Predictive Analysis of Wine Data Providing Solution to a Business Problem

An analytics workflow on predicting wine quality using their physicochemical features

Yao Luo[1]

[1] Farmer School of Business, Miami University, Oxford, OH, USA

E-mail: luoy25@miamioh.edu

Abstract

A business problem was brought up by the sales team of a wine company. The company is looking to cut costs by only bringing in experts for high-quality wine but are lacking the power of predicting the quality of the wine prior to bringing in experts. In this article, a complete analytics workflow will be presented, focusing on Understanding and processing the data, Analytical methods and tools, Solution and result to solve the proposed business problem. A multiple linear regression model will be presented as the solution and several statistical metrics will be used to evaluate the significance of the model and its prediction power. How the model will create gain and reduce pain to the sales team as well as how the model can be implemented and improved will be discussed. The inevitable limitations of the model will also be discussed, and recommendations on future improvements of the model will be given.

Keywords: UCI Wine data [1], Business Applications, Multiple Linear Regression, Stepwise, Evaluation, Limitations, Predictive Modelling, Analytics Workflow, Methods, Reproducibility

1. Background

The wine company's sales team has been finding ways to help increase the profit of sales. One of the suggestions is that they would like to reduce the cost of bringing in experts. They realized that some of the wine ended up with low expert ratings, meaning that these ratings are useless for advertising. This may not seem to be a problem for just one wine, but it gets very costly for a large amount of wine that the company produced. The sales team suggested to not bring in experts for the wine that is not likely to receive low ratings.

The issue for the sales team is that they lack the power in knowing and predicting how each wine will be rated by the expert. They have no way of knowing whether it will be rated high or low prior to getting an expert. However, they have historical data on the expert ratings on both red and white wines they produced as well as the chemical features of the wine. They wonder if it is possible to make predictions of the quality ratings only using the chemical features of the wine.

In this article, the wine data will be analyzed and a solution will be provided to the wine company's sales team targeting their needs. The analytics workflow will be introduced from preprocessing the data to evaluating the solution. Some discussions and the drawback of the solution will also be mentioned in the later section of the article.

2. Data & Pre-processing

The original data [2] given includes red and white wine data each in a separate CSV file. Both datasets have 12 variables including 11 input variables on chemical features of the wine, and one output variable representing wine quality with scores between 0 to 10. Missing data was checked by summing all blank rows in each column for both datasets using tidyverse

library and no missing value was found in both datasets. Therefore no additional imputation was needed.

Since both datasets have the same data structure without any missing values, they were combined into one dataset called *wine* using *rbind* function from the *dplyr* package [3]. A new dummy variable *wine.type* was created representing whether the wine was a red or white wine with white wine coded as the base. Variable *wine.type* was then recoded as a factor based on the categorical nature of the variable. The final dataset has 6497 observations of 13 variables with *wine.type* being a factor while all other variables being numeric.

Data Introduction

rows	6497
columns	13
discrete_columns	1
continuous_columns	12
all_missing_columns	0
total_missing_values	0
complete_rows	6497
total_observations	84461
memory_usage	627904

Figure 1. *The metadata for the finalized wine dataset*

To explore the data more, *DataExplorer* library [4] was used to generate scatter plots, histograms, and bar plots using *plot_scatterplot*, *plot_histogram*, and *plot_bar* functions respectively. No obvious trend or relationships were observed through the scatterplots generated. From the histogram, most numeric variables appeared to be right-skewed and unimodal which makes sense since these chemical variables can not have negative values. Both PH and quality variables appeared to be normally distributed. Total sulfur dioxide seemed to have a bimodal distribution.

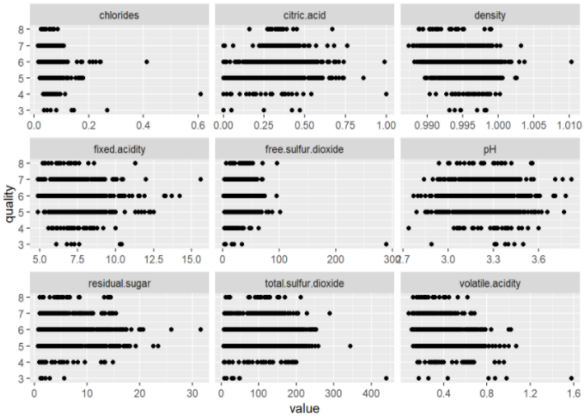

Figure 2. *Scatterplots for input variables against quality (1)*

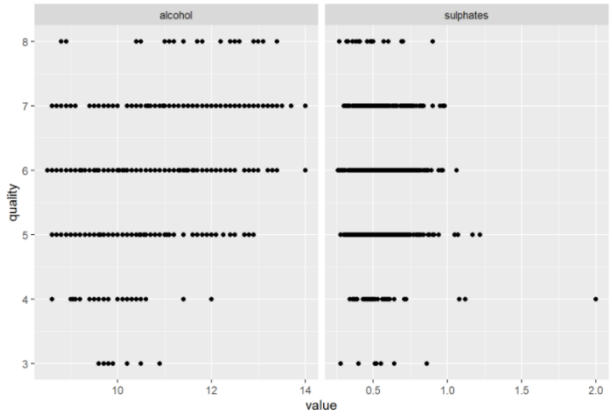

Figure 3. *Scatterplots for input variables against quality (2)*

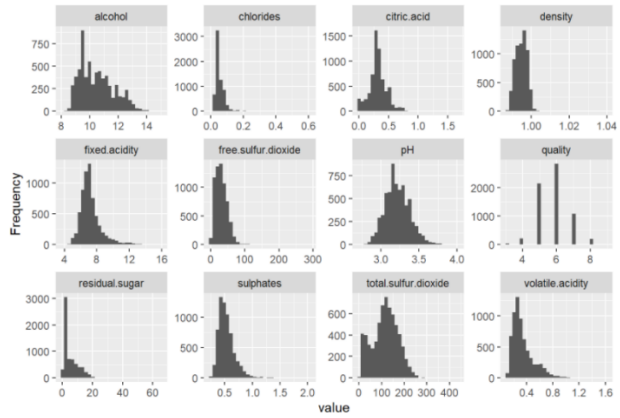

Figure 4. *Histogram for distribution of all numerical variables*

3. Methods

The method used modeling is a multiple linear regression as such models generally have low variance. Before modeling the finalized wine data, a training and validation split was performed so that the prediction capacity can be evaluated later on. The wine dataset was split into 75% training data and 25% testing data setting the seed to be 13. A training index was generated using the random generation and was used for the splitting process. The final training data was stored in a dataframe named *train* with 4873 rows. The final validation data was stored in a dataframe named *valid* with 1624 rows.

To model the finalized data, a regression model was created. Only the *train* data was used for the creation of the model. A full model using all variables was first generated using the multiple linear regression *lm* function. From there, a stepwise model was then derived from the full model using the *stepAIC* function from the *MASS* library [5] with the direction set to both directions and trace set to FALSE since we do not care about the elimination process.

3.1 Reproducibility

For both the data preprocessing steps and model evaluation process, some functions were created and used to enhance analytics workflow performance and for reproducibility. A function called *Metadata* was created with the aim to produce an overview of the data combining the *introduce* function from the *DataExplorer* library as well as *knitr::kable* function from *KableExtra* package [6] to generate a table with desired summary statistics.

Another function named *Evaluation* was created for evaluating model performance where *MLmetrics* library was used. The function inputs the created model as well as predicted y-value and actual y-value. Several evaluation metrics were calculated and reported when the function is called including Adjusted R-Squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) using *adjr2* function, *mse* function, *rmse* function, and *mae* function respectively from the *MLmetrics* library [7].

4. Results & Solution

A statistical model is provided to the sales team with some levels of prediction power on expert ratings using the chemical features of the wine. In this way, the sales team will only need to collect the chemical features of each new wine they produced and run it through the model to get an estimation of the predicted expert rating to help the sales team on deciding whether they are going to actually bring in the expert for that particular wine.

The final stepwise model was a multiple linear regression with 10 inputs and can be expressed by the equation below:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10}$$

where the following are values from β_0 to β_{10} and from x_1 to x_{10}:

(Intercept)	fixed.acidity	volatile.acidity
117.804843278	0.102714855	-1.530384932
residual.sugar	**free.sulfur.dioxide**	**total.sulfur.dioxide**
0.068461810	0.003898134	-0.001260643
density	**pH**	**sulphates**
-117.856650724	0.601265756	0.691825033
alcohol	**wine.type1**	
0.216869536	0.365954267	

Based on the summary below, the overall model appeared to be statistically significant due to the extremely small p-value (less than 0.001) from the F-statistics. All input variables used in the model were statistically significant based on the small p-values from the individual t-tests for each separate variable (All p-values less than 0.001). The model has a Root Mean Squared Error of 0.7342 on 4862 degrees of freedom. The Adjusted R-squared is 0.2977.

```
Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
    free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
    sulphates + alcohol + wine.type, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5440 -0.4690 -0.0480  0.4567  3.0005

Coefficients:
                       Estimate   Std. Error t value             Pr(>|t|)
(Intercept)          117.8048433   17.9777137   6.553  0.00000000006230879
fixed.acidity          0.1027149    0.0189541   5.419  0.00000006278223906
volatile.acidity      -1.5303849    0.0879938 -17.392 < 0.0000000000000002
residual.sugar         0.0684618    0.0072060   9.501 < 0.0000000000000002
free.sulfur.dioxide    0.0038981    0.0008880   4.390  0.00001158575072963
total.sulfur.dioxide  -0.0012606    0.0003756  -3.356              0.000796
density              -117.8566507   18.2523666  -6.457  0.00000000011717236
pH                     0.6012658    0.1065681   5.642  0.00000001775253255
sulphates              0.6918250    0.0878987   7.871  0.00000000000000432
alcohol                0.2168695    0.0229033   9.469 < 0.0000000000000002
wine.type1             0.3659543    0.0675387   5.418  0.00000006302367980
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7342 on 4862 degrees of freedom
Multiple R-squared:  0.2991, Adjusted R-squared:  0.2977
F-statistic: 207.5 on 10 and 4862 DF,  p-value: < 0.00000000000000022
```

Figure 5. *Stepwise regression summary output*

This model will be mainly used as a predictive model inputting wine chemical features to predict its quality score. To ensure the model predictive ability, some evaluation methods were applied. The metrics used are: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and adjusted R-squared. All metrics were calculated from the function created called Evaluation (See Section 3.1 for details). The evaluation was performed on both training and validation datasets to ensure that there is no potential overfitting issue.

When evaluated from the training dataset, an Adjusted R-Squared of 0.2977 was produced, with MSE being 0.5379, RMSE being 0.7334, MAE being 0.5686. When evaluated from the validation dataset, an Adjusted R-Squared of 0.2977 was produced, with MSE being 0.5357, RMSE being 0.7319, MAE being 0.57. When comparing the two sets of metrics, no overfitting issues were found since the metrics results appear to be similar between the training and validation data. The model has the same Adjusted R-Squared for both datasets being 0.2977.

5. Discussions & Business Value

As the sales team's main goal is to estimate the value of wine/expert rating by its chemical feature, the solution will be helpful for them. The model has an adjusted R-squared of 0.2977 meaning that on average 29.77% of the variation in wine quality can be interpreted by the model. Although this may not be a high number, it provides some levels of prediction power for the sales team rather than having no prediction power at all at the beginning. The model can be also used as a descriptive model to help the team understand what features of the wine are more impactful when it comes to getting a better or lower expert rating.

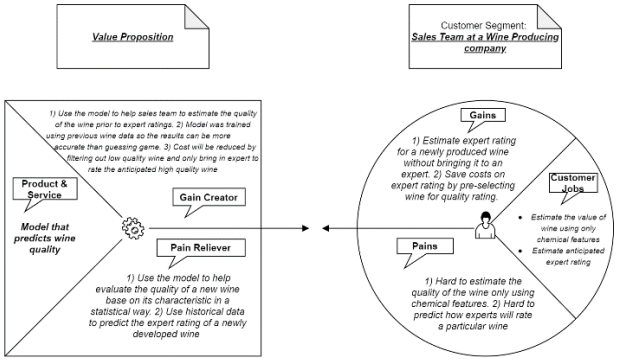

Figure 6. *Business Value Proposition*

The proposed business values were added based on the solution. The sales team's pain of having a hard time predicting how experts would rate a particular wine will be reduced by using the model. In addition, the model was trained using previous wine data so that results are more promising than guessing game. Therefore, the cost may be reduced by filtering out low-quality wine and only bring in experts to rate the anticipated high-quality wine.

With the model provided, the sales team can use it to make predictions for any newly produced wine using its chemical features. This can be scalable and be done in a short period of time for a large amount of wine since the model does not require a significant amount of computing power being a linear regression. Using the model to predict the quality score for all wine produced gives the sales team an estimation of their possible rating and helps them with making a decision on selecting which wine should be rated by experts.

5.1 Limitations & Improvements

Though it is clear that business value was brought to the sales team with the creation of the model, there are still some limitations and rooms for improvements for the current model.

During the data exploration process, plot_correlation functions from the DataExplorer package was used to generate a correlation matrix. From the plot, some relatively strong correlations were noticed:

- Between variable density and variable alcohol (-0.69)
- Between variable free.sulfur.dioxide and variable total.sulfur.dioxide (0.72)

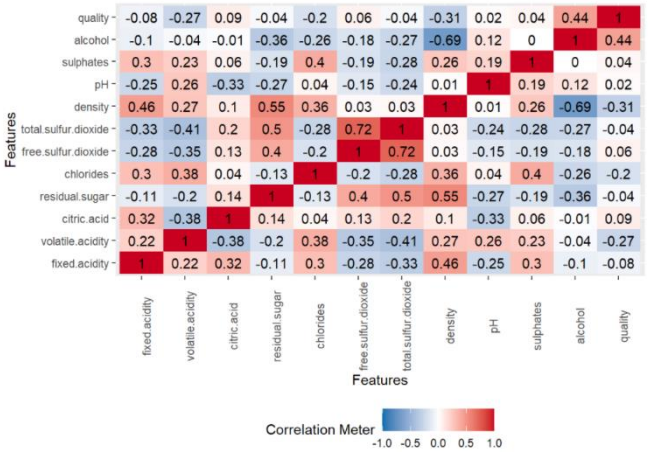

Figure 7. *Correlation Matrix*

These strong correlations may indicate possible multicollinearity which is something to take into consideration. The final model has used the variable that is highly correlated meaning that this could be a possible limitation of the model.

In addition, as it showed from the model evaluation, the Adjusted R-Squared being 29.77% can be relatively low. There may be other models such as random forests that can possibly improve the accuracy and these models can be attempted in the future. However, there is always the trade-off between bias and variance that needs to be taken into consideration. Our current model has relatively low variance meaning that the model would not change much when having a different training dataset. The drawback is that the current stepwise model seems to have a relatively high bias meaning that a larger error exists as we performed a simpler model to approximate a complex problem.

To maintain and improve the model, the sales team can continue to add newly collected wine data to the database each time when an expert was brought in for a rating. The model can be modified and retrained with the updated dataset. Some other future improvements of the model include but are not limited to adding interaction terms and higher-level terms to the existing linear model, conducting a PCA for correlated

variables to reduce dimensions, or try other types of machine learning models.

6. Conclusion

In this article, a business problem was brought up by the sales team of a wine company in need of a way to estimate wine quality before bringing in an expert. A complete analytics workflow was provided on how the data was pre-processed, analyzed modeled to provide the sales team a complete solution. A multiple linear regression model was presented as the solution and its performance was evaluated using different metrics. The model will improve the business processes and can be implemented by the company to help the sales team with their problem. There are some limitations of the model and some ways for the model to be improved were discussed in the article.

Acknowledgment

The datasets used in this article are obtained from the UCI Machine Learning Repository. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine [8].

References

[1] Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
[2] A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal, 2009
[3] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
[4] Boxuan Cui (2020). DataExplorer: Automate Data Exploration and Treatment. R package version 0.8.1. https://CRAN.R-project.org/package=DataExplorer
[5] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
[6] Hao Zhu (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra
[7] Yachen Yan (2016). MLmetrics: Machine Learning Evaluation Metrics. R package version 1.1.1. https://CRAN.R-project.org/package=MLmetrics
[8] Cvrvv. Vinho Verde, www.vinhoverde.pt/en/.