Determining the Importance of Variables for Crime Rates in Chicago by Random Forest

Dong Gai – Dec 14, 2022

University of Wisconsin – Madison

Introduction

How safe is Chicago? According to the article "What You Should Know About Chicago Crime Rates", R.E. says that "Chicago has a crime rate of 3926 per 100,000 people. That's 67% higher than the national average – and violent crime plays an outsized role in those numbers." For Chicago, we found that the violent crime, such as murder, rape, armed robbery, and aggravated assault, is one of the highest in the nation. According to Neighborhood Scout's analysis of FBI reported crime data, your chance of becoming a victim of one of these crimes in Chicago is one in 103. Neighborhood Scout's analysis also reveals that Chicago's rate for property crime is 25 per one thousand population.

However, Chicago is a very famous and attractive American metropolis. Whether it is the architecture of the city, the popular museums, or the food, people love Chicago but fear its high crime rate. In this case, I want to use the random forest to determine the importance of variables that contribute most to the high crime rate. Then police and residents can understand where crimes are more likely to occur and get better solutions for high crime rates.

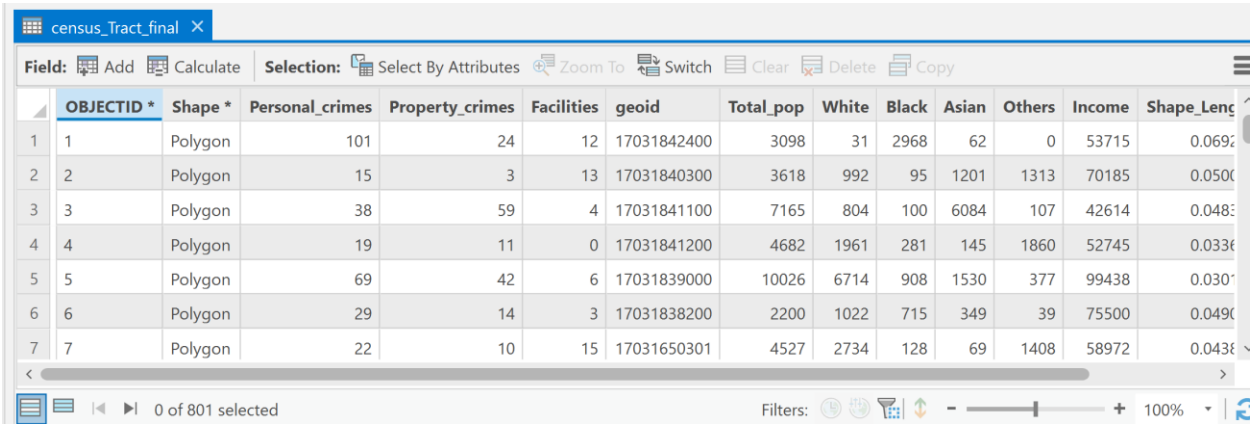
Data

Different type of crime rate for each crime case such as violent crime, property crime, the median income for household and population in census tract level and race distribution

from National Historical GIS, and number of facilities in public space (such as basketball court, public fitness area), and the Chicago census tract boundary from Chicago data Portal. These are all the variables we will use. Class: violent and property crimes. Including rape, murder, robbery and assault, theft. Variables: race, median income (house price), race distribution, number of facilities.

Method

Firstly, I integrated all the data in arcgis pro. The point data crimes and facilities data were aggregated into the census tract boundary layer using spatial join. Then the race population table and the median income table are also joined to the census tract boundary layer by joining table method. In this case, as we can see in Figure 1, all the data are aggregated into the attribute table of the census tract.



	OBJECTID *	Shape *	Personal_crimes	Property_crimes	Facilities	geoid	Total_pop	White	Black	Asian	Others	Income	Shape_Lenc
1	1	Polygon	101	24	12	17031842400	3098	31	2968	62	0	53715	0.0692
2	2	Polygon	15	3	13	17031840300	3618	992	95	1201	1313	70185	0.0500
3	3	Polygon	38	59	4	17031841100	7165	804	100	6084	107	42614	0.0483
4	4	Polygon	19	11	0	17031841200	4682	1961	281	145	1860	52745	0.0336
5	5	Polygon	69	42	6	17031839000	10026	6714	908	1530	377	99438	0.0301
6	6	Polygon	29	14	3	17031838200	2200	1022	715	349	39	75500	0.0490
7	7	Polygon	22	10	15	17031650301	4527	2734	128	69	1408	58972	0.0438

Figure 1

In addition, in R, I did multilinear regression analysis for crime table in order to find which variables have the strong correlation and which variables have the positive or negative correlation.

Finally, I separate the data to training dataset and the testing dataset. Then using the training dataset to build up the random forest model. Then using the testing dataset to train the model and find the least square errors.

Results

According to the figure 2, which shows the standardized regression coefficient for personal crimes. As we can see, in terms of personal crimes, the total population and the household income have the negative coefficient, which means the area with higher income and population will cause lower crime rates.

```
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[2] * sd(data$Facilities) / sd(data$Personal_c)
> print(std_coef_MedInc)
data$Facilities
     0.02965397
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[3] * sd(data$Total_pop) / sd(data$Personal_c)
> print(std_coef_MedInc)
data$Total_pop
    -0.2575915
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[4] * sd(data$White_alon) / sd(data$Personal_c)
> print(std_coef_MedInc)
data$White_alon
      0.3137058
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[5] * sd(data$Black_Amer) / sd(data$Personal_c)
> print(std_coef_MedInc)
data$Black_Amer
        0.83516
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[6] * sd(data$Asian) / sd(data$Personal_c)
> print(std_coef_MedInc)
data$Asian
 0.1753465
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[7] * sd(data$Others) / sd(data$Personal_c)
> print(std_coef_MedInc)
data$Others
  0.1967316
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[8] * sd(data$Median_Inc) / sd(data$Personal_c)
> print(std_coef_MedInc)
data$Median_Inc
     -0.1908156
```

Figure 2

For the property crimes, I also found that the total population and the household income have the negative coefficient, but the household income is less important and the coefficient is -0.05. In addition, the population of black and American Africa have the great impact on crime rate. (Figure 3)

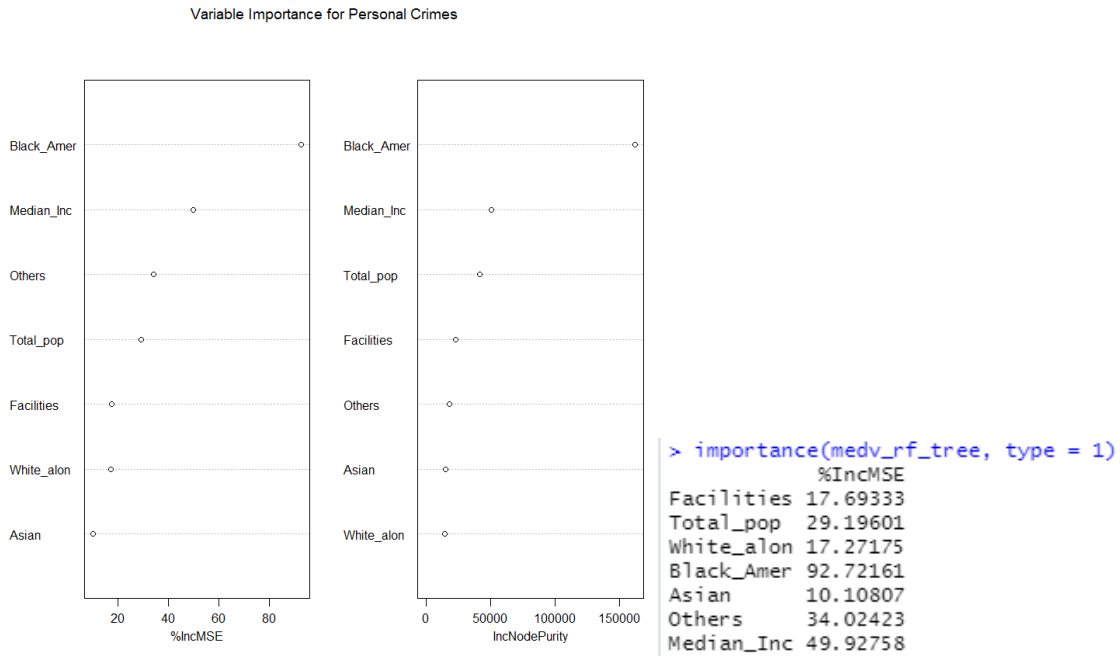
```
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[2] * sd(data$Facilities) / sd(data$Property_c)
> print(std_coef_MedInc)
data$Facilities
     0.03409551
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[3] * sd(data$Total_pop) / sd(data$Property_c)
> print(std_coef_MedInc)
data$Total_pop
    -0.3376055
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[4] * sd(data$White_alon) / sd(data$Property_c)
> print(std_coef_MedInc)
data$White_alon
      0.3422417
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[5] * sd(data$Black_Amer) / sd(data$Property_c)
> print(std_coef_MedInc)
data$Black_Amer
      0.7175413
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[6] * sd(data$Asian) / sd(data$Property_c)
> print(std_coef_MedInc)
data$Asian
 0.3135848
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[7] * sd(data$Others) / sd(data$Property_c)
> print(std_coef_MedInc)
data$Others
  0.1537551
> # Calculate the Standardized Regression Coefficient
> std_coef_MedInc <- coefficients_lm[8] * sd(data$Median_Inc) / sd(data$Property_c)
> print(std_coef_MedInc)
data$Median_Inc
    -0.05215922
```

Figure 3

After did the random forest model, I plot the importance of the variable. Figure 4 is the importance of variables for personal crimes. We can easily find that the black and American African population and the household income play an essential role in crime rates. Asian's population is least important.

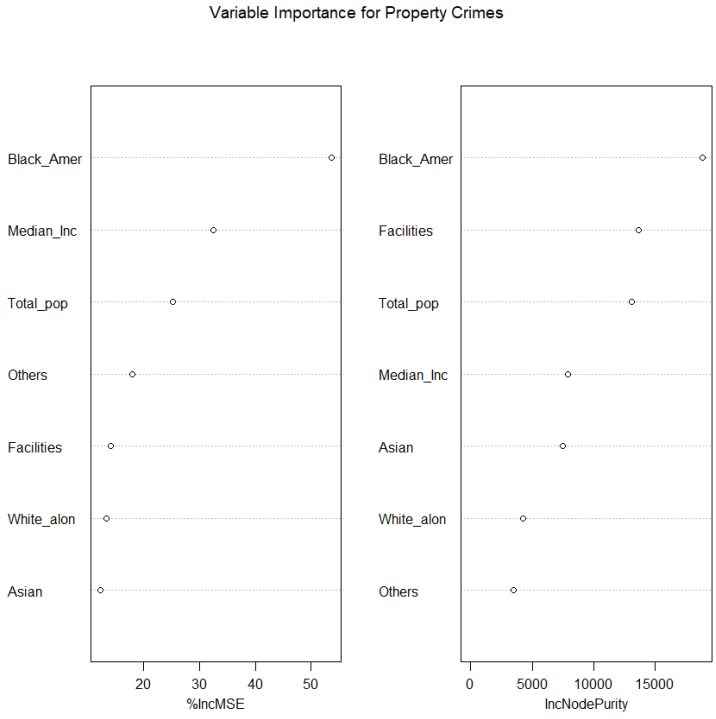
Figure 5 is the importance of variables for property crimes. We can see that the importance of each variable is roughly the same as that of another crime, only the total population is more important than other races, such as Hawaiian.

In conclusion, according to the two models, I found that an area with higher population of black people or American African, or other races (Hawaiian) will cause more crimes. And an area with higher household incomes and total population will cause lower crimes. These several variables play an essential role in the Chicago's crime rate.




```
> importance(medv_rf_tree, type = 1)
            %IncMSE
Facilities 17.69333
Total_pop  29.19601
White_alon 17.27175
Black_Amer 92.72161
Asian      10.10807
Others     34.02423
Median_Inc 49.92758
```

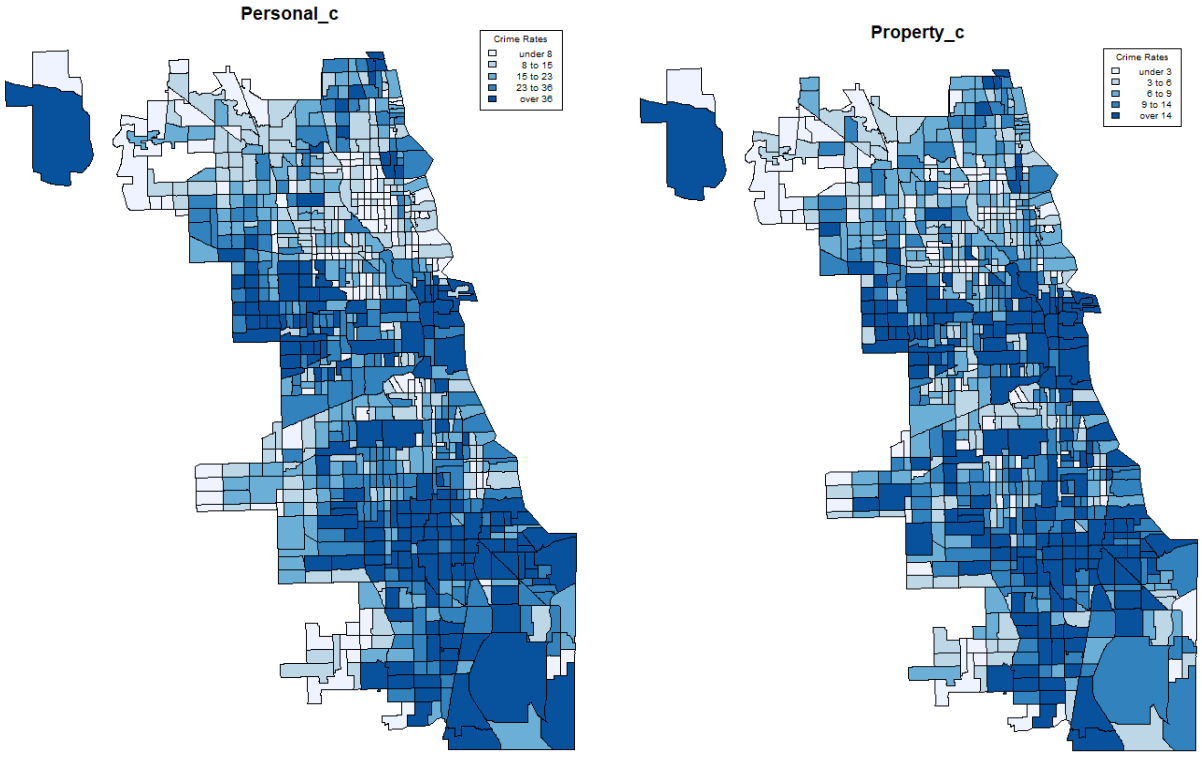
Figure 4




```
> importance(medv_rf_tree, type = 1)
            %IncMSE
Facilities 14.09123
Total_pop  25.22662
White_alon 13.35875
Black_Amer 53.73800
Asian      12.24309
Others     17.96712
Median_Inc 32.51486
```

Figure 5

The Choropleth Map for Person and Property Crimes



Reference

Fulton, R.E. “What You Should Know about Chicago Crime Rates.” *GetJerry.com*, Jerry, Inc., 29 Nov. 2022, https://getjerry.com/car-insurance/chicago-crime-rates.

“Chicago, IL Real Estate & Demographic Data.” *NeighborhoodScout*, https://www.neighborhoodscout.com/il/chicago.