

---

# COS 424 Assignment 1: Song Classification

---

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
Katie Hanss

Computer Science, 2017  
khanss@princeton.edu

Will Bertrand

Computer Science, 2017  
wmb@princeton.edu

## Abstract

Every day, the Seattle 9-1-1 center gets over 2,000 emergency calls which require response by EMS, Police or some other body [22]. In order to better understand crime and how the city might prevent it, it is helpful to classify the 9-1-1 calls into the type of crime ultimately reported. In this paper, we extract a number of call-related and city-related features to aid in classifying 9-1-1 calls into their crime type. Via a random forest classifier, classify 9 different crime types and achieved a zero-one loss error of 59.11%. We found that the location of the call, hour of the call, "Near Repeat Phenomenon" and the number of garages, private schools, traffic cameras and parks in the vicinity were most important in aiding this classification.

## 1 Introduction

Every day, the Seattle 9-1-1 center gets over 2,000 emergency calls [22]. When a call comes in, it is assigned to a primary calltaker, who then dispatches the appropriate response body (police, fire department etc) [22]. To close the report, the calltakers also record the type of crime responded to (Burglary, Larsson etc.). This interaction produces a huge amount of data about the time and location of different types of crimes – and Seattle's 9-1-1 calls are publicly available via data.gov [3].

In this project we hope to use Seattle's 911 data along with other metadata such as weather and traffic reports to predict the type of crime that occurs at a given location. We believe that an accurate prediction of crime-type would be very valuable to law enforcement and is therefore important to investigate.

## 2 Related Work

Crime prediction can be a highly valuable tool for police departments. Many police departments across the United States have turned to statistical crime analysis tools and experts to aid in their crime-prevention resources allocation, such as where patrol cars should spend time, where cameras should be in place, and where officers should be ready to respond to. John E. Eck, Spencer Chainey, James G. Cameron, Michael Leitner, and Ronald E. Wilson performed a piece of work for the U.S. Department of Defense in which they found that crime hot spot maps can effectively guide police action when including characteristics of reported crime (such as place, victim, street, or neighborhood).

As well as Walter Perry, Brian McInnis, Carter Price, Susan Smith, and John Hollywood of the Rand Corporation (a non-profit that aims to improve policy and decision making through research and analysis), who discussed in their work Predictive Policing multiple use cases of regression, classification, and clustering models operating on historical crime data in for predicting locations and hours of future crimes. [21]

Work at the University of California, San Diego, done by Junbo Ke, Xinyue Li, and Jiajia Chen, found that they were able to predict crime types to a slightly successful degree (2.7 log-loss) by

054 training a k-nearest-neighbors classifier on a crime dataset, where their features included the address,  
055 date, and time of occurrence of the crime. [19]

056 Also, work on the Near Repeat Phenomenon has shown that it is a useful characteristic in crime  
057 prediction. Tasha Youstin, Matt Nobles, Jeffery Ward, and Carrie Cook found that examining space-  
058 time clustering of crimes in Jacksonville Florida lead to an observation of a pattern of related crimes  
059 occurring. Their results suggested that this was due to Repeat Victimization, a victim of a crime is  
060 likely to be a victim of that crime again, where similar victims (whether they be houses in a similar  
061 neighborhood or stores of a similar type) were also targets of similar crimes over a relatively short  
062 amount of time.[18]

### 064 3 Methods

#### 066 3.1 Data

068 In this section, we will explain what data we used in our project and how we extracted features for  
069 classification via many different data sets available in Seattle.

##### 071 3.1.1 9-1-1 Call Data

073 Seattle has made a huge effort to make government data publicly available. As a result, we had ac-  
074 cess to 1,195,447 9-1-1 calls made from 2010 to the present [3]. Every entry contained information  
075 on a 9-1-1 call including the time the call was made, the longitude and latitude coordinates respon-  
076 ders were dispatched to and the type of crime that occurred (the "Event Clearance Group"). While  
077 the time of the call and longitude and latitude of the incident are available to the 9-1-1 call center as  
078 the call comes in, the crime type is reported back to the 9-1-1 call center by responders after they  
079 have investigated the call. With this in mind, our goal was to see if we could classify 9-1-1 calls into  
080 their crime type with only the information available to the 9-1-1 call center before they dispatched  
081 responders (namely latitude, longitude and hour of the day in which the call occurred). We believe  
082 that this sort of prediction could be useful both in helping the police allocate resources and in giving  
083 responders a better sense of what they might expect before they arrive on scene.

##### 084 3.1.2 Selecting Crime Types

086 After a preliminary investigation of the data, we  
087 found that there were 43 unique crime types.  
088 For the sake of classification, we decided to  
089 limit our analysis to a handful of crime types.  
090 In order to do this, we considered two factors:  
091 (1) crime types must have enough data points to  
092 train a classifier and (2) crime types should be  
093 specific, not catch-all terms such as "suspicious  
094 circumstances." As seen in Figure 2, distur-  
095 bances, suspicious circumstances and traffic re-  
096 lated calls have the most data points. However,  
097 we thought these were broad, "catch all terms,"  
098 which would be difficult to classify. With these  
099 categories eliminated and size in mind we chose  
100 to limit our dataset to the categories listed be-  
101 low. The proportion of each crime type in our  
102 subsetted dataset is shown in Figure 1.

- 102 1. BURGLARY
- 103 2. TRESPASSING
- 104 3. SHOPLIFTING
- 105 4. AUTO THEFT
- 106 5. NARCOTICS COMPLAINTS

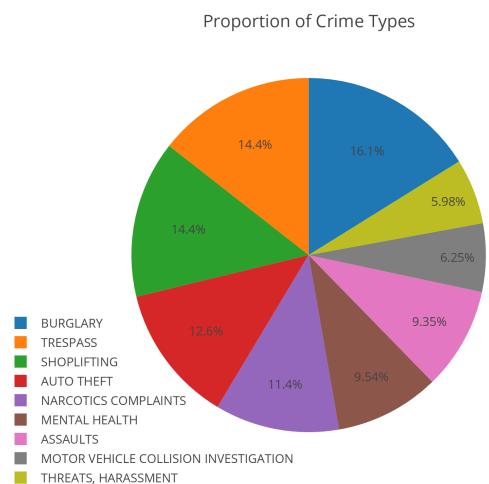


Figure 1: The proportion of each crime type in our subsetted dataset.

6. MENTAL HEALTH
7. ASSAULTS
8. MOTOR VEHICLE COLLISION INVESTIGATION
9. THREATS, HARASSMENT

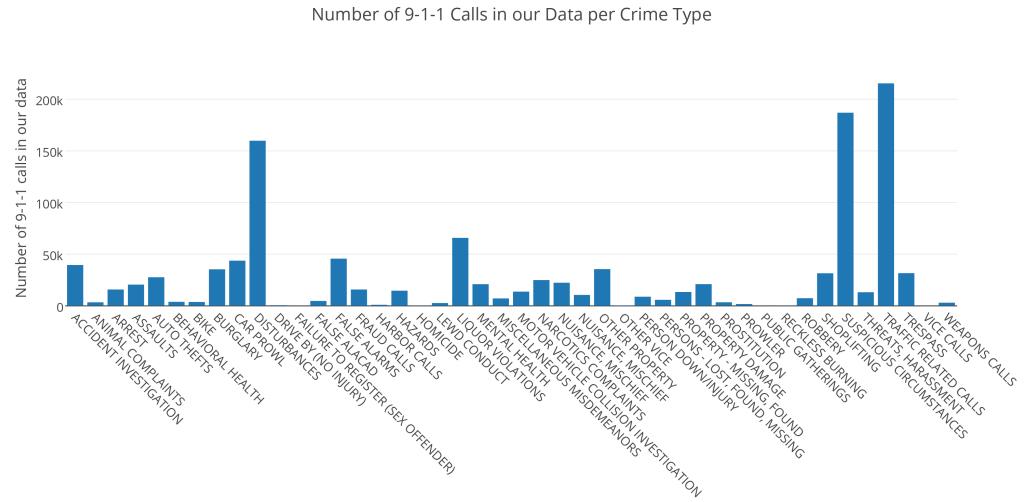


Figure 2: Number of 9-1-1 calls in our data per crime type. Clearly disturbances, suspicious circumstances and traffic related calls are the largest categories. We believe these are "catch-all" terms which encompass a range of incidents.

### 3.1.3 Seattle City Features Data

We quickly decided that classifying crimes based on three features (time of call, latitude and longitude) was insufficient. In order to extract features that would be accessible at the time of the 9-1-1 call, we once again turned to Seattle's open data initiative. Through data.gov, we found a number of Seattle's city features such as the location of baseball fields [4], basketball courts [5], fishing piers [6], garages [7], golf courses [8], movie theaters [9], parks [10], private schools [11], public schools [13], public restrooms [12], skate parks [15], Seattle Police Department precincts [14], track and fields [23], traffic cameras [16] and trails [17]. In order to incorporate these city features as 9-1-1 call features, we searched within a 0.5 mile radius of the crime location (given by it's latitude and longitude) and counted how many parks, baseball fields, private schools etc. were within this distance. Though the 0.5 mile radius was slightly arbitrary, we choose it because we thought that, 0.5 miles was within walking distance of the crime location. In this way, we added an additional 15 city features to each 9-1-1 call's features.

### 3.1.4 Near Repeat Phenomenon

In addition to the city features that we added, we also noticed that there were mini-clusters of a few crimes spread throughout the dataset (Figure 3). For instance, a 4-count burglary cluster can clearly be seen in the bottom left Figure 3. Moreover, not only do these mini-clusters include crimes of the same type, but they also occurred relatively near one another in time (on a scale of minutes up to hours). We suspected that we were not the first people to realize this phenomenon, and after some additional research, found that Eck et al. described the pattern and labeled it the "Near Repeat Phenomenon" [18]. They hypothesize that it occurs when the same assailant commits a string of crimes in the same area or when many people call in the same crime.

Regardless of cause, we decided to attempt to encode the Near Repeat Phenomenon as another feature in our dataset. For all the crime entries in the dataset, we first looked to see if there is a crime

162 occurring within the previous 60 minutes. If so, does that crime occurred within a 0.25 mile radius  
163 from the currently examined crime location? If a crime in our dataset matches both criteria, we  
164 include the integer label of that crime as an additional feature for the 9-1-1 call. If multiple crimes  
165 have occurred fitting these criteria, the most recent crime is used. If no crimes fit these criteria, -1 is  
166 added as the feature. With this additional feature, we generated a total of 19 features per 9-1-1 call.  
167 All of these features would (hypothetically) be available at the time a 9-1-1 call came into the call  
168 center.

### 169 3.2 Data Exploration

170 In order to get a better sense of the features we extracted and how they relate to crime type, we first  
171 performed an initial data exploration on some of our features. We began by plotting crime types on  
172 a map of Seattle using Esris ArcGis platform [1] to see if the location of the crime differentiated the  
173 type of crimes. We then examined the proportion of crime types which occurred per hour via plot.ly's  
174 API [2]. Finally, we examined histograms of the number of parks per crime type via plot.ly's API  
175 [2] to see if the number of parks in the area differentiated crime type.

176 While we did not, ultimately, include it as a feature, we also investigated "hot spots" – a phenomenon  
177 used by police departments to determine routes for patrol cars [21]. Hot spot analysis of our crime  
178 data was performed using Esris ArcGis platform [1] which computed hot spots by identifying grid-  
179 cells with higher than average crime rate using a kernel density function.

### 180 3.3 Crime Classification

181 We used 4 different classifiers from sci-kit learn's library [20] to classify crime type based on our  
182 19-dimensional feature vectors listed below.

- 183 1. Support Vector Classification (SVC) with a linear kernel
- 184 2. SVC with a RBF kernel
- 185 3. K-Nearest Neighbors Classifier (KNN)
- 186 4. Random Forrest Classifier (RF)

187 We used sci-kit learn's StratifiedKFold [20] to perform stratified 10-fold cross validation for each  
188 classifier and evaluated performance via sci-kit learn's zero-one loss error [20].

$$189 \text{zero-one loss} = \sum_{y=1}^N I(y_i, y'_i), \quad I(y_i, y'_i) = 1_{y_i \neq y'_i}$$

190 We initially classified with sci-kit learn's default hyperparameters to see which classification meth-  
191 ods worked best. We then optimized hyperparameters for the most promising classifier (RF). We  
192 examined the importance of each feature in our model. We realized that the radius with which we  
193 searches for parks, public restrooms etc. was also a hyperparameter of sorts. Therefore, we opti-  
194 mized this radius for the most important city-features. Finally we tested our model on our holdout  
195 set and examined the confusion matrix of our classification using sci-kit learn's confusion matrix  
196 function [20].

## 197 4 Results

### 198 4.1 Data Exploration

199 We first wanted to see how crime types were distributed throughout Seattle. As seen in Figure 3,  
200 none of the crime types are localized to a particular region. While the number of crime types and  
201 number of data points make it difficult to discern concrete patterns, we did notice that some crime  
202 types appear to be less common in certain parts of Seattle. For instance, there are few trespassing  
203 incidents (red) on the lower left landmass and narcotics complaints seem more common in the center  
204 of Seattle than on the outskirts. These patterns suggest that the latitude and longitude of a 9-1-1 call  
205 might be useful in classifying the call into a crime type.

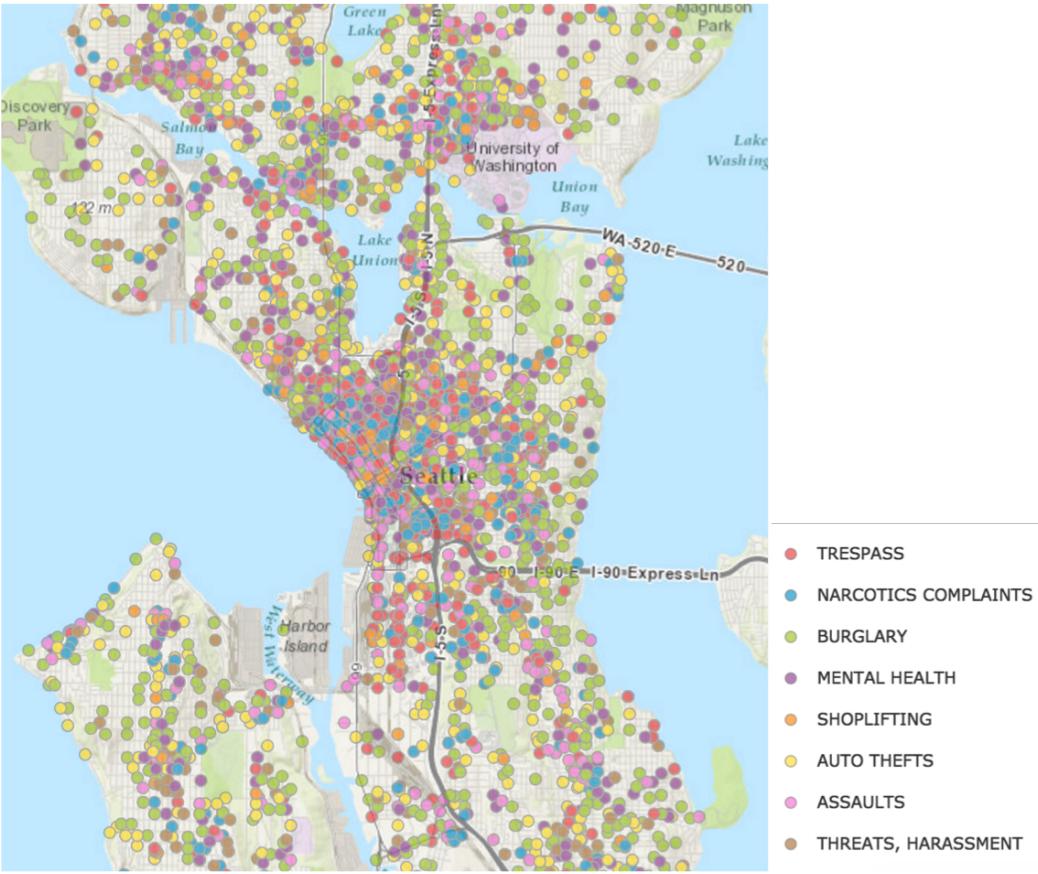


Figure 3: Crime locations for 10k 9-1-1 calls plotted by crime type. None of the crime types we looked at are localized to a particular region. However, there is some separation by location. For instance, it seems that burglary (green) and auto thefts (yellow) are more common on the outskirts of the map than trespassing (red). We also notice that there are mini clusters of crime types. For instance, a 4-count burglary (green) cluster can be seen in the top left.

Next, we wanted to understand how the hour of 9-1-1 calls were distributed across crime types. As seen in Figure 4, different crime types often occur at different times. For instance, assault 9-1-1 calls spike at 3am while many other crime types are less reported. Similarly, most shoplifting calls occur between 1pm and 5pm while other crime types are not reported as often during this period. Not only do these data suggest that the hour a call is placed is important in distinguishing crime type, but we also think that the trends in Figure 4 make sense. While we have no objective evidence to back these claims, we believe that the motor vehicle collision spike at 7pm makes sense given that many people are driving home from work at this hour. Similarly, the shoplifting spike between 1pm and 5pm coincides with the school day ending and the assaults spike at 3am confirms our notion that crimes of this type would be more frequent at this hour.

Finally, we wanted to understand if the city-features we added might help classify 9-1-1 calls into crime types. However, because 15 features are too many to discuss, we will focus on our result from the parks feature. As seen in Figure 5, burglaries often occurred in areas with fewer parks while trespassing often occurred in areas with more parks. While there is significant overlap in the histograms, these data suggest that the parks feature might help distinguish burglary from trespassing. When we added more crime types into the graph, it became difficult to interpret; however, we hypothesize that other crime types are similarly distinguished by parks.

Finally, as seen in Figures 6 and 7, we examined hotspots in the 9-1-1 call data. There was a lot of difference in hotspots when computed over the entire dataset (Figure 6) as compared to the hotspots computed over the 1000 most recent crime events (Figure 7). Further exploration of this

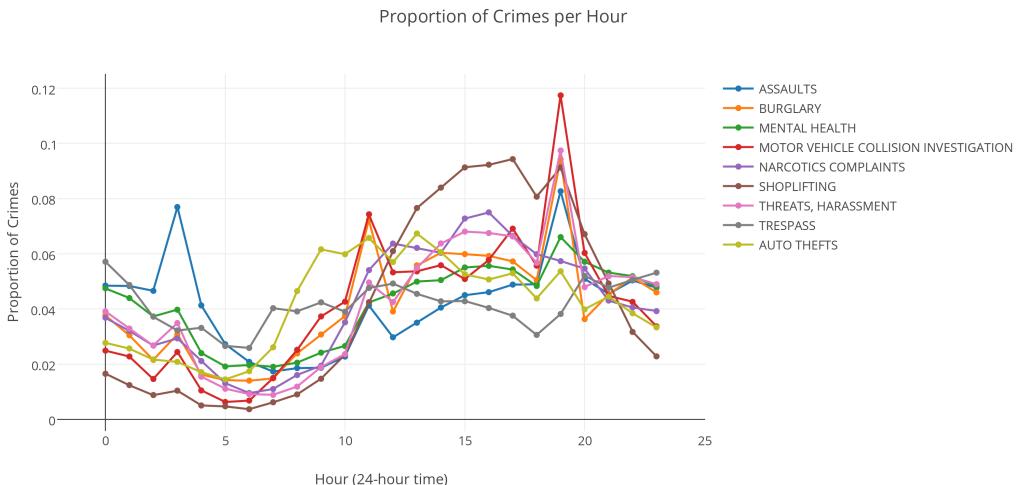


Figure 4: The proportion of each type of crime type which was called in at each hour. The graph suggests that certain crime types are more common at certain hours in the day.

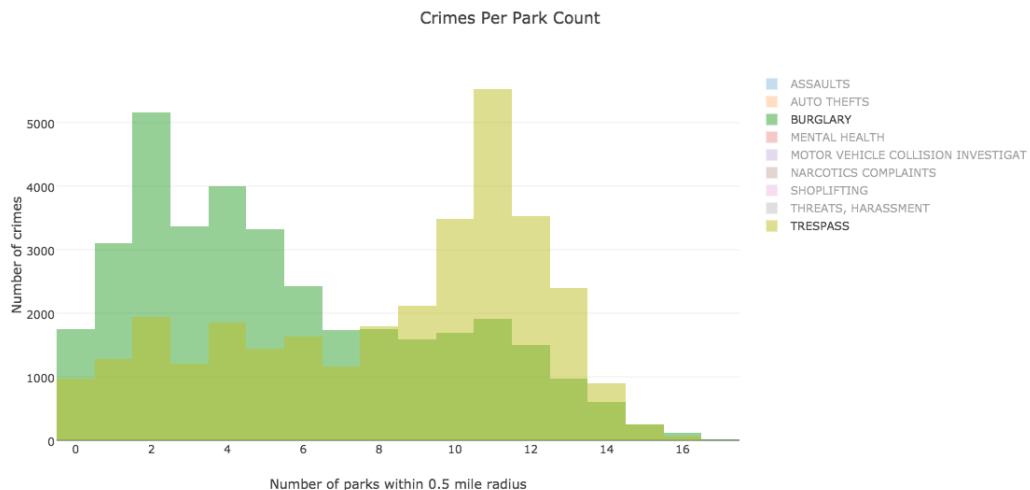
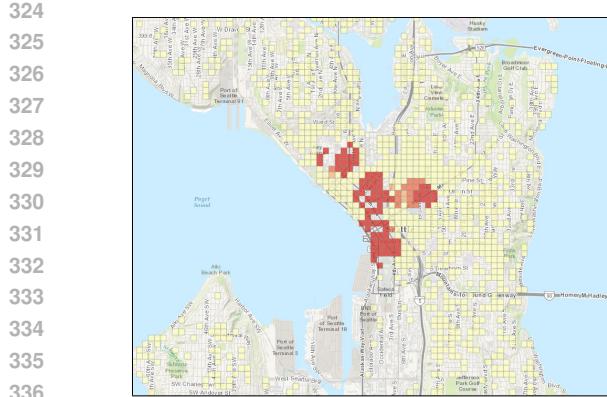


Figure 5: Histograms of the number of parks in burglary and trespassing crimes. The histograms suggest that burglary calls occurred in areas with fewer parks while trespassing calls occurred in areas with more parks.

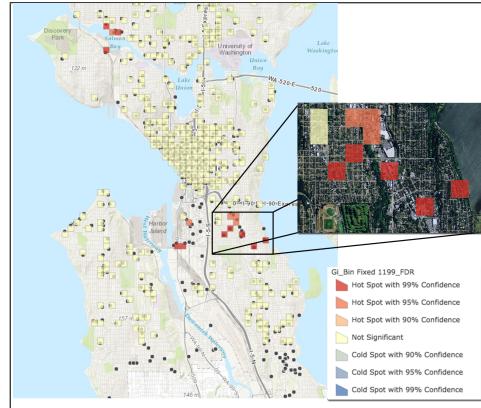
showed that the hot spots changed for any 1000 crime length window extracted from the dataset. In future work or an extension of this project, it would be interesting to use this hot spot analysis to aid our prediction method. If we were to compute hot spots for each of the different crime types and then encode either the nearest hot spot or the distance to hot spots of crime types as an extra feature in our feature vector. We believe that looking at hotspots in a window of recent crimes would be most likely to be effective in our prediction, as crime patterns are likely to change over time. At the least it would be very interesting to see how these hot spots relate to the crimes around them.

## 4.2 Crime Classification

With a better understanding of our features, we attempted to classify 9-1-1 calls into their crime types. As seen in Table 1, we found that, with no optimization of the hyperparameters, all classifiers



324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
Figure 6: Hotspots calculated over the entire dataset. See Figure 6 for the legend



342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
Figure 7: Hotspots calculated for the most recent 1000 crimes in the dataset.

performed better than random. (Here, we use "random" to mean the zero-one loss error we would expect if we implemented the best guessing strategy with know knowledge of features. In this case, since burglaries occurred 16.1% of the time in the dataset, we would expect 83.9% error if we always guessed this crime type.) That being said, the random forest classifier performed better than the rest of the classifiers we tested. Therefore, we chose to focus our efforts on this classifier.

Next we optimized the max depth of trees and number of predictor hyperparameters for these models. As seen in Table 2, we found that using 15 max depth and 160 predictors yielded the lowest zero-one loss error of 59.47%. While using 180 predictors yielded the same error, the time it took to train the model deterred us from increasing predictors any further.

Classifier	zero-one-loss error
Random	83.90
SVC (linear)	77.48
SVC (rbf)	67.20
K Nearest Neighbors	81.23
Random Forest	62.63

Table 1: The zero-one-loss error of 4 different classifiers on the classification of crime type given 19-dimensional feature vectors described in section 3

max depth	num predictors	error
5	50	69.58
10	50	62.30
15	50	59.63
20	50	59.99
30	50	60.83
15	20	59.82
15	40	58.68
15	80	58.52
<b>15</b>	<b>160</b>	<b>57.83</b>
15	180	57.83

Table 2: Optimization of max depth and num predictors hyperparameters in RF. The error column is zero-one loss error.

With a strong model in place, we decided to examine which features were important in classification. As seen in Figure 8, hour was by-in-large the most important feature in the model. Latitude, longitude and the "Nearest Repeat Phenomenon" were also important as were garages, public schools, traffic cameras and parks. In order to improve our model further, we then realized that we could optimize one final parameter: the radius with which we searched for city features. As seen in Figure 9, we found that varying the search radius changed our zero-one loss error slightly. While this method did not achieve the improvements we hoped, we updated the radii for the 4 most important city-features (garages, public schools, traffic cameras and parks) and retrained and tested our random forest classifier. This new feature space achieved a zero-one loss error of 57.88%. Finally, we notices that some features such as golf courses were not very important at all. We iteratively removed the most unimportant feature from our model and found that our zero-one loss error improved to 57.73% if we removed golf courses and track and fields.

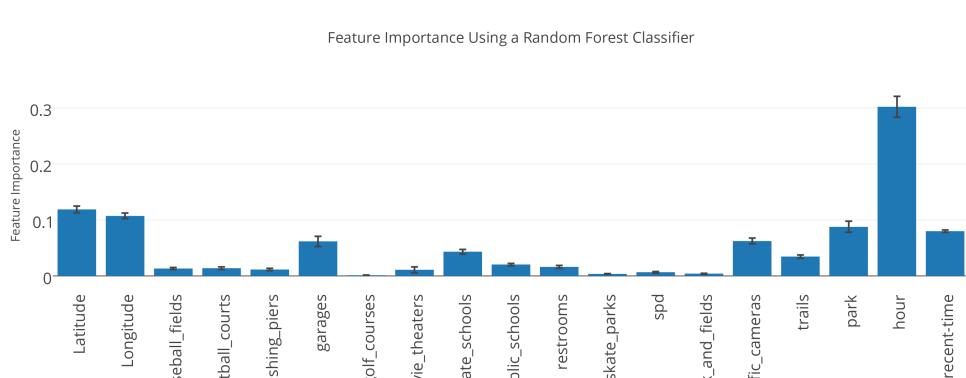


Figure 8: The importance of features in the Random Forest classifier when max depth is 15 and num predictors is 160. "Recent-time" is the "Near Repeat Phenomenon" metric.

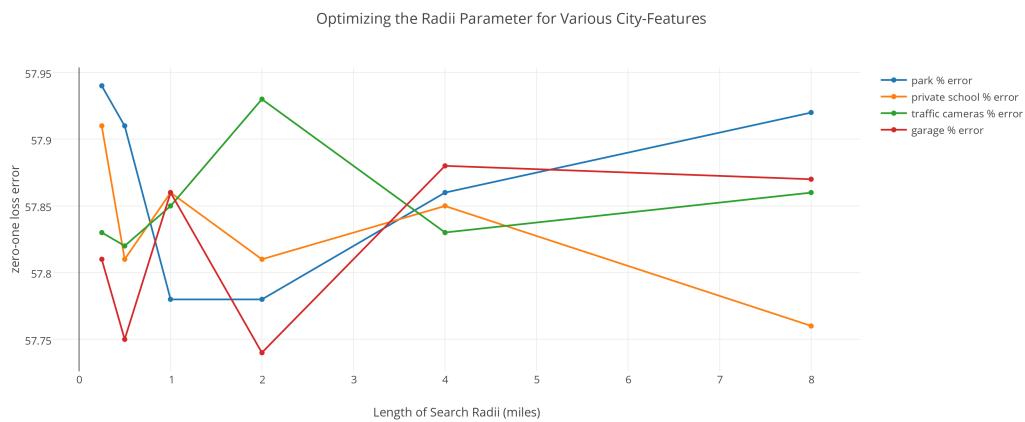


Figure 9: The percent error for different radii of the most important features.

With our model finalized, we applied the random forest classifier to the holdout set. We achieved a zero-one loss error of 59.11% – about 25% better than random. As seen in Figure 10, we were more successful at classifying some crime types than others. Our method is relatively good at classifying burglary, narcotics complaints, shoplifting, trespassing and auto thefts. It is not as successful at classifying assaults, mental health and motor vehicle collisions. It is bad at predicting threats and harassment and worse at predicting threats and harassment and motor vehicle collision investigations. It is interesting to note that the things we are good at predicting (burglary, narcotics complaints, shoplifting, trespassing and auto thefts) are the 5 most common crime types in our dataset while the type we are worst at classifying (threats, harassment) are the least common crime type in the dataset (Figure 1). However, our classification accuracy does not completely follow from representation in the dataset. For instance, we are better at classifying shoplifting than burglary despite the former’s higher representation in the dataset.

## **5 Discussion and Conclusion**

In conclusion, we were able to classify 9-1-1 calls into crime type with a 59.11% zero-one loss error. This is about 25% better than random on this dataset. We found that in this classification, the location of the call (Latitude and Longitude), hour of the call, "Near Repeat Phenomenon" and number of garages, private schools, traffic cameras and parks were most important in aiding this

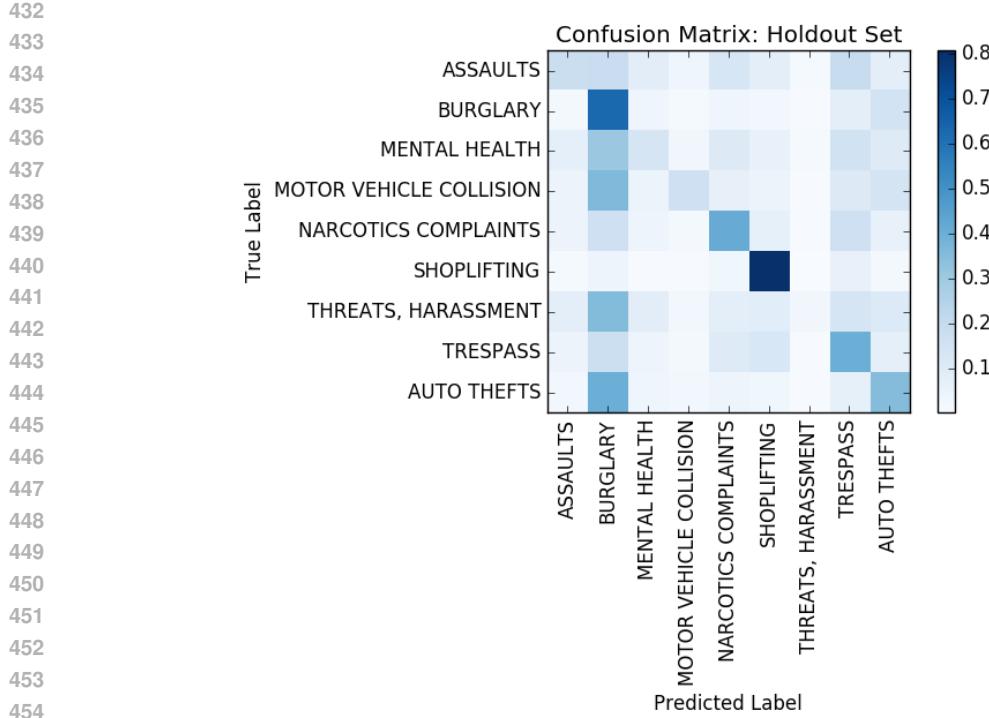


Figure 10: Normalized confusion matrix for the holdout set. Our method is better at predicting burglary, narcotics complaints, shoplifting, trespassing and auto thefts. It is not as successful at classifying assaults, mental health and motor vehicle collisions. It is bad at predicting threats and harassment.

classification. We were better at classifying burglary, narcotics complaints, shoplifting, trespassing and auto thefts and not as successful at classifying assaults, mental health, motor vehicle collisions and threats and harassment.

## Acknowledgments

Thanks to Lizzie for reading this over for us.

## References

- 471 [1] Esri 2011. arcgis desktop: Release 10. redlands, ca: Environmental systems research institute.
- 472 [2] Plot.ly's Python API. "<https://plot.ly/python/>".
- 473 [3] 9-1-1 Calls data.gov. "http://catalog.data.gov/dataset/seattle-police-department-911-incident-response-52779".
- 474 [4] Baseball Field Locations data.gov. "http://catalog.data.gov/dataset/seattle-parks-and-recreation-gis-map-layer-web-services-url-baseball-field-point".
- 475 [5] Basketball Court Locations data.gov. "http://catalog.data.gov/dataset/seattle-parks-and-recreation-gis-map-layer-basketball-court-outline".
- 476 [6] Fishing Pier Locations data.gov. "http://catalog.data.gov/dataset/seattle-parks-and-recreation-gis-map-layer-fishing-piers".
- 477 [7] Garage Locations data.gov. "http://catalog.data.gov/dataset/public-garage-or-parking-lot-includes-e-park".
- 478 [8] Golf Course Locations data.gov. "http://catalog.data.gov/dataset/seattle-parks-and-recreation-gis-map-layer-golf-courses".

- 486 [9] Movie Theater Locations data.gov. "http://catalog.data.gov/dataset/  
487 dea-movietheater-384e4".  
488 [10] Park Locations data.gov. "http://catalog.data.gov/dataset/  
489 seattle-parks-and-recreation-park-addresses-635d4".  
490 [11] Private School Locations data.gov. "http://catalog.data.gov/dataset/  
491 private-schools-6592f".  
492 [12] Public Restroom Locations data.gov. "http://catalog.data.gov/dataset/  
493 seattle-parks-and-recreation-gis-map-layer-public-restroom".  
494 [13] Public School Locations data.gov. "http://catalog.data.gov/dataset/  
495 public-schools-2d727".  
496 [14] Seattle Police Precinct Locations data.gov. "http://catalog.data.gov/dataset/  
497 spd-precincts".  
498 [15] Skate Park Locations data.gov. "http://catalog.data.gov/dataset/  
499 seattle-parks-and-recreation-gis-map-layer-skate-park".  
500 [16] Traffic Camera Locations data.gov. "http://catalog.data.gov/dataset/  
501 seattle-traffic-cameras-cb8ff".  
502 [17] Trail Locations data.gov. "http://catalog.data.gov/dataset/  
503 seattle-parks-and-recreation-gis-map-layer-trails".  
504 [18] John Eck, Spencer Chainey, James Cameron, and R Wilson. Mapping crime: Understanding  
505 hotspots. 2005.  
506 [19] Junbo Ke, Xinyue Li, and Jiajia Chen. San francisco crime classification.  
507 [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pret-  
508 tenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Per-  
509 rot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning  
510 Research*, 12:2825–2830, 2011.  
511 [21] Walt L Perry. *Predictive policing: The role of crime forecasting in law enforcement operations*.  
512 Rand Corporation, 2013.  
513 [22] Seattle.gov. "http://www.seattle.gov/police/work/911center.htm".  
514 [23] Track and Field Locations data.gov. "http://catalog.data.gov/dataset/  
515 seattle-parks-and-recreation-gis-map-layer-track-field".  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539