Transportation Recommendation

Team DC Criminalistics
Georgetown Certificate in Data Science (Cohort 14)

Greg Barbieri
Tara Brosnan
Dan Schorer

15 June 2019

Table of Contents

Abstract

The goal of the project is to use machine learning to predict crime rates. Regardless of available data, it is difficult, if not impractical, to predict whether a particular individual will be a victim of violent or non-violent crime in an area. The team hypothesized that it was possible to predict crime rates by block group in Washington DC using features such as weather, time of day, and location. The team ingested data from the US Census Bureau, DC Metropolitan Police Department, and Dark Sky website. After wrangling, feature generation and target rescaling, the team had about 180,000 instances and 26 features. Feature evaluation limited the selection from 26 to 11 and the team selected 6 features to model crime rates. The team used classification models to predict crime rate buckets of low, low-medium, medium, medium-high, and high. Overall, a bagging classification model with a decision tree estimator outperformed other models tested such as K-Nearest Neighbors and Random Forest models. Overall model accuracy was 82% percent, and all models, including those with economic and demographic data had trouble accurately predicting crime rates in the high category as measured by false negatives and visualized by the confusion matrix, while more certain in predicting crime rates in the medium, medium-high, and medium-low categories.

I. Introduction

According to national statistics on violent crimes from the FBI's Uniform Crime Report, Americans are safer now than they have been in the last half-century. Figure 1 shows the long standing upward trend in national violent crime rates, measured as crimes per 100,000 people, between 1960 and the early 1990's.[1] The trend reversed and the national violent crime rate fell 50 percent by 2014 from its peak in 1991.

Trends for US cities are similar to national trends in that violent crime rates have been falling since the early 1990's, but not all cities have seen decline evenly or at the same rate. A report published in 2016 discusses how crime rates by and trends vary wildly from city to city, with New York experiencing all-time lows of violent crime, but murder rates in Chicago remain persistently high.[2] Without making simplistic or misleading analyses about why various cities crime rates differ, DC maintains a violent crime rate above notable places such as Philadelphia, PA, New York, NY, Seattle, WA, and Los Angeles, CA.

Figure 1: National and City Violent Crime Rates

Figure 1: Violent crime rates, Federal Bureau of Investigation, Universal Crime Reporting Statistics Database, 1960 - 2014 (left), 1984 - 2014 (right). Extracted 5/25/2019.

A company that specializes in compiling and analyzing hyper-local real estate data ranked Washington, DC as 72nd in a list of 100 of the nation's most dangerous cities with a crime of 10.2 per 1,000 people, and only safer than 4% of U.S. cities included in the analysis.[3] These

[1] Violent crime rates, FBI Uniform Crime Reporting Statistics Database, Extracted 5/25/2019.
[2] Matthew Friedman, Ames Grawert, James Cullen, Crime in 2016: An Updated Analysis, Brennan Center for Justice, 2016, http://www.brennancenter.org/sites/default/files/analysis/Crime_in_2016_Updated_Analysis.pdf)
[3] Andrew Shiller, Top 100 Most Dangerous Cities in the US, Neighborhood Scout, 2019, https://www.neighborhoodscout.com/blog/top100dangerous)

statistics include both violent and property crimes defined by the FBI's Uniform Crime Report, concluding that the likelihood of becoming a victim of violent crime is 1 out of 100.[4]

Regardless of available data, it is difficult, if not impractical, to predict whether a particular individual will be a victim of violent or non-violent crime in an area. However, there is an expansive literature on modeling crime rates, types, and location to assess the validity of various crime-fighting policies and for "allocating resources for crime reduction initiatives."[5] The methods and data sources used to predict crime have varied, whether the goal was to predict incidence, area-levels, or types of crime. A more recent paper explored clustering of crime using multinomial logistic regressions[6], another used OLS models to model totals of property crime in an area[7]; and a study published in 1981 uses stochastic equations to model various crime rates using data from 1947 - 1972.[8] This team decided that it is possible to help inform residents by classifying and predicting crime rates at a specific time and in a specific area in Washington, DC using historical data.

II. Hypothesis and Motivation

The team's original hypothesis was that it is possible to predict the likelihood of a type of crime, such as violent or non-violent, given an area in Washington, DC. The team planned to prove the hypothesis by defining an individual instance as each reported crime, labeled by the DC Metropolitan Police as either violent or property, with numerous subcategories. However, the team realized that this idea was not a novel approach and prediction was only feasible with the conditional assumption of a crime taking place. This limitation on the original hypothesis is a result of the data, as the team only has access to data when a crime was reported. The team's alternative hypothesis is that it's possible to predict the crime rate in an area, defined as the number of crimes per 100,000 people for a particular year, month, day and time of day. The assumption is that the team can only predict crime rates in areas that have experienced some crime, otherwise, there is no historical data to use for prediction. The assumption is a similar assumption to the previous hypothesis, but a range of crime rates can be predicted, making it more useful to compare to crime rates and assess the risk relative of surrounding areas. As time progressed and ideas for the data product were refined, the team realized that relative risk is best predicted categorically, such as high, medium or low crime rates. To facilitate both the hypothesis that it is possible to predict crime rates and report predictions relative to crime rates overall, the team transformed the problem from one of regressions to classification. The final hypothesis is that it is possible to predict crime rate buckets, defined as low, medium or high.

The goal of the project is to show that machine learning techniques can predict categories of crime rates levels using historical data on reported crimes combined with weather and

---

[4] Ibid.

[5] Andromachi Tselon, Chris Kershaw, "Predicting Crime Rates, Fear and Disorder Based on Area Information, Evidence from the 200 British Crime Survey", International Review of Victimology, Vol.12, 2005, pp.293-311

[6] Andresen, Martin A. "Predicting Local Crime Clusters Using (Multinomial) Logistic Regression." Cityscape, vol. 17, no. 3, 2015, pp. 249–262. JSTOR, www.jstor.org/stable/26326975.

[7] Alan Trickett, Denise R. Osborn, Julie Seymour, Ken Pease, What is Different About High Crime Areas?, The British Journal of Criminology, Volume 32, Issue 1, Winter 1992, pp. 81–89, https://doi.org/10.1093/oxfordjournals.bjc.a048181.

[8] Cohen, Lawrence E. "Modeling Crime Trends: A Criminal Opportunity Perspective." Journal of Research in Crime and Delinquency, vol. 18, no. 1, Jan. 1981, pp. 138–164, doi:10.1177/002242788101800109.

socio-economic data. Based on the predictions, the team will make recommendations as to the safest transportation options in an area of Washington, DC.

Washington, DC ranks high nationally in public transit ridership. A 2014 analysis by FiveThirtyEight shows Washington, DC as ranking third in public transit trips per capita among the 54 largest urban areas analyzed.[9] High per capita transit ridership indicates that a program to recommend transit options based on predicted crime rate would be useful, as both residents and tourists use public transit but may want more information regarding safe times of day and locations to travel from.

The team expects this product to be potentially helpful to individuals new or visiting the city, to help long-time residents understand distribution of crime in their areas, or to help transportation providers understand crime surrounding their pick-up and drop-off locations.

For example, individuals new to the city may be unaware of the crime rates in their neighborhoods relative to surrounding areas at certain times that they travel most, say in the morning, evening, or later on the weekends. If the place they are travelling from has a higher crime rate than surrounding areas, then taking a taxi may be worth the higher price than accessing the nearest metro station or bus stop. In addition, as companies such as Airbnb continue to expand opportunities for homeowners to temporarily rent their homes to tourists, traditional hotel or tourist locations are shifting. As these locations shift, our product can help tourists be more aware of the relative crime rates in the area of their rental and the relative risk associated with accessing nearby transportation options at various times.

III. Data Sources

Crime Data

The first source of data for this project is the DC crime data, collected and maintained by the DC Metropolitan Police Department. The application used to extract the data was the MPD's Crime Cards application, which makes available comprehensive data on reported crimes over the last 8 years and is updated daily.

The team ingested data on crimes reported from January 1st, 2008 through December 31st, 2018 to begin the project, totalling 381,067 instances of crime. The application exported crime data as a CSV file, which was ingested and stored in a SQL database table. The crime data includes information on the geographic, date and time, and categorization of the reported crime. Geographic variables include neighborhood cluster, police service areas, Census tract and block group, as well as detailed geographic coordinates of the reported crime. The data includes the date and time information, such as year, month day, hour, minute, second, of the report of the incident, the start and end of the incident, and the police shift the crime was reported. Lastly, the data contains information on the type of crime, reported, such as property or violent crime, with further break-outs including robbery, theft, burglary, etc. The data was aggregated by

[9] Reuben Fischer-Baum, How Your City's Public Transit Stacks Up, 2014, https://fivethirtyeight.com/features/how-your-citys-public-transit-stacks-up/

counting incidents of crime over various time blocks by Census block group, which will contribute to the areas crime rate, the team's target variable. More information on how the data was wrangled, aggregated and the design of the target variable are explained in the Section IV Data Manipulation.

Initial data exploration of the raw crime data set showed that, as expected, both violent and property crime counts vary by month and day of the week, seen on Figure 2 and Figure 3, respectively. Figure 2 shows that on aggregate, count of crimes reported in our data is lower in winter months than in summer months. It also shows that relative to other fall months, more crimes are reported in October. Figure 3 shows that more crimes are reported taking place on Fridays and Saturdays than during the rest of the week. This initial exploration also demonstrates that while there are some clear differences in reported crimes by month and by day of week, some of these differences are slight and would require standardization prior to machine learning, which is covered in section IV.

Figure 2: Violent (left) and Property (right) Crime Counts by Month

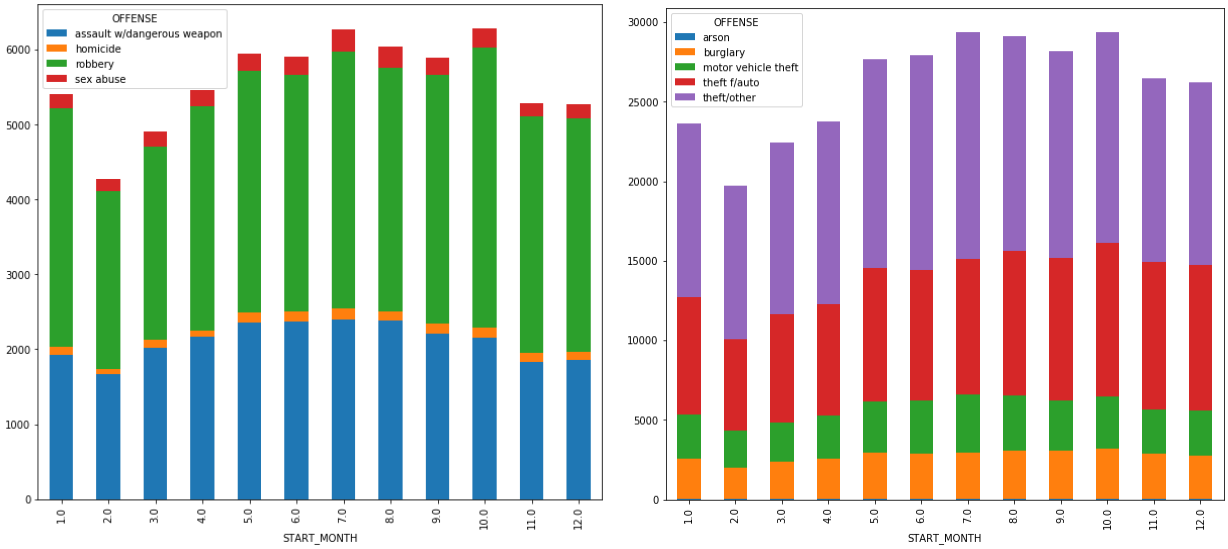
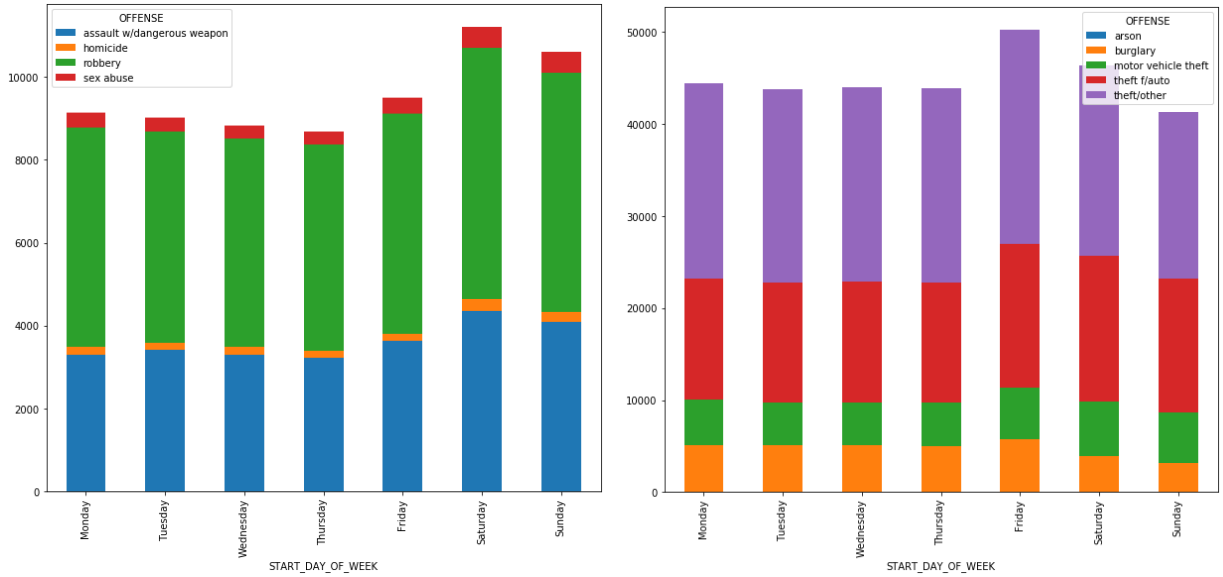


Figure 3: Violent (left) and Property (right) Crime Counts by Day of Week



Weather Data

Second, weather data was collected for each unique instance in the crime data. Unique instances in the crime data are defined by the geographic coordinates (latitude, longitude) and reported start date of the incident. The start date variable in the crime data includes the year, month, day, and time of the incident. In total, for the years 2008 to 2018, there are 379,610 unique instances of crime that the team attempted to collect weather data on.

Weather data was collected by sending the geographic coordinates, dates, and times associated with the reported crime to the Dark Sky API. Given the number of API requests required, the team used exception clauses in the ingestion program to reduce errors caused by gaps in internet connection, or missing elements in the crime data.

In total, the Dark Sky API returned 379,593 instances with at least one of our key weather variables populated, such as date, time, location, temperature, “feels like” temperature, and precipitation intensity, and percent chance are included on Table 2 below. Additional variables included with the weather data are visibility, UV index, cloud cover, dew point, wind speed, and wind gust. Dark Sky returned the data in JSON format and the team stored the JSON formatted as a SQL database table before wrangling and merging the weather, crime, and Census data.

The weather data was collected for each instance of crime data to support the original hypothesis where each instance of crime would be used to predict a crime type, violent or non-violent. As the hypothesis changed to predicting classified crime rates, the weather data was aggregated. How the weather data was aggregated is explained in Section IV Data Manipulation.

Table 1: Description of Key Weather Variables

	crime_rate	temperature	percip_probability	percip_intensity	uv_index	PerCapitaIncome	MedianHouseholdInc	HousingUnits
count	188,224.0	188,224.0	188,224.0	188,224.0	188,224.0	188,224.0	188,224.0	188,224.0
mean	80.7	59.9	0.1	0.0	1.6	48,634.0	77,993.8	787.2
std	46.2	17.6	0.2	0.0	2.3	30,614.1	44,770.5	402.8
min	18.5	3.0	0.0	0.0	0.0	3,787.0	6,548.0	10.0
25%	50.4	46.2	0.0	0.0	0.0	23,626.0	40,699.0	512.0
50%	68.7	62.0	0.0	0.0	0.0	42,186.0	72,467.0	692.0
75%	100.0	74.0	0.0	0.0	3.0	66,392.0	104,950.0	987.0
max	1,136.4	98.6	1.0	0.4	11.0	220,398.0	250,001.0	3,325.0

Census Data

Various metrics indicating the income, age[10], education, and population density[11] of an area are routinely tested as significant factors in predicting crime rates. For this project, census data was pulled from the US Census Bureau's American Community Survey (ACS) to support the calculation of crime rates as well as supplement modeling efforts. Population, income, age, and number of housing units were collected for the Washington D.C. area at both the Census Tract and Block Group geographic levels. Data was pulled for the years 2009-2017, however, block group data was not available for DC for years 2010, 2011, and 2012. Table 3 describes features of interest pertaining to income, population, and housing density. The estimated minimum population in any one block group with at least one person is 167 people and a max of 5,407. The minimally populated block group is an area that borders the national mall, an area likely to receive a higher number of people and interactions, despite the small official population estimate. Large variations in populations could generate outliers in locations that receive a high number of visitors and therefore may experience crime at a higher rate than their local population would suggest.

[10] Andromachi Tselon, Chris Kershaw, "Predicting Crime Rates, Fear and Disorder Based on Area Information, Evidence from the 200 British Crime Survey", International Review of Victimology, Vol.12, 2005, pp.293-311, used age, ethnicity, and poverty levels.

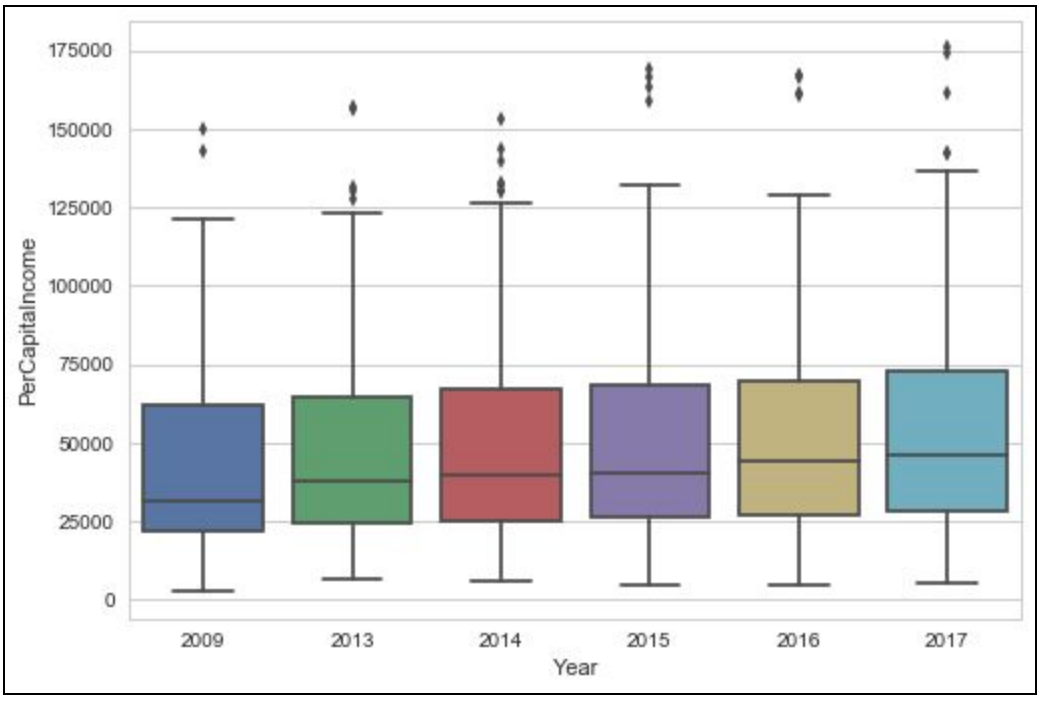
[11] Andresen, Martin A. "Predicting Local Crime Clusters Using (Multinomial) Logistic Regression." Cityscape, vol. 17, no. 3, 2015, pp. 249–262. JSTOR, www.jstor.org/stable/26326975, used characteristics on age, population, education, and number of parents in the household.

Table 2: Description of Key Census Statistics

	TotalPop	PerCapitaIncome	MedianHouseholdInc	MedianAge	HousingUnits	Year
count	2,586	2,586	2,586	2,586	2,586	2,586
mean	1,436	48,049	79,553	36	686	2,014
std	699	29,306	47,247	8	368	3
min	167	2,618	5,859	15	80	2,009
25%	932	24,522	41,586	31	427	2,013
50%	1,304	40,360	71,050	35	598	2,014
75%	1,756	67,972	105,376	41	872	2,016
max	5,407	175,725	247,222	66	3,325	2,017

Given the Census data is provided by year, this has the potential to significantly limit usage of the data together with the crime data. Specifically, the crime data is reported by the hour, but Census data only varies by year. With little variation in the Census data, machine learning models will not be able to learn from those data without serious considerations in the formation of test and training datasets. Figure 4 shows per capita income for the block group data by year. While there is significant differences among per capita across block groups, the picture shows less variation across time.

Figure 4: Per Capita Income by Year (Block Group)



WMATA, Capital Bikeshare Data

Lastly, transportation data was collected by the team in order to provide travel recommendations based on the predicted safety of a data users location, time of day, and the socio-demographic characteristics of the geographic location. If the location is deemed safe

using teams model developed in this project, then the team would recommend various transportation options in the surrounding geographic area. To recommend transportation methods, the team pulled the locations of all WMATA bus stops, rail stations and Capital Bikeshare stations in DC using the WMATA and Capital Bikeshare APIs, respectively. Each transportation method was stored as a SQL database table. Census tract and block group data was then appended to each rail entrance, bus station, and Capital Bikeshare station location using census geocode, a python wrapper for the US Census Geocoder API.

IV. Data Manipulation

Manipulation and wrangling of the data includes translating, transforming, and merging the various data sources together. First, the team downloaded crime data for incidents reported from 2008 through 2018, but the reported date-time is not necessarily the same as the start date-time, the latter being the preferred date-time for classifying and predicting crimes rates. Unique instances of crime defined by the start date-time and location of the reported crime and was used to collect weather data associated with each instance. Before the team merged the weather data and the crime data together, the team dropped non-numeric weather features correlated with other weather features. For example, the summary of the weather and an icon label for visual representation of the current weather are likely correlated with other weather features. In addition, the feature labeling the precipitation type, such as rain or snow, was also removed as seasonality and the remaining features on precipitation would likely capture the importance of the precipitation type.

Second, in order to calculate crime rates, the team needed to aggregate reported crimes. To facilitate aggregating crime data, the team labeled the reported start time of the crime in one of eight time-of-day categories, which are defined on Table 3 below. The goal of using time blocks was to capture the signal associated with the hypothesized asymmetric relationship between time of day and local crime rates, that is, crime rates may be higher in the evening. Then the team aggregated the weather and crime data by Census Block Group, year, month, and time of day to represent the frequency of crimes reported at the Census Block Group area level. The team aggregated to block group in order to calculate crime rates at the same geographic granularity made available by the US Census ACS data. While reported crimes were counted, the weather data associated with each reported crime was averaged over block group.

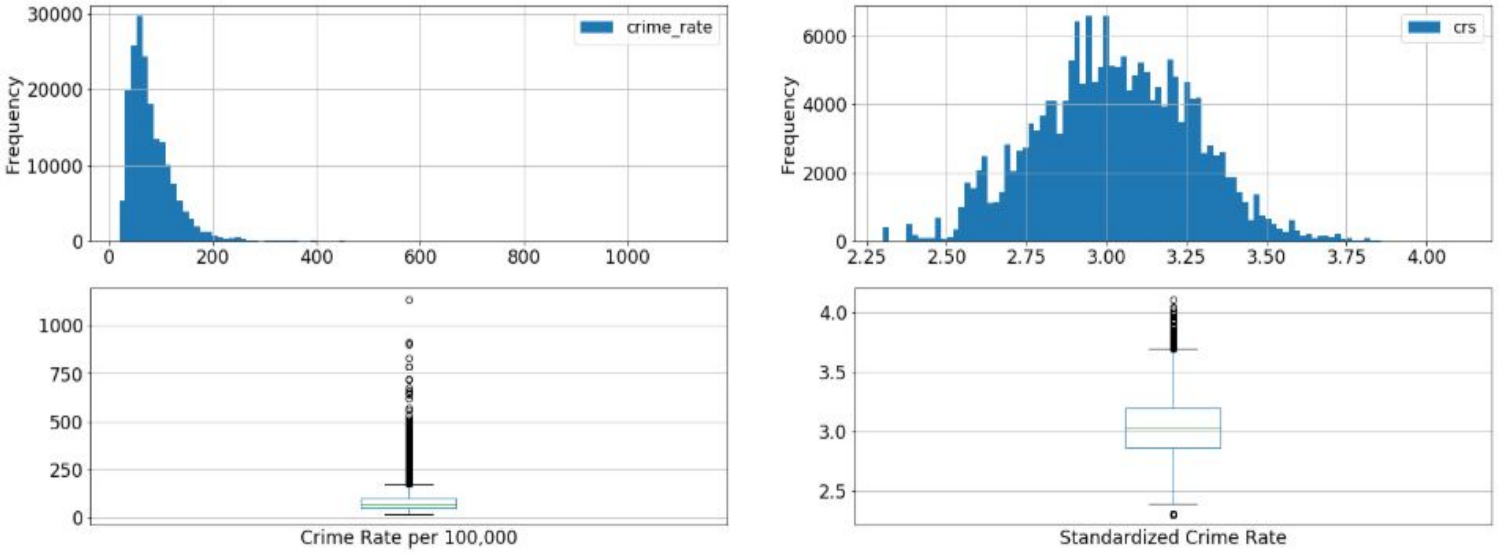
Table 3: Time of Day Categories

Time Category	Time
Midnight	11pm - 2am
Late Night	2am - 5am
Early Morning	5am - 8am
Morning	8am - 11am
Afternoon	11am - 2pm
Mid Afternoon	2pm - 5pm
Evening	5pm - 8pm
Night	8pm - 11pm

Third, the team merged aggregated crime and weather data with the Census data by block group. Crime rates were calculated as the number of crimes reported per 100,000 people by dividing the total count of crime at the block group level by the block group population and multiplying the results by 100,000. To prove the hypothesis that the labeled crime rates could be predicted, the team needed to classify the resulting crime rates into one of five ranges, low, medium-low, medium, medium-high, and high.

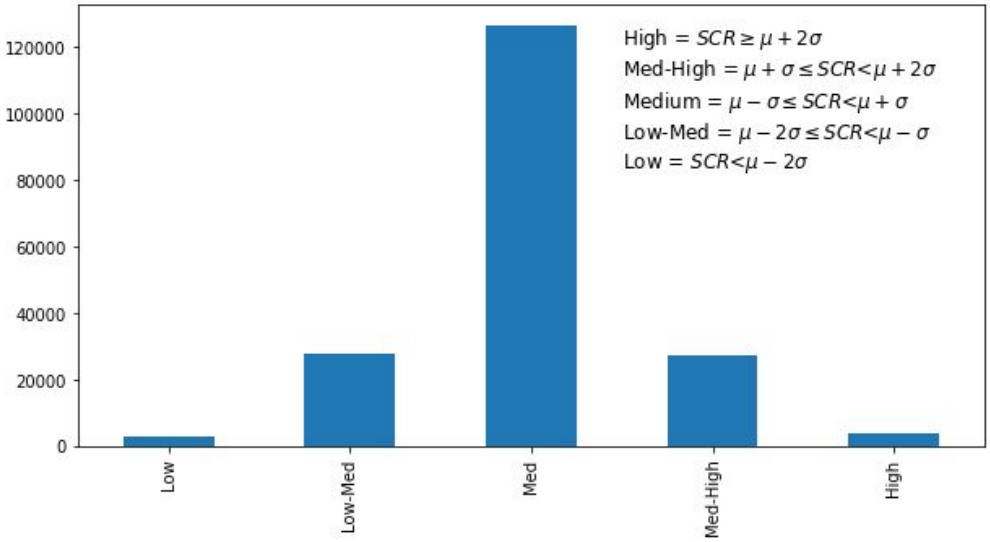
The left plot on Figure 6 below shows the distribution of crime rates as a boxplot and histogram before any adjustments were made. The plot reveals the asymmetric distribution of crime rates, highly clustered toward zero with a long tail. From this asymmetric distribution, team found it difficult to classify this distribution in a meaningful way, as it is not clear how to define crime rates as “high” or “low”. In order to classify crime rates, the team used a power transformer to convert the data into a more familiar Gaussian distribution, with the rescaled values displayed on the right plot of Figure 5.

Figure 5: Crime Rates per 100,000, Unadjusted (Left), Rescaled (Right)



Using the rescaled values, the team classified crime rates based their distance of the mean, measured in multiples of distributions standard deviation. Low and high labels are defined as below and above two standard deviations away from the mean, respectively. Low-Medium and Medium-High. Medium labels were assigned to values within one standard deviation of the mean, which following a normal distribution will contain 68% of the crime values.

Figure 6: Crime Rate Classifications



Lastly, rows of crime and weather data were dropped where Census block group population data was missing, which totaled 101,000 observations for the years 2008, 2010, 2011, and 2012. An alternative approach to missing Census data could have been imputation by regression analysis, but the team decided against it. A final table describing key elements of the 26 features in the data and the total number of instances after wrangling are below.

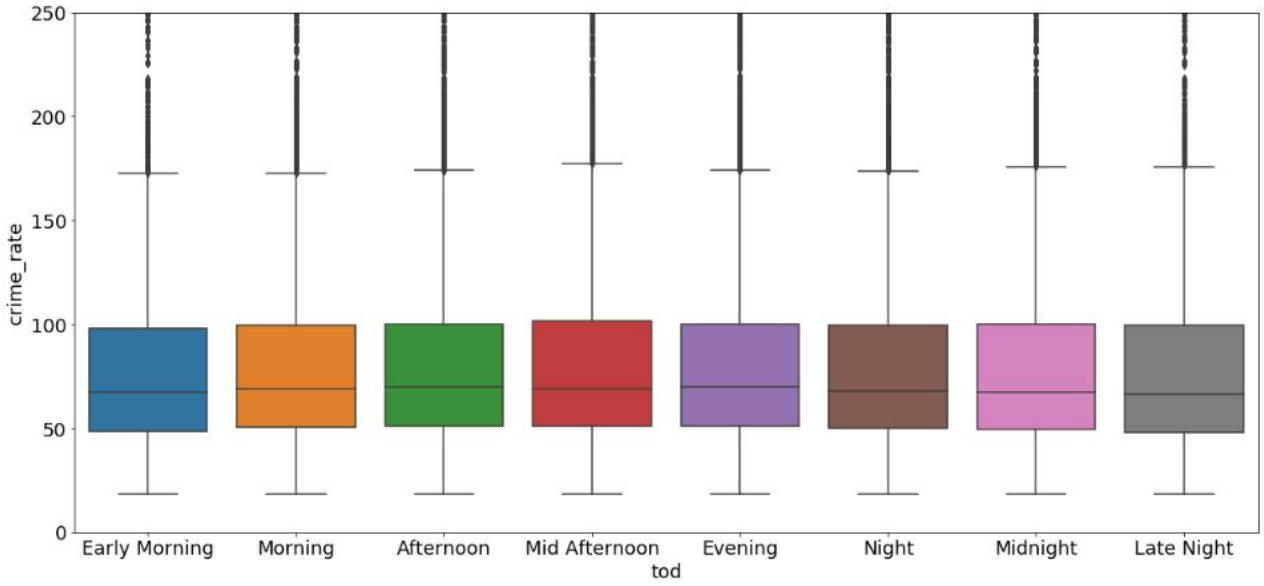
Table 4: Description of Key Numeric Variables, Final Dataset

	Crime Rate	Temperature	Precipitation Prob	Precipitation Intensity	UV Index	Per Capita Income	Median Household Income	Num Housing Units
count	188,224.00	188,224.00	188,224.00	188,224.00	188,224.00	188,224.00	188,224.00	188,224.00
mean	80.65	59.90	0.06	0.00	1.60	48,633.95	77,993.84	787.20
std	46.18	17.59	0.20	0.01	2.30	30,614.14	44,770.54	402.79
min	18.49	3.01	0.00	0.00	0.00	3,787.00	6,548.00	10.00
25%	50.35	46.21	0.00	0.00	0.00	23,626.00	40,699.00	512.00
50%	68.68	62.04	0.00	0.00	0.00	42,186.00	72,467.00	692.00
75%	100.00	74.04	0.00	0.00	3.00	66,392.00	104,950.00	987.00
max	1,136.36	98.60	1.00	0.40	11.00	220,398.00	250,001.00	3,325.00

V. Exploratory Data Analysis

After wrangling our data, our data set consisted of 179,382 instances and 26 features. Each instance is a crime rate for a time of day in a Census Block Group. As mentioned above, the team aggregated crime data by counting reported crime, calculating a crime rate, and rescaling the crime rate to easily classify crime rates with low, medium, and high labels. An assumption made when grouping crime into time-of-day categories was that crime rates were higher in the evening and night than in other parts of the day and these differences would be key features in the model. Figure 7 shows boxplots of violent and non-violent crime rates for each time-of-day category. The crime rates within each time-of-day have long tails, suggesting that the distributions of crime rates for all times of day are skewed and will need to be standardized before modeling. In order to make interpretations about the median, the y-axis was adjusted. Looking at the median, crime rates do increase throughout the day, but the difference in the magnitudes is slight. The time of day with the lowest crime rate is Late Night, between 2 am and 5 am, with about 66.5 crimes per 100,000 people, and that with the greatest median is the Evening between 5 pm and 8 pm, with about 70 crimes per 100,000 people, only a 5% increase.

Figure 7: Crime Rates by Time-of-Day

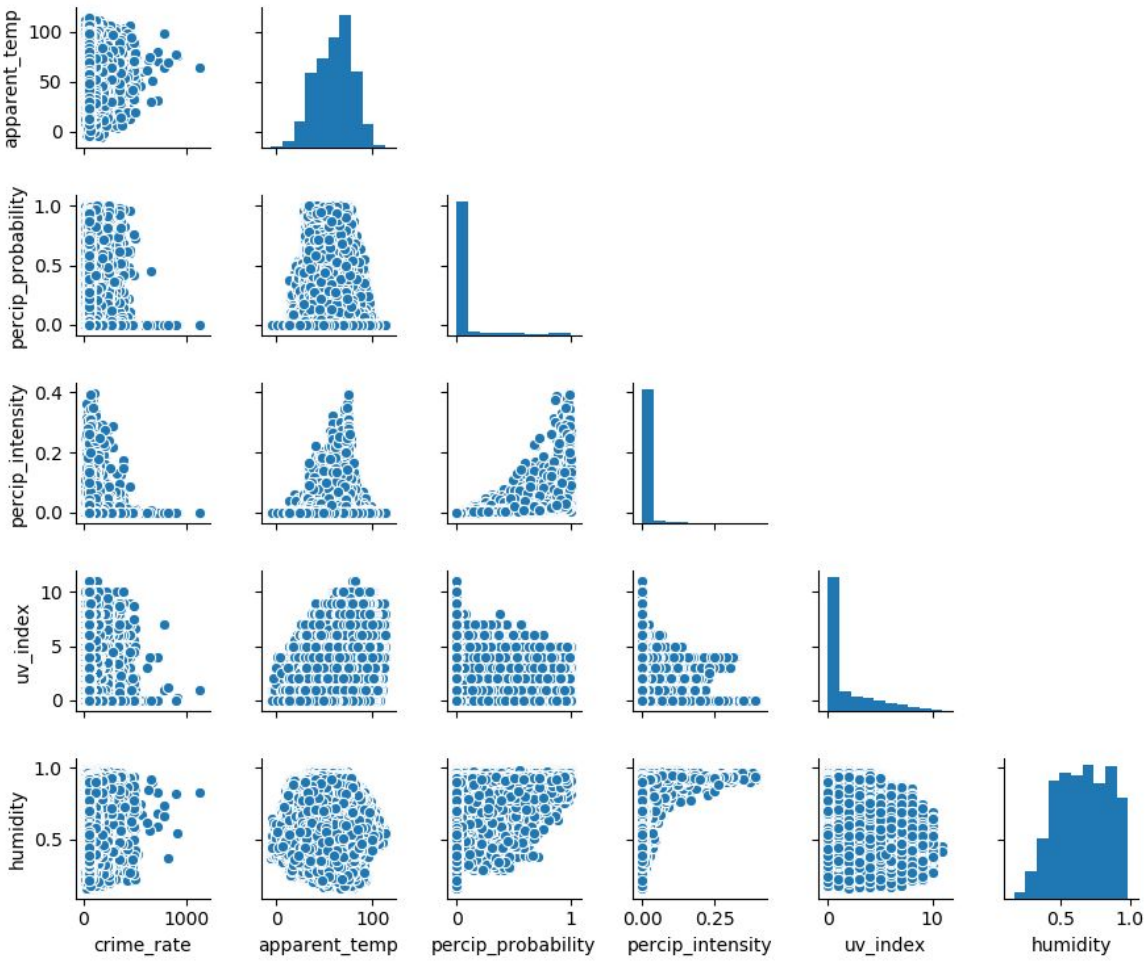


In addition to time of day, weather data was collected in association with the crime data, as research may suggest higher temperatures explain higher property and violent crime rates.[12] The Spearman correlation coefficient between crime rates and temperature, precipitation probability and precipitation intensity are statistically insignificant, suggesting that we cannot reject the null hypothesis that the three weather features are uncorrelated with our target, crime rates.

While low crime rates are in fact associated with lower temperatures, they are also associated with higher temperatures and after about 35 degrees fahrenheit, there appears to be no significant relationship at all. Associations between crime rates and precipitation intensity and probability are suggested visually in Figure 8, but as mentioned above, Spearman correlation tests suggest otherwise. Lastly, weather variables show strong relationships with one another, such as the relationship between probability of precipitation, the intensity of precipitation, humidity, and UV index. Some of these relationships were not mentioned in other literature on crime predictions and could be new avenues for predicting crime rates. When modeling, the team will need to be careful not to include multiple weather features that are associated with one another.

[12] Simon Field, The British Journal of Criminology, Volume 32, Issue 3, Summer 1992, Pages 340–351, https://doi.org/10.1093/oxfordjournals.bjc.a048222

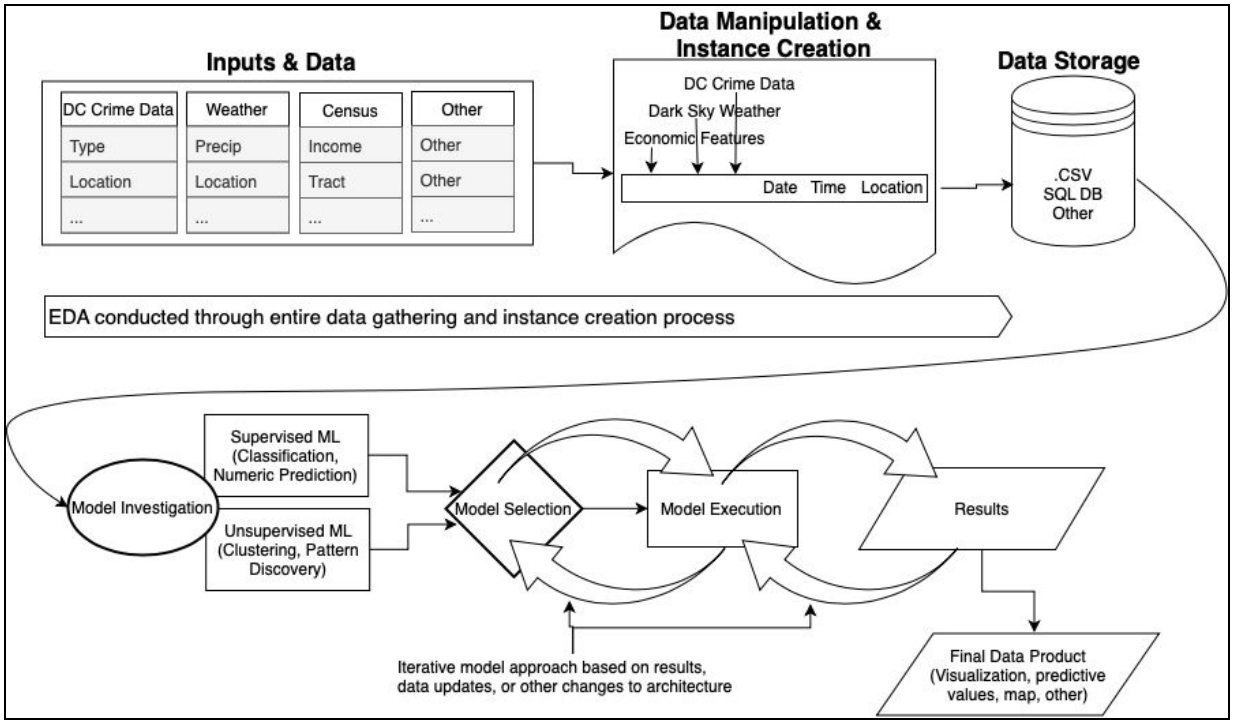
Figure 8: Crime versus Weather Pairplot



VI. Feature Selection and Modeling

Figure 9 below describes the team's approach to the data product architecture. A majority of the project duration was spent on data ingestion, manipulation, and wrangling. Moreover, the team's decision to pursue crime rate classification prediction resulted from preliminary data wrangling efforts.

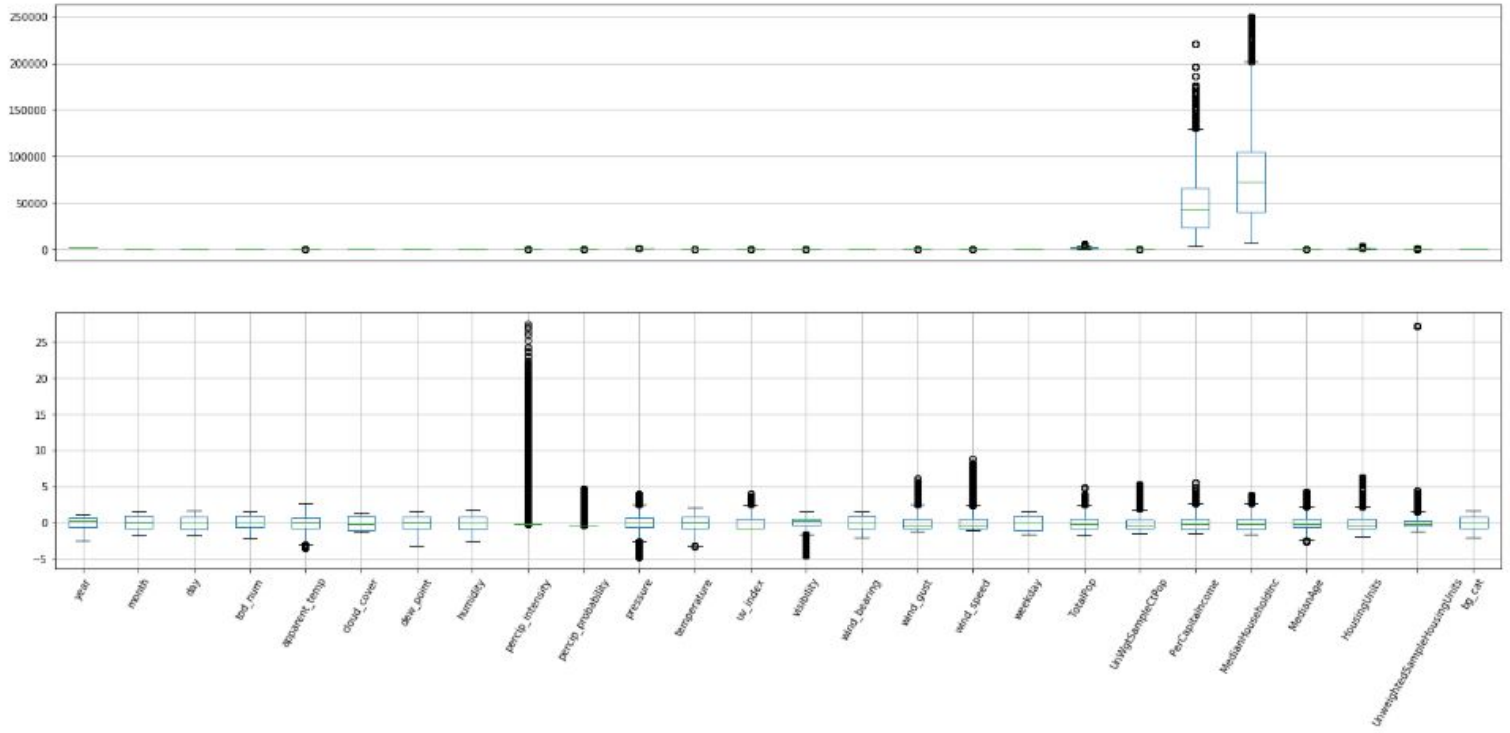
Figure 9: Architecture Diagram



Feature Selection

All 26 features are plotted on the top boxplot of Figure 10 below. There is significant variation, but it is dominated by the two income variables, median income and per capita income. In order to reduce variability, the team review the Standard Scaler, Robust, and Min-Max standardization functions available in Pythons scikit-learn library, finally choosing the standard scaler. The results of the standard scaler are displayed on the bottom figure of Figure 10.

Figure 10: Before Standardization (Top), After Standardization (Bottom)

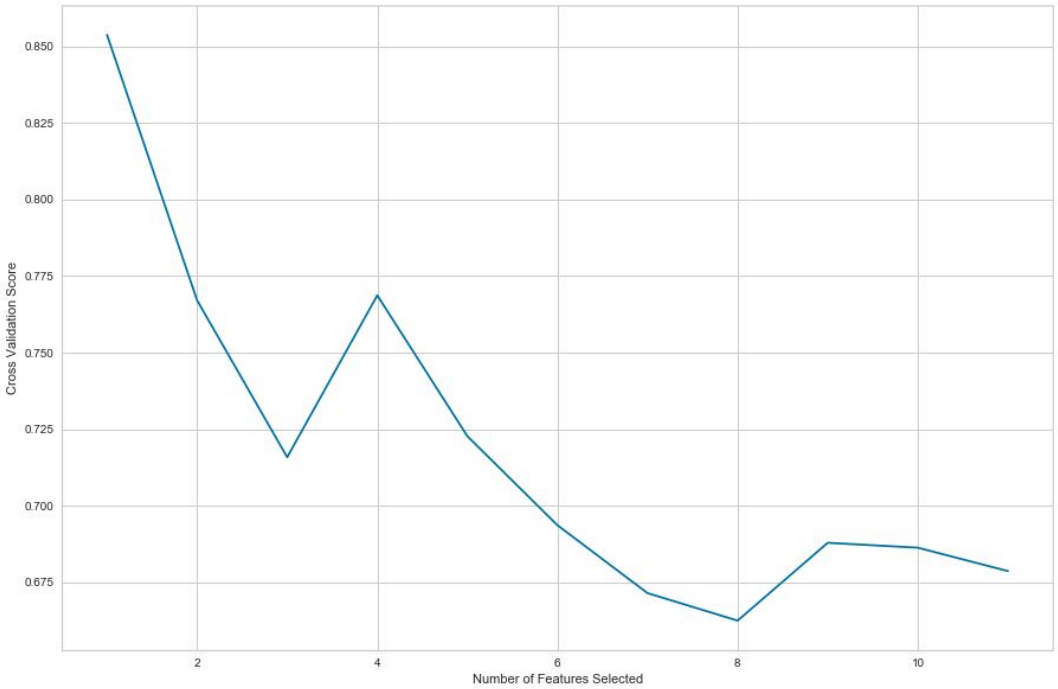


Several approaches were leveraged to identify the subset of features to be used in modeling. Starting with all 26 features, cross validation scores were suggesting the optimal number of features to be one. The team hypothesized that Census data was leaking information to the model, which was evident for features relating to population in a block group, such as housing units and other population estimates, as population was used to calculate the target. That is, there was a relationship between our target, which was calculated using Census block group population in the denominator and the other demographic data associated with the block group that resulted in model accuracy scores of 97%.

In addition to population variables, demographic and economic variables produced similar results. Specifically, a model provided the median age, per capita income, and block group could predict with high accuracy overall crime rate classifications. Upon review of the results, the team decided that in order to reduce leakage, but maintain user-friendliness of the data product, the final model would exclude median age, per capita income, and median income associated with the block group, but the block group would remain as a feature.

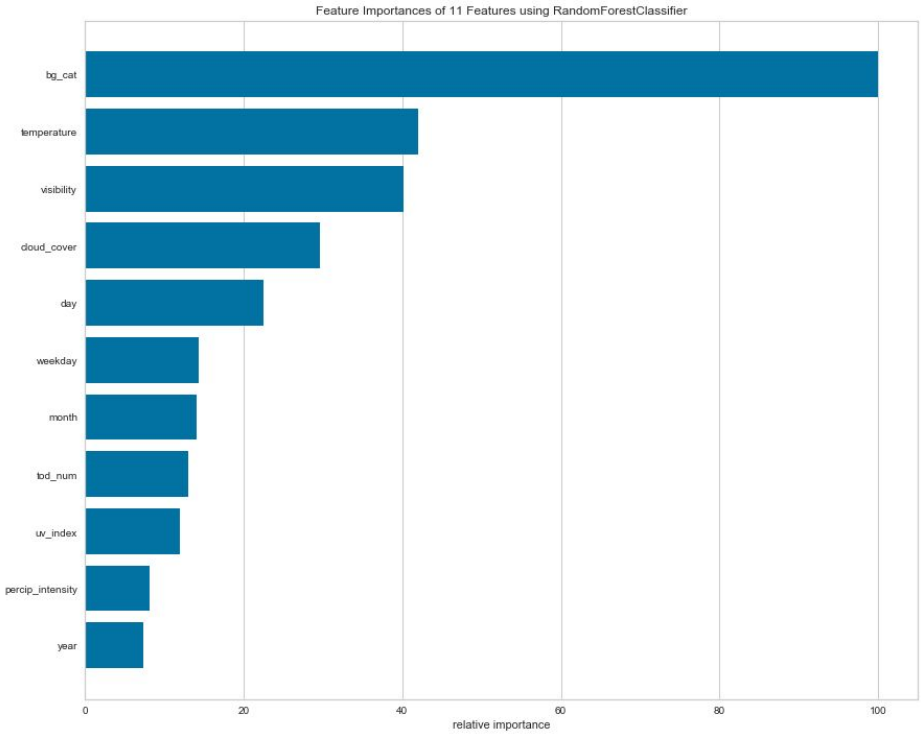
Figure 11 below shows the cross validation score of a Random Forest Classifier on all 11 features that remained after initial review of cross validation and modeling. Figure 11 indicates that the higher number of features produced worse results. Other models performed similarly and all models suggested between 3 and 5 features.

Figure 11: Cross Validation Score and Feature Optimization (Random Forest Classifier)



Another approach used by the team was the feature importance visualization tool provided by the yellowbrick python package. Figure 12 shows one example of feature importance visualization used in feature downselection. This particular example used a Random Forest model. These visualizations helped the team identify additional features to remove. For example, features like *Year* and *Month* continuously indicated low contribution to the model performance relative to the other features. Additionally, features with temporal characteristics tended to show high relative importance. Ultimately, there were similarities across many of these visualizations with respect to the features of least importance. Thus, the team leveraged these plots to eliminate features such as month, pressure, dew point, etc.

Figure 12: Feature Importance



The team narrowed down the model to 6 features, block group, the day of the week, day of the month, time of day, UV index, average temperature. Block group was expected to stand out as a feature, as did the other block-group level economic and demographic variables. The team chose to keep block group since (1) it is an important factor for predicting local reported crime rates, and (2) it is a key feature assumed in the data product, which suggests safest transit option based on predicted crime rates in a user-provided location.

Models Considered

As stated in Section IV above, crime rate was calculated per block group, and grouped by time of day. Subsequently, crime rate was categorized into five groups - low, low/med, med, med/high, and high. Thus, the team focused model investigation on the categorical target. Models explored include K-Nearest Neighbors, Random Forest Classifier, and Bagging Classifier with the Decision Tree Classifier estimator. The team implemented the models using the scikit-learn and YellowBrick libraries in Python. To support the classification of about 180,000 instances, the data was split such that 80 percent of the data represented the training dataset, and 20 percent represent the test dataset. In none of the models was resampling performed to adjust for the class imbalance, resulting in the low f1 scores seen in the high and low classes for even the best performing models.

The K-Nearest Neighbors classifier was implemented with the default parameters and the classification report is listed in the appendix as Figure A1. The poor performance of the model

discouraged the team from exploring it further and no tuning of the parameter settings, such as weights or leaf size were pursued. Next, random forest and decision tree classifications were explored, resulting in the most promising models for predicting crimes rates, despite the lower f1 score for rare crime rate classes. The classification reports for both models are in the appendix as Figure A3 and Figure A4, respectively. The most promising model was discovered using bagging classification with the decision tree classifier.

Bagging classification uses bootstrapping methods to sample the training dataset, hoping that by aggregating predictions made by classifiers on the various sub-sampled training data, the final prediction would be better than to classify the entire training dataset at once. Conceptually, the strength is that by sampling with replacement, sub-samples of the test data will be representative, suggesting that an average of multiple subsets could increase predictions of the model overall. The team utilized bagging classification in the scikit learn Python library, implementing bagging with the decision tree estimator, the default estimator. In addition, the team used the default equal probability with respect to selecting training data.

Modeling Results

The final model chosen was the Bagging Classifier using the Decision Tree estimator. The results for this model are described in this section, whereas results for the other models considered can be found in the appendix. The Bagging classifier produced the highest accuracy, precision, recall, and f1 scores. Overall accuracy was 82 percent. Figure 13 below shows the classification report for the chosen model. Precision, recall, and f1 scores are shown for each crime rate category. One observation by the team was that the high crime rate category did not perform as desired. Ultimately, the desire is to be able to accurately predict when a specific area may have a higher chance of reported incidents. The medium category was modeled most accurately. Medium-High, and Low-Medium categories were predicted favorably, with f1 scores of 0.65, and 0.76 - respectively. Given the nature of the target, it is more important to have better recall in model prediction. In other words, it’s more important to eliminate false negatives, as opposed to eliminating false positives. The classification report shows that precision and recall are relatively similar across each crime rate category

Figure 13: Bagging Classification Report

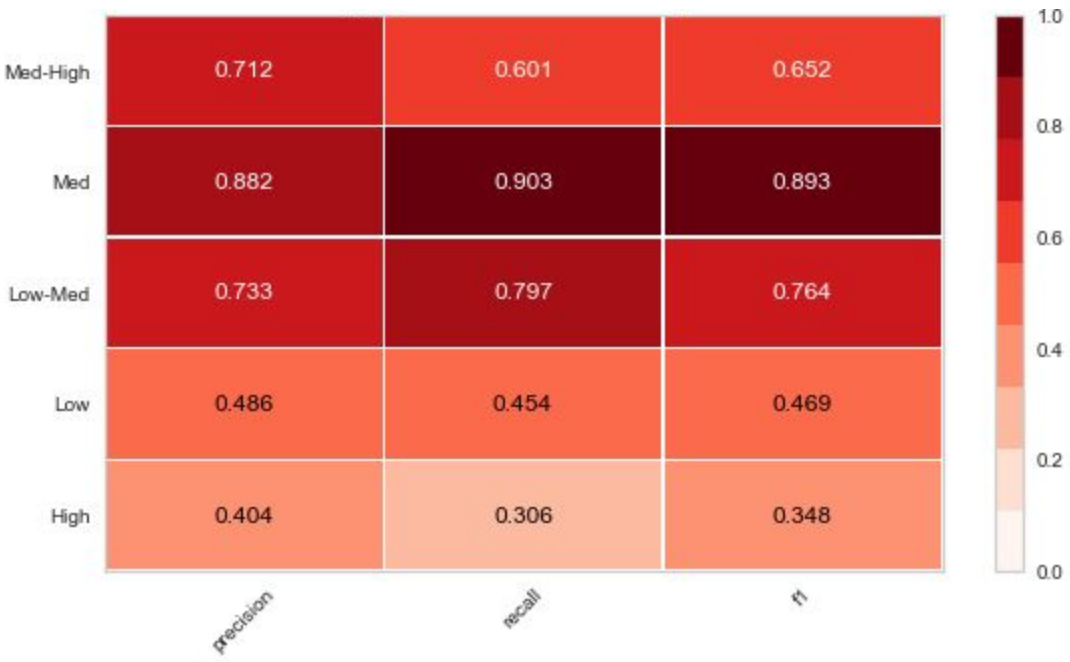
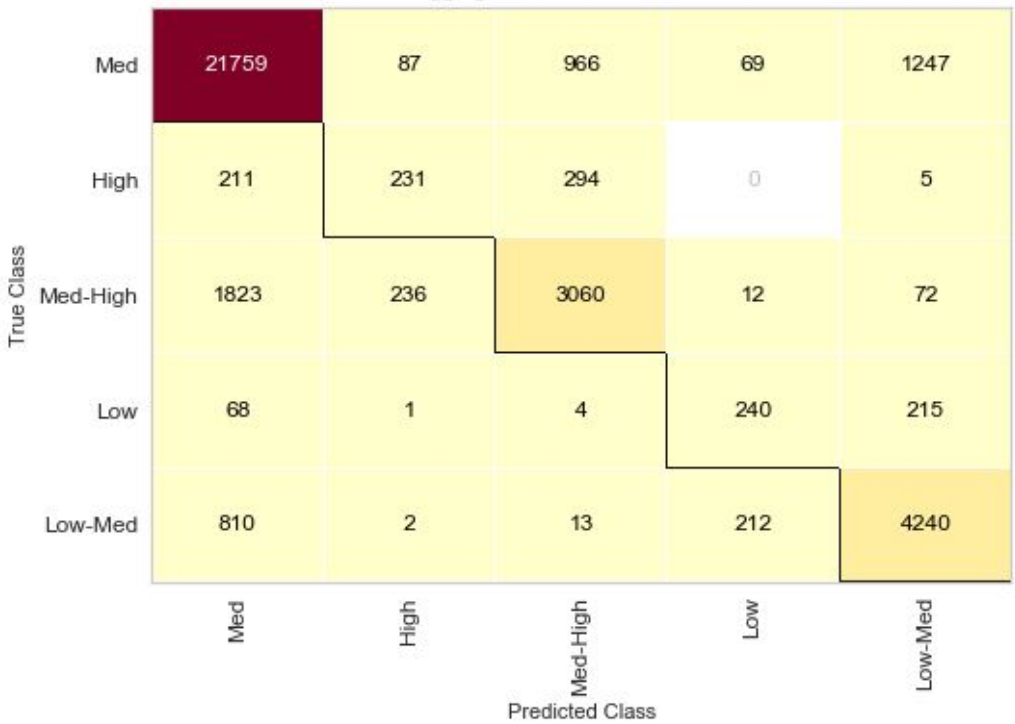


Figure 14 below shows the Bagging classifier confusion matrix. This graphic describes which categories of crime rate were correctly predicted. Moreover, for the model predictions that were incorrect, the graphic shows where the model predicted incorrectly.

Figure 14: Bagging Classifier Confusion Matrix



The confusion matrix shows that the Medium crime rate category performed the best. Conversely, low and high crime rates performed poorly. Visual inspection of the matrix shows that the incorrect predictions by the model were often placed in categories adjacent to the true class. For example, the model never predicted a high crime rate category into a low crime rate category. Moreover, the low crime rate category was incorrectly predicted into the high category one time. The confusion matrix also indicates imbalance within the target categories. Improving target category imbalance by stratification or undersampling would be a logical next step to improving model performance.

Model Output and Product

In order to make the crime prediction model available to users seeking transit recommendations, the group created an interactive Jupyter notebook. The notebook accepts user input for the address, date, and time of departure. Functions in the notebook use the inputted address to call the Census Geocoder API via the python census geocode wrapper to return the census tract and block group information of the address. Date and time data are reformatted in separate ways: one to input into a DarkSky API call to return current or forecasted weather data, and one to input the day and time of day into the model.

In order to use the data product to classify a crime rate and make a transit recommendation on a new data point, the new point must reflect the same encoding and standardization as the features included in the model. The team output the encoding, standardization, and model parameters after the final model was selected and imported them into the product notebook. The data product ingests these details and converts user input data to represent values expected by the model features. For example, when a user inputs their location, the data product converts it to the Census block group and subsequently encodes the values for input into the final model. If a user tried to put in a block group before encoding, or encoded the block group value in a way that was different than was used during model selection, the data product would either return an error or return a classified crime rate for the wrong area.

Once data is returned from the APIs used, it is encoded and standardized along with other model features, and the resulting dataframe is put into the model to generate a predicted crime rate category. If the category is predicted as “High”, the product returns a recommendation that the user take a rideshare or taxi. If the category predicted is any category below “high”, the product returns a list of Metro rail stations, bus stops, and Capital Bike Share station in the census block group of the input address.

VII. Future Research and Lessons Learned

Further iterations of this project could incorporate additional data points and data sources. For example, DC MPD updates the public on reported crime daily. Wrangling operations could be updated to consistently ingest crimes reported to inform the classifier model. Weather data associated with the time of day for newly reported crimes could be ingested.

Additional data sources that could assist in predicting crime were considered, but the team did not have the time or capacity to incorporate them into the model or the final data product. Additional data sources include a walkability index for neighborhoods, dates for local sporting events, marches, protests, and distance from employment, commercial, or residential centers. We are not aware of a comprehensive dataset that exists that would be appropriate to incorporate at this time. The incorporation of walkability data and large event data would, theoretically, identify geographies and times of high foot traffic that could potentially affect crime rate and explain outliers.

Future development on this project could also improve how transit recommendations are made. In its current form, the program returns transit options in the Census block group of the input address if the crime rate at the desired date and time of travel is predicted to be low, medium-low, or medium. To improve this utility, the recommendation system could take into account predicted crime rates of the users destination, either to reroute them to the nearest, safest location, or take it into account when suggesting transportations. In addition, the utility could be improved by predicting crime rates in adjacent Census block groups, and either return more transit options or alternatively return transit options in the “safest” nearby Census block.

The decennial US Census will be conducted in 2020. It would be interesting to incorporate new Census data when it is available to see how the updated data affects the model moving forward. While American Community Survey is useful, recently collected data from the 2020 Census will be the most up-to-date and accurate data available, and should be incorporated into future iterations.

VIII. Appendix

The final model chosen was the Bagging Classifier using the Decision Tree estimator. The figures below show the results for other models that were evaluated.

Figure A1: KNeighbors Classification Report

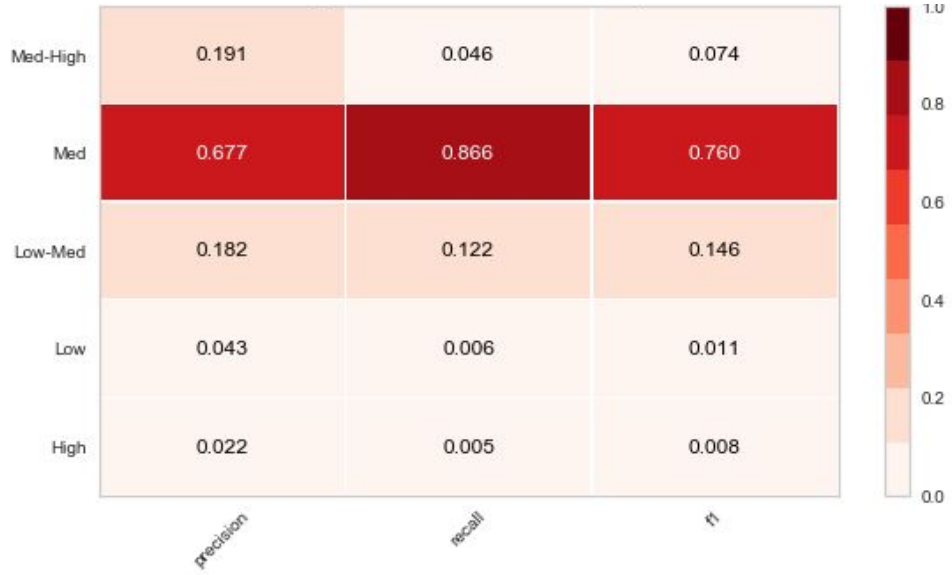


Figure A2: Extra Trees Classification Report

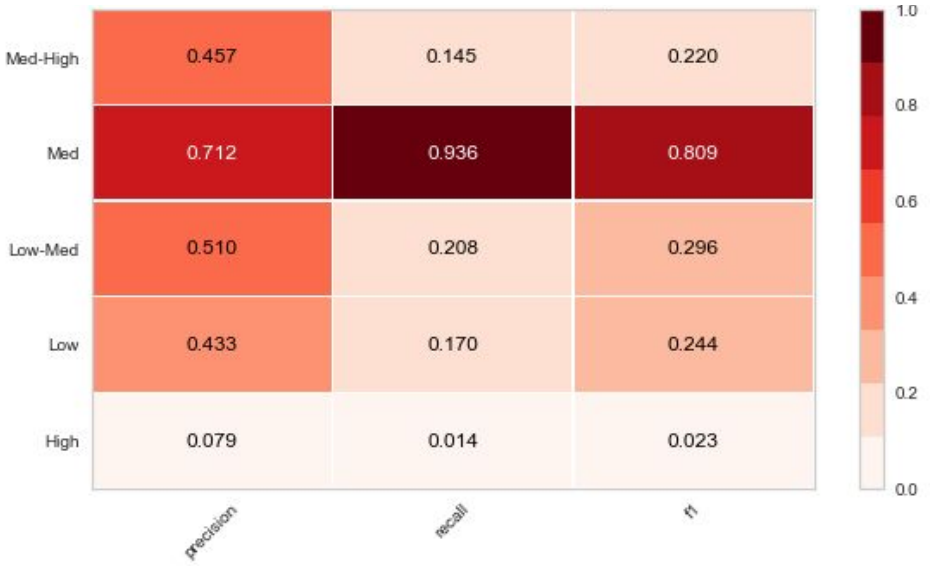


Figure A3: Random Forest Classification Report

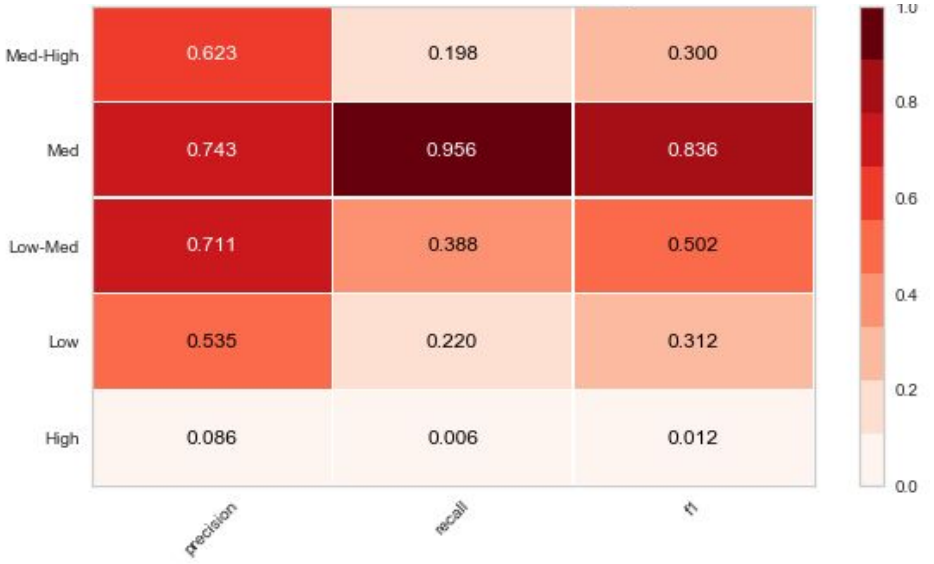


Figure A4: Decision Tree Classification Report