Predicting crime rates in Boston using Machine Learning models

Berend Markhorst (2598083) Esther Kuikman (2589803)
Dewy Koopman (2600015) Nina Malbašić (2594377)
Alex Mijatovich (2605863)

March 29, 2019

Abstract

This paper is aiming to predict crimes in Boston. With the help of different classification and regression models the number of crimes with and without a shooting is predicted. This involved statistical analysis of data that compared relations between features amongst others. For validation of the accuracy estimator, N-fold cross validation was used. With the classification model, results have been obtained that show that the 4-fold is the best model for the SVM model and that temperature influences the crimes. Furthermore, the results show that the linear regression model is the best regression method for the task to predict the number of shootings that will occur at a specific hour and day.

1 Introduction

For this research project, data is used from Kaggle, containing information about approximately 330.000 crimes between the years 2015 and 2018 in the city Boston, Massachusetts. The objective of this research is to predict crimes in Boston using machine learning algorithms and the aforementioned data.

1.1 Motivation

Machine Learning can give useful insights in big data. In this case, it is able to see patterns between incidents that occured in Boston that humans do not see. This information can be useful for the police. For instance, the daily surveillance routes by the police can be done more efficiently. Another instance where this information is useful is to build or move police stations to better locations. In this way, the application of machine learning can result in a safer environment for the residents of Boston and it could reduce costs for the Boston Police Department.

1.2 Hypothesis

Various models in Python in attempt to and predict crimes in Boston will be built. It is thought that a few features will have an affect on the crime rates. First, the temperature, since when its warmer, more people tend to go outside and therefore more likely to commit a crime. Second, the day of the week, it is thought that criminals will be more active during the weekends than on weekdays. At last, the month of the year. Some months are busier than others, December is usually busier due to the Christmas and holiday shopping at the end of the year. The summer months tend to be busier as well, because students have holidays in this period.

1.3 Literature Review

Research has been conducted on a few possible features that might have a correlation with crime. In this section, three such articles are reviewed. The first article that is about whether crime is influenced by time, the second article is about the influence of temperature on crime rates and the last article investigates the relation between time of year and crime rates.

According to Felson *et al.* [1], crime varies more by hour of day than by any other predictor we know. However, this variation is analysed rarely, one of the reasons for this is that the data produces too many categories. The article concludes that forecasting strategies based on annual, quarterly and maybe even

monthly data miss the essential of the dynamic in crime rate trends. It is time to recognize that crime has its own dynamics driven by day and hour.

However, the features *year*, *quarter* and *month* should still be used in forecasting because they contain information useful for prediction models. The article gives some simple indicators that can help apply the hourly observation of crime. "However, a larger problem needs additional work-how to think about hourly variations in crime" [1]. This suggests that hourly analysis is not capable in every situation. The paper only considered one crime and a limited range of cities.

The effect of global warming on the occurrence of violence and conflict have been researched extensively the last ten years. Some studies found a significant relation (e.g. suicide rates and ambient temperature). It is difficult to measure such a relation since a large number of factors are involved: such as decreased rainfall and other social, economical and political conditions. However, as can be seen in Figure 1 above, Tiihonen *et, al.* found a strong relation between the mean monthly ambient temperature and the number of violent crimes [2]. Overall, it was shown that during 1996-2013, the ambient temperature explained 10% of the variance in violent crime rates in Finland. Also, the results suggest that a 2 °C increase in average temperatures would increase violent crime rates by more than 3% in non-tropical and non-subtropical areas, if other contributing factors remained constant.

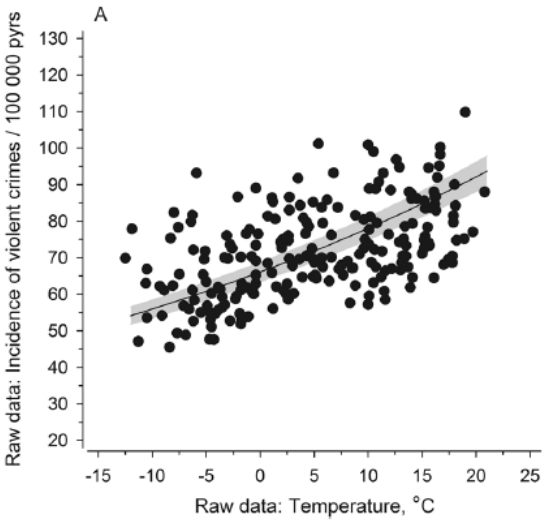

Figure 1: Scatter plot between mean monthly ambient temperature and the number of violent crimes in Finland.

In December, according to reports from the National Crime Victimization Survey (NCVS), two specific types of crime increased in the number of occurrences: Robbery and personal larceny [3]. This seems highly related to the desperation that people feel in the month December. Presents have to be bought for Christmas and many people face financial difficulties during the end of the year because of this. Next to this, a lot of people are carrying around expensive items, so it is a good time for criminals to strike. What is striking is that more violent crimes, like murder and rape, do not increase in December. Another observation is that during the summer months property crimes are the highest. It is assumed that this increase is strongly related to the schools being closed.

2 Data exploration and preparation

The original data set was retrieved from the Kaggle website. It exists out of 330.000 incidents that have occurred in Boston over the time period of May 2015 until December 2018. Using this entire data set would provide inaccurate results if analysis would be conducted on the number of incidents per month; the month May could show to have a higher number of incidents but this is relative as it is also a month that has a higher frequency of occurring in the data set. Therefore, the data set has been adjusted to only consist of dates that correspond to the year 2016 and 2017. This has shrunk the data set to 200.000 incidents. Each incident in the data set contains 19 distinct features. For further investigation

and accuracy of our analysis, two features are added to the data set: the weather conditions and baseball sport events for the day of each incident. The description and data type of each feature of the final data set can be found in Table 5 in the appendix.

To get a better understanding of the data, a more in-depth research is conducted that could provide insight for answering the research question. Therefore, the relation between shootings, date, type of crime, location and weather condition is investigated.

To start, the number of incidents in relation to the date and weather condition are determined. The relation between the number of incidents with shootings occurred per month and the temperature is shown in Figure 2. Also, the day and hour of incidents are investigated. This has given an interesting insight that on Thursday the lowest number of crimes occurred (see Figure 2). Moreover, a heat map was constructed that showed the relation between hour and weekday (see Figure 11 in the Appendix).

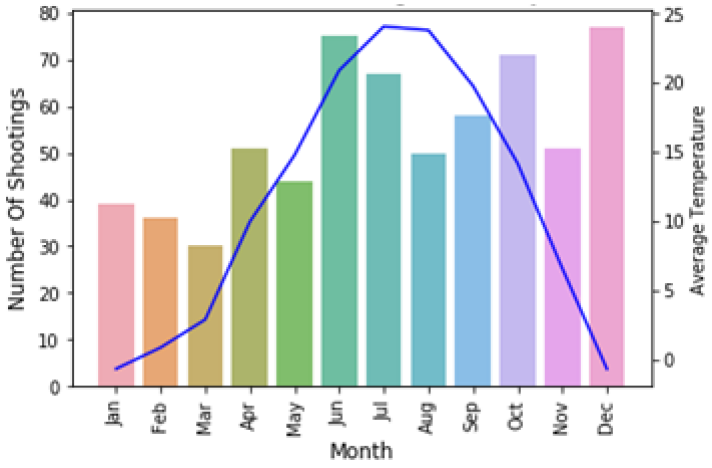

Figure 2: Relation between Shootings and Temperature

The next feature that is investigated, is the type of incidents that occurred. From the analysis the most common type of incident that can be determined. However, this did not belong to the serious type of offenses. Therefore, a selection is made of incidents that were considered serious offenses. The location of the top five incidents is shown on the map in Figure 12 in the Appendix. Also, a similar map is made for the top five incidents that involved shootings (see Figure 3 in the appendix).

Figure 3: Shooting counts per neighborhood

The location of these incidents is the next important feature to investigate. In Figure 3 the frequency of shootings per neighbourhood is shown. Each neighbourhood has been given a certain shade of blue. The darkness of the shade indicates the frequency of incidents with shootings: the darker the colour, the higher the frequency. This analysis has shown that the top two neighbourhoods with the highest frequency are: Dorchester and Roxbury.

The data exploration also extended to some statistical analysis of the data. This could be useful when choosing the prediction model. The data corresponding to the date are hour, day of the week and month that a shooting occurred. As Table 5 in the appendix shows, only the feature *temperature* is numeric, for which information can be gained about its distribution. This can be determined with the use of a histogram and QQ-plot. Figure 4 shows the normal QQ-plot of the feature temperature, that indicates that temperature follows a normal distribution that is lightly tailed (its histogram can be found

in the appendix, Figure 14). Figure 5 shows the number of crimes each day of the week. The error bars indicate the variance of the number of crimes per day, measured over all weeks of the year.

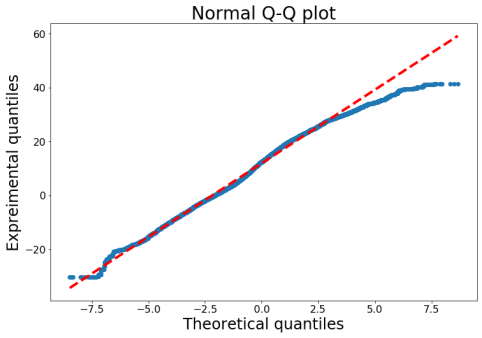

Figure 4: Normal QQ-plot

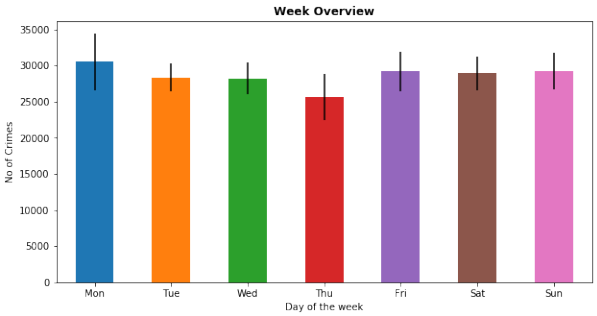

Figure 5: Week overview

Also, linear correlation is checked between incidents with shootings and the temperature. This was already taken into consideration in Figure 2, however, this can also be proven statistically. For this, the Pearson correlation coefficients can be used. The values are limited between 1 and -1, for which 1 indicates a strong positive relation, 0 indicates no relation while -1 indicates a strong negative relation. The calculated correlation value for the feature temperature is 3.38e−5.

3 Methods

3.1 N-fold cross validation

In this project n-fold cross validation is used to validate the accuracy estimation. N-fold cross validation, also known as rotation estimation, randomly splits the data set D into n mutually exclusive subsets, the folds, $D1, D2, \ldots, Dn$ of approximately equal size. A model is trained on every subset of the data set D with t element of $\{1, 2, \ldots, N\}$ [4].

Cross validation is used after the split between test and training data. Because of that, the cross validation only sees the training data and the test data can still be used for the final accuracy estimation. Cross validation can be a costly process because for every fold a new model has to be trained.

Using the cross validation, the hyper parameters of the model can be determined. After this, the final model can be evaluated on the test data. In Figure 6, an impression is given of the working of n-fold cross validation. N, or k in this case, is equal to four. All the data is divided in four equal parts and a model is trained four times on different training and test data.

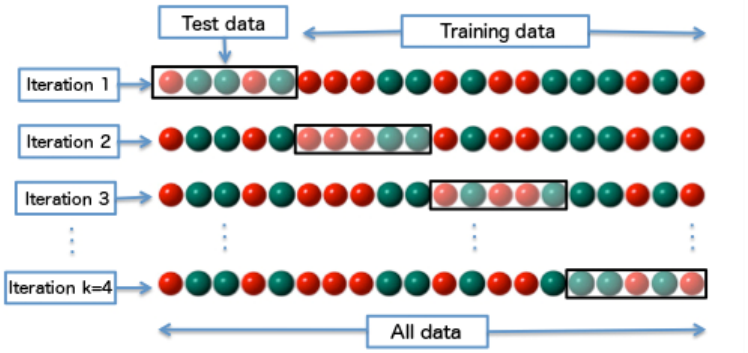

Figure 6: N-fold cross validation impression

3.2 Classification

Using the data set, the goal is to predict if an incident ends up with a shooting or not. In order to do this, a suitable classifier is needed. A support vector machine is used to do this classification, because the linear support vector machine has little assumptions of the input data which makes it easy to use.

3.2.1 Super vector machine

A support vector machine is a model that classifies observations using a decision boundary which tries to separate the different classes as good as possible. The kernel is set as a linear function, because a linear kernel is robust against outliers. This can be seen in Figure 7.

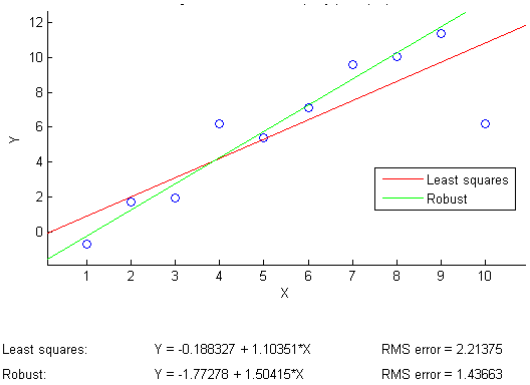

Figure 7: Example of linear robustness

In order to do so, the Support Vector Machine generates two support vectors which go through the two points that are the closest to the decision boundary. In Figure 8, the striped line is the decision boundary and the lines on both sides of this decision boundary at distance 1 are the support vectors. This distance between the support vectors and the decision boundary is the margin, is assumed to be the same on both sides. Figure 8 below is for a one-dimensional feature space. However, most of the time a higher dimensional space is used. The support vectors can be defined as:

$$W^t X + b \leq -1, \textit{ negative support vector}$$

$$W^t X + b \geq 1, \textit{ positive support vector}$$

The support vector machine tries to maximize twice the length of the margin such that the two equations hold. Maximize: $\frac{2}{\|w\|}$ such that $y_i(W^t x_i + b) \geq 1 for\ all\ x_i$

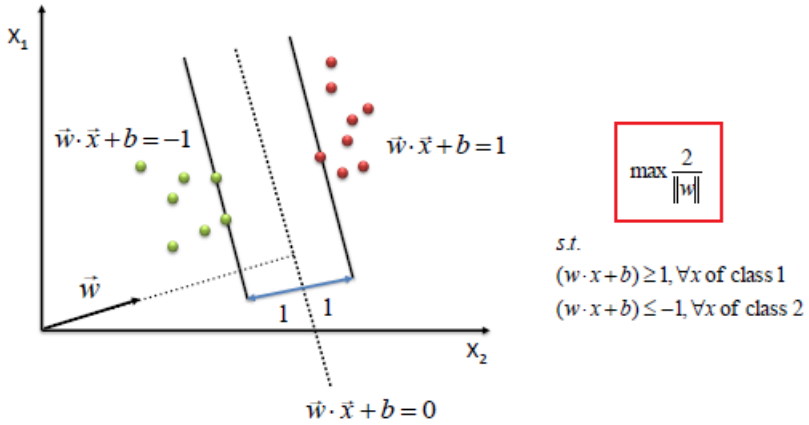

Figure 8: Support vector machine [5]

However, this works in the optimal state when both observations are linearly separated or if the data has little noise. When this is not the case, soft margin are used. This classifies the instances by the following equation.

$$minimize: \frac{1}{2\|w\|} + C\sum_i p_i \textit{ such that } y_i(W^t x_i + b) \geq 1 - p_i \textit{ for all } x_i \textit{ and } p_i \geq 0$$

3.2.2 Feature selection

The data set contains more than 40 features. In the section Data Exploration, correlations between shootings and other features were looked at. From the data set, a subset is used for the classifier with the features *temperature*, *hour*, *weekday* and *Shootings*.

3.2.3 Accuracy

The classifier is tested using the accuracy. The accuracy of a trained model is the number of correct assigned instances divided by the total number of instances.

3.2.4 Baseline

The accuracy of the trained classifier should be better than the baseline, the accuracy that is obtained when all instances are assigned to the same class. The data consist of 196.492 instances, only 645 instances contained a shooting, which resulted in a major class imbalance (0,33%). This means that the classifier has to beat the accuracy of $\frac{196.492-656}{196.492} = 0,9967$. This means that the classifier should have an accuracy higher than 0,9967, otherwise guessing that all instances don't become shootings is more accurate than the model.

3.3 Regression

For this regression model only the data from 2016 and 2017 are used (as can be seen in the section Data exploration and preparation). The goal of this regression model is to predict how many instances of shootings will occur on an hour of a specific day of the years 2016 and 2017. To be able to make this prediction the following features are used: *temperature*, *day of the week* and *hour*.

The Data Frame was grouped on the date such that every hour of every date was selected and next to this the average temperature per hour was used in the Data Frame. In Table 1 the first five lines of the Data Frame can be seen to give a visual representation of the grouped Data Frame. Here, the last column is number of shootings, which is the target value and indicates the number of incidents with shootings that occured that day and hour.

Row nr.	Temperature	Hour	Day	Month	#Shootings
0	4.23	0	Friday	January	40
1	3.73	1	Friday	January	29
2	3.31	2	Friday	January	24
3	3.13	3	Friday	January	26
4	3.01	4	Friday	January	12

Table 1: Data Frame

Since all the date and time features are categorical, these cannot be used for regression. For this reason, one-hot encoding was used. One-hot encoding is one of the most used methods for multi-class classification tasks due to its simplicity and effectiveness [6]. Every state has a separate bit of state in one-hot encoding. This is where the term *one-hot encoding* comes from, because at any time only one bit is "hot" or True. An example for three states is: 001, 010 and 100 [7].

The data was split such that the test set contained 10.000 instances. On this test, three kinds of regression are applied:

1. Linear regression
2. Decision tree
3. K-nearest neighbor

To calculate the accuracy of these models, the mean square error is used. The mean squared error (MSE) is defined as the expected value of the square of the difference between the estimator of a variable, $\hat{\theta}$, and the true value of that variable, θ: [8]

$$MSE = E[(\hat{\theta} - \theta)^2]$$

The mean square error takes the residuals (differences between the model predictions and actual data), squares them, and returns the average. In Figure 9 this can be seen as the difference between the purple points and the blue line. The square is used mainly to ensure that negative and positive residuals don't cancel out.

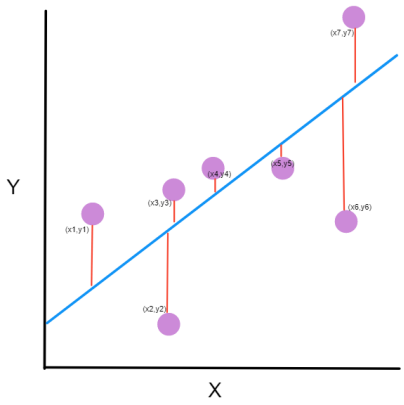

Figure 9: Mean Squared Error [9]

3.3.1 Linear regression

The linear regression model is $y = X\beta + \mu$ where $y = \begin{pmatrix} y1 \\ y2 \end{pmatrix}$, $X = \begin{pmatrix} X1 \\ X2 \end{pmatrix}$ and $\mu = \begin{pmatrix} \mu 1 \\ \mu 2 \end{pmatrix}$

- $y1$ is the column vector of known observations of the dependant variable.
- $y2$ is the column vector of the unknown values of the dependant variable.
- μ is the error term, which adds noise to the linear relationship (it can also be interpreted as the bias).
- X is the weight.

In a linear model the relationship between the dependent variable Y and the independent variable X is plotted. It focuses on the conditional probability distribution (the response given the values of the predictors) [10]. The X determines how much the line rises if its moved one step to the right and the μ determines where the line crosses the vertical axes ($x = 0$). The parameters of the linear regression model are the weight X and the error μ. These parameters are calculated using Python. For X, the slope is calculated (between dependent and independent variable). For μ, the bias is calculated (the intercept).

3.3.2 Decision tree

There are two main types of decision trees: Regression trees and Classification trees. For this model, the regression tree is used, in which the target variable can take continuous values. A decision tree is a symbolic learning technique that uses a hierarchical structure of nodes and ramifications to organize the information coming from a training data set. The tree is used to go from observations of an instance (represented in the branches) to a conclusion about the target value which is represented in the leaves [11]. When building a decision tree for a large number of features, it can lead to a large and complex tree with multiple splits. This can cause problems, as due to the large amount that needs to be memorized by the model; it can lead to overfitting. To reduce this, the maximal depth of the tree can be defined. This is the length of the longest path from the start node (root) to a leaf. For this value a default of value 3 is often used, which was chosen and indeed showed to create a good model (in regards to the loss function). For these regression models three features are used: *temperature*, *day of the week* and *month*.

3.3.3 K-nearest neighbor

The last model that is used is the K-nearest neighbor (kNN) regression. The input for kNN regression is the k-closest training examples in the feature space. The given test instances are compared with the

training set. For a test instance X, its distance d_i to every other instance X_i is calculated. The distances are ranked and the k nearest instances are taken. The output of this function is the property value of X, which is the average of the values of its k nearest neighbors (see Figure 10). For the kNN regression it can be efficient to add weights to the contributions of the neighbors, so that the neighbors which are closest to X contribute more to the average than the ones further away [12]. As parameter for the kNN-regression model, $k = 10$ is chosen. This means that for every node (point), ten neighbors are taken into account.

For determining an appropriate value for the parameter k in the kNN regression, cross-validation was used. That gave the following result, which can be seen in Table 2.

k	MSE
2	34.07
3	31.22
5	29.19
10	27.69
50	29.86
100	31.92

Table 2: k values with the corresponding mean squared error

As can be seen in Table 2, $k = 10$ gives the lowest mean squared error. Therefore, k is set equal to 10.

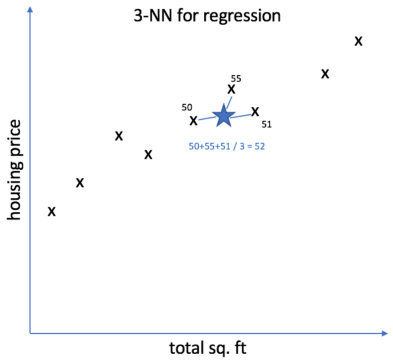

Figure 10: kNN regression, example $K = 3$

4 Evaluation

The accuracy that is used for the classifier is the number of correct assigned instances divided by the total number of instances. The accuracy of the model is calculated for the different n-folds, so that the best n-fold can be obtained to classify if an incident will end up with a shooting or not. Out of the results in Table 3 can be concluded that the 4-fold has the highest accuracy for the SVM model. This accuracy of value 0.9969 is better than the accuracy obtained with chance of 0.9974. So, our model predicts better than in the case that all the values for the instances would be guessed.

n-fold	Classification SVM
3-fold	0.996926
4-fold	0.996939
5-fold	0.996920
6-fold	0.996915

Table 3: Results accuracy n-folds SVM model

As mentioned before, the MSE is used as accuracy for the different regression methods. Table 4 4 shows that the average mean square error of the linear regression model is the lowest , so this is the best regression model to predict how many instances of shootings will occur on an hour of a specific day. Also, from the results of Table 4 can be concluded that the decision tree is the worst regression method for

the prediction. Different n-folds are used to predict which n-fold is the most optimal for every regression method. As can be concluded out of the table:

- The 10-fold is most optimal for the linear regression model.
- The 6-fold is most optimal for the decision tree.
- The 10-fold is most optimal for the kNN model.

n-fold	Linear Regression	Decision Tree	kNN
3-fold	19.30585	36.05716	28.95107
4-fold	19.24623	36.16320	28.29058
5-fold	19.28115	35.82459	28.44317
6-fold	19.29478	35.56366	28.05617
7-fold	19.26682	36.58932	28.28999
8-fold	19.33189	36.08117	27.93442
9-fold	19.28685	35.65972	27.91206
10-fold	19.24387	36.46087	27.69114

Table 4: Results MSE n-folds SVM model

5 Conclusion

Reflecting on the purpose of this research, four predictive models for crimes in Boston are built. These models can be distinguished as one classifier and three regression models. Looking at the Support Vector Machine (classifier), it can be concluded that the model classifies better than the baseline and therefore is useful for predicting if a call becomes a shooting or not. Using these parameters, a better than baseline classifier is constructed.

For regression, three different models are used: the linear regression, decision tree regression and the kNN regression. Looking at the mean squared error measurement for the regression, it can be seen that the linear regression outperforms the other two regression models. The linear regression model is therefore preferred in the task to predict how many instances of shootings will occur on an hour of a specific day. The selection of features has shown great results, therefore it reinforces the findings from the literature review concerning feature selection in regards to crime rates.

6 Discussion

During the data exploration, problems concerning categorical and numerical features were encountered. For example, for choosing linear regression it was required that the features and the target were not correlated. To determine such a correlation, one has to compare the values of a feature with the values of the target. How should that be done when the feature is categorical (e.g. day of the week) and the target is also categorical (1 or 0, shooting or no shooting, respectively)? Some literature showed that categorical data must be one-hot-encoded for regression and other literature showed that it is not possible at all for categorical variables with more than two categories. Therefore, it was difficult to show that the features and the target were uncorrelated.

Furthermore, another requirement for choosing linear regression is that the features are (multivariate) normally distributed. However, this cannot be checked for categorical features. Therefore, it was impossible to show that the model met this requirement. Another unfortunate result is that linearity cannot be checked with categorical variables (which is also a requirement of linear regression).

Since it was impossible to show that these three requirements of linear regression were met, there is room for discussion whether this model is applicable for these specific features and target.

References

[1] Marcus Felson and Erika Poulsen. Simple indicators of crime by time of day. *International Journal of Forecasting*, 19(4):595–601, 2003.

[2] Jari Tiihonen, Pirjo Halonen, Laura Tiihonen, Hannu Kautiainen, Markus Storvik, and James Callaway. The Association of Ambient Temperature and Violent Crime. *Scientific Reports*, 6543(7), 2017.

[3] Dan Carman. Understanding the reality of holiday-related crimes. https://www.hg.org/legal-articles/understanding-the-reality-of-holiday-related-crimes-49947.

[4] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[5] https://www.saedsayad.com/support_vector_machine.htm.

[6] Pau Rodríguez, Miguel Bautista, Jordi Gonzàlez, and Sergio Escalera. Beyond one-hot encoding: lower dimensional target embedding. *Image and Vision Computing*, 75, 05 2018.

[7] David Money Harris and Sarah L. Harris. *Sequential Logic Design*. 2013.

[8] Ismael Castelazo. On the use of the mean squared error as a proficiency index. *Accredication and Quality Assurance*, 17:95–97, 2012.

[9] Moshe Binieli. https://medium.freecodecamp.org/machine-learning-mean-squared-error-regression-line-c7dde9a26b93, 2018.

[10] Bernhard F. Arnold and Peter Stahlecker. Linear Algebra and its Applications. 354:3–20, 2002.

[11] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, and Yong Shi. Credit card churn forecasting by logistic regression and decision tree. 38:15273–15285, 2011.

[12] Yunsheng Song, Jiye Liang, Jing Lu, and Xingwang Zhao. An efficient instance selection algorithmfor k nearest neighbor regression. *Neurocomputing*, 251:26–34, 2017.

Appendices

Attribute name	Description	Data Type
INCIDENT_NUMBER	Unique ID-code of the incident	Categorical
OFFENSE_CODE	Code of the police for the type of incident	Categorical
OFFENSE_CODE_GROUP	Name of the OFFENSE_CODE	Categorical
OFFENSE_DESCRIPTION	Description of the police after the incident	Categorical
DISTRICT	District code for police station in Boston	Categorical
REPORTING_AREA	Area in which incident is reported from	Categorical
SHOOTING	Whether a shooting is occurred at the incident	Categorical
OCCURRED_ON_DATE	Date at which the incident took place	Numerical
YEAR	Year in which the incident took place	Categorical
MONTH	Month in which the incident took place	Categorical
DAY_OF_WEEK	Day of the week in which the incident took place	Categorical
HOUR	Time at which the incident took place	Categorical
UCR_PART	Part I offenses or Part II offense	Categorical
STREET	Street at which the incident took place	Categorical
LAT	Latitude of the location of the incident	Numerical
LONG	Longitude of the location of the incident	Numerical
LOCATION	Coordinates of the location of the incident	Numerical
TEMPERATURE	The outside temperature during the incident	Numerical
WINDCHILL	The windchill during the incident	Numerical
HUMIDITY	The moisture in the air during the incident	Numerical
PRECIPITATION_RATE	The amount of rain on the day of the incident	Numerical
RED_ SOX_PLAY	If Red Sox played on the the day of the incident	Categorical
SERIOUS_OFFENSE	If crime can be qualified as a serious offense	Categorical

Table 5: Attribute explanation

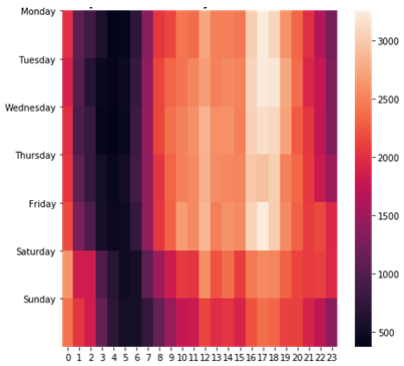

Figure 11: Heatmap of correlation Day and Time with Shootings

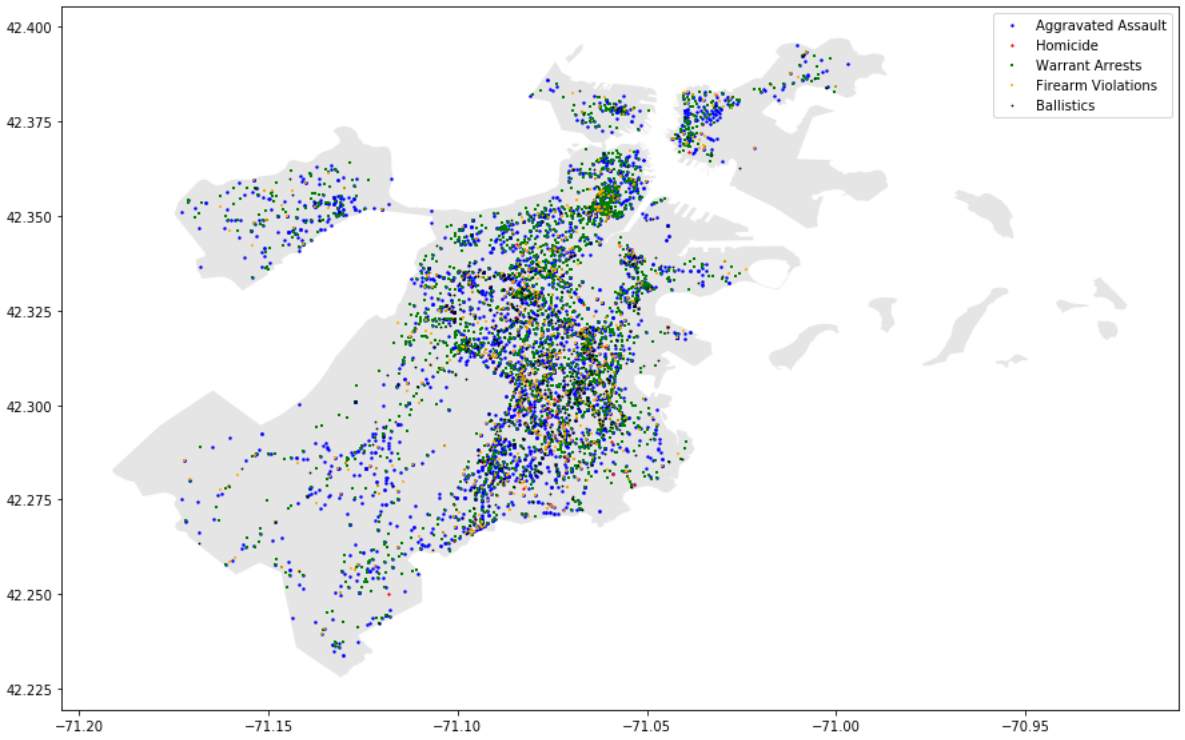

Figure 12: Location of top 5 offenses

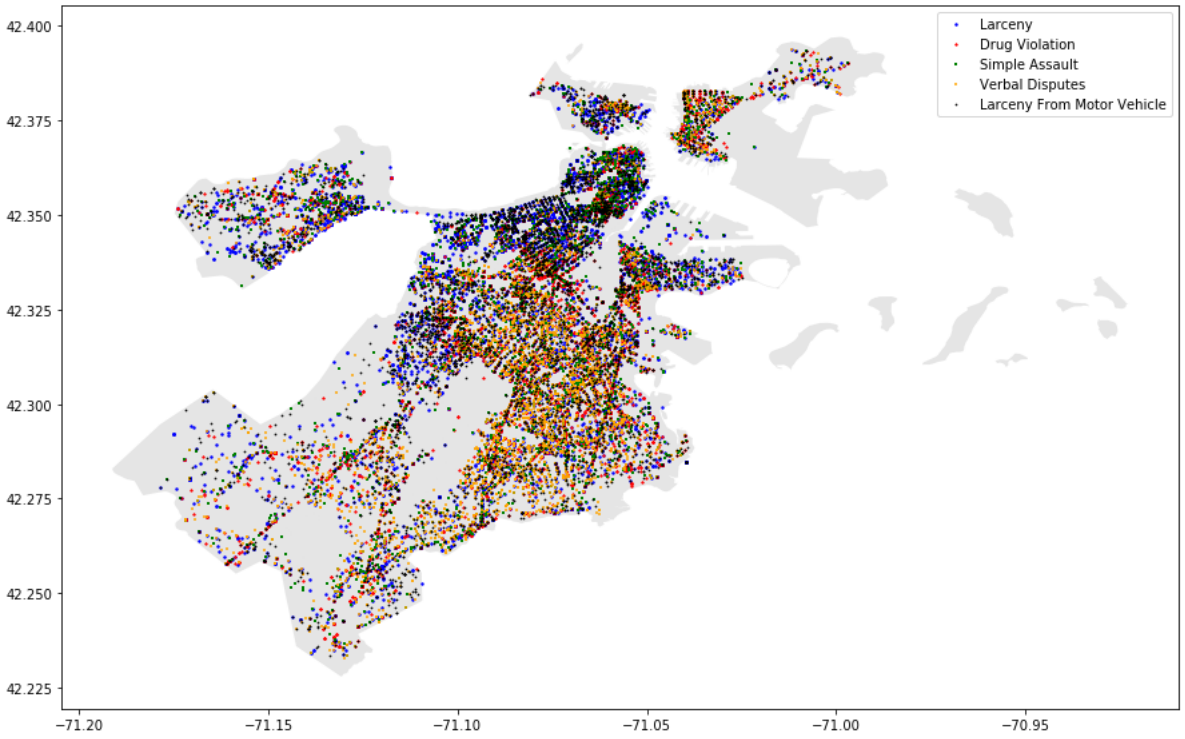

Figure 13: Location of top 5 shooting offenses

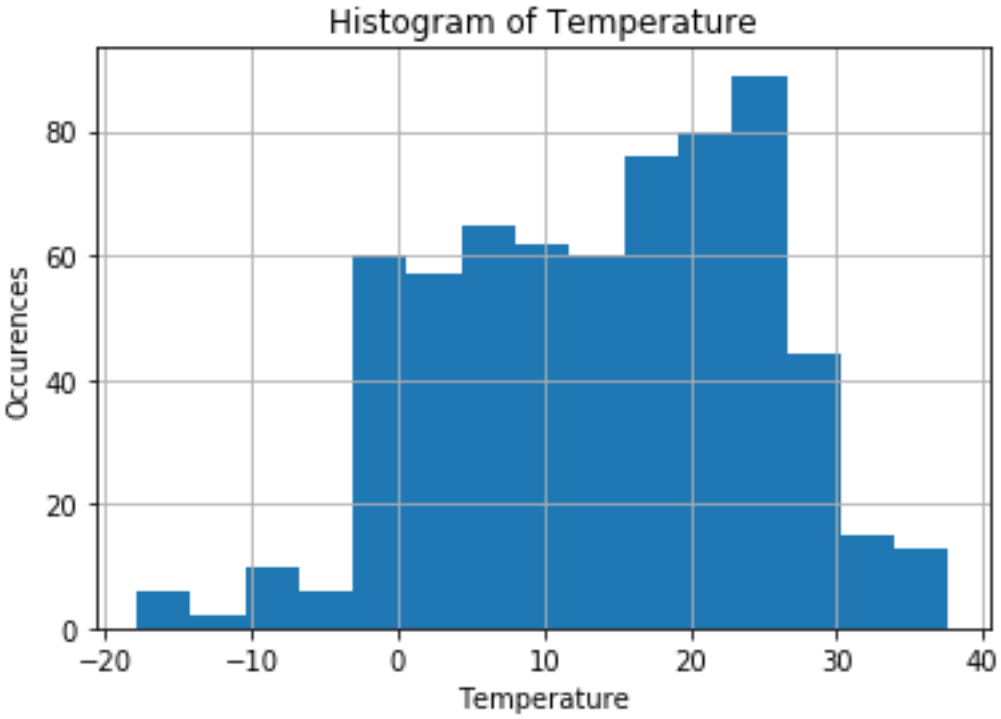

Figure 14: Histogram of temperature