# Random Forest vs Logistic Regression for Binary Classification

Kaitlin Kirasich, Trace Smith, and Bivin Sadler, PhD[1]

[1]Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
kkirasich@.smu.edu,traces@smu.edu,bsadler@smu.edu

**Abstract.** Selecting a learning algorithm to implement for a particular application on the basis of performance still remains an ad-hoc process using fundamental benchmarks such as evaluating a classifier's overall loss function, area under the curve (AUC) score, specificity, and sensitivity values. This work is aimed at addressing the difficulty of model selection by evaluating the overall classification performance between random forest and logistic regression for datasets comprised of various underlying structures. A model evaluation tool was developed in R for simulating a variety of datasets types in order to evaluate performance metrics such as true positive rate, false positive rate, and accuracy under specific conditions. Our findings indicate that when increasing the variance in the explanatory and noise variables, logistic regression consistently performed with a higher overall accuracy as compared to random forest. However, the true positive rate for random forest was higher than logistic regression and yielded a higher false positive rate. In all cases a paired two sample t-test indicates there is enough evidence to suggest the false positive rate for random forest is statistically different than logistic regression. The model evaluation application developed in this work is a proxy for answering other intruiguing questions related to model performance under various treaments.

## 1 Introduction

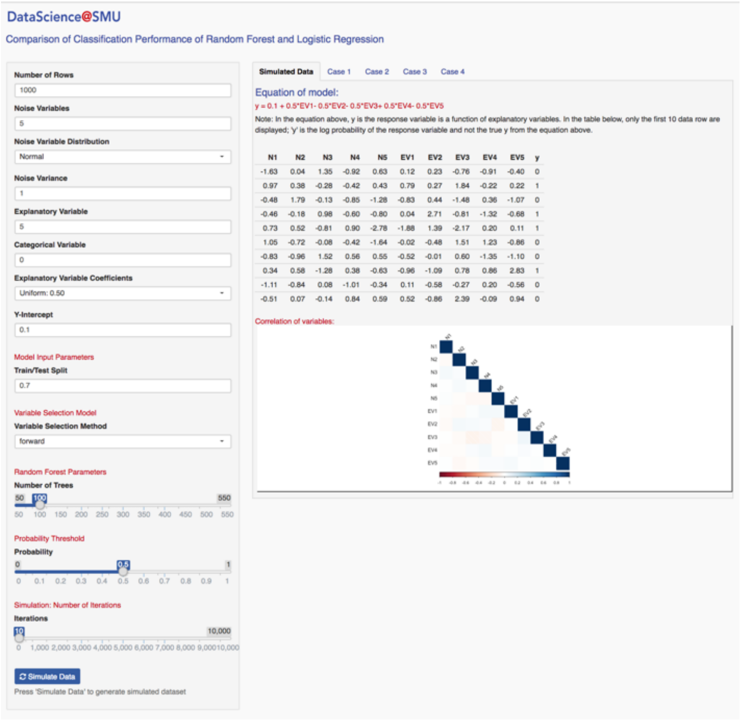
## 2 Motivation

## 3 Data Set

To conduct the statistical analysis, an interactive web application was developed using RShiny which allows end users to rapidly generate simulated datasets and evaluate performance metrics between machine learning models, random forest and logistic regression. For performing numerical simulations, creating synthetic datasets is pivotal for the analysis. Simstudy, a R package, was leveraged in this work as the method for producing datasets of various structures. Given this work is aimed at model performance for binary classification, the response variable 'y' is a function of only the explanatory variables 'x' included in the model euqation shown below. Binary response variable takes on the values of either 1 or 0; thus the formula represents the log of odd or probability of the response being a 1 or 0. As previously stated, the explanatory variable beta is related to the binary response, while the noise variables 'N' are not. The parameter estimates explain the relationship between independent variables 'X' and the dependent variable 'Y', and the 'Y' scale is known as the logit, or log of odds. For each simulation case study explored in the work, the default parameter estimate (e.g. beta) values are uniform at 0.50 and the input features, both noise and explanatory variables, are all continuous and normally distributed.

$$\log(y) = N_0 + \beta_1 X_1 + .....\beta_n X_n + N_n \tag{1}$$

The user interface for the RShinny application, shown in the figure below, allows users the ability to create multivariate datesets with several input configuraiton options such as specifying the distribution of input features as either gaussian or poisson. Moreoever, users can also modify the magnitudes of the parameter estimates to be non-uniform, allowing for a subset of the explanatory variables to be a more significant predictor of the response variable.



DataScience@SMU

Comparison of Classification Performance of Random Forest and Logistic Regression

Number of Rows: 1000

Noise Variables: 5

Noise Variable Distribution: Normal

Noise Variance: 1

Explanatory Variable: 5

Categorical Variable: 0

Explanatory Variable Coefficients: Uniform: 0.50

Y-Intercept: 0.1

Model Input Parameters

Train/Test Split: 0.7

Variable Selection Model

Variable Selection Method: forward

Random Forest Parameters

Number of Trees: 100

Probability Threshold

Probability: 0.5

Simulation: Number of Iterations

Iterations: 10

Press 'Simulate Data' to generate simulated dataset

Equation of model:

y = 0.1 + 0.5*EV1- 0.5*EV2- 0.5*EV3+ 0.5*EV4- 0.5*EV5

Note: In the equation above, y is the response variable is a function of explanatory variables. In the table below, only the first 10 data row are displayed; 'y' is the log probability of the response variable and not the true y from the equation above.

| N1 | N2 | N3 | N4 | N5 | EV1 | EV2 | EV3 | EV4 | EV5 | y |
|---|---|---|---|---|---|---|---|---|---|---|
| -1.63 | 0.04 | 1.35 | -0.92 | 0.63 | 0.12 | 0.23 | -0.76 | -0.91 | -0.40 | 0 |
| 0.97 | 0.38 | -0.28 | -0.42 | 0.43 | 0.79 | 0.27 | 1.84 | -0.22 | 0.22 | 1 |
| -0.48 | 1.79 | -0.13 | -0.85 | -1.28 | -0.83 | 0.44 | -1.48 | 0.36 | -1.07 | 0 |
| -0.46 | -0.18 | 0.98 | -0.60 | -0.80 | 0.04 | 2.71 | -0.81 | -1.32 | -0.68 | 1 |
| 0.73 | 0.52 | -0.81 | 0.90 | -2.78 | -1.88 | 1.39 | -2.17 | 0.20 | 0.11 | 1 |
| 1.05 | -0.72 | -0.08 | -0.42 | -1.64 | -0.02 | -0.48 | 1.51 | 1.23 | -0.86 | 0 |
| -0.83 | -0.96 | 1.52 | 0.56 | 0.55 | -0.52 | -0.01 | 0.60 | -1.35 | -1.10 | 0 |
| 0.34 | 0.58 | -1.28 | 0.38 | -0.63 | -0.96 | -1.09 | 0.78 | 0.86 | 2.83 | 1 |
| -1.11 | -0.84 | 0.08 | -1.01 | -0.34 | 0.11 | -0.58 | -0.27 | 0.20 | -0.56 | 0 |
| -0.51 | 0.07 | -0.14 | 0.84 | 0.59 | 0.52 | -0.86 | 2.39 | -0.09 | 0.94 | 0 |

Correlation of variables:

**Fig. 1.** R Shinny Application - Data Simulator

## 4 Methods and Experiments

The two machine learning algorithms studied in this work consist of random forest and logistic regression. Both models have been widely implemented in various disciplines for classification and regression purposes. Not only are these algorithms known for their success, but also their simplicity to implement and relatively straightforward to interpet. An overview of the two models are provided in the following sections.

### 4.1 Random Forest

Random forest is an ensemble based learner which is comprised of 'n' collection of de-correlated decision trees (Hastie, 2009). Built off the idea of bootstrap aggregation which with a method for resampling with replacement in order to reduce variance, random forest uses multiple trees to average (regression) or compute majority votes (classification) in the terminal leaf nodes when making a prediction. Presented by Leo Breiman and built off the idea of decision trees, random forest models resulted significant improvements in prediction accuracy as compared to a single tree by growing 'n' number of trees where each tree in the training set is sampled randomly without replacement (Breiman ,1966). Decision trees consist simply of a tree-like structure where the top node is considered as the root of the tree and is recursively split at a series of decision nodes from the root until the terminal node or decision node is reached.

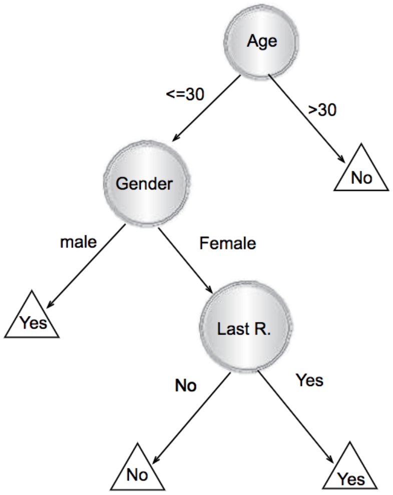


**Fig. 2.** Example of Decision Tree (Lior Rokach et el)

As illustrated in the tree structure, the decision tree algorithm is a top down greedy approach by partitioning the dataset into smaller subsets; the result is a tree with a series of decision nodes and leaf node. The decision node has has two or more branches where the features with the highest information gain is split. The predictor variable that yields the highest informatio gain is the root node. The leaf node is represented as a prediction, and in this case, classifying either 1 or 0. Decision trees can handle both categorical and numerical data. First, in order to determine the information gain which is based on the entropy after splitting on an attribute, entropy is computed. Entropy measures the homogeneity of the subset data; if entropy equals one then the class labels are equally divided while an entropy of zero means the sample is completely homogeneous.

$$Entropy = -p \log_2(p) - q \log_2(q) \tag{2}$$

Advantages of using tree like learning algorithms allow for training models on large datasets in addition to quantitative and qualitative input variables. Additionally, tree based models can be immune to redundant variables or variables with high correlation which may lead to overfitting in other learning algorithms. Trees have also very few parameters to tune for when training the model and performs relatively well with outliers or missing values in a dataset. However, trees are prone to poor prediction performance; decision trees themselves are prone to overfitting noise in a training set which ultimately leads to results with high variance. In other words, this means the model could accurately predict the same data it was trained on but may not possess the same performance on datasets without the similar patterns and variations in the training set. Even fully grown decision trees are notorious for overfitting and do not generalize well to unseen data; random forest solves the overfitting conundrum by using a combination or "ensemble" of decision trees where the values in the tree are a random, independent, sample. The idea of randomly sampling the without replacement is known as bagging and this results in a different tree being generated to train on; averaging the results from the 'n' number of trees will result in decreasing the variance and establishing a smoother decision boundary. For instance, while using random forest for classification, each tree will give an estimate of the probability of the class label, the probabilities will be averaged over the 'n' trees and the highest yields the predicted class label. In addition to bagging or bootstrap aggregation, in order to further reduces the variance in the decision boundary further, the trees must be completely uncorrelated and the method of bootstrapping alone is not enough. Breiman introduced the idea of randomly sampling 'm' number of features at each decision split in the tree as a way to decorrelate the trees in the random forest algorithm.

### 4.2 Logistic Regression

Linear models are composed of one or multiple independent variables that describes a relationship to a dependent response variable. Mapping qualitative or quantitative input features to a target variable that is attempted to being

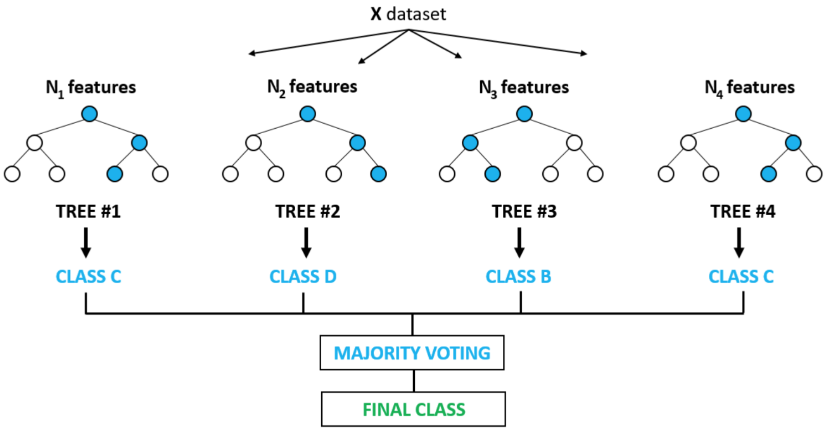


**Fig. 3.** Example of Decision Tree (Balazs Holczer)

predicted such as financial, biological, or sociological data is known as supervised learning in machine learning terminology if the labels are known. One of the most common utilized linear statistical models for discriminant analysis is logistic Regression.

$$\pi_i = \beta_0 + \beta_1 X_1 + .....\beta_n X_n \tag{3}$$

Simplicity and interoperability of logistic Regression can occasionally lead to outperforming other sophisticated nonlinear models such as ensemble learners or support vector machines. However, in the event the response variable is drawn from a small sample size, then linear regression models become insufficient and performs poorly for binary responses A number of learning algorithms could be applied to modeling binary classification data types, however the focal point of this work is to examine one linear model, logistic regression.

Unlike the response variable for Linear Regression which is quantitative, the target variable for logistic regression is the posterior probability of being classified in the ith group of a binary or multi-class response (Hastie, 2009). Logistic regression makes several assumptions such as independence, responses (logits) at every level of a subpopulation of the explanatory variable are normally distributed, and constant variance between the responses and all values of the explanatory variable. Intuitively, a transformation to the response variable is applied to yield a continuous probability distribution over the output classes bounded between 0 and 1; this transformation is called to “logistic” or “sigmoid” function where ‘z’ corresponds to log odds divided by the logit (Ng, 2008). The parameter estimates inform whether there is an increase or decrease in the predicted log odds of the response variable that would be predicted by one unit increase or decrease in one of the explanatory variables (e.g. x1), while holding all other explanatory variables constant.

$$\sigma(Z) = \frac{1}{1 + \exp^{-z}} \tag{4}$$

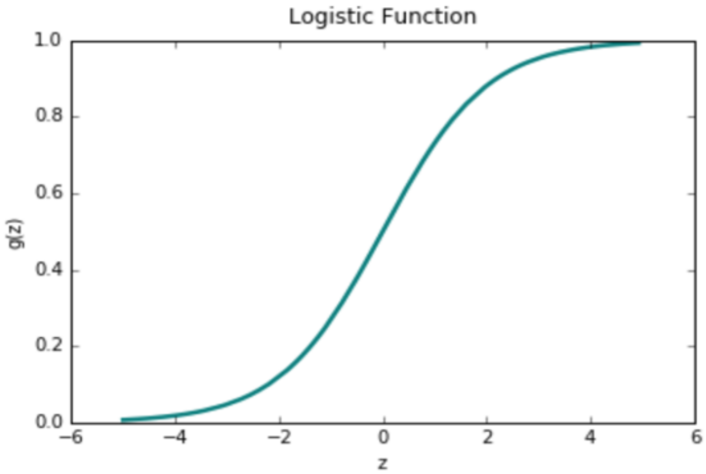


**Fig. 4.** Logistic Function

For a binary response, the logistic regression model can be expressed by summing over the linear combinations of input features and a corresponding weight plus a bias terms for each instance as shown below in equation (3) and (4).

$$p(y^{(i)} = 1|x^{(i)}, w) = 1 - \frac{1}{1 + \exp^{(w^T x^{(i)} + b)}} \tag{5}$$

$$p(y^{(i)} = 0|x^{(i)}, w) = 1 - \frac{1}{1 + \exp^{(w^T x^{(i)} + b)}} \tag{6}$$

The objective is to find a set of weights such that the negative log likelihood is minimized over the defined training set using optimization techniques such as gradient descent or stochastic gradient descent [3]. Minimizing the negative log likelihood also means maximizing the likelihood or probability the parameter estimate pi of selecting the correct class. The loss function that measures the difference between the ground truth label and the predicted class label is referred to as the cross-entropy. If the prediction is very close to the ground truth label, the loss value will be low. Alternatively, if the prediction is far from the true label, the resulting log loss will be higher.

$$J(\theta) = -\frac{1}{m} \sum p_i log(y_i) + (1 - p_i) log(1 - y_i) \tag{7}$$

## 5 Results

## 6 Analysis

## 7 Ethics

The disclaimer for utilizing this tool when selecting which machine learning model to implement in a production setting should done so with caution. Currently, the tool does not have certain functionalities to mimic real-world datasets, such as outliers or missing values to name a few. Failure to incorporate these types of characteristics in the simulated dataset can lead to inaccurate conclusions as to which algorithm yielded a better performance and thus any conclusions made from the RShiny tool may not generalize to these types of datasets. Moreover, the tool only considers random forest and logistic regression. As a result, other algorithms such as support vector machines or neural networks could produce higher prediciton accuracies and could be a better model to implement with datsets where the decision boundary is non-linearly seperable. Finally, in the context of the synthetic data generated in the application, there is no legal violations or security concerns. It should be clearly stated that the users should only consider the tool for educational purposes only as this application is still in the development phase. Any decisions drawn from the tool are not endorsed by the authors of this paper.

## 8 Conclusions