Optimizing Wine Quality Prediction Using Machine Learning Models

Qianshuo Wang STATS201
Instructor: Luyao Zhang

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

1. Background and Motivation

- **Problem or Research Gap:** Traditionally, wine quality prediction relies heavily on geographical factors, and consumers often believe that wines from certain regions are of higher quality. However, the chemical composition of wine, such as acidity, alcohol content, sugar levels, etc., may have a more direct impact on wine quality. Existing models often overlook this factor, focusing instead on regional advantages as the primary indicator of quality.
- **Research Significance:** The aim of this study is to highlight the importance of wine composition in predicting quality, challenging the idea that geographical origin is the only determinant of wine quality. By using machine learning models to analyze the relationship between the chemical composition of wine and its quality, we can more accurately predict wine quality without relying solely on its origin.

2. Research Questions

1. Which wine components (such as acidity, alcohol content, sugar levels, etc.) are the most important for predicting wine quality?
2. How can we improve the accuracy of wine quality prediction by focusing on chemical composition instead of geographical origin?

3. Dataset

Dataset Description:

This study utilizes a Wine Quality Dataset, which contains multivariate data about various chemical properties of red and white wines. The dataset is designed to support machine learning models focused on wine quality prediction. The goal is to demonstrate that wine composition, rather than geographical region, is the most critical factor in determining wine quality.

Dataset Characteristics:

- **Data Type:** Multivariate dataset containing chemical compositions of wines along with corresponding quality ratings.

Data Preprocessing:

- Missing Values: The dataset is relatively clean, with minimal missing values. Any missing values, if present, are handled by imputing with mean/median values based on the feature distribution.

4. Methodology

XGBoost (Extreme Gradient Boosting):

XGBoost is employed to predict wine quality, a multi-class classification task. It is highly suitable for this problem because it can capture the complex, non-linear relationships between features, such as the chemical composition of wine and its quality. XGBoost also offers advantages such as feature importance ranking and efficient handling of missing data. The model is trained using the default settings of XGBoost, but hyperparameter tuning can be done to improve performance.

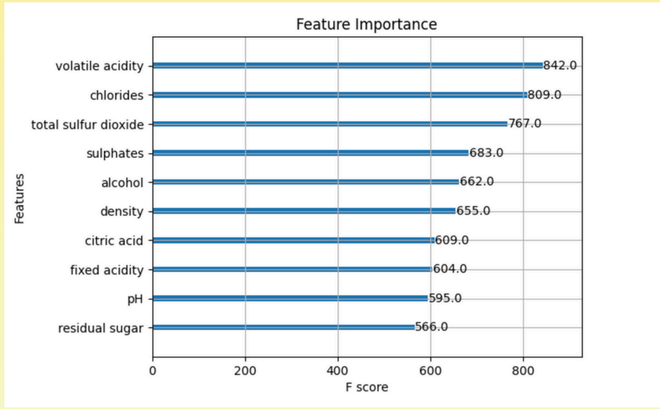
SHAP (SHapley Additive exPlanations):

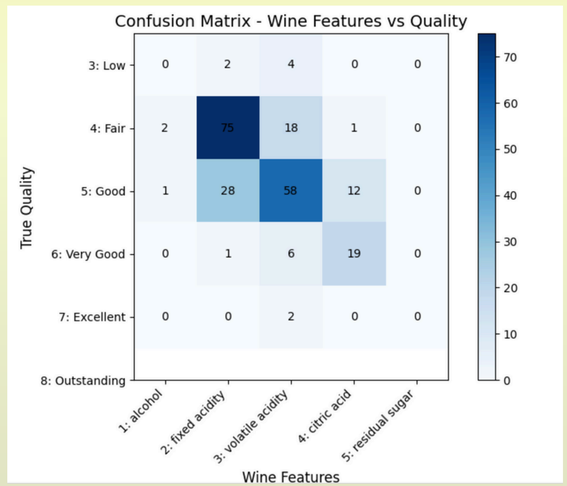
SHAP helps to determine which ingredients play a major role in predicting wine quality by assigning an independent contribution value to each characteristic. It helps to understand how each input characteristic (such as acidity, alcohol, pH, etc.) affects the prediction of wine quality.

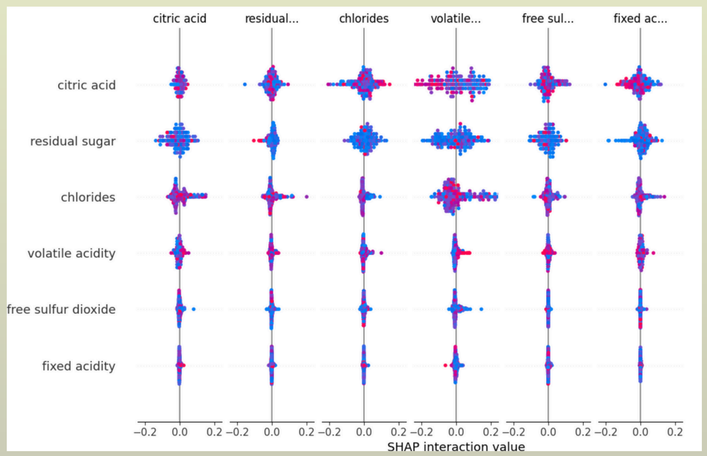
Confusion Matrix:

The confusion matrix will be used to evaluate the classification performance of our machine-learning models. Given that we are predicting wine quality (e.g., excellent, good, poor), the confusion matrix will show the number of true positives, true negatives, false positives, and false negatives for each class. This helps in understanding how well our model distinguishes between different quality categories.

Visualizations







Results

XGBoost: volatile acidity has the biggest impact on the quality of a wine, reaching 842, while **residual sugar** has the smallest impact, only 566. What's more, **chlorides, total sulfur dioxide** also show a relatively high impact on wine quality.

SHAP: Feature position: volatile acidity has a high SHAP value distribution, indicating that it is one of the important characteristics affecting the quality of red wine. **free sulfur dioxide** has a small distribution of SHAP values, indicating that it has a weak influence on quality.

SHAP value range: The distribution shows a wide range of SHAP values, which may range from -0.2 to 0.2, indicating that volatile acidity has a significant positive and negative effect on quality. While for **free sulfur dioxide,** SHAP ranges mostly around 0, and sometimes goes to 0.1, indicating this factor barely affects the quality of the wine.

Color distribution: Red dots indicate higher volatile acidity values and the corresponding SHAP values are mostly negative, indicating that higher volatile acidity has a negative impact on the prediction of wine quality. This is true, too high volatile acidity will make the wine taste poor, even acetic acid taste.

Discussion

Intellectual Merits

This study provides valuable contributions to the application of machine learning in the wine industry by integrating advanced techniques such as SHAP (SHapley Additive exPlanations) and XGBoost to analyze the relationship between wine chemical compositions and quality. Key intellectual contributions include:

- **Identification of Key Features:** Using XGBoost and SHAP to determine that volatile acidity has the greatest impact on wine quality, providing a more precise understanding of the chemical factors influencing wine quality compared to traditional methods that rely on regional origin.
- **Advancing Wine Quality Prediction:** By integrating machine learning methods into wine quality prediction, this study challenges traditional quality control practices in the wine industry, providing a foundation for future research on predictive analytics in the food and beverage sector.

Practical Impacts

- The findings from this study have significant practical applications for winemakers, quality control, and consumers:
- **Informed Wine Selection:** With the identification of volatile acidity as the most important factor affecting wine quality, consumers can make more informed choices when selecting wines. This knowledge can be used to assess wine labels, product descriptions, and reviews, helping consumers understand which chemical properties are most likely to result in a higher-quality product. For example, consumers looking for wines with balanced acidity or smoother textures can focus on wines that have optimized levels of volatile acidity.
- **Consumer Preferences and Market Insights:** Understanding the chemical composition's direct effect on wine quality enables winemakers to align their products with consumer preferences, particularly in targeting markets that prioritize certain flavor profiles. Moreover, the findings can help wine brands market their products more effectively by highlighting key ingredients that are scientifically proven to enhance quality.

References

1. Chen, Tianqi, and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. https://arxiv.org/abs/1603.02754.
2. Jain, K., Kaushik, K., Gupta, S.K. et al. Machine learning-based predictive modelling for the enhancement of wine quality. *Sci Rep* 13, 17042 (2023). https://doi.org/10.1038/s41598-023-44111-9.
3. Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems* 30 (2017): 4765–4774. https://arxiv.org/abs/1705.07874.
4. Markoulidakis, Ioannis, and Georgios Markoulidakis. 2024. "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis" *Technologies* 12, no. 7: 113. https://doi.org/10.3390/technologies12070113