Wine Quality Class Prediction

Presented by
Artem Ramus

Background

Data set created by: Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009

Past Usage: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties.
In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Link to the dataset at Kaggle: https://www.kaggle.com/danielpanizzo/wine-quality

Introduction

The goal of the project is modeling wine quality by data mining from physic-chemical properties with two separate data sets one for white and another for red wine.

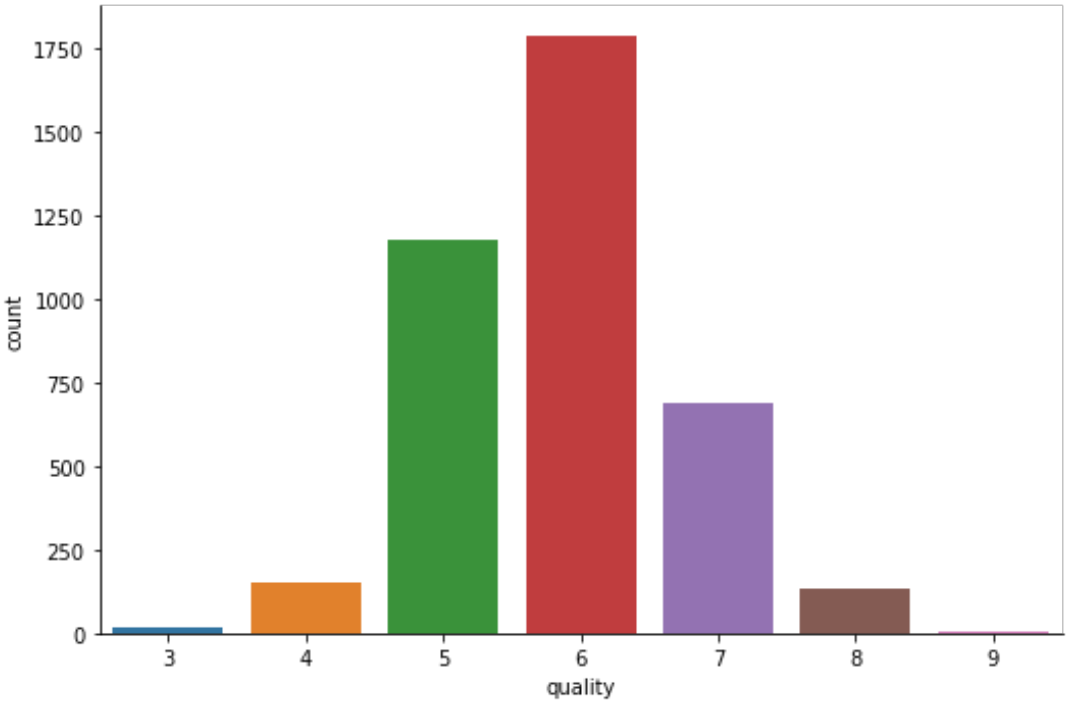
Both data sets were created, using red and white wine samples. The inputs include objective tests, like PH values, and the output is based on sensory data, median of at least 3 evaluations made by wine experts. Each expert graded the wine quality between 0very bad, and 10, excellent.

Both data sets has 11 numerical variables to predict wine quality.

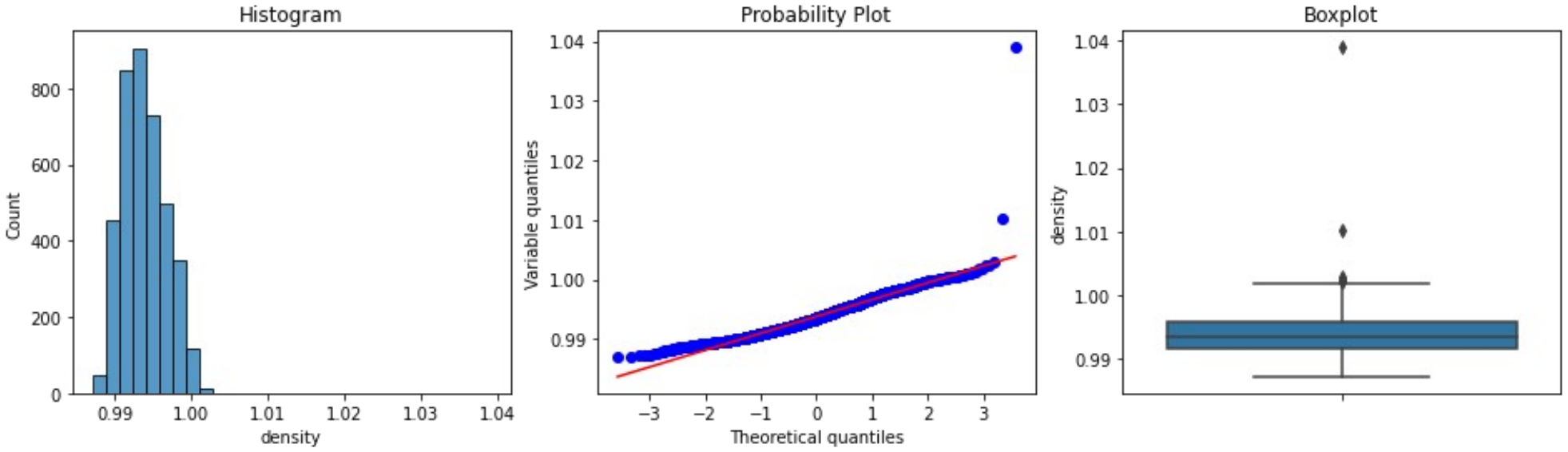
Exploratory Data Analysis - Quality

In both data sets, classes 3, 4 and 8 are under-sampled.

In white wine data set, class 9 contains only 5 observations.



Exploratory Data Analysis

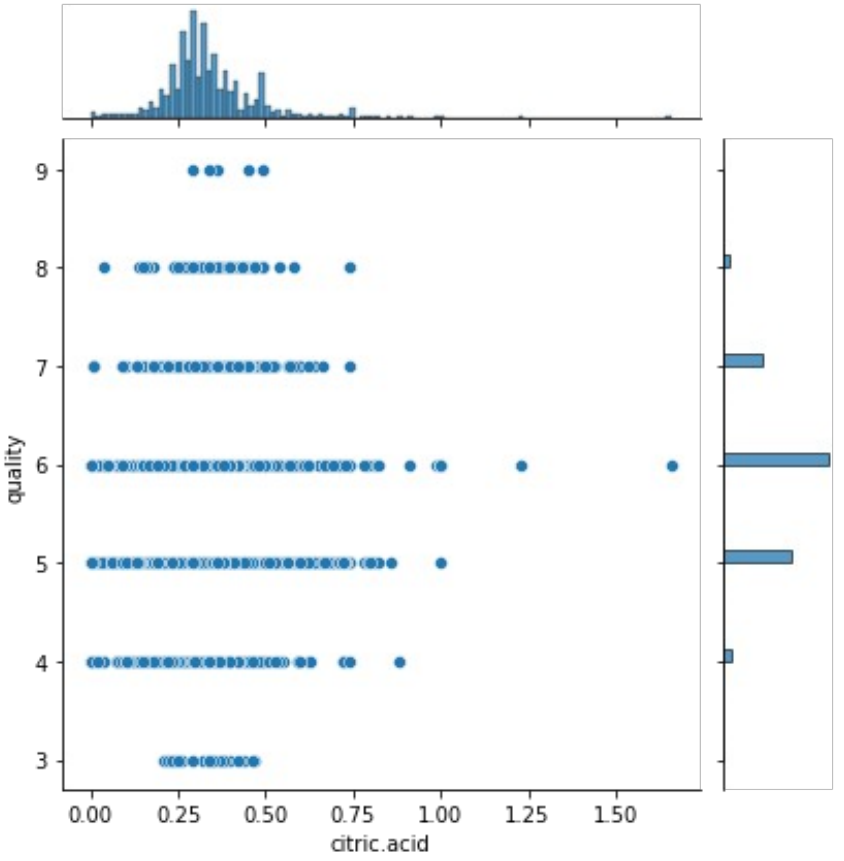


Some variables are skewed and include potential outliers

Exploratory Data Analysis

Some PH related variables and density are correlated

All the variables are distributed differently in classes



Feature Engineering and Selection

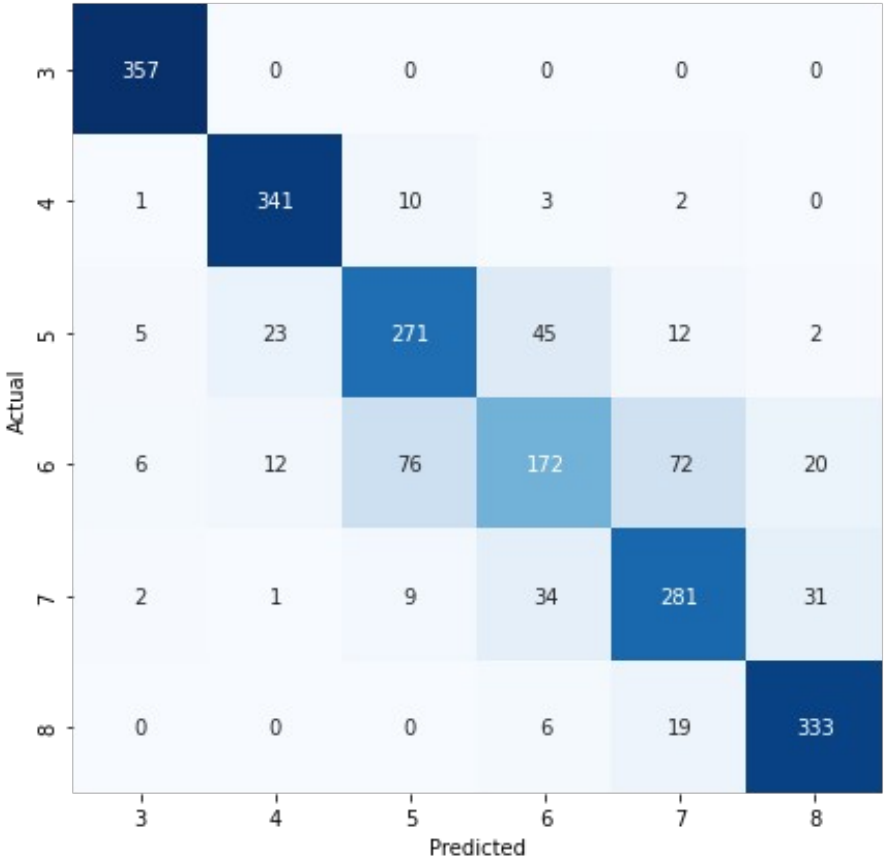
- Using all the features for modeling
- Original fearures' distribution
- Deletion of class 9
- Over-sampling of under-represented classes 3, 4 and 8 with SMOTE
- Modeling with Random Forest and XGBoost

Performance, white wines

Random Forest:

One-vs-One ROC AUC scores:
0.966332 (macro),
0.966307 (weighted by prevalence)

One-vs-Rest ROC AUC scores:
0.966304 (macro),
0.966275 (weighted by prevalence)

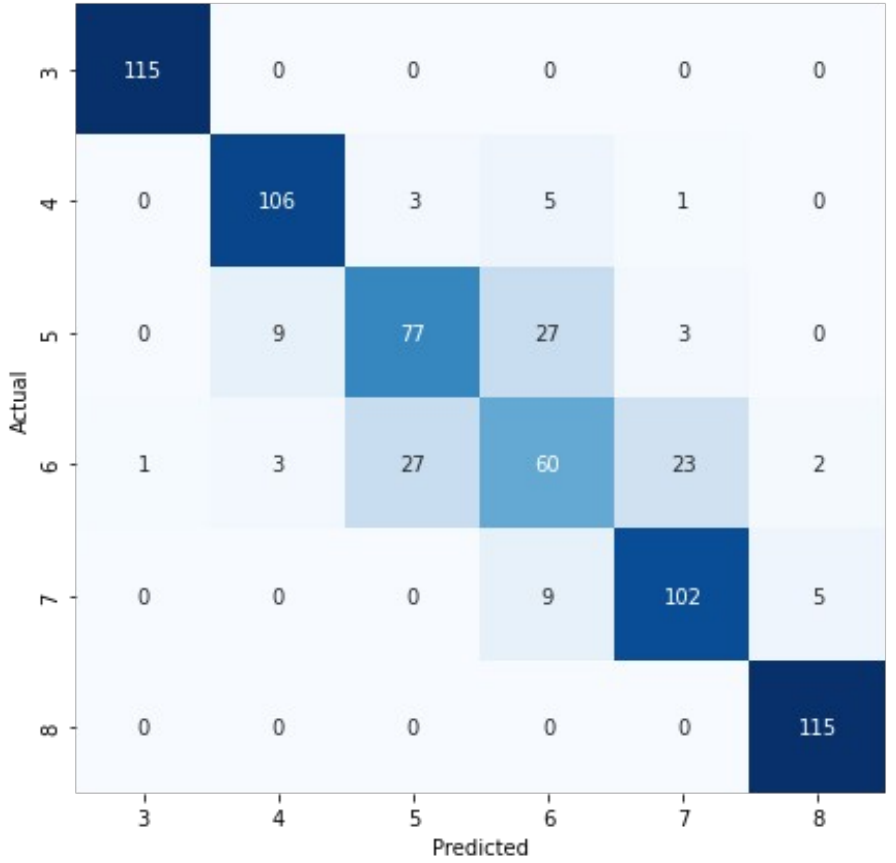


Performance, red wines

Random Forest:

One-vs-One ROC AUC scores:
0.970649 (macro),
0.970547 (weighted by prevalence)

One-vs-Rest ROC AUC scores:
0.970536 (macro),
0.970421 (weighted by prevalence)



Summary and Conclusions

White wines

Under-represented classes 3, 4 and 8 were hard to predict. After applying SMOTE over-sampling technique, ROC-AUC metric of XGBoost surpasses 0.90 and Random Forest – 0.96.

Red wines

Under-represented classes 3, 4 and 8 were hard to predict. After applying SMOTE over-sampling technique, ROC-AUC metric of XGBoost surpasses 0.94 and Random Forest - 0.97.

The end

Thank you for your attention!