Machine Learning Project Proposal

Sooihk Ro, Richard Shim, Jing Wang

Predicting the Quality of Red Wine

August 09, 2021

1. Problem Statement

Knowing which quality wine to pick for parties, dates or as gifts is imperative for setting a good impression in this day and age as social drinking is becoming more and more prominent. Therefore, our group decided to apply machine learning models to discover what makes wines good or bad.

With the wine market having a sale of over 39 billion dollars a year in the United States alone, many wine producers would also be interested in utilizing machine learning methods to discover how to improve the taste of their wines in such a competitive and saturated market. Knowing that the wine's chemical properties can be correlated with better tasting wines gives quality control and insights to improve their production.

We will be using machine learning methods to predict wine quality based on physicochemical measurements based on red wine samples from the north of Portugal.

2. Data Source and Prep

The detailed description and the dataset itself can be found under the following URL: https://archive.ics.uci.edu/ml/datasets/Wine+Quality. Out of two datasets available in this URL (each for red and white wines), we will be only using the wine quality dataset on red wine.

The wine quality dataset on red wine includes 12 independent variables: only physicochemical (inputs) and sensory (the output) variables are available due to privacy and logistic issues (e.g. there is no data about grape types, wine brands, prices, etc.)

Below is an example of the first 10 rows of the dataset.

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5

This dataset doesn’t require data cleaning and has no missing values nor the need for dummy variables. Only cleaning required is to convert semicolons separating values to commas (into csv file) so that it is easier to import data.

We expect the quality of wine to be mostly dependent on half of the variables such as alcohol percentage, pH, chlorides, residual sugar, and acidity. It might be interesting to create dot plots to see the general trend on how each variable is correlated to quality of wine.

3. Modeling and Assessment Strategy

The output of the model can be interpreted as a quantitative value between 0 to 10. A regression model will be used to quantify the nature of the relationship between the output (wine quality) and the input variables (physicochemical tests).

Start with multilinear regression with all input variables and then remove unwanted parameters that don’t heavily contribute to quality score. With the transformed dataset containing only wanted variables, we will perform multilinear regression, stepwise regression, partial least squares LASSO regression, Ridge regression, and Random Forest regression. The dataset will be split into training and test sets by a 70:30 ratio.