Project 4: VinoVista

Problem Statement

VinoVista, a renowned winery, is committed to producing consistently exceptional wines. However, they have observed variability in the quality of their white wine batches. To address this challenge, VinoVista seeks a predictive model that can accurately assess the quality of wine batches based on their chemical properties before bottling. ** VinoVista is a complete fictious company made up for ease of creating a case study and deliverables for a company

Project Goal

The primary objective of this project is to develop a robust machine learning model capable of predicting wine quality on a scale of 0-10. This model will utilize a dataset containing various chemical properties of wine, such as acidity, pH, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulfates, and alcohol content.

Our Features

1. Fixed Acidity

Effect: Represents non-volatile acids in wine, such as tartaric, malic, and citric acids. It contributes to the overall acidity and freshness of the wine.

Impact on Taste: Higher levels enhance tartness and crispness, which can be desirable in white wines. However, too much fixed acidity can make the wine taste overly sour.

2. Volatile Acidity

Effect: Mainly refers to acetic acid, which can turn into vinegar if present in high concentrations. It's a measure of the acids that can evaporate and contribute to the aroma.

Impact on Taste: Low levels can add complexity, but high levels result in unpleasant, vinegary flavors. Acceptable limits for volatile acidity are often higher in red wines compared to whites.

3. Citric Acid

Effect: Naturally found in small amounts in wine, it adds freshness and enhances the fruitiness of the wine.

Impact on Taste: Gives a slight sharpness and can brighten the flavor profile, especially in white wines, making them taste fresher.

4. Residual Sugar

Effect: The sugar left after fermentation. It's a key factor in determining the sweetness level of the wine.

Impact on Taste: High residual sugar leads to sweeter wines, while low residual sugar results in drier wines. It can also balance acidity, making the wine taste smoother.

5. Chlorides

Effect: Reflects the salt content in wine, often derived from the soil in which the grapes were grown.

Impact on Taste: Adds body and mouthfeel. However, high levels can make the wine taste salty or brackish, which is generally undesirable.

6. Free Sulfur Dioxide (SO_2)

Effect: Acts as an antioxidant and antimicrobial agent, protecting wine from oxidation and spoilage.

Impact on Taste: Adequate levels preserve the wine's freshness and prevent spoilage. However, excessive sulfur dioxide can impart a pungent, chemical-like taste and aroma.

7. Total Sulfur Dioxide

Effect: The sum of bound and free SO_2 in wine. It's a crucial preservative used during the winemaking process.

Impact on Taste: While necessary for stability, high levels can cause off-flavors and are a common cause of headaches for sensitive drinkers.

8. Density

Effect: Often used to estimate the sugar content in wine. A higher density indicates a higher sugar level, which can suggest sweetness.

Impact on Taste: Affects mouthfeel; wines with higher density often feel fuller or richer. It's also an indicator of the alcohol and sugar content balance.

9. pH

Effect: Measures the acidity level. A lower pH means higher acidity, contributing to the wine's stability and shelf life.

Impact on Taste: Lower pH wines taste sharper and more acidic, while higher pH wines can taste flabby or less vibrant. Most wines have a pH between 3.0 and 4.0, with whites generally being more acidic (lower pH) than reds.

10. Sulphates

Effect: Often added to wine as a preservative to prevent oxidation and microbial growth.

Impact on Taste: Can enhance the wine's flavor profile and longevity but, in excess, may lead to bitter or metallic notes.

11. Alcohol

Effect: Produced during fermentation, where sugars are converted to ethanol. Alcohol content is a key factor in the body and warmth of the wine.

Impact on Taste: Higher alcohol levels contribute to a fuller body and a warming sensation. However, if the alcohol is too high relative to other components, it can make the wine taste hot or unbalanced.

Extraction

We extracted data from UC Irvine (*Wine Quality* - Dataset from UC Irvine Machine Learning Repository - https://archive.ics.uci.edu/dataset/186/wine+quality) as a CSV and used pandas to read it in Jupyter Notebook.

Transform

We began by checking for null values in our dataset and found none across all 4,898 rows. Through research, we discovered that in wine production, Total Sulfur Dioxide levels are typically static and rarely exceed 150. Based on this, we identified any values above 150 as likely quality mishaps or data entry errors. Removing these outliers reduced the dataset to 3,090 rows. To improve overall accuracy, we decided to simplify the target variable by categorizing wine quality into two groups: "good" and "not good." To determine the cutoff between these categories, we analyzed the overall distribution of quality ratings to make a data-driven decision.

(insert Jacks graph)

We observed an overabundance of quality scores of 6 in our dataset. Based on this, we defined "good" wines as those with a quality score of 7 or higher. This categorization resulted in 2,246 entries classified as "bad" and 844 as "good." When comparing the means of various features between these two groups, we found that **residual sugar** (bad: 5.23, good: 4.41) and **free sulfur dioxide** (bad: 27.71, good: 31.92) showed the most significant differences. These findings suggest that the distinction between a "bad" and "good" wine is subtle and influenced by small variations in key chemical properties.

Some further exploration showed that wines pH should always be between 3 and 4.

(insert graph (Distribution of Wines by Ph Range)

(insert graph(Quality Distribution (True/False) of Wines with pH Outside 3-4)

Among the small number of wines with quality scores outside the 3–4 range, the majority were classified as bad. This suggests that such scores likely result from quality mishaps rather than being representative of production norms.

To explore multicollinearity, we examined the relationships between residual sugar, density, and alcohol.

(Insert graphs: Residual Sugar vs Density with trend line & Residual Sugar vs alcohol with trend line)

Our analysis revealed that residual sugar is positively correlated with density and negatively correlated with alcohol.

The Model

For our data, XGBoost provided the best results, achieving an overall accuracy of 84%. Breaking it down, the model performed better in identifying "bad" wines (True), with a precision of 0.88 and a recall of 0.90, leading to an F1-score of 0.89. For "good" wines (False), the precision was 0.71, recall was 0.68,

and the F1-score was 0.69. These metrics indicate that while the model is more effective at identifying "bad" wines, it still performs reasonably well for "good" ones.

The F1-score is particularly useful here because it combines precision (how many predicted positives were actually correct) and recall (how many actual positives were identified) into a single measure. This makes it a balanced metric for evaluating the model, especially when there's an imbalance in the classes, as seen in our dataset.

Considerations

Our dataset comes from the food industry, where quality ratings from tasters are inherently subjective. Despite this, we believe our model's results are very strong. Additionally, the relatively small size of our dataset may contribute to a slight loss in accuracy. The overrepresentation of bad wines in the data has also led to an accuracy imbalance, particularly when predicting good wines.

Conclusion

Our model excels at predicting bad wines. VinoVista, a well-established company, primarily produces wines with quality ratings of 6. In our effort to push for higher quality, the model could be leveraged for rapid, small-batch research and development. We also recommend using the model for quick packaging and customer decision-making. Quality ratings from tasters can take time to receive, but our model allows for swift detection and prediction of bad wines, which are most likely to receive a rating of 6. These wines could then be allocated as table wine batches, targeting high-volume, lower-quality customers more efficiently.

This approach could lead to increased productivity, enhanced customer service, and reduced delays in awaiting test results. While the slightly lower accuracy for good wines means confirmation from tasters is still necessary, the model's strategic application can significantly improve order fulfillment and streamline R&D processes overall.