Prediction Analysis Of Criminal Data Using Machine Learning

IOP Conf. Series: Materials Science and Engineering **852** (2020) 012164 doi:10.1088/1757-899X/852/1/012164

Prediction Analysis Of Criminal Data Using Machine Learning

Meiliana[1*], Dedi Trisnawarman[2], Muhammad Choirul Imam[3]
[1,2,3,4)] Information Systems Study Program in Universitas Tarumanagara, Jakarta Indonesia

* meiliana.825160021@stu.untar.ac.id

Abstract. The human-being life necessities have encouraged the commission of crime since ancient times. Nowadays, many crimes factor and techniques have developed in a more sadistic way, causing more victims and loss. The authorized parties need to find a way to minimize the commission of a crime. This research aims to utilize linear regression algorithm for analyzing crime data to generate predictions for crime, which shows a quite reliable result. The result can be used by the authorized parties to help them prevent and handle upcoming crime.

1. Introduction

Crime is an act that can be potentially harmful to some individuals, a community, society, or the state that is forbidden and punishable by law [1], [2]. In Indonesia, crime record shows a rising trend in May 2019 [3]. Those quite high crime records not only harmful for the victims but also making citizens feel insecure.

Machine learning exists as technological development driven by the human need to analyze big data. The combination of big data and machine learning will drive incredible innovation across pretty much every industry [4]. As for maintaining security, they can be used to analyze crime data accurately to help prevent and minimize upcoming crime. Linear regression algorithm is one of the algorithms that can be used to make analytic predictions. This study uses the linear regression algorithm to predict crime.

2. Previous Work

The utilization of machine learning for predicting crime patterns should be supported by the right algorithm to generate more accurate crime prediction. Previous researcher has tried to use the K-Nearest Neighbor and boosted decision tree algorithm. Still, the percentage of prediction accuracy only ranged between 39% to 44%, which is quite poor to rely on [5]. Commonly used algorithms such as K-Nearest Neighbor, NaïveBayesian, decision trees, support vector machines aren't considered to generate accurate predictions [6]. The comparison of linear regression, additive regression, and decision stump to prove the level of effectiveness and accuracy for analyzing crime patterns shows that the linear regression algorithm gives the best result overall [7].

3. Method

3.1. Data Source

The data used in this research are secondary, namely crime data from 2010 to 2019, located in Los Angles. The dataset used is data from the Los Angles Police Department (LAPD) [8], which has been collected since 2010 and is updated monthly. The dataset used has 2.036.897 rows and provide information about the type of occurred crime, area of the incident, time of the incident, the age and sex of the victim, as well as the weapons used by the perpetrators.

3.2. Preprocessing

The dataset needs to go through preprocessing to maintain data quality before further analysis is processed to produce more accurate predictions. In the preprocessing stage, the missing value will be filled, the data that contains noise will be changed, and data duplication will be deleted for the analysis process to run smoothly and generate better analysis. The preprocessing stage for each case can vary depending on the data [9]. In this study, the data obtained were quite consistent and had an appropriate format so that only cleansing was done to overcome the missing values and noises. Cleansing will be done using RapidMiner to resolve missing values and noises.

3.3. Prediction

Regression can be used to performs operations on a dataset where the target has numerical values and has been defined [10]. The data needed for regression are two-part, first section for defining model and the other for testing model [11]. The application of a linear regression algorithm to generate prediction will be made by Python.

In this study, the data will be divided into two, namely for training and testing. The training section will analyze and determine the model using 75% of the criminal record data, while the testing section to test the model will be done by the rest 25%. The model build will be tested by matching the prediction results between the prediction and original criminal record data. The linear regression algorithm is shown in the equation ($y = m\,x \pm c$) with m as the level of the relation of variable x with the prediction of y, and c as the bias value. To find out the accuracy of the prediction results with the original data, it can be calculated with the equation (1) and (2):

$$R^2 = 1 - rMSE \tag{1}$$

$$rMSE = \frac{n-1}{n} \times \frac{\sum_i^n \in_i^2}{\sum_i^n (x_i - E(x))^2} = \frac{MSE}{Var(x)} \tag{2}$$

4. Result

Before preprocessing is done, the data shows the presence of missing values and noises. Table 1 shows preview of data before preprocessing, and Table 2 shows the statistical data before preprocessing, which show missing values and some noises.

Table 1. Preview Data Before Preprocessing

Row No.	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	Mocodes	Vict Age
111	100100804	Apr 15, 2010	Apr 15, 2010	335	1	Central	155	1	330	BURGLARY F...	0433	32
112	100100807	Apr 15, 2010	Apr 15, 2010	1740	1	Central	153	1	440	THEFT PLAIN...	0344	54
113	100100809	Apr 16, 2010	Apr 15, 2010	2120	1	Central	105	2	920	KIDNAPPING...	?	18
114	100100811	Apr 16, 2010	Apr 16, 2010	510	1	Central	154	1	110	CRIMINAL H...	1100	41
115	100100812	Apr 16, 2010	Apr 16, 2010	800	1	Central	152	1	441	THEFT PLAIN...	1402	0
116	100100813	Apr 16, 2010	Apr 16, 2010	1800	1	Central	111	1	220	ATTEMPTED ...	0337 0355 0...	18
117	100100814	Apr 16, 2010	Apr 16, 2010	1815	1	Central	111	1	220	ATTEMPTED ...	0337 0400	56
118	100100816	Apr 16, 2010	Apr 16, 2010	1830	1	Central	155	1	210	ROBBERY	0344 0416	39
119	100100822	Apr 16, 2010	Apr 15, 2010	1400	1	Central	158	1	236	INTIMATE PA...	0416 0417 0...	36
120	100100826	Apr 17, 2010	Apr 17, 2010	1300	1	Central	124	2	745	VANDALISM -...	0329 1402	0

Tabel 2. Statistical Data Before Preprocessing

Name	Type	Missing	Statistics		Filter (28 / 28 attributes): Search for Attributes
Mocodes	Polynominal	220033	Least: 9999 2002 (1)	Most: 0344 (207170)	Values: 0344 (207170), 0329 (86516), ...[462814 more]
Vict Age	Integer	0	Min: -9	Max: 118	Average: 31.789
Vict Sex	Polynominal	189700	Least: - (1)	Most: M (940018)	Values: M (940018), F (857693), ...[4 more]
Vict Descent	Polynominal	189746	Least: - (3)	Most: H (699906)	Values: H (699906), W (492850), ...[18 more]

The missing value and noise are filled in using the average value or by the mode value of each missing attribute, depends on the datatype. The average value is used to fill numeric value, while the mode is used for non-numeric value. Table 3 shows the data display after going through preprocessing and Table 4 shows the statistic data after preprocessing.

Table 3. Preview Data After Preprocessing

Row No.	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Crm Cd	Crm Cd Desc	Mocodes	Vict Age
111	100100804	Apr 15, 2010	Apr 15, 2010	335	1	Central	155	330	BURGLARY F...	0433	32
112	100100807	Apr 15, 2010	Apr 15, 2010	1740	1	Central	153	440	THEFT PLAIN...	0344	54
113	100100809	Apr 16, 2010	Apr 15, 2010	2120	1	Central	105	920	KIDNAPPING...	0344	18
114	100100811	Apr 16, 2010	Apr 16, 2010	510	1	Central	154	110	CRIMINAL H...	1100	41
115	145947499	Apr 16, 2010	Apr 16, 2010	1346	11	Central	1153	506	THEFT PLAIN...	1402	38
116	100100813	Apr 16, 2010	Apr 16, 2010	1800	1	Central	111	220	ATTEMPTED ...	0337 0355 0...	18
117	100100814	Apr 16, 2010	Apr 16, 2010	1815	1	Central	111	220	ATTEMPTED ...	0337 0400	56
118	100100816	Apr 16, 2010	Apr 16, 2010	1830	1	Central	155	210	ROBBERY	0344 0416	39
119	100100822	Apr 16, 2010	Apr 15, 2010	1400	1	Central	158	236	INTIMATE PA...	0416 0417 0...	36
120	145947499	Apr 17, 2010	Apr 17, 2010	1346	11	Central	1153	506	VANDALISM -...	0329 1402	38

Table 4. Statistical Data After Preprocessing

Name	Type	Missing	Statistics			Filter (22 / 22 attributes): Search for Attributes
Mocodes	Polynominal	0	Least: 0418 0369 1305 (0)	Most: 0344 (427207)	Values: 0344 (427207), 0329 (86516), ...[462814 more]	
Vict Age	Integer	0	Min: 2	Max: 118	Average: 38.409	
Vict Sex	Polynominal	0	Least: - (0)	Most: M (1129719)	Values: M (1129719), F (857693), ...[4 more]	
Vict Descent	Polynominal	0	Least: - (0)	Most: H (889656)	Values: H (889656), W (492850), ...[18 more]	

The results of linear regression algorithm application for the crime data using Python are shown by using line chart. Figure 1 shows the comparison of original data (black line) with the predicted results (blue line) to test the accuracy of the prediction model. From the linear regression algorithm, the equation results obtained are $(\boldsymbol{y = 168.91428571\, x \pm c})$ and $(R^2 = 0.19)$.

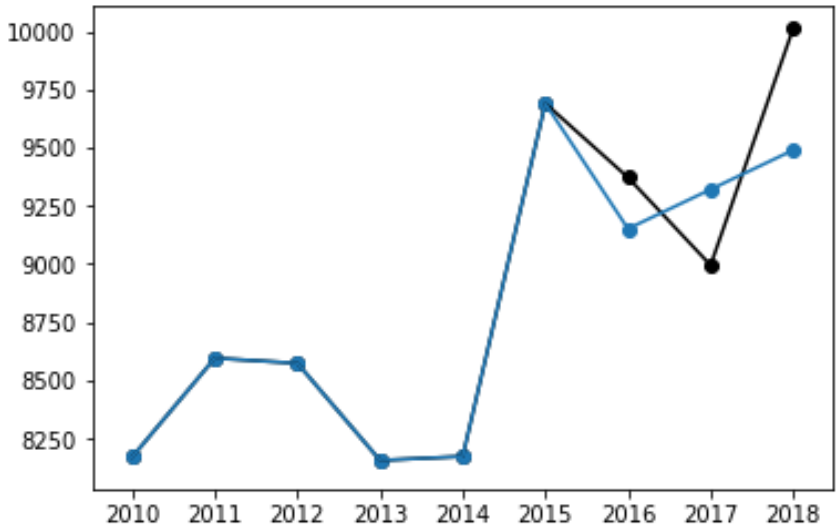

Figure 1. Comparison of Original Crime (Black) and Prediction (Blue)

5. Conclusion

Linear regression algorithm can be used to predict criminal data with a value of $R^2 = 0.19$. These results indicate that the linear regression algorithm is good enough to be used for prediction. The implementation of linear regression for analytical predictions can be applied in the Indonesian region provided that it has a criminal database that is feasible to analyze because the attributes used in this study are attributes that also exist in Indonesia.

6. References

[1] Oxford: Oxford University Press. 2009. Oxford English Dictionary Second Edition on CD-ROM.

[2] Farmer, L. 2008. Crime, definitions of. Cane and Conoghan (editors), The New Oxford Companion to Law, Oxford University Press.

[3] https://www.cnnindonesia.com/nasional/20190517062637-12-395609/angka-kriminalitas-naik-polri-fokus-empat-kasus-kejahatan

[4] Alfred, R., 2016, October. The rise of machine learning for big data analytics. In *2016 2nd International Conference on Science in Information Technology (ICSITech)* (pp. 1-1). IEEE.

[5] Kim, S., Joshi, P., Kalsi, P.S., & Taheri, P. 2018. Crime Analysis Through Machine Learning. 2018 IEEE 9th *Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 415-420.

[6] Kumar, K.S. and Bhalaji, N., 2016, August. A Study on Classification Algorithms for Crime Records. In *International Conference on Smart Trends for Information Technology and Computer Communications* (pp. 873-880). Springer, Singapore.

[7] McClendon, L. and Meghanathan, N., 2015. Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, *2*(1), pp.1-12.

[8] https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/63jg-8b9z

[9] Kontostathis, A., Edwards, L. and Leatherman, A., 2010. Text mining and cybercrime. *Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK*, pp.149-164.

[10] Brook, R.J., 2018. *Applied regression analysis and experimental design*. Routledge.

[11] Gharehchopogh, F.S., Bonab, T.H. and Khaze, S.R., 2013. A linear regression approach to prediction of stock market trading volume: a case study. *International Journal of Managing Value and Supply Chains*, *4*(3), p.25.