Wine Quality Prediction Based on Logistic Regression

Dante Lee, Leyang Li, Ethan Koran, Jose Pantaleon

Jun 15, 2023

Abstract

This paper explores the application of logistic regression in predicting wine quality and identifies the most important factors that contribute to wine quality. Logistic regression with L1, L2, and elastic net regularization techniques are employed for categorical and binary classification. The results are evaluated with accuracy and confusion matrices. The most satisfactory model is L1 regression in categorical classification, achieving an accuracy of 57.5% and 97.8% with tolerance scope of 1. The weights visualization of the models indicates that volatile acidity is the most important factor in determining wine quality.

Introduction

Wine, one of the most popular alcoholic beverages, requires understanding the key factors that determine its quality. This knowledge is crucial for consumers choosing wine and manufacturers striving for high-quality production. However, the standard quality verification process in the wine industry has drawbacks. It is expensive, reliant on human experts' subjective taste tests, and prone to risks. Machine learning models offer a solution by identifying important factors and their correlation with wine quality ratings. These models can determine the best combination of factors for producing high-quality wine.

Related Works

A group of researchers, namely Sunny Kumar, Kanika Agrawal, and Nelshan Mandan, predicted wine quality with three different classification algorithms: random forest, support vector machine (SVM), and naive Bayes techniques.

The random forest algorithm constructs decision trees based on various samples and determines the classification by taking a majority vote from the ensemble of trees. SVM is a supervised learning algorithm that creates an optimal decision boundary to effectively separate classes in a given space, making it easier to classify new data points. Naive Bayes, a simpler classification technique, uses the probability of likelihood to determine if a specific region belongs to a certain class and has an advantage in speed in making predictions. Among the three techniques, SVM yielded the best results for this group, achieving an accuracy of 67.25% (Kumar et al., 2020).

In a separate approach, S. Aich et al. used different feature selection algorithms such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) to identify the most relevant features. They then employed nonlinear decision tree-based classifiers including Recursive Partitioning Decision Tree (RPART), C4.5, PART, Bagging Classification and Regression Tree (Bagging CART), Random Forest, and Boosted C5.0 to analyze the performance metrics. Their results demonstrated accuracies ranging from 94.51% to 97.79% when utilizing different feature sets with the Random Forest classifier (S. Aich et al., 2018). The new model utilizes a classification algorithm that differs from the previous model, however, its objective remains unchanged.

Data

When thinking about the qualities that professional wine connoisseurs consider in determining what makes a good wine, the ones that typically come to mind include superficial qualities such as its taste, aroma, quality of the grapes used, color, etc. However, in this project, the model analyzed some of the physicochemical properties of wine that give it these qualities that professionals use to judge the drink all over the world. The data set used to accomplish this

was found on kaggle, and analyzes variants of the Portuguese “Vinho Verde” red wine. This dataset takes 11 chemical properties of each wine variant, and uses them as inputs to give each wine a “grade” on a scale of 0 to 10. It should be noted that a majority of the quality grades given fall between 3 and 8 (shown in fig. 1), which could have an effect on the model when attempting to predict wine quality. The parameters considered in the data set include: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol content. Each of these properties contribute to the taste, smell, color, and overall quality of each bottle of wine. For example, volatile acidity refers to the amount of acetic acid in the wine, which in too high of quantities can give the wine an unpleasant vinegar taste. On the other hand, small amounts of citric acid in a wine can add a hint of freshness to a bottle’s flavor. When each of these properties are put together in the right balance, the quality of the wine can be determined, which will be the goal of the machine learning model.

To visualize the data, bar distribution graphs were created for each chemical property vs the quality grade that each wine was given in the data set, a sample of which are shown in Fig. 1, below. In analyzing these data distributions, correlations between certain properties and their corresponding wines’ quality grade seemed to emerge. For example, it seemed as though as the volatile acidity of the wine increased, the quality of the wine also increased. Conversely, the bar distribution graphs showed an inverse relationship between the wines’ citric acid content and its quality grade. Thus, it can be reasonably predicted that properties such as these will have a significant impact in our determination of wine quality. On the other hand, some chemical properties explored in the data set seemed to have an arbitrary relation to the wines’ quality. These properties, such as residual sugar content and pH, were predicted to have little to no effect

on our model. Thus, regularization should be used to drive any unnecessary variables down to zero.

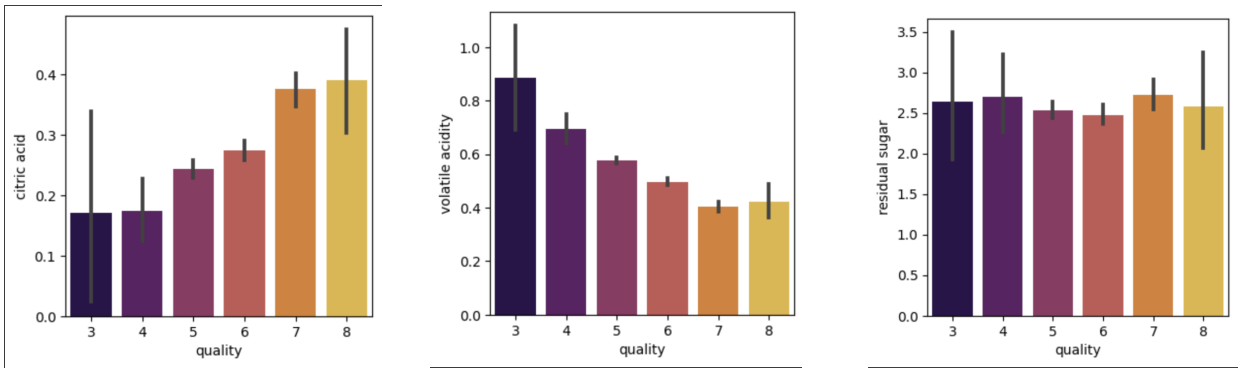


Fig. 1: Example Bar Distribution Graphs for the data set

Method

Logistic regression is applicable in the classification scenario (Pampel, 2000). The data is first downloaded as a dataframe, scaled, and split into test (20%) and train (80%) data. Then, regularization is used to avoid overfitting. The most accurate model can be found by comparing the performance with L1, L2, or Elastic net regularization in the same train and test data. The accuracy is calculated by dividing the correct prediction number by the total prediction number. Both categorical and binary models are trained. The results are visualized and analyzed using confusion matrices, and the best model can be selected. Finally, the weights of the model are visualized so that the most important factors are displayed.

The model is trained via sklearn library and all data visualizations are achieved via matplotlib library from Python. Stochastic Gradient Descent (SGD) optimizer is used for models with Elastic net of different L1 ratios. SGD iteratively updates the model parameters by computing the gradients of the loss function on randomly selected subsets of the training data (Bottou, 2010). The randomness of SGD allows the model to achieve a better performance.

Experiments

For categorical classification, lasso (L1), ridge (L2), and elastic net regressions. Model with L1 regression achieves 57.5% accuracy, outperforming L2 regression (57.2%) and elastic net regression (55.6%). The confusion matrices shown in Fig. 2 demonstrate more detailed results. The models achieve satisfactory accuracy with a tolerance scope of one: 97.8% for L1, 97.5% for L2, and 96.3% for elastic net. Different L1 ratios for elastic net are used, but all achieve the same accuracy, as Fig. 3 suggests.

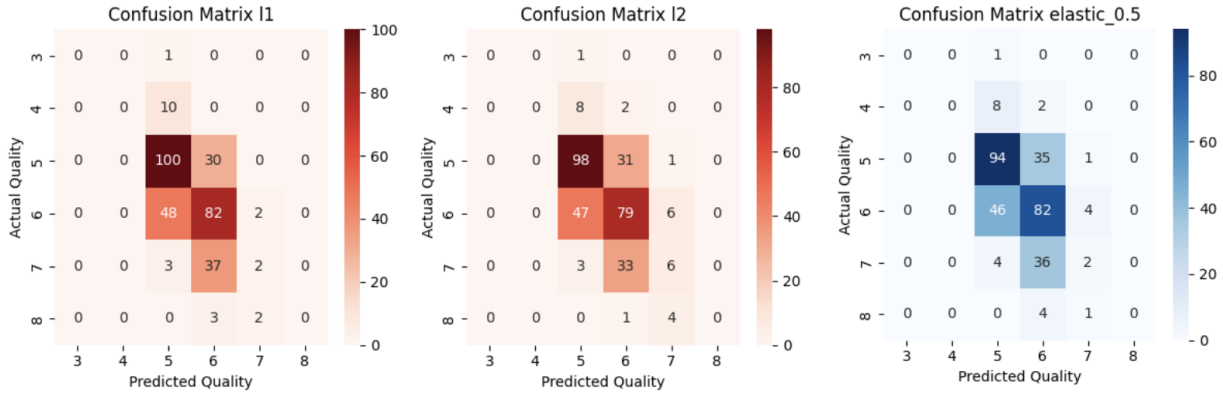


Fig. 2: Confusion matrices for categorical models with L1, L2, and elastic net regression

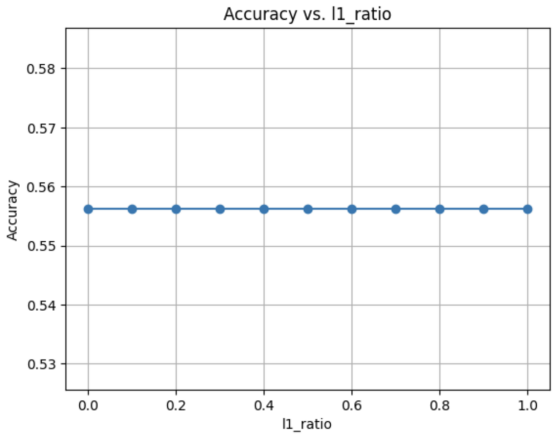
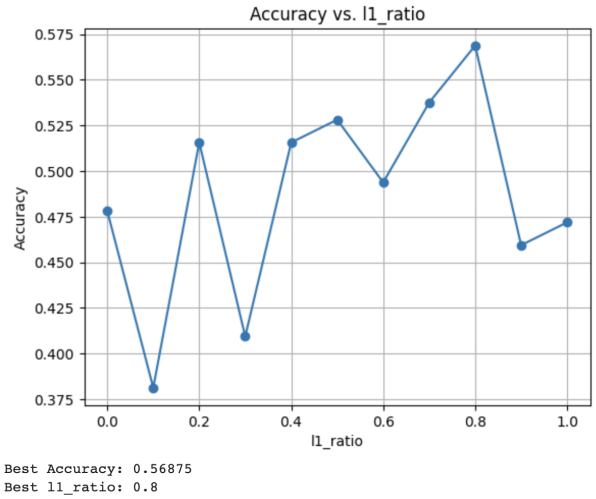


Fig. 3: Accuracy achieved with different L1 ratios

SGD optimizer is employed in elastic net regression models to achieve a randomized training process. Fig. 4 demonstrates the fluctuated accuracy achieved with various L1 ratios and

Fig. 5 shows the confusion matrix of the best model which achieves an accuracy of 56.9% and 96.3% with a tolerance scope of one.



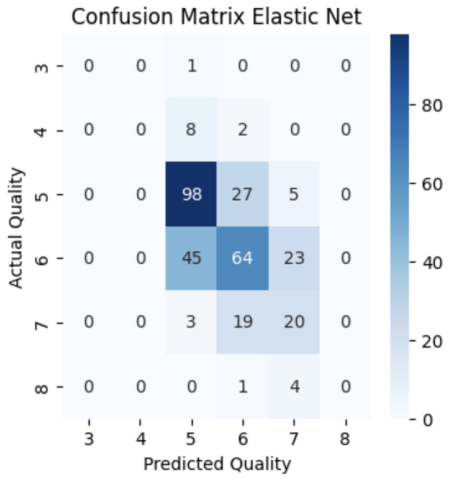


Fig. 4: Accuracy achieved with SGD different L1 ratios

Fig. 5: Confusion matrix

In terms of accuracy, L1 (Lasso regression 57.5%) outperforms L2 (Ridge regression 57.2%) and Elastic net regression (56.9% maximum). The confusion matrices show models tend to overestimate wines of quality 3 and 4 and underestimate wines of quality 7 and 8. This is speculated to be due to the distribution of corresponding data in the dataset..

The comparison of confusion matrices (Fig. 2 and Fig. 5) suggests that the SGD Classifier's top-performing model demonstrates superior performance in predicting quality 7. This could be attributed to the fortuitous alignment of randomly generated subsets, which tend to favor quality 7 predictions.

The same approach is used in binary classification with 6 as threshold (quality of 3, 4, 5 are labeled as “bad” and 6, 7, 8 as “good”). Models with L1, L2, and elastic net regularization (with and without SGD) are trained and compared. The results are shown in Fig. 6 and Fig. 7. L2 (74.1%) outperforms L1 (73.8%) and Elastic net regression (72.5% maximum with SGD).

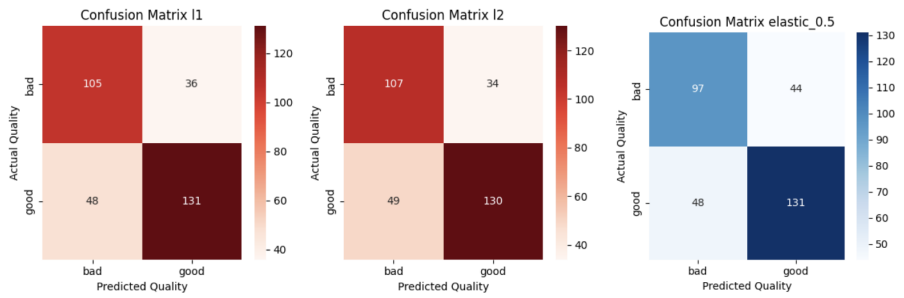


Fig. 6: Confusion matrices for binary models with L1, L2, and elastic net regression

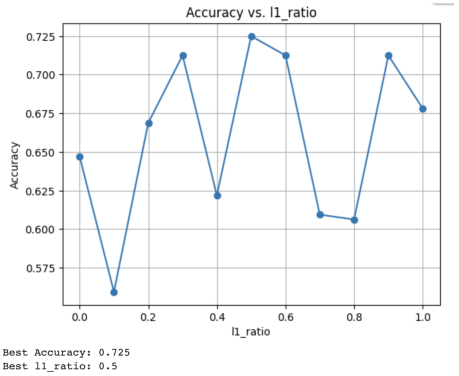


Fig. 7: SGD accuracy vs. L1 ratios

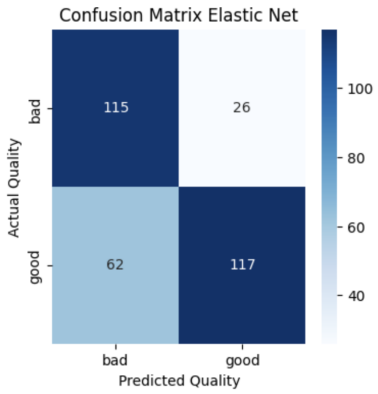


Fig. 8: Confusion matrix (L1 ratio 0.5)

Binary classification (L2, 74.1%) achieves a significant higher accuracy than categorical classification (L1, 57.5%, L2, 57.2%). However, categorical classification has more reference value with tolerance scope of 1 (L1, 97.8% and L2, 97.5%).

According to the weights visualization of the L1 and L2 categorical model (Fig. 9), volatile acidity is considered most important, followed by alcohol.

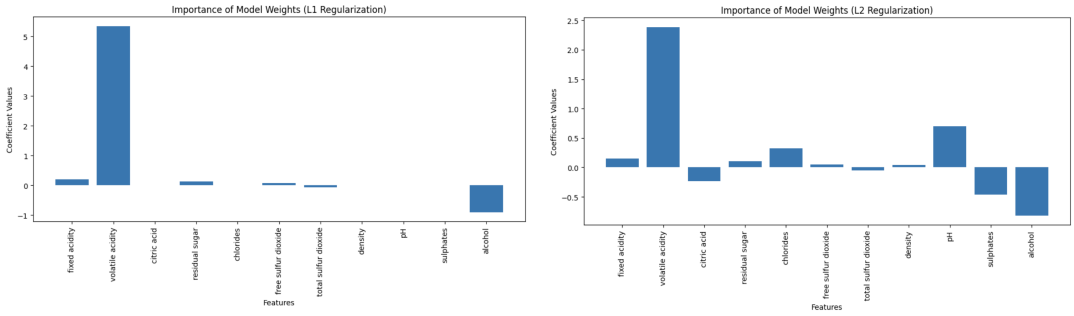


Fig. 9: Weights for categorical models with L1 and L2 regression

This matches the data set visualization as there is a significant correlation between volatile acidity and quality, shown previously in Fig. 1. Although a correlation is also present in citric acid, there are overlaps, especially between quality 5 and 6. Therefore, it's reasonable for the models not considering citric acid.

Conclusion

For wine quality prediction, different approaches of logistic regression (different regularization methods and SGD optimizer) are employed and compared. The models achieved satisfactory accuracy, with L1 regression outperforming other techniques in categorical classification, and L2 regression achieving the highest accuracy in binary classification. The weights visualization revealed that volatile acidity and alcohol were the most significant factors contributing to wine quality. These findings can provide valuable insights for wine producers and consumers in understanding the key determinants of wine quality.

The research has limitations in predicting wines of quality 3, 4, 7, and 8. Further research can use a data set with more wine data of these qualities. Other factors and models can also be considered to enhance the accuracy of wine quality prediction.

Works Cited

Dataset: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

https://www.diva-portal.org/smash/get/diva2:1574730/FULLTEXT01.pdf

https://link.springer.com/chapter/10.1007/978-981-99-0769-4_14

Bottou, Léon. "Large-Scale Machine Learning with Stochastic Gradient Descent." Proceedings of COMPSTAT'2010, Physica-Verlag HD, 2010, pp. 177–86, https://doi.org/10.1007/978-3-7908-2604-3_16.

Kumar, S., Agrawal, K., Mandan, N., 2020. Red Wine Quality Prediction Using Machine Learning Techniques, in: 2020 International Conference on Computer Communication and Informatics (ICCCI). Presented at the 2020 International Conference on Computer Communication and Informatics (ICCCI), IEEE, Coimbatore, India, pp. 1–6.

Pampel, Fred C. Logistic Regression : a Primer. SAGE, 2000.

S. Aich, A. A. Al-Absi, K. L. Hui, J. T. Lee and M. Sain, "A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 139-143, doi: 10.23919/ICACT.2018.8323674.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.