Data 670, Data Analytics

Wesley Clark

8/12/18

Final Report: Washington D.C. Crime

Table of Contents

Analysis Overview

An in depth analysis of crime was selected for the subject matter of this report. At the onset of this investigation, several cities were considered as potential sites. This list of cities included Philadelphia, Chicago, New York, San Francisco, Miami, Los Angeles, Baltimore and Washington D.C. Datasets were located from a variety of sources that show crime over the past decade at each of these locations. Initially these datasets were appealing. The level of detail was surprisingly specific and many cities. A dataset from Chicago alone contained over 6.8 million instances instances describing crime over time. Furthermore, these datasets were large, already compiled into one file and little imputation was needed before analysis.

There was however one stand out city - Washington D.C. It became increasingly apparent that this city was an excellent choice for this investigation. The first reason being that the data collected on this city was very high quality. A separate dataset for each year from 2008-2017 complete with all registered crime was publicly available. These datasets included several important key factors such as location, offense and method. The biggest problem with these D.C. datasets is that they were lacking in some crucial variables, such as violence classification. Sites such as Baltimore and Philadelphia contained this variable already. To resolve this, database imputation was required. The greatest asset these D.C. datasets displayed was their source. Ultimately, Washington D.C. was selected as the subject of this investigation because the origin of the data was highly reputable. Datasets were derived from opendata.dc.gov. Each dataset was generated from an official website. Having the data sourced from an official government origin carried a level of unquestioned integrity that other publishers were

lacking. It was clear that the District of Columbia presented a unique opportunity. As a district, it does not belong to any state and is classified differently. This has affected the ability for D.C. to gather and publicly publish official data. Initially, it was decided to study several cities, and compare how crime has changed over time among them. While high quality data existed for Chicago and San Francisco especially, it was not attached to a reputable source. This held true, for every major city, except D.C.

Project Scope

The scope of this investigation includes an analysis of violent crime in Washington D.C. from 2008-2017. Crime was classified as either violent or non-violent. Models were constructed to predict this classification. It was decided that data source integrity was simply too important to gloss over. The subject of this investigation, violent crime, carries ethical implications alongside. As such, only the highest quality data sources could be considered. Instead of comparing crime across multiple cities, the scope of the investigation was narrowed to a more in depth investigation of Washington D.C.

One issue D.C. has struggled with in since the 90's is how to deal with violent crime. In 1991, D.C. was dubbed the 'Murder Capital' experiencing 482 homicides(Telegraph, 2015). The professional basketball team was once titled the Washington Bullets. The name was changed to the Wizards in order to distance the franchise from inner city violence(Fact Check, 2018). While there has been overall improvement since that time, the level of violent crime continues to be well in excess of the national average(Washington D.C. Safety Organization, 2018). Every level of society from

community to national collectively carries the weight of crime. Cost generated by crime impact every citizen. In the United Stats alone, 23 million offenses were committed in 2017, resulting in $15 billion worth of economic loss directly passed to the victims. The economic loss on government expenditures associated with these crime cost during this time was a hefty $179 billion(McCollister, 2010) In other words, on a national level government spending outweighs economic loss from the victims perspective in same crimes at a ratio of 12:1. This striking difference highlights the inefficiencies associated with the process of retroactively fighting crime. These expenditures are necessary, and will in no way fully disappear. There is no corner that can be cut to reduce this ratio. Each step of a police investigation is crucial and necessary. Instead, the most effective way to lower the amount of money spent retroactively solving crime is to proactively prevent future crime. Every citizen stands to benefit from this.

In 2017 33,206 registered crimes were officially committed in Washington D.C. Crime rates vary tremendously around the world. The United States struggles with one of the highest homicide rates worldwide among developed countries(Hall, 2015). It is a tremendous burden on every police force many of which are understaffed and overworked. Furthermore, these unproductive policies are paid for by the public through taxes. I propose that in order to more efficiently deal with crime, a paradigm shift is imposed. This shift will employ preventative measures that aim to prevent crime from happening in the first place. When accomplished, the effect will be three fold:

1) Crime rates will decrease over time.

2) Police resource allocation will improve, resulting in increased efficiencies.

3) Society within Washington D.C. will benefit on a local and community level from living in a city with less crime.

This will be accomplished by better understanding what influences crime. As this proactive shift occurs, police will be more able to precisely target the specific instances of violent crime they are able to solve. An independent analysis of all registered crimes committed in each city over a multitude of years must be done. This is a continuous process that is different in each city. Some crime fighting principles that held true for 2017 invariably change in 2018. Crime principles will also change from city to city. The time period of this investigation is 2008 - 2017. The problem I am seeking to achieve is to achieve a lower the violent crime rate in Washington D.C.

Business Understanding

The purpose of this investigation is to enable the police force to lower crime and allocate resources more efficiently through data driven decision making. Specifically the Metropolitan Police of D.C. will benefit from this analysis. Many crimes only undergo a shallow investigation, simply because the amount of resources needed to solve a crime are not available. This process is costly. As a result a lot of solvable crimes remain unsolved and low on the priorities list. While preventative measures do not seek to solve crimes that have already been committed, they do attempt to gain a more comprehensive understanding of crime going forward. Preventative Predictive Policing(PPP) must be used to describe the act of police attempting to prevent a crime before it is committed

using data analytics(Avery, 2018). Furthermore, by separating violent and non-violent crimes within a city, the police will be able to concentrate on lowering violent crimes. Each individual precinct as well as MPDC will be affected by this report. Society on the whole stands to benefit from increased efficiencies within the police force. While some crime will always persist, the aim is to decrease the worst kinds of crime as much as possible. This is accomplished by police effectively allocating resources to correct locations and times. This feedback loop will continue, further promoting the cycle of decreased crime and taxes.

Stakeholders

The stakeholders for this project are those who fund the police force. Considering how the police force is funded through public money, the funding is therefore passed onto the public. Each tax payer is therefore a stakeholder. Furthermore, residents within the city are affected as well. Each resident of the city stands to gain from decreased crime throughout the city. Furthermore, as efficiencies improve, it is possible that the amount of funding necessary for this project will decrease.

Stakeholders

The stakeholders for this project are the people who fund the police force. The police department is funded through public taxes. As such, the stakeholders are the tax-payers. Each tax payer will be benefitting directly through this process, as efficiencies improve. It is possible that the police force would be able to maintain lower crime rates with less

funding. Certainly the residents within the city stand to benefit the most. A focus on PPP will enable MPDC to improve the quality of all of the services it provides. Additionally, a more comprehensive understanding of how crime moves within the city will be advantageous to MPDC. Reports developed through this analysis will be referenced in training programs. The business area in effect is all of the business within Washington D.C. As previously mentioned, the economic burden associated with crime is immense.

Business Objectives

The first and foremost objective of this study is to give city-wide and local law enforcement actionable data enabling practical data driven decisions. PPP is not the traditional approach to crime. Analytics must be used to compare patterns in crime in the city of Washington D.C. Many of these trends are ephemeral in nature and require continued maintenance and updates. This view of crime prevention is not static. It is a dynamic approach that attempts to keep up with data in real time in order to be able to more accurately predict future crimes that have not yet been committed.

The first business objective is to arm the police force with a new weapon in their fight against crime, PPP Models. The aim of these models is to give police a better understanding of how to predict violent crime.

The second business objective is decreased funding. This will be accomplished by increasing efficiencies within MPDC. As efficiencies increase, resources that have previously been used ineffectively will be needed to a lesser extent. Decreased funding is an objective that the stakeholder cares about, as it affects them directly. Lowering taxes

is something all stakeholders can agree on. The third business objective to allocate police units more efficiently. This is accomplished but gaining a higher level understanding of the factors that affect violent crime.

Business Success Criteria

This data will have a three Key Performance Indicators(KPI's). The first is development and maintenance of PPP models. These models must be accurate to provide value to the police force. In order for this to be true, a classification rate greater than **60%** must be derived. This is the minimum threshold required to influence decision making on a higher level within the police force. The second success criteria involves identifying the key variables that most influence violent crime. By determining which factors affect violent crime the most, police will have a concrete heir achy to reference on the go. Lastly, the third KPI is achieving lower than predicted yearly rates of violent crime. This is a process that can take time, and may change over time. It is expected that a lag is observed before receiving any improvement, as this is a proactive approach.

Data Preparation

For this analysis two main tools were used to prepare the data, Watson and SAS Enterprise Miner. Apache Spark enables big data to be more easily processed on a large scale. There is a lot of crime in D.C. As such, Spark was used in order to be able to work with such a large amount of data. Spark is conveniently listed as a service within Watson, and their coupling is a natural fit. Within Watson, Spark was employed to load and

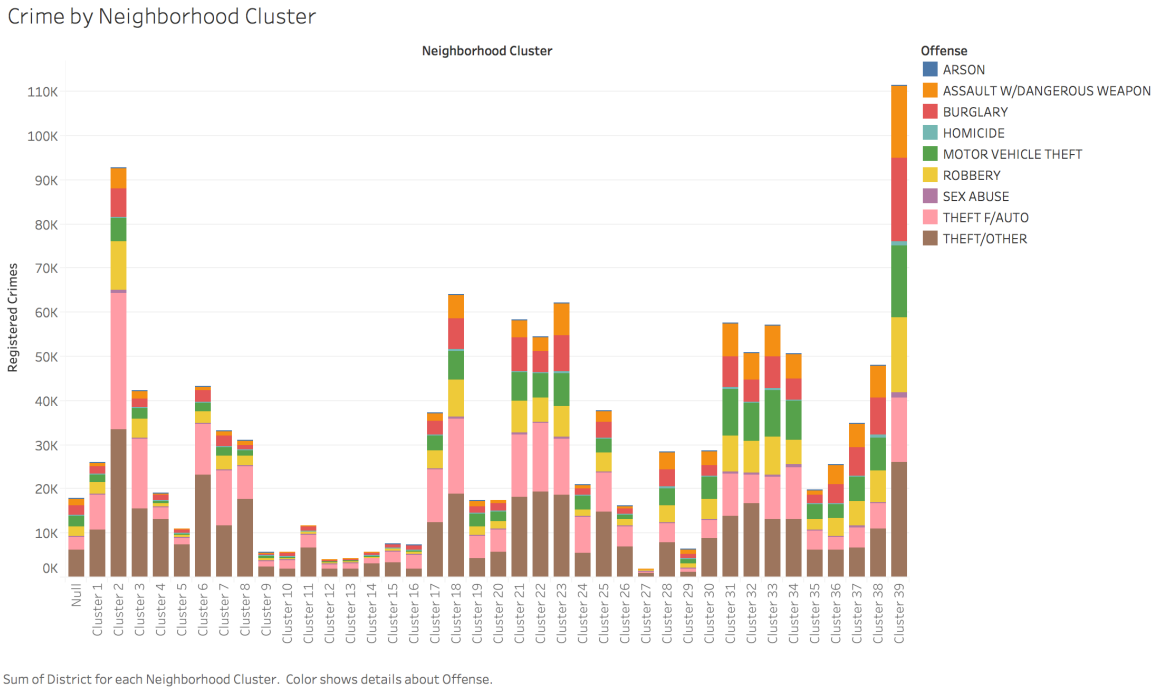
impute the data. Each dataset contained about 33,000 entries. Separately it would be possible to look at them in a standard environment using Python or R. For a complete analysis to be conducted, it was necessary to merge each dataset. This created a file too large for work in RStudio.

To begin, each dataset was loaded into SAS EM. Next the data was merged into one dataset using the merge node. Data was then exported and loaded into Watson for further imputation. In Watson, a new dataset that details voting percentages throughout the city of D.C. was joined to the single D.C. crime dataset. This dataset was joined at Voting Precinct, created a new variable, Trump Vote, that reflected the percent of people who voted for Donald Trump in the 2016 Presidential election. Next the results were imputed to reflect the Trump Vote for all years, not just in 2016. This involved editing the database to replace missing values with appropriate Trump Vote percent for each year. Finally the dataset was inspected to ensure the join occurred properly. The data was prepared and ready for cleansing.

Data Cleansing

Once merged, 24 columns within the single final D.C. dataset were present. Many of these columns were redundant and did not need to be repeated. The location of a registered crime was described in 9 different columns. Utilizing Watson, the following columns were removed: X String, Y String, Block, X Block, Y Block, ANC, BID, PSA.

It was decided that although it was possible to narrow down location to one single variable, it would be better to leave more than one(IBM, 2017). Block Cluster, Longitude and Latitude were next removed. The remaining location variables included Neighborhood Cluster, Voter Precinct, and Ward. It was decided that these were the three most logical choices for determining location. Washington D.C. was divided into 39 neighborhood clusters, 8 Wards, each of which is subdivided into multiple Precincts.

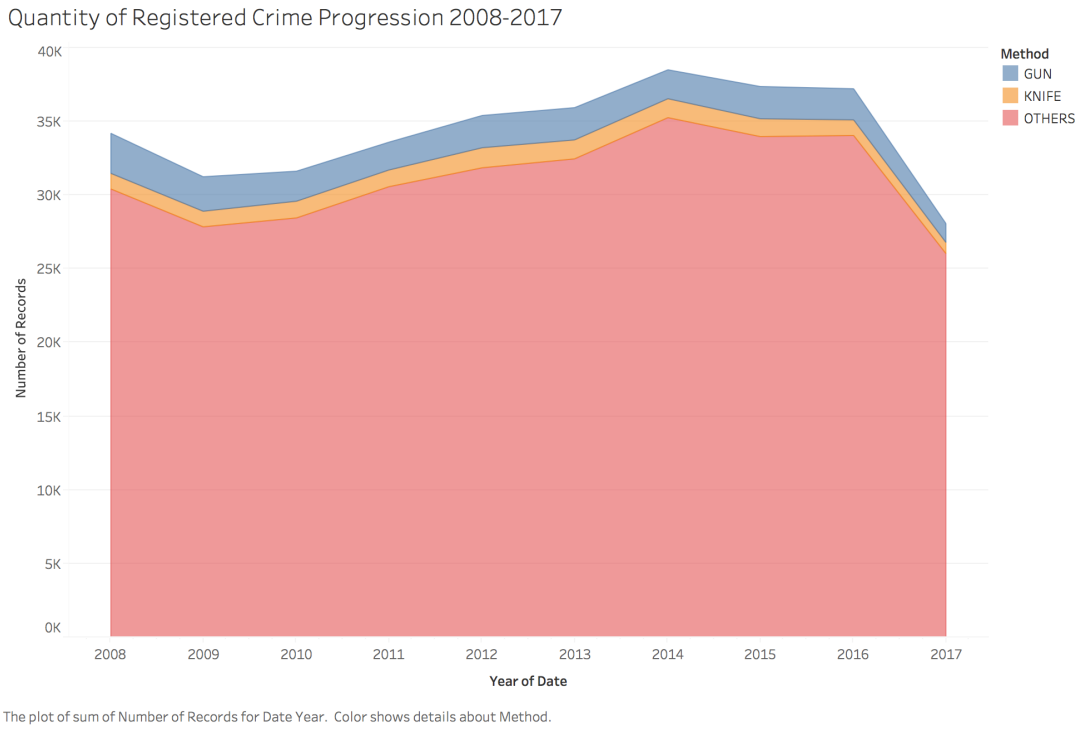


(Figure 1: Types of Crime per Neighborhood Cluster)

Data Transformation

A new column is created to display our target variable - Violent Crime. In order to accomplish this all crime was divided into either violent or non-violent sub categories. First, the OFFENSE column was duplicated. The new column received the name VIOLENT/NONVIOLENT. Next the strings are substituted. The substrings HOMICIDE, ASSAULT W/DANGEROUS WEAPON, SEX ABUSE, ASSAULT W/ DEADLY WEAPON, MURDER, MANSLAUGHTER are replaced with 1, indicative of a violent crime. All other values were non-violent. These included THEFT, MOTOR VEHICLE THEFT, BURGLARY, ROBBERY, ARSON and were replaced with 0. This process is very important to our investigation - our target variable has been created and is now set for analysis.

The other column that was imputed, Trump Vote, was derived in order to view the percentage of people who voted for Donald Trump in the corresponding location. It should be mentioned that D.C. is a historically democratic area. The range of votes received for Trump were from 1.66% - 7.32%. This was imputed to be 0.0166, 0.0732. Lastly duplicate entries were detected and deleted. Null values were left as is. It was decided that instead of trying to predict these values, it would be better to simply exclude any instances that contained a null value from the investigation entirely. Our fully merged dataset contained 337,000 instances. After deletion of duplicate entries and instances with null values ~310,000 instances remained. This was more than enough to conduct our analysis with confidence in our sample size.



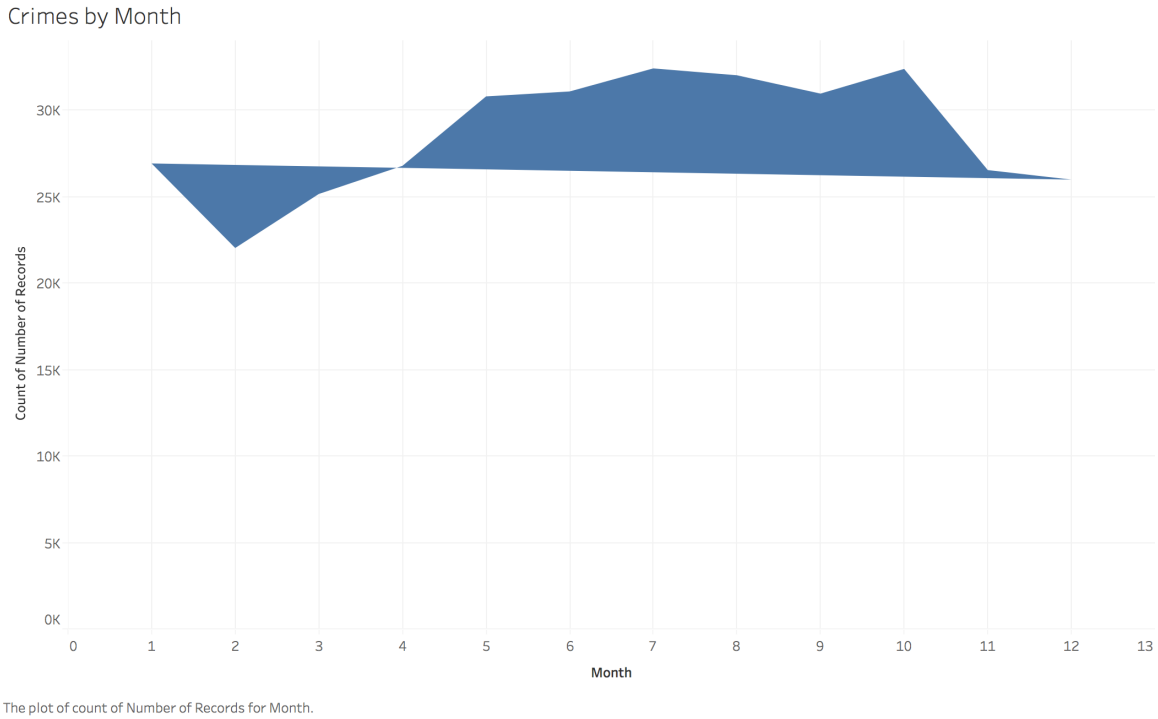
(Figure 2: Types of Crime per Year)

Data Analysis

The first tool that was be used for this analysis is SAS EM. SAS is the tool of choice for many data scientists. It is unique in its ability to create many kinds of models. Data was partitioned and Logistic Regressions created. SAS EM enables many models to be made with a smile drag and drop approach. Furthermore, the ability to compare different models is innate to SAS EM(SAS, 2015). Through SAS EM 3 different Logistic Regressions were created. Stepwise, Backward and Forward logistic regressions.

The next tool that was be used for analysis is Tableau. Tableau is similar to SAS EM in that it uses a drag and drop approach as well. Additionally, Tableau has the functionality to be able to handle large data sets. An initial investigation of the merged

data set was already conducted using Tableau. Tableau is able to visualize our data significantly better than Python or R, even with packages added on. Through Tableau, it was observed that there is a tremendous variation of crime on a monthly basis. This correlates with temperature. During the hotter months of summer there was increased time, as compared to the decreased crime over colder winter months.



(Figure 3: Crime by Month)

Predictive modeling was continued in Watson. Watson has built in functionality to enable R coding. In Watson 3 more Models were constructed. These included a Decision Tree, Random Forest Tree and Gradient Boosted Tree.

Modeling

Predictive Model 1: Logistic Regression

Logistic regressions are models are a widely used multivariable method used to model dichotomous outcomes. They are arguably the most common type of a model for binary classifications. This model works by measuring the relationship between a categorical dependent variable and independent variable(s) using the logistic function. Importantly, a standard logistic distribution of errors is assumed for this method. This differs from a profit regression, which assumes a normal distribution of errors. Logistic regressions are especially easy to explain as they predict probability of a particular outcome. Logistic regressions are also known as an alternative to linear regressions(which use a continuous independent variable). The two major drawbacks that encumber logistic regressions are 1) results improve significantly as sample size increases(poor performance with small sample size) and 2) other kinds of models typically yield more accurate results in many situations. Although the logistic regression may be outperformed by other models, our sample size of over 300,000 instances is certainly large enough to test the logistic regression out.

Three logistic regressions were conducted in SAS EM. All results were similar, but the best of which was the Stepwise Logistic Regression and achieved an Area Under PR of 0.8590. The area under ROC was 0.925. At first these results were surprising. With many highly advanced techniques, and the Logistic Regression being one of the older approaches it was interesting to see such a high area under ROC Curve. This area under

the ROC Curve represents the sensitivity corresponding to a decision threshold - violent or non-violent. In general, the higher the are under the ROC Curve, the more accurate the classifier. More specifically, this area can be interpreted as an aggregate measure of performance across all possible classification thresholds. The Logistic Regression performed the best of any model.

Logistic Regressions are used in a variety of fields from machine learning to medicine and even voter prediction. Because logistic regressions can be binomial, ordinal and even multinomial, there are a wide variety of situations to apply them. The key difference between logistic and linear regressions is the dependent variable. Predicting a continuous variable such as home prices will utilize a Linear Regression. Linear regressions will make nonsensical predictions for binary dependent variables, however. By applying the logarithmic function to the regression, binary variables are analyzed in logistic regressions. This makes logistic and linear regressions analogous in many ways.

There are however some issues with logistic regressions in general. First off, independent observations are required. In other words, observations cannot be related to each other. It is possible that there is some relationship amongst instances in our crime database. For example, two crimes may have been committed and registered at the same time. This was deemed irrelevant for our investigation. While it is possible that two people would commit the same crime and it could be registered twice, this would largely be an exception. An irregular occurrence of this would not take away from the independence of all other events. The second issue with logistic regressions is overfitting. There may be some overstating of the accuracy that is pushing the ROC

Curve over 90%. To compensate for this, the use of highly correlated variables was avoided. Admittedly, it was not expected that the logistic regression would yield the highest are under the ROC Curve. One thing that was learned as a result of this is that logistic regressions should be considered in most binary classification outcomes. Because of the success of this model, the scope of the expected results of this project are altered.

Achieving AUC >.90 is significant. Being able to differentiate between violent and non-violent crime to such a high degree is highly impactful. It is recommended that further study be done into each type of violent crime. Further investigation can include more specific for predictive analytics for each kind of violent crime(homicide, assault with deadly weapon, etc.).

Predictive Model 2: Decision Tree

For the second model a Decision Tree classifier was used. This type of learning repeatedly divides a plot of data using identifying lines. Each division can be thought of as a typical decision tree subdivision. This process may only take one single iteration if the data is simply divided by one line. In most cases, it takes several more steps in order to reach an equilibrium. There are two scenarios that cause the process to stop. The first is when the highest classification rates are met when it classes divided are pure. In this scenario, the data is perfectly classified. The other scenario in which this is met is when classifier attributes are received. Another way to think about this outcome is the best possible decision has been made.

The same featured columns and target variable were used as in the Logistic Regression(Target: Violent. Feature Columns: SHIFT (Integer), METHOD (Integer), WARD (Integer), DISTRICT (Integer), NEIGHBORHOOD_CLUSTER (Integer), TRUMP_VOTE (Decimal). Because we are attempting to decide whether the crime was violent or non-violent, a binary classification was conducted. In this model, 60% of the data was used to train the model, 20% to test the model and 20% of the data was left out. 20% of the data used to test the model is how we know how effective the model itself is. Once the model has been derived on the training data, it is tested on the testing data.

One of the reasons why this type of classification was chosen as a model is because it is easy to explain. Many people are familiar with decision tree's. Moving from that into a classifier is not a huge step, and one that can be put into words relatively easily. Another strength of this type of analysis is that it performs well in a large dataset, such as this one. Containing over 330,000 observations, there is ample data to train and test on. Never the less some limitations exist to this approach as well. A small change in data can result in a large change in the results. Additionally, this approach is known to be NP-complete, and as a result the greedy algorithm can weight locally optimal decisions heavily. When zooming out into many decisions, this can skew results.

The performance for the Decision Tree Classifier Model was good. An area under the ROC Curve of 0.87635 was achieved. Additionally, an area under the PR Curve of 0.83068 was realized. These numbers are both good, but neither is as good as was seen above in the Logistic Regression Model. One thing that was realized when performing

this analysis is that more columns could result in an increase of area under both curves. This model was done two times, once

Predictive Model 3: Random Forest

For the next model a Random Forest Model was done. This type of learning is a form of ensemble learning for classification, as well as regressions. RFM's work by constructing multiple decision trees during the training time. This is different than a typical decision tree, and corrects for overfitting. To differentiate from a decision tree, random groups of decisions are bagged together. This process is known as bootstrap aggregating. This process decreases the variance of the model without increasing the bias. The reason for this is that while a single tree may be very sensitive to noise in the data set, the overall average of the trees will not be. The process of randomizing the forests differ in that the modified tree learning algorithm selects and splits candidates in the learning process. The purpose of this differentiation is to avoid one or more features that may be highly correlated and over selected in the decision trees. In our case, each decision tree was bagged separately to predict whether or not the crime was violent. The aggregate of these trees collected and analyzed. My Random Forest Model achieved an Area Under the ROC Curve of 0.92474 as all as an Area under the PR Curve of 0.85627. Violent/Nonviolent was the target variable, and as usual - the dependent variables remained the same. I had expected the RFM to perform slightly worse than the Decision Tree Classifier. After some analysis, the ensemble nature of a RFM can lead it to a higher

accuracy. This has impacted the scope of expected results. It is important to note that this model did significantly better than the decision tree. It is possible that the decision tree approach was on the lower side naturally due to variance. The process of creating a forest and randomizing does allocate more resources with the goal of reducing variance in mind. The biggest downfall to this is that it is one more step to explain. While decision trees are especially easy to communicate, they have simply been outperformed.

Because the Random Forest Model achieved a an AUC of 0.92474, it was comparable to the logistic regression. Additionally, the AUPR was very similar as well. It appears these two models are working the best so far. This is again an appropriately high enough measure that it carries weight. The scope of expected results were exceeded in this model. Further, it is recommended that additional analysis go into specific type of violent crime in order to better understand how violent crimes differentiate. The scope of this project has been expanded.

Further RFM's are needed. To continue this investigation it is recommended that RFM's be developed for each type of violent crime. Once this has occurred a differentiation between each kind of violent crime can be conducted. Allowing police to be able to drill in with tremendous predictive specificity. It is possible that many of the insights from this kind of investigation will be common sense. Even still, being able to prove it statistically provides validation to that gut feeling. More to the point, being able to track how this changes over time provides a quantitative measure of change within crime itself.

Predictive Model 4: Gradient Boosted Tree

Because the random forest model greatly exceeded the decision tree classifier, it was decided to conduct a fourth model: Gradient Boosted Tree(GBT) classifier. GBT's are a machine learning technique that employ an ensemble of weaker prediction models. Most of the time, the are also decision trees. The key difference between RFM's an GBT's lies in the idea of gradient boosting. There are several different gradient boosting algorithms. The overall goal of these boosting algorithms is to increase classification accuracy. This is done by optimizing a cost function. If a direction is found to have a higher success, it will be selected in the individual decision trees that make up the ensemble more frequently. It was expected that this would yield the best results of all models. The AUC for this model was 0.876. It was disappointing to have such a low number. However it was observed that the AUPR(Area under Precision and Recall Curves) was the highest of any models at 0.891. The overwhelming majority of instances in this dataset were indeed non-violent crime. Because of this, class imbalance is present. When this is the case, it is important to consider both AUC and AUPR curves. Although this is not the highest model in terms of AUC, it was significantly higher than other models in AUPR. This led me to believe that it is a more well rounded model.

	Logistic Regression	Decision Tree	Random Forest	Gradient Boosted
Area Under ROC	0.925	0.87635	0.92474	0.8767
Area Under PR	0.8590	0.83068	0.85627	0.891

(Image 4: Model Comparison)

Results

All models performed reasonably well. Although the Decision Tree model performed the worst in both curves, it still yielded fairly high predictive accuracy. It does yield insight, and its ease of explanation is not to be overshadowed. The Random Forest and Logistic regressions performed very similarly well. They each have tremendous predictive accuracy. We are most interested in predicting violent crime. Because the overwhelming majority of the instances person in this database are non-violent, a cost function was employed to compensate for this(Chioka, 2017). Never the less, the true champion model of this analysis is the one that predicts violent crime the best. We are not interested predicting true non-violent crimes. Although ROC Curves are generally easier to explain, it less appropriate in this instance(Google ML, 2017). Our true negatives are not important. Because of this, the Area under PR curve is the most important measure of success. As such, the champion model is the Gradient Boosted Tree.

The Gradient Boosted Tree most accurately predicted violent crime of any model at a rate just shy of 0.891. Although other models achieved a higher AUC, that was of less importance. Being able to predict the value of violent crime is the focus of this investigation. It is especially interesting to note that the each tree model performed better in this regard as we moved forward. Decision Tree AUPR < Random Forest AUPR < Gradient Boosted AUPR. Further study is needed in order to decipher what differentiates each type of violent crime. It is recommended that more study be done in this regard, using all 4 above models.

Future Analysis

The results of this analysis concretely enable the police force to predict violent crime effectively. There is more room for improvement in this arena. Future study should include more specific PPP models. This can include predictive models for each kind of violent crime. By separating each offense, police are even more able to allocate resources efficiently. In this study violent and non-violent crimes were differentiated. Further study should differentiate among violent crimes. Lastly, it is recommended that a new variable, temperature be imputed into the database. While temperature was taken into consideration in this report, adding a temperature column to the data would likely impact the models positively. Increasing predictive classification through the addition of correlated variables is recommended.

Executive Summary

Washington D.C. was chosen to conduct a comprehensive analytic modeling and preventative policing report. This type of policing is labeled “Predictive Preventative Policing”. In this approach the focus is shifted from solving past crimes that have already been committed to allocating resources to more efficiently solve future crimes. Crime was classified as either violent or non-violent. The focus of this investigation was to lower crime over time. This is done by developing a more comprehensive understanding of violent crime. The Washington Metro Police Force is able to utilize this data to more efficiently deploying troops at the correct time and place. Once implemented people within Washington D.C. will benefit from lowered violent crime rates and increased safety. Additionally as police are more correctly positioned, response times to crime will decrease. In order to accomplish this, several models were constructed. Binary classification in the form of Logistic Regressions, Decision Trees, Gradient Boosted Decision Trees and Random Forest Trees were conducted. In the end, The Gradient Boosted Tree was the champion model, achieving >92% violent crime classification rate. Furthermore it is recommended that increased police presence be imposed during the warmer months.

References

1) Washington D.C. Homicide Rates (Avery). Retrieved July 17, 2018, from https://mpdc.dc.gov/page/statistics-and-data

2) Fact check/Is Washington, D.C., among the safest cities in the country? (n.d.). Retrieved July 15, 2018, from https://ballotpedia.org/Fact_check/Is_Washington,_D.C.,_among_the_safest_cities_in_the_country?

3) McCollister, K. E., French, M. T., & Fang, H. (2010, April 01). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2835847/

4) A. (2015, December 11). Mapped: Which countries have the highest murder rates? Retrieved from https://www.telegraph.co.uk/news/uknews/crime/12037479/Mapped-Which-countrieshave- the-highest-murder-rates.html

5) Hall (2015, December 11) Mapped: Which countries have the highest murder rates? Retrieved from https://www.telegraph.co.uk/news/uknews/crime/12037479/Mapped-Which-countrieshave-the-highest-murder-rates.htmlMpdc.

6) Refine Data. (2018, May 17). Retrieved from https://dataplatform.cloud.ibm.com/docs/content/refinery/refining_data.html Derived from IBMWatson Help

7) SAS Enterprise Miner: Imputing Missing Values. (n.d.). Retrieved from https://video.sas.com/detail/videos/sas-enterprise-miner_/video/2779613387001/sas-enterpriseminer:-imputing-missing-values

8) Fact check/Is Washington, D.C., among the safest cities in the country? (n.d.). Retrieved July 15, 2018, from https://ballotpedia.org/Fact_check/Is_Washington,_D.C.,_among_the_safest_cities_in_the_country?

9) Classification: ROC and AUC | Machine Learning Crash Course | Google Developers. (n.d.). Retrieved from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc Google Developers Inference, 2017

10). Differences between Receiver Operating Characteristic AUC (ROC AUC) and Precision Recall AUC (PR AUC). (n.d.). Retrieved from http://www.chioka.in/differences-between-rocauc-and-pr-auc/S. (2017, October 24).

Appendix

Figure 1: Types of Crime per Neighborhood Cluster

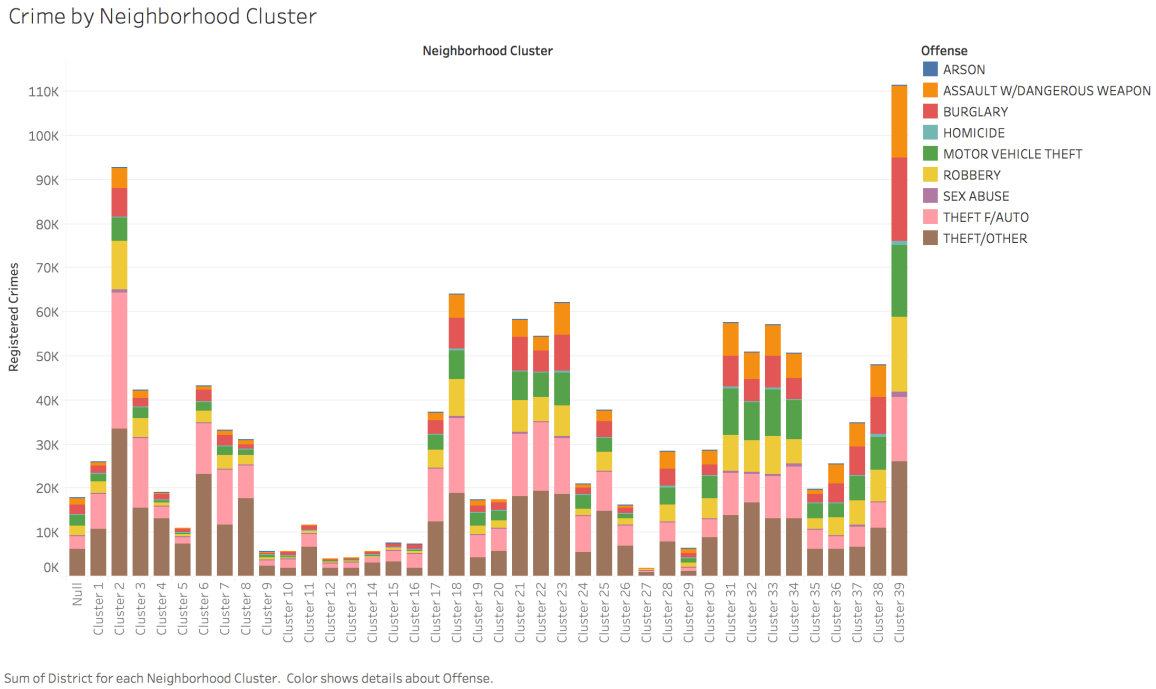


Figure 2: Types of Crime per Year

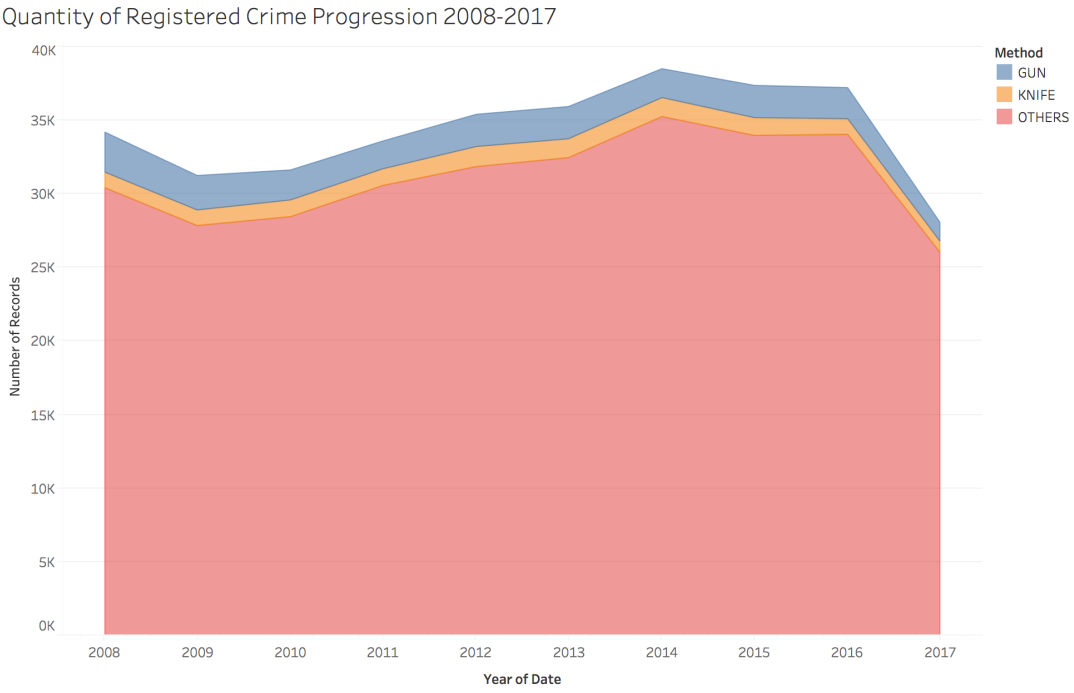


Figure 3: Crime by Month

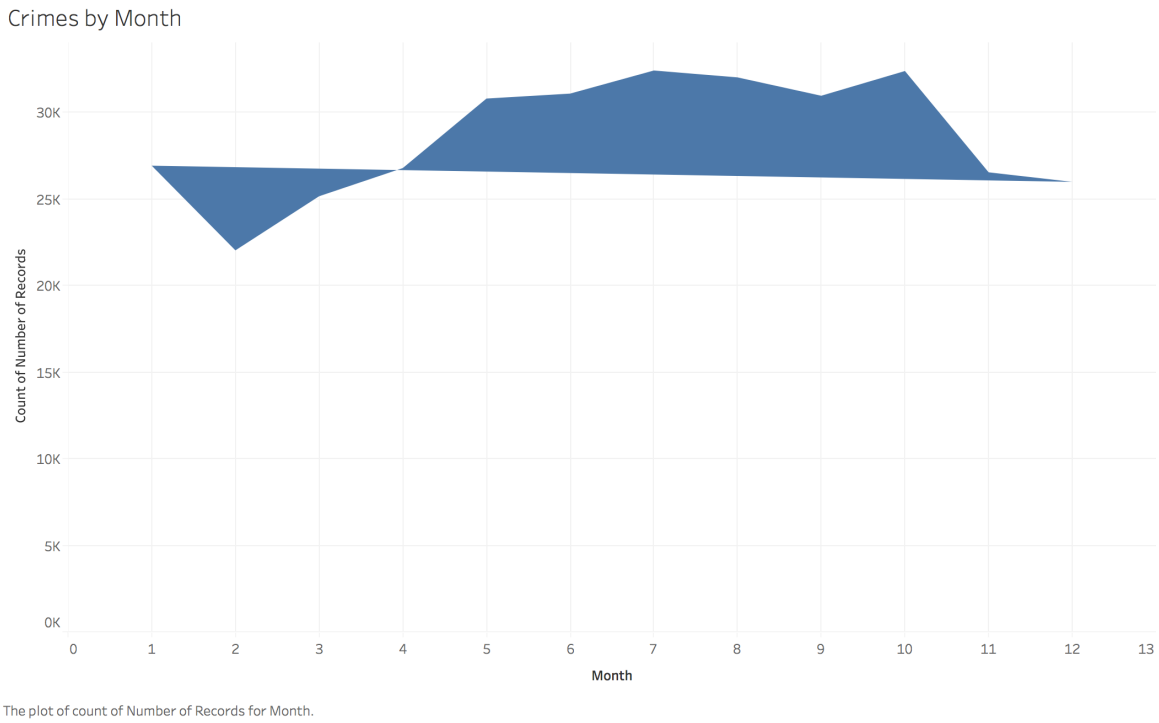


Image 4: Model Comparison

	Logistic Regression	Decision Tree	Random Forest	Gradient Boosted
Area Under ROC	0.925	0.87635	0.92474	0.8767
Area Under PR	0.8590	0.83068	0.85627	0.891