Wine Scoring Analysis

BAIS:9100 Data and Decisions

Professor David Nembhard

Olivia Hawkins, Munju Kam, Deanna Kwon

2020/12/02

1. Executive Summary

Purpose

The purpose of the project was to predict judge rating score of wines produced in Portugal after considering multiple indicator variables using data regression model. The project aims to discover the most and least important indicator variables that contribute to wine judge rating score, under the intent that the following analysis and report will inform potential wine buyers how to choose best quality wine only based on given indicator variables.

Data Preparation

The original dataset was retrieved from “BlogOsVinhos” which contains observations of 2,993 different wines produced in Portugal identified by its name and year ripened (appending B-1, D-1, and E-1). After the cleaning process, total observations were reduced to 2,917 rows and columns were reduced to the following variables: *Name, Produced Region, Produced Year, Color, Alcohol Percentage, Average Price in Dollars,* and finally *Judge Rating* (appending E-1-2). In this analysis project, *Judge Rating* is the only dependent variable that is being predicted after considering the possible indicator variables (appending E-1-3).

Method

The method of this project is divided into 3 major stages: Data cleaning and preparation, Regression Setup Stage, and Regression Modelling Stage. Excel and R were prominently utilized to proceed the following project. Refer to *2. Modeling Steps* page for more details on precise step measures of this project.

Results & Conclusion

As a result of trying various regression models, we have concluded that Model 5: Interaction regression with transformed X variables, is the most fit regression model for predicting wine’s Judge Rating score with given data. The following is the full model with coefficients included representing interaction of different colors with X variables ln(Alcohol Percentage), ln(Average Price in USD) (appending E-7):

JudgeRating = -7299.8+960.51 ln(Year)$_i$ -247.88 color(red)$_i$ -38.235 color(white)$_i$+ 2456.7
ln(AlcoholPercentage)$_i$+ 288.15 ln(AvgPriceUSD)$_i$+ 4.6299 C(Red)*ln(AlcPerc)$_i$+ 0.33403
C(Red)*ln(AvgPriceUSD)+ 30.962 C(Red)*ln(Year)$_i$+2.4922 C(White)*ln(AlcPerc)$_i$+0.16429
C(White)*ln(AvgPriceUSD)+4.1879 C(White)*ln(Year)$_i$-1.5641 ln(AlcPerc)*ln(avgPrice)$_i$-322.59
ln(AlcPerc)*ln(year)$_i$-37.272 ln(avgPrice)*ln(year)$_i$+2902

Note that all color (Rose) categories and its related interaction variables were considered dummy variables in this model, based on the result of regression (appending E-7). Not all x-variables complied to p-value <0.05, which resulted indication of some insignificance in x variables independently explaining the model. However, by looking at factors such as adjusted R^2 equaling to around 0.614 with F Statistics that has a p-value $0 < 0.05$, the model is sufficient to validate the variables’ joint significance when determining the dependent variable, Judge Rating Score. By looking at this regression model, we can conclude that the significance of x-variables by its p-value indicate its importance in predicting Judge Rating Score. The variables align in the following order from greatest to least significant when predicting the y-variable, Judge Rating: ln(Avg Price in US), red*ln(Alcohol Percentage), ln(Year), and red*ln(Avg Price in USD) (appending 2. Modeling Steps page 2-5).

Limitations of the findings retrace to following reasons: 1) Some variables that could have played a major role, such as the judges’ note was not translatable as the primary language used in the dataset was Portuguese, which could be compatible once translated or have used different dataset, 2) There were too many categorical variables which exceeded Excel’s limit of using up to 16 variables maximum, and finally 3) The data seemed to be old and outdated which questions the validity of the dataset.

2. Modeling Steps

2-1. Data Cleaning and Preparation Stage:

2-1-1. Stage 1: Using Excel

First, we started data cleaning process by eliminating any NA/Null/empty value from each column of data set. We also removed any irrelevant columns such as *Judge Note, Castes*, *Producer*, *MinimumPrice, MaximumPrice, Judge*, *Date*, *Label*, and *Link*. Instead of removing *MinimumPrice* and *MaximumPrice* variable, we added *AveragePrice* and convert that variable to *Average Price in Dollars* for easier interpretation (appending E-1-2).

2-1-2. Stage 2: Using R

We started data cleaning process using R by removing any outlier. For example, we assume that the *Year* 1000 and 1780 might be just a typo and simply removed the entire row that contains the data with typo or inaccurate information. The same process was repeated for any other outlier spotted in indicator variable columns, as well as most importantly the dependent variable Judge Rating. Then, categorical variable *Region* was cleaned by picking top 9 most common *Region* that appears in the data set and transformed any other region variable as *Region: Other*. Top 9 *Region* categories were selected to keep the minimum variety of variables possible, since Excel only allow up to 16 X-variables. Categorical variable has been transformed to binary values for both top 9 most common *Region*, *Region: Other* and *Color* (appending E-1-2).

2-2. Regression Set Up Stage:

Cleaned data set includes total of 2,917 rows with 18 columns. Dependent variable in this case is *JudgeRating* and independent variables are *Year*, *Color*, *Alcohol Percentage*, *Average Price In Dollars*, and *Region*.

2-2-1. Correlation Matrix:

Correlation Matrix was initially proceeded to find any multicollinearity and correlation amongst different X variables (appending E-2). Due to the enormous dataset consisting over 2,900 rows, highlight feature from conditional formatting was applied to the entire correlation matrix with assigning red to any value under -0.05 and over 0.05, to indicate existence of positive or negative relationship within each x-variables. According to the correlation matrix, we can observe that *JudgeRating* is highly correlated with *Alcohol Percentage* and *Average price in dollars* followed by *Color:Red* and wine produced from *Region:Espanha.* Additionally, there was an existence of lower correlation with *Region:Esporão S.A*. and *Region:França* suggesting similarity amongst those variables (appending E-2). In this step, direct correlation between X variable and Y variable was positively viewed as a possibility of indicating a relationship.

2-3. Regression Modelling Stage:

Following project was modeled around the sample model below:

$$\text{JudgeRating} = \beta_0+\beta_1\ \text{Year}+\beta_2\ \text{color(red)}+\ \beta_3\ \text{color(white)}_i+\beta_4\ \text{color(rose)}_i+\beta_5\ \text{Alcohol Percentage}_i+\beta_6\ \text{AvgPriceUSD}_i+\beta_7\ \text{Region: Esporao S.A.}_{i+}\beta_8\ \text{Region: DOC Douro}_{\ i}+\beta_9\ \text{Region: DOC Alentejo}_i+\beta_{10}\ \text{Region: Regional Peninsula de Setubal}_i+\beta_{11}\ \text{Region: Espanha}_i+\beta_{12}\ \text{Region: DOC Dao}_i+\beta_{13}\ \text{Region: DOC Vinhos Verdes}_i+\beta_{14}\ \text{Region: Argentina}_i+\beta_{15}\ \text{Region: Franca}_i+\beta_{16}\ \text{Region: Other}_i+\varepsilon_i$$

Here, note that β_0 is Y intercept, β_n are coefficients of multiple X variables, and ε indicate an error term or residuals.

2-3-1. Model 1: Linear Regression

Model 1 assumes the following equation based on linear regression model:

$$\text{JudgeRating} = \beta_0+\beta_1\ \text{Year}+\beta_2\ \text{color(red)}+\ \beta_3\ \text{color(white)}_i+\beta_4\ \text{color(rose)}_i+\beta_5\ \text{Alcohol Percentage}_i+\beta_6\ \text{AvgPriceUSD}_i+\beta_7\ \text{Region:Esporao S.A.}_{i+}\beta_8\ \text{Region:DOC Douro}_{\ i}+\beta_9\ \text{Region:DOC Alentejo}_i+\beta_{10}$$

Region:Regional Peninsula de Setubal$_i$+β_{11} Region:Espanha$_i$+β_{12} Region:DOC Dao$_i$+β_{13} Region:DOC Vinhos Verdes$_i$+β_{14} Region:Argentina$_i$+β_{15} Region:Franca$_i$+β_{16} Region:Other$_i$+ε_i

First model is based on the original dataset without any transformation on variable. When we ran the model with all the variables, we noticed that the model has low adjusted R^2 of 0.2935 (appending E-3). Also, by looking at the normal probability curve plot, we can see the errors do not align in a completely linear shape. Due to all these factors, we decided to try another regression using log transformation in model 2. Some variables such as *Color:Rosé*, *AlcoholPercentage*, *Region:Argentina*, and *Region:França* has #NUM! on the p-value section and *Region:DOC Douro*, *Region:Espanha*, and *Region:DOC Vinhos Verdes* has p-value higher than 0.05.

2-3-2. Model 2: Linear Regression with log-Transformed X Variables

Model 2 assumes the following equation based on linear regression model with log-transformation in X variables only:

JudgeRating = β_0+β_1 ln(Year)$_i$+β_2 color(red)i+β_3 color(white)$_i$+β_4 color(rose)$_i$ +β_5 ln(Alcohol Percentage)$_i$+β_6 ln(AvgPriceUSD)$_i$+β_7 Region:Esporao S.A.$_i$+β_8 Region:DOC Douro$_i$+β_9 Region:DOC Alentejo$_i$+β_{10} Region:Regional Peninsula de Setubal$_i$+β_{11} Region:Espanha$_i$+β_{12} Region:DOC Dao$_i$+β_{13} Region:DOC Vinhos Verdes$_i$+β_{14} Region:Argentina$_i$+β_{15} Region:Franca$_i$+β_{16} Region:Other$_i$+ε_i

Second model is based on the log transformation on X variable. In attempt to enhance the accuracy of model, we have applied the log transformation for Model 2. When we ran the model, we were able to observe that the adjusted R^2 has increased to 0.5967 from 0.2935 (appending E-4). However, most of *Region* variables has turned to p-value higher than 0.05. Also, the variables such as *Color:Rosé*, *AlcoholPercentage*, *Region:Argentina*, and *Region:França* display the #NUM! on p-value column.

2-3-3. Model 3: Linear Regression with log-log Transformed X and Y Variables

Model 3 assumes the following equation based on linear regression model with log-log transformation in both x and y variables:

ln(JudgeRating) = β_0+β_1 ln(Year)$_i$+β_2 color(red)i+β_3 color(white)$_i$+β_4 color(rose)$_i$ +β_5 ln(Alcohol Percentage)$_i$+β_6 ln(AvgPriceUSD)$_i$+β_7 Region:Esporao S.A.$_{i+}$$\beta_8$ Region:DOC Douro $_i$+β_9 Region:DOC Alentejo$_i$+β_{10} Region:Regional Peninsula de Setubal$_i$+β_{11} Region:Espanha$_i$+β_{12} Region:DOC Dao$_i$+β_{13} Region:DOC Vinhos Verdes$_i$+β_{14} Region:Argentina$_i$+β_{15} Region:Franca$_i$+β_{16} Region:Other$_i$+ε_i

Next, we applied the log transformation for both X and Y variable and ran the regression model. The adjusted R^2 has decreased to 0.5811 from 0.5967 suggesting the Model 2 log transformation on X is more effective model for our data set (appending E-5).

2-3-4. Model 4: Interaction Regression with Transformed X-Variables

Model 4 assumes the following equation based on interaction of multiple categories in linear regression model:

JudgeRating = β_0+β_1 ln(Year)$_i$+β_2 color(red)i+β_3 color(white)$_i$+β_4 color(rose)$_i$ +β_5 ln(Alcohol Percentage)$_i$+β_6 ln(AvgPriceUSD)$_i$+β_7 C(Red)*ln(AlcPerc)$_i$+β_8C(Red)*ln(AvgPriceUSD)$_i$+β_9 C(Red)*ln(Year)$_i$+β_{10} C(White)*ln(AlcPerc)$_i$+β_{11} C(White)*ln(AvgPriceUSD)$_i$+β_{12} C(White)*ln(Year)$_i$+β_{13} C(Rose)*ln(AlcPerc)$_i$+β_{14} C(Rose)*ln(AvgPriceUSD)$_i$+β_{15} C(Rose)*ln(Year)$_i$ +ε_i

This model was based of using interaction regression between categorical variables, which in this case was *Color*, on the following log transformed x-variables: *Alcohol Percentage*, *Alcohol Price in Dollar*, and *Year* (appending E-6). According to the result of the regression model, color (Rose) variable indicated itself as dummy variable, as well as its related interaction variables. Here, the following coefficient will be turned to 0 in the result: β_4, β_{13}, β_{14}, β_{15}. Additionally, some interaction variables had p-value higher than 0.05, which indicated its insignificance when predicting Judge Rating Score independently. However, adjusted R^2 of 0.601 and F statistics with p-value $0<0.5$ validated the variables' joint significance in determining the Judge Rating Score.

2-3-5. Model 5: Removing Dummy Variables and Interaction Terms

Model 5 uses the following equation after removing a dummy variable from the previous model:

$$\text{JudgeRating} = \beta_0+\beta_1 \ln(\text{Year})_i+\beta_2 \text{ color(red)i}+\beta_3 \text{ color(white)}_i+\beta_4 \ln(\text{AlcoholPercentage})_i+\beta_5 \ln(\text{AvgPriceUSD})_i+\beta_6 C(\text{Red})*\ln(\text{AlcPerc})_i+\beta_7 C(\text{Red})*\ln(\text{AvgPriceUSD})_i+\beta_8 C(\text{Red})*\ln(\text{Year})_i+\beta_9 C(\text{White})*\ln(\text{AlcPerc})_i+\beta_{10} C(\text{White})*\ln(\text{AvgPriceUSD})_i+\beta_{11} C(\text{White})*\ln(\text{Year})_i+ \beta_{12} \ln(\text{AlcPerc})*\ln(\text{avgPrice})_i+\beta_{13} \ln(\text{AlcPerc})*\ln(\text{year})_i+\beta_{14} \ln(\text{avgPrice})*\ln(\text{year})_i+\varepsilon_i$$

This model was different from Model 4 due to the elimination of the *Color:Rosé* dummy variable and its according interaction terms. This resulted in a model free from errors and the highest adjusted R^2 yet of 0.614. We eliminated this indicator dummy variable last since we wished to make sure there weren't any interaction terms that would be significant.

2-4. Conclusion

According to multiple regression attempt and analysis, highest adjusted R^2 among any other model suggests that the best model for wine scoring data set is Model 5 which is an interaction regression model with transformed X variables. Indicator variables such as *Average Price in Dollars*, *Color:Red* wine with *Alcohol Percentage* seems to have highest impact on *Judge Rating* while *Color:White* by itself does not have high impact on *Judge Rating*.

By looking at the coefficients located per each indicator variables, viewers can easily identify its positive, negative, or no-relationship status with the dependent variable. Positive relationship is identified by positive integer, negative relationship is identified by negative integer or minus sign, and finally no-relationship is identified by close-to-zero integer whether it be positive or negative. Note that up to 5 significant figures were used to make precise measure of the coefficients.

However, current regression model on wine scoring data ran into some limitations: Model 5 regression model did not take consideration of the indicator variables such as *Castes* and *Producer* because there is a limitation on the number of X variables that must be less than 16 variables in Excel. *Judge Note* was not considered for the regression model because it was written in Portuguese which could possibly have impacted on our conclusion.

3. <u>Appendix</u>

B-1 https://osvinhos.blogspot.com/

D-1 https://data.world/loliveira1999/portuguese-wine-dataset-from-blogosvinhos

(E) Excel Book

- E-1. Excel Book, Spreadsheet: Original Data
  - E-1-2. Excel Book, Spreadsheet: Cleaned Data
    - E-1-2-1. Excel Book, Spreadsheet: Transformed Data
  - E-1-3. Excel Book, Spreadsheet: Definitions
- E-2. Excel Book, Spreadsheet: Correlation
- E-3. Excel Book, Spreadsheet: M1. Regression 1
- E-4. Excel Book, Spreadsheet: M2. Regression + Transform (x)
- E-5. Excel Book, Spreadsheet: M3. Regression + Transform (x, y)
- E-6. Excel Book, Spreadsheet: M4. Interaction + Transform (x)
  - E-6-1. Excel Book, Spreadsheet: InteractionData_TransformedX
- E-7. Excel Book, Spreadsheet: M5. Interaction + Transform (x)dum