A Study of Crime Clearance Statistics in Austin, Texas

Kevin O'Connor, Elle Khun, Abigail Johnson

5/9/2022

Abstract

The purpose of this report is to predict whether or not a crime is cleared by arrest, based on various demographic factors across Austin, TX. To build our analysis, we utilize data from the Austin Police Department and the U.S. Census Bureau, which includes information on reported crimes and demographics by zip code. We employ logistic regression, stepwise regression, random forest, gradient-boosted tree models, and lasso to predict crime clearance status. From these models, we select the random forest model as our best model with an area under the curve (AUC) of 0.672 when validated against testing data.

Introduction

Whether in a large city or a small town, crime is an unfortunate, yet ever-present, fact of society. While crime may be a constant reality, unique social and demographic factors create asymmetry in the way crime is both committed and penalized in a given city. Such is the case for Austin, TX. As a large metropolitan area covering several zip codes, each area of Austin has a unique composition of various social and demographic factors, which ultimately influence the way crime is penalized across Austin. While the type of crime committed has a clear impact on the resulting repercussions, the location in which crime is committed also seems to influence criminal consequences. It begs the question of how much our social environment may skew legal decisions that should, ideally, be objective and just.

The motivation of this study is to build a model that accurately predicts the outcome of a criminal offense, based on the demographic factors of a crime's location. This type of predictive analysis can be difficult, as no criminal offense is the same as another. Two crimes may both be classified as "theft", but one offense may have more clear evidence of the crime than the other. Therefore, the consequences may vary based on the specific details of each case. Some cases reside in a more "gray area" of what the just consequence should be. It is these cases that may be more susceptible to bias based on the demographics of their environment, such as income and race. Without the granular detail of each criminal offense, it can be hard to predict when these external factors truly have an effect.

However, as more data regarding criminal action becomes publicly available, it is increasingly important that we build predictive models to understand the true magnitude of these external factors' influence on legal decisions. This paper is organized as follows. In Section 2, we describe the data used for our analysis and provide some initial visualizations. In Section 3, we present our statistical models, namely: Logistic, Stepwise Selection, Random Forest, Gradient Boosting, and Lasso. In Section 4 , we compare the performance of our models using confusion matrices and select the best performing model. Finally, we summarize our results and conclusions in Section 5.

Data

Feature Engineering and Data Loading

```
obs <- read_csv("austin_crime.csv")

# Encode the crime target variable of interest
# 1 if "Cleared by Arrest", 0 otherwise
obs$clearance_status <- ifelse(obs$clearance_status == "Cleared by Arrest", 1, 0)

# Omit any missing rows for clearance status
obs <- filter(obs, clearance_status != "NA")
# This result in ~40k observations being dropped

# We determine the arrest rate, and append it to our zip-code level data
t1 <- obs %>% mutate(number_crimes = n_distinct(unique_key)) %>%
    group_by(zipcode) %>%
    summarise(arrest_rate =(sum(clearance_status))/sum(number_crimes))

zip_list <- as.data.frame(t1$zipcode)
zip_list <- as.character(zip_list$`t1$zipcode`)

zip <- read_csv("austin_crime_zips.csv")
zip <- select(zip,
              c(zipcode, population_density, median_income, median_home_value,
                prop_white,arrest_rate))
zip$arrest_rate <- scale(zip$arrest_rate)
zip$zipcode <- as.character(zip$zipcode)


tx <- geojson_read("tx_zip_geo.json", what = "sp")


## Final cleaning steps: Encoding zip as a factor with 45 levels, and dropping NA values
## Resulting in 27459 observations remaining
table <- merge(obs, zip, by = "zipcode") %>% drop_na()
table$zipcode = as.factor(table$zipcode)
```

Data

Subsetting geojson

The GeoJSON we found has the boundaries of every zipcode in Texas. We subset this JSON based on the zipcodes included in our dataset.

Summary and Description

The data used in this analysis is from the Austin Police Department Crime Reports Data. This dataset only includes incidents during 2014-2015 that the Austin Police Department responded to and wrote a report. One incident could have several offenses associated with it, however, only the highest offense is recorded

Table 1: Table (1) Data Dictionary

Variable	Description	Type
clearance_status	whether or not crime is cleared by arrest	numeric
zipcode	postal zipcode in Austin, TX	numeric
population_density	people per square mile	numeric
median_income	median income of postal zipcode	numeric
median_home_value	median home value of postal zipcode	numeric
prop_white	proportion of postal zipcode that is white	numeric
arrest_rate	proportion of crimes cleared by arrest relative to total crimes	numeric

in the dataset. The dataset includes information about the exact location, zip code, time, and particular offense for each incident. Additionally, this dataset includes information about the clearance status for each recorded offense. The clearance status defines how or whether a crime was solved using three categories: Not cleared, Cleared by Exception, and Cleared by Arrest. For our study, the 'Cleared by Arrest' category is our chosen clearance status of interest. Therefore, we re-coded clearance status as a binary indicator, where Cleared by Arrest is equal to one, and zero otherwise. Additionally, we re-coded zip codes as factors in order to include each zip code in our models while also maintaining interpretability.

In addition to the Austin Police Department data, we used U.S. Census Bureau data to collect demographic information for each zip code in the crime report dataset. Specifically, we gathered information about the population density, median income, median home value, arrest rate, and racial composition for each zip code.

As an additional feature of interest, we create "arrest rate" as a new feature in the dataset. Arrest rate is the proportion of crimes cleared by arrest relative to the total number of crimes in each zip code. This helps us understand which areas in Austin tend to be high arrest areas, therefore adding predictive power to the likelihood a crime is cleared by arrest based on its zip code. Also, as this feature is relatively flat, we decided to z-score the arrest rate, so we can more easily understand relative differences between different areas.

As a final step to create a data set for modeling, we include latitude and longitude boundaries for each zip code of interest. This allows us to visualize group trends in Austin by each zip code, and understand the asymmetry of demographic factors in the Austin metropolitan area. Figure 1 and Figure 2 show median income by zip code and median home value by zip code, respectively. We can clearly see a correlation between income and home value, with most high income and high valued homes on the west side of Austin.

Visualizing Zipcodes

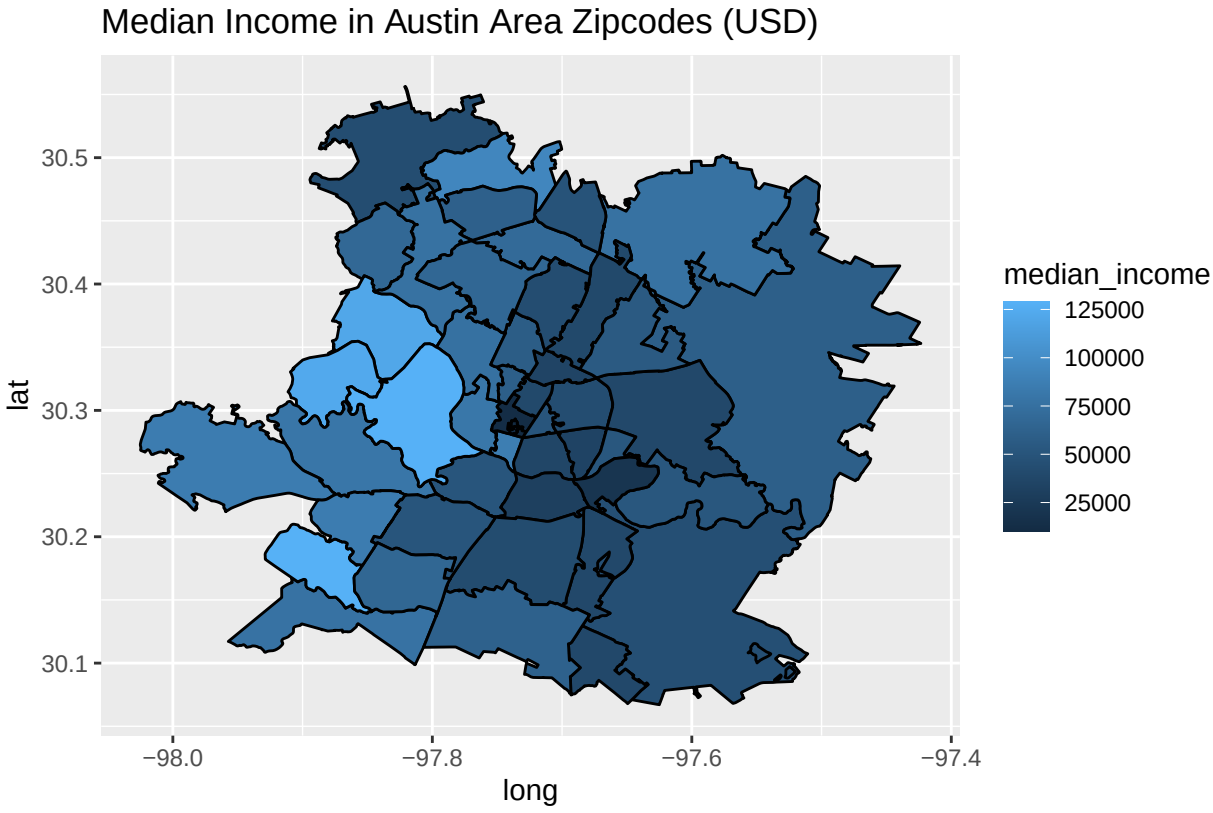


Figure (2)

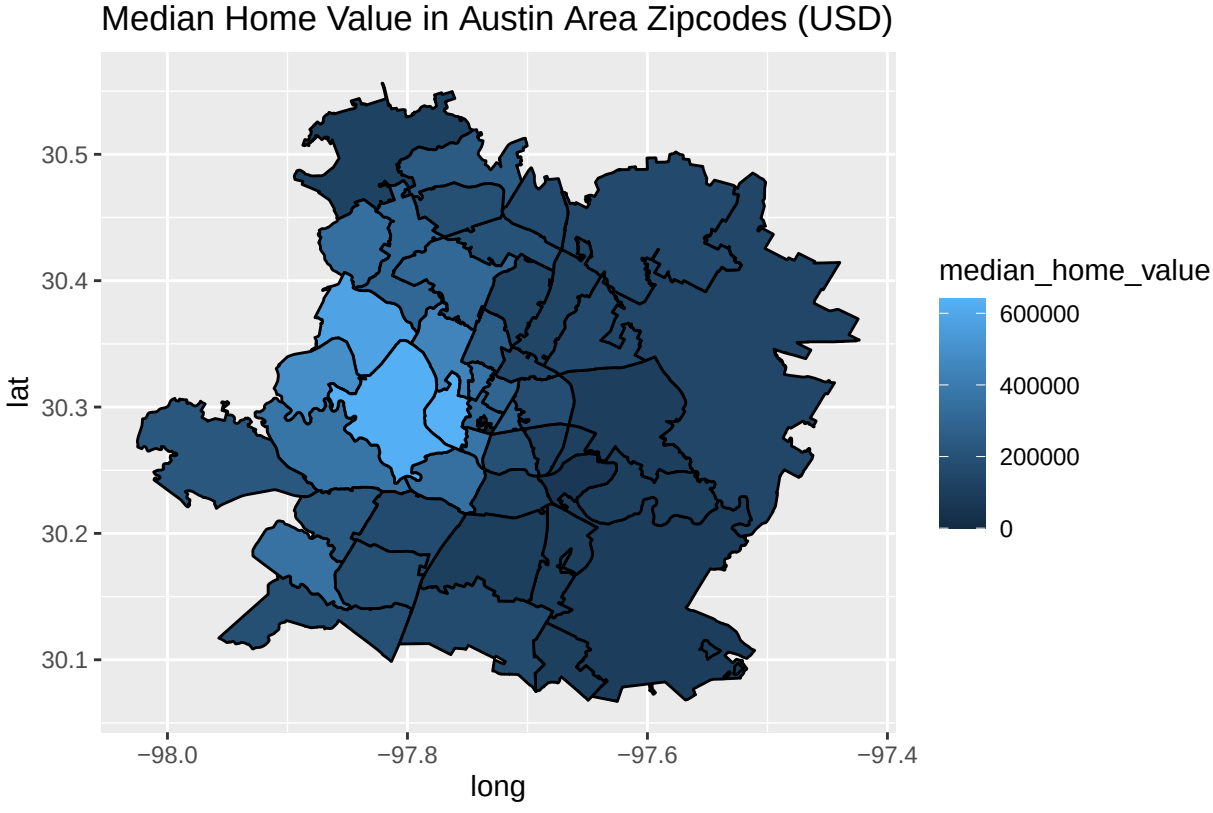


Figure (3)

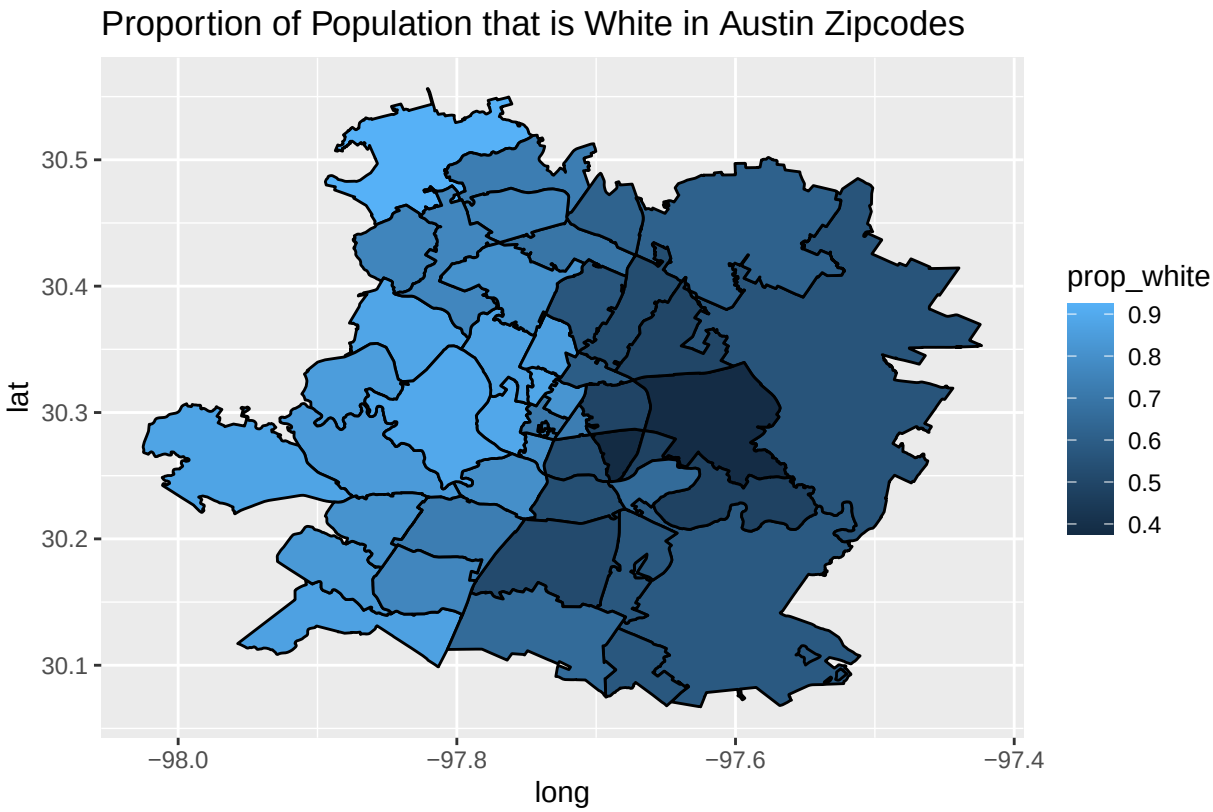

Proportion of Population that is White in Austin Zipcodes
30.5
30.4
30.3
30.2
30.1
lat
−98.0
−97.8
−97.6
−97.4
long
prop_white
0.9
0.8
0.7
0.6
0.5
0.4

Figure (4)

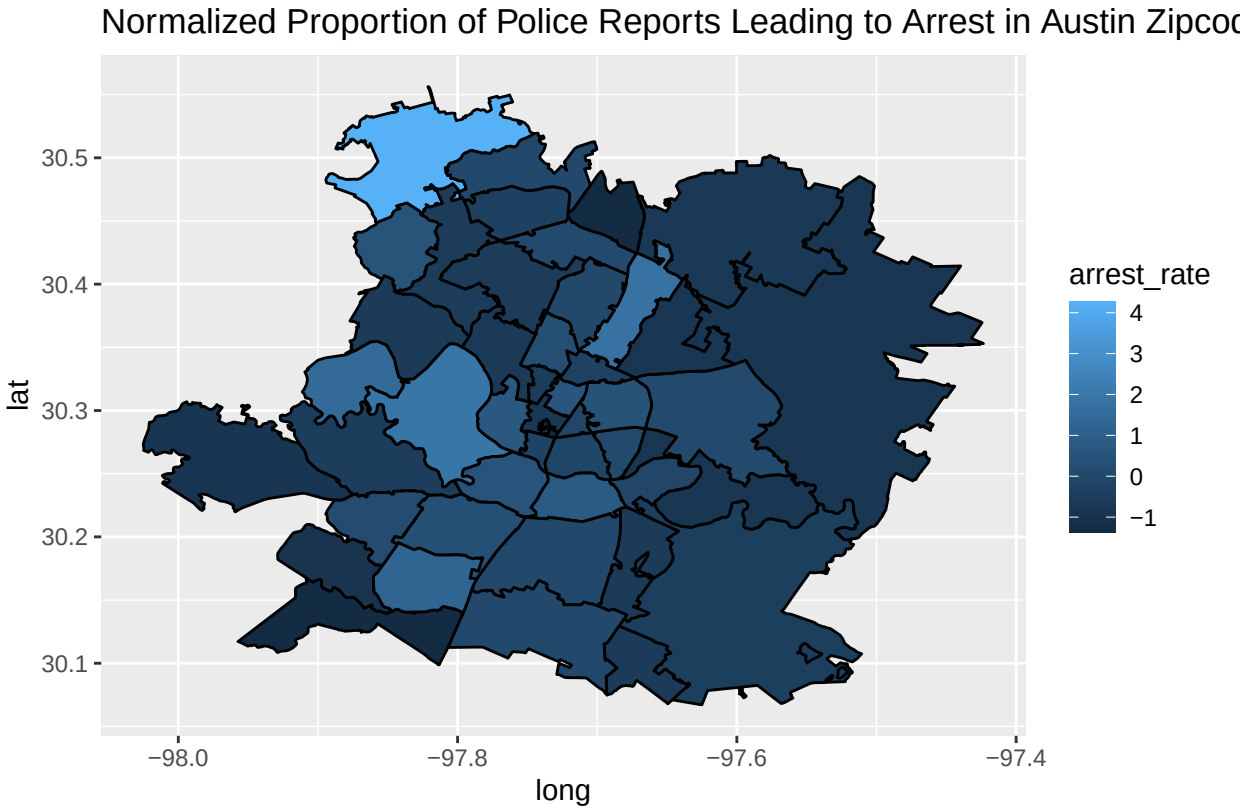


Figure (5)

Figure 4 shows the proportion of the population that is white by zip code. This plot shows a clear trend of a more white population on the west side of Austin, and a more minority population on the east side. When comparing this to Figure 2 and Figure 3, we can see that the west side of Austin is wealthier and more white, while the east side is less wealthy and less white. Do these trends correlate with arrest rates?

Figure 5 shows the average arrest rate by zip code. There does not appear to be a clear trend of certain areas with high or low arrest rates. However, one northwest zip code experiences a notably high arrest rate. This particular area of Austin is relatively more white, but of lower income status. However, this area has an extremely low population density, and only a few observations of crimes, so we have to understand that lower density areas with fewer observations may skew results in our analysis. When ignoring outliers, it would appear that median income may be one of the more important factors in determining the arrest rate. One might conclude that community law enforcement budgets are shaped by the prevailing wealth and taxes collected in those areas.

Question: Does Race play a part in the prevailing arrest rate?

A visualization

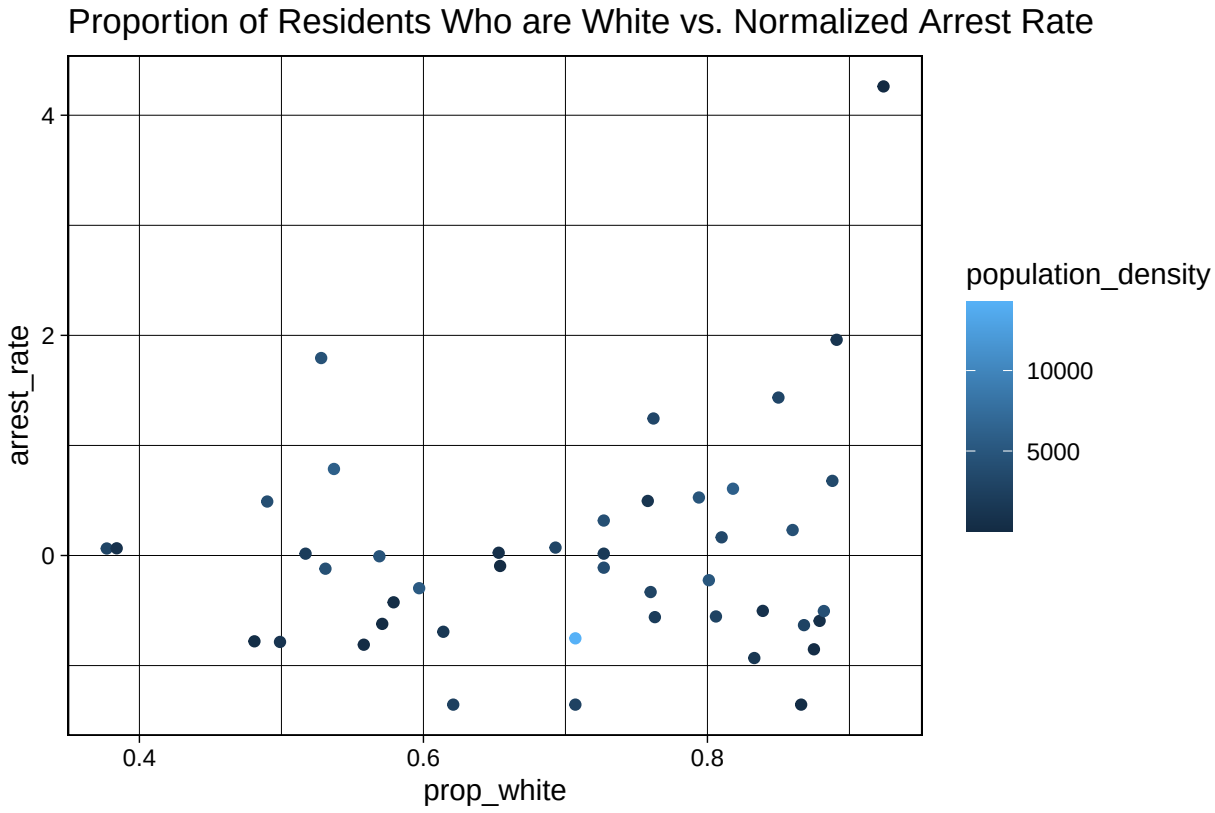


Figure (6)

Of high interest in today's political landscape is the implications of ethnicity in justice outcomes. What relationships does ethnicity have in the prevailing arrest rate in a community? If you are white, is your police report more likely to lead to an arrest? Figure 6, shows that, at least in Austin, there does not appear to be a clear trend. To understand whether or not there are outliers skewing results, the figure also accounts for population density, measured in people per square mile. We notice that there are several outliers in low density areas. There appears to be no readily obvious trend.However, we will run some models to see if such an assumption can be justified statistically.

Methods

Summary

To attempt to predict the outcome of 'clearance_status', we will run a model horse race, and select the model that returns the lowest validated RMSEout. To validate each model, we will calculate the RMSEout for a train/test split 10 times, and take the average from the ten samples to provide a more robust estimate for comparison. After we select the best model, we will generate ROC curves, and calculate AUC.

First Model: Logistic Regression

We started with a baseline logistic regression mode, with the specification of clearance status on everything else. The dependent variable `clearance_status` is re-coded to take on the unit value if crime is cleared by arrest and zero otherwise. Also, we dropped all nulls before creating a train/test split with 80 percent of the data going to the training set, while the remaining 20 percent ending up in the testing set data. In our analysis, we chose to include population density, median income, median home price, white population share, and arrest rate as independent variables. The resulting model is

$$P(clearancestatus = 1 \mid x_{i,t}) = \beta_0 + \beta_1 populationdensity_{i,t} + \beta_2 medianincome_{i,t} + \beta_3 medianhomevalue_{i,t} + \beta_3 propwhite_{i,t} + \beta$$

The t statistics reveal that `arrest rate`, `median home price`, and `white population share` are highly significant in predicting `clearance_status`, so they should be included in the model to help with prediction. Then, out of sample root mean square error (RMSE) is calculated as a measure of the model's out of sample performance. RMSE will be used to compare across all models, and the lower, the better.

Each coefficient shows a ceteris paribus effect of every feature on clearance status. For instance, $\beta_{prop_white} = -0.67$ is statistically significant at 1 percent. The interpretation is that as the share of the white population increases by 1 percent, the probability of crime being cleared by arrest declines by 67 percentage points, holding all other features constant.

Second Model: Stepwise Selection

Although the logistic model is simple, it fails to capture context-specific effects by ignoring interaction terms. Stepwise selection computes the best set of variables by including main effects, and pairwise interaction terms that result in the lowest RMSE. The resulting model chosen by stepwise selection is

$$P(clearancestatus = 1 \mid x_{i,t}) = \beta_0 + \beta_1 populationdensity_{i,t} + \beta_2 medianincome_{i,t} + \beta_3 medianhomevalue_{i,t} + \beta_3 propwhite_{i,t} + \beta$$

We notice an interaction term in our model. This term allows the effect of a unit change in *medianhomevalue* to depend on *arrest rate*. The coefficient of this interaction term measures the effect on clearance status of an additional dollar of median home value is greater, by the amount β_5, for each additional percentage point increase in arrest rate.

By including the interaction term, the model performance is enhanced. RMSE declined from 2.0318 to 2.0254. Other models will be considered in the following sections with the goal of reducing RMSE further.

```
## Start:  AIC=16842.09
## clearance_status ~ population_density + median_income + median_home_value +
##     prop_white + arrest_rate
##
##                                        Df Deviance   AIC
## + median_home_value:arrest_rate         1    16830 16814
## + population_density:median_home_value  1    16847 16831
## + median_income:median_home_value       1    16853 16837
## + median_income:arrest_rate             1    16854 16838
## + population_density:arrest_rate        1    16855 16839
## - median_income                         1    16830 16840
## <none>                                       16830 16842
```

```
## + population_density:prop_white         1    16859 16843
## + median_home_value:prop_white          1    16860 16843
## + prop_white:arrest_rate                1    16860 16844
## + population_density:median_income      1    16860 16844
## + median_income:prop_white              1    16860 16844
## - population_density                    1    16835 16845
## - median_home_value                     1    16841 16851
## - prop_white                            1    16841 16851
## - arrest_rate                           1    17788 17798
##
## Step:  AIC=16843.81
## clearance_status ~ population_density + median_income + median_home_value +
##     prop_white + arrest_rate + median_home_value:arrest_rate
```

Third Model: Random Forest

The next model performed is random forest. The highlight of tree is that it automatically detect nonlinearities and interactions. So there is no need to include interaction terms. The process involves resampling the data with replacement 500 times and fitting a tree to each one. Then, averaging the predictions of the 500 different trees. However, we can reduce covariance between each tree by using only a subset of the variables. Thus, the 500 trees are diversified and their predictions are less correlated. By introducing more randomness to the process, we can improve both accuracy and prevent over-fitting. The resulting RMSE is 0.3522 which is a significant improvement from the logistic model or stepwise selection.

Figure [figure number here] presents a variable importance plot. It shows that excluding zip code would increase mean square error (MSE) by 26 percent, suggesting that zip code should be included in the model. The process of the calculation involves comparing out-of-bag performance of the model when using the correct zip code versus permuting the zip code for all observations.

Fourth Model: Gradient Boosting

Similarly to the random forest model in the previous section, boosting combines many decision trees where each tree is fitted to the residual of the previous tree. However, the fit in each round will be scaled down to constrain it from explaining the full variation in the data. The resulting estimate is the sum of all crushed trees in each round. The highlight of the gradient boosting technique is that it keeps each tree in the ensemble from overfitting just like when random forest restricts the number of features to prevent overfitting.

Boosting requires choosing three main hyper parameters: (1) Number of trees: 500, (2) Shrinkage Parameter: 0.01, and (3) Interaction Depth: 4

Fifth Model: Lasso Model

Results

Model Performance Summary

Of the five models we tried to run, the random forest model outperformed all others, with a validated RMSEout of .352. From here, we will make predictions against the testing dataset, generate a ROC curve, and calculate AUC.

Table 2: Table (2) Modeling Results

Model	RMSEout
Logit	2.03
Stepwise Selection	2.04
Random Forest	.352
Gradient Boosting	1.984
Lasso	.9062

ROC Curve for best model by RMSEout

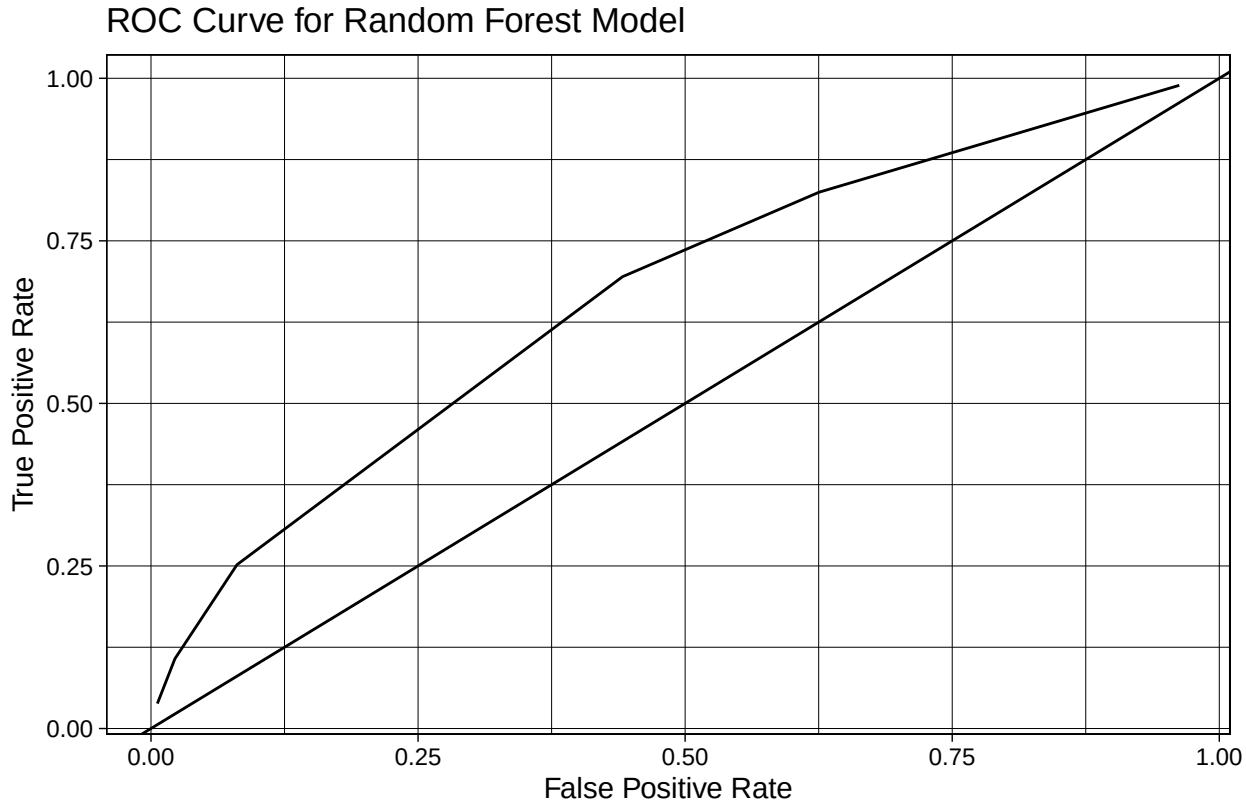


Figure (7)

As we see in figure 7, the RF model outperforms a 50/50 guess. When calculating AUC, we return .672. To make a comparison, does Random Forest outperform a more general model? When comparing to the results of the more simple logit model, we see that the random forest has 2.5% more AUC. (See appendix), and hence, as more predictive power. When we graph variable importance, we observe that the prevailing arrest rate in each zipcode is most important in regards to percent increase in MSE if omitted, followed by population density, and median income. Clearly, jurisdictional effects are of the greatest predictive power. We do observe a small effect of ethnicity, which could potentially be explained by systemic biases, but the effect is marginal in comparison, and we cannot state such with any meaningful degree of confidence. Population density also plays a significant part. One might infer that in higher density areas, there is a greater police presence.

Conclusion

After comparing five different methods, we select the best predictive model with an AUC measure of 0.672. This means our best model only correctly predicts about 67% of crime clearance statuses. While our rate of accurate prediction is better than a random guess, it still leaves a large margin for improvement. Our model uses demographic information at the aggregate level to predict crime clearance outcomes at the individual level. Therefore, our predictive power may be limited by our lack demographic information for each individual in the police report data set. Using data with both crime report and demographic information at the individual level, may improve the predictive performance of future studies. Moreover, future studies may want to include more features focused on income and economic factors, as these features seem to carry important predictive power for crime clearance status. With a more comprehensive data set and additional economic features, future studies likely will see improved model performance.

Appendix

Sanity-Check: ROC Curve for the logit model

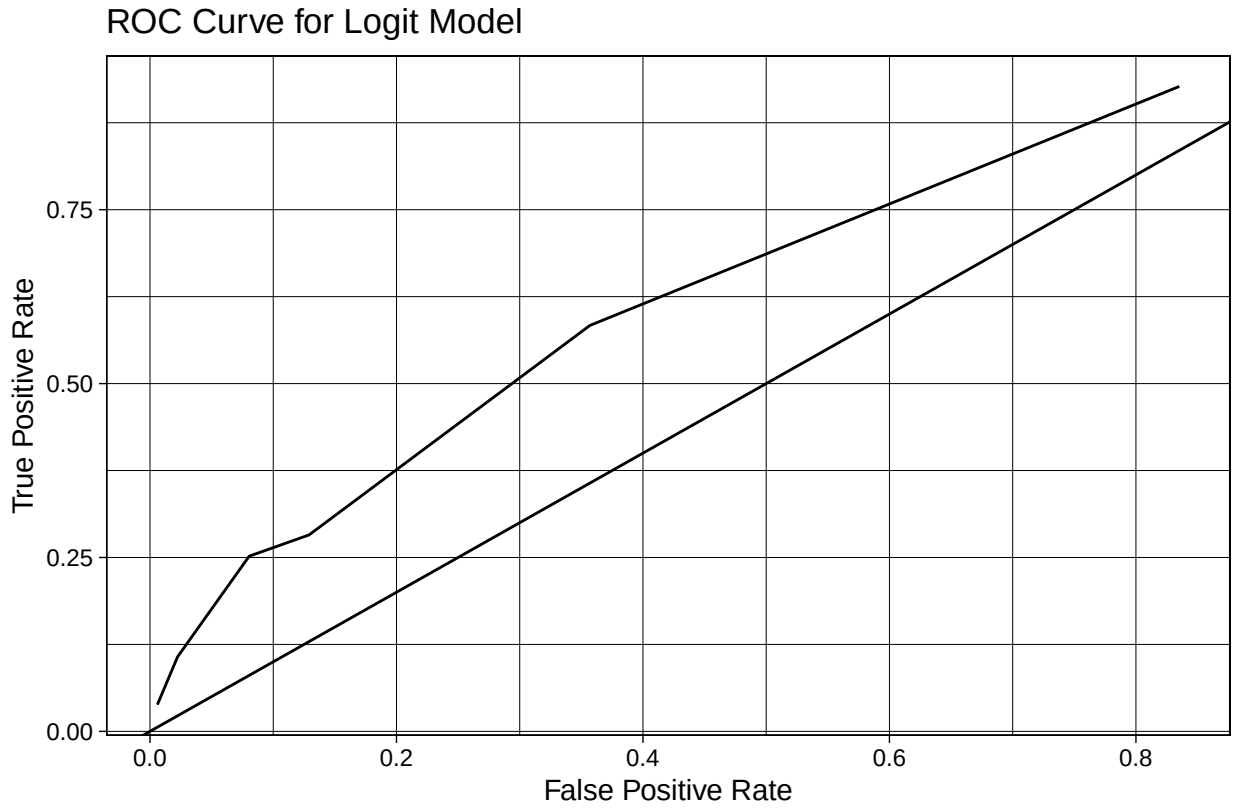
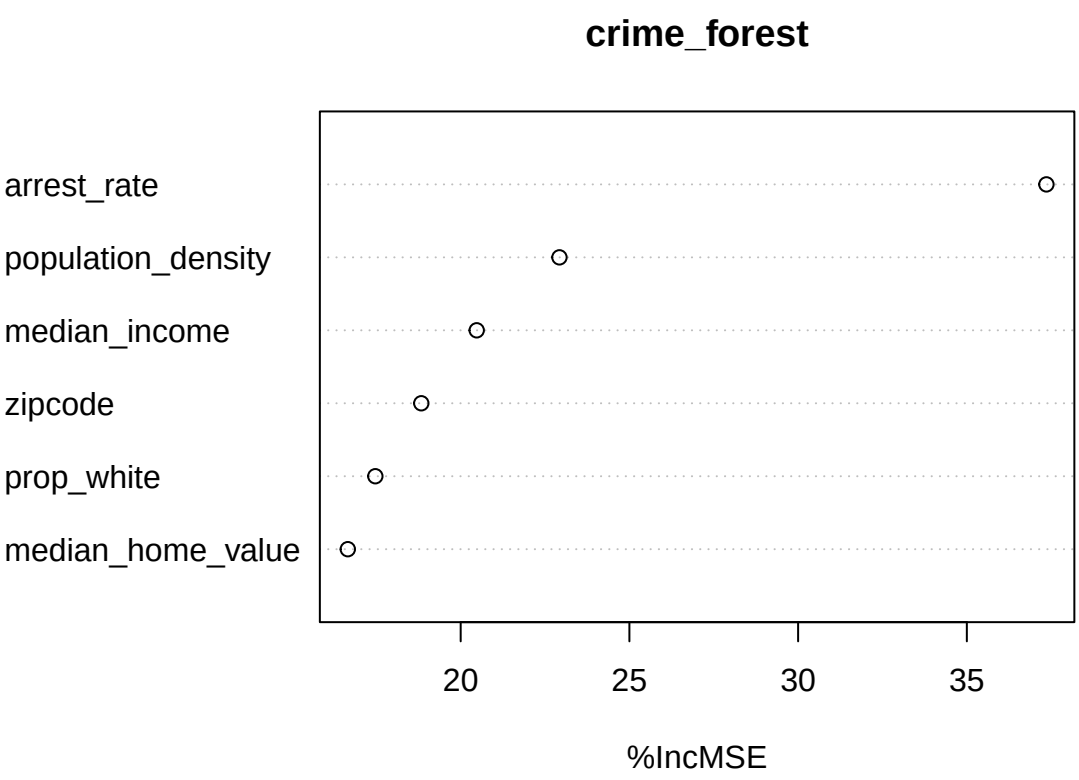


Figure (7)

##Variable Importance Plot for RF model


crime_forest
arrest_rate
population_density
median_income
zipcode
prop_white
median_home_value
20
25
30
35
%IncMSE