A Comparison of Linear Regression Model and Random Forest Model for Misdemeanour Prediction in Manhattan

July 2018

STUDENT NAME	S-NUMBER	COURSE
XINRAN WANG	s6036163	GFM
MAX KORIR	s6035760	NRM
QIAO REN	s6036341	GFM
UTSAV SONI	s6038387	GFM

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION
UNIVERSITY OF TWENTE, ENSCHEDE, THE NETHERLANDS

Table of Contents

1. INTRODUCTION

1.1 Background of the study

Crime is a major famous problem affecting the quality of societal and economic development. Studies depict that crime is often associated with slower economic growth at the local, national and regional level (Dong, Lepri, & Pentland, 2011). Criminal activities are a common occurrence in urban areas. City residents, therefore, have an urge of improving urban living through alleviation of these detrimental activities.

New York City is not safe as well from these grotesque occurrences. Although the overall crime is continually plummeting to historical lows, rape and murder cases have increased in the city this year. The city recorded 678 rape cases for the first half of the year 2017 as compared to the 903 cases for the first half of 2018 (Honan, 2018). In 2017 2,656 hotel crimes were reported as compared to 2,223 in 2015 (Calder, 2018). There is, therefore, an urgent need for state-of-the-art approaches to terminate potential crimes in the city.

Mastery of spatio-temporal patterns of criminal activities will help in combating this menace efficiently and effectively. In principle, it will enhance deployment of police at the right time and where they are most needed. It will boost crime prediction and combat by the police as well as timely response to crimes in progress (Saccilotto et al., 2011)

Modelling based on space and time is often more useful to provide more reliable predictions than purely spatial models. Because observations measured at different times can be also taken into consideration, thereby providing both spatial and temporal correlations (Gräler et al., 2016). In this study, linear regression models and random forest have been compared to find out the better one for predicting misdemeanour crimes using the seven complaints type cases in the New York City.

1.2 Objectives

The main objective of this study is to propose an effective and appropriate machine learning algorithm to find the relationships between complaints and crimes instances in space and time in MANHATTAN, New York City and further develop a model to predict the most probable region of crime growth for the year 2015.

1.3 Contribution of group members

First, we had a discussion about this assignment, and everyone gave their own inputs.

XINRAN WANG: data download, data selection, data processing, creating independent variable tables, and part of report writing

MAX KORIR: Excel data conversion to shapefiles, data cleaning, spatial joins and geoprocessing, and part of report writing

UTSAV SONI: model comparison, merging data, and part of report writing, attempt to build a model in google earth engine

QIAO REN: classify the complaint types into categories, data processing, coding to train and to validate the linear regression model and the random forest model, apply both the two models in the predictions, make a comparison of the two models with my teammates, analyse them with my teammates, make the flowchart with my teammates.

2. METHODS

2.1 Study Area

The study was carried out at New York City (NYC). It is the most populous city in the US with a population of 8,622,698 as recorded in 2017. The residents of the city increased by 316,000 between 2010 and 2014. NYC consist of five boroughs namely: Queens, Manhattan, The Bronx, Brooklyn and Staten Island. The city has a great degree of income differences. New York City Police Department remains the largest police unit in the city with over 35000 cops. In overall there has been a decrease in crime rates since the 1990s which the sociologists and criminologists attribute to the use of a new strategy employed by the NYPD such as CompStat.

Secondly, removal of lead from American gasoline the other highlighted reason behind crime reduction. The lead could increase aggression and lower intelligent level. Despite the overall decrease in criminal activities, the statistics show an increase in some individual crime types.

2.2 Dataset

This study focuses on using two different datasets, i.e. NYPD Complaint Data Historic and 311 Service Requests from 2010 to Present. The first one focuses on the Crime reports that were reported in NYC while the second one comprises of all kinds of complaints that were received by the 311 Service Requests. Due to the voluminous amount of data and difficulty in processing the data in raw formats, some pre-processing to the data is considered.

2.3 Data Pre-processing

The data being voluminous in nature had to be aggregated so as to make it possible to be processed. The aggregation was done both temporally as well as spatially. Since most of the retrieved complaints were from Manhattan region, only that area of the city was considered. There were three types of crimes classified in the raw dataset i.e. felony, misdemeanour, and violation. Cerdá et al., 2009 stated in their study that misdemeanour policing has proved to reduce homicide rates in the United States, hence this study focussed only the prediction of misdemeanour crimes considering they are the root of lethal crimes.

Similarly, there were approximately 200 categories of complaints that were retrieved, which were aggregated to 7 classes namely, Health, Living Condition, Noise, Others, Property, Social Assistance, and Transportation. The temporal aggregation is done by filtering the data to the months January to March for the years 2014 & 2015.

Post aggregation of data was done for the two datasets to make theme compatible with each other. To do so, first the two datasets are made up to same spatial level by spatially splitting Manhattan area by ZIP code boundaries and applying spatial join in QGIS to the point layers of crime locations and the polygon layer of ZIP boundaries.

2.4 Data Analysis

The data was explored to find out the correlation between dependent variables and the misdemeanour and results are a shown below:

Figure 1: The scatter plot showing the correlation between each complaint type and crime

The R^2 for the above scatterplots are as shown below:

Health	0.438006653
Living Condition	0.46201815
Noise	0.541775903
Others	0.180127613
Property	0.109604177
Social Assistance	0.12684125
Transportation	0.361190061

Table 1 R^2 for each complaint type groups

2.5 Workflow

To achieve the objective of the study the following steps were followed:

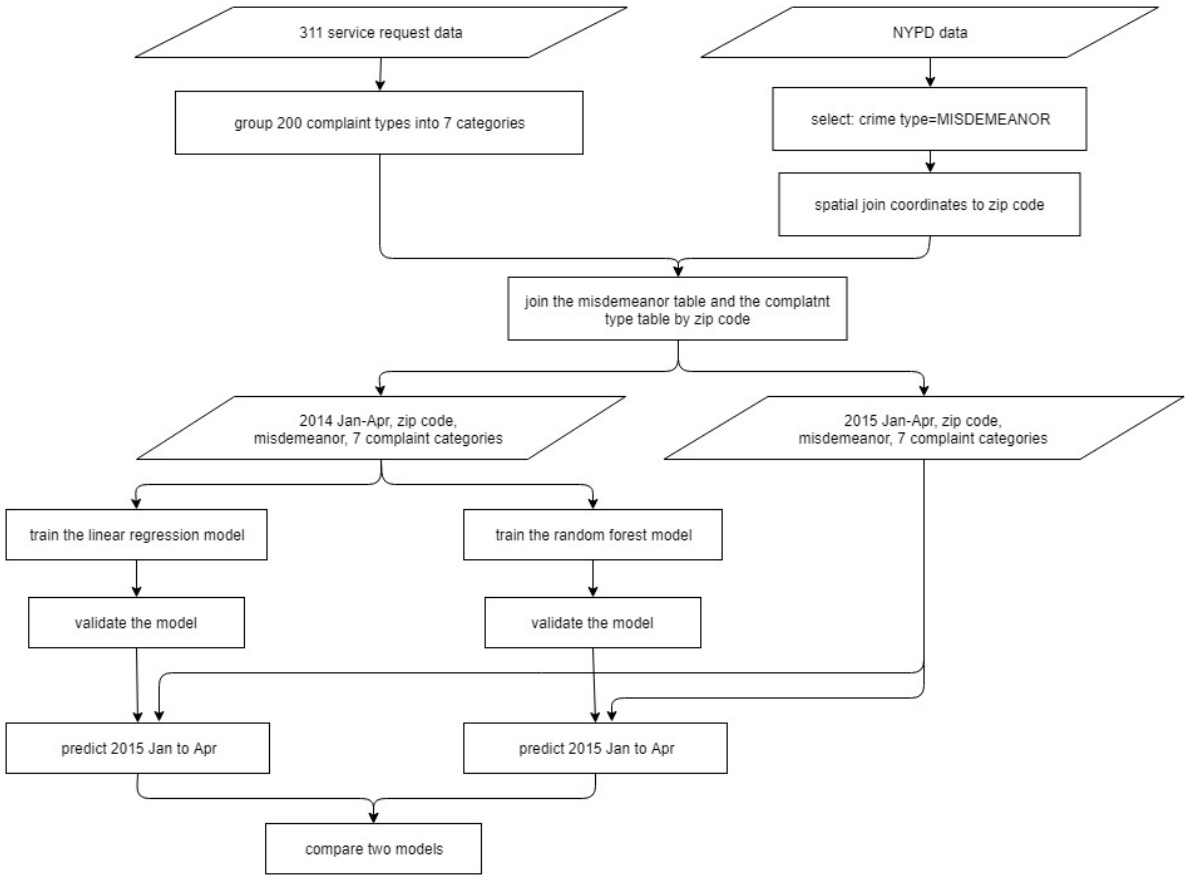


Figure 2: The flowchart showing the systematic procedure followed in the study

As explained in the flowchart, the two data sets are aggregated to a level where they can be used collaboratively. The combined dataset is further split it into two temporal zones, i.e. Jan - Apr 2014 and Jan - Apr 2015.

The Jan - Apr 2014 dataset is further split up in the ratio of 80%:20% following the Pareto Rule for the purpose of training and testing of the model. For modelling two different approaches are taken up, i.e. Random Forest & Linear Regression model. Accuracy assessment as well as comparison of both the models is done by running them on the 2014 data set. Once the training and testing of the models is done successfully, the models are run to predict the misdemeanour crimes for 2015 data. These predicted values are further compared with actual crime values so as to determine the accuracy of the models.

3. RESULTS

Two models have been trained: linear regression model and random forest model. Both these two types of model are not stable enough because the amount of training data is too small. However, the evaluation of these two can be done. In order to evaluate the fitness of the model, the coefficient of determination (denoted by R2) is used. The way of evaluation is as follows. Firstly, each type of the model has been built 1000 times. Secondly, by comparing with the crime frequency of the test data, each model has a corresponding R2. After that, the list of 1000 R2 values has been analysed.

The result of evaluation is shown in table [2]. From this table, it can be concluded that, random forest model has a better performance than linear regression model. Because, firstly, in random forest model, the R2 in the first quartile and the third quantile is between -0.1 and 0.37. This range is higher than the range of linear regression model. This can be seen in the histogram of R2 values as well. Secondly, the median of random forest model (1.8) is higher than the median of the linear regression model (0.8). Thirdly, variance of random forest model is much lower than the variance of linear regression model. This means that the random forest model is more stable compared with linear regression model.

Besides, the feature importance of random forest shows that the noise complaint type is the most important feature in the model. Complaints on health, living conditions, property and social assistance all have the feature importance around 0.1. This means that they are related to the crime frequency.

After running the model for 1000 times, the model which provides the highest R2 value has been chosen as the best model for prediction. The best model of linear regression type has 0.92 as R2. The best model of random forest type has 0.91 as R2. The figures of predicted misdemeanour against the observed misdemeanour in 2014 is shown in table [3]. In both the models, most of the blue dots are close to the red line. Some of the blue dots deviate from the red line. This illustrates that most of the predicted misdemeanour is close to the observed value.

The prediction of 2015 has been shown in table [4]. The two figures show the comparison between predicted crime value and the observed value. By comparing these two figures, it can be found that the best linear regression model fits better to the observed misdemeanour frequency in 2015, compared with random forest. The reason is that the R2 of the best linear regression model is higher than the best random regression model.

To sum up, although both the linear regression model and the random forest model have high variation in iterations, the random forest model has a better performance and it is more stable than the linear regression model. Complaint category of noise has the biggest influence on the model.

Training and Validation

	R2 of Linear regression model	**R2 of Random forest model**
min	**-9.78666**	**-5.30453**
1st Quantile	**-0.28958**	**-0.10384**
median	**0.08334**	**0.18904**
3rd Quantile	**0.30421**	**0.37321**
max	**0.92847**	**0.90793**
variance	**0.7394239**	**0.3620161**

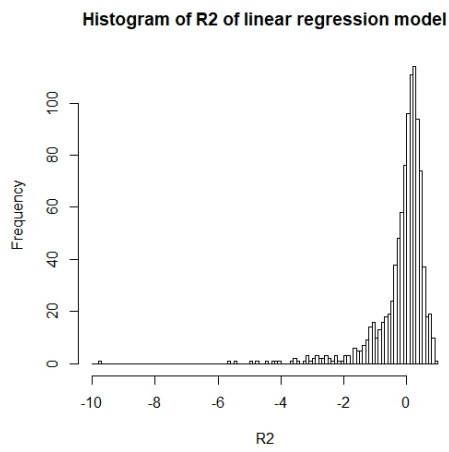
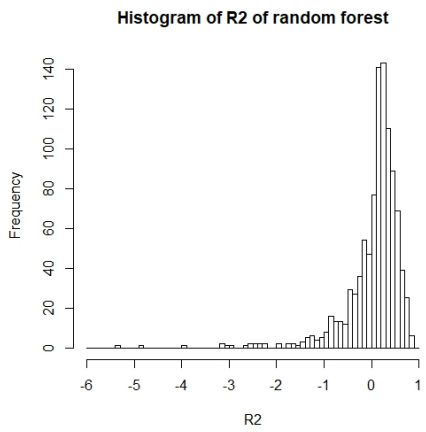
Histogram		
Feature importance		**when R2=0.90793, 'health', 0.1578558966773011, 'living condition', 0.12595359040375506, 'noise', 0.4201538439116046, 'others', 0.058468669554043055, 'property', 0.07430611554170226, 'social assistance', 0.09326129867807463, 'transportation', 0.07000058523351937**

Table 2 Comparison of coefficient of determination (R2) of LRM (left) and RF (right)

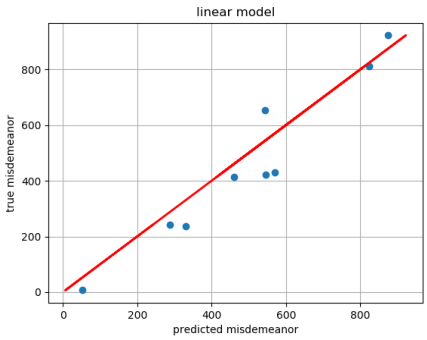
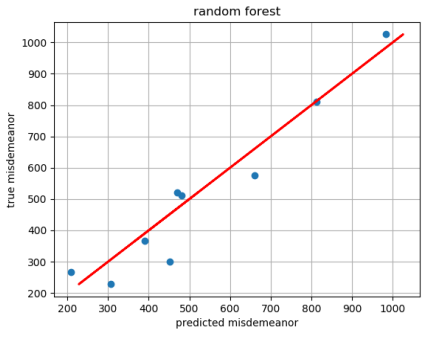
	Linear regression model	Random forest model
Predicted misdemeanour in 2014 against the observed misdemeanour in 2014	 **When R2=0.92847237**	 **When R2=0.9079354948**

Table 3 Comparison of LRM (left) and RF (right) in training and validation part

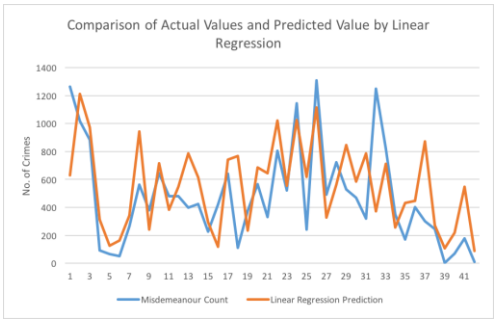
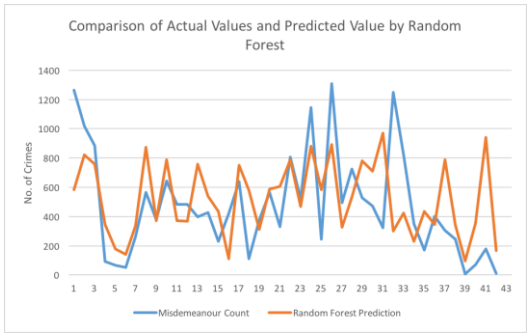
	Linear regression model	Random forest model
Comparison between predicted misdemeanour value and observed misdemeanour value in 2015		

Table 4 Comparison of LRM (left) and RF (right) in prediction in 2015

4. DISCUSSION

In the whole process, we have reasonable conditions, clear thinking and a logical model, which helps to find the most relevant complaint type of crime. Reducing numerous criminals throughout a community can create peaceful living environment.

However, there are some limitations in this study which make our model very unstable. A major limiting factor in this work is the size of the dataset, which is a common issue in both the two-modelling process. The number of input data is too small to train a model. The way to increase the amount of data can be either to collect data from a larger district or the whole New York city, or to split MANHATTAN into smaller units, like grids. In this way, we could create a more fit model.

The second limitation is that, in order to simplify the model, we aggregated about 200 complaint types into 7 groups. But the way of aggregation is mostly based on the common sense. It would be better to do the classification based on literature research and get theoretical support. Moreover, some kinds of complaint types do not have causal relationship with crimes. That means they hardly cause a crime. In the other words, we only used one-year data (2014) to train a model, of which the temporal range is not enough. So, some temporal relationship cannot be presented by this model.

5. CONCLUSION

The random forest model has a better performance and it is more stable than the linear regression model, although both the linear regression model and the random forest model have high variation in iterations. Moreover, complaints in the city are correlated with misdemeanour cases on different level. Noise is the highest predictor of the misdemeanour category of crimes in the random forest algorithm. This work could be extended to find out the crime hotspots in the city. This could be used to predict the temporal occurrence and timely decision making the fight the crimes.

REFERENCES

Alpaydın, E. (2014). *Introduction to machine learning*. *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-62703-748-8-7

Calder, R. (2018). Crime in New York City hotels has skyrocketed. Retrieved July 13, 2018, from https://nypost.com/2018/06/04/crime-in-nyc-hotels-has-skyrocketed-in-recent-years/

Cerdá, M., Tracy, M., Messner, S. F., Vlahov, D., Tardiff, K., & Galea, S. (2009). Misdemeanor Policing, Physical Disorder, and Gun-related Homicide: A Spatial Analytic Test of "Broken-Windows" Theory. Epidemiology. Lippincott Williams & Wilkins. https://doi.org/10.2307/25662699

Dong, W., Lepri, B., & Pentland, A. (2011). Modeling the co-evolution of behaviors and social relationships using mobile phone data. *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia - MUM '11*. https://doi.org/10.1145/2107596.2107613

Honan, K. (2018). New York City Murders on the Rise in 2018, NYPD Data Shows - WSJ. Retrieved July 13, 2018, from https://www.wsj.com/articles/new-york-city-murders-on-the-rise-in-2018-nypd-data-shows-1531262595

iHLS. (n.d.). AI Cutting Edge Tools To Fight Crime. Retrieved July 13, 2018, from https://i-hls.com/archives/81784

Saccilotto, R. T., Nickel, C. H., Bucher, H. C., Steyerberg, E. W., Bingisser, R., & Koller, M. T. (2011). San Francisco Syncope Rule to predict serious short-term outcomes: A systematic review. *CMAJ*. https://doi.org/10.1503/cmaj.101326

Wakefield, K. (n.d.). A guide to machine learning algorithms and their applications | SAS UK. Retrieved July 13, 2018, from https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html