death, or sexually dangerous person. The sentence of sexually dangerous person did not lead to time served in prison but did lead to that individual being placed on a registry for sexually dangerous persons, so during data cleaning this category was replaced with the numerical value 0. Individuals who received a death sentence were assigned a prison sentence of 15 years, as our research indicated this was the median amount of time an individual spent in prison while waiting for a death sentence to be enacted. More severe prison sentences were represented in the sentencing csv in a variety of ways. Some individuals received sentences with lengths of time longer than the human lifespan, including some sentences greater than 3,000 years. To standardize these numbers, prison sentences were capped at 100 years in the data cleaning process. Those who received a life sentence were also assigned a numerical sentence of 100 years to match this standardization.

Research Questions

- What predictors have the strongest effect on sentencing length?
- How accurately can we predict sentencing length?
- How accurately can we predict offence categories?

To investigate these research questions, we fit four different categories of models. Linear models, generalized additive models, and tree-based models were used to investigate sentencing length. K nearest neighbors based models were used to investigate offense category.

Methods

KNN

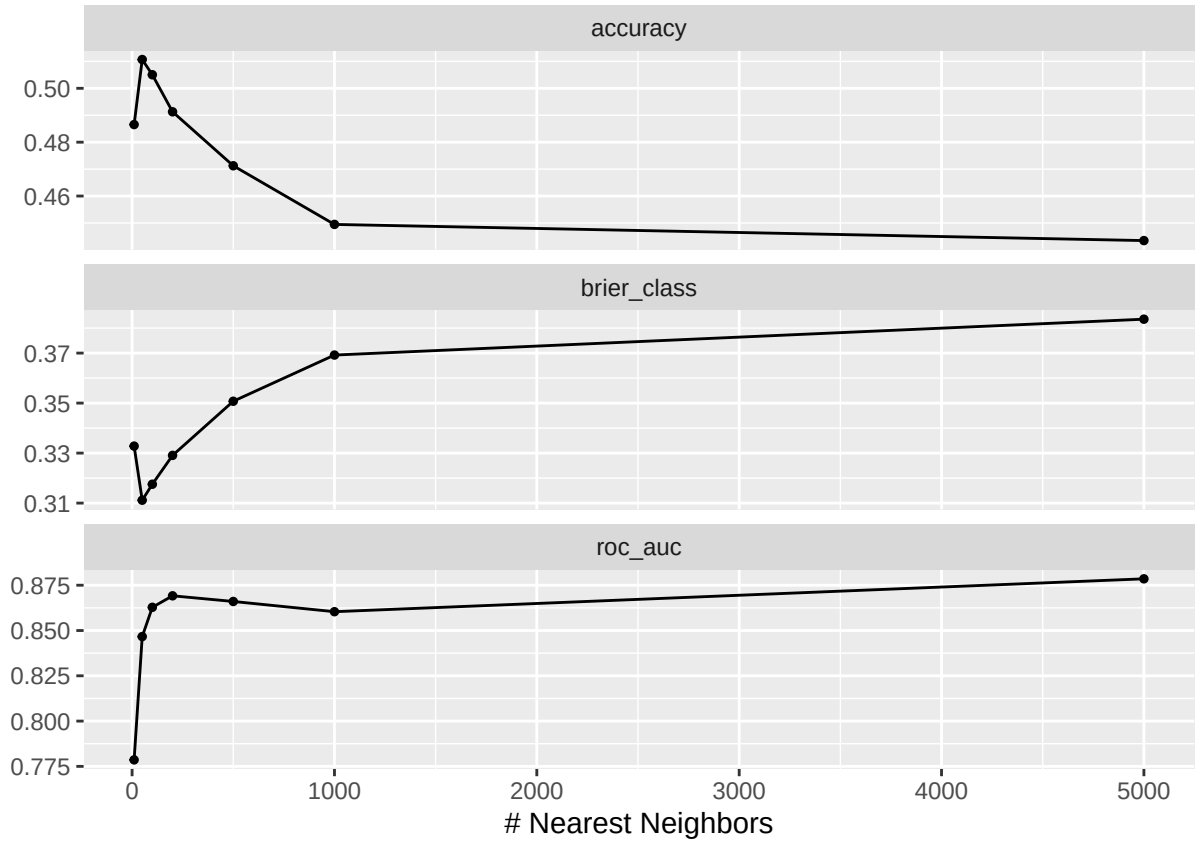
Within the criminal justice system, it is expected that some patterns will exist for sentencing type, prior offenses, counts and severity of a charge (i.e., was it armed or aggravated), among other variables. However, there are some patterns or correlations that would ideally not exist, in a perfect criminal justice system (such as offense category hypothetically being correlated with number of tattoos). With these seemingly conflicting ideas in mind, this section of analysis focuses on using K-nearest neighbors (KNN) to attempt to predict offense category within the data.

As such, in order to answer research question regarding prediction of offense category, a KNN model was chosen. KNN stands for k-nearest neighbor, which is an algorithm that cycles through each point in a dataset, and identifies the k points that are closest to that point. This k value is adjustable and helps determine the classifier itself, so tuning k is critical for optimizing classification performance. For this analysis, offenses were grouped into seventeen categories: controlled substance possession without a prescription, burglary, murder, armed robbery, theft (identity or property), battery, sexual assault, forgery, kidnapping, illegal firearm/weapon/handgun use or possession, harassment, bribery, drug/meth manufacturing, vehicular hijacking/theft, DUI, child porn, obstructing justice, home invasion, and other. It is important to note that these categories were created by reading through thousands of different individual charges, all unique and specific to each person. As such, these categories were the result of wide generalizations, and in no way serve as definite classifiers of charge or individual offense. With that said, the eighteen categories (besides other) captured a little more than 90 percent of all observations, meaning the vast majority of charges fell into one of the categories listed above.

Because this dataset is so large (even with an 80-20 training-test split), ten fold cross validation (which is the process of partitioning the entire data set in ten different folds, where each fold is the test set one time and the other nine folds are used for training) was extremely time-expensive (about 2 hours and 45 minutes to complete). An additional drawback to this approach is the inability to perform variable selection ahead of time, as can be done using best subset selection in various types of logistic regression.

The first step of building the model was to load the data, clean the predictors by ensuring all variables were correctly factors or numeric, and omit rows with NA values after transforming the necessary columns. Afterwards, the data was randomly split into an 80-20 training-test set. Then, a range of k-values to be used in tuning was selected. Because time-cost was so expensive for adding additional k-values to be trained, only seven were selected: 10, 50, 100, 200, 500, 1000, and 5000. The wide variety of k values was selected due to the size of the data, and anticipated complexity. A larger range allows for comparison of model performance with a greater variety of k-values selected for KNN. After this, ten folds for cross validation were fit, and the recipe of model fit (`offense_category` explained by all other variables) as well as the `tidymodels` KNN spec was set for KNN classification. Finally, cross-validation was performed to find the k-value with the highest accuracy in prediction.

After reading in the cross-validation results, we can plot the variation in accuracy for each k value examined.



Immediately, it is clear that predictive accuracy decreases substantially as k increases, with the best accuracy occurring when $k = 50$. Using this best model obtained through cross-validation, the next step of the analysis was to finalize the best model, and perform predictions on the test set.

Table 1: KNN Classification Recall

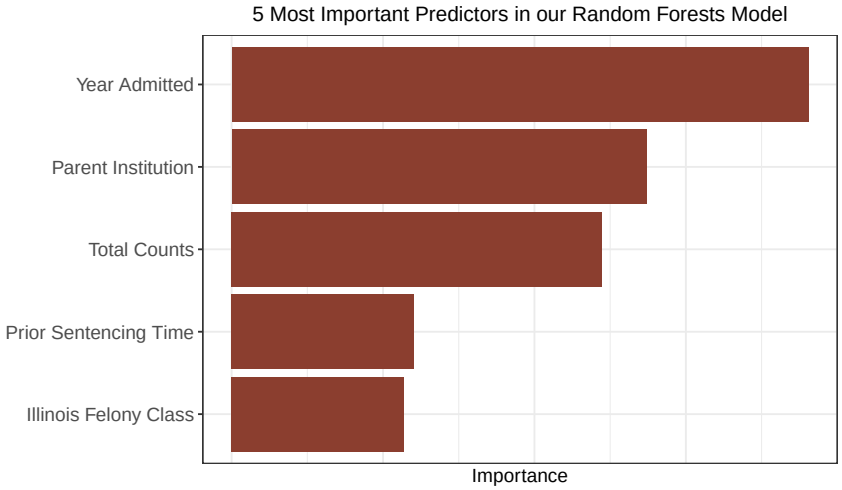
Truth	total	correct	prop_correct
BATTERY	1234	926	0.750
BURGLARY	877	500	0.570
DRUG MANUFACTURE	485	201	0.414
DUI	575	213	0.370
FORGERY	68	0	0.000
HOME INVASION	93	7	0.075
ILL. CONTR. SUBST. POSS	1536	1181	0.769
ILLEGAL WEAPON USE/POSS	1110	818	0.737
ILLEGAL/CHILD PORN	51	0	0.000
KIDNAPPING	11	1	0.091
MURDER	1359	1242	0.914
OBSTRUCTING JUSTICE	112	0	0.000
OTHER	1706	854	0.501
ROBBERY	916	555	0.606
SEXUAL OFFENSE	960	691	0.720
THEFT	582	185	0.318
VEHICULAR HIJACKING/THEFT	194	17	0.088

These test set prediction results indicate a clear trend: total proportion correctly predicted is significantly

higher in categories with more individuals. Logically, this makes sense because the model attempts to maximize accuracy by predicting the best within categories with more samples. For example, individuals marked as having a "sexual offense" were correctly predicted 72 percent of the time, those categorized for murder were correctly predicted 91.4 percent of the time, and those charged with illegal possession of a controlled substance were correctly classified 76.9 percent of the time. These classification rates are impressive, considering the number of categories that each individual could be predicted as, indicating similarity of predictor values (means) for individuals in these categories. The only categories with relatively significant sample sizes and poor predictive accuracy were burglary, theft, and drug manufacturing. A potential explanation for this poorer predictive performance would be greater variation in predictor variable values for these "lower-level" crimes (which might be classified as lower-level felonies or misdemeanors than murder or a sexual crime). Overall, this KNN model performed relatively well, achieving an overall accuracy of roughly 62.3 percent. Given that randomly classifying each individual into an offense category would result in an expected accuracy around six percent (1/17), this is a significant improvement, and likely indicates predictor mean similarity between individuals charged with the same categories of crimes (particularly for more severe charges).

Random Forest

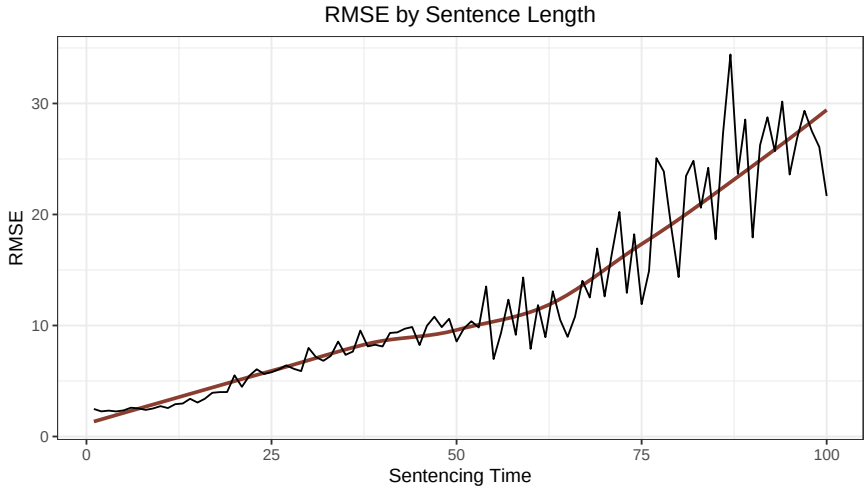
We made some decisions for this model, filtering out people with the hair color "Sandy" as there were very few observations (5). We also decided to not predict for people who were sentenced to zero years, as there were too many data entry issues around these rows which made no sense. This still left us with a large dataset at 59388 observations, where we decided random foresting would be a good model to run our data through for predicting a prisoner's sentencing time in years.The use of random forests allows for several trees to be fit with random subsets of our data and then combines results to better assess predictions. We found random foresting to be a good option because of its built-in out-of-bag feature which allows for us to run the model on all predictors and not worry about cross-validation. We didn't want the correlated trees you would expect in boosting, random forests don't care for the most important features it learns from tree to tree. It builds its trees on random subsets of predictors each time, while boosting begins to "boost" predictions by further finding optimal splits with the important features it learns as it goes. We went with 500 trees, and 4 predictors for each tree because it's common practice to choose the square root of how many predictors you have which is $\sqrt{20} = 4.47$. It's also important we note that we kept a similar training split for this model at 70:30 ($\sim$ 17 41,200:17,600). Once this model was fit, we were eager to look into the feature importance plot to see if it was any different to our previous models.



The predictor with the most importance in this case is the year the prisoner was admitted, which instantly tells us based off our model that the time period they were sentenced plays the biggest role in sentence

length. It's interesting to think about and hypothesize why this may be; Has the way certain charges have been punished drastically changed throughout the years? Another predictor that stood out from this variable importance plot was total counts, which makes sense looking at it from a judge's point of view. It makes sense that a person with more charges historically will often get a longer sentence.

We checked model performance first by looking at the out-of-bag RMSE which measures prediction error on the training split by evaluating each tree on the subset of data not used during its training. We got a value of 12.11, which up to this point was the best, but still not ideal at first glance. We also looked at the test RMSE which checks the model performance on the test split and we found a very similar value at 12.18, meaning our model generalized well. We took a further step to look into why our predictions were off that much by grouping by sentencing length predicted and then calculating RMSE per sentencing length. The motivation here was that we feared our model was no good if we were off by 12 years for people with sentencing lengths of 1-5 years. As you can see below, we were very happy to find that this was not the case. We don't get into that 12 RMSE region until just about 50 predicted sentencing years. With such a broad range from 1-100 on our response variable, I would actually consider our RMSE not that bad with all things considered.



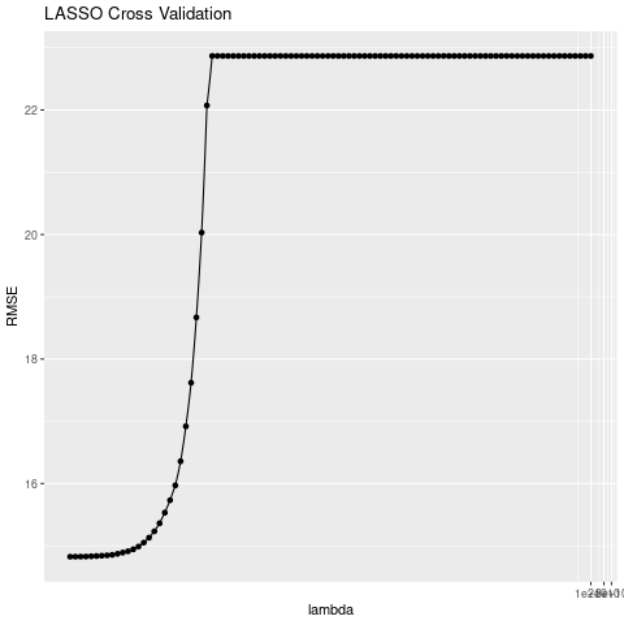
Linear Regression

To explore the relationship between prison sentence and our other variables, we chose to analyze a variety of linear models for their superior interpretability over non-parametric models. Because we were interested in identifying the most important variables for prison sentence prediction, we chose not to fit either an ordinary least squares regression model or a Ridge regression model. Both of these models would have assigned a non-zero coefficient to every predictor in the model, requiring every variable to be incorporated into the interpretation of a given prison sentence. We suspected that not every variable would be important so we focused on models that have a component of feature selection, including LASSO and subset selection models.

The first model we fit was the LASSO model, which selects optimal predictor coefficients by minimizing the combination of residual sum of squares and a penalty shrinking the size of those coefficients according to the formula below.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} \ + \lambda\sum_{j=1}^{p}|\beta_j|$$

The penalty parameter λ was chosen according to 10 fold cross validation using 100 possible values of λ between 0.01 and 10×10^10. As shown on the lambda vs. RMSE graph, cross validation selected $\lambda = 0.01$.



Unfortunately, due to the number of categorical predictors in our dataset, LASSO was not as helpful as we would have liked. Categorical predictors must be reference encoded for LASSO to compute coefficients, turning our 20 predictors into 88. Of these 88 terms, LASSO only selected out the following five variables:

term	estimate
height	0
hair_Salt.and.Pepper	0
sex_Male	0
race_Bi.Racial	0
race_White	0

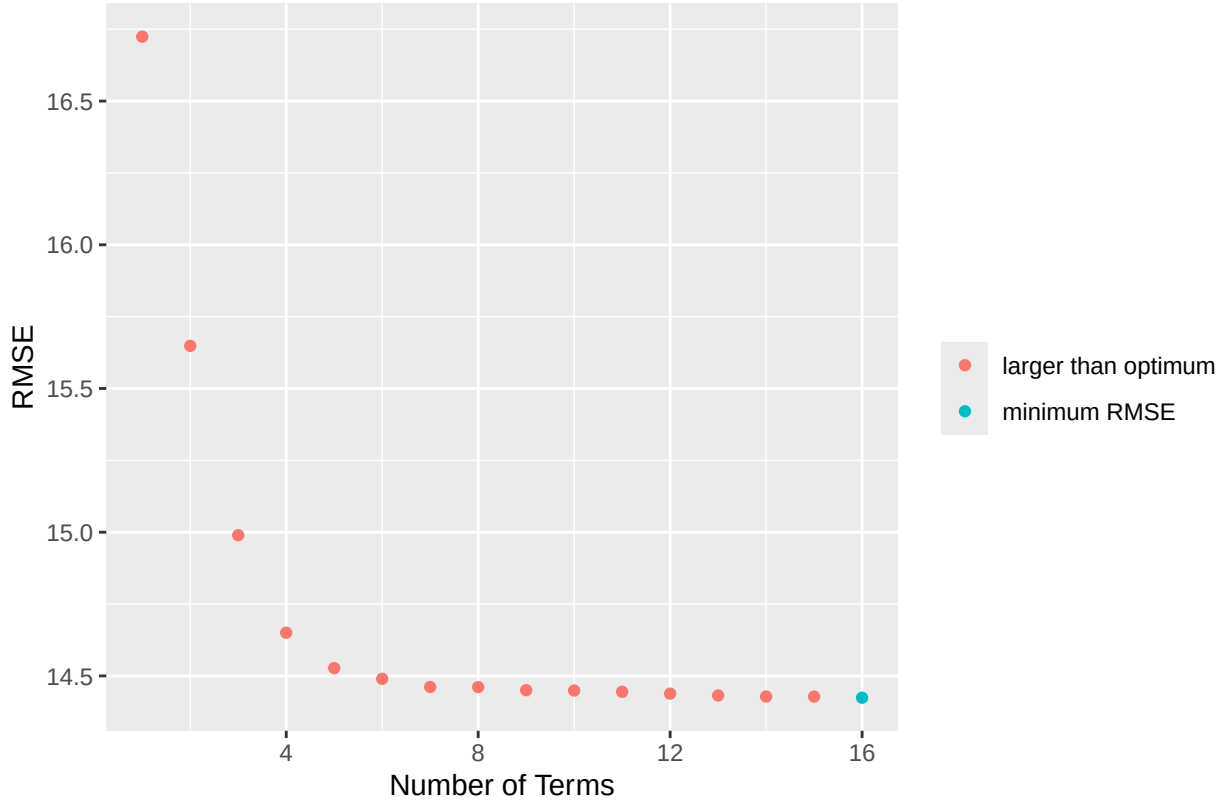
This tells us that height is not a significant predictor of prison sentence length. The other four variables, however, are individual categories from hair color, sex, and race respectively. These do not encapsulate whole categorical variables, so this information is not meaningfully interpretable, as it is not possible to only drop some categories from a categorical variable in linear regression.

Since LASSO did not produce as interprative a model as we had hoped, we moved on to subset selection. We elected not to perform best subset selection, as this method of subset selection requires fitting a model with every possible combination of variables and we decided that fitting approximately 2^20 models would be too computationally expensive. Instead we fit both forward and backward subset selection. Using the metric of adjusted R^2, forward and backward subset selection selected the same combination of variables, excluding the number of scars an individual has, their sex, number of other marks, and height, and including the following variables in order of importance.

step	variable	step (cont)	variable (cont)
1	year_adm	9	armed
2	class	10	attempted
3	parent_institution	11	race
4	total_counts	12	weight
5	offense_category	13	month_adm
6	aggravated	14	time_sentenced_prior
7	appx_age	15	hair

step	variable	step (cont)	variable (cont)
8	eyes	16	num_tattoos

Although both forward and backward subset selection indicated that the best model includes the above sixteen categories, an exploration indicates that smaller models can be fit with a minimal increase in RMSE, when the model is fit with a number of categories between five and fifteen. The graph below shows the RMSE for a model fit with the most important variable, followed by a model with the two most important variables, and so on up to sixteen variables, according to the ranking provided by the forward subset shown in the table above. This indicates that if interpretability is a priority, a model can be fit with five or six terms that will have nearly identical prediction results to the model with sixteen, but fewer variables would need to be explained.



Unfortunately, the best model fit according to subset selection has a RMSE of just under 14.5. This means that on average the linear model is off in its prediction of prison sentence length by about 14.5 years. Given that this data has a range of sentence lengths between 0 and 100 years, this is not an acceptable amount of error. The length of time an individual spends in prison has a substantial impact on their life, so we do not feel confident in applying the decisions made in this model to predictions made about the real world. Analysis of this model can help reveal the most important factors in understanding prison sentence length, but this model should not be used for prediction. The error from this model indicates that the relationship between prison sentence length and the predictors we investigated is likely not linear, so non-parametric models should be investigated instead.

GAM

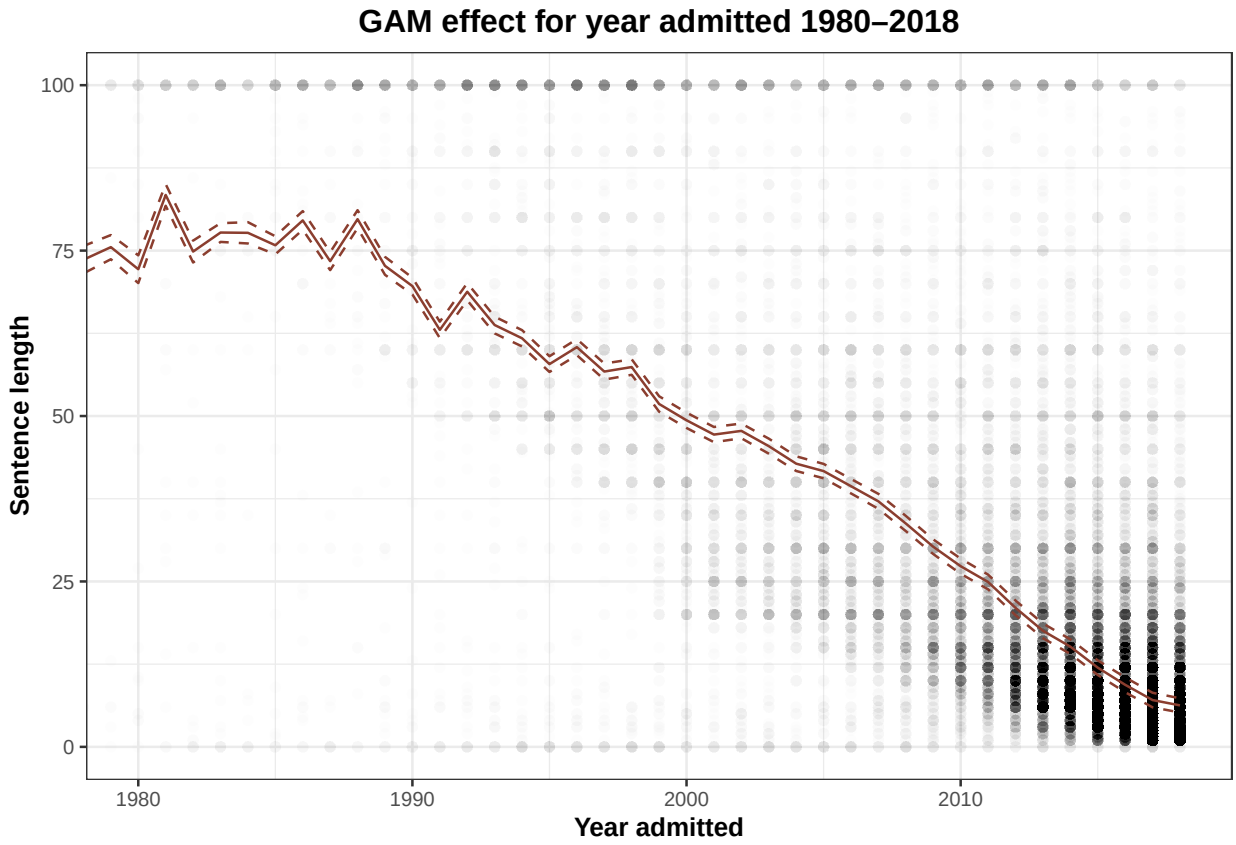
To see how accurately we can predict sentencing length, we fit a generalized additive model (GAM) with sentence length as the response. A GAM is an extension of linear regression that replaces straight-line effects with smooth functions of each predictor, allowing us to capture nonlinear relationships while retaining an additive and interpretable structure.

We used the merged Illinois DOC dataset and restricted our analysis to cases with complete information on sentencing length, basic demographics, physical descriptors, and key legal variables. We removed identifiers (ID, names) and the death_sentence indicator, and recoded all categorical variables (race, sex, hair color, eye color, felony class, offense category, parent institution, attempted/aggravated/armed) as factors. Numeric predictors in the GAM included number of scars, tattoos, and other marks, weight, height, year and month admitted, total number of counts, prior sentencing time, and approximate age at admission.

We randomly split the data into 80% training and 20% test sets. On the training set, we specified a linear regression model with natural spline terms for all numeric predictors and dummy variables for all categorical predictors. Degrees of freedom was treated as a tuning parameter and selected using 10-fold cross-validation to balance flexibility and overfitting. After choosing the optimal degrees of freedom, we refit the final GAM on the full training data and then evaluated performance on the held-out test set.

On the test set, the GAM achieved an $RMSE$ of approximately 12.2 and an R^2 of about 0.71. This means that the average prediction error is around 12 years, and the model explains roughly 71% of the variability in sentencing length on new cases. This indicates that the observed legal and physical variables capture a substantial portion of the structure in sentencing, while still leaving meaningful variation that is likely due to other factors not present in the dataset.

To understand which predictors had the most nonlinear effects, we generated partial effect plots from the fitted GAM.



In the figure above, we show the estimated nonlinear effect of year admitted on sentencing length, holding all other predictors constant. GAMs are easy to interpret as we can clearly see that as the years went on sentencing length has gone down holding all other predictors constant.

Conclusion

Looking at our models results compared to each other suggest that the predictors have a non-linear relationship with sentencing length. This is shown by the models respective $RMSE$ values, with the random forest performing the best followed by the GAM and then our linear regression model. When determining the best predictors for sentencing length we found that demographic data like race, sex, and weight are less helpful predictors while the best were year admitted, parent institution, total counts charged, prior sentencing time, and criminal class, and offense category. With our various R^2 values suggesting that while most of the sentencing length is explained by the predictors there is still unexplained variability that our data did not capture. In terms of offense category, the high predictive success rate for predicting higher-level charge category indicates increased similarity in the mean of all predictor variables for those individuals. This allowed the model to better generalize, compared to lower-level charges where larger variation in all predictors existed.

References

Illinois Compiled Statutes, Chapter 730, Article 4.5 General Sentencing Provisions. Justia, 2010, law.justia.com/codes/illinois/2010/chapter730/073000050HCh_V_Art_4_5.html.

"Time on Death Row." Death Penalty Information Center, deathpenaltyinfo.org/death-row/death-row-time-on-death-row.

Fisher, David J. Illinois DOC Labeled Faces Dataset. Kaggle, 2019, www.kaggle.com/datasets/davidjfisher/illinois-doc-labeled-faces-dataset/data.