Red Wine Regression Prediction

Tomás Sousa

March 8, 2021

Contents

Chapter 1

Introduction

There are two datasets available that describe the composition and the quality of two variants of Portuguese "Vinho Verde" wine. This wine originated in the north of Portugal and for a time, now, property of the Minho Region, one of the most beautiful regions in Portugal the one that I gladly call home. The main goal of this problem is to find which features of these kinds of wine are the ones that provide the most information about its quality. I will also try to make a prediction of a wine's quality and check if it matches with the real quality. Although this dataset can be viewed as a classification (multiclass classification) or a regression problem, we will solve it using regression techniques.

This dataset is available on the UCI machine learning repository.

The main objective of this analyse is to predict the quality of the wine in order to inform the producers, and as a result reduce the potential loss do to bad quality.

Chapter 2

Data Analyse

My model will focus on prediction, as I will try to predict the quality of the wine based on it's features

2.1 Description

The dataset has 11 features that together describe the quality of the wine being this the last feature of the data.

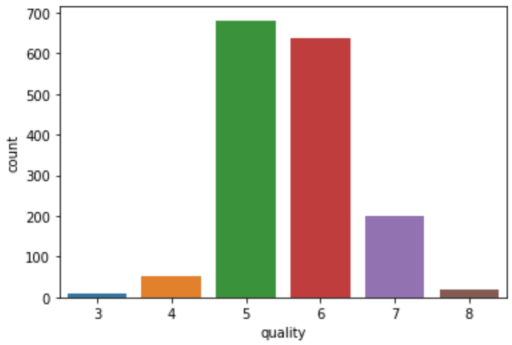
The features are:

- fixed-acidity;
- volatile-acidity;
- citric-acid;
- residual-sugar;
- chlorides;
- free-sulfur-dioxide;
- total-sulfur-dioxide;
- density;
- pH;
- sulphates;
- alcohol;
- quality;

2.2 Data exploration

After checking our dataset is fine and "ready to go" we are going to explore the data a little bit more, we are going to plot important information that will help us check how features behave and how they are correlated. We will also try to extract as much information as we can from it to help us understand the dataset better.

Knowing our target variable is "quality", we are now going to plot some information about it. Let's see which values this column contains and how many of them there are.



Now that we got information about our target variable we are going to study the correlation between quality and other features and see which are the ones that play an important role in deciding the quality of a wine.

From the features above, we are going to select the ones with bigger numbers since these are the ones that will give us more information. To do so we are going to stablish a minimum threshold of correlation approximately around 0.2 (absolut value) since we do not have to take into account features whose values might be redundant and not provide information at all.

```
quality                True
alcohol                True
sulphates              True
citric_acid            True
fixed_acidity         False
residual_sugar        False
free_sulfur_dioxide   False
pH                    False
chlorides             False
density               False
total_sulfur_dioxide  False
volatile_acidity       True
```

Figure 2.1: Caption

From all the values, we are selecting alcohol, sulphates, citric-acid and volatile-acidity in order to study them better and see the distribution of values

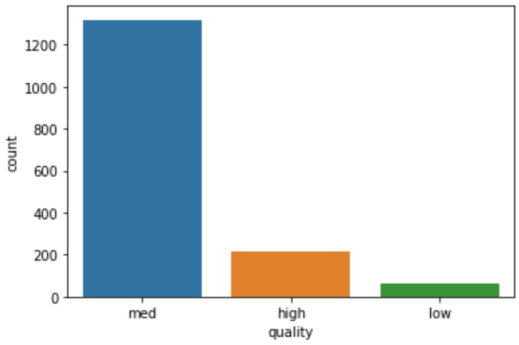
that separate the different qualities.

To have a better understanding of these let's see the correlation of these features with the quality overall:

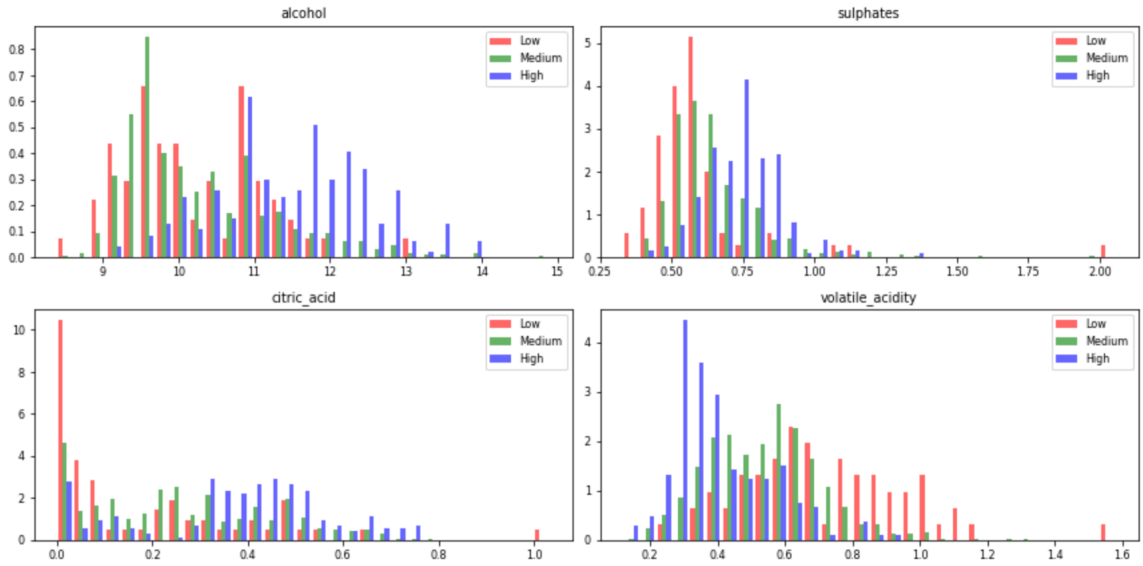
```
quality                1.000000
alcohol                0.375224
sulphates              0.162405
citric_acid            0.080146
fixed_acidity          0.053447
pH                     0.043065
residual_sugar        -0.018452
free_sulfur_dioxide   -0.060618
chlorides             -0.081813
density               -0.134559
volatile_acidity      -0.237193
total_sulfur_dioxide  -0.239067
```

For further investigations we are now going to plot histograms for each of those important features so we can see better the correlation between the distribution of values from each feature and quality. To do so, we are first going to separate the quality values in three different groups, so we can do things a little bit easier:

- Low: contains wines whose quality is 3 or 4.
- Medium: contains wines whose quality is 5 or 6.
- High: contains wines whose quality is 7 or 8.



Let's see these assumptions on a histogram:



As we can see in the histograms, higher values of alcohol, sulphates and citric acid seem to belong to higher quality wines while higher values of volatile acidity are present in lower quality wines.

2.3 Data Engeneering and Feature selection

Now that we have already studied our dataset through histograms and different graphics it's time to select some features we will use in our machine learning algorithms. In this specific case, what we are going to do is use the same columns we studied before, since those are the four ones that give us the most information between features and quality.

The unique values of the data set are:

```
fixed_acidity            96
volatile_acidity        143
citric_acid              80
residual_sugar           91
chlorides               153
free_sulfur_dioxide      60
total_sulfur_dioxide    144
density                 436
pH                       89
sulphates                96
alcohol                  65
quality                   6
```

After this we also introduced a new feature that differentiated if the wine was tasty or not bases on the quality overall above 6.

Chapter 3

Prediction using Regression

To have better predictive results after the regression we detected and handle some outliers using Z-score.

A Z-Score is a measure of position that indicates the number of standard deviations a data value lies from the mean. Any z-score less than -3 or greater than 3, is an outlier.

Note: From the empirical rule we see that 99.7% of our data should be within three standard deviations from the mean.

The first array is the list of row numbers and the 2nd array is the corresponding column number of the outlier. For example, the first outlier is in row 13, column 9. Once we calculated the Z-score, we can remove the outlier to clean our data, by performing the action

```
1     newdf = df[(z < 3).all(axis=1)]
```

3.1 Regression model

On this section, after having understood our data and dropped some useless features, we are going to make an estimation of quality based on the other features. To do so we are going to use Linear Regression, Decision Tree Rregressor and Random Forest Regressor. We will also plot the values of prediction and true quality and the confussion marices, so we can see how many of the predicted values are right (the diagonal of the matrix).

Linear Regression

- Linear regression is a statistical regression method which is used for predictive analysis. - It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables. - It is used for solving the regression problem in machine learning. - Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression. - If there is only one

input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression. - The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of the year of experience.

- The mathematical equation for Linear Regression is

$$Y = aX + b$$

Here, Y = dependent variables (target variables), X= Independent variables (predictor variables), a and b are the linear coefficients

Polynomial Regression

- Polynomial Regression is a type of regression which models the non-linear dataset using a linear model. - It is similar to multiple linear regression, but it fits a non-linear curve between the value of x and corresponding conditional values of y. - Suppose there is a dataset which consists of datapoints which are present in a non-linear fashion, so for such case, linear regression will not best fit to those datapoints. To cover such datapoints, we need Polynomial regression. - In Polynomial regression, the original features are transformed into polynomial features of given degree and then modeled using a linear model. Which means the datapoints are best fitted using a polynomial line.

- The equation for polynomial regression also derived from linear regression equation that means Linear regression equation Y= b0+ b1x, is transformed into Polynomial regression equation Y= b0+b1x+ b2x2+ b3x3+.....+ bnxn. - Here Y is the predicted/target output, b0, b1,... bn are the regression coefficients. x is our independent/input variable. - The model is still linear as the coefficients are still linear with quadratic

Ridge Regression

- Ridge regression is one of the most robust versions of linear regression in which a small amount of bias is introduced so that we can get better long term predictions. - The amount of bias added to the model is known as Ridge Regression penalty. We can compute this penalty term by multiplying with the lambda to the squared weight of each individual features. - The equation for ridge regression will be:

$$L(x, y) = Min(\sum_{i=1}^{n} (y_i - W_i x_i)^2 + \lambda \sum_{i=1}^{n} (w_i)^2)$$

- A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used. - Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization. - It helps to solve the problems if we have more parameters than samples.

Lasso Regression

- Lasso regression is another regularization technique to reduce the complexity of the model. - It is similar to the Ridge Regression except that penalty term contains only the absolute weights instead of a square of weights. - Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0. - It is also called as L1 regularization. The equation for Lasso regression will be:

$$L(x, y) = Min(\sum_{i=1}^{n} (y_i - W_i x_i)^2 + \lambda \sum_{i=1}^{n} |w_i|)$$

3.2 Score and Root Mean Square Error

After having prepared our models, it's now time to evaluate them. To do so we are going to use RMSE (Root Mean Square Error) which is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are so RMSE is a measure of how spread out these residuals are.

When deciding which regression algorithm is better by looking at RMSE we would better choose the one with smaller value, so for this problem, Random Forest Regression seems to be the best fitting algorithm.

Regressor Model	Precision	Root Mean Square Error
Linear Regression	0.7440008162319891	0.25265954803817475
Polinomial Regression	0.83818684005294	0.20087382257057018
Ridge Regression	0.74358633228472	0.25286400371374473
Lasso Regression	0.7412351715588991	0.2540206645011605

Chapter 4

Summary and Future Ideas

After having obtained all the results through our models and plots, these are some things we can say about this problem and solution:

- Out of the 11 features we examined, the top 3 significant features that help the producer to brew a delicious Red Wine are low levels of sulfur dioxide, and high levels of sulfates and alcohol, while volatile-acidity tends to impact Red Wine negatively.
- The vast majority of wines get a quality rating of five or six while having good and bad wines seem more unlikely. There seem not to be any excellent wines (¿8) on this database.
- Polynomial Regression algorithm yields the highest R2 value, 83%. Any R2 above 70% is considered good, but be careful because if your accuracy is extremely high, it may be too good to be true (an example of Overfitting). Thus, 83% is the ideal accuracy!
- In the future, it would be interesting to see the difference in prediction accuracy using other types of Regression models like SVM, Decision Tree, and Random Forest Regressions.