Empirical Comparison and Sample Efficiency of Supervised Learning Algorithms

Advaith Ravishankar

April 2, 2025

Abstract

Empirical comparisons of supervised learning algorithms are crucial for understanding their practical performance across diverse datasets and guiding model selection. This study evaluates the predictive capabilities of Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LogReg), and XGBoost (Gradient-Boosted Trees) on datasets from the UCI Machine Learning Repository and investigates sample-efficiency. Building on the foundational work of Caruana and Niculescu-Mizil [2006], which analyzed algorithm performance on the Adult, Cover Type, and Letters datasets, this study introduces the Wine Quality dataset to assess model behavior with limited data. Additionally, the study expands the scope of prior research by exploring the impact of various train-test splits, including 5000-rest, 20-80, 50-50, and 80-20 partitions. The findings reaffirm the robustness of Random Forest and Gradient-Boosted Trees, outperforming other models across all datasets.

1 Introduction

Supervised learning algorithms are central to many machine learning applications, with their performance often evaluated based on empirical comparisons across diverse datasets. Such comparisons are crucial for guiding model selection, as algorithm performance can vary significantly depending on the data characteristics. The study by Caruana and Niculescu-Mizil [2006] provides a comprehensive analysis of algorithmic performance on benchmark datasets from the UCI Machine Learning Repository, including Adult, Cover Type, and Letters. Their findings highlighted Boosted Trees (BST-Tree) and Random Forest (RF) as robust performers, demonstrating their effectiveness across multiple datasets.

Building on this foundational work, we aim to re-evaluate the performance of modern supervised learning algorithms while addressing an important gap in the original study: the exploration of different train and test splits. To this end, the study looks at original split of 5000 train and rest in test and extends

it by adding [20%, 80%],[50%, 50%], and [80%, 20%] train and test splits. The study evaluates, Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LogReg) and XGBoost (BST-Tree), a more recent gradient-boosting method known for its high accuracy and scalability.

The study not only extends the dataset scope with a small dataset (Wine Quality) but also revisits the conclusions of Caruana and Niculescu-Mizil in light of advancements in machine learning algorithms. By systematically evaluating these models across multiple datasets, the study aims to provide updated insights into their performance trends and practical applicability.

2 Datasets

This study evaluates the performance of supervised learning algorithms on four datasets from the UCI Machine Learning Repository: Adult, Cover Type, Letters, and Wine Quality. Each dataset has unique characteristics that allow us to assess the models' capabilities across a range of data complexities, feature types, and dataset sizes. Below, a brief description of each dataset and how it is used in a classification task is presented. For the sake of comparison, the study adopts the same task as in Caruana and Niculescu-Mizil [2006] to convert the classification from categorical to binary.

2.1 Adult Dataset

2.1.1 Description

The Adult dataset, Becker [1996], contains 48,842 instances and 14 features which are categorical (workclass, education, marital-status, occupation, realtionship, race, sex, native-country, income) and numerical (age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week)

2.1.2 Classification Task

The task is to predict whether an individual's income exceeds $50,000 per year based on demographic and employment-related attributes. The income is the outcome and is a binray class with values >50K, <=50K. The study maps >50k as the positive label (+1) and <=50k as the negative label (0).

2.2 Cover Type Dataset

2.2.1 Description

The Cover Type, Blackard [1998]. dataset consists of 581,012 instances and 54 features derived from cartographic variables such as elevation, aspect, slope, 4 relative distance metrics, 3 shade levels at time points (binary), 4 wilderness area data points (binary), 40 soil types (binary), and the label Cover Type

(Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, Krummholz).

2.2.2 Classification Task

The task is to predict the most common cover type (Pine) as the postive label (+1) and everything else is negative (0).

2.3 Letter Recognition Dataset

2.3.1 Description

The Letter Recognition dataset, Slate [1991], contains 20,000 instances with 16 features derived from pixel-level data. The features are (x-box, y-box, width, hight, onpix, x-bar, y-bar, x2xbar, y2bar, xybar, x2ybr, xy2br, x-ege, zegvy, y-ege, yegvx) representing the statistics from the image with label letter (A-Z)

2.3.2 Classification Task

The study adopts LETTER.p2 classification from Caruana and Niculescu-Mizil [2006] which labels A-M as positive (+1) and N-Z as negative (0)

2.4 Wine Quality Dataset

2.4.1 Description

The Wine Quality dataset, Cortez [2009], contains 4,898 instances and 11 features which are fixed aciditiy, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfer dioxde, desnity, pH, sulphates, alcohol content, color and the target quality (0 to 10).

2.4.2 Classification Task

For the classification task, the study focuses on predicting high quality wine (>=7) as positive and the rest as negative samples (0). The reason for this split is to ensure an equal distribution of wine in both samples for class balance.

3 Problem Description

For each dataset, the study will predict to distinguish between the positive and negative samples as described in the datasets section in a supervised training fashion with a Simple Vector Machine (SVM), Random Forest Model (RF), Logistic Regression (logReg), XGBoost (Gradient boosted Tree or GBTree) and K Nereast Neigbours Classier (KNN).

These methods cover 5 distinct types of classifier which are discriminative model (SVM), probabilistic model (logReg), ensemble Tree based models (RF),

ensemble booseted Tree based models (GBTree) and Lazy Learners (KNNs).

NOTE: The inclusion for XGBoost was to compare current state-of-the art naive modeling technique with the BST-DT in Caruana and Niculescu-Mizil [2006].

4 Method

4.1 Setup

For each dataset (4 sets), the study trains each model (5 models) on 4 different splits:

1. 5000 train and Rest in test (based on sampling strategy in literature)
2. 20% train and 80% test
3. 50% train and 50% test
4. 80% train and 20% test

For each iteration (a dataset for a model with a train test split), the study perform a 5 fold cross validation on the train set to choose the ideal parameters for the model. With the best performing model, the accuracy score on the train and test set is reported.

4.2 Parameter Tuning

To select each parameter, GridSearch was performed. Due to runtime constraints the following limited parameters were tuned for each model:

SVM: the regularization parameter C was explored with values [1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1, 10] and kernel was set t o "rbf". Scikit learn was used as the library with the fixed model parameters probability=True and random_state=42

RF: the number of estimators was explored with [100, 200, 300, 400, 500, 600, 700] and max_depth was explored with [10, 20]. Scikit learn was used as the library with the fixed model parameters random_state=42.

LogReg: the regularization parameter C was explored with values [1e-8, 1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1, 10, 100, 1000, 10000]. Scikit learn was used as the library with the fixed model parameters max_iter=1000, random_state=42.

GBTree: the number of estimators were explored with [100, 200], learning rate [0.01, 0.1, 0.2] and max_depth [3, 6, 9] was explored. XGBoost was used as the library with the fixed model parameters eval_metric='logloss, random_state=42.

KNN: the number of neighbors were explored with [2, 3, 4, 5, 6, 7, 8, 9, 10], and weights ["uniform", "distance"]. Scikit learn was used as the library.

This led to 4*5*4=80 reported results on train and test (total of 160). The runtime for this on datahub's largest server with parallel processing was around 52 hours.

5 Experiments

5.1 Overview

Table 1 has the reported scores for all the tests mentioned in the method section.

As the prior paper only has results on the test set reported for Letter, Adult and Cover on the 5000+Rest partition, the rest of the non study's tests are left blank. One major issue the study encountered was runtime with the Cover Type Dataset for the [20%, 80%], [50%, 50%], and [80%, 20%] partitions as the 5-fold cross validation on a full sized dataset of 581,012 instances and 54 features did not converge with 10 hours of training for the smallest partition of [20%, 80%]. Due to this computation constraint, the results for the rest of the partitions are left blank.

As the wine dataset only has 4989 instances, the first parition 5000-Rest does not have a train dataset. For this reason, only the train performance is reported.

Table 1: Model Results on Each Dataset

Dataset	Model	5000 + Rest		20% + 80%		50% + 50%		80% + 20%	
		Train	Test	Train	Test	Train	Test	Train	Test
Letter	SVM	-	0.954	-	-	-	-	-	-
	SVM (Ours)	0.914	0.921	0.911	0.918	0.935	0.940	0.943	0.949
	LogReg	-	0.446	-	-	-	-	-	-
	LogReg (Ours)	0.723	0.724	0.726	0.725	0.726	0.729	0.725	0.733
	RF	-	0.935	-	-	-	-	-	-
	RF (Ours)	0.939	0.946	0.931	0.943	0.958	0.965	0.971	0.976
	BST-DT	-	**0.976**	-	-	-	-	-	-
	GBTree (Ours)	0.950	0.950	0.935	0.948	0.963	0.968	0.973	**0.981**
	KNN	-	0.937	-	-	-	-	-	-
	KNN (Ours)	**0.955**	0.951	**0.941**	**0.948**	**0.967**	**0.969**	**0.974**	0.980
Wine	SVM	-	-	-	-	-	-	-	-
	SVM (Ours)	0.803	-	0.803	0.803	0.803	0.803	0.803	0.803
	LogReg	-	-	-	-	-	-	-	-
	LogReg (Ours)	0.817	-	0.820	0.816	0.817	0.819	0.817	0.820
	RF	-	-	-	-	-	-	-	-
	RF (Ours)	**0.884**	-	**0.870**	**0.868**	**0.885**	**0.873**	**0.885**	**0.885**
	BST-DT	-	-	-	-	-	-	-	-
	GBTree (Ours)	0.876	-	0.867	0.865	0.872	0.868	0.876	0.880
	KNN	-	-	-	-	-	-	-	-
	KNN (Ours)	0.848	-	0.835	0.836	0.841	0.838	0.850	0.859
Adult	SVM	-	0.886	-	-	-	-	-	-
	SVM (Ours)	0.838	0.840	0.845	0.846	0.840	0.842	0.847	0.848
	LogReg	-	0.886	-	-	-	-	-	-
	LogReg (Ours)	0.836	0.839	0.840	0.840	0.844	0.840	0.842	0.843
	RF	-	**0.934**	-	-	-	-	-	-
	RF (Ours)	0.861	0.857	0.861	0.862	0.855	0.859	**0.861**	0.861
	BST-DT	-	0.865	-	-	-	-	-	-
	GBTree (Ours)	**0.865**	0.862	**0.863**	**0.865**	**0.858**	**0.861**	0.8603	**0.863**
	KNN	-	0.785	-	-	-	-	-	-
	KNN (Ours)	0.840	0.845	0.846	0.846	0.842	0.846	0.848	0.849
Cover	SVM	-	0.765	-	-	-	-	-	-
	SVM (Ours)	0.730	0.725	-	-	-	-	-	-
	LogReg	-	0.625	-	-	-	-	-	-
	LogReg (Ours)	0.719	0.711	-	-	-	-	-	-
	RF	-	0.876	-	-	-	-	-	-
	RF (Ours)	**0.814**	0.812	-	-	-	-	-	-
	BST-DT	-	**0.938**	-	-	-	-	-	-
	GBTree (Ours)	0.823	0.819	-	-	-	-	-	-
	KNN	-	0.780	-	-	-	-	-	-
	KNN (Ours)	0.774	0.782	-	-	-	-	-	-

5.2 Analysis

Letter: For the letters dataset, the BST-DT from the comparison literature has the strongest score in the first partition with 0.976 accuracy, outperforming our GBTree by 0.026. The net difference between our implementation and the comparison paper is between 0.01 and 0.03 (excluding Logistic Regression) showcasing the study effectively reproduces the results. For the 20-80 train test split, 50-50 train test split, our best performing model was both a KNN (0.948, 0.969) and GBTree (0.948, 0.968) which still meet similar learning thresholds for BST-DT. However, on the 80-20, our GBTree and KNN outperform the papers work with scores of 0.981 and 0.980. Therefore, for the first dataset, GBTree and KNN are the strongest evaluators

Wine: As the prior work does not evaluate the wine dataset, our models are evaluated alone. For all partitions (ignoring the 5000-Rest), the Random Forest Model performs the strongest with 0.868 0.873 0.885 followed by the GBTree with 0.865, 0.868, 0.880.

Adult: For the adult dataset the best performing model is the random forest model with accuracy of 0.934 with our best models being GBTree (0.862) and Random forest (0.857), The gap between most of our models is 0.003-0.046 but for our KNN, our study performed 0.06 points better and our RF perfomed 0.077 worse. This divergence is due to our the differences in libraries as the study relies on sklearn.

Cover Type: For the cover type dataset, the best performing model from the literature is BST-DT with a score of 0.938. The best two works from our side are GBTree (0.819) and Random Forest (0.812). For KNN and SVM, our implementation performs similarly to the previous work and our Logistic regression performs 0.084 better. However, There is gap of 0.118 and 0.064 between the study and the literature performance Boosted tree For the boosted tree and RF. These divergences are also due to the differences in libraries as the study relies on sklearn.

(a) Sample Efficiency - Adult

(b) Sample Efficiency - Wine

(c) Sample Efficiency - Letter

Figure 1: Sample Efficiency Curves

5.3 Sample Efficiency

From the sample efficient curves we notice a consistent increase in performance as the sample increase for letter and wine but no consistent trend in adults dataset. As the Adult dataset is the largest, this indicates to us that sample efficiency starts plateauing for naive supervised models between 10,000 and 50,000 samples. However there are no conclusive results present.

5.4 Limitations and Improvements

One limitation the study encountered was computational resources. When training the Cover Type model on anything more than 20% of the dataset, the training failed to converge due to limited resources (8GB CPU). The study attempted to use parallel processing but this also did not yield to results due to the 5 fold cross validation. An improvement is to have more resources at disposal to test results.

The second limitation was the number of train-test partitions that were used. With only 4 partitions, the study could not construct a conclusive curve for sample efficiency. To get a better idea of the performance, taking more train-test partitions with every 5% interval will enable us to construct conclusive curves.

Finally, as the study only looked at 4 datasets with gaps in size (4,898, 20,00, 48,842 and 581,012), it is not indicative of generalization as it missing datasets between 50k-500k and >600k. By adding more datasets, a notion of generalization can be obtained.

6 Conclusion

The findings reveal that Gradient-Boosted Trees and Random Forest consistently rank among the top performers, demonstrating their robustness and scalability across different data partitions which matches the performances reported in Caruana and Niculescu-Mizil [2006] for Adult, Letter and Cover datasets. These

results are also validated by the additional dataset of Wine Quality

The sample efficiency analysis did not yield to conclusive results due to the 4 interval sample of 5000-Rest, 20-80, 50-50 and 80-20 train test splits. By increasing the intervals, one will be able to obtain a stronger notion of which train test splits lead to better performance.

In conclusion, this study not only validates prior findings but also highlights the continued relevance of Random Forest and Gradient-Boosted Trees in modern supervised learning tasks. Due to the works' shortfall in sample-efficiency, further research needs to be done to identify a strong train test split.

7 Bonus Points

The study investigates 5 models on 4 datasets and showcases sample-efficiency curves which highlights the depth of the study. One of the models was XGBoost which is a state-of-the-art boosting algorithm, highlighting the explorations complexity.

The study also performs a relative comparison with the literature, further validating the results. Finally, as parallel computing was attempted, it showcases the attempt to solve the runtime issues, further highlighting technical complexity.

References

R Becker, B. Kohavi. Adult [dataset], 1996. URL `https://doi.org/10.24432/C5XW20`.

Jock Blackard. Covertype [dataset], 1998. URL `https://doi.org/10.24432/C50K5N`.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, pages 161–168. ACM, 2006. doi: 10.1145/1143844.1143865. URL `https://doi.org/10.1145/1143844.1143865`.

Cerdeira A. Almeida F. Matos T. Reis J Cortez, P. Wine quality [dataset], 2009. URL `https://doi.org/10.24432/C56S3T`.

David Slate. Letter recognition [dataset], 1991. URL `https://doi.org/10.24432/C5ZP40`.