CS 777 – Final Project

Analysis and prediction of wine quality and type of wine

Laura Vidiella del Blanco
Big Data Analytics, December 2019

INDEX

1. *What is the dataset about?*

The dataset chosen to run several analyses gathers information about red and white wines. This data can be found in the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/. This project was inspired by the following

Although the size of the dataset is small, as it only has information of about 1,599 red and X white wines, the analysis can be implemented in large datasets executing the same code. Overall there are 10,000[1] different types of grape for wine in the world and at least 10 million wines[2] reviewed on the famous "Vivino" app.

Here is the definition[3] of the variables, what they are, their measurement and what are regular ranges:

- **Fixed acidity**. Does not evaporate rapidly. Tartaric acid (g/L)
- **Volatile acidity** (g/L). At too high of levels can lead to an unpleasant, vinegar taste
- **Citric acidity** (g/L). Found in small quantities, citric acid can add 'freshness' and flavor to wines
- **Residual sugar** (g/L). Amount of sugar remaining after fermentation stops. Values between 1g – 45g. Above 45g they are considered sweet wines.
- **Chlorides** (g/L). Measures the salt in wine.
- **Free sulfur dioxide** (g/L) or PPM 20-30 ppm. The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.

[1] Scott Washburn, How Many Different Types Of Wine Grapes Are There?, The Wine Guide (September 17, 2013) https://www.google.com/search?q=how+many+wines+are+there&oq=how+many+wines+&aqs=chrome.2.69i57j0l5.3968j0j1&sourceid=chrome&ie=UTF-8

[2] Brian Freedman, *The Launch Of Vivino Market Could Herald A New Era In Wine Buying*, Forbes (March 30, 2017), https://www.forbes.com/sites/brianfreedman/2017/03/30/the-launch-of-vivino-market-could-herald-a-new-era-in-wine-buying/#51a40f5c5ed1

[3] Definitions from: Daria Alekseeva, *Red and White Quality*, RPubs (N/A), https://rpubs.com/Daria/57835

- **Total sulfur dioxide** (gL). Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- **Density** (g/dL). It is close to that of water but it depends on the percent alcohol and sugar content.
- (Potential of Hydrogen) **pH**. Whether a wine is or not based on a scale from 0 (very acidic) to 14 (very basic); Normal ranges between 3 – 4 in the scale.
- **Sulphates**. A wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
- **Alcohol**. The percent alcohol content of the wine.
- **Quality**. Score 0 to 10 from experts.

Acidity is tightly related to pH. All wines lie on the acidic side of the pH spectrum and most range from 2.5 to about 4.5 pH (7 is neutral). Red wines have higher pH than white wines. Acidity defines the tart and sour taste of wines. Fixed acids are tartaric acid, malic acid, and citric acid.

Their respective levels found in wine can vary greatly but in general one would expect to see 1,000 to 4,000 mg/L tartaric acid, 0 to 8,000 mg/L malic acid, 0 to 500 mg/L citric acid, and 500 to 2,000 mg/L succinic acid.

Remember, pH is a logarithmic scale, so theoretically, a wine with a pH of 3 is 10 times more acidic than a wine with a pH of 4. The higher the wine acidity, the softer and smooth the taste.

Sweetness decreases the sensation of acidity (here is where residual sugar comes in). For instance, Coca-Cola has the same pH as Brut Sparkling wine, but our reaction when drinking the first is not the same as the later one, and that’s because of the high residual sugar of Coca-Cola and low in Brut Sparkling wine, which makes it dry.

Wines with high acidity and high residual sugar will age better, as acidity acts as a buffer to preserve wines longer.

Good wines must be in balance with four fundamental traits, which happen to be variables in this model: acidity, tannin, alcohol and sweetness. However, there is no apparent formula or mixture to make these four a perfect balance.

2. *What are we trying to learn?*

We are trying to learn whether (1) we can predict the quality of wine and (2) the type of wine (red or white) given the variables offered in the dataset: fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality.

In addition, we want to learn which is the best model to predict them using Binary Classification and Logistic Regression, as well as which could be the further steps to take if our model seems to have an inaccurate result or if it needs further research and implementation.

3. *What do we expect after the implementation?*

We expect to be able to be able to predict the quality of wine and the type of wine depending on the value for each variable. However, for the quality side of the model many research papers say it is hard to predict, as quality is relative to the person that drinks it and there is no perfect balance in between the main traits that make a wine good or not.

Following this, it might be had to predict which kind of wine each vector is. It is true that wines with a lower pH are generally white wines whereas wines with a higher pH are red wines. However, it also depends on the grape and there are 10,000 grapes which are unfortunately not defined in this dataset.

4. *How do you want to evaluate your project?*

We are trying to learn if we can apply Binary Classification to predict whether a wine is tasty or not. In order to do so, we are going to set a new column that gives 1 to wines with quality equal or bigger than 7, and 0 to the rest. We will analyze it first for red wines, and then for both datasets combined. We are going to apply the following models:

- Logistic classification
- Decision Tree Classifier
- Random Forest Classifier
- Gradient-Boosted tree Classifier

Then, we will read the results and proceed to do linear regression if needed, in order to continue the research to better understand the variables and extract conclusions. Linear regression will be performed separately on red wine, white wine and combined. In addition, we are going to analyze the variables with linear regression to do further research on the model and extract conclusions.

Finally, we will also apply Binary Classification to see if we can distinguish red wine from white wine given all the variables as feature vectors.

a. How to access the correctness of your model?

In order to access the correctness of our model, we are going to use the BinaryClassificationEvaluator() from Spark's ML Evaluation package, and we are going to look at the Area Under ROC as well as Test Error.

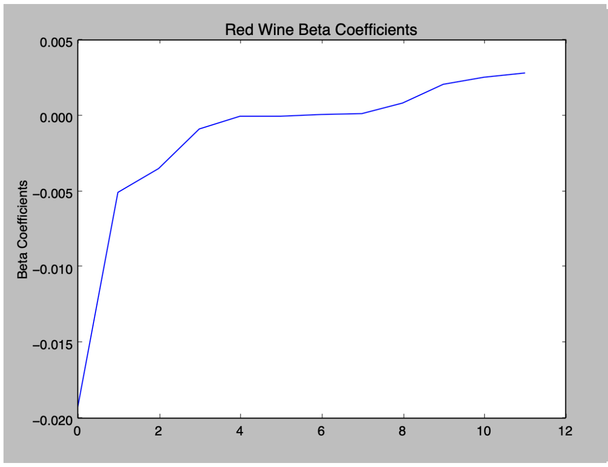
b. How well would you expect that the model will work?

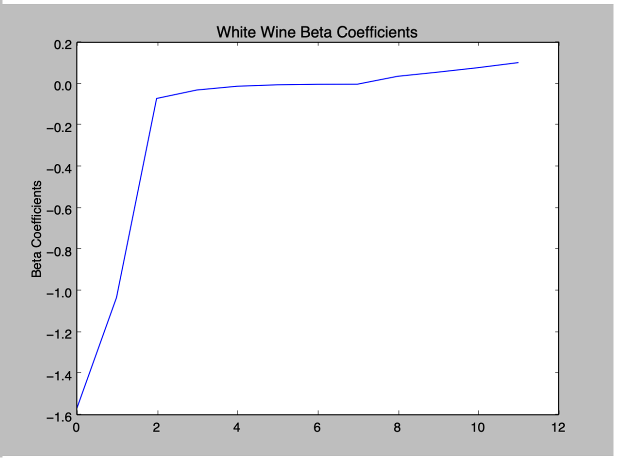
After reading several articles and papers regarding wine quality and what are the fundamental traits that make a good wine from a bad wine, I do not expect the model to get any accurate results. However, I hope to see one or two variables with bigger impact on quality over other ones, even if the impact is only between 20% - 40%. Nevertheless, we might be positively surprise.

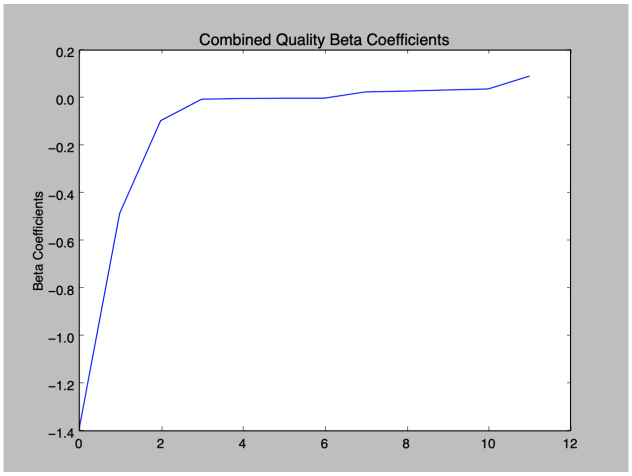
5. *Results*

- Let’s have a look at the results regarding wine quality:

Here are the beta coefficients for the Binary Classification using Logistic Regression for the combination of both wines and for white and red wine:







As we can see, all betas are pretty similar in terms of range above zero. Below zero, white wine coefficients tend to be more negative. The combined is simply a sum of both coefficients. What betas show is:

- βi > 0 implies eβi > 1 and the odds and probability increase with Xi
- βi < 0 implies eβi < 1 and the odds and probability decrease with Xi

The beta coefficients for white wine are:

```
[ −1.57355648e+00  −1.03178758e+00  −6.90557036e−02  −2.81130435e−02
  −9.97105081e−03  −2.98530420e−03  −4.47293700e−04  −1.13774462e−04
   3.80318263e−02   5.81352309e−02   7.96482699e−02   1.04049554e−01]
```

White wines have more negative values than red wines. Hence, with most of the variables, when any of the independent variables positively changes, we have a decrease in the probability of white wine quality, and vice versa. Very similar with the combined dataset. The contrary happens to red wines. For the red wine, this also shows the non-correlation between the independent variables and the target, alcohol.

However, the accuracy and area under ROC are almost 0 and 1 respectively for the three datasets (red, wine and combined), which means that the model can predict if a wine is good or not. On the other hand, I believe this happens because the dataset is too small and therefore predicting the results is easier than on a larger dataset. Hence, the same model should be tested on a larger dataset like the one the app “Vivino” has.

Another reason to believe the prediction of the quality of a wine it is not accurate is the results of the linear regression’s correlation variables. Let’s now have a look at the results of the correlation to quality for all variables in the combined dataset, with red and white wine:

```
('Correlation to quality for ', 'fixed acidity', -0.07674320790962014)
('Correlation to quality for ', 'volatile acidity', -0.26569947761146706)
('Correlation to quality for ', 'citric acid', 0.085531717183678)
('Correlation to quality for ', 'residual sugar', -0.0369804845857698)
```

('Correlation to quality for ', 'chlorides', -0.2006655004351014)

('Correlation to quality for ', 'free sulfur dioxide', 0.055463058616632414)

('Correlation to quality for ', 'total sulfur dioxide', -0.04138545385560937)

('Correlation to quality for ', 'density', -0.3058579060694188)

('Correlation to quality for ', 'pH', 0.019505703714435507)

('Correlation to quality for ', 'sulphates', 0.038485445876513875)

('Correlation to quality for ', 'alcohol', 0.44431852000765354)

('Correlation to quality for ', 'quality', 1.0)

All variables but alcohol, density or chlorides (the ones marked in bold) are very uncorrelated to the quality of the wine. Even if we take the three variables with the highest correlation, the metrics are not high enough to say that they have an impact on wine quality. Hence, their impact is very small.

- How about determining if a wine is red or white?

All the ROC results that we got from the Binary Classification of white or red are very low:

  - Logistic Regression – Area Under ROC: 0.47190294872930144
  - Decision Tree Classifier – Area Under ROC: 0.467254051182
  - Random Forest Classifier – Area Under ROC: 0.235479943328
  - Gradient-Boosted Classifier – Area Under ROC: 0.196745771717

Hence, the values that each variable offer for every wine collected, are not particular from one type of wine. The results might be more accurate if we were only testing for acidity, as we know that lower levels of pH are more in line with white wines and higher levels of pH are related to red wines, in general.

Finally, in for all the different linear regressions Gradient-Boosted tree regression was the one with the lowest RMSE in all cases (red, white or combined). Hence, the performance of Gradient-

Boosted tree was the best out of all of them, although the difference between one model or another one were very small so it hard to say if the results would change with a larger dataset.

For Binary Classification, when analyzing the quality, they all performed ‘excellent’ but logistic regression which instead of an area under ROC of 1 it had either 0.98 or 0.97. Given the small difference we can say that they all performed well. On the other hand, for the “red or white” prediction all the area under ROC were very low, which explains why it was not able to predict the output.

In conclusion, although many papers state that wine quality and type of wine are hard or almost impossible to predict since the human factor of relativity is very high, I believe that with further research or the right combination of variables results can be more accurate. However, with the current example, using all variables of the dataset, the results are not accurate enough so we cannot state that we have accomplished it.

BIBLIOGRAPHY

Doug Nierman, *Fixed Acidity*, Waterhouse (2004), http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity

Madeline Puckette, *Understanding Acidity in Wine*, WineFolly (December 9, 2015), https://winefolly.com/review/understanding-acidity-in-wine/

(N/A), Wine Acidity Guide: What Does Acidity In Wine Mean?, Winepair (N/A) https://vinepair.com/wine-101/acidity-wine-mean/

Chris Stamp, *Beginners Gudie to Striking Balance*, Wines&Vines (February 2009), https://www.winesandvines.com/features/article/62083/Beginners-Guide-to-Striking-Balance

Daria Alekseeva, *Red and White Quality*, RPubs (N/A), https://rpubs.com/Daria/57835

Bob Peak and Nancy Vineyard, *How To Use and Test Free SO2 in Wine*, The Beverage People (N/A), https://www.thebeveragepeople.com/how-to/wine/free-so2-in-wine.html