

New

Workspace

Recents

Catalog

Workflows

Compute

Marketplace

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Pipelines

AI/ML

Playground

Wine Quality

Python

Tabs: OFF

☆

File

Edit

View

Run

Help

Last edit was 16 minutes ago

Run all

Connected

Schedule

Share

12:55 PM (<1s)

Python

print(feature_importances.sum())

	importance
alcohol	0.160192
density	0.117415
volatile_acidity	0.093136
chlorides	0.086618
residual_sugar	0.082544
free_sulfur_dioxide	0.080473
pH	0.080212
total_sulfur_dioxide	0.077798
sulphates	0.075780
citric_acid	0.071857
fixed_acidity	0.071841
is_red	0.002134
importance	1.0
dtype:	float64

The two most important features are alcohol and density. The least important feature is whether or not it is red.

Intuitive sense? Maybe it makes intuitive sense. Density could potentially be a defining characteristic of quality, given mouthfeel.

Feature importance is calculated by the percentage of branches of a decision tree traversed by any one of the features. The feature importance doesn't make immediate intuitive sense. Maybe the density feature is the most likely to be important, given the effect of mouthfeel on taste. A curious finding: sulphates was shown to be the top feature in (Cortez et. al 2009), contrasted to the top two features predicted in the random_forest_model being alcohol and density.

New

Workspace

Recents

Catalog

Workflows

Compute

Marketplace

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Pipelines

AI/ML

Playground

Wine Quality

Python

Tabs: OFF

☆

File

Edit

View

Run

Help

Last edit was 17 minutes ago

Run all

Connected

Schedule

Share

Uploading artifacts: 100%

11/11 [00:00<00:00, 17.34it/s]

alcohol	-0.416504
total_sulfur_dioxide	-0.110571
free_sulfur_dioxide	0.481284
quality	0.680737
pH	0.754402
residual_sugar	0.930913
citric_acid	1.135369
volatile_acidity	1.477522
fixed_acidity	1.965486
chlorides	2.007598
is_red	2.030779
sulphates	2.296992
density	2.499226
dtype:	float64

+ Code

+ Text

Assistant

A lengthwise sum of the correlation matrix shows alcohol and density being the most correlated features to other features, alcohol being negatively correlated and density being positively correlated. The sum of correlations displays a sense about the data that intuition doesn't alone. The data suggests that low alcohol content and a higher density are the most important features for a highly rated wine. The bitterness of alcohol in combination with its astringent affect is probably the most important factor; the taste of the grapes and process is less overpowered. With this information, it does make intuitive sense.

Curiously, the feature importance calculated in the (Cortez et al., 2009) dataset differs from the feature importance calculated by the lab model.

[1] The example uses a dataset from the UCI Machine Learning Repository, presented in Modeling wine preferences by data mining from physicochemical properties [Cortez et al., 2009].