**Supervised Machine Learning Algorithms Analysis**

**Jungsoo Lee** [jul273@ucsd.edu](mailto:jul273@ucsd.edu)

*University of California*

*San Diego, CA 92023, USA*

**Abstract**

This paper is to replicate 2006 paper by Caruana and Niculescu-Mizil, focusing on analysis of performance of several supervised machine learning algorithms. The algorithms are performed on various datasets through multiple trials with a range of hyperparameter search. All of the datasets are binary classification problems. The performance of each algorithm is measured with various performance metrics. Additionally, the effect of training sampling size and the time complexity of performance are analyzed, as well. The results are investigated in detailed level offering multiple tables and appendices in between texts.

1. **Introduction**

As numerous machine learning algorithms have been developed since the mid 20th century, performance of each algorithm varies greatly depending on its principle and characteristics. The CNM06 explores ten supervised machine learning algorithms on eleven binary classification problems with eight performance metrics. While this paper is intended to use the same methodology, it is a small version exploring three algorithms on four binary classification problems with five performance metrics. The goal of this paper is to compare the performance of algorithms at a detailed level and analyze them.

1. **Method**

**2.1. Learning Algorithms**

Three different classifiers used in this paper: Logistic Regression, Random Forest, and Naive Bayes. All of the classifiers are used with varying parameters. MinMaxScaler from sklearn is used to normalize the attributes in all datasets, although Random Forest and Naive Bayes have no need of.

Logistic Regression predicts the probability of an event using a logistic function. Since it predicts the probability, its output values range from 0 to 1. The model is trained with ridge parameters ranging from 10^-8 to 10^4.

Random Forest is an ensemble method of decision trees. The model is trained with parameters of varying maximum depth of 5, 6, 7, and None.

Naive Bayes is based on Bayes’ theorem, considering each feature independently to predict the probability. The model is trained with parameters of varying variance smoothing of np.logspace(0, -9, num=5).

**2.2. Performance Metric**

The algorithms are measured in five different performance metrics: accuracy (ACC), F-score (FSC), precision (PRC), recall (RCL), and the area under the ROC curve (AUC). Overall, the models performed well having most of the metrics from 0.70 ~ 0.99.

ACC is the ratio of correct predictions to all predictions. It is a straightforward and intuitive metric which the cost of misclassification is especially critical for heavily imbalanced datasets. FSC is the weighted average of the PRC and RCL. It may be a better model than ACC but it is less interpretable because it is a metric PRC and RCL combined. PRC is the ratio of correct positive predictions to all positive predictions. It is preferred when false positive is crucial. RCL is the ratio of correct positive predictions to all real positives. It is preferred when false negative is crucial. AUC, the most widely used metric, is the area under the curve where the false positive rate and true positive rate meet on the graph.

**2.3. Datasets**

All of the three datasets are binary classification problems and collected from UCI Machine Learning Repository.

The purpose of ADULT dataset is to predict the income of an individual using fourteen attributes such as age, workclass, and education. Individuals with income greater than 50K are labeled as a positive class and among 30162 observations, 25% of the dataset represents a positive class.

The purpose of OCCUPANCY dataset is to predict the occupancy of an office using six attributes such as light, temperature, and humidity. Offices with occupied status are labeled as a positive class and among 8143 observations, 21% of the dataset represents a positive class.

The purpose of CREDIT CARD dataset is to predict probability of default of a client using twenty attributes such as amount of credit, gender and education. Clients with default payment are labeled as positive class and among 30000 observations, 22% of the dataset represents positive class.

The purpose of MAGIC dataset is to predict the image of primary gamma/ hadronic showers using ten attributes such as number of major/minor axis of ellipse and sum of content of all pixels. Images of primary gamma are labeled as a positive class and among 19020 observations, 65% of the dataset represents positive class. For technical reasons, the number of negative class is underestimated, but in real data, it represents the majority of the observations.[2]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 1. Description of Problems | | | | |
| Problems | # ATTR | TRAIN SIZE | TEST SIZE | % POZ |
| ADULT  OCCUPANCY  CREDIT CARD  MAGIC | 14/105  6  23  10 | 5000  5000  5000  5000 | 25162  3143  25000  14020 | 25%  21%  22%  65% |
|
|
|

1. **Experiment**

All of the factors are equally used for every algorithm except for parameters, since each algorithm has different parameters. 5000 of training samples are randomly collected from each dataset and all of the rest of the dataset are used as test sets. Through grid search, the optimal hyperparameter is selected. This setting is kept the same for five trials.

1. **Discussion**

The results of performance by algorithms are shown below in two tables. Table 2 shows the mean test set performance by algorithm/ data set combo and the best performance for each dataset is indicated as bold font. Table 3 shows the test set performance by algorithm/ metric combo and the best performance for each metric is indicated as bold font. \* indicates a performance that is not statistically significant from the best performance and it is investigated through t-tests. The threshold for t-test is set p=0.05 as CNM05 indicates.

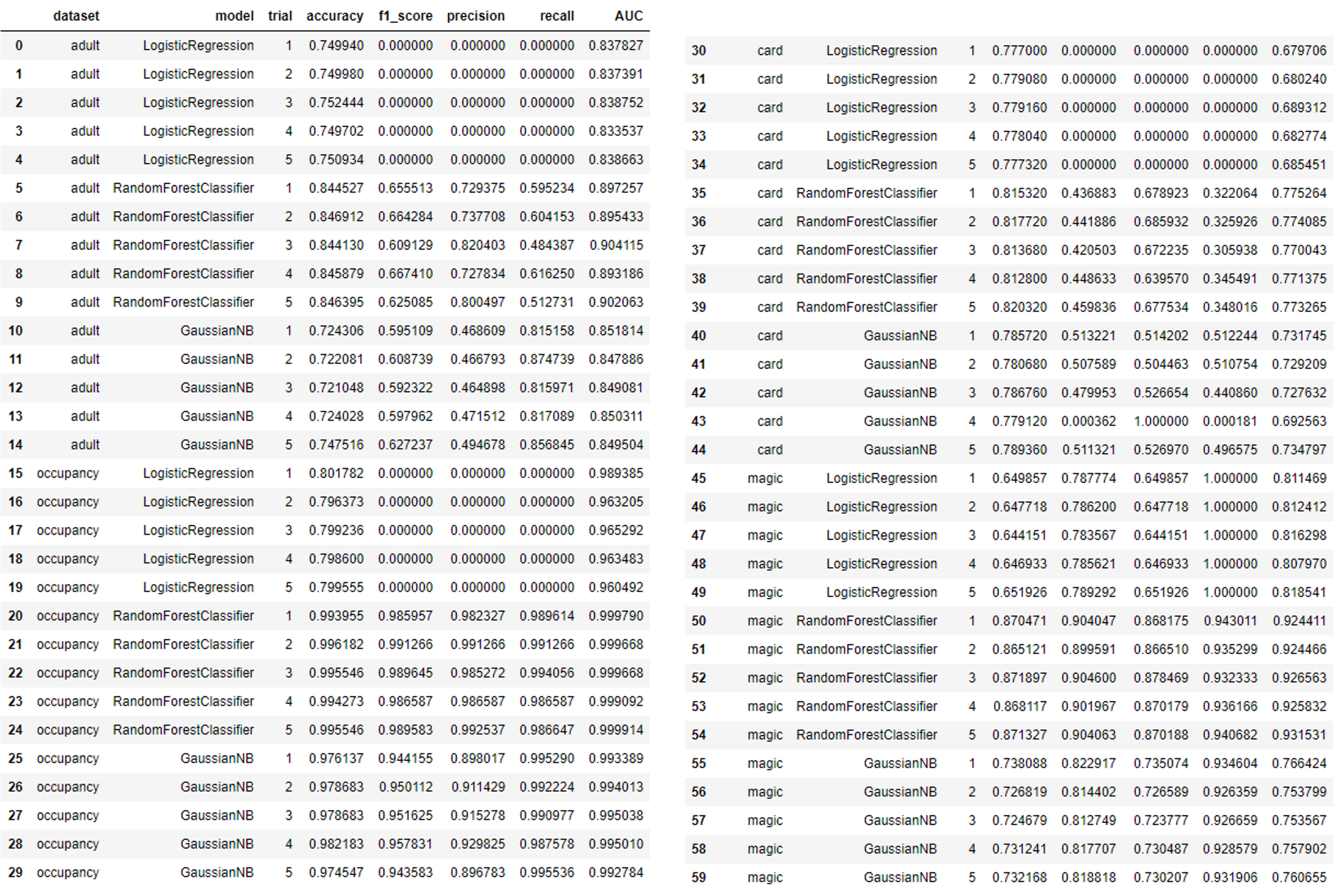
Overall, all of the algorithms perform relatively well on the datasets. The mean of test set performance of the Random Forest algorithm is the highest in ADULT, OCCUPANCY, and MAGIC datasets, but it is not significantly different from the highest in CREDIT CARD dataset, as well. The mean of test set performance of the Random Forest algorithm is the highest in every performance metrics except for RCL, in which Gaussian Naive Bayes is the highest.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 2. Test Set Performance Algorithm/ Data Set Combo | | | | |
| MODEL | ADULT | OCCUPANCY | CREDIT CARD | MAGIC |
| LogisticRegression  RandomForestClassifier  GaussianNB | 0.679876\*  **0.743908**  0.699049\* | 0.700777  **0.991897**  0.964431 | 0.379210  0.607837\*  **0.611524** | 0.809530  **0.902381**  0.793949 |
|
|

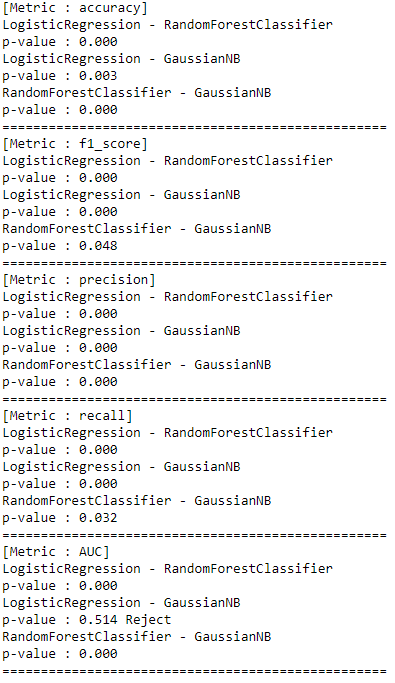
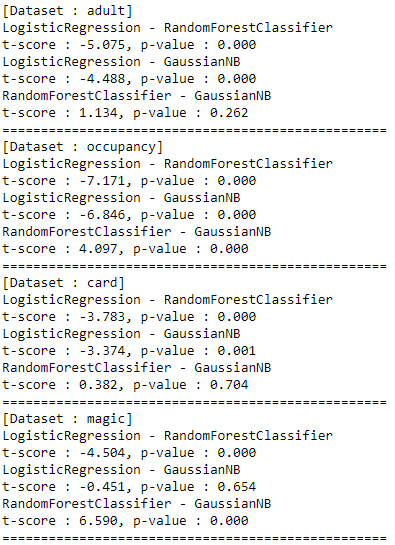
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3. Test Set Performance Algorithm/ Data Set Combo. | | | | | |
| Model | ACC | FSC | PRC | RCL | AUC |
| LogisticRegression  RandomForestClassifier  GaussianNB | 0.794524  **0.882063**  0.806772 | 0.456736  **0.746023**  0.720590\* | 0.703023  **0.821220**  0.661287 | 0.422209  0.708413  **0.812775** | 0.835249  **0.899811**  0.834768 |
|
|

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Appendix 1. Training Set Performance Algorithm/ Data Set Combo | | | | |
| MODEL | ADULT | OCCUPANCY | CREDIT CARD | MAGIC |
| LogisticRegression  RandomForestClassifier  GaussianNB | 0.682557  **0.949233**  0.6995912 | 0.7034434  **1.00000**  0.966424 | 0.3992838  **0.702199**  0.6119244 | 0.8099464  **1.00000**  0.7929318 |
|
|

Appendix 2. Raw Test Scores



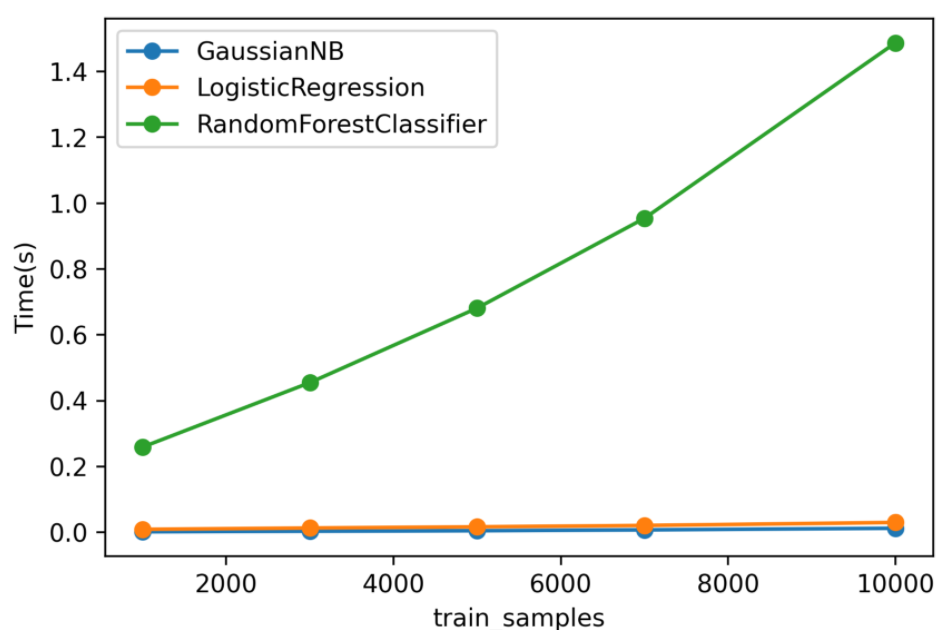
Appendix 3. P-values of the Comparisons across algorithms



1. **Bonus**

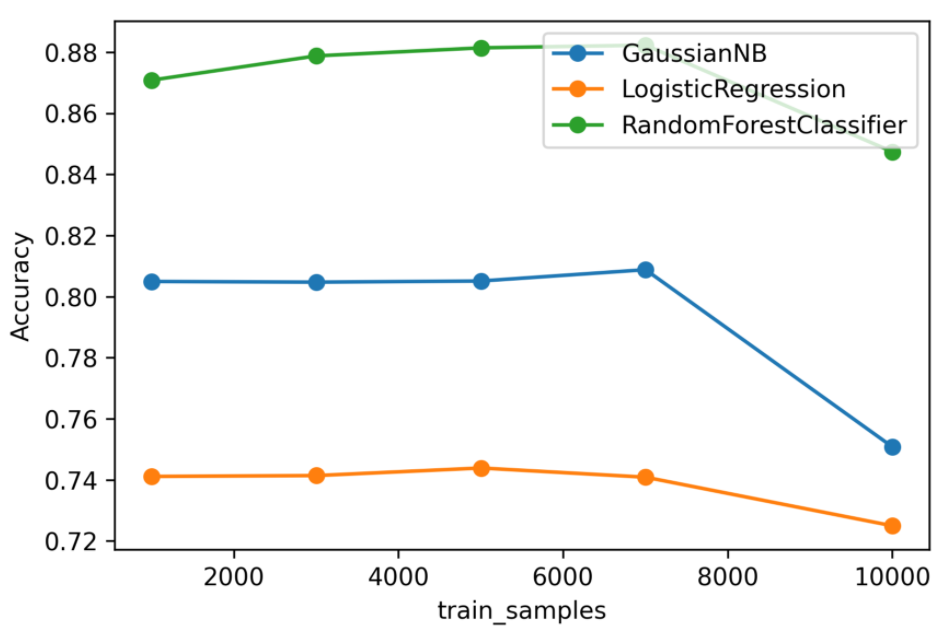
Appendix below is the time complexity curve of each algorithm varying the number of training samples. The Random Forest model has the longest training time and it increases the most drastically when the size of the training sample increases.

Appendix 4. Time Complexity Curve of Algorithms varying the Number of Training Samples



Appendix below is a learning curve that compares the test set performance for the best hyperparameter choice as varying the number of training samples. The performance for every three algorithms increases until the training sample size reaches ~7000, but significantly decreases after that.

Appendix 5. Learning Curve of Algorithm varying the Number of Training Samples

****

1. **Reference**

[1] Rich Caruana & Alexandru Niculescu-Mizil. (2006) An Empirical Comparison of Supervised Learning Algorithms. Ithaca, NY: Cornell University.

[2] UCI Machine Learning Data Repository. archive.ics.uci.edu/ml. Irvine, CA. University of California, Irvine.

[3] scikit-learn: Machine Learning in Python. scikit-learn.org.

[4] tensorflow:An open-source machine learning framework for everyone. tensorflow.org.