Intelligent Information Management, 2018, 10, 123-132
http://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

Comparison of Model Performance for Basic and Advanced Modeling Approaches to Crime Prediction

Yuezhexuan Zhu

Shanghai East Foreign Language School Affiliated to SISU, Shanghai, China
Email: zhuyuezx@163.com

How to cite this paper: Zhu, Y.Z.X. (2018) Comparison of Model Performance for Basic and Advanced Modeling Approaches to Crime Prediction. *Intelligent Information Management*, **10**, 123-132.
https://doi.org/10.4236/iim.2018.106011

Received: September 7, 2018
Accepted: November 25, 2018
Published: November 28, 2018

Abstract

A good machine learning model would greatly contribute to an accurate crime prediction. Thus, researchers select advanced models more frequently than basic models. To find out whether advanced models have a prominent advantage, this study focuses shift from obtaining crime prediction to on comparing model performance between these two types of models on crime prediction. In this study, we aimed to predict burglary occurrence in Los Angeles City, and compared a basic model just using prior year burglary occurrence with advanced models including linear regressor and random forest regressor. In addition, American Community Survey data was used to provide neighborhood level socio-economic features. After finishing data preprocessing steps that regularize the dataset, recursive feature elimination was utilized to determine the final features and the parameters of the two advanced models. Finally, to find out the best fit model, three metrics were used to evaluate model performance: R squared, adjusted R squared and mean squared error. The results indicate that linear regressor is the most suitable model among three models applied in the study with a slightly smaller mean squared error than that of basic model, whereas random forest model performed worse than the basic model. With a much more complex learning steps, advanced models did not show prominent advantages, and further research to extend the current study were discussed.

Keywords

Crime Prediction, Recursive Feature Elimination, Benchmark Model, Linear Regressor, Random Forest Regressor

1. Introduction

Numerous efforts have been made on crime prediction for cities around the

world. By predicting locations with a high crime rate, computers can help police departments distribute manpower more scientifically and efficiently, which may prevent severe crimes form happening. With the rapid development in approaches of machine learning, advanced learning models become popular tools for crime prediction. To make accurate predictions on when and where crimes happen, it is critical to identify which predictors and what types of model are optimal for crime prediction.

Based on former studies, the possibility of crimes may relate to various factors, neighborhood level socio-economic factors are one of the most studied. Raphael and Winter-Ebmer [1] analyzed the relationship between unemployment and crime using U.S. state-level data. The results consistently indicate that unemployment is an important determinant of property crime rates. Moreover, Sampson *et al.* [2] focused on relationship between social and organizational characteristics of neighborhoods and variations in crime rates, and the result showed that collective efficacy is negatively associated with variations in violence.

Performance for complex problems. For example, IAlBoni & Gerber [3] made a comparison between traditional kernel density estimation (KDE) models and area-specific hierarchical models. The result of the study showed that area-specific models have advantages in area-prediction and accuracy. In addition, Nguyen *et al.* [4] employed Support Vector Machine (SVM), Random Forest, Gradient Boosting Machines, and Neural Networks for prediction and focused on crime prediction. They concluded that Random Forest or Gradient Boosting turned out to be the two best models for dataset in which demographic features were employed. Moreover, Luiz *et al.* [5] utilized Random Forest Regressor and had a research on crime prediction through urban metrics and statistical learning. Utilizing Random Forest Regressor, their approach reached an accuracy of 97%.

This study focuses on predicting burglary crime rates for each census tract in the City of Los Angeles. The effectiveness of neighborhood-level socio-economic variables as predictors of burglary rate, and the effectiveness of linear regression and random forest models for crime prediction are assessed in this context. As neighborhood-level socio-economic factors are associated with crime, American Community Survey data at the census tract level are used in the prediction. Time-lag is taken into account by using prior year crime and ACS data as predictors, since most crime predictions bases on data of former. Advanced machine learning models such as random forest and SVM are often used in crime prediction. However, the effectiveness of such complex techniques is rarely evaluated against simple models. This study compares the performance of linear regression and random forest regressor to a simple benchmark model, which only uses prior year crime rate as predictor, to evaluate effectiveness of two advanced models.

The rest of the paper is organized as follows. The method section describes the

data source, detailed data processing procedures, and modeling approach for achieving the study objectives. The result section presents the feature selection and model comparison results. In the discussion section, effectiveness of various predictors and the effectiveness of the compared models are discussed. Conclusions and future directions are drawn.

2. Methods

2.1. Data Sources

In this study, we used data from three sources. The crime dataset is collected from LA official crime database [6], which covers detailed descriptions of crimes, including time, location and the crime type. The geographic dataset is obtained from Los Angeles Country GIS Data Portal [7]. The geographic data are shapefiles which illustrate the boundaries of each census tracts in Los Angeles. The American Community Survey (ACS) dataset is downloaded from Social Explorer [8], which contains population, education level, race and age distribution of residents in each census tracts.

2.2. Data Preprocessing

Data preprocessing turned raw downloaded dataset into specialized dataset for crime prediction by dropping the unrelated information, merging new datasets and splitting the original dataset into training, validation, and test datasets. The details of data preprocessing steps were illustrated in **Figure 1** and described as follows.

The first step was to remove irrelevant columns in the crime data and deal with missing data. Since the essential columns were Date Occurred, Location and Crime Code, other columns were dropped. The column "Location" was split into longitude and latitude columns. Longitude and latitude were plotted on a scatterplot, and a small number of locations were showing as (0,0), which were cases missing location. These cases of crime records were removed from the analysis. The current study focuses on prediction of burglary crime, so all other types of crime were removed from the analysis as well.

The second step was to calculate the rate of burglary for each census tract. Census tract IDs were attached to each crime record through spatially joining

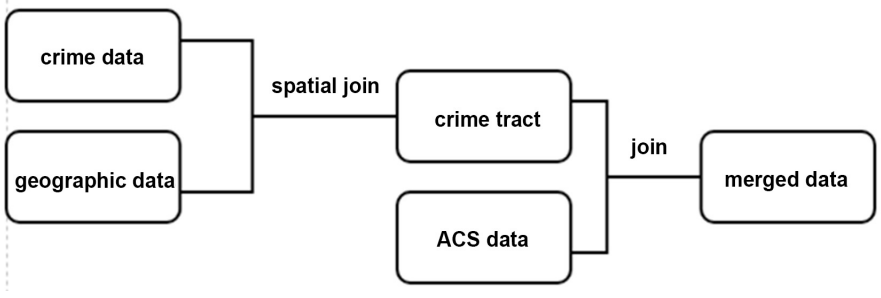

Figure 1. Flow chart for data merging.

the crime locations with the census tract shapefiles. The burglary records were then summarized into counts at tract level by year.

In the third step, the burglary counts data were transposed, in order to properly consider time-lag in prediction. The former data listed the year when the crime happened in one column, but data from past years were often used to make crime prediction for next year. Thus, it was crucial to transform burglary counts for each year as a separate column to allow previous years' rates to be predictors for the current year crime rate After transposing, Burglary counts for each year was separated into columns from BGLRY10 to BGLRY17, representing the burglary occurrence from 2010 to 2017.

At last, the transposed burglary rates dataset was split into training set, validation set and test set, and each was merged with American Community Survey data using TractID as the key. As a result, the training set included burglary rate for 2015 merged with 2011-2014 burglary rates and 2010-2014 ACS data; the validation set included burglary rate for 2016 merged with 2012-2015 burglary rates and 2011-2015 ACS data; and the test set included burglary rate for 2017 merged with 2013-2016 burglary rates and 2012-2016 ACS data.

2.3. Feature Engineering and Feature Selection

The process of feature engineering is shown in Figure 2. Feature engineering is an essential part for machine learning since it reduces overfitting, improves accuracy of the outcome, and reduces the training time by eliminating the number of features. In this study, a few features were dropped after reviewing the correlations and several new features were created by combining information from multiple features. Features were then ranked by applying feature selection method: recursive feature elimination. The final selection of features was determined according to the feature ranking and model performance.

First, correlation matrix of all the features from ACS and the outcome were examined. Features that had a correlation smaller than 0.05 in magnitude with the burglary occurrence was dropped, since these features were not likely to have a prominent contribution to the crime prediction. On the other hand, some features had an extremely high correlation with each other, with an absolute value higher than 0.9. These features were likely duplicated features that would not provide additional information. Therefore, these features with high correlations

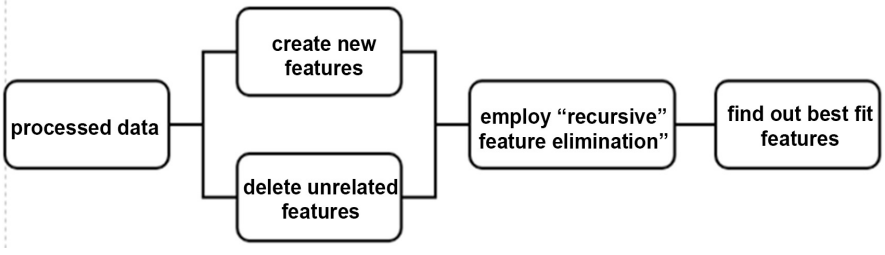

Figure 2. Flow chart for feature engineering.

between themselves were examined, and only one among each group of such features were retained. After this step, 32 features of 82 original features were.

The second step was to create new features by combining several original features. The features created are described below:

1) Neighborhood Disadvantage Index: This feature was made up of three separate features: population 25 years and over and less than high school, population 16 years and over in labor force who are civilians and unemployed, and population for whom poverty status is determined under 1.00. These 3 variables were standardized, and the index was created by averaging the standardized variables. The higher the index, the more socioeconomically disadvantaged the neighborhood was.

2) Racial diversity index: This feature contained eight racial composition features. It was created by first squaring each proportion, and then summing them together, and finally subtracting the sum from 1. The higher the index was, the higher racial diversity rate an area had.

3) Maximum: This feature was the highest burglary occurrence in past 5 years.

4) Minimum: This feature was the lowest burglary occurrence in past 5 years.

5) Mean: This feature was the average value of burglary occurrence in past 5 years.

6) Standard Deviation: This feature standard deviation of the burglary occurrence in past 5 years.

After this step, 57 features remained in the dataset, including 5 crime features from last 5 years, 46 ACS features and 6 constructed features. Since an excess of feature may lead to overfitting or low performance, the number of features in machine learning need to be tested.

Therefore, the last step of feature engineering was to apply the recursive feature elimination. Recursive feature elimination would form several subsets from the original dataset and eliminate the features with the least importance. As a result, the initially eliminated features were listed at the bottom of the ranking list. After eliminating the features one by one, the feature ranking was created. Model performance was evaluated to determine the optimal number of features by fitting models with different numbers of top ranking features.

2.4. Model Comparison

In this study, three regression models were compared: benchmark model, linear regressor, and random forest regressor. The first one, benchmark model, a model that makes prediction simply by employing crime data from last year, is the basic model. The intention of using this model is to judge those other models. If an advanced model performs worse, it means that that advanced model is not suitable for prediction. After comparing all the outcomes of models, the best model would be determined. We used two commonly used evaluation criteria to evaluate model performance: mean squared error, R Squared and adjusted R squared.

3. Result

3.1. Feature Selection

The study employs recursive feature elimination (RFE) to select features. The first Benchmark model, which is a basic model that employs crime data from the former year for prediction, does not require feature selection. The other two models, the linear regressor and random forest regressor, would benefit from feature selection. Ridge regressor was used as the model for RFE. Table 1 lists the 10 top ranked features in RFE.

3.2. Number of Features

To determine the number of features used in the advanced models, the performance metrics of the regression model and the random forest model were calculated using 1 to 20 features. Figure 3 is a line chart presenting the adjusted R squared of linear regressor and random forest regressor by number of features. In Figure 3, despite drastic increase in model performance when increasing the number of features within the first six features, the values of adjusted R squared become stable afterwards. Moreover, considering the increase of processing time and risk for overfitting when the number of features increases, creating a model that employs 6 top-ranked features would be the most efficient one.

3.3. Tuning Parameters

Several parameters can be tuned in random forest regressor, such as n_estimators, which is the number of trees in random forest model, max_features, which represents the maximum number of features considered for splitting a model. Since n_estimators has a big impact on model performance, we tuned this parameter as a variable and limited the range from 5 to 400 with an interval of 5.

Figure 4 illustrates the line chart of the accuracy of random forest regressor with growing numbers in n_estimators. It can be seen that there is a prominent

Table 1. Top 10 ranked features.

	Feature	Description
1	Mean	Mean of Burglary Occurrence in 5 Years
2	PCT_T033_002	Population 16 Years and Overin Labor Force
3	PCT_T033_004	Population 16 Years and Over in Labor Force, Civilian
4	PRY1	Burglary Occurrence ofLast Year
5	SE_T012_002	Median Age of Male Population
6	PRY3	Burglary OccurrenceThree Years Before
7	SE_T012_003	Median Age: Female Population
8	Max	the Highest BurglaryOccurrence in 5 Years
9	NhoodDisIdx	Neighborhood Disadvantage in Census Tracts
10	PCT_T033_007	Population 16 Years and Over not in Labor Force

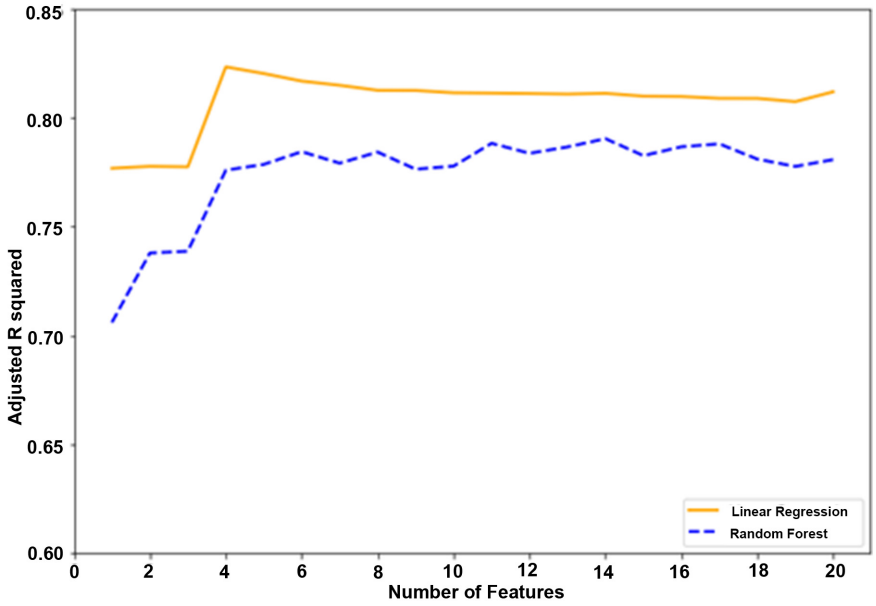

Figure 3. Outcome of models by using different numbers of features.

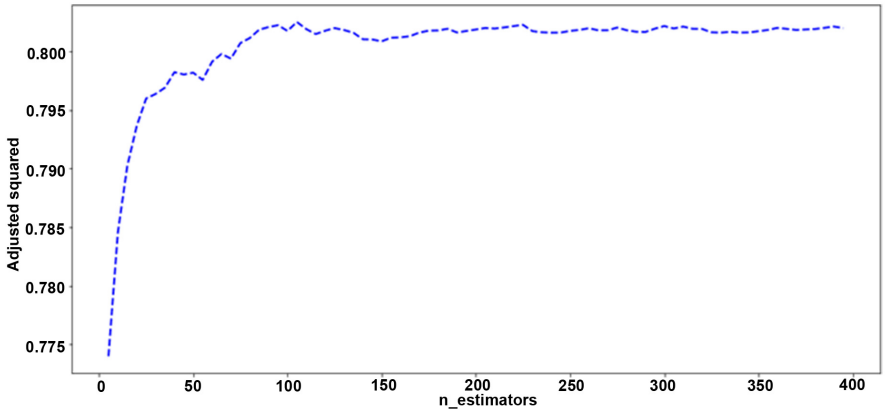

Figure 4. Outcome of random forest regressor with varied parameter (n_estimators).

rising trend of the adjusted R squared when n_estimators raised from 5 to 100. As the value of n_estimators keep increasing, adjusted R squared maintained around 0.803. As increasing the number of trees beyond 100 would not improve model performance, the value for parameter n_estimators was set at 100.

3.4. Model Comparison

Feature selection and parameter tuning were based on model trained on the training set and fit on the validation set. To examine whether advanced models have better performance and which model performs best, the three models, benchmark model, Linear Regressor and Random Forest Regressor with 6 top ranked features and chosen parameter settings were trained on the validation set, and fit to the test set to derive three performance metrics: mean squared error, R squared, and adjusted R squared. Table 2 shows the results.

Table 2. Comparison between three Models.

Model Method	Adjusted R squared	Mean squared error	R squared
Benchmark model	-	160.356292	0.800721
Linear Regression	0.802707	157.930205	0.803736
Random Forest	0.787329	170.240655	0.788437

R squared resulted from benchmark model is 0.800721, while that is 0.803736 from linear regressor and 0.788437 from random forest regressor. Linear regressor also has the smallest Mean squared error among the 3 models. Although linear regressor performed slightly better, linear regressor does not have obvious advantage over the benchmark model, whereas the random forest regressor performed slightly worse.

Linear regression was the best performing model among the three models. To illustrate the results of the prediction, predicted values from the linear regression model were plotted on a choropleth map in Figure 5. All the census tracts in LA were grouped into 5 categories and marked with different colors. Census tracts with high burglary occurrence can be easily identified on the map, which would be a great tool in real-life application.

4. Discussion

The feature selection results showed that the best fit feature was the mean burglary occurrence rate in last five years. Neighborhood level socio-economic factors contributed slightly above and beyond prior year burglary information in predicting current year burglary rate, but the impact was not significant. The comparison of the performance of linear regression and random forest regressor to the simple benchmark model, which only used prior year crime rate as predictor, showed that linear regressor was the model with highest performance, but only slightly. Advanced model such as Random forest model did not perform better than basic linear regression or benchmark model. So, to select models for further crime prediction, a benchmark model should be applied to determine whether they fit the requirement.

Although crime information is the most powerful in predicting future crime, features regarding to employment rate and average age of the population would help to raise the performance of the prediction. The neighborhood level socio-economic features in top 10 features were all related to two aspects: employment rate and the average age of the population. This indicates that these two types of factors have a high correlation with crime occurrence in specific areas since employment rate directly determines the average personal income. Other than the socio-economic environment, neighborhood environment features such as the physical environment might have higher correlation with burglary crime, which are worth exploring in future research.

In addition, in this study, burglary occurrence is the target outcome for prediction, and the findings are most applicable to prediction this particular crime

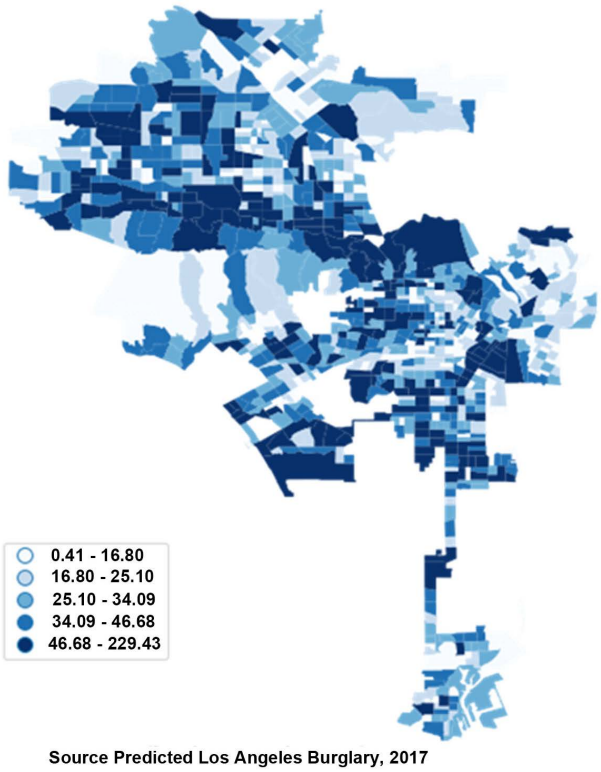

Figure 5. Map Visualization of the Burglary Prediction in LA.

type. However, each type of crime has its own characteristics and severity. In addition, each of them likely has different association with the neighborhood socio-economic environment. Therefore, it is worth exploring whether neighborhood socio-economic factors are more effective predictors for other types of crime.

Advanced model such as random forest performed adequately well in the study. However, it is not necessarily a better model than the naive benchmark model or linear regression model in predicting future crime. Adding more advanced models in the comparison may help to form a more comprehensive conclusion on whether other advanced models are suitable for the task. However, the findings in this study have one important implication for future studies on advanced machine learning methods. The results indicate the importance of including a basic benchmark model whenever comparing model performances. The cost-benefit of utilizing a more complicated advanced model will be more evident with reference to the simple benchmark model.

In general, the significance of this study can be separated into three sections. First, it indicates that neighborhood level socio-economic factors such as employment rate and the average age of the population contributes above and beyond prior year crime in predicting burglary occurrence and discussed future directions that may improve the prediction or help to generalize the approach to predicting other types of crime. Second, we have showed the importance of a benchmark model in evaluating predictive models. These findings would be

beneficial to other studies that relate to crime prediction through machine learning. Finally, we formed Predictions of burglary occurrence made for each census tract in Los Angeles City. The predictions illustrated in map would be helpful to resource allocation and crime prevention.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Raphael, S. and Winter-Ebmer, R. (2001) Identifying the Effect of Unemployment on Crime. *The Journal of Law and Economics*, **44**, 259-283. https://doi.org/10.1086/320275

[2] Sampson, R.J., Raudenbush, S.W. and Earls, F. (1997) Neighborhoods and violent Crime: A Multilevel Study of Collective Efficacy. *Science*, **277**, 918-924. https://doi.org/10.1126/science.277.5328.918

[3] Al Boni, M. and Gerber, M.S. (2016) Area-Specific Crime Prediction Models. 15*th IEEE International Conference on Machine Learning and Applications* (*ICMLA*), Anaheim, CA, 18-20 December 2016. https://doi.org/10.1109/ICMLA.2016.0118

[4] Nguyen, T.T., Hatua, A. and Sung, A.H. (2017) Building a Learning Machine Classifier with Inadequate Data for Crime Prediction. *Journal of Advances in Information Technol*ogy, **8**, 141-147. https://doi.org/10.12720/jait.8.2.141-147

[5] Alves, L.G.A., Ribeiro, V. and Rodrigues, F.A. (2017) Crime Prediction through Urban Metrics and Statistical Learning. *Physica A*: *Statistical Mechanics and its Applications*, **505**, 435-443.

[6] Crime Data from 2010 to Present. Los Angeles Police Department. https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq

[7] Los Angeles County GIS Data Portal. Census Tracts (2010) https://egis3.lacounty.gov/dataportal/2011/07/19/census-tracts-2010

[8] Social Explorer. American Community Survey-ACS 2012-2016, 2011-2015, and 2010-2014 ACS Data (5-Year Estimates). https://www.socialexplorer.com