Predicting Wine Quality from Chemical Components Using Linear Regression Machine Learning Models

Chase Hamelink and **Ian Rolls**

Abstract

For luxury markets, maintaining a high standard of products is necessary to uphold a luxury reputation. This paper uses supervised linear models to examine how the chemical properties of wine can be used to predict its quality. Specifically, we use L1 and L2 linear regression models on data sets for both red and white wine. We conclude that chemical properties do in fact affect wine quality and we find that our L2 regularization model when tuned was the best-performing model, increasing the likelihood of a correct quality prediction when compared to a dummy model.

1 Introduction

Wine making is an industry where there are many different types and qualities of wine purchasable. In order for a high-end winemaker to maintain their reputation, they must provide a certain level of quality in every bottle of wine. As food-related technology improves and these businesses advance, wineries can implement new technologies to better understand the chemical processes happening and optimally alter their production. This way they can save time and money when crafting batches that must meet certain quality thresholds. We aim to use data and machine learning approaches to better predict the quality of wine based on its chemical properties.

Prior work has been done on this topic utilizing an array of complex learning models, including neural networks, support vector machines, K-nearest neighbors, decision trees, and random forest models to name a few. Notably, in the literature, Dahal et al. concluded that a Gradient Boosting Regressor was the best-performing model they tested (Dahal et al. 2021). Using mean-squared error (MSE) as a metric to assess performance and running their models on red wine data, they reported an MSE of 0.3741 when run on the validation set. Similarly, Gupta looked to examine how the exclusion of variables improved predictive power when looking at wine quality (Gupta 2018). In this study, they made use of an 11-5-1 neural network architecture and reported a validation set error of roughly 0.243 for white wine and 0.170 for red wine when taking into account all chemical parameters. While this paper did report improved errors in their validation set when excluding variables, we made the informed decision to explore linear regressions' predictive power when supplied with all chemical variables available. By doing so we can assess the difference between basic L1/L2 regularization models to predict wine quality given all relevant chemical properties. The predictive power of these models will be scored and assessed with mean-squared error and compared to a benchmark dummy model.

2 Exploratory Analysis

Prior to model building and data processing, we generated histograms and heat maps to explore the relationship between each chemical property and the ultimate quality rating. In doing so we explored the distribution and correlation of the data and formed a better understanding of how our data will affect the model's learning.

From the histogram plots in figures 1 and 2, we can begin to visualize how our data is distributed across the features. Notably, the quality histogram shows that no ratings were provided below 3 or above 9. This should be accounted for when interpreting model accuracy as we do not have data that can train the model on extremely high or low-quality wine (quality < 3 or > 9). It is also evident that the distribution of quality data is skewed towards wines rated between 5 and 7. As such, model accuracy will likely be greater when predicting wines rated in this range

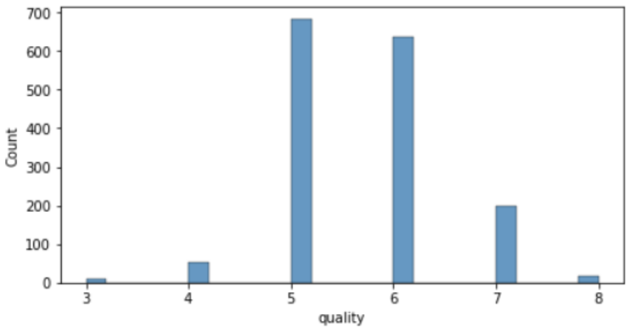

Figure 1: Histograms of features and target from red wine data.

To explore the correlation between our features and target, we generated correlation heat maps (figures 3-4). From these heat maps, we can see that alcohol was the most correlated with quality in both red and white wine (.48 and

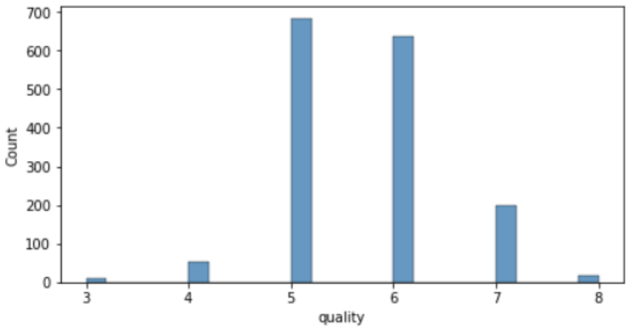

Figure 2: Histograms of features and target from white wine data.

.44 respectively). Although, density and chlorides seemed more correlated with quality in white wine (-.31, -.21), and volatile acidity and citric acid were more correlated with quality in red wine (-.39, .29).

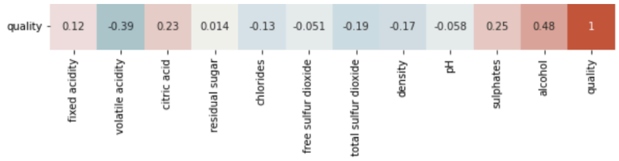

Figure 3: Heat map of the correlation between features and the target from red wine data. (More red being more positively correlated with quality and more blue being more negatively correlated with quality)

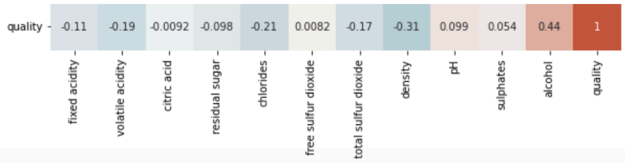

Figure 4: Heat map of the correlation between features and the target from white wine data. (More red being more positively correlated with quality and more blue being more negatively correlated with quality)

3 Dataset

Having performed an exploratory analysis of the data to determine the distribution and correlation of our features, we formatted our dataset to fit our linear regression models. We used the built-in min-max scaling function from sklearn to scale our features (while leaving the target unscaled). Once scaled, we used an 80/20 train-test split to separate the data into our training data and testing data.

The data set that is used to teach our models was obtained from Cortez et al. and pertains to red and white wine variations of the Portuguese ”Vinho Verde” wine (Cortez et al. 2009). For both the red and white wine data, there are 11 chemical compounds in the dataset that we are using as our features, and a quality rating for each batch of wine, which we are using as the target. Both red wine and white wine datasets are relatively robust, containing 1599 and 4898 unique entries respectively (Cortez et al. 2009).

The properties include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality rating. The quality rating was our target variable, which was determined by wine tasters and rated on a scale from 0-10, 0 being awful and 10 being excellent.

We separated the red and white wine data into two different models, as the average level of each chemical compounds sometimes vary between the two types of wine. Additionally, some of the chemical compounds have different correlations with quality.

4 Experiments

Our goal was to predict the quality of an unknown wine provided its chemical properties. We ran two types of supervised linear models to predict the quality: a Lasso regression (L1), and a Ridge regression (L2). Both methods use a cost curve and gradient descent to arrive at an optimal prediction for all θ values. Both linear regression models predict the θ values in equation 1:

$$\hat{y} = \theta_0 + \sum_{i=1}^{n} \theta_i \cdot x_i \tag{1}$$

Where $\hat{y}$ is our estimated target, each x is a feature, n is the number of features, and θ is the predicted coefficient. In multiple regression, the θ parameters are estimated by minimizing the mean squared error (MSE) defined by:

$$\text{MSE} = \frac{1}{n} \cdot \sum_{i=1}^{n} (y - \hat{y})^2 \tag{2}$$

where y is the actual value of the targets and $\hat{y}$ is the predicted value of the targets.

The Ridge regression (L2) uses gradient descent to minimize equation 3, which is a function of the MSE and a complexity cost.

$$\text{RR} = \text{MSE} + \lambda \sum_{i=1}^{n} \theta_i^2 \tag{3}$$

The λ parameter changes how highly the complexity cost is weighted towards the total cost. The complexity cost is useful for minimizing the size of the coefficients on each feature, as smaller coefficients are more likely to yield a convex cost function necessary for performing a gradient descent that always yields a global minimum cost.

The Lasso regression (L1), seen in equation 4, similar to the L2 regression is also a function of the MSE and a complexity cost.

$$\text{LASSO} = \text{MSE} + \lambda \sum_{i=1}^{n} |\theta_i| \tag{4}$$

However, the complexity cost of the Lasso regression encourages the total sum of all θ values to be minimized rather than the minimization of each individual θ value.

Additionally, for features more highly correlated with the target, we included x^2, $ln(x)$ and x^3 terms in the dataset to account for different patterns in correlation. Specifically, for red wine we did this for fixed acidity, volatile acidity, citric acid, chlorides, sulphates, and alcohol. For white wine, we did this for fixed acidity, volatile acidity, total sulphur dioxide, chlorides, density, and alcohol.

We used the sklearn package to implement our L1 and L2 linear regressions, specifically the RidgeCV and LassoCV functions. In order to predict the accuracy of our models we used 10-fold cross-validation on our training data. This enabled us to test our models on "unseen" subsets of the training data to obtain a better idea of how well our models were performing without compromising our final test data. Additionally, this allowed us to still train on a larger subset of the total data which strengthened the predictive power of our models.

For both L1 and L2 regressions, we tuned our algorithm's lambda parameter, which represents the weight of the complexity cost as a subset of the total cost function. Lambda was chosen for hyperparameter tuning to ensure that we avoided over-fitting or under-fitting our model. We did not manually tune our learning rate, as the sklearn package automatically optimizes the learning rate.

We tuned lambda by using a randomized search on basic Lasso and Ridge models. The randomized search returned MSE values for all selected lambda parameters based on their performance on the cross-validation data. The randomized search was run twice for each model. The first run was used to find an optimal order of magnitude for lambda (i.e. 0.001, 0.01, 0.1, 1, etc.) and the second was a more granular search used to further hone in on the optimized value (i.e. 0.1, 0.2, 0.3, 0.4, etc.). We then used the best-performing lambda values for each regression model. Additionally, the RidgeCV and LassoCV models take a list of lambda values and run a grid search to further tune your hyperparameter. As such, our best-performing lambda from the randomized search, as well as, three similar (untested) values were provided to the RidgeCV and LassoCV functions.

Lastly, we created a dummy model using sklearn's built-in dummy function. This function always predicts the mode of the set of targets, no matter what the features are. This allows us to create a baseline to compare our models to.

5 Results

This paper presents the use of two different types of linear regression models (Ridge regression and Lasso regression) to predict the quality of wine. Of these two models, we had anticipated the Lasso model to be more effective in accurately predicting wine quality. This is because the complexity cost in the Lasso regression is linear, so reducing the weight from 2 → 1 is the same as reducing the weight from 1 → 0. Therefore, the Lasso regression is more likely to weight features that are less correlated with the target as zero because the decrease in complexity cost will be larger than the subsequent increase in the MSE. This is very similar to Gupta's examination of excluding lowly correlated variables. Given the improvement in Gupta's model's predictive power when they performed manual feature selection, we expected a similar result upon implementing an L1 regularization model (Gupta 2018). On the other hand, the Ridge regression is likely to maintain small θ values for less correlated variables. This is because the complexity cost is a function of θ^2 values, so changing θ from 2 → 1 decreases the complexity cost much more than the change from 1 → 0. Therefore we predicted the Ridge Regression to overfit the data due to it's tendency to weight variables without statistically significant correlation to the target.

Cross Validation Results

Model	MSE Red Wine	MSE White Wine
Dummy Model	1.095	0.805
RidgeCV Model	0.402	0.542
LassoCV Model	0.406	0.550

Table 1: Mean squared error scores from cross-validation testing of our models using training data.

Final Test Results

Model	MSE Red Wine	MSE White Wine
Dummy Model	0.903	0.774
RidgeCV Model	0.380	0.533
LassoCV Model	0.374	0.537

Table 2: Mean squared error scores from final testing of our models using testing data.

However, as seen in table 2, the Lasso regression model only outperformed the Ridge regression model for red wine. Furthermore, the difference in MSE between the models for both white and red wine testing data was within 0.006. These results did not align with what we had expected, as we expected an L1 model to outperform the L2 model, however it seems the models were nearly equivalent in performance.

Despite the unexpected result, L1/L2 regularization performing relatively similarly, both models were effective when compared to the dummy model. The mean-squared error improved in both models significantly. When comparing our Ridge regression model to a dummy model, we saw a 137.6% improvement testing on red wine data and a 45.2% improvement testing on white wine data. Similarly, when comparing our Lasso and dummy models, we saw a 141.4% improvement testing on red wine data and a 44.1% improvement testing on white wine data.

From these results, we can conclude that both our L1 and L2 regression models were effective in predicting quality when compared to the dummy mode. However, the models were far more accurate when predicting on the red wine data than that of the white.

6 Broader Impacts

This model shows how applicable machine learning is to the luxury food and beverages industry. Understanding how the underlying compounds within a product affect its enjoyability is necessary to reliably develop popular products. One example that springs to mind is cheese-making. Cheese-making is an incredibly scientific process, which is the perfect opportunity to use machine learning, as a cheese-maker could gather data from tweaking individual steps in the process to better understand what each step does and how it contributes to the overall quality of the product. This way they could improve the quality of their cheese or tweak certain steps to more reliably produce better products.

Although the quality metric for this dataset was measured by expert wine tasters, it might be beneficial for winemakers to have ordinary consumers rate the quality of their products, as it could provide insight into the wine preferences of the average consumer. This wouldn't apply to a luxury brand, which needs to be rated by experienced wine tasters to obtain expert quality certifications. However, this would be helpful for a low to medium quality brand that is just trying to sell tastier wine to its consumer base.

This approach to optimizing products is applicable far outside of the food and wine industry, as companies seeking to sell more products whilst saving money in the development of new products can use data-driven machine learning to better understand customer preferences and predict future product quality.

7 Conclusions

The process of wine-making is highly scientific, and a data-driven approach provides insight into how each component affects wine. We sought to quantify these effects by running L1 and L2 regression models to predict the effect of physiochemical properties on the quality of different batches of red and white wine from Portugal. We found that these physiochemical properties do in fact have a large effect on quality given the increased predictive power of both regression models compared to the dummy model.

However, given the greater predictive power of each model on red wine data versus that of the white wine data, we concluded that more data is necessary to improve our model's accuracy. As discussed, the data provided by Cortez included nearly a third of the number of unique data points for red wine as that of white wine. Furthermore, as seen in the histograms of our exploratory analysis, the quality ratings pertaining to red wine were less diverse than that of white. Thus, in the future, it would be interesting to use a larger dataset that is more robust in quality ratings to assess the accuracy of these models on a larger scale. Furthermore, it would be interesting to include data on wine outside of the region of Portugal covered in the scope of this data. In doing so see we can observe if different varieties have similarly correlated physicochemical properties, or if there is a distinct difference.

8 Contributions

I.R. formatted the datasets and added any extra variables. C.H. made all of the graphics and tables. Both I.R. and C.H. worked on the implementation of sklearn solvers. C.H. wrote the Abstract, Introduction, Exploratory Analysis, Results, and Bibliography sections. Ian wrote the Dataset, Experiments, Broader Impacts, and Conclusion sections. Both I.R. and C.H. proofread all sections.

References

Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; and Reis, J. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decis. Support Syst.* 47(4):547–553.

Dahal, K. R.; Dahal, J. N.; Banjade, H.; and Gaire, S. 2021. Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics* 11(02):278–289.

Gupta, Y. 2018. Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science* 125:305–312.