Sophie Lyu, Michelle Tan, Michael Wu, Matt Zhang

MSCA 31008 2 Data Mining

March 18, 2022

**Identify the Most Relevant Features to Predict Wine Quality**

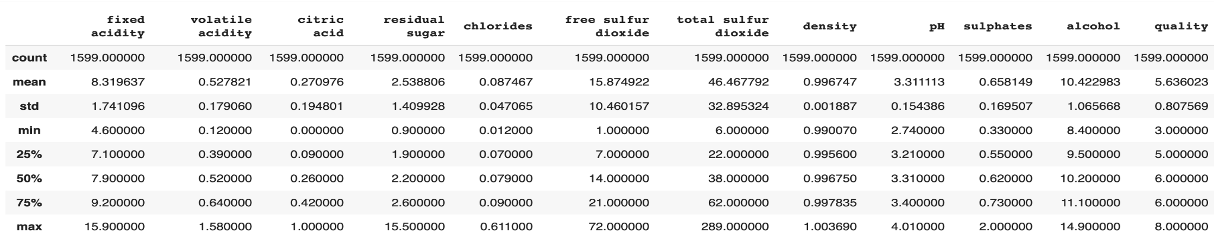
As the most popular adult beverage, red wine has formed a market that is worth $182 billion by 2020 and is expected to reach $278.5 billion in 2028.1 Not only is it intuitively make sense, but also mathematically proved that the price of red wine is related to its quality score in the positive way. 2 However, when taking a look at the checklist for wine scoring, these technical terms such as “bouquet”, “body”, and “room for improvement” usually fail to explain the physical features of a bottle of wine explicitly to amateurs. Moreover, agencies will be more than happy to estimate the quality of wine before they have to open a bottle of it and have a taste. Thus, being able to predict the wine quality given some chemical features that are measured ahead can be educational to individual consumers as well as profitably attractive to certificate agencies of wine. In order to identify the most relevant physicochemical features that influence the wine quality, we want to analyze a variety of physicochemical features of wine and each relation to the wine quality score. Ultimately, we intend to build a model to predict the wine quality with these physicochemical features as inputs.

The project intends to take four major steps, which will be discussed respectively in the following paragraphs. First, an exploratory data analysis of data will be conducted to detect if there’s any missing values, the distribution of each variable and potential correlation between each variable and the responsive output. Then, the dataset will be split into a training group and a test group. A number of models will be employed to study the patterns among the training data. Next, models that come up from the last step will be implemented on testing data and the performance of each model will be evaluated with the respects of accuracy scores and a number of other measurements. Lastly, the conclusion on our findings and a reflection on potential improvements will be presented.

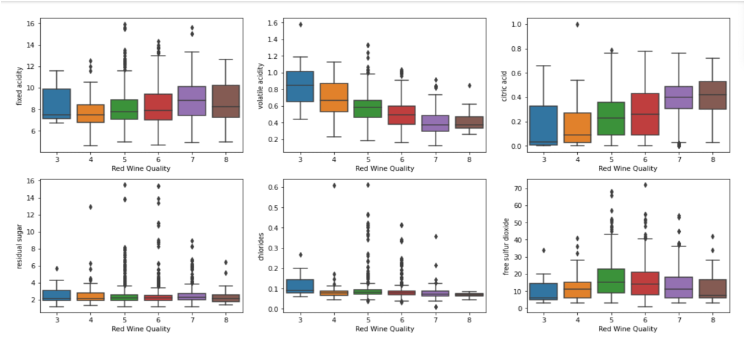
1. **Data Collection and Exploratory Data analysis**

To fulfill the goals of this project, we collect 1599 samples of wine produced in Vinho Verde, Portugue. 11 physicochemical features and quality scores are recorded in the dataset of these samples. These physicochemical features include density, pH value, fixed acidity, alcohol, etc. that are measured in numeric scale.

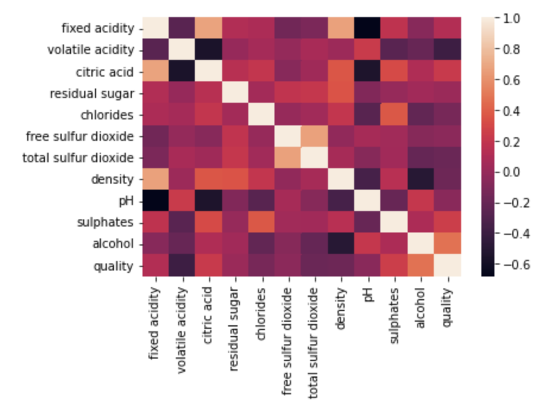
The process of exploratory data analysis starts with checking if there is any missing value in each variable column. With a confirmation of no missing value, statistical features of each variable are captured next. The output, wine quality score, ranges from 3 to 8 with a normal distribution centered on score 6, which suggests that our samples contain both below and above average quality wine. The range of pH value is pretty narrow, which is from 2.74 to 4, which covers the normal range that red wine falls into.3 Together, it can be inferred that the dataset contains a moderate range of wine with normal quality, but does not include many corner cases, such as wine includes extreme acidity, extremely low quality or extremely high quality (Chart 1). When looking at the distribution of each variable, it is found that most variables tend to skew toward the right (Chart 2). Correlations are clear between fixed acidity and density, fixed acidity and pH value, fixed acidity and citric acidity, free sulfates and total sulfates, density and alcohol, sulfates and pH value (Chart 3). In combination with the fact that the output variable as integers located between 0 and 10 is not continuous but discrete, it is determined that classification methods will be a better option for us instead of regression in this project.



(Chart 1)



(Chart 2)



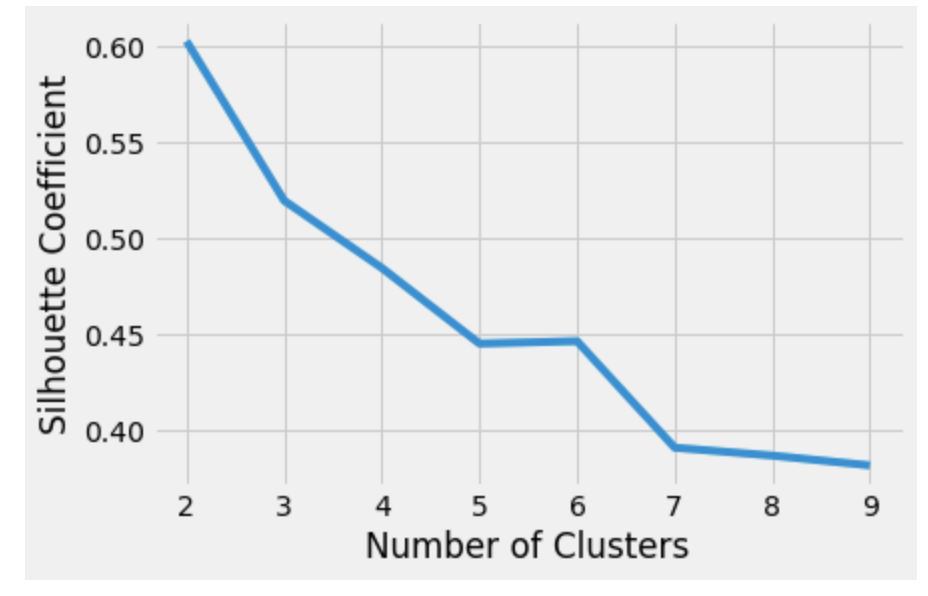
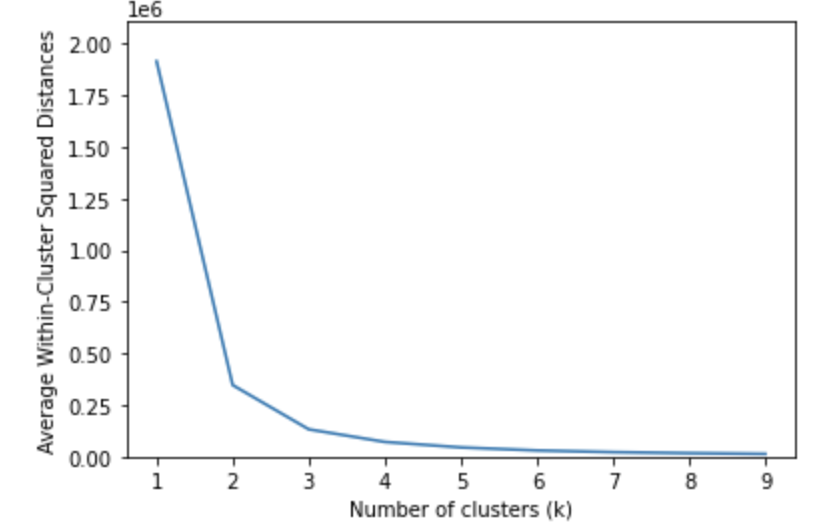
(Chart 3)

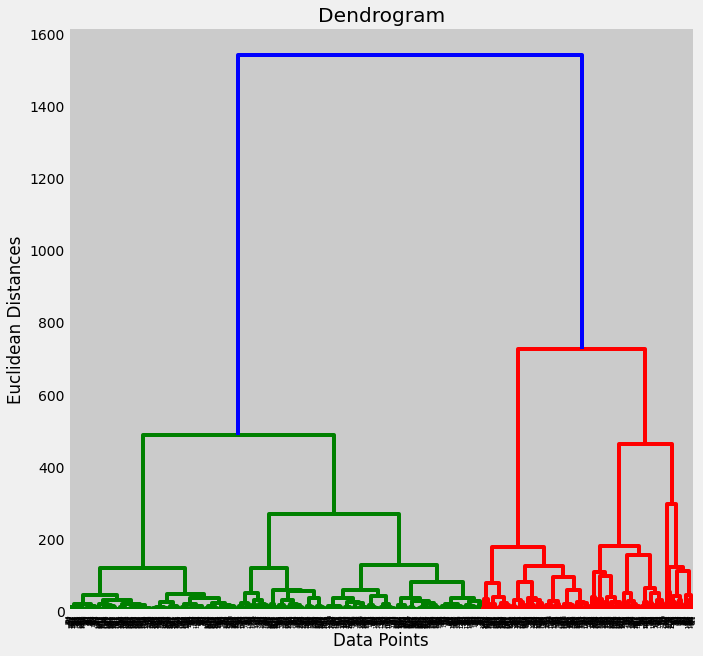
1. **Model Implementation**

**Unsupervised Model**

* 1. **K-Means and Hierarchical clustering algorithm**

We are first curious if we can try to make label clusters for the data set applying unsupervised learning models of K-Means and Hierarchical Clustering Algorithms. For the K-means clustering calculations, we applied both the elbow method, which conducts the optimal numbers of clusters by measuring WCSS ( Within-Cluster Sum of Square )’s decreasing trend, and the silhouette method, which computes silhouette coefficients of each point in order to measure if itis similar to its own cluster compared to other clusters. Both of the methods indicate an optimal number of clusters is 2, and therefore divide our red wine into two clusters of 1179 wines and 420 wines. Interestingly, the hierarchical clustering method also indicates two as the optimal clustering choice, with 1334 wines and 265 wines.





**Supervised Models**

1. **Linear Regression**

As mentioned above, linear regression is not an appropriate method in this case given high correlations among variables and discrete nature of the responsive variable. Thus, although we receive considerably low mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE) by running the linear regression algorithm, the model itself is not suitable to apply to the testing data.

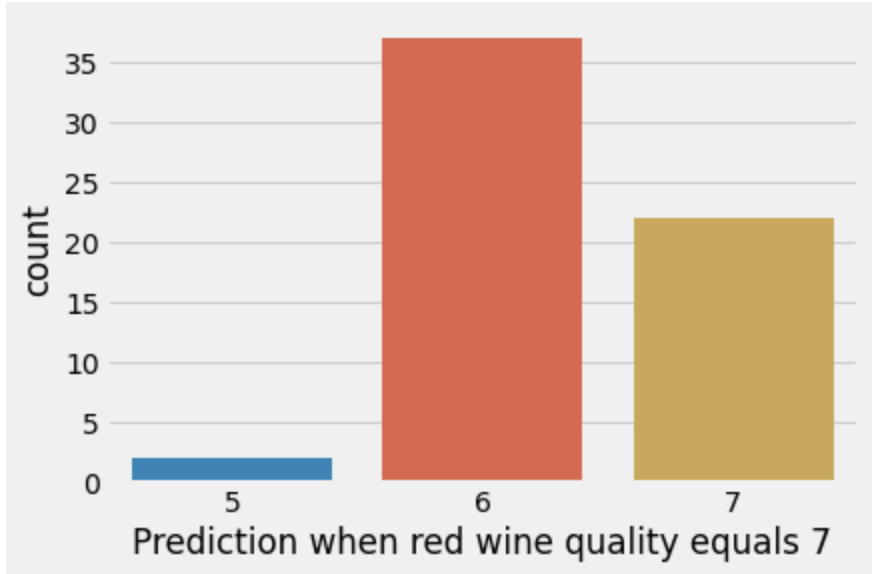
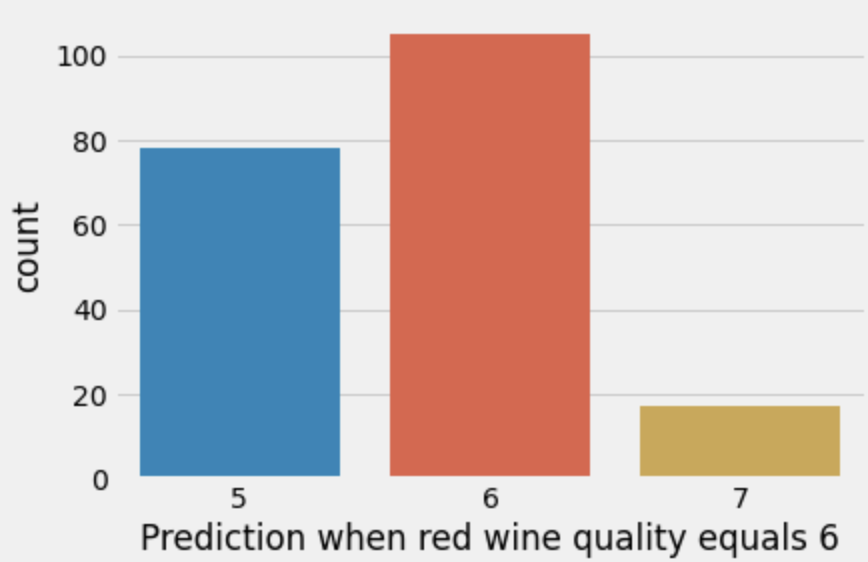
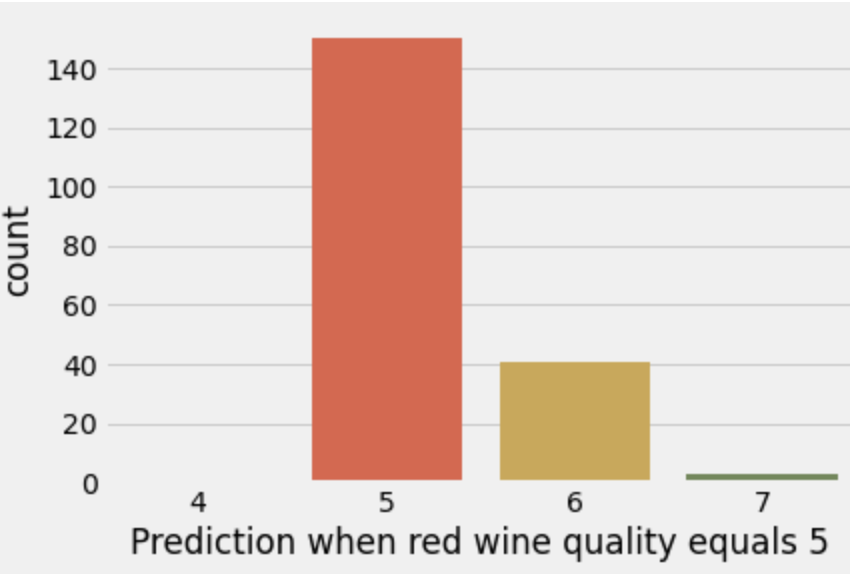
1. **Decision Tree**

Decision tree model is a good fit to this classification problem because it is computationally efficient, easy to interpret, robust to noises, and able to handle the categorical target variable well. The models are constructed with four different variable selections. The first one is the full model that utilizes all features in the dataset as inputs. The second one is the partial model that only uses six features that are shown to be correlated with the red wine quality. The third and fourth ones take even less features: it only contains the top three and five most important features obtained from the full decision tree model. Then, these models are optimized by adjusting the hyperparameters max\_depth based on MAE, MSE, and RMSE.

| Model | Accuracy | CV Accuracy | F MAE | MSE | RMSE |
| --- | --- | --- | --- | --- | --- |
| Full Model | 0.58 | 0.57 | 0.46 | 0.53 | 0.73 |
| Partial Model | 0.55 | 0.55 | 0.49 | 0.59 | 0.77 |
| Top-3 Features | 0.55 | 0.58 | 0.48 | 0.55 | 0.74 |
| Top-5 Features | 0.57 | 0.57 | 0.48 | 0.58 | 0.76 |

Based on the summary table above, the full model generates the highest accuracy with the lowest error metrics. The third model with 3 features can also achieve a reasonably good accuracy with much less inputs. But given that the overall accuracy is not that high, we are inclined to pick the full model as the final decision tree model.

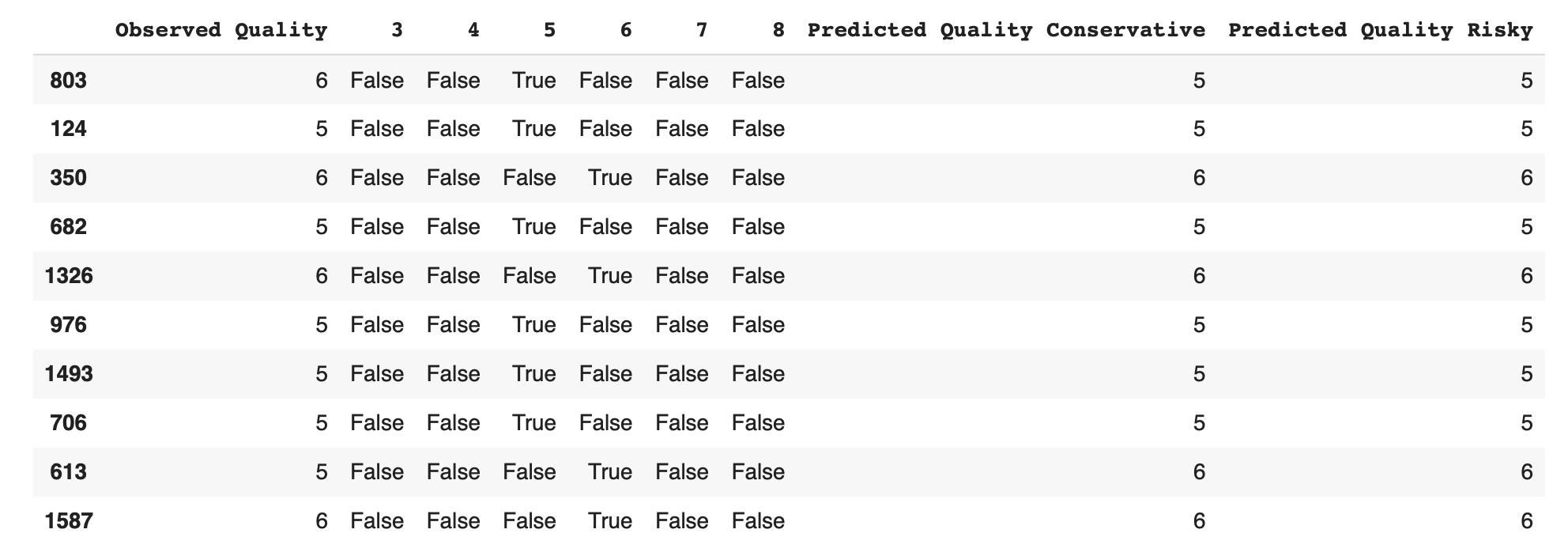
The final model, however, has relatively higher recall and lower precision, implying that the model is more likely to have type I error. In other words, the error the model inclines to make is to predict the red wine quality but it is in fact wrong. However, because our target variable has multiple categories, we need a more in-depth analysis.



When the model incorrectly predicts the quality to be 5, it is more likely to classify it to be 6 (over-predict). When the model incorrectly predicts the quality to be 6, it is more likely to classify it to be 5 (under-predict). When the model incorrectly predicts the quality to be 7, it is more likely to classify it to be 6 (under predict). This prediction error might be caused by either the size of the data (so that the model cannot learn much from the training set) or the closeness of the red wine with quality 5, 6, and 7.

1. **Support-Vector Machines**

Although the SVM model can only handle binary classification, we can convert the categorical red wine quality into multiple dummies and train SVM models on each dummy to solve this problem. At the same time, SVM is good at handling overfitting, irrelevant and redundant attributes, and robust to noise. The model construction process is more challenging compared to all other models. After converting the red wine quality 3-8 into 6 dummies and training 5 SVM on quality 3-7respectively, we can obtain the last red wine quality 8 binary prediction based on the previous 5 dummies (red wine quality 8 equals ‘True’ when all other predicted binary predictions equal ‘False’, and vice versa). The table below shows an example of the prediction result.



However, because we are creating multiple binary SVM predictions individually, it is possible that more than one SVM model predicts ‘True’ at one observation. For example, some observations have two ‘True’ values at both red wine quality 5 and 6 binary prediction. To solve this problem, our conversion from 6 binary outputs to 1 categorical output is conducted in two ways. The conservative way is to use the first ‘True’ value and the risky way is to use the last ‘True’ value in the row as the overall red wine quality prediction..

| Model | Accuracy | CV Accuracy | MAE | MSE | RMSE |
| --- | --- | --- | --- | --- | --- |
| Conservative | 0.54 | 0.54 | 0.69 | 1.27 | 1.13 |
| Risky | 0.53 | 0.54 | 0.69 | 1.23 | 1.11 |

The summary table above shows the model accuracy and error metrics of the two prediction results. It is obvious that they both have almost the same performances but are still not as good as the decision tree model.

1. **Random Forest**

Random Forest is an ensemble algorithm of decision trees, and thus has a better ability to generalize and reduce risk of overfitting than decision trees. GridSearch was used to fine-tune the hyperparameters. Gini was selected as the criterion, with minimum samples required to be at leaf node equals to 1, minimum samples required to split a node equals to 4, and number of estimators equals to 300. Multi-class model was built to predict original scores (3-8) of wines, and achieved highest accuracy and cv accuracy among all models trained, as well as lowest MAE, MSE, RMSE among all models. A binary model was also trained to classify wines as normal quality if quality score is between 3 and 6, and good quality if quality score is 7 or 8. The fine-tuned binary random forest model yielded 0.90 accuracy score. The reason why training a binary classification model yielded better results was primarily due to the subjectivity nature of institutions rating wine qualities. What separates a wine with rating = 7 from a wine with rating = 8 is hardly explainable from a quantitative perspective given the data we collected, but Random Forest does a much better job at differentiating a good (7-8) wine from a mediocre (3-6) wine.

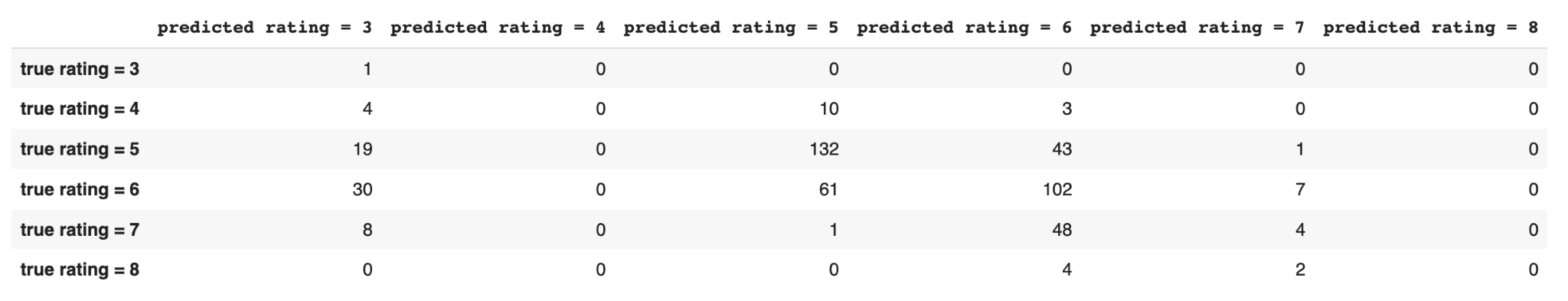
| Model | Accuracy | CV Accuracy | MAE | MSE | RMSE |
| --- | --- | --- | --- | --- | --- |
| Multi-class | 0.67 | 0.68 | 0.36 | 0.40 | 0.63 |
| Binary | 0.90 | 0.90 | 0.10 | 0.10 | 0.32 |

1. **Neural Network**

Neural Network was trained using the MLPClassifier package from sklearn. GridSearch was used to fine-tune the hyperparameters, including number of hidden layers, activation method, and number of neurons in each hidden layer. Fitting a neural network with 2 hidden layers (12,10) to the training set gives 0.57 accuracy on the test set and 0.582 cv accuracy.

| Model | Accuracy | CV Accuracy | MAE | MSE | RMSE |
| --- | --- | --- | --- | --- | --- |
| Neural Network | 0.57 | 0.582 | 0.742 | 1.458 | 1.208 |

While accuracy scores are approximately the same as the decision tree model, neural network’s MAE, MSE, and RMSE are the highest among all models, indicating that the neural network model tends to give more inaccurate results when predictions are wrong. The table below shows that the model predicts more 3’s when true rating = 5 and 6.



1. **Ensemble Method**

While all 6 models, both supervised and unsupervised, above can achieve accuracy as high as 67%, it is possible that an ensemble model that combines more than one classification model above can have a better performance. We use the soft voting ensemble method to put the ‘good models’ together.

| Model | Weight | Accuracy | CV Accuracy |
| --- | --- | --- | --- |
| Neural Network +  Random Forest | 1:3 | 0.65 | 0.68 |
| Neural Network +  Random Forest | 1:5 | 0.66 | 0.68 |
| Neural Network +  Random Forest | 1:10 | 0.66 | 0.68 |
| Neural Network +  Random Forest + Decision Tree | 1:3:1 | 0.65 | 0.67 |
| Neural Network +  Random Forest + Decision Tree | 1:5:1 | 0.66 | 0.68 |
| Neural Network +  Random Forest + Decision Tree | 1:10:1 | 0.67 | 0.68 |

Because random forest and decision tree models share the similar mechanism but random forest generally has better performance, we first try the combination of the neural network and random forest and try different weights. Based on the summary table above, we can see that the accuracy increases as we give random forest models a greater weight, and the accuracy finally reaches a steady state that is almost the same as the accuracy of the final random forest model. We then explore the ensemble method with the decision tree model and try different weights combinations. The results, unfortunately, stay the same: for ensemble methods to achieve a higher accuracy, we need to give random forest models more weight. But at the same time, the ensemble method will generate similar results as the original random forest model.

1. **Model Summary and Final Model Selection**

|  | Accuracy | CVAccuracy | MAE | MSE | RMSE |
| --- | --- | --- | --- | --- | --- |
| Decision Tree | 0.58 | 0.57 | 0.456 | 0.527 | 0.726 |
| SVM | 0.54 | 0.54 | 0.690 | 1.269 | 1.126 |
| Random Forest | 0.67 | 0.68 | 0.356 | 0.402 | 0.634 |
| Neural Network | 0.57 | 0.582 | 0.742 | 1.458 | 1.208 |
| Linear Regression | NA | NA | 0.520 | 0.430 | 0.656 |
| Ensemble Method (soft voting) | 0.67 | 0.68 | 0.352 | 0.402 | 0.634 |

Since Random Forest model has highest validated accuracy, cross validated accuracy, and lowest MAE, MSE, and RMSE, we chose Random Forest as the model we would like to use to accomplish our business objective.

1. **Conclusion**

After comparing all the indexes, we pick Random Forest as our best model to present. One might ask if such an accuracy score of 0.67 can be regarded as a good model fit. We, however, are confident that this is a reasonably good model fit of this data set after comparing with the legit literatures studying the same dataset. The accuracy score we obtained is above the average accuracy scores we observe from those academic and industrial literatures. (references listed in bibliography from number 4 to 8) We can indeed improve the accuracy score by turning the categorical red wine quality into binary one (good / bad red wine). But we assured that keeping quality as the original numerical target variable is more conformed with real-life red wine quality assessment rating process. With our machine learning algorithm implementation, we are able to provide an automation mechanism that generates substantial business profit for clients such as winery manufacturers and certification agencies. Our model helps them better understand the factors and easily use the predictions to control and improve wine quality and further decrease production cost. They can also categorize their wines more precisely and launch different marketing strategies towards various customer segments, which methods can effectively boost marketing returns and lower marketing costs.What’s more, such an automation mechanism can also reduce investments in building up huge expertise groups for wine testing and facilitate the entire certification process, which minimizes unnecessary manual errors during the operation of assessments and assurance.

However, we also find some limitations. First, the main problem came from the fact that our data set was unbalanced. A majority of the quality values were “regular” (5 and 6), which made it harder to identify each factor’s different influences on a “high” or “low” quality of the wine, So we need to collect more balanced data to represent different types. Another limitation worth mentioning is data insufficiency. From the existing dataset, we only have less than 1500 data points and 11 attributes, which might narrow down the accuracy of our predicting quality of red wine. The solution for this is to take the pre-existing data to generate more samples and include more relevant data features, like the year of harvest, brew time, location, or wine types. We also regard our dataset inaccurate. Features Ph and SO2, are having the correlations not conform to common winery production rules. SO we suggest tracing back to the original dataset and improve the quality of data.

To sum up all, our project successfully determines which features are the best quality red wine indicators and generates insights into factors to our model. And we believe that it can further bring up business values to our customers as winery manufacturers and certification agencies. We will expand this analysis to include feature development methods to test whether or not the model's predictive power may be increased. This is our presentation for today. If you have any questions, don’t hesitate.

Bibliography

1. <https://www.alliedmarketresearch.com/red-wine-market-A13400#:~:text=The%20red%20wine%20market%20size,dominance%20during%20the%20forecast%20period>.
2. <http://winegourd.blogspot.com/2019/08/the-relationship-of-price-to-wine.html>
3. <https://sensorex.com/2017/12/06/ph-improve-taste-color-wine/#:~:text=Understanding%20the%20basics%20of%20pH%20in%20wine%20making&text=Usually%2C%20a%20wine%20will%20fall,falling%20between%203.3%20and%203.6>.
4. <https://ieeexplore.ieee.org/abstract/document/9104095>
5. <https://www.scirp.org/journal/paperinformation.aspx?paperid=107796>
6. <https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377>
7. <https://www.researchgate.net/publication/340626215_Wine_Quality_Prediction_Using_Data_Mining>
8. <https://www.researchgate.net/publication/221612614_Using_Data_Mining_for_Wine_Quality_Assessment>