Final Project

Predicting Wine Quality

Chris Coker, Isaac Graham, Kate Kollman

Outline

Project Motivation:

- Wine quality is assessed using expert taste panels & chemical tests
- Expert reviews are subjective & expensive
- Chemical tests are objective but lack direct quality ratings

Research Objective: Use machine learning to link chemical properties to expert quality ratings

Analysis Plan:

- Exploratory Data Analysis (EDA)
- Baseline Model: Multinomial Logistic Regression
- Nonlinear Model: XGBoost

Problem Statement: Can physicochemical properties (X) and wine type predict expert-assigned wine quality scores(Y)?

Data Description

Target Variable (Y)

- Wine quality score (0–10), assigned by expert tasters

Predictor Variables (X)

- 11 physicochemical attributes (e.g., acidity, sugar, sulphates, alcohol)
- Wine type indicator (red or white)

Data Source

- Wine Quality Dataset (UCI Machine Learning Repository)
- Based on Cortez et al. (2009), Vinho Verde region

Preprocessing

- Median expert score used
- Measurement errors removed
- Red and white datasets merged
- Wine type added as binary variable
- No missing values in the final dataset; obvious outliers removed by original authors

Context & Limitations

- Data collected in 2009
- Single geographic region
- Predictors are continuous; quality is categorical

Hypothesis

Hypothesis 1: Alcohol & Quality

- Alcohol content is the main positive driver of wine quality

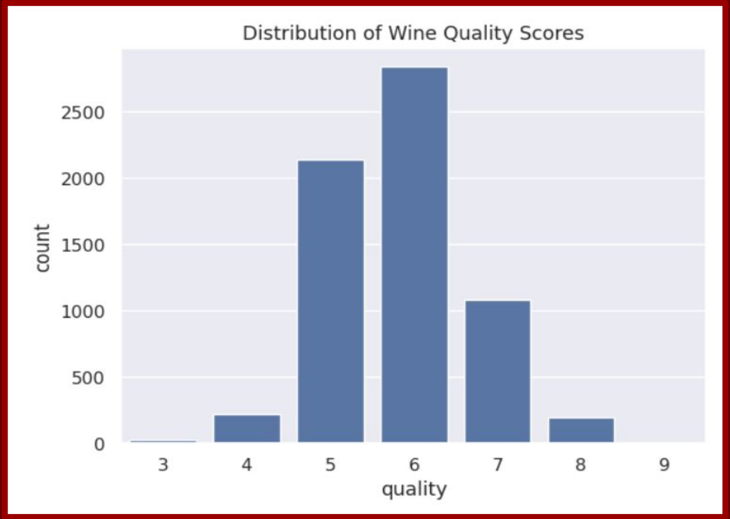
Hypothesis 2: Model Performance

- Nonlinear models (XGBoost) outperform logistic regression models
- Expected improvements in log loss score and F1 score

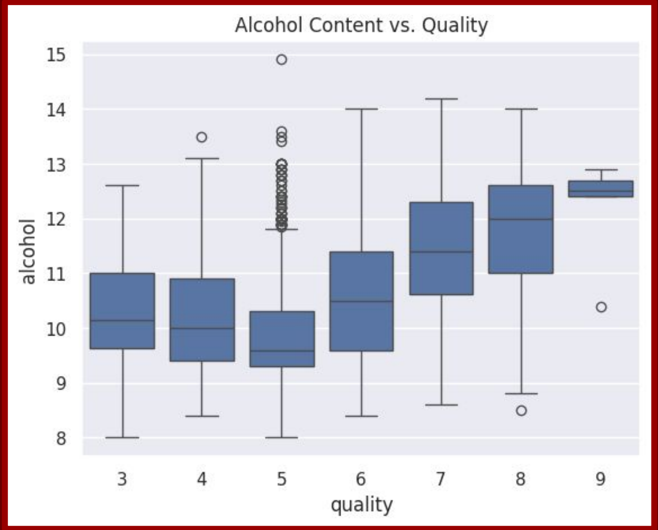
Hypothesis Testing

- Coefficients & feature importance assess driver significance
- Performance metrics compare predictive accuracy

Exploratory Data Analysis (EDA): Variable Distributions



- Wine quality scores are concentrated in the mid range (5-7), with a few extreme values



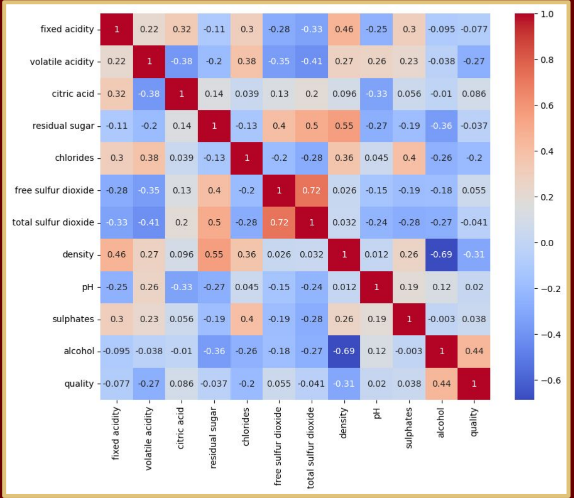
- Wine quality score increases as alcohol content increases

Exploratory Data Analysis (EDA): Correlations & Summary Statistics

```
wine_data.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	wine_type
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	115.744574	0.994697	3.218501	0.531268	10.491801	5.818378	0.246114
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	56.521855	0.002999	0.160787	0.148806	1.192712	0.873255	0.430779
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	0.220000	8.000000	3.000000	0.000000
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	0.430000	9.500000	5.000000	0.000000
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	0.510000	10.300000	6.000000	0.000000
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	0.600000	11.300000	6.000000	0.000000
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.010000	2.000000	14.900000	9.000000	1.000000

- Quality’s strongest positive correlation is with alcohol content
- Density is negatively correlated with alcohol & quality
- Correlations indicate potential nonlinear relationships among variables



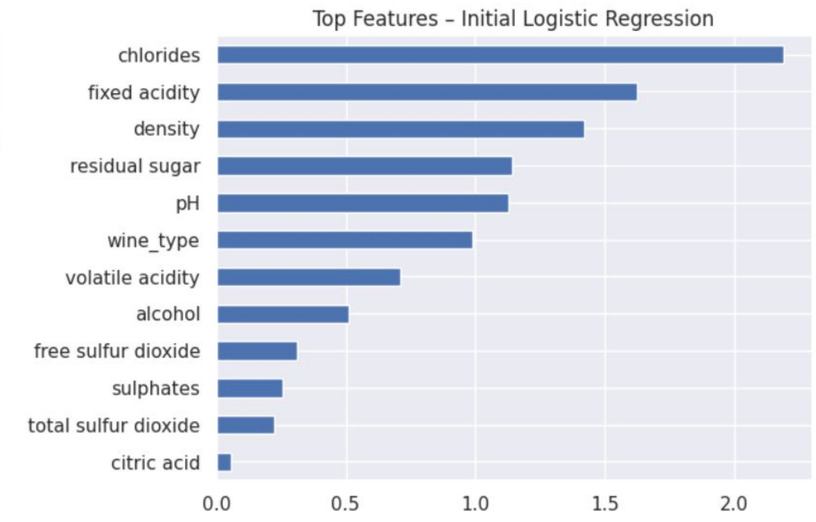
Multiclass Logistic Regression Model (Baseline)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	8
1	1.00	0.04	0.07	54
2	0.57	0.54	0.56	535
3	0.51	0.71	0.60	709
4	0.58	0.28	0.38	270
5	0.00	0.00	0.00	48
6	0.00	0.00	0.00	1
accuracy			0.54	1625
macro avg	0.38	0.22	0.23	1625
weighted avg	0.54	0.54	0.51	1625

```
Test Log Loss: 1.0832674714204464
Training Log Loss: 1.054018116931027
```

Performance:

- Accuracy ~54%
- Weighted Average f1-score: 0.51
- Test log loss ~1.08



Class imbalance observed: Poor performance on rare quality classes

Linear model captures general trends but struggles with minority classes

Logistic Regression with Hyperparameter Tuning (Ridge)

```
LogisticRegression
LogisticRegression(C=np.float64(1.0), max_iter=1000, multi_class='multinomial')
```

- Regularization Applied: Ridge (L2), tuned via C parameter
- Applied to reduce multicollinearity & improve coefficient stability
- Best C value: 1.0

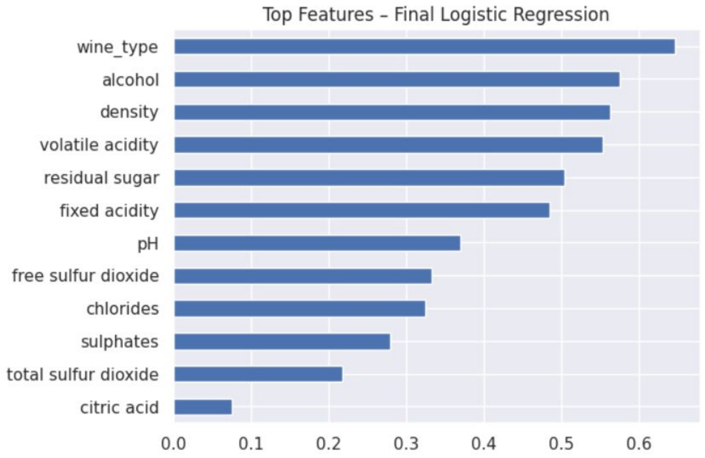
Final Logistic Regression – Classification Report

	precision	recall	f1-score	support
3	0.00	0.00	0.00	8
4	1.00	0.04	0.07	54
5	0.57	0.54	0.56	535
6	0.51	0.71	0.60	709
7	0.57	0.28	0.37	270
8	0.00	0.00	0.00	48
9	0.00	0.00	0.00	1
accuracy			0.54	1625
macro avg	0.38	0.22	0.23	1625
weighted avg	0.54	0.54	0.51	1625

Final Logistic Regression – Test Log Loss: 1.0710126051952216
Final Logistic Regression – Training Log Loss: 1.0559336686100274

Performance:

- Accuracy ~54%
- Weighted Average f1-score: 0.51
- Test Log Loss ~1.07



Regularization improved stability but did not significantly improve predictive performance

XGBoost Model: Gradient boosted decision trees

Why? Handles feature interactions automatically

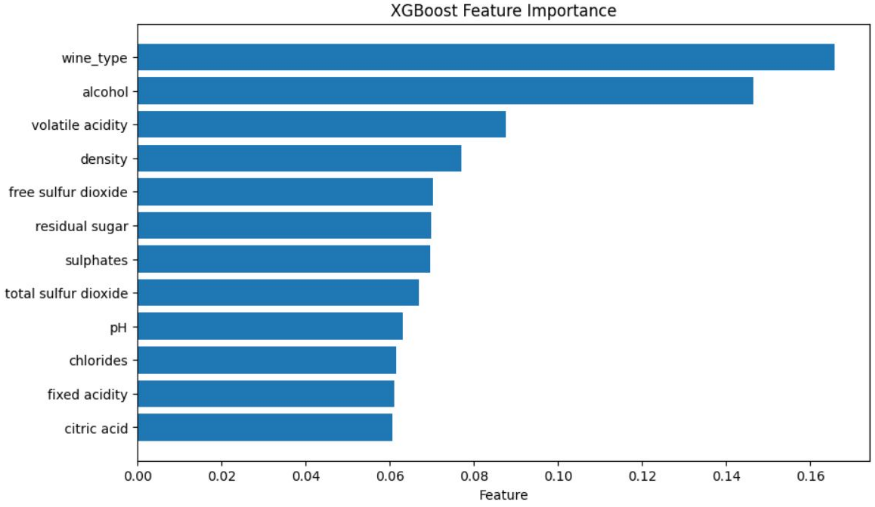
```
print(classification_report(y_test, y_test_pred_xgb))
```

	precision	recall	f1-score	support
3	0.00	0.00	0.00	8
4	0.53	0.19	0.27	54
5	0.71	0.69	0.70	535
6	0.66	0.76	0.71	709
7	0.68	0.62	0.65	270
8	0.94	0.35	0.52	48
9	0.00	0.00	0.00	1
accuracy			0.68	1625
macro avg	0.50	0.37	0.41	1625
weighted avg	0.68	0.68	0.67	1625

Training Log Loss: 0.1773491250523672
Test Log Loss: 0.8476340262813755

Performance
- Accuracy ~68%
- Weighted Average f1-score: 0.67
- Test Log Loss ~0.85

- Lower performance on rare classes due to limited data
- More flexible model significantly improves prediction accuracy



Conclusion

Final Model Decision: XGBoost was the best-performing model, showing superior accuracy, weighted F1, and log loss compared to logistic regression

Hypotheses Answered: Our results disproved the alcohol-content prediction hypothesis & showed non-linear models better capture wine-quality relationships

Limitations: Limited extreme-quality data, lack of interaction terms in logistic regression, and multicollinearity constrained interpretation

Future Research: Incorporating SHAP, adding interaction terms, expanding features, and using more diverse regional datasets would strengthen conclusions