Wine Quality Assessment: Logistic Regression on Chemical Properties

Josuke MINAMIGUCHI

1 Introduction

1.1 Motivation

赤または白ワインの化学的特性 (特徴量) に基づいて，そのワインが品質基準—品質カテゴリの値が 6 以上—を満たすかどうかを予測したい.

1.2 Information of the data

データセットとして，UC Irvine Machine Learning Repository から取得したワインの品質データを利用した．そのデータセットは，赤ワインおよび白ワインの化学的特性 (アルコール含量、pH、密度など) からなる特徴量の列と品質 (1 ∼ 10) を表す列とから構成される．赤ワインと白ワインのデータ数はそれぞれ，1599, 4898 であった.

2 Method

特徴量データを標準化したのち，主成分分析を行ってデータ分布の確認を行った.

scikit-learn ライブラリの linear model モジュールの LogisticRegression クラスを用いてロジスティック回帰分析を実行した.

その後, Accuracy, Precision, Recall, F1 score, Specificity から予測を評価した．また, Receiver operating characteristic curve (ROC) を図示し，Area under the curve (AUC) の値を計算した.

以上の分析を赤ワインと白ワインのそれぞれについて行った.

3 Results

3.1 Distributions of the data

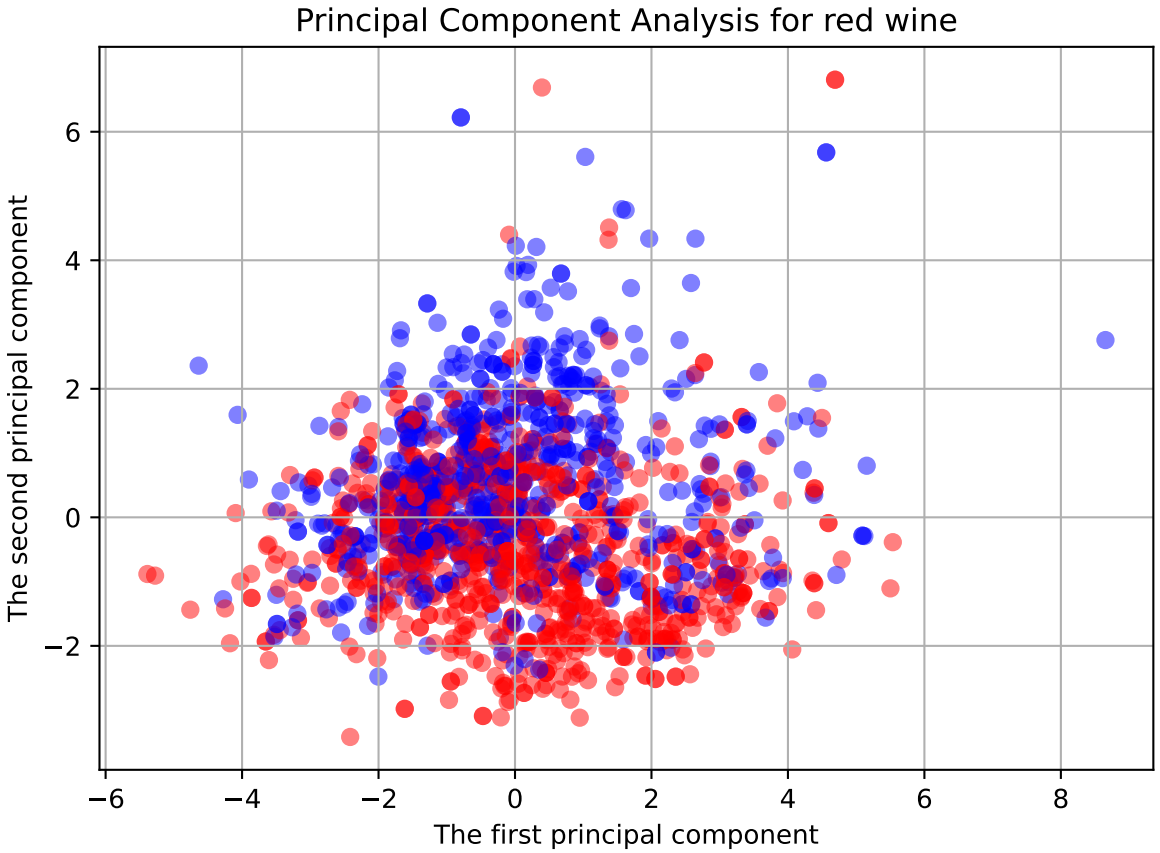

図 1: 標準化した赤ワインのデータの主成分分析

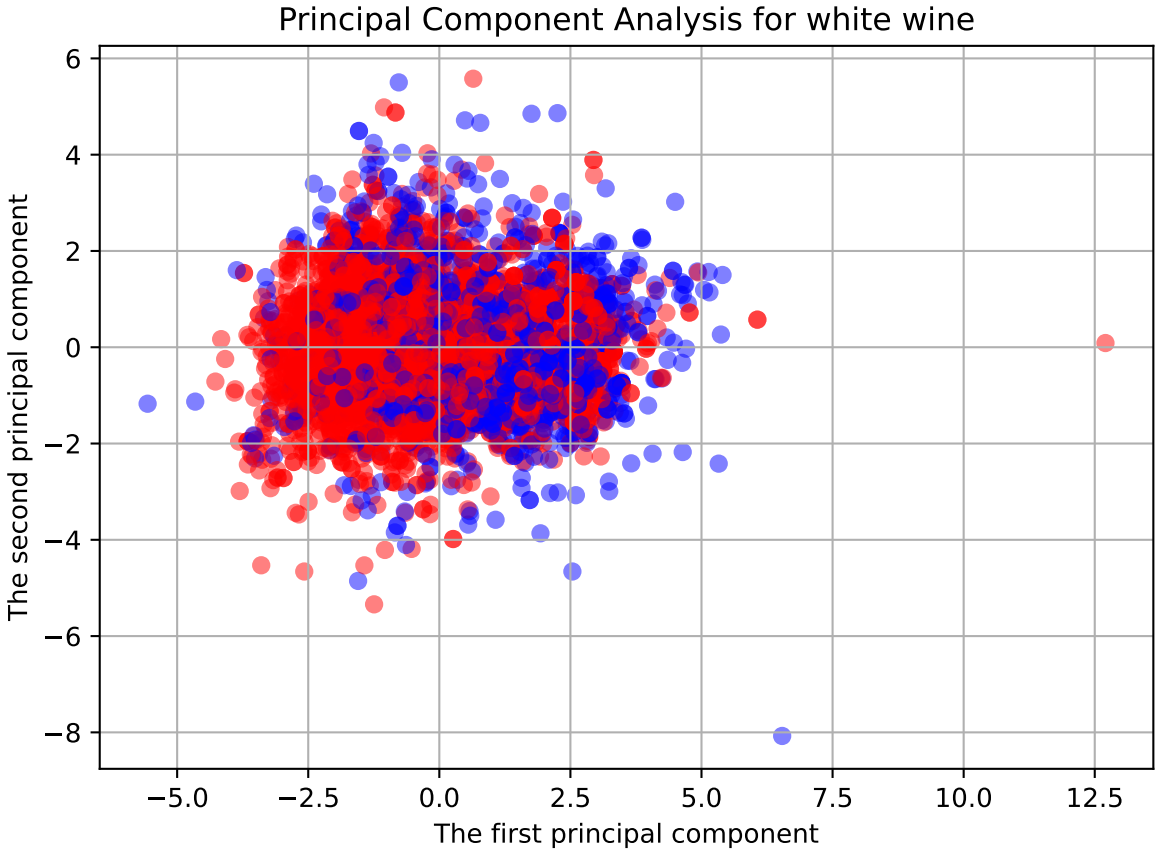

図 2: 標準化した白ワインのデータの主成分分析

3.2 Scores

表 1: 赤ワインのモデルの混合行列

	real 0	real 1
predicted 0	105	47
predicted 1	36	132

表 2: 白ワインのモデルの混合行列

	real 0	real 1
predicted 0	156	81
predicted 1	165	578

表 3: 赤および白ワインのスコア

	Accuracy	Precision	Recall	F1	Specificity
Red	0.741	0.786	0.737	0.761	0.745
White	0.749	0.778	0.877	0.825	0.486

3.3 ROC and AUC

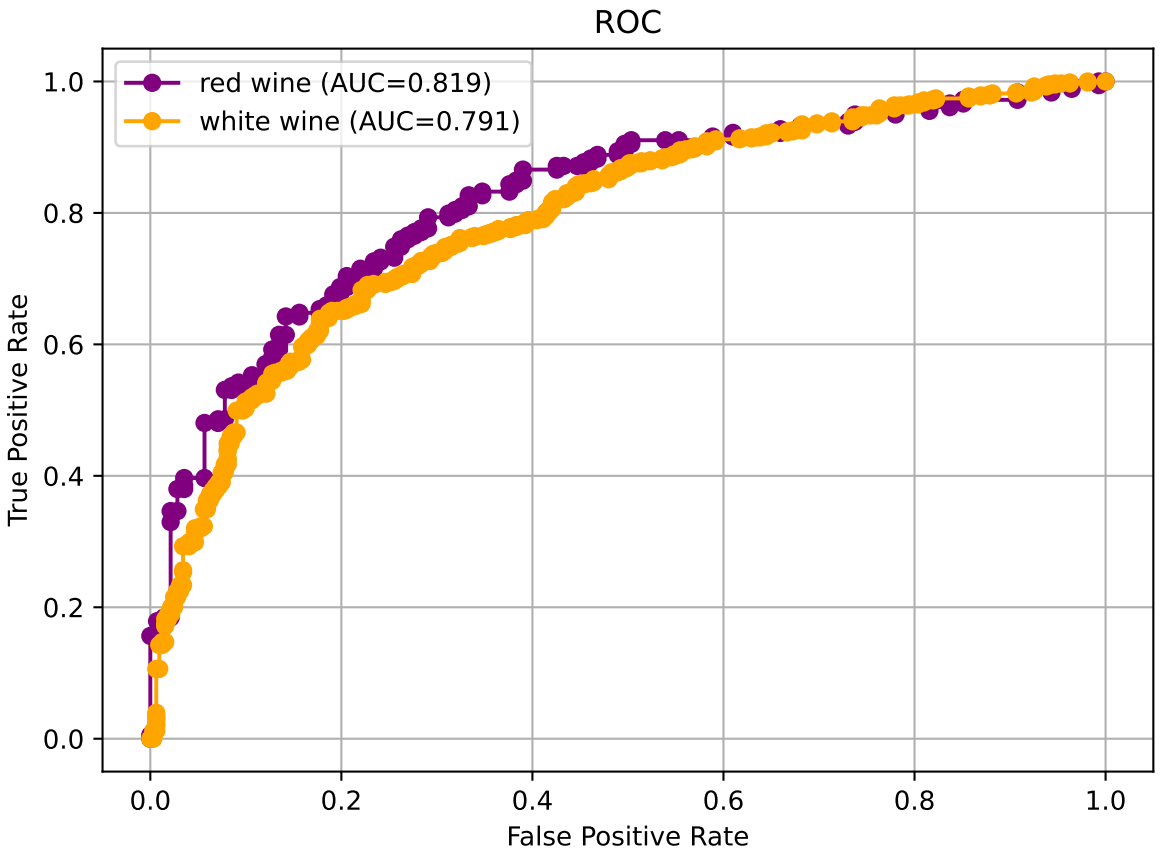

図 3: 赤および白ワインのモデルの ROC. 赤ワインの AUC は 0.819; 白ワインの AUC は 0.791.

4 Discussion

赤および白ワインについてロジスティック回帰分析により，その品質を予測した．どちらの場合でも Accuracy は 0.74 ∼ 0.75 程度であり，十分な正確さ (0.9 以上) とは言えなかった．

Recall と Specificity とについては，Accuracy と Precision とに比べて, 赤と白の場合とでスコアが大きく異なっていた．白ワインでは赤ワインより Recall が高いが Specificity の値が小さかった．原因としてはもともとのデータにおいて品質基準—品質カテゴリが 6 以上—を満たす白ワインのデータ数が，満たさない白ワインのデータ数に比べて多かったことが挙げられる．そのようなデータ数の偏りによって白ワインの Specificity

の値が赤ワインのものより悪くなったと考えた．AUC の違いについては，Accuracy と Precision とにおける赤と白の違いと同程度であった．AUC の赤ワインと白ワインの違いは，赤ワインのデータ数の少なさが影響していると考えられた．実際，ROC も赤ワインの場合はデータ数がすくないために，白ワインの場合と比べて，曲線の揺らぎが大きいことが確認できた．

以上のことから，本プロジェクトで用いたロジスティック回帰分析と同じ手法でより高精度の予測を実現するためには，(i) 赤ワインについてはデータ数を増やすことと，(ii) 白ワインについては品質基準を満たさないデータ数を増やすこととが必要だ—と考えた．

より厳密な赤ワインと白ワインの比較のためには，評価指標の誤差をつけるべきである．