Battle of Classifiers: Comparative Study of Different Classifier Models

Junghwan Kim
1212/21/24
junghwk11@gmail.com

Abstract

This project aims to evaluate and compare the performance of three distinct supervised learning classifiers—Support Vector Machines (SVM), Boosted Trees, and Neural Networks—on three datasets—Mushroom, Wine, and Income datasets—sourced from the UCI Machine Learning Repository. In order to evaluate the performance of each model, the study utilizes k-fold cross validation and test accuracy to ensure the effectiveness of each model on each datasets. The results are analyzed based on classification accuracy, highlighting the strengths and weaknesses of each model in different scenarios. The study has shown that the performance of the Boosted Trees model has shown a slight better performance than SVM and Neural Nets.

1. Introduction

In recent years, the field of machine learning has witnessed significant advancements, particularly in the development and application of supervised learning models. Among these, Support Vector Machines (SVM), Boosted Trees, and Neural Networks have emerged as popular choices due to their ability to handle complex classification tasks with varying degrees of success. This project seeks to explore the comparative performance of these three classifiers on a selection of datasets from the UCI Machine Learning Repository.

The primary objective is to assess how well each classifier performs in two-class classification problems, a fundamental task in machine learning that has applications across numerous domains such as finance, healthcare, and marketing. By employing cross-validation techniques for each model, this study aims to provide a comprehensive evaluation of their capabilities.

The datasets chosen represent a range of challenges, including variations in size, feature types, and class distributions. This diversity allows for a thorough examination of each model's adaptability and robustness. Additionally, by analyzing results across different data partitions,

this project seeks to identify trends in classifier performance relative to training data size and complexity.

Ultimately, this research contributes valuable insights into the practical utility of SVMs, Boosted Trees, and Neural Networks, guiding practitioners in selecting appropriate models for specific classification tasks.

2. Data

2.1 Mushroom Dataset

The Mushroom dataset consists of 8124 instances. It includes hypothetical samples of 23 species of gilled mushrooms in the Agaricus and Lepiota Family. With 22 categorical features, each of the species are mainly identified through whether it is edible, poisonous, not sure, and not recommended.

2.2 Wine Data set

The Wine Quality dataset consists of 4898 instances. It includes 11 features such as fixed acidity, volatile_acidity, density, pH, and more. The dataset is much suitable with multiple-class classification.

2.3 Income Dataset

The Income Dataset consists of 48842 instances. It includes 14 features that are either categorical or integer types.

3. Problem Statement

The primary objective of this research is to evaluate and compare the performance of three supervised classification models–Support Vector Machine (SVM), Boosted Trees, and Neural Nets–across three different datasets with varying characteristics. The study aims to identify which model performs better overall with respect to different data and different test/train partitions. Moreover, the study aims to highlight the importance of the data in model performance.

4. Method Description

The methodology of this study follows the nature of datasets and models. For each model, the study uses the same functions in deploying the model, evaluating the model, and KFold Cross validation. While for each data, the study uses the same data loading and preprocessing step for SVM models and Boosted Trees model. Due to complexity in coding and time constraints, the Neural Nets have slightly modified code for preprocessing. However, they are very similar with the SVM and Boosted Trees.

For data loading and preprocessing for the income data, the SVM model for Adult Census Income data took a long time to classify, which led to a halt in the study. It took thirty minutes to train, test, and evaluate the model for one partition. Due to the size of the data, feature scaling and dimensionality reduction with Principal Component Analysis had to be performed. After more rigorous preprocessing of the income data, the SVM model was able to classify the partitions in less than three minutes. For this reason, while SVM and Boosted Trees models have PCA in their preprocessing step, the Neural Net does not.

Although there were some limitations with data loading and preprocessing, the codes are as uniform as they can be. When it comes to writing the codes for data preprocessing and model deployment, reusing the codes for each model and data really improved the efficiency of the overall project. Moreover, it ensured much needed fairness with the model performances.

5. Experiments

The tables below show the performance of each model in the datasets. The first table for each model is the testing accuracy. The second table for each section is the K-Fold Cross Validation Training accuracy. The third table shows K-Fold Cross Validation, Validation accuracy.

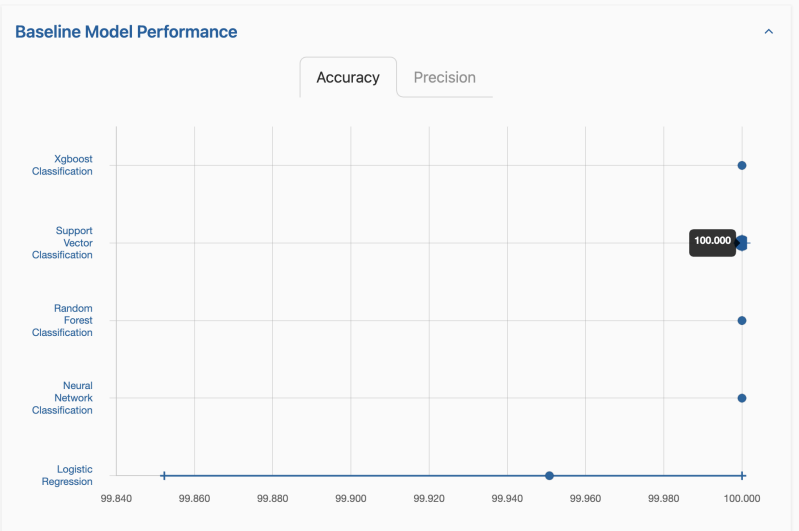
5.1 SVM

	Mushroom	Wine Quality	Adult Census Income
80/20 Split	100.00%	55.94%	85.03%
50/50 Split	100.00%	56.12%	84.82%
20/80 Split	99.88%	55.16%	84.46%
Averages of Splits	99.96%	55.74%	84.77%

	Mushroom	Wine Quality	Adult Census Income
80/20 KFold Training	100.00%	59.89%	85.22%
50/50 KFold Training	100.00%	60.73%	85.57%
20/80 KFold Training	100.00%	62.14%	85.77%
Averages KFold Training	100.00%	60.92%	85.52%

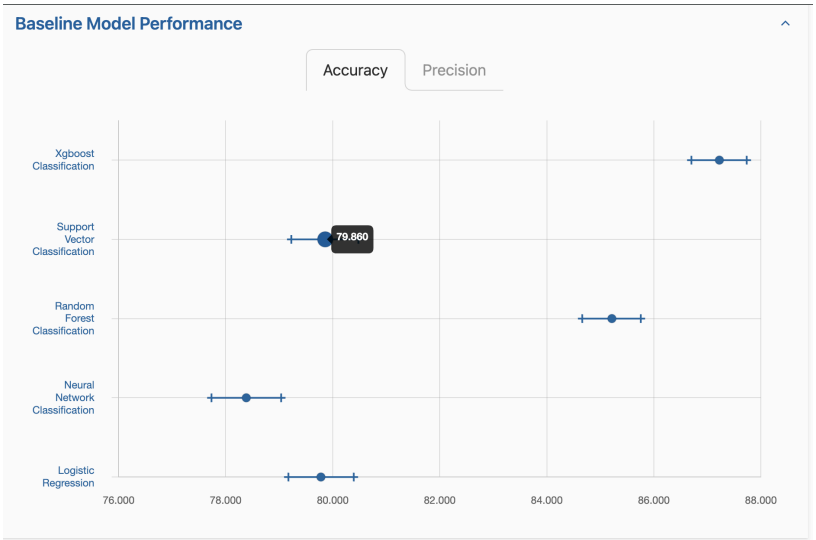
	Mushroom	Wine Quality	Adult Census Income
80/20 KFold Validation	100.00%	58.48%	85.02%
50/50 KFold Validation	100.00%	58.70%	85.25%
20/80 KFold Validation	99.94%	51.71%	85.07%
Averages KFold Validation	99.98%	56.30%	85.11%

For Mushroom sata, the result corresponds with the baseline model performance given by the UCI Machine Learning Repository. Though it might seem concerning to have near 100% accuracy, baseline model performance shows that it is the nature of the data rather than overfitting in the model.



For Wine Quality data, SVM model's average testing accuracy was 55.74%. This is similar to Paulo Cortez's paper, where SVM model testing accuracy was measured. In the paper, the SVM model had a testing accuracy of 62.4% for red wine and 52.6% for white wine in a 50/50 split. Looking at this paper's SVM's 50/50 split, it had 56.12%, which is very close to the average of red wine and white wine testing accuracy of Cortez et al paper.

For Adult Census Income data, SVM model had testing accuracy of 84.77%. Compared to the UCI Machine Learning Repository's baseline model performance of 79.86%, it is about 5% higher.



5.2 Boosted Trees

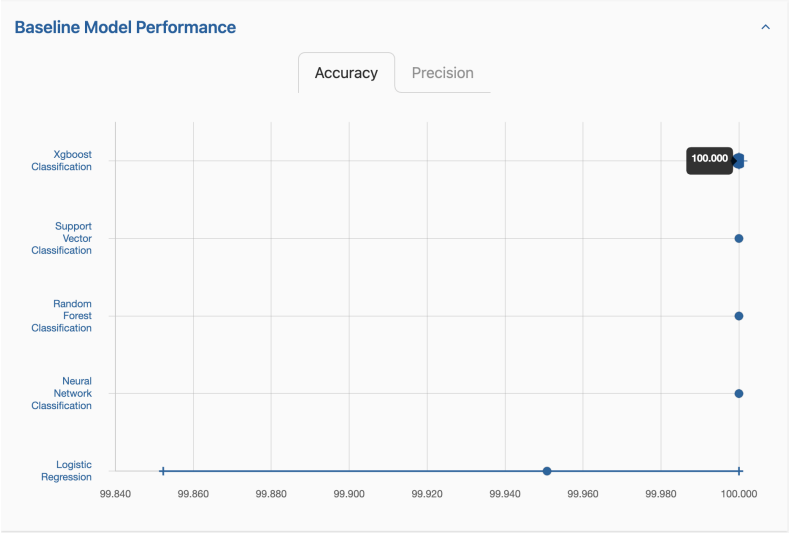
	Mushroom	Wine Quality	Adult Census Income
80/20 Split	100.00%	69.69%	87.21%
50/50 Split	100.00%	62.88%	86.51%
20/80 Split	100.00%	57.11%	85.87%
Averages of Splits	100.00%	63.23%	86.53%

	Mushroom	Wine Quality	Adult Census Income
80/20 KFold Training	100.00%	100.00%	90.96%
50/50 KFold Training	100.00%	100.00%	92.24%

20/80 KFold Training	100.00%	100.00%	95.28%
Averages KFold Training	100.00%	100.00%	92.83%

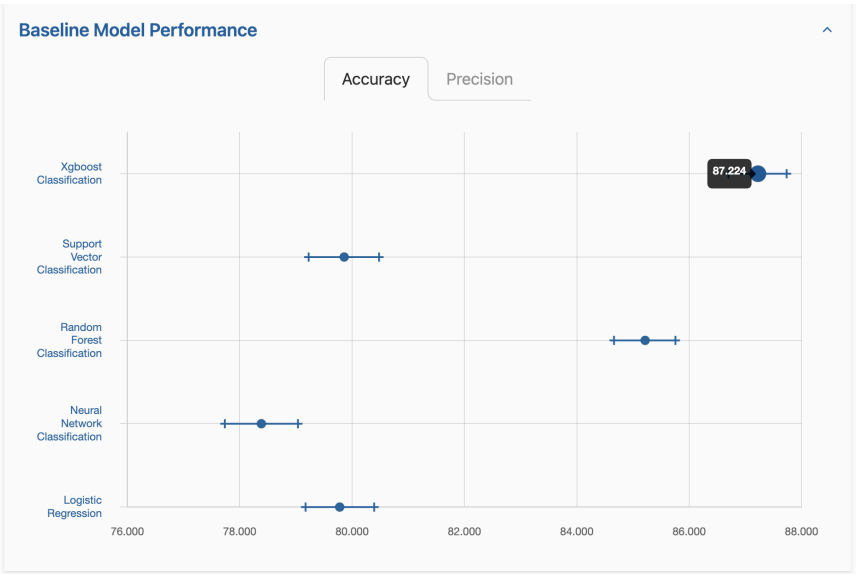
	Mushroom	Wine Quality	Adult Census Income
80/20 KFold Validation	100.00%	64.35%	86.90%
50/50 KFold Validation	100.00%	61.46%	86.76%
20/80 KFold Validation	99.94%	58.29%	85.50%
Averages KFold Validation	99.98%	61.36%	86.39%

For Mushroom data, Boosted Trees' 100% testing accuracy corresponds to the baseline model performance given by the UCI Machine Learning performance. Compared to the SVM model's performance above, both models seem to be extremely strong with the mushroom dataset.



For wine quality data, Boosted Trees' 63.23% testing accuracy was the highest out of the three models. Unfortunately, there are no comparisons found between the Boosted Trees' accuracy and selected references of this paper, but as mentioned in Caruana and Niculescu-Mizil, Boosted Trees model had the best performance in the wine data.

For Adulti Census Income data, Boosted Trees model had a testing accuracy of 86.53%. Compared to UCI Machine Learning Repository's baseline model performance of 87.224%, it was very close to the intended performance. Again, the Boosted Trees model has shown that compared to other two models, it has higher accuracy.



5.3 Neural Nets

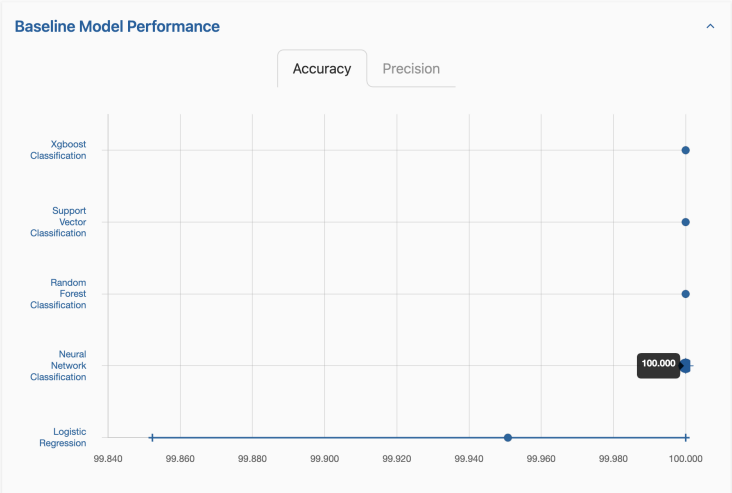
	Mushroom	Wine Quality	Adult Census Income
80/20 Split	100.00%	58.13%	84.68%
50/50 Split	100.00%	56.75%	84.33%
20/80 Split	99.86%	55.94%	83.89%
Averages of Splits	99.95%	56.94%	84.30%

	Mushroom	Wine Quality	Adult Census Income
80/20 KFold Training	100.00%	62.61%	87.83%
50/50 KFold Training	100.00%	64.52%	88.26%
20/80 KFold Training	100.00%	64.03%	90.29%
Averages KFold Training	100.00%	63.72%	88.79%

	Mushroom	Wine Quality	Adult Census Income

80/20 KFold Validation	100.00%	58.09%	84.56%
50/50 KFold Validation	100.00%	60.70%	84.69%
20/80 KFold Validation	99.94%	57.36%	84.23%
Averages KFold Validation	99.98%	58.72%	84.49%

For the mushroom data, the Neural Net testing accuracy corresponds with the baseline model performance given by the UCI Machine Learning Repository. Similar to the two models above, the Neural Net model also seems to be very much strong with the mushroom data.



For the Wine quality data, Neural net had testing accuracy of 56.94%. Comparing it with the testing result of Cortez et al., the result seems to be an accurate result for this data. The paper mentioned had an accuracy of 59.1% for red wine and 52.6% white wine in 50/50 split, which is a similar result to what the model achieved.

For Adult Census Income data, the Neural Net model had a testing accuracy of 84.30%. Compared to UCI Machine Learning Repositoty's baseline model performance of 78.386%, this study's Neural Net model seems to have better accuracy. Potential explanations for this difference could be in the process of loading and preprocessing the data. As mentioned above in the data section, the Neural Net model's data preprocessing functions were different from SVM and Boosted Trees for all three datasets. Moreover, for Adult Census Income data, the SVM and Boosted Trees models have undergone Principal Component Analysis to speed up the loading time. This drastic difference in the preprocessing may have affected the Neural Net's performance, not accurately performing.

6. Conclusions

As you can see in the section above, the Boosted Trees model had slightly better performance across all three datasets than other two models. Moreover, it had better performance in terms of K-Fold Cross Validation accuracies as well. The result for all three models in all three data seems to correspond with either the model baseline performance from UCI Machine Learning Repository or referenced papers. As mentioned in Caruana and Niculescu-Milzil's paper, this study has successfully shown that the Boosted Trees model has the best performance among many supervised learning classifiers.

References

- UCI Mushroom Dataset: https://archive.ics.uci.edu/dataset/73/mushroom
- UCI Wine Quality Dataset: https://archive.ics.uci.edu/dataset/186/wine+quality
- UCI Adult Census Income Dataset: https://archive.ics.uci.edu/dataset/2/adult
- Wine Quality paper: @article{Cortez2009ModelingWP, title={Modeling wine preferences by data mining from physicochemical properties}, author={P. Cortez and Antonio Lu{\'i}z Cerdeira and Fernando Almeida and Telmo Matos and Jos{\'e} Reis}, journal={Decis. Support Syst.}, year={2009}, volume={47}, pages={547-553}, url={https://api.semanticscholar.org/CorpusID:2996254}}
- Caruana and Niculescu-Mizil paper
  https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf