Hudson Finch-Batista

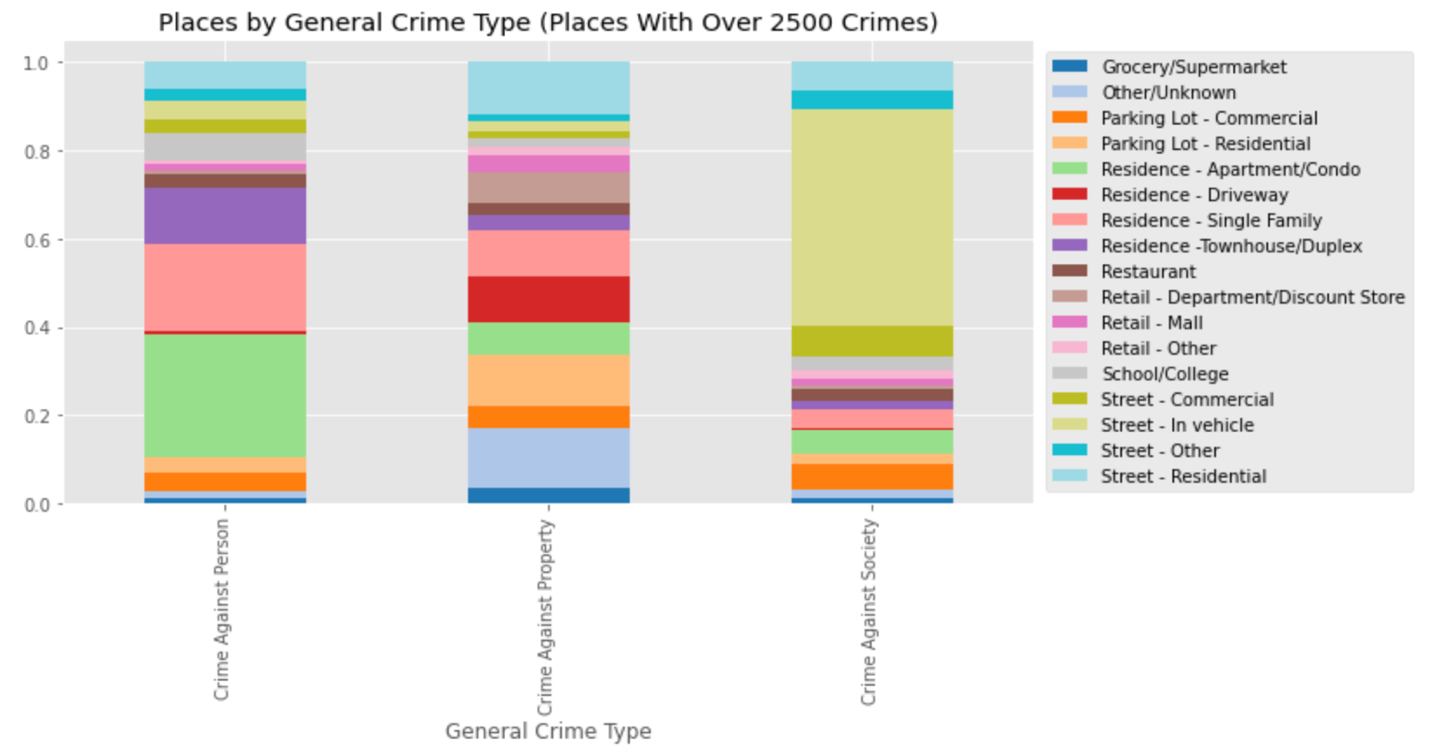
Data 205

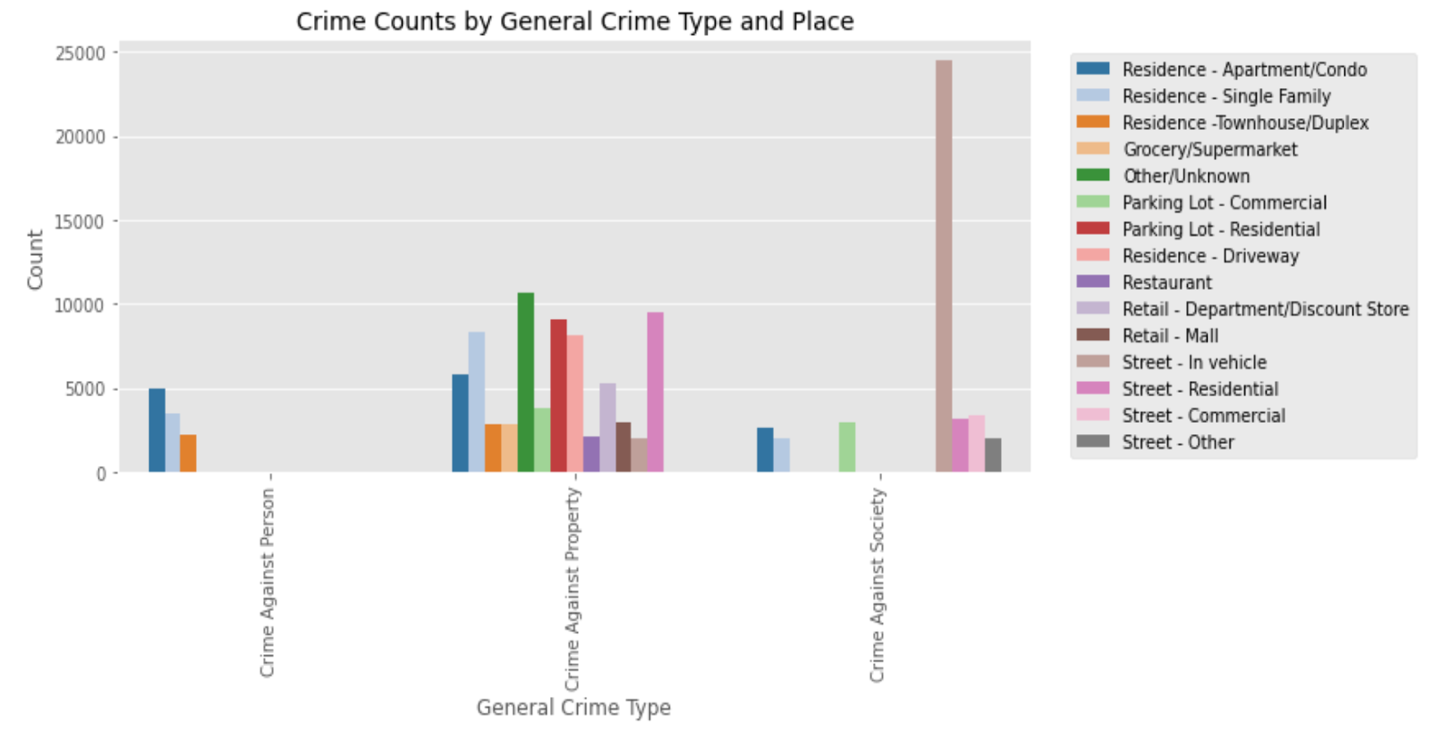
Predicting Crime in Montgomery County, MD

The following analysis of crime in Montgomery County, MD is based on a dataset from Data Montgomery entitled, “Crimes”, as well as data obtained from the U.S. Census Bureau and the Federal Communications Commission (FCC). While the “Crimes” dataset was downloaded as a comma-separated values (CSV) file directly from the Data Montgomery data catalog, the FCC data was obtained through the FCC Application Programming Interface (API) and the Census Bureau data was obtained through the Census API. An API is essentially a software interface that allows a client to connect to, request, and obtain data from a server in a structured format. The Montgomery County Census Bureau data was collected from the 2019 American Community Survey (ACS), which contained a plethora of demographic information at various geographical levels across the United States from 2015 to 2019. For this analysis, since the objective was to collect data that was as detailed as possible, data was collected at the census block level, which is the smallest geographical unit for which census data is collected. The FCC API, on the other hand, provides the Federal Information Processing Standard (FIPS) code, or block ID, corresponding to a given latitude, longitude coordinate given as an input. Since the FIPS code was not included in the “Crimes” dataset, it had to be derived from the latitude and longitude coordinates of each crime in the “Crimes” dataset via the FCC API. The census data for each block was then mapped to each crime that occurred by matching each FIPS code, or block ID, from the FCC API with the FIPS codes in the census data.

The “Crimes” dataset contains approximately 240,000 crimes that took place in Montgomery County from 2015 to the present. The primary variables of interest from the “Crimes” dataset were the type of crime, the number of victims, the type of locations where crimes occurred (e.g. residential street, grocery store, commercial parking lot, etc.), the city where crimes occurred, and the police districts where crimes occurred. The dataset, moreover, contains three different variables for classifying crimes, which follow the National Incident-Based Reporting System (NIBRS) nomenclature. The primary variables of interest from the 2019 ACS were the median income per block, the population per block, the median age per block, the number of unemployed individuals per block, the number of individuals by highest educational attainment per block, the number of individuals by race per block, the number of individuals by gender per block, and the number of individuals who use public transportation by block. Since Montgomery County has 615 blocks, this amounted to 615 possible values for each of the aforementioned variables.

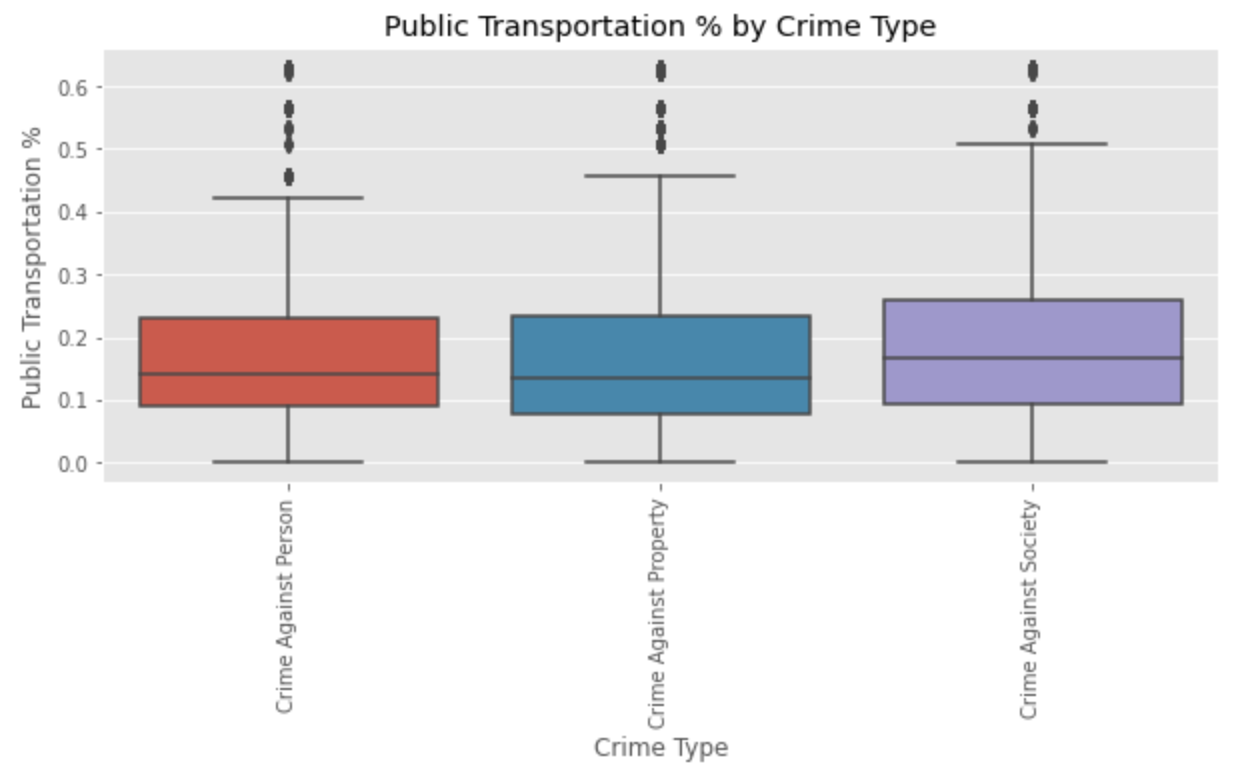
I began my exploratory data analysis (EDA) by identifying the number of crimes by the most general crime type variable. Over half of all crimes were classified as “Crimes Against Property” crimes while just below 10% of all crimes were classified as “Crimes Against Person” crimes. As for cities, while the vast majority of crimes occurred in Silver Spring, many crimes also occurred in Gaithersburg, Rockville, Germantown and Bethesda. The remaining cities accounted for a small number of crimes. With respect to place, while more crimes occurred in vehicles than any other specific place, the majority of crimes occurred in different types of residential places. A large number of crimes were not classifiable by place, but this seemed acceptable given that certain crimes might not necessarily be categorizable by place (e.g. identity theft). Moreover, since one would intuitively expect certain crimes to be associated with certain places, such as DUI’s with crimes in vehicles, this prompted me to look at how each crime type variable varied by place. To do so, I limited places to only those with 2,000 or more crimes. As can be seen from the stacked bar chart below, property crimes, though primarily occurring in residential streets, residential parking lots, and unclassified places, were approximately uniformly distributed across places, somewhat in contrast to the other crime types for which certain places comprised the majority of crimes. For instance, almost half of all crimes against society occurred in vehicles and crimes against people occurred predominantly in apartments, townhouses, and single-family residences.

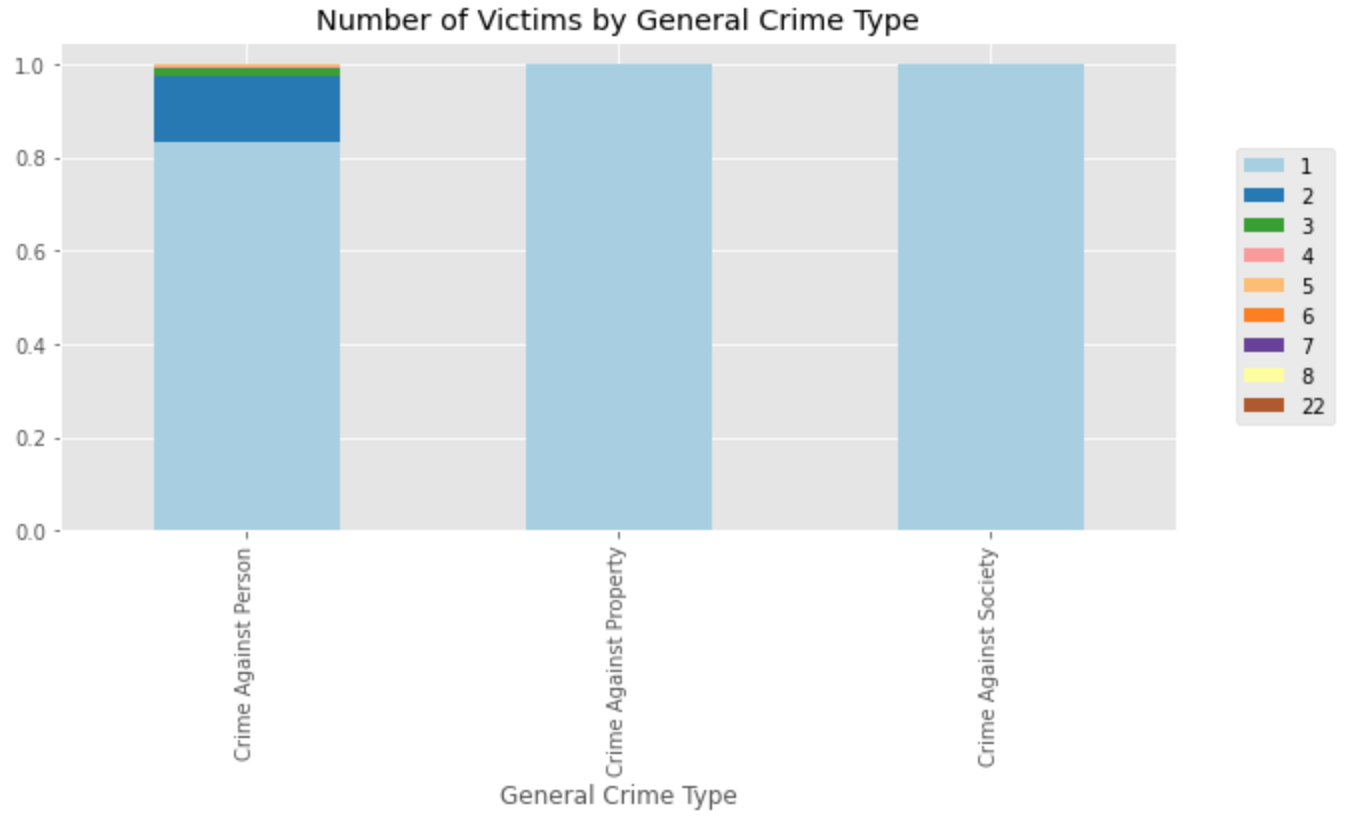




However, while the proportions of places that composed each general crime type were quite distinct, because property crimes accounted for over half of all crimes, it was necessary to observe the total number of crimes by place for each crime type as well. For instance, although a smaller proportion of property crimes took place in apartments and single-family homes compared to crimes against people, more property crimes took place in apartments and single-family residences than crimes against people as can be observed from the bar chart above.

The second section of the EDA was based on analyzing numerical variables by crime type, beginning with the ACS variables. While the crime categories generally did not vary by any of the ACS variables, they did vary by the percentage of individuals that use public transportation to some extent.





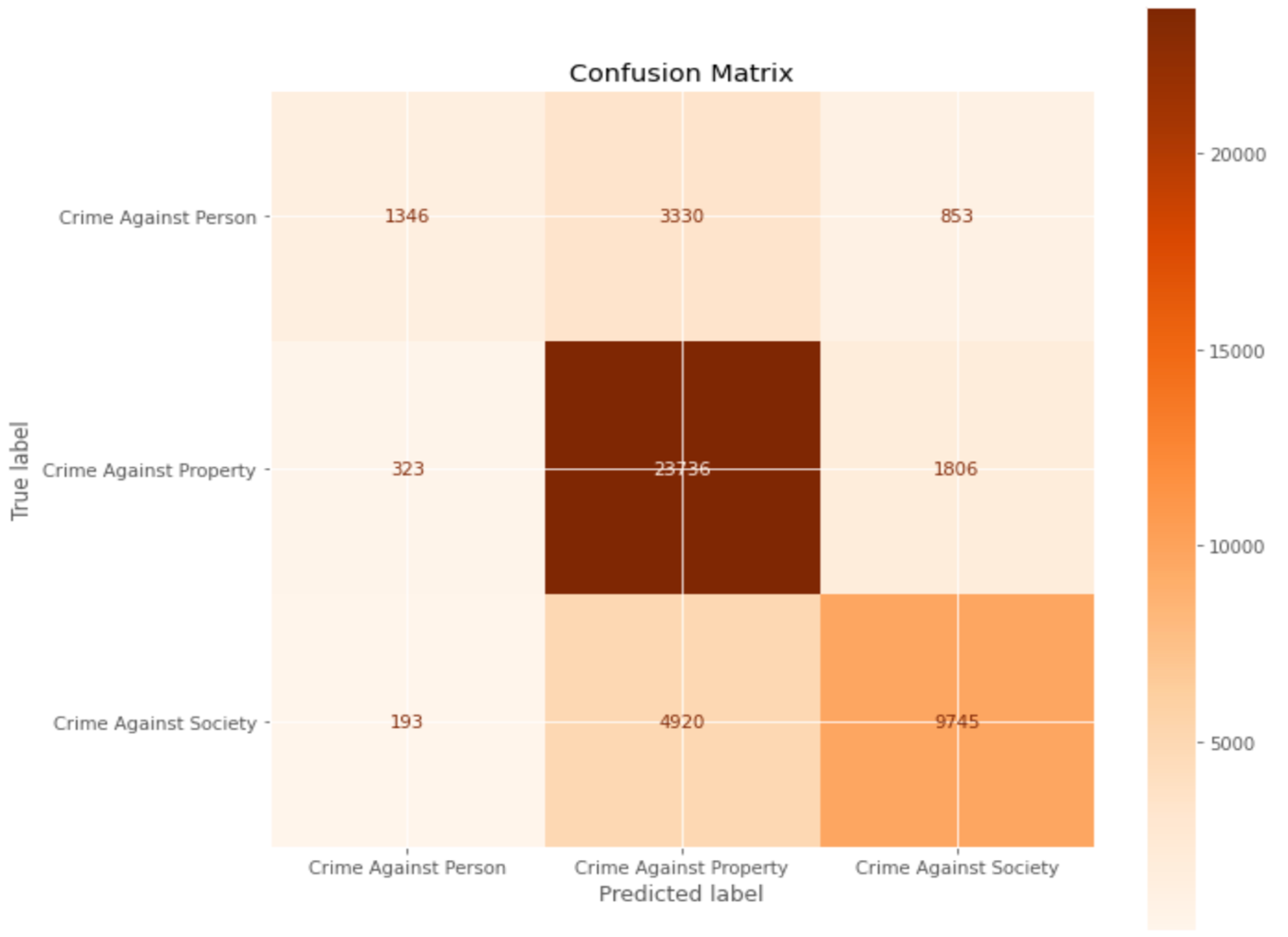
While the distribution of the percentage of individuals that use public transportation was almost identical across crimes against people and property crimes, it seemed to be noticeably different for crimes against society. That is, the distribution of crimes against society is noticeably shifted upward, indicating that crimes against society tended to occur in census blocks with a slightly higher public transportation percentage compared to census blocks where the other two crime types occurred.

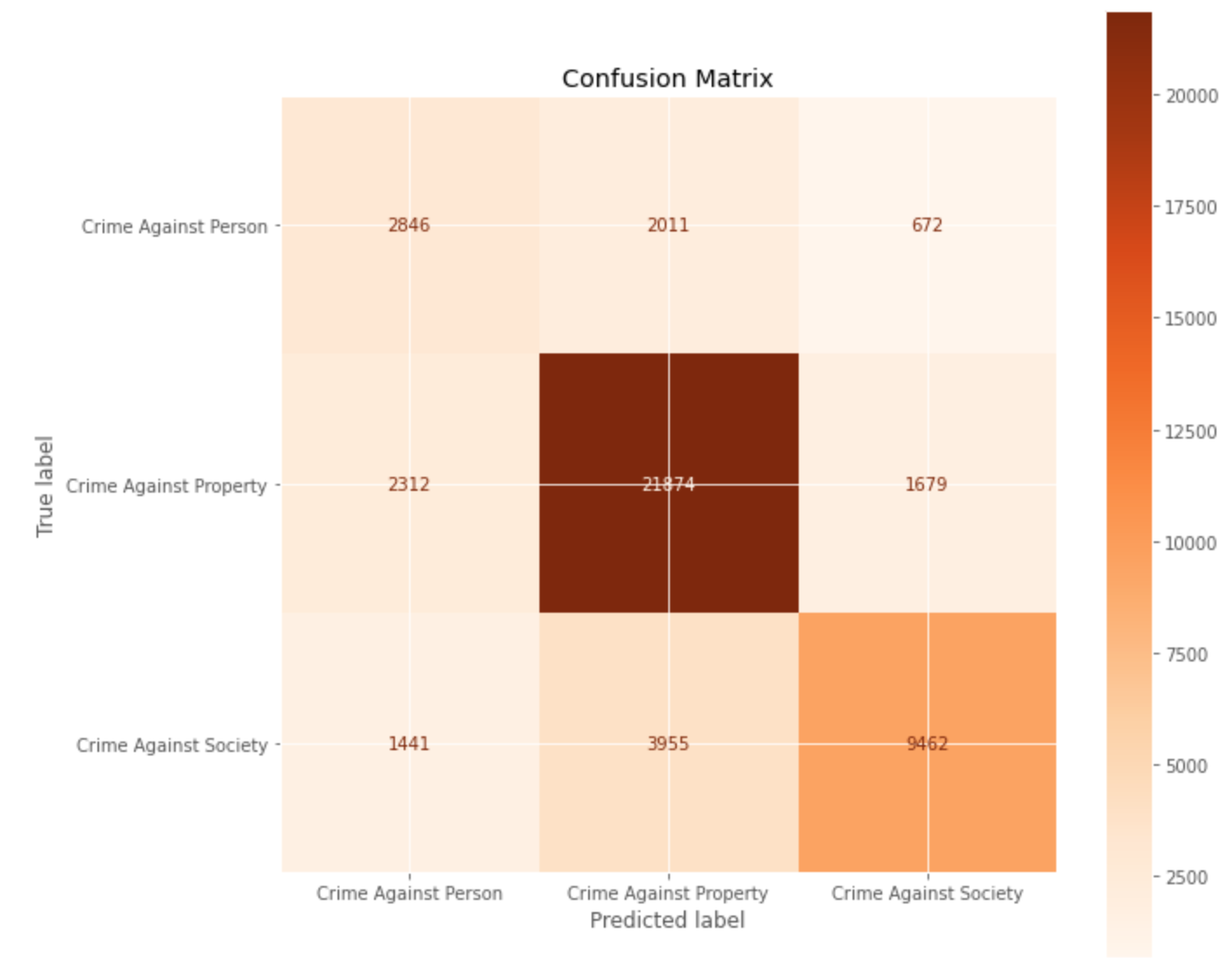
Unexpectedly, one of the primary numerical variables by which the general crime types varied was the number of victims’ variable. While the overwhelming majority of crimes had only one victim, a non-negligible fraction had more than one victim. This can be seen in the stacked bar chart above, which shows that the only general crime category that had any proportion of crimes for which the number of victims was greater than one was the crimes against person category. Although crimes with more than one victim accounted for less than 20% of all crimes against people, this could nonetheless be significant in differentiating crimes against people and property crimes, which had similar amounts of crimes that occurred in single-family homes and apartments.

Since my dataset consisted mostly of categorical variables, I decided to first address a classification problem. Specifically, I wanted to determine if I could model the general crime types as a function of place, victims, and ACS variables. However, since more property crimes occurred in apartments and single family-residences than crimes against people despite the fact that a greater proportion of crimes against people occurred in apartments and single-family residences, I thought that it would make sense to examine the effect of balancing the number of crimes against people and property crimes. In effect, I hypothesized that if the number of crimes against people that occurred in apartments and single-family homes outweighed the number of property crimes that occurred in apartments and single-family homes and, therefore, reflected the proportion of crimes against people that occurred in apartments and single-family homes, the model would be would be more adept at distinguishing property crimes and crimes against people. In order to test this hypothesis, I created an additional dataset in which the number of property crimes and crimes against people were equal (i.e. balanced) and trained one set of models on the balanced data and another set of models on the imbalanced (i.e. original) data. To synthesize an equal number of crimes against people and property crimes, I used a technique known as minority class oversampling, which involved repeatedly resampling random crimes against people until there were as many crimes against people as there were property crimes.

Additionally, in order to determine the absolute best model of the general crime categories and measure the effect of hyperparameter optimization on model performance, I decided to create a set of hyperparameter-optimized random forest and gradient boosted machine models for both the imbalanced and balanced data, the results of which were compared with a set of dummy model and models with arbitrary hyperparameters. Model hyperparameters are different than model parameters, like the y-intercept and independent variable coefficients, in that they can be optimized to find the best parameters for a given model. The hyper-parameter optimization was implemented using a randomized grid search approach, where discrete sets of hyperparameter values are randomly combined and cross-validated on training data.

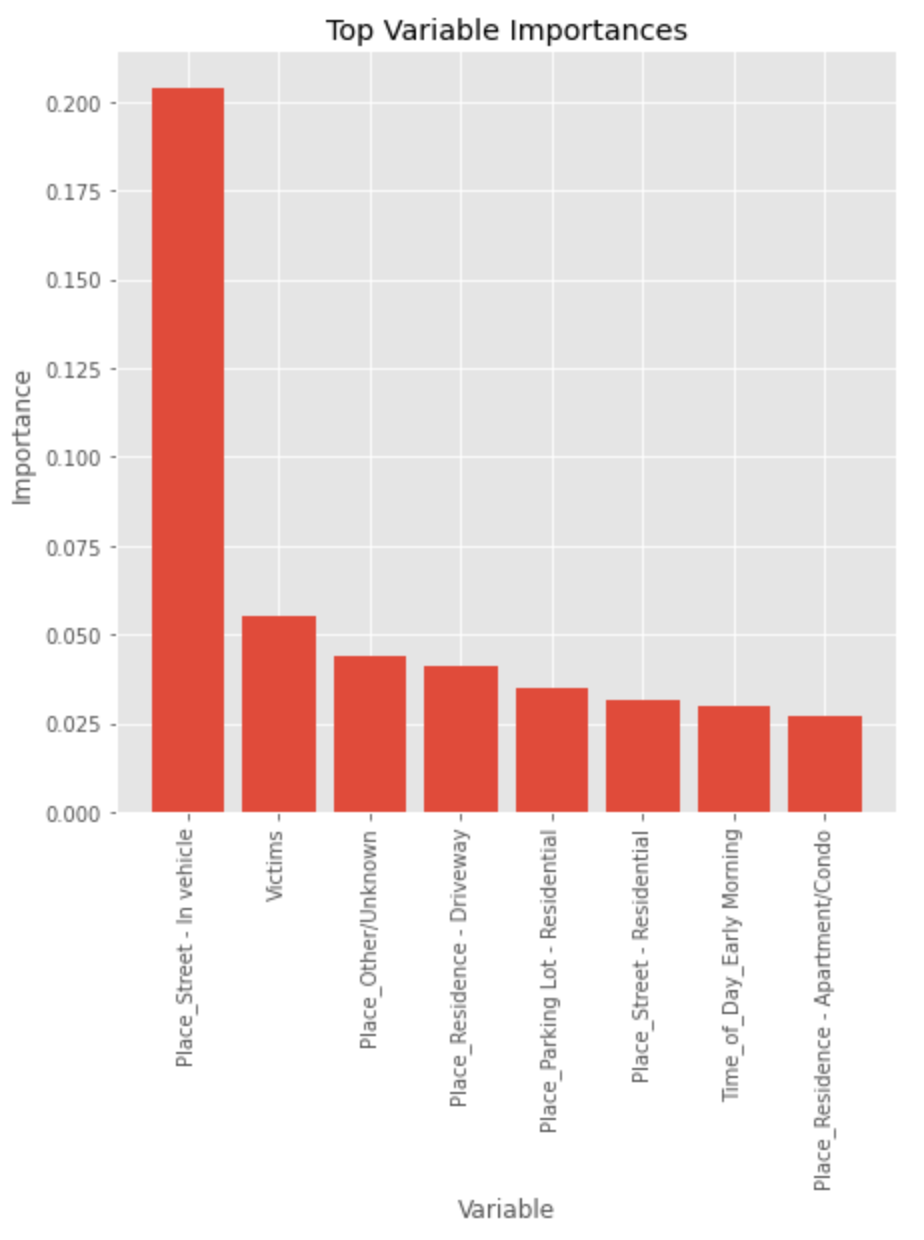
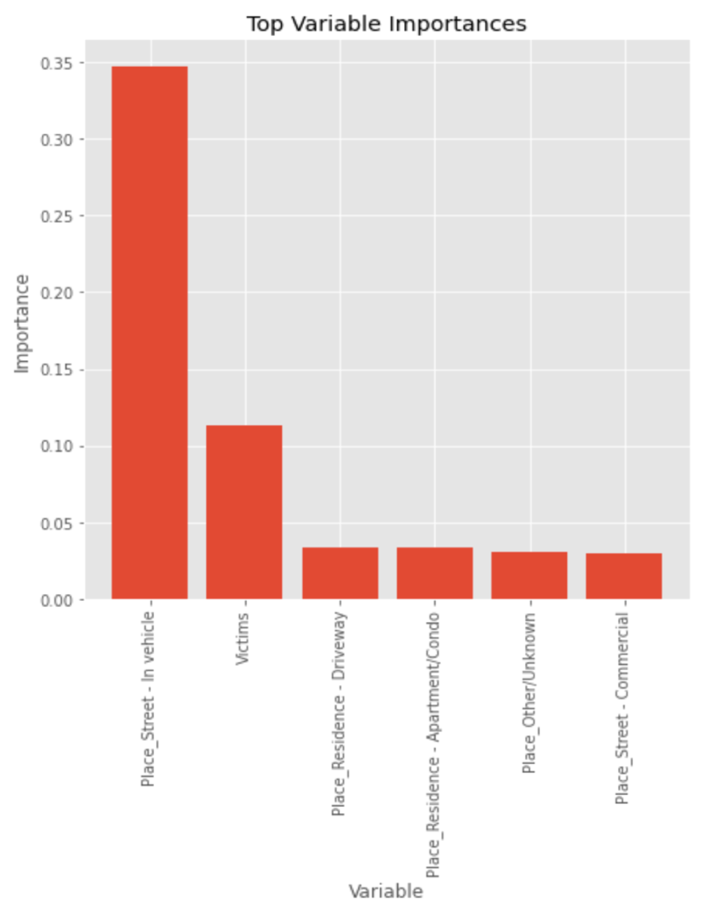
As can be observed from the top-most confusion matrix below, which shows the predicted labels and true labels for the best hyperparameter-optimized model trained on the class imbalanced data, the model does a poor job of differentiating crimes against people from property crimes, as was anticipated during EDA. Specifically, crimes against people were predicted incorrectly as property crimes more often than they were predicted correctly. As a result, the recall score for the crimes against people class was abysmal at about 24.3%.





The recall score of the crime against society class was also somewhat low at only 65.6%. However, despite the low recall scores of the crime against society and crime against person classes, the model still had a ROC AUC of 86.6% due to the fact that the recall score for the crime against property class, which made up the vast majority of observations in the test dataset, was so high at 91.8%. Similarly, since the crime against property class was predicted correctly so often, the precision score for the crime against property class was not affected significantly by the 3,330 crimes against people predicted as property crimes and was still fairly high at 74.2%. Moreover, since the precision scores for the crime against person class and crime against society class were 72.3% and 78.6% respectively, the average precision of the model based on the imbalanced data was 83.7%.

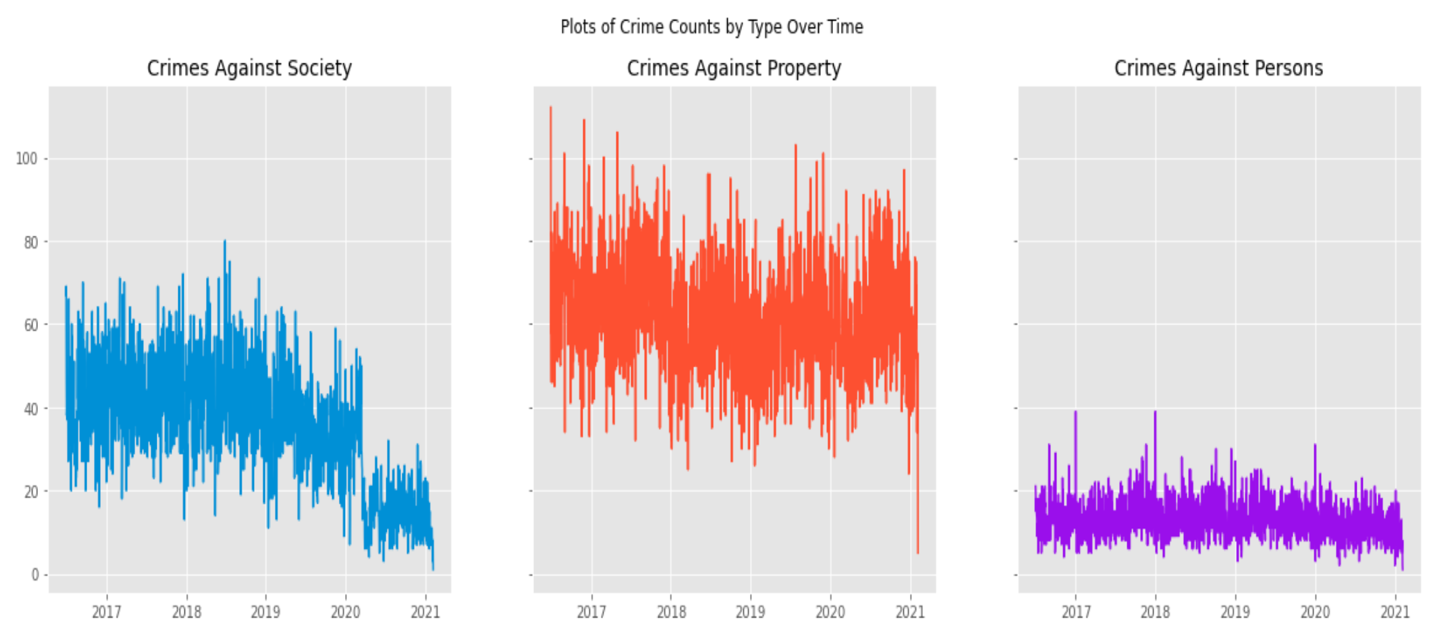
By contrast, the bottom-most confusion matrix above, which shows the predicted and true labels for the best model trained on the class balanced data, shows that the model does a much better job of differentiating crimes against people and property crimes. Specifically, crimes against people were predicted correctly as such more often than they were predicted incorrectly as property crimes. As a result, the recall score for the crimes against person class was now at about 51.5% and the precision score for the crime against property class, in turn, improved to 78.6%.



However, since the model also misclassified a greater number of property crimes as crimes against people, the recall score for the crime against property class decreased to 84.6% and the precision score of the crime against person class decreased to 43.1%. Consequently, the ROC AUC score of the model decreased slightly to 86.2% and the PR AUC, or average precision, also decreased to 82.3%. In sum, the best model based on the class balanced data performs slightly worse across all classification metrics, but it does a better job of correctly predicting crimes against people.

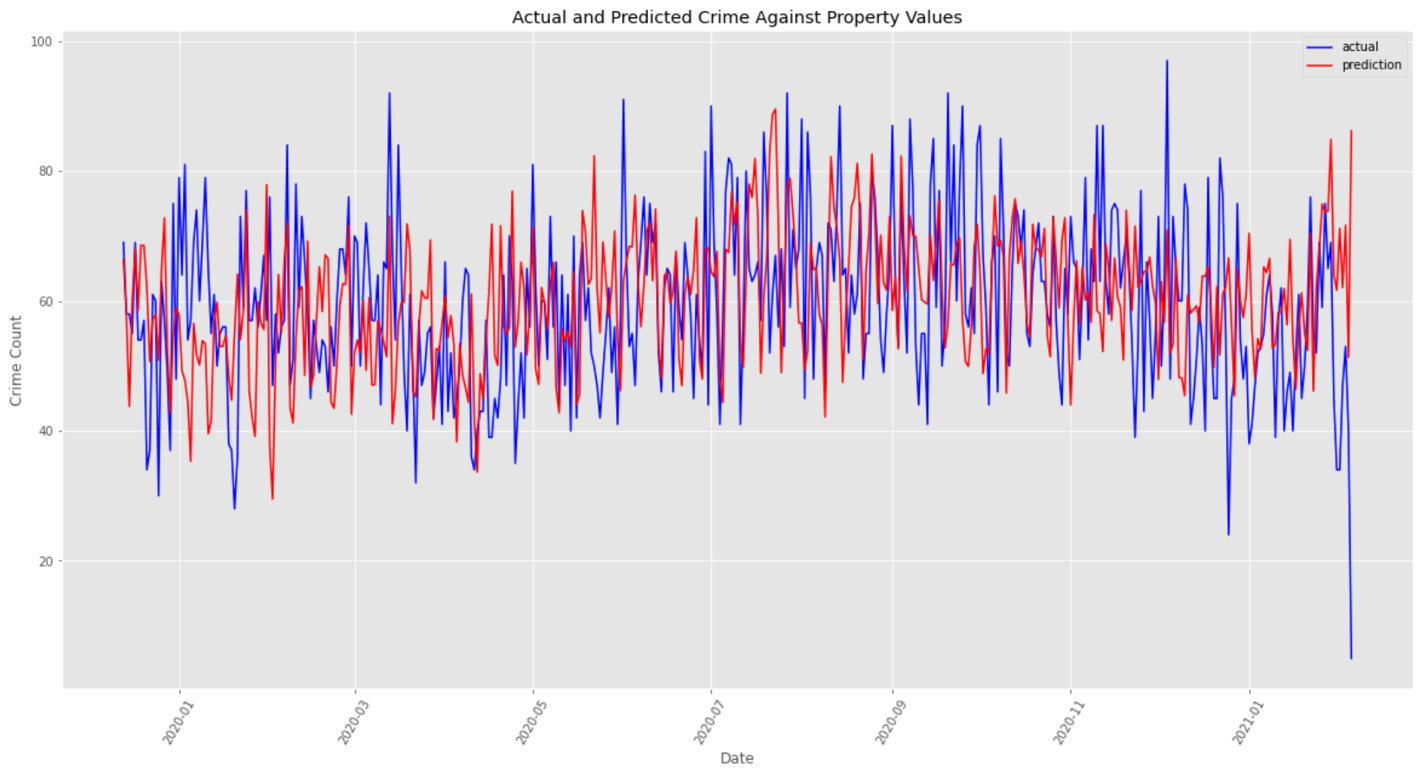
For the best models of both the imbalanced and balanced data, the most important variables were predominantly place variables, as can be seen in the bar charts above. The left-most bar chart shows the most important variables for the model trained on the class-imbalanced data while the right-most bar chart shows the most important variables for the model trained on the class-balanced data. The most important variable for each model was the in-vehicle variable, but this was not surprising given that most crimes occurred in vehicles and the crimes that occurred in vehicles were almost exclusively crimes against society. The second most important variable of both models was actually the number of victims, which also confirms our initial suspicion that crimes against people might be distinguishable from property crimes that occurred in the same places on the basis of the number of victims. The number of victims, however, was ascribed less significance by the model trained on the class-balanced data (shown on the right), presumably because this variable was no longer as important to distinguish property crimes and crimes against people that both occurred in single-family homes or residences given that the model was trained on data in which far more crimes against people occurred in apartments and single-family homes. The remaining variables of each model, though far less significant, were also interesting. For example, the best model trained on the class balanced dataset found more of the places where property crimes predominantly occurred to be important, presumably to distinguish them from crimes against people. However, while one might’ve expected this model to attribute more importance to the apartment place variable, because there were far fewer crimes against people than property crimes in the test dataset, this was not the case.

As for the models themselves, a set of arbitrary neural network and arbitrary random forest models were used to gauge the extent to which hyperparameter-optimization improved the models trained on both the balanced and imbalanced datasets. Surprisingly, the arbitrary neural networks outperformed, not only the arbitrary random forest models, but also the hyperparameter-optimized random forest models trained on both datasets with respect to every metric. As for the models trained on the imbalanced data in particular, the arbitrary neural network even outperformed the hyperparameter-optimized extreme gradient boosted machine model in terms of the log loss. In general, though, the gradient boosted tree hyperparameter-optimized models had the best performance of any hyperparameter-optimized model type with respect to both the imbalanced and balanced data for every classification metric. Specifically, the hyperparameter-optimized light gradient boosted tree models tended to perform just slightly better than the hyperparameter-optimized extreme gradient boosted tree models and the hyperparameter-optimized extreme gradient boosted tree models, in turn, tended to perform slightly better than the arbitrary neural networks which, as previously noted, outperformed the hyperparameter-optimized random forest models. This was true for the models trained on the balanced data in addition to the models trained on the imbalanced data. In sum, the hyperparameter-optimized models tended to improve the performance of the models using arbitrary hyperparameters, regardless of whether the models were trained on the imbalanced data or the balanced data, especially the light gradient boosted-tree models, which were the top hyperparameter-optimized model across every classification metric. In addition to having a superior ROC AUC and PR AUC, as previously mentioned, the light gradient boosted tree model trained on the imbalanced data was able to achieve 75.3% accuracy on the test data while the light gradient boosted tree model trained on the balanced data achieved a slightly lower accuracy of 73.9% on the test data. However, despite having a slightly worse performance with respect to all metrics, the light gradient boosted machine trained on the balanced data was much better at distinguishing crimes against people from property crimes. Lastly, it should be noted that a baseline model that simply predicted the most frequently occurring class regardless of the independent/predictor variables achieved 56.7% accuracy. Thus, both light gradient boosted tree models only improved the dummy model accuracy by just under 20%.

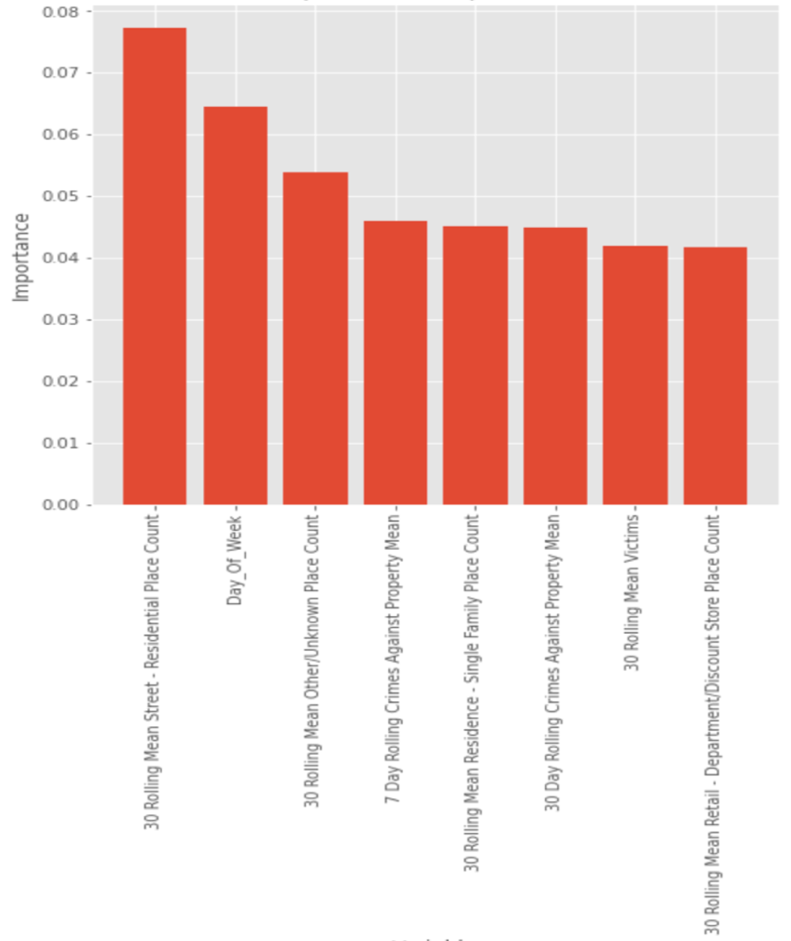


Since I was also interested in looking at aggregate crime numbers over time, I grouped and sorted crimes by day and created a plot of the number of crimes by crime type over time. Surprisingly, it appeared that the only crime type that was affected by Covid-19 was the crime against society crime type. As can be seen from the plot above, crimes against society dropped off steeply shortly after the start of 2020, while the other two crime types were unaffected. Moreover, it appeared that property crimes in Montgomery County were seasonal, as can be inferred from the periodic, sinusoidal pattern in the plot above. Consequently, I wondered whether it would be possible to model the number of crimes per day as a function of time. Specifically, I wanted to see whether or not it would be possible to predict the number of crimes that will occur one day in the future across all three crime types. To do so, I used two different types of models: a triple exponential smoothing model and a gradient boosted tree regressor. The gradient boosted tree regressor used the moving average number of crimes by crime type and place as well as the moving average number of victims up to the day before the day for which the number of crimes was to be predicted. Three different windows, or periods, were used to calculate the moving averages: a window of 2 days, 7 days, and 30 days. The day of the week of the day for which the number of crimes was to be predicted was also incorporated into the model, which seemed acceptable given that we would obviously know the day of the week for any given day well before the day in question.

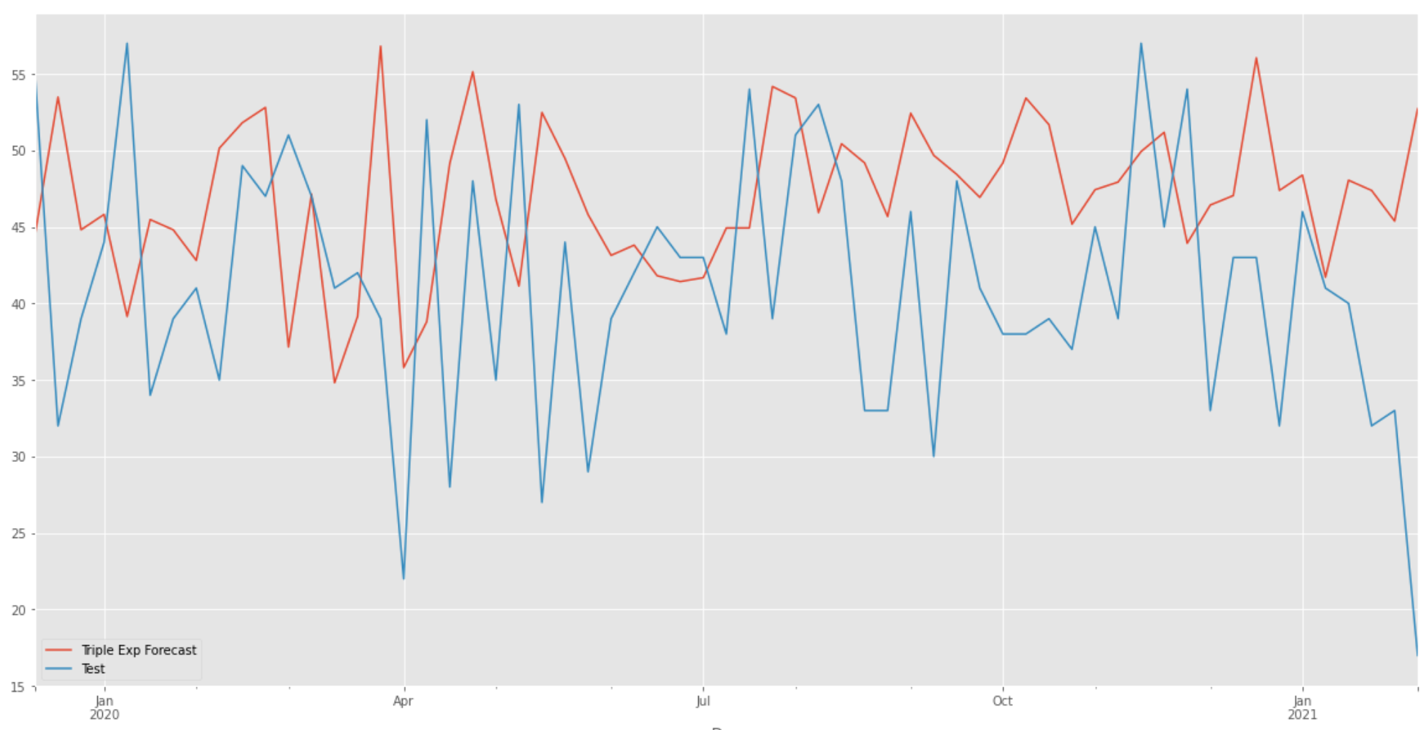
The extreme gradient boosted tree regressor model was trained on the first 75% of the data and tested on the last 25% of the data, which roughly corresponded to time period beginning in 2020 and ending in the present. As one would likely expect, the predictions of the number of crimes against society per day were very poor due to the abrupt change in the number of daily crimes beginning in 2020 as a consequence of Covid-19. The predictions of the number of crimes against people per day were also poor, meaning the moving average of prior crime counts wasn’t a good indicator of future crime counts. On the other hand, the moving averages of the number of crimes by place and the number of property crimes seemed to be good predictors of future property crimes as can be inferred from the superimposition of predicted and actual crimes over the test period in the plot below. The model had a mean absolute error (MAE) of 11.34 crimes, a root mean squared error (RMSE) of 14.52 crimes, and a mean absolute percentage error (MAPE) of 23.96%. While the model’s predictions did not capture many of the daily spikes and drops in property crime numbers, the model certainly captured some of them and seemed to do a god job of predicting the general trend of property crimes over the time period.



Moreover, as was suspected, the 30-day moving average and 7-day moving average of crimes and the day of the week variable were among the most important variables. This suggests that the number of crimes on a given day are highly dependent on the day of the week and the average number of crimes over the past 30 days and 7 days, as can be observed in the bar chart below.



Lastly, since it would ultimately be more useful to be able to predict the number of crimes by region in a city with an excess of crimes, I decided to divide Silver Spring, where the majority of crimes occurred, into 5 different clusters based on the locations of crime incidents and attempted to predict the number of crimes by crime type for each cluster for each day. The results across the board were poor and were, therefore, not included in this report. However, since it is profoundly difficult to predict any variable on a daily basis, particularly ones that have a high degree of volatility from one day to the next (like crime incident numbers), I figured it would make more sense to make predictions on a less granular scale by using weekly crime counts. Not surprisingly, the accuracy of the forecasting was improved when the data were grouped by and predicted by week rather than day. Like the previous forecasting task, both models did an inadequate job of predicting crimes against society, however, the forecasts of weekly property crimes by location were much more acceptable and were very similar across both models. Though forecasts for some regions in Silver Spring were worse than others, the models seemed to give a good approximation of the general trend of weekly property crimes for most regions. A graph of the actual and predicted forecast of weekly property crimes for one of the 5 regions is shown below.



This particular cluster had a MAE of 9.6 crimes, a RMSE of 11.86 crimes, and a MAPE of 27.9%. The results for the other clusters of property crimes were similar as well. As for crimes against people, while the predictions were not quite as accurate based on the MAPE scores, they did seem to reflect the overall trend of weekly crimes as well. For example, the model of weekly crimes against people for the same cluster shown above had a MAE of 2.75 crimes, an RMSE of 3.51 crimes, and a MAPE of 61.2%.

While the “Crimes” dataset certainly contained a fair amount of useful information, certain important variables, such as the end date time and the dispatch time, were missing a significant number of values. Additionally, the dataset could’ve benefitted from including more demographic/numerical data, especially given that the ACS data failed to reveal any patterns among crime types. For example, adding an age, gender, and race column for offenders (and even victims) would likely yield additional insights. If possible, data on how many times an offender has been arrested or how many times a victim has been a victim of crimes would also likely be useful. While I can’t guarantee that any of these suggestions would result in better modeling results, I suspect that they would. That said, the classification model developed from this dataset could be deployed immediately to aid in determining what type of crime occurred simply based on the type of place, police district, time of day, day of the week, and demographic variables based on the census block where crimes occurred. Likewise, the forecasting model developed from this dataset could also be deployed at once to determine the number of property crimes that will occur the next day in Montgomery County and the number of property crimes or crimes against people that will occur in any of the 5 different locations in Silver Spring a week in the future.

I would like to thank Dennis Linders of the Montgomery County Government, as well as Data Montgomery, for the opportunity to present my models. I would also like to thank my professors, particularly Professor Rachel Saidi, Professor Iapalucci, and Dr. Mohamed Abdirisak, for helping me to accumulate such a wealth of knowledge in the fields of statistics and machine learning during the course of the Data Science Certificate program at Montgomery College.