Memorandum From:

EVAN S. LEVINE, Director of Analytics

[redacted]

Redacted for Algotithm

Memorandum From:

EVAN S. LEVINE, Director of Analytics

[redacted]

Redacted for algorithm

Hot Spots and "Predictive Policing"

Using 500 ft by 500 ft grid, try to identify in advance the boxes where crime will occur.

Use training data set to identify 10 boxes. How much crime in test data set occurs in those boxes?

Training Data Set

- 28 day period: 5/4/2014 - 5/31/2014
- Complaints for Grand Larcenies, Robberies, and Burglaries
- Citywide
- Threw out 61's with bad geocodes

Test Data Set

- 7 day period: 6/1/2014 – 6/7/2014

Preliminary Analysis

Best possible – This is the theoretical best we can do by choosing 10 500 ft by 500 ft boxes in the grid

Random – This is the worst we can do, just picking 10 boxes at random

Hot spots – Choose the top 10 boxes based on the number of complaints during the test period

"Predictive" v1.0 – Choose the top 10 boxes based on the number of complaints during the test period, weighted by how long they occurred in the past

Algorithm	Robbery	Burglary	Grand Larceny
Best possible	61%	65%	44%
Random	<0.1%	<0.1%	<0.1%
Hot spots	8%	6%	14%
"Predictive" v1.0	10%	9%	14%

Algorithm Comparison

Parameters				Results							
Algorithm	Lookback (days)	Time weighting	Time blocking	Burglary	Fel Assault	Grand Larc	GL of Motor Veh	Murder	Rape	Robbery	Shooting
Hot Spots	[redacted]	[redacted]	[redacted]	8.1%	10%	15.4%	3.6%	0%	1.9%	8.4%	3.5%
Hot Spots b				8.7%	11.8%	17.0%	4.8%	0%	3.0%	9.8%	3.9%
Hot Spots c				9.0%	13.3%	18.3%	4.8%	0%	5.7%	10.7%	6.6%
Hot Spots d				9.6%	14.6%	19.9%	5.2%	1.9%	8.4%	12.8%	10.6%
Hot Spots e				10.8%	17.1%	21.9%	6.2%	5.8%	11.4%	14.3%	11.6%
Predictive v1.0				8.5%	10.1%	15.3%	3.9%	0%	1.9%	8.5%	3.5%
Predictive v1b				9.6%	11.5%	17.1%	4.8%	0%	3.0%	10.0%	4.6%
Predictive v1c				9.7%	12.7%	18.2%	5.6%	0%	5.7%	10.8%	6.4%
Predictive v1d				9.8%	14.6%	20.0%	5.4%	1.9%	7.9%	12.6%	10.1%
Predictive v1e				10.8%	16.5%	21.7%	6.2%	5.8%	11.2%	14.7%	14.3%
Predictive v2c				6.4%	8.8%	15.4%	2.3%	0%	0.3%	6.8%	0.2%
Predictive v2d				8.0%	10.6%	17.5%	4.0%	1.0%	0.8%	9.3%	1.4%
Predictive v2e				9.2%	12.9%	20.5%	5.1%	1.9%	2.5%	11.3%	3.1%
Predictive v3a				3.7%	4.1%	11.4%	1.0%	0%	0%	3.3%	0.2%
Predictive v3c				6.6%	8.4%	15.6%	2.3%	0%	0.3%	7.0%	0.2%
Predictive v4e				10.7%	16.0%	21.2%	6.4%	5.8%	11.4%	14.4%	14.7%

[redacted]: Shows algorithm Details of the

Algorithm Comparison (vis by Patrol)

Parameters				Results							
Algorithm	Lookback (days)	Time weighting	Annual boost	Burglary	Fel Assault	Grand Larc	GL of Motor Veh	Murder	Rape	Robbery	Shooting
Hot Spots	[redacted]	[redacted]	[redacted]	3.2%	7.5%	11.2%	3.1%	0%	0%	6.3%	3.5%
Hot Spots b				4.5%	9.7%	13.2%	3.5%	0%	0%	8.7%	3.9%
Hot Spots c				4.9%	10.8%	14.8%	3.8%	0%	0%	9.2%	6.6%
Hot Spots d				5.9%	12.0%	16.8%	5.0%	1.8%	0%	12.2%	10.6%
Hot Spots e				6.7%	14.7%	19.9%	5.4%	7.0%	0%	13.1%	11.6%
Predictive v1.0				3.2%	7.3%	11.1%	3.1%	0%	0%	6.4%	3.5%
Predictive v1b				4.7%	9.5%	14.0%	3.9%	0%	0%	8.7%	4.6%
Predictive v1c				5.1%	10.5%	15.0%	4.4%	0%	0%	9.3%	6.4%
Predictive v1d				6.8%	12.3%	17.2%	5.2%	1.8%	0%	11.6%	10.1%
Predictive v1e				7.0%	14.3%	20.2%	5.5%	7.0%	0%	13.4%	14.3%
Predictive v4e				7.2%	14.2%	19.2%	5.5%	7.0%	0%	12.9%	14.7%
Predictive v5e				6.5%	14.9%	20.2%	5.7%	7.0%	0%	12.9%	13.3%

shows algo detail rithm of the

Predictive Policing

July 14, 2015

Predictive Policing Vendors

- Several vendors sell predictive policing software:
  - PredPol (LA, Santa Cruz)
  - Hunchlab (Philadelphia)
  - Public Engines (Park City, Leesburg)

- These solutions have several weaknesses:
  - Not designed for verticality of NYC
  - Limited to a subset of crimes
  - Require sending NYPD data off-site to a vendor
  - Black-box algorithms are not compelling for CO's and cannot be adjusted by NYPD

NYPD's Predictive Policing

- We studied various algorithms to determine which could most accurately predict crime locations using 18 months of NYPD 61s
- We determined we could improve on accuracy of hot spot policing with simple weighting algorithms (more detail in last slide)
- The chart below shows the percentage of complaints that occurred in the top 10 highlighted boxes using the two methods:

	Burglary	Fel Assault	Grand Larc	GL of Motor Veh	Murder	Rape	Robbery	Shooting
28 Day Hot Spots	3.2%	7.5%	11.2%	3.1%	0%	0%	6.3%	3.5%
NYPD Predictive	7.2%	14.9%	20.2%	5.7%	7.0%	0%	13.4%	14.7%
% Improvement	125%	99%	80%	84%	-	-	113%	320%

- NYPD's predictive algorithm was more accurate than hotspots for each of the 7 majors and shootings (excluding rape, which is rarely visible by patrol)

Example (Week of July 8, 2014)

- We pulled 28 days of citywide complaints and shootings (6/10/2014 - 7/7/2014) to determine the top 10 500 ft by 500 ft hot spots for each precinct for each crime type
- During the next 7 days, the percentage of complaints and shootings that fell into those hot spots were:

	Burglary	Fel Assault	Grand Larc	GL of Motor Veh	Murder	Rape	Robbery	Shooting
28 Day Hot Spots	6.7%	6.4%	11.0%	2.9%	0%	0%	4.2%	7.1%

- We then used our predictive algorithm to determine 10 boxes for each precinct for each crime type. The percentage of complaints and shootigns that occurred in these predictive boxes were:

	Burglary	Fel Assault	Grand Larc	GL of Motor Veh	Murder	Rape	Robbery	Shooting
NYPD Predictive	8.3%	10.7%	21.5%	5.8%	25%	0%	11.3%	21.4%
% Improvement	24%	67%	95%	100%	-	-	169%	201%

Future Work

- NYPD's predictive policing algorithms can be improved by further work on:
  - Pulling key words from 61 narratives
  - Tuning the algorithms using machine learning
  - Expanding to include CIC and CIP 911 calls
  - Integrating with AVL (and eventually phones) to measure time spent in highlighted areas

Algorithm Details

- The score for each box is calculated using a weighted sum for each record in the box: $S_j = \sum_i n_{i,j} w_i$
- w_i is the weight associated with the ith event in the data set. The weighting formula varies depending on the crime type. d_i is how many days ago the event happened.
- For Grand Larceny and Robbery:

[redacted]

- For Burglaries, Shootings, Murder, and Rape: [redacted]
- For Felony Assaults and Grand Larceny of Auto:

[redacted]

NYPD's Predictive Policing (review)

- We studied various algorithms to determine which could most accurately predict crime locations using 18 months of NYPD 61s
- We determined we could improve on accuracy of hot spot policing with simple weighting algorithms
- The chart below shows the percentage of complaints that occurred in the top 10 highlighted boxes using the two methods:

	Burglary	Fel Assault	Grand Larc	GL of Motor Veh	Murder	Rape	Robbery	Shooting
28 Day Hot Spots	3.2%	7.5%	11.2%	3.1%	0%	0%	6.3%	3.5%
NYPD Predictive	7.2%	14.9%	20.2%	5.7%	7.0%	0%	13.4%	14.7%
% Improvement	125%	99%	80%	84%	-	-	113%	320%

- NYPD's predictive algorithm was more accurate than hotspots for each of the 7 majors and shootings (excluding rape, which is rarely visible by patrol)

NYPD's Predictive Policing (update)

- Received feedback on first production version from NYPD stakeholders, including Chief of Dept, DC Ops, OMAP
- Based on their interests, we focused on improving the shooting prediction algorithm
- Predictive v1 used the date of the shooting and weighted based on how far in the past the event took place
- Predictive v2 Beta adds:
  - Structured field data from the shooting (gang, club, domestic, primary motive)
  - Seasonal variations
  - 911 calls for shots fired and their disposition
  - Automatically tuned weighting (machine learning)

Algorithm Comparison (Shootings)
Preliminary Results

Algorithm	Shootings
28 Day Hot Spots	3.5%
Predictive v1 (in production now)	14.7%
Predictive v2 Beta	20.0%

- Predictive v2 Beta is 470% better than 28 day Hot Spots and 36% better than the predictive algorithm currently in production
- As a reminder:
    - We measure results by calculating the percentage of shootings that occur in the top 10 500 ft by 500 ft boxes, per precinct per week, from 4/1/14 to 3/30/15
    - Hot Spots simply totals the number of events during the lookback period

Predictive Policing Time Weighting Functions as of 10/20/2014

For all equations, w_i is the weight associated with the ith event in the data set. d_i is how many days ago this event happened.

1. For Grand Larceny and Robbery:

Redacted for Algorithm

2. For Burglaries, Shootings, Murder, and Rape:

Algorithm

3. For Felony Assaults and Grand Larceny of Auto:

Algorithm

We have a few questions that we need to get clarification on regarding the specs in the requirements document.

They are as follows :

1) What is the date range for the historical data 1/1/2013
2) Will the data geography cover the entire NYC area
3) Will each crime incident record have (*DateTime, precinct id, latitude, longitude, crime_category* on it) X,Y
4) What time of the day will NYPD provide us with the records they want predictions on for 0600 delivery 0100

5) [Point 8 in requirements doc] **Predictions should be made for 5 grid cells in EACH precinct, for Each shift for each crime type. We interpret as follows :**
- We will provide NYPD with 385 grid cells for the morning shift for the robbery crime type. (*77 precincts x 5 grid cells each = 385 grid cells*)
- The same will be done for each shift by crime type combination. Is our interpretation correct ?

6) [Point 9 in requirements doc] **In addition, predictions should be made for the top 1% of the land area of each borough for EACH shift, for EACH crime type.**
We interpret as follows:
- For the borough of Brooklyn, regarding robbery crime type in the morning shift, we will rank order all the grid cells from highest robbery score to lowest
robbery score. We will then select the top robbery score grid cells down to where the grid cells accumulate to 1% of the land mass of Brooklyn.
- The same will be done for each Borough by Shift by Crime type Cohort. Is this interpretation correct ?

7) [Point 2 in requirements doc] **Vendors will provide a secure FTP server for the NYPD to upload daily crime data. The NYPD will need the destination IP address, which should be static, port number, and credentials to login.**
- What is the file format for the files we will be receiving from NYPD
a) Crime History file CSV
b) Additional data layers files (liquor, facility locations etc)
- Can we get layouts for the files

8) [Point 7 in requirements doc] **Daily predictions, broken down for all shifts, must be received by 0600 in order for them to be counted for that day**
- What is the file format and layout you would like us to provide you at 0600 each day

ADDITIONAL QUESTIONS FROM NOVIANT

- Can you provide an estimate of the size of the initial historical data file? 20 MB
  - Will that file stay static for the demonstration period? If there are updates – can you describe approximate file size of the daily updates? 200K
  - The daily file transfers – can you provide an estimate of the daily file transfer sizing?
- Does NYPD anticipate any other data feeds or data sources to be used as part of the demonstration project?
  - Examples:
    - Newsfeeds
    - Social Media
    - Data feeds
    - Other Data Bases – Corrections, Parole, ICE, etc.
- Does NYPD have existing user licenses for Tableau Desktop?
  - Will the NYPD demonstration project license users need training for reporting and visualization?
    - Will NYPD share or discuss what the report templates will look like?
    - Will NYPD identify reporting requirement by user level?
- What type of training and support will NYPD require during the demonstration period?

Azavea's signed security control certification is attached for your records. We are also placing the original in the mail to your attention. In addition, we have some questions that we hope you can answer for us regarding the requirements document you provided. We want to be sure that everything proceeds to your expectations. Please see our questions below:

1. We are just confirming that the incoming data will contain a unique record identifier so that we can ensure that records are not counted twice if updates to records result in the record appearing in multiple day's upload files. Will that be the case? Yes
2. The directions state that our predictions must be received by 0600. By what time will crime data files be uploaded to our FTP server?
3. By default, our mission cells are presented in GeoJSON format within our application. Is this format acceptable for our submission to NYPD, or do you prefer another format?
4. By what mechanisms can we deliver the missions to NYPD?
5. We have created credentials for our FTP server. How do you prefer that we relay the password to you?