STA2101 Project: Crime Analysis and Prediction

December 15, 2023

Willem Atack

1 Introduction

In every community, reducing crime would equate to safer communities with increased well-being, health, engagement and economic output. Understanding the root problems that lead to crime allows us to more effectively prevent it. It is well known that factors like education level, income, unemployment and family structure affect crime rates. However, communities have limited resources to allocate toward crime reduction, so understanding what factors are the most significant is vital. In this project, a statistical analysis will attempt to understand what environmental factors or demographics are the most significant predictors of crime.

Of course, not all crime is created equal, and strategies to combat non-violent and violent crime differ. Therefore, it will be tested whether or not different types of crime have different significant predictors. To illustrate why this could be useful, consider if the analysis found that characteristic X tends to be a significant predictor of violent crime, but not non-violent crime. If community ABC exhibits characteristic X, they could take on a more focused strategy of reducing violent crime.

The demographics and characteristics of communities evolve over time, so it would also be useful to build a predictive model which generalizes well to unseen data. Therefore, this report will also focus on building a predictive model which is able to accurately estimate the rates of different crimes on a holdout dataset. This would allow communities to effectively adapt their crime reduction strategies dependent on changes in their demographics.

Ultimately, we are trying to answer the following two research questions:

1. What are the most statistically significant predictors of crime, and do these vary for different types of crime?
2. What is the optimal way to build a predictive model which generalizes well to unseen data?

2 The Dataset

2.1 Raw Dataset Description

The dataset in question combines socio-economic data from the 1990 U.S. census, law enforcement data from a 1990 Law Enforcement Management survey, and FBI crime data from 1995. It has 2,215 instances of data representing the demographics and crime rates in different communities across the United States.

Each instance has 5 non-predictive features which identify the community. They have 124 predictive features. This includes 102 community-wide socio-economic data features from the U.S. census which includes, but is not limited to, information on population, age distribution, income distribution, education levels, unemployment rates, family structure data, race, immigration, housing, and homelessness. The other 22 predictive features are from the 1990 Law Enforcement survey, which aims to provide information on the community's policing capabilities. For example data such as the community's number of police cars, and law enforcement department budget are included. An exhaustive list is not included due to the number of features, though a link to the dataset is included in the Appendix. It is important to note that the data is unnormalized, and must be standardized to utilize many statistical tests.

Each instance has 18 possible response variables. These represent the absolute number, and the rate of 7 different types of crime, and 2 aggregations of these crime types. The types of crime are "murder", "rape", "robberies", "arsons", "assaults", "burglaries", "larcenies", and "autoTheft". The former four of these crime types are aggregated into "Violent Crimes", while the latter three are aggregated into "Non-Violent Crime" response variables. In the analysis, the rates of "Violent Crimes" and "Non-Violent Crimes" will be treated as the response variables in separate models.

All three of the original data sources collected this data to accelerate their own missions. The U.S. govern-

ment collects U.S. census data in order to better understand its communities, and how to benefit its population. For example, the data may be used for determining where to allocate anything from new roads to new supermarkets. The FBI collects crime statistics to understand the current status of crime across the U.S., so that they can strategize to improve these numbers. Similarly, law enforcement agencies collect data to understand where their resources are currently allocated. These three sources were merged into a single dataset and published publicly by statisticians who used the dataset to perform analysis on the state of crime in the U.S., similar to this project.

2.2 Discussion of Data Issues

Firstly, it is worth noting that many features are redundant, and can be eliminated before beginning the analysis. In many cases, a feature representing an absolute value, and another feature representing the same quantity as a rate are included. For instance, ”number of people living in areas classified as urban” and ”percentage of people living in areas classified as urban” are separate features, and one may be eliminated. There are many other subsets of features which seem likely to have very strong correlations with each other. For example, it is likely that including both ”percentage of people that have immigrated in the last 3 years” and ”percentage of people that have immigrated in the last 5 years” would add unnecessary model complexity. The large number of predictors, and the associated redundancy and multi co-linearity among these possible predictors is the most significant challenge of using this dataset. Using methods to reduce the number of covariates is necessary to make meaningful insights and predictions about the data.

The dataset also contains many missing values. Upon further inspection, it was found that most of these missing values came from the columns originally from the Law Enforcement Survey. Among the 22 covariates from this survey, 20 of them had more than 500 missing values (out of a total 2,215 instances). This is because the majority of the survey only took place in police departments with over 100 police officers. Therefore, including these features would introduce a large sampling bias where a large amount of features are only available for large communities. Hence, these 20 features were omitted. Among the 102 features from the U.S census, there were no missing values.

In terms of the potential targets, the number and rate of rapes had around 200 missing values, because at the time of collection police departments in the Midwest considered this number to ”controversial” [4]. Therefore, the analysis will omit the number of rapes from the total of ”Violent Crime”, as including these numbers would introduce location bias into the aggregate number of total violent crimes. One of the other target variables had more than 3 missing values. Since this makes up such a small portion of the dataset (around 0.1% of total instances), and affects the aggregation of total crime in these communities, these instances were removed.

There are some additional issues with the data. Firstly, there is a time delay between when the predictive variables were collected (1990) and when the target variables were collected (1995). Therefore, the predictors from five years prior must be used as a proxy for the predictors at the time the crime rate numbers were reported. This is not completely unreasonable since demographics take quite a while for significant changes to occur, but does introduce quite a bit of unaccounted noise into the model. Also, the dataset does not contain every predictive feature of crime. For example, the FBI notes that the number of visitors into a community correlates strongly with crime rates, however is not included in the set of predictors[4]. Also, as a result of the law enforcement survey missing a lot of values in the dataset, the state of law enforcement in communities is mostly unaccounted for. In addition, communities are varied, and are likely to have existing differing initiatives to reduce crime rates, which is not reflected in the data. Overall, we must be aware that since this is an observational dataset with a large variety of features, confounding variables are likely to exist. Therefore, we must be careful when making conclusions with this data.

2.3 Exploration of Sampling Bias

To determine significant predictors of crime, or to build an effective predictive model, it is important to ensure that there are not any large sampling biases. For example, if each community sampled in the dataset are large

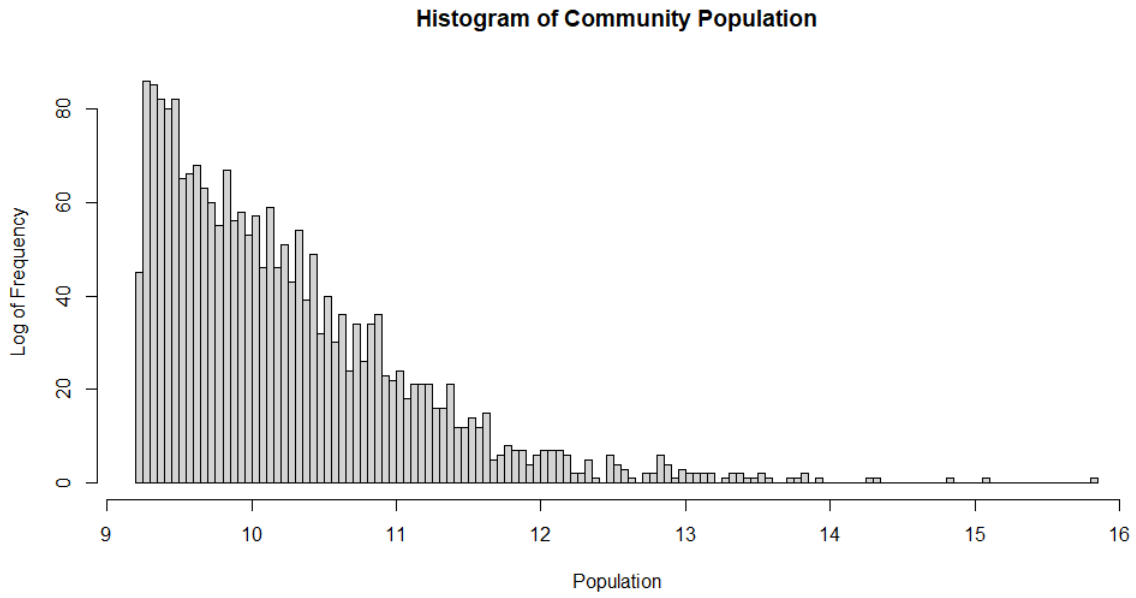

Figure 1: Histogram with the log-frequency of community populations

cities in California, it would not be meaningful to use a model built from this data to predict crime rates for a small town in Maine. Therefore, sampling bias was checked in three separate ways.

First, the number of communities in each state were checked to ensure that the instances are geographically diversified. In general this was not a significant issue, as 48/50 states have communities included in the dataset (there are no communities included in Vermont or Montana, which are small states). The states with the most communities are California (279), New Jersey (211), and Texas (162). Though New Jersey is likely over-represented, California and Texas are two of the largest states, so this makes sense. Notably, New York has just 46 communities in the dataset despite being the third largest state. Overall, the geographical distribution isn't perfect, but shouldn't cause an issue in the analysis. A table is included in the Appendix showing the full instance count by state.

Second, the size of the communities were explored to make sure that both small towns and large cities are included. The community with the largest population is New York City (7mm) while the smallest is Lake City (10k). The median population is 22k, while the mean is 53k. this shows that the dataset is mostly made up of small communities, with a small number of large cities skewing the mean population. This is in line with what is expected across the USA, so there does not seem to be a sampling bias problem here. See Figure 1 to see a histogram of the community populations.

Finally, the distribution of total violent crime rates was investigated to ensure that both areas with high and low crime rates are included. As viewed in the histogram in Figure 2, there appears to be a wide range of crime rates in the dataset, with a total range of 241 to 30,202 crimes/100k people, and quartiles of [3167, 6968]. There does not appear to be any sampling bias here.

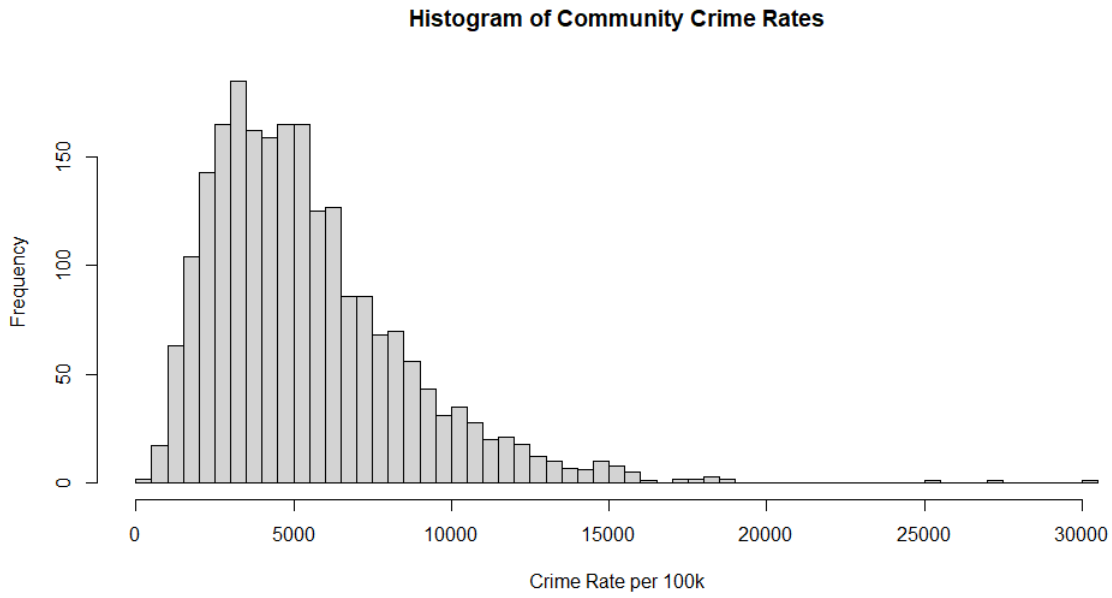

Figure 2: Histogram with the frequency of community total crime rates

3 Determining Statistically Significant Predictors

3.1 Methodology

As mentioned in Section 2.2, after removing features due to missing values, 104 features remained. Many of these features were clearly redundant and highly correlated. However, many of the covariates could reasonably be removed from the dataset when doing a manual pass over the data. For example, ”number of people living in rural areas” was removed, and ”percentage of people in rural areas” was kept. By using knowledge of demographic metrics (subject matter knowledge), an additional 63 potential covariates were eliminated as they were clearly redundant, reducing the size of the dataset to 41 features. For the rest of the report, this dataset with 41 features is referred to as the ”full dataset”, since this is the largest dataset with which models were built. It is crucial to note that this dataset was standardized, so that statistical tests would yield meaningful results.

Two tests were run to determine feature importance, each when predicting violent crime rates, and non-violent crime rates. First, after fitting a linear regression model to the dataset, the t-statistics were examined. In addition, a random forest feature importance test was performed, where the assumption that there is a linear relationship between covariates and predictors is removed. The random forest feature test measures feature importance based on GINI scores, The metric effectively measures how much worse the model will perform when each covariate is omitted (these scores are only significant on a relative basis, but a higher score indicates more importance) [2]. By running both types of feature importance tests, we may better determine if the most significant features are constant, irrespective of model choice.

When beginning the analysis, it was clear that when different samples were taken from the dataset, the feature importances varied greatly, when using both random forest and linear models. Therefore, to enable more stable results which could be meaningfully compared, non-parametric bootstrapping was used to compute these feature importance statistics. In particular, for each feature importance test on each dataset, 1000 datasets of 2215 instances were built by sampling from the original dataset with replacement. This results in 1000 sets of feature importance statistics, which are averaged. By using this non-parametric bootstrapping technique, the feature importance statistics remain stable, and thus more reliable conclusions can be made.

These tests were first run using the full dataset. However, with 41 features remaining, there was likely to be significant multi co-linearity remaining, as well as insignificant features. In addition, it is always preferable

to have simpler models, if both explain the same phenomenon, according to Occam's razor principle. Finally, reducing the size of the dataset should also later help in building a predictive model. Therefore, further model selection and reduction techniques were applied to reduce the size of the full dataset.

3.2 Model Selection

Various model selection techniques were applied to reduce the size of the full dataset, and the results were compared. Firstly, two criterion-based methods, the Akaike Information Criterion and the Bayesian Information Criterion were separately used to reduce the dataset size. These two methods were selected over significance-based methods, such as backward elimination, because criterion-based methods seek to optimize goodness-of-fit while penalizing the number of covariates, while significance-based methods consider only the more abstract notion of statistical significance and ignore goodness-of-fit. The goals of the report are to determine which are the important predictors of crime, and to build a predictive model, so optimizing goodness-of-fit is preferable.

Additionally, LASSO was used as a method for feature selection, as was done in this paper with similar goals [3]. By adding an $L1$-penalty term, some variables' coefficients converge to zero, and hence are eliminated from the model. Once these three techniques were applied, separate models were fit on the resulting datasets. The residual sum of squares and F-statistic were compared between the models. These results are summarized in Table 1.

Table 1: Model Comparison after Feature Selection

Model 1	Model 2	M1 - # Features	M2 - # Features	M1 - RSS (10^6)	M2 - RSS (10^6)	F-Statistic	p-value(F_stat)
Full Model	Reduced by AIC Step	41	19	298.9	299.7	0.256	0.998
Full Model	Reduced by BIC Step	41	14	298.9	301.6	0.698	0.874
Full Model	Reduced by LASSO	41	26	298.9	299.8	0.4178	0.9747
Reduced by AIC Step	Reduced by BIC Step	19	14	299.7	301.6	2.6658	0.021
Reduced by AIC Step	Reduced by LASSO	19	26	299.7	299.8	n/a	n/a

Rows 1-3 of Table 1 show the results after applying each of the three feature selection techniques. Using BIC resulted in a model with the fewest remaining features (14), but had the highest residual sum of squares (RSS). Meanwhile, using AIC resulted in a model with fewer features and a lower RSS than using LASSO, so clearly using the dataset resultant from AIC is preferred to the resulting dataset from using LASSO. It is worth noting that in all three of these cases, the F-statistic is quite low, representing that each of the reduced models do not represent a significantly different relationship than the full model. We also fail to reject the null hypothesis that the difference in fit between the two models is due to random chance, as seen by the p-values.

Next, to compare the difference between the resulting models from AIC and BIC, the F-statistic was computed between each of these two models. As shown in Row 4 of Table 1, the F-statistic is significant, as displayed by a p-value of 0.021. This means significant information is likely lost when dropping the five additional features (the BIC model had 14 features and the AIC model had 19). Therefore, the dataset resulting from the AIC reduction is preferred. Row 5 shows that the model resulting from the AIC reduction is clearly preferable over the model resulting from the LASSO reduction. Therefore, the model resulting from AIC will be used moving forward in the project, and is now referred to as the "reduced dataset".

Using this new reduced dataset, the additional step of fitting a LASSO regression to the dataset was applied, to determine if this would enable any further reduction. However, all the covariates remained non-zero when the fit was done. Finally, with this reduced dataset, the correlation matrix was inspected, as shown in Figure 3. Any covariate with a correlation over 0.8 was considered a candidate to be dropped from the model. However, prior to dropping any of these candidates it was tested if this would reduce the goodness-of-fit of the model. Though there were three pairs of candidate covariates to be dropped, it was found that the F-statistic was statistically significant when dropping any of these candidates from the model, and the RSS increased significantly. This indicates a loss in goodness-of-fit, so none of the candidates were dropped.

This whole model selection process was repeated for models fitted to violent crime rates, and non-violent crime rates, though the results were the same in each case. In summary, through this model reduction process, we

reduced the full dataset of 41 features, to a reduced dataset with 19 features. Though different techniques were applied, the AIC criterion reduction method proved to be the best at minimizing the number of covariates while optimizing goodness-of-fit.

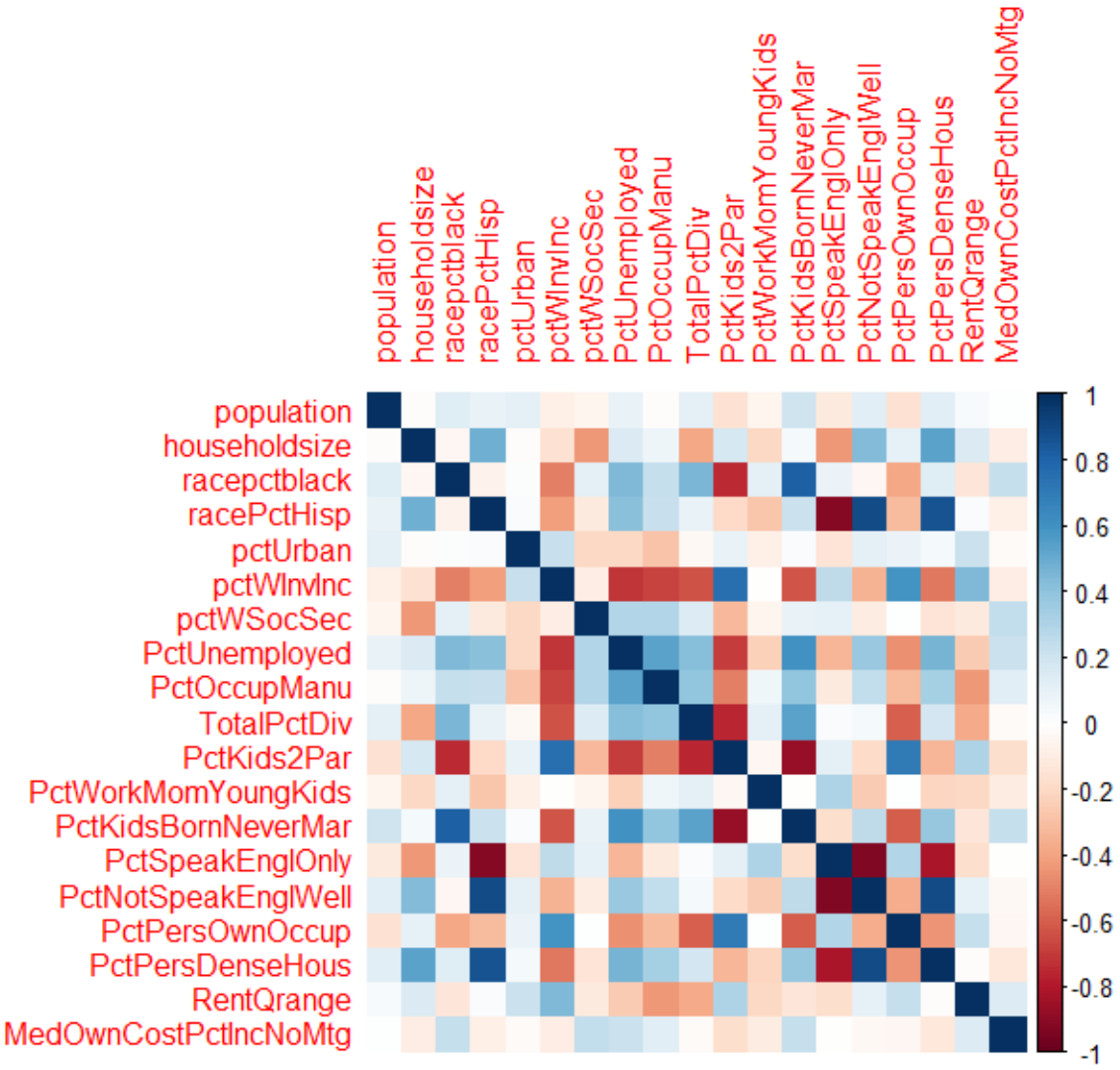

Figure 3: Correlation matrix of reduced dataset

3.3 Conclusions

In Section 3.1, we discussed the methods used to measure feature importance, and in Section 3.2, the process for reducing the model complexity was described. In order to test the stability of results, the feature importances were computed with both the full dataset, and the reduced dataset. The same covariate can have different effects on the model dependent on what other covariates are present in the model. If the most important predictors are the same, regardless of if the full or reduced dataset is used, we can be more confident that we have stable and meaningful results for estimating the most significant predictors. Remarkably, for both violent crimes and non-violent crimes, and when for generating feature importances using linear models, or random forest, the top five most significant predictors remain the same when both the full and reduced datasets were used (albeit in different orders amongst these top fives).

Figure 4 shows the top 5 most significant predictors of non-violent crime using linear models, while Figure 5 does so for violent crime. Meanwhile Figure 6 and Figure 7 do the same when generating feature importance with Random Forest. The full list of feature importances is in the Appendix.

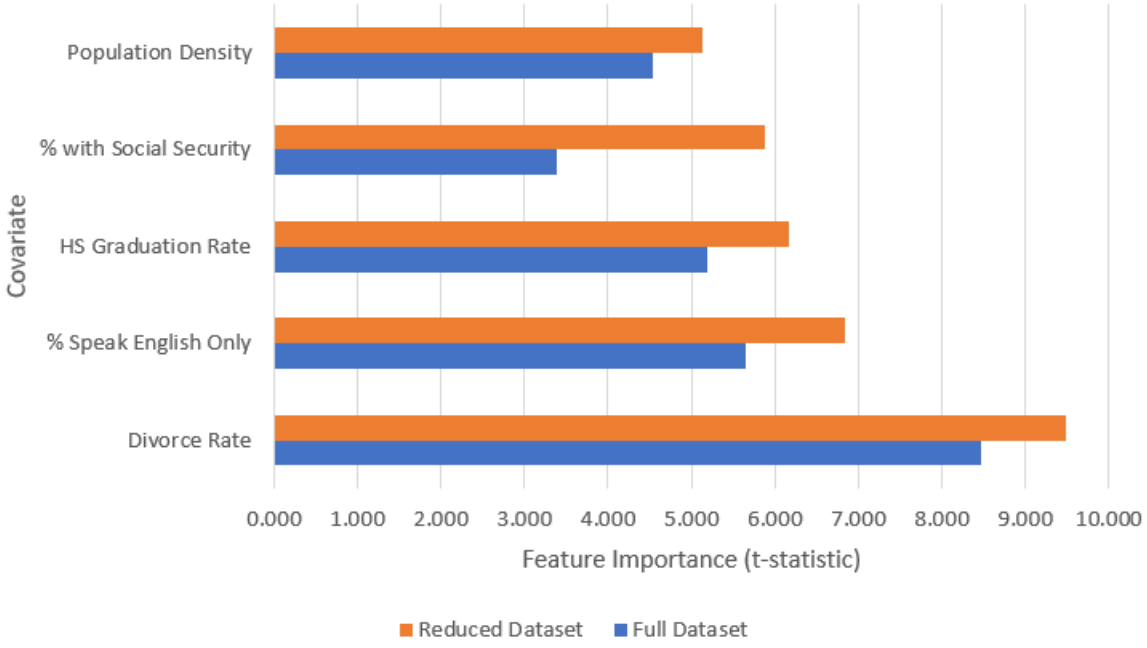

Figure 4: Feature Importance - Linear Regression Model on Non-Violent Crimes Rates

Figure 5: Feature Importance - Linear Regression Model on Violent Crimes Rates

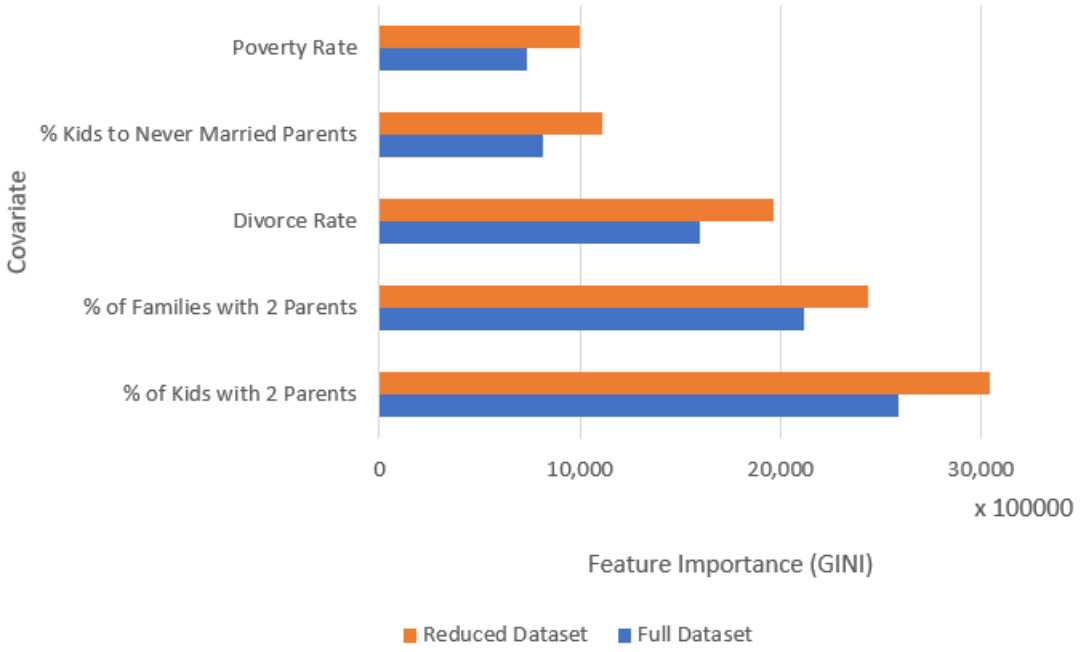

Figure 6: Feature Importance - Random Forest Model on Non-Violent Crimes Rates

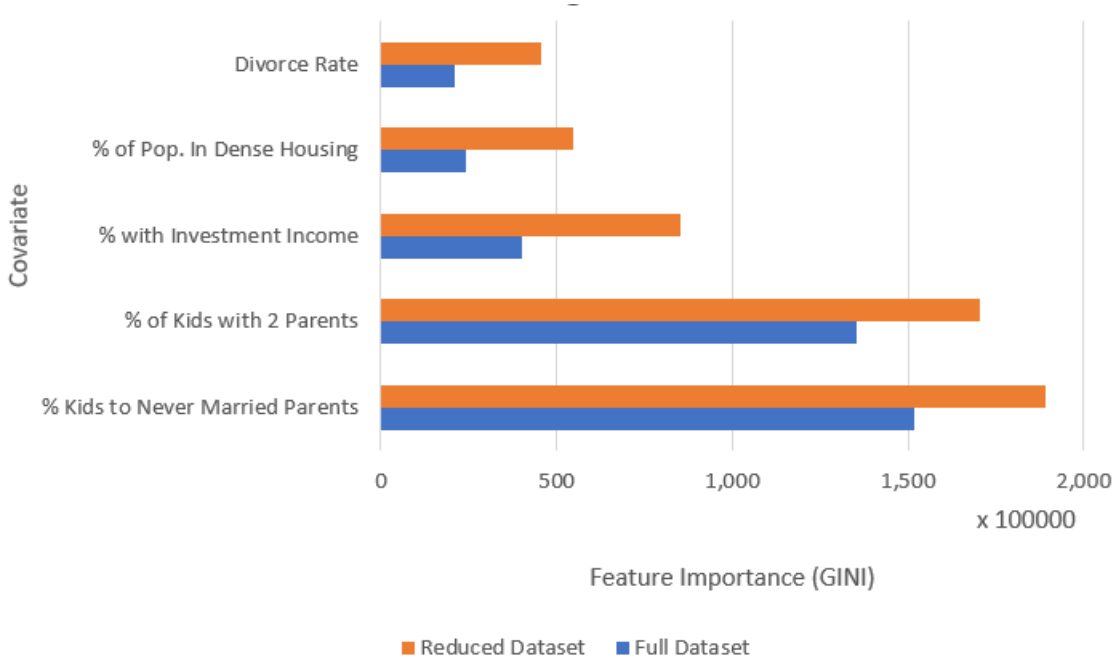

Figure 7: Feature Importance - Random Forest Model on Violent Crimes Rates

Looking at these results we notice a few things. Firstly, the most significant predictors of violent and non-violent crime appear to be different. Though there is some overlap, in both the linear and the random forest case, the list of the top 5 most significant predictors are different. Notably, both include family structure type variables, such as divorce rate. However, non-violent crime correlates stronger with some variables like the High School graduation rate, while violent crime tends to have a stronger correlation the covariates such as the population, and the proportion of the population in dense housing (defined by houses with twice as many people as bedrooms). This suggests that different strategies may be suitable to reduce different types of crimes.

We also note that the set of most important predictors depends on whether a linear or random forest model was used. Again, there was some overlap between the most important features when the two different models were used, but the results were not identical. Therefore, we can conclude that the most important features depend on model choice. Hence, we may not be able to make any absolute conclusions surrounding the most important features. Rather, we conclude that there are likely differences in the most significant features when we predict violent or non-violent crime, and these features also depend on model choice.

Finally, as a sanity check, the confidence intervals for the top five most significant predictors of non-violent and violent crime were plotted (using the results of the linear feature importance test). These were generated using non-parametric bootstrapping. Note that these confidence intervals come from the reduced models, as opposed to the full dataset models. This is shown in Figure 8 and Figure 9.

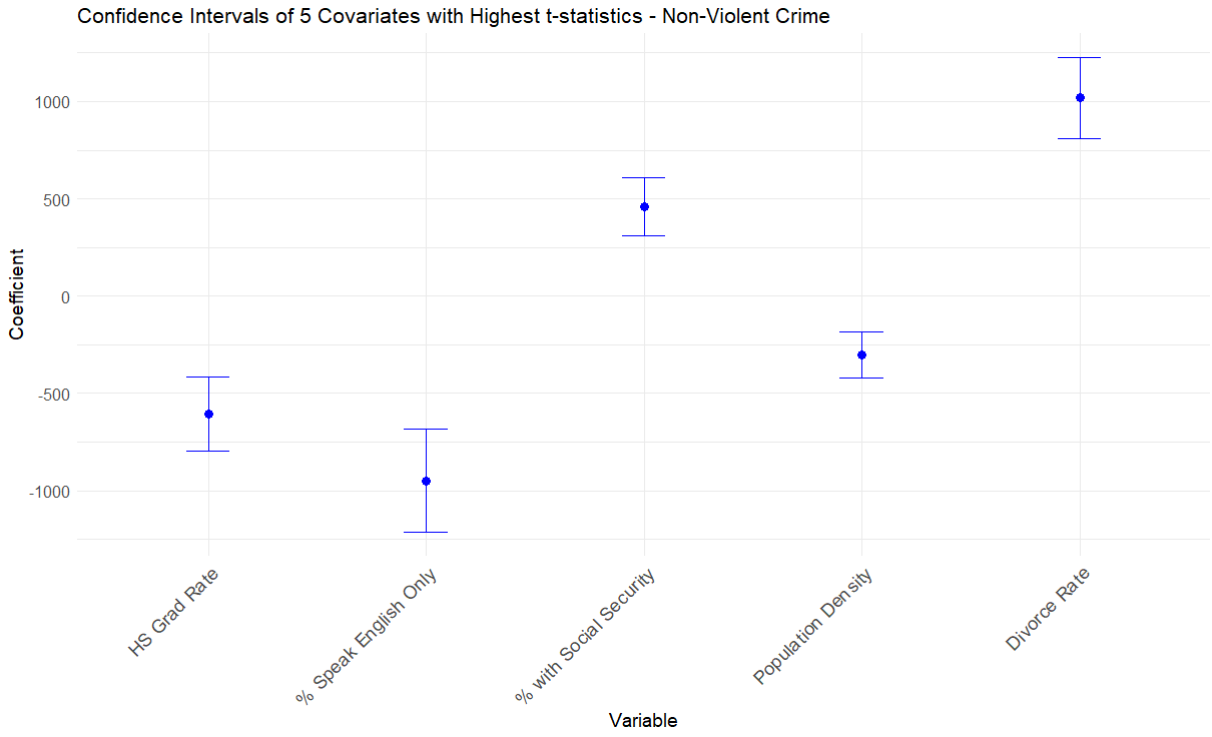

Figure 8: Confidence Intervals - Non-Violent Crime Model

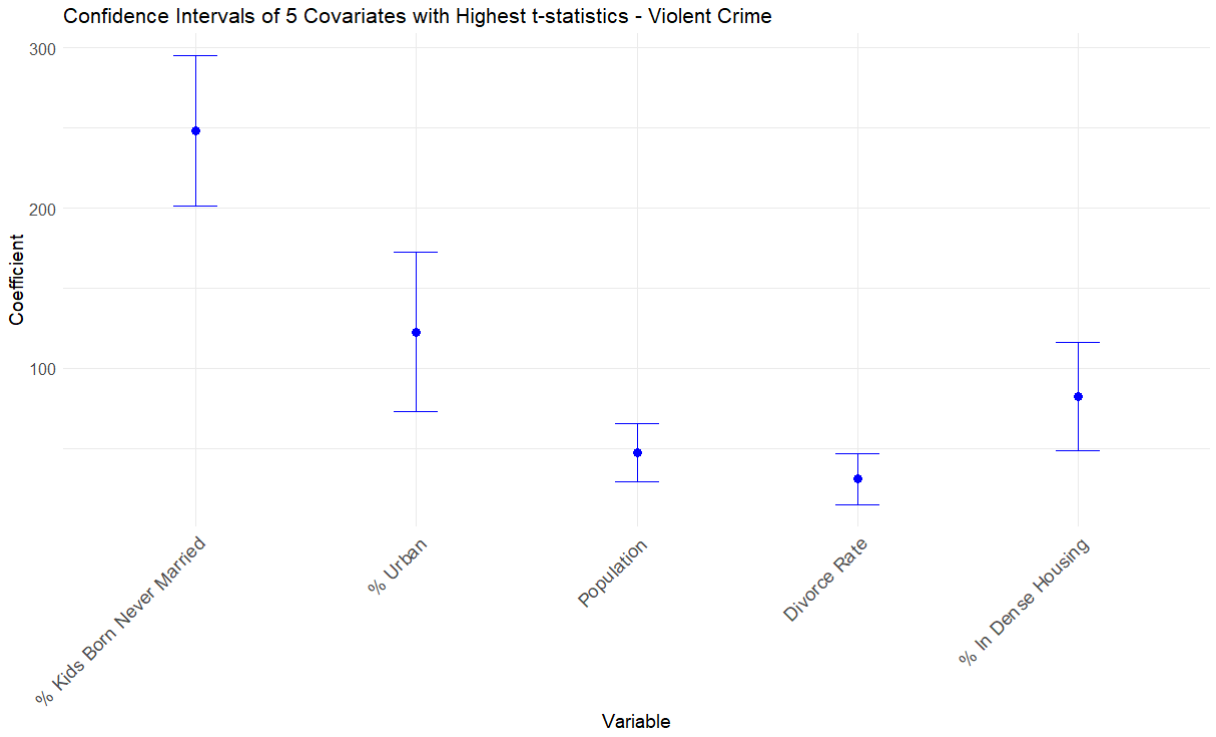

Figure 9: Confidence Intervals - Violent Crime Model

Figure 8 shows that non-violent crime rates are negatively correlated with High School Graduation rate, percentage of the population which speaks only English and population density. It is positively correlated with the proportion of the population with social security, and with divorce rate. Notably, none of these confidence intervals include zero, which emphasizes that these covariates are statistically significant.

In Figure 9, we see the top five most significant predictors of violent crime rates all have positive correlations - divorce rate, population, proportion of the population in dense housing, proportion of the community that is urban, and the percentage of kids born to parents that never married. Once again, these are all statistically significant covariates, as none of the confidence intervals include zero.

This concludes the project's section on estimating the most statistically important predictors of violent and non-violent crime, and we now move on to building a predictive model in Section 4.

4 Prediction

The objective of this section of the project was to build the best possible predictive model on an unseen holdout set. To do so, in this project both traditional linear methods were used, along with more modern tree-based methods. When reviewing past literature, it was found that for similar projects, tree-based methods, and specifically XGBoost had the best predictive performance [2]. For this section of the report, each model's predictive ability was measured by its Mean Squared Error (MSE) on the holdout dataset. When testing models, it was found that each model's MSE varied greatly, dependent on the split of test and training data. Therefore, 10-fold cross validation was used when assessing each model, to gain stability in our results. For each model built, the model was trained tested on both the full, and reduced dataset, as described in Section 3.

4.1 Linear Regression Methods

Four different linear regression models were tested. These included regular unbiased linear regression, as well as three penalized regression techniques: LASSO, which has an $L1-$ penalty term, Ridge Regression which has an $L2-$ penalty term and ElasticNet, which adds both penalty terms, each with weight 0.5 in this case. The

results are shown in Figure 9 and 10, for non-violent crime and violent crime predictive models, respectively. The linear results are on the left side of each Figure.

Non-Violent Crime Rate (MSE)	Linear Reg.		LASSO Reg.		Ridge Reg.		Elastic Reg.		Random Forest	XGBoost	Light-GBM
Full Dataset	1867		1860		1871		1863		1863	1900	1849
Simplified Dataset	1851		1851		1861		1851		1849	1918	1880

Figure 10: Model Evaluation Results - Non-Violent Crime Rates Prediction

Violent Crime Rate (MSE)	Linear Reg.		LASSO Reg.		Ridge Reg.		Elastic Reg.		Random Forest	XGBoost	Light-GBM
Full Dataset	377		375		376		375		378	378	371
Simplified Dataset	373		373		373		374		366	394	374

Figure 11: Model Evaluation Results - Violent Crime Rates Prediction

In terms of the regression results, for predicting both violent and non-violent crimes, it is clear that there are two factors which increase predictive performance. Firstly, the MSE is lower when using the reduced dataset as opposed to the full dataset. This is because simplifying the model tends to reduce overfitting, and allows the model to better generalize to unseen data by truly learning the relationship between the features and targets, as opposed to memorizing the training data. In the same manner, model loss was lower when using penalized regression (it was lowest when training with LASSO). Once again, this is because penalized regression biases the model to have lower coefficients, and thus tends to focus more on the most important features, and trims the importance of less important features. Hence, it is effective at preventing overfitting.

4.2 Tree-Based Methods

In addition to these regression models, tree-based methods were also tested - specifically, random forest, XGBoost and Light-GBM. Unlike linear regression, these models do not assume linearity, and seek to discover non-linear relationships by building decision trees. Another difference is that these algorithms require hyperparameters, such as the number of trees in a random forest model or the learning rate. These hyperparameters must be tuned for each model. The hyperparameters were tuned by testing a variety of possible values on a validation set. The hyperparameter value which minimizes MSE on the validation set is then tested on the test set.

For both types of crime, the random forest model has the lowest loss overall, and this loss is lowest when using the reduced dataset. This is unsurprising, since these models are able to learn non-linear dependencies. Once again, when using the reduced dataset, the model is better able to avoid overfitting, and generalize relationships.

Another interesting observation here is that XGBoost and Light-GBM models performed better when using all the data as opposed to the reduced dataset. This is because these are more sophisticated algorithms, which have additional means to prevent overfitting, such as pruning [1]. It also has more advanced ensemble learning mechanisms. For example, they use boosting, where decision trees are built sequentially, each correcting errors in the previous tree [1]. Therefore, when additional features are added to the dataset, these algorithms are able to avoid overfitting and improve performance, even if the additional features are not the most important. Since these algorithms do have more advanced mechanisms, it is surprising that they performed worse than random forest models. However, this may because boosting algorithms such as XGBoost and Light-GBM are more sensitive to outliers than Random Forest. As discovered in the EDA phase, this data does contain many outliers, hence reducing the performance of boosting algorithms.

4.3 Conclusions

Overall, it is clear that the overall performance of prediction was quite weak. The best model for predicting non-violent crime rates had an MSE of 1849, which is about 30% of the mean non-violent crime rate. Similarly,

the best model for predicting violent crime rates had an MSE of 366, which is about 35% of the mean violent crime rate. Hence, the prediction results are clearly quite inaccurate and non-meaningful. We may conclude the data used is quite noisy and insufficient for accurately predicting crime rates on unseen data. However, we were also able to conclude that using models which took advantage of non-linear relationships were able to outperform those which only considered linear relationships. In addition, we may conclude that in most cases, using a reduced dataset with only the most important features, or penalizing model complexity with regularization increases performance by reducing overfitting.

As a final note, one extension that could be made on the project would be to use SHAP values for explainability, as used in [5]. SHAP values are used to explain the output of a machine learning model by assigning each feature an importance value [5]. Particularly in more complex machine learning algorithms such as XGBoost, where the models are somewhat of a 'blackbox', SHAP values add an element of transparency and interpretability to the model. They also add a new measure of feature importance, in the new context of making prediction on unseen data with advanced algorithms. This new feature importance metric could be compared to determine if the most important features in this prediction are the same as those with the most importance when using the methods described in Section 3.

References

[1] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

[2] Dakalbab, F., Abu Talib, M., Abu Waraga, O., Bou Nassif, A., Abbas, S., and Nasir, Q. (2022). Artificial intelligence crime prediction: A systematic literature review. *Social Sciences Humanities Open*, 6(1):100342.

[3] Nitta, G., Rao, B., Sravani, T., Ramakrishiah, N., and Muthu, B. (2019). Lasso-based feature selection and naïve bayes classifier for crime prediction and its type. *Service Oriented Computing and Applications*, 13.

[4] Redmond, M. (2011). Communities and Crime Unnormalized. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5PC8X.

[5] Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., and Chen, J. (2022). Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, 94:101789.

5 Appendix

5.1 Zipped Files

The following files were submitted in a zipped folder and provide the code used for the analysis in this project, An excel file is also included with the full feature importance test results. Here are the file descriptions:

- feature_importance_selection.R: R source code for computing feature importance tests, and selecting features
- linear_models.R: R source code for testing predictive linear models
- tree_models.ipynb: Python notebook for training tree-based models
- cleaning_EDA.R: R source code for cleaning the raw dataset, checking sampling bias, and plotting EDA charts
- featureimportance.xlsx: Complete list of feature importance results

5.2 Link To Dataset

https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized

5.3 Total Instance Count by State

State	n instances	State	n instances	State	n instances
CA	279	WA	40	ME	17
NJ	211	GA	37	WV	14
TX	162	OK	36	MD	12
MA	123	TN	35	NM	10
OH	111	VA	33	SD	9
MI	108	OR	31	ND	8
PA	101	SC	28	ID	7
FL	90	KY	26	WY	7
CT	71	RI	26	NV	5
MN	66	AR	25	VT	4
WI	60	CO	25	AK	3
IN	48	UT	24	DC	1
NC	46	LA	22	DE	1
NY	46	NH	21	KS	1
AL	43	AZ	20		
MO	42	IA	20		
IL	40	MS	20		

Figure 12: State Counts