Predicting Wine Quality Using Advanced Tree-Based Methods

Authors: Karthik Kuppala

Github: https://github.com/KarthikKuppala/Wine-Quality-Prediction-XGBoost/tree/main

Abstract

This research intends to create predictive models for accurately evaluating wine quality through machine learning approaches. Although decision trees offer a fundamental insight, this project utilizes sophisticated tree-based methods, notably XGBoost, to improve predictive robustness and reduce overfitting

1. Introduction

Assessing wine quality is a complicated task shaped by the interaction of various chemical traits. The main goal of this project is to create a model that can effectively forecast these quality evaluations.

The research initiates with an extensive Exploratory Data Analysis (EDA) to comprehend the dataset's structure and detect possible problems like missing values or outliers. This stage includes visualizing data distributions and producing descriptive statistics to guide the modeling process. While decision trees provide an interpretable baseline, they frequently risk overfitting. Therefore, this study emphasizes XGBoost, a sophisticated ensemble method aimed at enhancing performance and consistency

2. Methodology

2.1 Data Acquisition and Preparation

The analysis utilized a pooled dataset consisting of both red and white wine samples. The data preparation process involved:

- **Data Integration:** Merging separate datasets for red and white wines (train_red, train_white) into a single training set, and similarly for the testing set.
- **Partitioning:** The target variable y was defined as the 'quality' column. [cite_start]The data was split using a partition where 80% was allocated for training (X_train) and 20% for validation (X_val) to ensure robust model evaluation.

2.2 Model Configuration

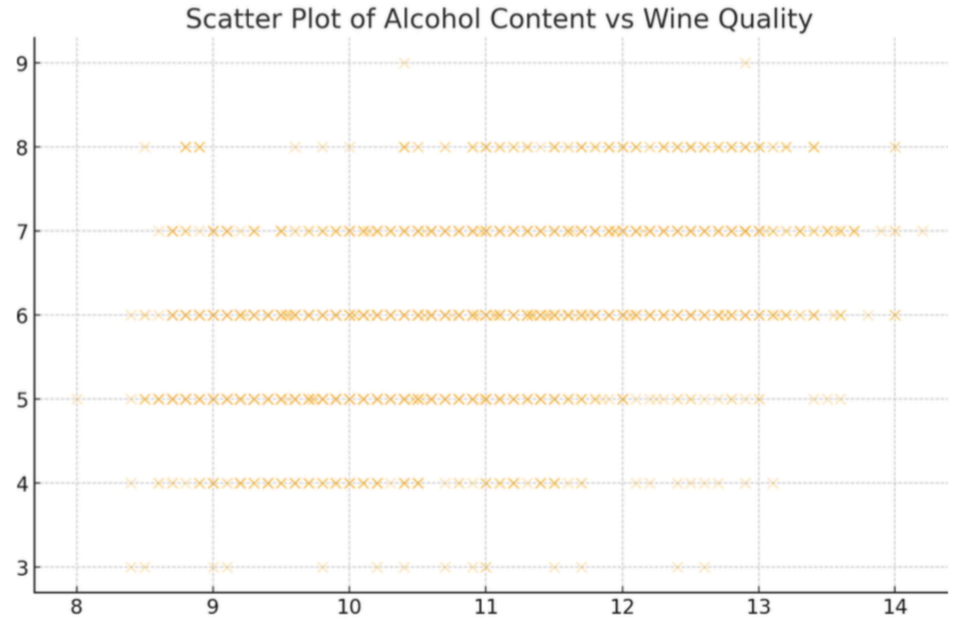
The project utilized the xgboost library in R for modeling. [cite_start]The model was configured as a regression task using the reg:squarederror objective function to minimize the root mean squared error.

Key hyperparameters for the XGBoost model were set as follows:

- **Maximum Depth:** 6
- **Learning Rate (Eta):** 0.1
- **Number of Rounds:** 50
- **Evaluation Metric:** RMSE (Root Mean Square Error)

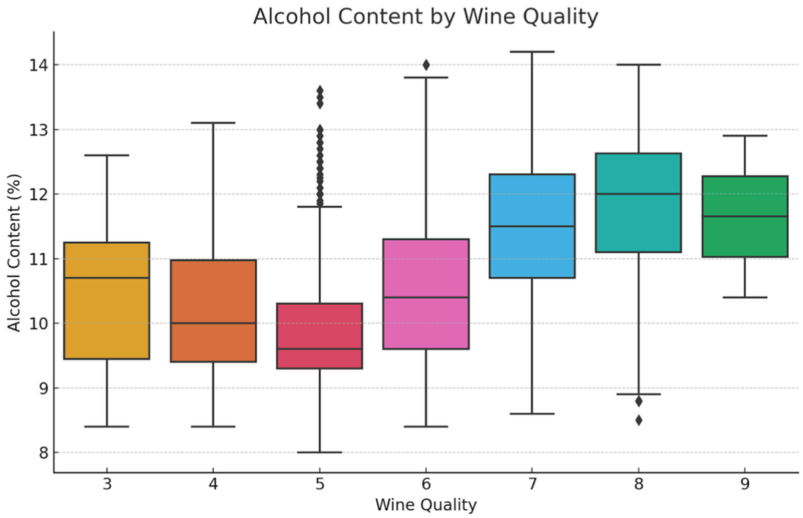
3. Results and Analysis

3.1 Performance of the Model



The last model submission achieved an accuracy score of 0.60652. This metric acts as a baseline measure of performance, showing that although the model can make correct predictions over half the time, there is substantial potential for enhancement. The computed RMSE on the validation dataset was used to measure the prediction error.

3.2 Feature Evaluation: Alcohol Concentration Compared to Quality



A major part of the research examined the relationship between wine quality and alcohol content.

- Distribution: Boxplots indicated that wines of higher grades typically have a slightly elevated median alcohol content.
- Variance: Although the median changed, the variability (variance) of alcohol content stayed fairly stable across all quality tiers.
- Outliers: Significant outliers were noted in wines with lower ratings, suggesting that certain low-quality wines might still have unusually high or low alcohol levels.
- Correlation: In general, the findings indicate a slightly positive connection between increased alcohol levels and improved wine quality.

4. Discussion

4.1 Analysis of Results

The accuracy near 0.60 indicates the intrinsic challenge of forecasting wine quality, influenced by a complicated interaction of chemical factors that are hard to fully encompass. Nonetheless, the application of visualizations effectively showcased important trends, like the favorable correlation between alcohol content and quality ratings.

4.2 Suggestions for Upcoming Tasks

To enhance predictive precision, forthcoming studies should concentrate on:

- Hyperparameter Adjustment: Modifying parameters for XGBoost more intensely to enhance performance.
- Feature Engineering: Exploring additional features or interaction terms that could more effectively capture the chemical intricacies of the wine.
- Enhanced Ensembles: Assessing the effectiveness of XGBoost against other advanced techniques, like Random Forests, to identify the best modeling strategy

5. Limitations

Predictive Accuracy: The ultimate model reached an accuracy of around 0.60652. This figure suggests that although the model offers a helpful starting point, it accurately predicts just over half of the time, indicating substantial opportunities for enhancing reliability.

Complexity of Chemical Interactions: The quality of wine is defined by a "complex interaction of various chemical properties." [cite_start]The existing modeling technique might not adequately reflect these intricate chemical details, restricting the model's effectiveness in differentiating between similar quality classifications.

Algorithmic Scope: While the research progressed from simple decision trees to employing XGBoost to mitigate overfitting, the examination suggests that this sole sophisticated technique might still be inadequate.

The report indicates that even more advanced ensemble techniques or different models such as Random Forest may be required to manage the complexity of the data more effectively.

Data Outliers: Exploratory analysis identified outliers, especially in lower-rated wines regarding alcohol levels. These extreme values suggest that certain wines have unusually high or low alcohol concentrations that deviate from the overall trend, likely introducing noise that the current model finds difficult to interpret.

6. Conclusion

Ultimately, this research employed a combined dataset of red and white wines to explore the relationship between chemical characteristics and wine quality. Utilizing Exploratory Data Analysis and XGBoost modeling, we discovered trends indicating a slightly positive association between increased alcohol content and enhanced wine quality. The implementation of visualizations, such as boxplots and scatter plots, was crucial in making these results understandable and showcasing the distribution of important variables.

Regardless of these findings, the ultimate accuracy score of 0.60652 highlights the challenges of reliably forecasting wine quality solely based on the existing feature set and model design. These findings establish a foundation for further investigation, suggesting that upcoming studies must emphasize feature engineering and hyperparameter optimization to more effectively capture the factors influencing wine quality. In the end, evaluating these outcomes alongside other models like Random Forest may result in the development and refinement of more effective predictive tools

7. References

https://github.com/KarthikKuppala/Wine-Quality-Prediction-XGBoost/tree/main