Credit Card Fraud Detection

Presented by
Artem Ramus

Introduction

Modeling wine preferences by data mining from physic-chemical properties. There is two separate data sets – white and red wine.

In the above reference, two data sets were created, using red and white wine samples. The inputs include objective tests (e.g. 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 0 (very bad) and 10 (very excellent). Several data mining methods were applied to model these data sets under a regression approach. The support vector machine model achieved the best results. Several metrics were computed: MAD, confusion matrix for a fixed error tolerance (T), etc. Also, we plot the relative importance of the input variables (as measured by a sensitivity analysis procedure).

Background

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.

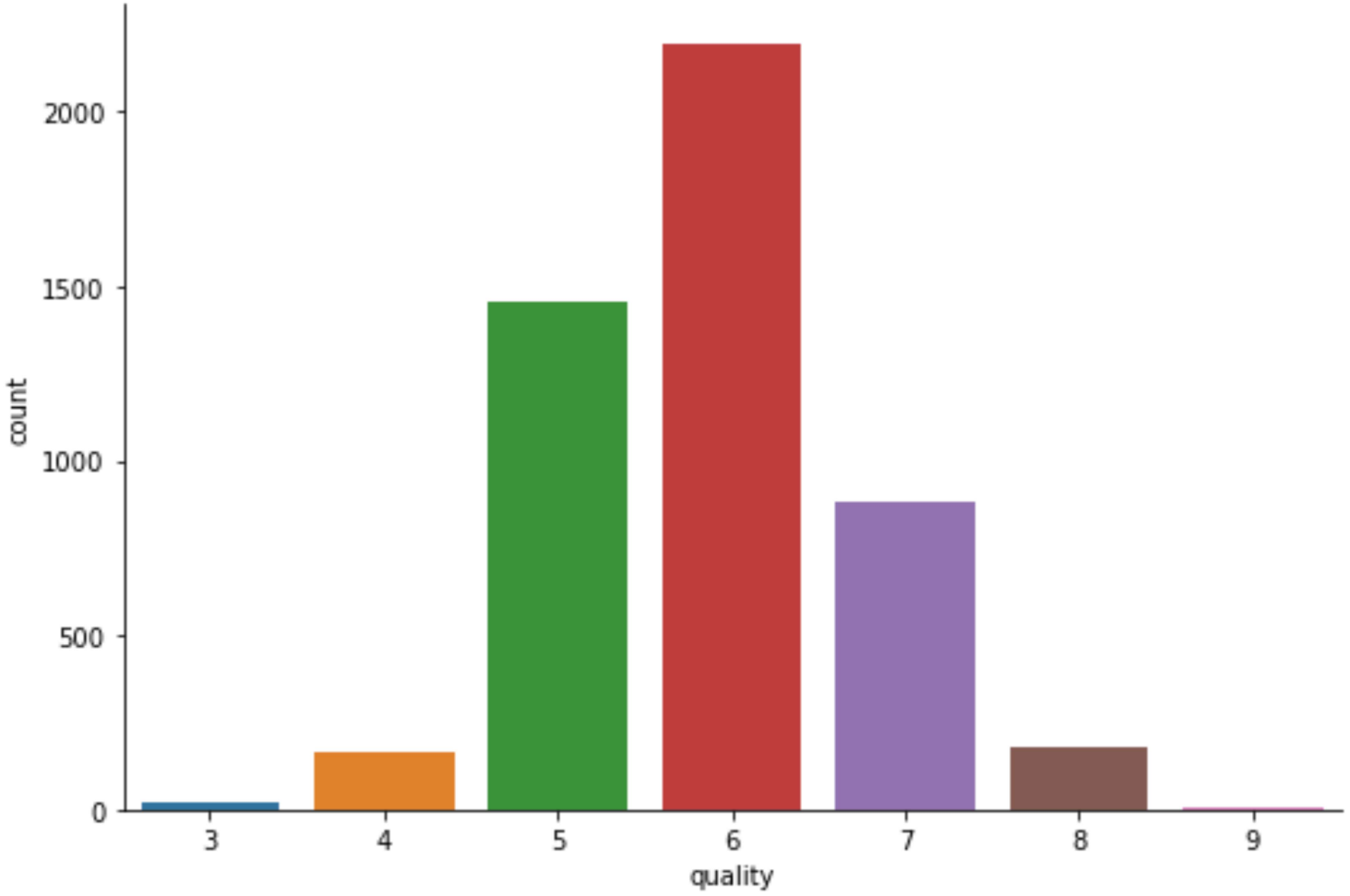
Link to hte dataset at Kaggle: https://www.kaggle.com/danielpanizzo/wine-quality

Methodology

- The data sets checked for duplicates and null values
- Outliers and skewness checked
- Constant, quasi-constant and highly-correlated features deleted
- Cross-validation used
- Rare target classes identified and deleted
- Hyper-parameters optimization technique grid search used

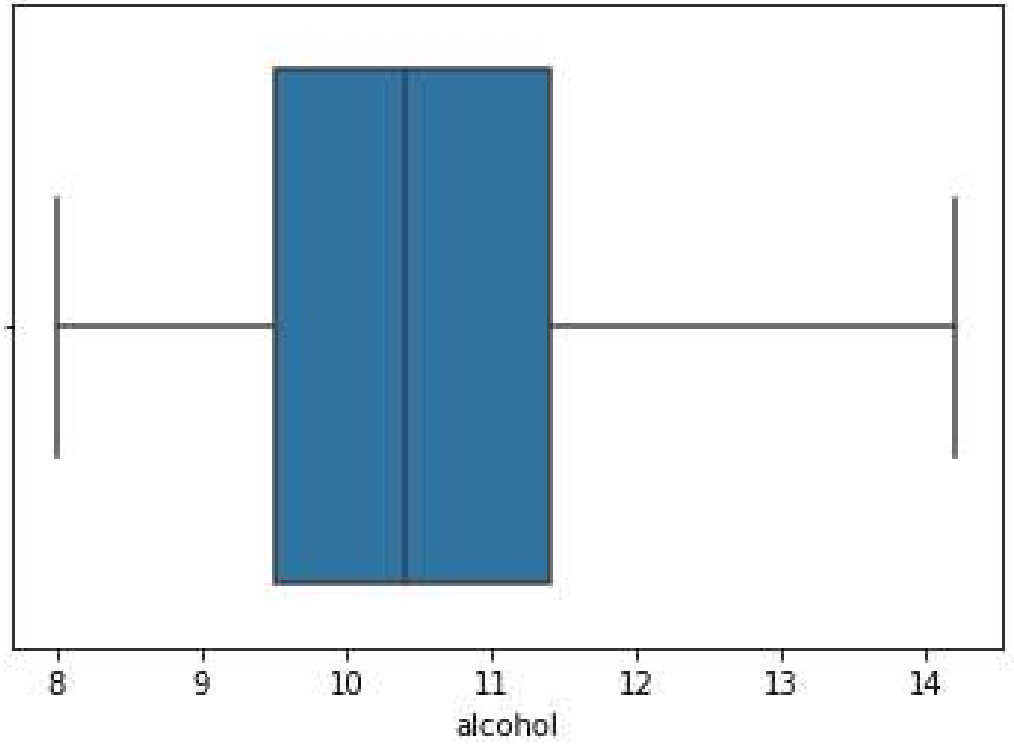
Methodology

- Identification of rare classes
- The rare classes excluded from the prediction



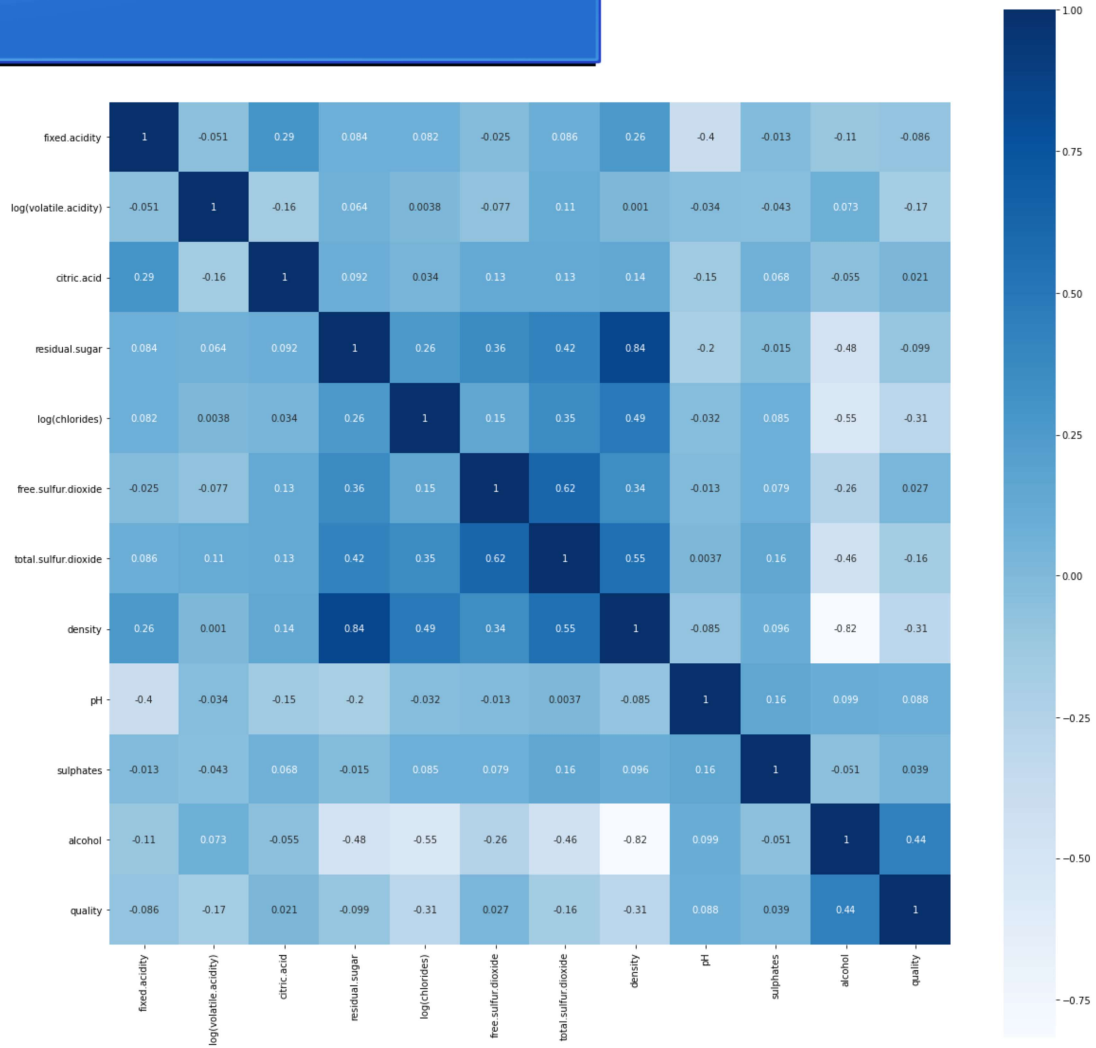
Methodology

- Box plot of the alcohol content
- Distribution of all the features verified and corrected



Methodology

- Correlation matrix
- Linear correlation of all the columns examined



Performance of the Model

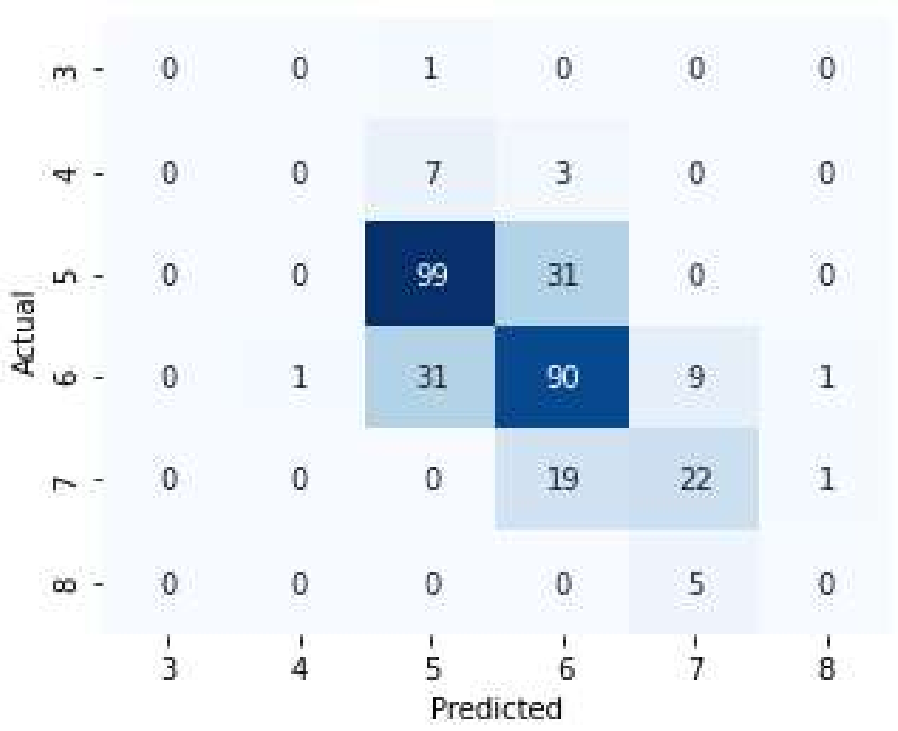
Following models were tested:

- Logistic regression
- KNN
- SVM/Kernel SVM
- Naive Bayes
- Random forest
- XGBoost
- Cat boost
- ANN

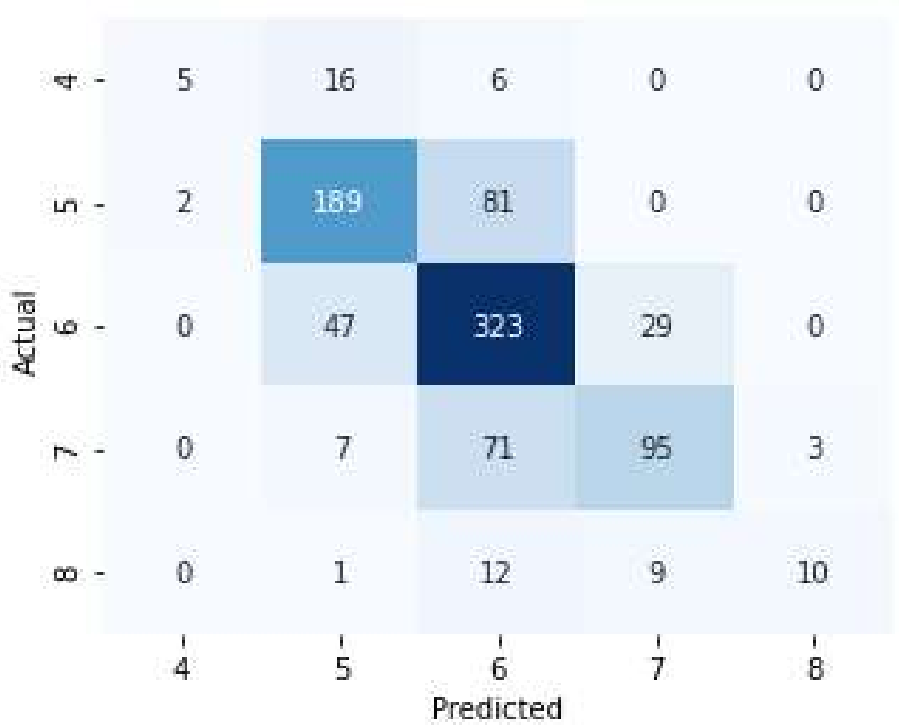
Hyper-parameters optimization technique grid search used with the random forest that gave the best performance in this case

Performance of the Model

Red wine 66%



White wine 70%



The end

Thank you for your attention!