

# Predicting Crimes in Berkeley

By Weijia Jin, Wenyu Zhang

## Contents

	Pg
1. Introduction	3
2. Data and Analysis	4
2.1 Description of Data	4
2.2 Retrieval of Data	7
2.3 Preliminary Analysis	8
3. K-Nearest Neighbors	12
3.1 All Crime Types	12
3.2 Petty Theft and Burglary	13
3.3 Other Crime Groups	16
4. Logistic Regression	19
4.1 Petty Theft and Burglary	19
4.2 Other Crime Groups	21
5. Linear Discriminant Analysis	23
6. SVM	24
7. Decision Trees	26
7.1 Boosting	29
8. Discussion of Results	30
8.1 Petty Theft and Burglary	30
8.2 Other Crime Groups	33
9. Conclusion	39
9.1 Limitations	39
9.2 Further studies	40

## **1. INTRODUCTION**

*The arrests were routine. Two women were taken into custody after they were discovered peering into cars in a downtown parking garage in Santa Cruz, Calif. One woman was found to have outstanding warrants; the other was carrying illegal drugs.*

*But the presence of the police officers in the garage that Friday afternoon in July was anything but ordinary: They were directed to the parking structure by a computer program that had predicted that car burglaries were especially likely there that day.*

*The program is part of an unusual experiment by the Santa Cruz Police Department in predictive policing — deploying officers in places where crimes are likely to occur in the future.*

The New York Times  
“Sending the Police Before There’s a Crime”  
August 15, 2011

The ability to predict crimes is appealing and desirable not only for the police departments, but more importantly, for the people living in the region. The main purpose of this project is to target the city of Berkeley, California, and to explore ways of predicting crime using past data. This choice is made because the main campus of UC Berkeley is situated in the region, such that the safety of the region is often of particular interest and concern to the students.

Our overarching research question is as follows:

- Using past data of crimes in Berkeley, predict the type of crime most likely to happen at a particular place, on a particular date and time, and under certain weather conditions.

Our initial goal is to build a classifier handling all different types of crime. However, due to the nature of the skewed crime type distribution and the limitations of our models, error rate is relatively high when we attempt to classify all crime types. This will be illustrated in the first algorithm introduced in this report, that is, k-nearest neighbors. Therefore, the report will mainly address the following two goals:

- Build a classifier for the top two crimes (petty theft and burglary);
- Build a classifier for the other crime types grouped into three broader categories (Part 1 offenses: violent, Part 1 offenses: property, Part 2 offenses).

We tackle the problem using a variety of models, such as k-nearest neighbors, support vector machine, logistic regression, and trees. The efficiency of the models are assessed based on the accuracy rate of classification.

## **2. DATA AND ANALYSIS**

### **2.1. DESCRIPTION OF DATA**

The complete dataset that we have describes the crimes that happened in Berkeley between October 1, 2012 and March 29, 2013. Due to the large number of crimes and the continuous nature of crimes across city boundaries, we focused on analyzing crimes within a 2 mile radius of the center of Berkeley, which is approximately taken to be the location as pinpointed by Google map. After removing erroneous and incomplete entries, we have a total of 4485 entries in the dataset. Of which, 596 entries for crimes between March 1 and March 29, 2013 are set aside as the test set. The remaining 3889 entries are used as the training set.

Fig 2.1.1 shows the spatial coverage of the dataset. The red pinpoint marks the center of Berkeley, and the blue pinpoints marks the Berkeley Police Department and UCPD which provided the crime information.

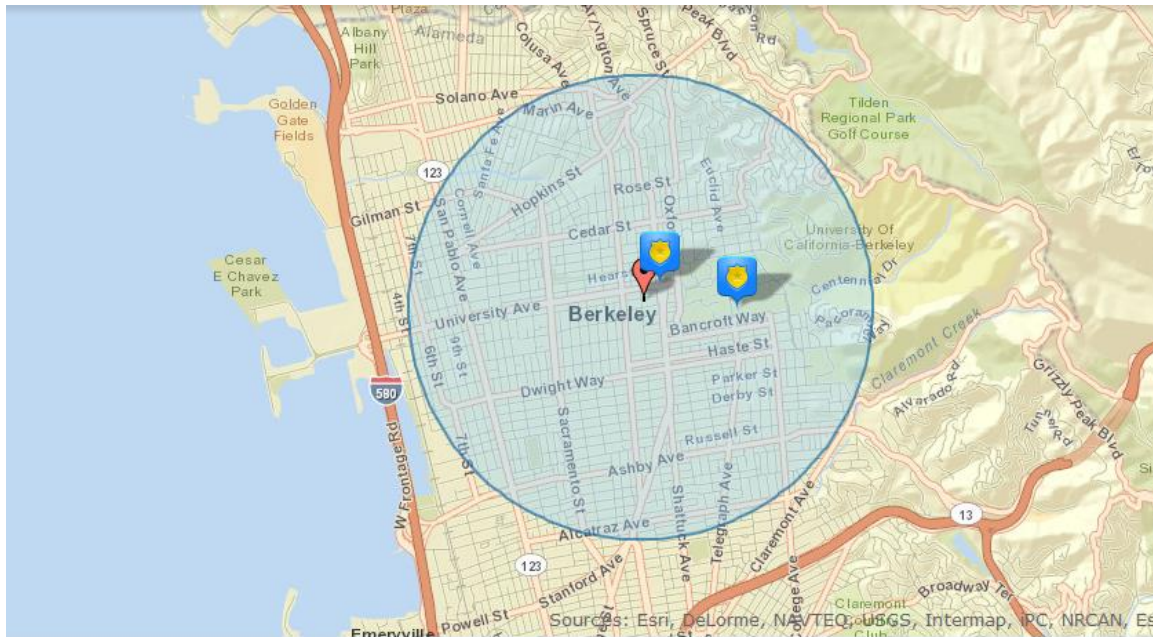


Fig 2.1.1 Data coverage

Fig 2.1.2 shows a portion of the unprocessed dataset. Each entry consists of 8 components, i.e. a brief description of the crime, date and time at which the crime happened, longitude and latitude coordinates specifying the location of the crime, and the corresponding temperature, visibility and weather conditions.

```
> update.data[1:20,]
  Description      Date      Time Latitude Longitude Temperature Visibility Conditions
1  BURGLARY AUTO 12/31/12  9:00:00 PM 37.86159 -122.2644      44.1      10.0 Scattered clouds
2  THEFT MISC. (UNDER $950) 12/31/12  8:00:00 PM 37.85808 -122.2692      46.0      10.0 Scattered clouds
3  DOMESTIC VIOLENCE 12/31/12  8:00:00 PM 37.85331 -122.2787      46.0      10.0 Scattered clouds
4  BURGLARY AUTO 12/31/12  7:00:00 PM 37.85545 -122.2635      46.4      10.0 Scattered clouds
5  THEFT FELONY (OVER $950) 12/31/12  4:35:00 PM 37.85839 -122.2532      50.0      10.0 Scattered clouds
6  NARCOTICS 12/31/12  4:04:00 PM 37.86988 -122.2705      52.0      10.0 Mostly cloudy
7  THEFT FELONY (OVER $950) 12/31/12  4:00:00 PM 37.87209 -122.2896      52.0      10.0 Mostly cloudy
8  THEFT MISC. (UNDER $950) 12/31/12  4:00:00 PM 37.86652 -122.2426      52.0      10.0 Mostly cloudy
9  THEFT FELONY (OVER $950) 12/31/12  3:40:00 PM 37.87647 -122.2689      52.0      10.0 Mostly cloudy
10 BURGLARY RESIDENTIAL 12/31/12 12:30:00 PM 37.85778 -122.2730      50.0      10.0 Scattered clouds
11 THEFT MISC. (UNDER $950) 12/31/12 12:27:00 PM 37.85839 -122.2532      50.0      10.0 Scattered clouds
12 ALCOHOL OFFENSE 12/31/12 12:05:00 PM 37.87104 -122.2683      50.0      10.0 Scattered clouds
13 ASSAULT/BATTERY MISC. 12/31/12 11:45:00 AM 37.87214 -122.2683      50.0      10.0 Scattered clouds
14 THEFT FROM AUTO 12/31/12 12:00:00 AM 37.85305 -122.2796      42.1      10.0 Scattered clouds
15 ALCOHOL OFFENSE 12/30/12 10:45:00 PM 37.85331 -122.2787      43.0      10.0 Clear
16 ROBBERY 12/30/12  9:45:00 PM 37.87803 -122.2714      48.0      10.0 Clear
17 ASSAULT/BATTERY MISC. 12/30/12  7:47:00 PM 37.84861 -122.2796      48.9      10.0 Clear
18 BURGLARY AUTO 12/30/12  6:35:00 PM 37.86980 -122.2862      50.0      10.0 Clear
19 VEHICLE STOLEN 12/30/12  6:00:00 PM 37.85854 -122.2666      50.0      10.0 Partly cloudy
20 BURGLARY AUTO 12/30/12  6:00:00 PM 37.87348 -122.2709      50.0      10.0 Partly cloudy
```

Fig 2.1.2 Unprocessed dataset

In the context of the classification problem we are addressing, the type of crime derived from the Description column will be the response variable, and the predictor features will be extracted from the remaining 7 columns of the unprocessed dataset.

Fig 2.1.4 shows the processed dataset. After keyword matching, 17 categories of crime are identified in the Description column, namely:

- Arson
- Assault
- Bike theft
- Burglary
- Disturbance
- Drugs / Alcohol Violations
- DUI (Driving Under Influence)
- Fraud
- Grand theft
- Homicide
- Motor Vehicle Theft
- Petty Theft
- Robbery
- Sex Crimes
- Vandalism
- Vehicle break-in
- Weapons

The predictor features extracted are the temperature (in degrees Fahrenheit), visibility (in miles) and weather conditions when the crime happened, the day of the week and the time of the crime, an indicator of whether the crime happened in UC Berkeley campus, the x and y coordinates (in km) mapping the location of the crime with respect to the center of Berkeley as (0,0), and the distance of the scene of crime from the nearest Bart station (in km). All features are continuous numerical variables, except for the weather conditions,

day of the week and campus indicator which are categorical/nominal variables. The Conditions variable has 12 categories, namely:

- Clear
- Fog
- Haze
- Heavy rain
- Light rain
- Mostly cloudy
- Overcast
- Partly cloudy
- Patches of fog
- Rain
- Scattered clouds
- Shallow fog

The DayofWeek variable has 7 categories, with category 1 being Monday and category 7 being Sunday. The IndCampus variable is binary, and has value 1 if the crime happened on the main UC Berkeley campus and 0 otherwise. The red rectangle in Fig 2.1.3 below is taken to be the boundary of the main campus.

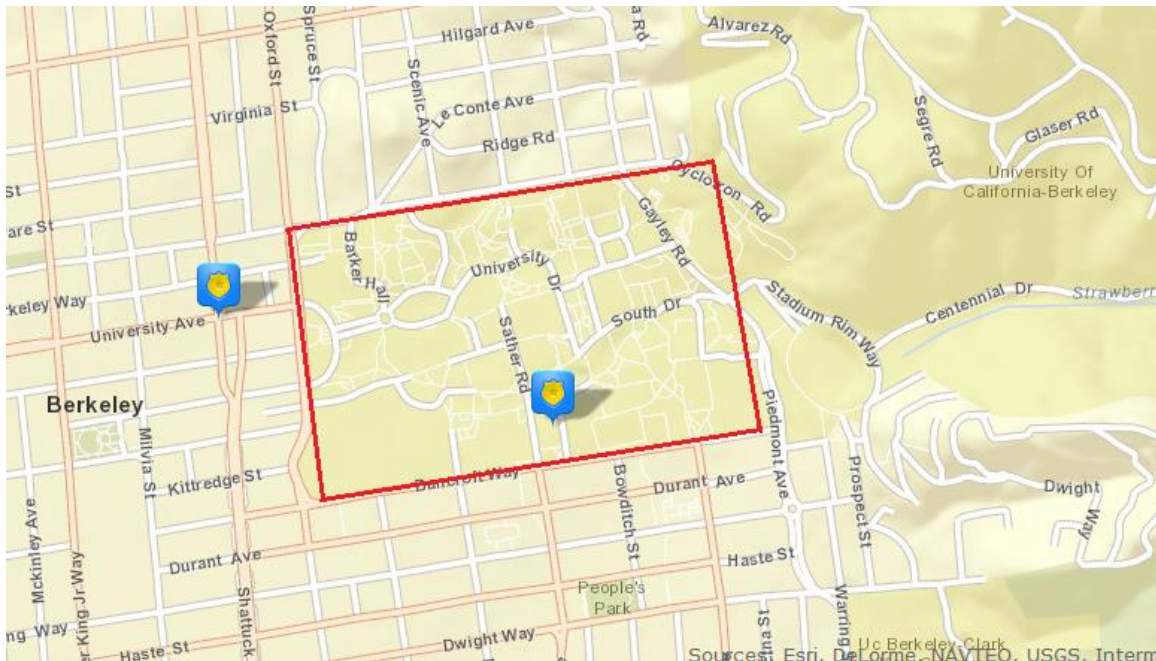


Fig 2.1.3 Boundary of main UC Berkeley campus



```

> TrainingSet[1:20,]
  Description Temperature visibility Conditions dayofweek Time IndCampus x_coord y_coord minBart
1    vandalism      63.0         10    Light Rain         3 22.08333      0 1.73968429 -0.25335772 1.91719449
2    Burglary      63.0         10    Light Rain         3 22.00000      0 0.07795672 1.45750213 0.80009040
3    Burglary      63.0         10    Light Rain         3 21.50000      0 -0.07267749 0.89091089 0.60714237
4 vehicle Break-in  63.0         10    Light Rain         3 21.00000      0 -1.83013520 0.59541882 0.85213865
5    vandalism      63.0         10    Light Rain         3 20.00000      0 0.77024056 -1.09647817 0.90769147
6 vehicle Break-in  63.0         10    Light Rain         3 20.00000      0 0.69536339 0.81872876 0.73285918
7 Motor Vehicle Theft 63.0         10    Light Rain         3 20.00000      0 -0.98141000 2.16036436 0.95514427
8 Motor Vehicle Theft 63.0         10    Light Rain         3 19.00000      0 0.63535607 -0.40005013 0.56015468
9      Fraud      63.0         10    Light Rain         3 19.00000      0 1.83039916 1.52303161 2.18333095
10 weapons      63.0         10    Light Rain         3 18.95000      0 -0.39207127 -1.48721372 0.61496729
11 Motor Vehicle Theft 64.0         7 Mostly cloudy      3 18.00000      0 0.34552601 0.73457173 0.41588224
12    Petty Theft  64.0         7 Mostly cloudy      3 18.00000      0 0.28877422 -1.09570202 0.51233750
13    vandalism  66.0         10 Mostly cloudy      3 17.08333      0 0.29519729 -0.68179258 0.38178261
14    weapons  66.0         10 Mostly cloudy      3 16.95000      0 0.21812044 -0.05444143 0.03764958
15      Fraud  64.9         10    overcast      3 14.41667      0 1.39319043 -0.93969600 1.55956870
16    Burglary  64.9         10    overcast      3 14.00000      0 -2.40434012 -0.57213369 1.66449604
17 Motor Vehicle Theft 64.9         10    overcast      3 13.96667      0 -2.38797449 -0.22419664 1.59324681
18    Burglary  64.9         10    overcast      3 13.50000      0 -0.49220079 -2.23309210 0.72685965
19    Assault  64.0         10    overcast      3 13.33333      0 -0.77006458 -0.23073848 0.60920357
20    Assault  64.0         10    overcast      3 12.80000      0 -0.62699728 -2.34330505 0.90697639

```

Fig 2.1.4 Processed dataset

The following subsections will provide more details on how the predictor features are extracted and selected from the unprocessed data.

## 2.2 RETRIEVAL OF DATA

Raw data is extracted from the following two sources:

- <http://www.crimemapping.com/>
- <http://www.wunderground.com/>

Crimemapping.com outputs a list of crimes within a specified timeframe and location boundary. Fig 2.2.1 shows a sample screenshot. The crime description, and date and time of crime are obtained directly. The location descriptor is further treated since R is unable to process the street addresses as provided on this website. <http://www.gpsvisualizer.com/geocoder/> is used to get the longitude and latitude coordinates corresponding to each street address.









Crime report for 3/1/2013 - 3/7/2013					
61 crimes found.					
✳ Click a crime to "Map It."					
Type:	Description:	Case #:	Location:	Agency:	Date:
	VANDALISM	13013039	WARRING ST / BANCROFT WAY	Berkeley Police	3/7/2013 11:07 PM
	484BR - PETTY THEFT -BIKE REPORT	13-01074	HEARST MINING BUILDING	UCPD Berkeley	3/7/2013 10:26 PM
	DISTURBANCE	13013036	1700 BLOCK FRANCISCO ST	Berkeley Police	3/7/2013 10:02 PM
	ARSON	13013014	1600 BLOCK EDITH ST	Berkeley Police	3/7/2013 07:10 PM
	BURGLARY RESIDENTIAL	13012976	1300 BLOCK HEARST AVE	Berkeley Police	3/7/2013 04:00 PM
	BURGLARY RESIDENTIAL	13012977	1300 BLOCK DELAWARE ST	Berkeley Police	3/7/2013 03:45 PM
	THEFT MISD. (UNDER \$950)	13012972	1800 BLOCK UNIVERSITY AVE	Berkeley Police	3/7/2013 03:35 PM
	THEFT MISD. (UNDER \$950)	13012948	2500 BLOCK DURANT AVE	Berkeley Police	3/7/2013 12:30 PM

Fig 2.2.1 crimemapping.com output

The temperature, visibility and weather conditions at the time of each crime is obtained from wunderground.com. A program is written to obtain the hourly weather information.

From the data obtained through the above methods, and in particular the date column, we extracted the day of the week on which the crime happened. Since our data spans half a year, we dropped the feature for month as that would make it difficult to handle entries for months not covered in the timespan. In addition, using the longitude and latitude coordinates, we determined whether the crime happened within the boundaries of the main UC Berkeley campus, the x and y coordinates with respect to the center of Berkeley, and the distances from each of the three Bart stations (i.e. Downtown Berkeley, North Berkeley, and Ashby) and police departments (Berkeley PD, and UCPD) in/near the spatial region covered. The minimum distance from a Bart station and from a police department is also calculated. A portion of the results can be found in Fig 2.2.2.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Descriptio	Latitude	Longitude	Temperat	Visibility	Condition	DayofWee	Time	IndCampu	x_coord	y_coord	Downtow	North	Ashby	BerkPD	UCPD	minBart	minPD
2	Vandalism	37.86771	-122.251	63	10	Light Rain	3	22.08333	0	1.739684	-0.25336	1.917194	3.687319	2.295508	2.489046	2.489046	1.917194	2.489046
3	Burglary	37.88314	-122.27	63	10	Light Rain	3	22	0	0.077957	1.457502	0.80009	1.660179	1.797523	0.851271	0.851271	0.80009	0.851271
4	Burglary	37.87803	-122.271	63	10	Light Rain	3	21.5	0	-0.07268	0.890911	0.607142	1.397561	1.502899	0.494863	0.494863	0.607142	0.494863
5	Vehicle Br	37.87536	-122.291	63	10	Light Rain	3	21	0	-1.83014	0.595419	2.623426	0.852139	2.74219	2.054184	2.054184	0.852139	2.054184
6	Vandalism	37.8601	-122.262	63	10	Light Rain	3	20	0	0.770241	-1.09648	0.907691	2.576987	0.990334	1.398204	1.398204	0.907691	1.398204
7	Vehicle Br	37.87738	-122.263	63	10	Light Rain	3	20	0	0.695363	0.818729	0.732859	2.356614	1.659477	1.235482	1.235482	0.732859	1.235482
8	Motor Vel	37.88948	-122.282	63	10	Light Rain	3	20	0	-0.98141	2.160364	1.917582	0.955144	2.545478	1.488613	1.488613	0.955144	1.488613
9	Motor Vel	37.86638	-122.263	63	10	Light Rain	3	19	0	0.635356	-0.40005	0.560155	2.31538	1.078297	1.112397	1.112397	0.560155	1.112397
10	Fraud	37.88373	-122.25	63	10	Light Rain	3	19	0	1.830399	1.523032	2.183331	3.828543	2.890559	2.717933	2.717933	2.183331	2.717933
11	Weapons	37.85658	-122.275	63	10	Light Rain	3	18.95	0	-0.39207	-1.48721	1.124616	1.417843	0.614967	0.848605	0.848605	0.614967	0.848605
12	Motor Vel	37.87662	-122.267	64	7	Mostly Clc	3	18	0	0.345526	0.734572	0.415882	1.911754	1.453482	0.810756	0.810756	0.415882	0.810756
13	Petty The	37.86011	-122.267	64	7	Mostly Clc	3	18	0	0.288774	-1.0957	0.599598	2.008969	0.512338	0.890848	0.890848	0.512338	0.890848
14	Vandalism	37.86384	-122.267	66	10	Mostly Clc	3	17.08333	0	0.295197	-0.68179	0.381783	1.936393	0.711284	0.762834	0.762834	0.381783	0.762834
15	Weapons	37.8695	-122.268	66	10	Mostly Clc	3	16.95	0	0.21812	-0.05444	0.03765	1.763356	1.004314	0.561305	0.561305	0.03765	0.561305
16	Fraud	37.86152	-122.255	64.9	10	Overcast	3	14.41667	0	1.39319	-0.9397	1.559569	3.313638	1.757992	2.11243	2.11243	1.559569	2.11243
17	Burglary	37.86483	-122.298	64.9	10	Overcast	3	14	0	-2.40434	-0.57213	3.345767	1.664496	3.201057	2.778595	2.778595	1.664496	2.778595
18	Motor Vel	37.86797	-122.298	64.9	10	Overcast	3	13.96667	0	-2.38797	-0.2242	3.312754	1.593247	3.227598	2.741968	2.741968	1.593247	2.741968
19	Burglary	37.84985	-122.276	64.9	10	Overcast	3	13.5	0	-0.4922	-2.23309	1.511895	1.663944	0.72686	1.268601	1.268601	0.72686	1.268601
20	Assault	37.86791	-122.279	64	10	Overcast	3	13.33333	0	-0.77006	-0.23074	1.270058	0.609204	1.380016	0.706595	0.706595	0.609204	0.706595

Fig 2.2.2 Complete set of extracted features

From this complete set of 17 extracted features, we selected a subset of 9 features as seen in Fig 2.1.4 in the previous subsection. The next subsection provides more details on this choice.

## 2.3 PRELIMINARY ANALYSIS

Fig 2.3.1 shows the counts of each type of crime in the training set. Assuming that the pattern of crime in terms of the number of occurrences remains approximately the same throughout time, the plot reflects the general distribution of crimes across type. As observed from the plot, petty thefts and burglary are the top two crimes, and their occurrences are much more frequent than the other types of crime. Therefore besides classifying the entire dataset, we will also explore handling two sub-problems:

- Singling out the entries for petty thefts and burglary, and build a classifier
- Build a classifier for the other crime types or a subset of the other crime types.



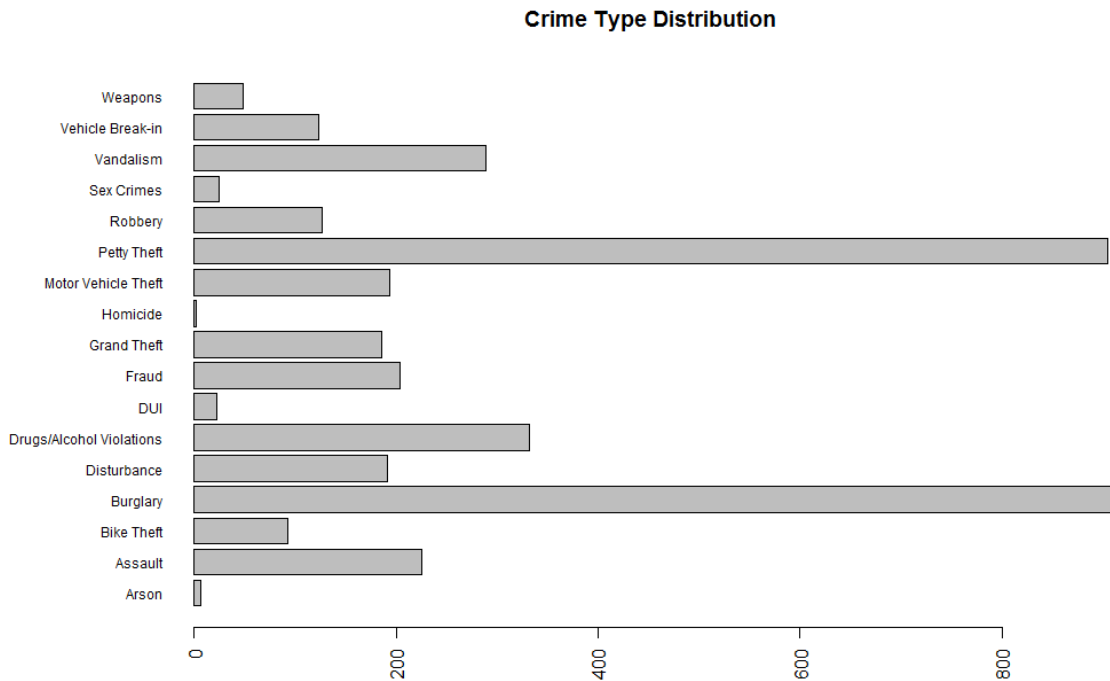


Fig 2.3.1 Crime type distribution

Looking at the complete set of predictor features in Fig 2.2.2 of the previous subsection, we want to select a subset of these to have a higher priority in testing our models. Since some features are derived directly from others, they have high correlations as seen in Fig 2.3.2. As a preliminary step, we select:

- Temperature
- Visibility
- Conditions
- DayofWeek
- Time
- IndCampus
- X\_coord
- Y\_coord
- minBart

We also test with the addition of the excluded features if that improves the accuracy rates. Depending on the classification technique used, this subset may also be further reduced.

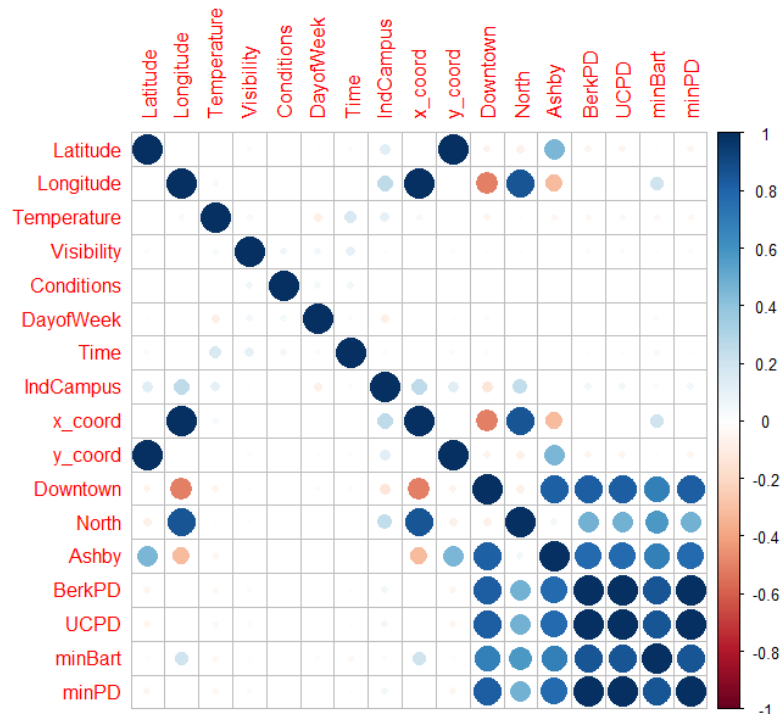


Fig 2.3.2 Correlation plot of the predictor features

Based on the subset of predictor features chosen, Fig 2.3.3 and Fig 2.3.4 reflects that the selected features are relatively uncorrelated. There are some weak correlations for Temperature and Visibility with Time, and IndCampus and minBart with x\_coord.

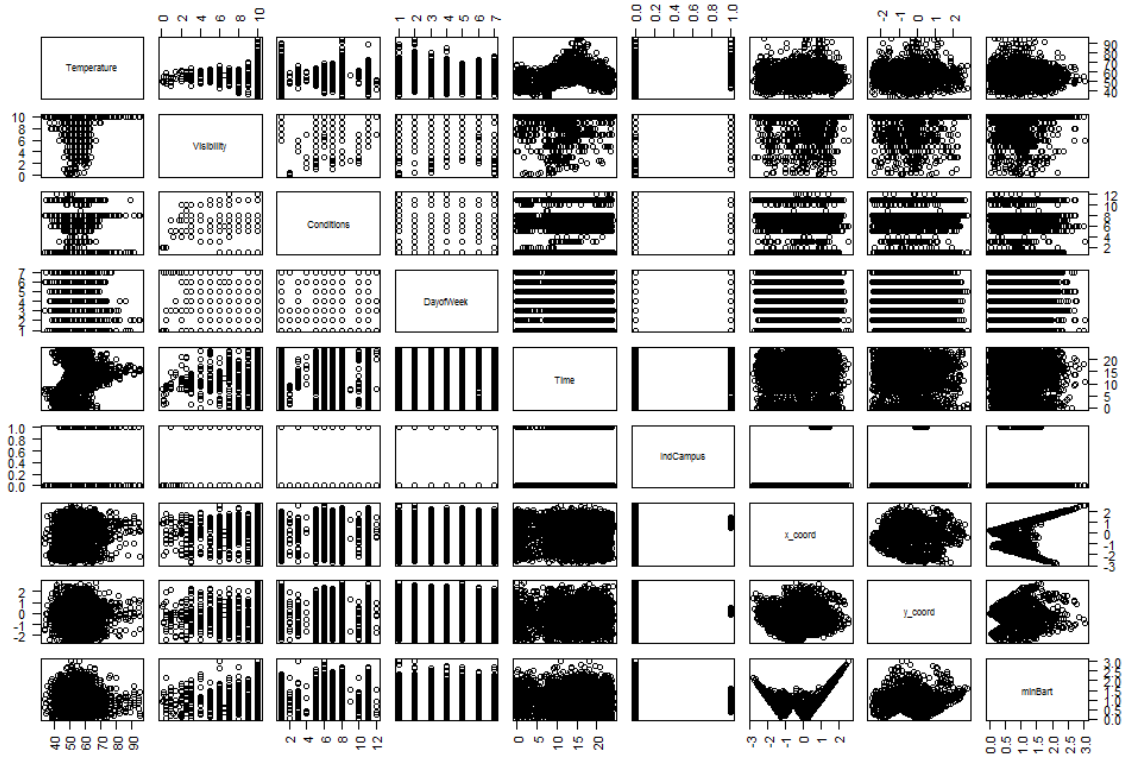


Fig 2.3.3 Pairwise plots of selected predictor features

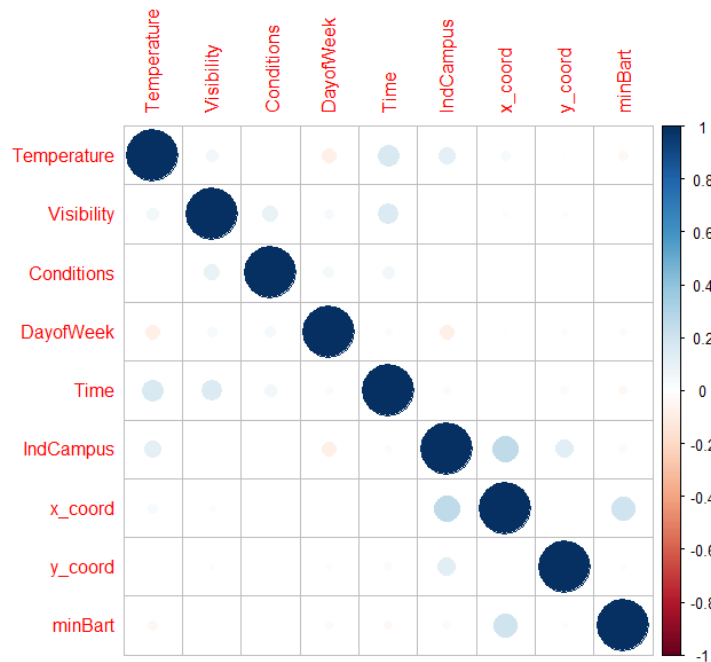


Fig 2.3.4 Correlation plot of selected predictor features

### **3. K-NEAREST NEIGHBORS [KNN]**

For each of the dataset below, we implement KNN classification with a modified Gower distance as the distance function and majority voting as the decision rule, with ties broken by picking the first candidate in an alphabetical order. We choose Gower distance since the predictors contain a mix of numerical and nominal variables. The Gower distance between two entries is the average of all the variable-specific distances, that is:

$$d(i, j) = \sqrt{\frac{1}{n} \sum_{k=1}^n |z(i, k) - z(j, k)|}$$

where  $n$  is the number of predictors, and  $|z(i, k) - z(j, k)|$  are the variable specific distances. For nominal variables, this is the dissimilarity coefficient, i.e. 0 if same and 1 if different. For numerical variables, this is the Euclidean distance or root sum-of-squares of differences. A modification is made for the Time column to take into account its continuous nature, i.e. the Euclidean distance between 23:00:00 and 1:00:00 is 2 instead of 22. The numerical variables are also scaled by Gower's standardization, based on range, to adjust for magnitude differences

As seen from the crime type distribution in the earlier Fig 2.3.1, the distribution is skewed, especially towards the top two crimes of petty theft and burglary. This potentially creates a drawback for the majority voting decision rule, since examples of a more frequent class tend to dominate the prediction of the new example, because they tend to be common among the  $k$  nearest neighbors due to their large number. One way we have addressed this is to divide the dataset into two more evenly distributed sets, one for the top two crimes (subsection 3.2) and one for the other crimes (subsection 3.3). Another adjustment we made to the algorithm is to multiply the class of each of the  $k$  nearest point by a weight proportional to the inverse of the distance from that point to the test point. The test point is classified into the class with the highest combined weight.

We carried out the algorithm for  $k = 1, 3, 5, 10, 15, 25, 35, 45, \dots, 165$ . Parameter selection is conducted by 10-fold cross validation.

#### **3.1 ALL CRIME TYPES**

The training set consists of 3889 entries, and the test set consists of 596 entries. There are 17 categories in the response column. Fig 3.1.1 shows the estimated test error by CV, and the test error for each parameter  $k$  when the 9 predictors are used. Judging from the estimated test errors, we may want to use a larger  $k$ . However, the error rates are generally high around 70%, since the top two classes still dominate the prediction of a new example. So we move on for now to see if the accuracy rate can be improved in the smaller datasets in the following subsections.

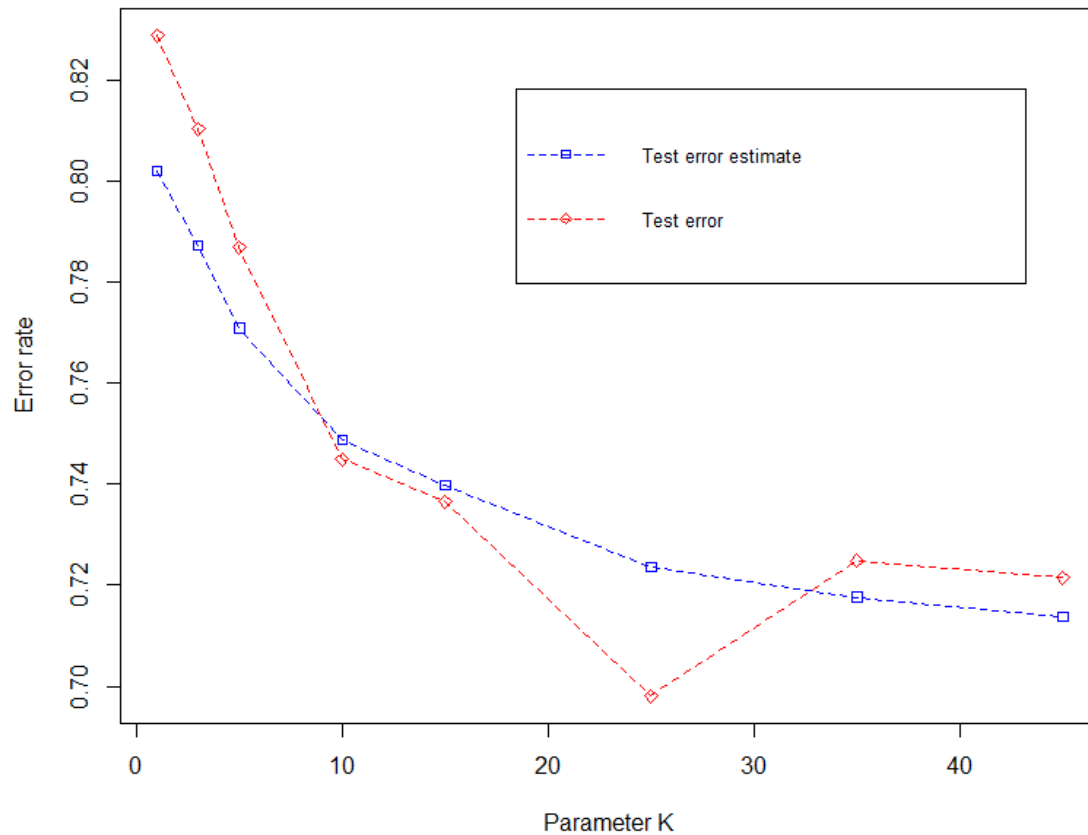


Fig 3.1.1 Error rate against parameter k

### **3.2 PETTY THEFT AND BURGLARY**

The training set consists of 1819 entries (914 burglaries and 905 petty thefts), and the test set consists of 262 entries (129 burglaries and 133 petty thefts). There are 2 categories in the response column. Fig 3.2.1 shows the estimated test error by CV, and the test error for each parameter k when the 9 predictors are used. Judging from the estimated test errors, we would want to use  $k = 45$ , since a larger k reduces the estimated test error slightly but increases the computation time.

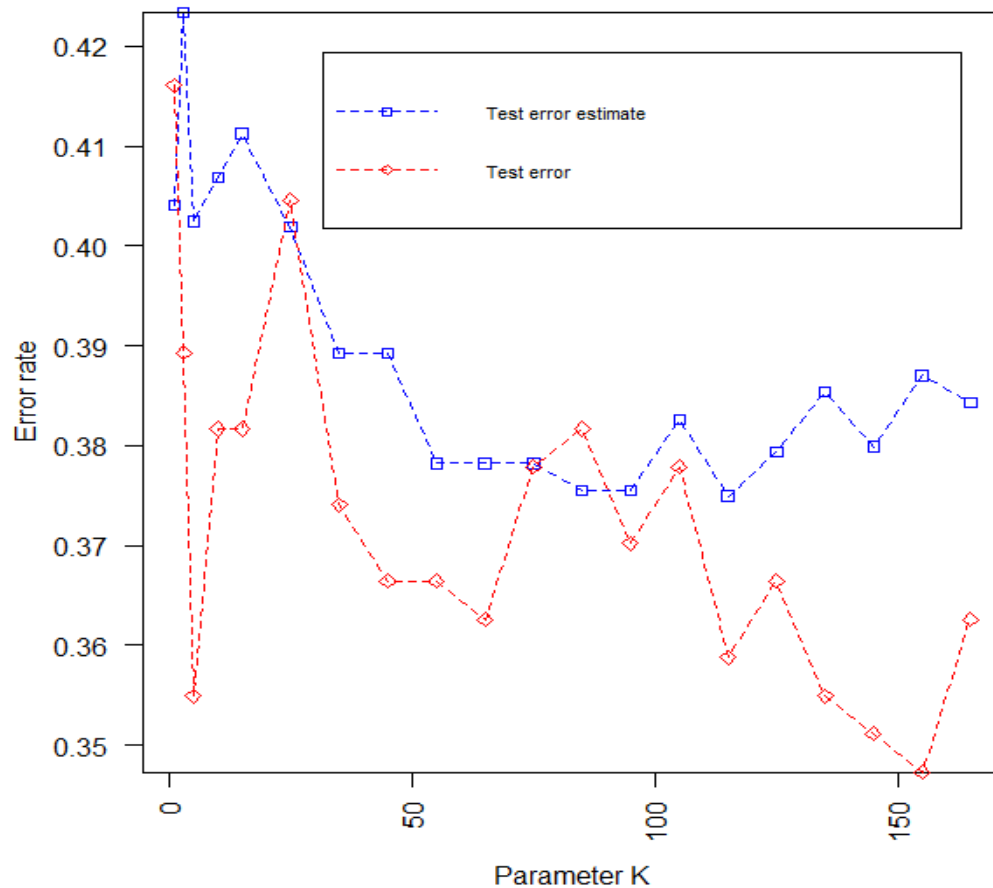


Fig 3.2.1 Error rate against parameter k

With  $k = 45$ , we attain a test error rate of 36.6%, or an accuracy rate of 63.4%. Table 3.2.2 shows the classifications made. The classifier is somewhat better than just taking a guess based on the crime type distribution in the training set, which would have an accuracy rate of approximately 50%.

	Classified as:		Correct
	Burglary	Petty Theft	
Burglary	77	52	60%
Petty Theft	44	89	67%

Table 3.2.2 Classification table



To improve the prediction, we used variable subset selection. The subset is chosen by cross validation on the training set with  $k = 45$ . Fig 3.2.3 shows the smallest 10 CV test error estimates. The subset corresponding to the smallest test error estimate consists of:

- Temperature
- Time
- IndCampus
- X\_coord
- Y\_coord
- minBart

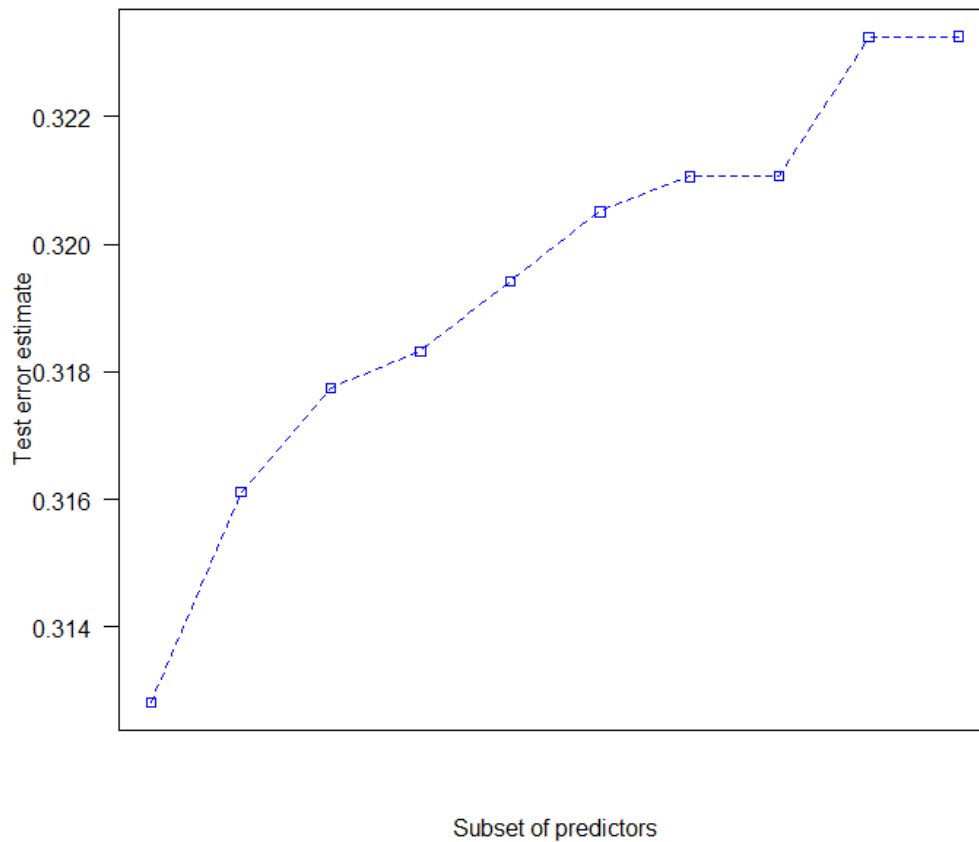


Fig 3.2.3 CV test error estimate for 10 subsets of predictors

Using the reduced set of 6 predictors, we attain a test error rate of 29.3%, or an accuracy rate of 70.7%. This is an improvement from the previous 63.4%. Table 3.2.5 shows the classifications made.

	Classified as:		Correct
	Burglary	Petty Theft	
Burglary	96	33	74%
Petty Theft	44	89	67%

Table 3.2.5 Classification table, using the reduced set of predictors

### 3.3 OTHER CRIME GROUPS

From the other crimes, we grouped the entries into broader categories based on the standards used in the Uniform Crime Reports (UCR). A subset is chosen for classification in a way such that the number of entries in each category is roughly similar.

- Part 1 offenses (violent): assault, robbery
- Part 1 offenses (property): motor vehicle theft, vehicle break-in
- Part 2 offenses: drugs/alcohol violations

The training set consists of 1002 entries (318 Part 1 property, 352 Part 1 violent, 332 part 2), and the test set consists of 188 entries (55 Part 1 property, 78 Part 1 violent, 55 part 2). There are 3 categories in the response column. Fig 3.3.1 shows the estimated test error by CV, and the test error for each parameter  $k$  when the 9 predictors are used. Judging from the estimated test errors, we would want to use  $k = 35$ , since a larger  $k$  only reduces the estimated test error marginally but increases the computation time.

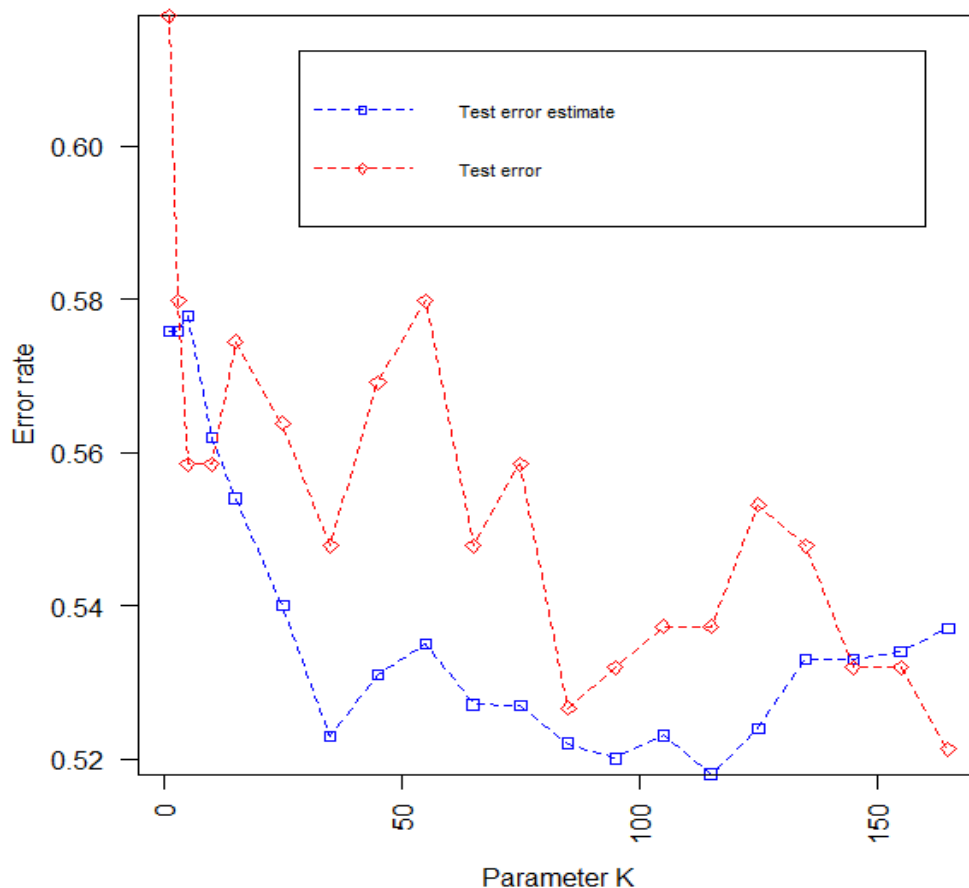


Fig 3.3.1 Error rate against parameter  $k$

Say we choose  $k = 35$ , we attain a test error rate of 54.8%, or an accuracy rate of 45.2%. Table 3.3.2 shows the classifications made. The classifier is somewhat better than just taking a guess based on the crime type distribution in the training set, which would have an accuracy rate of approximately 33%.

	Classified as:			Correct
	Part 1 Property	Part 1 Violent	Part 2	
Part 1 Property	23	21	11	42%
Part 2 Violent	19	41	18	53%
Part 2	9	25	21	38%

Table 3.3.2 Classification table

To improve the prediction, we used variable subset selection. The subset is chosen by cross validation on the training set with  $k = 35$ . Fig 3.3.3 shows the smallest 10 CV test error estimates. The subset corresponding to the smallest test error estimate consists of:

- Visibility
- Time
- IndCampus
- X\_coord

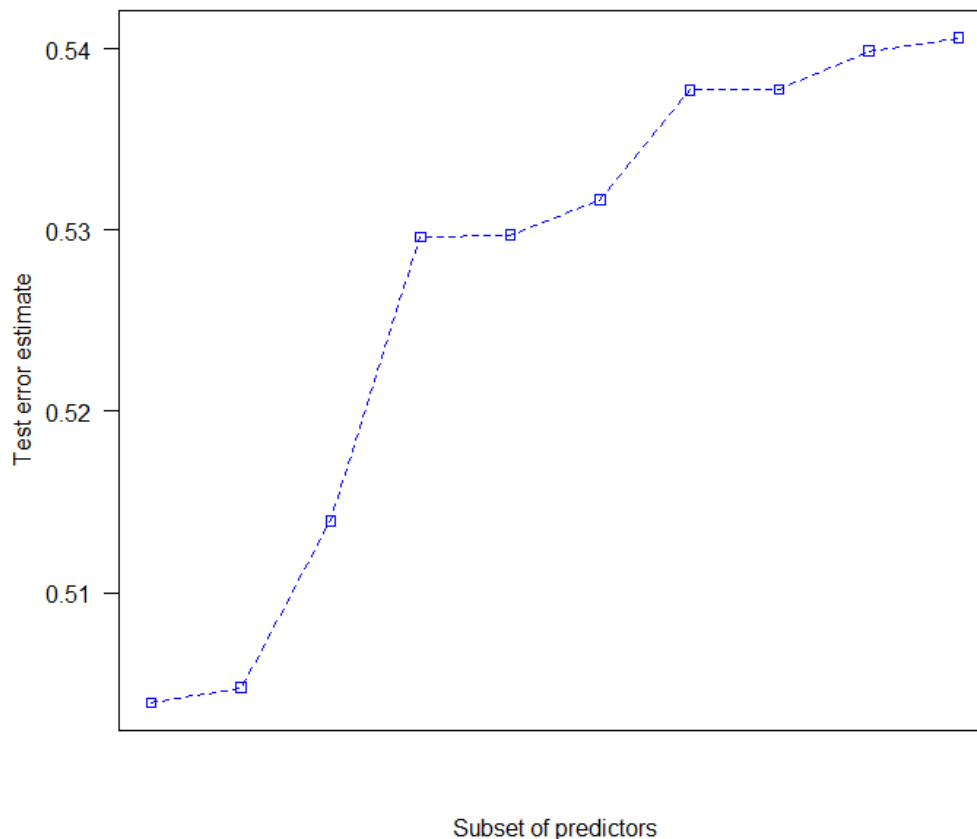


Fig 3.3.3 CV test error estimate for 10 subsets of predictors

Using the reduced set of 4 predictors, we attain a test error rate of 49.5 %, or an accuracy rate of 50.5%. This is an improvement from the previous 45.2%. Table 3.2.4 shows the classifications made.

	Classified as:			Correct
	Part 1 Property	Part 1 Violent	Part 2	
Part 1 Property	25	18	12	45%
Part 2 Violent	18	46	14	59%
Part 2	10	21	24	44%

Table 3.3.4 Classification table, using the reduced set of predictors

## 4. LOGISTIC REGRESSION

We implement binary logistic regression in classifying petty thefts and burglaries, and multinomial logistic regression in classifying the other crimes. The multinomial logistic regression is a simple extension of the binary logistic regression model where the response variable has more than two unordered categories. Here, we use the complete set of 17 extracted features as the predictors. We represent the categorical variables by indicators. That is, the 12-category Conditions variable is replaced by 11 indicators, and the 7-category DayofWeek variable is replaced by 6 indicators. The design matrix has 32 columns.

After the full model, we tried dimensionality reduction by principal component analysis. The columns are standardized before PCA. We determine the number of principal components to use by cross validation. Logistic regression is then re-implemented using this new set of variables.

We also tried variable selection using the R function *step*, a stepwise algorithm that choose a model by AIC.

### 4.1 PETTY THEFT AND BURGLARY

The full model has a training error of 37.6%, and a cross-validation test error estimate of 39.0%. We attain a test error is 39.1%, or an accuracy rate of 60.9%. Table 4.1.1 shows the classifications made.

	Classified as:		Correct
	Burglary	Petty Theft	
Burglary	88	40	69%
Petty Theft	62	71	53%

Table 4.1.1 Classification table

Next we applied PCA before re-implementing logistic regression. Fig 4.1.2 is a screenshot of the statistics of the 32 principal components outputted.

```
> summary(pca.data2)
Importance of components:
      Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6      Comp.7      Comp.8      Comp.9      Comp.10
Standard deviation  2.3335642  1.8065416  1.48479663  1.25798570  1.21185341  1.16533406  1.14129343  1.11292379  1.10244932  1.08468812
Proportion of Variance 0.1702544  0.1020363  0.06892753  0.04947778  0.04591546  0.04245801  0.04072428  0.03872484  0.03799934  0.03678481
Cumulative Proportion 0.1702544  0.2722907  0.34121819  0.39069597  0.43661143  0.47906944  0.51979372  0.55851856  0.59651790  0.63330271
      Comp.11      Comp.12      Comp.13      Comp.14      Comp.15      Comp.16      Comp.17      Comp.18      Comp.19      Comp.20
Standard deviation  1.08043313  1.06088068  1.05153835  1.00976228  1.00028733  0.99876416  0.98792305  0.96605234  0.94475130  0.9125733
Proportion of Variance 0.03649678  0.03518778  0.03457077  0.03187844  0.03128299  0.03118779  0.03051441  0.02917831  0.02790575  0.0260372
Cumulative Proportion 0.66979949  0.70498727  0.73955803  0.77143647  0.80271947  0.83390726  0.86442167  0.89359998  0.92150573  0.9475429
      Comp.21      Comp.22      Comp.23      Comp.24      Comp.25      Comp.26      Comp.27      Comp.28      Comp.29
Standard deviation  0.88199962  0.73738975  0.406245858  0.335268170  0.208950248  0.1704232439  0.0736575522  2.413376e-02  6.600036e-07
Proportion of Variance 0.02432179  0.01700016  0.005159845  0.003514337  0.001365037  0.0009080639  0.0001696264  1.820994e-05  1.361919e-14
Cumulative Proportion 0.97186472  0.98886488  0.994024725  0.997539062  0.998904100  0.9998121637  0.9999817901  1.000000e+00  1.000000e+00
      Comp.30      Comp.31      Comp.32
Standard deviation  1.296306e-08  3.802159e-09  0
Proportion of Variance 5.253807e-18  4.519800e-19  0
Cumulative Proportion 1.000000e+00  1.000000e+00  1
```

Fig 4.1.2 Principal components

We determine the number of principal components to use as the new set of variables through cross validation. Fig 4.1.3 shows the CV test error estimate for each number of components used. To minimize the test error estimate, we pick the first 24 components, since that is the point where the test error estimate graph reaches a minimum and flattens.

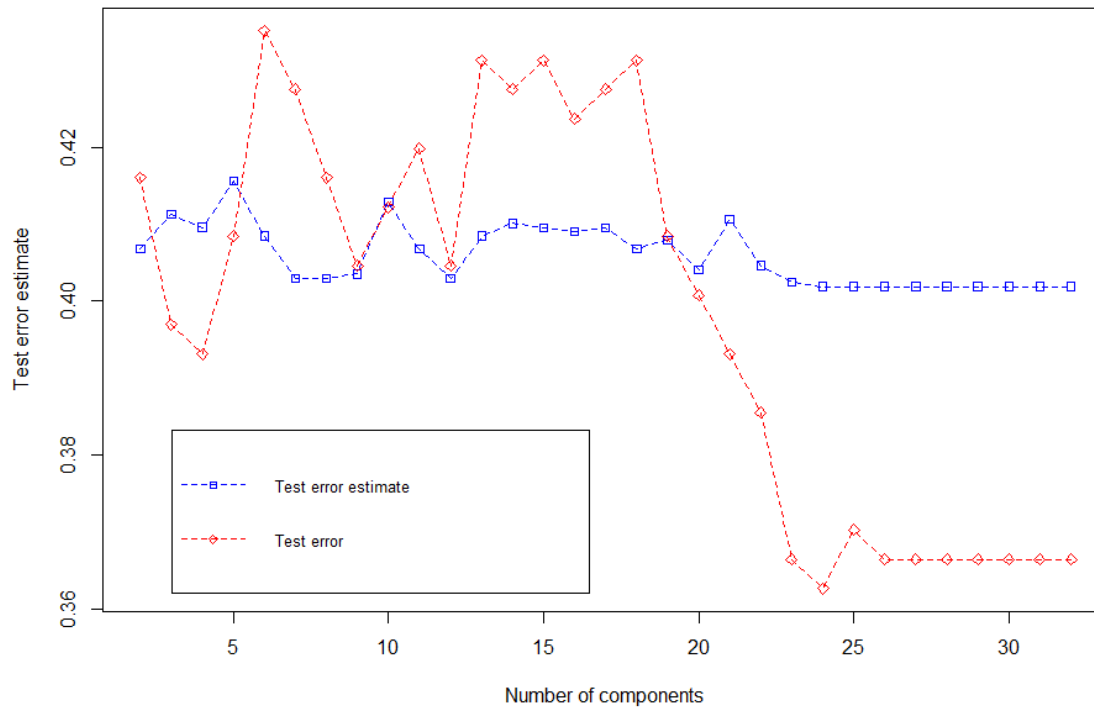


Fig 4.1.3 CV test error estimate for different numbers of principal components used

Implementing logistic regression with this new set of variables, we attain a training error of 38.4%, and a cross-validation test error estimate of 40.2%. We have a test error of 36.3%, or an accuracy rate of 63.7%. This is an improvement from the previous 60.9%. Table 4.1.4 shows the classifications made.

	Classified as:		Correct
	Burglary	Petty Theft	
Burglary	92	37	71%
Petty Theft	58	75	56%

Table 4.1.4 Classification table, using the first 24 principle components



Aside from PCA, we also tried dimensionality reduction by stepwise selection of variables. A combination of forward and backward stepwise selection is used to obtain the following 11 predictors:

- Longitude
- Temperature
- IndCampus
- X\_coord
- Y\_coord
- Downtown
- Ashby
- BerkPD
- UCPD
- minBart
- minPD

We obtain a training error of 37.6%, a cross-validation test error estimate of 38.1%, and a test error of 39.1%. The accuracy rate of 60.9% is lower than the previous model. Hence, PCA may be a better method for dimensionally reduction in this case.

## **4.2 OTHER CRIME GROUPS**

The full model has a training error of 49.9%, and a cross-validation test error estimate of 53.1%. We attain a test error is 52.7%, or an accuracy rate of 47.3%. Table 4.2.1 shows the classifications made.

	Classified as:			Correct
	Part 1 Property	Part 1 Violent	Part 2	
Part 1 Property	25	19	11	45%
Part 2 Violent	20	38	20	49%
Part 2	13	16	26	47%

Table 4.2.1 Classification table

Next, as in the previous subsection, we applied PCA before re-implementing logistic regression. We determine the number of principal components to use as the new set of variables through cross validation. Fig 4.2.2 shows the CV test error estimate for each number of components used. To minimize the test error estimate, we pick the first 12 components, since that is the point where the test error estimate graph reaches a minimum. However, using the first 12 principal components, the classifier has a 54.8% test error, which is higher than the 52.7% attained by the full model.

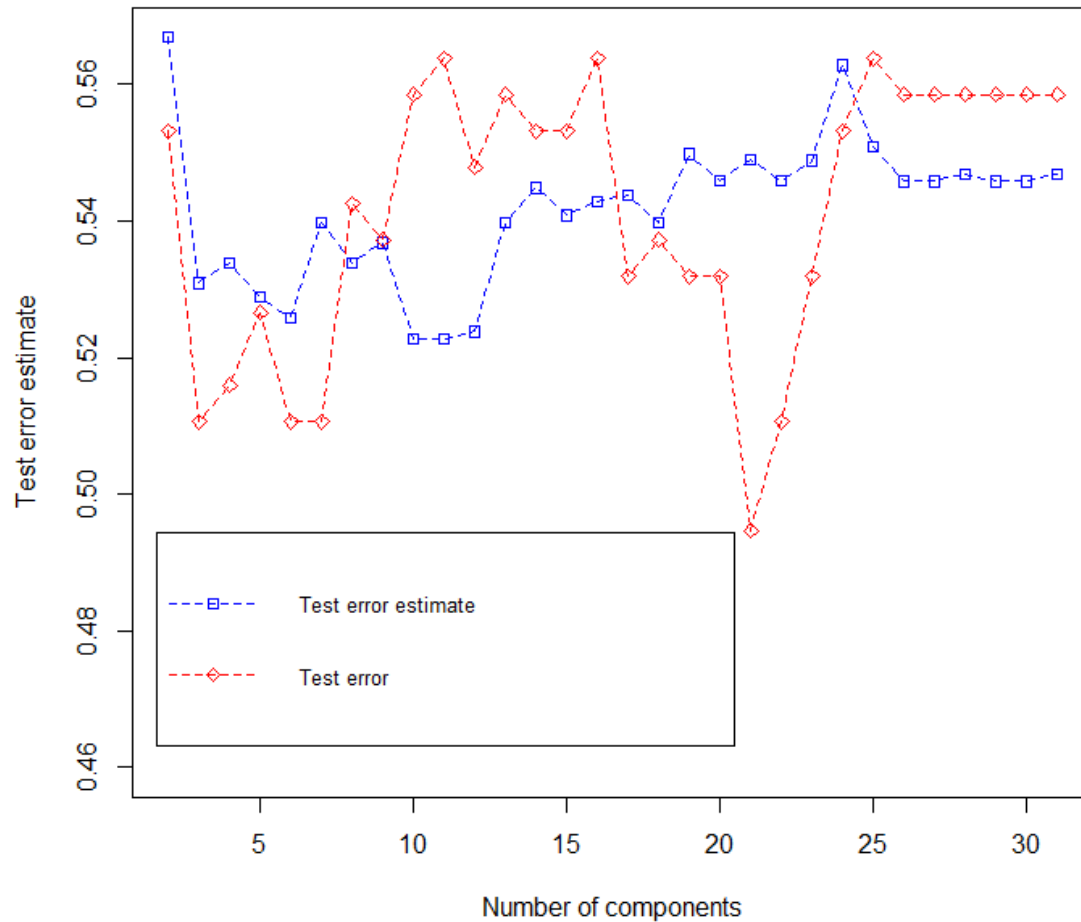


Fig 4.2.2 CV test error estimate for different numbers of principal components used

Aside from PCA, we also tried dimensionality reduction by stepwise selection of variables. A combination of forward and backward stepwise selection is used to obtain the following 11 predictors:

- Longitude
- Time
- IndCampus
- X\_coord
- Y\_coord
- North
- Ashby
- BerkPD
- UCPD
- minBart
- minPD

We obtain a training error of 49.9%, a cross-validation test error estimate of 54.0%, and a test error of 52.7%. The accuracy rate of 47.3% is the same as that of the full model. Hence, in this case, the two dimensionality reduction techniques do not seem to be useful in improving the classification accuracy.

## **5. LINEAR DISCRIMINANT ANALYSIS [LDA]**

The classification accuracy of LDA is quite low as compared to the other algorithms, possibly due to the following restrictive requirements:

- LDA works when the measurements made on independent variables for each observation are continuous quantities. However, our dataset has 3 categorical variables, i.e. Conditions, DayofWeek, and IndCampus. We excluded these 3 variables when applying LDA, and this loss of information may have compromised the accuracy of the classifier.
- The fundamental assumption of the LDA method is that the independent variables are normally distributed. We applied the Energy test of multivariate normality on the collection of entries in each category of crime, and find that the assumption of multivariate normality is violated. This may have undermined the optimality of the LDA algorithm.

Nonetheless, we state the results obtained from LDA. For the first sub-problem of classifying petty thefts and burglary, we have a training error of 40.1%, CV estimated test error of 40.6%, and test error of 39.4%. For the second sub-problem of classifying the 3 groups of crimes, we have a training error of 51.7%, CV estimated test error of 53.3%, and test error of 53.7%.

## 6. SVM

Support Vector Machines (SVM) are a supervised machine learning algorithm similar to nearest neighbors in that it represents each data sample as a point in  $p$  dimensional space, where  $p$  is the number of variables. Multivariable inputs are mapped into the high-dimensional feature spaces via a kernel function. Here we are using a C-support vector classification SVM where  $C$  is the regularization parameter, the cost of classification errors. We chose by CV a Gaussian radial based function as the kernel,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

Before training on our current dataset, we would first like to introduce some new features. Since Time is a periodic variable which fluctuates between 0 and 24, it is not as useful to treat it as a numeric variable. Hence we decided to add a feature, timeOfDay which captures which part of the day the crime occurs in. The four partitions are as follows:

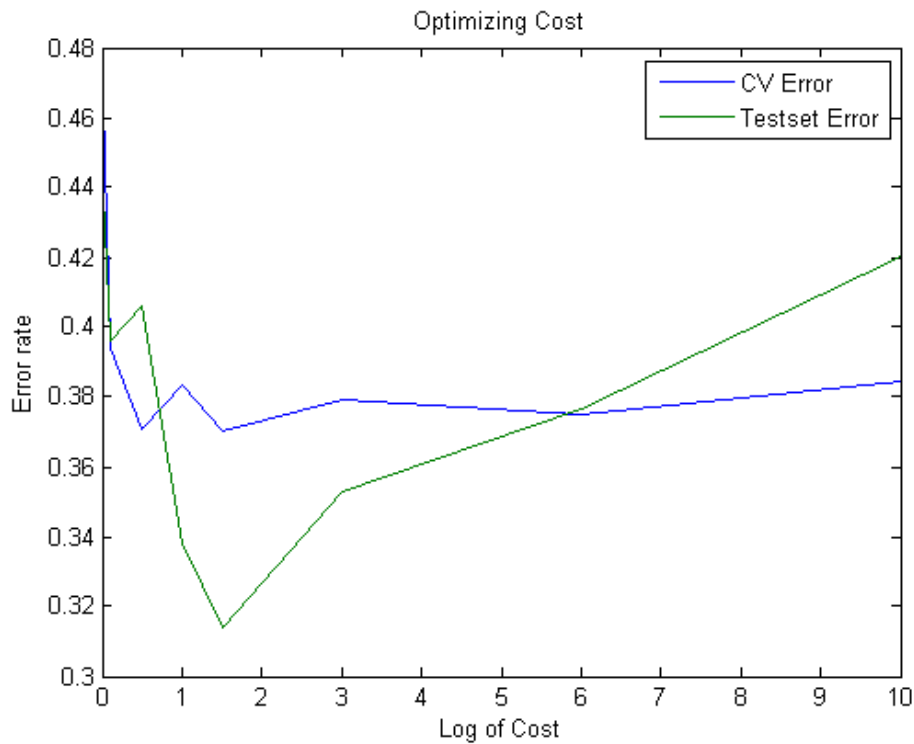
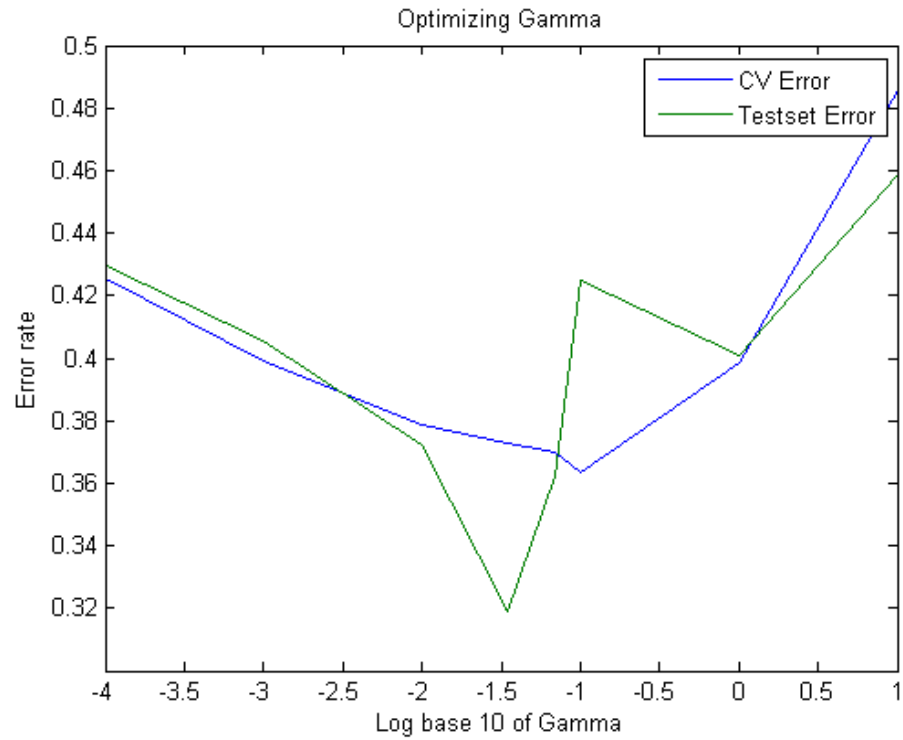
timeOfDay	Time
1	03:00 – 09:00
2	09:00 – 15:00
3	15:00 – 21:00
4	21:00 – 03:00

Another change we made is to divide the categorical variables into an indicator matrix. For example, DayofWeek is split into 6 separate feature columns as below,

mon	tue	wed	thur	fri	sat
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0

The last category, Sunday is left out since it is represented by all zeros in the other 6 variables. This procedure is repeated for the weather conditions.

Optimizing the accuracy for the SVM parameters gamma  $\gamma$  and cost  $C$  on the two crime data set, we obtain the following results,



Substituting in the optimized values of  $\gamma = 0.035$  and  $C = 1.5$ , we obtain 31.5% error rate when classifying the 2 crimes, petty theft and burglary.

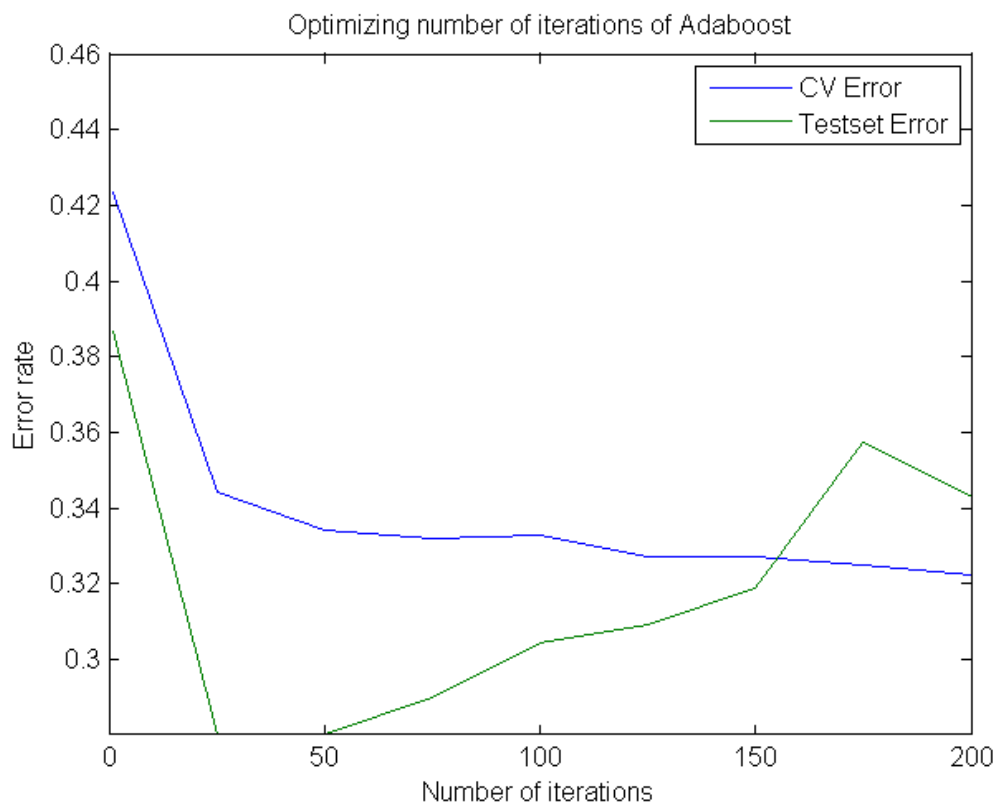




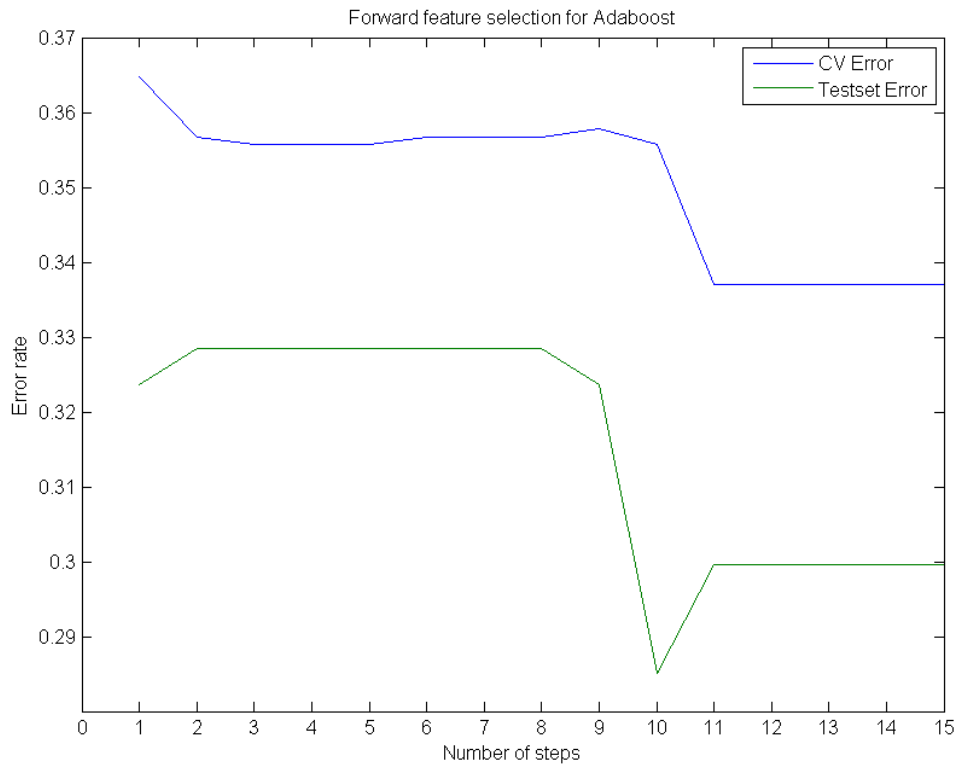
The variable remaining in the optimally pruned tree are x6: y\_coord (the relative location of the crime with respect to the center of Berkeley) and x8: minPD (the minimum distance from the crime to the nearest police department station). At the root node, we see that the tree splits on the variable y\_coord with decision threshold -0.358193. Samples with y\_coord less than the threshold are classified as burglary, suggesting that burglaries are more likely to occur on the south side of Berkeley. In the second level of the tree, we again split on y\_coord, indicating that burglaries are less common near the city center. This is consistent with the fact that residential buildings are more concentrated the further away from the center of Berkeley. Finally in the last level, we classify the crime as theft if minPD is less than approximately 2.5km and burglary otherwise, implying that burglars prefer to operate further away from police presence. This is also consistent with our previous observation because the Berkeley police department is situated slightly to the north-east of the center of the city.

## **7.1 BOOSTING**

Boosting is another example of model averaging alongside bagging and random forests. The stronger classification model constructed consists of an ensemble of basic threshold classifiers, which by themselves are relatively weak. We used adaptive boosting (Adaboost) to train on our data, with each iteration of the algorithm, the classifiers are adapted to the data by taking into account previous misclassifications. Using 10-fold CV to tune for the number of iterations of Adaboost,



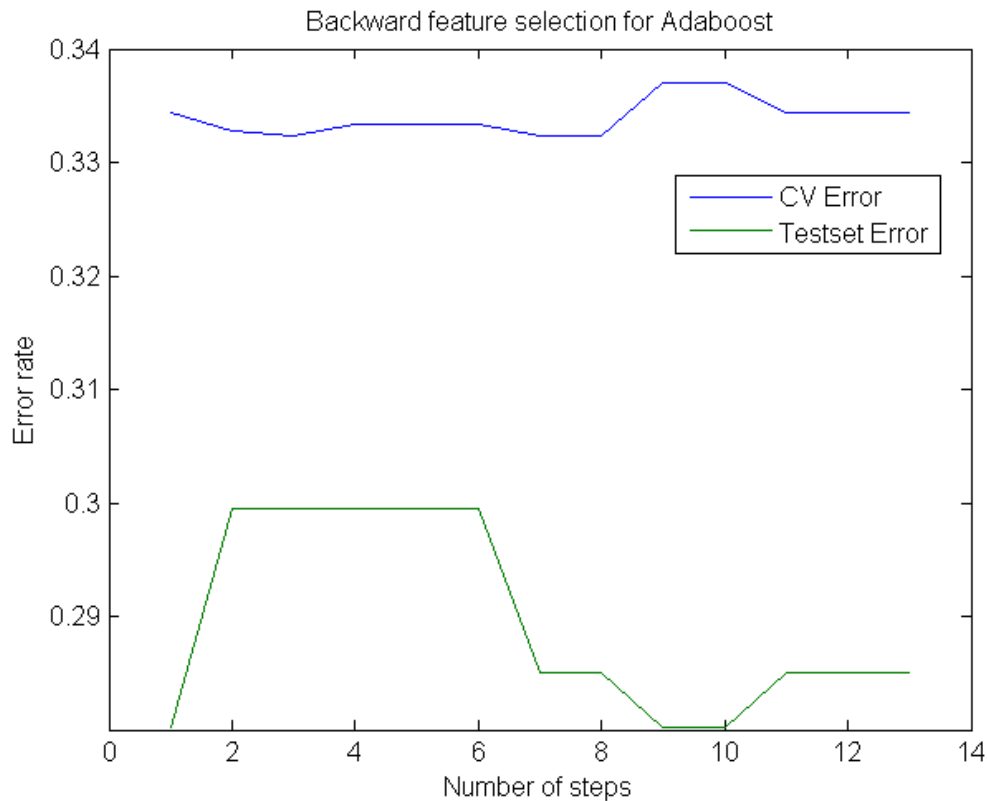
Adaboost is most effective for our data at 50 iterations, giving a test error rate of around 28%. Next, we attempt to optimize the performance further via variable selection. First we run forward stepwise feature selection on the training set with a threshold of  $1e-6$ .



The lowest test error rate was observed from the set of features extracted in the 10<sup>th</sup> step, the 5 most important variable names are the following (in order of selection):

- Y\_coord
- IndCampus
- Visibility
- Mon
- Tue

Similarly, for backward stepwise feature selection,



In this case, feature selection is not effective in minimizing the error rate of the algorithm, as the lowest test error obtained is still around 28%. One possible reason for this is with 50 iterations of model averaging, Adaboost by itself is already very efficient at picking out the most important features and adjusting the decision boundaries for them.

## **8. DISCUSSION OF RESULTS**

Table 8.0.1 below summarizes the results of our model testing. The cross validation test error estimates and the test errors corresponding to the most optimized version of each algorithm are listed here for comparison.

	Petty Theft & Burglary		Other Crime Groups	
	CV error	Test error	CV error	Test error
KNN	31.3%	29.3%	50.5%	49.5%
SVM	36.4%	31.9%	56.3%	51.2%
Logistic	40.2%	36.3%	53.1%	52.7%
LDA	40.6%	39.8%	53.3%	53.7%
Trees	37.3%	39.1%	49.3%	56.3%
Trees (pruning)	35.1%	32.3%	49.7%	37.8%
Adaboost	33.2%	28.0%	- <sup>1</sup>	-

Table 8.0.1 Summary of CV error and test error for all models

For each classification problem, we take the best model to be the one that gives the lowest test error. For the first sub-problem of classifying petty theft and burglary, this is Adaboost. For the second sub-problem of classifying the three crime groups, this is the pruned trees. Using the classification decisions made by the best model, we attempt to gain some insights on the factors that determine each type of crime. That is, for the first sub-problem, we study the characteristics of all the test set entries grouped under petty theft and burglary respectively. For the second sub-problem, we study the characteristics of all the test set entries grouped under Part 1 offenses: violent, Part 1 offenses: property and Part 2 offenses respectively.

### **8.1 PETTY THEFT AND BURGLARY**

We compare the relevant features of the 2 crimes dataset as classified by Adaboost. From the classified data, we see that petty theft tends to happen earlier in the day with the majority falling into the 09:00 – 15:00 slot. Burglars however are more active later in the day, with over half occurring between 15:00 and 03:00. This is consistent with the prior distribution of the time of day feature.

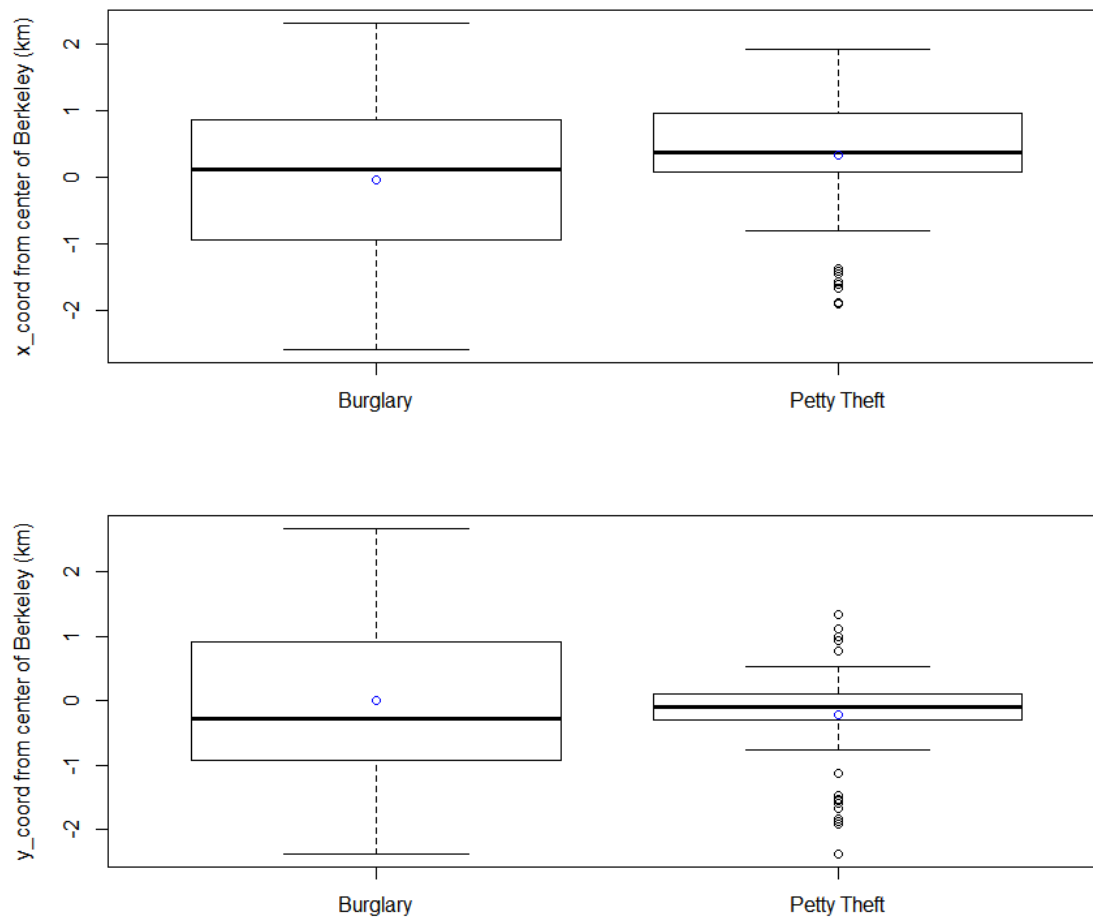
---

<sup>1</sup> For our project, we only manage to optimize Adaboost for a binary classification problem.

Table of distributions of timeOfDay variable in %:

timeOfDay	Petty Theft		Burglary	
	Posterior	Prior	Posterior	Prior
03:00 – 09:00	04	06	19	16
09:00 – 15:00	48	42	24	25
15:00 – 21:00	37	38	36	37
21:00 – 03:00	11	14	21	22

The distinction is even clear with the IndCampus feature, where 21% of crimes classified theft happened on campus whereas all burglaries had this indicator turned off. As we have observed from the pruned tree model earlier, burglaries tend to be spread out around the city while theft tends to be concentrated in the middle of Berkeley. The x coordinate for petty theft is skewed towards the positive (right side), which tells us that it is more likely in and around the campus grounds.

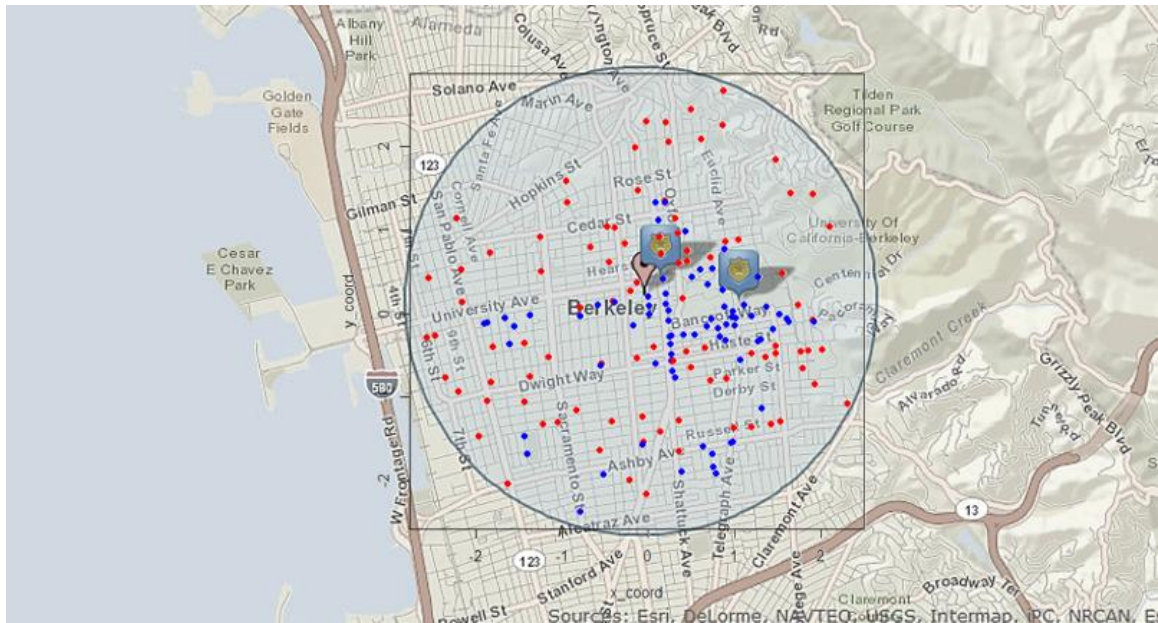


The 2 crimes are more likely in the vicinity of a BART station as opposed to a PD, with burglary being further away in both cases.

Table of distances on average to the nearest BART station and PD in km:

	minBart	minPD
Petty Theft	0.72	1.18
Burglary	1.06	1.54

Our observations are summarized in the scatter plot below:



Blue dots: Petty theft

Red dots: Burglary



## 8.2 OTHER CRIME GROUPS

For simplicity, we will refer to Part 1 offenses (violent) as violent crimes, Part 1 offenses (property) as property crimes, and Part 2 offenses as drug violations.

Fig 8.2.1 is a visualization of the pruned tree that is optimal in the classification of the three crime groups. The variables remaining in the optimally pruned tree are  $x_6$ :  $y\_coord$ ,  $x_5$ :  $x\_coord$ ,  $x_7$ :  $minBart$ ,  $x_8$ :  $minPD$ ,  $x_3$ :  $Time$ , and  $x_4$ :  $IndCampus$  in order of importance. At the root node, the tree splits on  $y\_coord$  with decision boundary 0.420961. Since samples with  $y\_coord \geq 0.420961$  are all classified as property crimes, we would expect the north side of Berkeley to consist predominantly of property crimes, and the south side to have more of a mix of crimes. We notice that except for  $Time$ , all the splits are made on location-related features. The model optimizes classification primarily by using the location features of each crime group. We will study in more detail the implications of these classification decisions.

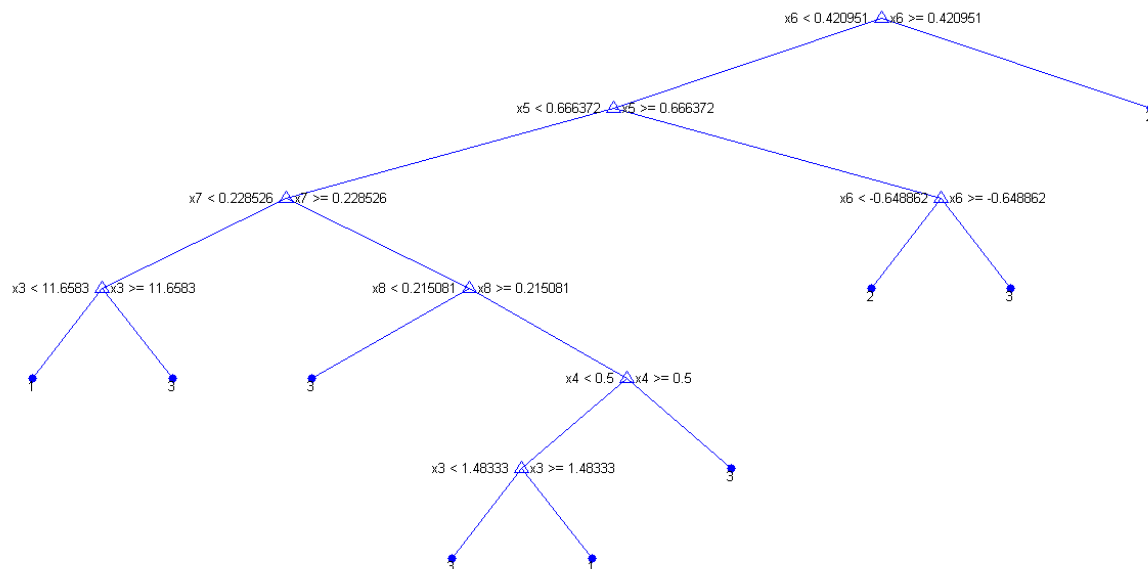


Fig 8.2.1 Pruned tree

- 1: Violent
- 2: Property
- 3: Drugs

First, we look at the weather information. Violent crimes are predicted to occur at a mean temperature of 57°F, while property crimes and drugs violations around 55°F. A one-way ANOVA test returns a p-value of 0.394, such that we cannot conclude that the mean temperature is different for the three groups. The mean visibility for all three groups is around 9.5 to 10 miles. 40-60% of crimes in each group is predicted to happen in slightly cloudy weather (i.e. partly cloudy and scattered clouds). In particular, 40% of property crimes are expected in 'Partly Cloudy' weather. We conclude that all three crime groups occur in similar weather conditions. This may also be due to the lack of extreme variability in Berkeley weather in the months from October to March.

Next, we look at the time and day at which a crime group is more likely to happen. For all three groups, 40-60% of crimes is predicted to happen between 3pm and 9pm. In particular, 56% of property crimes are expected in this timeslot. Considering the day of the week, Wednesday should see the largest number of crimes. For violent crimes, 30% of crimes is expected on Wednesday and 21% on Sunday. For property crimes, 22% on Monday and Wednesday each. For drug violations, 21% on Monday and Wednesday each. Thursday can be deemed the 'safest' day, with less than 10% of crimes predicted for all three groups.

Lastly, we look at the locations of crime. No violent and property crimes are expected on UC Berkeley campus, while around 10% of drug violations are expected on campus. Considering the map of Berkeley as a X-Y grid with the center of Berkeley as (0,0), we find that the x and y coordinates are significantly different for each group of crime. In Fig 8.2.2 and Fig 8.2.3 below, we can see that the x and y coordinates are distributed differently for different crime groups. In particular, the y coordinates for property crimes have a wide range, so that property crimes are expected in a bigger stretched area. In contrast, the y coordinates for drug violations have a small range, so that drug violations are expected in a more concentrated area. The blue point in each boxplot marks the mean. A one-way ANOVA test returns a p-value of 1.03e-12 and 0.000268 for the x and y coordinates respectively, such that we can conclude that the mean x and y coordinates are not the same for all three groups.

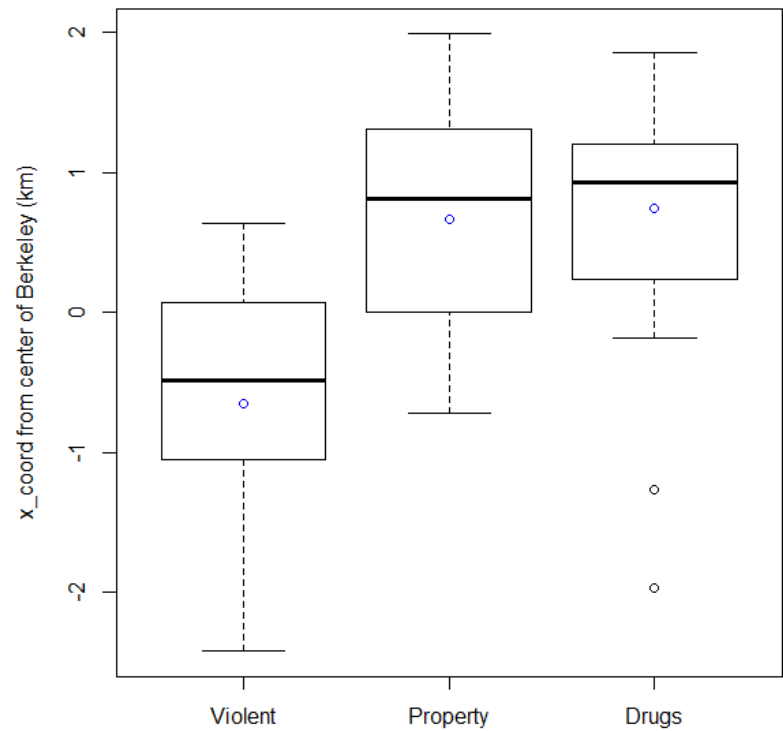


Fig 8.2.2 Boxplot of x\_coord for each crime group

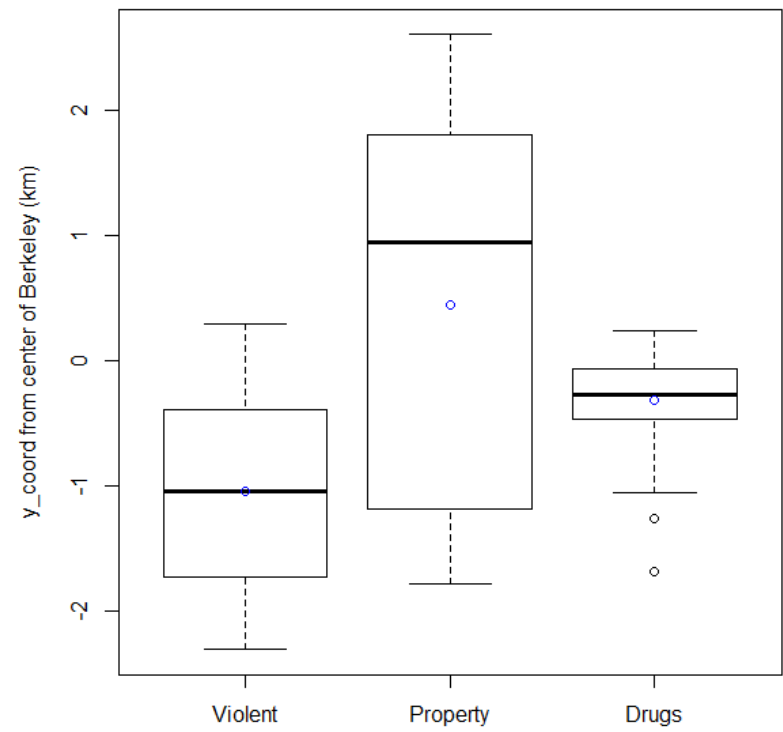


Fig 8.2.3 Boxplot of y\_coord for each crime group

Similarly, the three groups of crime tend to occur at different distances from landmarks such as Bart stations and Police Departments. Fig 8.2.4 and Fig 8.2.5 reflects the distributions of the minimum distances from Bart stations and from Police Departments for each group of crime. It seems that violent crimes are the most likely amongst the three to happen near Bart stations and Police Departments, while property crimes are the least likely. We also tested the equality of the means with one-way ANOVA. For minBart, the test returns a p-value of 0.0449. For minPD, the test returns a p-value of 0.0856. At 10% significance level, we would reject the null hypothesis of equal means in both cases. But at 5% significance level, we would not be able to reject the null hypothesis in the second case. Thus, we would conclude that minPD is correlated with the crime group, though possibly less than minBart.

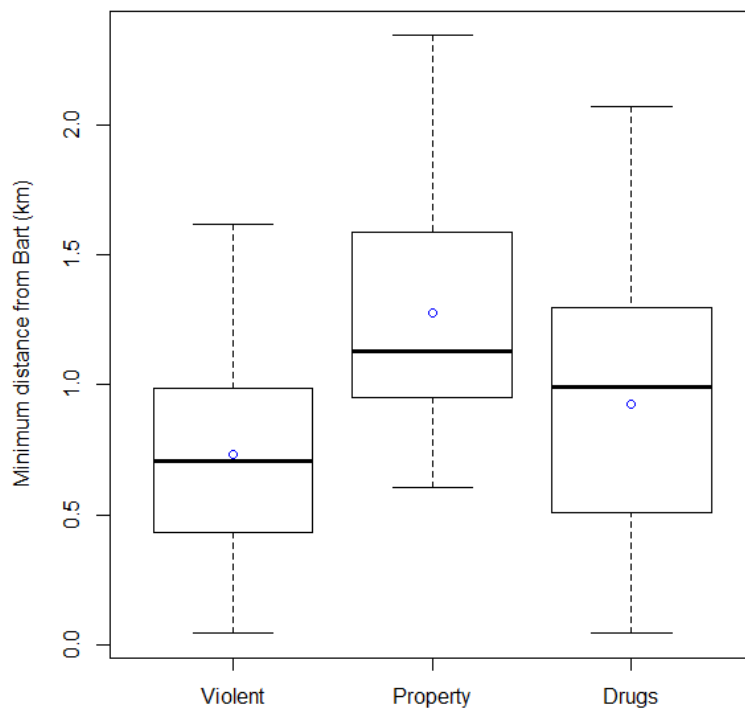


Fig 8.2.4 Boxplot of minBart for each crime group

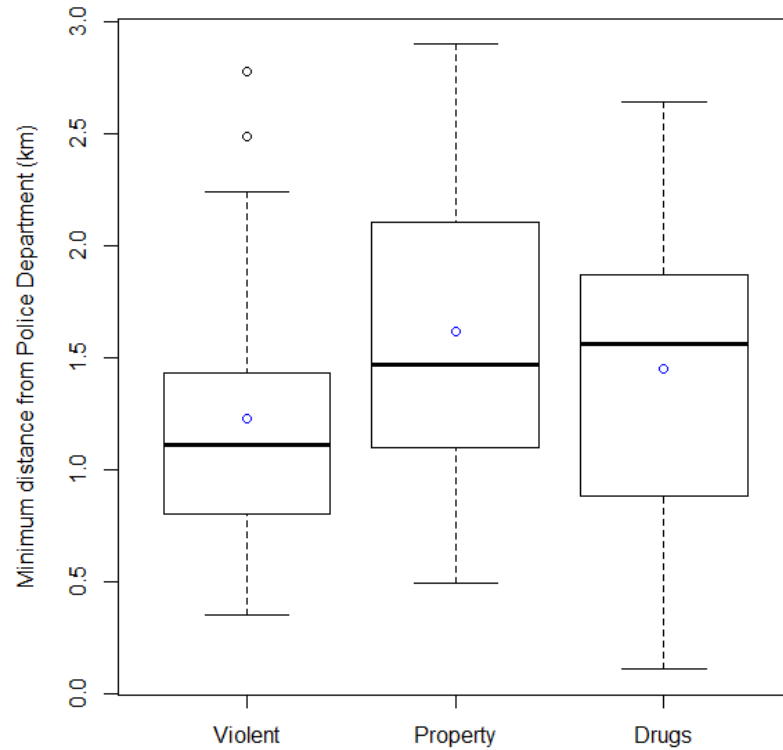
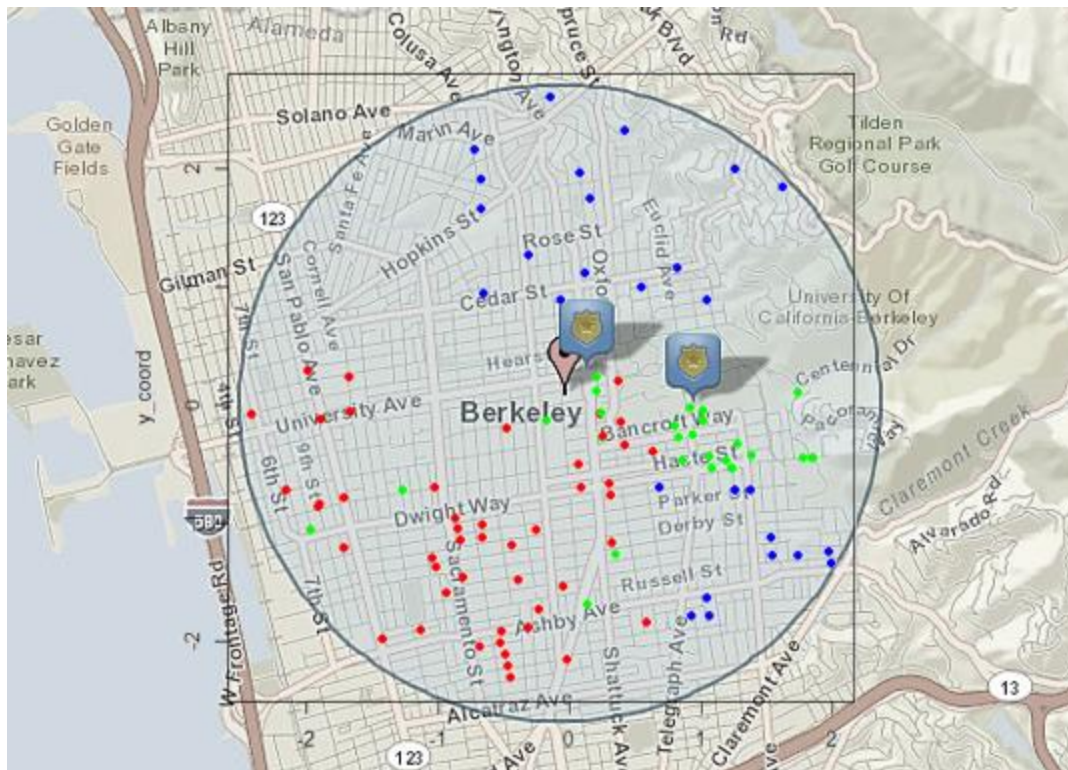


Fig 8.2.5 Boxplot of minPD for each crime group

To illustrate the relationship between the locations and the types of crime predicted, we created an overlay of the scatter plot over the map of Berkeley in Fig 8.2.6 below. As previously discussed, property crimes tend to spread along the y-axis, and drug violations tend to be concentrated near the parameter of the UC Berkeley campus.



- ♦ Violent
- ♦ Property
- ♦ Drugs

Fig. 8.2.6 Scatter plot of predictions for the three crime groups

## **9. CONCLUSION**

In this project, we have summarized several trends and characteristics of crime in Berkeley. Petty theft and burglary are the two most common crimes committed here, and while there are approximately equal numbers of reports for each crime over the past 6 months, there are significant differences in likelihood of them occurring in different parts of Berkeley. Petty theft is much more likely closer to the campus and the city center, on and near the main streets and transit. It also tends to occur in the mid-afternoon to early evening period, when the streets are most active. On the other hand, residential homes some distance from Berkeley center are most at risk of getting broken into, with the likelihood increasing the further the distance from the city center and the police departments. Burglars are active all periods of the day, with a small majority operating in the late afternoon to evening period.

For the three crime groups, violent crimes like assault and robbery are predicted to occur more often in the densely populated neighborhoods of south and west side. Property crimes, and in particular motor vehicle theft and vehicle break-in, are scattered across a wide region in the residential neighborhoods of both north and south side. They tend to be more likely in places further from human traffic. Drugs and alcohol violations are mostly cluttered near the city center, and especially at the periphery of the UC Berkeley campus. It is also the only crime group, out of the three tested, that is expected on campus.

In general, the accuracy rates we obtained in this project is not very high due to the lack of complete information regarding the crime. While characteristics of the victims and the culprits will be most relevant in classifying the type of crime, in our project we have set out to predict which kind of crimes are more likely to occur based on known variables: location and location features, time and weather conditions.

### **9.1 LIMITATIONS**

One limitation of this project is the range of our data. The crime samples used in our experiments range from October 2012 to March 2013, about half a year. The city of Berkeley has a population dominated by students, as evident from the neighborhood median age of 21 years. Therefore the population dynamics would have some significant changes from the summer holiday months from May to August, as compared to the semester months from September onwards. Also, general weather conditions also vary between the two halves of the year, with October-March being the cooler part of the year and April-September being the warmer part of the year. This lack in complete data prevents us from getting a full picture of the relationship between crime, location and weather conditions.

## **9.2 FURTHER STUDIES**

With regards to the above limitation, further experimentation on at least a year's worth of Berkeley crime data may reveal more helpful general observations for our research question. In particular, the higher weather variability between the 2 halves of the year would likely produce more interesting observations regarding crime and weather conditions. Also, optimizing decision tree boosting algorithms for multi-class classification would likely produce better results for the 3 groups of crime problem and also help us generalize our results to a larger number of crimes (top 5 crimes, top 10 crimes etc.). Finally, replicating this experiment in other university-oriented towns or regions could help us find common crime trends and risks on and near campus grounds.