Vino Virtuoso

Business Applications of Machine Learning: Final Report

# 

# Table of Contents:

# 

[**1.1 Background and Motivation**](#_ofpyxy2eonr3) **3**

[**1.2 Business & Statistical Question**](#_ytzgralmmnvb) **4**

[**1.3 Supervised Learning Workflow**](#_3py5yvy97xl) **5**

[Feature pre-processing](#_gmblu9lwlvpj) 5

[Data Cleaning](#_t8anv7ki7ftc) 5

[Categorizing response variable](#_8iv9c3js4dhw) 5

[Encoding categorical features](#_u5bbt25ggm8h) 5

[Training & Testing Sets](#_o261fdw5s8mx) 6

[Feature selection](#_bvoz983do2yy) 6

[Model 1](#_6da0b94xfrjb) 6

[Model 2](#_lw6sqg5qiauv) 6

[Model 3](#_m5ihp311doii) 7

[Model 4](#_5mei1njrmhkp) 7

[Hyperparameter Tuning](#_m5ekgayjxxq0) 7

[Model selection](#_11l822mg0iqp) 8

[Quantitative Consideration](#_4jh9vv8x4303) 8

[Qualitative Consideration](#_xcqjng43fmir) 8

[Human Choice Consideration](#_okiud9ydrt0h) 8

[Final Model](#_az98k03oo9y9) 9

[**1.4 Results & Recommendations**](#_ddx50m82auof) **10**

[**Appendix**](#_50q1ktyptmpj) **11**

# 

# 1.1 Background and Motivation:

Our client is a wine boutique looking to improve its profits by carrying wines that are more appealing. The business question we are aiming to answer is whether we can determine which wines will be popular amongst customers based on certain attributes of the wine. These attributes include description, country, province, region, winery, and variety.

The dataset we have used to address our business question is a dataset from WineEnthusiast Magazine’s online platform. WineEnthusiast Magazine was founded in 1988, with staff having an extensive knowledge of wines spanning across the globe. Their team of professional reviewers have reviewed 130,000 wines, which has been web scraped into a csv file on June 15th, 2017 for this project.

The dataset consists of the following columns:

* *country:* the country the wine is from
* *description:* written description of wine from reviewer
* *designation:* land within the winery where the grapes of the wine are formed
* *points:* rating of wine out of 100 points
* *price:* the price of the wine
* *province:* the province the wine is from
* *region\_1:* region where the wine is from
* *region\_2:* more specific region within the region where the wine is from
* *taster\_name:* name of the reviewer
* *taster\_twitter\_handle:* twitter user of the reviewer
* *title:* name of the wine
* *variety:* variety/type of the grapes the wine is made of
* *winery:* winery the wine is from

Please refer to **Appendix Figure 1.1.1**.

We note that all the columns consist of categorical variables except for columns “points”, which consists of integers, and “price”, which consists of floats. Please refer to **Appendix Figure 1.1.2**.

We also note there are many “NaN” values within the dataset. In fact, “region\_2”, “designation”, “taster\_twitter\_handle”, and “taster\_name” have the most missing values (in order). Please refer to **Appendix Figure 1.1.3**. If the columns are not relevant or provide little information, this is something we should consider removing from our dataset.

Lastly, we see that the average wine rating given is 88.45, with the minimum wine rating as 80, and the maximum as 100. Please refer to **Appendix Figure 1.1.4**. This means the ratings only have a range of 20.

# 1.2 Business & Statistical Question:

*Business question:*

“How do we accurately predict whether our customers will like a particular wine?”

*The corresponding statistical question:*

***X*** = Classifying a wine into four grade levels

***Y*** = Relevant features, chosen and refined from our wine attributes via the workflow

By successfully predicting the grade levels of a wine, our client can easily decide which are “good” wines that are worth buying and would thus be enjoyed by their customers. Our model with help in optimizing limited shelf space, time spent selecting wines, and costly build-up of unsold inventory. This ultimately helps maximize profit earning chances in an effective way.

However, we recognize that there are some limitations to our model. Whether a wine is liked by customers is a subjective issue, that varies depending on the person. Therefore, wines with high grades (assigned by the WineEnthusiast reviewers) might not not always result in the wines being popular with the majority of our wine boutique consumers. Other factors such as marketing campaigns could also play a role in the sales result of a wine. Our current model is unable to incorporate these external factors, and only incorporates characteristics of the wine. However, we believe that our model will serve as a useful reference point for the business decisions of our client. With more data about consumer preferences for our specific wine boutique, we may be able to improve our model to predict the popularity of a wine rather than simply predicting the grades, which would be a more applicable model.

# 

# 

# 1.3 Supervised Learning Workflow:

## Feature pre-processing

### Data Cleaning

We first drop columns “region\_2”, “designation”, and “taster\_twitter\_handle”, as they have a high proportion of “NaN” values, and are of less relevance and importance when predicting wine ratings (See **Appendix Figure 1.1.3**). Using dropna() on a dataset with columns that have a lot of “NaN” values could adversely affect the accuracy of our model because it would drastically reduce the amount and variety of data that are available for model training. This is a limitation to our dataset. In the future, additional data collection could provide a more complete picture and improve our prediction results.

Afterwards, we drop all rows which contain at least one “NaN” value with the dropna() function to obtain the final clean dataset. This reduces our dataset to 77,267 rows and 10 columns. Please refer to **Appendix Figure 1.3.1**.

### Categorizing response variable

Additionally, to determine whether a wine is of a superior quality, we converted the points rating of the wine from a quantitative variable (‘points’) to a categorical variable (‘grades’). The categories are as follows:

* Grade 1: 96-100
* Grade 2: 91-95
* Grade 3: 86-90
* Grade 4: 80-85

We notice that more than 50% of our wines fall within the Grade 3 category, and few fall into the Grade 1 category. Please refer to **Appendix Figure 1.3.2**.

### Encoding categorical features

We then encode categorical features such as country, province, and winery with One-Hot-Encoding to ensure numerical input for the algorithms used (used in Model 3 and 4). We also created a Sparse Matrix of dummy variables with One-Hot-Encoding in order to minimize the size of the data.

### Training & Testing Sets

Lastly, we split our data into a training (80%) and testing (20%) set, as this is standard practice. We will be running the training set to train the model, and the testing set to give us the best estimate of how our model will perform with future data. We will also be cross-validating our data with k-fold cross validation.

## Feature selection

Moving forwards with our cleaned dataset, we first obtain the error rate of the null model, which is 1 - (the mode of the rating categories). This turns out to be 1 - 0.583380 = 0.41662, or in other words the error rate is ~ 42% with the null model.

Note we decided not to include “price” of the wine in our model features, as the rating of a wine is not necessarily associated with the price. For example, a cheap wine can be rated highly because it is of worth its price tag, while an expensive wine can be rated poorly because the wine was overpriced. We believe that there are other features that will be better indicators of a highly rated wine.

### Model 1

For this model, we attempt to use the description of a wine to predict its grades. The expert reviews for a wine usually reflect the quality of a wine very accurately, and our data is from a renowned wine magazine where the reviews are given by professional editors who are experienced in wine tasting. We are interested to see whether we could train our model using these reviews, and possibly develop a list of keywords to use for predicting the quality of wines that are not in this dataset.

We chose to use a Naive Bayes model for description, as it is the most common algorithm used for text classification. To feed our description into the model, we use “bag-of-words” representation, which via the function CountVectorizer().

### Model 2

For our second model, we attempt to boost the accuracy of our prediction by incorporating more features other than description. We add province and region\_1 as we deem them as good indicators of the geographical location in which a wine is produced, which is an important factor when deciding the quality of the wine.

We continue to use a Naive Bayes model. To add the columns province and region\_1 to our model, we simply created a new column called “bundle”, that is a combination of description added to province added to region\_1. We then count vectorized the column “bundle” to feed into the Naive Bayes Model.

### Model 3

The third model attempts to predict the quality of a wine using only objective features such as country, province, region\_1, winery, and variety. Even though description could be a strong indicator of wine quality, it is relatively subjective and would be affected by the reviewing style of a tastor. Moreover, it could be costly to obtain a review for every wine the client is considering buying. Hence, we wish to explore the possibility of predicting the quality of a wine based on features that are easier to obtain.

We took our columns that had been one-hot-encoded, and tried them with various models: Multinomial Naive Bayes, Decision Tree Classifier, and Multinomial Logistic Regression via a pipeline.

### Model 4

For our last model, we would like to see whether the winery a wine is from (“winery” column in our dataset) is a strong indicator of a wine quality. Wineries are ranked as some of the most important features in Model 3, and we wish to compare the error rate of the simpler model of using only winery to that of the more complicated Model 3.

Again, we passed our “winery” column data through a pipeline containing models: Multinomial Naive Bayes, Decision Tree Classifier, and Multinomial Logistic Regression.

## Hyperparameter Tuning

We use a 10-fold GridSearchCV and pass a pipeline of various classifiers and hyperparameters to look for the best fit for each model we proposed.

For Model 1 and Model 2, we are only tuning the alpha of a MultinomialNB classifier. Based on the results, the best alpha for both models is alpha = 1.

For Model 3 and Model 4, we are passing three classifiers: MultinomialNB, DecisionTreeClassifier, and Multinomial Logistic Regression. We are also tuning the alpha, max\_depth, and C of the respective classifier. Based on the results, the best model fit for Model 3 is Multinomial Logistic Regression with hyperparameter C = 1, while the best model fit for Model 4 is Multinomial NB with alpha = 0.1.

## Model selection

### Quantitative Consideration

Our null model error rate is ~42%. The error rates for the 4 models we trained are as follows:

|  |  |  |
| --- | --- | --- |
| **Model Number** | **Training Error Rate** | **Testing Error Rate** |
| *Model 1* | 27% | 31% |
| *Model 2* | 27% | 31% |
| *Model 3* | 25% | 32% |
| *Model 4* | 27% | 35% |

From our testing results, we see that Model 1 and Model 2 give the lowest testing error. From a quantitative perspective, these two models are therefore the best option. Model 4, albeit being the simplest to implement, gives the highest testing error, making it quantitatively less desirable.

### Qualitative Consideration

From a qualitative perspective, the feature “description” provides a fuller picture of the wine in question, as compared to the other features (country, province, region, variety, and winery). We would therefore assume that “description” is the best qualitative feature for predicting wine ratings. We assume that “winery”, or “variety” would be the next best predictors. This is because “winery” is the most specific feature location-wise, and “variety” describes the main ingredients of the wine.

### Human Choice Consideration

From a human perspective, decision trees and Multinomial Logistic Regression are the easier of the models to interpret, as compared to Naive Bayes,or Random Forest Classifier. This is because Naive Bayes relies on more complicated mathematics and formulas, and Multinomial Logistic Regression relies on mathematical formula though it gives the weight for the each variable, which can help us understand the importance of each variable in deriving the decision of classification and hence making it more insightful and actionable. These concepts can be harder to fully understand, whereas a decision tree simply splits the data into two categories each time - a much easier concept to grasp. Moreover, decision trees only tells the importance order of parameters but does not give the magnitude of importance which can be easily derived from a multinomial logistic model.

Additionally, typically having a simpler model with less features is easier to understand. In this case, Model 4 with just “winery” as a feature, if modeled with the Decision Tree Classifier, should be the best option from a human choice perspective. However, when choosing a model, we must consider all three perspectives: quantitative, qualitative, and human choice - as well as feasibility and practicality.

### Final Model

After considering all three perspectives, we decided to choose Model 3, a Multinomial Logistic Regression, with features country, province, region\_1, winery, and variety. This is because of three main reasons. First, the model provides one of the lowest error rates, showing that it does well in a quantitative aspect. **Appendix Figure 1.3.3** also shows the resulting confusion matrix of the model for detailed number of correct and wrong predictions in each grade. Secondly, the features are also more objective compared to a description/review of the wine, so the results should be more consistent.Thirdly, the model is easy to interpret.Lastly, the feature information fed into the model is more easily obtainable. More specifically, Models 1 and 2 use the wine description provided by the reviewers at WineEnthusiast Magazine as a feature. However, the feasibility of obtaining a legitimate expert review of each wine can be both time-consuming and costly - especially for resources available to a local wine-boutique owner. Models 3 and 4 use features of a wine that are more easily attainable, and are often found on the wine bottle itself.

# 

# 1.4 Results & Recommendations:

Our recommendation is for our client, the wine boutique owner, to incorporate our model into his wine selection process. Our model can accurately predict 68% of wines and their grades with simple input of information such as wine variety and winery - information that is often easily found on the bottle of the wine itself. Doing so will improve the wine selection process and in turn, by accurately predicting wines that will be well-received, more than likely the owner should see an increase in store popularity, favourability, and profit.

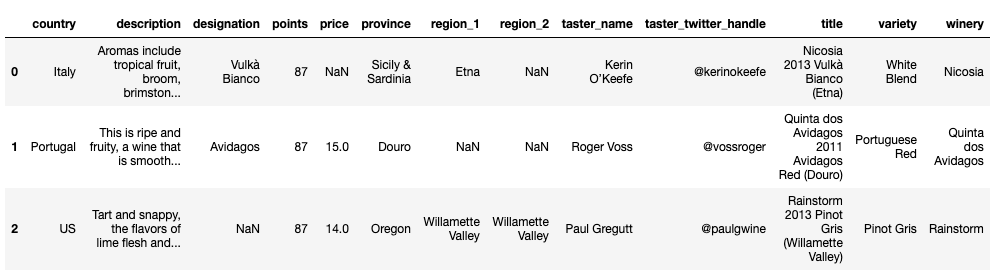
Our model uses a Multinomial Logistic Regression to predict wine grades. As we have four levels of wine grades to be decided, the algorithm would create four different models which would classify one type of grade with respect to others. For example, the first model will be deciding Grade 1 versus the rest of the grades, the second model will be Grade 2 versus the rest of the grades, and so on.The final classification will be made on the bases of combined output from all the four models, where each model will produce a probability value for each grade and the grade that gets the highest probability will be the final grade for that wine. While each model is making predictions, it assigns some weights to each explanatory variable, which are known as beta coefficients of that variable. The variable with the highest beta coefficient is known to be the most important factor in classifying the grade of the wine. In the future, we could collect data on variables having high beta coefficients which would help in making correct predictions while keeping the complexity and data input to the model low.

One limitation of the model is that it is not extremely user-friendly. With more time and resources, a potential suggestion would be to incorporate the model into a wine ratings phone application. The application would allow wine owners to scan a wine bottle label with information about the wine, which would be fed into the model automatically as features, and provide an accompanying grade result.

Additionally, it is also important to acknowledge that the taste of wine is subjective, which is another limitation to our model. We are assuming in our model that in general, the majority of people have similar tastes to professional sommeliers. To further personalize our model, we could look into collecting information on the wine boutique’s customer preferences and look to incorporate that into the model.

Lastly, we are working on the assumption that expert wine reviews can be hard to obtain. However, if these reviews are readily available and of low cost (For example available for free on the Internet), then we may also consider employing Model 1, which uses a Multinomial Naive Bayes algorithm with the feature “description”, as the accuracy rate of this model is 69%, and it requires less features.

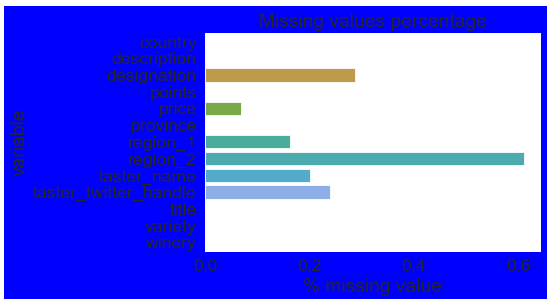
# Appendix



**Figure 1.1.1**



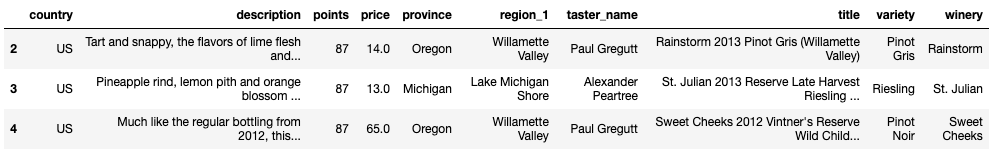
**Figure 1.1.2**

****

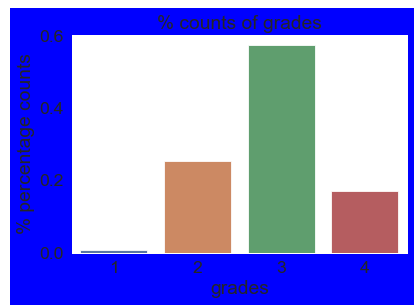
**Figure 1.1.3**

****

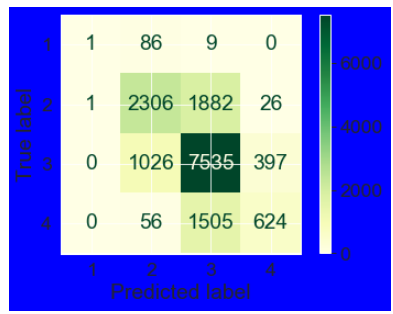
**Figure 1.1.4**

****

**Figure 1.3.1**

****

**Figure 1.3.2**

****

**Figure 1.3.3**