Data Science Professional Certificate
HARVARDX and EDX

Predicting Wine Quality using Data
- FULL REPORT

by Sophie Bassnagel
June 19th 2020

I. (Personal) Introduction and background

This project is an assessment within the Data Science Capstone class, final module of the Data Science Professional Certificate from HarvardX and EDX.

This assessment challenges us, attendees, to put in practice everything we have learnt after nine different modules: from R Basics to Machine Learning.

I started this certificate at a low point in my life. Unsatisfied with my life, with my job, I wanted and needed to make a big change. Passionate about data, I decided to start this to possibly change career, from business analyst to data analyst.

When I took back control of my life, I got extremely lucky. I was approached by a big tech company for a job. It was all a dream, and I forced myself to be pragmatic. I was going to do my best at this interview process, without expecting to get in. My luck stayed with me and couple of weeks after, I resigned at my job and got ready to change city, change job, change life.

That's why I pressed the pause button on this certificate for a year. It was very hard for me to keep this going. I was struggling to make it fit with a full-time job and a very active lifestyle. It is finally during the Covid-19 pandemic that, forced to be stuck at home, I managed to find the time to bring my focus back on these studies. I am a strong believer of finishing what I've started, even if the motivation and reasons evolve.

When I first started this certificate, my main objective was to change career from business analyst to data analyst. Moving forward two years later, while my passion for data is intact, I find myself more attracted now by program management. I am looking to evolve in a position where I could be the link between technical and more business-oriented people.

You will find below the results of these two years of back and forth with this certificate: my first ever machine learning algorithm!

II. Project Executive Summary

The topic I choose for this project if the following: Wine. As a French woman, I am fully in love with wine, and I found the perfect data set on Kaggle which allowed me to do an in depth analysis of how is wine quality affected by other factors.

We have 12 variables, 11 of them are explanatory. We will use them to build models predicting whether a wine is good or low quality.

To do so, we are using the following statistical tools: PCA, logistic regression, Distribution Tree, Random Forest, etc. The data set is large enough to be able to have a training sample and a test sample.

The main objective here is the following: if we add new wines into the data set, are we correctly able to assume if it is going to be a good one or a bad wine depending on the explanatory variables? I want to minimize the error rate.

The most effective one is Random Forest as we have the highest value of the Area Under the Curve (AUC). Indeed, we reached an error rate of 15.36% in the case we add new data.

III. Detailed Methodology

The table we are using has 1599 observations, and 12 variables. Here we can have a quick look at our data set.

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulf
1	7.4	0.7	0	1.9	0.076	
2	7.8	0.88	0	2.6	0.098	
3	7.8	0.76	0.04	2.3	0.092	
4	11.2	0.28	0.56	1.9	0.075	
5	7.4	0.7	0	1.9	0.076	
6	7.4	0.66	0	1.8	0.075	

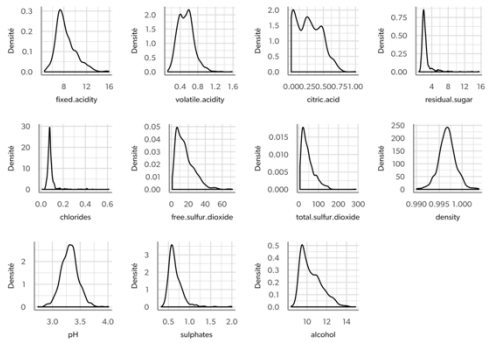
For each wine, we have its characteristics, from acidity to sugar. The first 11 variables are numeric, and the last one is binary. It assesses if the quality of a wine is good (1) or bad (0). As such, the 11 first variables are explanatory to the last one in the models we will build.

We first want to understand whether or not we have any missing values.

Variable	Valeurs Manquantes
fixed.acidity	0
volatile.acidity	0
citric.acid	0
residual.sugar	0
chlorides	0
free.sulfur.dioxide	0
total.sulfur.dioxide	0
density	0
pH	0
sulphates	0
alcohol	0
qualiteY	0

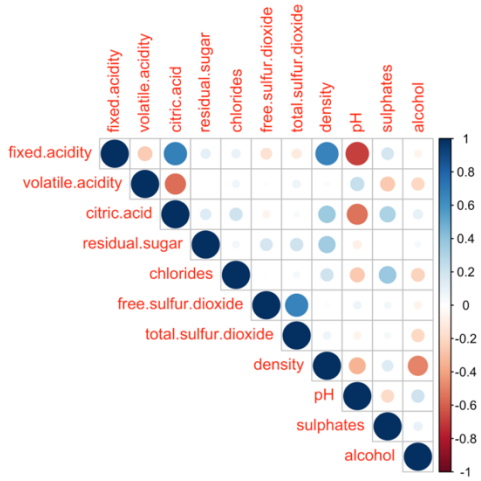
As we can see here, there is not missing values.

Let’s deep dive into the data set and conduct some exploratory analysis. Let’s start with the variable distribution:



As we can see here, most of the variable follow a normal distribution, except for two: residual sugar and chlorides which are skewed to the left.

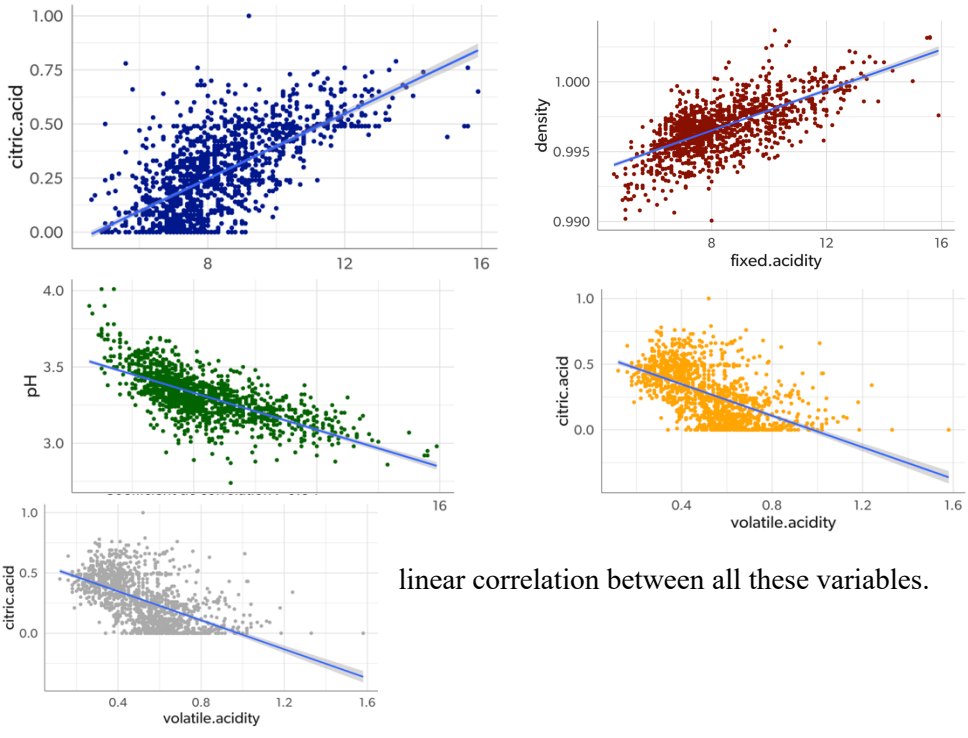
Let's now understand if they are correlated between themselves. To do so, we will build a correlation matrix (Pearson).



There is a strong negative correlation between pH and fixed acidity, between citric acide and volatile acidity, and between pH and citric acid.

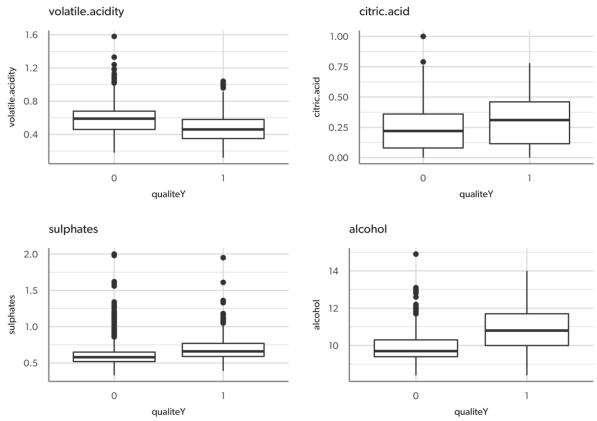
There is a strong positive correlation between fixed acidity, citric acid and density.

Let's have an even closer look at the relations between these variables :



As we can see here, there is a linear correlation between all these variables.

We now want to understand if these variable have an influence on wine quality. I have seen that the distribution are very different between quality =1 qnd quality = 0 for these four variables, which is why we are focusing on these now.



We can see here that it looks like they do.

To statiscally prove this intuition, we realised a test of the means for each explanatory variable, between `qualiteY = 0` and `qualiteY = 1`.
The hypothesis H0 of the means test is the situation were the means are equal.

Let's start for instance with `volatile.acidity`, the test will make us able to check if the mean of `volatile.acidity` for the wine quality 0 is equal to the mean of `volatile.acidity` for the wine quality 1.

We fix the threshold at 5%. The alternative hypothesis will be rejected if the p-value is above 5%.

By applying the means test on these 4 variables, here is what we conclude:

The good quality wines have:

- a level of volatile acidity below bad wines because the P-value is 1.710^{-39} < 5%.

- a level of citric acid higher than bad wines because the P-value 10^{-10} < 5% < 5%.

- a level of sulphates higher than bad wines because the P-value is 10^{-18} < 5%.

- a level of alcohol higher than bad wines because the P-value is 10^{-77} < 5%.

In order to go further in the analysis, we are going to perform a principal component analysis (PCA) which will allow us to see if we can flag individuals groups and variables by lowering the dimensions.

The analysis in principal components allow to analyse and visualise a dataset with individuals described by multiple quantitative variables.
It is therefore possible to study the similarities between individuals with a view on all variables and to understand individual profiles by lowering the dimensions.

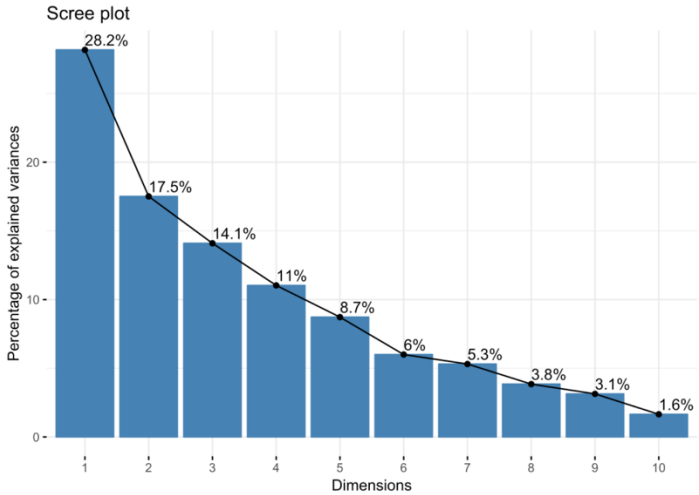
I do an PCA on this data set to understand if there is a combination of these 11 explanatory variables than can explain the wine quality.

Analysis of proper values:

The variance represent the information within a dataset. The idea is to reduce the number of dimensions while not losing too much information.
We choose to keep 70% of the information from the data set and to reduce the number of dimensions from 11 to 4.

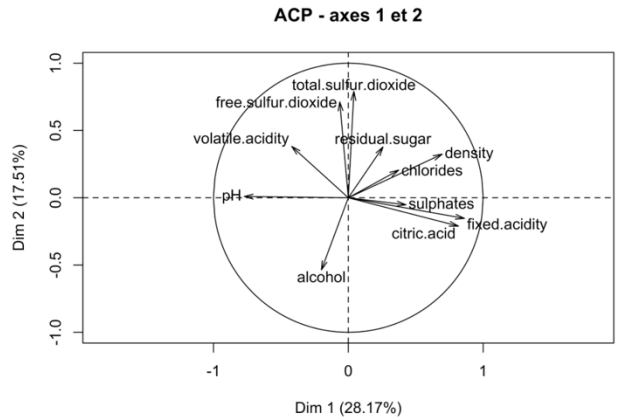
```
##         percentage of variance cumulative percentage of variance
## comp 1              28.1739313                           28.17393
## comp 2              17.5082699                           45.68220
## comp 3              14.0958499                           59.77805
## comp 4              11.0293866                           70.80744
## comp 5               8.7208370                           79.52827
## comp 6               5.9964388                           85.52471
## comp 7               5.3071929                           90.83191
## comp 8               3.8450609                           94.67697
## comp 9               3.1331102                           97.81008
## comp 10              1.6484833                           99.45856
## comp 11              0.5414392                          100.00000
```

Scree plot:



There is no strong uncoupling on the scree plot, except between the first and the second dimension. We will stay with the analysis of the first 4 axis.

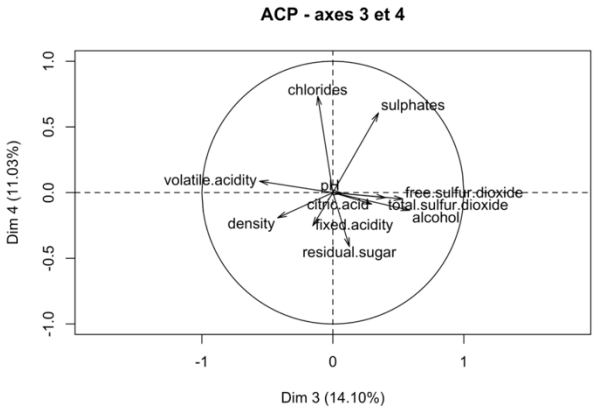
With a PCA, each axis is a linear combination of the variables.



The variance explained with the 2 first axis is 45%.

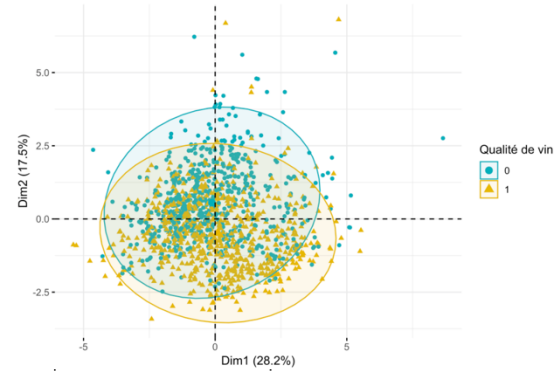
- AXE 1 : Axis 1 represents wine acidity. It set against two variables very correlated (citric acid and fixed acidity), with the pH. Un vin acide aura un pH faible pour une mesure de fixed.acidity élevée. We already saw in the correlation matrix that these variables were negatively correlated.

- AXE 2 : Axis 2 represents the sulfur in the wine (Free sulfure and total.sulfure, positively correlated). These variables are negatively correlated with alcohol.

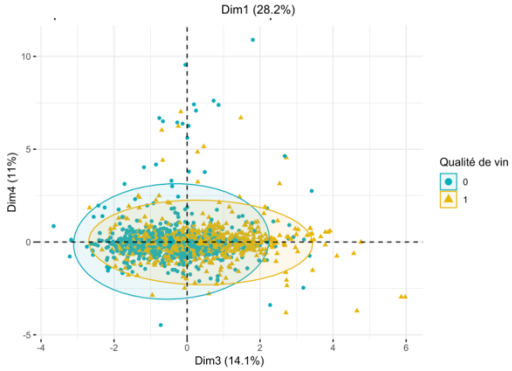


No variable is well represented on the axis 3, we can still see a negative correlation between `alcohol` and `volatile.acidity`.
Axis 4 is represented by chlorides and sulfats which are positively correlated.

Individual analysis:



Good quality wine have a tendancy to have a lower sulfat rate vs bad quality wines.
However, acidity doesn't seem to impact wine quality.



Good quality wines have an alcohol percentage higher and a lower volatile acidity vs lower quality wines.
Some individuals are standing out : 152,1436,1477.

These points being very represented, let's have a closer look at them to understand why they stand out. We can see that the data residual sugar (except for 152) and free sulfur are high for these 3 individuals - a lot higher than the median.

Logistic regression:
We are trying to understand which variable are best able to explain a wine quality. The function step allows to select a model with a step by step procedure based on minimalising the AIC criteria. It allows me to keep only the relevant variables for my model and to delete the variable that do not contribute to it or add noise only.

The model keeps these variables:
`fixed.acidity - volatile.acidity - citric.acid - chlorides - free.sulfur.dioxide`
`total.sulfur.dioxide - sulphates - alcohol`

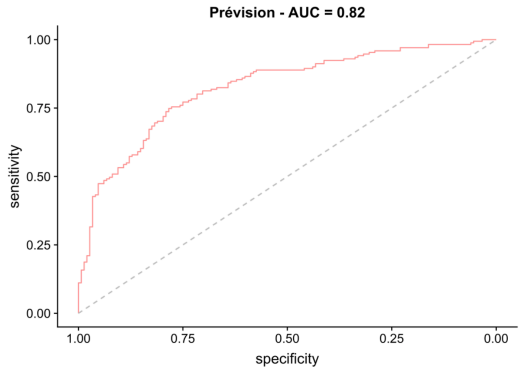
```
##
## Call:
## glm(formula = qualiteY ~ fixed.acidity + volatile.acidity + citric.acid +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     sulphates + alcohol, family = "binomial", data = vin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3340  -0.8488   0.3242   0.8294   2.3493
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -9.216919   0.949966  -9.702  < 2e-16 ***
## fixed.acidity         0.127271   0.051081   2.492  0.01272 *
## volatile.acidity     -3.379881   0.477983  -7.071 1.54e-12 ***
## citric.acid          -1.260357   0.560972  -2.247  0.02466 *
## chlorides            -3.529121   1.509122  -2.339  0.01936 *
## free.sulfur.dioxide   0.022082   0.008184   2.698  0.00697 **
## total.sulfur.dioxide -0.015645   0.002811  -5.565 2.62e-08 ***
## sulphates             2.686254   0.432624   6.209 5.32e-10 ***
## alcohol               0.905412   0.073423  12.331  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2209.0  on 1598  degrees of freedom
## Residual deviance: 1657.8  on 1590  degrees of freedom
## AIC: 1675.8
##
## Number of Fisher Scoring iterations: 4
```

Alcohol is the most significant variable to forecast a wine quality. It is the variable that brings the most information. The more alcohol increases, the more the probability of having a good quality wine is increasing (positive estimate). On the contrary, the more volatile acidity increases, the more the probability of having a good quality wine is low (negative estimate).

Predictive model:

If we need to predict a wine quality on new data, we will build a predictive model on a training model which will be tested on a test dataset to understand our error rate. The sample contains enough data, we can divide it in 2 samples for test and training.

After this, we check the ROC to understand errors: the ROC curve (Receiver Operator Characteristic Curve) represents the ratio of true positive on the y axix vs the ratio of false positives on the y axis.



The AUC (Area under the curve) gives the classification rate without error compared to a logistic model, on the test sample. It allows to compare the ROC curves between multiple models.

Matrice de confusion avec un seuil à 0.5

	0	1
0	112	41
1	36	130

If I get new data, I can expect an error rate of **24.14%** with this predictive model, and by picking a threshold of 0.5.

Optimal threshold: We are going to at the threshold that will allow us to minimize this error rate.

```
##   threshold specificity sensitivity
##   0.5139985   0.7837838   0.7485380
```

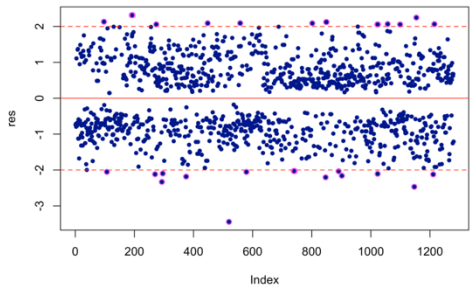
Let's fix the threshold to 0.51 to create the confusion matrix.

qualiteY	Fréquence
0	46.56
1	53.44

At the threshold of 0.51, the error rate is **23.51%**.
If I get new data, I can expect an error rate of **23.51%** with this predictive model. When we receive new data, we'll need to re-calibrate it.

Residual analysis: a good residual is a residual without any expored structure. Residuals need to be independent from observations.

We will need to re start the model without line 653.



In theory, 95% of the residual from Student are within the interval [-2,2]. It is the case here as 30 residuals are outside of the interval (so 2.34%).

Comparison with other models: I will compare with other models to understand if they give better results than the logistic regression model.

Interaction model: the AUC with this model is **0.82**.
Model without correlated variables: the AUC with this model is **0.82**.
Decision tree: the AUC with this model is **0.77**.
Random Forest: the AUC with this model is **0.92**.

It is the model that allows for the best results (highest AUC).

	0	1
0	122	23
1	26	148

With a random forest, I can expect an error rate of 15.36% on new data.

IV. Conclusion

Random Forest is here the best performing model, allowing us to have the best prediction on wine quality. While the other models are also performing, they are not reliable enough.

In any case, this model needs to be recalibrated each time new wines are added to the data set to assure the best accuracy possible.