Predicting Assaults in Raleigh, North Carolina

Hotspot Analysis

Alexander Ng

12/10/2021

The Plan

We'll cover

- Crime prediction: ML in hotspot analysis
- Literature review
- Methodology: hotspot analysis
- Data sources
- Findings
- Interactive leaflet demo

Punchline

- Random Forest and Cubist models produce similar and effective hotspot forecasts of assaults on a 20,000 cell grid decomposition of Raleigh, NC.
- Both models were more effective in forecasting 2019 assaults than 2020 assaults possibly caused by COVID-19 and social unrest.
- This study contributes to the literature:
  - crime ML study of Raleigh
  - use of Potholes location data as a predictor of assaults
  - use of Cubist model in a spatial grid context. Although others have done city-level prediction.
  - new Cubist parameter visualization

Literature Review

- There is an extensive literature on spatial crime prediction but not standard approach to model validation or agreement on metrics. (Kounadi et al., 2020)
- Criminologists agree that the distribution of crime incidents in urban areas is spatially clustered. (Chainey et al., 2008; Cichosz, 2020; Drawve, 2016; Wheeler & Steenbeek, 2020)
- Hotspots are generally defined as relatively small urban areas where crime rates are higher than average. One objective of the ML literature is predicting crime hotspots based on historical crime data. Typically methods include kernel density estimation, prior counts, STAC ellipses. (Eck et al., 2005)
- The top 4 methods used by ML practitioners in crime prediction are random forests, multilayer perceptron, kernel density estimation and support vector machines. (Kounadi et al., 2020)
- Crime prediction researchers has developed its own performance metrics. Predictive Accuracy Index (PAI), Recapture Rate Index (REI), Predictive Efficiency Index (PEI) are often used. (Chainey et al., 2008; Drawve, 2016; Wheeler & Steenbeek, 2020) Joshi et al. (2021) has proposed Penalized PAI measure to handle its shortcomings.

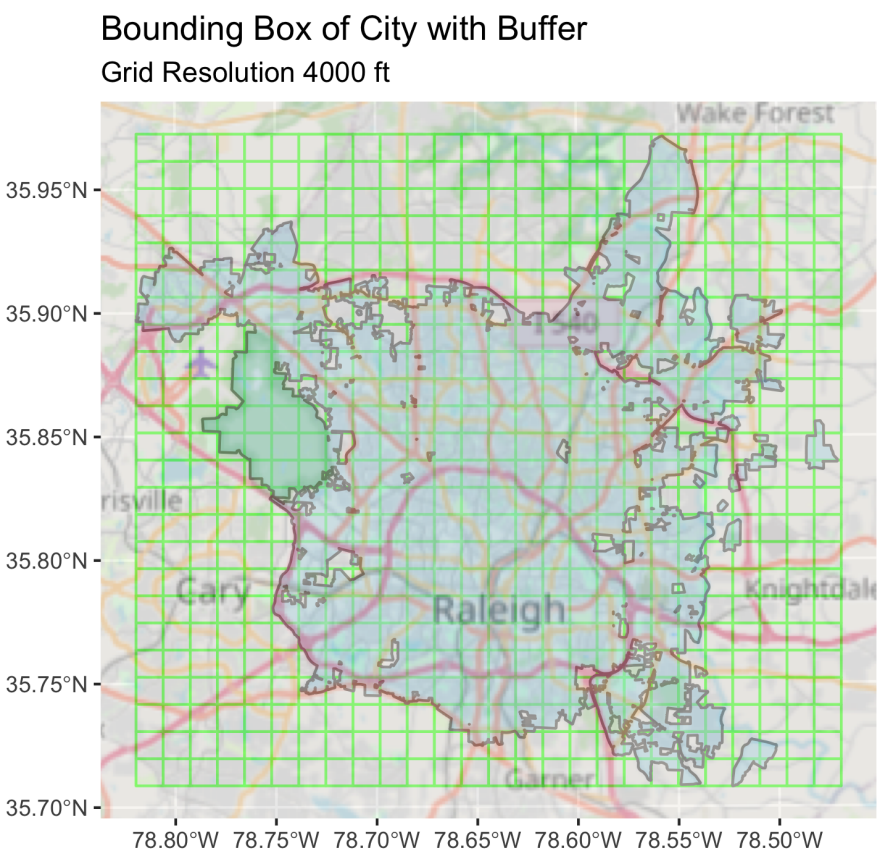
Methodology: Hotspot Prediction

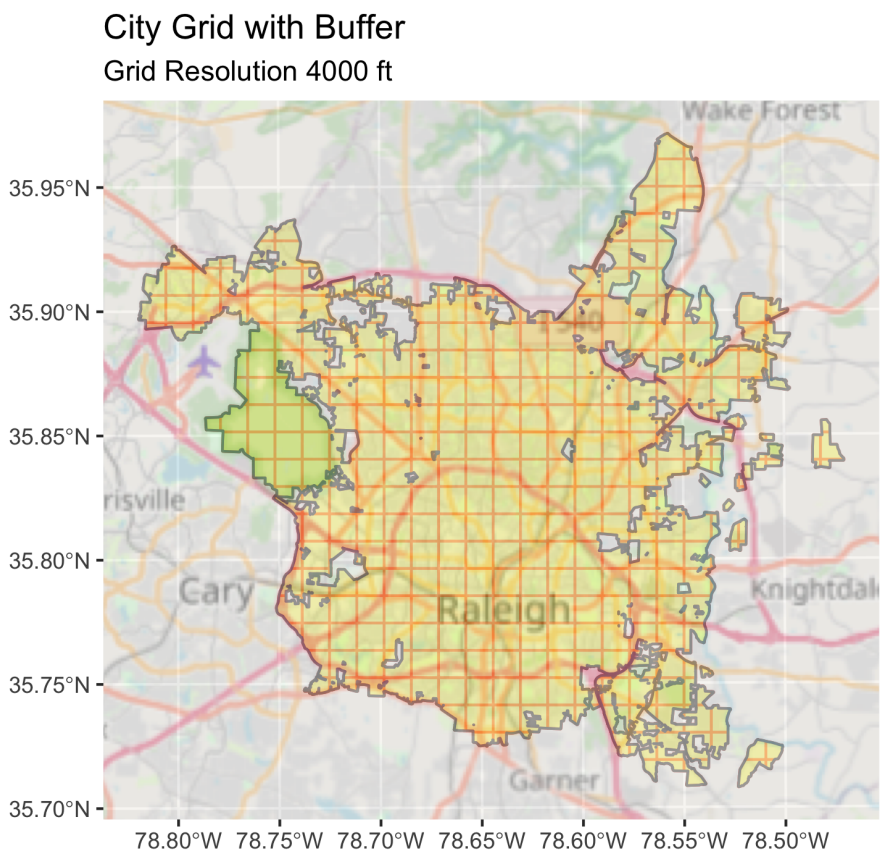
- Divide the city territory into a uniform grid. (e.g. squares, hexagons)
- Assemble datasets that help predict crime like:
  - Historical crime incident
  - Locations of crime generators: bars, retail stores, bus stations, etc.
  - Demographic variables: socioeconomic and demographics
  - Cellphone or social media data
- Measure the influence of each predictor at each cell.
  - Count the incidents or generators in each cell
  - Spatial interpolation of areal predictors: income, unemployment -
  - Measure the density or distance to a crime generator
- Train a model to predict crime rates at each cell.
- Rank the cells by predicted crime rate.
- Choose a threshold N so that the top N ranked cells are called *hotspots*
- Evaluate hotspot performance (e.g. PAI, RRI, etc.)

See Wheeler & Steenbeek (2020) for details.

Methodology: Forecast Crime on a Grid

- A very large grid at 4000 feet square resolution for illustration. Most cells are squares. Cells at the boundary are irregular.
- We use a finer resolution of 490 feet squares for the model.

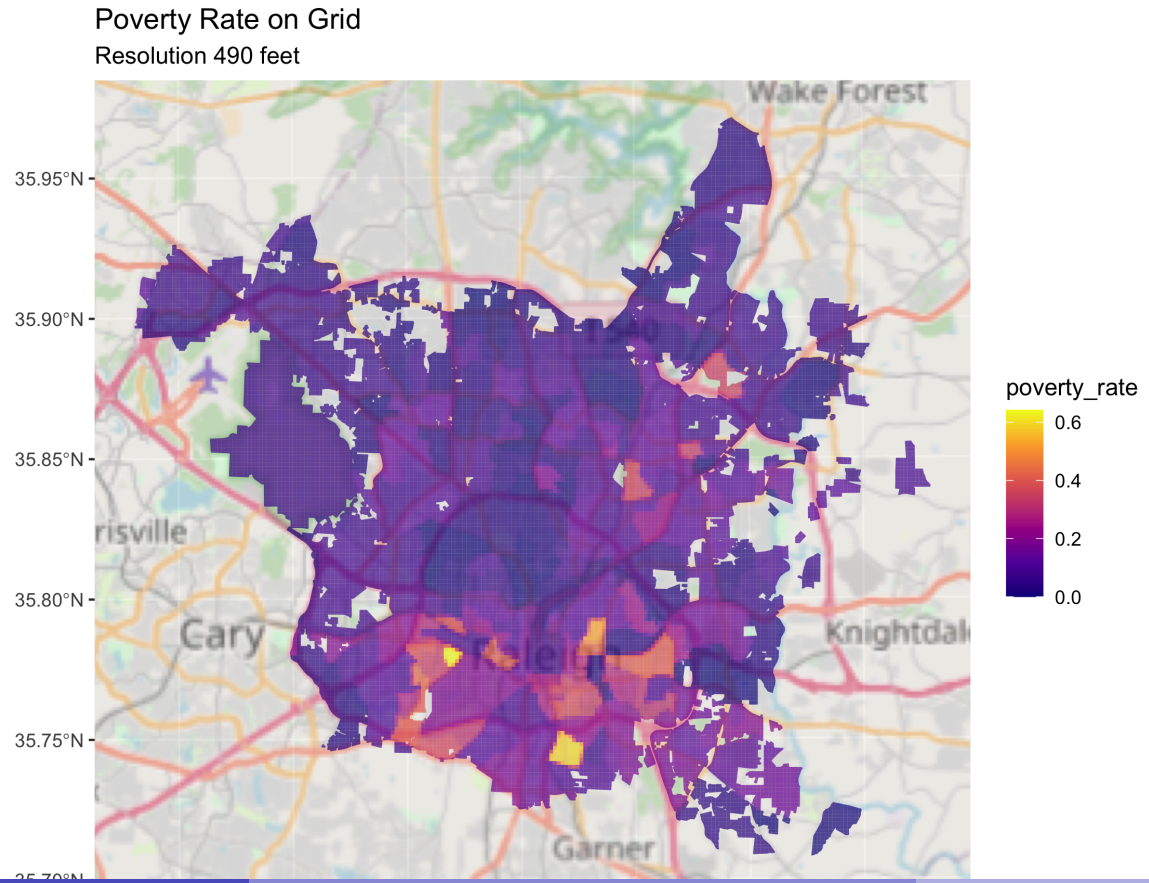




Methodology: Actual Grid Used

- At 490 feet resolution, static plots are hard to see.
- 19782 cells were used. We generate a crime forecast on each cell.
- Interactive plots help overcome this issue.

Example: Poverty Plot at 490 feet.



Data Sources

Content	Source	Format
Police Crime Incidents	City of Raleigh, NC	geojson
Income, Poverty, Unemployment, Population Density	US Census ACS Survey 2019	geojson
Points of Interest Data	OpenStreetMap project	geojson
City of Raleigh boundary	City of Raleigh, NC	geojson
Potholes	City of Raleigh, NC	geojson

Methodology

- Random Forests - ensemble tree method widely used in crime spatial prediction
- Cubist Model - rule-based approach
  - committees - akin to boosting
  - neighbors - adjust preliminary forecast by committees
  - linear regression model at each node for each committee
- Incident and point data is counted within each grid. No kernel smoothing.
- Define top $N = 100$ cells by predicted assaults as *hotspots*.

Prediction Setup

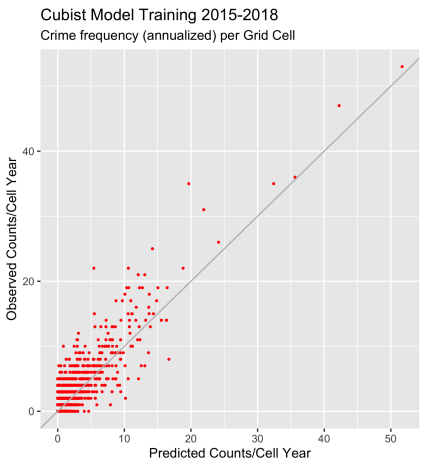
	Period 1 (2019)	Period 2 (2020)
Test Period	Jan 1, 2019- Dec 31, 2019	Jan 1, 2020 - Dec 31, 2020
Num. Assaults in Test Period	6289	5659
Training Period	Jan 1, 2016 - Dec 31, 2018	Jan 1, 2017 - Dec 31, 2019
Demographics	ACS 2019	SAME
Potholes	2018	SAME
Num. Potholes	721	SAME
Points of Interest	unknown	SAME
Num. Types POI	37	SAME
Num. POI	2324	SAME

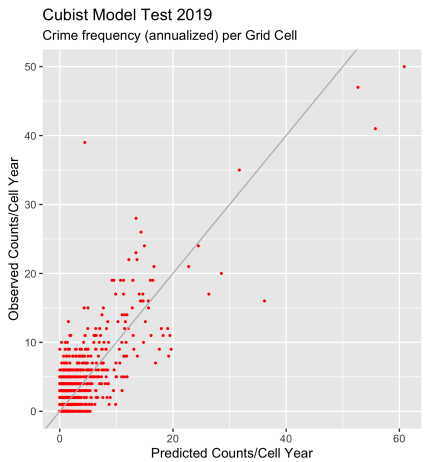
Results

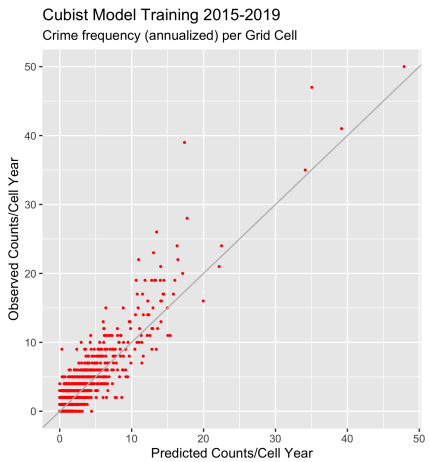
Model	Test Year	Train RMSE	Train R-sq	Test RMSE	Test R-sq
ranger	2019	.8072	66.67%	0.8209	67.60%
cubist	2019	.8035	66.68%	0.8530	66.40%
range	2020	.8034	68.21%	0.8959	55.36%
cubist	2020	.8092	67.27%	0.9035	54.84%

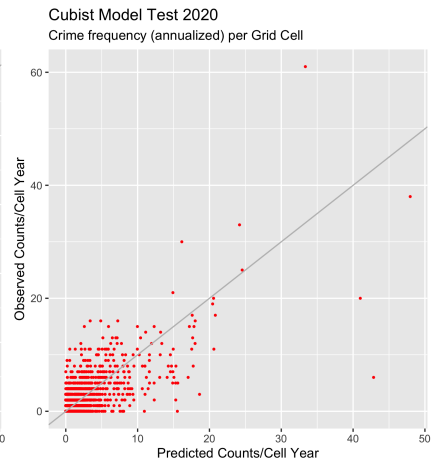
Model performance was comparable in 2019 test year, but degraded at comparable levels in 2020.

Results: Cubist



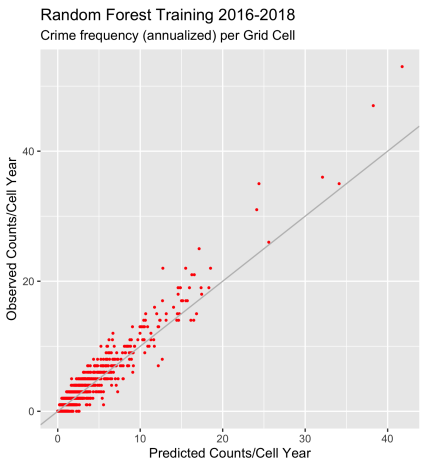


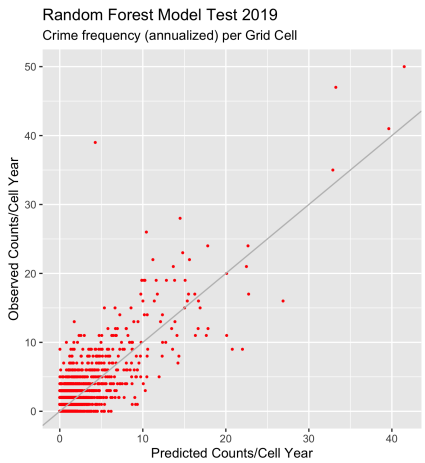


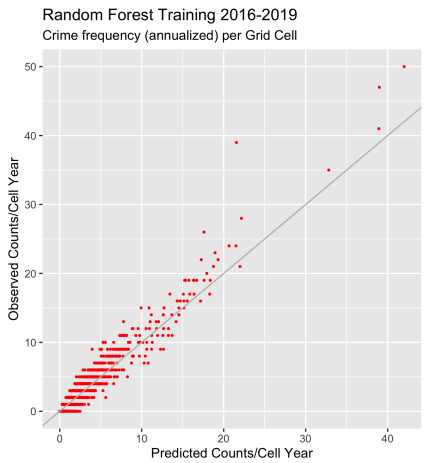


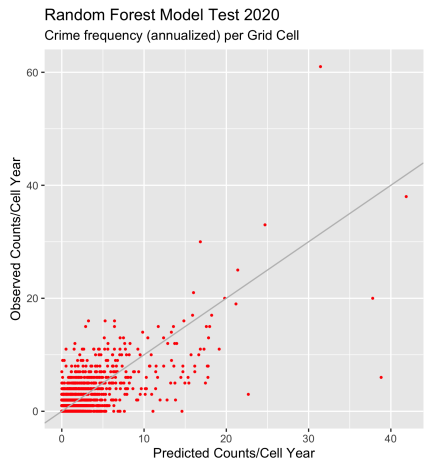
A small number of hotspots significantly exceeded expectations in 2020. Notice the outlier with 61 observed assaults in 2020. This occurs during demonstrations in downtown near the Old State Capitol.

Results: Random Forest



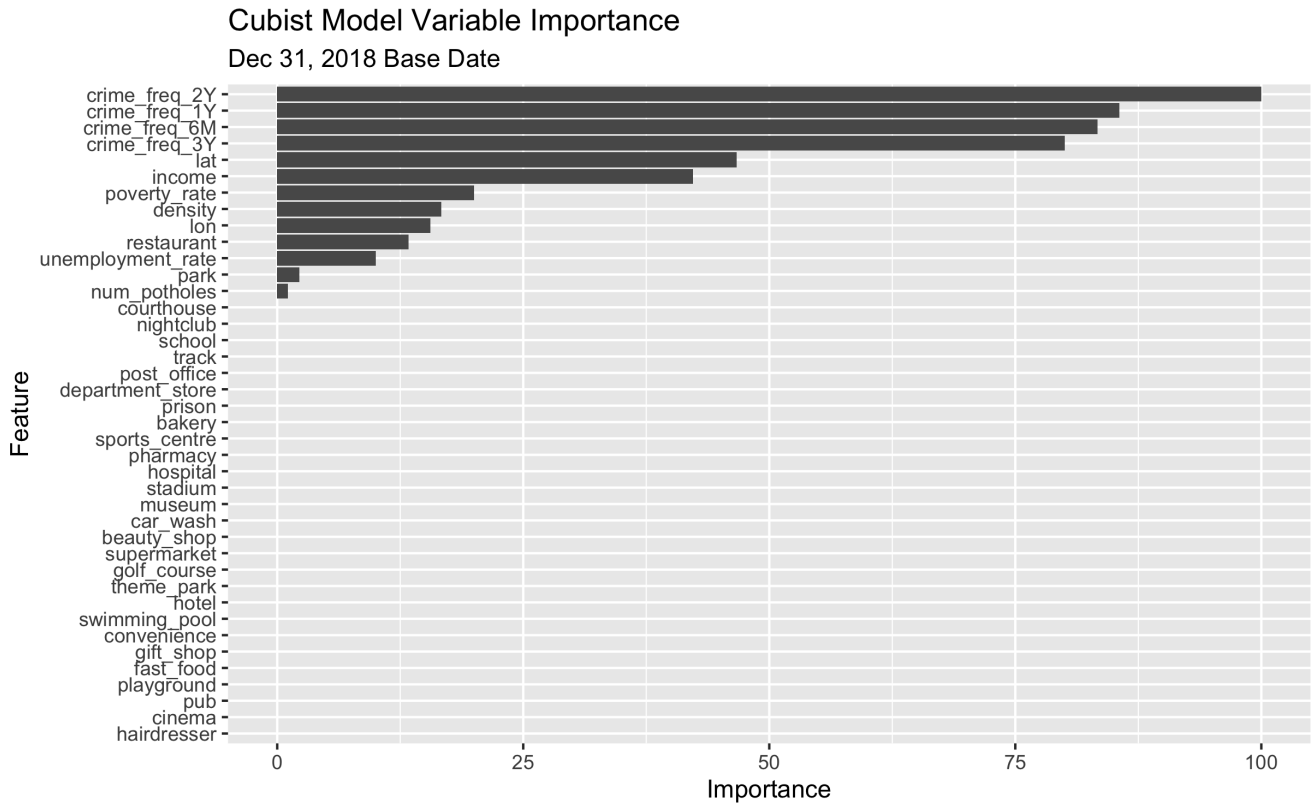






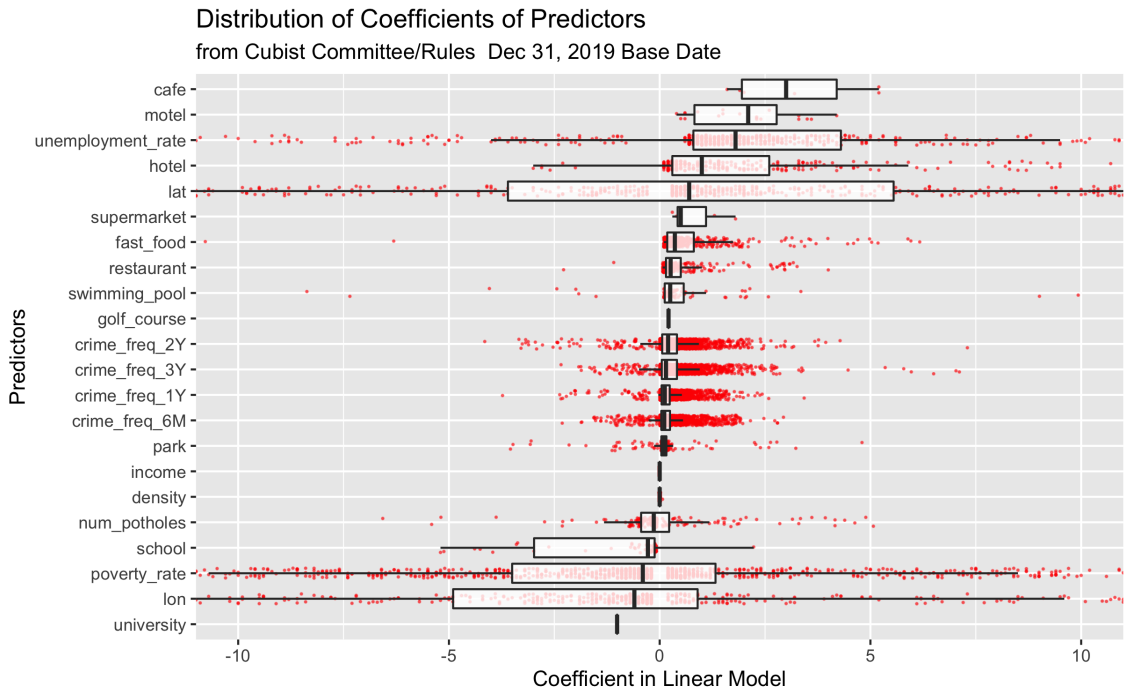
The outlier cell with 61 assaults in 2020 affected Random Forest model too.

Variable Importance: Cubist



2020 test year variable importance is similar. Past crime predictors are the most important. Location and demographics are next.

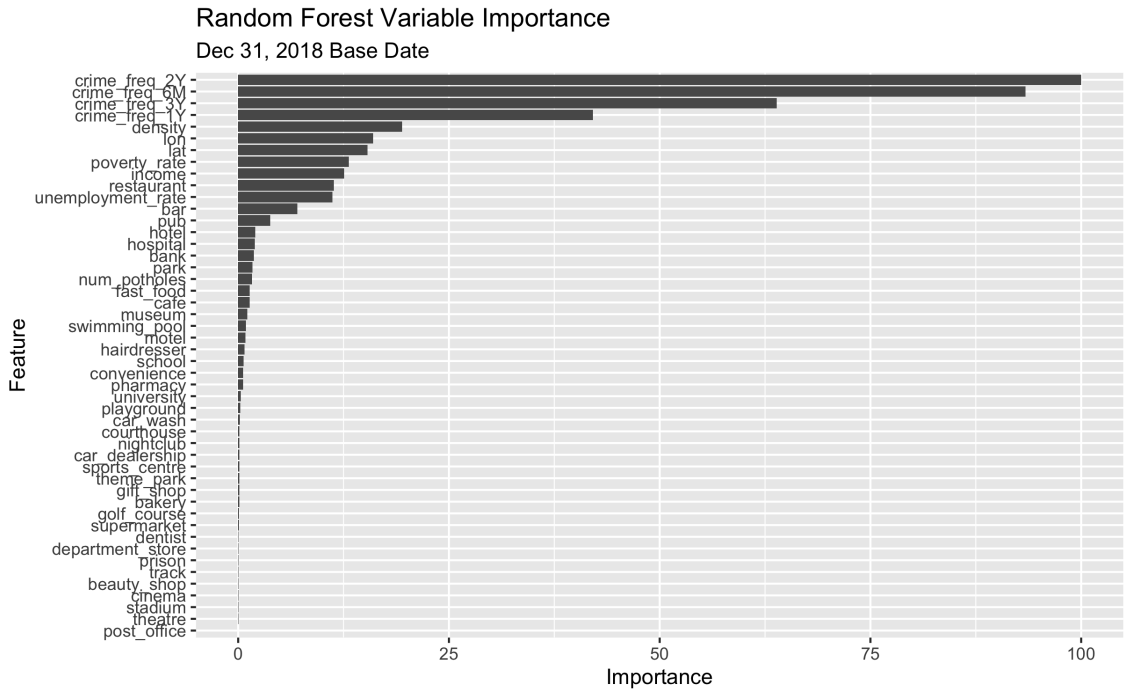
Interpretation of Cubist Parameters



Shows the distribution of predictor coefficients in linear regression model constructed by the Cubist tree. Using the sign of the median.

- Longitude and latitude suggest more crime in the South East.
- Cafes, restaurants, motels, unemployment rate associated with higher crime.
- Schools, potholes associated with less crime.

Variable Importance: Random Forest



Random Forest assigns a small weight to more points of interests.

Hotspot Metrics Defined

Assume a city C is partitioned into disjoint grid cells $A_1, A_2, \ldots, A_N$.

- Predictive Accuracy Index (PAI):

$$PAI(A) = \frac{N(A)}{\mu(A)} \frac{\mu(C)}{N(C)}$$

where $A \subset C$ and A is made up of a subset of the A_i in the entire city C. $N(A)$ is the number of crime incidents in A. $\mu(A)$ is the area of A. PAI measures the ratio of crime rate in a set of hotspots over the overall city crime rate.

- Predictive Efficiency Index (PEI):

$$PEI(A) = \frac{PAI(A)}{PAI(A^*)}$$

where $\mu(A) = \mu(A^*)$ and A^* is the union of the highest crime rate cells in the city partition. PEI is the ratio of the observed PAI versus the best possible PAI if the model has perfect foresight. A value near 1 is good.

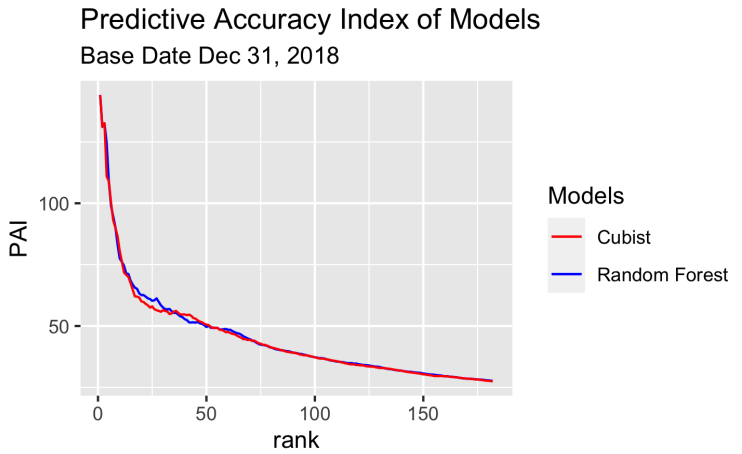
Hotspot Metrics Defined

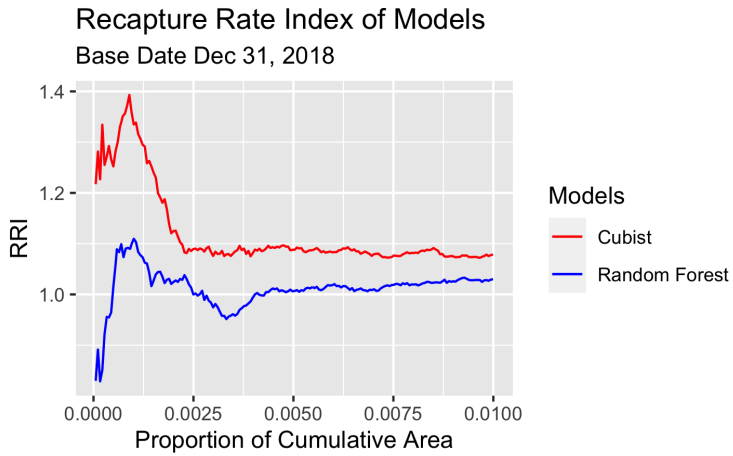
- Recapture Rate Index (RRI):

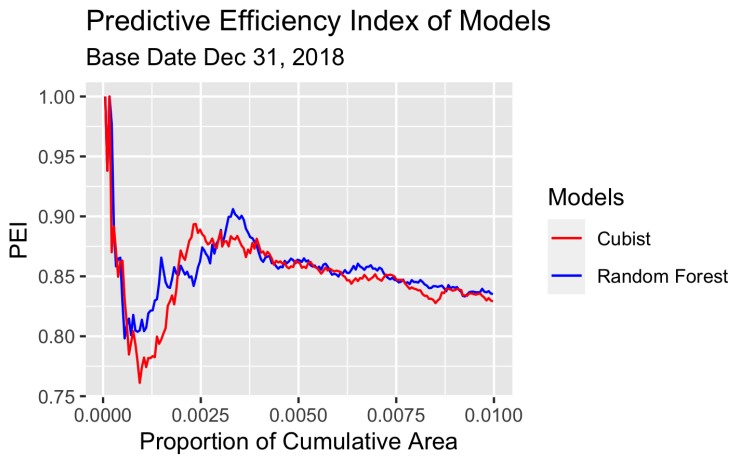
$$RRI(A) = \frac{N^*(A)}{N(A)}$$

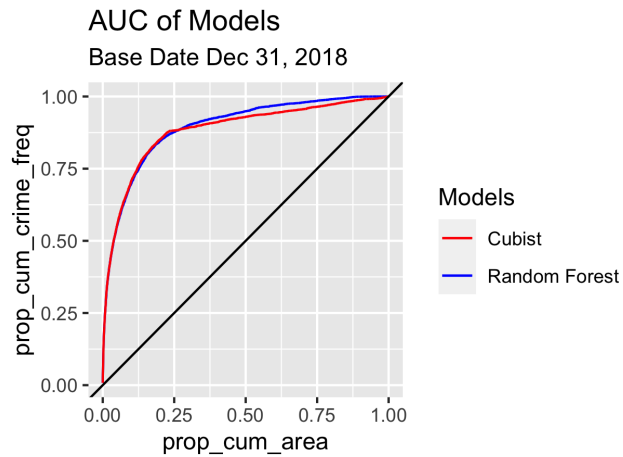
  where $N^*(A)$ is the predicted number of crimes in A. A value of RRI near 1 is good.
- Area Under the Curve (AUC): The area under the cumulative proportion of crime in A versus C divided by the proportion of area of A versus C. A value of 1.0 means the model has perfect power. Area close to 0.5 means the model has no predictive power. A value less than 0.5 means the model predicts the opposite.

Hotspot Metrics: 2019 Test Year



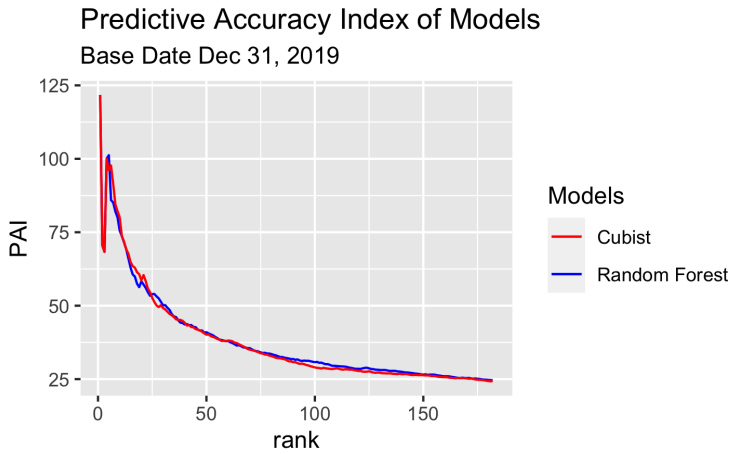


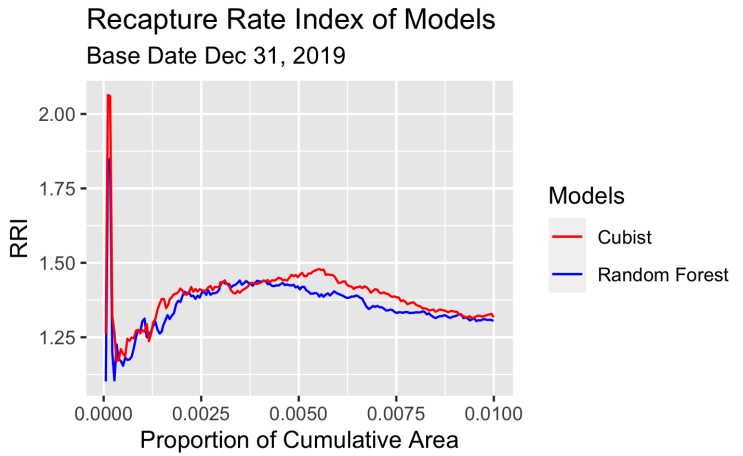


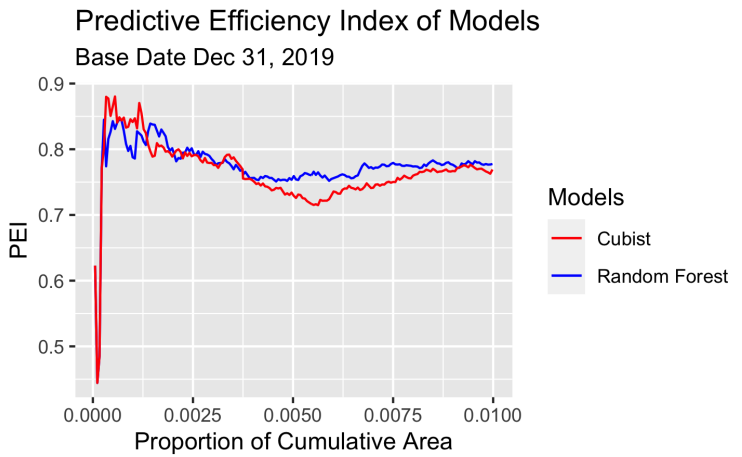


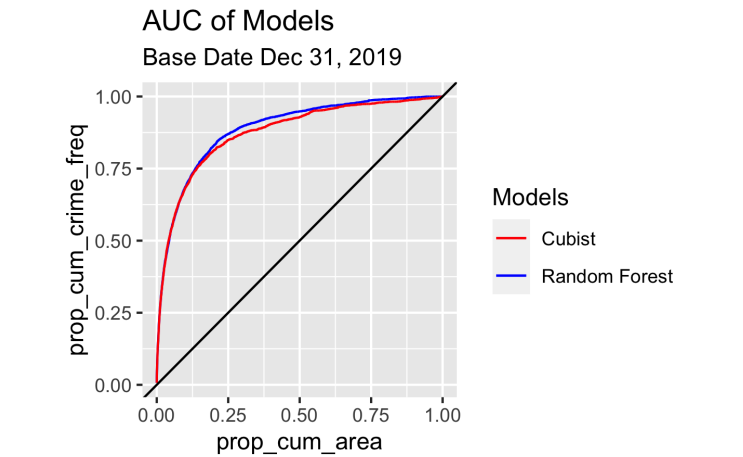
- Random Forest equals Cubist on PAI - the most important metric.
- Random Forest beats Cubist slight on AUC and PEI.
- Random Forest beats Cubist on RRI since the target level is 1.

Hotspot Metrics: 2020 Test Year





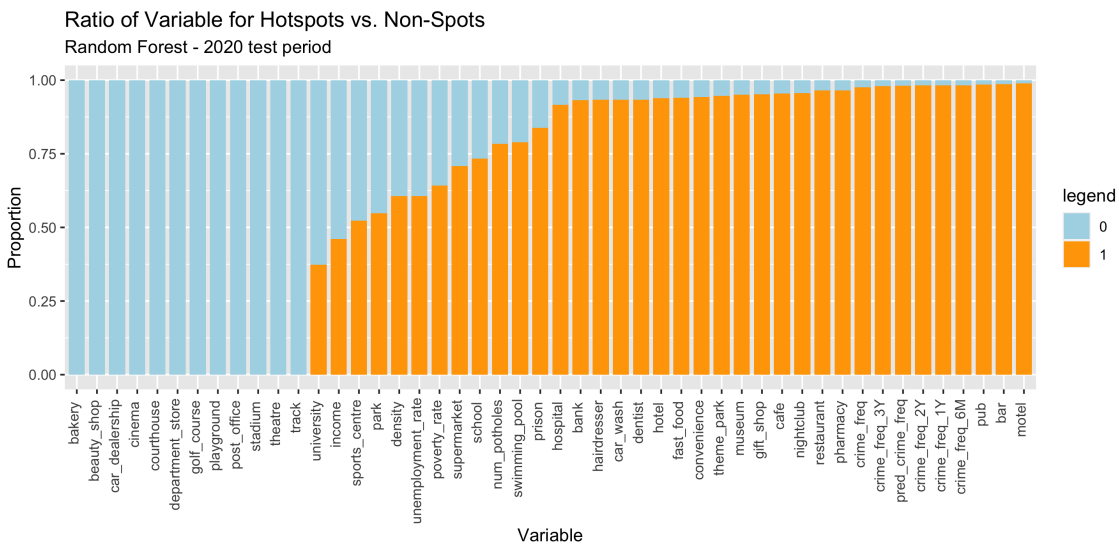




- Random Forest equals Cubist on PAI - the most important metric.
- Random Forest beats Cubist slightly on AUC and PEI.
- Cubist beats Random Forest slightly on RRI but both overestimate the actual

Interpretation of Random Forest

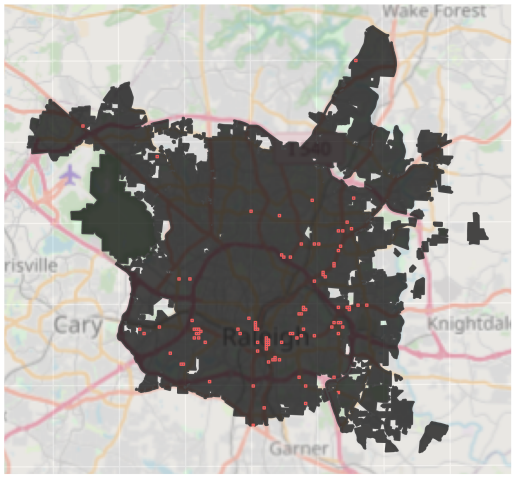
Shows the ratio of each predictor's mean conditional being a hotspot (orange) or not (blue).



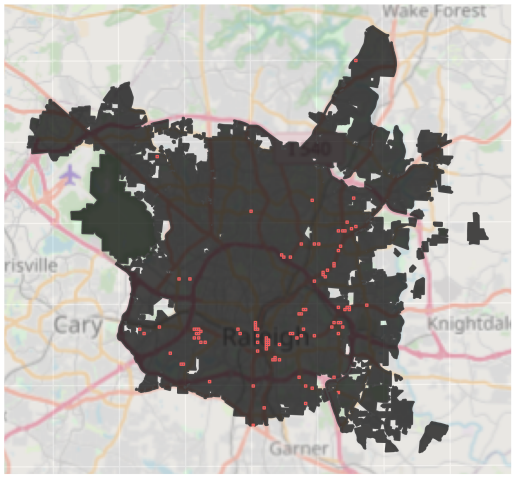
Example: Hotspots are associated with higher frequency of motels, bars, pubs, high past crime rates, restaurants, fast food, potholes than non-hotspots.

Hotspots Predicted 2019/2020

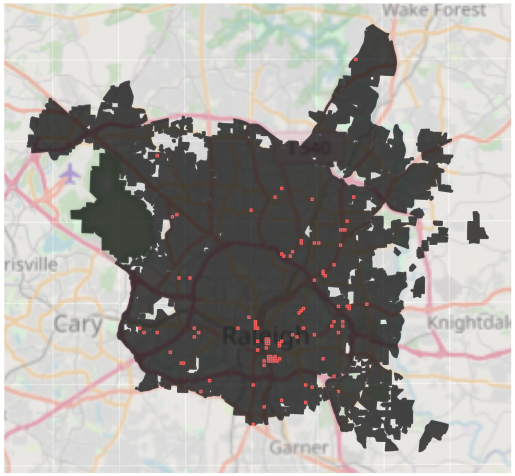
Top 100 Predicted Hotspots by Cubist 2019
Resolution 490 feet



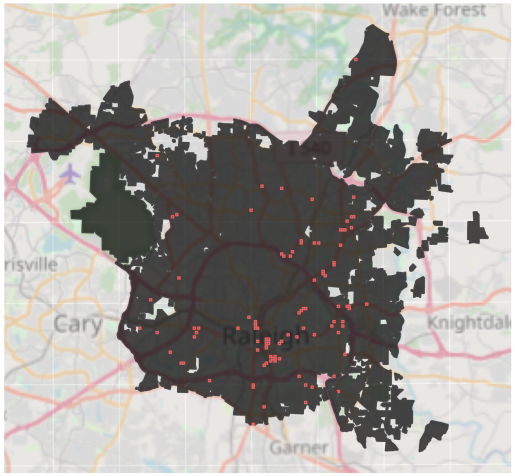
Top 100 Predicted Hotspots by Random Forest 2019
Resolution 490 feet



Top 100 Predicted Hotspots by Cubist 2020
Resolution 490 feet



Top 100 Predicted Hotspots by Random Forest 2020
Resolution 490 feet



Conclusion

- Hotspot prediction works in Raleigh
- Random Forest and Cubist models deliver reasonable PAI
- Future work ought to consider other crime types
- Forecasts for 2021
- Kernel density estimation may improve predictive performance
- Potholes have a slight predictive power

The github repo for this project: the paper and supporting materials:

https://github.com/completegraph/Raleigh

Thanks!

References I

Chainey, S., Tompson, L., & Uhlig, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, *21*(1-2), 4–28. https://doi.org/10.1057/palgrave.sj.8350066

Cichosz, P. (2020). Urban Crime Risk Prediction Using Point of Interest Data. *ISPRS International Journal of Geo-Information*, *9*(7), 459–483. https://doi.org/10.3390/ijgi9070459

Drawve, G. (2016). A Metric Comparison of Predictive Hot Spot Techniques and RTM. *Justice Quarterly*, *33*(3), 369–397. https://doi.org/10.1080/07418825.2014.904393

Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M., & Wilson, R. E. (2005). *Mapping Crime: Understanding Hot Spots* (No. 209393; p. 77). National Institute of Justice, U.S. Department of Justice.

Joshi, C., Curtis-Ham, S., D'Ath, C., & Searle, D. (2021). Considerations for Developing Predictive Spatial Models of Crime and New Methods for Measuring Their Accuracy. *ISPRS International Journal of Geo-Information*, *10*(9), 597. https://doi.org/10.3390/ijgi10090597

References II

Kounadi, O., Ristea, A., Leitner, M., & Araujo Jr, A. (2020). A systematic review on spatial crime forecasting. *Crime Science*, *9*(1), 7–7. https://doi.org/10.1186/s40163-020-00116-7

Wheeler, A. P., & Steenbeek, W. (2020). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology*. https://doi.org/10.1007/s10940-020-09457-7