Different Machine Learning Algorithms for Wine Dataset

Di Xiao | 002930356
Jay Shukla | 0002619454
Emmanuel Chibua | 002799484
College of Engineering
Northeastern University
Toronto, ON
Xiao.di1@northeastern.edu
Shukla.j@northeastern.edu
Chibua.e@northeastern.edu

Abstract

The quality of wine is a crucial aspect for both consumers and the wine industry. However, the conventional method of measuring wine quality is time-consuming. As machine learning models have gained significance in replacing human tasks, we aim to explore the essential wine features to achieve desirable results using these models. To accomplish this, we have employed Decision tree regressor, XGboost, and Random forest regressor for evaluation of relevant features. The dataset will be split into training and testing sets, with the training set being used to train the model and the testing set to evaluate its performance. We will also implement Grid search hyper parameter adjustment to improve the model's accuracy. This project emphasizes the importance of selecting the appropriate machine learning algorithm for a particular dataset to obtain optimal outcomes and identify new features. The results of this study will be valuable for researchers and professionals in the field of data analysis and machine learning.

1 Introduction

Datasets used in this report is from Kaggle.com named "Wine Review", the information about the dataset is as follows:

Table 1: The basic feature of dataset.

	Data Set Characteristics	Attribute Characteristics	Associated Tasks	Number of Instances(train-test)	Number of Attributes(train-test)
Wine Review	Multivariate	Real	Regression	150930	11

1.1 Data characteristics

This time, we can showcase the practical applications of machine learning techniques on the wine dataset. The Wine dataset provides a valuable opportunity to explore the application of machine learning techniques in a practical setting.

Screenshots of the data are provided in Figure 1-1.

		country	description	designation	points	price	province	region_1	region_2	variety	winery
1	0	US	This tremen	Martha's Vin	96	235	California	Napa Valley	Napa	Cabernet Sa	Heitz
2	1	Spain	Ripe aromas	Carodorum S	96	110	Northern Sp	Toro		Tinta de Tor	Bodega Carm
3	2	US	Mac Watson	Special Sele	96	90	California	Knights Valle	Sonoma	Sauvignon Bl	Macauley
4	3	US	This spent 2	Reserve	96	65	Oregon	Willamette \	Willamette \	Pinot Noir	Ponzi
5	4	France	This is the to	La Br??lade	95	66	Provence	Bandol		Provence re	Domaine de l
6	5	Spain	Deep, dense	Numanthia	95	73	Northern Sp	Toro		Tinta de Tor	Numanthia
7	6	Spain	Slightly gritt	San Rom??n	95	65	Northern Sp	Toro		Tinta de Tor	Maurodos
8	7	Spain	Lush cedary	Carodorum ?	95	110	Northern Sp	Toro		Tinta de Tor	Bodega Carm
9	8	US	This re-name	Silice	95	65	Oregon	Chehalem M	Willamette \	Pinot Noir	Bergstr??m
10	9	US	The produce	Gap's Crown	95	60	California	Sonoma Coa	Sonoma	Pinot Noir	Blue Farm
11	10	Italy	Elegance, co	Ronco della	95	80	Northeaster	Collio		Friulano	Borgo del Tig
12	11	US	From 18-yea	Estate Viney	95	48	Oregon	Ribbon Ridg	Willamette \	Pinot Noir	Patricia Greer
13	12	US	A standout e	Weber Viney	95	48	Oregon	Dundee Hills	Willamette \	Pinot Noir	Patricia Greer
14	13	France	This wine is	Ch??teau Mc	95	90	Southwest F	Madiran		Tannat	Vignobles Bru
15	14	US	With its soph	Grace Viney	95	185	Oregon	Dundee Hills	Willamette \	Pinot Noir	Domaine Sere
16	15	US	First made ir	Sigrid	95	90	Oregon	Willamette \	Willamette \	Chardonnay	Bergstr??m
17	16	US	This blockbu	Rainin Viney	95	325	California	Diamond Mc	Napa	Cabernet Sa	Hall
18	17	Spain	Nicely oakec	6 A??os Rese	95	80	Northern Sp	Ribera del Duero		Tempranillo	Valduero
19	18	France	Coming from	Le Pigeonnie	95	290	Southwest F	Cahors		Malbec	Ch??teau Lag

Figure 1-1: Wine train data.

1.2 What problems you want to explore?

In this project, we aim to investigate the significance of wine features in achieving desirable outcomes using machine learning models as a substitute for human tasks. We employed three different algorithms to perform this analysis and will compare their performance to determine which algorithm is most effective. Ultimately, our goal is to identify the most important wine features that can be used to obtain promising results through the use of machine learning techniques.

1.3 Why are these problems or dataset interesting?

The wine dataset is interesting because it provides a rich source of information for exploring the world of wine and developing machine learning models for wine recommendations. This dataset is interesting for two reasons below:

1)The dataset contains a large number of observations, including different varieties of wine, points, and descriptions.

2) The dataset has been cleaned and preprocessed, which makes it easier to work with and reduces the amount of time required for data cleaning and preparation.

2 Experiment

2.1 Interesting Facts

Based on the findings presented in Figure 2-1, this figure give a quick look of the points count distribution, we could see the most points are majorly concentrated between approximately 86-90 points. However, there are some wines which have received points less than 82 as well. In addition, the points seem to follow a bell shaped normal distribution curve.

In Figure 2-2, It can be seen that there is significantly strong positive correlation of price with the points the wine gets. however, there are few cheaper wines who are successful in getting good points.

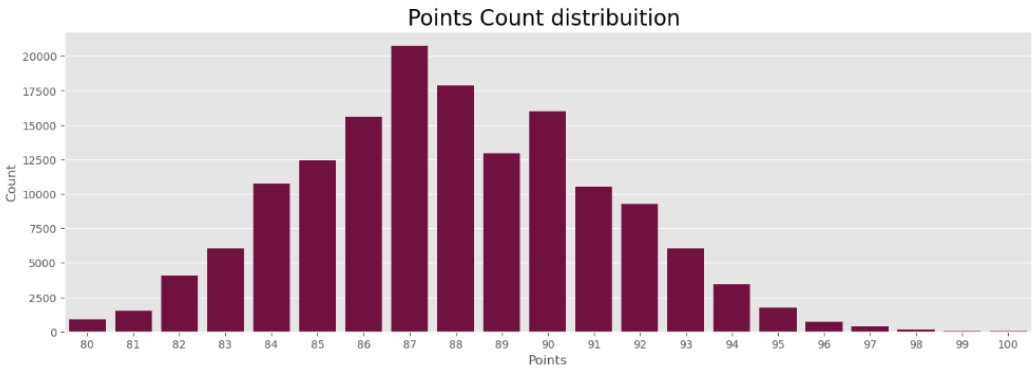

Figure 2-1: Wine- points count distribution data.

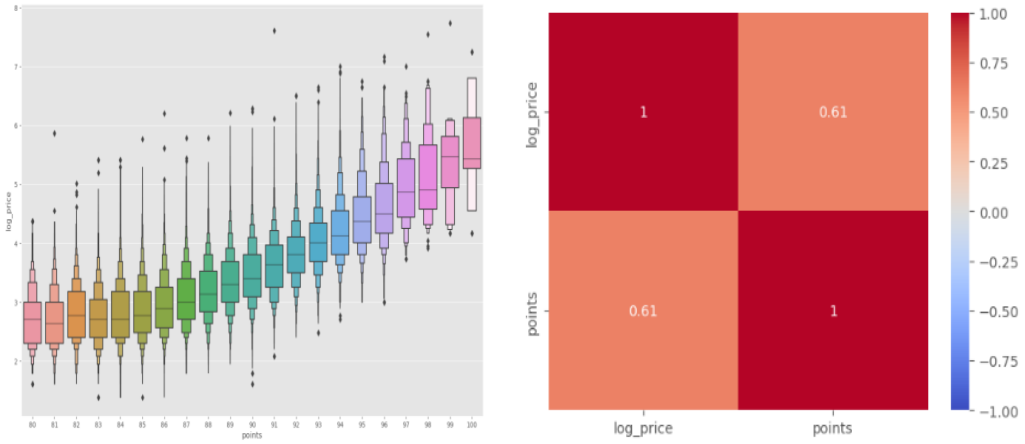

Figure 2-2: Wine – log price data and heatmap for logprice and points data.

3 Problem Description

Before starting to write the report, I found the following problems:
1) Can we use the algorithm to predict the price of some kind of wine? Using points? And how is the relevant algorithm's performance like?
2) How to compare the results? If the results are not as match as we expected. what should we do to next step? Is that enough as we only select three algorithms?

4 Model Implement and Methodology

4.1 Decision Tree Regressor

The decision tree regressor is a powerful and interpretable algorithm that can handle both numerical and categorical features, it is a type of supervised machine learning algorithm used for regression tasks. It is based on the concept of a decision tree, which is a tree-like model of decisions and their possible consequences. Please check below Figure 4-1, it showed us the decision tree regressor learning curve.

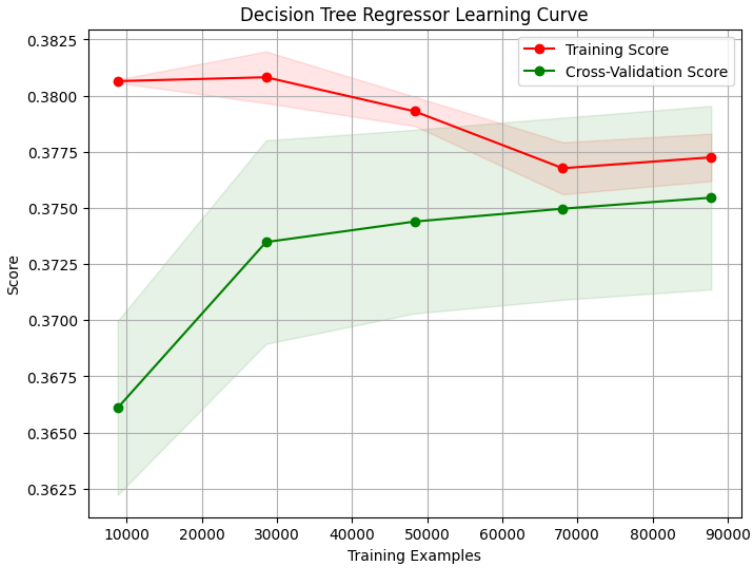

Figure 4-1: decision tree regressor learning curve data.

4.2 XGboost

XGBoost (Extreme Gradient Boosting) is a powerful and popular machine learning algorithm for regression, classification, and ranking tasks. It is an optimized implementation of gradient boosting that leverages parallel computing and tree pruning techniques to improve performance and prevent overfitting. Please check below Figure 4-2, it showed us XGboost learning curve.

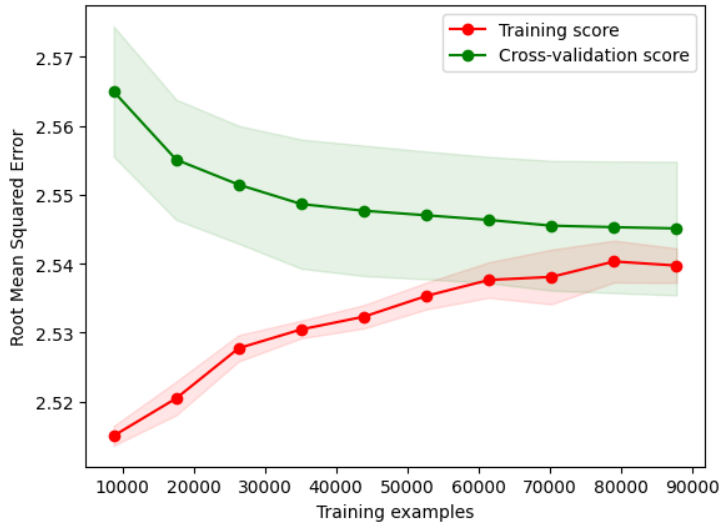

Figure 4-2: XGboost learning curve

4.3 Random forest Regressor

Random Forest is a popular machine learning algorithm for both regression and classification tasks. It is an ensemble learning method that combines multiple decision trees to improve accuracy and prevent overfitting. Please check below Figure 4-3, it showed us random forest learning curve.

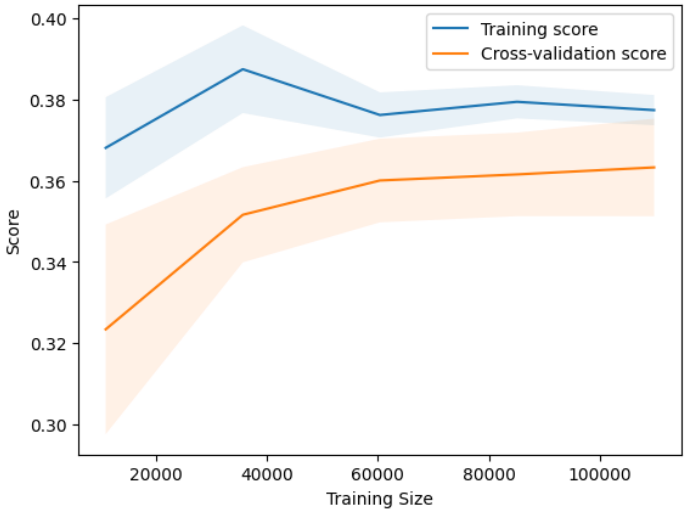

Figure 4-3: random forest learning curve

4.4 Summary three mentioned algorithms performance

Method	Accuracy
Decision Tree Regressor	0.377191346
Random Forest Regressor	0.377264352
XGBoost	0.169599592

5 Conclusions

Based on the evaluation of the model performance using the same dataset in this project, it can be concluded that the Decision Tree Regressor and Random Forest Regressor have similar accuracy. However, the Random Forest Regressor outperforms the Decision Tree Regressor in terms of accuracy. The XGBoost model, on the other hand, has the lowest performance compared to the Decision Tree Regressor and Random Forest Regressor. Therefore, based on the given dataset, the performance order of the models would be Random Forest Regressor with the highest accuracy, followed by Decision Tree Regressor, and then XGBoost with the lowest accuracy.
We also can not make some conclusions that Decision Tree and Random Forest Regressor algorithms will predict the price of some kind of wine.

5.1 Conclusion and future work

It is important to note that the analysis presented above is not sufficient to draw a definitive conclusion on which algorithm is better for regression tasks. The reason for this is that the analysis only considers a single dataset, and only a few algorithms are compared. To have a more comprehensive understanding of the performance of different algorithms for regression tasks, it is necessary to analyze multiple datasets with different characteristics, and to compare more algorithms. This will provide a more robust and reliable evaluation of the strengths and weaknesses of each algorithm, and help identify the most suitable algorithm for a given problem.

In future, when selecting an algorithm to analyze a particular dataset, it is important to consider the specific requirements and characteristics of the problem at hand. After knowing more machine learning algorithms, we can start to compare or analysis the performance of the algorithms. It is important to keep in mind that the performance of a model can be influenced by many factors,

including the quality and representativeness of the data, the feature selection and engineering process, and the choice of evaluation metrics. Therefore, it is crucial to carefully consider these factors and choose the most appropriate algorithm for the specific task at hand.

Overall, machine learning algorithms are definitely worth investing time and effort in, as they provide valuable insights and help extract meaningful patterns from complex and unstructured datasets, which can be used for various applications in research and industry.

6 Acknowledgments

The learning code is adapted from

https://www.kaggle.com/code/ayushnitb/wine-review-indepth-eda-sentiment-analysis

https://www.kaggle.com/code/hetulmehta/wine-recommendation

https://www.kaggle.com/code/halimedogan/red-wine-quality-prediction

and learning code algorithms materials from week6 and week7 hands on session.

7 References

[1] Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

[2] Chawla, N.V., 2005. Data Mining for Imbalanced Datasets: An Overview, in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA, pp. 853–867. https://doi.org/10.1007/0-387-25465-X_40

[3] X. Li, L. Wang, and E. Sung, "AdaBoost with SVM-based component classifiers," Engineering Applications of Artificial Intelligence, vol. 21, pp. 785-795, Aug. 2008.

[4] Lee, S., Park, J., Kang, K., 2015. Assessing wine quality using a decision tree, in: 2015 IEEE International Symposium on Systems Engineering (ISSE). Presented at the 2015 IEEE International Symposium on Systems Engineering (ISSE), IEEE, Rome, Italy, pp. 176–178. https://doi.org/10.1109/SysEng.2015.7302752

[5] Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458), 611-631.

[6] Huang, J. Z., & Ling, C. X. (2005). Using AIC and BIC criteria to determine the number of clusters in mixture modeling and hierarchical clustering. IEEE Transactions on Knowledge and Data Engineering, 17(2), 182-191.