Milestone report: Capstone project 1 (Wine Project)

Title: predicting the quality of wine by using physicochemical properties

Problem:

What makes good wine good? I always have this question in my mind when having a glass with friends. Normally, I select my favorite bottle from beautiful label, deepness of the bottom and rating website in the internet. However, it would be much more legit and informative to investigate this problem through physicochemical properties which are the fundamental of flavor and taste of wine. After eliminating all the psychological biases such as price, popularity, year, branding, packaging etc., we can ask something like what physicochemical properties will affect the quality of wine? Is there any magic mixture of acid and sugar that makes wine guru give a thump up?

Client:

By digging deeper, it can be beneficial to both supply side and demand side. For supply side, this is obvious that the wine makers want to brew the best wine for market. Getting known more about physicochemical properties will give them more resource to achieve their goal. In demand side, wine customers will learn more about physicochemical properties and how they affect wine quality. They will have more scientific indicators for selecting their own favorite bottle. In this particular research, we will focus more on supply side because we won’t see it every day that some random guys go to the liquor store with testing equipment to measure residual sugar or pH level of wine before buying them. On the other hand, the one who is fully equipped with all the expensive equipment is the brew house. By making a better wine or higher rating, the brew house will earn higher profit. Therefore, in this sense of getting physicochemical properties, we will target at the supply side. How do we improve the rating? What should be added in wine to make it more appearing to the tester?

Data set:

The data set is from Kaggle website. This datasets is related to red variants of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). Input variables (based on physicochemical tests): 1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 - residual sugar 5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide 8 - density 9 - pH 10 - sulphates 11 - alcohol Output variable (based on sensory data): 12 - quality (score between 0 and 10). For more information, check

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, 2009.

Relevant publication

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, 2009.

Cleaning Process:

First, we checked the missing value in the data first to make sure that there is no anything missing. After checking, our data looks complete with legit type of information (float 64) which is the type of storing number in python. However, there is some duplicated row in our data. For purpose of convenience and meaningful result, we decide to delete the duplicated row since the wine with exactly the same physicochemical properties has a higher chance to come from the same origin or even the same tank. In this study, we want to study more about how physicochemical properties affect wine quality, so there is no reason to keep the duplicated row in our data. Then, we looked for outlier by checking basic statistic parameters such as mean, standard deviation, max, min, etc. The result looks legit. Therefore, we are ready to use this data set in the further process.

Potential data:

For the similar data set in Kaggle, there are some such as <https://www.kaggle.com/akram24/wine-pca/data>

<https://www.kaggle.com/zynicide/wine-reviews>

<https://www.kaggle.com/residentmario/most-common-wine-scores/data>

<https://www.kaggle.com/danielpanizzo/wine-quality>.

As we can see from the data, we have some logistic issue with our data. There is no label nor name nor year of the wine in our data set. . It would be much more interesting, if we can investigate more in the relationship between physicochemical properties and price or year, so we might find answer to some myths in winery industry.

Initial finding:

First, the psychochemical properties of wine can be very extreme. For example, total sulfur dioxide level of 270 is still possible; even though the level is almost triple the mean of the feature. However, they are still acceptable according to the meaning of parameters. Second, by virtually checking, most features are not normally distributed. We investigate this by plotting histogram. Next, after trying to apply some EDA process to our data, these are the interesting findings. Variables that have correlation more than 0.15 are volatile acidity, citric acid, total sulfur dioxide, density, sulphates, and alcohol. For convenient, we will categorize these features to have a direct effect to wine quality. However, for other variables, the correlation between them and wine quality is weak, but they have a relationship with the direct effect features. For example, pH is strongly related with acid properties (0.67, 0.25,-0.55). Also, density is strongly related with alcohol (-0.5). Hence, we will categorize our features into two categories direct features and indirect features. Then, we investigate more in the behavior of direct features and indirect features for each rating of wine. After plotting scatter plot, for each quality level, the behavior of feature is the same among different features. For example, the slope of correlation between fixed acidity and pH is negative no matter what quality level it is. The same behavior happens in other scatter plot that we illustrated too. Therefore, to make it easier for applying logistic regression to our data, we decide to create other variable called rating. The wine quality will be split into two categories (0 and 1). Zero parameter represents lower quality of wine range from 3 to 5. One parameter represents 6 to 8 quality rating. By doing this, we will equally separate our data into above average and below average category (aka good wine and bad wine). This separation will bring meaning to our result at the end. If the wine is classified as good wine, it means that the wine is at least above average. On the other hand, the wine that is classified as bad wine is below average. Also, we can apply logistic regression directly to the data. We will talk about the reason of using logistic regression in the next procedure. Next, we run the same EDA process to our new variable to see whether there is any change in direction of relationship or not. The result is that the degree of relationship is somehow weaker than using original rating variable, but the direction of relationship remains the same. Then, we try to apply some inferential statistical technics to test the difference in mean of each feature whether there is any statistical difference between two types of wine. Our test will be consisting of two steps such as plotting ECDF of each group, using bootstrapping method. As we can see from the results, all tests that we conducted by using bootstrapping method has significant result meaning that the mean of each feature of each group is statistically significant. However, we want to dig deeper on every feature. By doing that, we use t-test the distribution of each feature is unknown. Moreover, it will be much faster and easier to use t-test to test on a lot of feature. From the results, only residual sugar and pH fail to reject the test which means that the difference in mean of those features is not statistically significant. To sum up, after using ECDF, bootstrapping, and t-test, the difference of mean of features between good wine and bad wine is statistically significant. Only two variables (residual sugar and pH) is not statistically significant.