CYO Project Report

Magdalena Borisova

6/15/2019

Abstract

This report presents a machine learning algorithm for violent incidents recorded during police 9-1-1 calls. The algorithm provides a model for violent incident predictions. The model was evaluated using overall accuracy, along with sensitivity and specificity. The highest overall accuracy achieved was 0.59, while sensitivity and specificity were 0.77 and 0.41, respectively. Despite the relatively low evaluation quantities attained, the model highlights some interesting patterns related to violent incidents.

Introduction

A violent incident is defined as an incident that involves a threat or an attack on a person [10, 11]. An example of a violent incident is an assault, while an example of a non-violent incident is property damage [7]. National trends in the United States indicate an increase in violent crime rates beginning in 2014 [4]. Research done on data obtained from the Milwaukee Police Department 9-1-1 calls, demonstrated that broadcasted violent incidents on African American men affected the public's proactive response to reporting crime and trust in law enforcement agencies [2, 6]. In Seattle, the Police Department's 9-1-1 incident response data is readily available to the public [8]. The Police Department's online database is updated regularly and contains data from 2001 to present. The data includes information on incident description, location and time. The goal for this project was to create a violent incident prediction system by building a machine learning algorithm on the Seattle Police Department 9-1-1 Incident Response data from the year of 2017. The dataset was obtained on Kaggle's website, which contained datasets that were ready to use for machine learning analysis [5]. The dataset was split into two sets, a train set and a test set. A machine learning algorithm that predicts violent incidents was developed and used to train a model on the train set and was then tested on the test set.

Methods

Sample

The Seattle Police Department 9-1-1 Incident Response online database contains response activity data updated daily. The current database includes 3.97 million observations starting in 2001 [5]. The data used in this analysis was obtained from the Kaggle website, on which part of the original version of the data was available for learning purposes [7]. The dataset used for this analysis originally included 2135 observations for the year 2017. Some observations were removed because of missing data, which lowered the number of observations to 1773. The dataset contained 4 categories of variables. The first category included variables, which served as identifying characteristics for each observation in the dataset: CAD CDW ID, CAD Event Number and General Offense Number. The second category contained variables, which were related to the type of incident and were filled out by the primary officer on scene: Event Clearance Code, Event Clearance Description, Event Clearance Subgroup, Event Clearance Group and Event Clearance Date. The third category included variables, which presented incident location information: Hundred Block Location, District Sector, Zone Beat, Census Tract, Longitude, Latitude and Incident Location. And the fourth category had variables, which classified the type of incident at the time of the 9-1-1 call: Initial Type Description, Initial Type Subgroup, Initial Type Group and At Scene Time. The dataset did not include any demographic information, the time of the 9-1-1 call and the call recordings. The dataset contained 1773 CAD CDW ID's, 1773 CAD Event Numbers, 1773 General Offense Numbers, 10 distinct Event Clearance Codes, 10 distinct Event Clearance Descriptions, 6 distinct Event Clearance Subgroups, 5 distinct Event Clearance Groups, 1750 distinct Event Clearance Dates, 1211 distinct Hundred Block Locations, 17 distinct District Sectors, 51 distinct Zone Beats, 1048 distinct Census Tracts, 1105 distinct Longitude locations, 1127 distinct Latitude

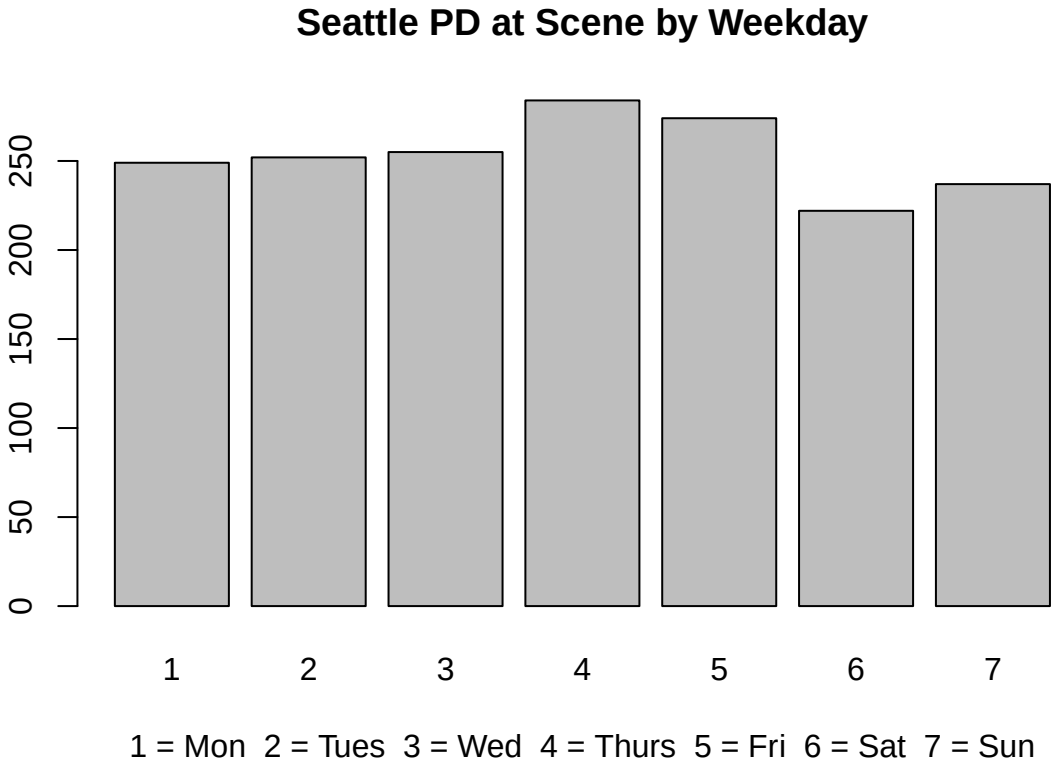
locations, 1188 distinct Incident Locations, 52 distinct Initial Type Descriptions, 20 distinct Initial Type Subgroups, 19 distinct Initial Type Groups and 1762 distinct At Scene Times.

Table: Sample characteristics

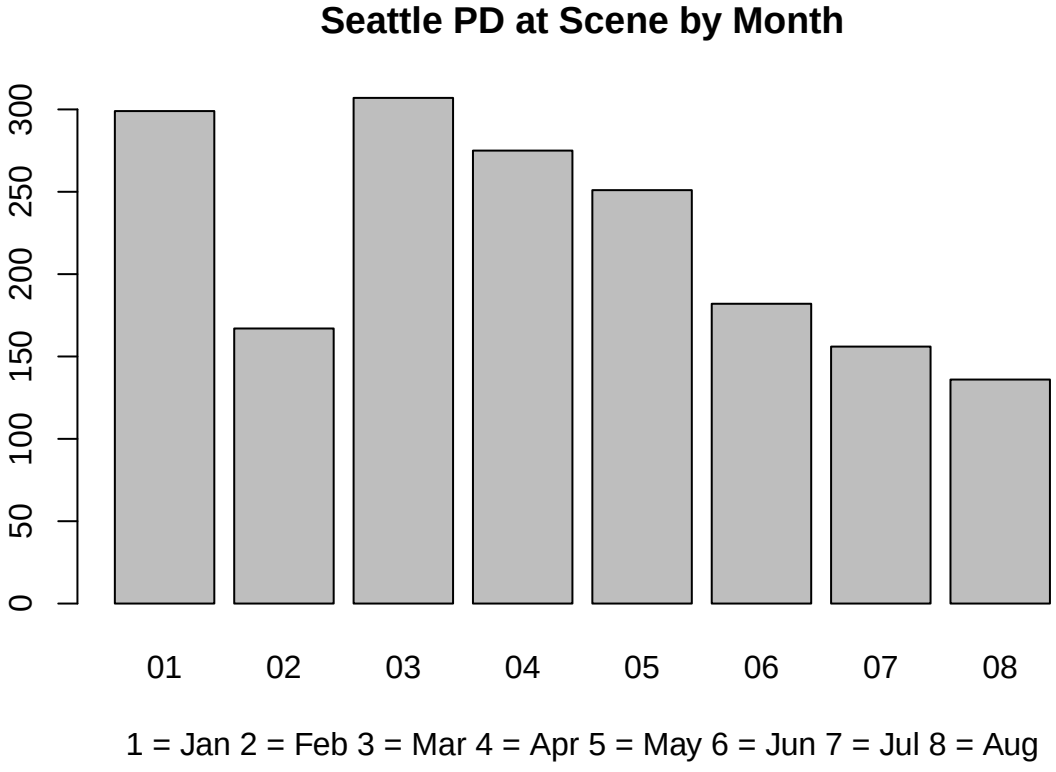
variable	distinct	missing
CAD CDW ID	1773	0
CAD Event Number	1773	0
General Offense Number	1773	0
Event Clearance Code	10	0
Event Clearance Description	10	0
Event Clearance Subgroup	6	0
Event Clearance Group	5	0
Event Clearance Date	1750	0
Hundred Block Location	1211	0
District Sector	17	0
Zone Beat	51	0
Census Tract	1048	0
Longitude	1105	0
Latitude	1127	0
Incident Location	1188	0
Initial Type Description	52	0
Initial Type Subgroup	20	0
Initial Type Group	19	0
At Scene Time	1762	0

Data Cleaning, Exploration and Visualization:

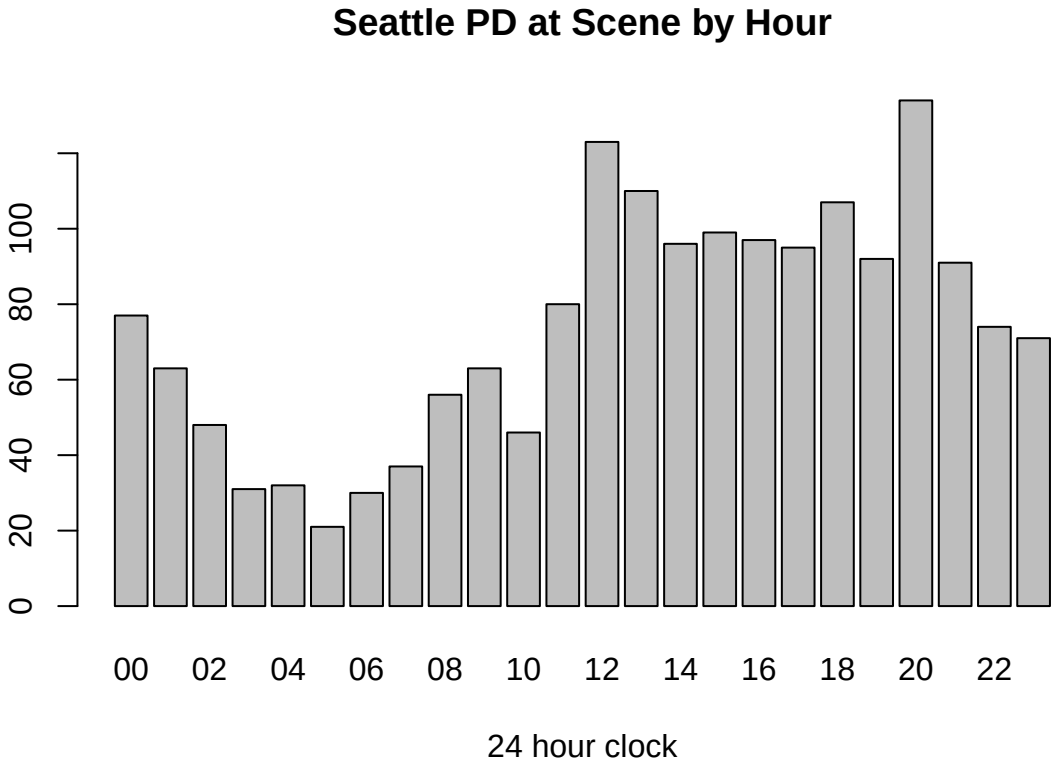
Data exploration and visualization included confirming the structure and dimensions of the dataset, inspecting the first few observations in the dataset, checking for missing and distinct values for each variable in the set and producing graphs for some of the variables. The data originally contained 19 variables. The "At Scene Time" variable had missing values. Because the variable was going to be used in the data analysis, the missing observations were dropped from the dataset. The variable was originally recorded as date and time. For this analysis, it was formatted and split into hours, minutes, weekdays and months to aid in future analysis. The dataset did not have observations recorded between the months of September and December. The "Initial Type Group" variable was used to create a new categorical variable called "violent incident", which was assigned a value of 1 if the incident was coded as violent and a value of 0 if the incident was coded as non-violent. The criteria for violence were taken from the U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics [10, 11]. The appearance of the graphs suggested that not all hours, weekdays and months had the same number of incidents. Thursdays and Fridays had a higher number of incidents when compared to other days of the week.



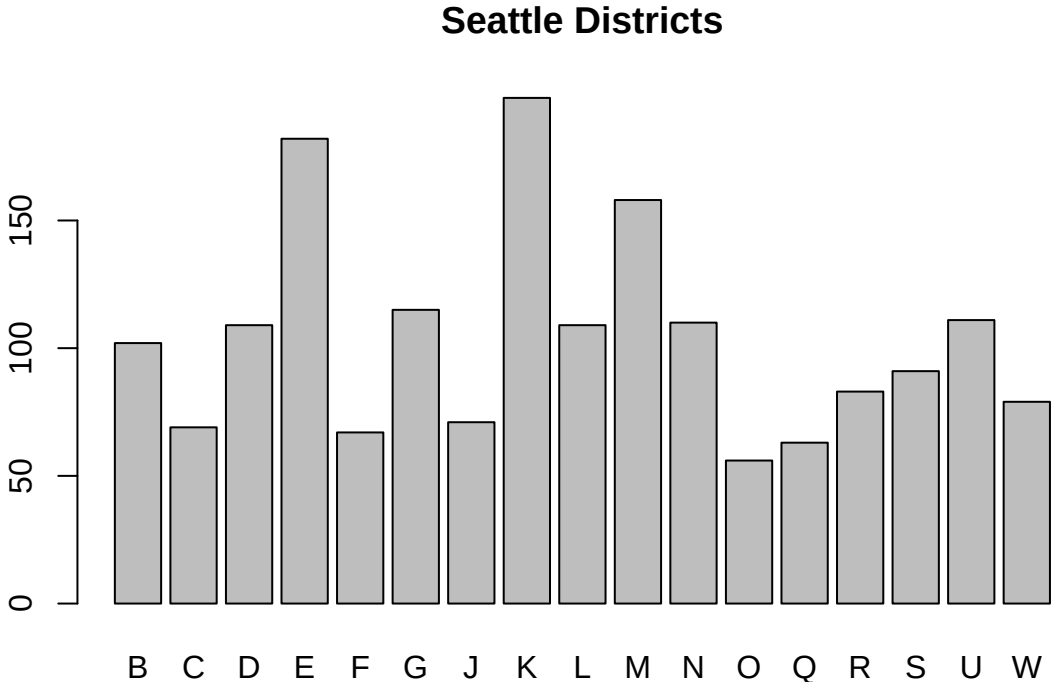
It appeared that January, March, April and May had a higher number of incidents when compared to other months of the year.

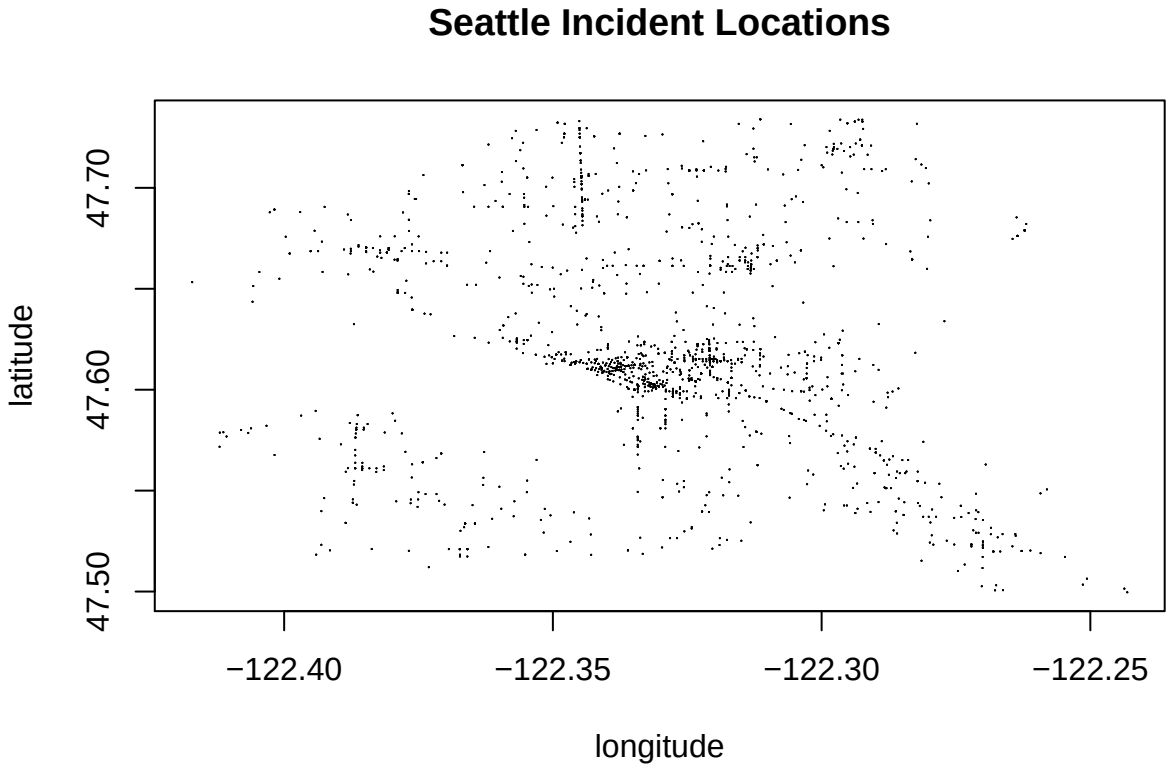


Incident frequency started increasing around noon and stayed high until 9 pm.

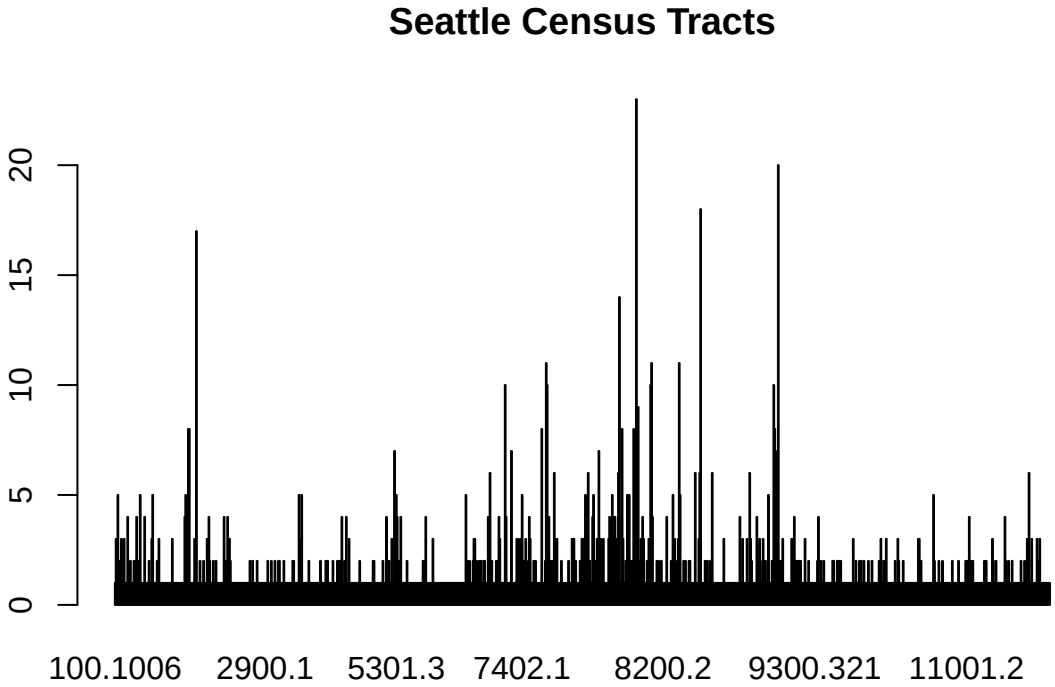


The graphs also suggested that not every location had the same number of incidents. Incidents were more common in districts E, K and M when compared to other districts.

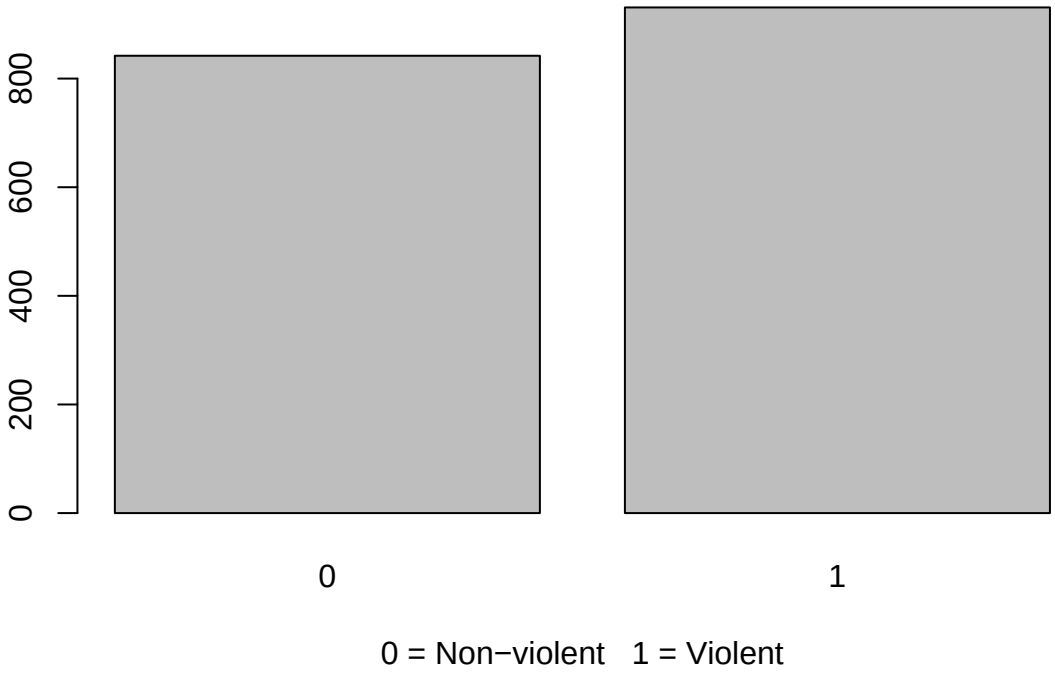




Number of incidents was also higher in Census tracts with population between 1200 and 1800 people, and 6700 and 9500 people.



A plot of number of incidents showed a higher number of violent than non-violent incidents.



Insights:

The visual insights gained from the graphs illustrate a possible relationship between violent incidents and time and location. Different hours of the day along with different days of the week and different months of the year appear to have an impact on the number of incidents [1]. This appears to be true for the location of the incidents as well. Research suggests a link between location and violence, which is worth exploring for the Seattle Police Department Database [2, 6]. The research was focused on African American neighborhoods and police activity. Although Caucasians are the highest percent of the population of Seattle, the location/neighborhood appear to have an impact on incident frequency and should be included in the analysis to follow [9].

Training and test set:

The Seattle Police Department Incident Response dataset was split into a train set named "train" and a test set named "test". The test set was 10% of the dataset.

Modeling approach:

A machine learning algorithm was built to predict violent incidents. The accuracy of the trained model was measured by computing overall accuracy, sensitivity and specificity. A k-nearest neighbors' machine learning algorithm was implemented with all predictors included in the model. An optimal k was selected first and then each predictor was removed from the model, one at a time, to test whether that improved the accuracy and sensitivity. The model was optimized by maximizing the overall accuracy and sensitivity in the test set. The final model, which included "hours", "weekdays", "months", "Incident location", "Hundred block location", "District Sector", "Zone beat" and "Census tract", was selected based on its high accuracy and sensitivity.

Results

Different models were trained to predict violent incidents in the test set. The goal was to train a model that would improve overall accuracy and sensitivity. The model with all of the time and location predictors included, produced the highest accuracy and sensitivity. Although the overall accuracy did not reach more than 0.59, the sensitivity of the trained model improved and reached 0.77. Specificity of 0.41 was low but, due to how the violent incident variable was coded, sensitivity was a more important predictor. It was more important to positively identify all violent incidents as such regardless of whether some of the non-violent incidents would also fall into the category of violent.

Conclusion

Key Findings:

The final model opened up a possibility to explore the relationship between violent incidents, time and location. It appeared that the predictors had an effect on the outcome even though the size of the measured effects was not as desirable. It is also important to mention that just because the overall accuracy for the model was low, the sensitivity was not. Sensitivity is an important predictor of how well an algorithm is able to predict a positive outcome when the outcome is truly positive [3]. In this machine learning algorithm, it was important to correctly classify violent incidents as violent.

Limitations:

The dataset contained missing data for the time predictor. The observations with missing values of the "At Scene Time" variable, which was then separated into hours, weekdays and months, were not used in the analysis. Information for the months of September through December was not available. The violent incident variable might have been coded incorrectly. Demographic information was not made available along with 9-1-1 call recordings and time of 9-1-1 calls. If time of calls was available, one could have incorporated the call frequency in the analysis and controlled for demographic characteristics in the model [2]. This report does not include data from different years. It is possible that including additional years in the modeling approach may improve overall accuracy and sensitivity predictions.

Future areas of research:

The model creates a possibility to explore the relationship between violent incidents, time and location more by finding and adding additional predictors, such as demographic information, 9-1-1 call recordings and temporal features of the 9-1-1 calls.

References

[1] Comparing Offending by Adults & Juveniles. (2016). U.S. Department of Justice, Office of Justice Programs. Retrieved from https://www.ojjdp.gov/ojstatbb/offenders/qa03401.asp?qaDate=2016

[2] Desmond, M., Papachristos, A. V. & Kirk, D.S. (2016). Police Violence and Citizen Crime Reporting in the Black Community. American Sociological Review, 81(5), 857–876. Retrieved from https://www.asanet.org/sites/default/files/attach/journals/oct16asrfeature.pdf

[3] Irizarry, R. (2019). Introduction to Data Science [Book]. Retrieved from (https://rafalab.github.io/dsbook/large-datasets.html#recommendation-systems)

[4] James, N. (2018). Recent Violent Crime Trends in the United States. Congressional Research Service. Retrieved from https://fas.org/sgp/crs/misc/R45236.pdf

[5] Kaggle's curated list of datasets for machine learning analysis. Retrieved from https://www.kaggle.com/annavictoria/ml-friendly-public-datasets?utm_medium=email&utm_source=intercom&utm_campaign=data+projects+onboarding

[6] Lantigua-Williams, J. (2016, September 28). Police Brutality Leads to Thousands Fewer Calls to 911. It might be making neighborhoods less safe. The Atlantic. Retrieved from https://www.theatlantic.com/politics/archive/2016/09/police-violence-lowers-911-calls-in-black-neighborhoods/501908/

[7] Seattle Police Department 911 Incident Response dataset. Retrieved from https://www.kaggle.com/sohier/seattle-police-department-911-incident-response

[8] Seattle Police Department Public Datasets (Original Version of CYO dataset) Retrieved from https://www.seattle.gov/police/information-and-data/public-data-sets

[9] U.S. Census Bureau, Quick Facts. Retrieved from https://www.census.gov/quickfacts/fact/table/seattlecitywashington,US/PST045218

[10] U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics Fact Sheet. (2004). Retrieved from https://www.bjs.gov/content/pub/ascii/pnoesp.txt

[11] U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. Retrieved from https://www.bjs.gov/index.cfm?ty=tp&tid=31