Metropolitan City Crime Analysis

Angela Baltes

March 28th, 2019

Background

New Mexico History

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]

New Mexico Poverty and Crime

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 100,00 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]

Purpose of Research

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 it with analysis.

Problem Statement

Crime data is regularly generated for the city of Albuquerque. The size of the sources can be difficult for a small team to analyze manually. Predictive analytics would greatly aid in the decision making process and have provide 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 area.

Datasets and Inputs

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]

Table 1: Albuquerque Crime Incidents

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 2: Albuquerque Incidents Codes

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

Ideal Dataset for Analysis

Table 3: Ideal dataset for use after combining Census dataset to Incidents

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
NIBRS Group			Grouping the offenses into categories.
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			Assigning to each address the appropriate census tract it belongs in.
Median Income			The median Income of the particular census tract.
Median Age			

Evaluation 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 is a multiple linear regression model. 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

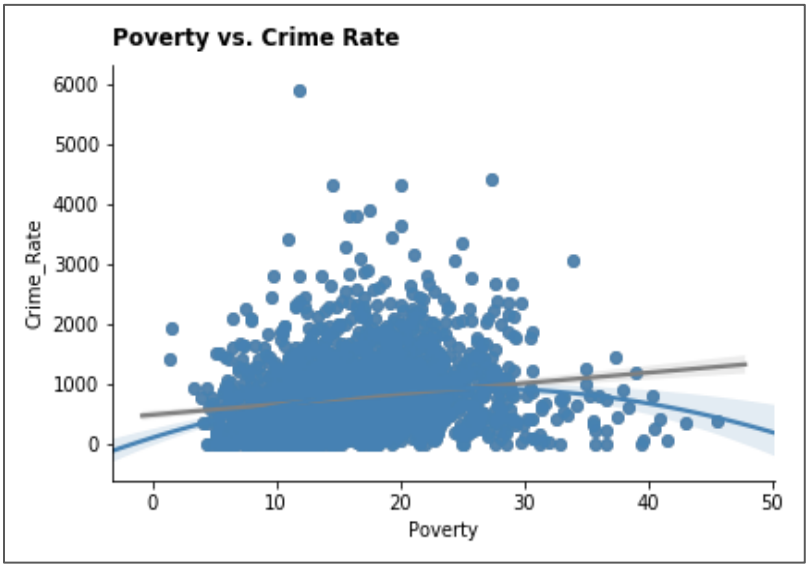
Simple linear regression: $Yi = \beta_0 + \beta_1 xi + \varepsilon i$

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 β0, β1, β2, ..., βk in the model. In regression, R-squared is the goodness of fit. The performance of the

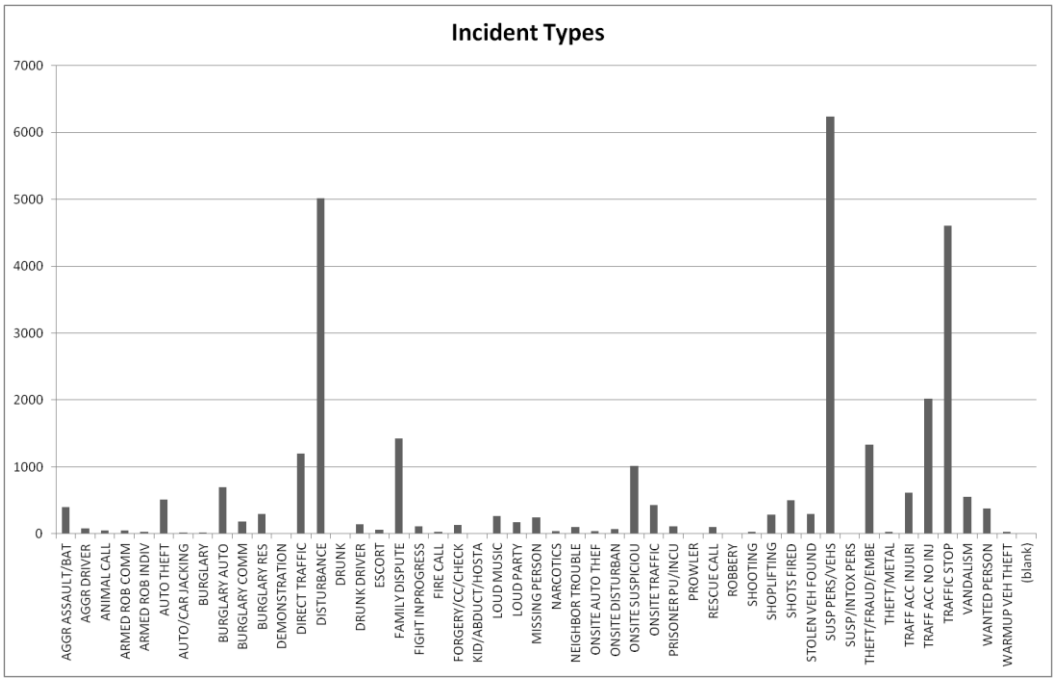
regression model within this project will be measured on R squared and coefficients. For a multiple linear regression model, ideally it would look like the example plot I have provided. [10]

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



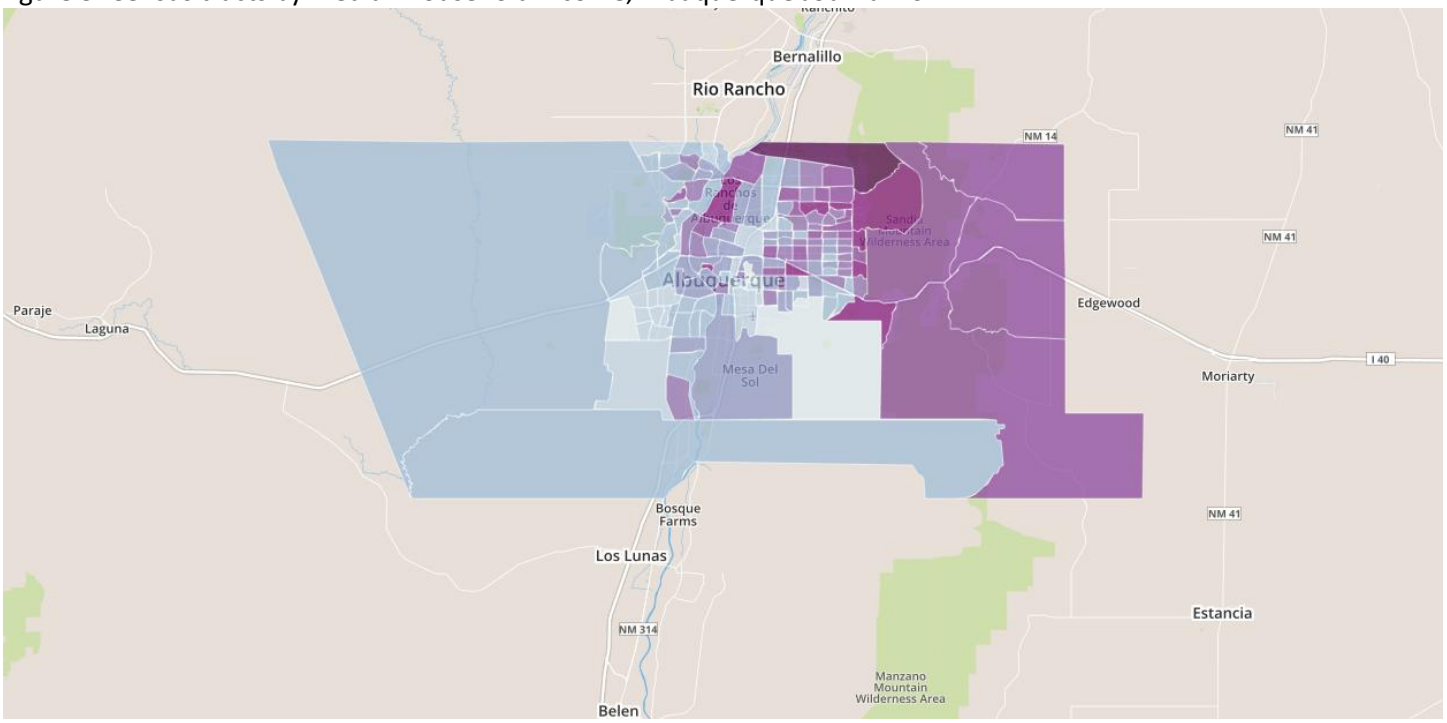
The Incident dataset has 30,000 records from the 180 rolling days in Albuquerque. There are 49 incident types associated within, with the highest being: SUSP PERS/VEHS at 6239, DISTURBANCE 5016 and TRAFFIC STOP at 4601. Below is a chart showing the distribution.

Fig 2: Albuquerque Police Incidents



Below is an example of a map that will likely be incorporated that has each census tract by median income and age. The original map will be interactive, but a screenshot was provided in this case.

Figure 3: Census tracts by median household income, Albuquerque Journal 2014

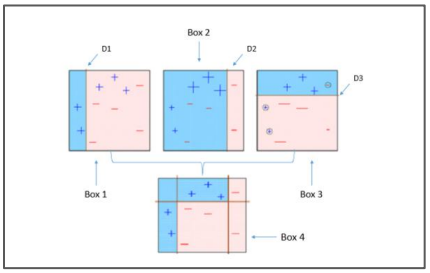


Solution Statement

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 incidence calls by crime type for that census tract. Each crime group will be analyzed separately rather than crime as a whole. Since crime and crime type is not distributed equally across the city of Albuquerque, that is why this approach is most necessary. As explained earlier, linear regression would be an appropriate solution for this type of problem. However, there are other options that may also be appropriate as well.

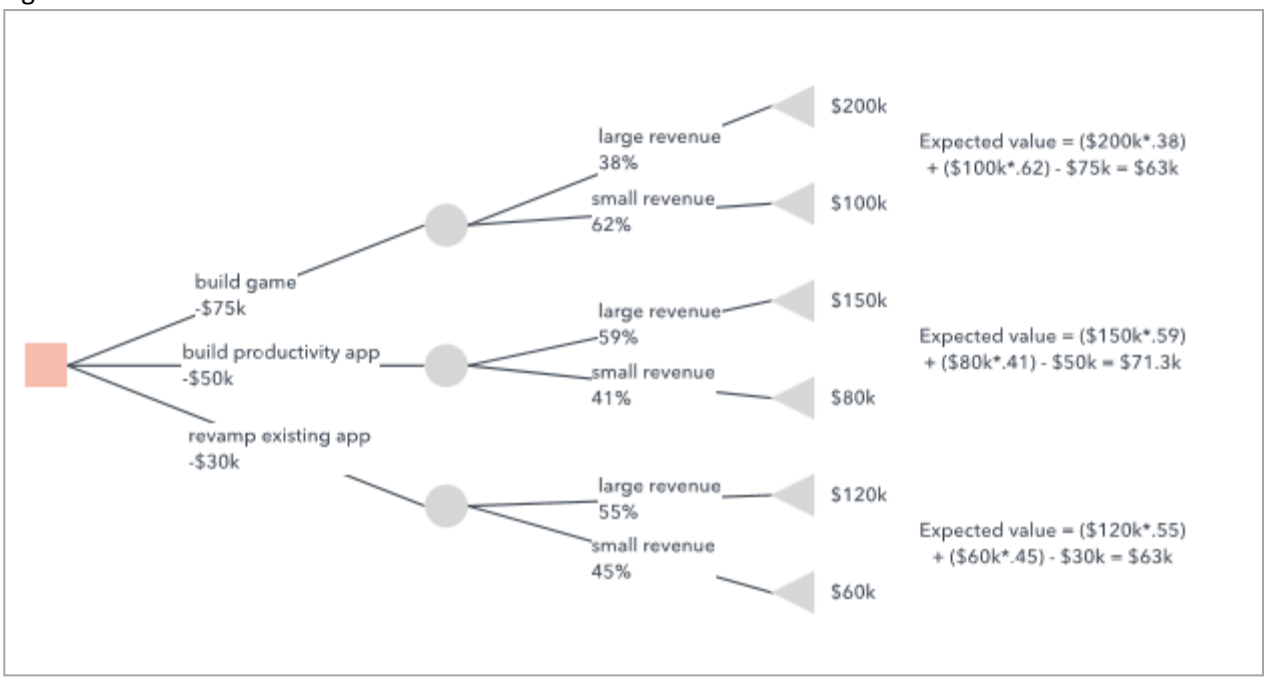
Extreme Gradient Boosting (XG Boost): XGBoost belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core. It is an optimized distributed gradient boosting library.[12]

Figure 4: XGBoost



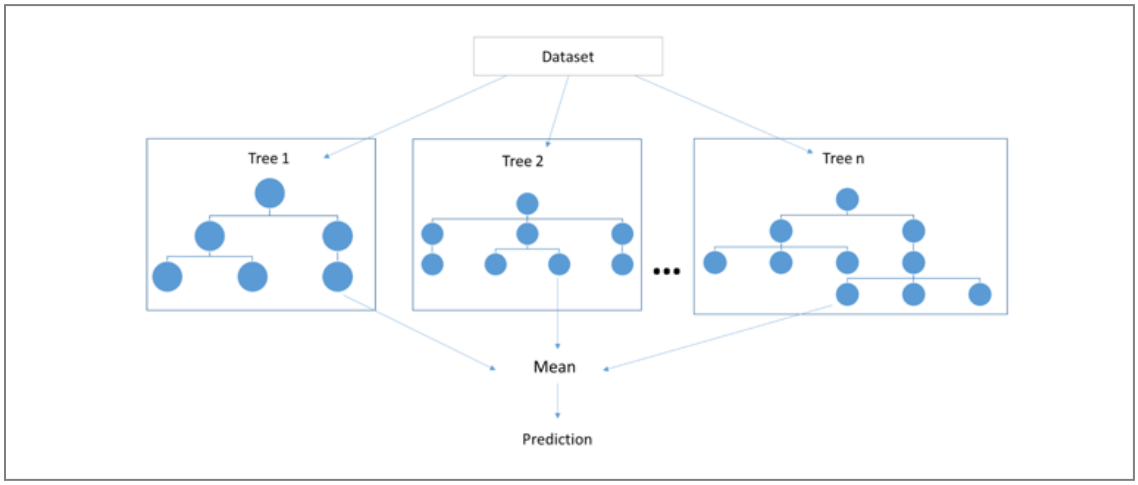
Decision Trees: This algorithm is compared to as a map of possible outcomes. [13] It is able to handle categorical and numerical data. Doesn't require extensive data pre-processing, and can handle data which has not been normalized, or encoded for Machine Learning Suitability. Complex Decision Trees do not generalize well to the data and can result in over fitting. 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. Can create biased trees if some classes dominate others.[13]

Figure 5: 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 6: 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.

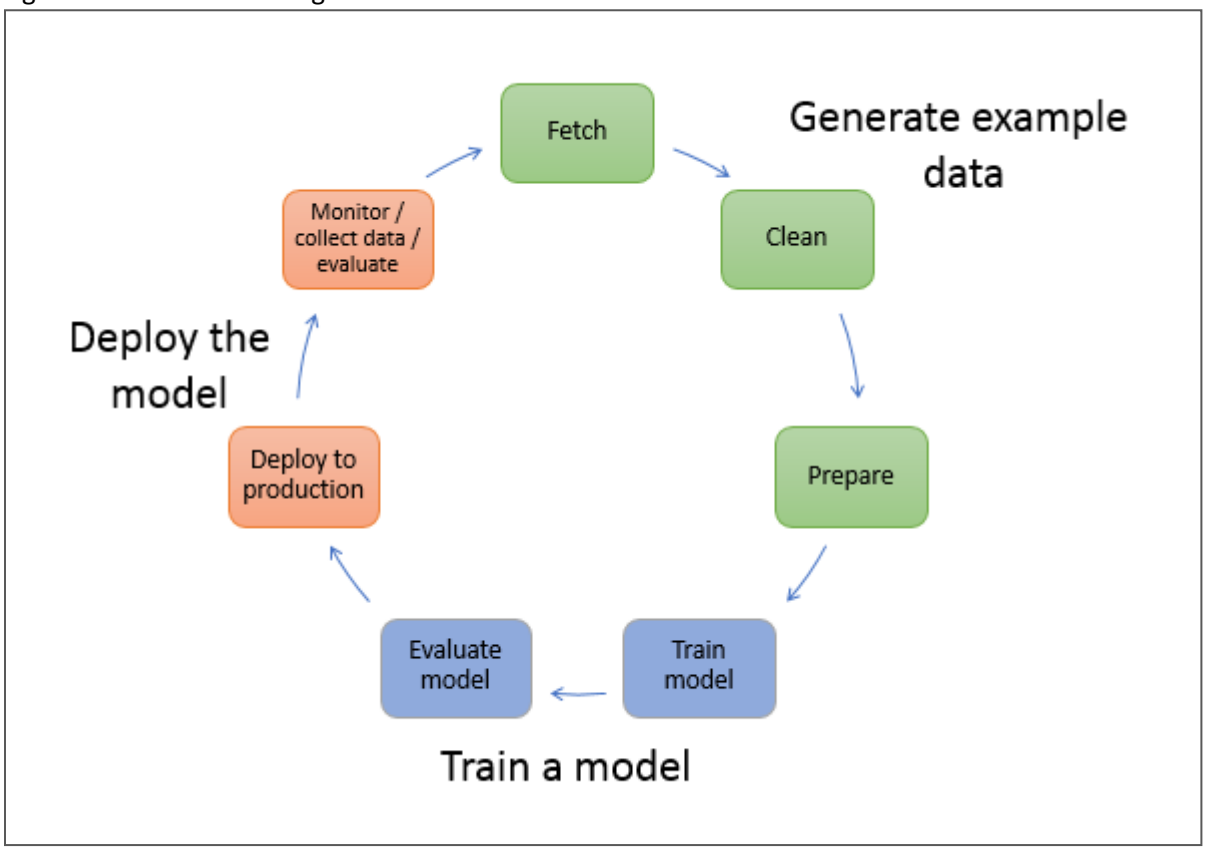
Benchmark Model

In the benchmark model, the standard Extreme Gradient Boosting (XGBoost or XGB) performed 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
Random Forest	0.90698	.007137
AdaBoost	0.91399	.007840
XGBoost	0.91425	.007850
Support Vector Machines	0.91168	.006821
Decision Trees	0.91251	.007059

Project Design

Figure 7: Machine Learning Workflow [16]



This project will follow a common machine learning workflow. Following figure 7, fetching the datasets that will be used for analysis is the absolute first step in which two datasets have been chosen. There will be a level of cleaning and combining that will be performed in order to have a "master set" for analysis. The next step is to validate the accuracy of the model that was chosen, thus, the data will be split into training and validation sets (70% train, 30% validation).[17]

Three columns will be added to the master data file which originally contains police incident information. The combined dataset will include these columns from The United States Census- census area, median age and median income. An interactive map similar to figure 3 will be created that will display the number of police incidents by crime category per a particular census area, median age and median income. A gradient in a single color palette will be for ease of viewing incident volume per census area. After benchmarking the accuracy of the untuned chosen model, then will perform hyper parameter tuning in batches using GridSearchCV. This will allow for keeping track of best outcomes for each parameter and apply to the next batch. After this step, and viewing results, the final goal will be to create a new map that displays similar information as before, but with projected police incidents by crime category by census area.

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. Using XGBoost in Python. (n.d.). Retrieved March, 2019, from https://www.datacamp.com/community/tutorials/xgboost-in-python
13. What is a Decision Tree Diagram. (n.d.). Retrieved March, 2019, from https://www.lucidchart.com/pages/decision-tree
14. 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
15. Schapire, R. (n.d.). *Explaining AdaBoost* (pp. 1-16, Tech.).http://rob.schapire.net/papers/explaining-adaboost.pdf
16. Machine Learning with Amazon SageMaker. (n.d.). Retrieved March, 2019, from https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-mlconcepts.html

17. 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