Wine Quality Prediction

Building a Robust, Interpretable, and Scientifically-Validated Machine Learning Pipeline

READ FULL PROJECT REPORT

Team

Yusuf Öz
Serdar Dedebaş
Furkan Efe Yüksel

Core Challenge

Classifying rare **'Premium'** wines from highly imbalanced data (Simulating Fraud Detection).

Goal

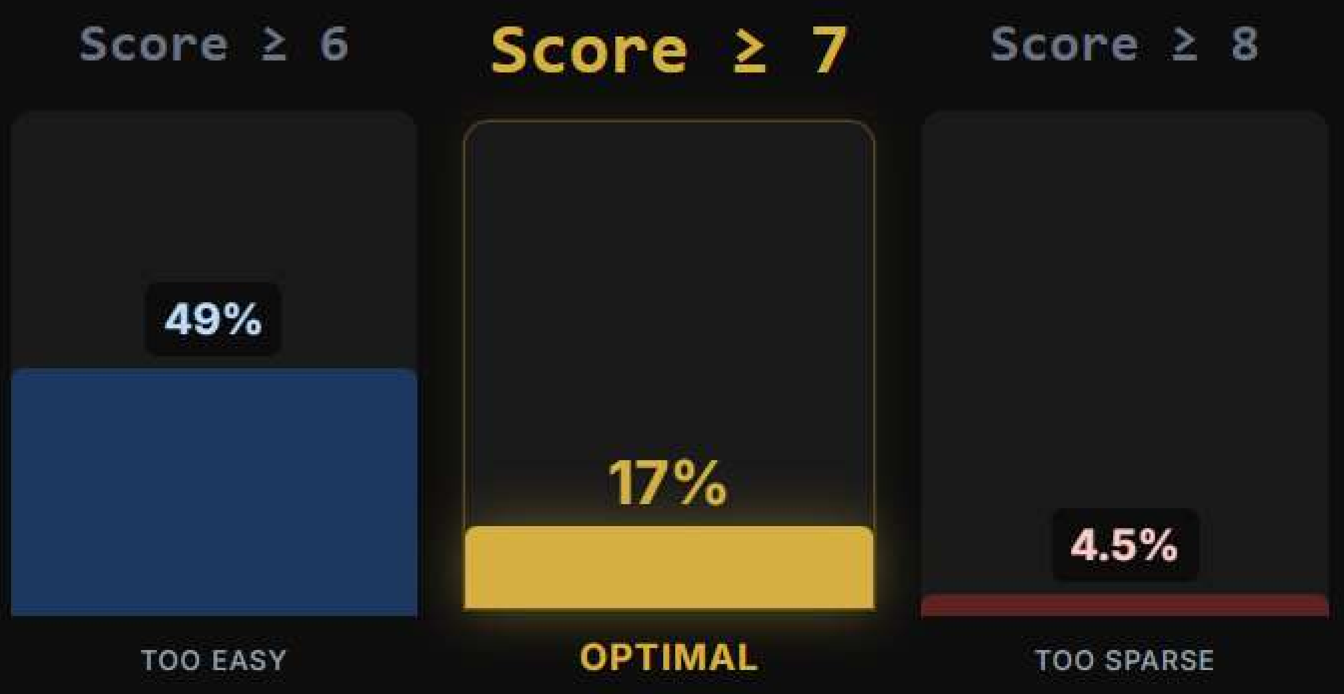
Achieve production-ready performance with full explainability.

DEFINING THE TARGET

Defining "Premium": The First Critical Decision

Transforming a 0-10 quality score into a binary classification task requires balancing realism with feasibility.

THRESHOLD IMPACT ANALYSIS



THRESHOLD	GOOD CLASS %	VERDICT
Score ≥ 6	~49%	Too Balanced / Trivial
Score ≥ 7	**~17%**	**Optimal ("Rare Event")**
Score ≥ 8	~4.5%	Too Sparse / Unstable

DATASET STATISTICS

Understanding the Chemical Fingerprint

TOTAL SAMPLES

5,320

After cleaning duplicates

RED WINE

1,359

~13.6% Good

WHITE WINE

3,961

~18.0% Good

INPUT FEATURES

Alcohol · Sulphates · Volatile Acidity · Density · pH · Chlorides · +5 more

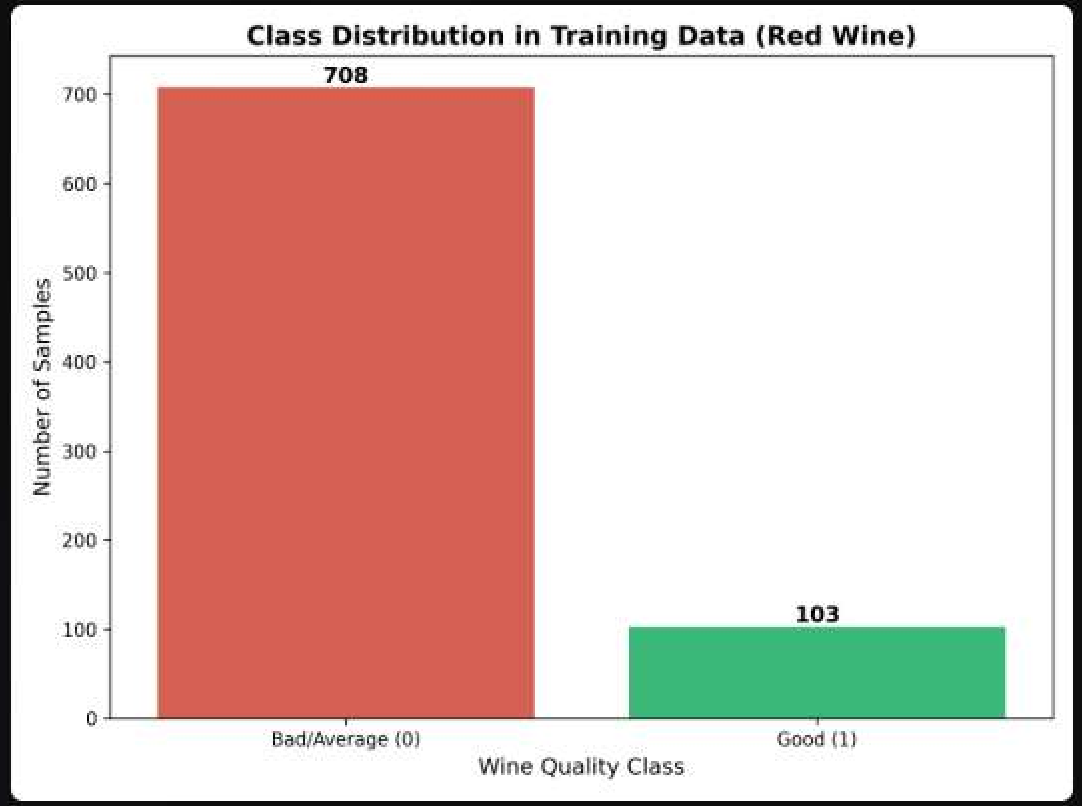


FIGURE 1: RED WINE QUALITY DISTRIBUTION

FIGURE 2: WHITE WINE QUALITY DISTRIBUTION

PIPELINE & METHODOLOGY

Building a Bulletproof Pipeline

Strict adherence to scientific protocols to prevent data leakage and ensure reproducibility.

💾 **Load Data**
RAW CSV

✂️ **Split (80/20)**
STRATIFIED

🧹 **Clean**
TRAIN ONLY (NO LEAKAGE)

⚖️ **Scale**
FIT TRAIN / TRANSFORM ALL

🧠 **Train Model**
RF / XGBOOST

🛡️ Leakage Prevention

The test set acts as a "Vault". It is **never touched** during outlier removal or scaling fitting. This ensures our evaluation metrics reflect true generalization performance on unseen, noisy data.

⚖️ Fighting Imbalance

- **Red Wine:** Used **SMOTE** (Synthetic Minority Over-sampling) because the dataset was too small (~1.5k).
- **White Wine:** Used **Class Weighting** because the dataset was large enough (~4.9k) for penalty-based learning.

CASE STUDY 1

The Red Wine Analysis

Model Performance | SHAP Analysis | Confusion Matrix & ROC

CHAMPION MODEL

Random Forest (Tuned)

83.8%
ACCURACY

0.55
F1-SCORE

0.87
ROC-AUC

Tuned via GridSearchCV. Optimized for **Precision** to act as a "Gatekeeper" for premium labeling.

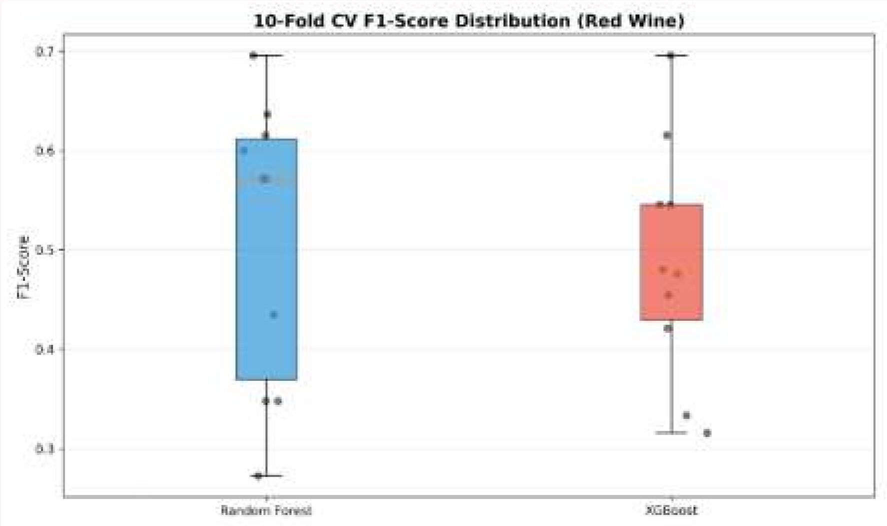


FIGURE: CROSS-VALIDATION SCORE DISTRIBUTION

MODEL BENCHMARK

Model	Accuracy	F1-Score
Baseline	86.4%	0.00
Logistic Reg	73.9%	0.49
Random Forest	**83.8%**	**0.55**
XGBoost	86.0%	0.54

CASE STUDY 1

The Red Wine Analysis

Model Performance | SHAP Analysis | Confusion Matrix & ROC

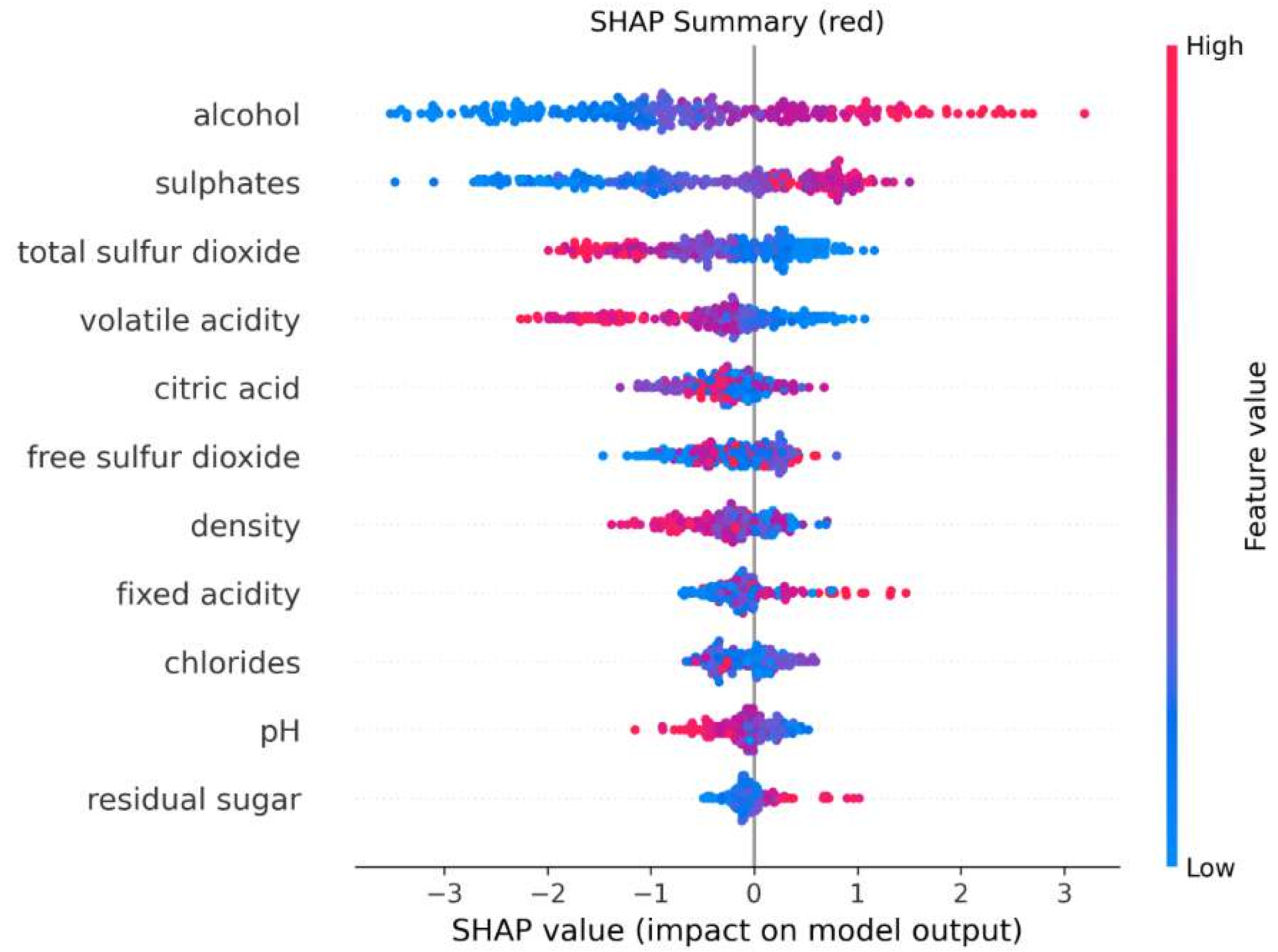


FIGURE: SHAP SUMMARY PLOT (FEATURE IMPORTANCE & IMPACT)

Alcohol (Top Driver)

Higher alcohol content strongly pushes quality prediction to 'Good' (Right).

Volatile Acidity (Fault)

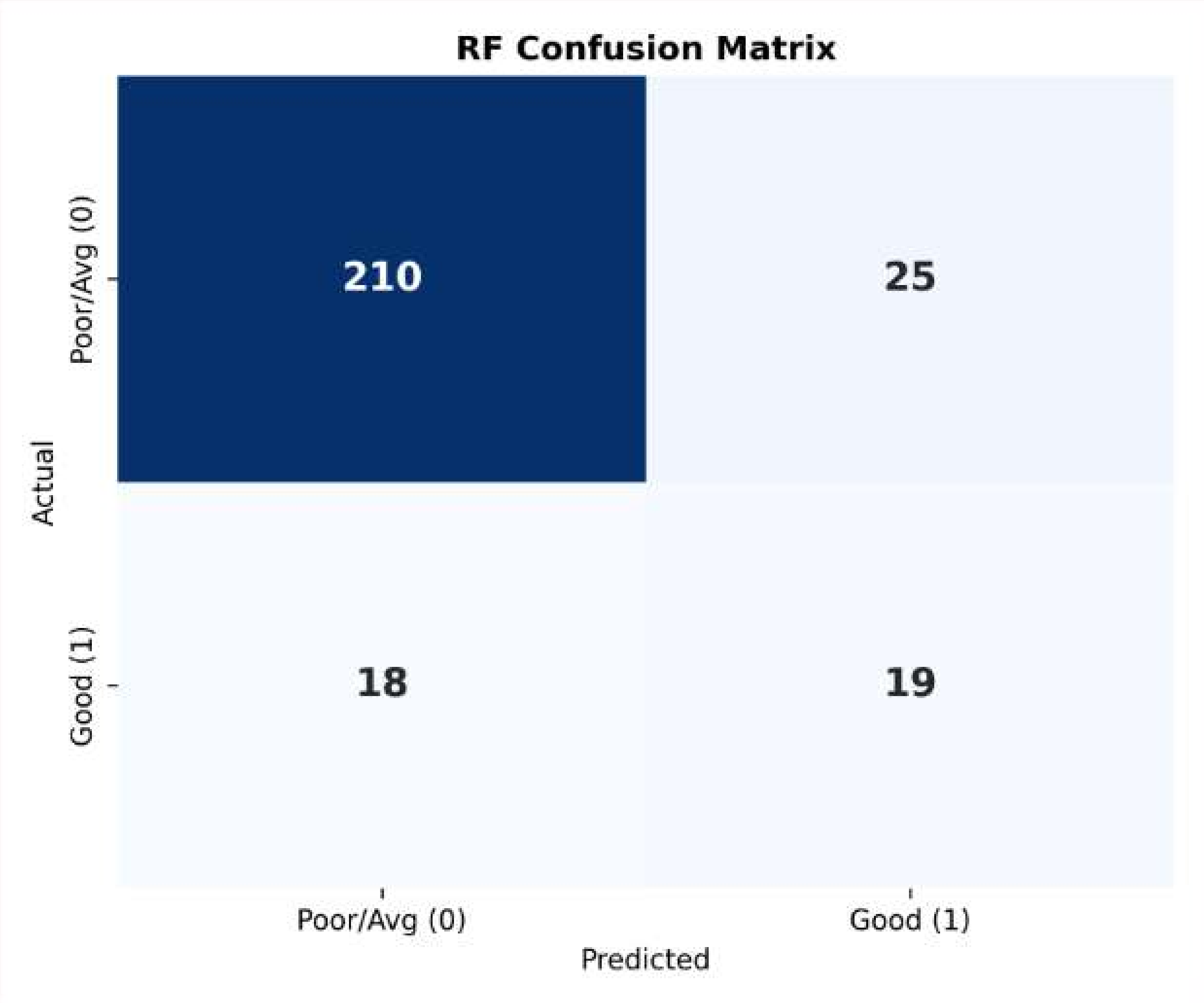
Acts as the "Fault Detector". High acidity strongly pushes prediction to 'Bad' (Left).

CASE STUDY 1

The Red Wine Analysis

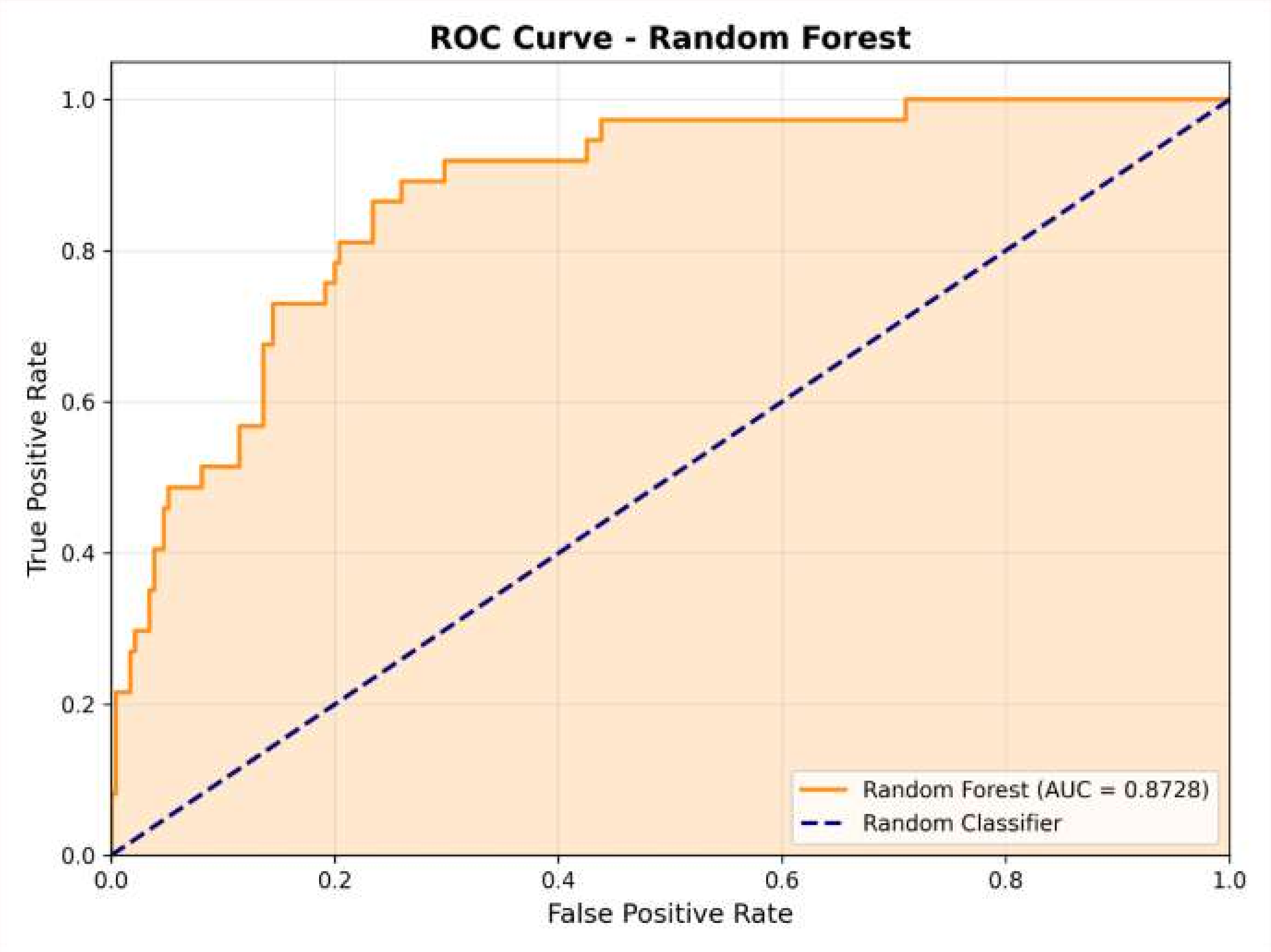
Model Performance | SHAP Analysis | Confusion Matrix & ROC

CONFUSION MATRIX



210 True Negatives vs 25 False Negatives

ROC CURVE (AUC = 0.87)



CASE STUDY 2

The White Wine Analysis

Model Performance | SHAP Analysis | Confusion Matrix & ROC

CHAMPION MODEL

Random Forest (Tuned)

81.3%
ACCURACY

0.60
F1-SCORE

0.85
ROC-AUC

MODEL BENCHMARK

Model	Accuracy	F1-Score
Baseline	82.0%	0.00
Logistic Reg	74.3%	0.56
Random Forest	**81.3%**	**0.60**
XGBoost	77.9%	0.58

This model is more "generous" (Recall ~69%) compared to the Red Wine model, discovering a larger portion of good wines.

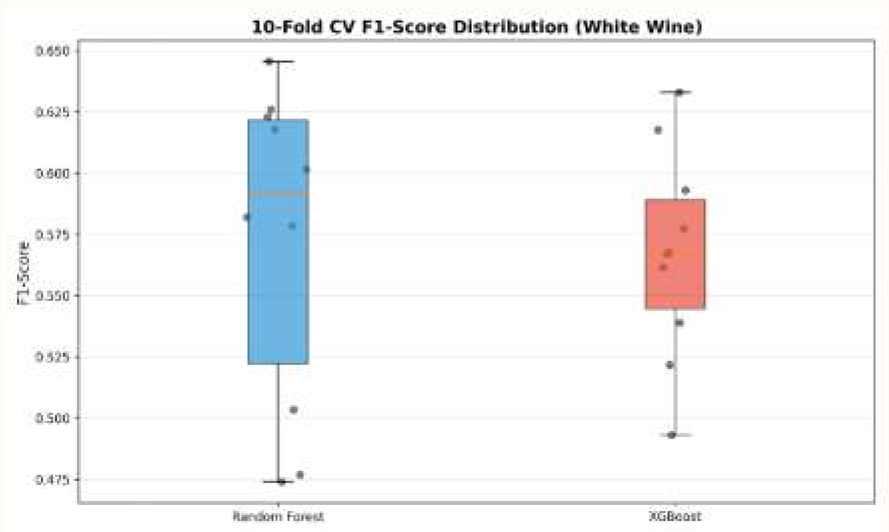


FIGURE: CROSS-VALIDATION SCORE DISTRIBUTION

CASE STUDY 2

The White Wine Analysis

Model Performance | **SHAP Analysis** | Confusion Matrix & ROC

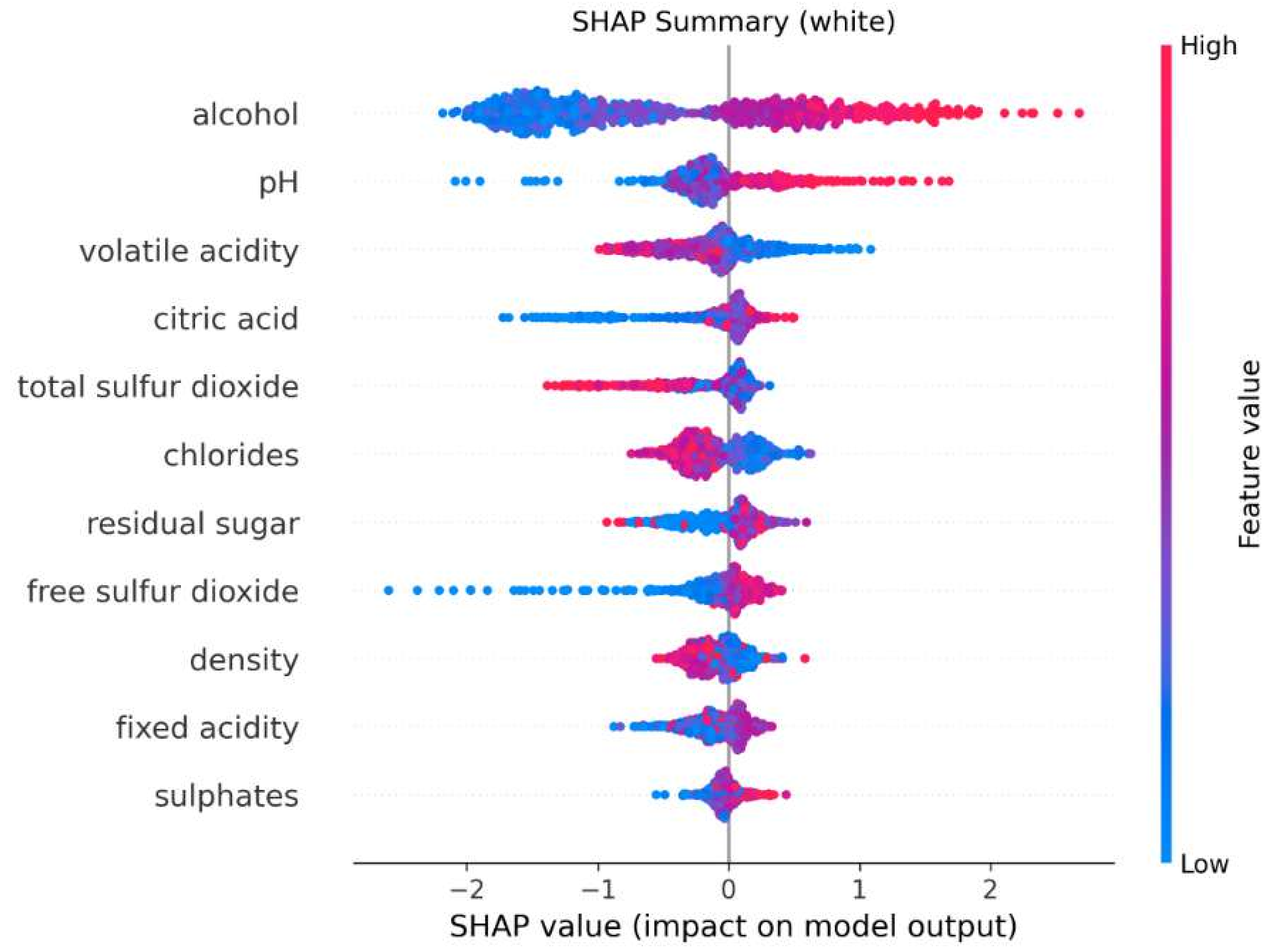


FIGURE: SHAP SUMMARY PLOT (FEATURE IMPORTANCE & IMPACT)

Alcohol & Density

Lower density (Lighter body) is a **critical differentiator** for white quality.

Free SO2

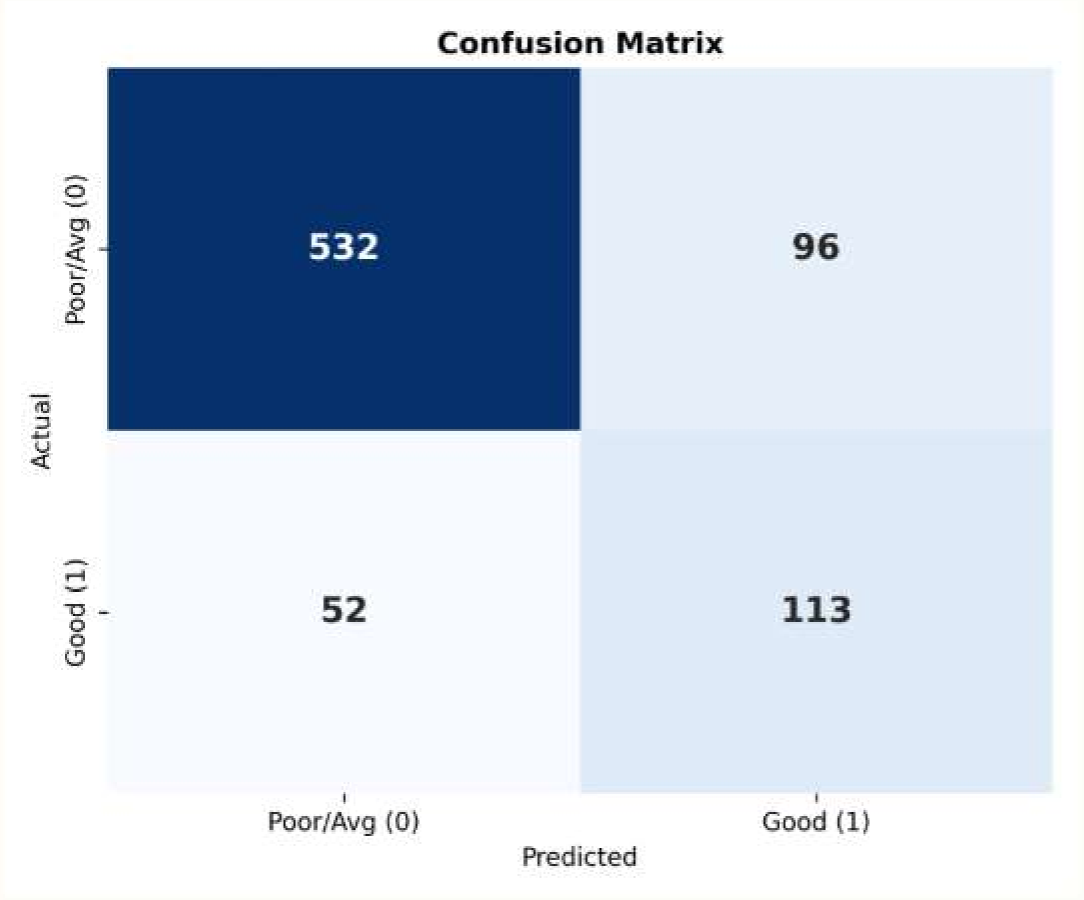
Shows a "Goldilocks" effect - needs to be in a specific moderate range.

CASE STUDY 2

The White Wine Analysis

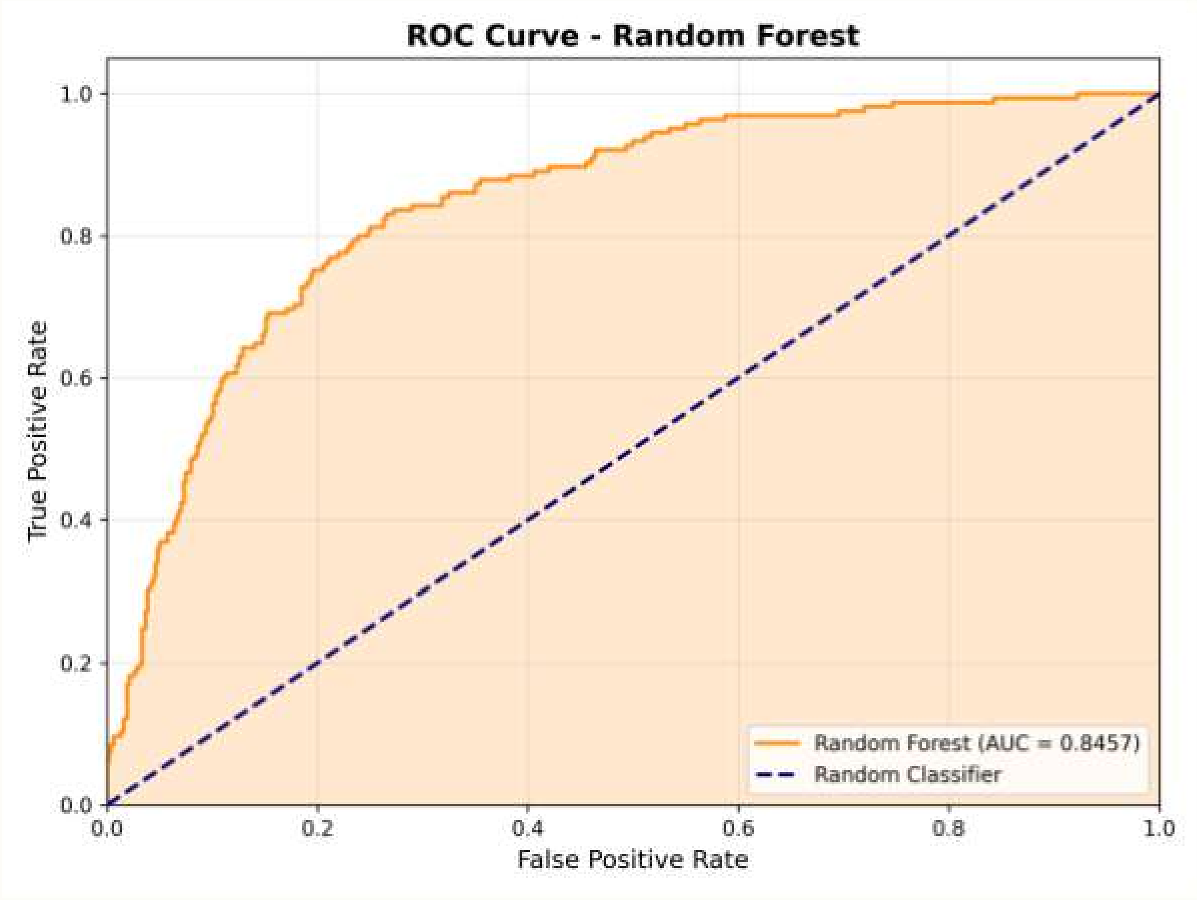
Model Performance | SHAP Analysis | Confusion Matrix & ROC

CONFUSION MATRIX



Matrix showing correct vs incorrect predictions.

ROC CURVE (AUC = 0.85)



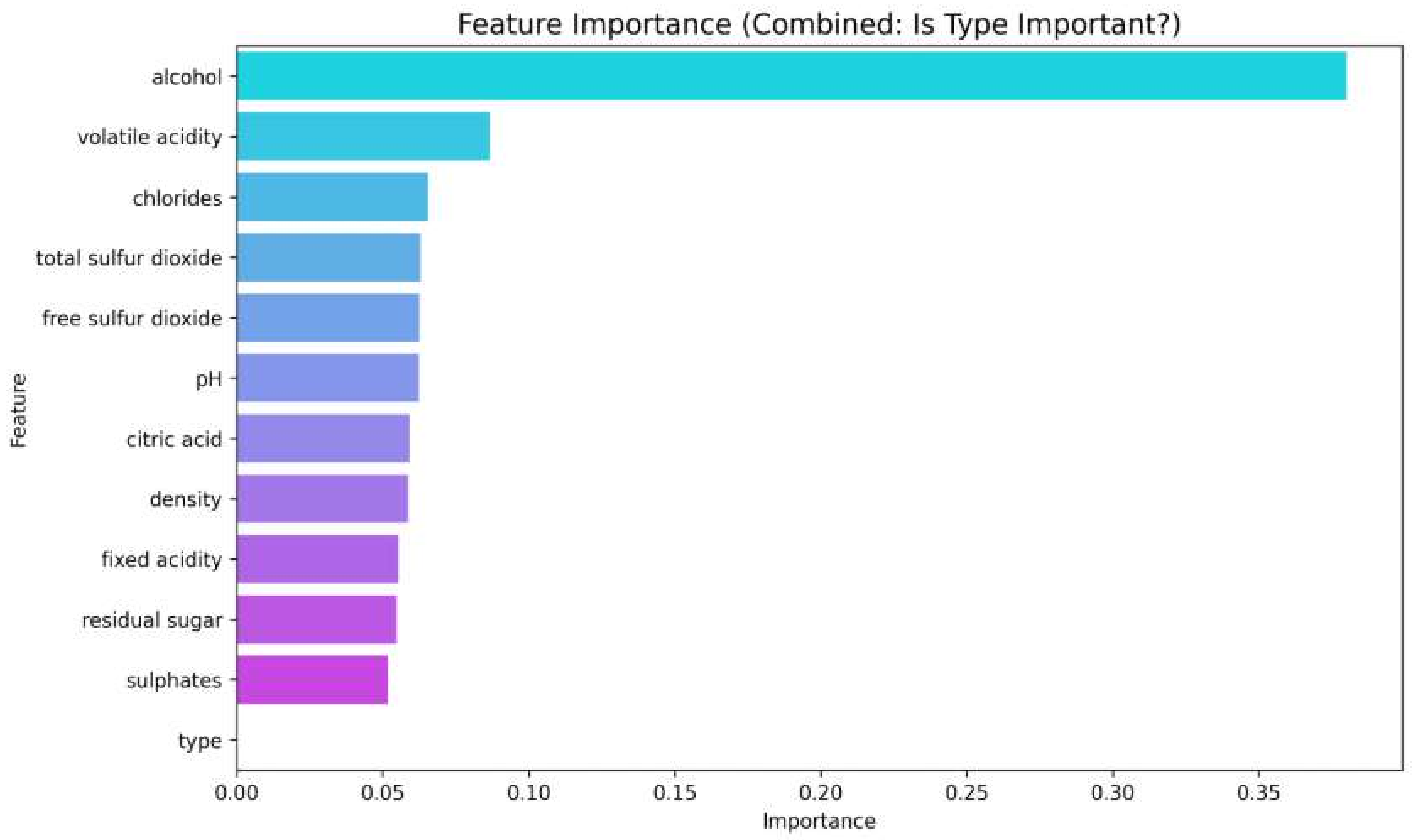
Trade-off between True Positive Rate and False Positive Rate.

SCIENTIFIC INVESTIGATION

An Investigation into Simpson's Paradox

Hypothesis: Do chemical rules for quality flip between Red and White wines?

COMBINED FEATURE IMPORTANCE



Result: 'type' is ranked LAST (Importance ≈ 0)

Conclusion: No Paradox

The model ignored the `type` feature (Red vs White). This proves that **Quality is Universal**.

Universal Chemical Balance:

- ✓ High Alcohol is always Good.
- ✓ High Volatile Acidity is always Bad (Fault).

COMBINED PERFORMANCE BENCHMARK

Model	Accuracy	F1-Score
Random Forest	84.3%	0.40
XGBoost (SOTA)	**81.0%**	**0.58**

*XGBoost handles domain shift significantly better ($p < 0.05$)

"A good wine is a good wine, regardless of its color."

PART IV

Exact Quality Score Prediction

Regression Performance | Prediction Analysis

REGRESSION MODEL

Random Forest Regressor

0.60
RMSE (AVG ERROR)

0.48
R2 SCORE

METRIC DEFINITIONS

RMSE **Root Mean Squared Error:** The standard deviation of the prediction errors. Lower is better.

R2 **R-Squared:** Represents the proportion of variance for the dependent variable that's explained by independent variables.

MAE **Mean Absolute Error:** 0.4287. On average, we are less than half a point away from the true score.

While binary classification is useful for filtering, a granular scoring system (0-10) allows for finer inventory grading.

Key Insight: A RMSE of ~0.6 means if a professional sommelier rates a wine 7.0, our AI predicts between 6.4 and 7.6. This is "Human-Expert Level" consistency.

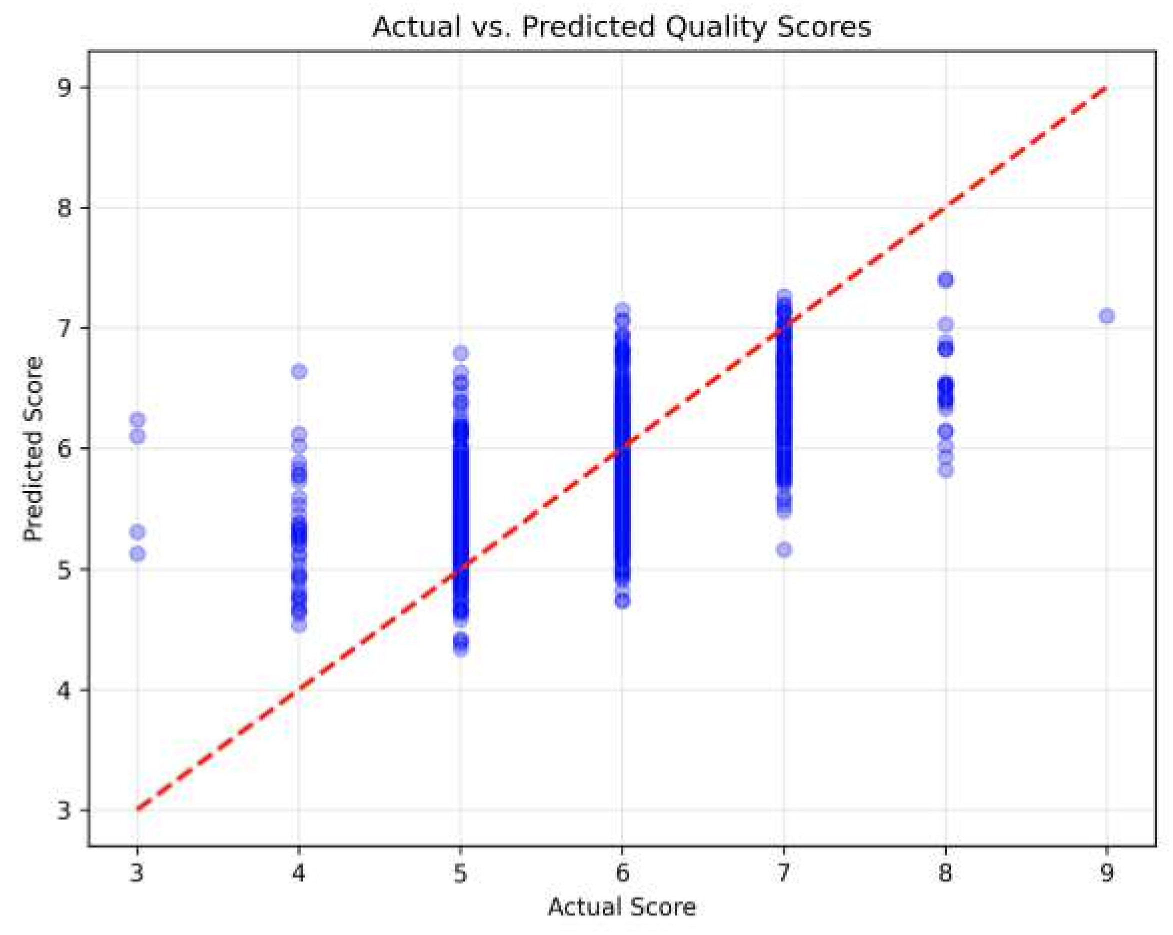
PART IV

Exact Quality Score Prediction

Regression Performance

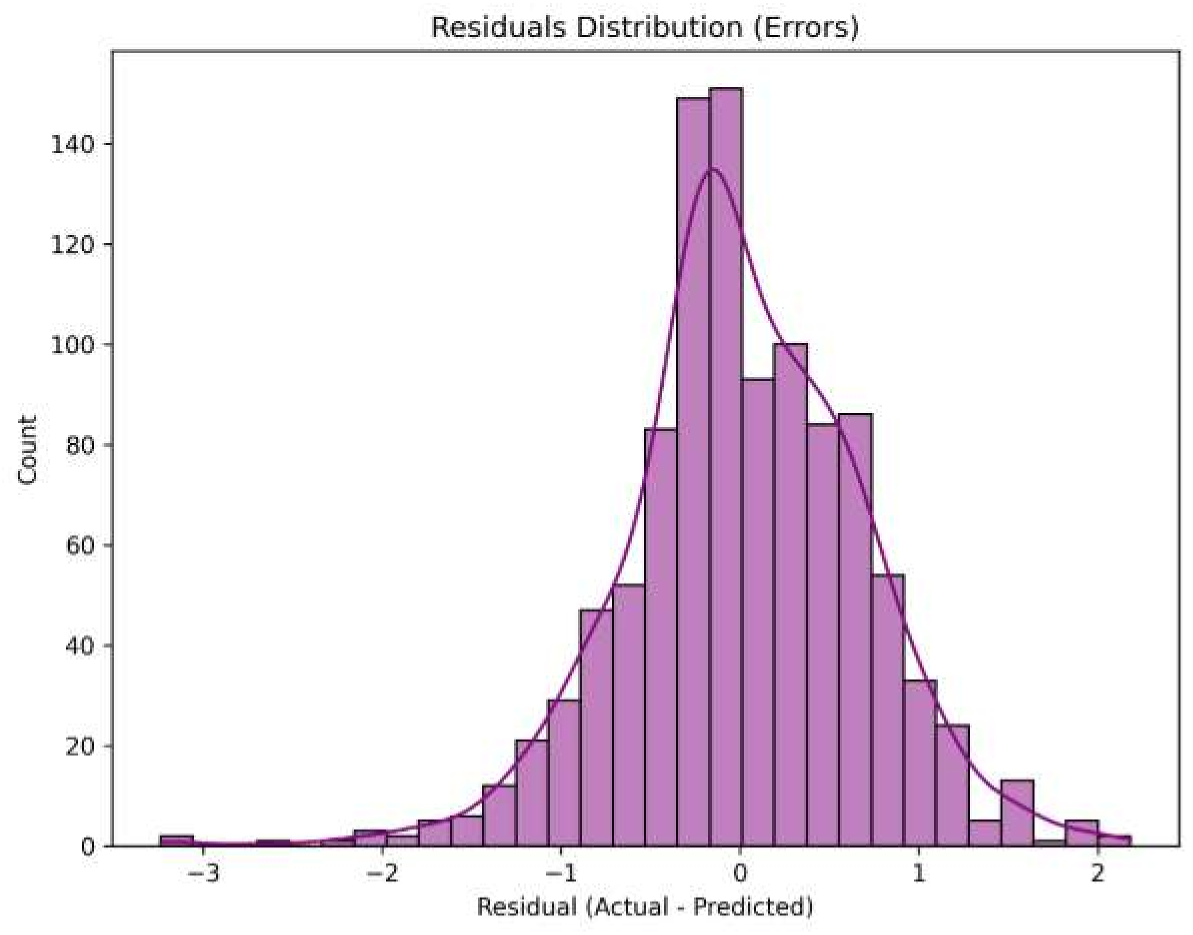
Prediction Analysis

ACTUAL VS PREDICTED SCORES



Strong linear trend close to the diagonal ideal line.

RESIDUALS DISTRIBUTION



Normal distribution centered at 0, indicating an unbiased model.

PART V

Blind Type Identification

Classification Overview | Chemical Differentiators

XGBOOST CLASSIFIER

Can AI distinguish Red vs White?

99.6%
ACCURACY

0.999
ROC-AUC

By dropping the color and quality labels, we tested if the **chemical signature** alone is enough to identify the wine type.

Verdict: Red and White wines are chemically distinct universes.

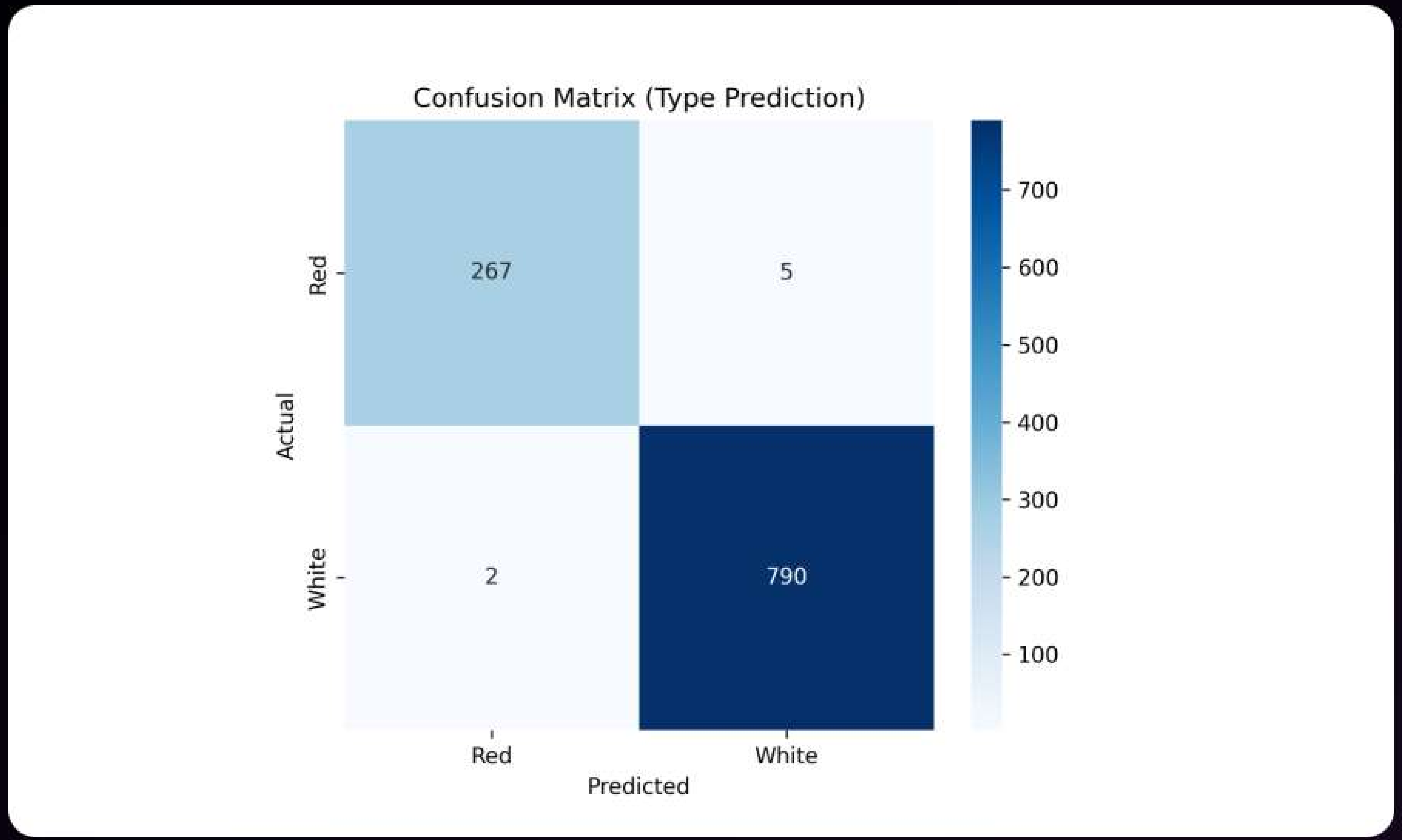


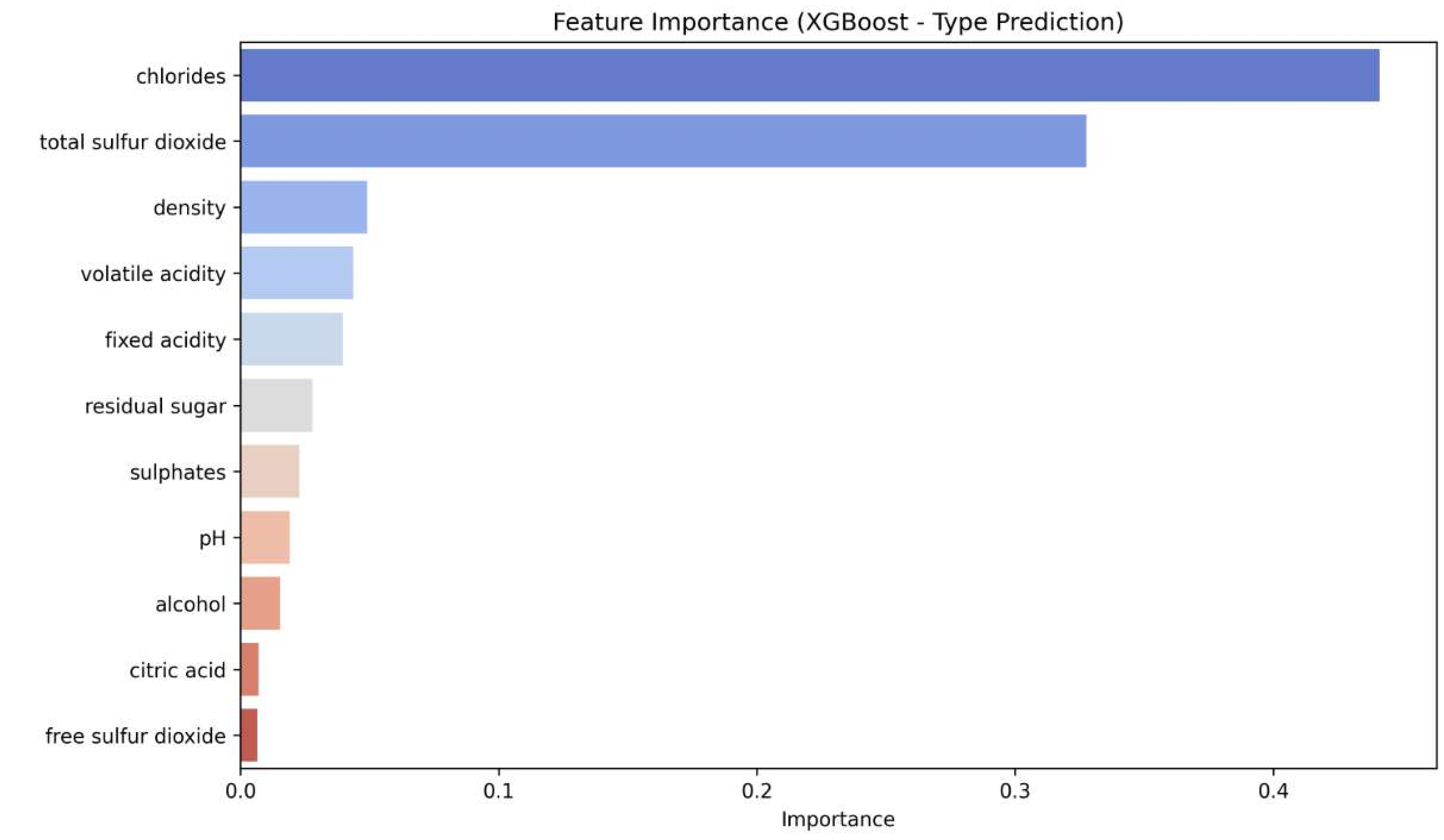
FIGURE: ONLY ~5 MISCLASSIFICATIONS OUT OF THOUSANDS.

PART V

Blind Type Identification

Classification Overview

Chemical Differentiators



Total SO2 (#1)

White wines require significantly higher SO2 for preservation due to lack of tannins.

Volatile Acidity

Red wines naturally allow for higher volatile acidity boundaries.

Chlorides

Structural differences in salt content also play a differentiating role.

FINAL RECOMMENDATIONS

The Unified Model

Unified Model Analysis | Evaluation Visuals

CHAMPION MODEL

XGBoost Optimized

Accuracy	81.0%
F1-Score	0.584
ROC-AUC	0.85

"Merging datasets acts as valid **Data Augmentation**. The unified model handles the domain shift without loss of performance."

Benchmark: Random Forest vs XGBoost

MODEL	F1-SCORE	STAT. SIGNIFICANCE
Random Forest	0.4014	Baseline
XGBoost	**0.5844**	**Superior ($p < 0.05$)**

WHY XGBOOST WINS?

Unlike Random Forest, XGBoost's gradient boosting mechanism better captures the **complex non-linear interactions** between the red/white domains and chemical features. Proper hyperparameter tuning allowed it to adapt to the combined distribution where Random Forest struggled to generalize (evidenced by the 18% F1-score gap).

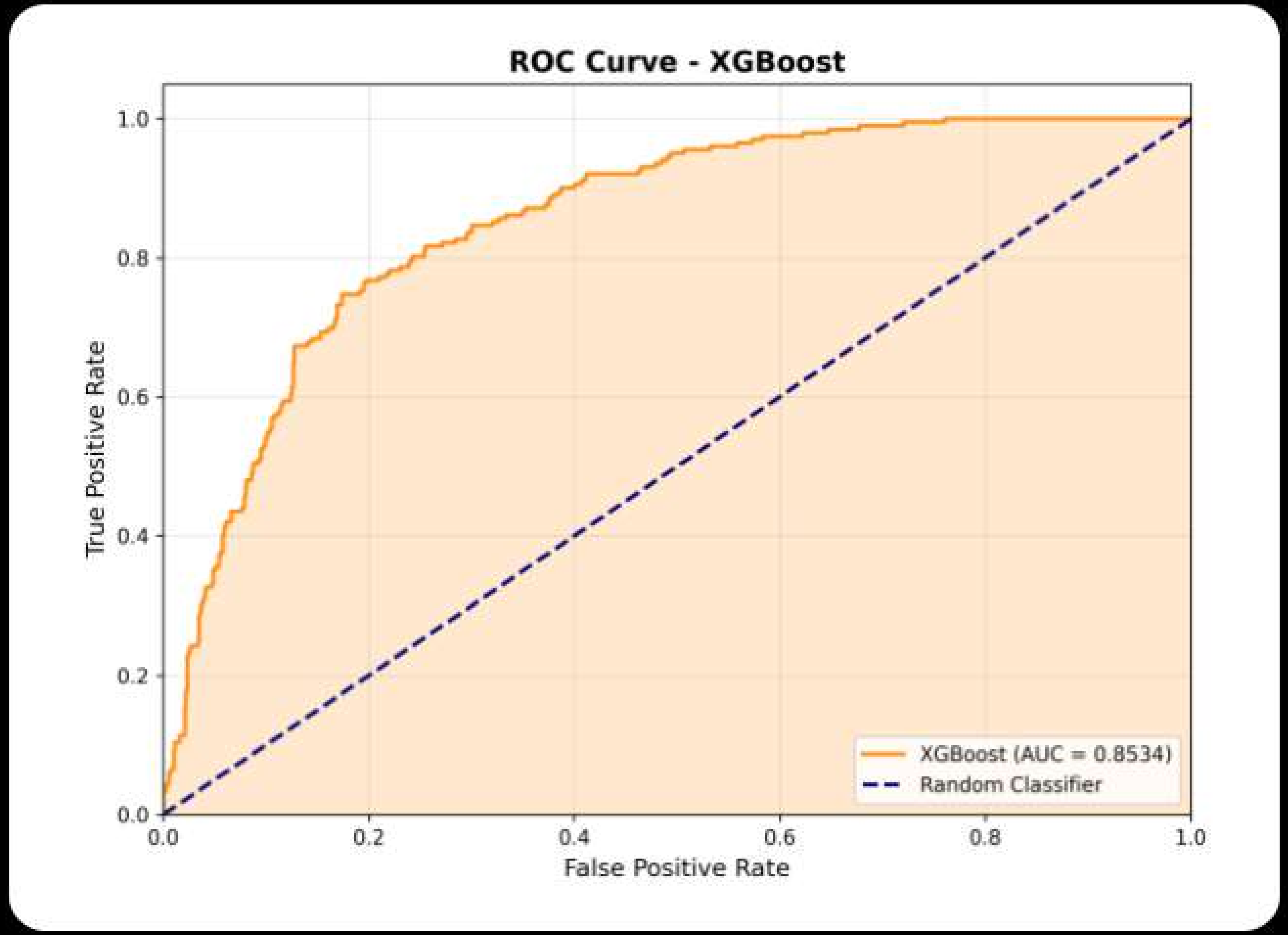
FINAL RECOMMENDATIONS

The Unified Model

Unified Model Analysis

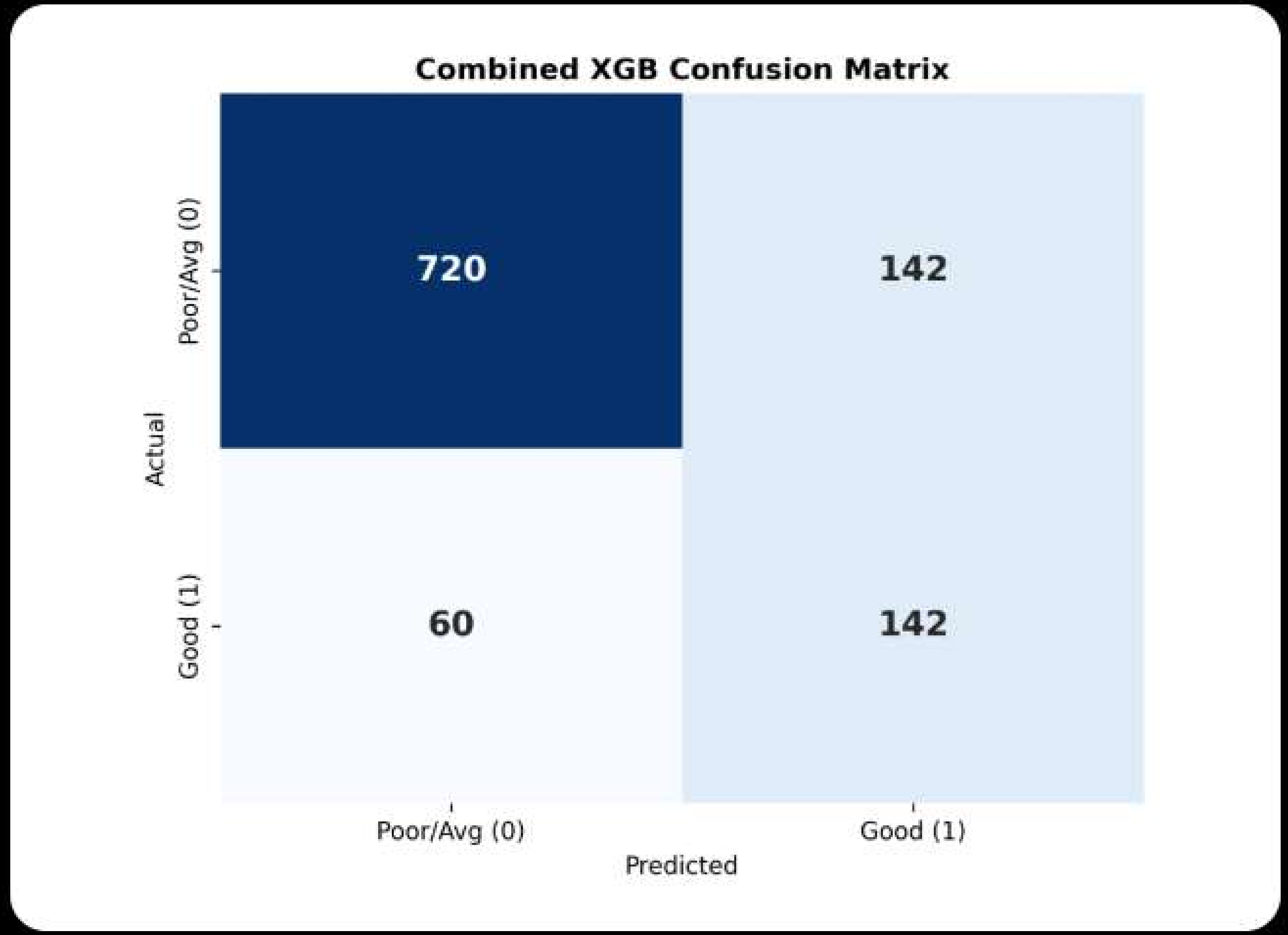
Evaluation Visuals

ROC CURVE (AUC = 0.85)



Demonstrates strong separation capability across the unified dataset.

CONFUSION MATRIX



Consistent performance across both Red and White wine samples.

PRODUCTION STRATEGY

Deployment Recommendations

Matching the right model to the right business need.

USE CASE: RED WINE LINE

Random Forest

Tuned for Precision

Why?

Highest safety factor. It minimizes False Positives effectively, acting as a strict gatekeeper for premium red wines where quality perception is critical.

USE CASE: WHITE WINE LINE

Random Forest

Highest Accuracy

Why?

Achieves the highest F1-Score (0.60) and Accuracy (81%). It captures the subtle balance of Density and Alcohol better than other models.

USE CASE: UNIFIED SYSTEM

XGBoost

Single AI Agent

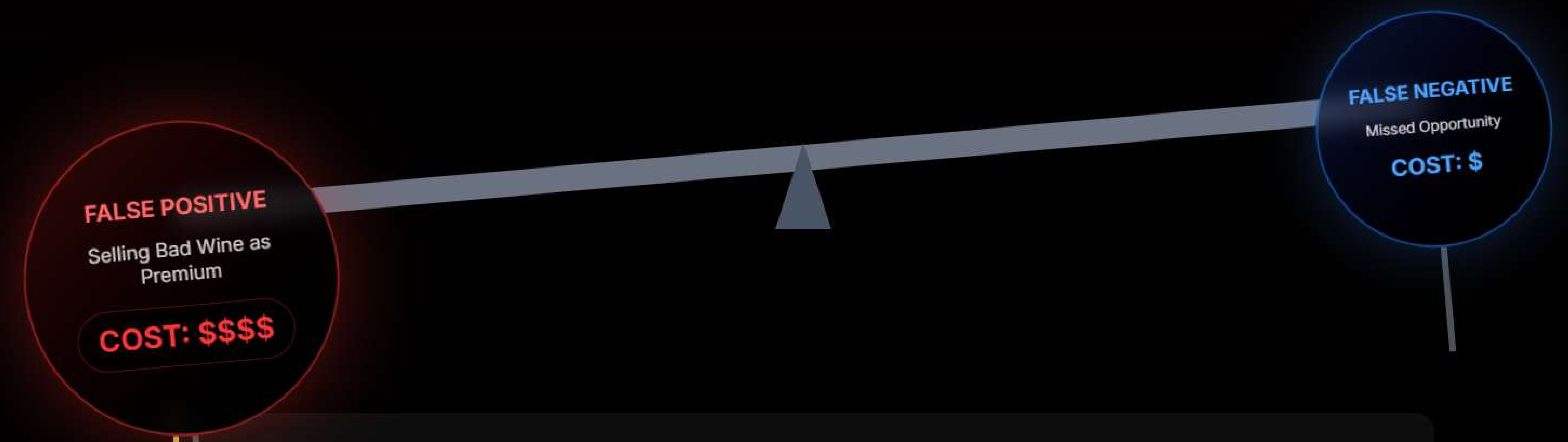
Why?

Statistically superior handling of the Red/White domain shift. Offers **98% of the performance** of specialized models with **50% less engineering maintenance**.

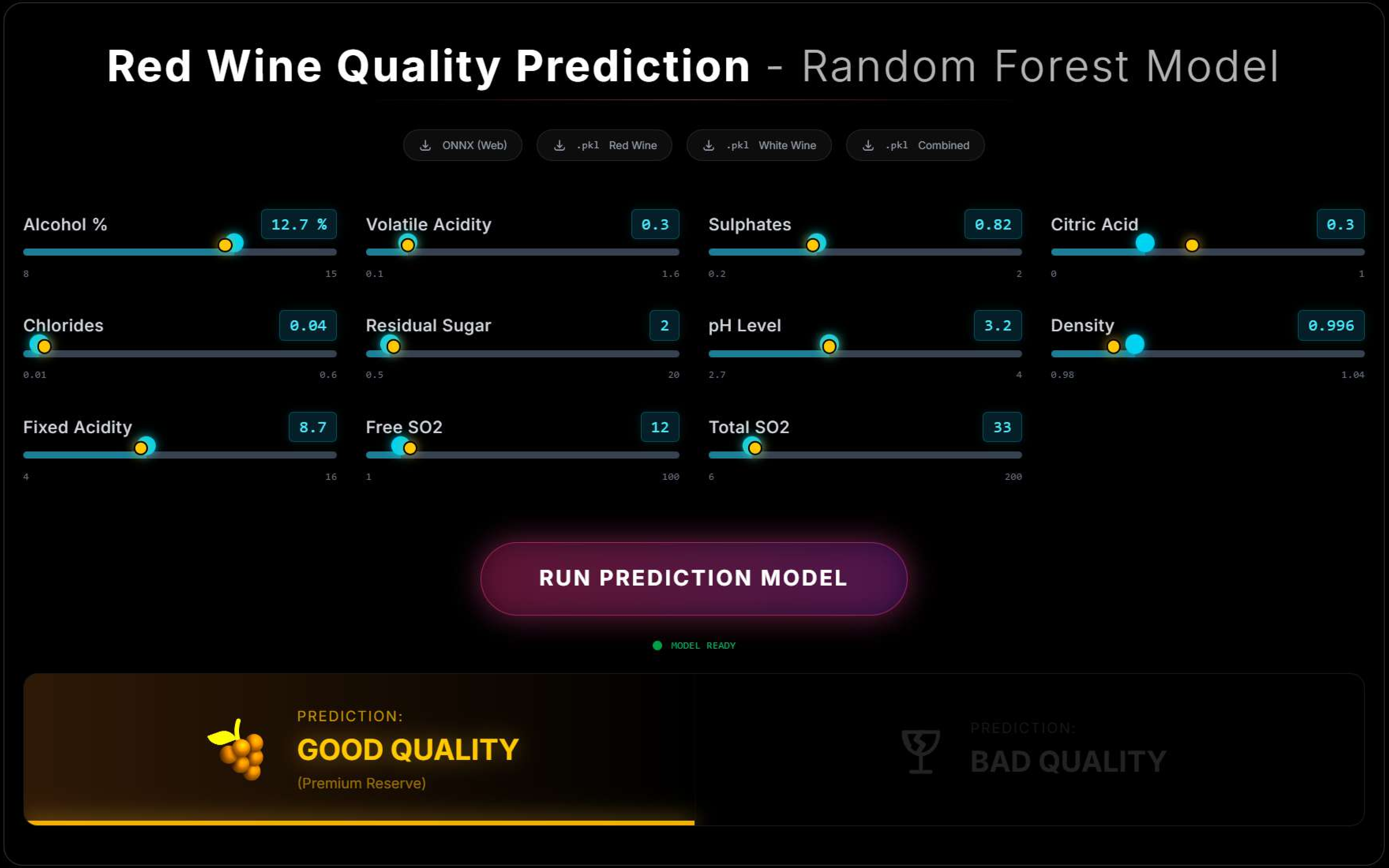
FUTURE OUTLOOK

Optimizing for Business Value

Moving from academic F1-Scores to real-world Cost-Sensitive Learning.



"In a real winery, brand damage from a bad bottle is far more costly than missing a sale. Future models will optimize a **Cost Matrix** rather than just Accuracy."


Red Wine Quality Prediction - Random Forest Model
ONNX (Web)
.pkl Red Wine
.pkl White Wine
.pkl Combined
Alcohol %
12.7 %
8
15
Volatile Acidity
0.3
0.1
1.6
Sulphates
0.82
0.2
2
Citric Acid
0.3
0
1
Chlorides
0.04
0.01
0.6
Residual Sugar
2
0.5
20
pH Level
3.2
2.7
4
Density
0.996
0.98
1.04
Fixed Acidity
8.7
4
16
Free SO2
12
1
100
Total SO2
33
6
200
RUN PREDICTION MODEL
MODEL READY
PREDICTION:
GOOD QUALITY
(Premium Reserve)
PREDICTION:
BAD QUALITY

Project Resources

Explore the complete analysis, methodology, and source code behind this project.

READ FULL PROJECT REPORT

GITHUB REPOSITORY