



WHITE WINE QUALITY PREDICTION

Data Mining for Business (BYGB-7967-V01) | Instructor: Professor Lin Hao

GROUP 3

- Wanshan Mao
- Muhetaer Mayila
- Yankun Song
- Yu Xiao
- Yue Wang
- Duygu Torun

Fordham
University

December 2020

EXECUTIVE SUMMARY

This project report consists of eight components: Introduction, Variables, DIDA Framework, Python Implementation of Prediction Models, Prediction Model Comparison and Result, The “Perfect Wine”-Scenario, Conclusions, and Limitations. The dataset contains 4899 observations and 14 chemical variables (Data).

To remain competitive in the wine industry, wine companies must ensure reliable wine quality. Therefore, it is beneficial to gain insights about the wine quality before launching it to the high-end white wine market (Insight).

In our definition, a white wine whose score is above six is considered “good” and should therefore be launched to the high-end white wine market (Decision).

Our predictions help wine company executives make efficient decisions leading to cost-cutting, profit-maximizing, and brand awareness strengthening (Advantage).

To reach the goal, this report deep dives into four prediction model approaches:

- I. Logistic Regression
- II. Classification Tree & English Rules ,
- III. k-NN
- IV. Neural Network

In this context, Accuracy and AUC values are defined and compared. The results show that each model reaches a reliable performance (around 80% each) with insignificant differences. Therefore, the model offers reliable results and the essential information to wine producers, which can lessen their workload and costs, gain more reputations for their brand, and make more company profits and grow business in the future.

Nevertheless, there are typical data-mining limitations like the trade-off between Accuracy vs. Interpretability. Since the accuracy of predictions gets better the interpretability, however, becomes more complex. Therefore, even though our k-NN and Neural Network have the best AUC results compared to Logistic Regression and Classification Tree, there is still a tradeoff that our interpretability is low.



TABLE OF CONTENTS

Executive summary	1
Introduction	3
Variables	3
DIDA Framework	5
Python Implementation Of Prediction Models	6
Logistic Regression	6
Classification Tree & English Rules	7
Neural Network	8
k-NN	9
Prediction Model Comparison and Result	10
The “Perfect Wine” - Scenario	11
Conclusion and Limitations	12



INTRODUCTION

Is it wine o'clock yet? No matter if red or white - There are wine enthusiasts all around the world. However, people often wonder how they can ensure that the bottle they bought is a good quality wine. Can they rely solely on the price or the year of the wine? Or looking at it from the other perspective, how can I, as a wine producer, ensure that my wine quality will meet the consumers' expectations? To find more certainty about these questions, our group has decided to choose this topic.

We want to examine the expectation if a wine's quality will be good enough to launch it to the high-end wine market in the role of the R&D department of a wine-producing company. In this context, the wine quality can vary from a scale of 1 to 10 and measured based on physicochemical variables such as alcohol, sulfates, volatile acidity, density, etc. We will focus on white wine data. Our target is the high-end white wine market. The segmentation for good wine starts with a score above 6. The data can be retrieved from Kaggle.

VARIABLES

The dataset contains 4899 observations of a variety of wine selections from different countries worldwide. It is divided among 13 other variables that can be used as measurements to predict wine quality. We will focus on the alcohol level, the percent alcohol content of the wine. The sulphates, a wine additive that can contribute to sulfur dioxide gas (SO₂) levels, help as an antimicrobial and antioxidant. The volatile acidity (VA) is a combination of compounds — primarily ethyl acetate and acetic acid present in small amounts in pretty much every wine. When a wine has a lot of VA, it can have acetone-y, vinegary, kombucha-esque notes. The density measurements can be used to measure the sugar content. Finally, the citric acid, which is often added to wines to increase acidity, complements a freshness's flavor to wines.

List of variables: fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates, alcohol, quality, good_quality

The following list shows each predictor and its respective description:







Predictor	Description
Fixed_acidity	Main contributor of the flavor of the wine
Volatile_acidity	<p>A combination of compounds — primarily ethyl acetate and acetic acid present in small amounts in pretty much every wine.</p> <p>Grapes contain acid itself but by adding more acid the wine could taste fresher. Too low the taste will be plain, too high it will taste rough</p>
Citric_acid	The most common acid, could be contained in grapes itself or added by man, adds to freshness
Residual_sugar	<p>The left sugar.</p> <p>Wine classification, for example sweet wine and dry wine is decided by residual_sugar. Also, it could neutralize the sourness of the wine</p>
Chlorides	Chlorides, along with salts of mineral acids and organic acids are major contributors to saltiness of the wine
Free_sulfur_dioxide(SO2)	The main use of sulfur_dioxide is to prevent the wine oxidizing into vinegar, so it nearly effects no taste
Total_sulfur_dioxide	The main use of sulfur_dioxide is to prevent the wine oxidizing into vinegar, so it nearly effects no taste
Density	The thickness of wine which is also essential to the taste
pH	The main scale of sourness of the wine
Sulphates	Sterilize, use as an antiseptic, exist very little in wine, so it also nearly effects no taste
Alcohol	The main scale of liquor strength, also essential to the taste



DIDA FRAMEWORK

The DIDA framework is a framework to identify and translate real-world problems into data mining problems¹.

The following table summarizes our DIDA framework in the context of this project:

Data 	<p>The dataset we will use contains the chemical attributes of white wines, each row represents a specific wine, so it is of individual-level data and historical. We have five predictors, and they are all ex ante, which means they are already a fact before the wine is graded. This dataset contains 4899 observations, with five predictors to use, so it satisfies the portrait-shape requirement pretty well.</p> <p>In our case, we want to create a binary variable named "quality" as the dependent variable, a binary with two possible answers: good or not good. It indicates if a wine is good enough BEFORE launching or not launching it on the high-end wine market. The third quartile for the quality score is 6.0, so if the score is above 6.0, it would be considered good, otherwise not.</p>
Insights 	<p>The insight is a probability: how likely the wine will be perceived as good on the market.</p> <p>The insight within the DIDA-Framework is the predicted value of a variable of interest. In other words, it is something that has NOT happened yet.</p>
Decision 	<p>Based on the insight, we could make decisions about whether to launch the wine to the high-end market or not.</p> <p>If the probability > 50%, it would be considered “good” (score above 6.0), and the company will launch the white wine on the high-end market, giving more resources on advertising and marketing to it.</p>
Advantage 	<ol style="list-style-type: none"> 1. It is cost-cutting because there is no need to prepare samples and send them to wine experts/customers to rate new wines. 2. Using these conclusions, companies can increase their odds of winning on the high-end wine market, gaining market share and presence. More precisely, wine companies can maximize

¹ Hao, L., Data Mining - Lecture 1, Fordham University, 2020.



	their profit and raise their brand awareness as a high-end wine producer.
--	---

PYTHON IMPLEMENTATION OF PREDICTION MODELS

The following chapters deep dive into our four prediction approaches: Logistic Regression, Classification Tree & English Rules, Neural Network, and k-NN. By doing so, we use Accuracy and AUC as a goal in order to compare and evaluate both methods eventually.

LOGISTIC REGRESSION

In order to make an appropriate analysis, we can use regression analysis to conduct results when the dependent variable is dichotomous. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

Procedures: After distinguishing between categorical variables and numerical variables, we drop the quality column, which is the benchmark for judging whether it is good quality. Then we import logistic regression to get coefficients for our variables using the penalty of 10. To get the best performance, we maximize AUC to get a different penalty and get a set of different coefficients. Also, we compare it with the best performance with accuracy.

Results	
Based on optimal AUC:	Based on optimal accuracy:
Alpha: 0.0736846	Alpha: 7.69159
AUC: 0.80077	Accuracy: 0.7887755



Conclusion: The top 5 predictions are density, residual_sugar, pH, fixed_acidity, volatile_acidity.	Conclusion: The top 5 predictions are density, residual_sugar, alcohol, volatile_acidity, chlorides.

CLASSIFICATION TREE & ENGLISH RULES

To choose a classification technique that performs well across a wide range of situations without requiring much effort from the analyst while being readily understandable by the consumer of the analysis, a strong contender would be the tree methodology.²

Procedures: After distinguishing between categorical variables and numerical variables, we drop the quality column, which is the benchmark for judging whether it is good quality. Then based on the AUC and accuracy, we import GridSearchCV to get different trees.

Results: It follows the root of alcohol, volatile acidity, density, pH, free sulfur, dioxide.	
Based on optimal AUC:	Based on optimal accuracy:
The level of depth of the best pruned tree is 5 and the AUC is 0.8249.	The level of depth of the best pruned tree is 15 and accuracy is 0.82857.
AUC: 0.80077	Accuracy: 0.7887755

² Shmueli, G., Bruce, P. C., Patel, N. R., & Gedeck, P. (2020). Data mining for business analytics: Concepts, techniques, and applications in Python. Hoboken: John Wiley & Sons.

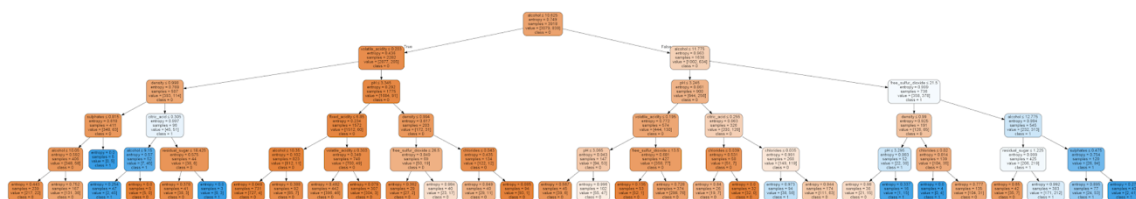


If we pick leaf **node ID = 44**,

Path = ['alcohol <= 10.625', 'volatile_acidity <= 0.20250000059604645', 'density > 0.9978799819946289', 'citric_acid <= 0.3050000071525574', 'alcohol <= 9.150000095367432', 'fixed_acidity > 6.450000047683716']

sample = 45 value = [0, 45] class = 1. Then we can get the predicted probability given by this node equals 100%.

The top 6 rules are if alcohol <= 9.15 and volatile_acidity <= 0.2025 and density > 0.998 and citric_acid <= 0.305 and fixed_acidity > 6.45, then it is high quality wine.



NEURAL NETWORK

Neural Network is a computer program that acts like a brain. It imitates how humans think when they learn something new. Neural Network involves the idea of nodes and layers. The input layer consists of nodes that could receive new information. The output of the input layer is the input of the next layer. The hidden layer is located between the input layer and the output layer. The hidden layer is like a black box, and it computes the data within the black box. Therefore, the Neural Network is flexible but hard to interpret.³

After preprocessing the dataset, use pandas to read the cleaned data. In order to perform the Neural Network model, we need to drop the quality column first. Otherwise, we will have a very high AUC that is almost close to 100%. Split the dataset and set the test part size to 0.2. Import the MLPClassifier, drop the dependent variable “good_quality_1”, set the alpha level to 0.1, which is the same as the logistic regression model, and set the hidden levels to 3. In this step, we get the weight for 11 predictors and three levels for each predictor. To find out the best Neural Network model, we need to run the model

³ Shmueli, G., Bruce, P. C., Patel, N. R., & Gedeck, P. (2020). Data mining for business analytics: Concepts, techniques, and applications in Python. Hoboken: John Wiley & Sons.



from the hidden layer 1 to 10, and alpha from 0.0001 to 10. The alpha number is 10. Use the test dataset to find out AUC and accuracy.

Results:	
Optimal penalty level	5.5556
AUC	0.842225123550468
Accuracy	0.8193877551020409

K-NN

k-NN model is an algorithm that measures the similarity of the sample data for each predictor and utilizes the similarity to classify new data. K (the nearest neighbor) represents a class in which the new data will be recorded. If the new record is closest to the k, it will be put in the same class as the nearest neighbor. Based on the existing data in the class, we can determine or predict the new record.⁴

The procedures of performing k-NN models are similar to the Neural Network model at the setting up stages. The significant difference is how to calculate the AUC and accuracy. After splitting the dataset, we will import the KNeighborsClassifier and set the `n_neighbors = 5` as pre-specify k to get the AUC score. Calculating the best AUC score runs the model with "Euclidean" from range 1 to 200. Then we could get the optimal k and highest AUC. We need to change the wording slightly and use the best `kNN.score` to come up with the accuracy score for accuracy. The highlight of coding is to measure each data's distance with the "Euclidean" function since the k-NN model uses closeness to measure the similarity between each data.

⁴ Shmueli, G., Bruce, P. C., Patel, N. R., & Gedeck, P. (2020). Data mining for business analytics: Concepts, techniques, and applications in Python. Hoboken: John Wiley & Sons.



Results:	
Optimal k	49
AUC	0.837974472245572
Accuracy	0.85

PREDICTION MODEL COMPARISON AND RESULT

Comparison:		
Techniques	Accuracy	AUC
Logistic Regression	0.7887	0.8007
Classification Tree	0.8285	0.8249
k-NN	0.8500	0.8279
Neural Network	0.8204	0.8421

Results:

Based on the comparison, we can conclude that logistic regression has the lowest accuracy of 0.7887 and AUC 0.8007. The classification tree's accuracy and AUC are both a bit higher than 0.82. KNN has the highest accuracy of 0.8500, but the AUC is just 0.8279.

The neural network is the opposite, because it has an accuracy of solely 0.8204, but the AUC 0.8421 is the highest among all four models. The accuracy and AUC of the four models are mainly the same, but the accuracy of the k-NN 0.8500 is much higher than 0.7887 of the logistic regression. Regarding AUC, the k-NN model, and the classification tree have nearly the same AUC 0.8249 and 0.8279, while Neural network's AUC 0.8421 is 0.0414 higher than logistic regression.



THE “PERFECT WINE” - SCENARIO

To virtually produce a "perfect wine" , which means it is predicted to be good by all our prediction models, we used the logistic regression model's insights. We checked the coefficients under the optimal penalty level and filtered all the wines' records getting a quality score of 9.

If the coefficient is positive, then this predictor has a positive impact on quality. The larger the predictor, the better the quality. Furthermore, if the coefficient is negative, then this predictor has a negative impact on quality. In this case, the rule of thumb is: "The smaller the predictor, the better the quality."

For a predictor with a positive coefficient, we make it the maximum number among the wines.

For a predictor with a negative coefficient, we make it the minimum number among the wines.




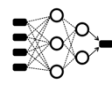
In the end, we came up with a wine with the following index, and it was predicted as good quality by all our models. So that is a perfect wine!

fixed_acidity(+)	volatile_acidity(-)	citric_acid(-)	residual_sugar(+)
9.1	0.24	0.29	10.6
chlorides(-)	free_sulfur_dioxide(+)	total_sulfur_dioxide(+)	density(-)
0.018	57	139	0.98965
pH(+)	sulphates(+)	alcohol(+)	
3.41	0.61	12.9	



CONCLUSION AND LIMITATIONS

In summary, we developed four prediction models for this dataset.

Techniques	Accuracy	AUC	Running time
Logistic Regression 	0.7887	0.8007	43s
Classification Tree 	0.8285	0.8249	26s
k-NN 	0.8500	0.8379	54s
Neural Network 	0.8204	0.8421	106s

From the absolute values perspective, each of our models achieves a good performance of around 80%. By cross comparing each model's result, each model only shows little difference. Overall, we believe our model offers reliable results and essential information to wine producers, which can lessen their workload and costs, gain more reputations for their brand, and make more company profits and grow business in the future.

However, there are still limitations on Accuracy vs. Interpretability to this dataset. We need both high accuracy and high interpretability algorithms for our high-stake scenario "if a wine's quality will be good enough to launch it to the high-end wine market."

While the accuracy of predictions raises in terms of accuracy, the interpretability gets more complicated. Therefore, even though our k-NN and Neural Network have the best AUC results compared to Logistic Regression and Classification Tree, there is still a tradeoff - the better accuracy of machine learning algorithm output, the less interpretable it becomes.

