Wine Quality Analysis Midterm Report

Yu Jin (yj465) Jinyuan Yu (jy478) Heru Wang (hw743)

1. Dataset

The dataset used by the team describes the sensory quality of white variants of the Portugal 'Vino Verde' wine from May 2004 to Feb 2007, which was published by P. Cortez as part of his research in decision support systems.

- **Sample Size**: The dataset has 4898 entries in total.
- **Output**: The response is the quality of wine based on sensory data (median of at least 3 evaluations made by wine experts using blinding test). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent) with an integer.
- **Input**: The features are 11 physicochemical variables that could be used to determine wine preferences. We have decided to include all the physicochemical variables in our preliminary analysis.

The overall quality of the dataset proves to be great, with only a few missing data on some of the entries as is shown in the table below.

Table 1. Attributes Description and Missing Values

Attribute	Count	Description
Fixed Acidity	4890	Most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
Volatile Acidity	4891	The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
Citric Acid	4896	Found in small quantities, citric acid can add 'freshness' and flavor to wines
Residual Sugar	4896	The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
Chlorides	4896	The amount of salt in the wine
Free Sulfur Dioxide	4898	The free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
Total Sulfur Dioxide	4898	The amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
Density	4898	The density of water is close to that of water depending on the percent alcohol and sugar content
pH	4891	Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
Sulphates	4896	A wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
Alcohol	4898	The percent alcohol content of the wine

2. Exploratory Data Analysis

The variables presented are solely real variables. To obtain a general idea on which physicochemical characteristics are relevant on white wine's sensory, we adopted few methods to generally detect the distribution of values of different physicochemical variants respectively. Here, we present histograms of three typical variables: density, pH and alcohol, along with the output variable 'quality' which is an ordinal variable that ranges from 1 to 10.

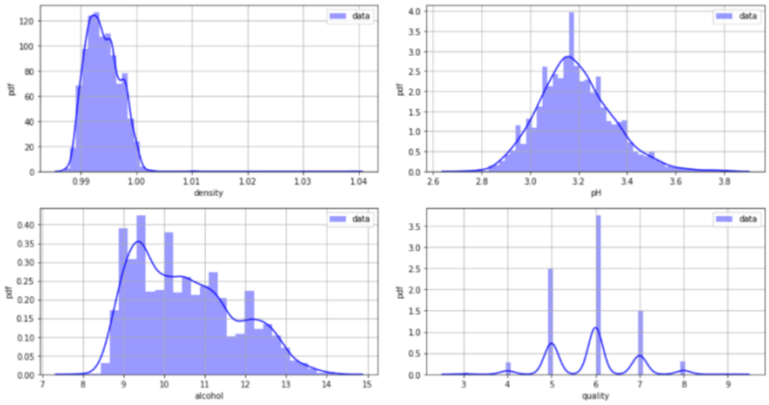

Figure 1. Data Distribution of Features

Based on past endeavors of researchers in wine industry, we suggest that some of the variables contribute more to the sensory than the others. For example, physicochemical characteristics such as density and pH are certain to have a more determinant effect on the overall flavor of white wine, whereas the level of alcohol has been proven to have less influence on sensory perception.

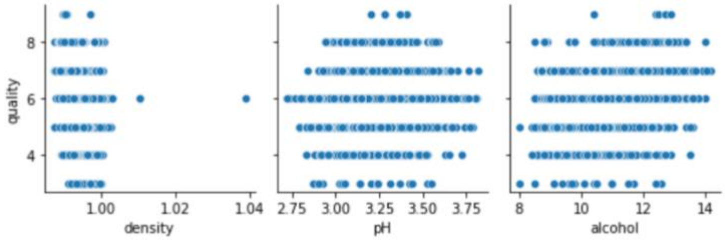

Figure 2. Scatter Plot of Quality Scores and Features

3. Methodology

Since the dependent variables are all ordinal integers in this problem, we adopt the regression approach instead of classification to preserve the order of the wine preference. In our preliminary analysis, we have selected several kinds of regression models and train each of them separately.

Table 2. Models for Learning

Family	Model	Description
Linear	Least Squares	Standard linear model (with least-squares estimators)
	Polynomial	Linear model with 2nd degree polynomials of features
	Lasso	Linear model with L^1-norm penalty function
	Ridge	Linear model with L^2-norm penalty function
Trees	Decision Tree	CART
	Random Forest	Bootstrap samples and select features at random to grow a group of trees for prediction
Others	K-Nearest Neighbors	Mean of the responses of k nearest neighbors in the train set

Our objective is to fit the model using train-set (in-sample) data and test the model’s performance with test-set (out-of-sample) data. The error metric we chose for evaluating the models was Mean Squared Error (MSE). Details of the methodology are as follow:

- **Feature Engineering:** We discarded all missing values in our dataset (since there are only a few of them). For models like Lasso, Ridge and K-Nearest Neighbors, we standardized the variables. For other linear models, we added an offset to the features. For Polynomial model, we added 2nd degree polynomials of features in regression.
- **Hyperparameter Tuning:** We applied the 5-folds cross validation technique to tune the hyperparameters in the train set with grid search. For regularized linear regression models, hyperparameters are the regularization coefficients. For tree models, we needed to tune the maximum of depth for each tree and the number of trees (in random forests). For KNN, we chose the optimal "K" for our model.
- **Overfitting Avoidance:** We attempted to prevent overfitting in following ways: (a) Separated the train set and the test set for generalization. (b) Applied cross validation to tune the hyperparameters. (c) Introduced regularization techniques like penalty functions in linear models. (d) Controlled the maximum depth for tree models.

4. Preliminary Analysis

The error metrics of different models in the test set is shown in the following figure.

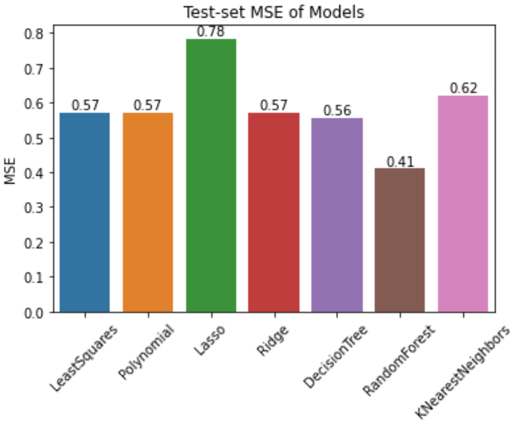

Figure 3. Test-set MSE of Models

For linear models, adding polynomials of features didn't increase too much test error. For regularized regression, L^1-norm penalty function seemed to perform worse, while ridge regression worked well. Two tree models outperformed the linear models. Random forests method has the lowest test error among all the models, which indicated that ensemble methods may be effective in our problem. Similarity-based methods like K-NN didn't work well compared with models above.

5. Further Steps

- We are going to focus on the application of these technologies in the industry background. We will interpret the models and analyze the feature importance, so that the methodology can help to provide expert advice for the wine industry.
- For the rest of the semester, we also plan on adding more advanced models for solving the prediction problem, introducing other feature engineering and regularization techniques, and evaluating the models with additional error metrics.