Predicting Wine Quality With XGBoost and SVM

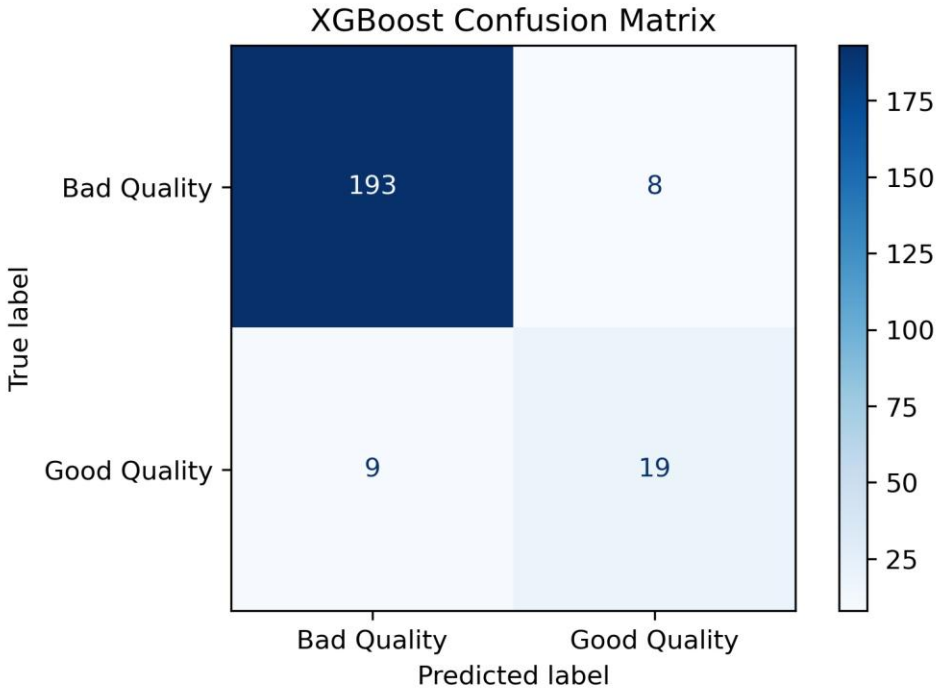
Maxwell Snodgrass

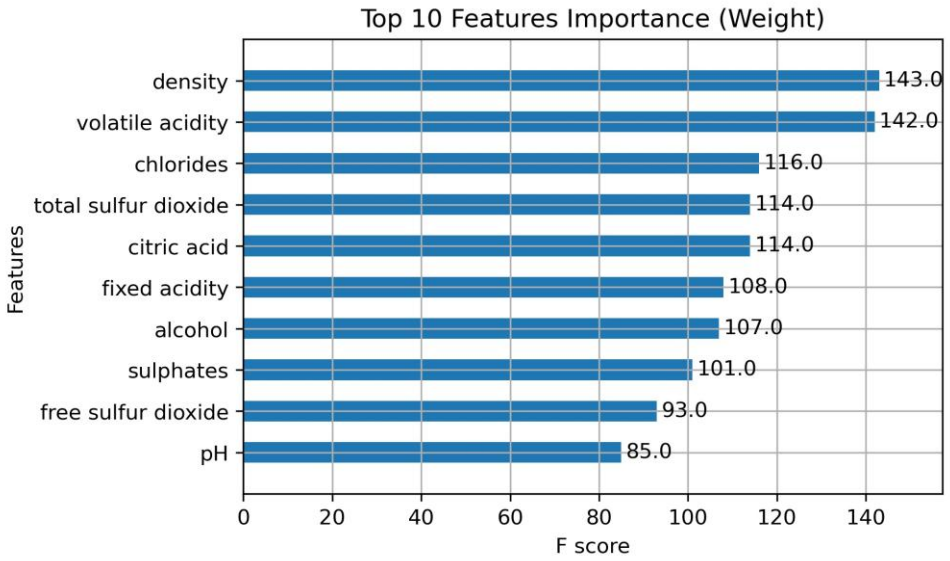
Data Cleaning and Preparation

- 1. Convert text file to CSV
- 2. Check for missing values
- 3. Check predictor variable data types
  - All are continuous
- 4. Feature engineering
  - Want binary classification: “good” vs. “bad”
    - Convert quality ratings of 7-8 to 1 and 3-6 to 0
  - Standardize all variables
- 5. Remove unnecessary columns
  - Observation Id
  - Raw quality variable

XGBoost Results

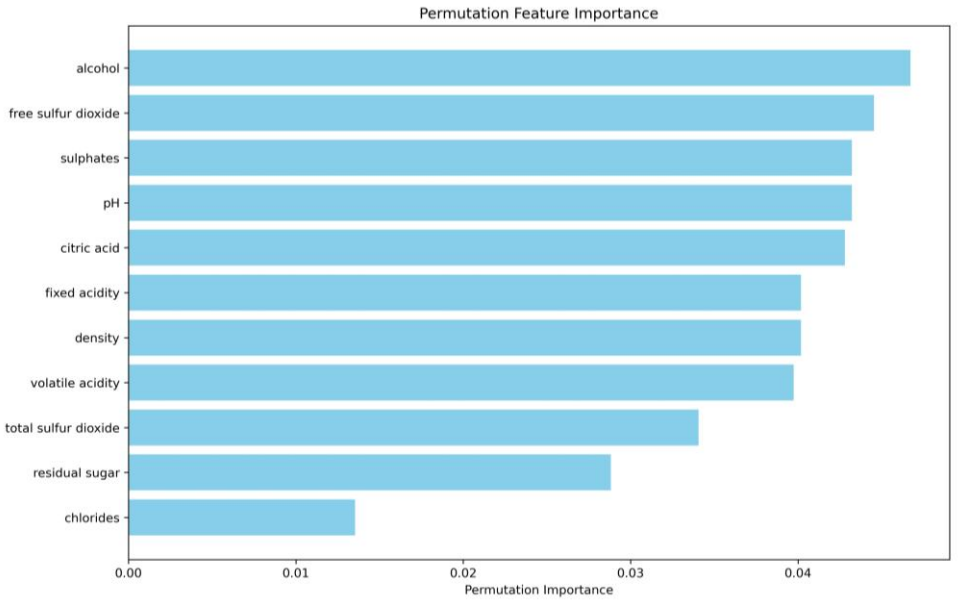
- Accuracy: **93%**
- Precision of “Good” Quality: **70%**
  - Of the wines that were predicted “good,” 70% were actually “good”
- Recall of “Good” Quality: **68%**
  - Of the wines that were actually of “good” quality, the model predicted 68% of them as “good”
- Feature Importance Method: F-Score
  - How many times the feature is used for a split.
  - Top 2 Features: **Density** and **Volatile Acidity**





SVM Results

- Accuracy: **95%**
- Precision of “Good” Quality: **90%**
  - Of the wines that were predicted “good,” 90% were actually “good”
- Recall of “Good” Quality: **64%**
  - Of the wines that were actually of “good” quality, the model predicted 64% of them as “good”
- Feature Importance Method: Permutation
  - How much the model is affected upon a shuffle.
  - Top 2 Features: **Alcohol** and **Free Sulfur Dioxide**



Model Comparisons

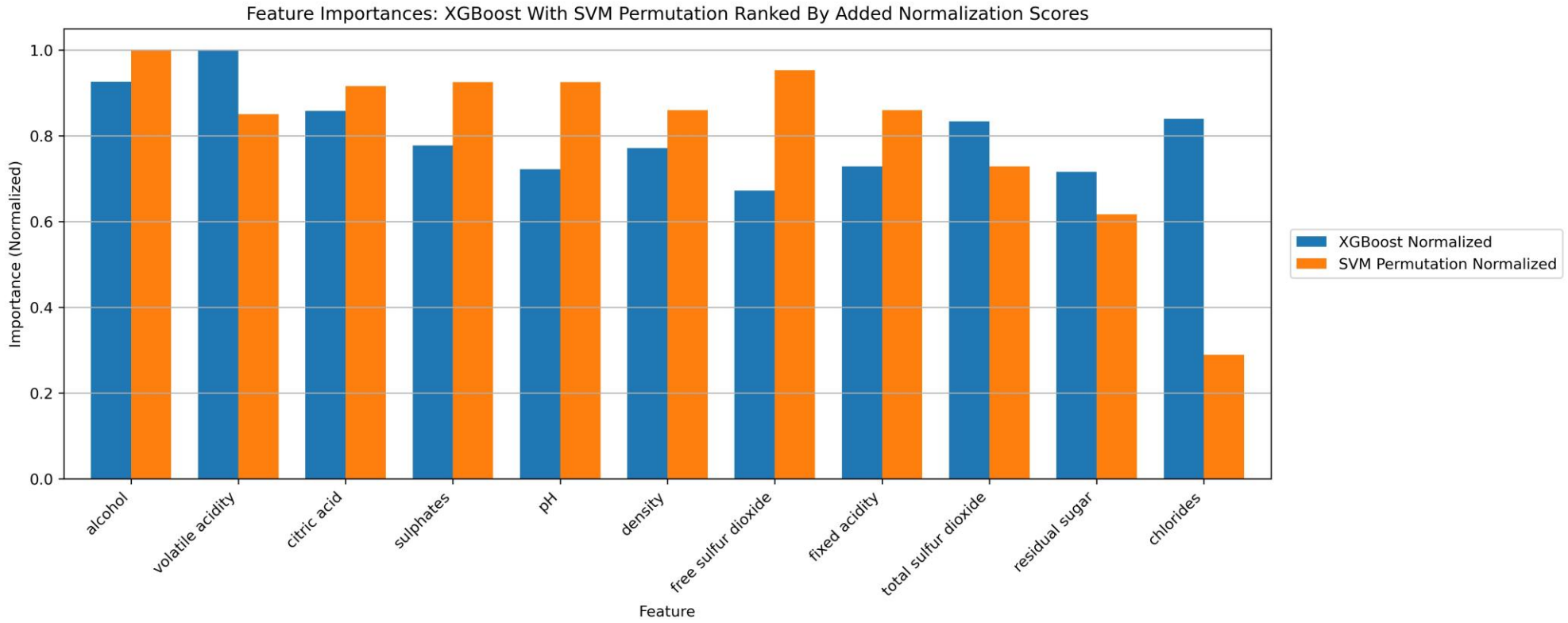
Metric	XGBoost	SVM
Model Accuracy	93%	95%
Precision of “Good” Quality	70%	90%
Recall of “Good” Quality	68%	64%
Feature Importance Method	F-Score	Permutation
Top 2 Important Features	Density and Volatile Acidity	Alcohol and Free Sulfur Dioxide

- The SVM performs better overall
- XGBoost tends to be more balanced in its predictions
- SVM tends to be stricter in predicting “good” quality
  - Hence a higher precision
  - At a very slight cost to recall (correctly predicting actual “good” quality wines)
- Feature Importances are quite different

Aggregated Feature Importance

- Normalize the feature importances of both models
  - This makes the feature importances of both directly comparable and keeps relativity to their own model
- Each feature has a normalized F-score and permutation importance
- Add each normalized score and create a ranking

Aggregated Feature Importance



- From the aggregated feature importance, the top three features of a “good” quality wine are **alcohol, volatile acidity, and citric acid**
- A correct combination of these three features plays the most important role in creating “good” wine