RED WINE QUALITY PREDICTION USING REGRESSION MODELING

Linear Regression Model
Project 1

Nicole Aguilera

DSC 680-T302 – Summer 2021

Abstract

This case study will look at the components of what makes a good red wine. Each variable is a scientific component of wine. Some of the variables we will be dealing with are alcohol level, residual sugars, quality, sulphates, total sulphur dioxide and volatile acidity. These variables were hypothesized to be a key determinant of wine quality based on each variable's scale. Based on this data set, the main research question is which variables determine the quality of the wine. We would also like to know if any of the variables are correlated, which would continue to further explain the relationship(s) between the independent variables and the quality variable, which is the dependent variable of this data set.

Because the data set was already clean, the project was clear to move on to the analysis portion. The log function of several variables was added into the data set to adjust for the skewed data in the variables. A Pearson correlation visual was generated to visibly see the correlation between the variables in the set. The main method being used for the model for this project is regression, more specifically, linear regression with multiple variables that are being run through it. After selecting to best features, or variables, the data was then split into independent and dependent variable sets and assigned their own variable. There were four linear regressions total that were run on this data set. Of these four, two resulted in a r-squared value of over 99%. This means that there were two sets of independent variables explained over 99% of the linear model, which is a good result of the model. The other two models had results of 0.3466 and 0.3839, which are fairly low, but not far off what was to be expected.

Introduction

Wine is as much science as it is an enjoyable adult beverage. It takes a long time to perfect a good wine, studying the proportions of flavors to put in, how long to age the grapes, etc. Hopefully with this project, wine producers will use the findings as a guide to know the physicochemical components of

a good quality wine, and therefore, be able to charge more if they know that their wine is of better quality. Consumers will know what to look for when searching for a good quality wine, as well.

This project will look at the components of what makes a good red wine. Each variable is a scientific component of wine. The outcomes will determine which variables have the highest impact on wine quality. There are no brands or prices in this dataset. Some of the variables we will be dealing with are alcohol level, residual sugars, quality, sulphates, total sulphur dioxide and volatile acidity. These variables were hypothesized to be a key determinant of wine quality based on each variable's scale. Hopefully, some of these variables will line up in order to tell us which physiochemical properties will make up a good quality wine. No data cleaning will be necessary due to this data set already being clean. All data is numeric. The data set was obtained from Kaggle: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009?select=winequality-red.csv

Exploratory Data Analysis

The 12 variables in this dataset are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH value, sulphates, alcohol content, and a quality score between 0 and 10. Quality is the dependent variable being used for this project. A log transformation was used to adjust for the skewed data in some variables, found by exploring the histograms of each variable. These variables were alcohol, residual sugar, total sulfur dioxide, sulphates, and chlorides.

A feature selection through RFECV, which is Recursive Feature Elimination and Cross-Validation Selection, was run on the variables to see if some were desirable over others for a model. Quality was not included in this process due to the fact that it is the dependent variable for this project and cannot be excluded from the data set. After this evaluation, a data set with the desirable variables was created. This new data set included the citric acid, chlorides, density, pH values, sulphates, alcohol, the log

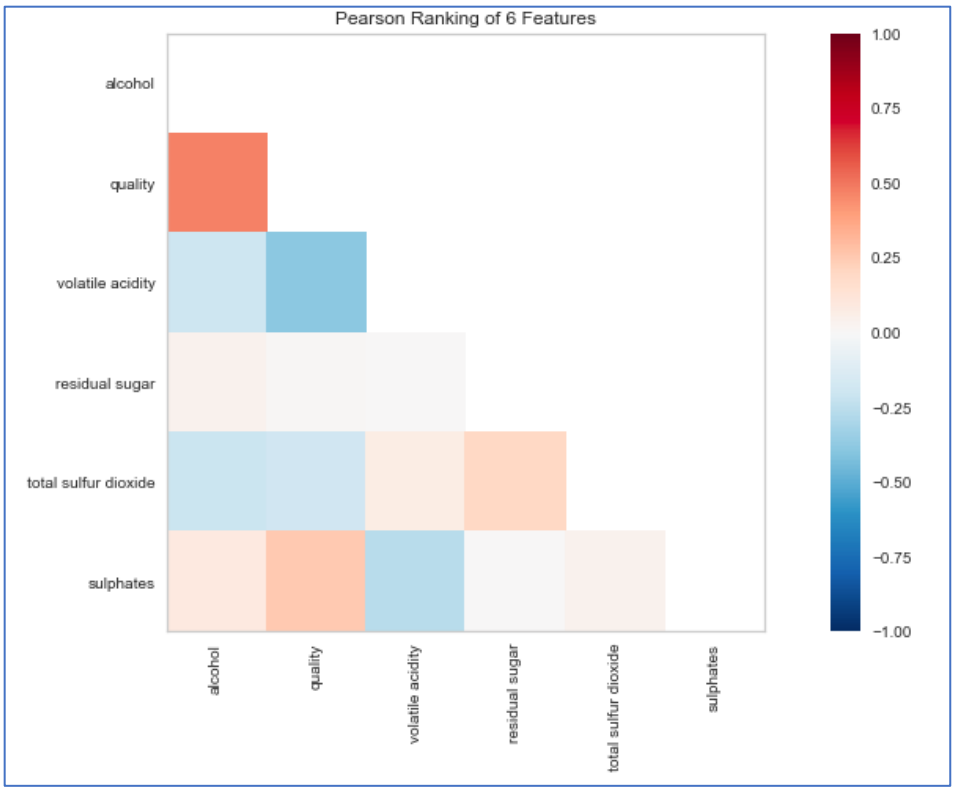
transformation of alcohol, and the log transformation of sulphate variables, and will be used in a linear regression model.

Methods

There are two major methods used for this project, based on the two research questions being asked. They are the Pearson correlation and the linear regression model. One Pearson chart was generated with the data set, and four linear regression models were run. All linear regression models used 'quality' as the dependent variable. The first linear regression model was fit through a least squares model with all of the independent variables. The second regression used the desirable variables chosen by the RFECV feature selection as the independent variables, while the third used all of the variables, including the log transformations of select variables. This is a full linear regression on the entire data set. The independent variables of the final linear regression were made up of the hypothesized variables, which were explained at the beginning of the project to be the most influential on the 'quality' variable.

Results

Both of these models conducted in this project showed to be very effective in producing tangible results. The variables that were the strongest correlated in the Pearson chart are quality and alcohol, which are highly positively correlated, as well as quality and volatile acidity, which are highly negatively correlated. A few other variables are somewhat correlated in the following Pearson chart shown:



The results of the linear regression were somewhat split in half. The first and last regressions run, which were fit through ordinary least squares and the hypothesized variables, respectively, both had high r-squared values that were over 0.99. The value of the first regression was 0.9968, and the value of the last regression run was 1.0. This means that these independent variables explain 99.68% and 100%, respectively, of the data set, which in turn explains why the dependent variable of quality is the values that it is in the set. At this point in the project, it is uncertain whether the r-squared value of 1.0 is a good result or not because it may be caused by too much correlation between the independent and dependent variables.

The second and third linear regressions run were the desirable variables found through feature selection and the collective list of independent variables of the original data set, including the log transformation variables, respectively. These produced lower r-squared values of 0.3468 and 0.3839. This makes somewhat logical sense for the second regression because it was not the entire list of

features from the original data set. The third linear regression run, however, does not make as much sense as to why it has a low r-squared value because it does represent the entire data set.

Conclusion

With the results of the linear regressions run, alcohol level, residual sugars, sulphates, total sulphur dioxide and volatile acidity are all variables that have a significant impact on the quality of a wine. The alcohol and sulphate variables have a positive impact on the wine quality, and the volatile acidity and total sulfur dioxide variables have a negative impact on the quality of wine. The residual level of sugars do not have a noticeable positive or negative impact on the quality of wine, however, it is significant, nonetheless.

References

Red Wine Quality. (2017). Kaggle. https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Questions to be asked from presentation:

1. What business or industry benefits from this project?
    a. The wine industry benefits from this through finding out what components contribute to the quality of the wine. The general public of consumers who purchase wine will also know what to look for in their research to know what makes a good quality wine.
2. Why is this project significant?
    a. Along with the benefits of this project, this project proves to be significant because wine is a very large industry. It has actually grown since before the pandemic of COVID-19 due to people staying home more. This broadens the wine audience and increases the total amount of wine consumption.
3. What are the risks?
    a. There are no risks to this project because the data set is already provided and is not implementing any code into the real world. If more data were to be added to this set, it would be a riskier outcome, unless we were 100% confident in the trustworthiness of the new data. It's a risk because other people could possibly use that information as a baseline for the quality of data, whether they be consumers or producers.
4. Is there a way to drill down the data?
    a. There is not a way to drill down the data past the data set. We are not provided with names of wines or types due to privacy concerns of the original owners of the data.
5. Why is this only one type of wine? How many wines are being analyzed?
    a. We do not know how many types of wine are being analyzed, which is also due to privacy concerns of the data. We do, however, know that this data set only analyzes red wine due to the explanation provided on Kaggle.
6. Would this benefit retailers that sell the wines?

    a. Personally, I do not believe that these conclusions would assist retailers that sell wines, unless they are the wineries themselves or specialty stores. Most retailers are not debriefed on the wine they are given to sell, so they may not be providing the right information to consumers about the wine. Wineries and specialty stores would, by nature, be knowledgeable about the contents of the wines they are selling.

7. Would other data be beneficial in this analysis?
    a. Other data would be beneficial on multiple conditions. It should be fact checked and sourced properly. Data should also be along the same types of red wine that is already in this data set, however, since we do not know this, data of this kind may be hard to find.