Jim Haskin
GA_Data Science
December 1, 2015

Analysis and prediction of Crime data as it relates to weather conditions in the city of San Francisco

Hypothesis

It is often said that people lose their temper more when it is hot out and that all the crazies come out when there is a full moon. But does the data prove these hunches are true?

If it is true, can we look at the weather forecast for the next week and predict if the number or intensity of Police Incidents will be higher or lower than usual.

Using this Report

This report is a summery of the project. For more details on the implementation please see the accompanying ipython notebooks.

- **SF_final_crime_and_weather_JFH** : Contains the project statement along with the modeling and prediction code. It will also guide you to the other notebooks used.
- **2_clean_data_sf_crime :** Processes all the crime data and summarizes by the day.
- **3_clean_weather_combine_with_crime** : Cleans the weather data adds engineered features and combines it with the crime data.
- **5_analysis_cleaning** : Investigates the data and looks into outliers and other issues.
- **8_get_weather_forcast** : Uses two weather APIs to collect the forecast for the next ten days. Also adds any engineered features that were added to the modeling data.

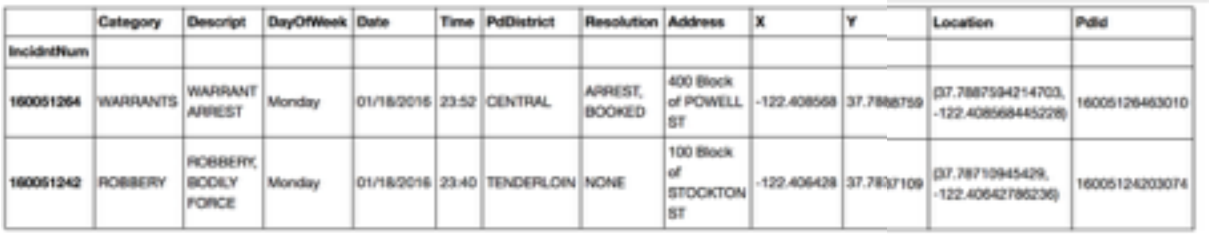
The Data

The Crime Data

Crime Data - This was done in the **2_clean_data_sf_crime** ipython notebook.

I collected the incident reports of the San Francisco Police Department from the SF OpenData website. https://data.sfgov.org/data?category=Public%20Safety.

I have the records from January, 2003 until the beginning of 2016.



	Category	Descript	DayOfWeek	Date	Time	PdDistrict	Resolution	Address	X	Y	Location	PdId
IncidntNum												
160051264	WARRANTS	WARRANT ARREST	Monday	01/18/2016	23:52	CENTRAL	ARREST, BOOKED	400 Block of POWELL ST	-122.408568	37.7888759	(37.7887594214703, -122.408568445228)	16005126463010
160051242	ROBBERY	ROBBERY, BODILY FORCE	Monday	01/18/2016	23:40	TENDERLOIN	NONE	100 Block of STOCKTON ST	-122.406428	37.787109	(37.78710945429, -122.40642786236)	16005124203074

The features that I am interested in are the Category, Description, Date and Time.

What Am I Predicting

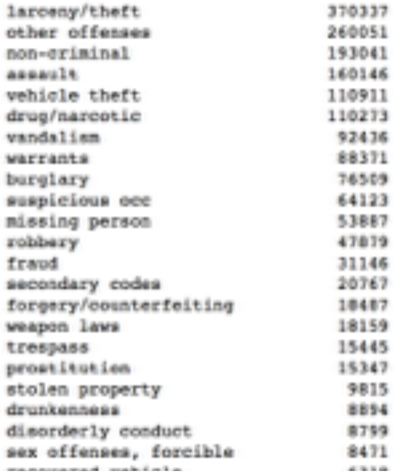
I am interested in predicting if the amount of crime reported on a given day is effected by the weather. So I aggregated the data for each day.

I created several different daily statistics. They were all saved in the data file to be analyzed later. The simplest variable was just the count of the number of Incidents on the day.

That was called **crime_count.**

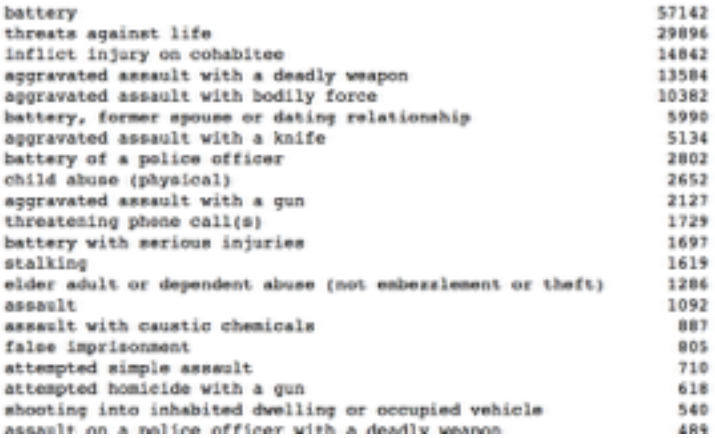
But the data also contains Categories and Description the can be used to categorize the type of crime.

Categories



```
larceny/theft                   370337
other offenses                  260051
non-criminal                    193041
assault                         160146
vehicle theft                   110911
drug/narcotic                   110273
vandalism                        92436
warrants                         88371
burglary                         76509
suspicious occ                   64123
missing person                   53887
robbery                          47879
fraud                            31146
secondary codes                  20767
forgery/counterfeiting           18487
weapon laws                      18159
trespass                         15445
prostitution                     15347
stolen property                   9815
drunkenness                       8894
disorderly conduct                8799
sex offenses, forcible            8471
```

Descriptions in the Assault category



```
battery                                                            57142
threats against life                                               29896
inflict injury on cohabitee                                        14842
aggravated assault with a deadly weapon                            13584
aggravated assault with bodily force                               10382
battery, former spouse or dating relationship                       5990
aggravated assault with a knife                                     5134
battery of a police officer                                         2802
child abuse (physical)                                              2652
aggravated assault with a gun                                       2127
threatening phone call(s)                                           1729
battery with serious injuries                                       1697
stalking                                                            1619
elder adult or dependent abuse (not embezzlement or theft)          1286
assault                                                             1092
assault with caustic chemicals                                       887
false imprisonment                                                   805
attempted simple assault                                             710
attempted homicide with a gun                                        618
shooting into inhabited dwelling or occupied vehicle                 540
assault on a police officer with a deadly weapon                     489
```

These allowed me to separate the different type of crimes. I am looking for the ones are most effected by the weather. The thought is that the weather effects peoples moods and emotions. So crimes that are more violent or spur of the moment will be effected most.

- **violent_count** - Incidents from the ['assault', 'sex offenses, forcible', 'secondary codes'] categories.
- **COP_count** - Crimes of Passion - Incidents that have any of these key words in the Description.

```
cop_words = ['assault', 'battery', 'drunk', 'abuse', 'forced', 'rape', 'shooting',
        'violence', 'harassing', 'threat', 'threatening', 'threats', 'resist', 'resisting',
        'destruction', 'weapons', 'gun', 'knife', 'armed', 'deadly', 'drunkenness',
        'bomb', 'bombing', 'influence', 'looting', 'disorderly', 'force', 'forcible',
        'fighting', 'injuries', 'nuisance', 'homicide', 'alcohol', 'rape', 'mayhem',
        'abuse', 'cruelty', 'lewd', 'molest', 'disturbing']
```

The summarized data was written to file. **sf_crime_clean.csv**

The Weather Data

Weather Data - Done in the **3_clean_weather_combine_with_crime** ipython notebook

I collected historical weather data for the San Francisco area from January 2003 until December 2015.

The data comes from Weather Underground. http://www.wunderground.com/history/

There were 22 features related to the Daily weather. Max Temperature, Max Humidity, etc. Please see the python notebook for details. Most of the data was clean. Only a couple of field needed adjustments.

More Features

To that data I added new calculated features that I thought might effect peoples emotions.

The Sun and the Moon

Using the date field I was able to use. the **ephem** module in a script (**sf_sun_moon.py**) to calculate the phase of the moon and the length of the day. I had to see if the full moon really does make the lunatics come out. The amount of sun each day can also have a large effect on people.

What does It Feel Like Out There

When talking about weather we always talk about how cold it really feels out there. There are many formulas used over the years to try and get this right. Accuweather has several patents dealing with this.

I used the **pywws.conversions** module to calculate the following:

- Heat index
- Wind Chill Max and Mean
- Dew Point Max and Mean
- Apparent Temperature

I also calculated the change in temp and humidity during the day. The thought is that big swings will effect people more.

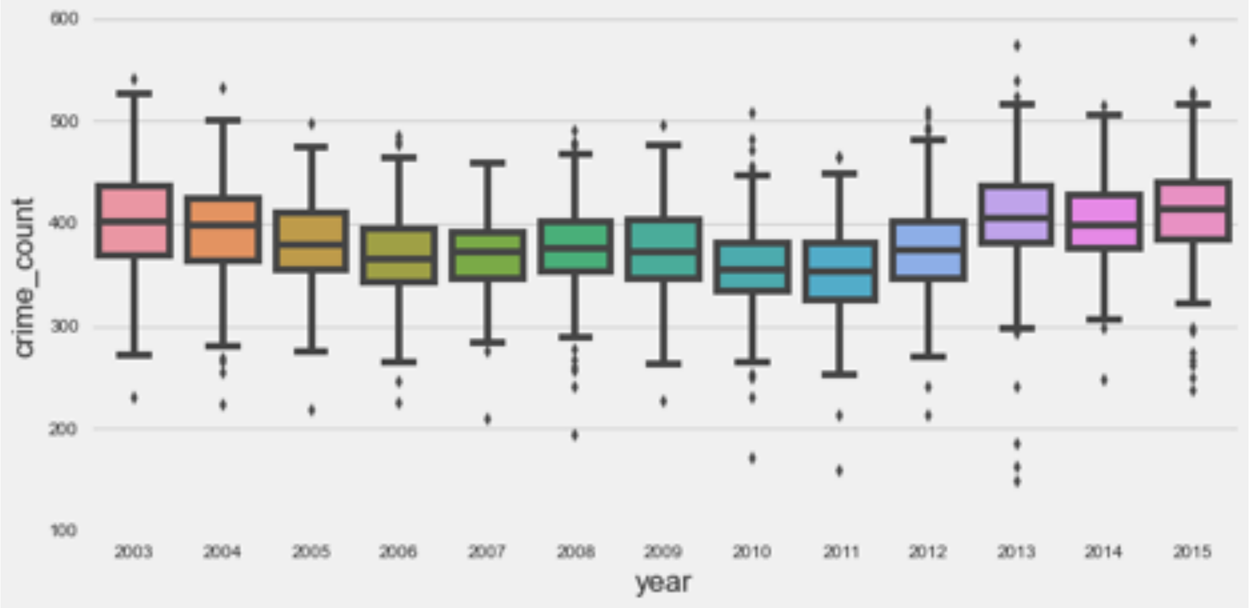
Data was then written to file. **sf_weather_clean.csv**

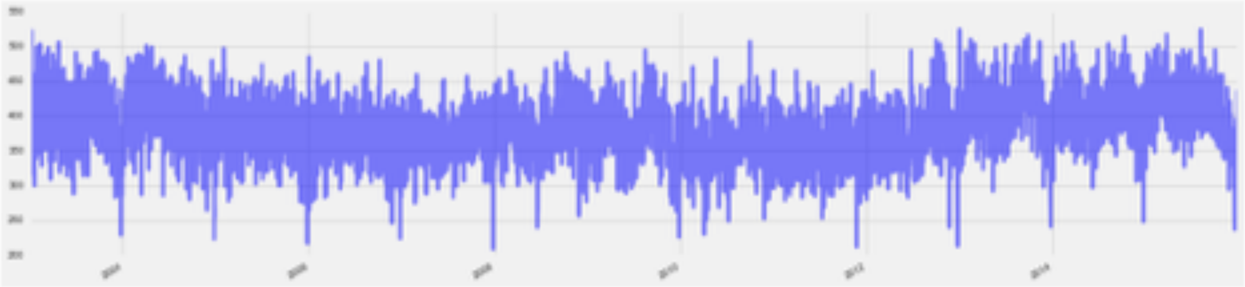
The weather data and crime data were merged and written to the file. **sf_crime_weather.csv**

Looking at the Data

The analysis was done in the **5_analysis_cleaning** ipython notebook.

First looked at the yearly stats to see if crime levels have changed much over the last 13 years.



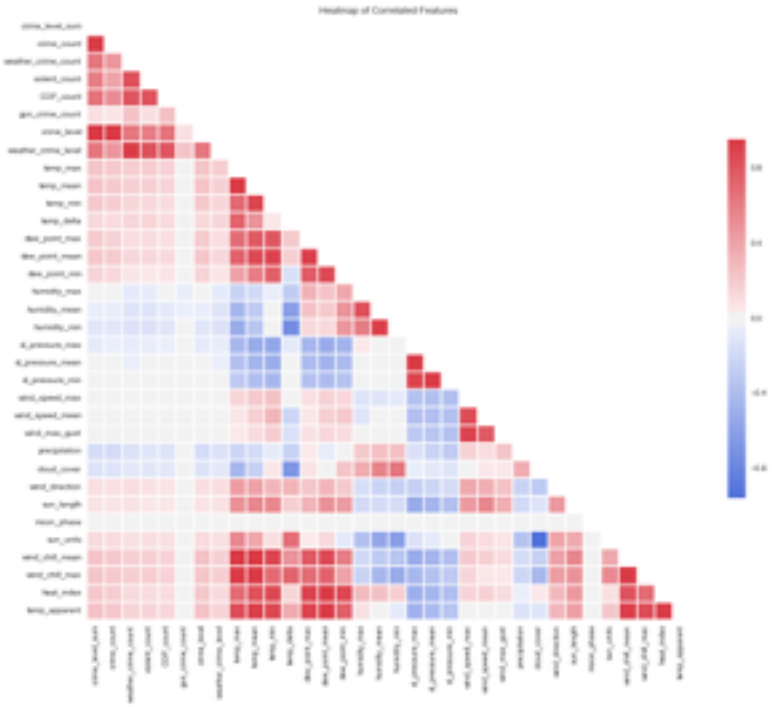


There has been some up and downs (Slight dip in 2010-2011). But not that great. I was concerned that changes in the way the police department was run might have effected the way crimes were reported.

Can also see repeated drops near end of year.

Correlations

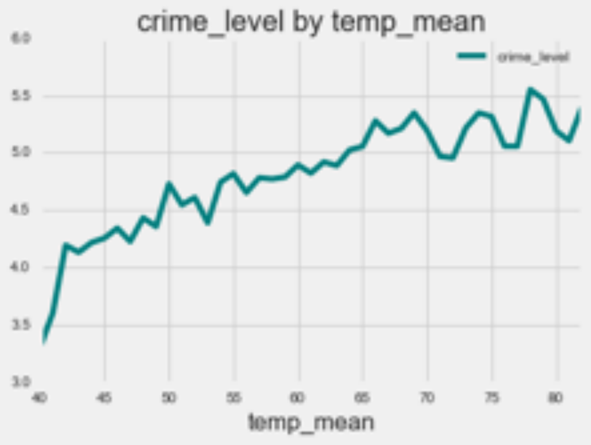
I was very concerned about correlated features. The max, mean and min values for Temperature and Humidity etc. will be heavily correlated. The calculated features such as Wind Chill and Heat Index are all based on Temperature, Humidity and Wind Speed. Wanted to know if they have a halo effect that helps or just they just add more variance.

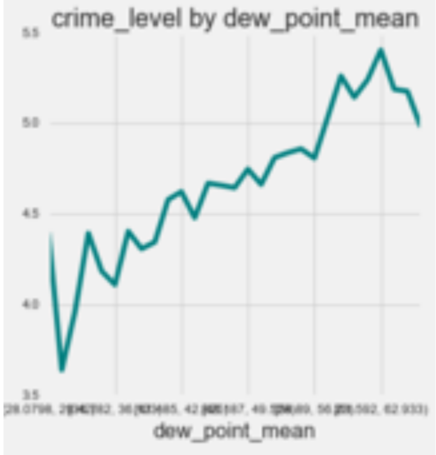


Relations to Crime levels

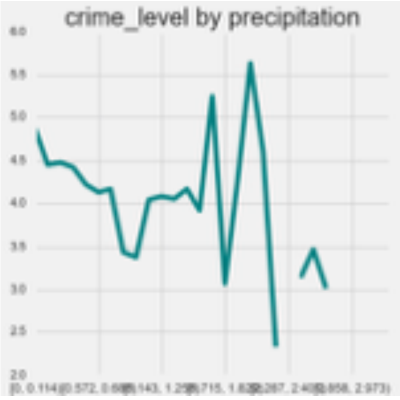
Next I looked at the individual features to see if any looked like they were correlated to the crime levels.

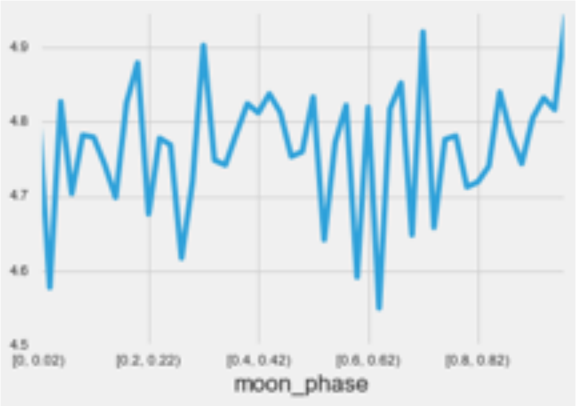
Several look promising. The temperature features, cloud cover, dew point, wind chill.





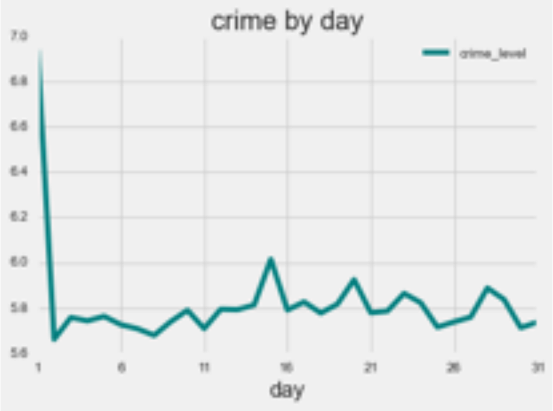
Others did not look promising. Including the moon phase. Perhaps if I went back and pulled out only the night time crimes I could find something.





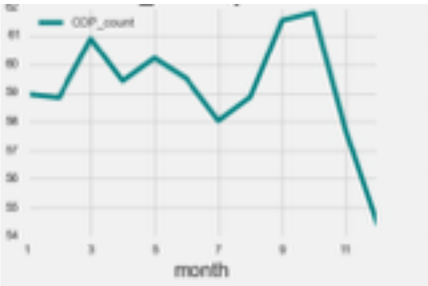
Problems

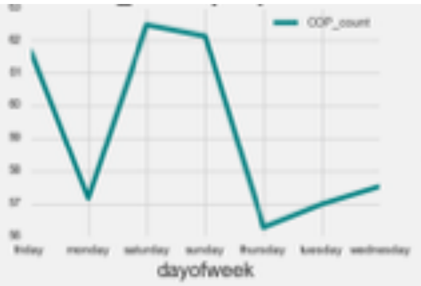
While looking at the date features I found a large spike in crimes that occurred on the first day of the month. I went back to the raw data to investigate. There were many incidents on those first days that had a time stamp of 00:01. My belief is that some incidents were not recorded at the right time and were put in just the month. I removed these.

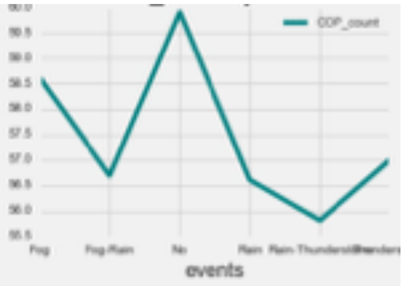


Categorical Features

I checked the categorical feature to see if they should be expanded with dummy features or consolidated into one or two features.







Decided to create several booleans to replace these features. One for winter months and one for rain events. Crime appears to go down then. Makes sense. Crime goes up near the weekend. Friday and Saturday are when everyone is out and about.

Feature and Model Selection

Models

After the collection, cleaning and engineering I ended up with 34 features and six different crime variables to model.

For model selection I had the following concerns:

- Correlated data - Many of the features were based on temperature and humidity.
- Outliers - I was not sure if all the data was entered correctly.
- Complex/Nonlinear - Humidity can make the hot feel hotter and the cold seem colder. How these effect human emotions is also very complex.
- Ease of interoperability - I wanted to see which features canceled each other out and which ones had the greatest effect.

These concerns pointed me toward SVM or Random Forests. But because of the interoperability I kept a Lasso Regression in mind.

Features

For the first run of the data I tried just putting in every feature. (I had to give it a try.) The results barely beat the dummy model which just chooses the mean.

Because of my concern with all the correlated features, I started over with just the minimum weather related features. Temperature , humidity, wind speed, dew point, pressure. I then started adding features. First the calculated features and then the date features.

Crime Variable selection

I did run model on all of the different variables, but I focused on just the general number of Crimes and Crimes Of Passion count. (COP found crimes that matched a keyword list)

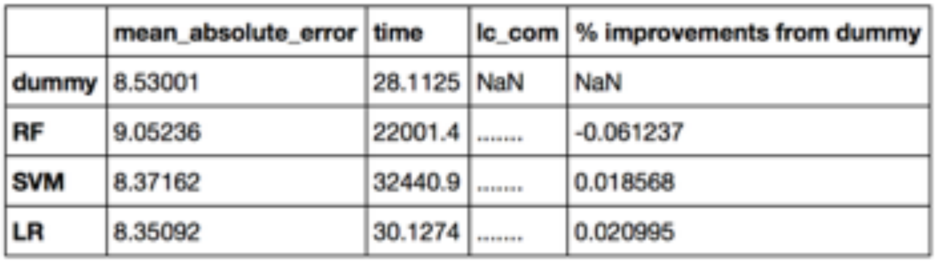
Evaluation parameters

For evaluation I chose Mean Absolute Error. MAE would allow me to talk in terms of number of crimes. I was also not concerned with penalizing large errors more than small.

Findings

Features

Starting with just a few of the basics had almost no effect. Only a couple Percentages better than the dummy.

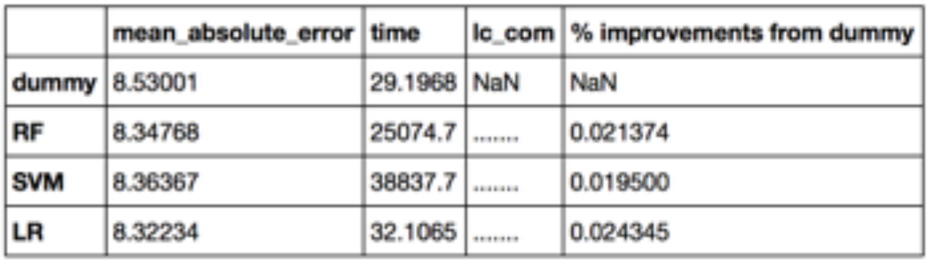


	mean_absolute_error	time	lc_com	% improvements from dummy
dummy	8.53001	28.1125	NaN	NaN
RF	9.05236	22001.4	....	-0.061237
SVM	8.37162	32440.9	....	0.018568
LR	8.35092	30.1274	....	0.020995

I added in more features like the cloud cover and sun length. There was a tiny improvement in a couple of models, but no overall improvement.

I added in the engineered features and also say very little change. Looking at LR coefficients showed that as I added some of the engineered features (heat index) the regular ones, (Temperature) were zeroed out.

The best I was able to get with the weather features was:

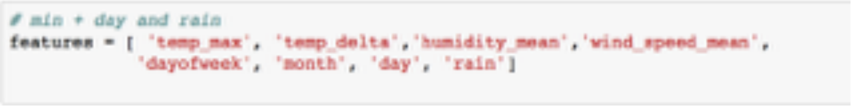


	mean_absolute_error	time	lc_com	% improvements from dummy
dummy	8.53001	29.1968	NaN	NaN
RF	8.34768	25074.7	....	0.021374
SVM	8.36367	38837.7	....	0.019500
LR	8.32234	32.1065	....	0.024345

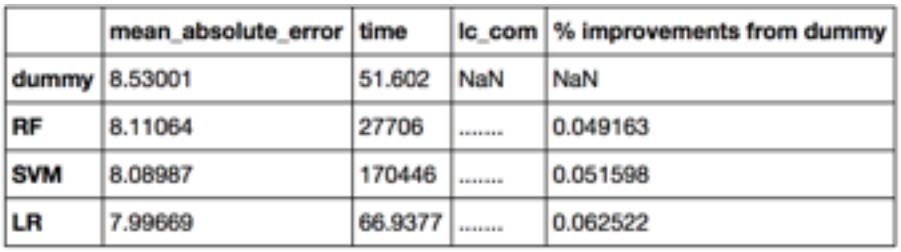
Next I added in the date features and found a larger impact. I ran the model with just the date features and got the improvements of 4% - 5%.

I tried replacing the 7 sparse features for the day of the week with a single weekend feature and found the performance suffered. The extra information for the date features seems more important than reducing the features.

Putting them all together looked like this:



```
# min + day and rain
features = [ 'temp_max', 'temp_delta','humidity_mean','wind_speed_mean',
            'dayofweek', 'month', 'day', 'rain']
```



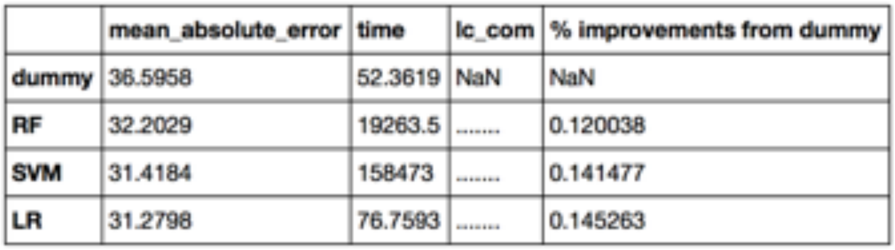
	mean_absolute_error	time	lc_com	% improvements from dummy
dummy	8.53001	51.602	NaN	NaN
RF	8.11064	27706	....	0.049163
SVM	8.08987	170446	....	0.051598
LR	7.99669	66.9377	....	0.062522

I also played with removing several levels of outliers, but did not find a significant difference.

I tried SVA feature reduction and also found a drop in performance.

Crime Variables

Also went back and ran models on the total crime levels and saw better performance than with the more focused COP crime level.



	mean_absolute_error	time	lc_com	% improvements from dummy
dummy	36.5958	52.3619	NaN	NaN
RF	32.2029	19263.5	....	0.120038
SVM	31.4184	158473	....	0.141477
LR	31.2798	76.7593	....	0.145263

This was probably due to the fact that the non weather features had more influence. The more regular data there was the more these features could contribute and the better trained the model could be.

You should also remember that the officers are people too. The weather could also effect them when it comes to non violent crimes. When it is raining and cold they may be out less. When it is hot and humid they may be less likely to let people off with a warning.

Models

Was surprised to find that in my final configuration all the models (Random Forests, SVM and Lasso Regression) came in at basically the same MEA. But since the performance is so low I should not be surprised that they all preformed at the same low level. I did depend on Lasso to help select the features, so if I went back and worked on features with one of the more complex models I might get better results.

But from what I have now it seems that Lasso Regression is the best model. It has the same results, is much faster and gives me the coefficients which can help me explain the results.

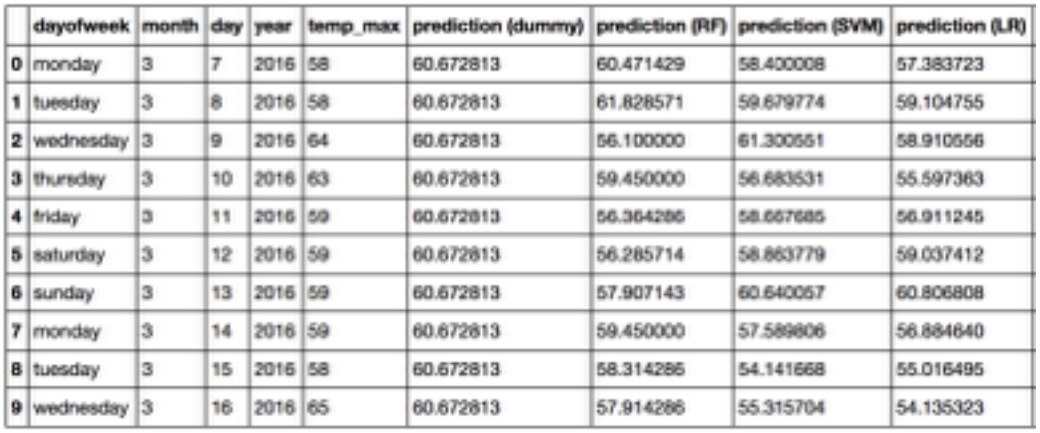
Predictions

I collected information using the Weather Underground APIs with json and request.

I collected several other features not included from above from Open Weather Map APIs using the pyowm wrapper.

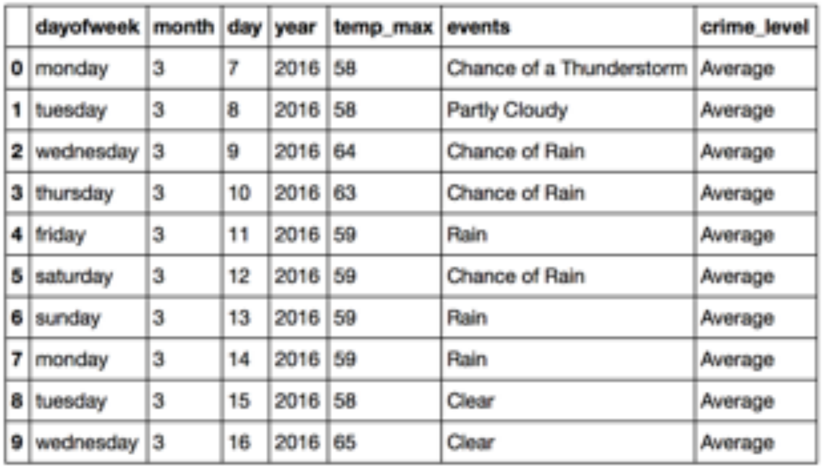
I then had to add in all the Engineered features as I did for the Modeling data.

With these I was able to make my predictions based on the 10 day forecast. But with the large MAE predicting an accurate number of crimes is not possible. The best we can get is a general range. Perhaps if the day will be better or worse than average.



	dayofweek	month	day	year	temp_max	prediction (dummy)	prediction (RF)	prediction (SVM)	prediction (LR)
0	monday	3	7	2016	58	60.672813	60.471429	58.400008	57.383723
1	tuesday	3	8	2016	58	60.672813	61.828571	59.679774	59.104755
2	wednesday	3	9	2016	64	60.672813	56.100000	61.300551	58.910556
3	thursday	3	10	2016	63	60.672813	59.450000	56.683531	55.597363
4	friday	3	11	2016	59	60.672813	56.364286	58.667685	56.911245
5	saturday	3	12	2016	59	60.672813	56.285714	58.863779	59.037412
6	sunday	3	13	2016	59	60.672813	57.907143	60.640057	60.806808
7	monday	3	14	2016	59	60.672813	59.450000	57.589806	56.884640
8	tuesday	3	15	2016	58	60.672813	58.314286	54.141668	55.016495
9	wednesday	3	16	2016	65	60.672813	57.914286	55.315704	54.135323

Then based on the range of the variables I turned the prediction into a general class of ‘High’, ‘Average’ and ‘Low’.



	dayofweek	month	day	year	temp_max	events	crime_level
0	monday	3	7	2016	58	Chance of a Thunderstorm	Average
1	tuesday	3	8	2016	58	Partly Cloudy	Average
2	wednesday	3	9	2016	64	Chance of Rain	Average
3	thursday	3	10	2016	63	Chance of Rain	Average
4	friday	3	11	2016	59	Rain	Average
5	saturday	3	12	2016	59	Chance of Rain	Average
6	sunday	3	13	2016	59	Rain	Average
7	monday	3	14	2016	59	Rain	Average
8	tuesday	3	15	2016	58	Clear	Average
9	wednesday	3	16	2016	65	Clear	Average

This caused me to go back and look at the actual prediction numbers from the test set and realize that the models predictions barely varied from the mean. So all predictions were ‘Average’.

Challenges and Unknowns

There are several factors that concerned me when taking on this project. There were also some that showed up once I got started.

Policing Policies

I do not know how accurate the reporting by the police department was. 13 years is a long time and many things may have changed.

- Police recording procedures. - Technology has certainly changed over this time.
- New elected officials - New Police Chiefs and Mayor may have different agendas that change the arrest patterns
- New Policing Techniques - Broken Windows - Stop and Frisk - They can certainly change patterns

Additional factors in Crime

Even at the beginning I knew that the weather would only play a small and subtle part in crime patterns.

- Economic Conditions - Crime usually goes up in bad times.
- Political Environment - Changes in benefits and policies can effect people very much.
- Longer term : state of the schools in SF - effects peoples outlook for the future.

Using SF as Test city

San Francisco may not have been the best choice for this analysis. Although San Francisco has more varied weather than cities in Southern California, it is still pretty mild. A city like New York has much more extreme weather and would have been a better choice. (I was not able to get the same level of crime detail for New York)

How complex human emotions are

The biggest problem is that humans are complicated. Trying to judge what they will do on any one factor is a big challenge.

I also realized that Police Officers are human also. No matter how well trained they are they can be effected by the weather also.

Going Forward

I feel that there is much more that could be done with this project. There are several thing that could be improved and many that

Analysis on the weather features

I would like to spend more time and get a better understanding of the effects of the individual features.

Cumulative effects

Anger and emotions are not always instantaneous, they can build up over time. It would be interesting to look back over the last few days to look for cumulative effects. The data is already available and for the prediction would only take one more API call to get the last few days history.

Change Test city

A city with weather swings such as New York, Boston or Chicago would be better for finding the weather relationships.

Investigate factors other than weather

If the concern is to make a general crime predictor there are other factors that could be added. Economic factor and others may help with predictions, but that is a much bigger project.

Refining the data and selecting crime variable

There is much that could be looked at in the crime data.

- Further investigation into the types of crimes would help in knowing what type of officers are needed. (Street Cop, Detective, Desk Officer)
- Looking at location would be helpful in precinct staffing and possible closings
- Time of day analysis could help with shift staffing. Looking at night time crimes may finally reveal the ‘lunar’ effects.

Conclusions

I was not able to get a very definitive crime prediction using the weather data. It does give some indication of if the day will be better or worse than average, but it may not be any better than the intuition of an experienced police officer. I still feel that I found that there is a relationship between weather conditions and crime. I believe that the problem has many other factors that are more dominate than the weather. In my tests I found that the day of the week had a much stronger correlation than the weather. I'm sure that economic conditions and the political environment also play a large roll.

I did see connections with rain and higher temperatures, but I'm sorry to say that, I did not find that the moon caused all the lunatics to come out.