Predicting the Quality of Wine with Machine Learning models

Bryce Anthony and **Christos Koumpotis**
{Branthony,Chkoumpotis}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

In this paper we use explore the effectiveness machine learning regression models for the task of predicting the quality of wine based on its chemical components. Using data-sets for red and white wines, we came up with separate models for red and white wines that both proved to be effective in predicting the quality of wines when compared to the baseline models for their respective data-sets. The data-sets for each type of wine consists of 11 features that were used to predict the quality of a given wine. Our research attempts to expand upon existing literature by implementing feature selection and ultimately yields two models that each have a mean squared error of roughly .2 points lower than the baseline models.

1 Introduction

Within the multi-billion dollar global wine industry, One of the main focuses of critics and fanatics is the quality of wine. Despite the quality of wine often being determined by sommelier's, there seems to be a lack of consensus as to the effects of various characteristics on the overall quality of a wine. This lack of consensus combined with the overall subjectivity of wine evaluation creates a unique opportunity for the use of Machine learning in the wine evaluation process.

In this paper, we implement two machine learning algorithms to predict the quality of wines using the 11 characteristics; fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates and alcohol contents, provided in the data-sets. Previous research on the topic, Gupta (2018) and Bhardwaj et al. (2022), identified statistically significant chemical components and used them to create powerful algorithms that were highly effective in predicting the quality of wines.

In our research we found that omitting non-statistically significant specific attributes, results in greater errors, and therefore adjusted our model accordingly. In the remainder of the paper; we provide background information on the machine learning algorithms we implemented, our approach and experimental setup, the results of our approach, as well as the broader impacts of our research and recommendations for future researchers.

Overall, the aim in this section is context-setting: what is the big-picture surrounding the problem you are tackling here?

2 Background

To predict the quality of wines we used data-sets consisting of their chemical components along with their corresponding qualities. Both the red and white wine data-set had 11 variables in addition to the variable for wine quality. Our models used the fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and the alcohol content of wines to predict their quality. The data-sets used to train our models were randomly shuffled and we used the leave-one-out-cross-validation process on 80% (2005 observations for the white wine data-set and 1280 observations for the red wine data-set) of the shuffled data-sets to train and validate our models. The remaining 20% of the data-sets (502 observations for the white wine data-set and 320 observations for the red wine data-set) were reserved for testing our final models after training was completed. The models that we trained were all variations of a linear regression model that used polynomial features of n degrees as inputs or variations of a linear least squares model where the loss function was the l2-norm. To determine the effectiveness of our best performing models their errors were compared to the error of a dummy model that simply averages the results of the training data and uses this average as the prediction for each of the wines.

3 Experiments

In total we tested 16 variations of a linear regression model that used polynomial features as inputs and 16 variations of a linear least squares model, where the loss function is the l2-norm, for each data-set. These linear regression models differed in the number of folds used in the leave-one-out-cross-validation process as well as the degree of the polynomial features to which the predictors were transformed. The linear regression models for both data-sets used predictors that had been transformed to have n degree polynomial features where n = 1, 2, 3, or 5 and the number of folds used for the leave-one-out cross validation process were either 5 or 10 for each polynomial feature variations.

Similarly the linear least squares models varied in the number of folds used in the leave-one-out-cross-validation process but also in the level alpha we used when training the models. For these linear least squares models, we tested alphas 2, 1, 0.5, 0.2, 0.01, 0.001, 0.0001 and 0.0000001 and used both 5 and 10 folds during the leave-one-out-cross-validation process for the models of each level of alpha. Each of these model variations was evaluated using mean squared error as the metric. The mean squared error measures the quality of our models predictions based on the average of their differences from the true value and we chose mean squared error as our metric because our outputs were continuous variables rather than discrete predictions corresponding to the quality scores present in the red and white wine data-sets. The mean squared error of each model variation was computed by averaging the mean squared errors produced from each round of testing when using the leave-one-out-cross-validation process.

–image–

In addition to the 16 variations of linear least squares and 16 variations of linear regression models mentioned above, based on the research of Yogesh Gupta we also tested each of the models using data-sets that only contained features deemed significant at the .05 level in his linear regression. For the red wine model this reduced the data-set to one only contain 7 features (Volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulfates, and alcohol) and for the white wine model this reduced the data-set to one containing 8 features (fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, density, pH , sulfates, alcohol). A similar method of feature selection was used by Bhardwaj et al. where they identified 10 features of the 24 features in their data-set to be used by their models. The effectiveness of these 2 models was also evaluated using mean squared error as the metric and compared to the dummy models made for each data-set.

4 Results

Although feature selection yielded the best performing models for Gupta and Bhardwaj et al., combining feature selection with our models yielded slightly worse results in all of our model variations. For the red wines data-set the dummy model, which averages the wine quality ratings of the training data and uses this average as the prediction for each sample, had a mean squared error of 0.64; meaning that on average the dummy is over a half point away from the true quality rating for a given wine. In comparison our best Red Wine model was linear least squares model with l2 regularization and had a mean squared error of 0.41 on the test data-set meaning that , on average, it was closer to the correct wine quality rating than it was to the incorrect rating. This difference of 0.23 appears to be notable due to its proximity to an error of .50. For the white wine model, the dummy model for the white wine data-set had a mean squared error of .78 while the best performing white wine model that we trained yielded a mean squared error of .51 on the test data-set for white wines. In addition to our model having a lower mean squared error than the dummy model, the 0.27 difference between the error of the dummy model and our model appears to be a notable one Though we can't compare the effectiveness of our models directly to the effectiveness of the models presented in Bhardwaj et al. and Gupta, as Gupta used different metrics for evaluation and Bhardwaj research had a different focus. The results of our model training suggests that there may be some variables in the white wine data-set that aren't as important to the results of the regression as others. We reached this conclusion because the linear least squares regression with l2 normalization is a method used when the independent variables are highly correlated, performed worse than our linear regression with degree 2 polynomial inputs. The linear regression model allows for the interactions between variables to be weighted more heavily than others or even zeroed out which leads us to believe that the better performance of the linear regression model may be due to certain features and their interactions being more important for the regression than others. The white and red wine data-sets have the same variables, so the effectiveness of the different model types on each data-set could be due to the differing number of samples present in the data-sets or the performance of the different models could suggest that the components that affect the quality of red and white wines may be different from each other.

5 Broader Impacts

Overall, the main issue with using technologies such as machine learning algorithms is that they simply learns to predict certain aspects, in relation to the specific information it has been trained on. Therefore, one issue that could arise from using our's or similar models to predict the quality of a wine, is the complete lack of consideration to innovation or unique, and special wine productions. More specifically, such models can be very good at predicting the quality of a common or average wine, with qualities similar to the hundreds of others that already exist and have been tried and produced for many years, however, it would completely fail at predicting the quality of a wine that incorporates new and unique qualities. This could pose a real threat to the industry, if producers and/or consumers over-rely on such predictions to guide their decision making, as it would result in making the wines produced and preferred to be less and less diverse and therefore uniform. Many economists support that lack of innovation and diversity of products in any given market, results in less growth and eventually less choices for the consumers, and therefore could have a great negative effect on the market.

We created two machine learning models that that use information about 11 characteristics of a wine to predict its quality using regression. We, also, investigated the methodology of previous literature, where specific characteristics have been statistically determined to be more important variables when it comes to their influence on wine quality and therefore models using only those were more accurate, however for our models, eliminating certain characteristics proved to increase the error of our predictions. We came up with unique models for red and white wines, with the best performing model for red wines outperformed our best performing model for white wines. Overall, our models were

relatively accurate in predicting the quality of the wines, for both white and red,as in comparison to the dummy model, they both were more than 0.2 accurate. However, for future research, we suggest investigating the effect the differences between our model and experimental process and previous literature's models and processes, such as Gupta's (2018) and Bhardwaj (2022), so to identify what is more accurate for a greater set and variety of data. It would also be interesting to test the models with more unique wines, whose characteristics might be different than the norm. It is important to note, that the predictions of their models were generally more accurate than ours, which is attributed to their model using classification, a method that has been demonstrated to perform better than regression.

6 Conclusion

We created two machine learning models that use information about 11 of the chemical components of a wine to predict its quality using linear and least squares regression techniques. We also investigated the methodologies of previous literature attempting to perform feature selection only using the features deemed statistically significant, by past researchers, when it comes to their influence on wine quality. We hoped that it would yield more accurate models however, for our models, removing certain features from the datasets proved to increase the error of our individual predictions and led to worse model performance overall. We came up with two different models for predicting the quality of red and white wines, with our best performing red wine model slightly outperforming our best white wine model. Overall, both our white and red wine models were relatively accurate in predicting the quality of the wines when compared to our dummy models, both the white and red wine models had an average error of 0.2 points less than their respective dummy models. For future research, we suggest investigating the differences between our model and experimental process compared to the models and experimental processes of existing literature (i.e. Gupta's (2018) and Bhardwaj et al. (2022)) to identify methods of more accurate and more generalizable models. It would also be interesting to test the models with more unique wines, whose characteristics might be different than those of a white or red wine. It is important to note, that the predictions of their models were more accurate than ours, which we believe can be attributed to their use of a classification model, which has been demonstrated to perform better than regression models for classification problems.

7 Contributions

Bryce and Christos pair programmed all of the models and experiments that were conducted but split up the actual writing portion of the paper. The Introduction, broader impacts, conclusion, and references were mostly written by Christos while the Background, experiments, and results were mostly written by Bryce

References

Bhardwaj P, Tiwari P, Olejar K, et al (2022) A machine learning application in wine quality prediction. Machine Learning with Applications. doi: 10.1016/j.mlwa.2022.100261

Gupta Y (2018) Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science 125:305–312. doi: 10.1016/j.procs.2017.12.041