Research on the Criminal Recidivism Prediction Based on Machine Learning Algorithm

Jiaxin Zhang[1*]

[1]Chinese University of HongKong (Shenzhen), Shenzhen, China

*120090580@link.cuhk.edu.cn

Abstract. Criminologists and social security personnel around the world have found that the risk of released criminals is much higher than that of people who have not committed crimes, and preventing people with criminal records from recommitting crimes should be one of the strategic priorities of social crime prevention. Therefore, risk assessment of criminal recidivism has been used to improve social security by predicting the criminal recidivism of offenders. In order to predict criminal recidivism, this article applied machine learning (ML) algorithms models (KNN, random forest, support vector machine and logistic regression) on the data set of the basic information about 10,000 criminal defendants in Broward County, Florida and their recidivism within two years. The predictive accuracy of models used in this article was between 0.64 and 0.67, with AUC ranging between 0.65–0.72. The AUC value of logistic regression is highest with 0.713 while support vector machine has the highest accuracy reaching to 0.671. This study provides a reference on selecting best method to predicting criminal recidivism.

Keywords: Machine Learning Algorithm, Recidivism Prediction

1 Introduction

Criminology is a negative phenomenon that endangers social welfare and safety of residents. It is a big issue that human beings always struggled with. Minimizing crime is one of the most important conditions for maintaining the sustainability of a society, thus enabling people to live peacefully and positively. Without peace, a society cannot achieve social and economic prosperity [1]. Therefore, the analysis of crime reports and statistics to prevent crime is very important to establish the security of residents.

The concept of algorithms risk assessment is first proposed in 1920s [2]. For nearly a century, researchers studying justice and crime have concluded a variety of outcomes [3]. In the past 20 years, there has being a growing debate about the application big data and machine learning algorithms in criminal justice and criminology. But it is hard to say how accurate they are because most of the predictions are not properly assessed [3]. With the rise of risk assessment, people are starting to focus on the topic of whether risk assessment should be applied into the justice department [4]. Proponents argue that

V. Escudero et al. (Eds.): BADS 2022, AHCS 11, pp. 1297–1306, 2023.
https://doi.org/10.2991/978-94-6463-102-9_134

risk assessment can act as a crime prevention tool to reduce recidivism rates. The other people argue that the accuracy of prediction can not be guaranteed.

Machine learning encompasses computer science and statistics. It contains data science and artificial intelligence. It is one of the fastest growing technology tools today [5]. It has been widely used in Medicine, justice, manufacturing and education. There has been a large amount of research in the past on the use of ML methods to predict criminal recidivism [4]. In the early days of ML application, Duwe and Kim (2017) predicted the recidivism of homicide offender released from Minnesota prisons and compared 12 supervised learning algorithms [6]. In recent years, more and more researchers are interested in the application of big data in predicting recidivism rates. Ghasemi at el (2021) applied ML algorithms, including DT, RF, and SVM to two data sets provided by Correctional Services (MCSCS) and the Ontario Ministry of Community Safety[7]. Wang et al (2022) studied interpretable recidivism prediction with ML models and analyzed their ratings for predictive power, sparsity, and fairness [4]. Risk assessment has been treated as sensible.

This article predicts the recidivism of the criminals in Florida with machine earning algorithms, including KNN, SVM, logistic regression and random forest. It provides a help and method for the future study of recidivism prediction. Furthermore, this research also provides basis for the relevant public security organs and judicial departments to take appropriate measures to effectively prevent crimes.

2 Data and Model

2.1 Data

COMPAS is a popular commercial algorithm. It is usually used by judges and parole officers to score the likelihood of recidivism of a criminal defendant recidivism. The data set contains variables used by the COMPAS algorithm when scoring defendants, as well as how they scored more than 10,000 offenders from Florida, in the two years following the verdict. 2 subsets of the data are provided, including a subset of only violent crime and another subset with information about offenders and their recidivism condition within two years.

The first data set (D1) includes three COMPAS scores received by each pretrial defendant: Risk of Failure to Appear', 'Risk of Violence' and 'Risk of Recidivism'. Each defendant was compared by COMPAS algorithm on a scale of 1 to 10, with 10 being the riskiest. The score of 1 to 4 is marked as "low" by comparison; The score of 5 to 7 is labeled "moderate"; The score of 8 to 10 is labeled "high." Figure 1a and figure 1b provide summary statistic of data sets D1, including three kinds of scores.

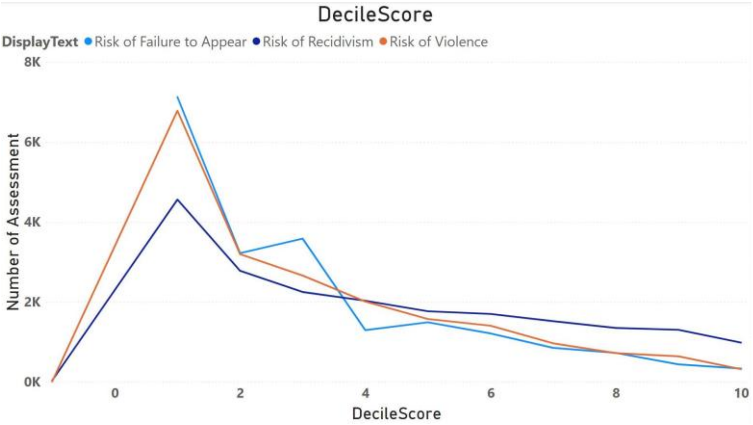

Fig. 1. a Decile score. The count of assessments for different decile score (self-painted)

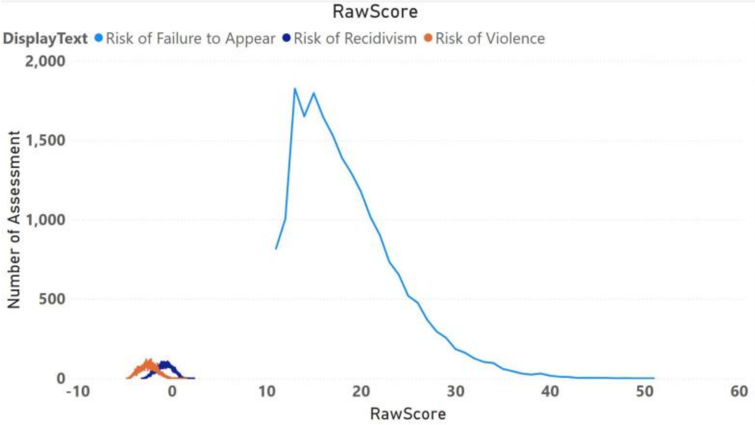

Fig. 1. b Raw score. The count of assessments for different raw score (self-painted)

The second data set (D2) includes age, gender, region, number of priors, scores factor and recidivism within two years of each criminal. In most analysis, this article defined recidivism as a new arrest within two years. It relies on Northpointe's guidelines, which state that its recidivism score is to predict 'a new misdemeanor or felony offense within two years of the date'. Figure 2 shows simple distribution of D2.

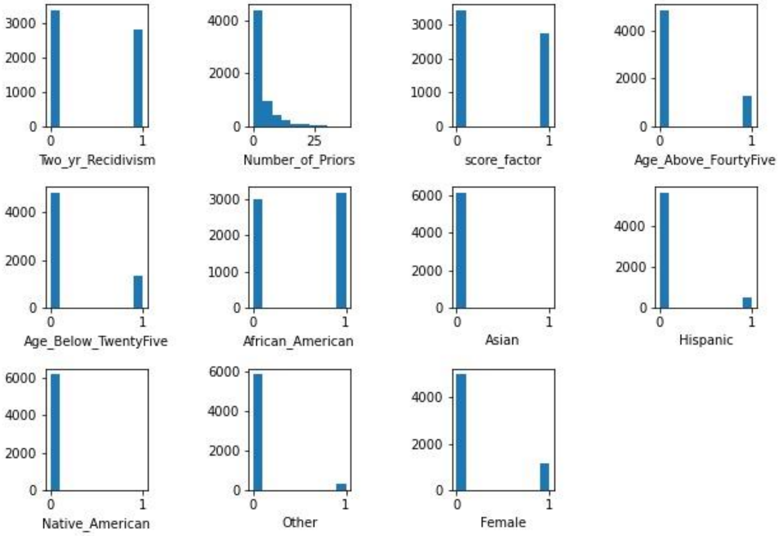

Fig. 2. Data distribution of D2. (self-painted)

2.2 Methodology

Prediction method.

According to Samuel (1959), Machine learning is about learning of some of the computer's data and then making the prediction and judgment of some other data. [8]. That is, the acquired data is used by computer to fit some models and then the suitable models are applied to new data and making predictions. This is similar to the way humans acquire new knowledge in some degree. Machine learning mainly focuses on making computers to learn features of given things without being directly programmed, and it is a branch of computer science [9]. Machine learning is used to 'teach' machine handle large amounts of historical data and identify patterns in the data by machine learning algorithms. There are various kinds of machine learning algorithms can be used to fit the training data set. For the project, supervised learning is used to make the predictions. Supervised learning requires to learn the relationship between input data and output data, through which the prediction of invisible data can be accomplished [10]. The supervised learning algorithms used in this program is briefly introduced below.

1. KNN. The full name of KNN is K-nearest neighbors. The KNN method is only relevant to a very small number of neighboring samples when it involves category decision making. Therefore, KNN method is more suitable than other methods for sample sets with intersecting or overlapping class domains. The basic idea of k-Nearest Neighbors (KNN) is to determine the category of a given query based not only on the docu-

ment that is closest to the point to be classified in the given space, but on the k categories that are closet to it [11].

2. SVM. SVM (support vector machine) provides an advanced learning method, which has achieved great success in a variety of applications [7]. There has been a lot of interest in using kernels in various machine learning problems in recent years, especially SVM model [12]. The basic motivation of a support vector machine is to find a decision hyperplane that maximizes the interval between two data types, construct the objective function based on its interval maximization, and then transform it into its dual problem for solution [4]. It solves problems in small samples. SVM has many unique advantages in nonlinear and high-dimensional pattern recognition problems. And it largely avoid 'overfitting' [13].

3. Random Forest. Decision tree is a classical machine learning algorithm. As a tree model, the tree structure is intuitive and interpretable, for which is widely used in the field of data analysis. Although decision tree has the advantages of simple, intuitive, strong interpretability, it is easy to overfit. Therefore, the random forest algorithm was developed to address this problem. The Random forest algorithm is a typical parallel integrated learning method, where a set of individual decision tree learners without strong dependencies on each other is first constructed in parallel, and then some strategy is used to combine them. In solving the classification problem, the random forest method selects the majority as the final result according to the classification result of each tree. In solving the regression problem, the random forest method calculates the mean value of each tree as the result. Random forest algorithm is easily parallelizable and improves the tolerance of the noise for the algorithm, so it has the potential to deal with large real-life systems [14].

4. Logistic Regression. Although called regression, logistic regression is often used for binary classification and is actually a classification model. Logistic regression is favored by industry for its simplicity, parallelism, and interpretability. Linear regression is a widely used prediction model, but it is not appropriate when the correct model is parametric nonlinearity [15]. There are only two predicted results in this study: 0 for a relatively low possibility of recidivism within two years, and 1 for a relatively high possibility of recidivism within two years. So logistic regression is a reasonable classification approach for criminal recidivism prediction.

Evaluation method.

1. Confusion Matrix. The article conduct four types of machine learning algorithms to make prediction. To evaluate the performance of these models, there must be a nature method can be applied to all the models. So the author randomly choose 20 percent of the data set (D2) to be the test set and compare the outcomes with available data. Confusion matrix is a measurement method often used to solve classification problems. It can be used for binary and multi-class classification problems [17]. The following form is the one it is often summarized into:

$$\binom{TP\,FP}{FNTN} \qquad (1)$$

In the machine learning field, the confusion matrix is also known as the likelihood matrix. It is a visualization tool, especially for supervised learning. The number of cases correctly predicted as positive is abbreviated as ‘TP’; the number of cases correctly predicted as negative is abbreviated as ‘TN’ ; the number of cases incorrectly predicted as negative is abbreviated as ‘FN’; the number of cases incorrectly predicted as positive is abbreviated as ‘FP’. The following formula calculates the accuracy of algorithm:

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} \quad (2)$$

2. K-Fold Cross-validation. The training set data into K parts in K-fold cross-validation divides. K-1 of them as testing and another one part as training. All the samples in the training set are bound to become the training data and the pages are bound to become the test set once. It is a great advantage for K-fold cross-validation. The basic form of cross-validation is k-fold cross-validation [16]. The four machine learning models are built with k-fold cross validation (K=5). K-fold cross-validation can avoid overfitting and underfitting effectively. This article uses the mean of the k-fold cross-test as an indicator of model prediction accuracy.

3 Result

3.1 Accuracy of prediction model

To make a prediction of the recidivism of offenders, the article analyses the second data set (D2) through setting up SVM, KNN, logistic regression and random forest to fit the data.

Problem setting.

These prediction problems are treated as binary classification problems in analysis because of the characteristic of binary nature for recidivism task. The prediction results includes only ‘0’ and ‘1’. ‘0’ means the likelihood of reoffending within two years is relatively low, ‘1’ means the likelihood of reoffending within two years is relatively high. In this paper, SVM, KNN, logistic regression and random forest are used to forecast criminal recidivism, and the software used is Python.

Accuracy of predictions.

The mean scores value of k-fold validation for the four models set up through python are: KNN, 0.652632; RF, 0.648583; LR, 0.655870; SVM, 0.671255. The scores are directly returned by score function in sklearn. It represents the coefficient R^2 of this prediction. The closer the score is to 1, the better the model performs in the test set.

Table 1. Mean value of k-fold cross-validation. (self-painted)

Machine learning algorithms	Mean value of k-fold cross-validation
KNN	0.652632

Random Forest	0.648583
Logistic Regression	0.655870
SVM	0.671255

3.2 Results of evaluation on models

By calculating the performance evaluation methods of recidivism prediction for the KNN, RF, Logistic regression, and SVM methods. Figure 3 shows ROC curve of each model. It can be observed that the AUC value of all the models are not in a big difference (AUC values are generally around 0.69). Logistic regression model performs a little better than other models, and its AUC value reaches to 0.713.

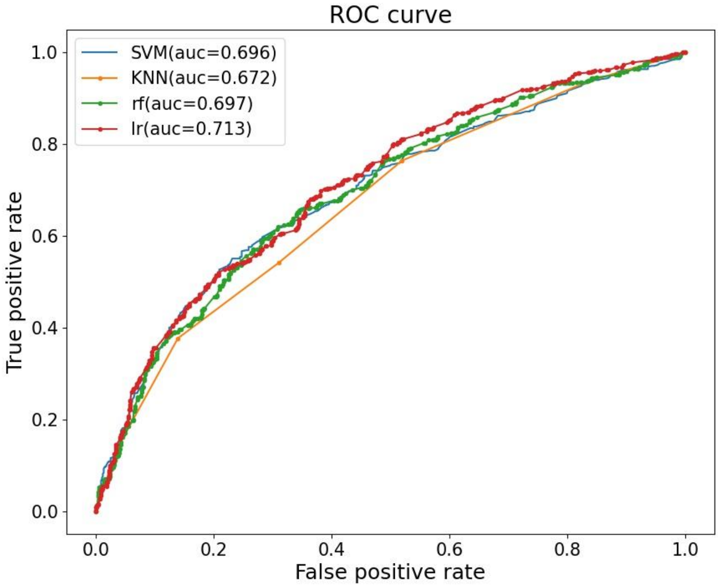

Fig. 3. ROC curve and AUC value of the four models. (self-painted)

An ROC curve shows the performance of classification model at all different thresholds. ROC curve refers to the line of each point drawn with the false positive probability and true positive probability as abscissa and ordinate under specific stimulus conditions. In the ROC graph, the horizontal axis is false positive rate and the vertical axis is true positive rate. The area under the ROC curve enclosed by the coordinate axes is the definition of AUC value. It is easy to know that the value of this area will always be smaller than 1. In addition, the AUC value is bigger than 0.5 because the ROC curve is always above the line y=x. The closer the AUC is close to 1.0, the higher the reliability of the detection method is. Generally speaking, the AUC results are considered excellent for AUC values between 0.9-1.0 and failed for AUC values under 0.5.

4 Discussion

From this analysis, the article concludes that machine leaning as a tool can be used to forecast criminal recidivism. It is noteworthy that the support vector machine model essentially outperforms the other three models in terms of predictive accuracy. According to the mean value of k-fold cross-validation, the four models are not in a very big difference for accuracy. SVM model has the highest accuracy in fitting the test sets. The trend of ROC curve and AUC value are also similar for these four models. Logistic regression preforms a little better with higher AUC value among all the models. In particular, the logistic regression performs the best in terms of AUC and ACC for this data set (D2). Although the accuracy of these four methods is not very different, all around 0.65, it is still not very ideal. This is probably because there are too few features in the data used to make predictions.

Many researchers have proved that machine learning approaches is a tool for effective recidivism risk prediction. Initially, Liu et al. (2011) analyzed the the accuracy of prediction for classification, neural networks, logistic regression, and regression tree models. The data they used was about a prospective sample of 1225 male prisoners in UK. The accuracy of the three models was between 0.59 and 0.67, with AUC value ranging from 0.65 to 0.72 [18]. Ozkan (2017) fitted data including recidivism of released prisoners in 1994 from the Bureau of Justice Statistics. The random forests, XGBoost, neural networks and SVM models were used in his research. The researcher found neural networks and XGBoost performed better than other models when comparing (AUC betwen 0.79-0.83). [20]. Currently, Wang and his team (2022) studied interpretable recidivism prediction with ML algorithms and analyzed the sparsity, predictive power, and fairness (AUC between 0.65-0.72) [4]. It can be found that the prediction results of this paper are basically within the same range as those of previous studies. (AUC between 0.68-0.72).

Indepth analysis by ProPublica can be found in the data methodology article by researchers. The researchers analyzed that black defendants were twice as likely as white defendants to be misclassified as being at risk for violent recidivism. It stands to reason that researchers need more detailed delineation of offender characters, such as race and area, to achieve higher predictive accuracy. Researchers can set up different models for different races and different areas.

The primary purpose of risk assessment should not be only prediction. Instead, the main orientation is prevention [7]. The reasonable use of machine learning for recidivism prediction and risk assessment can more effectively reduce the recidivism.

5 Conclusion

Machine learning has been widely used in various fields and there has been an increasing demand of risk assessment. The use of risk assessments has had a large impact on corrective classification over the past few decades [19]. Machine learning is encouraged to be used correctly for further assessment and prediction. In this paper, the recidivism problem of criminals is studied by several ML methods, including KNN, SVM, RF and logistic regression. The recidivism problem was predicted as a classification problem in the analysis.

This article compares the accuracy of KNN, SVM, random forest and logistic regression method in predicting recidivism and found that logistic regression preforms better in D2. In perspective of accuracy, SVM performs the best and the accuracy of it has reached to 0.671. According to AUC value and ROC curve, the logistic regression has the highest AUC value with 0.713. All in all, logistic regression performs the best among these four models, although it has no significantly improvement than others.

However, the accuracy of all the models are not very ideal. It is possible that different regions, different races need different models to fit. This conjecture also provides an idea for the future research, where researchers can build different models for different regions and different races to improve the accuracy of machine learning models. Further research might be needed to investigate the best way to fit the data and make predictions.

References

1. Safat, W., Asghar, S., & Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. IEEE Access, 9, 70080-70094.
2. Bureau of Justice Assistance (2020) History of risk assessment. Bureau of Justice Assistance. DOI: https://psrac.bja.ojp.gov/basics/history337–351.
3. Berk, R. (2008). Forecasting methods in crime and justice. Annual review of law and social science, 4, 219-238.
4. Wang, C., Han, B., Patel, B., & Rudin, C. (2022). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. Journal of Quantitative Criminology, 1-63.
5. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
6. Duwe, G., & Kim, K. (2017). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. Criminal Justice Policy Review, 28(6), 570-600.
7. Ghasemi, M., Anvari, D., Atapour, M., Stephen Wormith, J., Stockdale, K. C., & Spiteri, R. J. (2021). The Application of Machine Learning to a General Risk–Need Assessment Instrument in the Prediction of Criminal Recidivism. Criminal Justice and Behavior, 48(4), 518-538.
8. Samuel, A. L. (1959). Machine learning. The Technology Review, 62(1), 42-45.
9. Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. American journal of epidemiology, 188(12), 2222-2239.
10. Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In Machine learning techniques for multimedia (pp. 21-49). Springer, Berlin, Heidelberg.
11. Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application, 7(1), 61-70.
12. Tsang, I. W., Kwok, J. T., Cheung, P. M., & Cristianini, N. (2005). Core vector machines: Fast SVM training on very large data sets. Journal of Machine Learning Research, 6(4).

13. Ding, S. F., Qi, B. J., & Tan, H. (2011). A review of support vector machine theory and algorithm. Journal of University of Electronic Science and Technology, 40(1), 2-10.
14. Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.
15. DeMaris, A., & Selman, S. H. (2013). Logistic regression. In Converting Data into Evidence (pp. 115-136). Springer, New York, NY.
16. Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5, 532-538.
17. Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In data democracy (pp. 83-106). Academic Press.
18. Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. Journal of Quantitative Criminology, 27(4), 547-573.
19. Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2001). Risk/need assessment, offender classification, and the role of childhood abuse. Criminal Justice and Behavior, 28(5), 543-563.
20. Ozkan, T. (2017). Predicting recidivism through machine learning (Doctoral dissertation).