Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets

Kaitlin Kirasich, Trace Smith, and Bivin Sadler, PhD

Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
{kkirasich,traces,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 and misclassification metrics. In this paper we address the difficulty of model selection by evaluating the overall classification performance between random forest and logistic regression for datasets comprised of various underlying structures: (1) increasing the variance in the explanatory and noise variables, (2) increasing the number of noise variables, (3) increasing the number of explanatory variables, (4) increasing the number of observations. We developed a model evaluation tool capable of simulating classifier models for these dataset characteristics and performance metrics such as true positive rate, false positive rate, and accuracy under specific conditions. We found 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 for dataset with increasing noise variables. Each case study consisted of 1000 simulations and the model performances consistently showed the false positive rate for random forest with 100 trees to be statistically different than logistic regression. In the four simulated cases studies presented in this work, the relative classification scores for logistic regression and random forest yielded varying results for datasets with distinct characteristics.

1 Introduction

Datasets are composed of various dimensions and underlying structures. The relative performance of machine learning algorithms on datasets with varying data characteristics is not well documented. Most published work compares overall performance between several models on a single dataset as opposed benchmarking overall model performance for datasets that are comprised of various dimensions, multicollinearity, input feature types (e.g. continuous and categorical), and distributions of numerical variables. The performance of machine learning algorithms also weighs heavily on the algorithm selected for implementation. For instance, if the target variable is not linearly separable in n-dimensional space,

either continuous or categorical, then a more complex model may be needed to achieve higher prediction scores. Complex models like decision trees or other non-parametric algorithms can have decision boundaries with high variability in predictions but low bias, often leading to overfitting if not properly tuned. Overfitting is the result of a model with a high classification score on a training set while generalizing poorly on out of sample datasets. On the other hand, parametric based models like logistic regression are less complex, resulting in a linear decision boundary, but can result in a higher bias. Moreover, this can translate to under fitting as the model fails to adequately learn patterns in the data for accurate predictions if not tuned properly. Balancing the bias vs variance trade-off is driven by the complexity of the algorithm, which is crucial for deploying successful models for practical applications.

Depending on the structure of the dataset, deciphering which algorithm to deploy in order to achieve the highest performance scores still remains an ad-hoc process. This prompts the questions, under what circumstances does one model begin to outperform another model? For instance, when increasing the number of noise and explanatory variables in a dataset, at what point does the relative model performance begin to deviate between models? To answer these questions, our work consisted of building an analytical tool that simulates various data complexities to directly observe the classification performance of two machine learning algorithms by averaging metrics for 1000 random generations of specified multivariate datasets. For the sake of interoperability and computation time for model training, we considered one parametric and non-parametric machine learning model for binary classification, logistic regression and random forest, respectively.

Logistic regression and random forest are two very common and widely studied machine learning models. Machine learning is the process of mathematical algorithms learning patterns or trends on previously recorded data observations and then makes a prediction or classification. In this work, we are examining only binary classification (e.g. Y = 1,0), which is a form of supervised learning in which an algorithm aims to classify which category an input belongs to. Supervised learning can be described as taking an input vector comprised of n-features and mapping it to an associated target value or class label. The term "supervised" originated from the concept that the training and testing datasets contain a response label and the algorithm observes the input vector and attempts to learn a probability distribution to predict 'y' given 'x' [7]. The algorithms learn a series of input weight parameters that determine how the input feature vector affects the prediction. The objective of the algorithm is to learn a set of weights on a subset of the data that minimizes the error or loss between the ground truth and predicted value in order to precisely classify the input to the associated label. Classification metrics like accuracy, true and false positive rates, and area under the curve are examined on the portion of data that was held out during training for evaluating how well the model classifies the input feature vector.

For binary classification model evaluation between random forest and logistic regression, our work focused on four distinct simulated datasets: (1) increasing

the variance in the explanatory and noise variables, (2) increasing the number of noise variables, (3) increasing the number of explanatory variables, (4) increasing the number of observations. To benchmark and compare classification scores between random forest and logistic regression, metrics such as accuracy, area under the curve, true positive rate, false positive rate, and precision were analyzed. To provide statistical quantification as to whether a difference in model performance is conclusive enough to state the difference is significant or if the observed difference is by random chance, a pairwise two-sample t-test is also conducted at the end of each simulation case study.

2 Background

A dataset is a collection of an arbitrary number of observations and descriptive features which can be numerical, categorical or a combination of the two. Characteristics of a dataset can be comprised of missing values, outlier, highly correlated variables, concave or convex shapes, or subsets of the data that can be represented as clusters. The data examined in this study is only for continuous variables that have a normal distribution, similarly to Figure 1. As shown in (Figure 1), the decision boundaries learned from both logistic regression (left) and random forest (right) are able to effectively segment the two classes for the two clusters.

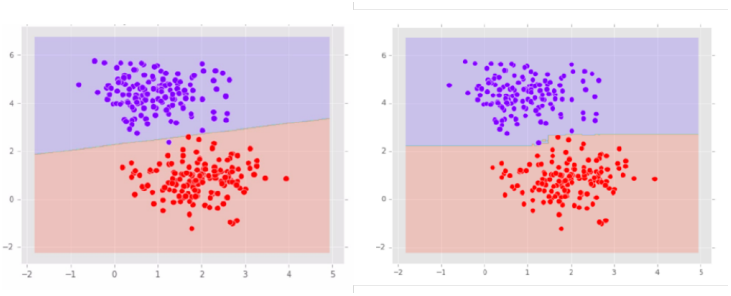


Fig. 1. Decision boundary between binary classes for Logistic Regression (left) and Random Forest (right) with compacted clusters

For the cases of more complex datasets, linear-based algorithms may not be sufficient in segmenting the class labels, leading to poor accuracies. More sophisticated algorithms may then be required like random forest, which can learn a non-linear decision boundary and thus can achieve higher accuracy scores. For instance, a toy dataset is shown below in Figure 2 consisting of concave and convex shapes. As illustrated in this figure, logistic regression (left) poorly segments the two classes while the more flexible decision boundary learned from the random forest model produces a higher classification accuracy. This example

raises a profound question as to which data characteristics constitutes one model achieving an overall better classification score. It should be noted this work only investigates random forest and logistic regression, however generalization of the current application can be adapted to other linear and nonlinear models. Therefore, performance of machine learning classifiers can yield varying results depending on the shape and structure of the data.

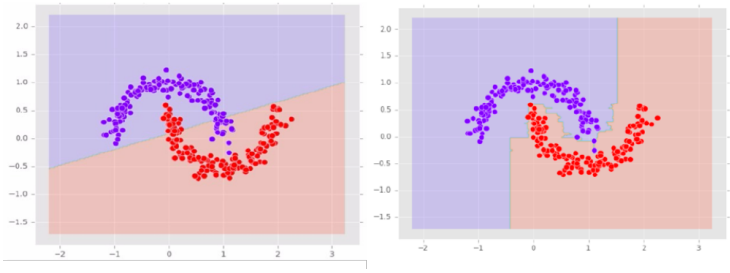


Fig. 2. Decision boundary between binary classes for Logistic Regression (left) and Random Forest (right) with complex data structures (e.g. concave and convex

One approach to inferring underlying complexities of high-dimensional dataset is Topological Data Analysis (TDA). TDA is an evolving method that utilizes topological and geometric tools to identify relevant features in the data. TDA can be described as method that helps identify structures in noisy and incomplete datasets like clusters or other hidden shapes that can provide a more accurate representation of the dataset [6]. Models can then be trained on the new representation of the data that has been reconstructed, which has shown promising results. While TDA looks at the proximity of data points and connectivity that can be mapped to a 1-dimensional plane for representing the shape of the data [16], our analysis is aimed at creating various complexities in the data and evaluating model performance on the raw structure in a multidimensional space. For instance, altering the variance in the explanatory and noise variables, changing the number of observations, and varying the number of continuous features included in a multivariate dataset was considered. We investigated why models like logistic regression or random forest perform differently for simple and complex data characteristics.

Numerous studies have been published that compare random forest and logistic regression algorithms however, most research experiments consisted of either a single dataset or multiple datasets from the same source. In these scenarios, sometimes logistic regression performed better while in other cases random forest performed better. For example, one experiment used several neuropsychological tests to predict dementia stated that with respect to specificity and overall classification accuracy, random forests and linear discriminant analysis rank first among all the classifiers including logistic regression [9]. In this case study, data

for 921 elderly non-demented patients with complaints that were referred for neuropsychological evaluation at three different institutions between from 1999 to 2007 was collected. At the follow-up of each evaluation, subjects were either classified as being diagnosed with dementia or not. Metrics such as sensitivity, specificity, ROC, and accuracy were evaluated using fivefold cross validation. Results indicated Random Forests and Linear Discriminant Analysis proved to have the highest accuracy, sensitivity, specificity compared to other models. Median overall accuracy and ROC for Random Forest is 0.73, respectively.

Contrastingly, another publication analyzed Twitter tweets surrounding the 2016 United States election. A data set of 850 observations and 299 features extracted from tweets obtained during the election period were utilized to classify voter sentiment and whether a tweet was either political or non-political. In addition to analyzing feature importance, random forest yielded the highest overall accuracy score of 95%. Several cases were considered, such as including all of the explanatory variables in the model and performing dimensionality reduction by applying Principal Component Analysis. In all cases, random forest consistently produced higher accuracy scores compared to logistic regression and support vector machines [2].

Random forest and logistic regression are commonly reported in published work on specific datasets and have shown varying performance results. The work presented by Ruiz-Gazen on classifying satellite measurements of cloud systems as either convective or non-convective systems [17]. With 41 numerical variables built from satellite measurements to train models on, the class labels were heavily skewed with only less than five percent of the labels being non-convective. The overall model results indicate virtually identical performance, however the authors recommended using logistic regression due to the interpretation of parameter estimates of the explanatory variables in addition to quicker computation time to train models [17]. The type of data and data sources used in the studies above are drastically different from each other and each algorithm performs differently due to the type of data it was utilizing to train each classifier. This analysis aims to provide a method of evaluating random forest and logistic regression models by simulating a variety of data characteristics and then evaluate which model yielded better performance under certain conditions.

Many researchers have set out to find one algorithm that performs better than another. In Data Science, there is the idea of a no-free-lunch theorem for supervised algorithms. The no free lunch theorem tells us that if one algorithm outperforms another in one metric, it will lose in another metric. Research has shown that a better performance over one class of problems is equivalently paid for in performance of another class of problems [20]. "In particular, if algorithm A outperforms algorithm B on some cost functions, then loosely speaking there must exist exactly as many other functions where B outperforms A."[20]. One of the motivations behind this work with the no free lunch theorem was to examine performance metrics like AUC, ROC, true positive rate, and false positive rates and determine under what conditions constitutes random forest or logistic regression performing better for different data characteristics.

3 Machine Learning Algorithms

The two machine learning algorithms studied in this work consist of random forest and logistic regression. Both models have been widely implemented successfully in various disciplines for classification and regression purposes [1]. The functionality of logistic regression, a parameter based model, and random forest, a non-parametric model, are summarized in the following section.

3.1 Random Forest

Random forest is an ensemble-based learning algorithm which is comprised of n collections of de-correlated decision trees [10]. It is built off the idea of bootstrap aggregation, which is 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. Built off the idea of decision trees, random forest models have resulted in significant improvements in prediction accuracy as compared to a single tree by growing 'n' number of trees; each tree in the training set is sampled randomly without replacement [4]. Decision trees consist simply of a tree-like structure where the top node is considered the root of the tree that is recursively split at a series of decision nodes from the root until the terminal node or decision node is reached.

As illustrated in the tree structure (Figure 3), the decision tree algorithm is a top down "greedy" approach that partitions the dataset into smaller subsets. Greedy algorithms are those that take the simplest solution rather than the most optimal solution, which is often more complex. The top of the decision tree is known as the root node and this corresponds to the best predictor variable. At each decision node, the features are split into two branches and this process is repeated until the leaf nodes are reached, which is used to make the final prediction.

To determine which feature to split on at each node, the 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 (Eq. 1). As in the case of binary classification with only two labels, if the split resulted in the class labels being all 1 or 0, then the entropy will be zero. Likewise, if half of the labels are 1 or 0, then a higher entropy of 1 is observed. The entropy is computed for each variable and then the difference between the entropy prior to the split (e.g. parent node) and after the split (e.g. child node) is performed on each variable. The variable that is most useful in segmenting the class labels will yield the highest difference in entropies. This difference is also referred to the information gained which is a method for quantifying how important an input attribute in the model is.

$$Entropy = -p\log_2(p) - q\log_2(q) \tag{1}$$

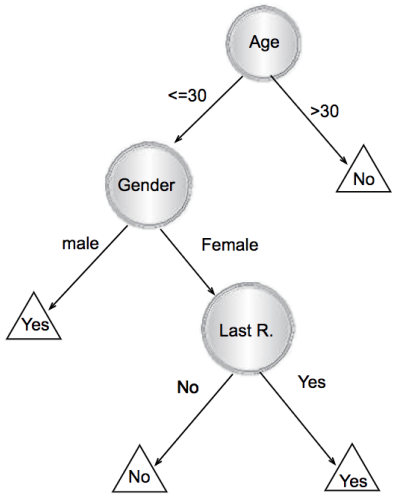


Fig. 3. Example of decision tree with "age" as root node; "gender" and "last r" are the two decision nodes [12]

Advantages of using a tree-like learning algorithm 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 also have 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.

Randomly sampling with 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 [10]. 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 (Figure 4). 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 de-correlate the trees in the random forest algorithm [4].

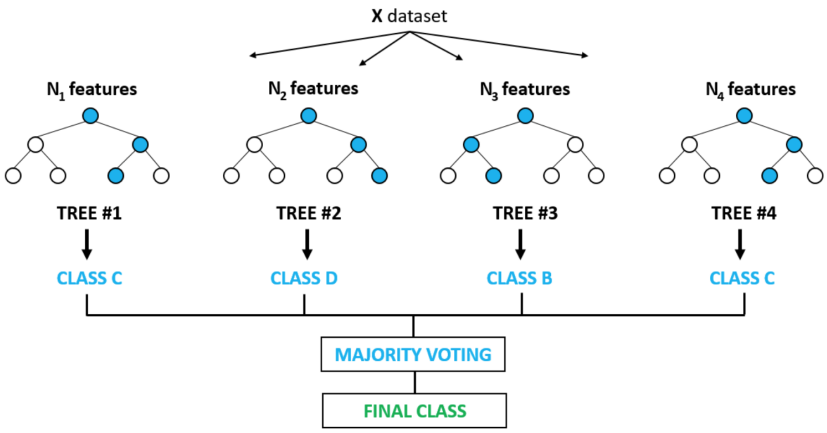


Fig. 4. Example of ensemble of decision trees (random forest)

3.2 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 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.

$$\pi_i = \beta_0 + \beta_1 X_1 + ...\beta_n X_n \tag{2}$$

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 logistic regression models become insufficient and performs poorly for binary responses[15]. 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.

In the case of logistic regression, the response variable is quantitative. For logistic regression, the response variable is the log of the odds of being classified in the ith group of a binary or multi-class response [10]. 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 [14]. 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.

$$\sigma(Z) = \frac{1}{1 + \exp^{-z}} \tag{3}$$

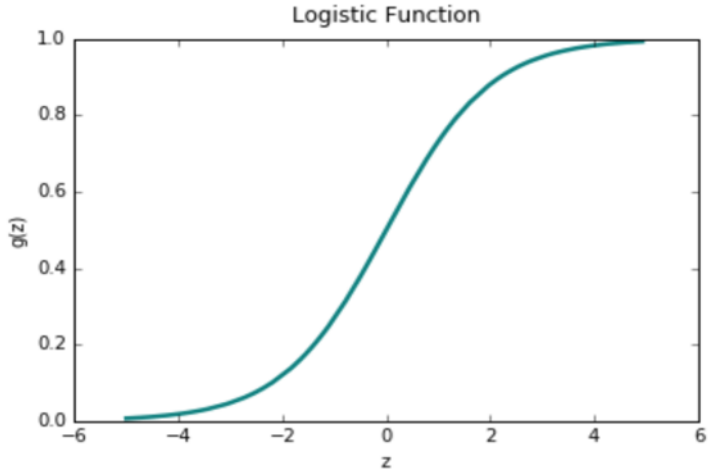


Fig. 5. Logistic Function

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

$$p(y^{(i)} = 1|x^{(i)}, w) = 1 - \frac{1}{1 + \exp^{(w^T x^{(i)} + b)}} \tag{4}$$

$$p(y^{(i)} = 0|x^{(i)}, w) = 1 - \frac{1}{1 + \exp^{(w^T x^{(i)} + b)}} \tag{5}$$

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 [14]. 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.

$$J(\theta) = -\frac{1}{m}\sum p_i log(y_i) + (1 - p_i) log(1 - y_i) \tag{6}$$

4 Analytical Tool

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 the machine learning models. This application is shown in Figure 6 and allows several input options to be configured prior to generating a multivariate dataset of an arbitrary length such as specifying number of observations in the dataset, variances in the features, amount of noise and explanatory variables, and the distribution of input features as either Gaussian or poisson. Moreover, the user has the ability to choose how the parameter estimates are weighted (e.g. uniform weights or unbalanced), allowing for a subset of the explanatory variables to be a more significant predictor of the response variable than others. The left column of the tool also has configurations for the machine learning models such as setting the number of trees for random forest or specifying the percentage of the training and testing set splits.

For performing numerical simulations, creating synthetic datasets is pivotal for the analysis. SimStudy is an open source package in the R programming language. The open source package was leveraged in this work as the method for producing datasets of various structures. To compare model performance for binary classification, we needed a response variable, which we will call 'y'; the response variable is a function of only the explanatory variables 'x' included in the model equation and displayed in the tool in Figure 7. Figure 7 is a screenshot of the first tab in the center content of the tool. It displays the equation as well as the first 10 rows of the dataset with all columns of 'EV', 'N', and the response 'y'. The explanatory variable 'EV' is related to the binary response, while the noise variables 'N' are not. The binary response variables take on the value of either a 1 or 0; thus, the formula represents the log of odd which is the probability of the response being a 1 or 0. The parameter estimates explain the relationship between independent variables 'X' and the dependent variable 'Y', and the 'Y' scale is known as the logit, log of odds. For each simulation case study explored in the work, the default parameter estimate, beta, is set to a uniform 0.50 and the input features, both noise and explanatory variables, are all continuous and normally distributed.

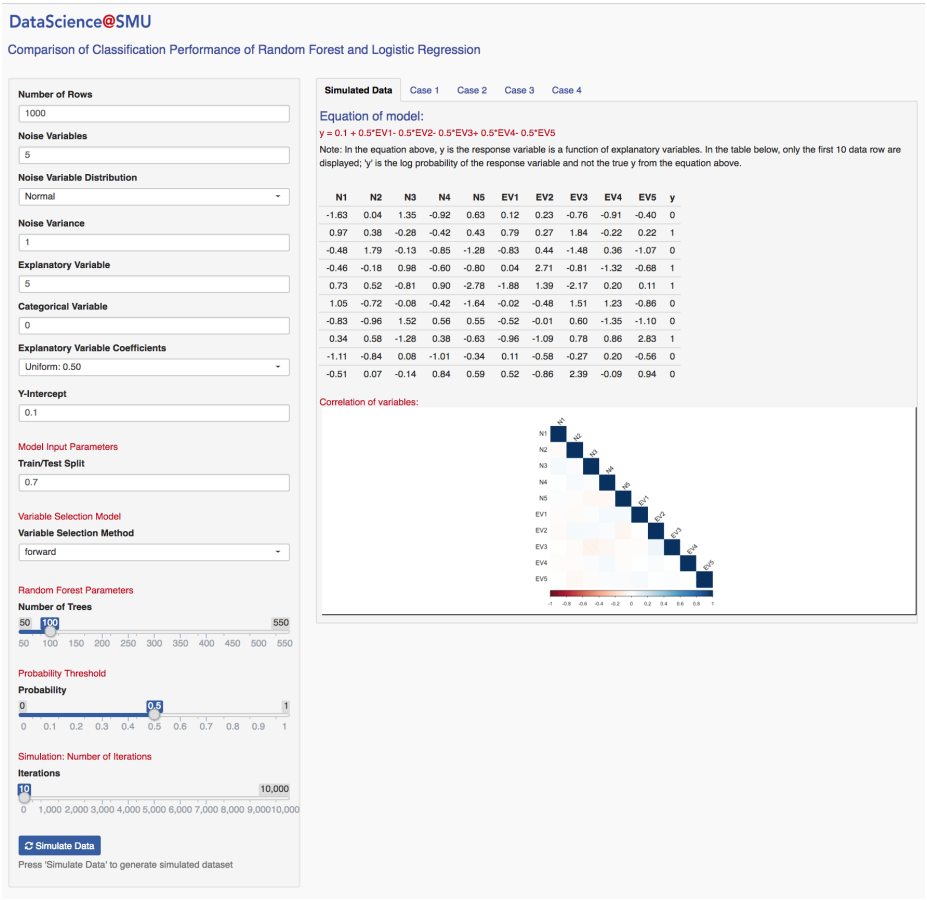
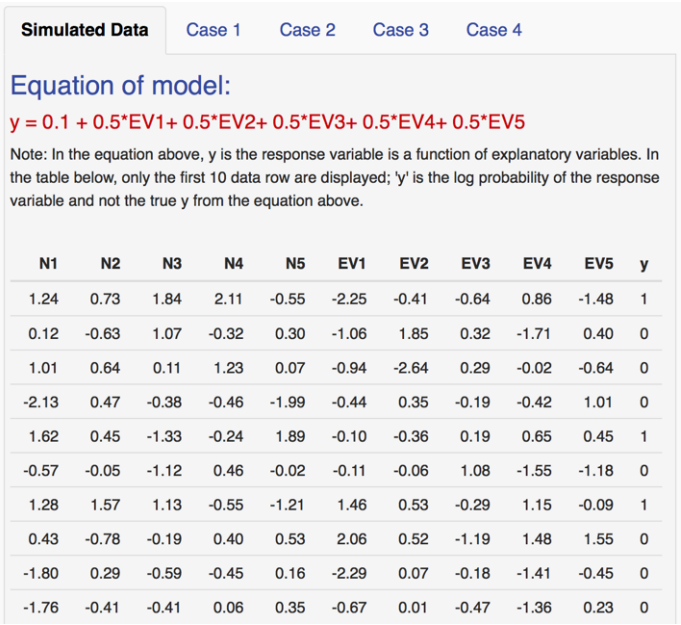


Fig. 6. Data Simulator



Simulated Data | Case 1 | Case 2 | Case 3 | Case 4

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
1.24	0.73	1.84	2.11	-0.55	-2.25	-0.41	-0.64	0.86	-1.48	1
0.12	-0.63	1.07	-0.32	0.30	-1.06	1.85	0.32	-1.71	0.40	0
1.01	0.64	0.11	1.23	0.07	-0.94	-2.64	0.29	-0.02	-0.64	0
-2.13	0.47	-0.38	-0.46	-1.99	-0.44	0.35	-0.19	-0.42	1.01	0
1.62	0.45	-1.33	-0.24	1.89	-0.10	-0.36	0.19	0.65	0.45	1
-0.57	-0.05	-1.12	0.46	-0.02	-0.11	-0.06	1.08	-1.55	-1.18	0
1.28	1.57	1.13	-0.55	-1.21	1.46	0.53	-0.29	1.15	-0.09	1
0.43	-0.78	-0.19	0.40	0.53	2.06	0.52	-1.19	1.48	1.55	0
-1.80	0.29	-0.59	-0.45	0.16	-2.29	0.07	-0.18	-1.41	-0.45	0
-1.76	-0.41	-0.41	0.06	0.35	-0.67	0.01	-0.47	-1.36	0.23	0

Fig. 7. Equation of the response variable and first 10 rows of the simulated dataset

The rest of the tabs in the tool are for the case studies that we conducted in this work. On each tab, there is a detailed description of each simulated case and the results of running random forest and logistic regression predictions on the simulated data. The mean of each evaluation metric is summarized in a table and line charts and a spread of the evaluation metrics are shown in a boxplot. Figure 8 is an example of the summary, table of average evaluation metrics for each model.

5 Criteria for Model Comparison

When comparing overall model performance, accuracy, true positive rate, false positive rate, precision, recall, and AUC were considered as the core metrics. Accuracy, true, and false positive rates are classic classification metrics while precision, recall, and AUC are functions of true and false positive rates.

For each simulation case, the dataset was randomly partitioned into 70% being utilized to train the model while the remainder 30% is left out of training the model and used to test the model. To determine how well a model predicts on the training data, we used a few different metrics. The first metric is accuracy, which is the percentage of correct classification. If the data point is actually a success, how often does the model predict success and if it was a failure, how often is a failure predicted. Accuracy is a nice overall average of how well a model can predict and simple to compute. However, if there is a class imbalance

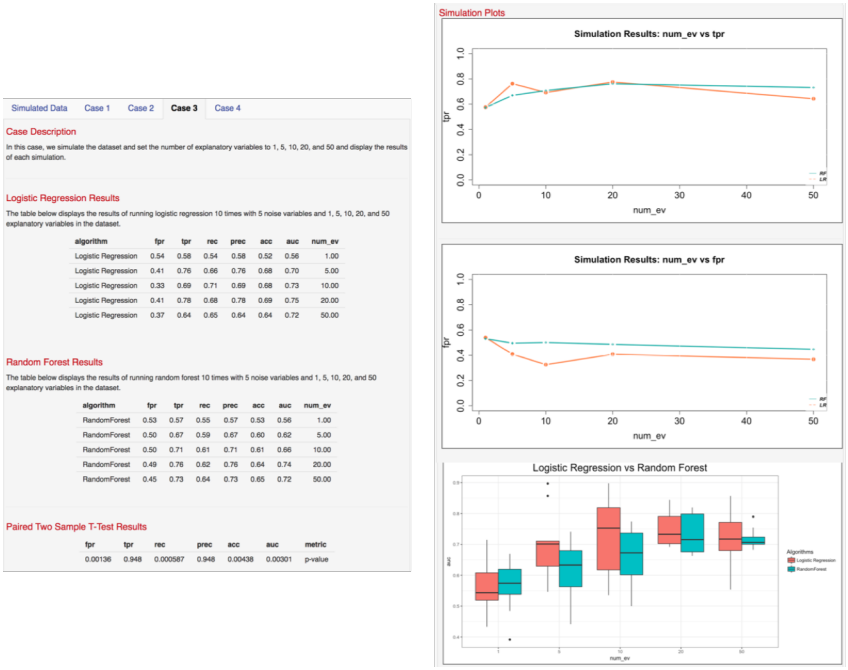


Fig. 8. Results of running each model of simulated data in each case study

Table 1. Evaluation metrics for comparison of model performance

Metric	**Formula**
Accuracy	(TP + TN) / (TP + TN + FP + FN)
True Positive Rate (TPR)	TP / (TP + FN)
False Positive Rate (FPR)	FP / (FP + TN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
Area Under the Curve (AUC)	Integral area of plotting TPR vs FPR

meaning 99% of my data is a success and only 1% of the time it is a failure, the model could predict success 100% of the time and have a very high accuracy of 99%. This causes an illusion that a model is performing very well but when implemented and used in the real world it may not be useful. In cases where there is a large class imbalance we may want to look at other evaluation metrics such as true positive rate and false positive rate.

True positive rate, also known as sensitivity, is calculated as the portion of positives or successes that are correctly identified. On the other hand, false positive rate is the portion that was incorrectly identified as positive or success but is actually negative [18]. Depending on the application and domain, one may care about incorrectly classifying a positive more than incorrectly classifying a negative. For example, when dealing with anything medical or health related, such as predicting if a patient will have dementia, it is extremely important to have a low false positive rate because telling someone they have dementia when they do not can cause a lot of emotional stress amongst other issues. False positive rate is also important when determining quality where the cost of a misclassification is high. For instance, to test a silicon wafer for defects, a machine goes through and returns a report with an outline of a wafer and places a dot on the area of the wafer where there could be a defect in the material or conductivity. If there are too many defects, a tester will often throw away out the entire wafer. However, one wafer could cost upwards of $10,000 so throwing away a wafer that could be perfectly fine because of a false defect can be very costly mistake. When dealing with an automated event or airport security, it may be okay to have a higher false negative rate because it is a relatively cheap and non-life-threatening task to confirm an automated alert as actually positive.

After running the simulated dataset through each model, the results can be graphically represented using the receiver operating characteristic curve or ROC curve as seen in Figure 9. The ROC curve is a graph with the x axis from 0 to 1 of the false positive rate, and the y axis from 0 to 1 of the true positive rate at various threshold settings. A perfect predictor would have a false positive rate of 0 and a true positive rate of 1. When graphed over a series of thresholds, the area under the curve (AUC) can provide a single value for providing insight into how well the model is classifying the labels. For interpretation, the higher the AUC, the better the model performs. The AUC is more descriptive than accuracy because it is a balance of accuracy and false positive rate.

Both recall and precision are often reported for classification performances. Recall is the ability to find all relevant instances while precision is the proportion of the data points the model considers relevant that are actually relevant. Precision is the number of true positives divided by the number of true positives and false positives. This provides an indication of the ability of a classification model to identify only relevant data. For example, if running a preliminary test to predict if a patient has a disease or not, precision would be equal to the number of patients who have the disease and were predicted correctly divided by the number of patients who have the disease and were predicted correctly plus the patients incorrectly predicted as having the disease. Recall is the number of true

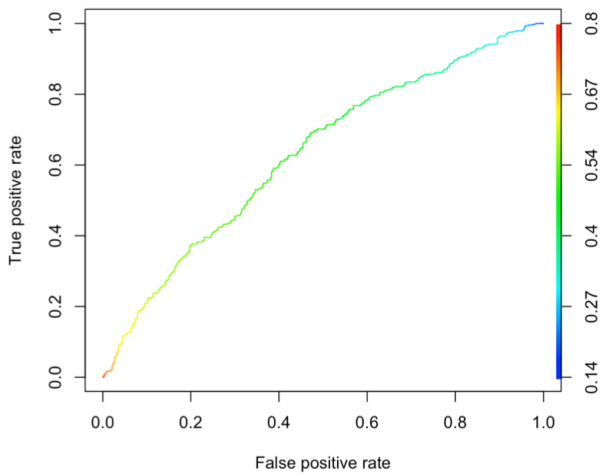


Fig. 9. Example ROC curve

positives divided by the number of true positives and false negatives. Recall tells us the ability of a model to find all relevant cases within a dataset. Using the example above, recall would be equal to the number of patients who have the disease and were predicted correctly divided by the number of patients who have the disease and were predicted correctly plus the patients incorrectly predicted as not having the disease. Thus, a high recall is desired to find all patients who actually have the disease and can allow a lower precision if the cost of a follow up check is low.

6 Analysis and Results

6.1 Case 1

The first case investigated was comparing model performance with respect to change in variance in the explanatory and noise variables. The hypothesis was that an increase in variance would strengthen the accuracy for both models. For this simulation case, the application was configured to run 1000 simulations for 1000 observations. In the top row of Figure 10, the results display the accuracy for varying levels of variance in 10 noise and 5 explanatory variables. There is both visual evidence from Figure 10 and statistical evidence from the paired t test (p-value less than .05) to suggest that, on average, the accuracy of the logistic regression model is greater than that of the random forest model.

The bottom row of Figure 10 displays the results of the true positive rate on the left and false positive rate on the right. The true positive rate for both models are nearly the same at each variance level. However, one can see that the false positive rate for random forest is not significantly higher than logistic regression (p-value = 0.63). Even though both models have the same performance

in terms of correctly classifying a true value as true, the false positive rate for random forest is higher than logistic regression. This causes logistic regression to outperform random forest in terms of overall accuracy at each level of variance.

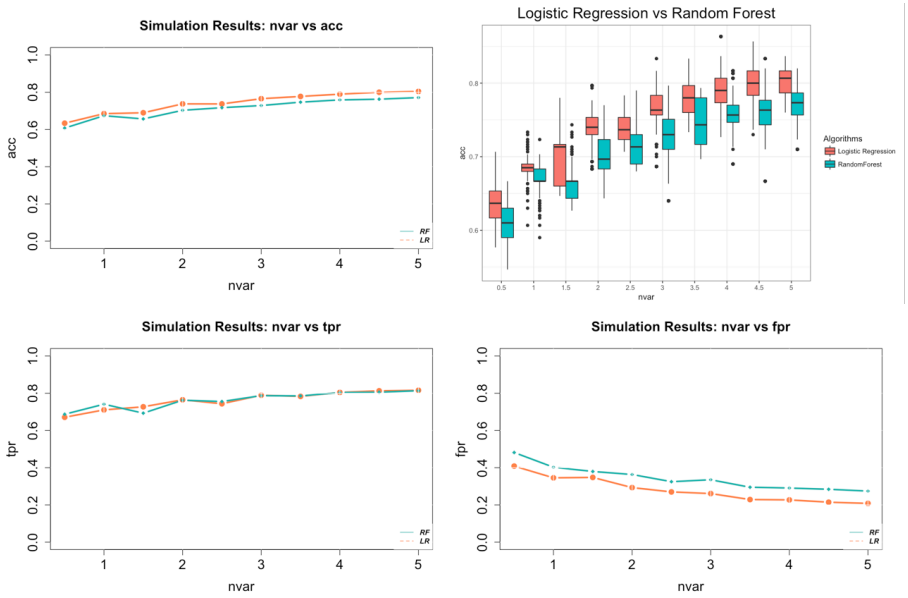


Fig. 10. Case 1 Simulation Results: 5 noise and 10 explanatory variables

The simulation is conducted again, but now adding more noise variables (noise = 100) and the results of the accuracy is shown in the top row of Figure 11. By looking at Figure 10 and Figure 11, a similar trend is observed. With a p-value less than 0.05, there is strong evidence to suggest that a significant difference in accuracy between the two models exists. The boxplots for this simulation is also comparable with 10 noise variables where minimal overlap in the boxplots for each model at each level of variance is observed. However, with 100 noise variables, the boxplots are much more consistent in that they are all about the same size for each level of variance. In Figure 11, the boxplots for each model are noticeably different sizes.

The bottom row of Figure 11 displays the true positive rate on the left and false positive rate on the right with 100 noise variables and 5 explanatory variables over increasing levels of variance in the variables. Interestingly, for variance 0.5 to 2.5 and a lot of noise, random forest has a higher true positive rate. At around variance = 3.0 and higher, random forest still has a higher true positive rate, but it is not as large of a difference from logistic regression than variance less than 2.5. The false positive rate for random forest is again higher

than logistic regression so the gap in higher true positive rate is not enough to make overall accuracy higher for random forest.

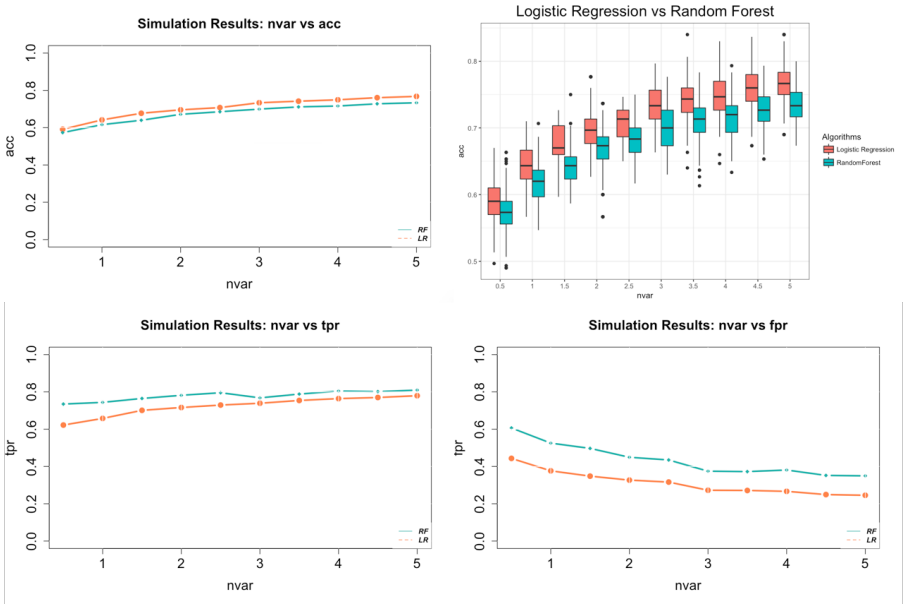


Fig. 11. Case 1 Simulation Results: 100 noise and 5 explanatory variables

6.2 Case 2

In case 2, we compared model performance with respect to change in the amount of noise in the dataset. We first did this by running 1000 simulations on a dataset with the number of noise variables = 1, 5, 10, 20, and 50. In Figure 12, we see the results of the accuracy for each model when the number of explanatory variables is 5 with 1000 observations. As we expected, as the amount of noise in the dataset increases, we see the accuracy start to decline for both models. However, we were not able to get the full picture by stopping at 50 noise variables, so we ran this again increasing the noise further. Figure 13 shows the results of accuracy when setting the number of noise variables to 1, 5, 10, 20, 40, 60, 80, 100, 150, 200. Accuracy is still slightly declining as we increase the noise past 50 noise variables and logistic regression is still performing with a higher accuracy (p-value = 9.559e-07).

The bottom rows of Figures 12 and 13 show the true positive rate on the left and false positive rate on the right. For true positive rate, when the number of noise variables is the less than or equal to the number of explanatory variables in the dataset, logistic regression is higher. However, once the number of noise

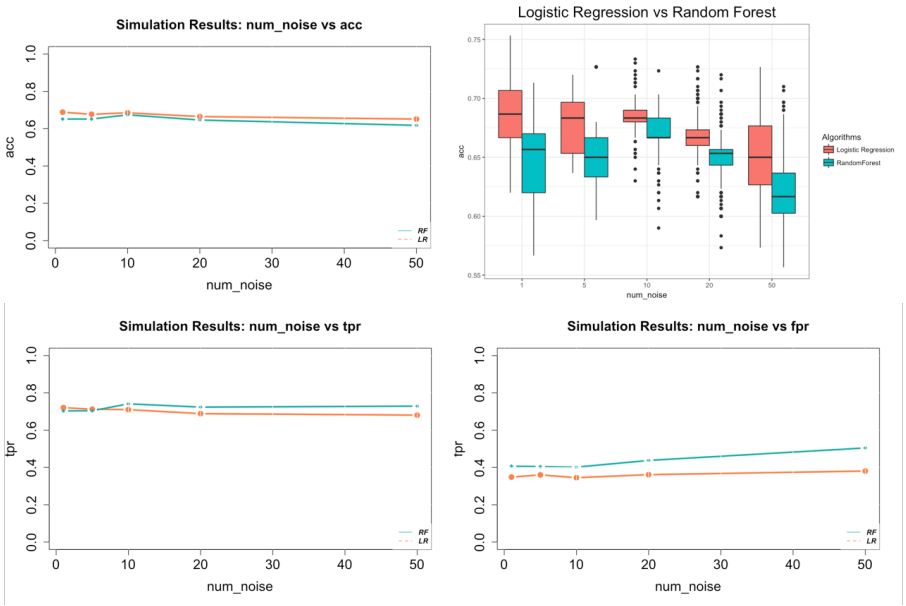


Fig. 12. Case 2 Simulation Results: 1 to 50 noise variables

variables exceeds the number of explanatory variables, random forest begins to have a higher true positive rate than logistic regression. As the amount of noise in the data increases, the false positive rate for both models also increase. However, the rate of increase in false positive rate for random forest is greater than the rate of increase in false positive rate for logistic regression as noise increases. Logistic regression does not have much change in true or false positive rate as noise increases past 50, but random forest false positive rate noticeably increases past 50 noise variables.

6.3 Case 3

In case 3, we compared model performance with respect to change in the number of explanatory variables in the dataset. In other words, the number of variables that relate to the response variable we are predicting. We did this by running 1000 simulations on a dataset with the number of explanatory variables = 1, 5, 10, 20, 30, 40, 50. In the top row of Figure 14, we see the results of the accuracy for each model when the number of noise variables is 50 with 1000 observations. With 30 or less explanatory variables, as the number of explanatory variables in the dataset increases, the accuracy increases as well. When the number of explanatory variables is above 30, random forest begins to taper off whereas logistic regression continues to increase in overall accuracy.

The bottom row of Figure 14 displays the true positive rate on the left and false positive rate on the right. When we look at the true positive rate for

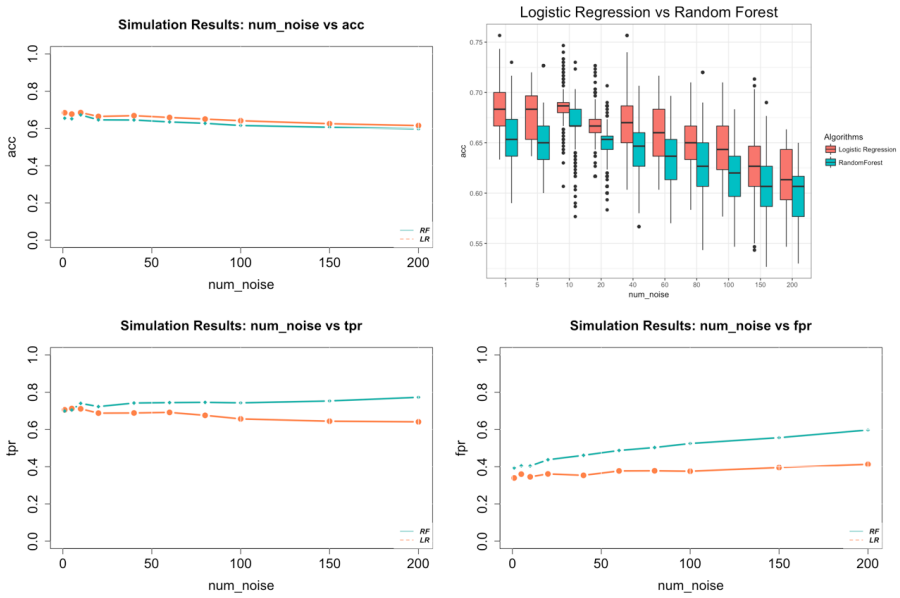


Fig. 13. Case 2 Simulation Results: 1 to 200 noise variables

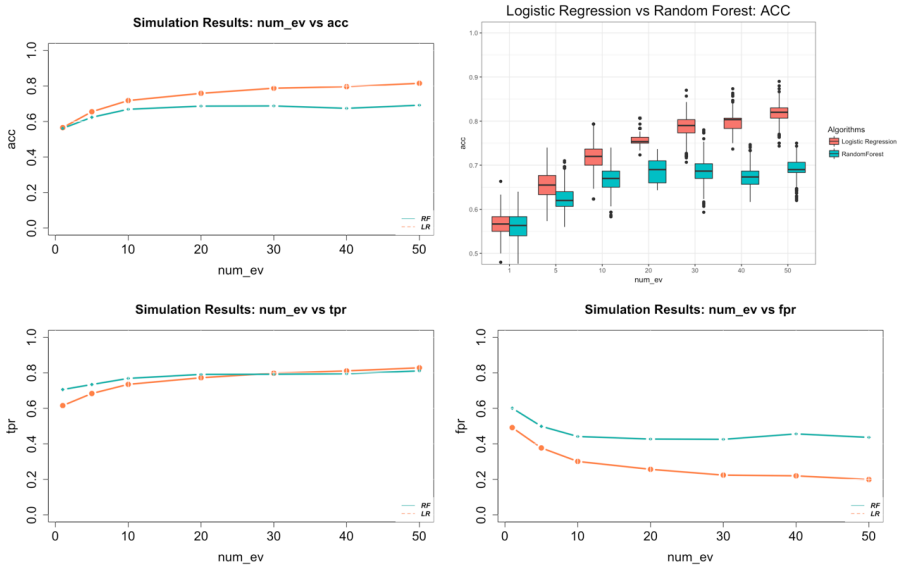


Fig. 14. Case 3 Simulation Results

the models, below 30 explanatory variables, random forest had a higher true positive rate. At 30 explanatory variables, logistic regression crosses over and continually increases to have a higher true positive rate than random forest. When we look at false positive rate, logistic regression decreases as we add more explanatory variables. Random forest false positive rate initially decreases from 1 to 10 explanatory variables but from 10 to 50 explanatory variables, there is not much change. The crossover point in true positive rate at 30 explanatory variables and the continued decrease in false positive rate for logistic regression as explanatory variables increases is evident in the overall accuracy plots. From the accuracy plots, we can see the drastic gap in performance of the two models after 20 to 30 explanatory variables.

6.4 Case 4

This simulation case study looked at iteratively increasing the number of observations in the dataset from 10 to 10000 while holding the number of explanatory variables in the model constant. A total of four different subcases were evaluated in which the number of explanatory variables ranged from 1,10,20, and 50; each case comprised of 10 noise variables. Given the computational complexities and completion time to train and validate a model for 1000 simulation as described in the previous cases and increasing the overall size of the dataset, the total number of simulations for specific case study is 10. Hence the moderate variance observed in Figure 15 and Figure 16.

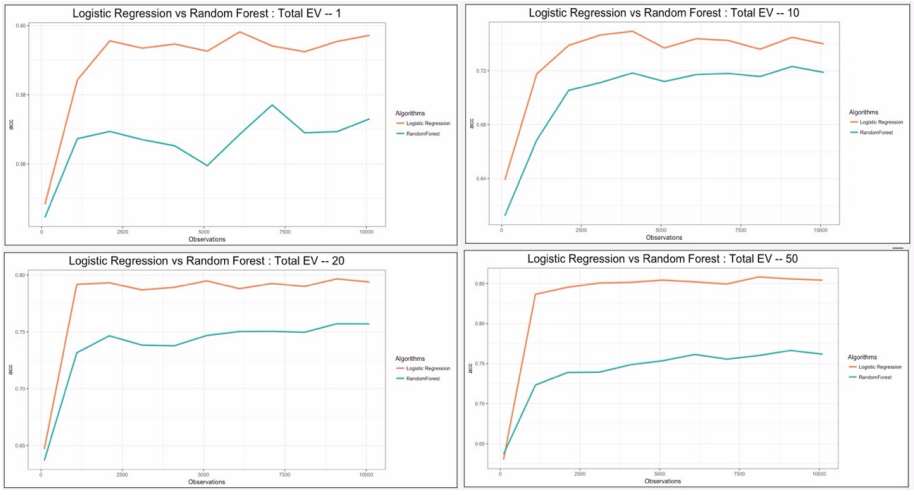


Fig. 15. Case 4 Simulation Results - Accuracy

One of the interesting findings in this simulation case is random forest and logistic regression perform nearly the same up until approximately 1000 obser-

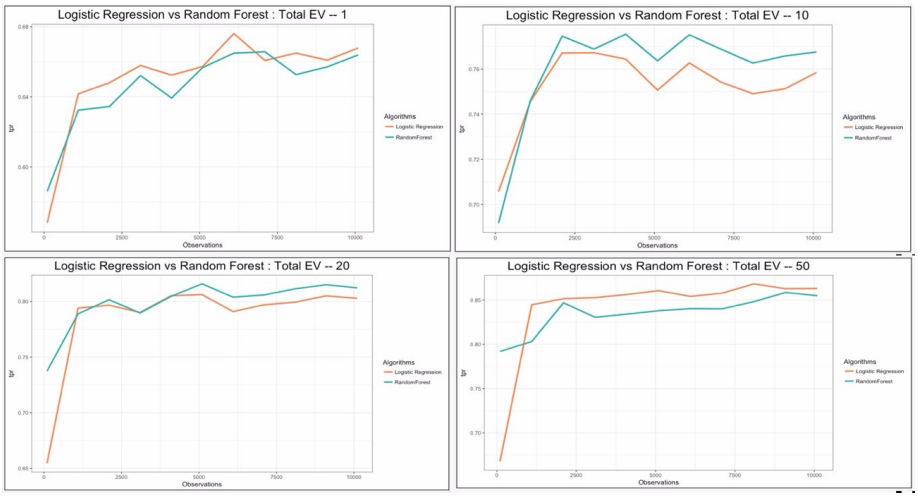


Fig. 16. Case 4 Simulation Results - True Positive Rate

vations in the dataset before diverging or in some instances crossing over as shown in Figure 16. Secondly, in Figure 15, the overall accuracy for logistic regression is consistently higher than random forest with 100 trees as both the number of explanatory and observations increases. With 10 and 20 explanatory variables included in the model respectively, the difference in overall accuracy is at a minimal compared to the case where 50 explanatory variables are included.

Additional analysis will need to be performed to determine if these observations are consistent when random forest is trained with a larger number of trees in addition to increasing the number of simulations to have a better understanding of the true accuracies.

6.5 Summary of Results

In summary, we found 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 for dataset with increasing noise variables. In all simulated case studies, we consistently found that the false positive rate for random forest with 100 trees was statistically different than logistic regression. In general, logistic regression performs better when the number of noise variables is less than or equal to the number of explanatory variables and random forest has a higher true and false positive rate as the number of explanatory variables increases in a dataset. Logistic regression and random forest are comparable for smaller datasets with less than 1000 observations.

7 Ethics

Data science is a rapidly evolving field and along with its success comes data privacy concerns. With massive amounts of data being collected today more so than ever before, our professional requirement is to conduct responsible and ethical innovation. In doing so, adhering to such ethical practices can continue fostering progress in data science and protect individual and group rights [5]. Our data is synthetically generated via the analytical tool developed for our analysis. Therefore, there is no liable concerns of the availability, integrity, usability, and security of the data. Moreover, the tool does not have the functionality to allow users the ability to upload their own dataset and thus does not raise any data privacy or security concerns. If a user intends to utilize this tool for simulating data that mimics a particular domain such as health or social media, then the user is responsible for any conclusions drawn from the tool. When it comes to simulating sensitive data such like medical records, two ethical problems arise. The absence of consent and deception of the research subject(s) [11]. For instance, in the event of having access to confidential records, the user of this tool should uphold ethical principles when simulating datasets that can closely mimic real-world data. One recommendation is to obtain informed consent from the research subjects conditioned on protection from research expose risks by the researcher.

8 Conclusions

We developed a machine learning model evaluation tool to simulate a range of different dataset characteristics. The tool allows users to specify characteristics of a dataset which then generates the simulated data and runs logistic regression and random forest on that data, returning the performance metrics of each model. We looked into the effects of changing four specific characteristics and each machine learning model's performance as that characteristic varies. Specifically, these characteristics were: (1) increasing the variance in the explanatory and noise variables, (2) increasing the number of noise variables, (3) increasing the number of explanatory variables, (4) increasing the number of observations. To compare classification performance between the two models, we used accuracy, area under the curve, true positive rate, false positive rate, and precision as metrics. We found 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 for dataset with increasing noise variables. In all simulated case studies, we consistently found that the false positive rate for random forest with 100 trees was statistically different than logistic regression.

To provide a statistical quantification as to whether a difference in model performance was conclusive enough to state the difference is significant or if the observed difference is by random chance, we used a pairwise two-sample t-test. We found that there is a statistically significant difference in classification

metrics when varying data characteristics using an alpha of 0.05. Increasing the variance in the explanatory and noise variables will cause a statistically significant difference in accuracy, true, and false positive rate. When varying the amount of noise in the dataset or the number of variables related to the response (explanatory variables), there is a statistically significant difference in accuracy and false positive rate but not true positive rate.

9 Future Work

In the current development version, the application provides the ability to answer in-depth statistical questions and evaluate classification performance of two machine learning models, random forest and logistic regression. Future development is to incorporate other algorithms such as Naive Bayes, XGBootst, and Artificial Neural Networks. Also, the application can be expanded beyond binary classification to multi-labeled datasets and evolving to include regression. Specifying the for the number of trees in the random forest model is an input the user will need to tune for in attempt to improve performance. Rather than hard coding the number of trees, the user could select an apply grid search option, which is an exhaustively optimization method that scans all possible parameter combinations in order to find the best estimators that yields the highest accuracy or other specified metrics. Lastly, one of the motivations behind this work is having the ability to open the door for other questions of interest to help address, like what is the average Type I or Type II error for logistic regression using forward selection criteria under similar data structures outlined in the results section of this paper. Just one example, but this application is intended to provide a foundation and the ability to expand well beyond the scope of work presented in this analysis.

References

1. Couronné, Raphael. Probst, Philipp.Boulesteix, Anne-Laure. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018
2. Beğenilmiş, Erdem; Üsküdarlı, Suzan. Organized Behavior Classification of Tweet Sets using Supervised Learning Methods. eprint arXiv:1711.10720. 11/2017.
3. Bertrand Michel. A Statistical Approach to Topological Data Analysis. Statistics [math.ST]. UPMC Université Paris VI, 2015.
4. Breiman, L. Random Forests. 2001
5. Floridi L, Taddeo M. What is data ethics? Phil.Trans.R.Soc.A 373: 20160360. 2016
6. Frédéric Chazal; Bertrand MichelAn introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. 2017
7. Goodfellow, Ian; Bengio,Yoshua; Courville, Aaron. Deep Learning. MIT Press. 2016
8. Graham, Dunn. Regression Models for Method Comparison Data. Journal of Biopharmaceutical Statistics 17:4, pages 739-756. 2007
9. Guerreiro, Manuela; Maroco, João; de Mendonça, Alexandre; Rodrigues, Ana; Santana, Isabel; Silva, Dina. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Research Notes20114:299. https://doi.org/10.1186/1756-0500-4-299.
10. Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning: data mining, inference and prediction. Springer. 2009
11. PLACEHOLDER
12. Rokach, Lior. Maimon, Oded. Data Mining with Decision Trees; Theory and Applications. 2nd ed. World Scientific Publishing Co.
13. Olson, Randal S.; Moore, Jason H. Identifying and Harnessing the Building Blocks of Machine Learning Pipelines for Sensible Initialization of a Data Science Automation Tool. eprint arXiv:1607.08878. 07/2016.
14. Ng, Andrew. CS229 Lecture Notes. Stanford University. 2012
15. Anastasiy Motrenko, Vadim Strijov, Gerhard-Wilhelm Weber: Sample size determination for logistic regression. Journal of Computational and Applied Mathematics Volume 255, Pages 743-752. 2014
16. Munch, Elizabeth. A User's Guide to Topological Data Analysis. University at Albany. 2017
17. Ruiz-Gazen, Anne; Villa, Nathalie Storms Prediction: Logistic Regression vs Random Forest for Unbalanced Data.
18. Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.
19. Thanh Lam, Hoang; Thiebaut, Johann-Michael; Sinn, Mathieu; Chen, Bei; Mai, Tiep; Alkan, Oznur. One button machine for automating feature engineering in relational databases. eprint arXiv:1706.00327. 06/2017.
20. Wolper, D.H. "No free lunch theorems for optimization," in IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 67-82, April 1997. doi: 10.1109/4235.585893
21. Zoran Bursac, C Heath Gauss, David Keith Williams, David W Hosmer: Purposeful Selection of Variables in Logistic Regression. Source Code for Biology and Medicine. 2008