Report Midterm Statistical Learning

Cesar Conejo Villalobos

12/18/2020

Introduction

For this project, we consider two data sets related to two variants of the Portuguese Vinho Verde wine. The first variety is called *red* and has 1599 observations. The second variety is *white* and has 4898. Both data sets are available in the repository data.world.

The biggest differences between both products are the following:

1. Red wines are produced with black grapes and white wines with grapes. Moreover, red wines are fermented with the grape seeds and skins. On the other hand, white wines are not fermented with these parts.
2. In the oxidation process, red wine varieties require an increase in oxygen. For archiving that, winemakers use oak barrels. On the other hand, while white wine requires to reduce exposure to oxygen. This can be done using stainless steel vats.

Second, the data set shows the chemical properties (continuous variables) for each observation. Also, an ordinal variable *quality* is used as a ranking by tasters. In the case of the data set, each variety is tasted by three independent tasters and the final quality rank is the median rank given by the taster on a scale from 0 (worst) to 10 (best).

Although some crucial information is hidden in this data set such as wine brand and wine selling price, the price of wine depends on the quality and appreciation by wine tasters. As a result, the objective of this project is to predict the quality of the wine as a function of the other variables. Moreover, the focus of the predictions will be to detect high-quality wine in order to increase the selling price of the wine products.

The raw `R` code is attached to this report. However, if possible, more details about the code used in this project are available in my repository on Github. Also, to guaranty *reproducibility* of the results, the seed used in all codes is 42.

Dataset

A quick review of the data set is given with the function `summary()`.

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923
##  Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949
```

```
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality       Category
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000   Red  :1599
##  1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   White:4898
##  Median :3.210   Median :0.5100   Median :10.30   Median :6.000
##  Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

The most important variable for this data set is `quality`. Before the graphical description of the variables, we create a `train` and `test` data set using 80% and 20% of the observations respectively.

The distribution of the dependent variable for the `train` set is given by figure 1. We can observe how the quality values are concentrated in values 5, 6, and 7. Also, we notice that there are no wines ranked in the values 1 and 2, and a small proportion of wines are qualified as quality 3 and 4. On the right side, again, there is a small proportion of ranked wine as 8 and 9, and there is no wine qualified as 10. Moreover, the proportion of quality wine by category is similar. Then we can no conclude that the category of wine (red or white) define previously their quality.

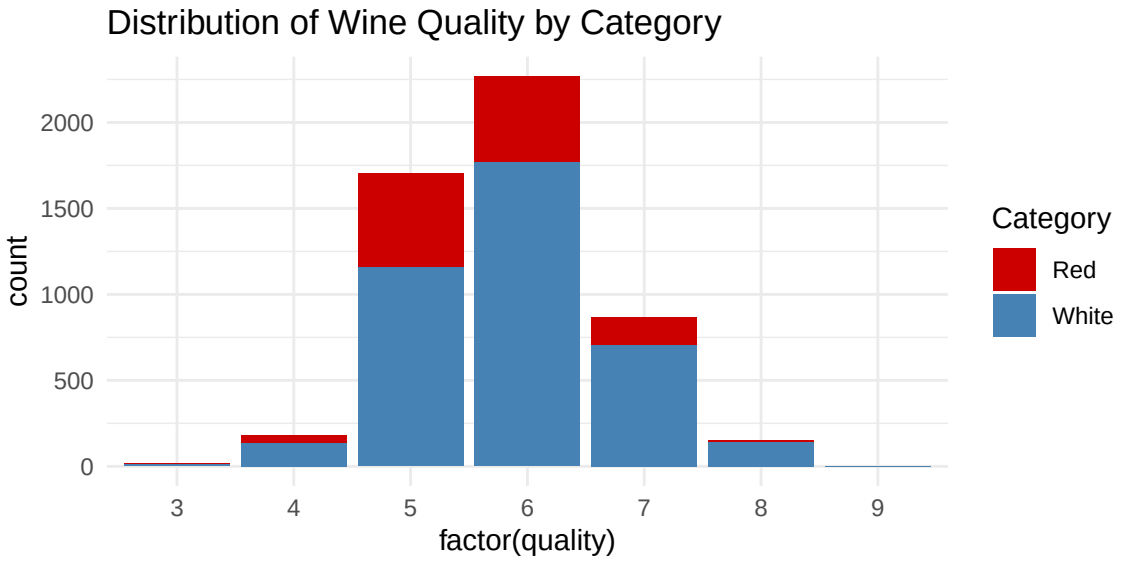

Figure 1: Distribution of the quality variable scale 1 - 10 (Train data)

Figure 2 provide the relation of wine quality vs alcohol. The graphic shows that there is no clear linear relationship between these two variables. However, we notice that white wines tend to have higher levels of alcohol.

Regression Problem

First, in figure 3 we review the relation of the independent variables with the dependent variable `quality`. If we proceed with a simple linear regression of the most relevant predictor `alcohol` using the function `lm(quality ~ alcohol, data = wineQuality.Train)` with obtain a model with a $R^2 = 0.21$ in training test and $R^2 = 0.17$ for testing set. To improve this model, we can add more variables to the model. Figure 4 reflects the relation between the variables (With the idea of identifying multicollinearity). Except for the

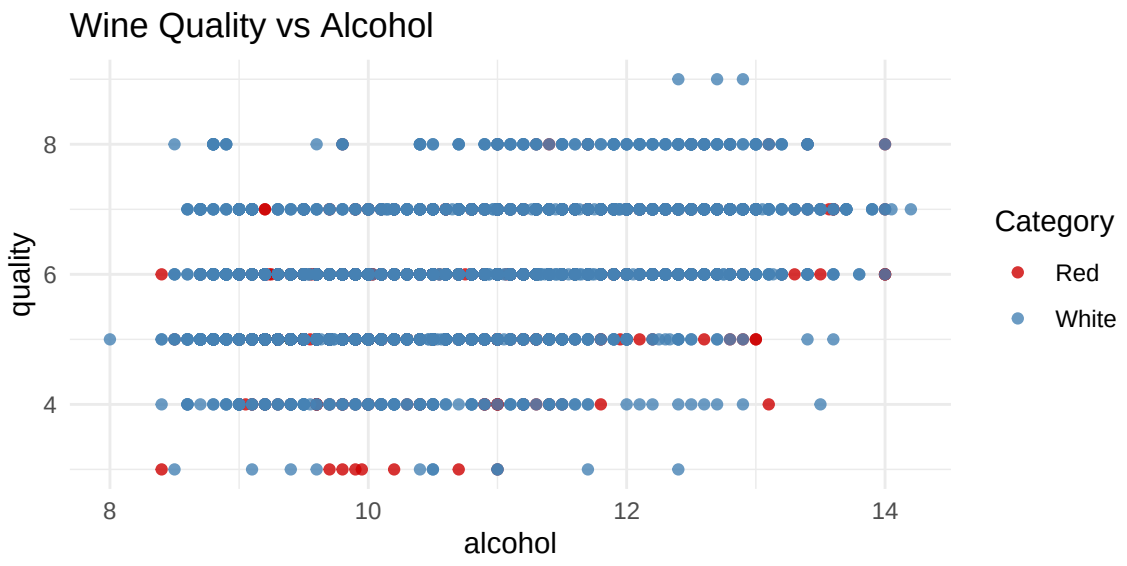

Figure 2: Distribution of wine quality vs alcohol level (Train Data)

relation between *density* and *alcohol*, most of the variables are not related. This observation will be important in the future models when we will see how the *traditional* and *penalized* frameworks have similar performance.

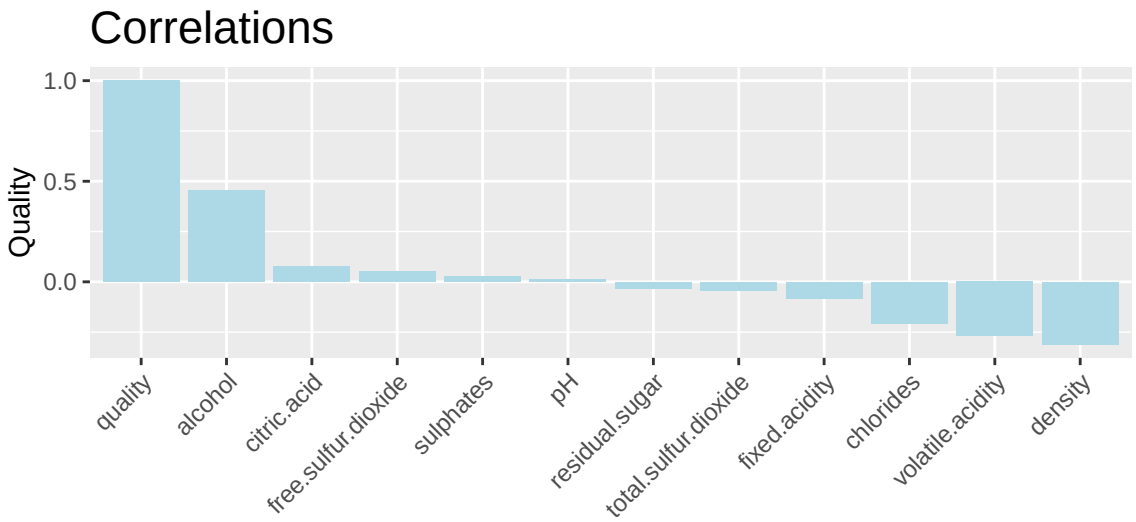

Figure 3: Relation between the independent variables and quality (Train Data)

With this analysis, we proceed with the formula `lm(quality ~ ., data = wineQuality.Train)`. In this case, we have a $R^2 = 0.31$ for the training set and $R^2 = 0.26$ for the testing. We have an improvement of almost 10% if we consider all the variables. However, this prediction can be considered as noisy if the purpose is to predict the exact quality value. Table 1 shows the proportions of observations for each category is unbalanced, especially in the cases of too low (less or equal than 4) or too high (greater or equal than 8).

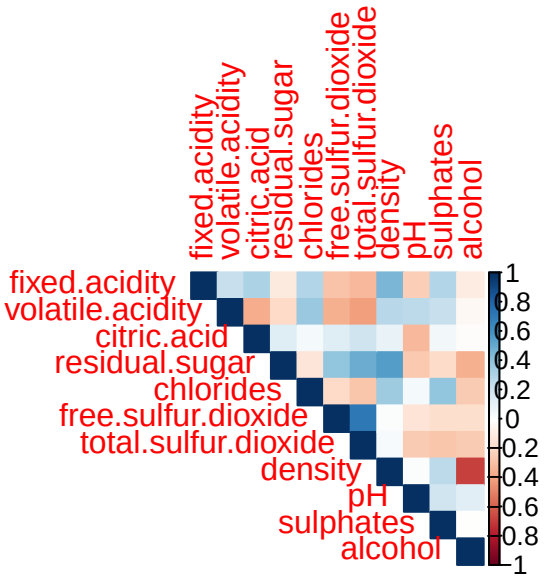

Figure 4: Relation between the independent variables (Train Data)

Table 1: Proportion (in %) of quality category

3	4	5	6	7	8	9
0.4618	3.3246	32.9075	43.6509	16.6077	2.9706	0.0770

As a result, instead of predicting the *real quality* values, wine quality can be analyzed into three categories with the following characteristics:

- `Low`: Low quality wine, values of quality less than 5. (2384 observations, proportion: 36.7%)
- `Medium`: Medium quality wine. Values with quality equal to 6. (2836 observation, proportion: 43.6%)
- `High`: High quality class. Values of quality greater or equal than 7. (1277 observation, proportion: 19.7%)

Classification Problem

We create a new variable called `quality.class` based on the previous scheme. In this case, we will focus our attention on detecting the `high` quality class that represents approximately 20% of the observations. Although the proportions are a little further from being perfectly balanced, the proportion of 20% gives the opportunity of expecting better results under this new scenario.

Figure 5 show the relation under this new approach to the problem. Again, `alcohol` is the variable that best identifies the relationship between the independent and dependent variables.

Then, we can divide our analysis based on the tools:

1. **Logistic Regression:**

a. Simple logistic regression. Independent variable: `alcohol`
b. Multiple logistic regression. Using all the independent variables.
c. Penalized logistic regression. Method: `glmnet`.

2. **Bayes classifiers:**

a. Linear discriminant analysis. Methods: `LDA`, `sparseLDA`, `stepLDA`.
b. Quadratic discriminant analysis. Methods: `QDA`, `stepQDA`.
c. Naive Bayes. Method `nb`.

The simple and multiple logistic regression follows the traditional scheme of 80% for training and 20% for testing. (Specifically, we have 476 low observations, 567 as medium quality, and 255 as high quality for the testing set.) The rest of the methods are calibrated using the `train()` function of the `caret` package under the scheme of cross-validation with 5 repeats of 10-fold cross-validation.

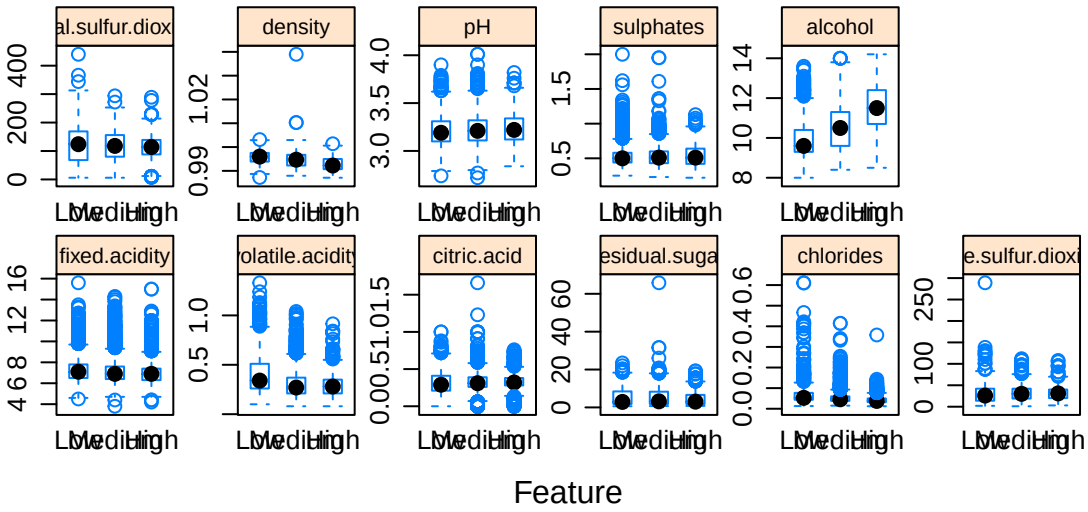

Figure 5: Relation indepedent variables with dependent variable (Train data set)

The key metrics for this classification problem will be:

1. **Global Accuracy:** General accuracy of the predictive capability of the model. (The greater, the better, with a perfect score of 100%.)
2. **High-quality Sensitivity:** Percentage of high-quality wine correctly predicted as high-quality wine. (The greater, the better, with a perfect score of 100%.)
3. **Number of low-quality misclassification:** Nominal value of low-quality wine predicted to be high-class. It is the worst error in this model. (The lower, the better, with a perfect value of 0.)

Logistic regression

In the case of the logistic regression approach, we have 3 groups. Therefore, it is necessary to apply 2 regression models. The performance of the three logistic models is given in table 3. In general, accuracy is slightly similar in the range between 53 and 55 percentage. In this sense, the simple logistic model will provide a parsimonious and fairly accurate model if the purpose is to provide explanations and statistical inference. For example, taking as reference level the low-quality wine, for every unit in alcohol, the *log odds* of medium-quality (vs low quality) increases by 0.70 units. Similarly, for each unit increase in alcohol, the *log odds* of high quality (vs low quality) increases 1.35.

Table 2: Coefficient simple logistic Regression

	x
(Intercept):1	-6.9695125
(Intercept):2	-14.9509382
alcohol:1	0.7009949
alcohol:2	1.3535353

However, the models that use all the independent values have a better high-quality wine detection (with a cost of 5 low-quality wine predicted has high-quality). It is important to observe the effect of the few correlations between the independent variables. This gives, as a result, the *multiple log. regression* and *penalized log. regression* has a similar performance.

Table 3: Metrics for Logistic Regression methods

Method.lr	Accuraracy.lr	High.Sensitivity.lr	Worst.Error.lr
Simple.Log.Reg	53.39	27.06	8
Multiple.Log.Reg	55.16	31.37	13
Penalized.Log.Reg	55.16	30.98	13

For the logistic models, the best approach will be the *penalized* version using `caret` becasue this approach will be more robust. Also, `caret` provides the importance of the variable in the calibration process (Figure 6). We can see how the variables *density* and *residual.sugar* are also prominent in the detection of high and low-quality wine. (together with *alcohol*)

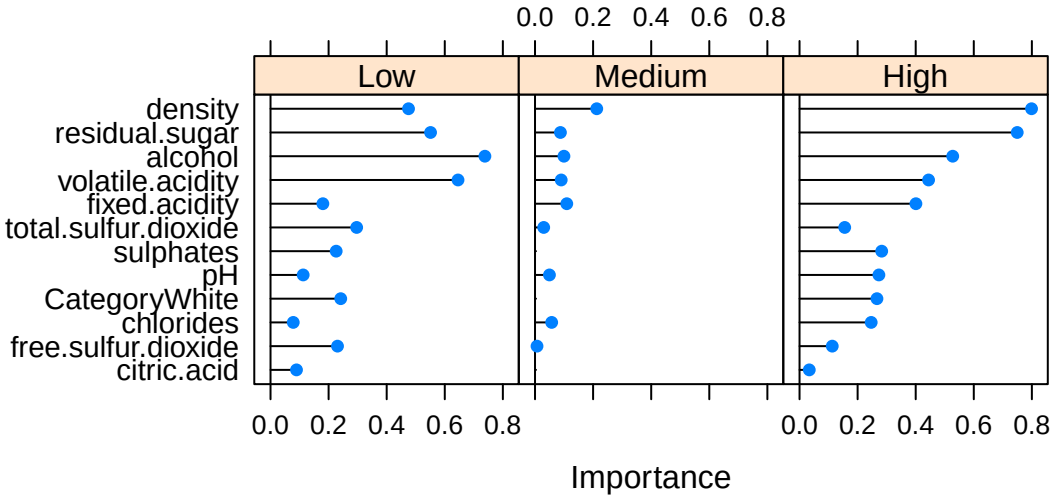

Figure 6: Variable importance for Penalized Logistic Regression

Bayes Classifiers

a) Linear Discriminat Analysis

Table 4 provides the key metrics for these linear models. We have a range of accuracy from 52% to 55%. As a result, we have a similar performance as the logistic models. However, high-quality sensitivity detection is superior for any linear discriminant than logistic. On average, we increase the right prediction of high-quality wine in 13 observations with a cost of an increase of 4 observations low-quality wine fitted as high quality. In summary, for each bad prediction of high-quality wine, we will have 3.25 observations correctly predicted as high-quality. Again, all three models perform similarly due to a lack of correlation between the independent variables.

Table 4: Metrics for LDA methods

Method.lda	Accuraracy.lda	High.Sensitivity.lda	Worst.Error.lda
lda	54.78	36.08	17
sparseLDA	54.70	36.08	17
stepLDA	52.93	36.47	16

b) Quadratic Discriminat Analysis

Table 5 reflects the results in the case of the quadratic methods. As expected, the performance of these models in large data sets is lower than their linear variant (Although not so much in terms of general accuracy). Also, it is important to show the high sensitivity of the `qda` model. However, the increase of the low-ranking wine predicted as high is considerably higher than the other models.

Table 5: Metrics for QDA methods

Method.qda	Accuraracy.qda	High.Sensitivity.qda	Worst.Error.qda
qda	51.93	70.59	43
stepQDA	52.47	27.06	8

c) Naive Bayes

Finally, Table 6 shows the result of the naive Bayes methodology. This method has the lowest general accuracy. Additionally, naive Bayes has the same effect over the high sensibility percentage with a costly increase of low-ranking wine classified as high-quality.

Table 6: Metrics for nb (naiveBayes) method

Method.nb	Accuraracy.nb	High.Sensitivity.nb	Worst.Error.nb
nb	50.92	49.8	28

Best model unders these metrics?

A graphical summary of the performance of the models in terms of accuracy and kappa is given in figure 7 using the function `resamples` from `caret`.

In this case, the figure reflects how in general terms, the penalized logistic regression and the sparseLDA have better efficiency. Also, both approaches are more conservative in terms of high-quality sensibility and low-quality misclassification.

In the case that we need to choose a model, the election on this stage will be the `sparseLDA` framework (Notice how higher values of accuracy and kappa are considered as outliers in the logistic regression).

The confusion matrix for `sparseLDA` model is given in table 7.

Table 7: CM sparseLDA Method

	Low	Medium	High
Low	281	137	13
Medium	178	337	150
High	17	93	92

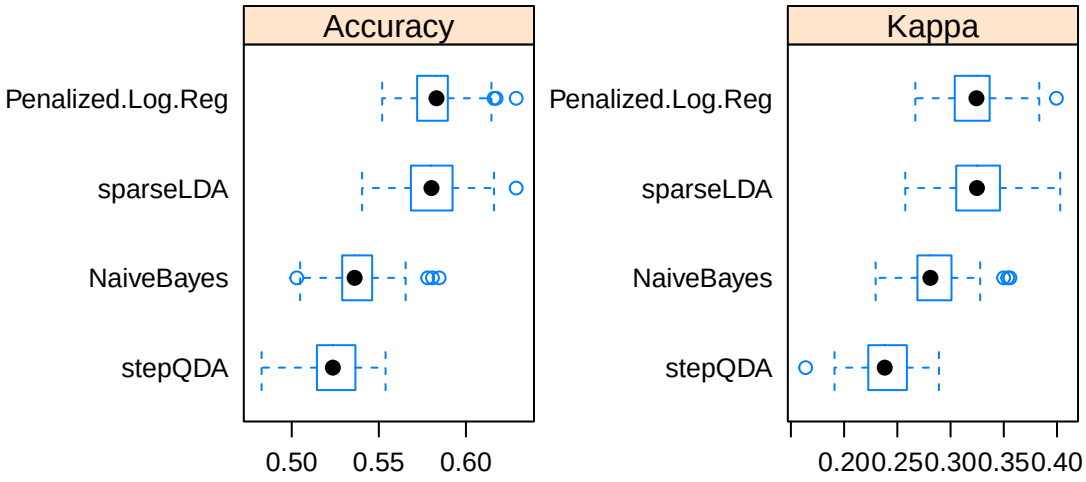

Figure 7: Relation indepedent variables with dependent variable (Train data set)

Cost-Sensitive Learning:

Considering a rigorous scenario where predicting high-quality wine is crucial for the interests of the company, the previous framework has several difficulties in predicting precisely the quality of the wines. As a result, we can simplify the amount of variance increasing the bias in the modeling process. Continuing with this line, we can proceed in two directions:

a. **Reduce noise:** In this case, we will identify *High* wine quality from *No High* quality. As a result, we will combine the categories `Low` and `Medium` as a class `No High`. This new class has 5220 observations, with a proportion of 80.3% of the data set. The high-quality wine continues with the same number of observations (1277) from the previous models.

Although it might be thought that the problem of the imbalance of high-quality wines continues, we have two aspects that could be considered as benefits with this new paradigm:

1. We reduce the variance of the problem transforming the problem into a binary classification task.
2. We can make use of the probabilities included in these modeling tools. The default configuration is to apply the Bayes rule, so we can modify this threshold.

b. **Cost-sensitive learning:** In previous sections, we mention the relation between the price and the quality of the wine (High-quality wine is more expensive in the markets). Also, we can add the crucial importance of detecting high-quality wine for certificate the production of wine.

In relation to the first approach, if we transform the problem into a binary problem, the algorithm `sparseLDA` will have a general accuracy of 80.99%. However, we must be cautious, because the sensitivity for this model is 30.98%. (This result is similar to all previous approaches). Moreover, table 8 shows the confusion matrix under this strategy.

Table 8: CM sparseLDA Binary Problem

	No High	High
No High	973	176
High	71	79

It is important to remember that the optimization criteria used for calculating the model that results in the confusion matrix of table 8, the threshold used was 0.5 in the probabilities. So the problem becomes to determine an adequate threshold for resolving this framework. In the following table, the left side is the confusion matrix when the threshold is 0.25. On the other hand, the right side is the confusion matrix when the threshold is 0.75.

	No High	High
No High	806	85
High	238	170

	No High	High
No High	1040	239
High	4	16

When we decrease the threshold from 0.5 to 0.25, we can see how high-quality detection increases from 79 to 170 (An increase of 115%). However, the worst error increases considerably from 71 to 238 (235%). The accuracy under this approach is 75.13%.

In the case of the increasing of the threshold from 0.5 to 0.75, the numbers of low and medium wine fitted as high-quality reduce from 71 to 4 (Decrease of 94%), but the correctly high-quality wine predicted to decrease from 79 to 16 (80%). The accuracy in this case is 81.29%.

To solve this problem, we can use the *cost-sensitive* approach. Consider the following economic scenario:

In wine markets, the price of high-quality wine is double in comparison with regular wine. The wine company wants to forecast the economic profit in case that all the wine products are predicted correctly and as result, it can be sold at the right price, taking as reference the price of the regular wine. (Consider the case that company makes their economic profit forecast under the assumption that all their wine has regular-quality.)

As a result, if a regular quality wine is predicted as regular (`No High`), the profit will be zero. If a high-quality wine is predicted correctly, then the profit will increase by 1 unit. In the case of misclassification, consider the following scenarios:

1. If a regular quality wine is predicted as high quality, the company will lose 0.5 units of the original price for high-quality wine.
2. If a high-quality wine is predicted as regular-quality, the company eventually will win 0.5 monetary units. However, there is an opportunity cost of 0.4 monetary units. As a consequence, the company will obtain only 0.1 monetary units.

A summary of these assumptions is given in table 9. The economic profit under the scenario using a threshold of 0.5 is 61.1. With a threshold of 0.25, the economic profit will be 59.5 . In the case of a threshold of 0.75, the economic gain will be 37.9.

Table 9: Cost-sensity assumptions: Profit matrix

Prediction_And_Reference	No_High	High
No High	0.0	0.1
High	-0.5	1.0

Figure 8 shows the distribution of economic profits for different threshold values. We see that the maximum on the distribution is reached with a threshold value of 0.35.

With this value, the confusion matrix is given in table 10. The economic profit will be 67.4. General accuracy of the model is 78.75%, with a sensitivity of 51.37%.

Table 10: CM sparseLDA with 0.35 threshold

	No High	High
No High	892	124
High	152	131

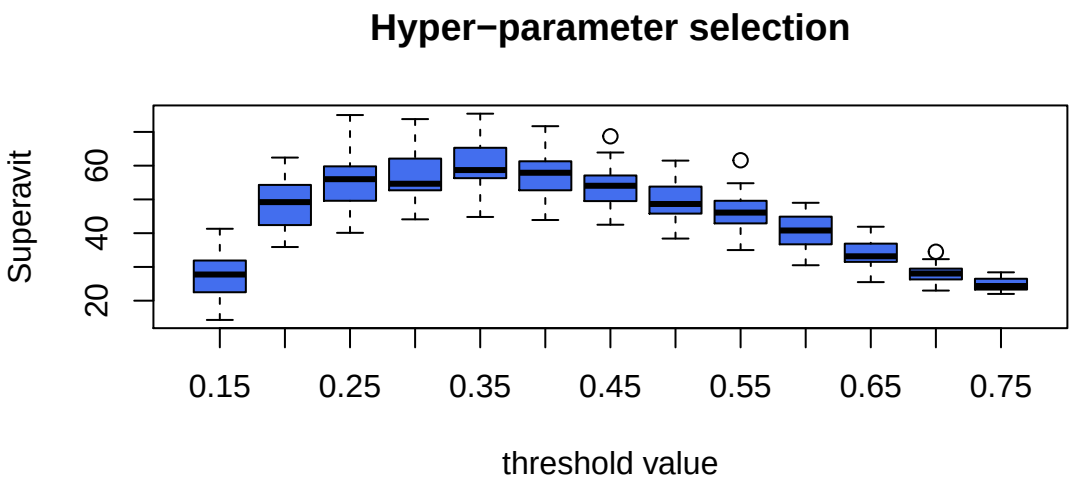

Figure 8: Distribution threshold value (Based on 30 iterations in cv.)

Conclusions

The goal of this project was to predict high-quality wine based on the chemical properties of the two categories of wine. However, the model failed to predict with a high precision the range of the wine. These failures can be explained possible for three factors:

1. At each quality level, the variability of the independent variables is high. Based on figure 5, we can observe several outliers, especially in the variables *Fixed acidity*, *volatile acidity*, and *sulphates*. Also, we can see how *alcohol* has an irregular distribution. For the `low` quality category, there is a lot of observations considered as outliers. As a result, the net effect of the distribution of `low`, `medium`, and `high` quality wine is the same.

2. The quality wine groups are not well separated for the predictor variables. Again, the boxplots in figure 5 show how the distribution of the chemical properties are relatively the same for each wine quality category.

3. Probably, the wine quality is also related to other variables, beyond the chemical properties of the wine. For example, the variety of grapes, time, and materials used for fermenting the product can also affect the quality of the wine and its flavour. This data set does not have variables related to these factors. Then, there is a hidden variance provided by the data generating process that the models used in this project will not be able to take into account.

Finally, increasing the predictive capacity of the models has the effect of increasing the bias of the results, obtaining as a result *precise* models with almost 80 of accuracy. However, these results can be taken with caution. In the last part of the project, the criteria used to determine the best model was the economic profit of the company. Notwithstanding, these approaches cause the increase of the regular quality wine fitted as high quality. Maybe this can be useful from the perspective of the company, but on the other side of the market, customers can receive the wine of regular quality promoted as high quality. As a result, in the long run, it can affect the image and reputation of the company.