Statistical Analysis of the *Communities and Crime* Data Set

Angelina Kolomoytseva

April 27, 2021

1 Abstract

Investigation of data that links crime rates and different demographic and socio-economic factors becomes increasingly important in supporting more informed decision-making when it comes to policy changes and police funding cuts as demanded by the protesters and supporters of the “Black Lives Matter” movement. The current analysis attempts to investigate which demographic and socio-economic factors are related to an increase in violent and non-violent crime rates. While the model explanatory power is not very high, predicting violent crime rates may be feasible. Classification analysis to determine features, associated with a higher percentage of educated people in communities is also performed. All the applied methods except for the Logistic Regression produced well-performing models.

2 Introduction

The current project has been inspired by the “Black Lives Matter” movement against discrimination and police brutality. Most recently the slogan "Black Lives Matter" has won a lot of public support as it has been raising an important question of equality and justice for historically vulnerable and disadvantaged populations, communities, and individuals. Meanwhile, some of the counter-protests claim that “Police Lives Matter” due to the unsafe nature of the job and exposure to crime daily. Insufficient data-driven research has been making it more difficult to address historical tension between protesters and law enforcement.

The two objectives of the current project are: 1) investigate the relationships between various demographic and socio-economic factors and violent and non-violent crime rates in 2215 communities across the U.S. (regression analysis); 2) identify factors, associated with a higher percentage of educated people in the 2215 communities (classification analysis).

3 Methods

Data Source

Lack of scientific research may be attributed not only to the complexity of the issue but also to the scarcity of publicly available data. The researcher has been able to identify only one relevant data set that is publicly available, authentic, and involves multiple attributes. The data set “Communities and Crime” is acquired from the UCI machine

learning repository website [1]. It combines socio-economic data from three sources: the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats (LEMAS) survey, and crime data from the 1995 FBI UCR. One important advantage is that the data source includes a detailed explanation of the variables.

The data set involves 2215 observations and 147 variables. There are 4 non-predictive variables and the variable *State* that have not been used in the analysis, 124 potential predictors, and 18 crime variables. All the variables utilized in the analysis are continuous. The complete data set and attribute information can be found on the UCI machine learning repository website. Table A1 summarizes the variables in the dataset.

Data Preprocessing

The data set contains 44592 missing observations. 22 of the law enforcement variables have 1872 missing observations (approximately 85%) due to the LEMAS survey limitation described on the data source website. All such variables along with the remaining rows with missing values (there are 963 of such values) are excluded. Non-predictive variables, the variable *State*, and individual crime variables are also removed. I also exclude *numbUrban* and *NumUnderPov* while keeping *pctUrban* and *PctPopUnderPov* since those respective variables contain the same information (i.e., number/percentage of people living in areas classified as urban and number/percentage/of people under the poverty level). Finally, due to the singularity issue when applying the OLS method, two additional variables are removed at that time: *OwnOccQrange* and *RentQrange*.

The two response variables for the regression analysis, *Viol.Rate* (violent crime rate) and *nonViol.Rate* (non-violent crime rate) are generated based on the total number of violent and non-violent crimes per 100K population. All the percentage type variables in the data set are provided in the percentage form (not in the decimal form). Thus, I define the *Viol.Rate* and *nonViol.Rate* as (*ViolentCrimesPerPop*/100,000 x 100) and (*nonViolPerPop*/100,000 x 100), respectively.

The second part of the analysis focuses on determining the factors, associated with a larger percentage of educated people in a community. The target variable for the classification analysis, *Edu* is generated using the *PctBSorMore* (the percentage of people 25 and over with a bachelor's degree or higher education). Communities in the 66.7^{th} percentile are considered "more educated" and those below the 66.7^{th} percentile are "less educated". For this part of the analysis, the other two education-related variables, namely *PctLess9thGrade* (the percentage of people 25 and over with less than a 9^{th}-grade education) and *PctNotHSGrad* (the percentage of people 25 and over that are not high school graduates), are not included.

For both types of analysis, the data is randomly split into training (1301 observations) and test (600 observations) sets. The test set contains roughly 30% of the data. For the regression analysis, I also standardize the response and the predictors, which all are continuous variables, and remove the three leverage points after checking the model assumptions, which improves model performance for predicting both *nonViol.Rate* and *Viol.Rate*.

Data Exploration

Prior to modeling it is important to explore the data by looking at summary statistics, linear associations, and plots. However, some of these steps are difficult to execute when working with high dimensional data. For instance, examining pairwise linear associations between all variables is not very practical. Yet, investigating strong linear correlations that appear to be significant may be helpful. Figure A1 demonstrates that such linear associations are present in the data. The dark blue circles on the plot indicate some of the strong positive linear associations while the dark red circles display some of the strong negative associations among the variables.

Since many variables are provided as a total number as well as a percentage of the population, or per capita, or per 100K population value, or contain similar information, one can expect some of the predictors to be highly correlated. As such, the percentage of immigrants who immigrated within the last 8 years has a strong linear positive correlation with the percentage of immigrants who immigrated within the last 10 years. The same is true for other immigration-related variables, housing and rent variables, etc. Some interesting correlations include: *population* (population of a community) and *NumUnderPov* (the number of people under the poverty level) with the correlation coefficient of 0.9892; *NumUnderPov* (the number of people under the poverty level) and *NumKidsBornNeverMar* (the number of kids born to never married) with the correlation of 0.9828; *NumImmig* (the total number of people known to be foreign-born) and *NumStreet* (the number of homeless people counted in the street) with the correlation of 0.9355; *racePctHisp* (the percentage of the population that is of Hispanic heritage) and *PctSpeakEnglOnly* (the percent of people who speak only English) with the negative correlation coefficient of -0.915; *pctWWage* (the percentage of households with wage or salary income) has a negative correlation of -0.902 with *pctWSocSec* (the percentage of households with social security income).

Many of the crime variables have high significant positive correlations among themselves (i.e, *burglaries* and *larcenies*, *robberies* and *autoTheft*, etc.). Several crime types, including *robberies*, *autoTheft*, *murders*, *burglaries*, *assaults*, *larcenies*, and *rapes* have a strong positive linear association with one or more of the following variables: *NumUnderPov* (the number of people under the poverty level), *NumKidsBornNeverMar* (the number of kids born to never married), *population* (population of a community), *numbUrban* (number of people living in areas classified as urban), *NumImmig* (the total number of people known to be foreign-born), *NumStreet* (the number of homeless people counted in the street), *NumInShelters* (the number of people in homeless shelters), and *HousVacant* (the number of vacant households).

The variable *Viol.Rate* has a moderate to strong positive correlation with all the crime types variables and some other predictors, including *PctKidsBornNeverMar* (the percentage of kids born to never married, 0.74) and *racepctblack* (the percentage of the population that is African American, 0.63). *Viol.Rate* is negatively associated with *PctKids2Par* (the percentage of kids in family housing with two parents, -0.73), *racePctWhite* (the percentage of the population that is Caucasian, -0.68), etc.

The variable *nonViol.Rate* has a moderate to strong positive correlation with most of the crime types variables, especially with *burglPerPop* (the number of burglaries per 100K population, 0.81) and *larcPerPop* (the number of larcenies per 100K population, 0.94) and the following predictors: *PctPopUnderPov* (the percentage of people under the poverty level, 0.51), *MalePctDivorce*/*FemalePctDiv*/*TotalPctDiv* (the percentage of males/females/total who are divorced, 0.59, 0.6, 0.6), *PctKidsBornNeverMar* (the percentage of kids born to never married, 0.55). *nonViol.Rate* is negatively associated with *PctKids2Par* (the percentage of kids in family housing with two parents, -0.67), *PctPersOwnOccup* (the percent of people in owner-occupied households, -0.5), etc.

Some of the moderate to strong positive linear associations with *Edu*, which is based on the data set variable *PctBSorMore,* include *PctOccupMgmtProf* (the percentage of people 16 and over who are employed in management or professional occupations, 0.779), *medFamInc* (the median family income, 0.618), *pctWInvInc* (the percentage of households with investment/rent income, 0.605), *whitePerCap* (per capita income for Caucasians, 0.589), *medIncome* (the median household income, 0.560). *Edu* has negative linear associations with *PctOccupManu* (the percentage of people 16 and over who are employed in manufacturing, -0.651), *pctWPubAsst* (the percentage of households with public assistance income, -0.486), *PctHousNoPhone* (the percent of occupied housing units without a phone, -0.454), etc.

Some of the listed linear relations are demonstrated in Figures A2-A3.

The summary statistics indicate skewness in variables' distribution and some potential outliers. Several of the variable distributions are displayed in Figures A4-A5.

Modeling

A total of three statistical methods are used for the first part of the analysis: the OLS, the Stepwise Selection, and the LASSO regression. Since there is a large number of predictor variables, the OLS is applied to gain a general idea about the model fit while the two other methods are more useful since they have the ability to select the parameters to include in the final model. The LASSO regression analysis on the training data includes plotting the solution path, plotting the cross-validation errors, and selecting the best tuning parameter that minimizes the cross-validation error.

For the regression modeling, two cases are considered separately: 1) when the response variable is *nonViol.Rate;* 2) when the response variable is *Viol.Rate.* Each model's performance is evaluated in terms of the test error (MSE). I also report the R^2_{adj} for the OLS and Stepwise models. R^2_{adj} measures the goodness of fit of a model and takes into account its relative simplicity.

The classification analysis methods involve the LDA, Logistic Regression, Classification Tree, and the three ensemble classifiers: Bagging, Random Forest, and Boosting. By checking the data set graphically before conducting the analysis, and using just two variables at a time, I notice a clear possibility for a linear separation between communities with a higher percentage of educated people and those with a lower percentage of educated people. The plots are shown in Figure A6.

4 Results

Regression Analysis

When *nonViol.Rate* is used as the response variable, the OLS model determines a total of 17 out of 98 predictors to be statistically significant at .05 level of significance (e.g., *pctWSocSec* - the percentage of households with social security income, *pctWRetire* - the percentage of households with retirement income; *PctPopUnderPov* - the percentage of people under the poverty level, *PctEmploy* - the percentage of people 16 and over who are employed, *MalePctNevMarr* - the percentage of males who have never married, *PctPersOwnOccup* - the percent of people in owner-occupied households; *PctKids2Par* - the percentage of kids in family housing with two parents, *HousVacant* - the number of vacant households, *LemasPctOfficDrugUn* - the percent of officers assigned to drug units).

The Stepwise Selection includes 55 variables in the final model. In addition to the variables listed above, there are several variables related to ethnicity, income, employment (e.g., the percentage of people employed in manufacturing/in professional services), education, immigration, marital status (e.g., the percentage of males who are divorced), housing (e.g., the median year housing units built), etc.

The LASSO model has the effect of forcing some of the coefficient estimates to be exactly 0, yielding a sparse model, which includes only a subset of variables as well. Surprisingly, the solution path for the LASSO model determines 73 predictors to be included in the model.

Table 1 summarizes the results of the analysis for *nonViol.Rate*.

Table 1. *Regression Results when the Response is nonViol.Rate (non-violent crime rate).*

Model	MSE	R^2_{adj}
OLS	0.354	0.604
Stepwise	0.351	0.612
LASSO	0.339	

The test MSE improves for the Stepwise selection approach compared to the OLS, and it further improves for the LASSO method. However, the R^2_{adj} of the OLS and Stepwise models suggest that linear models do not approximate very well the true function regardless of the type of model. Perhaps, a more flexible model or the addition of some other predictor variables could lead to better estimation and a lower test MSE value.

When *Viol.Rate* is used as the response variable, the OLS model determines a total of 20 predictors to be statistically significant at .05 level of significance (e.g., *HousVacant* - the number of vacant households, *racepctblack* - the percentage of the population that is African American, *LemasPctOfficDrugUn* - the percent of officers assigned to drug units, population - population for the community, *PctLess9thGrade* - the percentage of people 25 and over with less than a 9th-grade education, *PctWorkMom* - the percentage of moms of kids under 18 in the labor force, *PctKidsBornNeverMar* - the percentage of kids born to

never married, *NumImmig* - the total number of people known to be foreign-born, *RentLowQ* - rental housing - lower quartile rent, etc.).

The Stepwise Selection includes 46 variables in the final model. Additional variables describe ethnicity, age, income, education (e.g., the percentage of people 25 and over with less than a 9th-grade education/that are not high school graduates), employment, family, and marital status (e.g., the percentage of moms of kids under 18 in labor force, the percentage of males who are divorced/who have never married), immigration (e.g., the total number of people known to be foreign-born, the percentage of immigrants who immigrated within last 3 years), housing and rent, population density, etc.

The solution path for the LASSO model determines only 21 predictors to be included in the model shown in Figure A7.

Table 2 summarizes the results of the analysis for *Viol.Rate*.

Table 2. *Regression Results when the Response is Viol.Rate (violent crime rate).*

Model	MSE	R^2_{adj}
OLS	0.308	0.671
Stepwise	0.304	0.678
LASSO	0.299	

The test MSE and R^2_{adj} suggest that linear models work better when predicting *Viol.Rate* than when the goal is to predict *nonViol.Rate*. The MSE marginally improves for the Stepwise selection approach compared to the OLS, and there is an additional small improvement for the LASSO regression. The LASSO model is the most parsimonious as it includes only 21 predictors and thus it seems to be the best performing model out of the ones presented in the current analysis.

Classification Analysis

The LDA method seems to be a good fit for the data. This is expected since all the predictors are continuous, and this method is robust to the violation of the multivariate normal assumption. It imposes an assumption on X (marginal distribution) and requires equality of variance-covariance matrix assumption for the two groups, however, these are less strict. LDA is preferred over QDA for these data as the number of predictors *p* is large. Figure A8 shows the LDA classification results using two inputs at a time on both training and test data sets. The plots show that most of the misclassified observations are close to the border.

Both the Logistic Regression and Regularized Logistic Regression, which uses L_1 penalty for variables selection and shrinkage to fit a reduced Logistic Regression, did not converge due to the partial separation issue. The model setup requires some overlap between the groups and does not work for well-separated groups. The existence of the partial separation is very difficult to identify when the number of predictors is large. However, it was revealed by the error messages based on the model fit.

The Classification Tree methods enjoy automatic stepwise selection and impurity reduction; they are robust to outliers due to the feature truncation which reduces the effect of the extreme values. This is imperative for the current project. Considering the importance of pruning and controlling the complexity of the parameters, I apply the "tree" and "rpart" packages. The "tree" package model provides a smaller misclassification error on the testing data set and uses 6 features to build a tree: *PctOccupMgmtProf* (the percentage of people 16 and over who are employed in management or professional occupations), *agePct16t24* (the percentage of the population that is 16-24 in age), *pctWPubAsst* (the percentage of households with public assistance income), *pctWRetire* (the percentage of households with retirement income), *PctSameCity85* (the percent of people living in the same city for 5 years as in 1985), *PctUsePubTrans* (the percent of people using public transit for commuting). The Classification Tree results can be found in Figure A9.

The main goal of all the ensemble classifiers is to improve the performance of individual "weak learners" that suffer from instability. They allow to reduce the variability and to improve the prediction accuracy. The Bagging method aggregates predictions from multiple individual tree models built from the bootstrap samples and requires individual classifiers to be independent. This method utilizes many additional predictor variables compared to a single Classification Tree which can be seen in Figure A10. The most important turning parameter is the number of bootstrap samples. Using 25 bootstrap replicates provides the lowest misclassification rate. The "ipred" package model provides a smaller test error compare to the "rpart" package. The out-of-bag estimate of the misclassification error from the training sample is higher than the test error. This could mean that the method generalizes well.

The Random Forest (RF) extends Bagging by decreasing the correlation among individual classifiers. The method subsets individual observations (by bootstrapping) and samples the features at each step which allows to de-correlate the classifiers. As a result, classifiers from the bootstrap samples may contain different subsets of the predictors. The number of variables tried at each split is roughly a square root of the total number of features. RF allows assessing the relative importance of the features in terms of the overall prediction and further reduces the misclassification error. Figure A11 provides numerical measures and relative importance plots based on the mean decrease in terms of the different impurity measures (Accuracy, Gini Index, etc.) *PctOccupMgmtProf* is by far the most important variable. Increasing the number of variables tried at each node from 9 (default) to 18 provides the lowest test error.

While Bagging and Random Forest both build multiple classification trees in parallel, the Boosting uses ensemble learning in a sequential way instead of bootstrapping: it re-weights the original observations, giving the misclassified observations higher weights before constructing the next tree. One of the concerns with the ensemble methods is overfitting. Boosting is known to be resistant to overfitting. Once again, as shown in Figure A12, *PctOccupMgmtProf* is the most important feature in the analysis.

Table 3 summarizes the results of the classification analysis.

Table 3. *Classification Analysis Results: Model Comparison.*

Model	Training error	Test error
LDA	4.38%	6.17%
CART	5.61%	6.33%
Bagging	6.38%	4.83%
RF	5.53%	4.33%
Boosting	4.30%	4.67%

The Random Forest model performs the best thanks to its smallest misclassification rate on the test data set.

5 Conclusion

In the proposed project the relationship between the predictor variables and two generated outcome variables (separately) - violent crime rate and non-violent crime rate - has been examined. The statistical methods included the OLS, the Stepwise selection, and the LASSO regression analysis. The linear models work better when predicting *Viol.Rate* rather than *nonViol.Rate*. The LASSO model for predicting *Viol.Rate* is the most parsimonious as it includes only 21 predictors and is selected as the best model in the current analysis. As suggested by the coefficient estimates in the final model (Figure A7), numerous features, including marital status, family characteristics, housing, ethnicity, age, etc. may play a role in the prediction of the violent crime rate.

The second part of the analysis focuses on classification, investigating factors, associated with a higher number of educated people in a community. Due to the clear evidence of the linear separation between the two classes, the LDA method performs expectedly well while the Logistic Regression does not converge. The LDA's test error rate is slightly over 6%. The misclassification error is further reduced to less than 5% by all the ensemble methods with Random Forest performing the best. Relative importance measures have determined that *PctOccupMgmtProf* (the percentage of people 16 and over who are employed in management or professional occupations) is by far the most important variable in a classification model development.

6 Discussion

The current research project attempts to investigate the relationship between multiple demographic and socio-economic factors and violent and non-violent crime rates in 2215 U.S. communities. It also investigates the factors, associated with a higher percentage of educated people in each community. The obtained data set can be used to address numerous research questions thanks to many useful variables and a detailed explanation of the features. The major limitation of the analysis is the lack of more recent publicly available data sets of similar quality. Obtaining current data would allow drawing more relevant conclusions.

Appendix

Table A1. *Summary of the Variables in the "Communities and Crime" data set.*

Variables	Number of Variables & Variable Type
General Information (non-predictive) • Community name (string) • U.S. state (categorical, not used in current analysis) • County Code (numeric) • Community Code (numeric) • Fold (numeric)	– **5 variables total**
Demographic (population, age, ethnicity, etc.)	– **10 variables (continuous)**
Income (median household income, per capita income, etc.)	– **15 variables (continuous)**
Community (the total number or percent of the population considered urban, the total number percentage of people under the poverty level, number of people in homeless shelters, number of homeless people counted in the street, etc.)	– **10 variables (continuous)**
Education (percentage of people 25 and over with a bachelor's degree or higher education, etc.)	– **3 variables (continuous)**
Employment Status (percentage of people 16 and over who are employed, percentage of people 16 and over who are employed in management or professional occupations, etc.)	– **6 variables (continuous)**
Marital Status (percentage of males/females who are divorced or have never married, percentage of population who are divorced)	– **4 variables (continuous)**
Household/Family Characteristics (mean number of people per family, percentage of kids in family housing with two parents, percentage of kids born to never married, percent of family households that are large, mean persons per household, etc.)	– **12 variables (continuous)**
Foreign born/Immigrated (total number of people known to be foreign born, percent of population who have immigrated within the last 5 years, percent of people who do not speak English well, etc.)	– **12 variables (continuous)**
Housing (mean persons per owner occupied/rental household, median number of bedrooms, number of vacant households, percent of vacant housing that has been vacant more than 6 months, median year housing units built, owner occupied housing - median value, rental housing - median rent, median gross rent as a percentage of household income, etc.)	– **26 variables (continuous)**
Law Enforcement (number of sworn full time police officers, percent of sworn full time police officers on patrol, total requests for police, police average overtime worked, police operating budget, etc.)	– **23 variables (continuous)**
Other Law Enforcement Variables (population density in persons per square mile, percent of people using public transit for commuting, land area in square miles**)**	– **3 variables (continuous**
Crime Variables (number of robberies in 1995, number of robberies per 100K population, number of auto thefts in 1995, number of auto thefts per 100K population, total number of violent crimes per 100K population, total number of non-violent crimes per 100K population, etc.)	– **18 variables (continuous)**

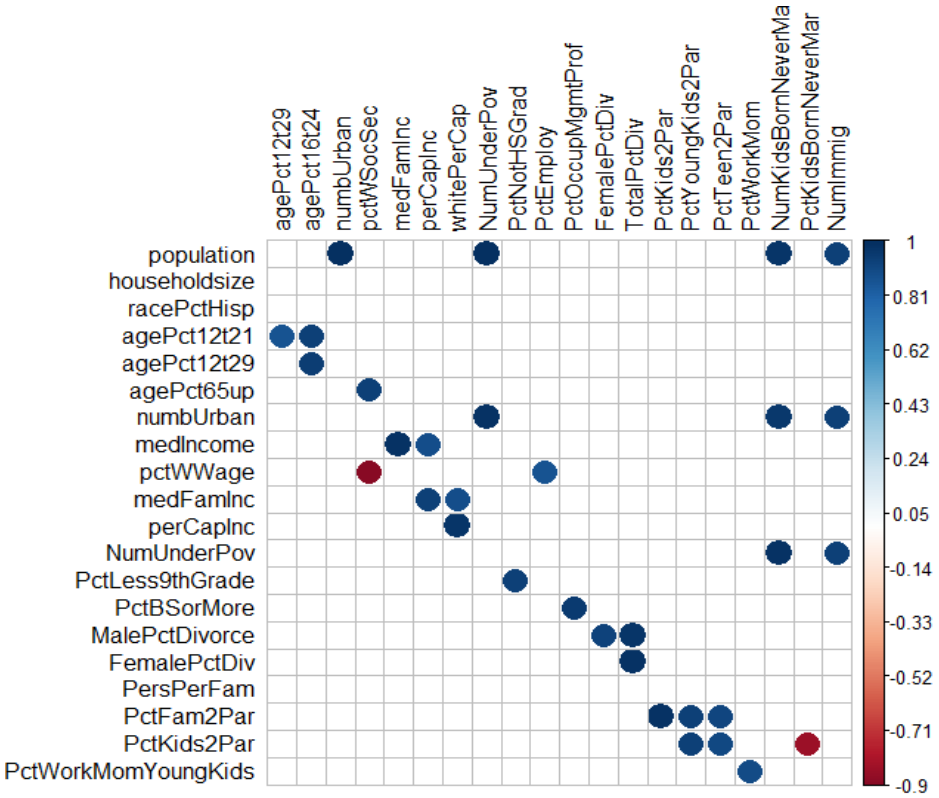


Figure A1. Some of the strong linear significant associations present in data. Dark blue circles indicate strong positive correlations. Dark red circles indicate strong negative linear associations.

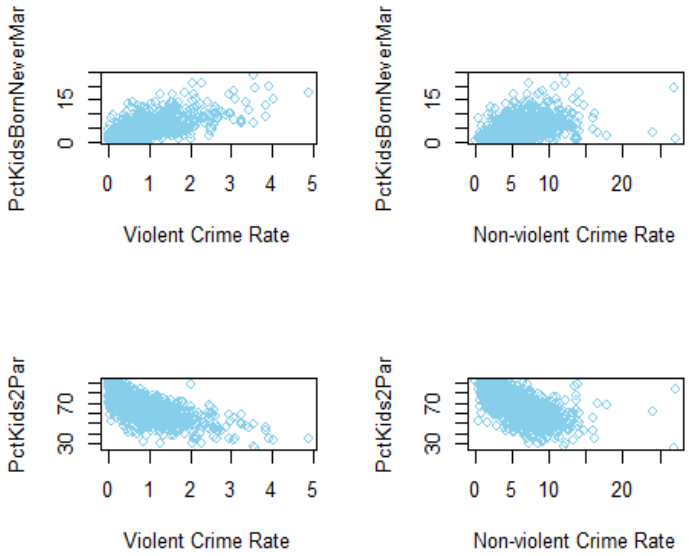


Figure A2. Top two plots demonstrate moderate positive linear relationship between Violent Crime Rate & PctKidsBornNeverMar (the percentage of kids born to never married) and Non-violent Crime Rate & PctKidsBornNeverMar. Bottom two plots show moderate negative linear association between Violent Crime Rate & PctKids2Par (the percentage of kids in family housing with two parents) and Non-violent Crime Rate & PctKids2Par (the percentage of kids in family housing with two parents).

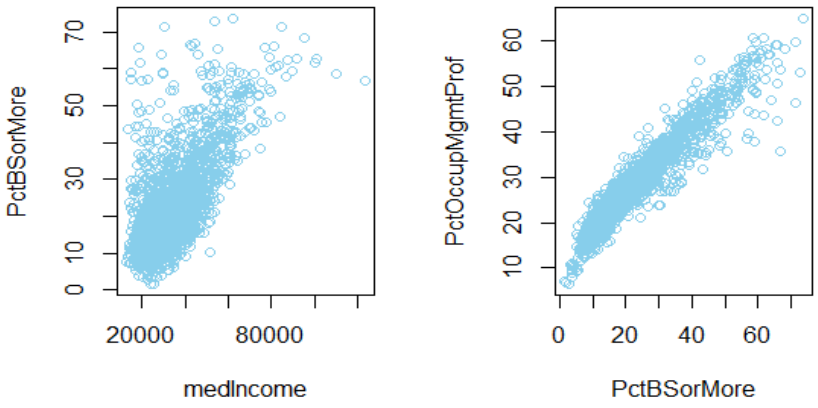


Figure A3. The first plot shows that PctBSorMore (the percentage of people 25 and over with a bachelor's degree or higher education) & medIncome (the median household income) have a moderate to strong linear relationship while the second plot displays a strong linear association between PctBSorMore and PctOccupMgmtProf (the percentage of people 16 and over who are employed in management or professional occupations).

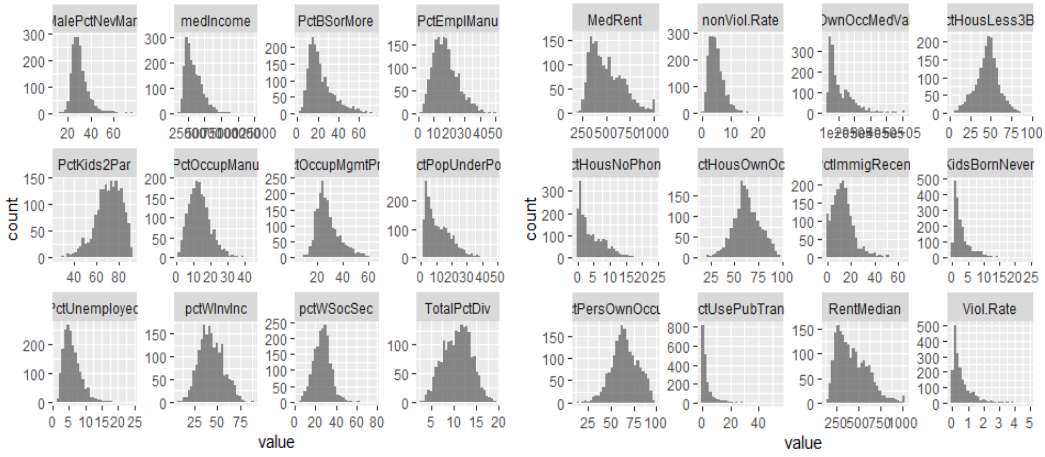


Figure A4. Selected variables: histogram plots. Most are skewed (right or left). Several distributions appear approximately symmetric (pctWSocSec & PctHousLess3BR).

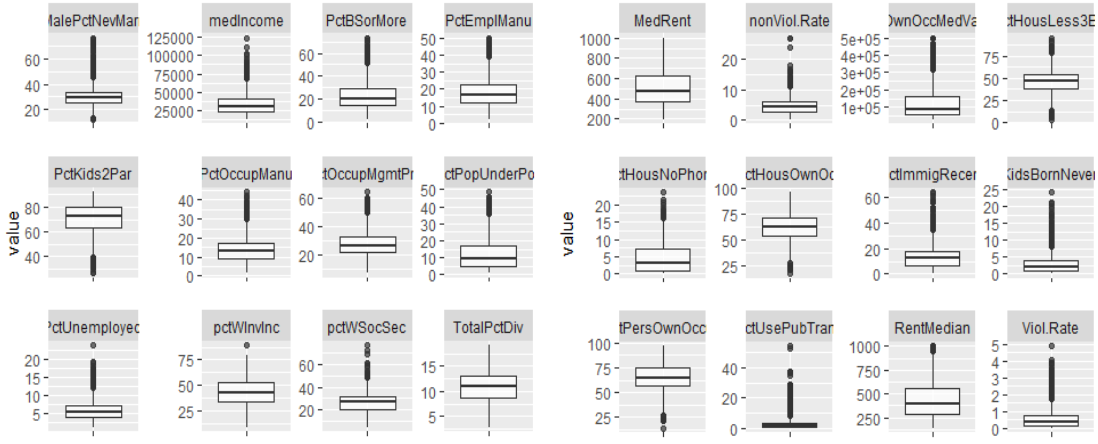


Figure A5. Selected variables: box plots. Mostly skewed data with outliers (Violent Crime Rate and Non-violent Crime Rate, PctKidsBornNeverMar, PctUsePubTrans, and medIncome among others).

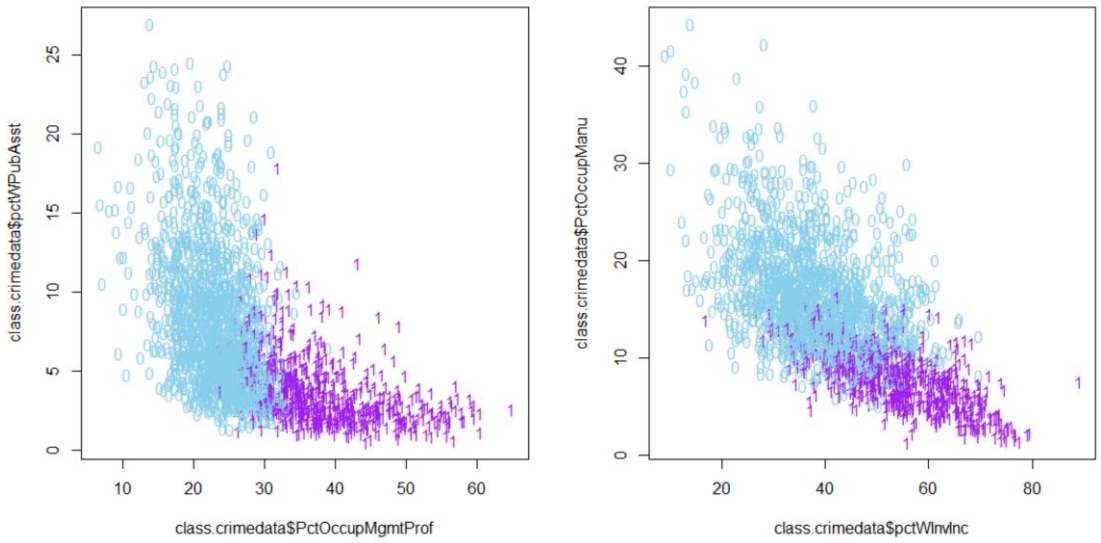


Figure A6. Considering two inputs at a time and assessing the possibilities for the observations to belong to either of the two classes. There is a possibility for a linear separation between communities having more educated people and those having fewer educated people.

21 predictors:

- Racepctblack
- racePctWhite
- agePct12t29
- pctUrban
- pctWPubAsst
- AsianPerCap
- PctEmplManu
- MalePctDivorce
- PctKids2Par
- PctWorkMom
- PctKidsBornNeverMar
- PctPersDenseHous
- PctHousLess3BR
- MedNumBR
- HousVacant
- PctHousOccup
- PctVacantBoarded
- MedRentPctHousInc
- MedOwnCostPctIncNoMtg
- PctForeignBorn
- LemasPctOfficDrugUn

```
99 x 1 sparse Matrix of class "dgCMatrix"
                          1
(Intercept)         0.00974
population          .
householdsize       .
racepctblack        0.05900
racePctWhite       -0.12509
racePctAsian        .
racePctHisp         .
agePct12t21         .
agePct12t29        -0.02605
agePct16t24         .
agePct65up          .
pctUrban            0.04230
medIncome           .
pctWWage            .
pctWFarmSelf        .
pctWInvInc          .
pctWSocSec          .
pctWPubAsst         0.00289
pctWRetire          .
medFamInc           .
perCapInc           .
whitePerCap         .
blackPerCap         .
indianPerCap        .
AsianPerCap         0.00023
OtherPerCap         .
HispPerCap          .
PctPopUnderPov      .
PctLess9thGrade     .
PctNotHSGrad        .
PctBSorMore         .
PctUnemployed       .
PctEmploy           .
PctEmplManu        -0.00495
PctEmplProfServ     .
PctOccupManu        .
PctOccupMgmtProf    .
MalePctDivorce      0.11647
MalePctNevMarr      .
FemalePctDiv        .
TotalPctDiv         .
PersPerFam          .
PctFam2Par          .
PctKids2Par        -0.18379
PctYoungKids2Par    .
PctTeen2Par         .
PctWorkMomYoungKids .
PctWorkMom         -0.04474
NumKidsBornNeverMar .
PctKidsBornNeverMar 0.25801
NumImmig            .
PctImmigRecent      .
PctImmigRec5        .
PctImmigRec8        .
PctImmigRec10       .
PctRecentImmig      .
PctRecImmig5        .
PctRecImmig8        .
PctRecImmig10       .
PctSpeakEnglOnly    .
PctNotSpeakEnglWell .
PctLargHouseFam     .
PctLargHouseOccup   .
PersPerOccupHous    .
PersPerOwnOccHous   .
PersPerRentOccHous  .
PctPersOwnOccup     .
PctPersDenseHous    0.08008
PctHousLess3BR      0.00287
MedNumBR           -0.03952
HousVacant          0.05014
PctHousOccup       -0.03341
PctHousOwnOcc       .
PctVacantBoarded    0.02333
PctVacMore6Mos      .
MedYrHousBuilt      .
PctHousNoPhone      .
PctWOFullPlumb      .
OwnOccLowQuart      .
OwnOccMedVal        .
OwnOccHiQuart       .
RentLowQ            .
RentMedian          .
RentHighQ           .
MedRent             .
MedRentPctHousInc   0.01512
MedOwnCostPctInc    .
MedOwnCostPctIncNoMtg -0.01688
NumInShelters       .
NumStreet           .
PctForeignBorn      0.00696
PctBornSameState    .
PctSameHouse85      .
PctSameCity85       .
PctSameState85      .
LandArea            .
PopDens             .
PctUsePubTrans      .
LemasPctOfficDrugUn 0.04268
```

Figure A7. The LASSO regression results when the response variable is Viol.Rate (violent crime rate). The solution path determines a total of 21 predictors to be included in the model.

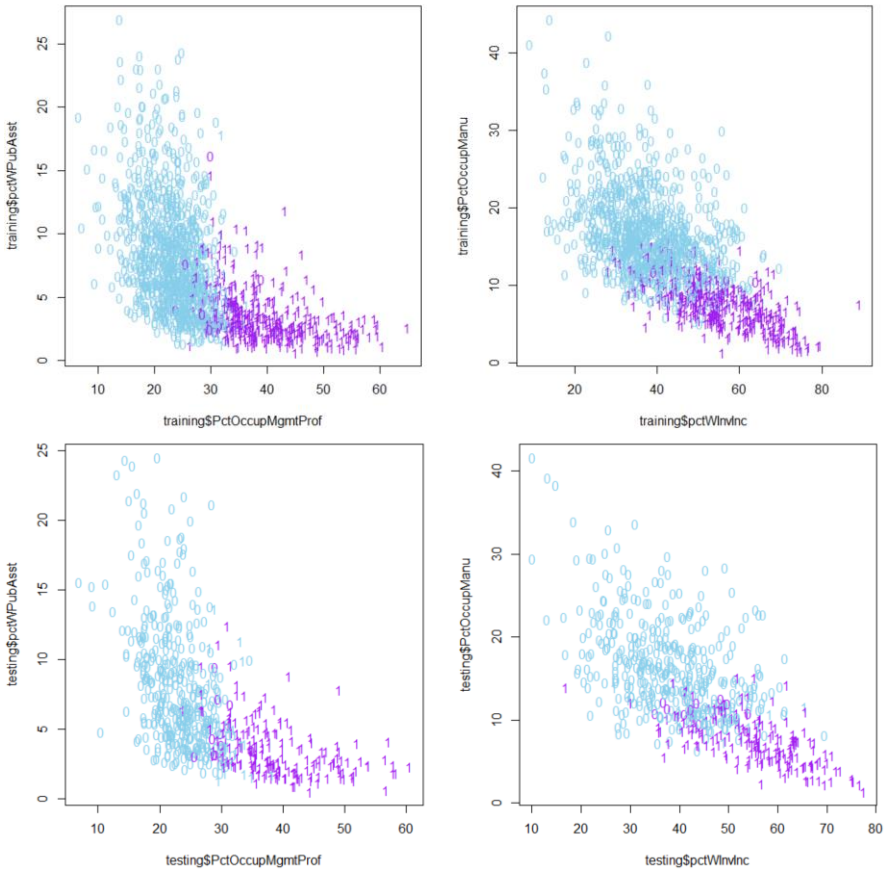


Figure A8. The LDA classification results using two inputs at a time on training set (top two plots) and test set (bottom two plots).

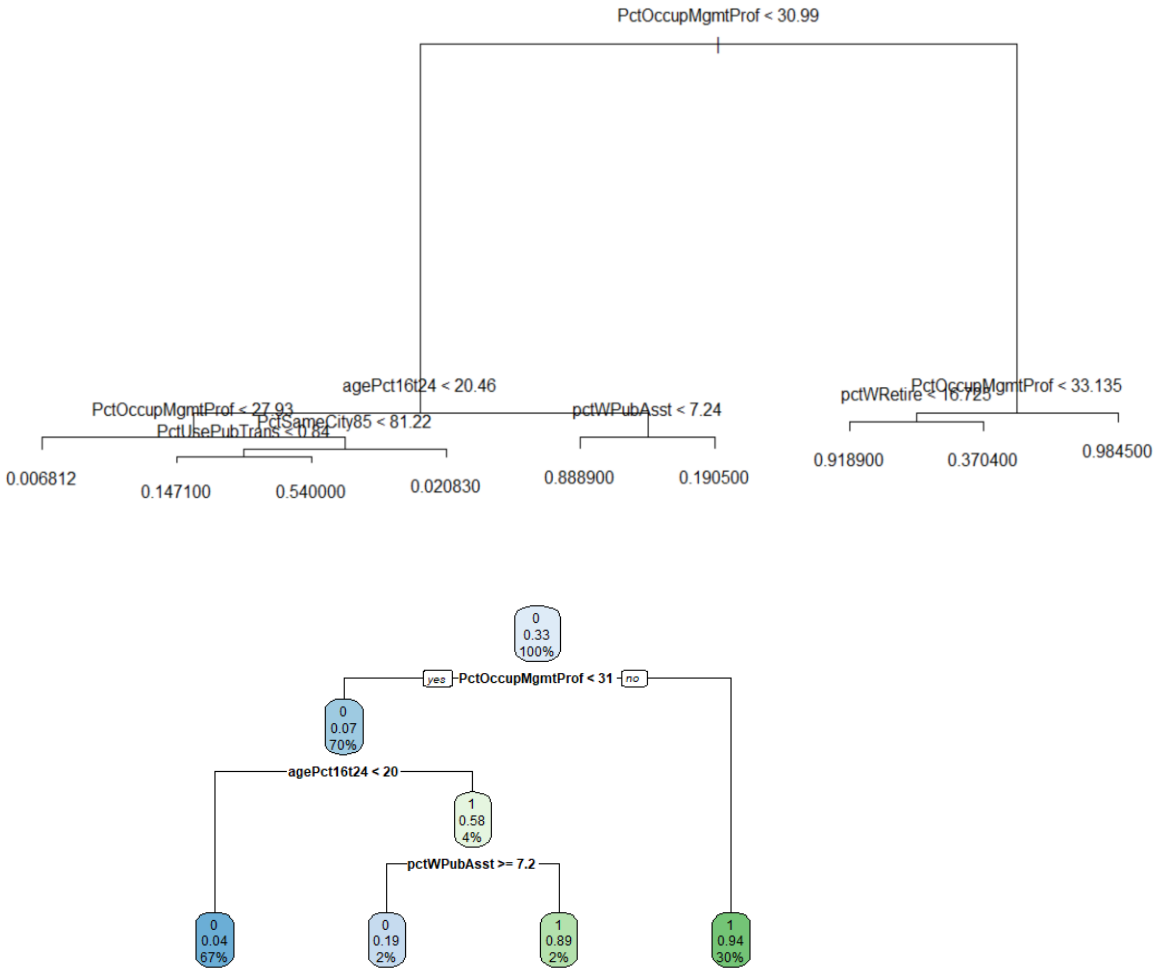


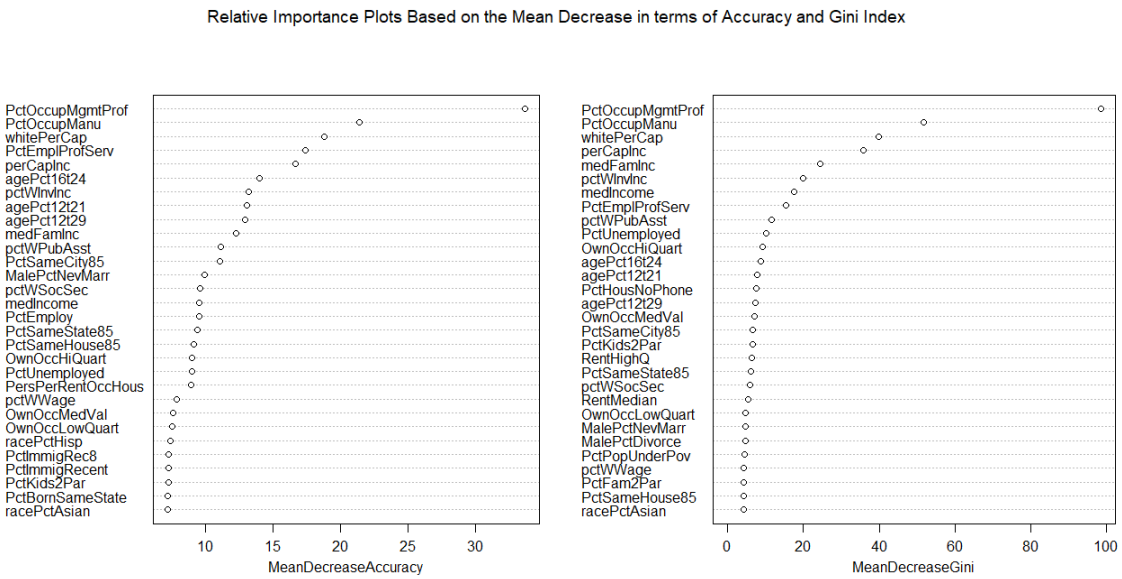
Figure A9. The Classification Tree method's partial output: "tree" package (top output and plots) and "rpart" package (bottom plot)

```
n= 1301

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 1301 461 0 (0.645657187 0.354342813)
    2) PctOccupMgmtProf< 30.32 857  62 0 (0.927654609 0.072345391)
      4) agePct12t21< 19.605 813  35 0 (0.956949569 0.043050431)
        8) PctOccupMgmtProf< 28.075 697   3 0 (0.995695839 0.004304161)
         16) PctHousOccup>=57.2 694   0 0 (1.000000000 0.000000000) *
         17) PctHousOccup< 57.2 3   0 1 (0.000000000 1.000000000) *
        9) PctOccupMgmtProf>=28.075 116  32 0 (0.724137931 0.275862069)
         18) PctSameHouse85>=46.31 75   9 0 (0.880000000 0.120000000)
           36) pctWRetire>=15.32 57   1 0 (0.982456140 0.017543860)
             72) agePct12t21>=10.83 56   0 0 (1.000000000 0.000000000) *
             73) agePct12t21< 10.83 1   0 1 (0.000000000 1.000000000) *
           37) pctWRetire< 15.32 18   8 0 (0.555555556 0.444444444)
             74) agePct16t24>=12.025 10   1 0 (0.900000000 0.100000000)
               148) agePct16t24< 16.435 9   0 0 (1.000000000 0.000000000) *
               149) agePct16t24>=16.435 1   0 1 (0.000000000 1.000000000) *
             75) agePct16t24< 12.025 8   1 1 (0.125000000 0.875000000)
               150) householdsize>=2.96 1   0 0 (1.000000000 0.000000000) *
               151) householdsize< 2.96 7   0 1 (0.000000000 1.000000000) *
         19) PctSameHouse85< 46.31 41  18 1 (0.439024390 0.560975610)
           38) PctBornSameState< 43.265 12   1 0 (0.916666667 0.083333333)
             76) HispPerCap< 11441.5 11   0 0 (1.000000000 0.000000000) *
             77) HispPerCap>=11441.5 1   0 1 (0.000000000 1.000000000) *
           39) PctBornSameState>=43.265 29   7 1 (0.241379310 0.758620690)
             78) PctWorkMomYoungKids< 59.335 8   2 0 (0.750000000 0.250000000)
               156) population< 103055 6   0 0 (1.000000000 0.000000000) *
               157) population>=103055 2   0 1 (0.000000000 1.000000000) *
             79) PctWorkMomYoungKids>=59.335 21   1 1 (0.047619048 0.952380952)
               158) population< 11701 1   0 0 (1.000000000 0.000000000) *
               159) population>=11701 20   0 1 (0.000000000 1.000000000) *
      5) agePct12t21>=19.605 44  17 1 (0.386363636 0.613636364)
        10) pctWPubAsst>=6.965 17   2 0 (0.882352941 0.117647059)
          20) PctSameHouse85>=39.95 14   0 0 (1.000000000 0.000000000) *
          21) PctSameHouse85< 39.95 3   1 1 (0.333333333 0.666666667)
            42) householdsize>=3.115 1   0 0 (1.000000000 0.000000000) *
            43) householdsize< 3.115 2   0 1 (0.000000000 1.000000000) *
        11) pctWPubAsst< 6.965 27   2 1 (0.074074074 0.925925926)
          22) householdsize< 2.56 1   0 0 (1.000000000 0.000000000) *
          23) householdsize>=2.56 26   1 1 (0.038461538 0.961538462)
            46) householdsize>=3.77 1   0 0 (1.000000000 0.000000000) *
            47) householdsize< 3.77 25   0 1 (0.000000000 1.000000000) *
    3) PctOccupMgmtProf>=30.32 444  45 1 (0.101351351 0.898648649)
      6) OwnOccHiQuart< 83350 28  10 0 (0.642857143 0.357142857)
```

Figure A10. The Bagging method's partial output: features used in the model development.



	MeanDecreaseAccuracy	MeanDecreaseGini
population	4.015030	1.67369362
householdsize	4.401126	1.00558392
racepctblack	4.292726	0.88888896
racePctWhite	5.111720	1.18514404
racePctAsian	5.291214	2.08900748
racePctHisp	7.612488	1.77981023
agePct12t21	13.630015	8.52737903
agePct12t29	13.096527	8.87935332
agePct16t24	16.075864	10.47817879
agePct65up	7.391787	2.49950136
pctUrban	1.698495	0.22885814
medIncome	6.780676	13.33681568
pctWWage	7.713565	2.95931771
pctWFarmSelf	3.181993	1.58412615
pctWInvInc	9.995827	14.67901639
pctWSocSec	9.254551	4.52943771
pctWPubAsst	9.482455	11.69279914
pctWRetire	8.560130	3.16450451

Figure A11. The Random Forest method's partial output: the numerical measures and relative importance plots based on the mean decrease in terms of the Accuracy and Gini Index.

	var	rel.inf
PctOccupMgmtProf	PctOccupMgmtProf	86.13894330
agePct16t24	agePct16t24	2.00506474
whitePerCap	whitePerCap	1.82825694
agePct12t29	agePct12t29	1.82138377
PctOccupManu	PctOccupManu	1.57992706
agePct12t21	agePct12t21	1.51035972
PctEmploy	PctEmploy	0.85576614
pctWSocSec	pctWSocSec	0.61066332
PctSameState85	PctSameState85	0.50022505
PctEmplProfServ	PctEmplProfServ	0.45940675
perCapInc	perCapInc	0.40084760
PctSameCity85	PctSameCity85	0.32942034
HousVacant	HousVacant	0.31150087
pctWPubAsst	pctWPubAsst	0.27536900
PctBornSameState	PctBornSameState	0.21242786
PctVacMore6Mos	PctVacMore6Mos	0.20370994
population	population	0.18058417
PctImmigRec5	PctImmigRec5	0.14281555
PctHousOwnOcc	PctHousOwnOcc	0.09558133
PctHousOccup	PctHousOccup	0.09487887
pctWWage	pctWWage	0.08841822
PctImmigRecent	PctImmigRecent	0.08229854
LandArea	LandArea	0.07903266
PctImmigRec8	PctImmigRec8	0.06761793
PersPerRentOccHous	PersPerRentOccHous	0.06709862
MedOwnCostPctInc	MedOwnCostPctInc	0.05840171
householdsize	householdsize	0.00000000
racepctblack	racepctblack	0.00000000
racePctWhite	racePctWhite	0.00000000

Figure A12. The Boosting method's partial output: relative feature importance.

References

[1] Redmond, M. (2011). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Communities%20and%20Crime%20Unnormalized]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Gallagher, R.J., Reagan, A.J., Danforth, C.M., Dodds, P.S. (2018.) Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. PLoS ONE 13(4): e0195644. https://doi.org/10.1371/journal.pone.0195644

[3] Williams, C. (2020). How to Create a Correlation Matrix with Too Many Variables in R. *Towards Data Science.* https://towardsdatascience.com/how-to-create-a-correlation-matrix-with-too-many-variables-309cc0c0a57

[4] Wei, T., Simko, V. (2021). R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.90), https://github.com/taiyun/corrplot

[5] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.

[6] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. New York: Springer.

[7] Friedman, J., Hastie, T., Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software, 33(1), 1–22. https://www.jstatsoft.org/v33/i01/.

[8] Venables, W. N., Ripley, B. D. (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

[9] Ripley, B. (2021). tree: Classification and Regression Trees. R package. Version: 1.0-4. https://CRAN.R-project.org/package=tree

[10] Therneau, T., Atkinson, B., Ripley, B. (2019). rpart: Recursive Partitioning and Regression Trees. Version: 4.1-15. https://CRAN.R-project.org/package=rpart

[11] Peters, A., Hothorn, T., Ripley, B. D., Therneau, T., Atkinson, B. (2019). ipred: Improved Predictors. Version: 0.9-12. https://CRAN.R-project.org/package=ipred

[12] Greenwell, B., Boehmke, B., Cunningham, J., GBM Developers. (2020). gbm: Generalized Boosted Regression Models. Version: 2.1.8 https://CRAN.R-project.org/package=gbm

[13] Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[14] Wickham H., François, R., Henry, L., Müller, K. (2021). dplyr: A Grammar of Data Manipulation. Version: 1.0.7. https://CRAN.R-project.org/package=dplyr

[15] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

[16] Wickham, H. (2021). tidyr: Tidy Messy Data. Version: 1.1.3. https://CRAN.R-project.org/package=tidyr

[17] Wickham, H., Bryan, J., et al. (2021). readxl: Read Excel Files. Version: 1.3.1. https://CRAN.R-project.org/package=readxl

[18] Wickham, H., Hester, J., François, R., et al. (2021). readr: Read Rectangular Text Data. Version: 2.0.1. https://CRAN.R-project.org/package=readr