RESEARCH

Supplementary material for: Mining large-scale human mobility data for long-term crime prediction

Cristina Kadar* and Irena Pletikosa

*Correspondence: ckadar@ethz.ch
ETH Zurich, D-MTEC,
Weinbergstrasse 56/58, 8092
Zurich, Switzerland
Full list of author information is available at the end of the article

1 Descriptive Statistics

In this section, we provide further exploratory analysis of the dataset.

First, Figure 1 provides the matrix of Spearman's rank correlation coefficients between all factors and the different types of crime incidents within a census tract. In terms of correlations, as expected, we notice some clusters of co-related features: the number of venues by type, the number of checkins by type, the number of popular venues by time, the average numbers of metro exits/entries, the average numbers of taxi drop-offs/pickups. Notably, these features are also positively and relatively strongly correlated with the crime indicators. We also notice negative correlations, like the demographic diversity indexes and some of the venues fractions and offering advantages.

Second, Figure 2 presents scatter plots and OLS regressions between crime counts and some exemplary metrics of the resident and ambient population, all in logarithmic scale. Amid some large leverage points, we notice weak positive linear relationships between the crime indicators and the considered attributes, of up to $adj.R^2 = 44\%$ for larcenies as a function of shops. These initial results support our assumption that attributes of the ambient population of a neighborhood are related to the crime levels.

2 Model Assessment

In this section, we investigate the optimal range for hyper-parameter tuning of the models by means of validation curves, and the data requirements of the models by means of learning curves.

2.1 Validation Curves

Figure 3 presents validation curves, i.e. training and cross-validated MSE of the ensemble models as a function of the hyper-parameters. If the complexity of the model is too low, we will underfit the data, and (both training and cross-validated) MSE will be very large because the model is too simple/biased to describe the underlying phenomena. If the complexity of the model is too high, we will overfit the training data, and the cross-validated MSE will be very large because the supposed patterns that the method found in the training data simply do not exist in the test data. Hence, to find the best combination of number trees and maximum depth of the trees (and in the case of GBR also learning rate), we employ a grid-search algorithm which exhaustively searches for the best combination of parameters and selects that combination which delivers best cross-validated MSE, and use the curves above to inform our values ranges and avoid underfitting and overfitting. In case of RF and ET, we use values ranging from 50 to 400 for number of trees, and values ranging from one third, to one half, to the full set of features for the depth; in case of GBR, we use values ranging from 100 to 400 for number of trees, and values ranging from 1 to 4 for the number of boosting iterations/depth.

2.2 Learning Curves

Figure 4 presents learning curves, i.e. training and cross-validated MSE of the models as a function of training set size. From these curves we are able to obtain a deeper insight into the data requirements of the different algorithms. Most notably, we conclude that cross-validation error keeps decreasing with the number of samples considered for model training, hence we should use all available data in the final model. Also, by comparing the different models, we notice that the linear models exhibits the worse MSE and also high variability, while the GBR delivers the best MSE.

3 Additional Model Specifications

In this section, we provide results of additional model specifications: census + FS, census + Subway, census + Taxi.

We also present here the geographical error of the additional model specifications. As presented in Table 5, the census + FS model achieves 1847 samples with a low absolute error, while census + taxi model achieves 1758 samples, while the census + subway model achieves only 1677 samples.

4 Feature Importances across Models

Figure 6 plots the results on the complete set of features for the top performing crime categories: total incidents, grand larcenies and assaults. The size of the bar represents the features' importance in the respective ensemble. The importance of a feature is computed as the impurity decrease this feature brings, averages across all trees in the ensemble. Impurity is the measure based on which the (locally) optimal condition is chosen at each node in a tree, and for regression trees it is variance.

Comparing the different tree-based ensemble methods, we observe that the Random Forest models tend to assign very high predictive power for a few chosen features, while the Gradient Boosting models tend to distribute feature importance more evenly. This is due to the difference in the ranking of features within the two models: the impurity based ranking of the random forest is typically aggressive in the sense that there is a sharp drop-off of scores after the first few top ones, while for GBR, all features are used in the boosting process (iterative fitting to minimize the residuals). This fact, paired with its top performance on the most relevant crime categories, makes GBR a better model to do feature interpretation on.

	Census + FS		Census + Subway		Census + Taxi	
	MSE	R^2	MSE	R^2	MSE	R^2
2015						
Total incidents						
Random Forest	0.46±0.05	0.58±0.07	0.52±0.09	0.45±0.14	0.49±0.07	0.52±0.08
Extra-Tree	0.44±0.03	0.61±0.07	0.51±0.07	0.48±0.09	0.50±0.09	0.51±0.12
Gradient Boosting	0.44±0.04	0.62±0.06	0.51±0.07	0.49±0.09	0.51±0.12	0.48±0.19
Grand larcenies						
Random Forest	0.52±0.04	0.53±0.09	0.64±0.14	0.32±0.10	0.67±0.04	0.42±0.10
Extra-Tree	0.51±0.04	0.54±0.09	0.63±0.12	0.33±0.08	0.58±0.12	0.43±0.06
Gradient Boosting	0.51±0.04	0.54±0.10	0.63±0.12	0.33±0.08	0.61±0.14	0.38±0.10
Robberies						
Random Forest	0.64±0.05	0.46±0.10	0.65±0.05	0.44±0.10	0.59±0.12	0.42±0.06
Extra-Tree	0.64±0.04	0.47±0.08	0.65±0.04	0.45±0.08	0.68±0.06	0.39±0.14
Gradient Boosting	0.62±0.04	0.49±0.07	0.65±0.05	0.45±0.10	0.69±0.11	0.35±0.24
Burglaries						
Random Forest	0.56±0.03	0.30±0.06	0.59±0.03	0.23±0.04	0.60±0.03	0.20±0.09
Extra-Tree	0.55±0.03	0.32±0.04	0.58±0.03	0.25±0.05	0.59±0.04	0.22±0.06
Gradient Boosting	0.55±0.03	0.31±0.04	0.56±0.03	0.29±0.04	0.59±0.04	0.21±0.05
Assaults						
Random Forest	0.61±0.03	0.56±0.07	0.65±0.03	0.51±0.06	0.65±0.02	0.49±0.06
Extra-Tree	0.60±0.04	0.57±0.06	0.64±0.02	0.52±0.04	0.65±0.02	0.50±0.04
Gradient Boosting	0.61±0.03	0.57±0.05	0.63±0.02	0.53±0.04	0.64±0.02	0.52±0.04
Vehicle larcenies						
Random Forest	0.62±0.07	0.09±0.10	0.61±0.08	0.11±0.12	0.58±0.03	0.19±0.04
Extra-Tree	0.62±0.05	0.09±0.04	0.61±0.06	0.11±0.07	0.61±0.05	0.13±0.04
Gradient Boosting	0.63±0.07	0.07±0.08	0.62±0.08	0.09±0.12	0.58±0.04	0.19±0.04
2014						
Total incidents						
Random Forest	0.46±0.06	0.56±0.09	0.53±0.09	0.42±0.14	0.51±0.09	0.45±0.15
Extra-Tree	0.45±0.05	0.58±0.09	0.53±0.06	0.42±0.09	0.53±0.11	0.42±0.20
Gradient Boosting	0.45±0.05	0.57±0.07	0.53±0.07	0.42±0.10	0.55±0.15	0.36±0.31
Grand larcenies						
Random Forest	0.51±0.06	0.54±0.09	0.64±0.12	0.28±0.12	0.62±0.17	0.34±0.15
Extra-Tree	0.51±0.06	0.54±0.08	0.63±0.11	0.31±0.09	0.63±0.17	0.32±0.16
Gradient Boosting	0.51±0.06	0.54±0.07	0.64±0.11	0.28±0.10	0.64±0.21	0.30±0.25
Robberies						
Random Forest	0.66±0.07	0.43±0.13	0.66±0.04	0.43±0.09	0.68±0.04	0.39±0.11
Extra-Tree	0.64±0.06	0.45±0.12	0.65±0.04	0.44±0.10	0.73±0.14	0.26±0.37
Gradient Boosting	0.66±0.06	0.43±0.11	0.65±0.05	0.44±0.10	0.70±0.08	0.33±0.21
Burglaries						
Random Forest	0.59±0.03	0.30±0.06	0.62±0.03	0.23±0.03	0.62±0.04	0.23±0.04
Extra-Tree	0.58±0.03	0.32±0.06	0.62±0.03	0.24±0.04	0.62±0.05	0.23±0.08
Gradient Boosting	0.59±0.04	0.30±0.04	0.60±0.03	0.29±0.01	0.62±0.04	0.23±0.03
Assaults						
Random Forest	0.64±0.05	0.52±0.09	0.67±0.04	0.49±0.08	0.68±0.05	0.47±0.09
Extra-Tree	0.62±0.04	0.55±0.08	0.65±0.03	0.52±0.07	0.68±0.08	0.46±0.13
Gradient Boosting	0.65±0.04	0.52±0.06	0.65±0.04	0.51±0.07	0.68±0.06	0.46±0.10
Vehicle larcenies						
Random Forest	0.62±0.05	0.08±0.10	0.61±0.04	0.11±0.08	0.59±0.02	0.19±0.05
Extra-Tree	0.62±0.03	0.08±0.06	0.62±0.03	0.09±0.03	0.63±0.05	0.06±0.13
Gradient Boosting	0.63±0.04	0.07±0.07	0.62±0.05	0.08±0.09	0.61±0.04	0.12±0.07

Table 1: Geographical out-of-sample results of the regressors using different subsets of the features: for each year, repeatedly trained on 80% of the census tracts, and tested on 20% of the census tracts.

	Census + FS		Census + Subway		Census + Taxi	
	MSE	R^2	MSE	R^2	MSE	R^2
Total incidents						
Random Forest	0.07	0.89	0.09	0.85	0.09	0.85
Extra-Tree	0.07	0.89	0.07	0.88	0.07	0.88
Gradient Boosting	0.08	0.86	0.12	0.80	0.15	0.75
Grand larcenies						
Random Forest	0.13	0.82	0.19	0.74	0.15	0.79
Extra-Tree	0.15	0.79	0.22	0.70	0.15	0.80
Gradient Boosting	0.18	0.75	0.23	0.68	0.22	0.69
Robberies						
Random Forest	0.23	0.75	0.26	0.72	0.27	0.70
Extra-Tree	0.28	0.70	0.27	0.71	0.36	0.61
Gradient Boosting	0.27	0.71	0.34	0.63	0.38	0.59
Burglaries						
Random Forest	0.25	0.48	0.25	0.48	0.24	0.49
Extra-Tree	0.32	0.32	0.25	0.47	0.26	0.45
Gradient Boosting	0.25	0.48	0.28	0.41	0.29	0.40
Assaults						
Random Forest	0.23	0.77	0.24	0.75	0.24	0.76
Extra-Tree	0.28	0.72	0.25	0.75	0.26	0.74
Gradient Boosting	0.27	0.73	0.32	0.67	0.36	0.63
Vehicle larcenies						
Random Forest	0.30	0.32	0.29	0.35	0.28	0.37
Extra-Tree	0.40	0.12	0.29	0.35	0.32	0.28
Gradient Boosting	0.29	0.34	0.30	0.32	0.31	0.30

Table 2: Temporal out-of-sample results of the regressors using different subsets of the features: trained on 2014 and tested on 2015.

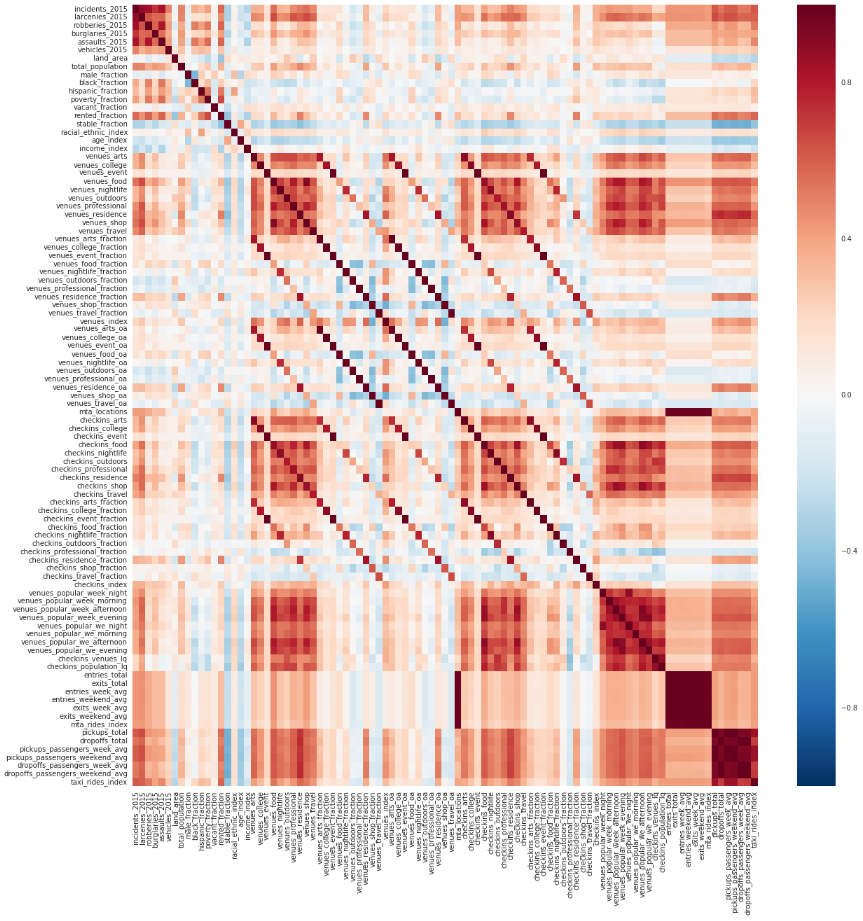

Figure 1: Correlation matrix.

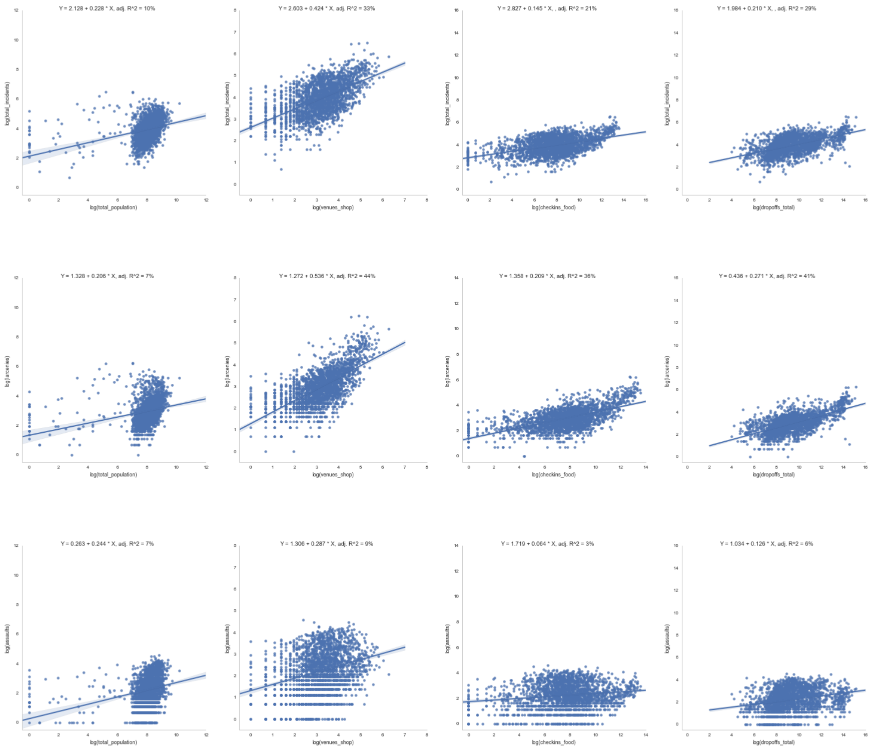

Figure 2: Scatter plots and univariate regressions.

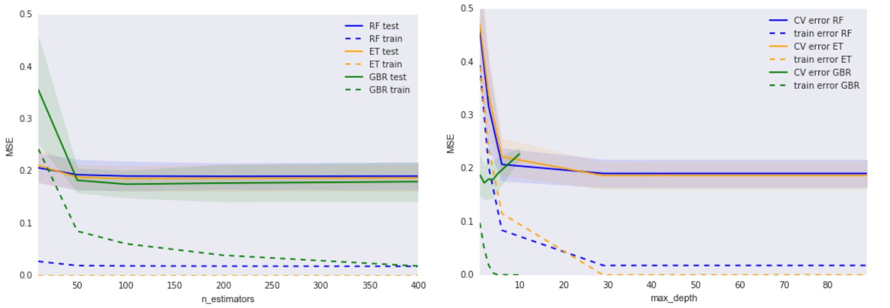

Figure 3: Validation curves of the models (full specification) for total incidents 2015.

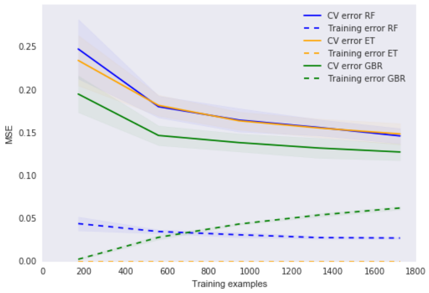

Figure 4: Learning curves of the models (full specification) for total incidents 2015.

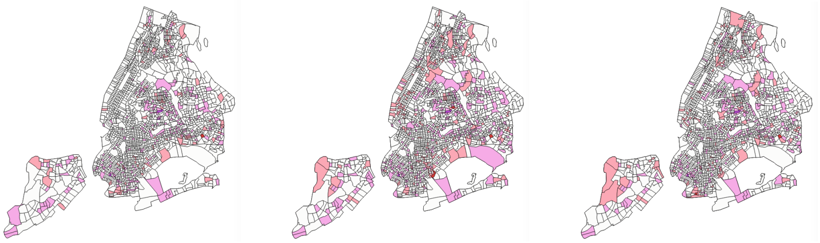

Figure 5: Absolute error of predicted vs actual values for the 2015 larcenies counts per census tract. From left to right: census + FS, census + subway, and census + taxi.

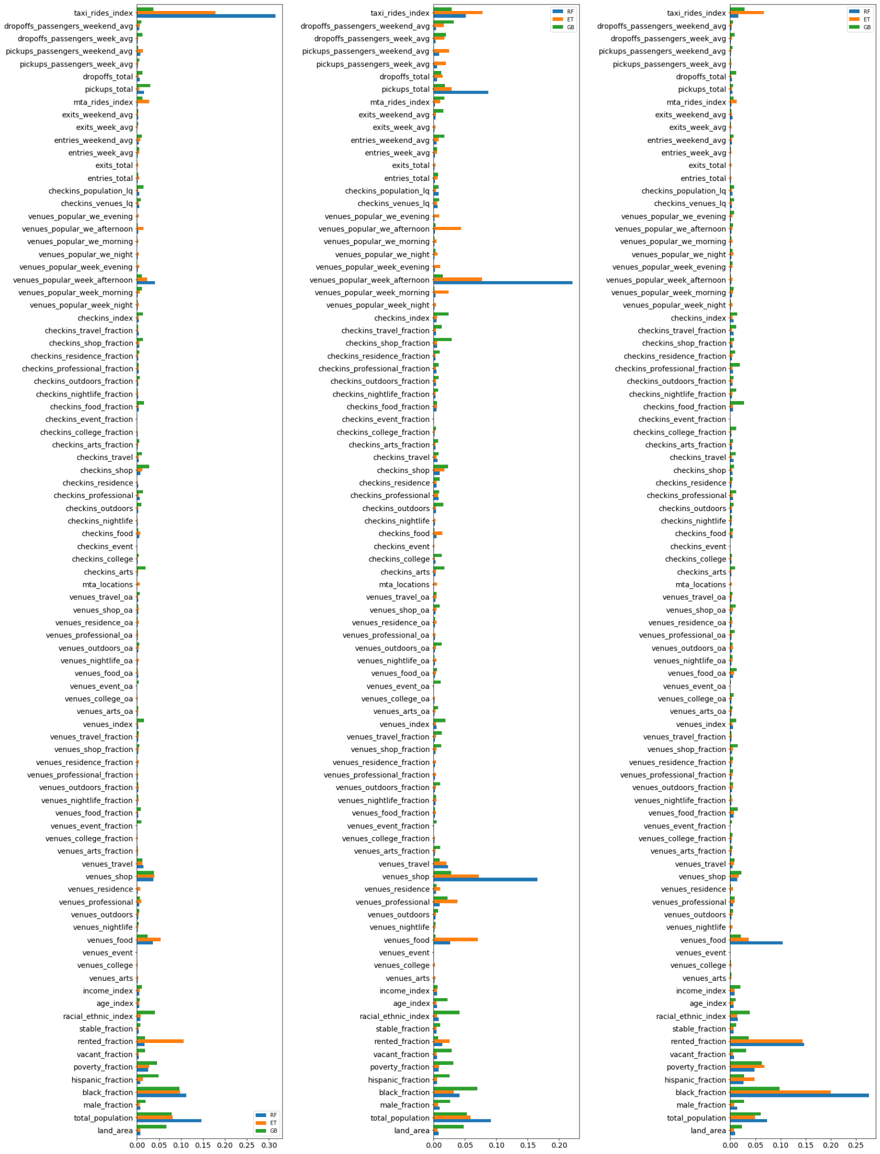

Figure 6: Variable importance plots across all three models (RF = Random Forests, ET = Extra-Trees, GB = Gradient Boosting). From left to right: 2015 total incidents, 2015 grand larcenies, and 2015 assaults.