A Comparison of Logistic Regression(LR) and Random Forest (RF) Machine Learning algorithms for Prediction of Recidivism

INM431 Machine Learning Coursework | RAJANI MOHAN JANIPALLI | City University of London

Brief description and motivation of the problem

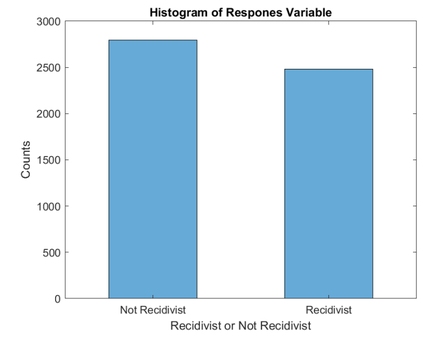
- An attempt was made to predict the chance of Recidivism from the COMPAS data set, which is basically a task of binary classification of whether or not a convict will commit a crime again.
- To do this, two models were built using Logistic Regression and Random Forest machine learning algorithms to perform the said classification task.
- Analyse and compare two models in terms of predicting recidivism by using various model performance metrics.
- Methods similar to those used by Scott Cole & Thomas Donoghue (2017) [1] for binary classification on a different data, using the above two machine learning algorithms, to be considered for model building and analysis.

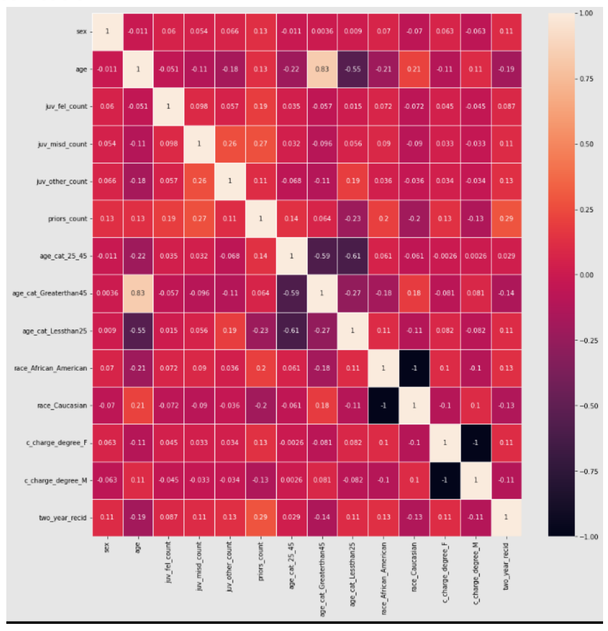
Initial analysis of the data set including basic statistics

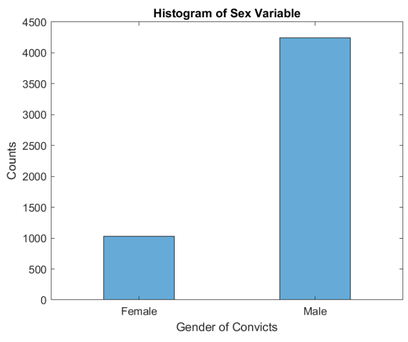
- The data set, which has been obtained from OpenML website [2], is a pre-processed version of original COMPAS dataset and has been reduced to contain features that have relevance to the classification.
- The original data collected by PROPUBLICA [3][4] was for the years 2013 and 2014, which contained COMPAS scores of convicts in Broward County, Florida in the USA.
- The software makes predictions based on the answers given by the prisoners to a questionnaire, after they have been arrested and put into jail.
- Data set used for this task consists of various details of prisoners like age, race, counts of crimes committed and different types of degrees under which they have been charged.
- The data set consists of 5278 observations, 13 features and one target column or response variable.
- Features such as age and count of various types of crimes committed are continuous variables of numerical data and others are categorical variables in binary form.
- The target column or response variable is a categorical column of 0s and 1s, indicating whether or not an individual was a recidivist after two years.
- There were no missing values in the dataset used for the task.
- With 53% observations being not recidivist and 47% being recidivist, the data is almost balanced with respect to response variable.
- Looking at the descriptive statistics of the data set, it appeared that the maximum values of continuous variables were very much practically possible and not outliers.
- Since all the continuous variables had numerical values less than 100 and the remaining variables were categorical data, normalization or standardization of the data was not done.
- Some categorical features were completely related, but that may be due to the fact that they are categorical columns consisting of 0s and 1s.

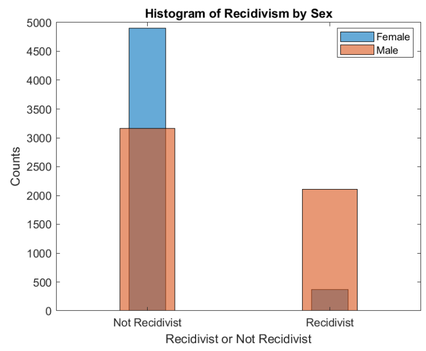
Descriptive Statistics of Data Set

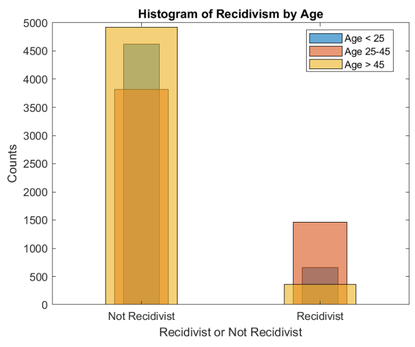
	sex	age	juv_fel_ count	juv_misd _count	juv_other _count	priors_ count	age_cat_ 25_45	age_cat_ Greaterthan45	age_cat_ Lessthan25	race_African _American	race_ Caucasian	c_charge_ degree_F	c_charge_ degree_M	two_year _recid
count	5278	5278	5278	5278	5278	5278	5278	5278	5278	5278	5278	5278	5278	5278
mean	0.8047	34.4494	0.0614	0.0985	0.1192	3.4615	0.5733	0.2077	0.219	0.6016	0.3984	0.6518	0.3482	0.4704
std	0.3965	11.7326	0.4102	0.5171	0.482	4.8759	0.4946	0.4057	0.4136	0.4896	0.4896	0.4765	0.4765	0.4992
min	0	18	0	0	0	0	0	0	0	0	0	0	0	0
25%	1	25	0	0	0	0	0	0	0	0	0	0	0	0
50%	1	31	0	0	0	2	1	0	0	1	0	1	0	0
75%	1	42	0	0	0	5	1	0	0	1	1	1	1	1
max	1	80	10	13	7	38	1	1	1	1	1	1	1	1

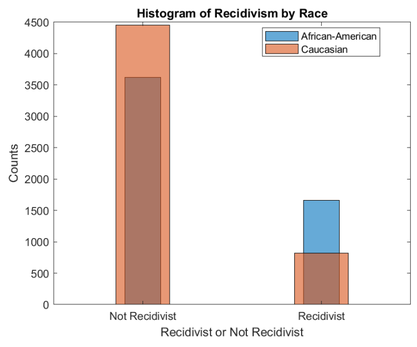


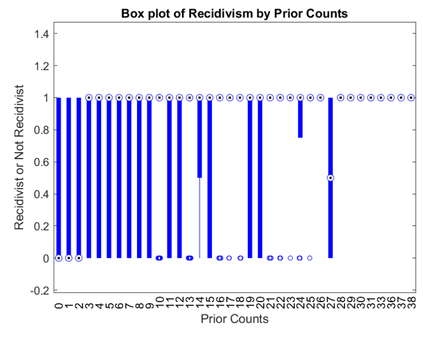












Logistic Regression

- One of the supervised machine learning algorithms, where the probability of a distinct outcome for a given input is modelled[5].
- Often such models produce a binary outcome which can have values such as true or false, yes or no, 0 or1, etc.
- Linear combination of features are applied on to a nonlinear Sigmoid function to produce a binary output. [6]
- The predicted values of outcome lie between 0 and 1, separated by decision boundary which can be determined by using a threshold. [7]
- Can also classify multiple classes as multinomial logistic regression.
- Gradient descent method and its variants are known for tuning the parameters of the model. [7]

Pros

+ Lower chances of Overfitting. [8]
+ Simple to apply, understand and train.
+ Useful for both classification and probability prediction.[9]
+ Needs lesser time to train.
+ Gives good accuracy with simple and basic data sets.
+ No need to tune hyperparameters. [9]

Cons

- Susceptible to overfitting in case of complete separation. [10]
- Possible underfitting for complex datasets.
- May lead to overfitting for small datasets.
- Sensitive to outliers.

Hypothesis statement

- Both Logistic Regression and Random Forest would perform well in the binary classification task.
- Random Forest with its optimization advantages would perform better than the Logistic Regression.
- For a related data, an accuracy of 66% was achieved by ProPublica [3], which will be tried to be achieved in this task.

Choice of parameters and experimental results

Logistic Regression

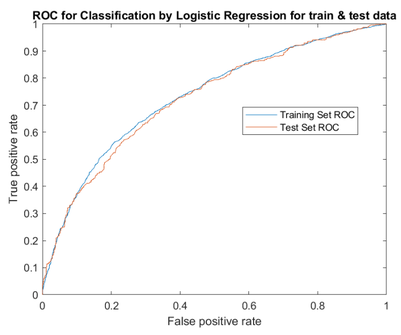
- Since the features were a mixture of continuous numeric variables and binary values in 0 or 1, the distribution was considered as binomial and correspondingly logit was assigned as the link function.
- For determining the decision boundary, a threshold of 0.52 was chosen.
- No features were excluded for training the model, based on the diagnostic plots of the baseline model.

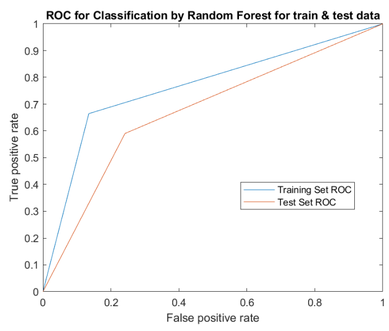
Random Forest

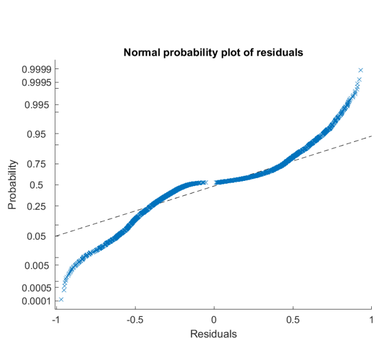
- The baseline model was trained without assignment of any specific values to parameters, so as to check which values are considered best by MATLAB for different parameters.
- In the first iteration to improve the model, Bag was assigned as the method.
- In the second iteration, different values of number of learning cycles were assigned.
- In the third iteration, Hyperparameter optimization was set to auto.
- In the forth iteration, best observed values of the parameters during the previous iteration were assigned to the model.
- In the fifth iteration, 'curvature' was assigned as the predictor selection method instead of the default 'CART'.

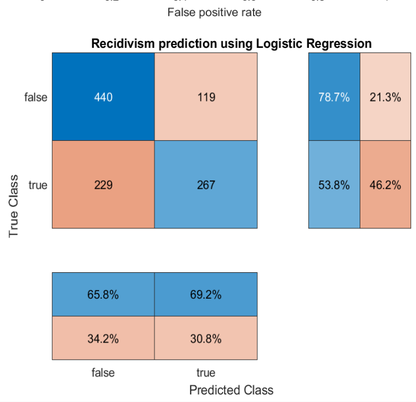
Performance Metrics

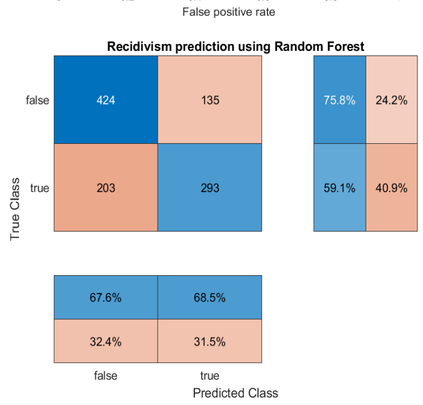
	Logistic Regression	Random Forest
Avg. Train Time	0.081	1.081
Avg. Test Time	0.0087	0.067
Avg. Train Error	0.3171	0.3327
Avg. Test Error	0.3299	0.3735
Train AUC	0.7316	0.7649
Test AUC	0.7242	0.6746
True Negative	440	424
True Positive	267	293
False Negative	229	203
False Positive	119	135
Sensitivity	0.5383	0.5907
Specificity	0.7871	0.7585
Precision	0.6917	0.6846











Lessons Learned

- Bagging in RF may reduce variance but may tend to overfit on training data in some cases.
- Hyperparameters optimization may not always give the best accuracy to a model under training.

Future Work

- Prepare models based on other machine learning classification algorithms with the same data for comparison.
- Work on original COMPAS data to compare how LR and RF would perform with data which is not binary.
- Try methods to reduce overfitting of RF.
- Instead of Bagging, try Boosting to create a RF model and compare its performance with LR.

References

1. Cole. S. and Donoghuge. T, "Predicting departure delays of US domestic flights ", 2017.
2. https://www.openml.org/d/42193
3. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
4. https://github.com/propublica/compas-analysis/
5. Thomas W. Edgar and David O. Manz "Research Methods for Cyber Security", 2017, chapter 4, page 125

Random Forest

- Another supervised machine learning algorithm in which group of individual Decision Trees are utilized for prediction of the outcome [10].
- Decision Trees are chosen in random and the output is obtained by taking a mean of their predictions or a class that has been majorly voted by them, in a process called Bagging. [11]
- The idea is that aggregating the predictions of a collection of weak classifiers will produce a better performance due to the reduction of generalization error. [10]
- Has a capability to create levels for predictors, based on the importance measure of variables. [12]

Pros

+ Less prone to Overfitting and class imbalance. [10]
+ Ensemble techniques reduce potential error.
+ Overall variance is reduced.[11]
+ Good generalization is shown.
+ No need for scaling features.

Cons

- Difficult to interpret. [11]
- Computations can become complex for larger data sets[13].
- No correlation should exist among predictions of trees[9].

Description of the choice of training and evaluation methodology

- Split data set into training and test sets in a proportion of 80% of the data for training set and 20% of the data for testing set.
- Keep the test set unseen to the models till it is ready for testing.
- Use the hold out method to partition the data so that the training set, after the modelling, can be cross validated in whichever algorithm it is possible to cross validate.
- Try to improve the models using parameters optimization, wherever possible.
- Identify the most accurate and stable model for each algorithm, based on various performance metrics, after all the iterations to improve the model.
- Predict the outputs for test set using the most accurate and stable model for each algorithm.
- Evaluate the test performance of both algorithms using various performance metrics.
- Observe the time taken for training model and predicting the output with test set, for both algorithms.

Analysis and critical evaluation of results

- Starting with the baseline models, the error for RF at 0.2643, was lesser than LR which was 0.3171. This would be very much expected due to the inherent advantages of RF. By default, the *fitcensemble* function of MATLAB choses a boosting method, in which the individual trees are trained in such a manner that the misclassification from one learning cycle are weighted heavily in the next learning cycle in order to increase the accuracy of the model [15]. But this may lead to reduction in precision of the outcome.
- In order to minimize the variance, the RF model was then trained with bagging method, which reduced the error further to 0.2292. This may be because in bagging the decision trees are trained by using subsets of training data that have been randomly chosen with replacements and then the average of predictions by all the decision trees trained in the ensemble is used to produce the final output, which will be better than an individual decision tree [15].
- Thereafter the RF model was tried to improve further with the changes in various parameters and also using optimization of hyperparameters. But in spite of all these trials of improvising the RF model, the second RF model turned out to be the best model for the training data, considering the consistency of its error.
- For logistic regression, the diagnostics plot of the baseline model indicated that all the features were useful for training the model and hence no feature was removed in order to optimize the model. Also from the plot of normal probability of residuals shown on the left, it can be seen that the data is almost forming the s-shaped logistic function curve. So, no regularization technique was applied as well.
- Since logistic regression is a generalised linear model [16], it's error for the baseline training model was considered as the average training error. Whereas for RF a cross validation of the trained model would give a generalization error [17] and at 0.3327 this error was higher than the average error of LR.
- The performance of both Logistic Regression and Random Forest were surprising as they are very different for train data and test data. This may be because of the size of the data set. LR performs better as the size of the data set gets bigger [8].
- From the performance metrics displayed on the left, it can be seen that the average test error for logistic regression is 0.3299. So, it can be said that the accuracy of the model for test data was as close as what ProPublica[3] achieved using logistic regression with a related data set.
- For RF the average test error value of 0.3735 which is in line with the expectation from cross validated error of the model and is more than that of LR. This can also be correlated by comparison of ROC plots for RF training and test data sets.
- The ROC of RF train data shown here is not that of the cross validated model but of the best RF model without cross validation. There is a considerable gap between the ROC plot of training data and test data for the RF. Quantitatively these can be indicated by the Area Under Curve (AUC) values of training and test data which are 0.7649 and point 0.6746 respectively. This basically means that the model was an overfit which might have been caused due to the model picking up random noise and fluctuations from the training data [18].
- On the contrary, the ROC plot and the AUC values of training and test data for LR shown on the left, are very much consistent and indicate a much better fit to the data. This means that the predictions using LR would be not just good but also consistent. This is similar to results obtained by Scott Cole & Thomas Donoghue (2017) [1] .
- From the performance metrics shown on the left, LR has better specificity and precision, whereas RF has better sensitivity [19]. So LR would predict those who are not recidivists better and RF will predict those who are recidivists better.
- The training and test time for both the models were as expected, that is LR took lesser time.

References (continued)

6. Stephanie Kay Ashenden "The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry", chapter 7, page 123-124
7. V.N.Gudivada,M.T.Irfan,E.Fathi,D.L.Rao "Handbook of Statistics" Volume 35, 2016, Pages 169-205
8. https://www.analyticsvidhya.com/blog/2021/10/building-an-end-to-end-logistic-regression-model/
9. https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6
10. Stephanie Kay Ashenden "The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry", chapter 7, page 123-124
11. Siddharth Misra, Hao Li, "Machine Learning for Subsurface Characterization, 2020", chapter 9, page 265.
12. Weiyan Zong,Junyi Zhang,Ying Jiang, "Transport and Energy Research A Behavioral Perspective , 2019", chapter 15, page 379]
13. https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04
14. https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb
15. https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/
16. Christopher M Bishop, "Patter Recognition and Machine Learning", chapter 4, page 205
17. https://www.crosstab.io/articles/bates-cross-validation
18. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/
19. https://classeval.wordpress.com/introduction/basic-evaluation-measures/