Wine Study

Team 2: M. Borunda, K. Childers, and F. Valdez

Southern Methodist University

Data Science Bootcamp (09/2023)

**Introduction**

This study explores wine variants of “Vinho Verde” wine from Portugal. Team 2 was interested to find any correlation between these wine variants and their features which included fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality score. A database of these features was developed by combining two datasets for machine learning. The machine learning would help with developing a predictive model on wine type (red or white) based on the features. The results would educate the consumer on what aspects make a high-quality wine and its wine type. A third data set was used to bring in cost, region of origin, and rated points to enhance findings from the first and second datasets and determine price points and regions of high-quality wine.

# Data Cleaning

The first two datasets were transformed for machine learning by first cleaning and combining them. The first data set brought in red wines and the second one consisted of white. To combine them, team two had to reformat the second set to match the first. Then a column had to be added to each data set to label the wines as red or white.

A screenshot of a data

Description automatically generated

**Figure 1: Original read in of dataset 1 & 2**

**A screenshot of a computer

Description automatically generated**

**Figure 2: Added wine type columns to dataset 1 & 2**

**A screenshot of a computer

Description automatically generated**

**Figure 3: Combined datasets 1 & 2**

The range of the data was not extreme, and scaling was not needed. It did look like there was a large skew due to the number of white wines. After analysis, about 93% of the data fell between the quality rating 5-7. The features “fixed\_acidity” and “alcohol” seemed to better determents of wine type. When looking for correlation between the wine features and quality the most correlation that could be seen (it was from low to moderate) was “alcohol”, “density”, “volatile\_acidity”, and “chlorides” as a close runner up with -0.2.  
A colorful squares with white text

Description automatically generated

A computer screen shot of a number of objects

Description automatically generated

**Figure 4: Correlation between wine features**

**Findings/Analysis**

Machine learning was performed on the dataset to create a predictive model. Team 2 first performed various analysis to determine the best model to predict quality of wine. The first analysis utilized linear regression. The values that came out showed that it was not the best model to use. The R2 value was low (below 0.4), meaning there was low correlation. The MSE value was slightly higher than team 2 would prefer because the closer to 0, the better the model. In this case it was around 0.54. In addition, the RMSE was kind of high, the model is better the lower it is. Finally, the MAE value should be lower for a better fit model, but it was kind of high as well.

A number on a white background

Description automatically generated

A graph with blue dots

Description automatically generated

A diagram of a plot

Description automatically generated  
**Figure 5: Linear regression analysis to predict wine quality**

Random Forest was utilized next to determine if it would be a better fit model to predict wine quality. It had an accuracy score of 68%. It did verify that alcohol, density, and volatile were the best determinants of wine quality as seen in figure 7 below.

**A screenshot of a graph

Description automatically generated**

**Figure 6: Random Forest analysis**

**A screenshot of a computer program

Description automatically generated**

**A graph with text on it

Description automatically generated**

**Figure 7: Feature importance**

KNN analysis was the next analysis and was only a 54% accuracy. PCA was the next technique and came out okay as a predictive model.

**A number on a white background

Description automatically generated**

**Figure 8: PCA analysis**

Decision Tree Analysis was performed, and accuracy was about 61% as seen in the figure 9 below.

A screenshot of a graph

Description automatically generated

**Figure 9: Decision tree analysis**

In conclusion, team two decided that quality did not have strong enough correlations to the other wine features to use in a predictive model. Therefore, attention was switched to focus on predicting wine type as red or white based on wine features (jupyter notebook title “machine\_learning\_classification”). One-hot encoding was implemented to id the wine type as either 1 for red or 0 for white. Predicting wine type had much stronger correlations as seen in figure 11.

A screenshot of a computer

Description automatically generated

**Figure 10: Converting wine types to 1 for red and 0 for white**

A colorful squares with white text

Description automatically generated with medium confidence  
**Figure 11: Correlation with wine\_type**

Team 2 implemented SMOTE and NearMiss techniques to balance out the data due to the higher count of white wines (white: 4,898; red: 1,599). After performing SMOTE and re-sampling   
the data with logistic regression analysis the accuracy skyrocketed to 98%.

A number of numbers in a row

Description automatically generated with medium confidence

**Figure 12: Analysis after SMOTE**

Team two explored some neural network analysis for fun and found it came out with an accuracy level of 96% for one node, 97% for two, and 99% for three. Further analysis showed that specifically, “total\_sulrfur\_dioxide”, “volatile\_acidity”, and “chlorides” were the best indicator of wine type.

A screenshot of a cell phone

Description automatically generated  
A number of numbers on a white background

Description automatically generated

**Figure 13: Stats on volatile acidity and wine type**

The XGBClassifier model was the best fit overall and Team 2 decided to use that technique as the final predictive model to determine wine type. It did seem like a bit of an overfit, but overall, it was the best model to use after comparing Logistic Regression (no imbalance of data techniques used), SMOTE, and NearMiss.

A screenshot of a computer code

Description automatically generated

**Figure 14: XGBClassifier results**

**Dashboard**

The Dashboard was created to have an approachable way for people to explore wine and its varying features that make it either a great or bad wine. The Tableau dashboard also incorporates wine regions, prices, and ratings to make the average drinker knowledgeable on what aspects make a highly rated wine. Team 2 organized Tableau into four main dashboards.

The first dashboard focused on wine types, exploring the distinct difference with alcohol, density, and citric acid that can determine if a wine is white or red.

A screenshot of a computer

Description automatically generated

**Figure 15: Wine type dashboard**

The second dashboard provided an overview of wine quality. A user can interface with the dashboard by selecting a quality to see how alcohol, volatile acidity, pH, and sulphates effected it. Although there were not strong correlations, some could be found with these features to determine a wine quality.

A screenshot of a computer

Description automatically generated

**Figure 16: Wine quality dashboard**

The third dashboard covered wine ratings from the third dataset. A user of the interface can view the average wine points per Country and which countries had the highest points for wine.

A screenshot of a computer

Description automatically generated

**Figure 17: Wine ratings dashboard**

The final dashboard displayed wine price, point ratings, and average cost of wine per country. A user can enter a price range and points to see which wine would best fit their criteria.

A screenshot of a computer

Description automatically generated

**Figure 18: Wine prices dashboard**

**Conclusion**

Overall XGBClassifier was used as the final predictive model. The logistic regression performed fine with the raw data. The team was not comfortable using it though, due to the high imbalance of red wine data points compared to white.

A screenshot of a computer

Description automatically generated

**Figure 19: Logistic regression results**

Taking the imbalance into consideration, SMOTE was used to create synthetic data points to make-up the red wine data and match the white wine amount. That equated to an oversampling amount of 3,673 datapoints each. The logistic regression performed with that data came out better as seen in the figure below.

A screenshot of a computer

Description automatically generated

**Figure 20: Model using SMOTE data**

The Near Miss technique was then used which deleted the white wine datapoints to match the red wine ones. The team was not comfortable with this approach due to the deletion of so many datapoints.

A screenshot of a computer program

Description automatically generated

**Figure 22: Model using Near Miss data**

Finally, looking at the XGBClassifier utilizing the SMOTE sampled dataset, it came out as the best model for predicting wine types.

A graph with a line

Description automatically generated

**Figure 23: XGBClassifier model using SMOTE data**

In conclusion, team 2 found the best predictive module and created a dashboard interface to guide the user on understanding what aspects effect wine. Naturally this study is intriguing because wine is within the top 10 beverages consumed globally. This study educated the average drinker on the various features of wine and how they can be used to predict wine type, price, and point ratings. Not every wine has to be super expensive to be a great drink and a multitude of aspects come together to create a perfect bottle of vino.

Works Cited

Bootswatch:

[https://bootswatch.com/flatly/](https://bootswatch.com/flatly/" \t "_blank)

Prediction of quality of Wine

<https://www.kaggle.com/code/vishalyo990/prediction-of-quality-of-wine>

Red Wine Quality:

[https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009?resource=download](https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009?resource=download" \t "_blank)

White Wine Quality:

<https://www.kaggle.com/datasets/piyushagni5/white-wine-quality>

Wine Reviews:

<https://www.kaggle.com/datasets/zynicide/wine-reviews>

W3:

<https://www.w3schools.com/html/default.asp>

Other Resources:

SMU Data Analytics Boot Camp Lessons

[hhttps://agrovin.com/en/techniques-for-correcting-wine-acidity/#:~:text=Fixed%20acidity%20is%20the%20set,of%20tartaric%20acid%20per%20litre. /](https://agrovin.com/en/techniques-for-correcting-wine-acidity/" \l ":~:text=Fixed%20acidity%20is%20the%20set,of%20tartaric%20acid%20per%20litre.          " \t "_blank) [https://www.wineenthusiast.com/basics/drinks-terms-defined/volatile-acidity-wine/](https://www.wineenthusiast.com/basics/drinks-terms-defined/volatile-acidity-wine/%0a%20%20%20%20%20%20%20%20%20%20) [https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid](https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid%0a%20%20%20%20%20%20%20%20%20%20) [https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/](https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/%0a%20%20%20%20%20%20%20%20%20%20) [https://wineamerica.org/the-magic-of-wine/wine-facts/](https://wineamerica.org/the-magic-of-wine/wine-facts/%0a%20%20%20%20%20%20%20%20%20%20) [https://www.randoxfood.com/why-is-testing-for-free-sulphite-so2-important-in-winemaking/#:~:text=The%20sulphur%20dioxide%20ions%20that,to%20health%20if%20drank%20excessively.](https://www.randoxfood.com/why-is-testing-for-free-sulphite-so2-important-in-winemaking/#:~:text=The%20sulphur%20dioxide%20ions%20that,to%20health%20if%20drank%20excessively.          )

Terminology

**Fixed Acidity: Is the set of the wine’s natural acids. It preserves the wine’s natural qualities, as well as its color.**

**Volatile Acidity: Is a measure of a wine’s gaseous acids. The amount of VA in wine is often considered an indicator of spoilage.**

**Citric Acid: It can be added to finished wines to increase acidity and give a “fresh” flavortive or additive to food or drink to add a sour taste.**

**Residual Sugar: It is the natural grape sugars leftover in a wine after the alcoholic fermentation finishes.**

**Chlorides: Depend on the geographic, geologic and climatic conditions of vine culture. The content is increased in wines coming from vineyards which are near the seacoast.**

**Free Sulfur Dioxide: Sulfur Dioxide Ions that are not chemically bonded to other chemicals. Too much sulfur in wine can cause danger to health if drank excessively.**

**Density: Mass/Volume – How “heavy” the wine is**

**pH: Influences the taste of the final product, but also its color, oxidation and chemical stability.**

**Alcohol: Chemical found in beverages such as wine, liquor, and beer.**