Partial Dependence - Wine Quality Dataset

Partial dependence (PD) plots illustrate the relationships between one or more input variables and the predictions of a black-box model.  They are both visual, model-agnostic techniques. For example, a PD plot can show whether estimated car price increases linearly with horsepower or whether the relationship is another type, such as a step function, curvilinear, and so on.

The dataset here Is Red Wine Quality dataset from the UCI Machine Learning repository

The task is to predict the quality of the wine using classification or regression

The dataset has the following features.

fixed acidity

most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

volatile acidity

the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

citric acid

found in small quantities, citric acid can add 'freshness' and flavor to wines

residual sugar

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

the amount of salt in the wine

free sulfur dioxide

the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

total sulfur dioxide

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

the density of water is close to that of water depending on the percent alcohol and sugar content

pH

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

a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

alcohol

the percent alcohol content of the wine

quality

output variable (based on sensory data, score between 0 and 10)

fixed acidity

most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

volatile acidity

the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

citric acid

found in small quantities, citric acid can add 'freshness' and flavor to wines

residual sugar

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

the amount of salt in the wine

free sulfur dioxide

the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

total sulfur dioxide

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

the density of water is close to that of water depending on the percent alcohol and sugar content

pH

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

a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

alcohol

the percent alcohol content of the wine

quality

output variable (based on sensory data, score between 0 and 10)

Here we aim to study the Partial Dependence of variance machine learning models using the wine quality dataset.

Following machine learning models have been used to study the partial dependence.

* Random Forest Classifier
* Gradient Boosting Classifier
* Decision Tree Classifier

1. Random Forest Classifier

I tried a simple Random Forest Classifier on the Wine Quality dataset

Following was the classification report

A screenshot of a cell phone

Description automatically generated

Partial Dependence for label 3

A screenshot of a cell phone

Description automatically generated

For the label 3 the partial dependence of Random Forest is almost zero. This is because the Random Forest Model is not able to recognize the class with label 3 dues to very a smaller number of instances of class 3.

Partial Dependence for label 5

A close up of a map

Description automatically generated

The partial dependence plot shows some good results for label 5.

The label 5 has the maximum number of instances in the dataset.

We observe that for the following variable the prediction probability decreases as there value increases.

* Alcohol
* Sulphates
* pH
* Residual Sugar

It increases for the following variables

* Total Sulfur dioxide
* Volatile Acidity
* Density

It remains almost constant for

* Fixed acidity
* Free sulfur dioxide
* Chlorides
* Citric Acid

This indicated that variable like fixed acidity, free sulfur dioxide, chlorides and citric acid don’t have a major effect on the model

Partial Dependence for label 6

A close up of a map

Description automatically generated

Here we see the reverse trend as that of label 5

We observe that for the following variable the prediction probability decreases as there value increases.

* Total Sulfur dioxide
* Volatile Acidity
* Density
* Citric acid

It increases for the following variables

* Alcohol
* Sulphates
* Free Sulphur dioxide
* pH

It remains almost constant for

* Fixed acidity
* Residual Sugar
* Chlorides

Here again the more variables like fixed acidity, residual sugar and chlorides seem of less importance to the model.

Partial Dependence of class 7

A close up of a map

Description automatically generated

Here we see a pattern similar to label 5

The variable that were had increasing probability of prediction for label 5 have also have increasing probability of prediction for label 7.

The probability increases for alcohol, sulphates, and fixed acidity and decreases for total sulfur dioxide, volatile acidity and density

It almost remains the same for pH, density, chlorides, sulfur dioxide, citric acid and residual sugar.

In conclusion we can say that the most important vairables for Random Forest model are

* Alcohol
* Sulphates
* Volatile Acidity
* Total Sulfur dioxide and
* Fixed Acidity
* Density

The same can be observed from the plot below which was trained on a Random Forest Model with the subset of above features.

A close up of a map

Description automatically generated

Gradient Boosting Classifer

I also trained a gradient boosting classifier on the dataset to observe its partial dependence on the features.

A close up of a map

Description automatically generated

This is for label 5

We observe that for the following variable the prediction probability decreases as there value increases.

* Alcohol
* Sulphates
* Residual Sugar
* pH
* fixed acidity

It increases for the following variables

* Total Sulfur dioxide
* Volatile Acidity
* Density
* Chlorides

It remains almost constant for

* Free sulfur dioxide
* Citric Acid

Partial Dependence for label 6

A close up of a map

Description automatically generated

We see a opposite pattern here than that of label 5

The probability increases for alcohol, residual sugar, total Sulphur dioxide and density.

But decreases for volatile acidity, fixed acidity, chlorides and sulphates

These are the patterns similar to the Random Forest Model.

Binning

I used the ranking function to get the rank of variables according to variance for 3 bins

Here I have used the same Random Forest Model as before.

Following was the ordering.

The order of columns for 3 bins according to variance in the plots

free sulfur dioxide

citric acid

pH

residual sugar

volatile acidity

fixed acidity

density

sulphates

chlorides

alcohol

total sulfur dioxide

Binning on citric acid

A close up of a map

Description automatically generated

We see much variance in the plots when citric acid is binned

We observe that for the following variable the prediction probability decreases as there value increases.

* Alcohol
* Sulphates
* Fixed Acidity
* Residual Sugar

It increases for the following variables

* Total Sulfur dioxide
* Volatile Acidity
* Density

It remains almost constant for

* Free sulfur dioxide
* Residual Sugar
* Chlorides
* pH

I also did binning on total sulfur dioxide… where I don’t see a specific pattern

A close up of a piece of paper

Description automatically generated