WINE UP!

Developing a model to predict wine quality

Jadav Bucktowar
Rafael Nazareno
Taimoor Khan

Problem Statement

Are we able to predict red and white wine quality based on its physicochemical characteristics?

Dataset Description

The dataset used from the UCI website for Wine Quality depicts quality on a number of physicochemical properties - which is split into 11 attributes listed below

- **Fixed Acidity**
- Volatile Acidity
- **Citric Acid**
- Residual sugar
- Chlorides
- **Free sulfur dioxide**
- **Total sulfur dioxide**
- Density
- pH
- Sulphates
- Alcohol

The target variable in this case is the quality score for each wine - which is a numerical value between 0 and 10.

Data Preparation

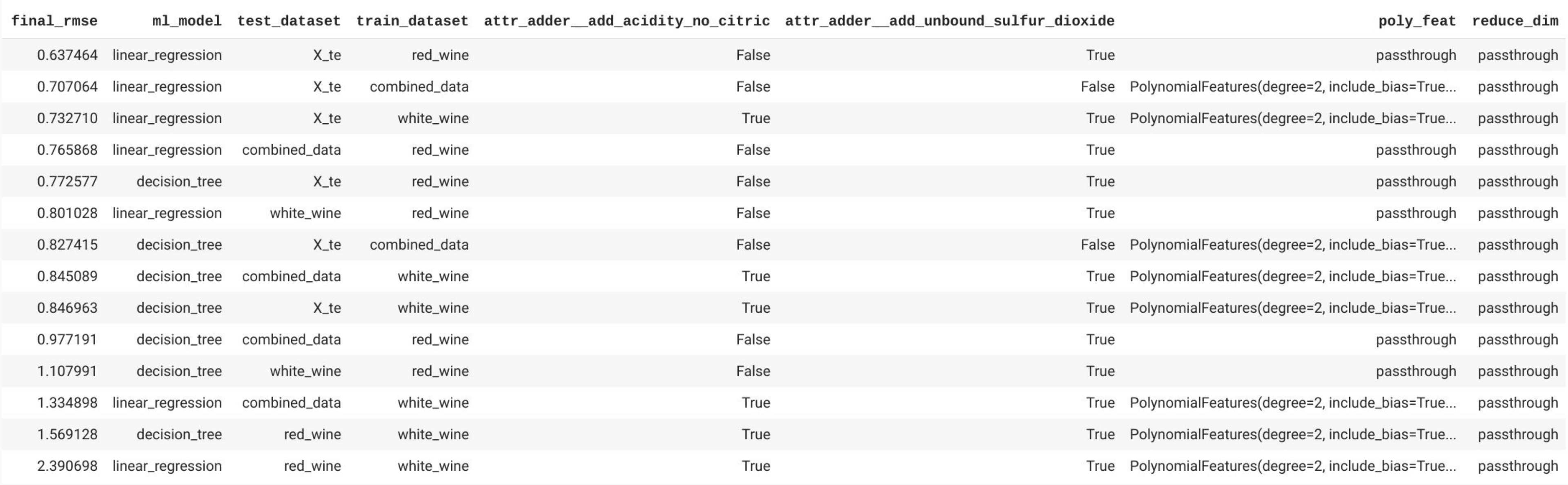
- Fixed acidity and citric acid
    - UC Davis article, mentions Fixed Acidity already contains citric acid concentration
    - New column **acidity_no_citric**
- Free sulfur dioxide and total sulfur dioxide
    - Iowa State University article, total sulfur dioxide includes free sulfur (SO_2) concentration already
    - New column **unbound_sulfur_dioxide**

Model Selection

- GridSearchCV
  - 'attr_adder__add_acidity_no_citric': [True, False]
  - 'attr_adder__add_unbound_sulfur_dioxide': [True, False]
  - 'reduce_dim': ['passthrough', PCA(n_components=0.90, random_state=42), PCA(n_components=0.95, random_state=42)]
  - 'poly_feat': ['passthrough', PolynomialFeatures(degree=2), PolynomialFeatures(degree=3)]

- Machine learning models chosen
  - Linear Regression
  - Decision Tree Regressor

Results



final_rmse	ml_model	test_dataset	train_dataset	attr_adder__add_acidity_no_citric	attr_adder__add_unbound_sulfur_dioxide	poly_feat	reduce_dim
0.637464	linear_regression	X_te	red_wine	False	True	passthrough	passthrough
0.707064	linear_regression	X_te	combined_data	False	False	PolynomialFeatures(degree=2, include_bias=True...	passthrough
0.732710	linear_regression	X_te	white_wine	True	True	PolynomialFeatures(degree=2, include_bias=True...	passthrough
0.765868	linear_regression	combined_data	red_wine	False	True	passthrough	passthrough
0.772577	decision_tree	X_te	red_wine	False	True	passthrough	passthrough
0.801028	linear_regression	white_wine	red_wine	False	True	passthrough	passthrough
0.827415	decision_tree	X_te	combined_data	False	False	PolynomialFeatures(degree=2, include_bias=True...	passthrough
0.845089	decision_tree	combined_data	white_wine	True	True	PolynomialFeatures(degree=2, include_bias=True...	passthrough
0.846963	decision_tree	X_te	white_wine	True	True	PolynomialFeatures(degree=2, include_bias=True...	passthrough
0.977191	decision_tree	combined_data	red_wine	False	True	passthrough	passthrough
1.107991	decision_tree	white_wine	red_wine	False	True	passthrough	passthrough
1.334898	linear_regression	combined_data	white_wine	True	True	PolynomialFeatures(degree=2, include_bias=True...	passthrough
1.569128	decision_tree	red_wine	white_wine	True	True	PolynomialFeatures(degree=2, include_bias=True...	passthrough
2.390698	linear_regression	red_wine	white_wine	True	True	PolynomialFeatures(degree=2, include_bias=True...	passthrough

Conclusion

- Dimensionality reduction has no effect
- When trained against the red wine dataset, linear regression produces better results than a polynomial function
- Training on single dataset and testing on combined dataset
    - Difference in # instances
- Highlights importance of having more data
    - More variance
    - More stratification
    - White wine points (~5000) vs red wine points (~1500)