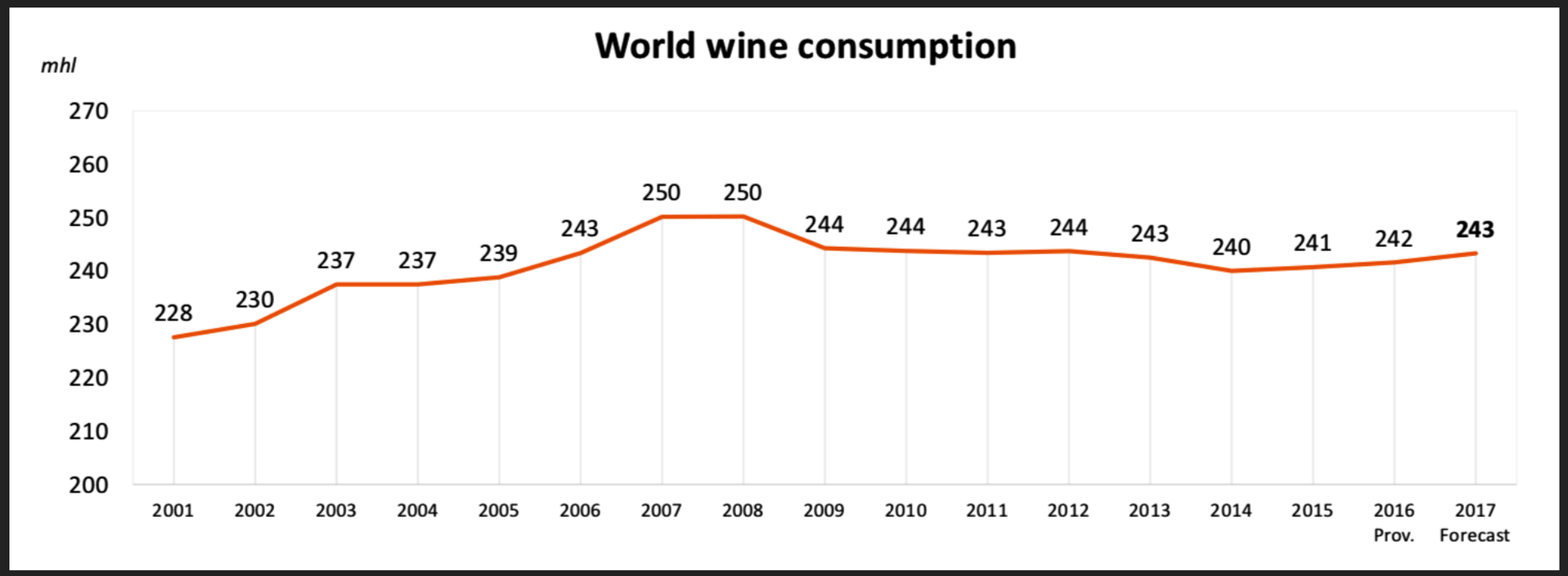
Predict Wine Quality Without Sensory Analysis

White Wine Quality Classification

WINE CONSUMPTION

- According to the International Organization of Vine and Wine, the world wide wine consumption has growing steadily each year.
- More than 20 countries with wine production above 1 million hectoliter.



State Of The Vitiviniculture World Market, April 2018

WINE CLASSIFICATION AND RATING

- Wines are classified by its production region or grape varieties.
- Some bottled wines are given numerical score by sensory tasting. The raters can be individuals or panels.
- Both classification and rating can influence pricing of wine and transaction.

THE PROBLEM

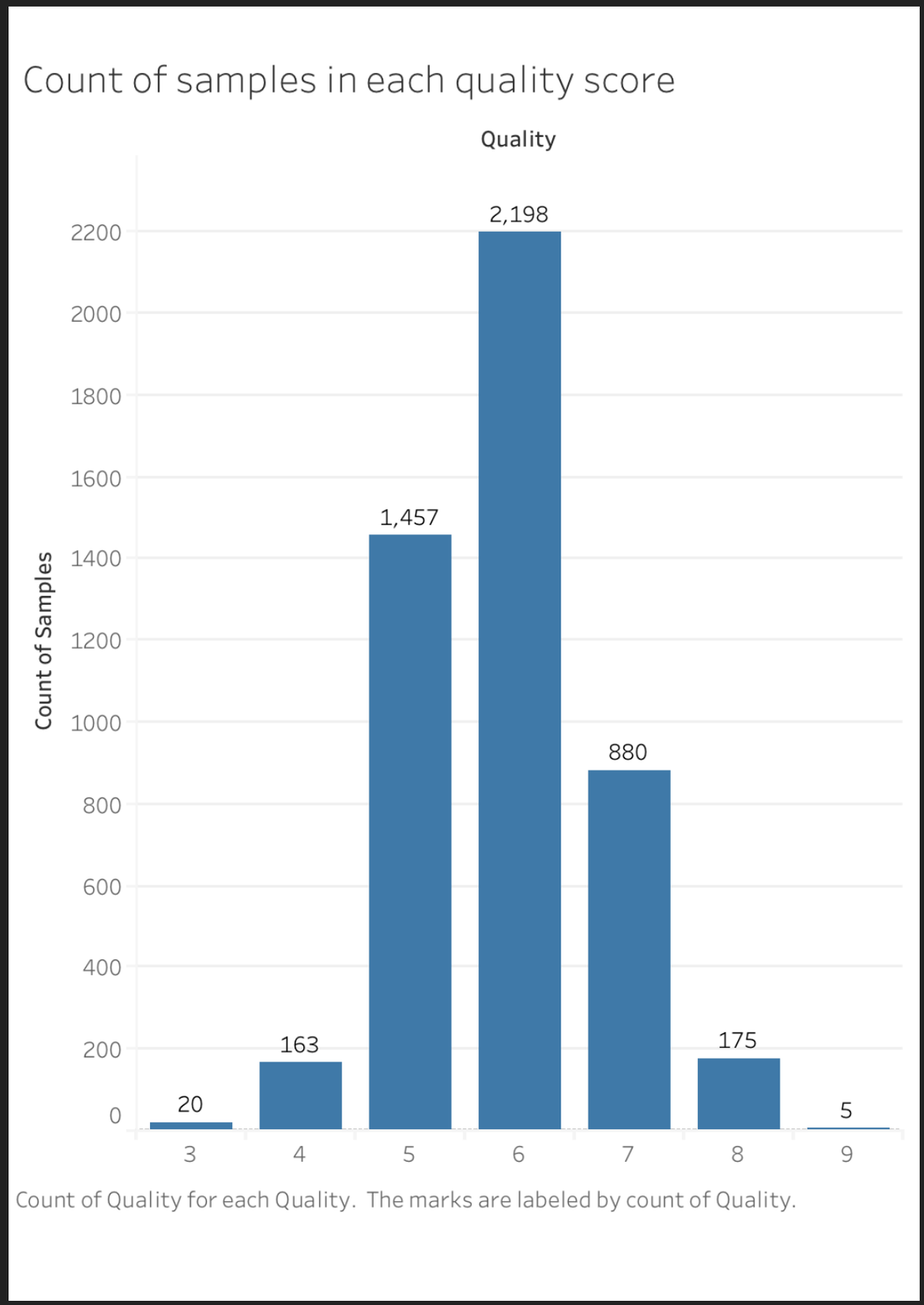
- Wine rating can be great marketing strategy, but wine critics and professional tasters can expensive and limited.
- Predictive modeling can help winemakers to market their wines by simulating professional tasting using data.
- Wine traders and wholesalers can also predict the wine ratings before purchasing from wineries overseas.

DATASET

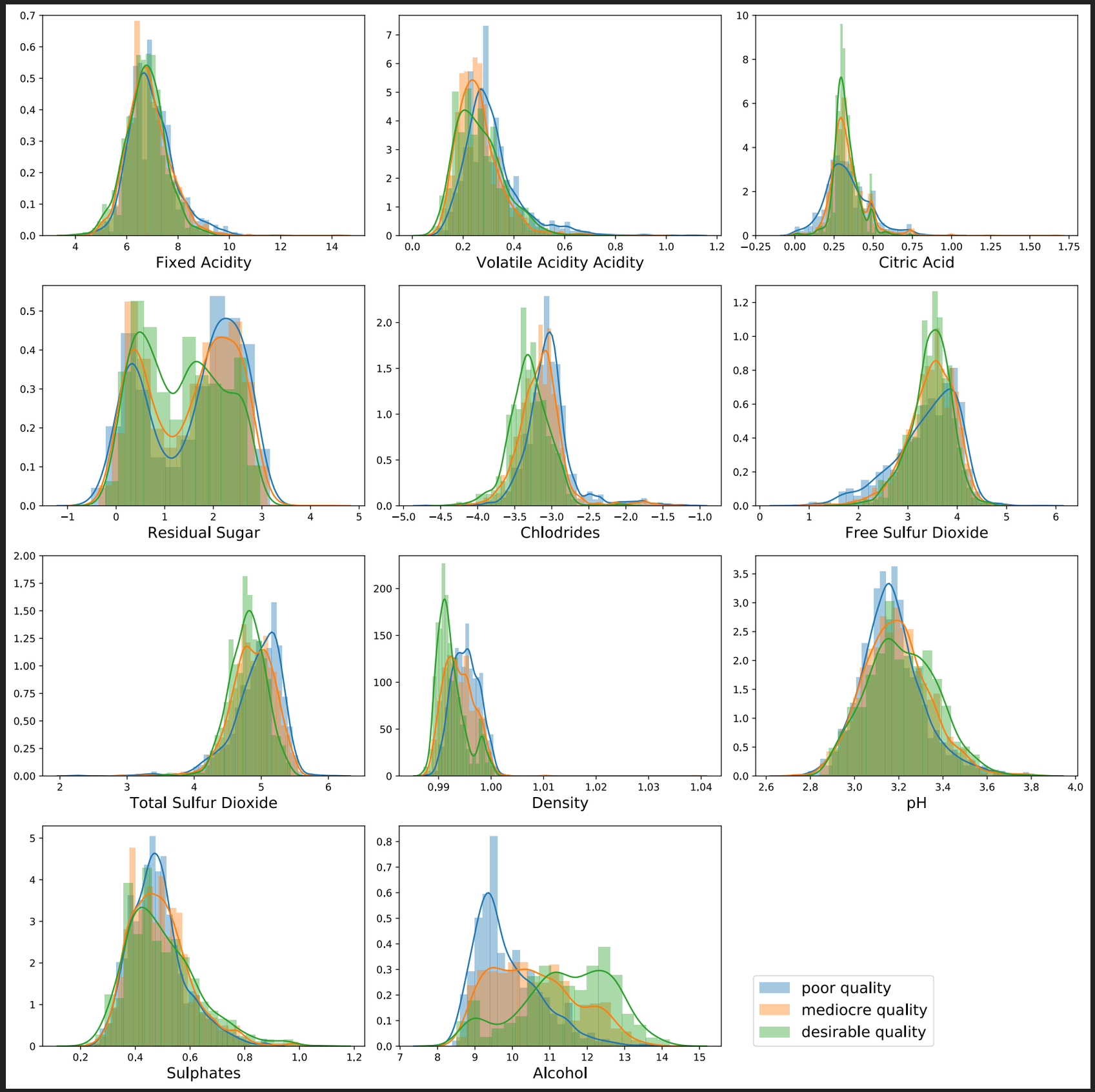
- Wine Quality Data Set from UCI Machine Learning Repository.
- Contains 4898 samples of white wine.
- 11 attribute variables of physicochemical measurements.
- Sensory preference by assessors as target variable.

TARGET VARIABLE DISTRIBUTION

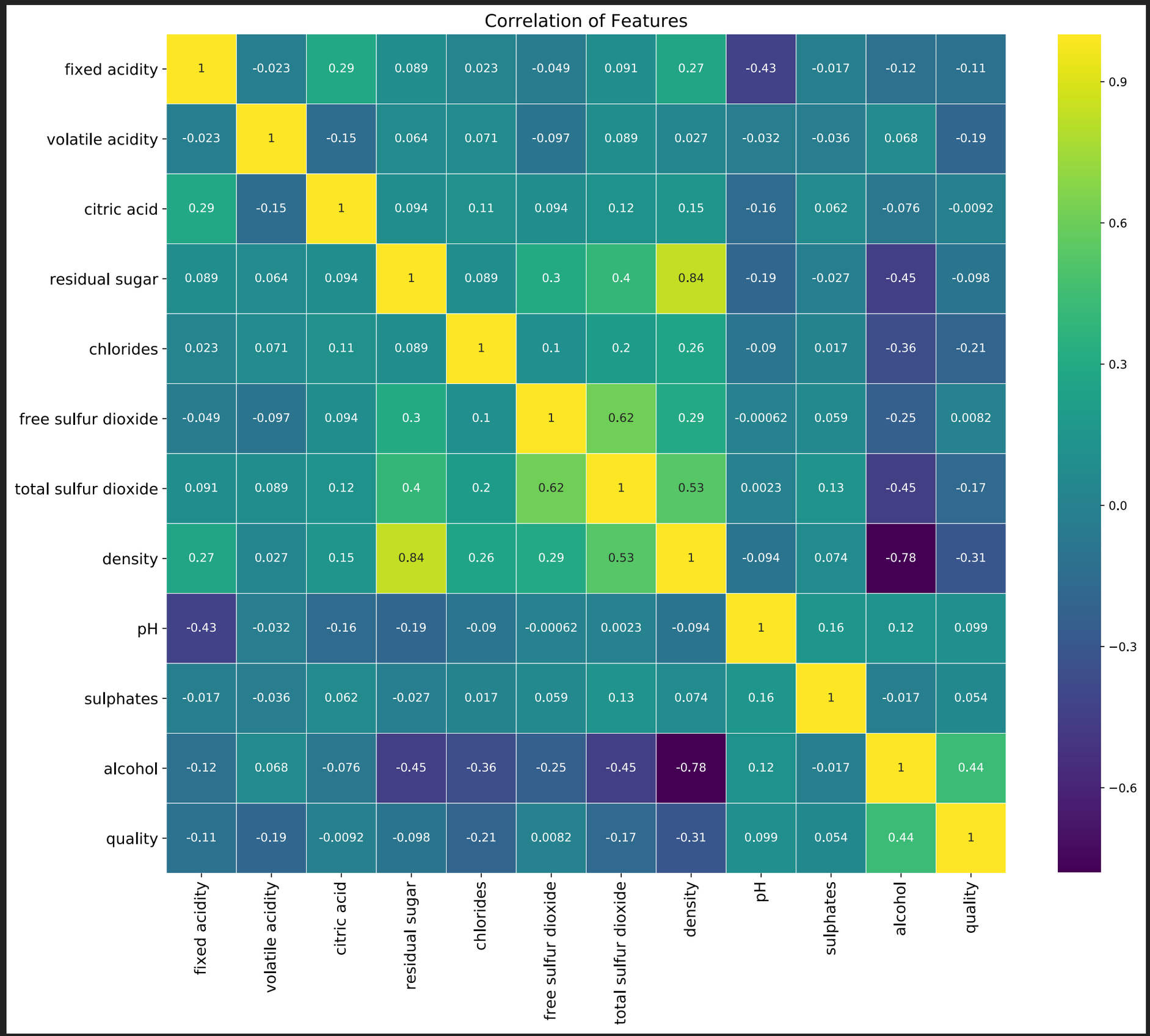
- The distribution is resembling normal distribution.
- Entire data are in 7 quality scores, and can be divided into 3 target variable labels:
  - Quality < 6: poor (33.48%)
  - Quality = 6: mediocre (44.88%)
  - Quality > 6: desirable (21.64%)



DISTRIBUTION OF FEATURES



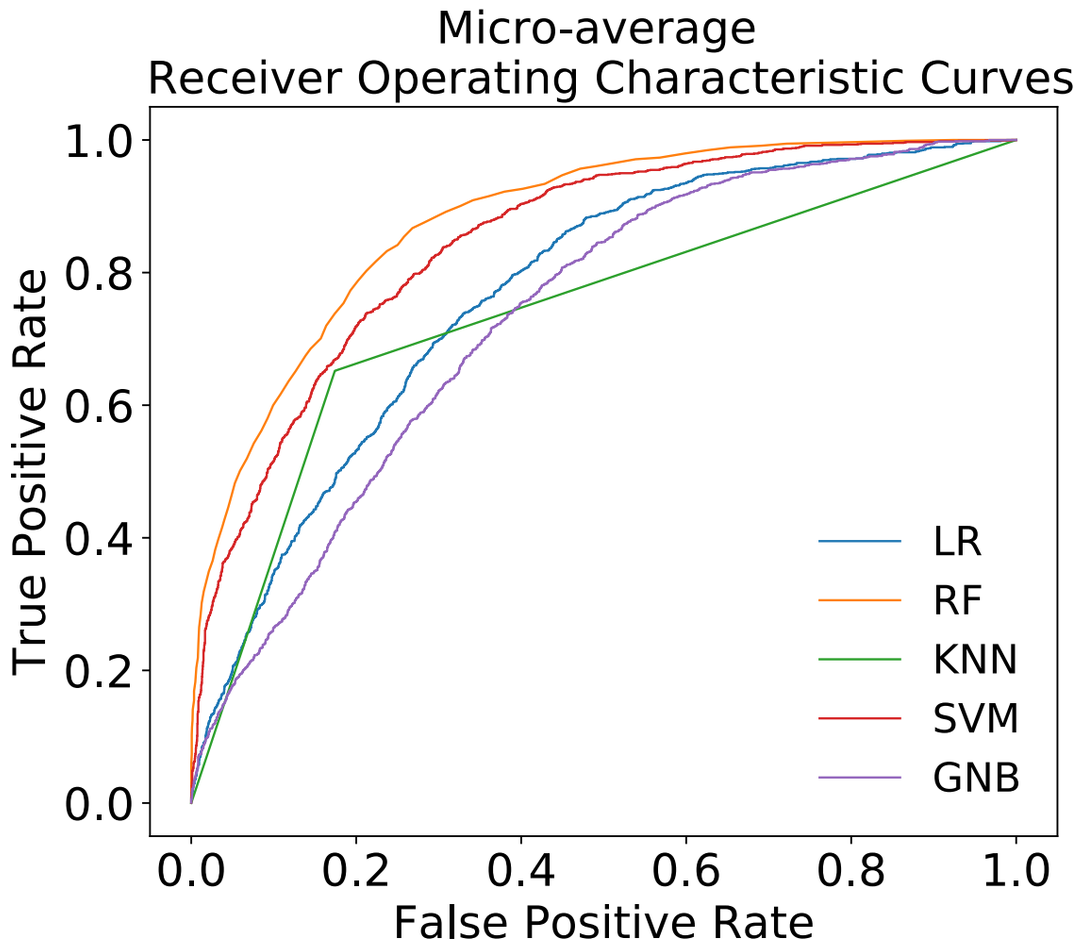
FEATURE CORRELATION

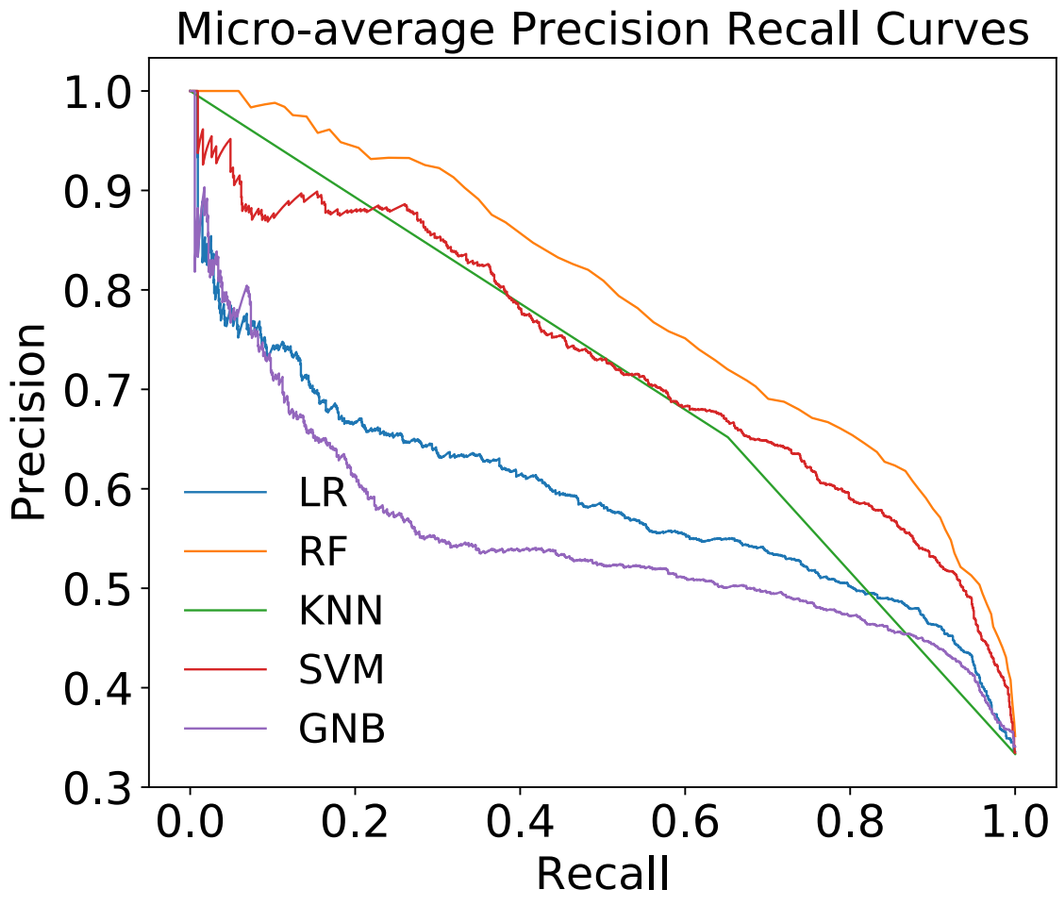


APPROACH

- Type: supervised learning
- Classification: Multi-class
- Data assumption: all samples measurements are independent from others
- Tool: scikit-learn, a machine learning module in Python
- Learning algorithms:
  - Logistic regression
  - Random forest
  - K-nearest neighbors
  - Support vector machine
  - Gaussian naive Bayes

COMPARING MICRO-AVERAGE SCORES





- Random Forest model has highest performance in making accurate predictions among other models. Precision recall curves shows that the model has significantly superior accuracy.

PREDICTIVE MODELING

- Provides an alternative tool for stakeholders in wine business to conduct wine rating without the limitation of professional sensory assessors.
- Can be used to sort wine sample into several tiers:
  - Top tier (~22%)
  - Mid tier (~45%)
  - Bottom tier (~33%)

RECOMMENDATION FOR IMPROVEMENT

- Use dataset from other designated origin other than Vinho Verde.
- Include additional features of different production methods like malolactic fermentation and time of barrel aging.
- Include raw data of sensory assessment to compare misclassified data points and score variation among assessors.