Predicting criminal recidivism: Comparing Statistical and Machine Learning Predictive Model

Nurul Najibah Sabrina Ibrahim (16b9034)

Universiti Brunei Darussalam

Abstract

In the field of predicting recidivism, there is no standard that a modeling strategy should be pursued to achieve an optimum predictive model. The aim of this research is to compare statistical and machine learning predictive model in predictive accuracy regarding recidivism. For this purpose, 10 types of statistical and machine learning models were investigated in terms of their performance by finding the error rate and Brier score of the train and test for every model. A range of statistical and machine learning models is fitted on National Corrections Reporting Program data collected from 1991 to 2014. For predicting recidivism neural network is the best predictive model. Although the neural network performed well in this research relative to the other models, more research is also needed to define more predictive variables and models.

Keyword: Recidivism, prediction, machine learning

Introduction

Recidivism refers broadly to reoffending, and common measurements of recidivism include rearrests, reconviction, reincarceration, or a supervision violation (for individuals on probation or parole). Recidivism rates measure the frequency with which individuals reengage with the criminal justice system. Recidivism measures can provide policy makers with information regarding the relative threat to public safety posed by various types of offenders, and the effectiveness of public safety initiatives in deterring crime and rehabilitating or incapacitating offenders. There are multiple ways to measure recidivism and each measure relies on a

somewhat different definition of reinvolvement, hence providing a different picture of the problem at hand. Reducing recidivism is important as it improves public safety, reduces taxpayer spending on prisons, and helps formerly incarcerated people successfully resume family and community responsibilities. However, a lack of data has complicated efforts to understand the collective effects of countless federal, state, and local efforts to reduce reoffending. Despite recidivism measure (or combination of measures) a jurisdiction employs, it is important to collect and analyze these data regularly and consistently to examine system functioning, effectiveness, costs, and trends.

According to Bureau of Justice Statistics (2000) in 1999 there were 6.3 million people on probation or parole, in jail or in prison. This figure represents an average annual increase of 5.8 percent since 1990. Due to the increase of budgetary constraints and public pressure over crime rates, criminal justice officials have sought out new ways to deal with offenders more effectively. The development of effective methods for predicting whether an individual released from prison eventually returns or not is a major concern in criminology. A simple model for predicting parole outcomes was propose by Burgess as early as 1928 and was followed by the introduction of a variety statistical models for classifying recidivist. (Caulkins, et. al., 1996).

The objective of recidivism prediction is to gain insight into criminal behavior through analysis of historical criminal recidivism data. Algorithms for predicting recidivism are commonly used to assess a criminal defendant's likelihood of committing a crime. Proponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans. However the performance of these models has been uneven at the best due to the lack of relevant variables and the limitations of the models designed.

With more accurate predictive tools, the criminal justice system would be better at correctly classifying high-risk offenders and adjusting their stay in prison accordingly. This may prevent future crimes of these offenders, which otherwise may have occurred if they had not been incapacitated. For this reason, identifying these high-risk offenders at stages such as parole decisions and treatment allocation can reduce recidivism rates. Moreover, a more accurate prediction can be of help by allowing probation officers to adjust their time and energy on the high-risk offender group (within released population) informed by actuarial analyses.

Consequently, designing tools that can assess future recidivism risk more accurately than what is already being used is of utmost importance in crime control and public security. The need for better, more accurate predictive models is real and, it is wiser to seek for superior prediction for the criminal justice system (Bushway, 2013).

Objectives

This research will be conducted in order to answer the following questions:

1. How well is it possible to predict recidivism committed by offenders released from prison based on (1) the offense type, (2) characteristics of the offender, such as sex, race and age group, (3) time served?

2. Which machine learning and statistical method provides us with the best possible prediction?

Data

The data used in this research were retrieved from the National Corrections Reporting Program collected from 1991 to 2014 by the United States Department of Justice, Bureau of Justice Statistics. The data gathered are selective variables and are used to monitor the nation's correctional population and address specific policy questions related to recidivism, prisoner reentry, and trends in demographic characteristics of the incarcerated and community supervision populations. Participation in the data collection is voluntary. Not all states participate and not all states have participated for each year. Some records were excluded due to noticeable defects, such as the inclusion of samples where the individual's date of release was in fact not during the period of time that defined the data set and missing information on one or more variables. Altogether, there are n = 10,907,333 observations and p = 21 predictors.

Variables		Recidivism rate (%)
Offence Type	Violent	23.0
	Property	32.5
	Drugs	28.9
	Public Order	14.2
	Other	0.8
Sex	Male	90.9
	Female	9.1
Race	White	34.7
	Black	38.6
	Hispanic	15.6
	Other	2.1
Age Admitted	18 – 24	22.2
	25 – 34	37.0
	35 – 44	26.8
	45 – 54	11.7
	≥55	2.3
Time Served	< 1 year	58.6
	2 – 4.9 years	17.8
	5 – 9.9 years	13.0
	≥ 10 years	3.2

Table 1: Recidivism rate and the variables used to predict recidivism

After conducting a simple statistical analysis, it is discovered that there are 6,981,739 repeat offenders out of the 10,907,333 cases. In order to answer the question of this research, only 5 predictor variable were used in the model.

Models

Machine learning is when a dataset was loaded into a software program and select a model that helps the machine to come up with predictions that match the data. Via algorithms, the way the machine makes the model will range from a simple equation to a very complicated logic or mathematical system that brings the computer to the best predictions.

The following 10 models are compared with respect to their respective predictive performance.

1. **Linear Regression (LM)**: an analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The motive of the LM algorithm is to find the best values for the intercepts and coefficient of the variables to fit line for the data points. In general such a relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables; the unobserved deviations from the above equation the errors. Suppose that n data pairs were observed and called $\{(x_i, y_i), i = 1, ..., n\}$. It can describe the underlying relationship between y_i and x_i involving this error term εi by:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (1)$$

2. **Logistic regression (LR):** a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In binary outcomes, it has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". LR is a linear combination of predictors that computes fitted probabilities. It can be understood simply as finding the β parameters that best fit. where ε is an error distributed by the standard logistic distribution:

$$y = \begin{cases} 1 & \beta_0 + \beta_1 x + \varepsilon > 0 \\ 0 & otherwise \end{cases} \quad (2)$$

3. **Linear discriminant analysis (LDA)**: a tool for classification, dimension reduction, and data visualization. It is used to reduce the number of features to a more manageable number before the process of classification. LDA is the classical method of predicting the two-class model. It is a standard linear regression with a dummy independent variable and has more assumptions than logistic regression, regarding the underlying class distribution and equality of the class covariances. It generates a linear decision boundary.

4. **Lasso regression (L1 Logistics):** is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It is often used in machine learning to select the subset of variables. Consider a sample consisting of N cases, each of which consists of p covariates and a single outcome. Let y_i be the outcome and x_i: $(x_1, x_2, \dots, x_p)^T$ be the covariate vector for the *i*th case. Then the objective of lasso is to find, where *t* is a prespecified free parameter that determines the amount of regularization:

$$\min_{\beta_0,\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2\} \; subject\ to \; \sum_{j=1}^{p}|\beta_j| \leq t. \quad (3)$$

5. **Ridge regression (L2 Logistics):** is a way to create a tight model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables). This method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias.

6. **Support vector machine (SVM):** a model that forms vectors from feature space and constructs a hyperplane in multidimensional space. It uses this hyperplane to classify any new data points. That is, SVM searches for a hyperplane that divides one class from another, and it does so by maximizing the distance between the line and the data points. The more distance from a given training point results in better prediction, and these distances are called support vectors.

7. **Random forest (RF):** a decision-tree-based models classify individuals associated with different outcome categories by breaking up the data into subsets of individuals with similar characteristics. Random forests is an inductive procedure averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. Taking the teamwork of many trees thus improving the performance of a single random tree.

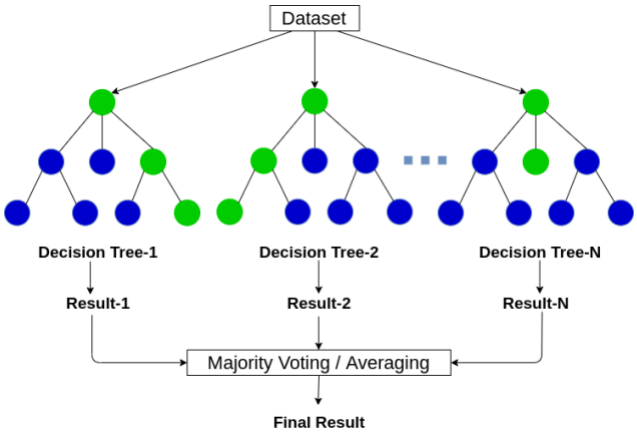

Figure 1. Random forest algorithm (Sharma, 2020)

8. **Neural Network (NN):** a machine learning system that uses a network of functions to understand and translate a data input of one form into a desired output, usually in another form. The concept of the artificial neural network was inspired by human biology and the way neurons of the human brain function together to understand inputs from human senses. Neural network can be used to approach any function of arbitrary form. It also use the hidden layer to make predictions more accurate.

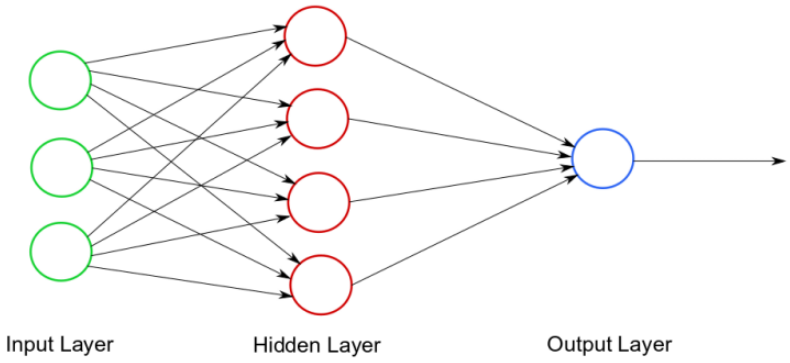

Figure 2: Simple neural network (Mayo, n.d.)

9. **K-Nearest Neighbors (KNN):** K-nearest neighbors is a rather simple method of classification. This algorithm finds a group of k observations closest to the test observation in terms of Euclidian distance. The proportion of the votes for the positive class is then returned as a probability.

10. **Gaussian Process for Classification (GPC):** is a non-parametric classification method. It assumes some prior distribution on the underlying probability densities that guarantees some smoothness properties. The final classification is then determined as the one that provides a good fit for the observed data, while at the same time guaranteeing smoothness.

Software used

All calculations and analysis were performed using R Mathematical Software. In the circumstance that of a simple calculation and graphing both Microsoft Excel and Google Sheets were used.

Methods

Due to computational restrictions, it is not feasible to fit the models using the entire dataset. Alternatively a random subsample of dataset (n = 2000 for training and 200 for testing) were used as an independent validation dataset to obtained the error rate and Brier score. Error rate is used to calculate the probability of an error occurring during the completion of a task. Meanwhile, Brier score is used to calculates the precision of probabilistic predictions. The process was repeated a total of B = 150 more times and the results were averaged. The inclusion criteria for the cases were as follows: (i) the general offense; characteristics of the offender, such as (ii) sex, (iii) race and (iv) age admitted; (v) time served in the prison.

To measure predictive ability, we fit the models on a random subset of the data and obtain the out-of-sample error rate and Brier score from the remaining held-out observations. We then

compare the results against statistical and machine learning predictive model namely: 1) linear regression; 2) logistic regression; 3) linear discriminant analysis; 4) lasso regression; 5) ridge regression; 6) support vector machine; 7) random forest; 8) neural network; 9) k-nearest neighbors; and 10) gaussian classification process. The experiment is set up as follows:

A random sample of 2200 cases were observed with the intention for a convenient assessment.

1. Form a training set by sub-sampling 2000 observations.
2. The remaining unsampled data (200) is used as the test set.
3. Fit model on training set, and obtain error rate and Brier score defined as

$$\text{Error rate} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \quad (4)$$

Where:

N = the number of data.

y_i = actual output value.

$\hat{y}_i$ = predicted output

Σ = sum of all the values.

$$\text{Brier Score} = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \quad (5)$$

Where:

N = the number data.

f_t = the prediction probability.

o_t = the outcome (1 recidivate, 0 not recidivate).

Σ = sum of all the values.

4. Repeat the process 150 times then get the average error rate and Brier score.

Results

Model	Error Rate (%)		Brier Score	
	Train	Test	Train	Test
Linear Regression	0.331	0.334	0.215	0.246
Logistic Regression	0.331	0.350	0.215	0.519
Linear Discriminant Analysis	0.331	0.334	0.216	0.248
Lasso Regression	0.335	0.644	0.216	0.218
Ridge Regression	0.335	0.640	0.217	0.336
Support Vector Machine	0.308	0.328	0.207	0.215
Random Forest	0.253	0.325	0.196	0.244
Neural Network	0.306	**0.317**	0.188	**0.193**
K-Nearest Neighbors	0.310	0.330	0.202	0.215
Gaussian Process Classification	0.292	0.326	0.197	0.210

Table 2: Mean out-of-sample error rate and Brier score for 2000 runs of various training(s) and 200 test(s) sizes for the recidivism.

Table 2 shows the performance of all 10 models for predicting recidivism on both train and test data. The error rate is almost the same for every model ranging from 0.30-0.33 except for Lasso (0.644) and Ridge Regression (0.640) which has the highest error rate differences. The best values are shown in bold-face across the rival models. Specifically, neural network had the lowest test error rate (0.317), lowest Brier test score (0.1.93). Support vector machine and random forest yielded identical results (0.215) for the Brier test score.

Following the Brier test Score, neural network model seems to outperform the other models. It is surprising that logistic regression has the highest Brier test score but low non-comparable error rate to neural network, which indicates that the prediction made using this model is not accurate.

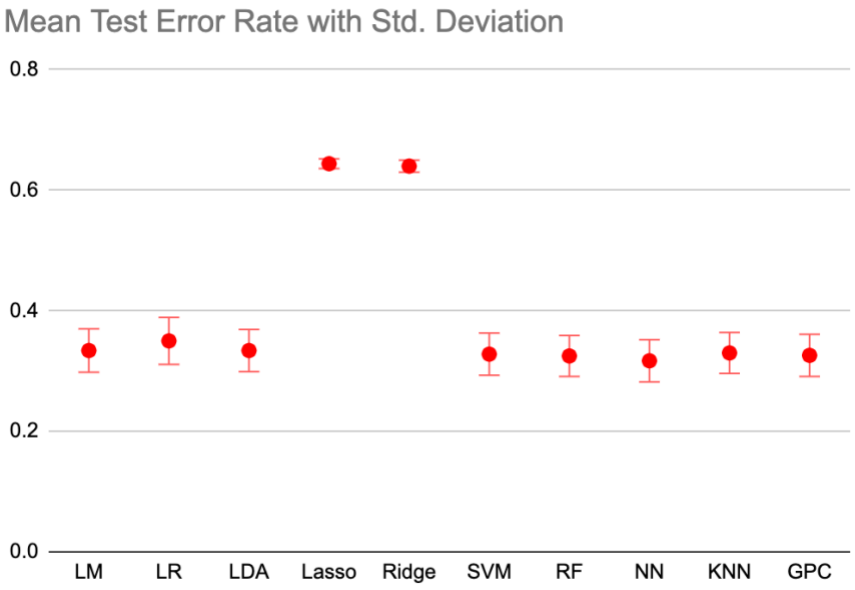

Figure 3: Mean test of error rate with standard deviation error bars

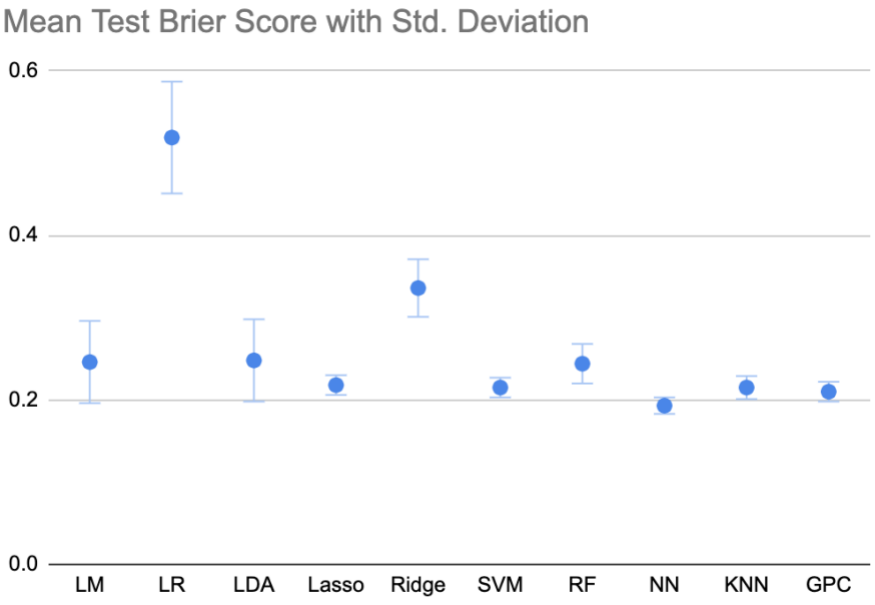

Figure 4: Mean test of Brier score with standard deviation error bars

Figure 4 can further shows that neural network is indeed the best model out of all the predictive model. Even though the differences of the mean Brier test score between neural network and the other model is not that high, the low standard deviation (error bar) shows that the prediction made by neural network is better compared to the rest of the model.

Brier score attempts to capture this phenomenon, where if the model predicts the outcome of 1 with probability 0.99 compare to another model predicting outcome 1 with probability 0.51, it is clear that the first model is better because it is more confident about its prediction. Therefore the lower Brier score indicates more accurate predictions while a higher Brier score indicates less accurate predictions (Rufibach, 2010).

Discussion

The purpose of this study is to find the best predictive model using the general offense, characteristics of the offender (sex, race and age) and time served in the prison as the predictive variables. There was a big difference in the predictive performance between logistic regression and neural network. According to Yang et al. (2010) for its accuracy and convenience, logistic regression can be chosen as the best predictor, while neural networks can be used with a large number of predictors with limited effect sizes in considerable sample sizes. In algorithmic decision making, there are two reasonable concerns. One is concealed algorithmic decision-making (data processing and storage black-box mode), and the other is obscurity, which suggests probable incoherence to human reasoning (Danaher, 2016).

Practitioners may prefer a model that is more intuitive and less opaque or black box. Black box models are the ones which are not implicit about selecting appropriate variables, weighting their influences and iteration process. Neural networks are certainly superior due to its hidden layers. It also use the hidden layer to make predictions more accurate. Neural network outperform most of these model but it does make sense since neural network performs very well if a big data was provided.

However judges, parole boards, jury or prosecutors want to understand what is going on within the black box. This can contribute to a choice over interpretable models compared to the best

predictive model (Rudin, 2019). For example, random forests can provide more information, or its relatively tree structure, which is reasonably simple to grasp and can be a factor in the model selection process.

Hastie et al. (2009) stated that essentially, it is important to note that no form of classification is universally superior than the other. The change would be focused on the available data and target function of each model.

Limitation

In the current study, only basic demographics, type of offence and time served were used. Needless to say, there were other factors that matter and were not included. If other significant variables were available, it is possible that an even more higher accuracy could be obtained. Lastly, there were many missing data that were not hold into account. A more comprehensive dataset could result in better predictions. This study also raises a lot of ethical questions. How successful can predictive efforts be to justify the use of models to take convincing steps that require others' freedoms. In the case where the algorithm is used for actual decision making, there is a very ethical problem that needs to be thought about.

Conclusions

This study attempt to find the best predicting model for criminal recidivism. Variables used were the one that are simple and can be fit in all of the models. It was determined that the best predictive model is the neural network with a test error rate of 31.7% and 0.193 Brier test score. The differences in terms of performance between the best and the follow up models are generally very small. While the neural network in this study performed well compared to the other models, there is still a need to further research to identify with more predictive variables and models.

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Dr Haziq Jamil Assistant Professor in Statistics. Mathematical Sciences, Faculty of Science (FOS), University of Brunei Darussalam (UBD), for the continuous support during this year. His invaluable guidance, patience, motivation and immense knowledge help me in all time of doing this research and writing this report. I learn a lot for someone who are not really familiar with statistics and now I am able to use the R software. I could not have imagined having a better supervisor for my final year paper.

Last but not least, I would like to thank my mother for her love, prayers and sacrifices for educating and preparing me for the future. Her advices and support meant a lot to me when I encounter some struggles throughout the journey of doing this research.

References

Bushway, S. D. (2013). Is there any logic to using logit: Finding the right tool for the increasingly important job of risk prediction. Criminology and Public Policy, 12(3), 563–567.

Caulkins, J., Cohen, J., Gorr, W., & Wei, J. (1996). Predicting criminal recidivism: A comparison of neural network models with statistical methods. Journal of Criminal Justice, 24(3), 227-240.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. Philosophy & Technology, 29(3), 245-268.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Mayo, M. (n.d.) Simple neural network [Diagram]. Retrieved from https://www.kdnuggets.com/2017/10/neural-network-foundations-explained-gradient-descent.html

National Corrections Reporting Program (1991-2014). Selected Variables ICPSR 36404 United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. Retrieved from www.icpsr.umich.edu

Perkins, C. (1993). National Corrections Reporting Program, 1990. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Rufibach, K. (2010). Use of Brier score to assess binary predictions. Journal of clinical epidemiology, 63(8), 938-939.

Sharma, M. (2020). Random forest algorithm [graph]. Retrieved from https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Yang, M., Liu, Y., & Coid, J. (2010). Applying Neural Networks and other statistical models to the classification of serious offenders and the prediction of recidivism. Ministry of Justice Research Series, 6(10).