Machine Learning Methods to Predict Wine Quality

Joseph Despres

April 12, 2022

Contents

Introduction

Anyone looking to purchase a bottle of wine will be facing a number of options so large that only possible for a computer to search through. Wine varies signnificantly from region to region. Therefore, this is an ideal problem for machine learning to handle. In particular, this projct aims to speficy a function that maps chemical properties of varies wines to the judges quality scores. This report begins with a dicsussion of the problem followed up by a description of the data. To model I use the sklearn package (Pedregosa et al., 2011) in the Python Programming Language (Van Rossum & Drake, 2009). After that, I devote several paragraphs to comparing different feature scaling strategies against logistic regression because there is a significant difference in runtime and performance. After that, I discuss the useof an elastic net regression, XGBoost, and Support Vector Machines in an ensemble to reach a prediction accuracy of . From there, I will go back into the dataset and attempt to improve the model's the accuacy through some more feature engineering.

Problem Description

This project aims to discover the relationship between wine quality and chemical properties. The motivation for this project is that one purchasing wine seriously must consider different regions, grapes, growing climate, seasons, storage time, and substantial variations in quality and price. This is a vast and complex search space with the potential to find something of significant value. This is a problem well suited to Machine Learning because of the search The benefits of an accurate predictive wine model are that one could select excellent wine at a low price, lower the risk of purchasing low-quality wine, or even determine undervalued wine at auctions. I am not the first person to think of this, hence the published dataset.

Methodology

To determine the quality score based on chemical properties of a wine we will use a dataset hosted by the University of California Irvines Machine Learning reposotory (Lichman, 2013). These data are from a study conducted by the portugese governmnt studying the ability to perdict human wine taste (Cortez, Cerdeira, Almeida, Matos, & Reis, 2009). Using support vector machines, they achieve an out of sample accuracy of 0.33. Given this is in 2009, the purpose of this study is to employ a mixture of newer methods and attempt to outperform the publushed by Cortez. First, I do not have the domain knowledge to asses the validity of the data. I have no idea what is an unreasonable level of acidity, for instance. However, given that this is a published dataset, I assuming all the opservatins are actual obeservations.

This is a structured, curated, tabular dataset conviently in First Normal Form. The outcome we wish to predict is the quality as rated by a professional wine judge. One thing that was not obvous from the UCI reposotory is the fact that there are duplicated rows, which means multiple judges rated the same wine the same quality. However, judges were annonomoys in this study, therefore there is no feature indicating this. I will come back to this point after discussign model specifications. There are 11 additional features, which are listed in Table 1.

the target is wine quality as scored by a professional judge. Wine quality is an ordinal categorical variable. The class assigments is quire unballanced. However, the nature of this problem is that very few wines will be of very high quality and many will be average. See Figure 2 for a plot of the class distributions. At this point, I will split the data and stash 30% of the data to reserve for testing and I will not touch those data until it is time to report a final result.

After some dificulties with fitting initial models and performance plateaus associated with the min max scaling method. I decided to test and document the results of using several diffferent scaling strategies. I decided to run a 2,500 different partitians of training and testing data with a 50% split to investigatiing

Table 1: Descriptive Statistics

Feature	Mean	Std
Fixed Acidity	7.215	1.296
Volatile Acidity	0.340	0.165
Citric Acid	0.319	0.145
Residual Sugar	5.443	4.758
Chlorides	0.056	0.035
Free Sulfur Dioxide	30.525	17.749
Total Sulfur Dioxide	115.745	56.522
Density	0.995	0.003
Ph	3.219	0.161
Sulphates	0.531	0.149
Alcohol	10.492	1.193
Quality	5.818	0.873
Is Red	0.246	0.431

Note: 6497 observations

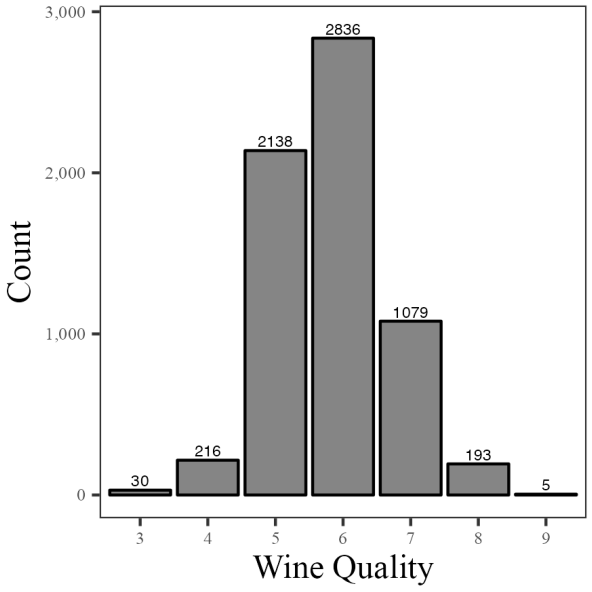

Figure 1: Comparing Resampled Prediction Accuracy by Transformationf

the effects diffeerent scaling schemes. There was a significant difference in results as well as runtime. This is due to the design matricies haveing dificulties inverting. The runtimes will differ based on packages, hardware and datasets so I don't list them.

This paper will consider two performance metrics. First accuracy measured as the percentage of correct predictions. Note, this is an unballanced assignment problem, accuracy should not be compared to the number of classes rather the frequency of the most frequent class. The most freqent class is a quality score of 6. Which occurs in about 43% of the observations. This This metric is intuitive, but does not capture our true intentions. I will also use a ballance between precision and recall known as AUC. AUC is simply the area under the ROC cuve.

Table 2: Demo table

Scaling Method	Mean	Std	Median
No Transformation	0.590	0.024	0.592
Dropped Outliers	0.645	0.026	0.641
Min Max Scale	0.690	0.018	0.691
Rescaled to Uniform	0.707	0.015	0.707
Standard Scale	0.726	0.017	0.725
Yao Johnson Transform	0.733	0.015	0.732

The Yao-Johnson outperforms the Standard scaling methods however, not by much and is a far more complicated transformation process. It is a Box-Cox transformation mathod modifed to take values less than 0. The Box-Cox Method is for linearizing vectors by some exponent that will convert the variable to a linear scale. I do not see any indication that wine chemical properties are anything other than linear in practice. My reasioning for this is that one does not add an exponentially increasing or decreasing amoutn of sugar for instance. Although it has slightly better accuarcy and lower standard deviation, both are desireable properties, The invrease did not justify the additioal complexity and a deviation from what I felt is reasionable. Therefore, I will use the standard scaler.

After studying the features and exploratory plotting, we are now ready to specify the models. The plan is to use an ensemble of models to predict

wine quality. The idea is that a collection of viting classifiers will be more accurate than a single one. The reasioning is that one model could be wrong in many ways however it would be correct in only one. Add a collection together and they will be collectively right more than wrong.

The first model is the multinomeal logistic regression used during the feature scaling trials. This model is model is based on estimating the cumulative log odds of belonging to a given category based on the features. I add some regularization parameers L^1 and L^2 norms. commonly refered to as elastic net regression. A gridsearch over a five fold cross valudation shows that an optimal $L^1 = 0.0315$ and $L^2 = 0.013$. This achieves a training set accuarcy of 54.53% and and AUC of 0.775. These are not impressive results, so we will fit a Support Vector Machines.

A support vector machine model aims to find the maximimum margin hyperplane between featurs that seperate points. I use the radial basis function kernal and tune only the regularization parameter squared L^2 norm. After crossvalidating I find the best regularization parameter to be 6.58. This reaches an insample accuracy of 60% and an AUC of 0.801. A slight improvement.

After that I use XGBoost (Chen & Guestrin, 2016) an effecient implimentation of gradent boosted trees. There are many parameters to search through, however I will only tune max tree depth, column sample by tree, and learning rate. After tuning, I found column sample to be 50% to be the best a learning rate of about 0.034 and a max depth of 13. These yield an astonishing in sample accuracy of 99.5% and an AUC of 0.998. This a substantial improvement over the past two models. From here I will wrap all three in an ensemble. Test the accuaracy. Given these matrics, it may be better to not even ensmemble and just use XGBoost.

Then I save these models and wrap them into one ensemble model. Then fit that to the training data and compare different approaches. I want to reengineer some features and report some results.

Results

The XGBoost model wildly outperforms the linear I found it has an accuacy of the final testing set of. The ensemble. The model perdictions are in Figure . This shows that the model could be used to

Conclusions

The main take-away of this stuyd

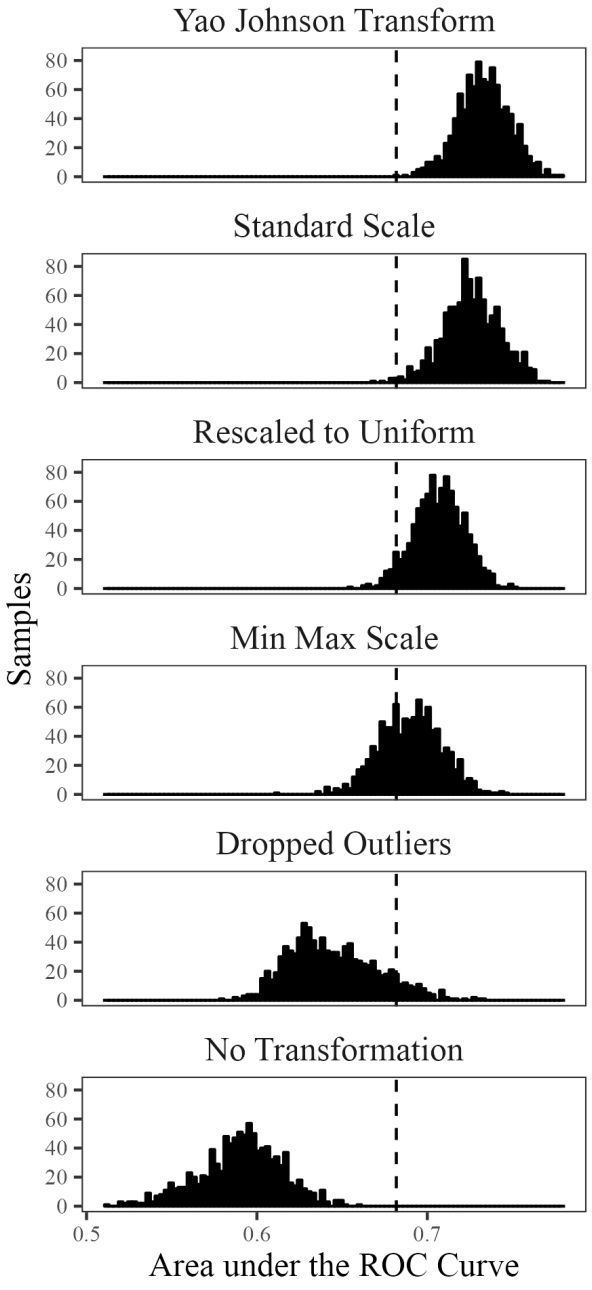

Figure 2: Comparing Resampled Prediction Accuracy by Transformationf

References

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2939672.2939785` doi: 10.1145/2939672.2939785

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547-553. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167923609001377` (Smart Business Networks: Concepts and Empirical Evidence) doi: https://doi.org/10.1016/j.dss.2009.05.016

Lichman, M. (2013). *UCI machine learning repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual.* Scotts Valley, CA: CreateSpace.

Appendix

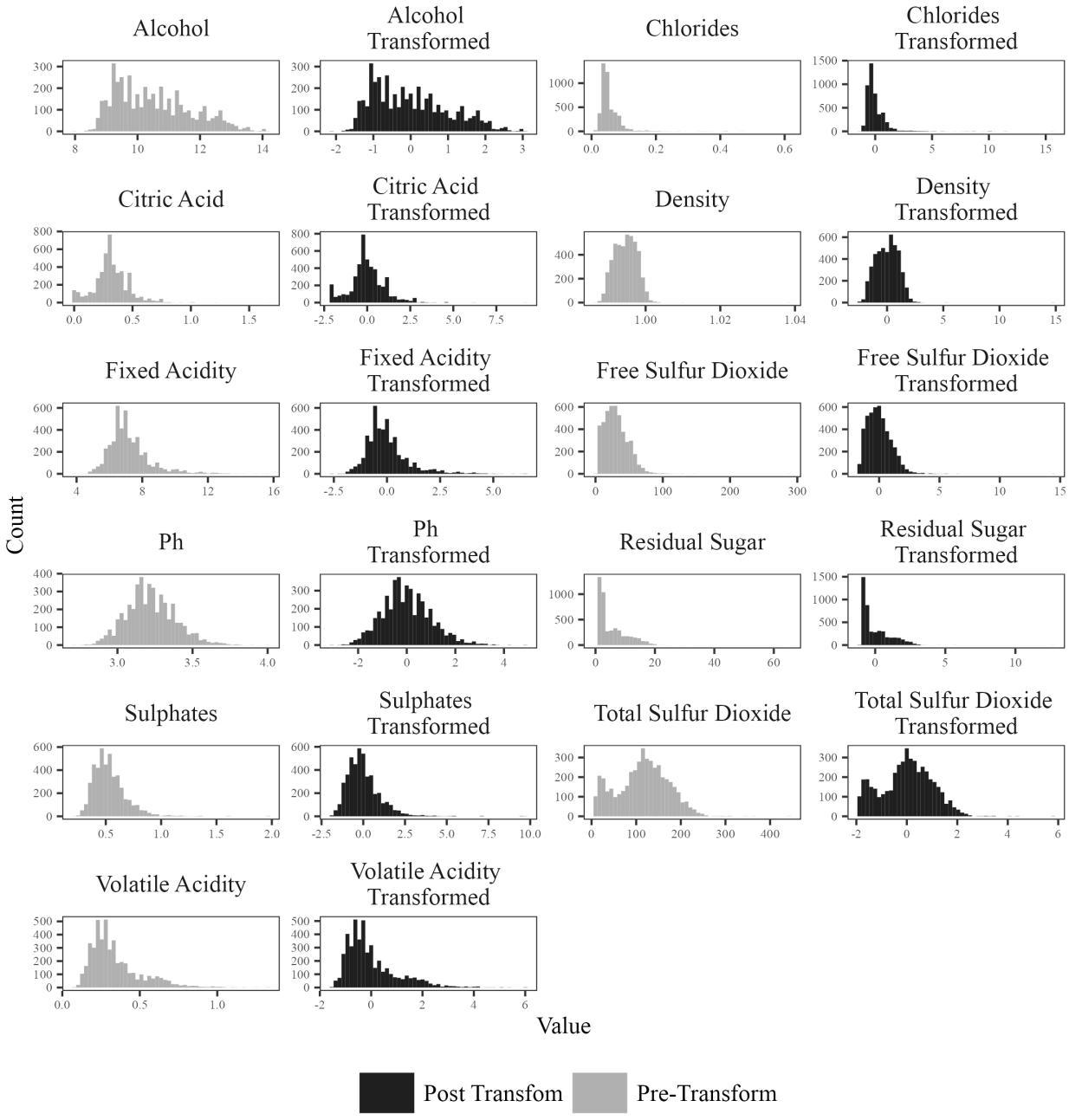

Figure 3: Comparing Resampled Prediction Accuracy by Transformation