Conduent Public Safety Solutions
Department of Justice – Office of Justice Programs

Real Time Crime Forecasting Challenge

Post-Mortem Analysis
Challenge Performance

Contents

Contents **2**

Executive Summary **3**

Project Success begins with Sponsor Leadership 3

Challenge Problem **4**

Solution Approach **5**

Ground Truth 5

Algorithm 5

Feature Set 7

Challenge Performance **8**

Conclusion and Next Steps **10**

Why Should DOJ-OJP Consider Conduent as a Partner of Choice? 10

Next Steps 10

Contact **11**

Executive Summary

Conduent deployed functionality from the Operational Analytics module within the Conduent Business Intelligence platform (CBI) to the challenge. CBI is a highly configurable flexible platform that provides a user friendly interface for running machine-learning based analytics. This challenge focused on crime data exported from a Record Management System (RMS). We further refined our predictive models with Census tract data which provides population, education, and OpenStreetMap data that provides points of interest. Our mission is to highlight fact based, yet often unintuitive, actionable insights. Some of the analytical features used for the challenge include (1) Clustering of law enforcement agencies with similar patterns of crimes, effectively peer ranking predictive models, (2) benchmarking of agencies based on relevant Key Performance Indicators, (3) identifying the most important domain-relevant attributes (such demographics, functional nature of the city, etc.) that influence the crime patterns of the regions, (4) hotspot prediction to forecast the time and type of the next likely crime event.

In this document we provide the technical details of the last feature to provide context to our methods, i.e., the hotspot prediction.

Project Success begins with Sponsor Leadership

We wanted to take a moment to thank the National Institute of Justice for sponsoring this very important challenge. We hope that you were able to accomplish your goal of advancing place-based crime forecasting. We have learned valuable lessons using new data and have been able to evolve our core product CBI because of the lessons learned in the challenge. Your leadership made this possible, Thank-you!

Challenge Problem

The NIJ challenge gave the participants the option of predicting the hotspots in Portland, OR for a window of 1 week, 2 weeks, 1 month, 2 month or 3 months. The prediction could be for an individual crime type (burglary, street crime or motor vehicle theft), or for all the crime types combined. The data provided was the Call For Service (CFS) records from the Portland Police Bureau for the period of five years, from January 2012 to February 2017. The predictions were to be made for the chosen period starting from March 1 2017. A hotspot does not have a well-defined definition. The region of interest (Portland, OR) can be divided into cells of uniform (but any) shape and area. Each cell had to be in within the range of 62,500 – 360,000 ft^2. Out of the total area of around 148 $miles^2$, only an aggregate area within 0.25-0.75 $miles^2$ could be forecast as hotspots. The evaluation was in terms of two metrics, namely (i) Prediction Accuracy Index (PAI) and (ii) Prediction Efficiency Index (PEI)[1].

[1]Formulas can be found here https://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge-document.aspx#judgingcriteria

Solution Approach

The initial set up around our solution included aggregating the data at a monthly level, and selecting square-shaped cells of the largest allowable size, i.e., 360,000ft^2. This choice of size for a cell was data-driven, where we observed that any grid size smaller than this yielded more number of cells with zero crimes, or flipped frequently from month to month from being a high crime area to a low crime area. A cell size of 360,000ft^2 yielded a relatively optimal density of crimes and crime distribution across all grids, which helped bring out the consistent variations in crime densities across the grids. With this size, the challenge translated to forecasting between 19 to 158 cells out of a total of 11450 cells.

Ground Truth

The first step in our solution was to determine the ground truth, i.e., what made a cell a hotspot for any given month. We use the straightforward heuristic of ranking the cells in the order of crimes and taking the cells in the top 19 ranks as the hotspots. The rationale behind this approach was that this yielded the optimum values for both PAI and PEI - taking the cells with highest number of crimes optimized the PEI and assigning the minimum allowable number of cells (19) as hotspots optimized the PAI.

Algorithm

It was observed that some cells were consistently cold-spots, and some were consistently hotspots, across all months. There were also some cells that flipped from being a hotspot to not being one, each month as shown in Figure 1. These varying cells were the ones that are difficult to predict. Hence, we cast this as a binary classification problem. We used support vector machines (random forest classifier yielded similar results), to train and test the data from January 2012 to December 2016 with five-fold cross validation, to predict March 2017. We included the following optimizations to improve the classification accuracy:

- We did not use the last two months' data, i.e., January and February 2017, in order to avoid over-fitting[2].
- We removed those cells that were, and all their neighboring cells were, always a hotspot or always a coldspot.

---

[2] Burnham, K. P.; Anderson, D. R. (2002), *Model Selection and Multimodel Inference*(2nd ed.), Springer-Verlag

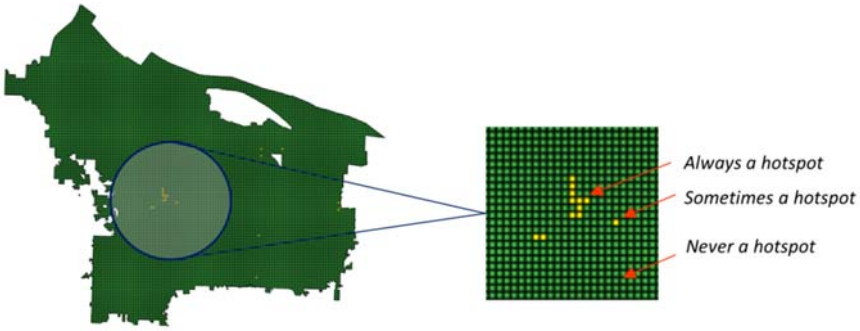

Figure 1

When more than 25 cells were predicted as hotspots, we used the probability of positive prediction obtained from the SVM model to rank the outputs and subsequently prune the result set. The value 25 was derived by allowing a 5% error beyond the optimal number of hotspots. The minimum number of hotspots, 19, is the optimal number if they are all accurately predicted because that would yield the highest possible PEI and PAI. However, mispredicting hotspots would adversely impact the PEI, while allowing some misprediction margin by increasing the number of hotspots would adversely impact the PAI. Keeping this tradeoff in mind, we allowed a 5% increase in the number of hotspots, to balance the loss in PAI and gain in PEI. The steps in our solution as illustrated in Figure 2.

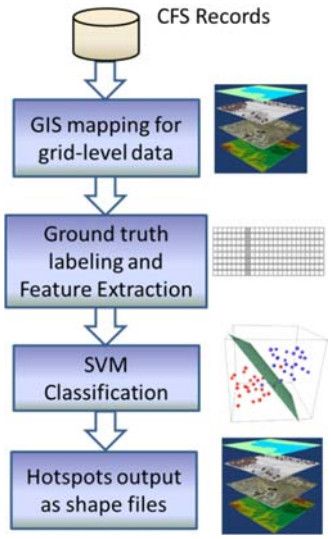

Figure 2

Feature Set

Next, we describe one of the most important aspects of our solution, which is the feature set. The feature set that we used are of three categories (i) statistical features, (ii) landmark features and (iii) neighborhood features. For statistical features, we computed the following metrics, each for time windows of (the past) one month, two months, three months and one year:

(i) *Mean number crimes*
(ii) Frequency of hotspots, i.e., the number of times the cell has been a hotspot in each time window.
(iii) Normalized number of crimes
(iv) Rank of the cell, with the cell with highest number of crimes being rank 1 (ties were broken with the average rank in the past)

For landmark features, we computed the following metrics, which are all categorical features and we represented them using one hot encoding (i.e., binary values taking values true or false) for better efficiency:

(i) Police stations
(ii) Schools or colleges
(iii) Airport
(iv) Train station
(v) Residential areas
(vi) Industrial areas
(vii) Downtown or Active Nightlife area
(viii) River side
(ix) Highway

For neighborhood features per cell we computed the following metrics as the average over all neighboring cells. A neighboring cell is defined as a cell that shares a common boundary with the cell:

(i) Average number of crimes for this month
(ii) Average number of crimes in the past one, two and three months and one year.
(iii) Average rank of neighboring grids in the previous one, two, three months and one year.

Challenge Performance

We submitted the forecast for the month of March 2017 for the category of street crimes. We predicted 20 hotspots. When the ground truth, i.e., the actual number of crimes results, was released by NIJ, we observed that based on our definition of hotspot, we got 14 hotspots correct. Our PAI score was 7.4 (out of 7.8), while the PEI score was 0.94 (out of 1). We won the challenge under the category of PEI. We note that we predicted all the top 13 cells (i.e., those cells with the 13 highest numbers of crimes for March 2017) correctly. The remaining cells were mostly within the top 20. However, one of our mispredicted hotspots was according to the ground truth, in rank 51 (Grid ID 13759 in the downtown area, as shown in Figure 3) , which affected the PAI score the most. In other words, we gained on the PEI because of correctly predicting the higher rank grids, but lost on PAI because of incorrectly calling out lower rank grids as hotspots. The table in Figure 3 shows the grids with highest number of crimes for March 2017, according to the ground truth, i.e., the ground truth rank is listed in the first column. The last column is our forecast. (D), (C), (M), (L) indicates the geographical location of the cells, standing for Downtown, Centennial, Montavilla and Lents areas, respectively.

Rank	GRID ID	NUMBER OF CRIMES	HOTSPOT PREDICTION
1	13434(D)	28	1
2	13271(D)	22	1
3	12944(D)	21	1
3	14081(C)	21	1
5	13834(D)	19	1
5	14080(C)	19	1
7	13596(D)	17	1
8	14161(C)	16	1
9	13107(D)	14	1
10	13270(D)	13	1
10	13272(D)	13	1
12	13765(D)	12	1
13	16082(L)	11	1
13	10052(M)	11	0
13	13597(D)	11	0
13	14323(C)	11	0

Rank	GRID ID	NUMBER OF CRIMES	HOTSPOT PREDICTION
13	19015(L)	11	0
18	**13433(D)**	10	1
18	10541(M)	10	0
18	12781(M)	10	0
24	13432	9	1
24	13756	9	1
28	12943	8	1
28	13435	8	1
37	13492	7	1
51	13759	5	1

Figure 3

Conclusion and Next Steps

The challenge successfully demonstrated the technology works well. We recorded high precision and recall in a new application for the technology. We had a solution that was new and novel but proving its viability was extremely important to us and we look to expand CBI based on lessons learned with this challenge. Thank-you for the opportunity to participate.

In conclusion, this challenge gave us important insights into the practical aspects of hotspot detection in real-time crime analytics. We have been able to incorporate some of these features into our own OA-PBI platform and we see that it makes improvements in the hotspot prediction. Again, we thank NIJ for providing this opportunity and invaluable learning experience.

Why Should DOJ-OJP Consider Conduent as a Partner of Choice?

The Conduent team pushed the boundaries of possibility based on customer-centered product development. Enabling you to do more with less. We remain true to the Conduent mission which lines up with your goals of multiplying operational efficiency via crowd sourcing and data-driven decision making.

Next Steps

We look forward to future workshops, demos, or collaboration across companies or subject matter experts within the department to deploy technology in new and creative ways to help reduce crimes impact on society.

Contact

Contact	Todd Redmond Director: Strategy & Innovation Conduent 606 South Olive St, 23rd Floor Los Angeles, CA 90016 todd.redmond@conduent.com