Jason Dean

DSC680-T301 (2241-1)

November 17, 2023

Project 3 Final Paper

Business Problem

In addressing the challenges faced by the wine industry, I focused on the critical issue of quantifying wine quality. The industry has long grappled with subjective methods of quality assessment. My objective is to explore the application of data science techniques to predict wine quality, aiming to provide a more objective and consistent approach to quality assessment.

Background/History

Traditionally, wine quality has been assessed by sommeliers and experts, relying heavily on sensory evaluation. Over the years, as wine production has incorporated more technology and science, there has been growing interest in correlating objective physicochemical properties of wine to its perceived quality.

Data Explanation

The dataset underpinning this analysis comprises physicochemical attributes of various wines, each paired with a quality rating. The preparation of this data involved thorough cleaning, normalization, and partitioning into training and test sets. This preparation was critical to ensure the reliability and validity of the subsequent analysis.

Methods

I attempted multiple predictive models and after fine tuning the best results came from the random forest classifier. This model was chosen for its proficiency in handling datasets with a mix of variable types and capturing intricate patterns. To address the class imbalance in wine quality ratings, I implemented oversampling techniques.

Analysis

Through detailed analysis, I identified that certain attributes, notably alcohol content and volatile acidity, played pivotal roles in determining wine quality. The Random Forest model, after addressing the class imbalance, achieved an impressive accuracy of 94.1%.

Conclusion

My analysis revealed that certain features, particularly alcohol content and volatile acidity, were significant predictors of wine quality. Utilizing these insights, I was able to enhance the Random Forest model's accuracy. The final model achieved an accuracy of 94.1%, indicating a high level of reliability in predicting wine quality.

Assumptions

- The physicochemical attributes selected encompass the major factors affecting wine quality.
- The quality ratings in the dataset are assumed to be representative of broader consumer perceptions.

Limitations

The primary limitation of this model lies in its dependency on the characteristics of the dataset, which may not encompass all wine varieties. Furthermore, the inherent subjectivity in individual taste preferences can lead to discrepancies between the model's predictions and personal evaluations.

Challenges

- The dataset exhibited a class imbalance, with fewer samples of low and high-quality wines compared to medium quality.
- Distinguishing between close quality ratings (e.g., 5 vs. 6) proved more challenging than broader differences.

Future Uses/Additional Applications

Beyond its current application, this model could be adapted for other beverages or foods where quality is paramount. Additionally, combining this model with sensory data could lead to even more robust quality prediction systems.

Recommendations

For wine producers seeking consistent quality evaluations, integrating this model into the production process could be invaluable. Additionally, focusing on key attributes like alcohol content could guide production decisions.

Implementation Plan

- Integrate the model with the existing quality assessment infrastructure in wineries.
- Continuously collect data on new wine batches, refining and retraining the model for enhanced accuracy over time.

Ethical Assessment

While data-driven predictions offer objective assessments, it's imperative to treat them as complementary to traditional methods, ensuring wines with unique profiles aren't marginalized. Transparency in how quality scores are derived can also help in gaining consumer trust.

References

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

UCI Machine Learning Repository: Wine Quality Data Set.

https://archive.ics.uci.edu/dataset/186/wine+quality