paper

Group 5

2025-12-04

Abstract

have a summary of the paper

Background

here's our data and why we chose it here's where the data is from here are our research questions

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. (Talk about data cleaning and the work put into making one final dataframe)

Research Questions

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

Methods

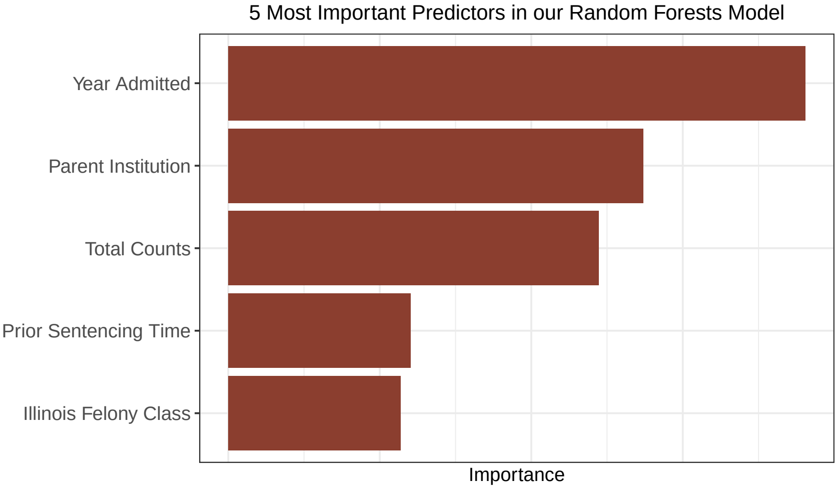
KNN

- basic explanation of model In order to answer our second two research questions regarding prediction of offense category, as well as determining which predictors have the strongest effect on offense category, we decided to fit a KNN model. 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. I grouped offenses/charges into 18 categories: controlled subst poss w/out 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.

- results Because our dataset is so large (even with a 70-30 training-test split), I expect the cross-validation to take a very long time to run, and I have temporarily commented out that code. I may need to make modifications to the training-test split, as well as the number of possible k-values, to increase efficiency. Once complete, I will create a confusion matrix to assess the model (with the optimal k) on the test set.

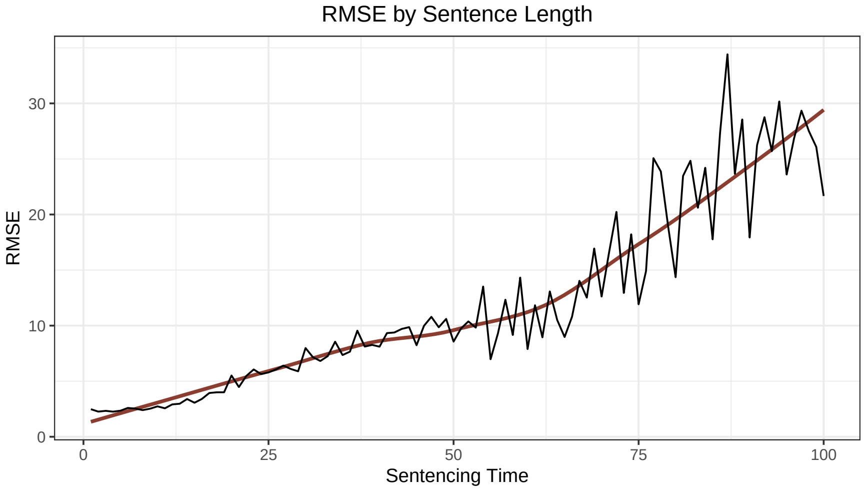
Random Forest

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 just sometimes 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 prisoners sentencing time in years. The use of random forests allows for several trees to be fit with random susbsets 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 is 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. Since we have 20 predictors and 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 judges 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 it was 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 dont 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

- basic explanation of model quantitative response: current sentence

if the relationship is approximately linear, least squares will have low bias if we have many data points, n » p, also have low variance – do have many data points limitations to linear model: linearity is an approximation use LASSO, forward, backward to obtain more interpretable model, drop predictors not associated with the response

not doing best subset because there are over 20 predictors and fitting approximately $2^2$0 models would be too computationally expensive; not doing principle component regression because interpretability again; ridge doesn't eliminate does not eliminate predictors and I care about interpretation

Ridge regression: RSS + $\lambda \sum_{j=1}^{p} \beta_j^2$ where λ is the tuning parameter - variable selection - LASSO - Ridge - results

GAM

We decided to use fit a generalized additive model(GAM) to see how well it can predict sentencing length and what predictors have the strongest effect on sentencing length. GAMs are an extension of linear regression that are used to show relationships between the predictors that can be better explained non-linearly while being easily interpretable. We then used 10 fold Cross-Validation to evaluate the model

When fitting a model that includes physical descriptors and a model without any of the physical descriptors. The model that was able to best predict sentencing length was:

This model shows that:

- results ## Conclusion here's why we pick x model over y model

References

kaggle data set