Project Milestone 4

Harlan Wittlieff

Data Science, Bellevue University

DSC 630: Predictive Analytics

Dr. Brett Werner

May 15, 2022

Introduction

With the almost infinite options winemakers have at their disposal when creating a new wine many find themselves asking, “What incorporates a successful wine?” Is it alcohol content, sweetness, another metric, or a combination of multiple characteristics? Knowing the foundation of a good wine can turn a failing winery into a thriving one. In today’s competitive environment, unlocking the formula for a successful wine is the advantage that can set a winery above the competition.

To investigate this question, a dataset was selected from UCI’s machine learning repository. The “Wine Quality Data Set” contains 4898 instances of white wine and 1599 instances of red wine of the “Vinho Verde” variety. The dataset contains eleven total characteristics for each individual wine. These features are listed below.

- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide
- Density
- pH
- Sulphates

- Alcohol

In addition to the features listed on the previous page the data set contains an output variable, quality. The quality metric is built from sensory data from wine experts. These experts graded each wine based on a scale from 0-10 with 0 equating to a “very bad” score and 10 being “excellent.” The final quality metric is the median from at least three of these scores for each wine.

From the combination of features a model will be built to predict a wine’s quality. The initial model to be evaluated will be k-nearest neighbors. This model was selected based on the assumption that features of a good wine will fall into groups. In other words, good wines will have similar characteristics. Based on the success of this model, regression and random forest may also be considered. All models will be built with a training and test set of data. The existing data will be randomly split placing eighty percent of the data into the training set and twenty percent into the test set to evaluate the model.

After the model is completed, it must then be evaluated. The primary tool for understanding the model’s success will be a confusion matrix of the model’s predictions on the test data set. The confusion matrix will be used to understand the model’s overall accuracy in addition to identifying the individual areas where the model succeeds and fails. Priority will be given to models that successfully identify high ranking wines (wines with a quality score of 7 or higher) as these wines contain the characteristics that need to be identified by the model. Misclassification of a poor-quality wine (below 6) as a high-quality wine will also be considered more harshly.

Based on the success of this model, additional types of wines may be investigated to further expand the overall scope and applicability of the model. This additional insight will allow the model to be more relevant to wineries that create wines outside of the initial project's scope.

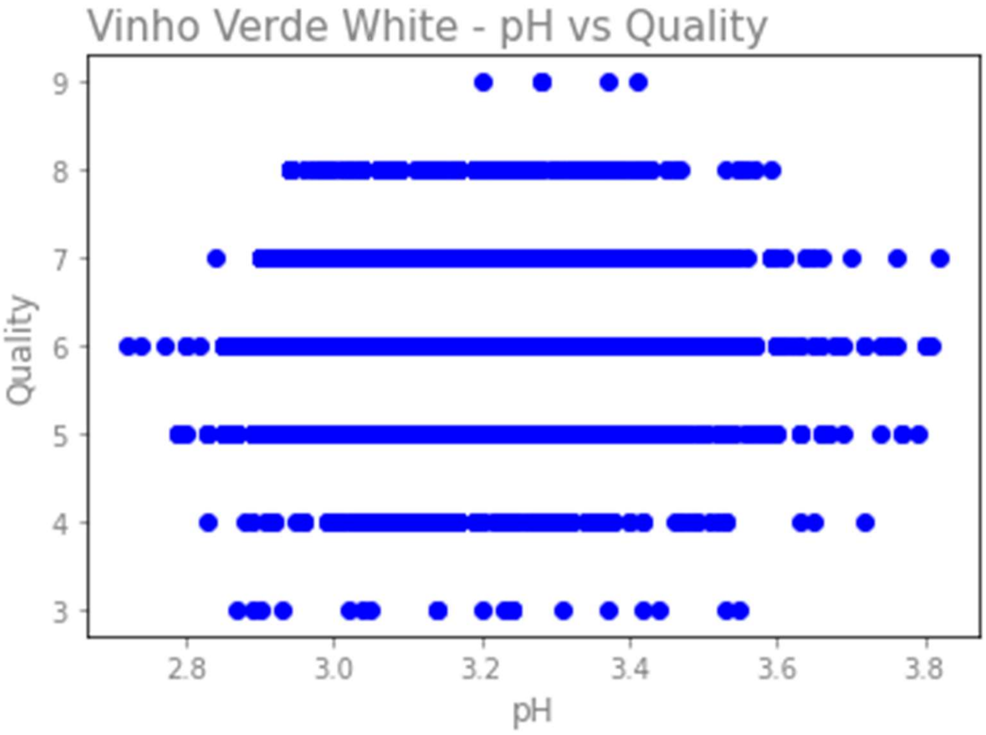
If the model successfully predicts the quality within the test data set, there are still a few risks that warrant investigation. The first risk is that the quality metric is subjective. This metric is based on the opinion of an expert and may not translate well to other populations. Secondly, the subjective nature of the quality metric brings the risk of inconsistency. If the quality metric was not created in a consistent fashion, any model may have difficulty in predicting it. Additionally, just because a wine is considered high quality does not mean the wine will be a financial success. Other factors (marketing, economic conditions at the time, etc.) will play a role in determining an individual's wine's overall acceptance within a population.

Should this original project plan for the evaluation of wine quality not work out as anticipated, a contingency plan is in place. This contingency plan consists of investigating additional dataset to uncover new features of wine and potentially additional correlations through the investigation of the relationships between the new and existing features and their impact on the wine's quality. The lack of unique identifiers could cause this contingency plan to be more challenging.

Preliminary Analysis

I am confident that I should be able to identify the key characteristics that make up a superior Vinho Verde wine using the Wine Quality Data Set. As I investigate the relationships between specific variables and their impacts on the overall quality of wine, plotting the variables against wine quality on a scatter plot will be most useful as this visualization technique allows for immediate visual cues to identify correlation. Additionally, I plan to gain insight from the overall distribution of key variables through leveraging histograms.

As an example, insight can be gained into pH's impact on quality by creating a scatter plot of the two variables. As the below chart demonstrates, white wines that achieved an extremely high rating of 9 for quality had a narrow pH range of approximately 3.2-3.5.



At this time I do not anticipate needing to adjust the data or my driving questions. I may however need to adjust my model and evaluation choices. Based on the success of the KNN model, additional models may be explored such as decision trees and regression models.

Data Preparation

To begin prepping the data, I checked for null values within the dataset. Fortunately, this dataset did not contain any null values. I then checked the distribution of the quality values for both red and white wines. Since this analysis is primarily focused on identifying high quality wines, I then decided to convert the quality metrics to a binary high or poor quality field. Quality ratings of a seven or higher were assigned a value of high and anything lower than a seven was assigned a quality value of poor.

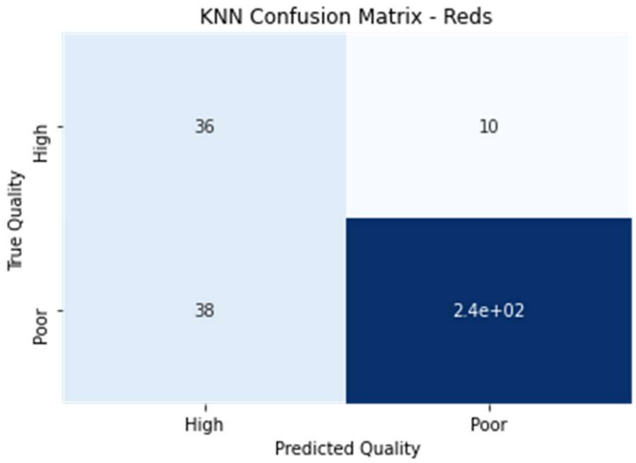
After the initial data was cleaned, it was then necessary to build training and test sets for both the red and white wine databases. Utilizing Sklearn's train_test_split, both databases were split into train and test sets with twenty percent of the data retained for testing. This allowed for eighty percent of the data to be used to train the models.

Following the training and testing data split, I then checked the distribution of the training data. After checking the distribution, I found that the classes in the red dataset were unbalanced with 171 examples of high quality and 1108 examples of poor quality. Additionally, the white wine training dataset was also unbalanced with 852 examples of high-quality wines and 3066 examples of wines of poor quality. In order for the models to properly learn from the low quantity of high-quality examples in both datasets, the training sets were rebalanced with SMOTE. The final step of the data preparation was to scale the features.

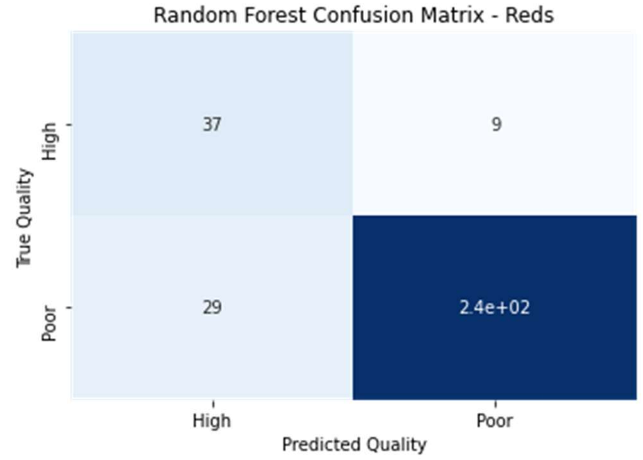
Model Building & Evaluation

The first KNN model for the red wines was set with a baseline of four neighbors. This model had a total accuracy of seventy-five percent. To validate if the model's accuracy could be improved by adjusting the number of neighbors, a grid search was created to evaluate neighbor values of one through twenty-five. The grid search found that having one neighbor provided the best results. The one neighbor KNN model improved the accuracy to eighty-five percent.

To further evaluate the quality of the model predictions, a confusion matrix was created. The KNN model successfully identified thirty-six high quality wines. Unfortunately, the model missed ten high quality wines and misclassified thirty-eight low quality as high.

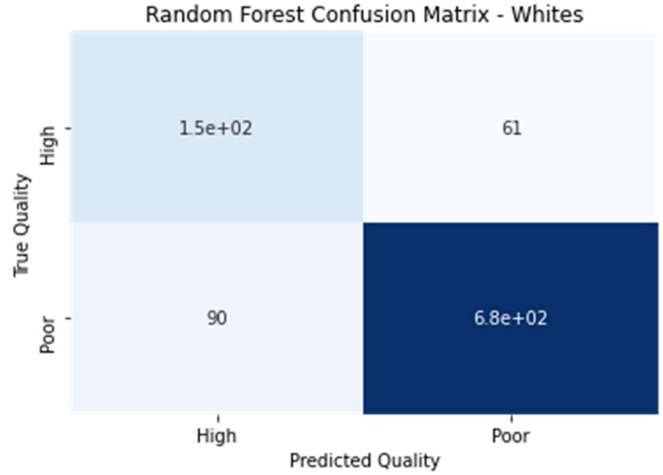


To investigate if further improvements could be gained by leveraging another type of model, I then proceeded to build a random forest model. This model utilized the same data as the KNN model with the exception of the features not being scaled. Once again, to evaluate the quality of the model predictions, I created another confusion matrix.



The random forest model successfully identified thirty-seven out of forty-six of the high quality wines. This is an improvement of one high quality wine when compared to the KNN model. The largest gain was seen in the reduction of false positives. This model only falsely identified twenty-nine wines compared to the KNN model results of thirty-eight.

Due to the increased success of the random forest model, I chose this model to be utilized on the white wine database. Again, a confusion matrix was built to evaluate the results.



This model successfully identified one-hundred and fifty of the high-quality white wines. However, it misidentified sixty-one high-quality wines as poor-quality wines. The model also had ninety false positives.

Recommendations & Conclusions

The models will be a useful tool for evaluating red and white wines prior to being mass produced. The models could also be leveraged prior to wines being tested and rated by critics. However, due to the rate of false positives and false negatives within the models, additional research into other characteristics that could potentially increase the model's accuracy may be warranted.

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (n.d.). *Wine Quality Data Set*. UCI Machine Learning Repository: Wine quality data set. Retrieved March 25, 2022, from https://archive.ics.uci.edu/ml/datasets/Wine+Quality

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.