Udacity Machine Learning Nanodegree

Metropolitan City Crime Analysis

Angela Baltes

May 7th, 2019

I. Definition

Project Overview

The state of New Mexico is located in the Southwestern United States, situated between Arizona, Colorado, Texas and sharing a small border with Utah. The area where the corners of the four states meet is aptly named "Four Corners". New Mexico , a state rich in culture and history, was colonized by Spain and became a United States territory as part of The Gadsden Purchase in 1853. Later, New Mexico became the 47th state in 1912. [1]During World War II, New Mexico became a hub of science and technology establishing Los Alamos National Laboratories -the site of the Manhattan Project. [2] The goal of The Manhattan Project was to design and build the first atomic bomb. Two years and 3 months later, The first atomic bomb was detonated near Alamogordo, New Mexico July 16th, 1945. Since the establishment of Los Alamos and then later Sandia National Laboratories, there has been a significant federal presence and investment in this state, and home of three air force bases: Kirtland, Holloman and Cannon. Given that a large majority of New Mexico's work is concentrated in federal government, with little growth in other industries, this leaves fewer options for many residents for viable employment . [3]

The state of New Mexico unfortunately suffers from issues that are complex and multiplicative in nature. Percentage of overall poverty in New Mexico is close to the highest in The United States at nearly 20% in year 2017, coupled with high child poverty rates. [4] School graduation rates are also the lowest in The United States. Considering these factors, the state of New Mexico is home to particularly high crime per

100,000 persons in the areas of property crime and burglary-close to the highest in The United States. Albuquerque is the largest city in the state of New Mexico, with a population of roughly 550,000 inhabitants. [5] Overall crime within the city is higher than other cities in the state due to its size, accounting for a large percentage of the overall crime rate. It appears that there is a direct relationship between crime and poverty in the state of New Mexico. [6]

According to a paper by von Lampe, Kurti and Johnson (2015), the link between poverty and crime has been a contested issue for many years, and there is not a clear consensus. [7] According to prior research, members from lower socioeconomic groups are over-represented among crime suspects and convicted offenders, and there may be a fair distribution of crime distributed evenly along the socioeconomic spectrum. However, some self-reported data have shown a positive relationship between socioeconomic status and crime for certain types of crime. Interestingly, in a study conducted by Foley (2009), there is a suggestion that crime is a way to account for a lack of income. It was found that toward the end of the month, there was an increase in financially motivated crime in cities among some welfare recipients. [7]

Although it is widely known that crime and poverty is elevated in New Mexico, there is a lack of literature that explicitly addresses the issue for this state. The cultural dynamics here in New Mexico are unique and can be difficult to understand in general terms. My interest in wanting to investigate Albuquerque's crime stems from my time as a resident. Although I can visibly see what is happening, my experience is purely anecdotal and others may have a very different experience as a resident. I am doing this research to give an objective viewpoint to what is occurring within the city of Albuquerque and confirm or disprove it with analysis.

Problem Statement

Crime data is regularly generated for the city of Albuquerque. The volume and variety of the data sources is at times difficult for a small team to analyze manually, and continuously provide actionable information. The purpose of this research is to develop a predictive crime analytics solution that will aid law enforcement in data-driven decision making. Predictive analytics would be a valuable tool and provide a greater

understanding of the dynamics of crime within Albuquerque. In designing a solution, first step is to construct and test a model based on various algorithms (decision tress, random forests, logistic or linear regression), to discover a relationship between factors that may affect crime in the area relating to the major types of crime: property, sexual, robbery, violent and others. With experimental results we will demonstrate the performance of the models based on particular metrics. The models that score highest will be used for predicting Albuquerque crime within a given census area.

Metrics

The evaluation metrics proposed are appropriate given the context of the data, the problem statement and the intended solution. There will be several different models that will be tested in this scenario.
The first model that will be tried and is ideal is a multiple linear regression model. This model in particular is ideal because we are looking to predict the output of incidents considering median age, income and number of crime incidents historically for the census tract in question. In a simple linear regression, a single predictor variable (X) is used to model the response variable (Y) . In many cases, there are more than one factor that influences the response. [10] Multiple regression models describe how a single response variable Y depends linearly on a number of predictor variables. Below is the equation for simple linear regression as well as a multiple linear regression model with k predictor variables X1, X2, ..., Xk and a response Y , can be written as

Simple linear regression : $Yi = \beta_0 + \beta_1 xi + \varepsilon i$

Multiple linear regression: $Y = \beta_0 + \beta_1 x_1 + B2\ x2 + ... B_k\ x_k + \varepsilon$

The ε are the residual terms of the model and the distribution assumption that is placed upon the residuals that will allow later to perform inference on the residual (remaining) model parameters. Interpret the meaning of the regression coefficients $\beta 0, \beta 1, \beta 2, ..., \beta k$ in the model. In regression, R-squared is the goodness of fit. The performance of the regression model within this project will be measured on R squared and coefficients. For a multiple linear regression model, ideally it would look like the example plot provided [10]

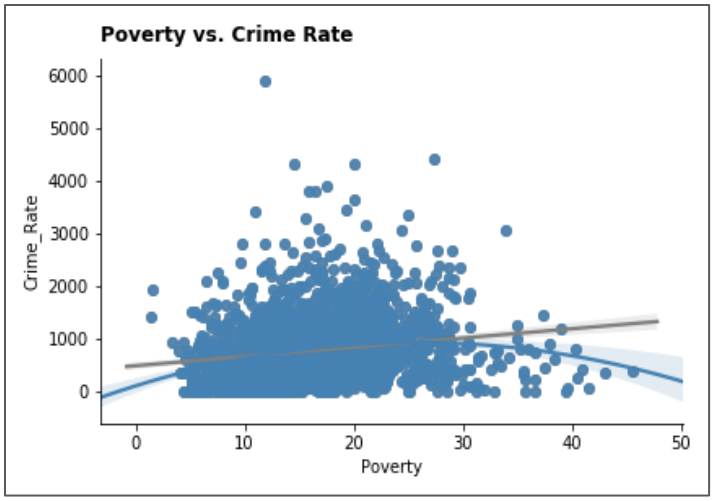

Fig 1: Example of Regression Output between Crime Incident and Income [11]

II. Analysis

Data Exploration

This dataset contains the block location, case number description and date of calls for service received by The Albuquerque Police Department. The incidents have been entered into the Computer Aided Dispatch system and closed. This dataset contains 180 rolling days of incidents. Accompanying the Incidents table is a codes table for describing each type of incident. The data is available in JSON and KML format which will allow for ease of generating a frequency location map. [8] Census tracts will be assigned to each address by proximity using census data from 2006-2010 (TIGER/Line®) and in addition a median income and age for the area (census tract) will be utilized as predictive measures, in particular for a linear regression model. The various types of crime (NIBRS Group)will be tested individually against median income. [9]

The census information was extracted from the "2013-2017 American Community Survey 5-Year Estimates" dataset contains median age and median income by each census tract in Bernalillo County. Two tables were used to obtain age and income values: Table S0101 - AGE AND SEX and S1903 and Table S1903 - MEDIAN INCOME IN THE PAST 12 MONTHS (IN 2017 INFLATION-ADJUSTED DOLLARS) The median income and median age was then binned to a census tract, which resulted in a one-to-one match.

Incidents			
Field Name	**Field Alias**	**Format**	**Description**
OBJECTID	Object_ID	esriFieldTypeOID	ESRI generated OBJECTID – This is not the case id. There will be one unique ID per record, although the ID itself is not guaranteed across different requests for the dataset (i.e. a separate request for the same dataset *could*, potentially, have different OBJECTID's).
BlockAddress	Location	esriFieldTypeString	Neighborhood block level location of incident.
IncidentType	Description	esriFieldTypeString	Description/type of incident. Based on NIBRS incident types used by Albuquerque Police Department. It will be one of the values found in nibrAbqCodes.csv. More information on NIBRS can be found at http://www.fbi.gov/about-us/cjis/ucr/nibrs/nibrs-user-manual
Date	Date	esriFieldTypeDate	Date that the incident was reported as milliseconds since Jan 1st 1970 UTC. Note that this might be different to the date on which the incident took place.

Table 1: Albuquerque Crime Incidents

nibrAbqCodes.csv (National Incident-Based Reporting System used by Albuquerque)		
Field Name	**Format**	**Description**
NIBRS Group	Text	Reported Offense Categories
NIBRS Heading	Text	Reported Offense Category Headings
CABQ Offense	Text	Offense in the CABQ field CVINC_TYPE found in the JSON file
NIBRS Code	Text	Code associated with the Reported Offense

Table 2: Albuquerque Incidents Codes

Field Name	**Field Alias**	**Format**	**Description**
HC01_EST_VC37	Total; Estimate;- Median age (years)	Integer	Calculated median income per census tract.
HC03_EST_VC02	Median income (dollars); Estimate; Households	Integer	Calculated median age per census tract.

Table 3: 2013-2017 American Community Survey 5-Year Estimates

Incidents			
Field Name	**Field Alias**	**Format**	**Description**
OBJECTID	Object_ID	esriFieldTypeOID	ESRI generated OBJECTID – This is not the case id. There will be one unique ID per record, although the ID itself is not guaranteed across different requests for the dataset (i.e. a separate request for the same dataset *could*, potentially, have different OBJECTID's).
BlockAddress	Location	esriFieldTypeString	Neighborhood block level location of incident.
Longitude	Long(x)	Location String	X Coordinate for location
Latitude	Lat(y)	Location String	Y Coordinate for location
IncidentType	Description	esriFieldTypeString	Description/type of incident. Based on NIBRS incident types used by Albuquerque Police Department. It will be one of the values found in nibrAbqCodes.csv. More information on NIBRS can be found at http://www.fbi.gov/about-us/cjis/ucr/nibrs/nibrs-user-manual
Date	Date	esriFieldTypeDate	Date that the incident was **reported** as milliseconds since Jan 1st 1970 UTC. Note that this might be different to the date on which the incident took place.
Census Area	Census Tract	String	Assigning to each address the appropriate census tract it belongs in.
Counts	Counts	Integer	This provides an overall count of incidents per census tract
Median Income	Median income (dollars); Estimate; Households	Integer	The median income of the particular census tract.
Median Age	Total; Estimate;- Median age (years)	Integer	The median age of the particular census tract.

Table 4: Ideal dataset for use after combining Census dataset to Incidents

The ideal dataset that will be used for this project will use nine fields. Only two fields from the before mentioned census datasets will be analyzed. Each census tract will receive one value of median income and median age. The total incidents will be tabulated into one value per census tract for ease of processing. The result will be more than 150 census tracts. The Incident dataset has 28,886 records from the 180 rolling days in Albuquerque. There are 49 incident types associated within, with the highest being: SUSP PERS/VEHS at 6239, DISTURBANCE 5016 and TRAFFIC STOP at 4601. Median age and median income distributions were

tabulated to understand the range in both fields. The classification goal is to predict the number of incidents per each census area.

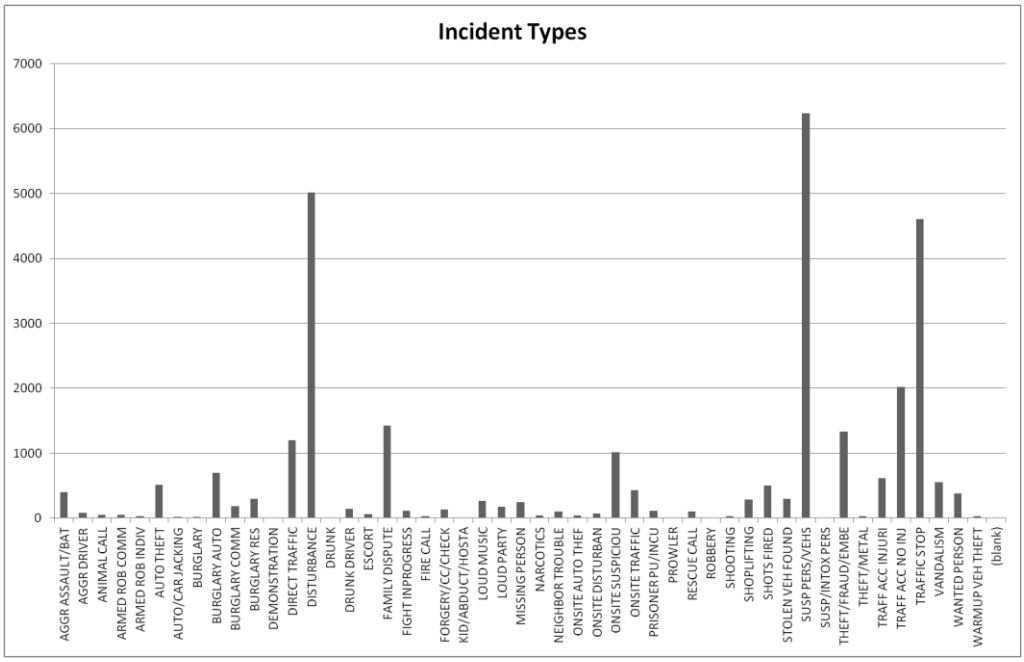

Fig 2: Albuquerque Police Incidents

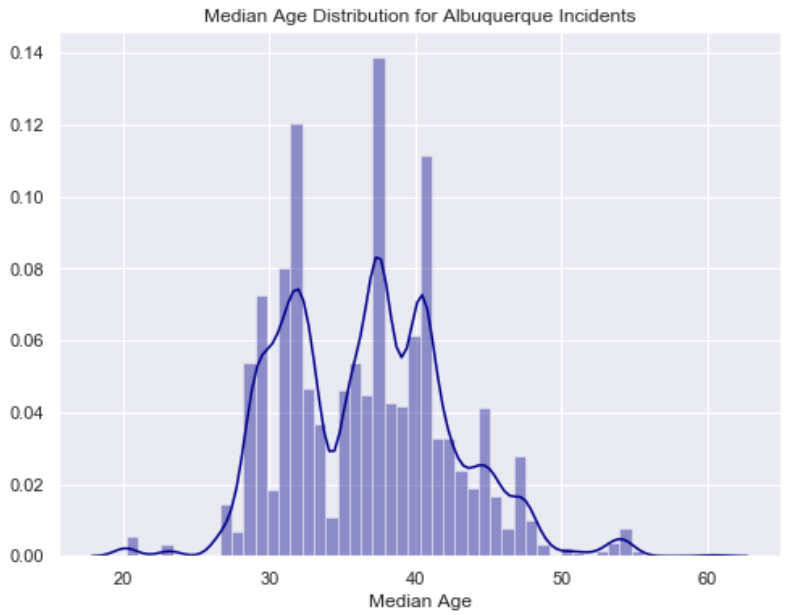

Figure 3: Median Age Distribution

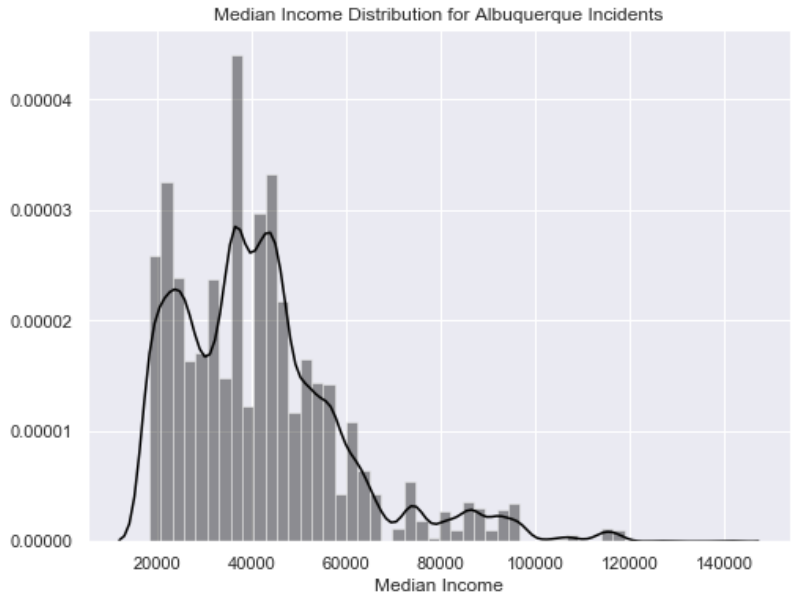

Figure 4: Median Income Distribution

Once the dataset was joined for analysis, it was clear that the date column was presented in milliseconds, which then needed to be transformed into a year, month , day and time . The location coordinates present in the file, were a projected web coordinate, that was unrecognized as a coordinate by common location tools. Using pyproj, a python library that performs cartographic transformations between geographic (lat/lon) and map projection (x/y) coordinates, we were able to transform the coordinates from a Pseudo-Mercator[12] to WGS84 [13]. Each census tract, was given another column, that was its extended version of the integer. For instance Census tract 1.07 was given a long integer of 000107. This was done in order to associate this dataset to a public geocoded dataset to visualize as a map. Median income is right skewed, but that indeed represents median income in the state of New Mexico, especially Albuquerque. It has been accepted that incomes are lower than other cities, and this data proves that assumption. There is a fairly balanced distribution of ages being that many median ages fall between 20 and 60 years of age. In reviewing initial statistics, only columns 'MED_AGE', 'MED-INCOME', and 'COUNTS' provide information that is useful for statistical analysis. There is a wide range in income, as well as incident counts, which vary greatly upon the census tract.

	CENSUS_TRACT	MED_AGE	MED_INCOME	COUNTS	LONGITUDE	LATITUDE	LONG_CENSUS	CENSUS	STATE_ID
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000
mean	213.637933	38.962667	54470.280000	192.566667	-103.748722	35.128191	21363.793333	213.637933	75.500000
std	1317.718534	7.246740	23290.457153	198.275533	24.530972	0.055839	131771.853399	1317.718534	43.445368
min	1.070000	20.200000	18306.000000	1.000000	-106.753181	34.939182	107.000000	1.070000	1.000000
25%	7.047500	33.500000	37710.500000	54.250000	-106.665293	35.091717	704.750000	7.047500	38.250000
50%	34.505000	38.650000	48765.500000	124.500000	-106.582616	35.121983	3450.500000	34.505000	75.500000
75%	46.850000	42.975000	66326.500000	257.250000	-106.516961	35.169194	4685.000000	46.850000	112.750000
max	9407.000000	60.400000	140833.000000	941.000000	106.568750	35.248188	940700.000000	9407.000000	150.000000

Table 5: Descriptive Statistics for Combined Dataset

Exploratory Visualization

In reviewing incidents, it was clear that there were more reported incidents in October 2018, than any other month. When looking more closely at the month of October, the 2nd and 8th were the highest days of reported incidents. October 8th, 2018 was a holiday and at the start of The International Balloon Fiesta. It is likely there were many visitors and many had an extended weekend due to the holiday. It is unknown why October 2nd, 2018 had an elevated number of incidents, as it was not a holiday. It may be possible that visitors were entering Albuquerque at that time in preparation for Balloon Fiesta.

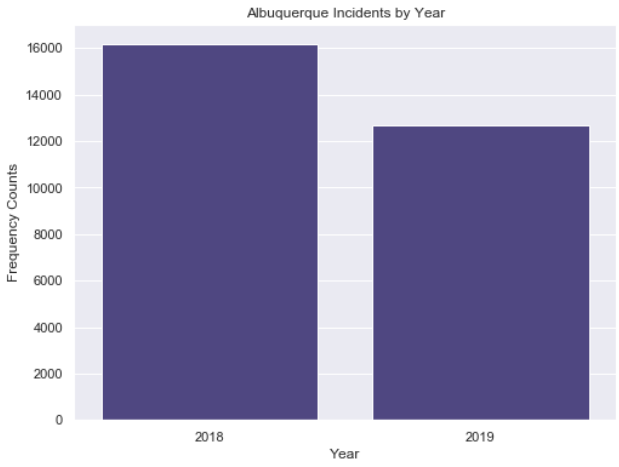

Figure 5: Overall Incidents by Year

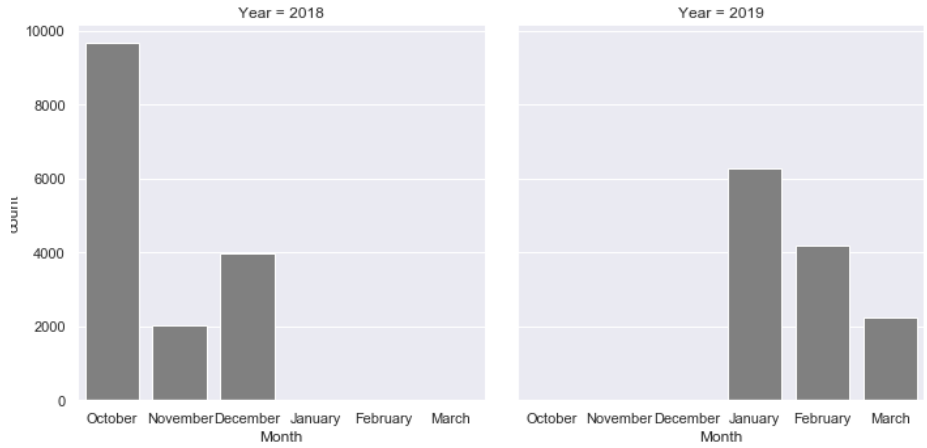

Figure 6: Overall Incidents by Month and Year

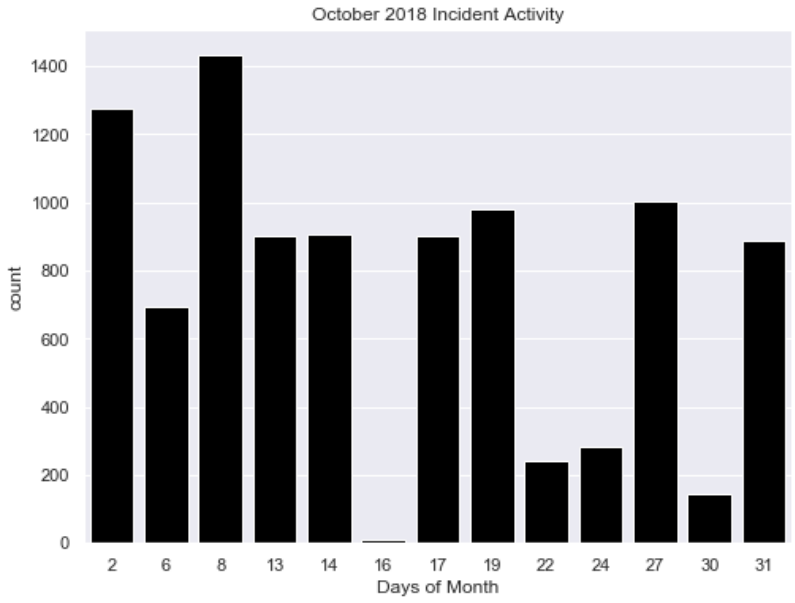

Figure 7: Overall Incidents by October

Algorithms and Techniques

Based on this data, a supervised learning algorithm would be the most appropriate solution for prediction. The goal is to use two input variables to predict crime per census tract: median age, and median income . The ideal output would be the number of expected incidents for that census tract. Census tract information may prove to be more valuable rather than analyzing crime in Albuquerque as a whole. As explained earlier, linear regression would be a likely solution for this type of problem. However, a decision tree

model may also be appropriate. There are concerns that a linear regression model will perform poorly sue to lack of inputs and a potential for not having enough data. .

Extreme Gradient Boosting (XG Boost): XGBoost, belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core. It is an optimized distributed gradient boosting library.[12]

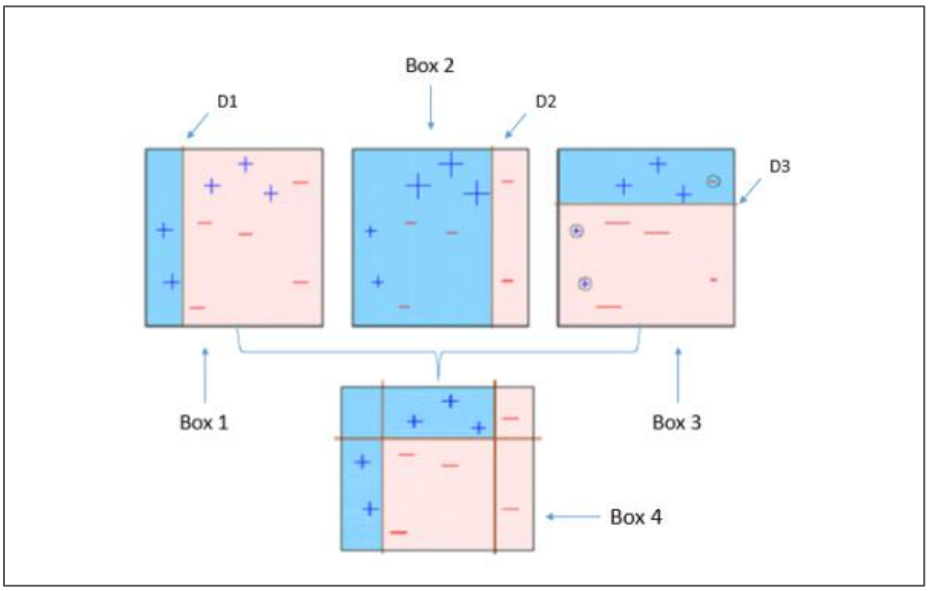

Figure 8: XGBoost

Decision Trees: This algorithm is compared to as a map of possible outcomes. [13] It is able to handle categorical and numerical data. This algorithm doesn't require extensive data pre-processing, and can handle data which has not been normalized, or encoded for Machine Learning Suitability. Complex Decision Trees do not generalize well to the data and can result in over fitting. This algorithm can be unstable, as small variations in the data can result in a different decision tree. This is why they are usually used in an ensemble (Random Forests) to build robustness. This algorithm can create biased trees if some classes dominate others.[13]

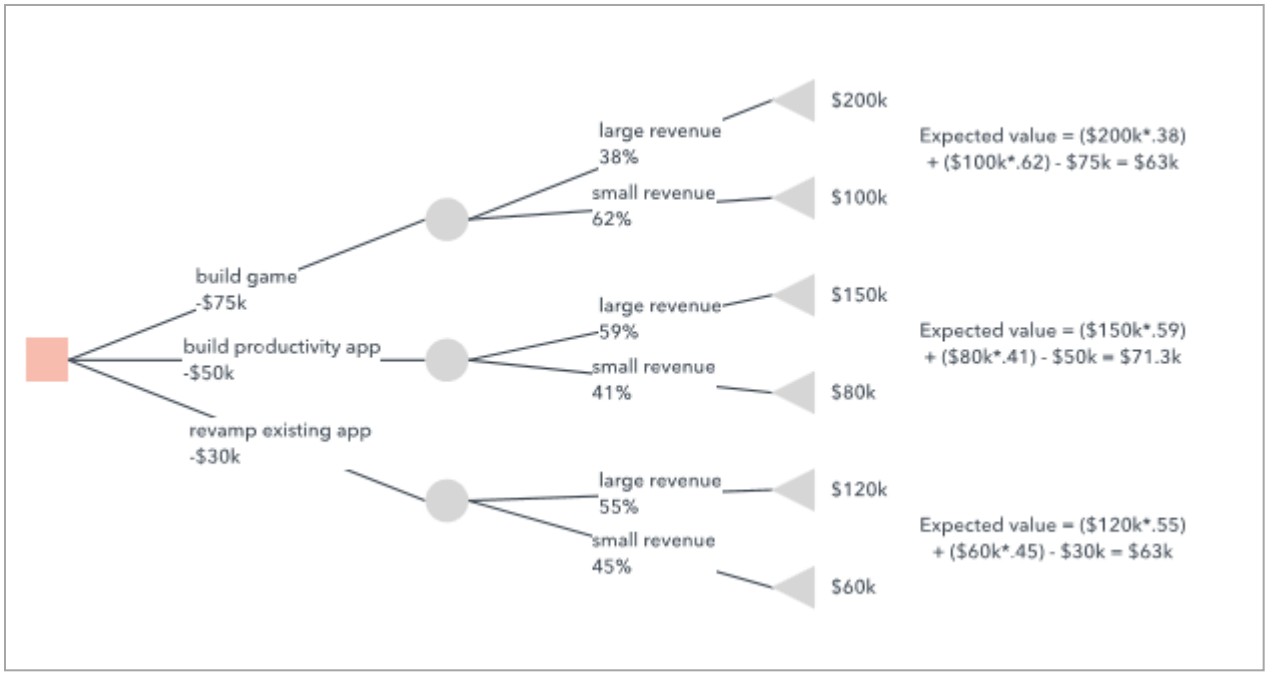

Figure 9: Decision Trees

Random Forests: Is a collection of decision trees to produce a more generalized model by reducing the notorious over-fitting tendency of decision trees. [14]

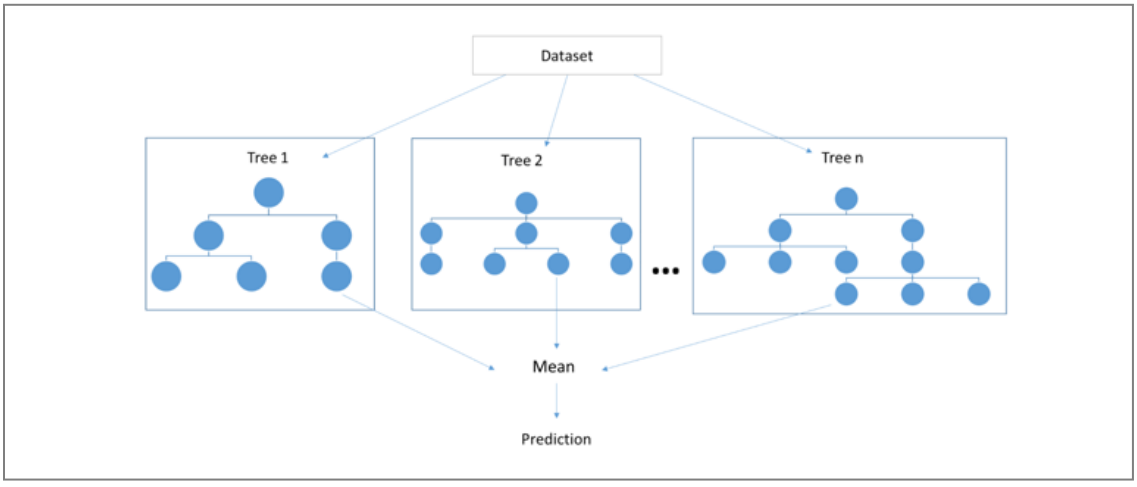

Figure 10: Random Forests

Adaptive Boosting(Adaboost):This is a type of an ensemble method. Ensemble methods, including Adaboost are more robust than single estimators, and can boost the performance of a machine learning model.

But however, best used with binary classification models. Simple models can be combined to build a complex model, which is computationally fast. If we have a biased underlying classifier, it will lead to a biased boosted model. If using complicated "weak learning" models, there is a higher chance of over fitting. [15.]

Support Vector Machines (SVM): Effective in high dimensional spaces, (computationally efficient) and can handle multiple features. Kernel functions can be used to adapt to different cases, and can be completely customized if needed. SVMs are fairly versatile. The improper choice of the kernel can negatively affect results. Doesn't directly provide probability estimates.

Benchmark

In the benchmark model, Decision trees performed fairly well. In this case, it will be considered as one of the model to attempt the outperform the benchmark in the hyperparameter tuning phase of the project. This is not an accurate representation of the performance of the final algorithm , and is only to display potentially that may be encountered. The below table highlights the performances of the various types of models attempted with their accuracies and standard errors.

Algorithms	Accuracy	Std. Error
Linear Regression	0.90587	.006182
Support Vector Machines	0.91168	.006821
Decision Trees	0.91425	.007850

III. Methodology

Data Preprocessing

The data was first prepared by transforming the longitude and latitude that was present in the incidents dataset from a WGS 84 / Pseudo-Mercator format to WGS 84. This format is essential in order to associate incidents to a location in Albuquerque by common geolocation tools. Using pyproj, a python library, this was

easily accomplished. The variable date was originally available in milliseconds and was transformed to a year, month, day and timestamp. Each block location address was associated to a census tract location using Census 2010 information. Overall counts for each census tract was calculated. Median age and Median income by census tract for Bernalillo county from The United State's census site, was downloaded and joined to incident data by its respective census tract. This resulted in a one-to-one match to the 152 census tracts associated with Albuquerque, New Mexico. One census tract did not associate to any other tracts in the census data and was subsequently dropped for being a nan value. Lastly, each census tract had a "long census" associated with it in order to join to a public geolocation dataset to visualize incidents per census tract in Albuquerque. Which was performed by dividing the census tract by 100 to generate a value that included leading zeros. This preprocessing resulted in a dataset where each census tract had a longitude/latitude, median age, median income, overall incident count and long census associated with it. There were unfortunately three census tracts in Albuquerque without data points. This will unfortunately result in unplotted census tracts to the map.

Implementation

Following the format of the predicative analytics workflow depiction as shown below, we first performed exploratory analysis by viewing the distributions of the features and target variables. Analyzing other variables such as incident frequencies during year, month, and incident types were reviewed as well. Using a plot, we also evaluated the relationship between the feature and target variables. We then performed data cleaning steps to identify nan values and drop them if necessary. Once the data was properly cleaned, it was then prepared for model implementation.. The data was split into 80/20 ratio. At that time two models were prepared for analysis based upon benchmark model results: linear regression and decision trees. Given the nature of the data it is assumed that these models will likely be appropriate for analysis.

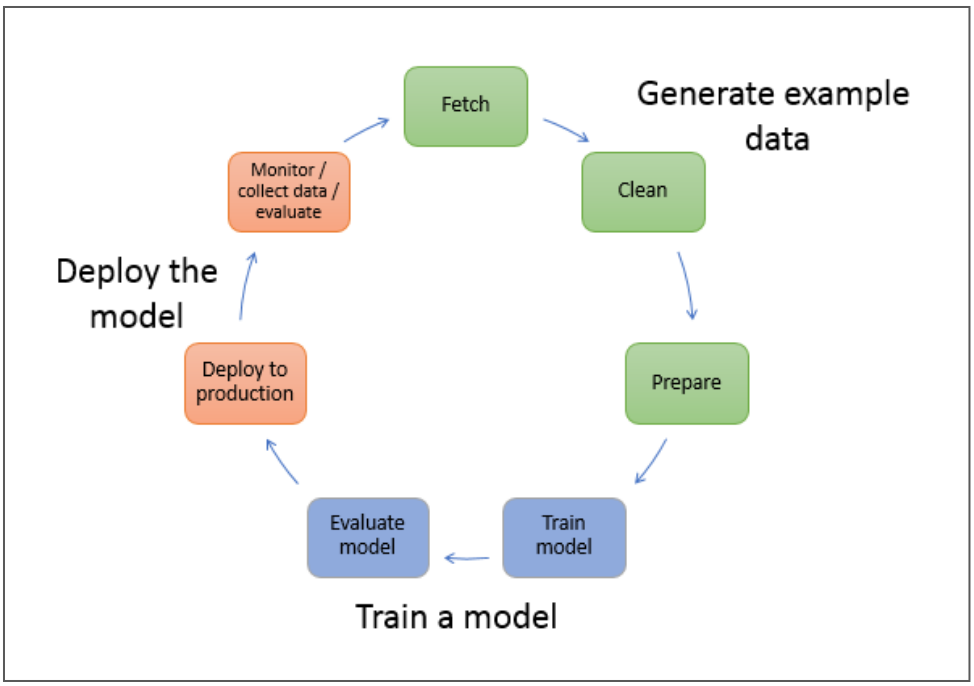

Figure 11: Machine Learning Workflow [16]

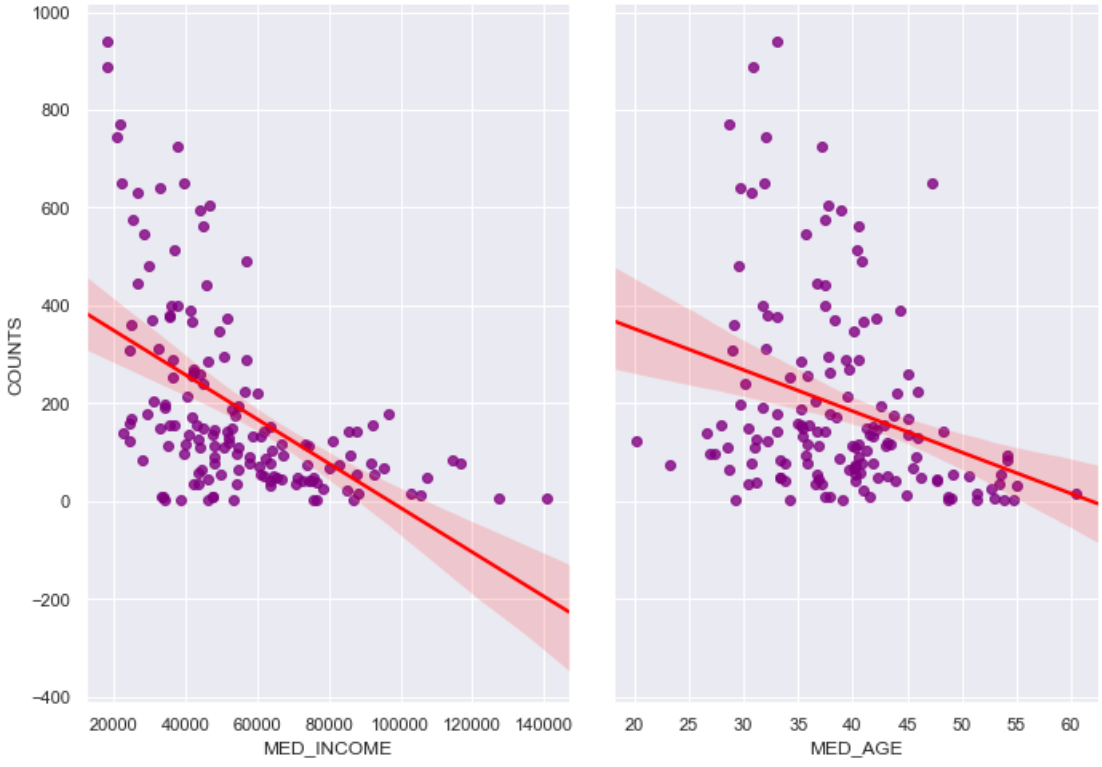

Figure 12: Relationship between features (Median Age and Income) to Target Variable (Counts)

Refinement

The data was split by features and target variables and checked for quality of data. To evaluate the models, the data was split in to 80% training and 30% testing to ensure accuracy. The max_depth was 2 based upon the parameters set. The initial results of a linear regression model proven to be less than favorable. To note, the features and target variables needed to be transformed in order to fit into a model. Originally, the array for X-train was: 120,2 and y_train: 120,1. The algorithm was unable to process it. By passing y_train=y_train.transpose() for the training variable, it resulted in a format that could fit into the model. The reason for this very poor result is due to binning each age and income to one census tract. It resulted in the linear regression model performing poorly.

	Actual	Predicted
0	35	219.492158
1	137	210.866400
2	221	162.872019
3	36	219.256395
4	640	253.077457

Figure 13: Linear Regression Actual vs. Predicted

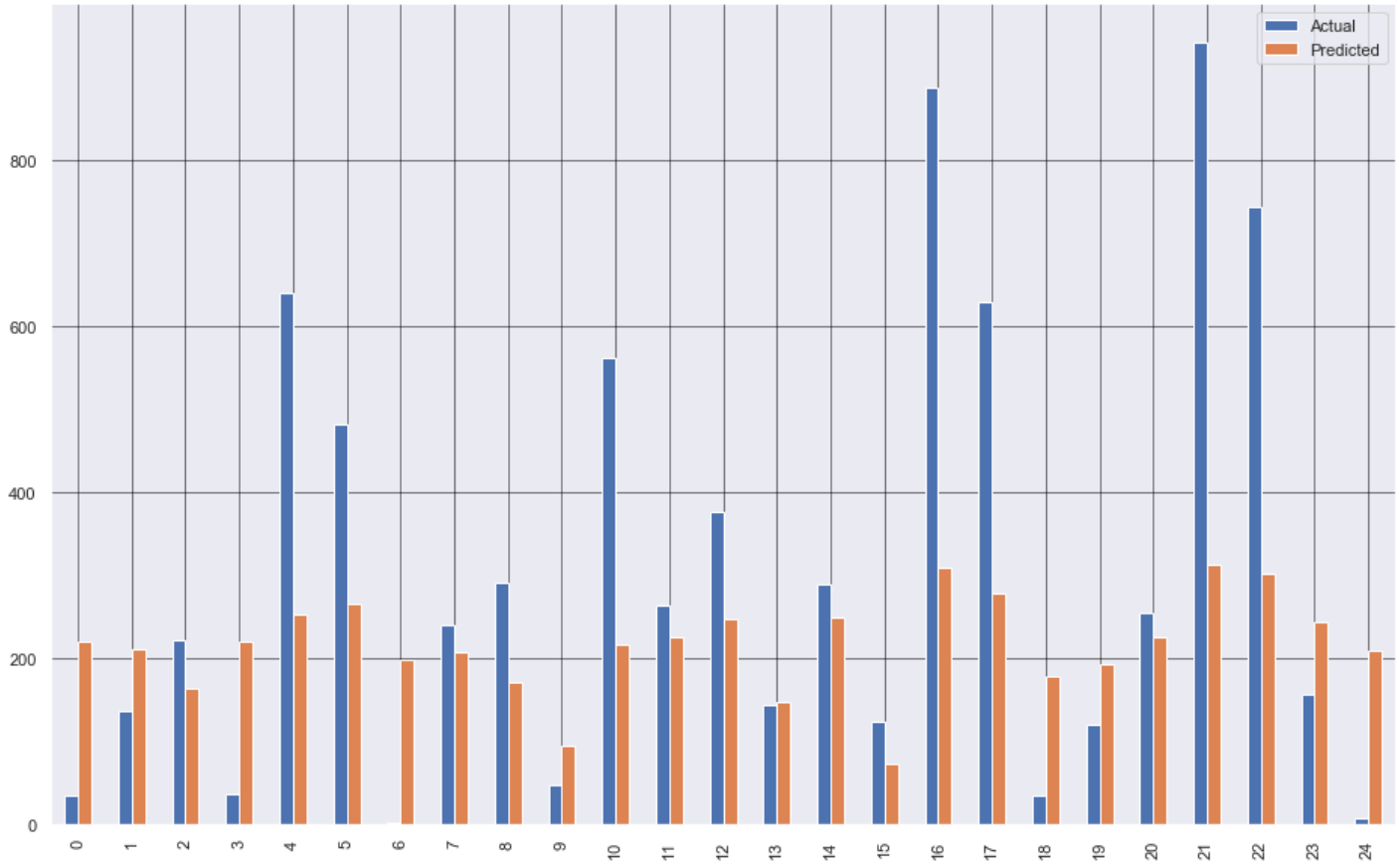

Figure 14 Linear Regression Visualization of Actual vs. Predicted

IV. Results

Model Evaluation and Validation

After testing linear regression and decision trees, the latter proven to be the best method for the data. Given the data, and the few inputs the model had to learn from, any model would have likely had poor predictive results. We applied out of the box decision tree to the data. Our end goal was to have a tuned model that would outperform the initial benchmark and unfortunately that did not occur. The results of the prediction resulted in very rigid output. For instance to test the power, we put into the model various ages and a range of income. It was clear that after an income of $65,000 annually we would receive the same result in number of incidents. The model's predictive power is not accurate and therefore is not a good representation of predicting crime incidents per census tracts in Albuquerque, New Mexico.

Justification

There is absolutely room for improvement on the final results, and that is largely due to the few data points that was available for the dataset. As mentioned earlier, each census tract had a median age and income associated with it, which resulted in 152 records. We do not feel that is enough data points for a model to truly perform well. In order for a model to perform more favorably, there needs to be a unique age and income associated with each incident. From there, a census tract can be associated with each incident. Using a binned age and income resulted in a rigid model that only gave a set number of responses.

V. Conclusion

Free-Form Visualization

The map below is a visualization of incidents per census tract in Albuquerque, New Mexico. The darker areas indicate census tracts with higher frequency of incidents. This map is the best representation of this data and can be used as a tool to gain understanding of the level of crime in Albuquerque. Contrary to the belief that Albuquerque suffers from high crime, it is actually that the frequency of crime in concentrated in only certain areas of the city. The interactive map is located at: https://arcg.is/1849zD

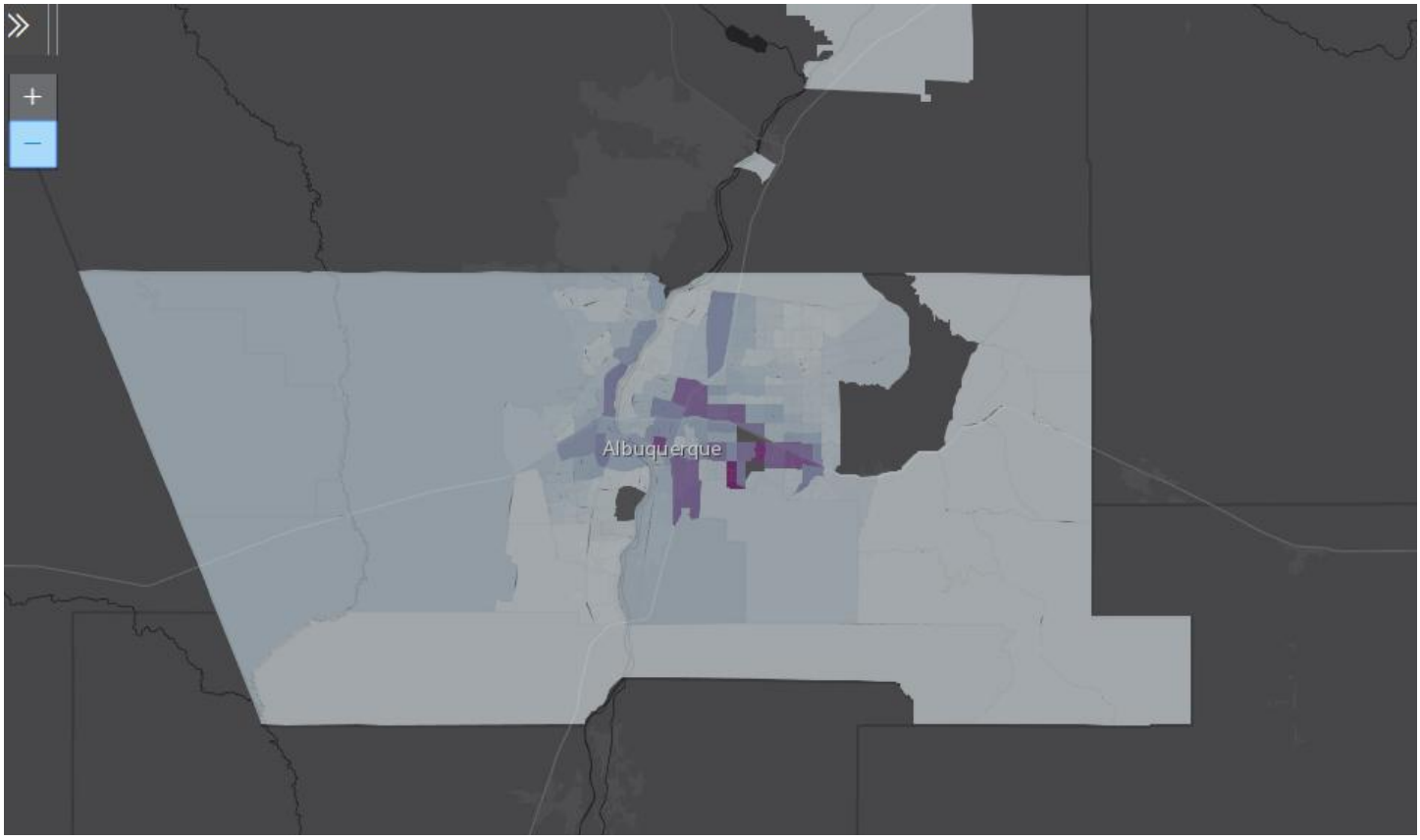

Figure 13: Map of Incidents by Census Tract

Reflection

The most important and time -intensive part of this project was data preprocessing. Transforming the coordinates, associating each location to a census tract and then joining the data to census median age and income tool a great deal of time and effort. The next most intensive process was ensuring the data was in a format that could be ingested by a cartographic program, such as ESRI. Once the data was prepared, it was clear given that each value was binned to a specific median age and median income, that any model that was chosen would not perform as well as anticipated. In the end, there was only 152 records for the algorithm to learn. Median income does have an association between incidents in Albuquerque, but age did not have a strong relationship with incidents. In hindsight, I see that in addition to have too few data points, I also used a weak feature in addition to a strong feature.

Improvement

A major improvement to the prediction result is to use a dataset that has a variety of incomes and ages associated with each incident. Using a binned median age and income resulted in rigid predicted results. Once that is resolved then it would not be necessary to bin records therefore resulting in a more accurate view of incidents per census tract. It would be ideal if Albuquerque Police Departments began incorporating census information into each incident for greater insight and produce actionable information that may be used for collaboration.

VI. References

1. Unknown. Author (2009, November 09). New Mexico. Retrieved March, 2019, from https://www.history.com/topics/us-states/new-mexico
2. Los Alamos National Laboratory, Los Alamos National Security, Llc, & U.S. Department of Energy. (n.d.). Our History. Retrieved March, 2019, from https://www.lanl.gov/about/history-innovation/index.php?row_num=0
3. New Mexico Economic Development Department. (n.d.). Retrieved March, 2019, from https://gonm.biz/site-selection/economic-statistics
4. Documentation. (n.d.). Retrieved March, 2019, from https://www.ers.usda.gov/data-products/county-level-data-sets/documentation/
5. FBI Table 6. (2017, September 18). Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-6/
6. Rio Grande Agenda (n.d.). Retrieved March, 2019, from http://riograndefoundation.org/downloads/
7. Duyne, P. C. (2015). *The relativity of wrongdoing. Corruption, organised crime, fraud and money laundering in perspective*. Nijmegen: W.L.P. (Wolf Legal).
8. ABQ Data. (n.d.). Retrieved March, 2019, from https://www.cabq.gov/abq-data
9. Branch, G. P. (2012, September 01). TIGER/Line® with Data. Retrieved March, 2019, from https://www.census.gov/geo/maps-data/data/tiger-data.html
10. *Math 261A - Spring 2012*. (n.d.). Retrieved March, 2019, from Cornell University website: http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement 5 - multiple regression.pdf
11. Data Analysis & Interpretation 3.3: Testing a Multiple Regression Model. (2018, October 24). Retrieved March, 2019, from https://datapyguy.com/2018/09/30/data-analysis-interpretation-3-3/
12. WGS 84 / Pseudo-Mercator, 2018. https://epsg.io/3857
13. WGS 84, 2018. Retrieved from https://epsg.io/4326
14. Using XGBoost in Python. (n.d.). Retrieved March, 2019, from https://www.datacamp.com/community/tutorials/xgboost-in-python
15. What is a Decision Tree Diagram. (n.d.). Retrieved March, 2019, from https://www.lucidchart.com/pages/decision-tree

16. Kumar, V., & Kumar, V. (2018, October 23). Random forests and decision trees from scratch in python. Retrieved March, 2019, from https://towardsdatascience.com/random-forests-and-decision-trees-from-scratch-in-python-3e4fa5ae4249
17. Schapire, R. (n.d.). *Explaining AdaBoost* (pp. 1-16, Tech.).http://rob.schapire.net/papers/explaining-adaboost.pdf
18. Machine Learning with Amazon SageMaker. (n.d.). Retrieved March, 2019, from https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-mlconcepts.html
19. Emblem, H., & Emblem, H. (2018, July 07). Training and test dataset creation with dplyr. Retrieved March, 2019, from https://medium.com/@HollyEmblem/training-and-test-dataset-creation-with-dplyr-41d9aa7eab31