Cornell
Engineering

ORIE 4740 Final Project

Predicting Crime in Boston with Weather Data

Zoe Wang
Thomas Pendock
Thomas Serafin

1. Abstract

The Boston Police Department is leading the effort to combat crimes by keeping a database containing every recorded crime in Boston since 2014. Our project explores avenues for predicting daily crimes rates using temporal and weather data provided by the National Center for Environmental Information (NCEI). We employ linear regressions, GAM models, and decision trees to predict the number of crimes on a given day solely based on temporal as well as weather variables. We discovered that month, temperature, precipitation, and day of the week are the most powerful variables for predicting daily crime rates in Boston. Because our models rely only on such trackable and predictable variables, we believe that our results may provide Boston administrators with some insight on how to manage their resources more efficiently only by looking at the date or their weather app, and inspire others to explore further into the relationship between crime and weather.

2. Introduction

Crime is one of the largest social issues we face: it can influence neighborhoods we choose to live in, schools we choose to attend, and it can be a deciding factor in our everyday lives. The easily obtainable weather and time data could potentially help the Boston Police Department predict daily crimes. We have collected and compiled incident reports from the Boston Police Department and weather data from the NCEI between August 2015 and February 2020 into a table containing the number of crimes and the weather of each day.

3. Basic Statistical Analysis

Before any analysis, the data that had to be cleaned and compiled into a daily counts data set. We performed data cleaning and data extraction on the crime and weather data sets, which in total contain over 1 million data points. After compiling them into a data set of 1,568 rows for each day only containing the number of crimes on each day and the date, we then combined this compiled data set with the daily weather data set that provided us with wind speed, temperature, precipitation etc. We decided to split the date into categorical variables (eg. month, day of the month, day of the week) and continuous variables (time_delta: days since Jan 1st). We removed outliers, any row with 3 standard deviations from the mean, and high leverage points using studentized residuals. Below (Table 3.1) are a couple of sample rows and columns of our table.

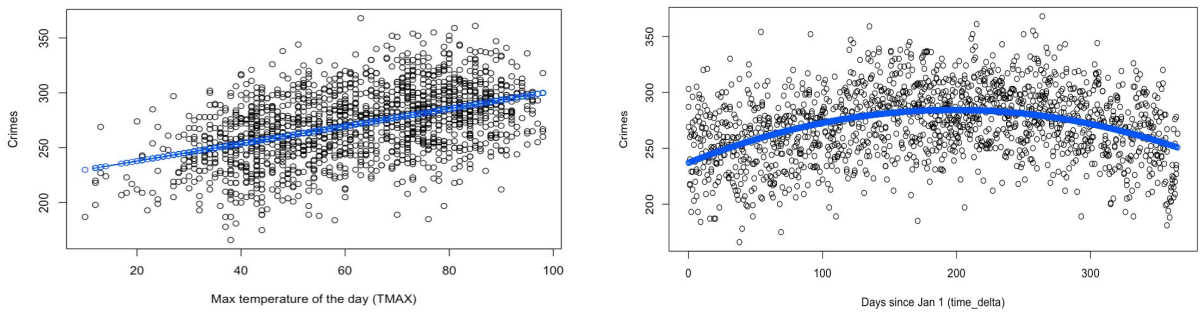
Table 3.1 Sample rows and columns of cleaned data set

count	AWND	PRCP	SNOW	TAVG	TMAX	TMIN	WSF2	WSF5	month_cat	weekday	time_delta	TMIX
250	8.28	0.00	0.0	24	34	18	18.1	23.9	Dec	Tuesday	354	612
289	7.61	1.33	0.0	76	82	69	17.0	19.9	Jul	Wednesday	192	5658

The crime rates are highly dependable on human behavior, therefore it is difficult to explain most of the variance in crime with any predictors. The adjusted R^2's for our models are below 0.5.

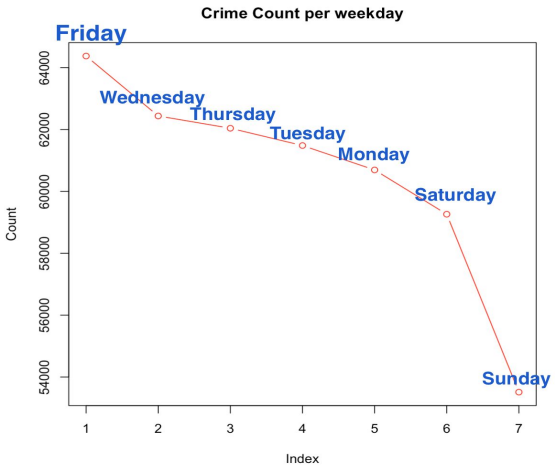
Since our data set is imbalanced in that the days with "normal weather conditions" outnumber the days with special weather conditions (precipitation, snow, WT09, WT04 and WT01) by about four times on average. Therefore we applied the random undersampling method which eliminates observations from the majority class randomly until the data set gets balanced. Specifically, we reduced the number of days without special weather conditions to 300, which is roughly equal to the average number of days with special weather. Figure 3.2 are two sample plots of regression models using only one variable.

Figures 3.2 - Examples of simple regression models only using one variable



In Figure 3.2, the first model indicates the crime rises linearly with temperature, and the second model is a quadratic polynomial with the predictor being time_delta. The regressions show that the number of crimes peaks in summer, around mid June and early July, and drops to around 250 in December and January. Perhaps this is because people tend to stay outside and are more active on warmer days than colder days.

Figures 3.3 - Crime count per weekday



In addition, we also looked into the aggregate number of crimes for each weekday. Figure 3.3 on the left shows the total number of crimes in Boston per weekday from August 2015- February 2020. We

can see that Friday has the highest crime count followed by Wednesday and a sharp decline on Sundays. (For all other basic statistical analysis please see graphs attached in the Appendix. We didn't show them here because of space limit.)

4. Methods

This paper explores a number of different supervised learning techniques to achieve our goal of predicting the number of crimes in a given day based on the weather data. For each approach, we will analyze the model accuracy (MSE, MAE and adjusted R^2, etc) and model interpretability. MAE is the mean absolute error and is used because it is more interpretable. We will start with linear regression models and GAM, and then move to the more complex decision tree models including random forest and boosting. Finally, we apply the best subset selection model to our cleaned data set. Overall, the best subset selection model appears to be the best model among all the models, given its high test accuracy and interpretability. In the conclusion, we will discuss the application as well as limitations of these models and any further improvements on them.

5. Models

5.1 Linear and non-linear models

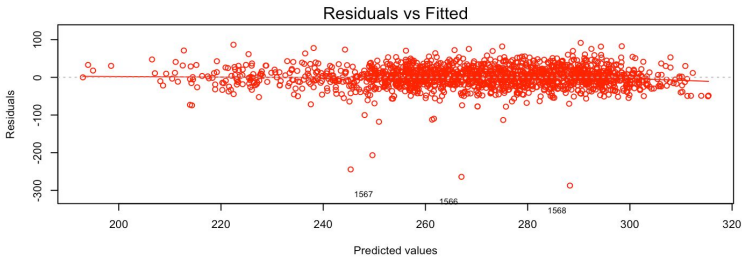
We start at a preliminary level by training a simple linear regression model on all variables using a training/test split of 4:1 to test the model accuracy. Our very first model has an adjusted R^2 of 0.3792 and an MSE of 929. We felt that we could improve upon this by only including 8 variables, which have significant p-values at a confidence level of 0.9, to fit another linear model that may generate better results. The adjusted R^2 and MSE of the new linear model

are 0.40 and 688, respectively. This is a huge improvement increasing our adjusted R^2 and bringing down our MSE.

Next, we extend the standard linear model by implementing GAM models. Using GAMs, we are able to fit different models to individual predictors, therefore modeling non-linear relationships automatically while still examining the impact of each predictor on the outcome individually. In the simple linear regression, the p-values for WT01, WT04 and WT09 (WT01 = Fog, ice fog, or freezing fog; WT04 = Ice pellets, sleet, snow pellets; WT09 = Blowing or drifting snow) are relatively large compare to the p-values of the other predictors, therefore we employ GAM and analysis of variance (ANOVA, using an F-test) to further quantify the extent to which the more complicated models are superior to the simpler linear models.

With the balanced data set, we analyze a few nested models using ANOVA function and sequentially compare the simpler model to the more complex model. The ANOVA table shows that the most complex model, the one that contains all WT01, WT04 and WT09, seems to be far superior to the other models given its p-value is virtually zero. Its MSE by using the validation set approach is 843.24, and the model's training error and test error are similar and show no sign of overfitting. Further, the residual plot of the linear model shows no fitted discernible pattern (Figure 5.1), which means linear relationship is an appropriate assumption.

Figure 5.1 Residual plot of linear regression



5.2 Random Forest and Boosting

We consider the random forest technique to be another appropriate method here because it tends to overfit less and can further reduce the correlation among individual trees. Initially, we trained a random forest model based on all predictors. We randomly sampled 4 predictors out of 8 based on the lowest MSE and found the best split based on them. Compared to the previous result, the MSE of random forest technique decreased to 836 and MAE (Mean absolute error) decreased to 20.94 . In order to further reduce the test error, we select 8 variables with the lowest p-values in the linear regression, while m is still equal to 4, to construct an even better random forest model. The result contained an MSE and MAE of 810 and 20.90, respectively. Figure 5.2 and 5.3 in Appendix show the percent increase in MSE for the two Random Forest models we built. Practically, our random forest model complimented our preliminary assumptions that crime is with weekday, temperature and time of the year. For instance, the crime rates would be higher on a warm Friday summer night as compared to a cold snowy Sunday evening.

In addition to random forest, we also tried the boosting technique. To begin, we added a new column to our data that enhanced our boosting trees accuracy significantly. This column, called TMIX, is TMAX*TMIN. Our boosting model used 1000 subtrees. We iterated over a set of learning rates ranging from 2^{-1} to 2^{-8}, and chose the lambda that resulted in the lowest test error. We then looked at the influence plot and removed the bottom half of the variables that had little influence in the tree model, and then refitted the mode with a subset of the original predictors.

The table below (Table 5.4) shows the relative influence of the variables that were used in the final model. It is clear that the model relies heavily on weekday and all temperature variables. Interestingly, the boosting model, unlike other models discussed in this paper, does not

rely very heavily on the month and time_delta_sq. This model uses a validation set of 25% of the data and has an MSE of 588, an MAE of 18.78, the learning parameter of 0.01457, and an adjusted R^2 of 0.452. We calculated this model's accuracy using Accuracy = 1 - mean(abs(prediction - actual)/actual), which is 93.21%.

Table 5.4-Boosting relative influence

	var	rel.inf
weekday	weekday	40.1585916
TMIX	TMIX	17.8910935
TMAX	TMAX	14.1681033
TAVG	TAVG	6.3602578
PRCP	PRCP	5.8549201
TMIN	TMIN	5.7389738
time_delta	time_delta	4.4304201
month_cat	month_cat	3.3664476
SNOW	SNOW	0.9777965
AWND	AWND	0.7823191
WSF2	WSF2	0.1594326
WSF5	WSF5	0.1116441

5.3 Best Subset Selection

Although boosting produced the lowest test error compared to other models we analyzed so far, its low interpretability could limit the model's application in real life. Therefore, we try a presumably more interpretable model –the best subset selection approach –to predict the number of crimes, which consists of testing all possible combinations of the predictor variables, and then selecting the best model according to adjusted R^2. The best subset selection approach chooses the model with 16 variables (Table 5.3), which produces an MSE of 617 and an adjusted R^2 of 0.428. The accuracy of the best subset model was 92.6%. Overall, this model has relatively high accuracy and also sufficiently high interpretability.

Table 5.3 Regression Analysis of Best Subset Selection

(Intercept)	PRCP	SNOW	TAVG	WT011	WT041	day	month_catAug	
252.46	-7.76	-2.66	0.93	-3.31	-16.43	-1.31	-5.41	
month_catJul	month_catMay	weekdayMonday	weekdaySaturday	weekdaySunday	weekdayThursday	weekdayTuesday	weekdayWednesday	day_sq
-11.36	5.74	-17.12	-21.67	-49.33	-10.90	-13.22	-11.42	0.03

For example, in Table 5.3 we can see that temperature has a positive relationship with crime count, and precipitation and snow are negatively correlated with the crimes. Sunday displays the lowest number of crimes throughout the week and it has the most negative coefficient in comparison to all the other days of the week.

6. Conclusion

In summary, the best subset selection model using 16 predictors proved to be the most superior model because it produces a relatively low test error, shows no sign of overfitting, and is highly interpretable. Although we did get a lower MSE using Boosting, the complex interactions between the independent variables are difficult to understand. Therefore the Boosting method lacks interpretability and is not user-friendly.

Our findings surprised and amazed us; we can use only weather and time to explain just under half of the variance of crime, predict it with about 93% accuracy. We can conclude that month, temperature, precipitation, and day of the week are the most important variables in our models. Crime is generally lower with bad weather and during winter, possibly because people are less active. Crime peaks on Fridays possibly because people are active and partying, and is low on Sundays.

This project by no means is a good way to reduce crime by itself or prove causation, but we hope that this opens up doors for others to explore and extend our ideas. We can truly put our models to the test by comparing our predictions with future weather reports and crime counts that are added to the database everyday. If given more time, we could look into weather and crime relationships for specific types of crimes, or specific districts. In addition, we could improve our models by including more predictors such as a binary variable that indicates a holiday, or synergy variables such as snow and precipitation. We can also try other models like time series models or other functions that have memory. We hope that more powerful and specific models will be developed in the future and will be paired with police department policies to mitigate crime more effectively.

Bibliography

Boston Crime Data Source:

https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

Weather Data Source:

U.S. Department of Commerce: Record of Climatological Observation from National Centers for Environmental Information

Textbook Source:

An Introduction to Statistical Learning (ISLR) by James, Witten, Hastie and Tibshirani.

Online Source:

https://towardsdatascience.com/the-balance-accuracy-vs-interpretability-1b3861408062

Data Dictionary

WSF2: Fastest 2-min wind speed
WSF5: Fastest 5-second wind speed
SNOW: Snowfall
AWND: Average wind speed
PGTM: Peak gust time
WT01: Fog, ice fog, or freezing fog (may include heavy fog)
WT02: Heavy fog or heaving freezing fog (not always distinguished from fog)
WT03: Thunder
WT04: Ice pellets, sleet, snow pellets, or small hail
PRCP: Precipitation
WT05: Hail (may include small hail)
WT06: Glaze or rime
WT08: Smoke or haze
WT09: Blowing or drifting snow

TAVG: Average Temperature
TMIN: Minimum temperature
TMAX: Max temp
Count: Number of Crimes per day
month_num: Numerical value for Month
day: Day of the month
month_cat: Text value for month
weekday: Day of Week
time_delta: 0-364. Number of days after Jan 1st
time_delta_sq: Time_delta^2
day_sq: Days into the month squared
date: [M/DD/YY 12:30] format for date
year: Numerical, year
month: Month of year
day_of_week: Day of week

Appendix

Figure 1 - Crime types vs Count

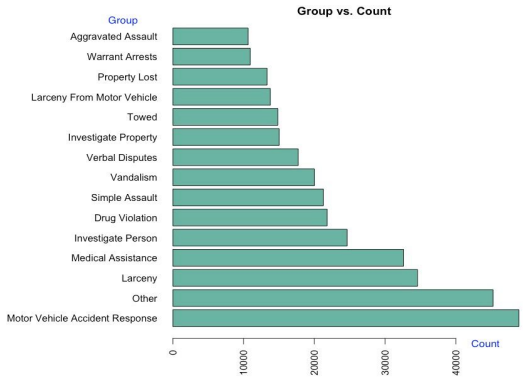


Figure 2 - Crime count vs. hour of the day

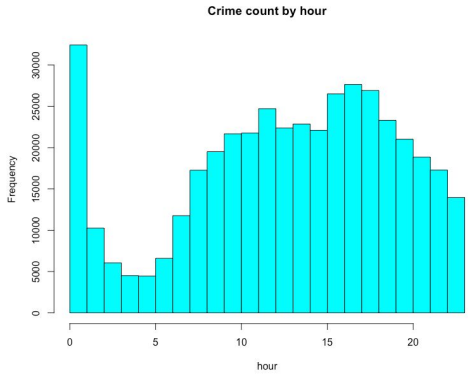


Figure 3 - Crime count vs. district

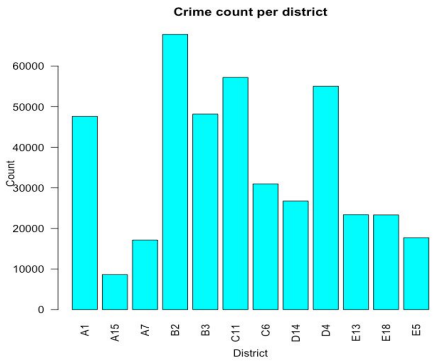


Figure 3.3 - Crime count vs. weekday

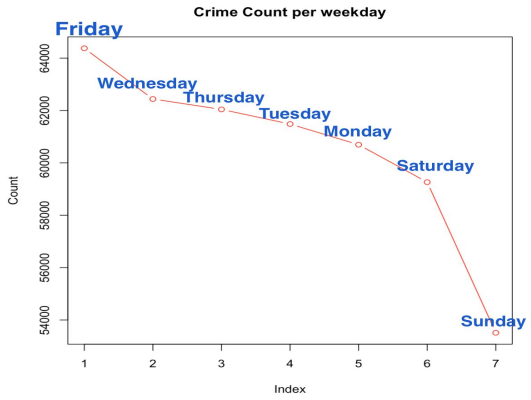


Figure 5.2 - Random forest with all variables

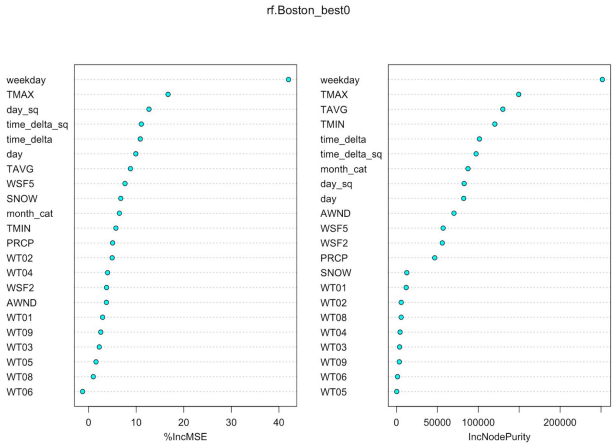


Figure 5.2 - Random forest with 8 variables