Chicago Crime Dataset with Census Information

October 30, 2018

Introduction

Datasets:

- Chicago crime dataset
  - It includes incidents for crime that occurred in the city of Chicago from 2001 to August 2018
  - The total number of cases is over 6.7 million, and the number of cases that has at least one NULL value is about 0.7 mil.
- Chicago census information
  - It is grouped by 77 Chicago community areas
  - It is gathered between 2011 and 2015

❑ Purposes

Use the Chicago crime dataset and the Chicago census information for Chicago community areas to

- Explore the relationship between the crime rate and important demographic factors
- Build a predictive model with these factors
- Cluster community areas based on their important demographic factors

❑ Clients

The Chicago city government would be the clients for this project. The results from the project can be utilized:

- Understand the correlation between the crime rate per community area and an individual demographic factor independently
- Understand the correlation between the crime rate per community area and all of the available demographic factors together
- Use the predictive model to produce a crime rate
- Use the clustering model to identify if a community area potentially falls into a group of community areas with the highest average crime rate
- Mobile resources more efficiently and effectively and improve areas in reducing the crime rate by community area

❑ Data Sources

- The Chicago crime data is available for download from the Chicago Data Portal:

  https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

- The Chicago census information by community areas is available for download from the below link:

  https://datahub.cmap.illinois.gov/dataset/1d2dd970-f0a6-4736-96a1-3caeb431f5e4/resource/8c4e096e-c90c-4bef-9cf1-9028d094296e/download/ReferenceCCA20112015.csv

Datasets and Data Wrangling

Data Science Techniques Utilized:

- Data Wrangling
- Inferential Statistics
- Linear Regressions
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
- Clustering - KMeans

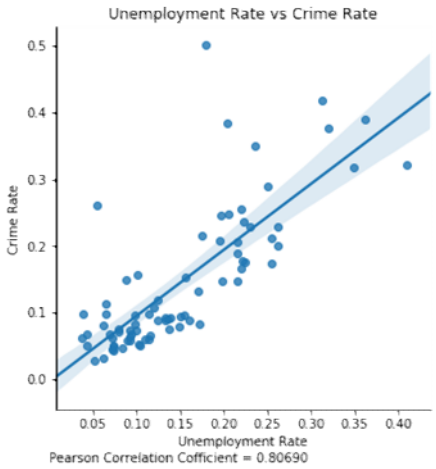
❑ Data Wrangling

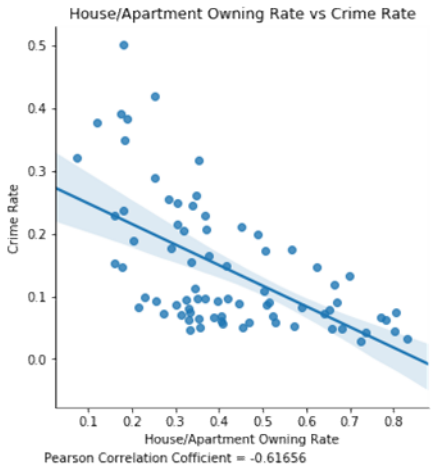
- Overall, the Chicago crime dataset is quite clean. With data exploration, some cleanups were conducted
- Remove rows containing at least one null value (less than 10% of rows are removed, Majority of the NULL values happen for the fields 'Ward' and 'Community Area')
- Format data – for instance, converted the Date column from the string format to the date-time format
- Remove duplicates
- Remove outliers
- The Chicago census information is clean
- Focus on the crime data in 2015 only because the census information was gathered between 2011 – 2015
- The number crime cases in 2015 are grouped by community area in order to combine them with the census information
- Convert some factors from actual numbers to ratios
  - Crime number to Crime Rate over population 19 years or old per community area
  - High school diploma or higher number to high school diploma or higher rate over population 19 years or old per community area
  - Bachelor's degree or higher number to Bachelor's degree or higher rate over population 19 years or old per community area
- Make two factors less than 1.0 on par with other factors
  - Median age per community area divided by 100.0
  - Median income per community area divided by 200000.00

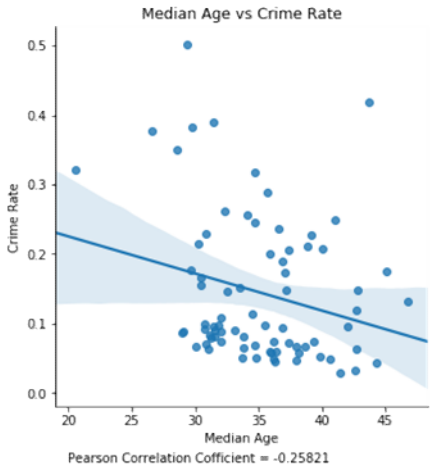
Crime Rate vs Individual Demographic Factor (I)

Observations:

- By reviewing the Crime Rate and a demographic factor independently, the Unemployment Rate has the strongest correlation with the Crime Rate.
- The Unemployment Rate is the only factor having the positive correlations. The rest factors have a negative correlation with the Crime Rate for each Community Area.
- With this analysis, it seems that reducing unemployment rate is the most effective way to lower the crime for a community area.







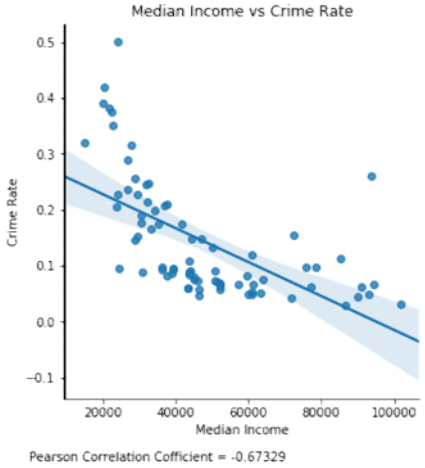
Notes:

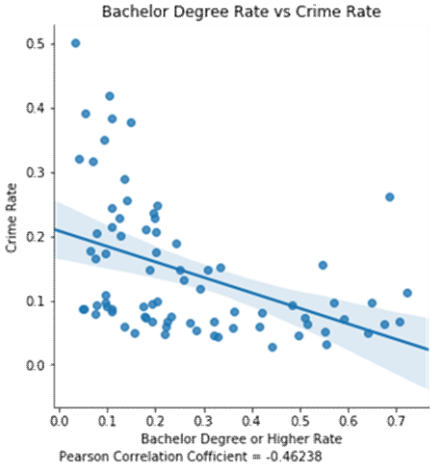
1. In the data analysis, some factors may inter-dependent.
2. The Pearson correlation coefficients for some factors may be stronger if their logarithmic values are utilized.

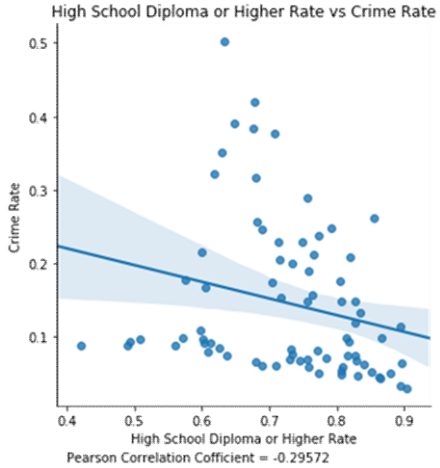
Crime Rate vs Individual Demographic Factor (II)

Observations:

- Median Income for a community area seems to play a quite important role for its Median Income.
- High School Diploma or Higher Rate has slightly negative relationship with the Crime Rate for a community area.
- With this analysis, it seems that reducing unemployment rate is the most effective way to lower the crime for a community area.







Notes:

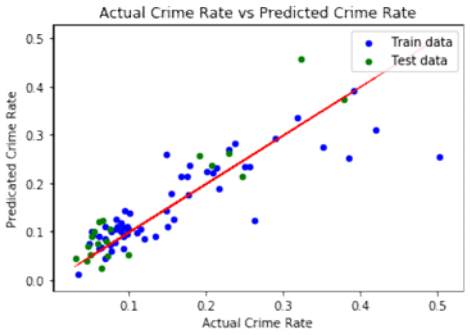
1. In the data analysis, some factors may inter-dependent.
2. The Pearson correlation coefficients for some factors may be stronger if their logarithmic values are utilized.

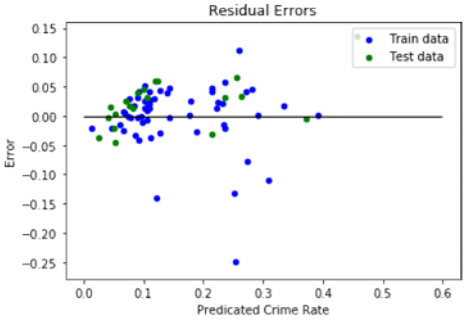
Crime Rate vs All of Demographic Factors

Observations:

- The estimated coefficients in the table and the graphs are produced through Ridge regression.
- The model for predicting the Crime Rate by using all of six demographic factors works well. There is a small number of outliers.
- Based on the coefficients, the Unemployment Rate plays an very import role in the Crime Rate. Please note that these factors have different average values and variances.

Factors	Estimated Coefficient
(Intercept)	-0.054676
HS_RATE	0.297241
BACH_RATE	-0.338422
UNEMP_RATE	0.828602
OWN_OCC_HU_RATE	-0.396490
MED_AGE_N	-0.035262
MEDINC_N	0.558467





Notes:

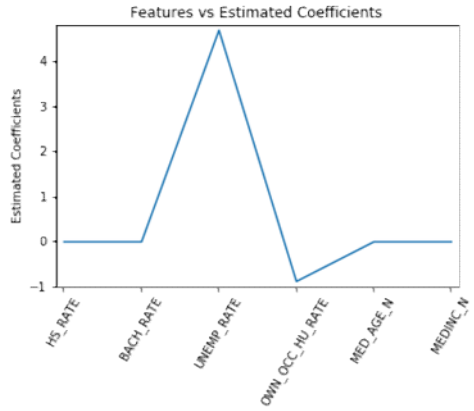
1. Both LinearRegression and Ridge regression are utilized. The latter yields better results, which is the reason here that the results from Ridge regression are presented here.
2. 5-fold cross validation is performed for the regression.
3. The 77 community area data is split into two sets: one for training and one for test. From the above, the predicated crime rate based on the test data set performs well.

Crime Rate vs Dominant Factors

Observations:

- The purpose here is to select dominant factors that have the largest impact on the crime rate.
- From the table and graph, it is obvious that the Unemployment Rate plays the most role in the Crime Rate. The House/Apartment Owning Rate is the second important factors.
- In order to reduce the Crime Rate effectively for a community area, it is sensible to lower the Unemployment Rate and increase the House/Apartment Owning Rate.

Factors	Estimated Coefficient
HS_RATE	-0.000000
BACH_RATE	-0.000000
UNEMP_RATE	4.683478
OWN_OCC_HU_RATE	-0.876674
MED_AGE_N	-0.000000
MEDINC_N	-0.000000

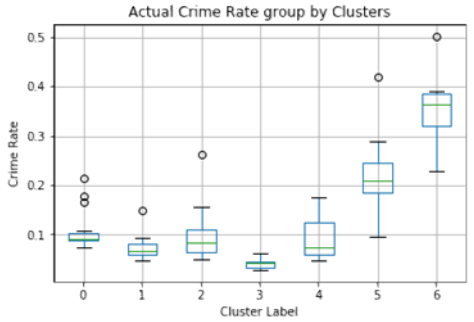


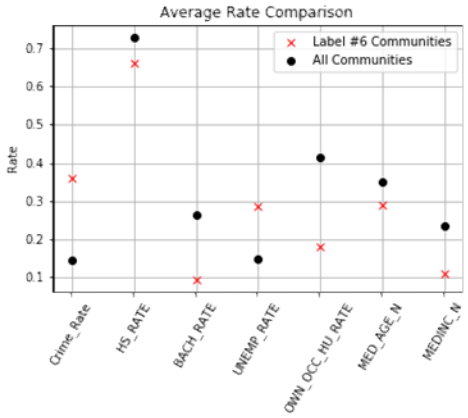
Notes:
1 . Lasso regression is utilized. The normalize is set to True for the training.

Clustering Community Areas

Observations:

- The purpose is to group community areas based on their six demographic factors and identify which one has the highest average crime rate.
- Based on the graph in the middle, the average for all of factors except Median Age for the #6 group are worse than that for all community areas.
- The table on the right lists the community areas in the #6 group. If a community area has similar factors in values, it is very likely it will fall into the #6 group.





Community Area	GEOG	Crime Rate
26	West Garfield Park	0.502303
27	East Garfield Park	0.383683
29	North Lawndale	0.350840
40	Washington Park	0.377099
42	Woodlawn	0.228956
54	Riverdale	0.321468
67	West Englewood	0.317384
68	Englewood	0.390983

Notes:
1. Unsupervised training – KMeans is used. Through multiple tries and comparison, using 7 clusters is the best.
2. The training takes into account standardization because each factor has different mean and variance.

Conclusions and Future Considerations

Deliverables:

- Jupyter Books: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20Data%20Wrangling.ipynb

  https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20Data%20Storytelling.ipynb

  https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20MachineLearning.ipynb

- Project Summary: https://github.com/DavidFlanders/Capstone/blob/master/Capstone%20Project%20Summary.pdf

❑ Conclusions

- It is strongly recommended to the Chicago City Government that the unemployment rate is the top factor impacting the crime rate for a community area. They should strive to lower this rate as top priority in order for them to reduce the crime rate.
- The Chicago City Government should also to improve other house/apartment owning rate, and encourage college eduction.
- For the high crime rate community areas (in the #6 group), they may want to deploy more policemen there as interim solutions for reducing their crime rate.

❑ Future Considerations

- If the Chicago City government can provide more demographic information for other factors for a specific year, the data analysis would be more comprehensive and potentially identify more factors for safety improvement.
- If the Chicago City government can provide the demographic information year by year, the data analysis can be conducted accordingly to understand the impact of factor changes on the rise and decline of the crime rate.