Capstone Project

Impact Analyses of Census Features on Chicago Crime Rates through Machine Learning

October 2018

Contents

I. Introduction

The Chicago Data Portal publishes a dataset reflecting reported incidents of crime that occurred in the City of Chicago from 2001 to August 2018. The CMAP Data Hub provides Chicago community area census information aggregated through the period of 2011 to 2015.

In this project, I examined the data, conducted data wrangling, utilized different charts to identify general trends for the crime cases in Chicago, and applied inferential statistics techniques to the data. Further, I conducted broader correlation analyses between the crime cases and census features such as population, eduction, unemployment etc. In the last part of the project, I have applied some supervised machine learning techniques to the data and built a predictive model by using demographic features to predict the crime rate per community area. In addition, I have also utilized an unsupervised machine learning technique to label community areas in Chicago and identify a cluster having the highest average crime rate.

The Chicago city government would be the clients of this project. They can use my analysis report to understand the impact of demographic features on the crime rate, and which features play critical roles in impacting the crime rate. They can also use the built predictive model to identify the crime rate for a community area, and focus on a group of community areas with the highest average crime rate. Subsequently, they can mobilize resources more efficiently and effectively, and improve relevant areas in reducing the crime rate for a Chicago community area.

II. Data Sources

1. The Chicago crime case data is available for download from the Chicago Data Portal as follows:

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

2. Chicago community area data covering population, eduction, income etc. 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

III. Data Exploration and Data Wrangling

1. Data Wrangling:

Overall, the Chicago crime data is quite clean. With further exploration, some cleanups, mapping and transformation have been conducted. The Jupyter Notebook is available in Github: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20Data%20Wrangling.ipynb

- Removed rows containing at least one null value

The total number of cases is over 6.7 million, and the number of cases that have at least one NULL value is about 0.7 mil. Majority of the NULL values are for the fields 'Ward' and 'Community Area'. Since the latter is only about 10% of the total cases, their removal have little impact on the data analysis for the whole dataset.

- Formatted date

  a. Some field names such as 'Community Area' include a space. I replaced space with '_'.
  b. The fields of 'District', 'Ward', and 'Community_Area' are float64 data. Actually, they are just numeric codes containing no decimal point. I converted them to the int64 format.
  c. I changed the fields of 'IUCR', 'Primarhy_Type', and 'FBI_Code' to the categorical variables.
  d. I converted the field 'Date' from the string format to the date-time format. Since the field 'Year' can be easily extracted from the 'Date', I removed the field for 'Year' only.

- Removed duplicates

The field 'ID' is unique, and the field 'Case_Number' included 378 duplicates. I removed the latter from the dataset.

- Removed outliers

I used the scatter plot to examine 'X_Coordinate' and 'Y_Coordinate', and find 107 outliers. I removed these outliers from the dataset.

- Saved the clean file

The dataset after the data wrangling was saved to a new file for data analysis.

2. Data Story Telling:

I utilized different plots to identify patterns of the crime cases. The Jupyter Notebook is available in Github: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20Data%20Storytelling.ipynb

Below is the list of major trends and patterns identified:

- The total number of crime cases decreased over the years from 2003 to 2015. However, the total number of crime cases stays flat between 2015 and 2018.
- The number of crime cases drops to the lowest point around January or February each year, and the number of crime cases rises to the highest point around July each year.
- The number of cases is the lowest between 4:00 – 6:00am. The peak hours for the largest number of crime cases varies with primary types of crime cases.
- The top three primary types of crime cases are THEFT, BATTERY and CRIMINAL DAMAGE.
- The Community Area #25 has the largest number of crime cases.

3. Data Utilized for Inferential Statistics and Machine Learning

- The Chicago census information is clean.
- Focus on the crime data in 2015 only because the census information was aggregated between 2011 and 2015.
- The number crime cases in 2015 are grouped by community area in order to combine them with the census information.
- Converted some features from actual numbers to ratios:
  a. Crime number to Crime Rate over population 19 years or old per community area
  b. High school diploma or higher number to high school diploma or higher rate over population 19 years or old per community area
  c. Bachelor’s degree or higher number to Bachelor’s degree or higher rate over population 19 years or old per community area
- Made two features less than 1.0 on par with other features
  d. Median age per community area divided by 100.0
  e. Median income per community area divided by 200000.00
- The Crime Rate is the dependent variable.
- Table-1 below lists the six features extracted from the census data, which may impact the Crime Rate. They were utilized for the subsequent data analysis.

Table-1

Feature Name in Jupyter Book	Actual Feature Name
MED_AGE	Median age
HS_RATE	High school diploma or higher rate
BACH_RATE	Bachelor degree or higher rate
UNEMP_RATE	Unemployment rate
MEDINC	Median income
OWN_OCC_HU_RATE	House/Apartment owning rate

IV. Initial Statistics Analyses

1. Correlation between the crime rate and an individual demographic feature

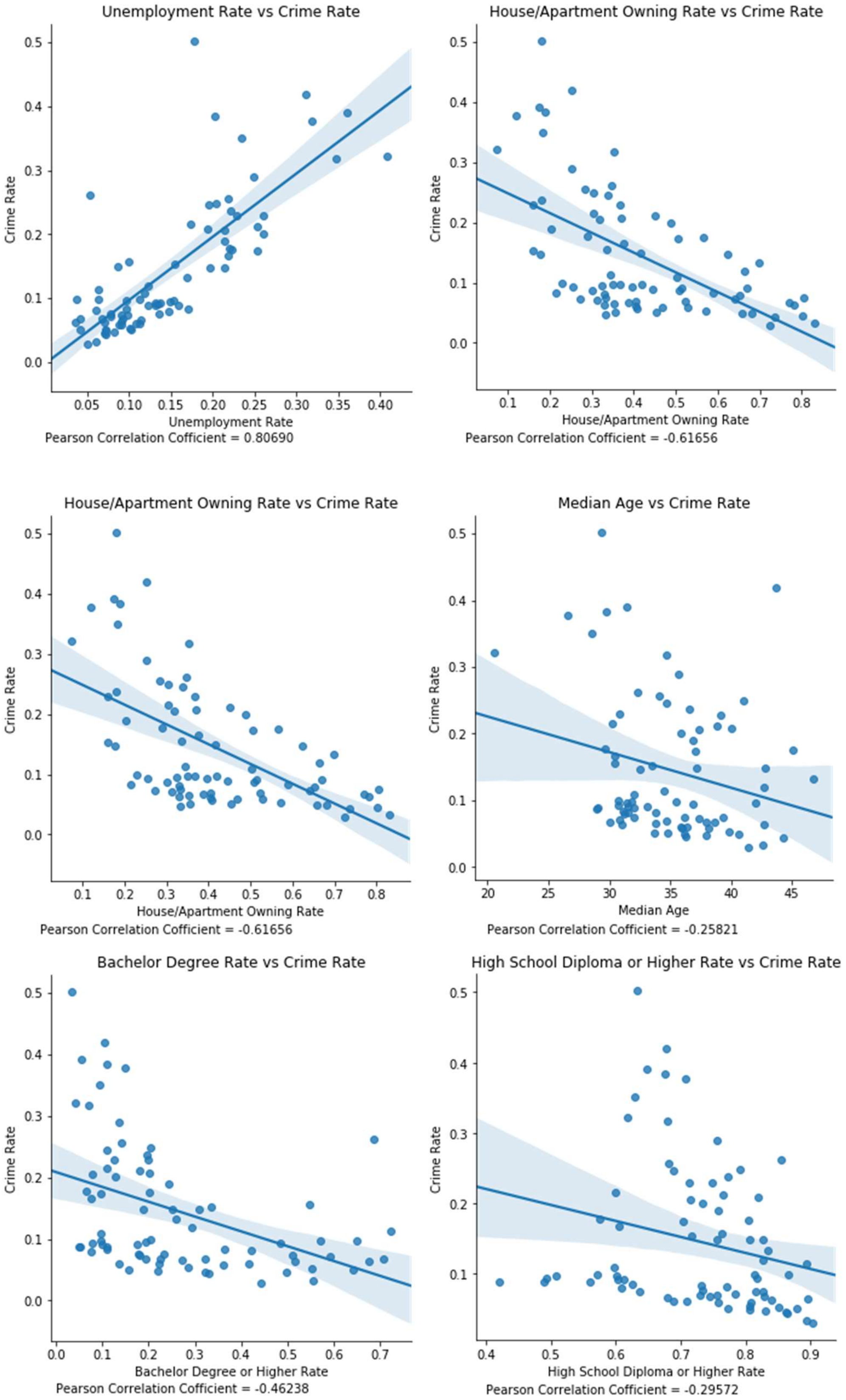
Firstly, the statistics technique was applied to explore the relationship between the crime rate and each of the available demographic features when other features remain unchanged. Table-2 below lists the Pearson correlation coefficients. Pearson Correlation Coefficient or the Bivariate Correlations, is a measure of the linear correlation between two variables.

Table-2

Feature	Pearson Correlation Coefficient between Crime Rate and a Feature
MED_AGE	-0.25821
HS_RATE	-0.29572
BACH_RATE	-0.46238
UNEMP_RATE	0.80690
MEDINC	-0.67329
OWN_OCC_HU_RATE	-0.61656

Figure-1 includes six charts illustrating the linear relationship between the crime rate and each of the demographic features.

Figure-1



Pearson Correlation Cofficient = 0.80690

Pearson Correlation Cofficient = -0.61656

Pearson Correlation Cofficient = -0.61656

Pearson Correlation Cofficient = -0.25821

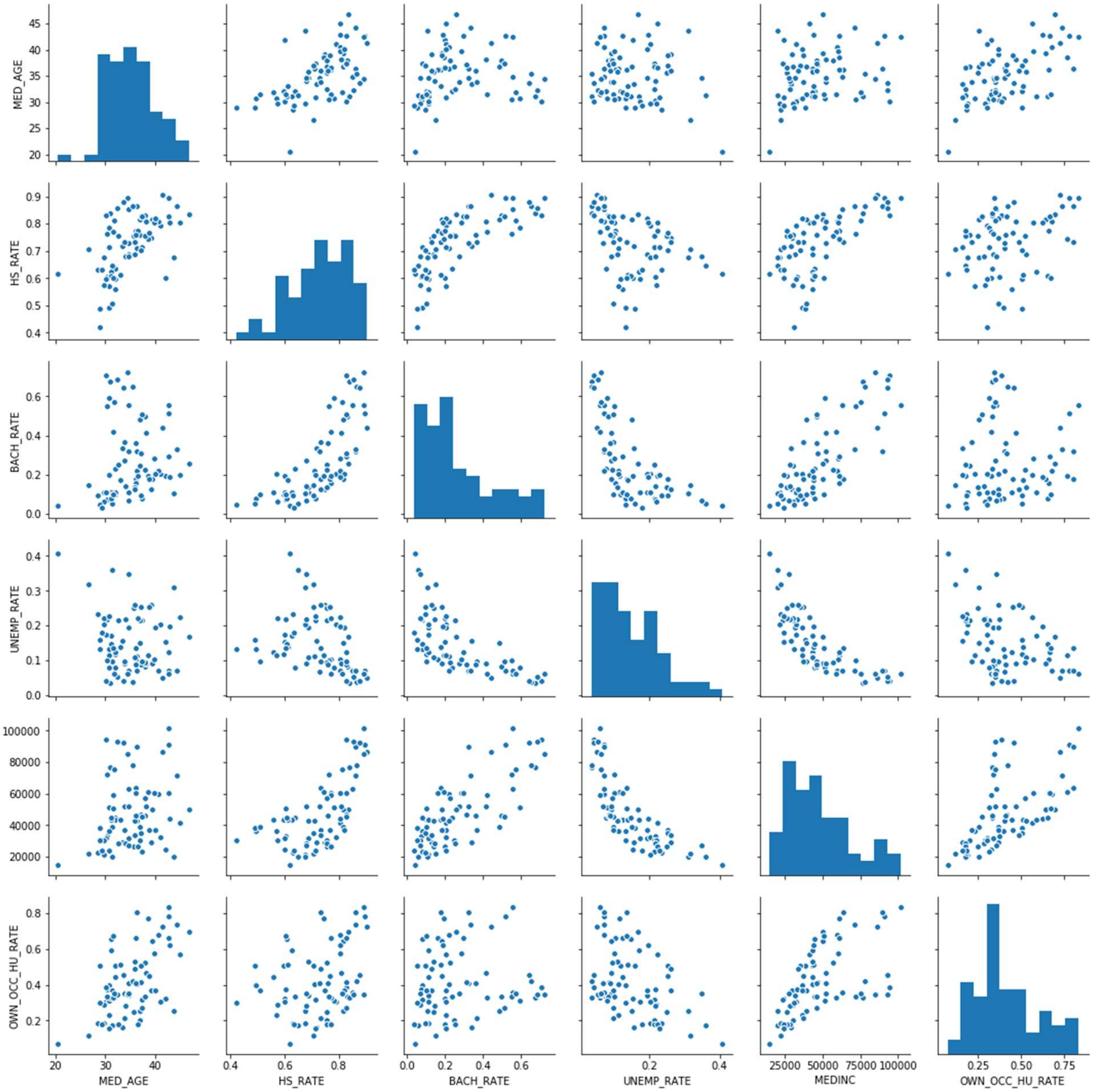
Pearson Correlation Cofficient = -0.46238

Pearson Correlation Cofficient = -0.29572

2. Inter-dependencies among demographic features

Figure-2 display the potential dependencies among demographic features.

Figure-2



3. Observations:

- By reviewing the Crime Rate and a demographic feature independently, the Unemployment Rate has the strongest correlation with the Crime Rate.
- The Unemployment Rate is the only feature having the positive correlations. The rest features 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 Rate for a community area.
- 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 Rate for a community area.
- The Pearson correlation coefficients for some features may be stronger if their logarithmic values are utilized.
- Some features have some correlations to some degree. For example, the Bachelor degree or higher rate and the High school diploma or higher rate have some overlaps.

V. Building Predictive Model with Supervised Machine Learnings

Several Linear regression methods were utilized for analyzing the Crime Rate and all of the aforementioned demographic features.

1. Linear Regression

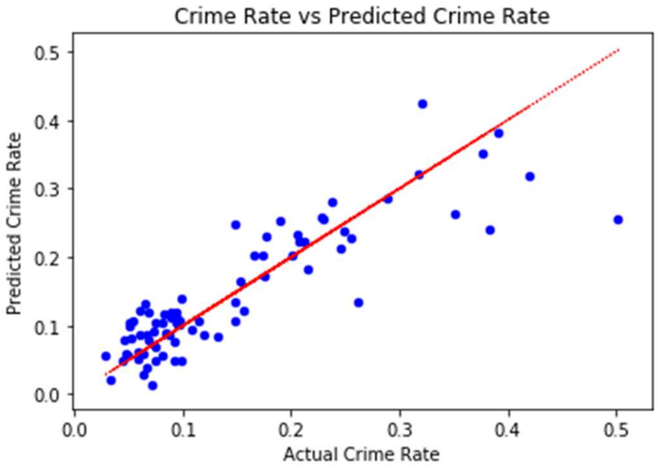
1.1 Linear Regression fits a linear model with coefficients w= (w1,...,wp) to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation. Linear Regression was applied to the Crime Rate and the six demographic features. Table-3 below lists the estimated coefficients for the linear model.

Table-3

Feature	Estimated Coefficient
(Intercept)	-0.116248
HS_RATE	0.280622
BACH_RATE	-0.465539
UNEMP_RATE	0.778362
OWN_OCC_HU_RATE	-0.514992
MED_AGE_N	0.216100
MEDINC_N	0.854811

Figure-3 illustrates the relationship between actual Crime Rate and the predicted Crime Rate.

Figure-3



1.2 The R-squared for this regression is 0.75986, which means that about 75.99% of the Crime Rate change can be explained by these six features. F-statistic=44.300176 indicates that the model built is statistically significant overall.

1.3 5-fold Cross Validation is used. K-Fold Cross-Validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called K that refers to the number of groups that a given data sample is to be split into. The Cross Validation score for the Linear Regression varies quite largely as indicated below.

```
[0.58752454 0.824689   0.19062981 0.54555085 0.50916479]
Average cross validation score = 0.531512
```

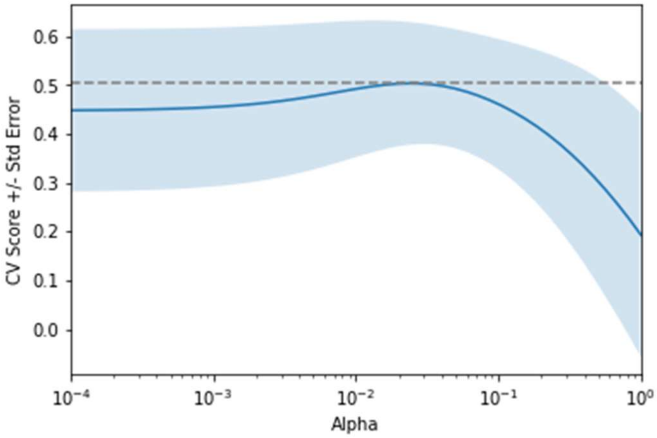
1.4 Observations:

- After putting all features together, the High School Diploma or Higher Rate has now positive correlation with the Crime Rate.
- After putting all features together, the Median Income has positive correlation with the Crime Rate. If considered independently, both have strong negative relationship.
- Based on the scatter plot between the actual Crime Rate and the predicated Crime Rate, it seems that the model works well.
- R-squared is a statistical measure of how close the data is to the regression line. In this case, about 75.99% of the Crime Rate change can be explained by these six variables or features.
- F-statistic indicates that the model is overall statistically significant.
- Through the Cross Validation, the average score is 0.531512. By checking the results, the score varies largely probably because of the small data volume.

2 Ridge Regression

2.1 Ridge regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares and help avoid overfitting. Firstly, display the relationship between Cross-Validation score +/- Standard error and alpha for Ridge Regression. Alpha is a complexity parameter that controls the amount of shrinkage: the larger it is, the greater the amount of shrinkage and thus the coefficient becomes more robust to collinearity. Figure-4 is chart.

Figure-4



2.2 From the chart above, use alpha=0.03. Split the dataset into two parts: training and test, and apply the Ridge regression to the training set. Table-4 below lists the estimated coefficients for the predictive model.

Table-4

Feature	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

The predication score for the test set is 0.78409.
Figure-5 illustrates the actual Crime Rate and the predicted Crime Rate.

Figure-5

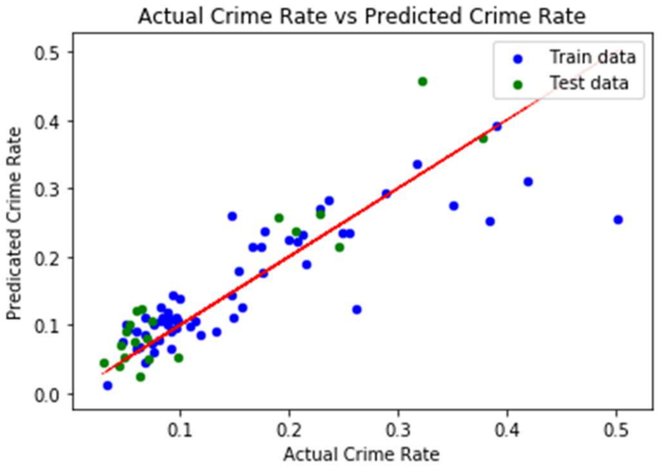
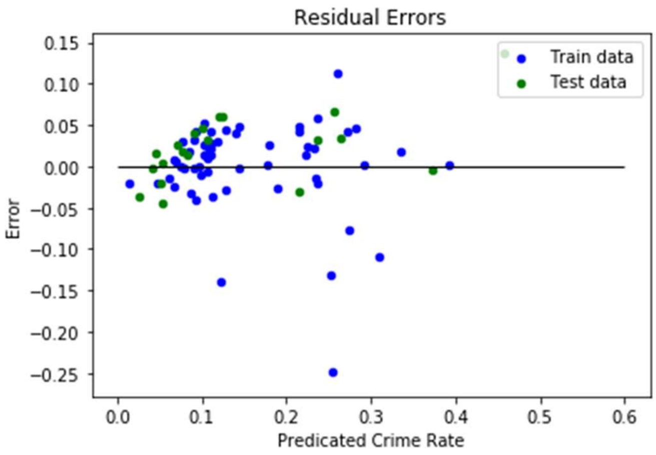


Figure-7 below shows the residual errors for the predicated Crime Rates:

Figure-7



2.3 With the 5-fold Cross Validation, the Cross Validation score varies in a much narrower range than that for the previous Linear Regression. The average score is 0.58675. This regression technique yields better results (model) than the previous Linear Regression.

```
[0.69274765 0.79999792 0.40765601 0.5929476 0.44038259]
Average cross validation score = 0.586746
```

2.4 Observations:

- In comparison with LinearRegression, the Ridge regression seemingly performs better. For example, for the 5-fold Cross Validation, the scores for the former vary from 0.19063 to 0.82469 and the same for the latter vary from 0.40766 to 0.80000. The latter is more consistent. In addition, the mean score for the latter is also higher. The results from the Ridge regression should be used.

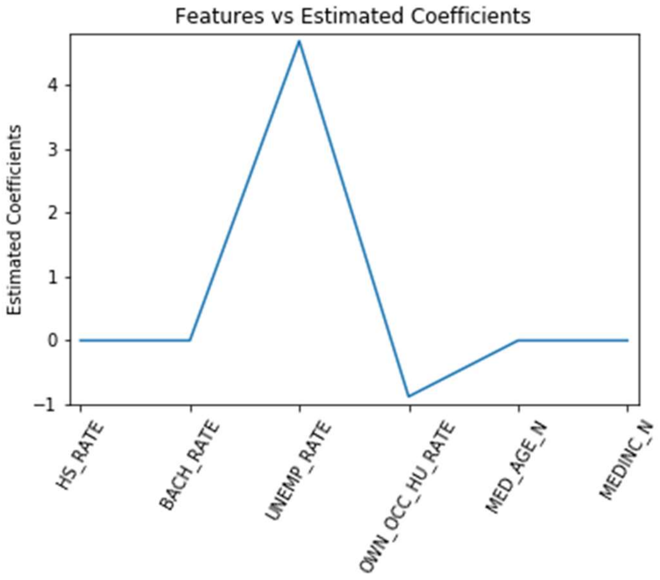
3 Lasso Regression

3.1 The Lasso is a linear model that estimates sparse coefficients. It is useful in some contexts due to its tendency to prefer solutions with fewer parameter values, effectively reducing the number of variables upon which the given solution is dependent. For this reason, the Lasso and its variants are fundamental to the field of compressed sensing. Although the dataset in this project has only has six features, we would like to identify dominant ones. Table-5 lists the estimated coefficients, and Figure-8 shows the features and their estimated coefficients.

Table-5

Feature	Estimated Coefficient
HS_RATE	-0.000000
BACH_RATE	-0.000000
UNEMP_RATE	4.683478
OWN_OCC_HU_RATE	-0.876674
MED_AGE_N	-0.0000000
MEDINC_N	-0.000000

Figure-8



3.2 Observations:

- With this technique, we can conclude that the Unemployment Rate is the dominant feature in determining the Crime Rate, and the House/Apartment Rate is the second important feature impacting the Crime Rate.

VI. Clustering Community Areas with Unsupervised Machine Learnings

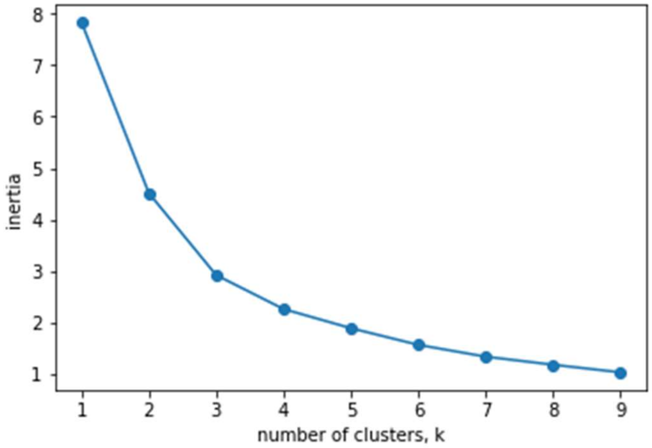
KMeans Clustering technique is utilized only.

1. Unsupervised Machine Learning – KMeans Clustering

1.1 Identify the optimized number of clusters

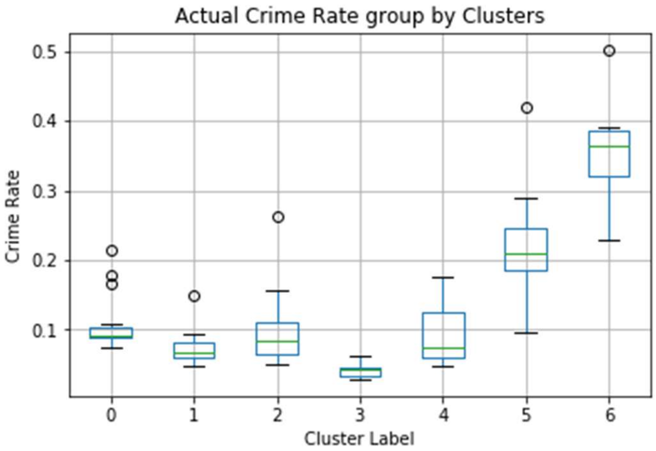
Figure-9 is the graph between inertia and the number of clusters for KMeans, from which it seems k = 5 is the optimized. After several tries (k=5, 6, 7, 8, 9), k=7 is the optimized cluster number based on their boxplots.

Figure-9



1.2 With k = 7 and by taking into account the standardization because each feature has different mean and variance, the technique produces six clusters. Figure-10 shows the actual average Crime Rate by clusters.

Figure-10



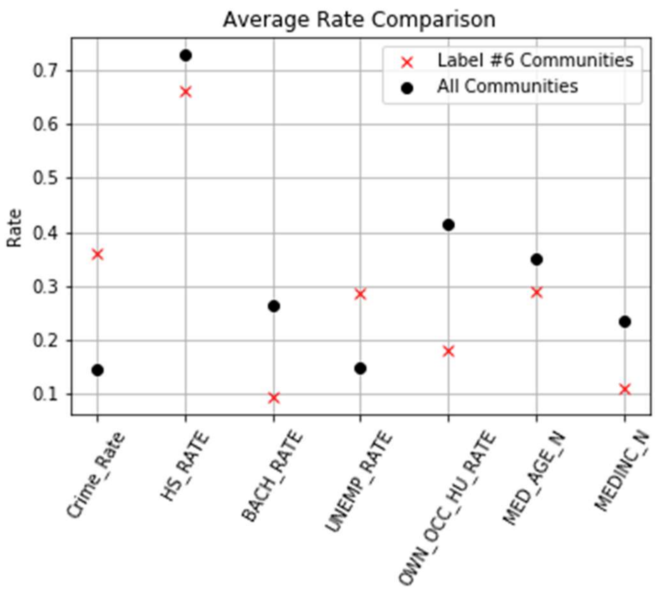
With the above, the cluster label #6 has the highest average Crime Rate. Table-6 lists the community areas with that label.

Table-6

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

Figure-11 is a chart comparing the average rates for crime and demographic features for the community areas labeled as #6 against the average rates for all of community areas.

Figure-11



1.3 Observations:

- Community areas labeled as #5 or #6 have highest average crime rate, and these as #3 have the lowest.
- If a community area falls in the cluster of #6 with its demographic information, its crime rate is very likely high. Suggest that the Chicago government should pay attention to the community areas in this cluster.
- By comparing the features between the community areas labeled as #6 and all of community areas, the following can be reached:
  - For the features of High School Diploma or Higher Rate, Bachelor Degree or Higher Rate, House/Apartment Owning Rate, Median Income, the community areas labeled as #6 have lower average values than the average community areas;
  - For the Unemployment Rate feature, the community areas labeled as #6 have much higher average values than the average community areas;
  - The community areas labeled as #6 have younger median age than the average community area.

VII. Conclusions and Future Considerations

1. Conclusions

- It is strongly recommended to the Chicago city government that the unemployment rate is the top feature impacting the crime rate for a community area. They should do every effort to lower this rate as top priority in order for them to reduce the Crime Rate for a community area in Chicago.

- The Chicago city government should also to improve other house/apartment owning rate, and encourage college eduction.

- For the high crime rate community areas (labeled as #6), they may want to deploy more policemen there as interim solutions for reducing their Crime Rate before they resolve the high unemployment rate for such community areas.

2. Future Considerations
   - If the Chicago city government can provide more demographic information for other features for a specific year, the data analysis would be more comprehensive and potentially identify more features 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 feature changes on the rise and decline of the Crime Rate.

VIII. Appendix A: Deliverables

1. Jupyter books for data analysis and machine learning
   - Data Wrangling: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20Data%20Wrangling.ipynb

   - Data Story Telling:

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

   - Inferential Statistics Techniques:

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

   - Supervised Machine Learning - Linear Regressions: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20MachineLearning.ipynb

   - Unsupervised Machine Learning – KMeans Clustering: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20MachineLearning.ipynb

2. A document describe the project (this document)

3. A slide deck on the whole project and conclusion for end users: https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeDataset_Project_UserPresentation.pdf

IX. Appendix B: Description of Chicago Crime Dataset

Column Name	Description	Type
ID	Unique identifier for the record.	Number
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	Plain Text
Date	Date when the incident occurred. This is sometimes a best estimate.	Date & Time
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.	Plain Text
IUCR	The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e.	Plain Text
Primary Type	The primary description of the IUCR code.	Plain Text
Description	The secondary description of the IUCR code, a subcategory of the primary description.	Plain Text
Location Description	Description of the location where the incident occurred.	Plain Text

Arrest	Indicates whether an arrest was made.	Checkbox
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	Checkbox
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74.	Plain Text
District	Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.	Plain Text
Ward	The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.	Number
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.	Plain Text
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.	Plain Text
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Number

Year	Year the incident occurred.	Number
Updated On	Date and time the record was last updated.	Date & Time
Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.	