

Predicting White Wine Quality



Geoffrey Hart

TaeHoon Kim

Yingfen Huang

Monica Montano

IST 348 – BIG DATA

06/29/2018

1. Abstract

A predictive model is developed by utilizing machine-learning to estimate white wine quality using Microsoft Azure. This paper introduces several data preparation methods, such as data normalization and SMOTE. To achieve the best results, two different types of training models were constructed: one analysis approach is regression to predict the wine quality value from 0 to 10; another preferred model is the multiple-class classification which predicts quality categories.

2. Literature Review

The wine industry has undergone major changes in the past seventeen years. Prior to 2001, the wine gallons drunk per resident in the United States was about 2.01 gallons. Beginning in 2002, the total number of wine gallons drunk began to increase significantly, with a reported total of 2.94 gallons per resident (total of 949 million gallons of wine) drunk in 2016 (Wine Institute, 2017). This increase in wine consumption correlates with the millennial generation (also known as the Y-generation) becoming of legal age to drink in the United States. After the Baby Boomer generation, millennials are the fastest growing wine consumers in the United States. Research conducted by The Wine Market Council in 2016, reported 33% of millennials (individuals aged 23-40) were wine drinkers, undoubtedly higher than the 21% of Generation X (individuals aged 41-52), and only slightly lower than the 36% of the Baby Boomer generation (individuals aged 53-71) (The Wine Market Council, 2017). To keep up with the increase in demand, wine makers worldwide have increased production. Portugal is a top ten wine exporting country, with exports increasing 45% from 2011 to 2015 (Wines of Portugal, 2018). To support its growth, the wine industry has invested in new technology for the wine making and wine

selling process (Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J., 2009). Wine certification and quality assurance are the measures used to prevent the illegal corruption of wine and ensure quality wine enters the market whereas, quality evaluation is used to enhance wine by identifying significant factors. There are two methods employed to determine the quality of wine, sensory tests and analyzing the physicochemical properties of wine. The first method uses the senses of sommeliers to determine the quality of wine. This method can be problematic, not only because it is subjective to each person, but there are no standard procedures to measure the overall sensory quality of wine as it lacks scientific and statistical foundation (De Mets, G. , Goos, P. , Hertog, M. , Peeters, C. , Lammertyn, J. and Nicolaï B., 2017). The other method of wine certification is objective and focuses on laboratory tests to characterize the physicochemical properties of wine (Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J., 2009).

There are a couple of research papers that used the same dataset to try and predict the quality of wine. The first research paper by M. Horak, uses Matlab to run a regression tree algorithm to predict the quality of wine. The author set three parameters training fraction, splitmin, and nFolds with a result of 88% accuracy (Horak, M., 2009). The second research paper by Nachev, A. and Hogan, M. explores four predictive techniques neural networks (NN) a.k.a. multilayer perceptrons (MLP), cascade-correlation neural networks (CCNN), general regression neural network (GRNN), and support vector machines (SVM) to predict the quality of wine. Results indicated that SVM could be a better alternative of prediction models based on neural networks (Nachev, A. & Hogan, M., 2013).

3. Problem Definition

The increasing demand for wine and the increase in worldwide wine consumption has led to many newcomers in the wine producing market. The market has become highly competitive, with each producer looking to find a niche. As previously stated, millennials have become the main consumers and target market for wine producers. Millennials are also known to be very price conscious. As such, we believe that focusing on quality and reducing price will be the best way to cater to millennials, as supported by the Marginal Utility Theory.

The marginal utility theory is a concept that places a qualitative value for the satisfaction someone receives when purchasing a product, and the change in satisfaction when additional products are added to the purchase while considering the price of the product purchased. Regarding wine, research states that those who purchase wine of a higher quality have a higher average marginal utility, therefore there is more satisfaction in purchasing a good quality wine. (Gibbs, M., Tapia, M., & Warzynski, F., 2009). This is an important feature for wine makers and distributors to consider when releasing wine into the market, wine makers can focus on producing better quality of wine for a lower price and increasing the marginal benefit of the consumers. Hence, we wish to answer the following:

1. Can the quality of wine be predicted by analyzing its physicochemical properties?
2. What properties are the most important in predicting the quality of wine?

Also, we would like to know if we could answer these questions while not including any features that may influence the subjective result? (i.e. price, label, sustainability).

The goal of the research is to determine which physicochemical properties are strongly correlated with the subjective quality score and to make a predictable model that determines the

quality of wine based on the properties, using machine learning algorithms. Wine producers will then be able to focus on these properties and cater to the consumers.

4. Data Selection and Acquisition

The dataset utilized provides details of Portuguese white wine consisting of a total of 4898 cases and 12 variables. The independent variables are all continuous numeric values that reflect the physicochemical features of the wine. The dataset has one dependent variable named 'Quality' measured as ordinal integer and contains a value range from 0 to 10. A score of zero is poor quality wine and a score of 10 is excellent quality of wine.

Fortunately, the dataset was accessible via the Pennsylvania State University website. It was provided in the .csv format that was compatible with Microsoft Excel and Azure. Therefore, we encountered no problems when uploading to Azure. There were also no missing values within the dataset, and no additional issues regarding the selection.

5. System and Tools

The data analysis was completed using Azure Machine Learning Studio, the cloud-based computing platform. A majority of the writing was done on both PC laptops running Microsoft Windows 10 operating system and Apple laptops running OS High Sierra. All project documents were accessible to each group member via two cloud-based platforms, Google Drive and Microsoft One Drive. The document sharing platforms were chosen to enable the group to work on the paper and presentations simultaneously. Azure Machine Learning Studio was the required program to use for the data analysis and model building. The benefits to this application allowed

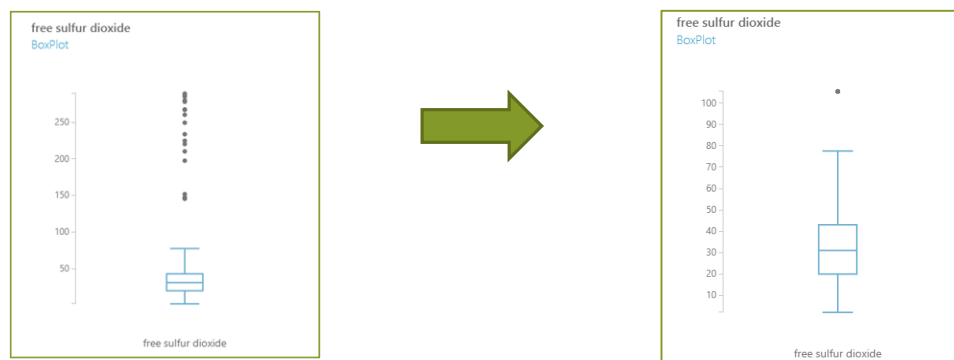
for open, flexible, enterprise-grade cloud computing platforms. Models were easy to construct, and the application enabled sophisticated visualization of the results.

6. Methodology

- **Outliers**

Most of the features are negatively skewed. To ensure the outliers did not negatively influence the data analysis, the ‘clip values’ module was used to remove the outliers. The upper percentile threshold was set at 99%, and the upper substitute value was set at a threshold which converted the top 1% of outliers into the value of threshold. This process allowed for the replacement of most outliers.

The following images show examples of the change:



- **Duplicates**

In the dataset, about 25% of the cases were thought to be duplicates. Upon further research, a separate study stated that the dataset has only distinct rows [1]. Thus, the duplicate cases were in fact wines of different brands that had the same feature values. This fact confirmed the underlying theory that the physicochemical properties are correlated to the quality of wine, as the same physicochemical features of the cases also had the same corresponding quality value. It

was further discovered that by keeping the duplicate rows, only about a 4% difference in accuracy was found, thus not materially impacting the results of the predictive model. The figures below reflect the results of both analyses. The model on the left shows metrics that include the duplicate rows, and the model on the right shows the result without the duplicate cases.

Metrics

Overall accuracy	0.903537
Average accuracy	0.935691
Micro-averaged precision	0.903537
Macro-averaged precision	0.908288
Micro-averaged recall	0.903537
Macro-averaged recall	0.903025

Confusion Matrix

		Predicted Class		
		med	high	low
Actual Class	med	90.7%	3.7%	5.7%
	high	7.5%	91.7%	0.8%
	low	10.9%	0.6%	88.5%

Metrics

Overall accuracy	0.862627
Average accuracy	0.908418
Micro-averaged precision	0.862627
Macro-averaged precision	0.875664
Micro-averaged recall	0.862627
Macro-averaged recall	0.85647

Confusion Matrix

		Predicted Class		
		med	high	low
Actual Class	med	88.1%	3.7%	8.2%
	high	17.8%	81.7%	0.5%
	low	12.6%	0.4%	87.1%

• Normalization

Each of the features in the dataset have different units of measurement. For instance, when the feature ‘Chlorides’ was to be compared to the feature ‘Total Sulfur Dioxide’, the difference between the both measurements was significant. A concern was that the inconsistency may mislead and impact of the analysis negatively, thus reducing the accuracy of the prediction models. To standardize the features, the module ‘normalize data’ was used. The module ensured the features were all adjusted to a standard of having a mean of 0 and a standard deviation of 1. The first figure shows the features prior to the standardization and the second figure shows the features after normalize model, indicating the features have been standardized.

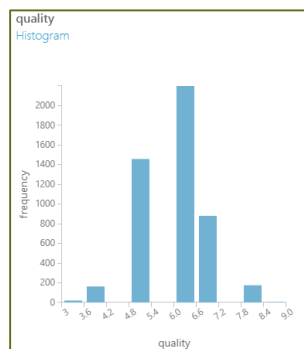
FEATURE	MIN	MAX	MEAN	STANDARD DEVIATION
FIXED ACIDITY	3.800	14.200	6.893	0.887
VOLATILE ACIDITY	0.080	1.100	0.305	0.121
CITRIC ACID	0.000	1.660	0.325	0.116
RESIDUAL SUGAR	0.600	65.800	5.728	4.615
CHLORIDES	0.009	0.346	0.045	0.022
FREE SULFUR DIOXIDE	2.000	289.000	33.269	21.058
TOTAL SULFUR DIOXIDE	9.000	440.000	132.412	45.518
DENSITY	0.987	1.039	0.994	0.003
PH	2.720	3.820	3.195	0.142
SULPHATES	0.220	1.080	0.487	0.116
ALCOHOL	8.000	14.200	10.732	1.292

FEATURE	MIN	MAX	MEAN	STANDARD DEVIATION
FIXED ACIDITY	-3.562	3.013	0.000	1.000
VOLATILE ACIDITY	-1.922	3.137	0.000	1.000
CITRIC ACID	-2.874	3.247	0.000	1.000
RESIDUAL SUGAR	-1.127	2.656	0.000	1.000
CHLORIDES	-1.947	5.868	0.000	1.000
FREE SULFUR DIOXIDE	-1.659	3.903	0.000	1.000
TOTAL SULFUR DIOXIDE	-2.806	2.537	0.000	1.000
DENSITY	-2.294	2.335	0.000	1.000
PH	-3.394	2.541	0.000	1.000
SULPHATES	-2.318	3.272	0.000	1.000
ALCOHOL	-2.118	2.070	0.000	1.000

- **Class Imbalance**

There was a significant class imbalance among the dependent variable, ‘quality score’.

The level of measurement of this variable was ordinal and scores ranged from 0 to 10. Upon

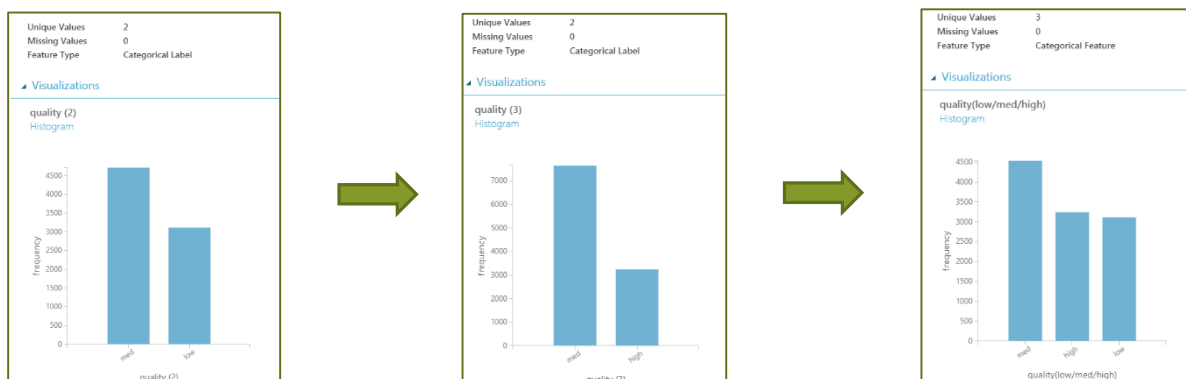


inspection of a histogram, it was noted that a large proportion, about 90%, of the values were reported as 5, 6, and 7. Moreover, some of the lower end values and high-end values were not recorded. For example, there were no scores listed for 0, 1, 2, and 10 (as seen in the figure to the left). Therefore, it would be very difficult to build a model that

correctly predicts the quality score. Due to this disparity of not having as many lower values in the training dataset, the model would not be able to learn the situation when it comes in contact to an output of zero. To alleviate this issue, the dependent variable was transformed into a

variable with three categories titled low, medium, and high. Wines with a quality score of 0 to 4 were reclassified into the low category, wines with a quality score of 5 to 7 were reclassified into the medium category, and lastly wines with a quality score of 8 and above were reclassified into the high category.

After the reclassification of the dependent variable ‘quality score’, SMOTE was used to balance out the cases within the new groups. Because most cases belonged to the ‘medium’ class, the prediction model would have high accuracy by just predicting every wine as a ‘medium’ class. The number of minority classes ‘low’ and ‘high’ were increased. Since SMOTE can only run two classes at a time, it was used twice. At first, the data was divided into two groups (low and the others). SMOTE increased the amount of ‘low’ cases and then the data was saved as a new dataset. Afterwards, the new dataset was imported, the SMOTE procedure was repeated this time separating the data into high and the others which increased the amount of ‘high’ wines, the data was once again saved as a new dataset. The figures below show the changes in the data after running SMOTE.



After running SMOTE, the total number of rows were increased from 4,898 to 10,886. This oversampled the rows including those that had the same features which lead to an increase in duplicates. The proportion of increase was compared, and it was determined it was necessary

to delete the duplicates if the proportion went up significantly. However, the ratio eventually went down from 25% to 20% of the total rows. Therefore, the rows were not deleted.

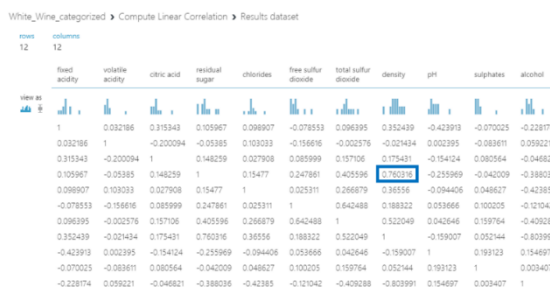
A researcher named Oleg Leyzerov, analyzed the same dataset and recoded the dependent variable, 'quality score', into three categories Low (scores 3-5), Average (score of 6), and High (scores of 7-9). By categorizing the variable in this manner, the class imbalance was addressed, however since the Average class only included the value of 6, the model would have a hard time defining which wine would fall in the Average class if used with real data in the future (Leyzerov, 2017).

- **Feature Selection Methods**

To select the optimal number of features for having accurate and generalized model, three different modules in Azure. The first model created was the Compute Linear Correlation, the second was Filter Based Feature Selection and the third was the Permutation Feature Importance.

Compute Linear Correlation

The Compute Linear Correlation module was used to find features that were highly correlated to each other. Two features that were highly correlated were Density and Residual

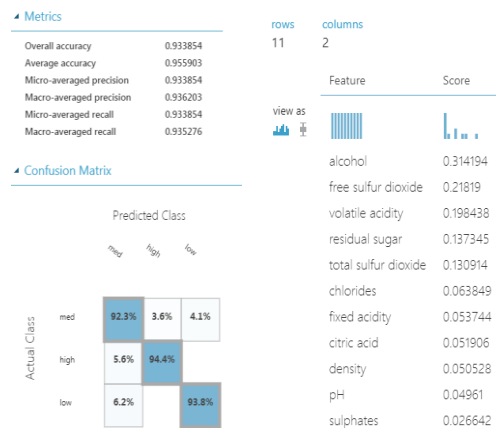


Sugar with value of 0.760. The Principal Component of Analysis (PCA) was used and integrated to remove any redundant features. When the analysis was run with all 11 features, Residual Sugar had a high impact on the results and Density had a lower impact. When the analysis was run with less features, 3 and 6 features, the results indicated something different. Interestingly, the data showed that Density had a high impact on the model and Residual Sugar did not. Due to

this difference this correlation was abandoned. The figure shows the correlation of Density and Residual Sugar.

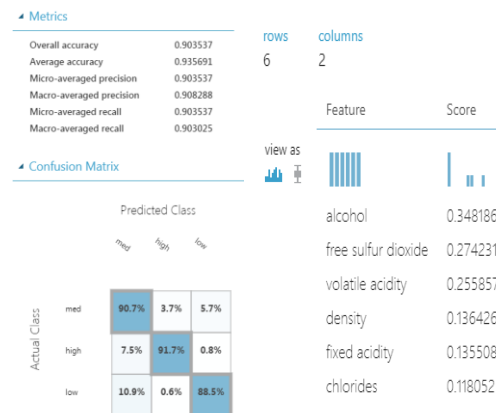
Filter Based Feature Selection & Permutation Feature Importance

To find the optimal number of desired features the model was run multiple times. When



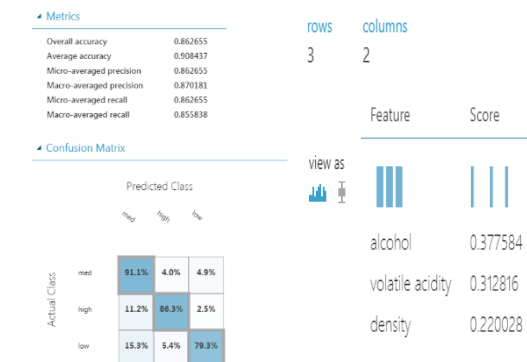
all 11 features were included the overall accuracy was very high, 93.39%. It seemed that using all 11 features seemed excessive because 6 features had a less than 0.1 impact on the model. The concern was having an over fitted model so models with less features were run. The figure shows the model with all 11 features included.

When 6 features were selected, the overall accuracy decreased slightly but was still very



high at 90.35%. Also, every feature had an importance of more than 0.10, which is fairly high. To have a more generalizable model, the model was rerun with less features. The figure on the left shows the output of the model with 6 features selected.

The last model was run with only 3 features. The overall accuracy decreased to 86.27%.



Although the result was still high, the decision was to keep the model with 6 features because the accuracy over 90%. The concern was that three features were not enough to explain the quality score since the sum

of importance of three features was about 0.3 less than the sum of importance of six features which was 1.25. This may indicate that there are unknown features that influence the model significantly. The figure on the left shows the model output with 3 features selected.

- **Model Training**

Regression Model vs Multiclass Classification Model

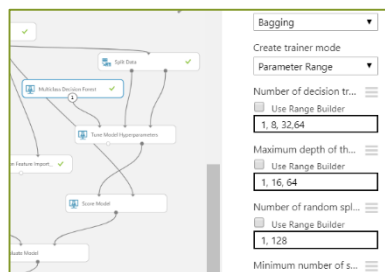
The first model built was a Boosted Decision Tree that predicted the quality score ranging from 0 to 10. Although the coefficient determination of 0.886 was high and the relative squared error was low 0.134, the impact of the missing scores of 0, 1, 2, and 10 and a sizeable class imbalance led us to prefer utilizing categorical methods within the classification model. Other models could have achieved high accuracy rates by estimating every case that belonged in the middle category and had a score of 5, 6, or 7 due in part that 90% of the scores were in this range. However, wines with low and high scores would not be correctly predicted. Although the Boosted Decision Tree is more elaborate because it gives specific quality score of wine, the problems listed above cannot be perceived as accurate to the classification model.

Many models could have obtained high accuracy by just estimating that every case has one of the scores of 5,6 or 7 which takes 90% of the whole cases even though the rest scores would not be correctly predicted. Also, the model would not easily have the score either close to 0 or 10 because those missing numbers cannot be trained in the training set.

Multi-classification Model

After analyzing the Boosted Decision Tree, the next step was to categorize the data and build classification models that predict whether the quality of wine was low, medium, or high.

Not knowing which one would provide the best prediction level, three multi-classification models were built, results were compared to see which model provided the best results. The first model, Multiclass Decision Forest model (as seen in the table below) provided not only the highest accuracy, but also both the highest precision and recall, which indicated that there are less possibilities that the wine may be misled to wrong class. For this reason, the Multiclass Decision Forest was chosen as the prediction model.



To improve the model, the “Tune Model Hyper -parameters” was used and no significant changes were found. The following figures show the Tune Model Hyper parameters and the metrics. The metrics on the left side are from the single parameter and the metrics on the right side are with the parameter range.

Metrics	
Overall accuracy	0.903537
Average accuracy	0.935691
Micro-averaged precision	0.903537
Macro-averaged precision	0.908288
Micro-averaged recall	0.903537
Macro-averaged recall	0.903025



Metrics	
Overall accuracy	0.905834
Average accuracy	0.937222
Micro-averaged precision	0.905834
Macro-averaged precision	0.906528
Micro-averaged recall	0.905834
Macro-averaged recall	0.90936

	Class	Predicted as "med"	Predicted as "high"	Predicted as "low"	Average Log Loss	Precision	Recall
Multiclass Logistic Regression	med	571	185	145	0.80251	0.568725	0.63374
	high	217	401	24	0.855184	0.650974	0.624611
	low	216	30	388	0.812738	0.696589	0.611987
Multiclass Neural Network	med	570	245	86	0.777263	0.649943	0.63263
	high	79	542	21	0.51093	0.663403	0.844237
	low	228	30	376	0.89231	0.778468	0.59306
Multiclass Decision Forest	med	817	33	51	0.474105	0.874732	0.90677
	high	48	589	5	0.504349	0.940895	0.917445
	low	69	4	561	0.452178	0.909238	0.884858

7. Results

Using Permutation Feature Importance module, it became clear what physicochemical properties significantly impacted the prediction model and calculated the average of each feature by quality class to see if there is any noticeable pattern. The results indicated the following, wines receiving a low-quality score had lower alcohol and free sulfur dioxide. As the quality of wine increased, the properties of alcohol and free sulfur dioxide increased as well and in addition properties such as density, fixed acidity, and chlorides needed to be lower. The property of volatile acidity decreased from the lower quality of wines to the medium quality of wines, however there was a slight increase in the property from medium quality of wines to high quality

	low	med	high
alcohol	10.1251	10.4829	11.663322
free sulfur dioxide	26.3773	35.60573	36.616166
volatile acidity	0.37152	0.274308	0.2830563
density	0.99435	0.994087	0.9922253
fixed acidity	7.15124	6.848633	6.7076077
chlorides	0.05121	0.045887	0.038096

of wines. The table to the left shows the average of each of the physicochemical properties as categorized by Low, Medium, and High. consumers but also

the producers. Millennials are the largest generation purchasing wine however they are also price sensitive. This model can provide information to the producers which physicochemical properties benefit the quality of wine. By focusing on the properties wine makers can produce higher quality of wines for less cost.

8. Conclusions

Thus, we were able to discover with high accuracy what wines would be of best quality by analyzing various physicochemical features. We were also able to determine which features were the most important in determining the quality of wine per the results described. It must be emphasized that our data does not take into consideration other subjective factors that may

influence wine purchase such as price, packaging, and the charity of the company. We have created a tool that instead will only look at the objective features to accurately predict a subjective response.

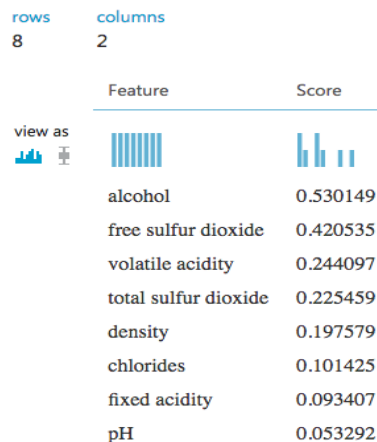
We do hope that a model such as this is utilized by wine producers to study the features of good quality wine and focus less on packaging and other subjective factors that may negatively influence price for the consumer. Currently, the main roadblock would be obtaining the chemical properties of the various wines, as this data is not typically available to the public, and rarely analyzed by wine producers on a regular basis. Only the alcohol percentage is typically shown on wine labels. We hope our suggestion to become more competitive on price would be motivation enough for producers to adopt these procedures and to possibly participate in the future in these types of analytical business models.

Appendix:

- **Regression Model: output score scale 0~10**

1) **Data:** *Refined White Wine.csv*, the dataset was cleaned and SMOTE.

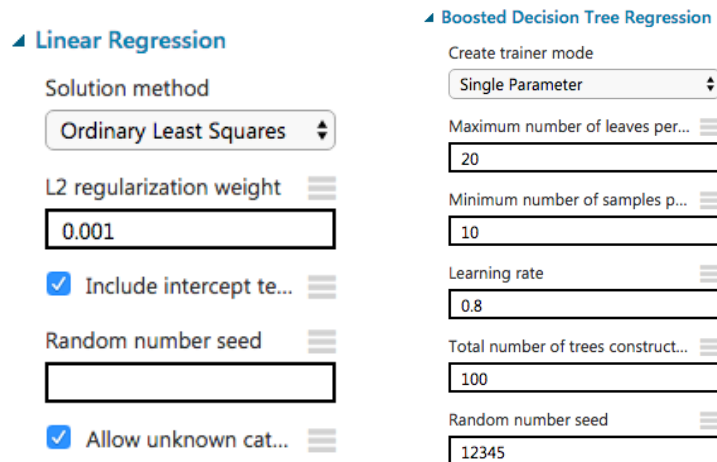
2) **Feature Selection Analysis:** First, we use *Filter Based Feature Selection*, the 8 most informative features toward the output were chosen; Second, we use *Permutation Feature importance*, ranking these 8 features. The rank results as following:



rows	columns
8	2
view as	
Feature	Score
alcohol	0.530149
free sulfur dioxide	0.420535
volatile acidity	0.244097
total sulfur dioxide	0.225459
density	0.197579
chlorides	0.101425
fixed acidity	0.093407
pH	0.053292

We chose top 6 features: *alcohol*, *free sulfur dioxide*, *volatile acidity*, *total sulfur dioxide*, *density*, *chlorides*.

3) **Regression Training Model:** after compared several regression algorithms, we chose *Linear Regression*, Ordinary Least Squares, and weight is 0.001
Boosted Decision Tree Regression, Single Parameter

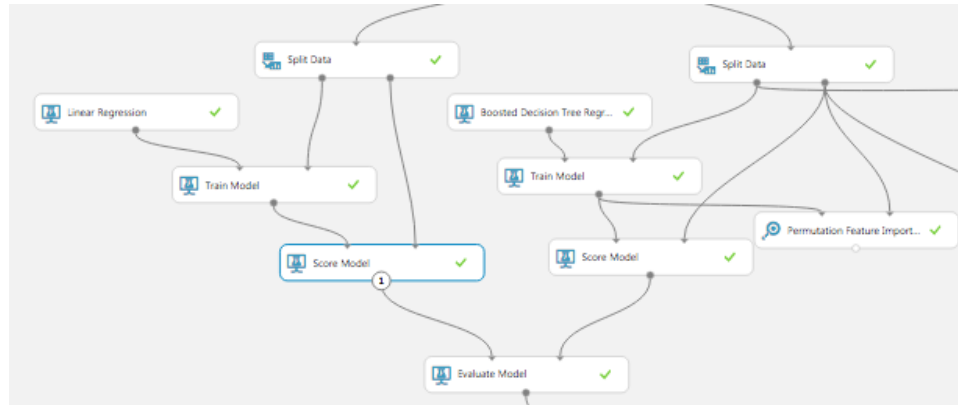


Linear Regression
Solution method
Ordinary Least Squares
L2 regularization weight
0.001
☒ Include intercept te...
Random number seed

☒ Allow unknown cat...

Boosted Decision Tree Regression
Create trainer mode
Single Parameter
Maximum number of leaves per...
20
Minimum number of samples p...
10
Learning rate
0.8
Total number of trees construct...
100
Random number seed
12345

The experiment in Azure as following:



4) Evaluate Models: the results as following

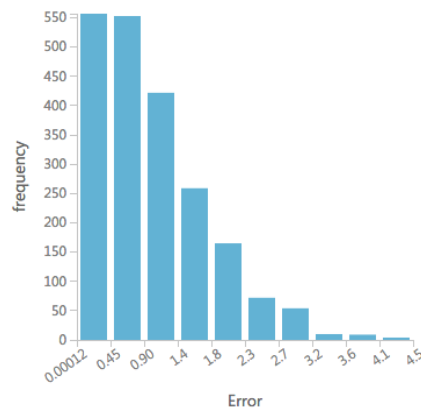
For the liner regression, the MAE is 1.007, Coefficient of Determination is 0.396;

For the Boosted Decision Tree, the MAE is 0.439, Coefficient of Determination is 0.8257;

Metrics

Mean Absolute Error	1.00752
Root Mean Squared Error	1.257994
Relative Absolute Error	0.718398
Relative Squared Error	0.603904
Coefficient of Determination	0.396096

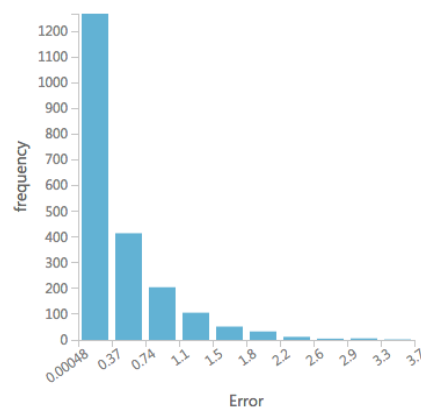
Error Histogram



Metrics

Mean Absolute Error	0.439512
Root Mean Squared Error	0.675794
Relative Absolute Error	0.313388
Relative Squared Error	0.174277
Coefficient of Determination	0.825723

Error Histogram

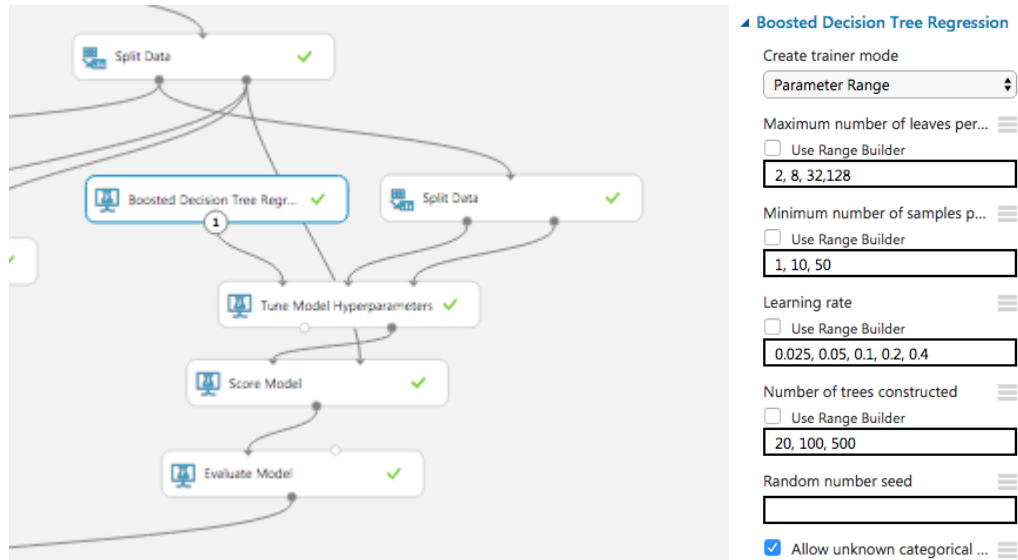


The Boosted Decision Tree Regression has a much better performance than Liner Regression.

5) Model Improvements: *Tune Model Hyperparameters*

Tune Model Hyperparameters set the create trainer mode option to Parameter Range and use the Range Builder to specify a range of values to use in the parameter sweep. It finds optimal model parameters using a parameter sweep and Perform cross-validation during a parameter sweep. [5]

We use Tune Model Hyperparameters with Boosted Decision Tree Regression. The parameters ranger builder setting and as following:

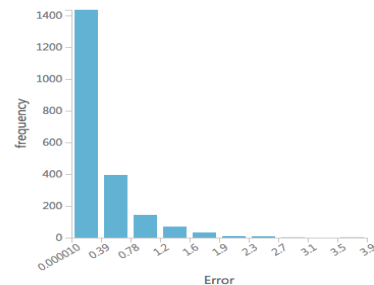


Improvement Results: the MAE is 0.338, Coefficient of Determination is 0.8821;

Metrics

Mean Absolute Error	0.338796
Root Mean Squared Error	0.555652
Relative Absolute Error	0.241574
Relative Squared Error	0.117819
Coefficient of Determination	0.882181

Error Histogram



The Boosted Decision Tree Regression with Tune Model Hyperparameters has a much better performance, which improve the Coefficient of Determination from 0.8257 to 0.8821.

Regression Model Result Conclusion:

liner regression, the MAE is 1.007, Coefficient of Determination is 0.396;

Boosted Decision Tree, the MAE is 0.439, Coefficient of Determination is 0.8257;

Tune Model Hyperparameters, the MAE is 0.338, Coefficient of Determination is 0.8821;

Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
1.00752	1.257994	0.718398	0.603904	0.396096
0.439512	0.675794	0.313388	0.174277	0.825723
0.338796	0.555652	0.241574	0.117819	0.882181

References

- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547-553.
- De Mets, G. , Goos, P. , Hertog, M. , Peeters, C. , Lammertyn, J. and Nicolaï, B. (2017), *Sensory quality of wine: quality assessment by merging ranks of an expert-consumer panel*. Australian Journal of Grape and Wine Research, 23: 318-328. doi:10.1111/ajgw.12287
- Gibbs, M., Tapia, M., & Warzynski, F. (2009). Globalization, superstars, and reputation: Theory & evidence from the wine industry. *Journal of Wine Economics*, 4(1), 46-61.
- Horák, M. (2009). Prediction of wine quality from physicochemical properties. *Alcohol*, 8(5), 10.
- Leyzerov, O. (2016). Exploratory data analysis of the white wine based on physicochemical properties. Retrieved from <https://olegleyz.github.io/>.
- Nachev, A., & Hogan, M. (2013, January). Using data mining techniques to predict product quality from physicochemical data. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
- Przyswa, E. (2014). Protecting Your Wine. *Wines and Vines*, 38-48.
- Wine Institute. (2017). *Wine consumption in the United States*. Retrieved from <https://www.wineinstitute.org/resources/statistics/article86>.
- Wine Market Council. (2017). *2017 Wine market council wine consumer segmentation slide handbook*. Retrieved from http://winemarketcouncil.com/wp-content/uploads/2017/10/2017_WMC_Wine_Consumer_Segmentation_Slide_Handbook_2.pdf.
- Wines of Portugal. (2018). *Portugal wine exports to USA in 2015*. Retrieved from <http://www.winesofportugal.com/us/press-room/statistics/performance-us/>.