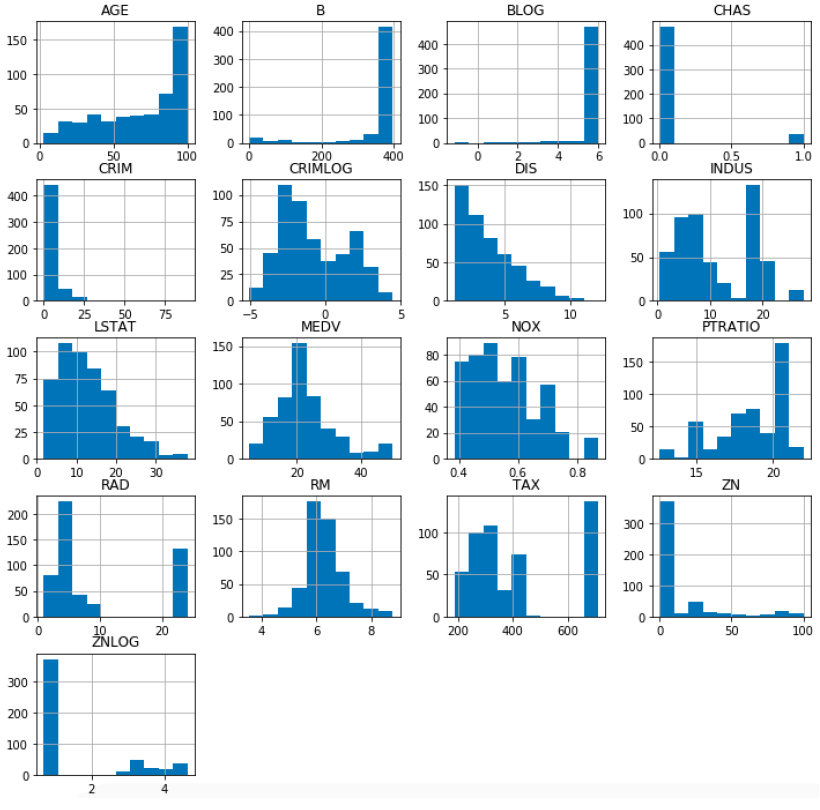
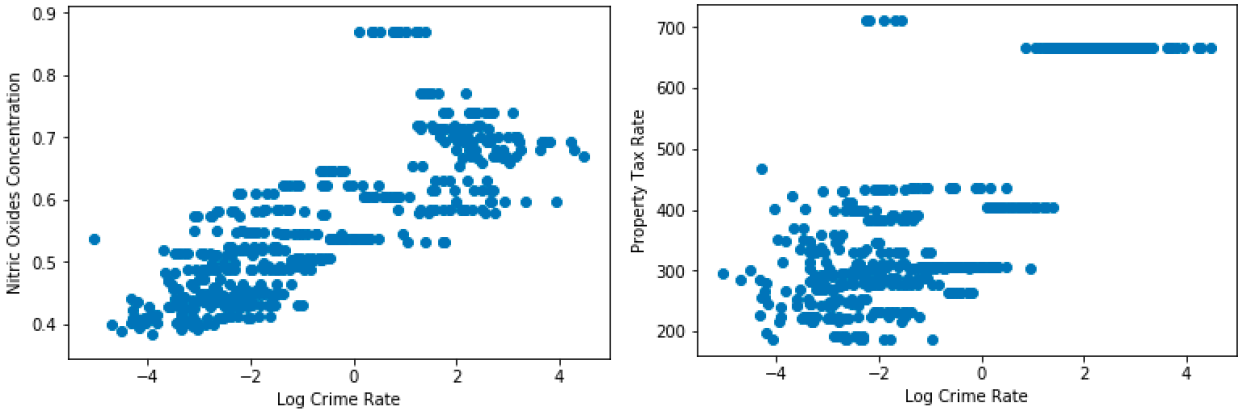
NAME: Galen Byrd
DATE: 04/11/19
Homework: 07
STAT/CS 387

Problem 1

The dataset contains 506 observations of 14 columns with no missing cells regarding housing valuations in Boston from 1993. I started by looking at the distributions of each column (line 21-22), specifically their histograms and summary statistics including count, mean, standard deviation, min, max, and the quartiles. As we can see, many of the variables, like B, ZN, CHAS, and CRIM are heavily skewed, while others like MEDV and RM are normally distributed. For the skewed variables, I tried putting them on a log scale to see if their distributions would be on a smaller scale with higher variance, which worked for CRIM but not for B and ZN.



I then looked at the Pearson correlation coefficients of the data frame (line 23). There were not many significant correlations, but the following correlations (not including CRIMLOG) were higher than 0.7: NOX and INDUS, NOX and AGE, NOX and DIS, DIS and INDUS, TAX and INDUS, MEDV and LSTAT, MEDV and RM. There were no correlations of 0.8 and the only correlation larger than 0.9 is TAX and RAD, having the largest correlation of 0.91. As we will be looking at predicting the crime rate, I investigated this variable more. The log transformation highlights many correlations that do not exist with the non-log-transformed data. There are two correlations greater than 0.8, RAD and TAX, with INDUS and NOX larger than 0.7. This means that in



relation to predicting the crime rate, we would expect these variables to have the most relevance or predictive power in our model. Just based on this simple exploration I would say that the Charles river dummy variable CHAS is not very important, as it is not correlated with any other variables and has a large imbalance within the levels.

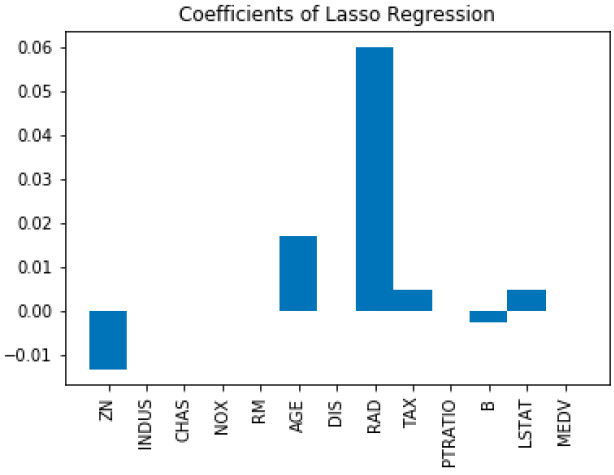
Problem 2

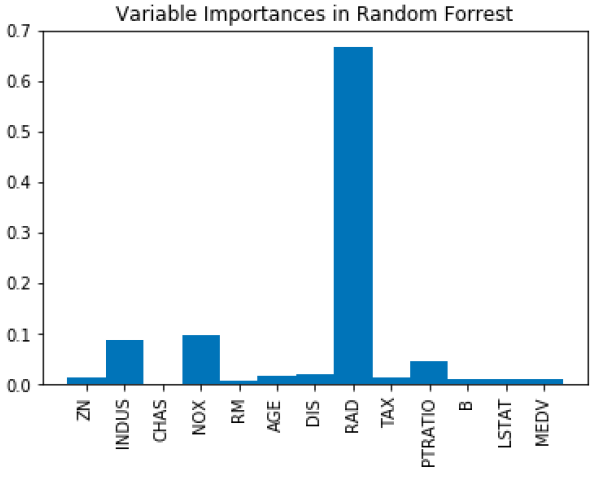
I proceed to build a Lasso Regression model and a Random Forrest as my linear and nonlinear methods respectively. I decided to implement a Lasso model because it seems like the best linear method as it deals with variable selection and is highly interpretable. I decided to implement a random forest as I have not built a model using this method before and in class Professor Bagrow says it is his preferred method for nonlinear methods. These two methods have been implemented in detailed functions in scikit-learn that make applying them to our dataset very easy. I simply build a base model with default parameters, then fit the model to our data (Lines 49-50 and 65-66). These methods both report the R^2 score to evaluate model performance and can be used when implementing cross-validation.

Before fitting the models I split the data into 70% training data and 30% testing data to evaluate model performance and overfitting. For Lasso regression, I was able to build a model with a test-R^2 of 0.819, and a train-R^2 of 0.837. My random forest built a model with a test-R^2 of 0.941, and a train-R^2 of 0.989. Based on this metric, we know that our nonlinear model is a much more accurate predictor of the crime rate as it is a much more flexible method.

Lasso Regression not only builds a linear model but also deals with feature selection in order to minimize the complexity of the model. In terms of interpretability, this means the columns of data that don't add any predictive power to the model will end with a coefficient of 0, and the largest absolute value of the coefficients is the variable with the most weight in the

model. This process is known as shrinkage, as the coefficients shrink to zero. While Random Forrest does not explicitly do feature selection like Lasso, we do get feature importance, which is somewhat similar but less interpretable. The variables that account for the most variance in, or are most highly correlated with the crime rate will be the more important variables, as we will be able to split the data more accurately along that dimension. The most important variable will likely be the first split in each tree as it will separate the data well. I made histograms of these quantities to visualize the impact on the model of each variable (lines 55-57, 69-71).





As we can see from the histograms, both of these methods rely heavily on the index of accessibility to radial highways. In addition, my prediction that CHAS was not going to be a relevant predictor was correct, as neither method relies heavily on this covariate. For the lasso, we can interpret this as increased access to radial highways leading to a higher probability of a high crime rate in this area. This might make sense, as increased access to highways typically means more people with more action and interaction than locations far from radial highways, but this is simply conjecture. I was surprised to find both models relying so heavily on this predictor and almost no others. While the random forest is less interpretable at this step, if we were simply concerned with prediction accuracy it would be the optimal method as it is much more flexible to different relationships within the data.

To check for overfitting of my models I used 10-fold cross-validation to estimate the actual predictive accuracy of each model. My true prediction accuracy for the lasso method was an R^2 of 0.776 (line 60) and a random forest with an R^2 of 0.942 (line 74). In terms of overfitting, we can see that our random forest is not overfit, as the cross-validated accuracy is very close to our estimated accuracy. The lasso model might be slightly overfit, as the cross-validated accuracy decreased slightly from our estimate. To avoid overfitting we wish to create a model where the predictive accuracy is as close to the true accuracy of the model, avoiding an

overly bias or overly variable model. This bias vs variance tradeoff is the most essential part of building a flexible model as we wish to allow for irreducible error in our model, but wish to account for all of the reducible error so we accurately predict unseen observations.

For predictive modeling, it is important to split data into testing and training sets to be able to evaluate performance on real predictions. If we did not take this step, and simply trained our models on 100% of the data, we would have no way of knowing the true prediction accuracy. Also, when it comes to evaluating predictive models it is important to evaluate performance based on the context, as opposed to simply the R^2. We might want to build a model that may be slightly less accurate, but more interpretable, based on the question being asked.

Another important aspect of model building as applied to these data are variable transformations. After I completed my data exploration, I attempted to build a model using the non-transformed crime rate but was extremely unsuccessful. I could not get an R^2 of above 0.5, and continually getting negative values, which meant the regression model was doing worse than simply predicting the mean value. Once I realized I needed to use the log transformation to more appropriately capture the spread in the variable, my models' prediction accuracy spiked dramatically.

Problem 3

Attempting to predict future crime levels, while beginning with good intention, can have massive ethical concerns. The risks associated with implementing these predictive models are currently far higher than the potential rewards, and with such a complex problem space we would not be able to say with much certainty if a drop in crime rate would be associated with the use of the predictive model. I have found that there are two types of models, one using demographic-based data to predict individuals that are more likely to commit crimes, and another using location-based data to predict crime hotspots.

Individual-based predictions can be heavily biased using demographic data that might, for example, predict that black Americans are significantly more likely to commit a crime than white Americans. This could lead to police targeting specific groups, who tend to be under-served, under-represented minorities. Giving this data back to the model confirming that it is making accurate predictions, will only reinforce these biasses and result in a positive feedback loop. Predicting that an individual who has not committed a crime is more likely to, simply based on their demographics can end up getting reasonable people caught up in a system that they may never be able to escape. This can lead to police making more inferences or assumptions about a situation than they already do that could be even more biased or misconstrued. The consequences of this are the police likely taking more serious action than is necessary, potentially resulting in injury or fatality if the situation escalates uncontrollably.

While location-based predictions can be less biased than individual-based, they still have their downfalls. Due to redlining and the institutional racism that has impacted the US its whole existence, many of our communities remain highly segregated with poorer communities having more minorities and higher crime rates. These models, while not designed to target these communities, end up doing so due to the biased nature of our reality and the data the model receives as input. As with the individual-based models, these too can fall into positive feedback loops under the assumption that the data collected for the model is a true unbiased estimate of crime in the area. If more patrol cars are sent to a targeted area with a high crime rate, they are likely to see more crime occurring as they spend more time there, leading the model to send more patrols to cover the area, leading to more crimes being reported.

While it might be seen as “good” that people are getting caught for committing crimes, we must also consider the severity of the offense. If predictive models are sending more patrols to areas with high rates of vandalism, drug or non-violent crime, these offenders will be caught more frequently. But if the resources of the local police department are focused on these offenses, there might be homicides occurring in other areas that are not being responded to as quickly or effectively. Assuming we are predicting violent offenses, even if a police team arrives

at the scene where a crime is about to be committed, there are so many variables and complexity to be sure that they could even have more of an impact than if they were on their previously traditional route. For example, there were officers on the scene of the Florida high school shooting, but they did not respond until more backup came and students had been evacuated. While these officers were not on scene due to a predictive model, this is an instance in which being present on the scene may not have impacted the outcome all that much.

Once these models are implemented, evaluating the models' prediction accuracy is yet another challenge. We can obtain the precision of our model, evaluating how many times it correctly sent resources to a person/location where a crime was about to be committed out of the total locations the model sent resources. This is helpful, but tells us nothing about the rest of the problem space, or if a crime was committed elsewhere and the model did not predict it. Of course, it is impossible to collect this data as we cannot have an observation for an event we do not know occurred. In addition, data can be riddled with confounders that we may not be able to account for. For example, if the crime rate drops is that due to the predictive model, or some alternate tactics being used? One predictive model, PredPol evaluates their models against hot-spot policing and targeted interventions, but these cannot account for false positives or false negatives.

Even if people agree on the fact that these models should be put into place, there are disagreements about the impacts they should have. Many feel that they should be used to prevent crime in the first place, stopping it before it can even happen, while others feel they should be used solely to catch individuals who have already committed a crime. For these different goals, we would then need varying evaluation metrics to account for the intentions behind the model.

When different non-profit organizations have run evaluations comparing districts using predictive models with those who have not yet adopted them, there was "no statistical evidence that crime was reduced more in the experimental districts than in the control districts." If crime reduction is the true goal of these models, we must acknowledge their shortcomings and adapt them in a way that serves their true purpose.

In theory, predictive policing is logical and would provide safety and peace of mind to law abiding citizens. As I have discussed, these models may have good intentions but have real negative impacts. Law-abiding citizens could get caught up in a system that does not care about them, and pre-existing biases in our society can be reinforced through positive feedback loops. While the potential benefits exist, we must operate under the constraints and assumptions of our reality and take concrete steps toward making our communities safer through education and unity.

Bibliography

Fitzpatrick, Dylan, et al. “Keeping Score: Predictive Analytics in Policing.” *Annual Reviews*, 7 Nov. 2018, www.annualreviews.org/doi/full/10.1146/annurev-criminol-011518-024534.

Hunt, Priscilla, et al. “Evaluation of the Shreveport Predictive Policing Experiment.” *RAND Corporation*, 1 July 2014, www.rand.org/pubs/research_reports/RR531.html.

Robinson, David, and Logan Koepke. “Stuck in a Pattern: Early Evidence on ‘Predictive Policing’ and Civil Rights.” *Upturn*, Aug. 2016, www.upturn.org/reports/2016/stuck-in-a-pattern/.

Shapiro, Aaron. “Reform Predictive Policing.” *Nature News*, Nature Publishing Group, 25 Jan. 2017, www.nature.com/news/reform-predictive-policing-1.21338.