Math. Appl. **8** (2019), 173–188
DOI: 10.13164/ma.2019.11

AN EMPIRICAL DEMONSTRATION OF THE NO FREE LUNCH THEOREM

EZEKIEL ADEBAYO OGUNDEPO AND ERNEST FOKOUÉ

Abstract. In this paper, we provide a substantial empirical demonstration of the statistical machine learning result known as the No Free Lunch Theorem (NFLT). We specifically compare the predictive performances of a wide variety of machine learning algorithms/methods on a wide variety of qualitatively and quantitatively different datasets. Our research work conclusively demonstrates a great evidence in favor of the NFLT by using an overall ranking of methods and their corresponding learning machines, revealing in effect that *none of the learning machines considered predictively outperforms all the other machines on all the widely different datasets analyzed.* It is noteworthy however that while evidence from various datasets and methods support the NFLT somewhat emphatically, some learning machines like Random Forest, Adaptive Boosting, and Support Vector Machines (SVM) appear to emerge as methods with the overall tendency to yield predictive performances almost always among the best.

1. Introduction

Throughout the relatively young yet tremendously fascinating history of statistical machine learning, data science and artificial intelligence, there has constantly been a fascination and interest in the possibility of creating/finding the holy grail in the form of a learning machine and hypothesis/function space that uniformly outperforms all other machines on all possible datasets. The so-called no free lunch theorem (NFLT) of which many different formulations and incarnations exist [3–5], is an intriguing and sometimes controversial result, that establishes that such a holy grail does not exist, namely that no learning machine exists that outperforms all other possible learning machines on all datasets. The goal of this paper is neither to rehash the debate nor to re-ignite some of the controversies surrounding NFLT but instead to harness the richness and easy availability of a wide variety of datasets [2] in a substantial comparison of the predictive performances of some of the most prominent and most commonly used statistical learning machines, with the finality of finding if there might be a conclusively empirical evidence in favor of NFLT. The very existence of NFLT stems from the fact that in statistical machine learning, the fact that the probability distribution of the generator of the data is

MSC (2010): primary 62F15; secondary 62F07.
Keywords: learning machine, generalization, Bayes risk, predictive performance, no free lunch theorem (NFLT), empirical evidence, statistical learning, data science, dataset, function space, random split, score function.

unknown in practice, makes it impossible to ever attain the universal best learning machine for any given task. Typically, the learning task consists of building functions $f : \mathcal{X} \longrightarrow \mathcal{Y}$ mapping elements of some input space $\mathcal{X}$ to those of some output space $\mathcal{Y}$. Specifically, one defines a theoretical risk functional $R(f)$ which is essentially the expected loss given by

$$R(f) = \mathbb{E}[\mathcal{L}(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(\mathrm{x}, \mathrm{y}) dP(\mathrm{x}, \mathrm{y}),$$

where $\mathcal{L}(\cdot, \cdot)$ is the so-called loss function, and ideally set out to achieve the goal of finding the universal best function

$$f^\star = \underset{f\in\mathcal{Y}^{\mathcal{X}}}{\operatorname{arginf}} \Big\{ \mathbb{E}[\mathcal{L}(Y, f(X))] \Big\} = \underset{f\in\mathcal{Y}^{\mathcal{X}}}{\operatorname{arginf}} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(\mathrm{x}, \mathrm{y}) dP(\mathrm{x}, \mathrm{y}) \right\}.$$

Unfortunately, the joint distribution $P(\mathrm{x}, \mathrm{y})$ of X and Y, is never known in practice, making it impossible to ever know $f^\star$. If the universal best function $f^\star$ were practically realizable, NFLT would make no sense. Since $f^\star$ is indeed not practically realizable, a substantial part of the research effort in statistical machine learning is dedicated to approximation theory in the sense of coming up with function spaces that hopefully have as strong a representation power as possible to help build function or learning machines that approximate $f^\star$ as accurately and precisely as possible specifically using data assumed to have been generated by the unknown distribution $P(\mathrm{x}, \mathrm{y})$. Indeed, given a data set $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ independently and identically drawn from $P(\mathrm{x}, \mathrm{y})$, along with a loss function $\mathcal{L}(\cdot, \cdot)$ and a function space $\mathcal{H}$ from which one can select members $f \in \mathcal{H}$, one can define a realizable empirical risk functional $\widehat{R}(f)$ as in Equation (1.1) given by

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(Y_i, f(X_i)), \tag{1.1}$$

and obtain the empirical best within the function space $\mathcal{H}$, namely

$$\widehat{f}_n = \underset{f\in\mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(Y_i, f(X_i)) \right\}. \tag{1.2}$$

It turns out that there are several practical, theoretical and philosophical challenges inherent in the empirical best defined in Equation (1.2), the details of which are far beyond the scope of the present paper. A more detailed account of the theoretical underpinnings of this vast body of results can be found in [1]. *Intuitively, one of the foundational challenges in statistical machine learning lies in the fact that making the empirical risk $\widehat{R}_n(f)$ is far from being ideal, since it is first of all achieved only for a given function space $\mathcal{H}$ that has a built approximation error, and even worst of all, it is constructed not on the whole population but solely on a fragment therefrom.* As a matter of fact, practical statistical machine learning has continually fed, triggered and animated the creativity, ingenuity and imagination of statisticians and computer scientists, many of whom never cease to come up with a wide variety of algorithms and learning machines vying to provide the best predictive performances imaginable. By predictive performances here, one means the average loss or risk on samples that were not used to build $\widehat{f}_n$. Before

delving into our gigantic empirical exploration, let it clear once again that we are not doubting the proofs given by the authors of the various incarnations of NFLT, and somehow trying to provide counters through evidences from the data. No! We are simply herein providing an empirical account to help readers gain insights into NFLT. Our work has the potential of serving the pedagogical and educational purposes of helping students understand and appreciate the niceties and subtleties of the construction of statistical learning machines and methods. Now, to date, there are several incarnations of NFLT. However, one of the earliest can be rightly attributed to [5], where the following claim is made by the authors: *It is shown that one cannot say: if empirical misclassification rate is low, the Vapnik-Chervonenkis dimension of your generalizer is small, and the training set is large, then with high probability your off-training set (OTS) error is small.* This statement was given in [5] to counter the pervading belief and use of the so-called bounds on the generalization error among statistical machine learning researchers, believed to be the ultimate way to deciding which machine is the best. *Indeed, whether the function space $\mathcal{H}$ from which the realized learning machine $\widehat{f}$ is selected, is explicitly or implicitly defined, it cannot consistently yield the universal best learning all possible learning tasks of its type, simply because no $\mathcal{H}$ can be the whole universe $\mathcal{Y}^{\mathcal{X}}$, and high probability is not the same as almost sure convergence.* In other words, no learning machine predictively outperforms all other possible learning machines on all problems of a given type. A learning machine works better than all other learning machines on a given task only when the explicit or implicit assumptions underlying the workings of that nicely performing machines are inherent in the data generator for that task. If the generator is changed, that learning machine will not longer have the best performance. The assumptions of a model for one problem may not hold for another. Therefore, it is not uncommon in machine learning, in fact it is typical to try several learning methods for any given task with the finality of empirically finding one that works best for that particular problem at hand. Practically speaking, rather than choosing a "favorite" learning machine to be resort all the time, it is better to always try several algorithms (parametric and/or nonparametric) and assess the trade-offs among speed, accuracy, and complexity of different models and find a model that works best for that particular problem [6, 7]. Please note that we have deliberately elected not to provide a formal mathematical version of NFLT, and we have done so to avoid distracting the reader from our decision to provide an intuitive take on it. Hence our repeated instances of intuitive aspects of NFLT throughout this paper. More formal presentations of NFLT can be found in [5].

2. Empirical demonstration of NLFT

We live in an era of data abundance or even data opulence. So abundant indeed is data in present times that the emerging field or discipline of Data Science is quickly becoming one of the most desirable paths for a lucrative and deeply fulfilling career. As we said earlier, the No Free Lunch Theorem (NFLT) has been around for quite some time now, and has been formulated in several different ways, and rigorously proven on sound mathematical foundations. Given the blessings of data

abundance, we have deemed up of practical and educational use to provide an empirical account of NFLT via a substantial comparison and ranking of the predictive performances of several existing learning machines and many quantitatively and qualitatively different data sets. Interestingly, we are not the first to embark on such an empirical study. Similar work, albeit of a lesser magnitude and coverage than ours, have been attempted before by several authors. To demonstrate empirically with data that there is no best machine learning algorithm (classification or regression) on every instance, [8] compared four classification learning methods (J48, CART, Random Forest, and Bayesian Network) on the students' academic performance data at colleges of Assam in India before decided to choose random forest due to its performance using classifier error metric. Realizing that the true probability of default on credit card debt by clients is unknown, [9] compared six data mining techniques which included k Nearest Neighbors (kNN), Logistic Regression, Discriminant Analysis, Naïve Bayesian, Artificial Neural Networks (ANN), and Classification Trees on the real card holders' credit risk data in Taiwan. Among the methods used for the study, only an ANN model achieved the best performance based on area ratio and a relatively low error rate. As a result, [9] concluded that an ANN was the best model that could accurately estimate the real probability of default among the 6 classification methods compared. The reference [10] compared regression learning methods which are based on tree structures (Decision trees (DT), Random Forests (RF)) and non-linear functions such as Neural networks (NN) and support vector machines (SVM) to predict the burned area of forest fires from the northeast region of Portugal meteorological data. Out of the five learning methods compared, only SVM was capable of predicting the burned area of small fires, which are more frequent. The reference [11] applied data mining classification algorithms viz. C4.5, C-RT, CS-MC4, Decision List, ID3, Naïve Bayes, and RndTree in predicting vehicles collision patterns on the road accident training dataset obtained from the Fatality Analysis Reporting System (FARS), University of Alabama. The experimental results indicated that the RndTree classification algorithm achieved better accuracy than other algorithms in classifying the manner of the collision which increases the fatality rate in road accidents. Although all the authors cited above used the out of training sample prediction to elect the best method for their particular task, one cannot use their work as a basis of evidence for NFLT because the comparisons do not involve the crucial cross product of several learning machines with several data sets. Our work remedies their limitations by carefully considering the cross product of several learning machines with several qualitatively and quantitatively different data sets. Like we said earlier, one would ideally like to compare the true theoretical performances of the methods measured by the risk function

$$R(f) = \mathbb{E}[\mathcal{L}(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(\mathbf{x}, \mathrm{y}) dP(\mathbf{x}, \mathrm{y}).$$

Due to the fact that $P(\mathbf{x}, \mathrm{y})$ is unknown in practice, we will instead compute many instances of the error on the out-of-training set (test set), and perform several statistics on those values in order to measure, compare and rank all the methods considered for each of the datasets considered. The stochastic hold out approach used here will typically create R random splits of the provided realized data set

$\mathcal{D}_n = \{(\mathbf{x}_1, \mathrm{y}_1), \cdots, (\mathbf{x}_n, \mathrm{y}_n)\}$, into a training set and a test set, for each task, and for each learning machine $\widehat{f}$, predictive performance quantities like the average test error $\texttt{AVTE}(\widehat{f})$ of $\widehat{f}$ over the R random splits [15] will be computed and then used as the score of $\widehat{f}$ for comparison purposes.

$$\texttt{AVTE}(\widehat{f}) = \frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{s}\sum_{t=1}^{s}\mathcal{L}(\mathrm{y}_t^{(r)}, \widehat{f}_r(\mathbf{x}_t^{(r)}))\right\}$$

where $\widehat{f}_r(\cdot)$ is the r^{th} realization of the estimator $\widehat{f}(\cdot)$ built using the training portion of the split of $\mathcal{D}_n$ into a training set and test set, and $(\mathbf{x}_t^r, \mathrm{y}_t^r)$ is the t^{th} observation from the test set at r^{th} random replication of split of $\mathcal{D}_n$. For each of the datasets considered, $\texttt{AVTE}(\widehat{f})$ is computed for each of the learning machines of interest, and the values are ranked.

3. Implementation and applications

The benchmark datasets used in this study are rich in terms of varieties and dimensions. We have binary class, multi-class and regression datasets. Table 1, 2 and 3 has $k = \frac{n}{p}$ which represented the measure of the information (data) richness. When k is very small, we said we are in an information poverty regime, if otherwise, we are in an information opulence regime. Data opulence is good for the weak law of large numbers and the strong law of large numbers.

Table 1. Structure of multi-class datasets.

SN	Dataset	n	p	G(Number of classes)	k = n/p
1	Balance scale	625	5	3	125
2	Cars	1728	7	4	246.86
3	Indonesia contraceptive method choice	1473	10	3	147.30
4	Iris	150	5	3	30
5	Red wine quality	1599	12	6	133.25
6	Seeds	210	8	3	26.25
7	Vechicle silhouett	846	19	4	44.53
8	White wine quality	4898	12	7	408.17
9	Wine recognition	178	14	3	12.71

The predictive performance of machine learning models depends on the structure of the dataset and proper data preparation will ensure the models work optimally. Since the best machine learning method on the dataset cannot be known beforehand, in this section, we carry out an empirical study through benchmark datasets to demonstrate that no universally best machine learning algorithm exists for all datasets. This study used datasets at the UCI Machine Learning Repository[16]. Additional datasets were retrieved from MLData and GPA dataset [12]. Statistical analyses were run in R studio with R version 3.6.1. Packages used included tidyverse for data analysis and visualization [13] and caret for regression and classification training [14]. Raw data and scripts are available on GitHub at `https://github.com/gbganalyst/NFLT-journal` for scientific reproducibility.

We performed various data pre-processing activities, such as data imputation for data sets that were found to have missing values. We also scaled numerical features and did some appropriate encoding of categorical features that needed it. We specifically used 80% of data for model training, and the remaining 20% for model evaluation or performance. Wherever needed, we also perform the appropriate tuning of all the hyperparameters of models in the caret package using 10-fold cross-validation.

Table 2. Structure of binary class datasets.

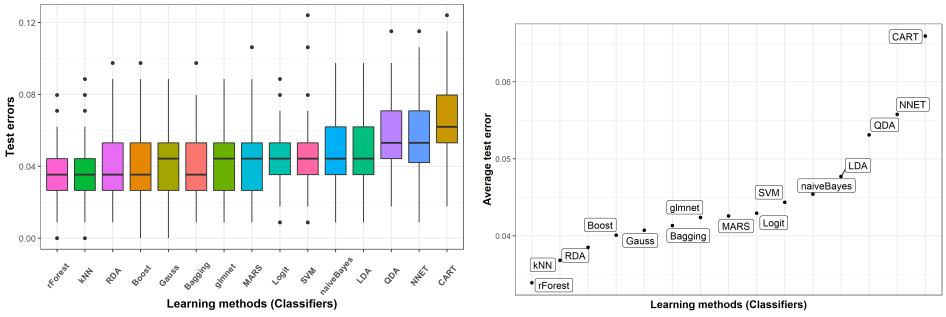
SN	Dataset	n	p	C(Number of classes)	k = n/p
1	Asthmatic	405	11	2	36.82
2	Breast cancer	569	10	2	56.90
3	Congressional	435	17	2	25.59
4	Cryotheraphy	90	7	2	12.86
5	Diabetic retinopathy debrecen	1151	20	2	57.55
6	Gender voice	3168	21	2	150.86
7	HTRU2	17899	9	2	1988.78
8	Indian liver patient	583	11	2	53
9	Monk problem	1711	8	2	213.88
10	Social network advertisement	400	4	2	100
11	Sonar dataset	208	61	2	3.41

Table 3. Structure of regression datasets.

SN	Dataset	n	p	k = n/p
1	Abalone	4178	9	464.22
2	Airfoil self-noise	1503	6	250.5
3	Attendance	680	3	226.67
4	Auto mpg	392	8	49
5	Boston housing price	506	14	36.14
6	Charity	4268	5	853.6
7	Combined cycle power plant	9568	5	1913.6
8	Computer hardware	209	7	29.86
9	Concrete comprehesion strenght	1030	9	114.44
10	Diabetes	442	11	40.18
11	Ducan MBA	203	7	29
12	Forest fire	517	13	39.77
13	GPA	141	5	28.2
14	Hprice 2	506	6	84.33
15	Insurance	1338	7	191.14
16	Istanbul stock	536	8	67
17	Mortality rate	992	4	248
18	Red wine quality	1599	12	133.25
19	Servo	167	5	33.4
20	Wage	935	4	233.75
21	White wine quality	4898	12	408.17
22	Yacht hydrodynamics	308	7	44

3.1. Empirical demonstration of NFLT in binary classification

To show tangible practical evidence that the No Free Lunch Theorem is backed by several real life data sets, we trained fifteen (15) different classifiers chosen from linear and/or non-linear, parametric and/or non-parametric learning machines on different binary class datasets. The distribution of the test errors of all the learning machines over 100 replications is shown in Figure 1a and Figure 1b for comparison. As shown in Figure 1, Random Forest has the smallest test error on the breast cancer dataset when compared to other learning methods under consideration. One could therefore conclude empirically that Random Forest is the winning method (minimum misclassification rate) for predicting whether a tumor is benign or malignant.



(a) Distribution of test error over 100 replicates.

(b) Average test error over 100 replicates.

Figure 1. Example of NFLT of ML methods on breast cancer dataset.

The mean, median, and standard deviation rank of each classifier on each dataset is shown in Table 4, 5 and 6, respectively.

Table 4. The rank of the mean score of method M on binary data S.

SN	Dataset	Bagging	Adaboost	CART	Gauss	glmnet	kNN	LDA	Logit	MARS	naiveBayes	NNET	QDA	RDA	rForest	SVM
1	Monk	1	3	6	10	11	2	13	14	12	15	5	9	8	4	7
2	Indian liver patient	9	10	12	2	3	11	6	1	7	15	13	14	8	5	4
3	Gender voice	5	3	13	6	9	7	11	8	4	14	12	-	10	2	1
4	Cryotherapy	7	13	15	3	2	9	4	5.5	11	14	12	5.5	8	1	10
5	Breast cancer	8.5	5	15	3	8.5	2	12	11	7	10	14	13	4	1	6
6	Diabetic retinopathy debrecen	11	9	13	6	3	12	5	2	1	14	4	-	8	10	7
7	Congressional voting	5	1	4	12	2	13	10	6	7	14	11	-	9	3	8
8	Social network advertisement	12	5	3	4	14.5	2	14.5	13	8	11	10	6.0	9	7	1
9	Sonar	8	3	14	6	9	7	11	13	10	15	5	12	4	1	2
10	HTRU2	9	4.5	11	6	2	10	12	3	1	15	4.5	14	13	8	7
11	Asthmatic	14	11	3	12	4	10	1	6	2	8	13	-	5	9	7
	Mean of the mean rank	**8.14**	**6.14**	**9.91**	**6.36**	**6.18**	**7.73**	**9.05**	**7.5**	**6.36**	**13.18**	**9.41**	**10.5**	**7.82**	**4.64**	**5.45**
	Overall rank	**10**	**3**	**13**	**5.5**	**4**	**8**	**11**	**7**	**5.5**	**15**	**12**	**14.**	**9**	**1**	**2**

General note for all tables:
Performance of classifiers are ranked on each dataset row-wise

For instance on Table 4, the mean test error of Bagging was ranked 1^{st} while Naïve Bayes was ranked 15^{th} on Monk dataset (SN: 1). On breast cancer dataset (SN: 5), the mean test error of Random Forest was ranked 1^{st} while Decision Tree (CART) was ranked 15^{th}. It can be seen clearly that no classifier exists that outperformed all possible learning machines on all datasets. Training Quadratic

Discriminant Analysis (QDA) classifier on gender voice, diabetic retinopathy debrecen, and asthmatic datasets could not work due to rank deficiency in the classes of their labels. The mean of the mean rank of each machine learning method was calculated for all the datasets and it can be seen that random forest, SVM, and naiveBayes were ranked 1st, 2nd and 15th, respectively.

Table 5. The rank of the median score of method M on binary data S.

SN	Dataset	Bagging	adaboost	CART	Gauss	glmnet	kNN	LDA	Logit	MARS	naiveBayes	NNET	QDA	RDA	rForest	SVM
1	Monk	1	3	7	10	11	2	13.5	13.5	12	15	5	8.5	8.5	4	6
2	Indian liver patient	9.5	9.5	12	3	3	12	7	3	3	15	12	14	7	7	3
3	Gender Voice	5.5	3.5	13	5.5	8.5	7	11	8.5	3.5	14	11	-	11	2	1
4	Cryotherapy	4	10.5	14.5	4	4	10.5	4	4	10.5	14.5	10.5	4	10.5	4	10.5
5	Breast cancer	6.5	6.5	15	6.5	6.5	1.5	13	6.5	6.5	11	13	13	6.5	1.5	6.5
6	Diabetic retinopathy debrecen	11	9	12.5	6	3	12.5	5	2	1	14	4	-	8	10	7
7	congressional voting	5.5	2.5	2.5	12	2.5	13	9	9	5.5	14	9	-	9	2.5	9
8	Social network advertisement	12	3	3	3	14	3	14	14	8.5	8.5	8.5	8.5	8.5	8.5	3
9	Sonar	8	4	14	6	9.5	6	11	12.5	9.5	15	6	12.5	2	2	2
10	htru2	6	6	10.5	6	6	10.5	12.5	1.5	1.5	15	6	14	12.5	6	6
11	Asthmatic	14	10	5.5	12	5.5	10	1.5	5.5	1.5	5.5	13	-	5.5	10	5.5
	Mean of the median rank	**7.55**	**6.14**	**9.95**	**6.73**	**6.68**	**8**	**9.23**	**7.27**	**5.73**	**12.86**	**8.91**	**10.64**	**8.09**	**5.23**	**5.41**
	Overall rank	**8**	**4**	**13**	**6**	**5**	**9**	**12**	**7**	**3**	**15**	**11**	**14**	**10**	**1**	**2**

Recalling that the mean as a measure of central tendency is always affected by outliers, we considered comparing the median ranks of the test errors for all the learning machines on all the datasets as shown in Table 5. For instance, on the monk dataset (SN: 1), the median test error of Bagging was ranked 1st while Naïve Bayes was ranked 15th. The median test error of Multivariate Adaptive Regression Splines (MARS) was ranked 1st while Logistic Regression was ranked 2nd in predicting whether an image contained signs of diabetic retinopathy or not on diabetic retinopathy debrecen dataset (SN: 6). This again shows that no classifier won on all datasets as shown in the median rank of test error. The mean of the median ranks of the learning machines was calculated for all the datasets and it can be seen that Random Forest, SVM, and Naïve Bayes were ranked 1st, 2nd, and 15th respectively.

Table 6. The rank of the standard deviation (std) score of method M on binary data S.

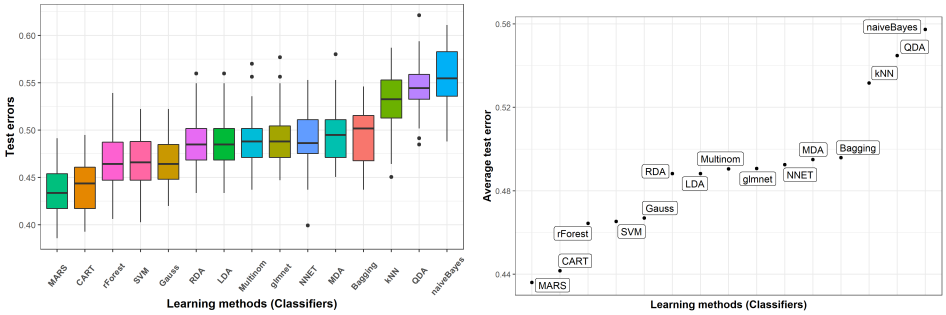
SN	Dataset	Bagging	adaboost	CART	Gauss	glmnet	kNN	LDA	Logit	MARS	naiveBayes	NNET	QDA	RDA	rForest	SVM
1	Monk	3	4	6	9	13	2	11	12	14	10	15	7	8	1	5
2	Indian liver patient	11	14	12	5	2	8	4	7	3	9	15	13	6	10	1
3	Gender Voice	6	2.5	12	10.5	9	10.5	8	5	4	13	14	-	7	2.5	1
4	Cryotherapy	6	13	14	1	3	8	5	12	15	10	11	4	9	2	7
5	Breast cancer	8	13	15	9	12	4	5	11	2	1	14	10	6.5	6.5	3
6	Diabetic retinopathy debrecen	13	8	11	6	4	7	10	3	12	1	5	-	14	9	2
7	congressional voting	5	2	8	10	13	1	14	3	11	6	15	7	12	4	9
8	Social network advertisement	7	5	2	6	14	1	13	15	9	12	8	10	11	3	4
9	Sonar	13	4	12	8	7	10	3	2	14	15	11	5	9	1	6
10	htru2	2.5	1	9	5	9	5	12	11	5	15	7	14	13	2.5	9
11	Asthmatic	13	9	1	12	2	10	3	4	7	5.5	14	-	11	5.5	8
	Mean of the std rank	**7.95**	**6.86**	**9.27**	**7.41**	**8.0**	**6.05**	**8**	**7.73**	**8.73**	**8.86**	**11.73**	**8.75**	**9.68**	**4.27**	**5**
	Overall rank	**7**	**4**	**13**	**5**	**8.5**	**3**	**8.5**	**6**	**10**	**12**	**15**	**11**	**14**	**1**	**2**

The predictive stability of the 15 classifiers was compared by taking the rank of the standard deviation (std) of the test errors of each classifier on a given dataset as shown in Table 6. The standard deviation test error of k Nearest Neighbors (KNN) was ranked 1st while Artificial Neural Network (NNET) was ranked 15th on the USA congressional voting dataset (SN: 7). This shows that a neural network model is not stable in its prediction across 100 replications. Evidence of NFLT can be seen, for example, while the std test error of Naïve Bayes was ranked 15th

on Sonar and htru2 datasets, its std was ranked 1st on breast cancer and diabetic retinopathy Debrecen datasets. The mean of the std rank of each machine learning method was calculated for all the datasets and it can be seen that random forest, SVM, and neural network were ranked 1st, 2nd, and 15th, respectively.

3.2. Empirical demonstration of NFLT in multicategorical classification

In this section, we show practical evidence that the NFLT is indeed valid by training 15 different classifiers on 9 different multi-class datasets. The classifiers were chosen from linear and/or nonlinear, parametric and/or non-parametric. Figure 2a is the distribution of test errors of each learning method over 100 replications on the Indonesia Contraceptive Method Choice dataset and Figure 2b is its the average test error. The main idea of running all the methods on each dataset is to vividly demonstrate that no classifier/learning method won on all the datasets (see Tables 7–9).



(a) Distribution of test error over 100 replicates.

(b) Average test error over 100 replicates.

Figure 2. Example of NFLT of ML methods on Indonesia Contraceptive Method Choice dataset.

As shown in Figure 2, MARS has the least test error on the balance scale dataset when compared to other learning methods under consideration. This means that MARS is suitable in predicting which way the scale tips (left, balance, or right) accurately. It can be seen that Naïve Bayes has the least accuracy therefore, it is not advisable to use it in this problem.

Table 7. The rank of the mean score of method M on multi class data S.

SN	Dataset	Bagging	CART	Gauss	glmnet	kNN	LDA	MARS	MDA	Multinom	naiveBayes	NNET	QDA	RDA	rForest	SVM
1	seeds	11	15	13	4	12	1.5	8	3	5	14	6	7	1.5	10	9
2	Wine recognition	14	15	5	7	13	3	12	4	9	11	10	2	1	8	6
3	Contraceptive method	7	3	2	11	6	9	1	8	10	14	13	15	12	5	4
4	Red wine quality	2	5	4	6	9	11	8	13	7	15	12	14	10	1	3
5	Cars	1	5	11	7	14	9	10	12	6	13	2	-	8	3	4
6	Balance scale	14	15	3	11	7	12	9	8	10	6	5	1.5	1.5	13	4
7	white wine quality	2	9	3	4	7	11	8	12	5	14	15	13	10	1	6
8	Vehicle Silhouette	11	14	12	4	13	7	6	3	5	15	9	2	1	10	8
9	Iris	13.5	12	10	5	8	1.5	15	3	6	7	9	4	1.5	13.5	11
	Mean of the mean rank	**8.39**	**10.33**	**7**	**6.56**	**9.89**	**7.22**	**8.56**	**7.33**	**7**	**12.11**	**9**	**7.31**	**5.17**	**7.17**	**6.11**
	Overall rank	**10**	**14**	**4.5**	**3**	**13**	**7**	**11**	**9**	**4.5**	**15**	**12**	**8**	**1**	**6**	**2**

The rank of the mean test error of each classifier is shown in Table 7. For instance, on wine recognition dataset (SN: 2), the mean test error of Regularized Discriminant Analysis (RDA) was ranked 1st while Decision Tree (CART) was ranked 15th in predicting the three types of wines accurately. Evidence of NFLT can be seen, for example, while the mean test error of MARS was ranked 15th on the Iris dataset (SN: 9), it was ranked 1st on Indonesia contraceptive method choice dataset (SN: 3). As shown in Table 7, QDA could not work on the cars dataset due to the rank deficiency in classes of the label. The mean of the mean rank of each machine learning method was calculated for all the datasets and it can be seen that RDA was ranked 1st.

Table 8. The rank of the median score of method M on multi class data S.

SN	Dataset	Bagging	CART	Gauss	glmnet	kNN	LDA	MARS	MDA	Multinom	naiveBayes	NNET	QDA	RDA	rForest	SVM
1	Seeds	10.5	14.5	10.5	5	10.5	1.5	10.5	5	5	14.5	5	5	1.5	10.5	10.5
2	Wine recognition	11.5	15	4	4	11.5	4	11.5	4	8	11.5	11.5	4	4	11.5	4
3	Contraceptive method	8	2.5	2.5	11	5.5	8	1	8	10	14	13	15	12	5.5	4
4	Red wine quality	2	4.5	6.5	4.5	9	10.5	8	13	6.5	15	12	14	10.5	1	3
5	Cars	1	5	11	7	14	9	10	12	6	13	2	-	8	3	4
6	Balance scale	14	15	4.5	11	8	12	8	8	10	4.5	4.5	1.5	1.5	13	4.5
7	white wine quality	2	6.5	3	6.5	6.5	11	6.5	12	6.5	14	15	13	10	1	6.5
8	Vehicle Silhouette	11.5	14	11.5	4	13	6.5	6.5	3	5	15	8	1.5	1.5	10	9
9	Iris	12.5	12.5	12.5	5	5	5	12.5	5	5	5	5	5	5	12.5	12.5
	Mean of the median rank	**8.11**	**9.94**	**7.33**	**6.44**	**9.22**	**7.5**	**8.28**	**7.78**	**6.89**	**11.83**	**8.44**	**7.38**	**6**	**7.56**	**6.44**
	Overall rank	**10**	**14**	**5**	**2.5**	**13**	**7**	**11**	**9**	**4**	**15**	**12**	**6**	**1**	**8**	**2.5**

We compared the rank of the median test error of each classifier has shown in Table 8. The median test error of random forest was ranked 1st while artificial neural network (nnet) was ranked 15th in predicting the rating of wine quality on the white wine quality dataset (SN: 7). Overall, RDA has the least test error across the 9 datasets.

Table 9. The rank of the SD score of method M on multi class data S.

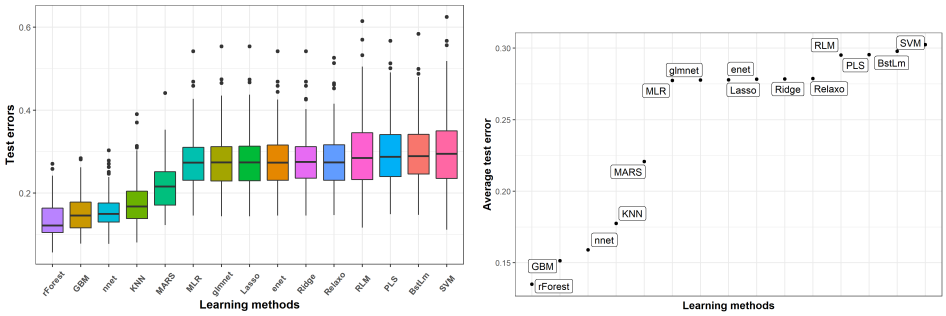
SN	Dataset	Bagging	CART	Gauss	glmnet	kNN	LDA	MARS	MDA	Multinom	naiveBayes	NNET	QDA	RDA	rForest	SVM
1	seeds	9	14	8	6	13	1	7	3	5	15	10.5	4	2	12	10.50
2	Wine recognition	14	15	3	6	13	1	11.5	5	11.5	10	9	7	2	8	4
3	Contraceptive method	11	7.5	9	2	3	4	6	14	1	7.5	15	10	12	5	13
4	Red wine quality	7	8	5	3	9	10	4	13	6	15	12	14	11	2	1
5	Cars	1	6	3	7.5	10	9	11	14	5	12	13	-	7.5	2	4
6	Balance scale	13	14	2	8	3	5	7	6	9	4	15	10.5	10.5	12	1
7	White wine quality	11	7	5	3	1.5	10	4	12	6	14	15	13	9	8	1.5
8	Vehicle Silhouette	10	8	12	1	11	2	7	13	5	14	15	4	3	6	9
9	Iris	15	14	11	7	5	1.5	13	3	8	12	6	4	1.5	10	9
	Mean of the std rank	**10.11**	**10.39**	**6.44**	**4.83**	**7.61**	**4.83**	**7.83**	**9.22**	**6.28**	**11.5**	**12.28**	**8.31**	**6.5**	**7.22**	**5.89**
	Overall rank	**12**	**13**	**5**	**1.5**	**8**	**1.5**	**9**	**11**	**4**	**14**	**15**	**10**	**6**	**7**	**3**

Table 9 is the rank of the standard deviation test errors of each classifier on a given dataset. On vehicle Silhouette dataset (SN 8), the std test error of GLMNET was ranked 1st while Neural Network was ranked 15th in predicting vehicle type (Opel, Saab, Bus, and Van). This means that the neural network model is not stable in its prediction across 100 replications. The standard deviation of the test errors of Bagging was ranked 1st on the car dataset and was ranked 15th on Iris dataset. The mean of the std rank of each machine learning method was calculated for all the datasets and it can be seen that the neural network was ranked 15th.

3.3. Empirical demonstration of NFLT in regression learning

Fifteen (15) different regression function spaces which were chosen from parametric and/or non-parametric to model the functional patterns underlying the data were

considered. We compared the test errors (MSE) of each learning method over 100 stochastic hold out subsamples with the training set representing $\frac{2}{3}$ of the data. The main idea of running all the methods on each dataset is to tangibly demonstrate that no learning method won on all the datasets (see Tables 10–12). The distribution of the test errors of each learning method on the Boston dataset was shown in Figure 3. Out of the 15 learning methods compared in Figure 3, only the decision tree (CART) has the least mean square error on the Boston housing price test set.



(a) Distribution of test error over 100 replicates

(b) Average test error over 100 replicates

Figure 3. Example of NFLT of ML methods on Boston median housing price dataset.

The mean rank of the test error of each machine learning method is shown in Table 10. For instance, on the power plant dataset (SN: 1), the mean test error of Random Forest was ranked 1^{st} while Partial Least Square (PLS) was ranked 15^{th} in predicting the net hourly electrical energy output (EP) of the plant. Evidence of NFLT can be seen, for example, while the mean test error of Boosted Linear Model (BstLM) was ranked 15^{th} on Istanbuck stock dataset (SN: 16), it was ranked 1^{st} on attendance figures dataset (SN: 4). The mean of the mean rank of each machine learning method was calculated for all the datasets and Artificial Neural Network (NNET) was ranked 1^{st}.

The median rank of the test errors of each learning method on a given dataset is shown in Table 11. For example, on the auto mpg dataset (SN: 6), the median test error of SVM was ranked 1^{st} while PLS was ranked 15^{th} in predicting engine miles per gallon of cars from 1970s and 1980s. The evidence of NFLT can be seen for example, while the median test error of Random forest was ranked 1^{st} on datasets such as powerplant (SN: 1) and Boston housing price (SN: 5), its median test error was ranked 15^{th} on both diabetes (SN: 9) and wage (SN: 22) datasets. The mean of the median rank of each machine learning method was calculated for all datasets and it can be seen that nnet was ranked 1^{st}.

Table 12 is the rank of the standard deviation of test error of each machine learning method over 100 hold out sub-sample. With Istanbul Stock (SN: 16) dataset, the std test error of Multiple Linear Regression (OLS) was ranked 1^{st}

Table 10. The rank of the mean score of method M on regression data S.

SN	Datasets	BstLm	enet	GBM	glmnet	KNN	Lasso	MARS	MLR	nnet	PLS	Relaxo	rForest	Ridge	RLM	SVM
1	Powerplant	12	9	4	9	2	13	6	9	5	15	14	1	9	9	3
2	Abalone	15	7	5	9.5	12	9.5	3	7	1	13	14	4	7	11	2
3	Air foil	14	9	5	7	4	12	6	9	2	13	15	1	9	11	3
4	Attendance	1	4.5	9	3	12	6	7	2	10	11	15	14	4.5	8	13
5	Boston	15	11	2	8	5	7	6	10	4	14	9	1	12	13	3
6	Auto mpg	8	14	4	13	6	10	5	8	1	15	12	3	8	11	2
7	Computer hardware	5	9	13	10	14	8	3	6	2	11	4	1	7	12	15
8	Concrete	14	9	2	11	6	12	5	9	3	15	7	1	9	13	4
9	Diabetes	9	6	10	5	14	4	13	2	1	8	11	15	7	3	12
10	Charity	9	13	1	11	3	10	5	12	2	8	7	4	14	15	6
11	Ducan MBA	2	7	13	1	15	4	8	6	14	9	5	11	3	10	12
12	MSU GPA	8	7	10	1	13	5	12	4	11	2	9	14	3	6	15
13	Forest fire	6	4	1	7	13	5	8	11	9	2	3	14	12	10	15
14	Housing price	12	9	2	7	5	11	6	8	3	13	15	1	10	14	4
15	Insurance	12	7	1	9	13	10	5	7	2	11	15	3	7	14	4
16	Istanbuck Stock	1	5	12	3	14	7	11	2	9	10	6	13	8	4	15
17	Red wine quality	8	11	3	12.5	15	7	4	12.5	5	14	6	1	9	10	2
18	white wine quality	15	12	3	11	4	8	6	10	5	14	9	1	13	7	2
19	Yacht	14	5	2	8	10	9	4	7	3	15	12	1	6	11	13
20	Servo system	14	12	2	9	5	10	6	8	3	7	15	1	11	13	4
21	Mortality rate	14	8	4	10	5	13	6	8	1	11	15	2	8	12	3
22	Wage	2.5	6	12	2.5	14	7	10	4.5	9	8	11	15	4.5	1	13
	Mean of the mean rank	**9.57**	**8.39**	**5.45**	**7.61**	**9.27**	**8.52**	**6.59**	**7.36**	**4.77**	**10.86**	**10.41**	**5.55**	**8.23**	**9.91**	**7.5**
	Overall rank	**12**	**9**	**2**	**7**	**11**	**10**	**4**	**5**	**1**	**15**	**14**	**3**	**8**	**13**	**6**

Table 11. The rank of the median score of method M on regression data S.

SN	Datasets	BstLm	enet	GBM	glmnet	KNN	Lasso	MARS	MLR	nnet	PLS	Relaxo	rForest	Ridge	RLM	SVM
1	Powerplant	12	9	4	9	2	13	6	9	5	15	14	1	9	9	3
2	Abalone	15	9	4	7	12	6	3	10	1	13	14	5	8	11	2
3	Air foil	14	8.5	5	8.5	3	12	6	8.5	2	13	15	1	8.5	11	4
4	Attendance	3	6	9	7	12	8	2	4.5	1	11	15	14	4.5	10	13
5	Boston	14	12	3	7	5	9	6	11	4	15	10	1	8	13	2
6	Auto mpg	13	8.5	4	14	6	8.5	5	8.5	3	15	12	2	8.5	11	1
7	Computer hardware	6	11	13	8	14	5	3	9	2	12	7	1	10	4	15
8	Concrete	14	11	2	9	6	7.5	5	13	3	15	7.5	1	11	11	4
9	Diabetes	10	2	9	6	14	3.5	13	3.5	1	5	11	15	8	7	12
10	Charity	9	13	1	12	7	10	3	4	2	8	11	14	6	5	15
11	Ducan MBA	1	5	13	2	15	8	3	9	14	4	10	12	6	7	11
12	MSU GPA	8	6	10	1	12	2	13	3	11	5	9	14	4	7	15
13	Forest fire	7	5	2	6	13	4	8	10	9	3	1	14	11	12	15
14	Housing price	12	9	2	7	4	11	6	9	3	13	15	1	9	14	5
15	Insurance	12	6.5	1	9	13	6.5	5	10	2	11	15	3	8	14	4
16	Istanbuck Stock	1	8	12	2	14	4	11	3	9	10	5	13	7	6	15
17	Red wine quality	14	8.5	3	12	15	8.5	5	6	4	13	7	1	10	11	2
18	white wine quality	15	12	3	10	4	8	6	12	5	14	9	1	12	7	2
19	Yacht	14	6.5	3	5	10	9	4	8	1	15	11	2	6.5	12	13
20	Servo system	14	12	2	7	5	11	6	10	3	9	15	1	8	13	4
21	Mortality rate	14	9	4	7	5	13	6	9	1	11	15	2	9	12	3
22	Wage	8	5.5	12	4	14	1	10	5.5	9	3	11	15	2	7	13
	Mean of the median rank	**10.45**	**8.32**	**5.5**	**7.25**	**9.32**	**7.66**	**6.14**	**7.98**	**4.32**	**10.59**	**10.89**	**6.09**	**7.91**	**9.73**	**7.86**
	Over all	**13**	**10**	**2**	**5**	**11**	**6**	**4**	**9**	**1**	**14**	**15**	**3**	**8**	**12**	**7**

while Support Vector Machine (SVM) was ranked 15^{th} in predicting engine miles per gallon of cars from 1970s and 1980s. This means that SVM is not stable in its prediction across 100 replications. The std of the test error of random forest which was always ranked 1^{st} was later ranked 15^{th} on wage dataset (SN: 22). The mean of the std rank of each machine learning method was calculated for all the datasets and it can be seen that the Gradient Boosting Machine (GBM) was ranked 1^{st}.

4. Post-processing of predictive ranking of results on classification and regression datasets

So far we have created a battery of potential very rich scores for each of the learning machines considered. Indeed, using the datasets and the internal characteristics of attributes, we can view each learning machine as a sampling unit in some hypothetical population, and then seek to investigate if there might be some groupings among the learning machines. This makes sense in the presence of the plurality

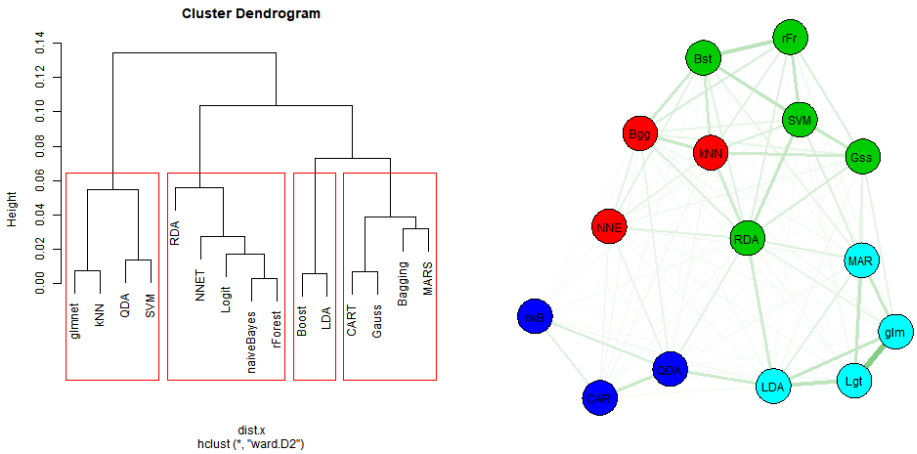
Table 12. The rank of the SD score of method M on regression data S.

SN	Datasets	BstLm	enet	GBM	glmnet	KNN	Lasso	MARS	MLR	nnet	PLS	Relaxo	rForest	Ridge	RLM	SVM
1	Powerplant	5	5	13.5	5	5	5	11	5	13.5	13.5	9.5	1	5	9.5	13.5
2	Abalone	13	10	4	12	6	8	2	10	1	14	15	5	10	7	3
3	Air foil	5	9	3	7	2	6	15	9	14	11	12	1	9	13	4
4	Attendance	4	2	11	3	9	7	6	1	15	10	14	13	5	8	12
5	Boston	14	9	1	11	7.5	12	4	7.5	2	13	6	3	10	15	5
6	Auto mpg	9	15	4	8	12	7	2.5	5.5	2.5	13	14	1	5.5	10	11
7	Computer hardware	6	2	12	8	14	7	13	3	9	4	5	10	1	11	15
8	Concrete	6	11.50	1	8	7	9	5	11.50	2	14	11.50	3	11.50	15	4
9	Diabetes	7	3	11	1.5	15	1.5	8.5	4.5	10	4.5	14	13	6	8.5	12
10	Charity	10	13	1	11	3	9	7	12	2	8	6	4	14	15	5
11	Ducan MBA	7	8	13	5.5	4	3	1	10	15	12	2	5.5	9	14	11
12	MSU GPA	6	5	10	1	12	2	15	7	11	3	9	14	8	4	13
13	Forest fire	8	10	2	7	6	12	3	15	11	1	4.50	4.50	14	13	9
14	Housing price	8	9	1	6.5	5	11	4	6.5	3	12	13	2	10	14	15
15	Insurance	1	5	8	3	15	2	10	5	11	7	13	9	5	14	12
16	Istanbuck Stock	6	3	12	7.5	14	7.5	11	1	10	9	5	13	4	2	15
17	Red wine quality	3.5	10.5	1	6	15	5	2	9	14	12.5	7	10.5	8	12.5	3.5
18	White wine quality	6	15	3	10	4	8.5	5	13	7	8.5	11	1	12	14	2
19	Yacht	13	6.5	1	8	11	9	4	5	3	12	15	2	6.5	10	14
20	Servo system	11	10	3	9	4	7	2	6	15	5	12	1	8	13	14
21	Mortality rate	13	9	3	11	5	15	6	9	1	12	7	2	9	14	4
22	Wage	6	6	4	8.5	12.5	2	12.5	10	3	6	1	15	8.5	11	14
	Mean of the std rank	**7.61**	**82**	**5.57**	**7.16**	**8.55**	**77**	**6.8**	**7.52**	**7.95**	**9.32**	**9.39**	**67**	**8.14**	**11.25**	**9.59**
	Overall rank	**7**	**9**	**1**	**5**	**11**	**4**	**3**	**6**	**8**	**12**	**13**	**2**	**10**	**15**	**14**

of learning machines, as one can understandably hypothesize that some learning machines are more similar than others. We pursued this idea of groups (clusters) of learning machines by performing cluster analysis on the learning machines used. We further visualized the results of our cluster analysis by plotting the mean rank of test errors using both the corresponding dendrogram and the corresponding graph-theoretic plot. The graph-theoretic plot was created and displayed with the finality of hopefully revealing any potential network structure among the learning machines. As shown in Figures 4, 5 and 6, the pattern of association among the learning methods appears to reveal that learning machines with similar overall predictive performances across all the datasets tended to fall within the same cluster. It would be very revealing to explore the discovered associations in greater details, perhaps to find out if methods can be clustered according to the foundational building blocks like kernel methods together, tree based methods together or ensemble methods together. That might require using different scores other than average ranks on the learning machines. Our work in this paper has contributed a little bit towards a more tangible grasp of the so-called No Free Lunch Theorem, and it is re-assuring to know that there is no one size fits all learning machine out there, at least not yet, and that the ingenuity of machine learning researchers is still in as greater a need as ever. So please go right ahead and finalize that great learning machine idea you had and bless us with a nice package so that we can add your learning machine to our data science arsenal.

5. Discussion and conclusion

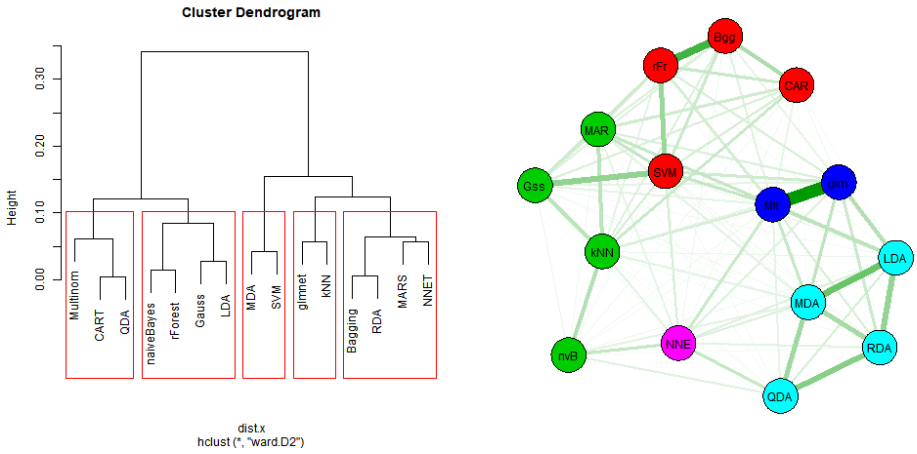
Many experts have contributed various instances of a result known as the no free lunch theorem (NFLT) in machine learning that says essentially that no functional representation (function space) along with its accompanying algorithms can yield the best prediction performance on all possible datasets. For any typical statistical machine learning approach to a typical data science task, one can and should consider a wide variety of possible and plausible function spaces to model the



(a) Dendrogram of methods.

(b) Graph theoretic plot of methods on datasets.

Figure 4. Clustering of methods on binary classification datasets.

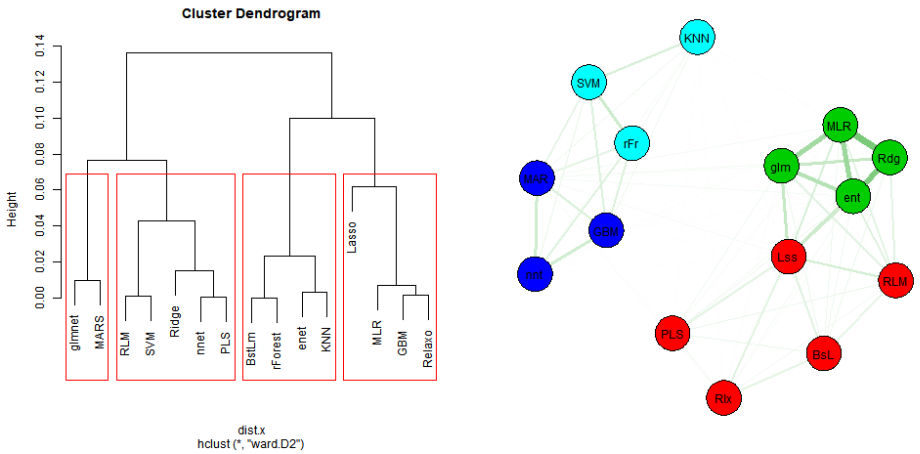


(a) Dendrogram of methods.

(b) Graph theoretic plot of methods on datasets

Figure 5. Clustering of methods on multi classification datasets.

functional pattern underlying the data. The theoretical result of NFLT though somewhat intuitive for some could be made intuitive by a substantial empirical demonstration on various datasets. The idea of running both parametric learning methods and nonparametric learning methods on different qualitative and quantitative benchmark datasets is to vividly show that NFLT is really true (see Tables



(a) Dendrogram of methods.

(b) Graph theoretic plot of methods on datasets.

Figure 6. Clustering of methods on regression datasets.

4–6, Tables 7–9 and Table 10–12, respectively). This is useful to data science practitioners as it serves as a recommendation to carefully consider as many function spaces and algorithms as possible for serious statistical machine learning tasks. In this paper, we consider a wide variety of datasets that are suitable for classification and regression learning problems (see Tables 1, 2 and 3) and we show tangible practical evidence that the NFLT is indeed valid. After all, commonsense seems to suggest that it would actually be very strange if a learning machine did indeed exist in practice that were always superior to all other learning machines on all possible data sets. Indeed, the existence of such a learning machine would apply the knowledge of the origin of reality, a knowledge we have not yet, and most likely are very far from attaining.

References

[1] E. Fokoué, *Foundational aspects of the theory of statistical function estimation and pattern recognition*, Bulletin of PFUR, Series Mathematics, Information Sciences, Physics **3** (2008), 40–54.

[2] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, *Mad skills: New analysis practices for big data*, Proceedings of the VLDB Endowment **2** (2009), 1481–1492.

[3] D. H. Wolpert, W. G. Macready, *No Free Lunch Theorems for Search*, Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

[4] D. H. Wolpert and W. G. Macready, *No free lunch theorems for optimization*, IEEE Transactions on Evolutionary Computation **1** (1997), 67–82.

[5] D. H. Wolpert, The lack of a priori distinctions between learning algorithms. Neural Computation **8** (1996), 1341–1390.

[6] J. Brownlee, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*, Machine Learning Mastery, 2016.

[7] J. Brownlee, *A Data-Driven Approach to Choosing Machine Learning Algorithms*, 2014, https://machinelearningmastery.com/a-data-driven-approach-to-machine-learning/ (accessed on December 18, 2018).
[8] S. Hussain, N. A. Dahan, F. M. Ba-Alwi and N. Ribata, *Educational data mining and analysis of students' academic performance using WEKA*, Indonesian Journal of Electrical Engineering and Computer Science **9** (2018), 447–459.
[9] I. C. Yeh and C. H. Lien, *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*, Expert Systems with Applications **36** (2009), 2473–2480.
[10] P. Cortez and A. J. R. Morais, *A data mining approach to predict forest fires using meteorological data*, in: J. Neves, M. F. Santos and J. Machado (eds.), New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 – Portuguese Conference on Artificial Intelligence, 2017, pp. 512–523.
[11] S. Shanthi and R. G. Ramani, *Classification of vehicle collision patterns in road accidents using data mining algorithms*, International Journal of Computer Applications **35** (2011), 30–37.
[12] J. M. Shea and K. H. Brown, *wooldridge: 111 Data Sets from "Introductory Econometrics: A Modern Approach, 6e" by Jeffrey M. Wooldridge*, 2018, R package version 1.3.1.
[13] H. Wickham, *tidyverse: Easily Install and Load the 'Tidyverse'*, R package version 1.2.1., R Core Team: Vienna, Austria, 2017.
[14] M. Kuhn, *Building predictive models in R using the caret package*, Journal of Statistical Software **28** (2008), 1–26.
[15] S. Prabhakaran, *Caret Package – A Practical Guide to Machine Learning in R*, 2018, https://www.machinelearningplus.com/machine-learning/caret-package/attachment/caret_package_a_practical_guide_to_machine_learning_in_r/ (accessed on December 12, 2018).
[16] D. Dua and C. Graff, *UCI Machine Learning Repository*, Irvine, CA, University of California, School of Information and Computer Science, 2019.

Ezekiel Adebayo Ogundepo, Data Science Nigeria (DSN), Lagos, Nigeria
e-mail: ogundepoezekiel@gmail.com

Ernest Fokoué, School of Mathematical Sciences, Rochester Institute of Technology, 98 Lomb Memorial Drive, Rochester, New York 14623
e-mail: epfeqa@rit.edu