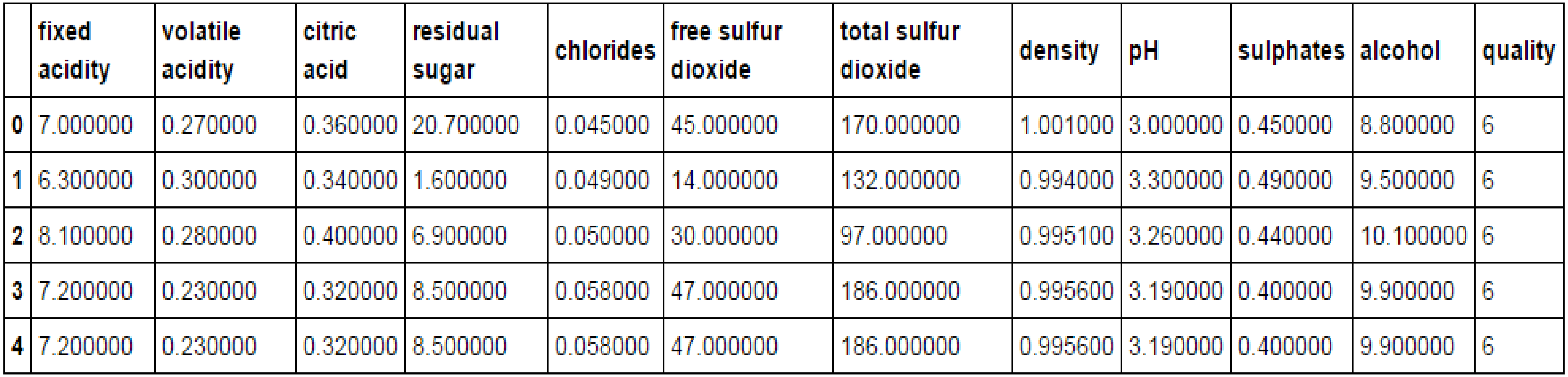
In vino veritas

GA Data Science Class
Final Project
Wen Lu
Mar 16 2016

About the Data

- Two datasets – Red & White Wine
- Source: UCI Machine Learning Repository
- 11 physiochemical attributes
- No missing values
- In different units of measurement



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.000000	0.270000	0.360000	20.700000	0.045000	45.000000	170.000000	1.001000	3.000000	0.450000	8.800000	6
1	6.300000	0.300000	0.340000	1.600000	0.049000	14.000000	132.000000	0.994000	3.300000	0.490000	9.500000	6
2	8.100000	0.280000	0.400000	6.900000	0.050000	30.000000	97.000000	0.995100	3.260000	0.440000	10.100000	6
3	7.200000	0.230000	0.320000	8.500000	0.058000	47.000000	186.000000	0.995600	3.190000	0.400000	9.900000	6
4	7.200000	0.230000	0.320000	8.500000	0.058000	47.000000	186.000000	0.995600	3.190000	0.400000	9.900000	6

Looking into the features

Findings

Outliers

Expected levels for some of the physiochemical attributes

1,500 - 14,500 mg/L tartaric acid; 0 - 500 mg/L citric acid; 0 - 3 g/L volatile acid; 10 - 350 mg/L sulphates;

Overlapping features

"The predominant fixed acids found in wines are tartaric, malic, citric, and succinic.“

total SO_2 = free SO_2 + bound SO_2

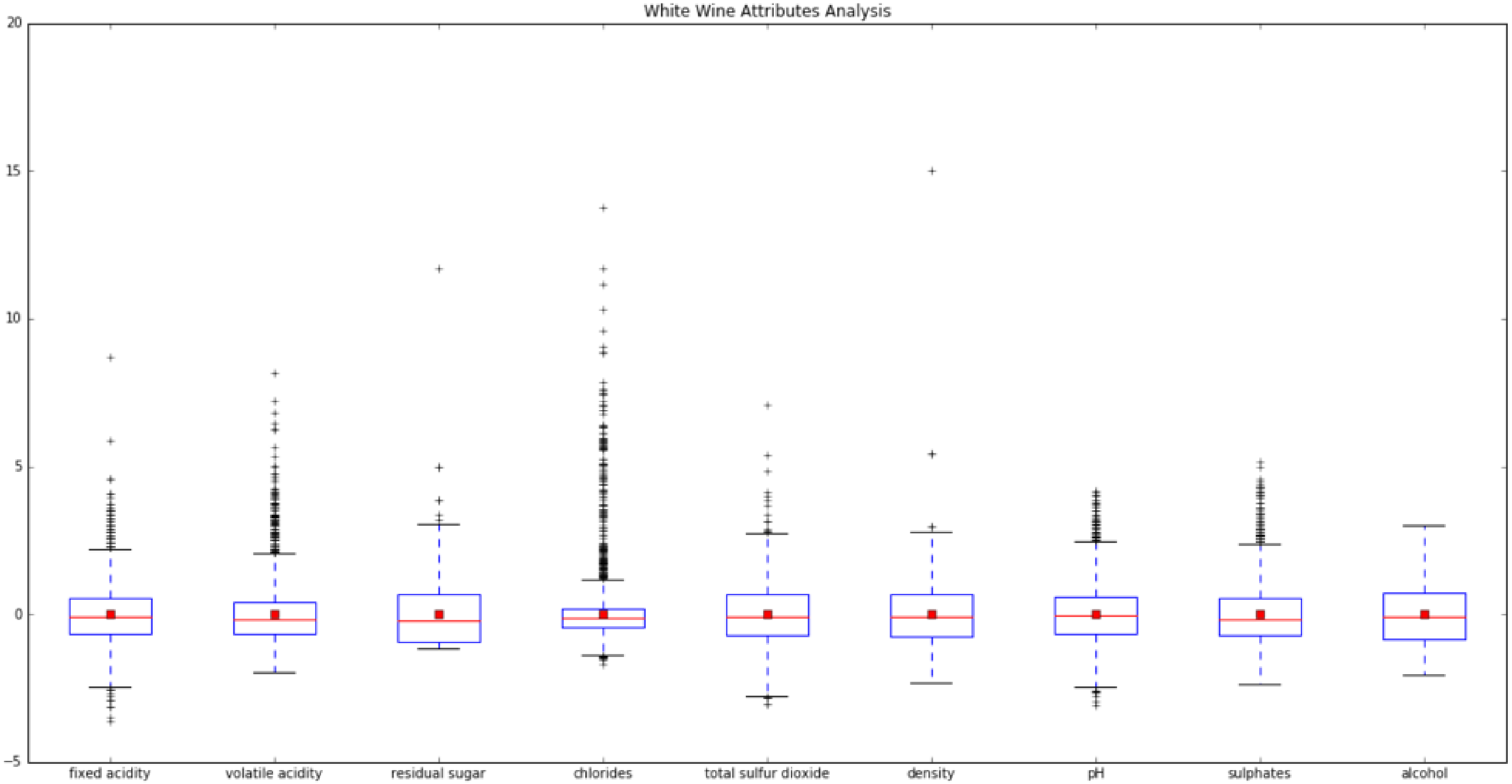
What I did

Drop 'citric acid‘ and 'free sulfur dioxide'

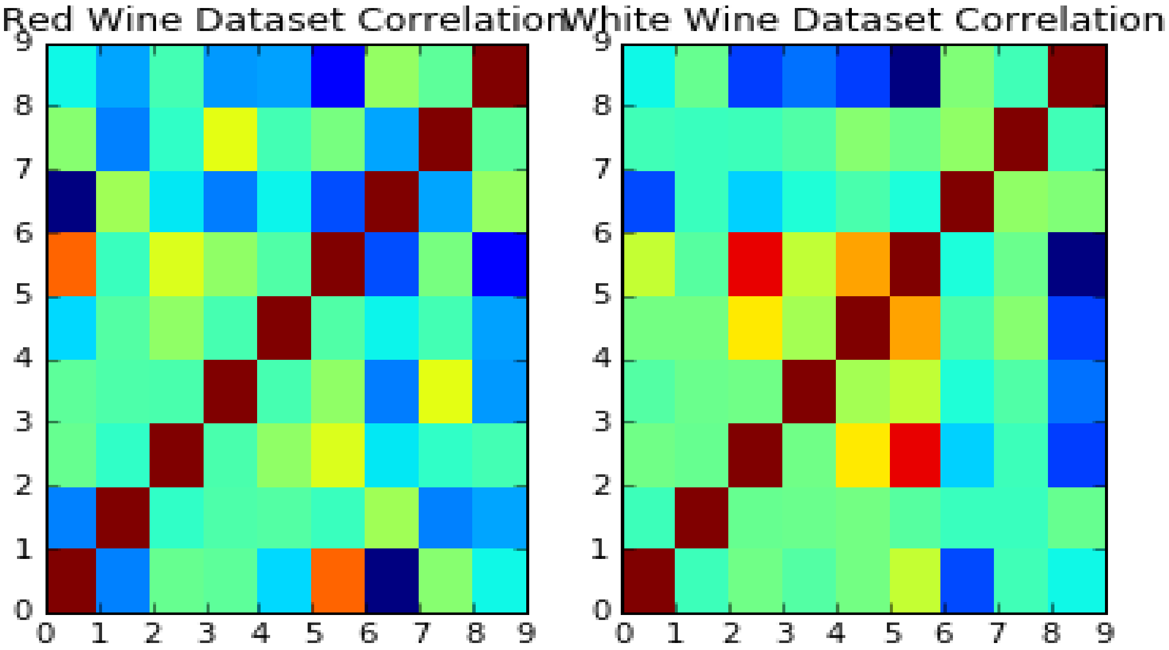
Replace with a binary column for quality

Feature Standardization

Attribute Analysis using Box Plot

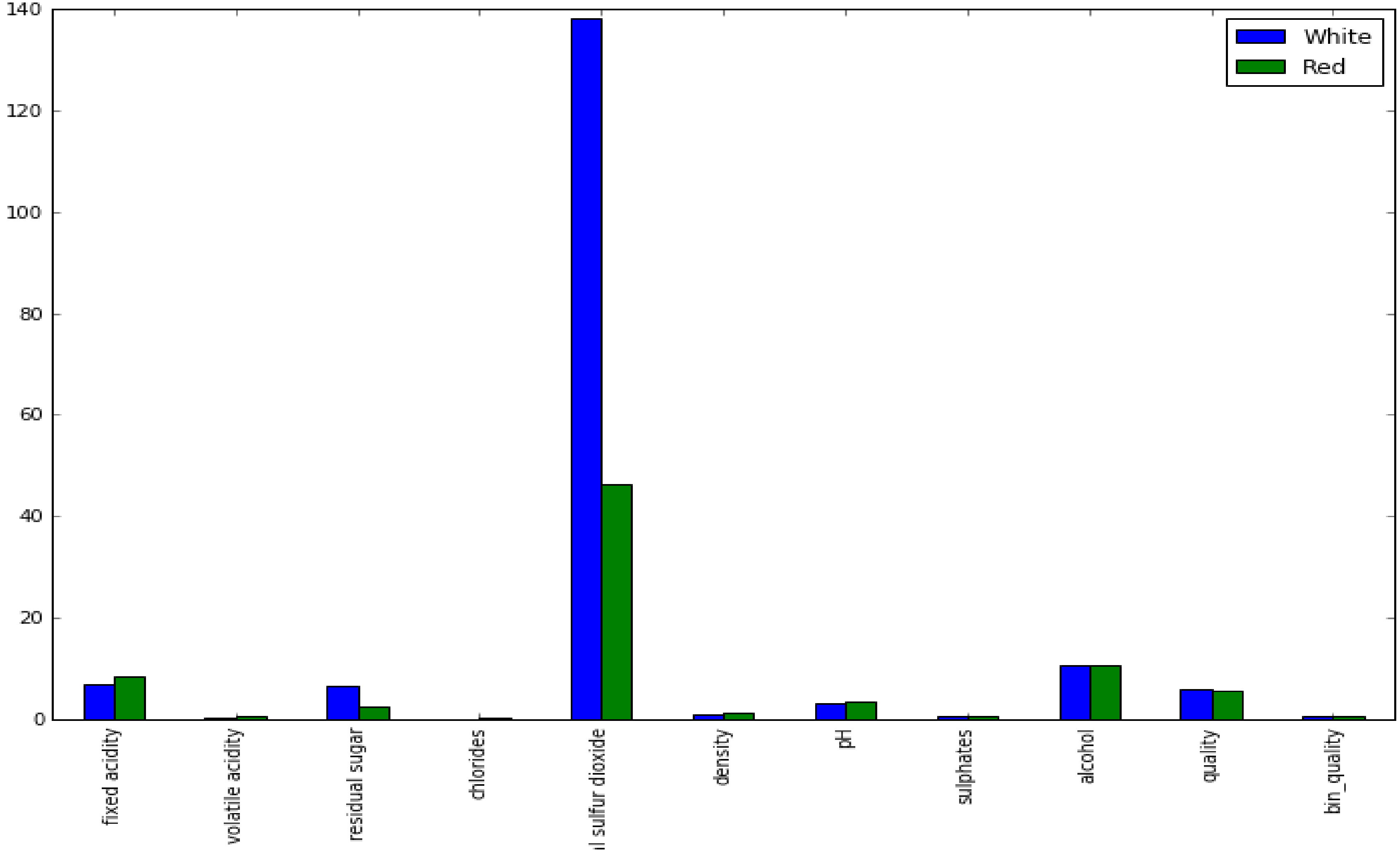


Feaure Correlation



- In red wine dataset, fixed acidity is correlated to a certain degree with density.
- In white wine dataset, there is strong correlation between residual sugar and density.

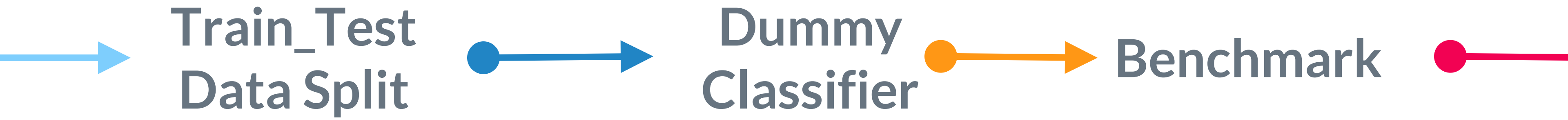
Feature Comparison



Apply Both Supervised and Unsupervised Learning Models

Using Dummy Classifier

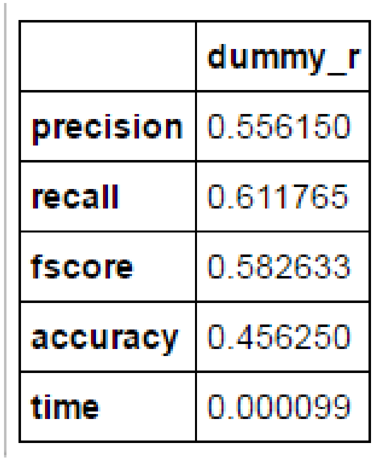
Process



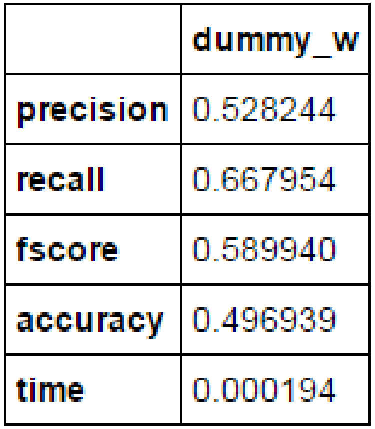
Train_Test Data Split

Dummy Classifier

Benchmark



	dummy_r
precision	0.556150
recall	0.611765
fscore	0.582633
accuracy	0.456250
time	0.000099



	dummy_w
precision	0.528244
recall	0.667954
fscore	0.589940
accuracy	0.496939
time	0.000194

Logistic Regression

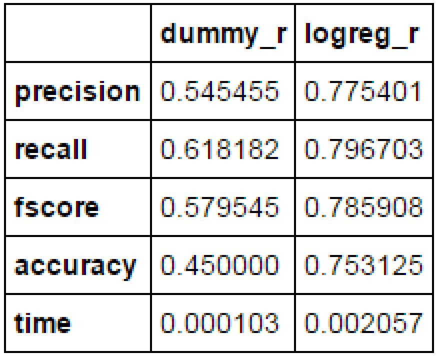
L2 - Ridge Regularization

Randome Forest

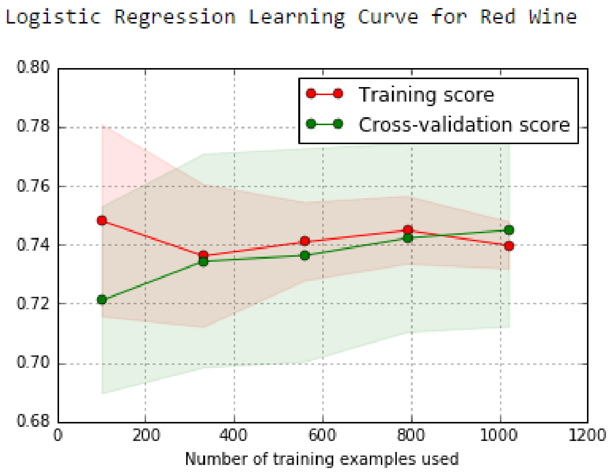
GridSearchCV

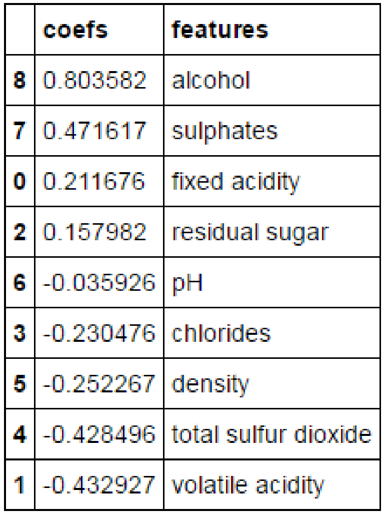
K-Means Clustering

Logistic Regression Performance

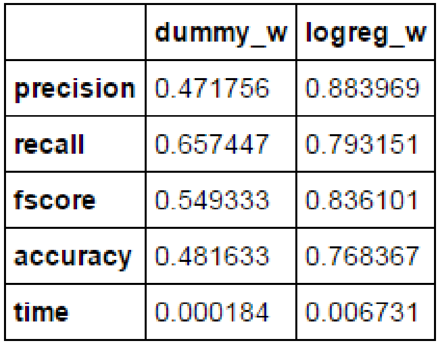


	dummy_r	logreg_r
precision	0.545455	0.775401
recall	0.618182	0.796703
fscore	0.579545	0.785908
accuracy	0.450000	0.753125
time	0.000103	0.002057

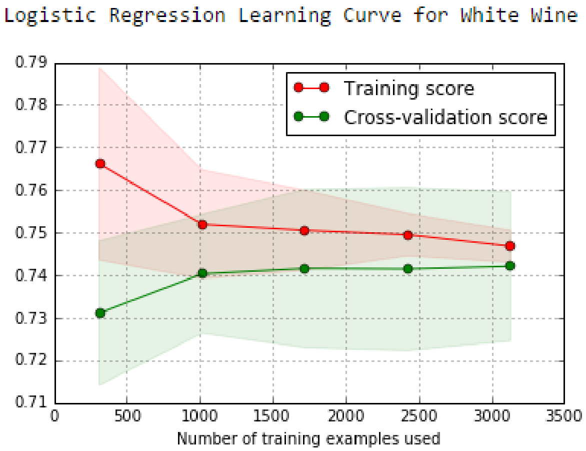

Logistic Regression Learning Curve for Red Wine

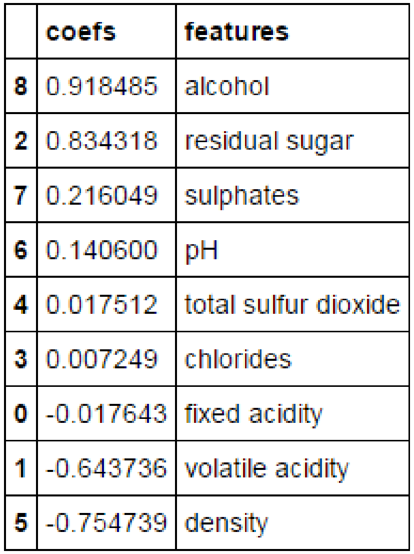


	coefs	features
8	0.803582	alcohol
7	0.471617	sulphates
0	0.211676	fixed acidity
2	0.157982	residual sugar
6	-0.035926	pH
3	-0.230476	chlorides
5	-0.252267	density
4	-0.428496	total sulfur dioxide
1	-0.432927	volatile acidity



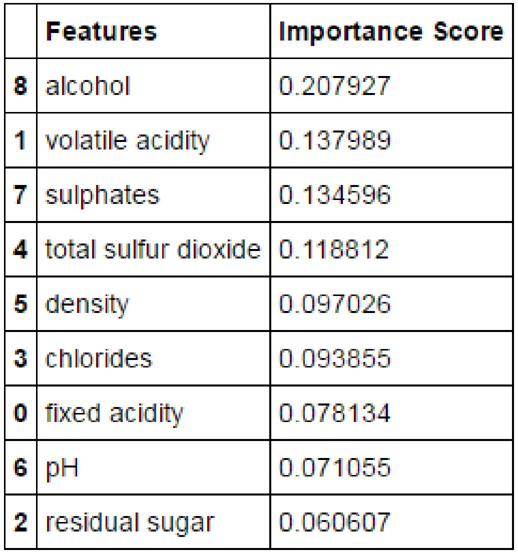
	dummy_w	logreg_w
precision	0.471756	0.883969
recall	0.657447	0.793151
fscore	0.549333	0.836101
accuracy	0.481633	0.768367
time	0.000184	0.006731


Logistic Regression Learning Curve for White Wine

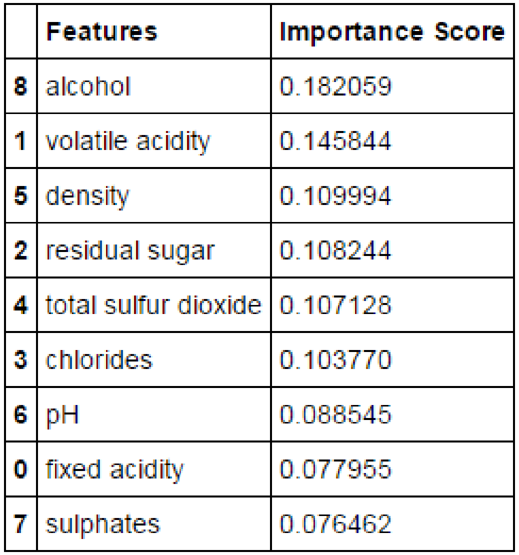


	coefs	features
8	0.918485	alcohol
2	0.834318	residual sugar
7	0.216049	sulphates
6	0.140600	pH
4	0.017512	total sulfur dioxide
3	0.007249	chlorides
0	-0.017643	fixed acidity
1	-0.643736	volatile acidity
5	-0.754739	density

Randome Forest Performance

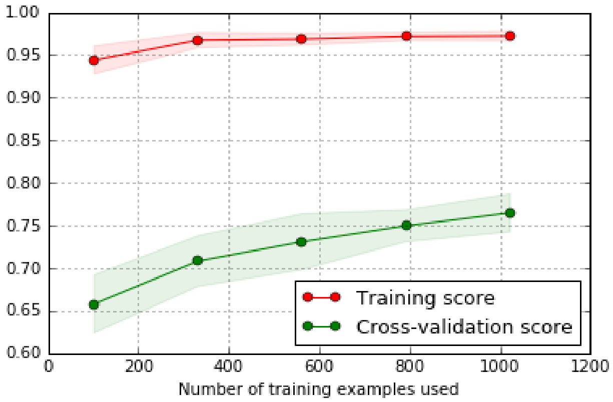


	Features	Importance Score
8	alcohol	0.207927
1	volatile acidity	0.137989
7	sulphates	0.134596
4	total sulfur dioxide	0.118812
5	density	0.097026
3	chlorides	0.093855
0	fixed acidity	0.078134
6	pH	0.071055
2	residual sugar	0.060607

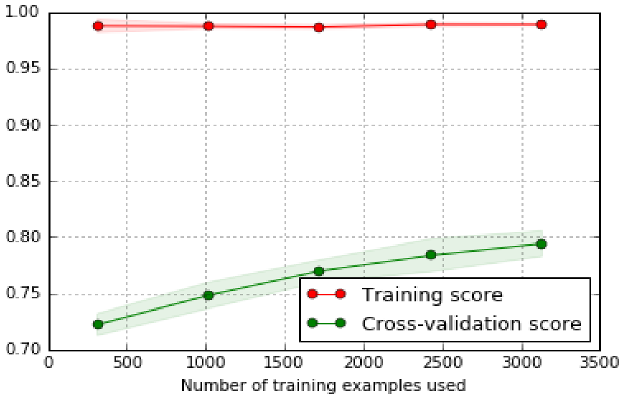


	Features	Importance Score
8	alcohol	0.182059
1	volatile acidity	0.145844
5	density	0.109994
2	residual sugar	0.108244
4	total sulfur dioxide	0.107128
3	chlorides	0.103770
6	pH	0.088545
0	fixed acidity	0.077955
7	sulphates	0.076462

Random Forest Learning Curve for Red Wine



Random Forest Learning Curve for White Wine



K-Means Clustering

StandardScaler

- On both datasets

Create a 'Type' column

- To distinguish red and white wines after they are concatenate
- Type 1 is White Wine; Type 2 is Red Wine.

Concatenate

- Into a master 'Wine' dataset

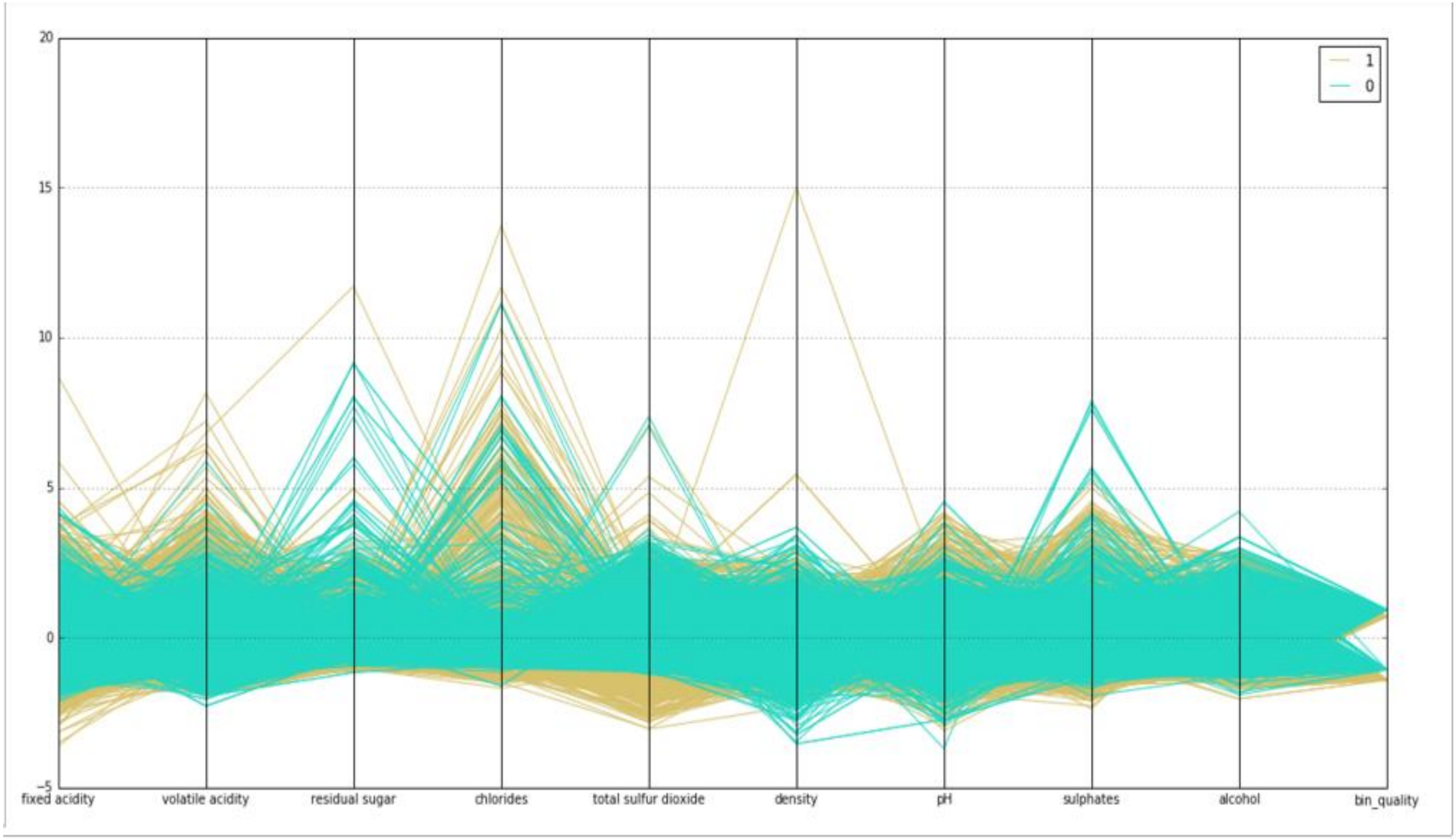
Visualize

- Parallel coordinates

Inertia and silhouette score

- Determine the best n_clusters
- N=2

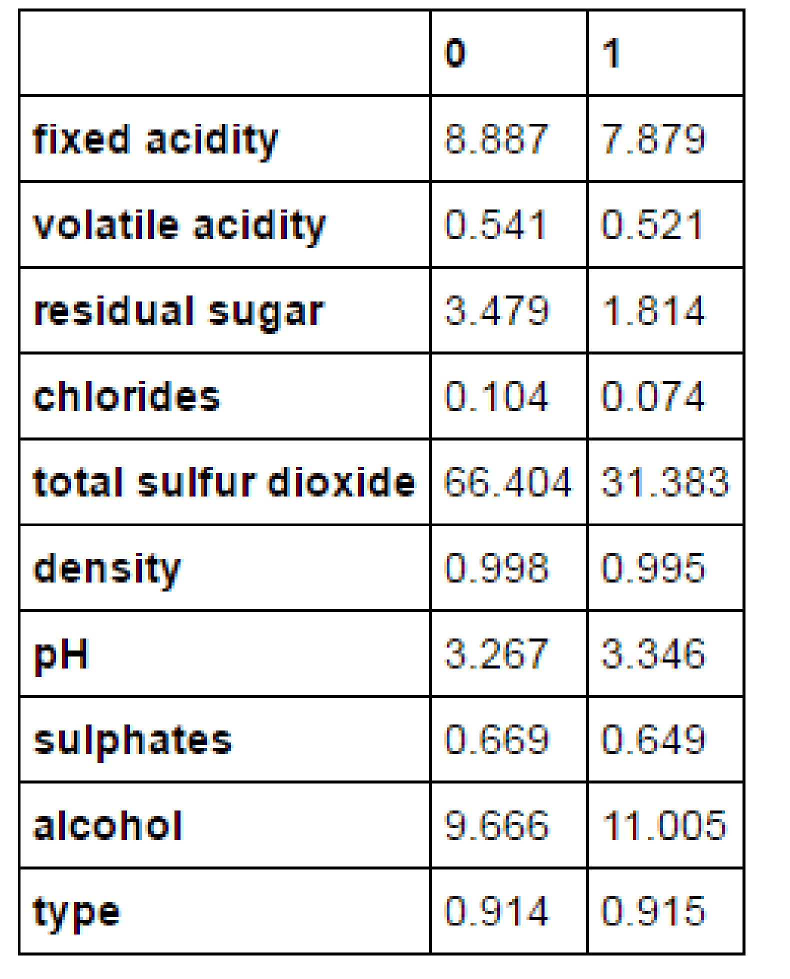
Clustering Result



1 = White Wine   2= Red Wine

Clustering

Result



	0	1
fixed acidity	8.887	7.879
volatile acidity	0.541	0.521
residual sugar	3.479	1.814
chlorides	0.104	0.074
total sulfur dioxide	66.404	31.383
density	0.998	0.995
pH	3.267	3.346
sulphates	0.669	0.649
alcohol	9.666	11.005
type	0.914	0.915

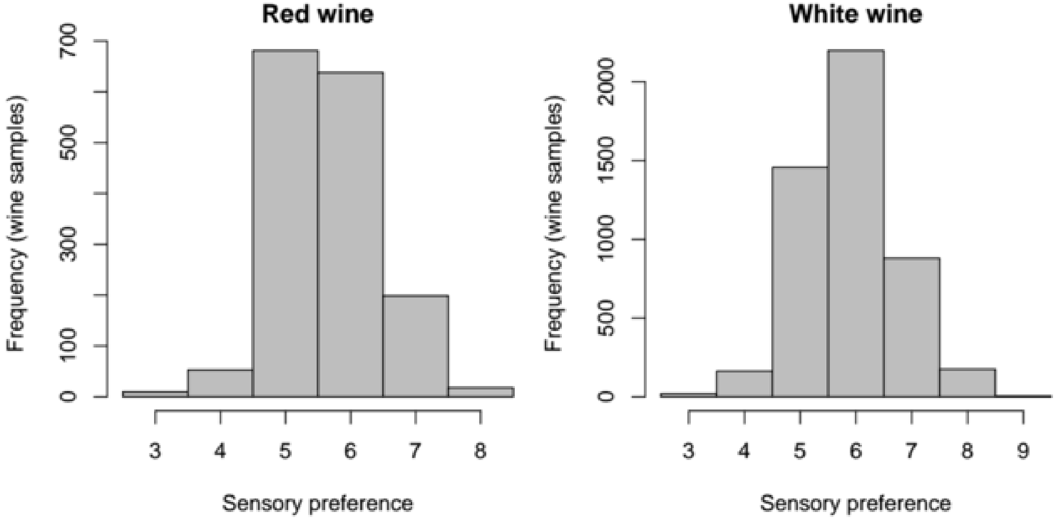
Limitations and Extensions

Limitations:

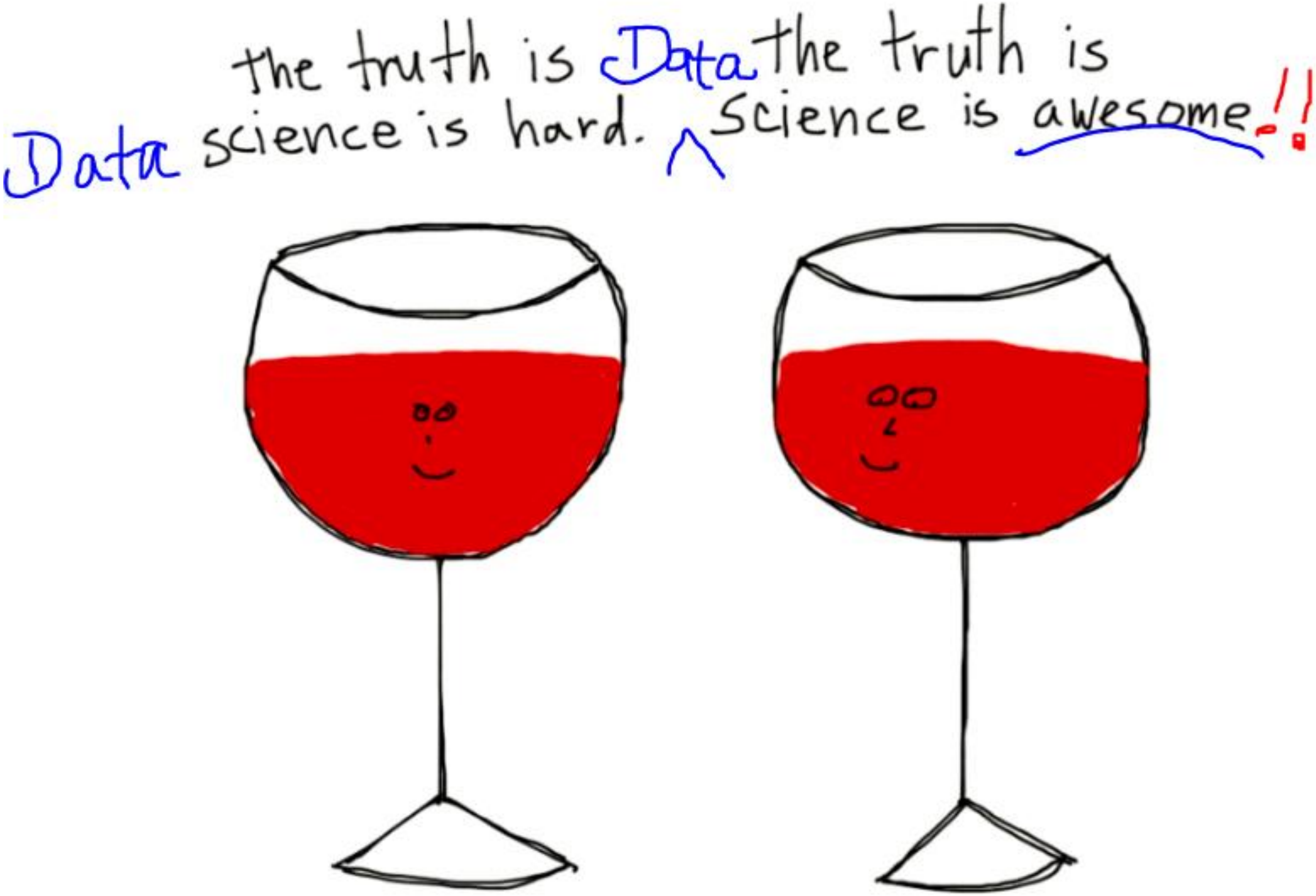
- Production year and evaluation year
- White wine data size is 3X of red wine
- 5 cut-off line (most wines get scores 5 and 6)

Extended Project:

- Predict wine geographic orgins based on both physiochemical characters and chemical components



Questions


the truth is Data the truth is
Data science is hard. ^ Science is awesome!!