

Predicting Wine Quality: A Conundrum

Would you like some cheese with that?

Kalbi Zongo, Song Hoa Choi, Gina Shellhammer, Matt Edwards

June 2, 2014

Outline

- 1 Introduction
- 2 Machine Learning Methods
- 3 Findings
- 4 Discussion

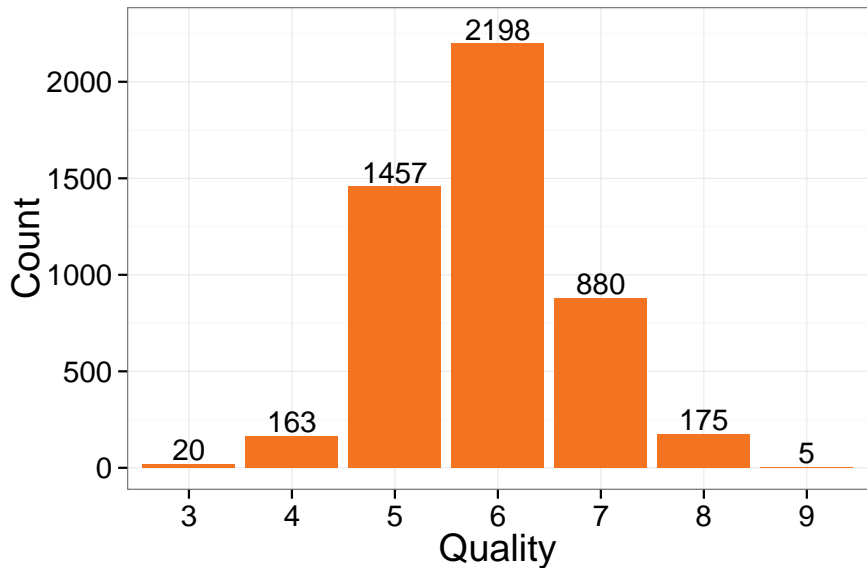
Task

Predict the blind taster quality score of a wine based on chemical tests.

Data

- Two Datasets: Red & White vinho verde wine samples from northern Portugal
- 1599 & 4898 rows, respectively
- 11 Explanatory variables: measurements from various phytochemicals in wine
- Response variable "quality" is discrete variable on ordered scale from 0 (worst) to 10 (best)

White Wine Quality Scores



Training and Testing Sets

- Training and Testing set constructed through stratified sampling.
- Quality variable was the strata
- Why: Ensure representation of all quality categories in both Training & Testing datasets.
- How: 37.5% of items (rounded up) in strata were randomly selected to be in the testing set. Remaining 62.5% were the training set.
- Same training & testing sets used for each analysis type.

Regression

content...

Classification

content...

Random Randomness is Random

- 75% of Quality ratings were either 5 or 6.
- Is randomly assigning 5 or 6 to everything as good as, or better than, our other methods?
- Using `rbinom(1,1,0.6014)`, 1s were predicted as quality 6, 0s as quality 5
- Probability of 60.14% because from Training Set, considering only 5s and 6s, 6s were 60.14% of total observations
- Our base line success rate to compare other methods.

Regression: 50% Success Rate

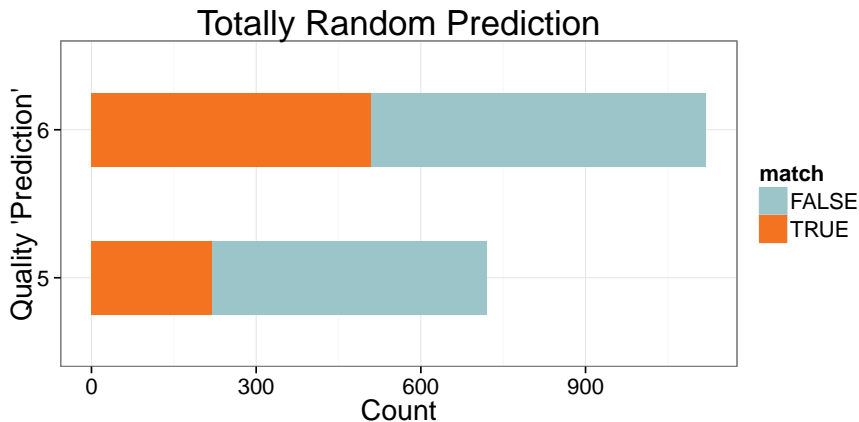
content... change success rate in title

Classification: 50% Success Rate

content... change success rate in title

Random 'Prediction': 39.67% Success Rate

Turns out, that's not really a great 'prediction' method. Who knew?



content...