LUISS

Advanced Statistics - Final Project

Fabrizio Borrelli - 789121 | Filippo Navarra - 782571 | Joshua Brauner - 778931

---

1. Problem description

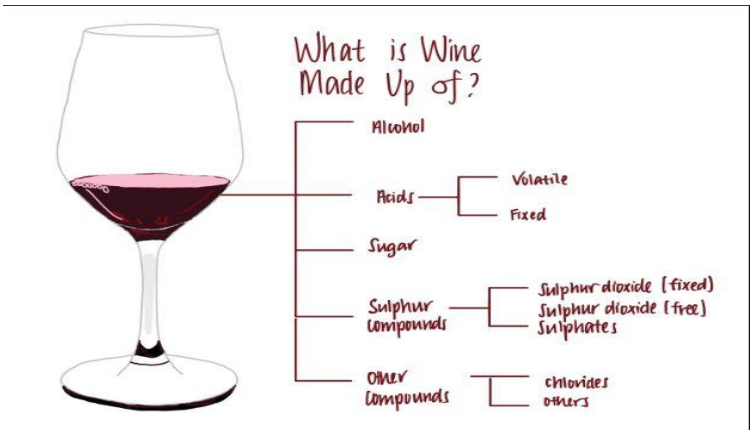
Understanding the determinants of wine quality has become a priority for producers and consumers. Quality is influenced by various factors, such as chemical components like acidity and alcohol. This study aims to understand the complex relationship between the chemical properties of wine and its perceived quality. Leveraging data analysis techniques on a detailed wine dataset, the team attempts to identify key predictors that can effectively predict wine quality. The overall analysis could benefit winemakers in making superior wines, assist distributors in selecting quality products, and enhance consumer knowledge in selecting wines that align with their preferences.

2. Data

Our dataset, titled "Wine Quality Prediction," is made up of a total of 5,427 observations and includes 13 variables, of which 12 are independent features and 1 is a dependent variable, denoted as *“quality”*. Focused on the *"Vinho Verde"* variant of Portuguese red wine, this dataset was acquired from Kaggle.com, a popular platform for data science competitions and exploration.

Given the public nature of Kaggle.com, this dataset has been widely utilized for various analyses, employing both linear regression and classification models. The independent variables within the dataset characterize different chemical aspects of the red wine, encompassing fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, Sulphates, alcohol, and ID. Notably, we opted to exclude the ID variable from our analysis, as it merely serves as an indicator for individual wines.

Our initial data exploration revealed the presence of 38 missing values instances (NA), and 2179 duplicates, which were promptly removed.



3. Method

We started our data analysis by checking and fixing the dataset's integrity, removing duplicates, and missing values. This preliminary process allowed us to get a better understanding of the dataset while making sure that our subsequent analyses are reliable. Consequently, we delved into the exploration of correlations among variables, where we found some important correlation between predictors, but we chose to keep them as they didn't make any significant change in our results.

During our analysis, we made a crucial decision about our modeling approach. Having a dataset composed exclusively of numerical variables, we faced the choice between constructing a linear regression model or opting for a classification model. This decision was informed by the nature of the data and our overarching analytical goals. Ultimately, we chose to adopt a classification analysis, with a focus on predicting wine quality based on its chemical characteristics. To facilitate this analysis, we splitted wines into two groups: "bad wines" with a score below 6 and "good wines" with a score of 6 or higher. This transformation allowed us to convert the target variable, quality, into a binary variable.

Considering our dataset, we identified the variable ID as irrelevant for our analysis and consequently chose to exclude it. Following a visual exploration of the variables, we initiated the classification process.

Opting for logistic regression, we leveraged its suitability for modeling the probability of a binary outcome. In the context of our wine quality analysis, logistic regression enabled us to model the probability of a wine being of good quality based on its chemical characteristics, facilitating classification into the "good" or "bad" quality categories.

4. Implementation

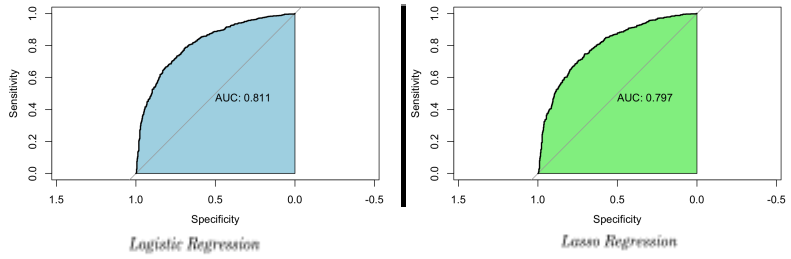
The dataset, originally comprising 7640 observations, underwent a reduction to 5427 following the removal of instances with missing values (NA) and duplicates. Subsequent visualization, including density plots, unveiled outliers, prompting their removal, and resulting in a refined dataset of 4556 total observations.

To facilitate model training and evaluation, we employed a 70/30 split, segregating the data into training and test sets with a predetermined seed for reproducibility.

Performing the Logistic Regression with *quality* as a response variable, and including all the features, we found that a lot of predictors weren't significant for the model given their related high p-value. Not being satisfied by the *full model* (the model which includes all the features as predictors), we performed a Forward Selection, running the stepAIC function, a systematic approach to selecting the most informative predictors for our model. Beginning with a null model, stepAIC tested each additional predictor, choosing to include the one that most improved the model according to the AIC. This function seeks to find the model that best explains the data with the fewest parameters, thus avoiding overfitting. The forward selection process continued until no further predictors improved the model by a notable margin. The Forward Selection provided a model with an AIC score of 3270.2, while the full model scored an AIC of 3273, enhancing the model fit by 2.8.

As our goal was to refine the predictive accuracy of our model while addressing potential overfitting, we implemented a Lasso Regression. To begin, we prepared our data by creating a matrix of predictors, excluding the response variable quality, and encoded our binary outcome quality into a dichotomous variable, with 'Good' as 1 and otherwise 0. We utilized the cross-validation function cv.glmnet, which allowed us to identify the optimal lambda parameter (the parameter on which is based the strength of the penalty). This function performs cross-validation to prevent overfitting. The lambda value that minimized the cross-validation error, lambda.min, was identified and used to fit the final Lasso model. With the optimal lambda determined, we fitted the Lasso model using glmnet, specifying the binomial family to accommodate our binary outcome. The coefficients of the resulting model were extracted, revealing the variables retained in the model and their respective weights after the Lasso's shrinkage process.

The ROC curve has been plotted for both models.



5. Evaluation

Our logistic regression model, refined through forward selection, identified Alcohol, Volatile Acidity, Sulphates, Residual Sugar, Total Sulfur Dioxide, and Free Sulfur Dioxide as the most significant predictors of wine quality. The model demonstrated a training accuracy of 74.5% and a test accuracy of 74.17%, indicating a robust fit. Notably, the model achieved a sensitivity of 0.759, implying it correctly identified 75.9% of the high-quality wines, and a specificity of 0.71, correctly identifying 71% of the lower-quality wines.

Comparatively, our Lasso Regression model yielded a slightly lower accuracy of 71.8%. Despite the modest decline in overall accuracy compared to the logistic regression, the Lasso model was instrumental in pinpointing key predictors impacting wine quality. The magnitude of the coefficients revealed that Alcohol, Volatile Acidity, and Total Sulfur Dioxide were the most influential factors, listed in order of increasing impact. The model achieved a sensitivity of 0.821, implying it correctly identified 82.1% of the high-quality wines, and a specificity of 0.585, correctly identifying 58.5% of the lower-quality wines.

6. Interpretation

The analysis made in this study yields several insights into the factors that contribute to wine quality, through their chemical characteristics. Logistic regression, enhanced via forward selection, provided key predictors that are statistically significant in forecasting wine quality. The role of Alcohol in determining quality, highlights the importance of balance in wine's alcoholic strength, while Volatile Acidity's negative influence reconfirm its generally negative impact on sensory appeal.

The slight diminishment in accuracy observed in the Lasso Regression model, compared to the logistic regression model, invites further scrutiny. Despite its lower overall accuracy, the Lasso model's higher sensitivity suggests it may be more conservative in predicting a wine's quality as 'Good', while the specificity measures in both models indicate a relatively lower ability to identify lower-quality wines accurately.

In conclusion, the study exploits Logistic Regression and Lasso Regression to discover the complexity of wine quality prediction. The resulting models not only work as a tool for winemakers and distributors, but also contribute to the understanding of quality determinants in enology.

References

1) M.YASSER - *Wine Quality Dataset*

   https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

2) https://www.decanter.com/learn/volatile-acidity-va-45532/

3) G. James, D. Witten, T. Hastie, R. Tibshirani, et al.

   An introduction to statistical learning, volume 112. Springer, 2013.