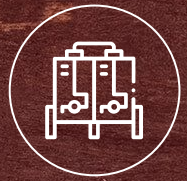
Red Wine Quality Prediction Model

Group 6
Lok Lam Wong
Shangzhou Xia
Yiwen Ma
Ziming Qin

TABLE OF CONTENTS

01. BUSINESS OBJECTIVES

BUSINESS OBJECTIVES



SUPPLIER

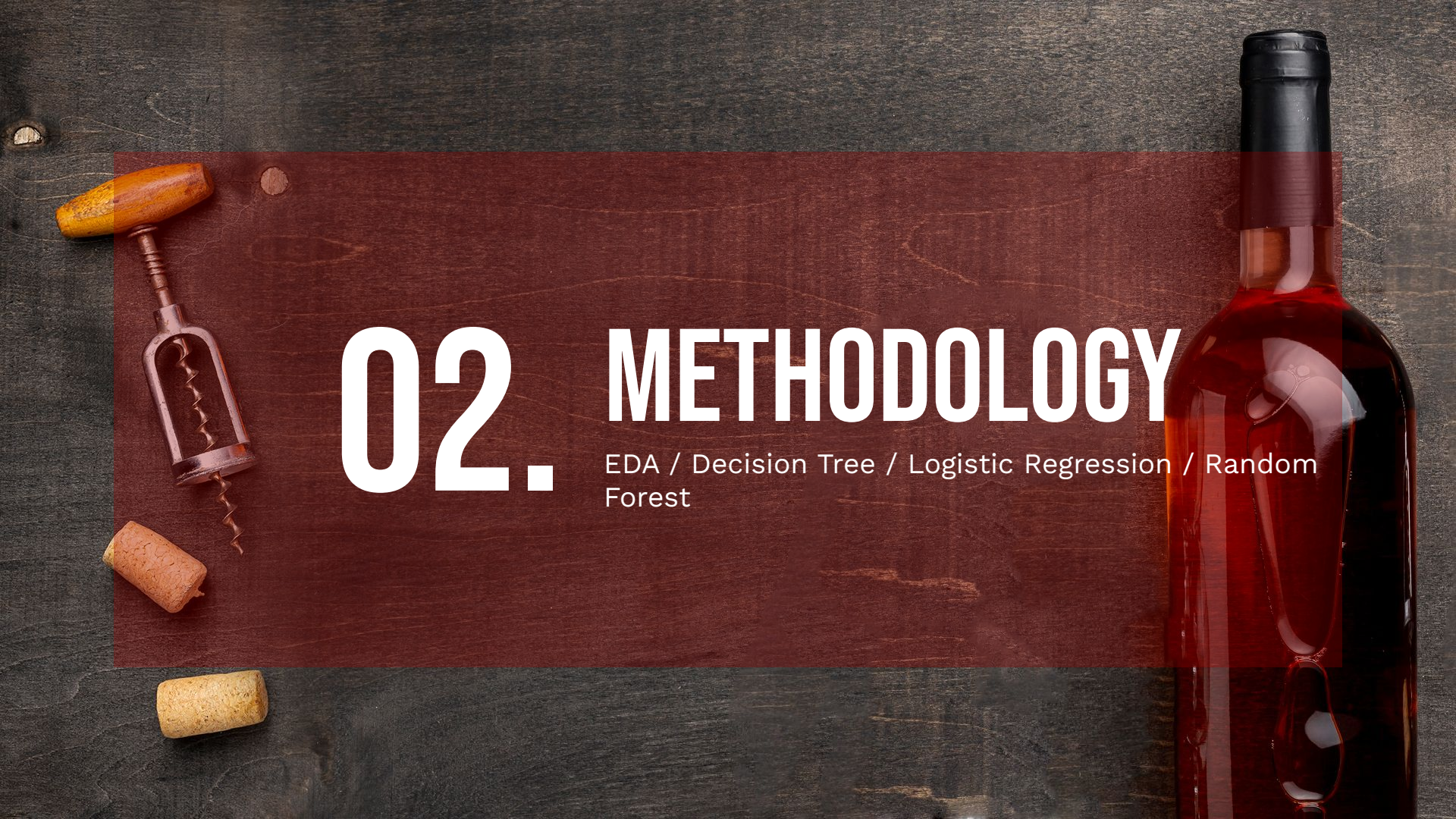
Ensure a certain wine quality; enhance the supplier's reputation and product value

CUSTOMER

Make informed purchasing decisions

RETAILER

Optimize stock levels and reduce the risk of unsold inventory



02. METHODOLOGY

EDA / Decision Tree / Logistic Regression / Random Forest

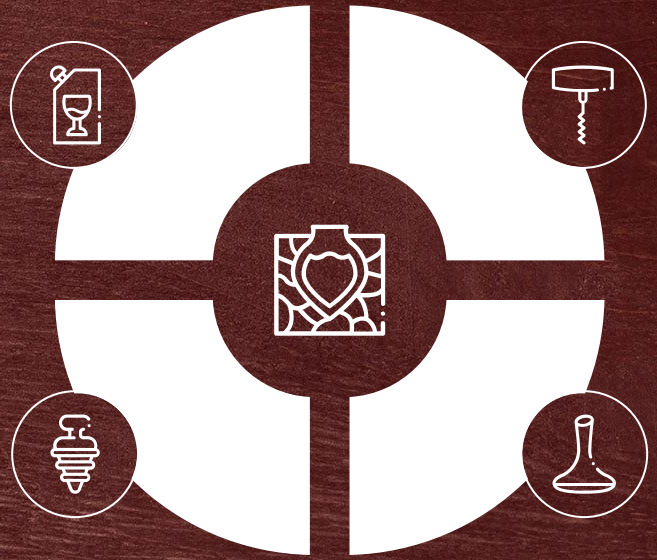
METHODOLOGY

EDA

Descriptive Status & Plots

LOGISTIC REGRESSION

Accuracy:0.747



DECISION TREE

Accuracy:0.756

RANDOM FOREST

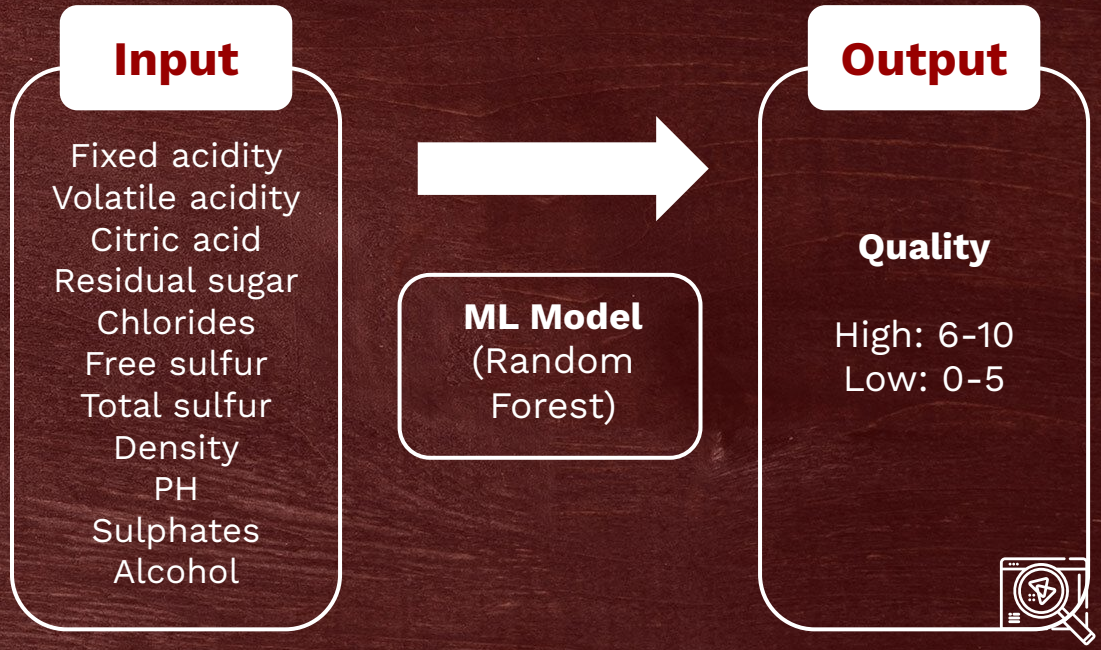
Accuracy:0.822

03. EDA

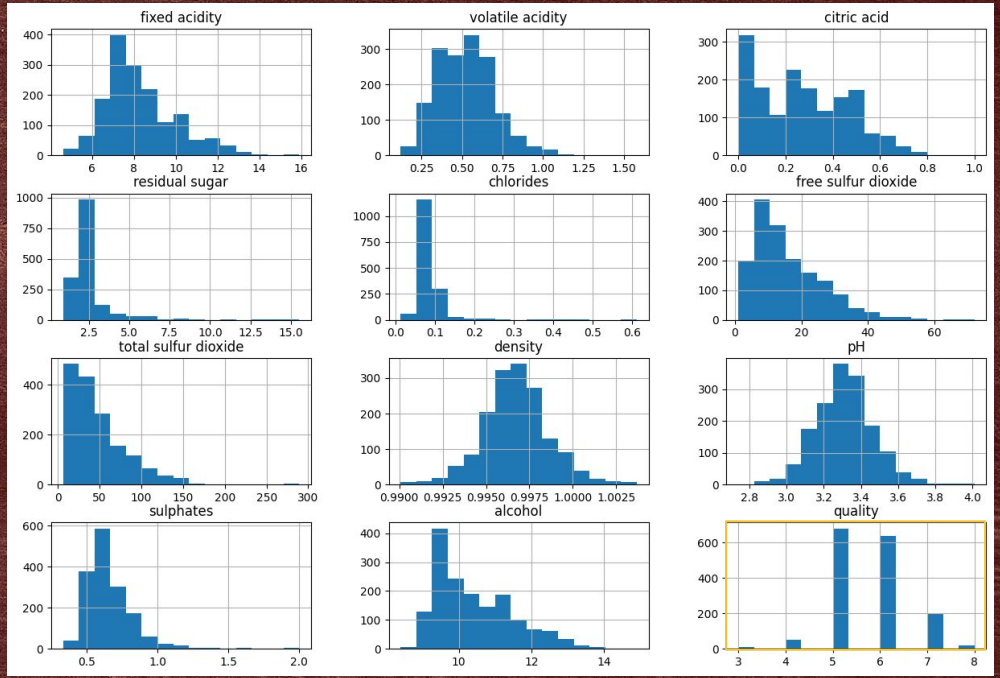
Descriptive Data & Histogram

DATA SET

- North Portugal Red Wine sample data based on physicochemical tests
- 13 Variables with no missing value
- Consider this dataset as classification tasks, given that the wines were being subjectively rated on a scale from 1 to 10



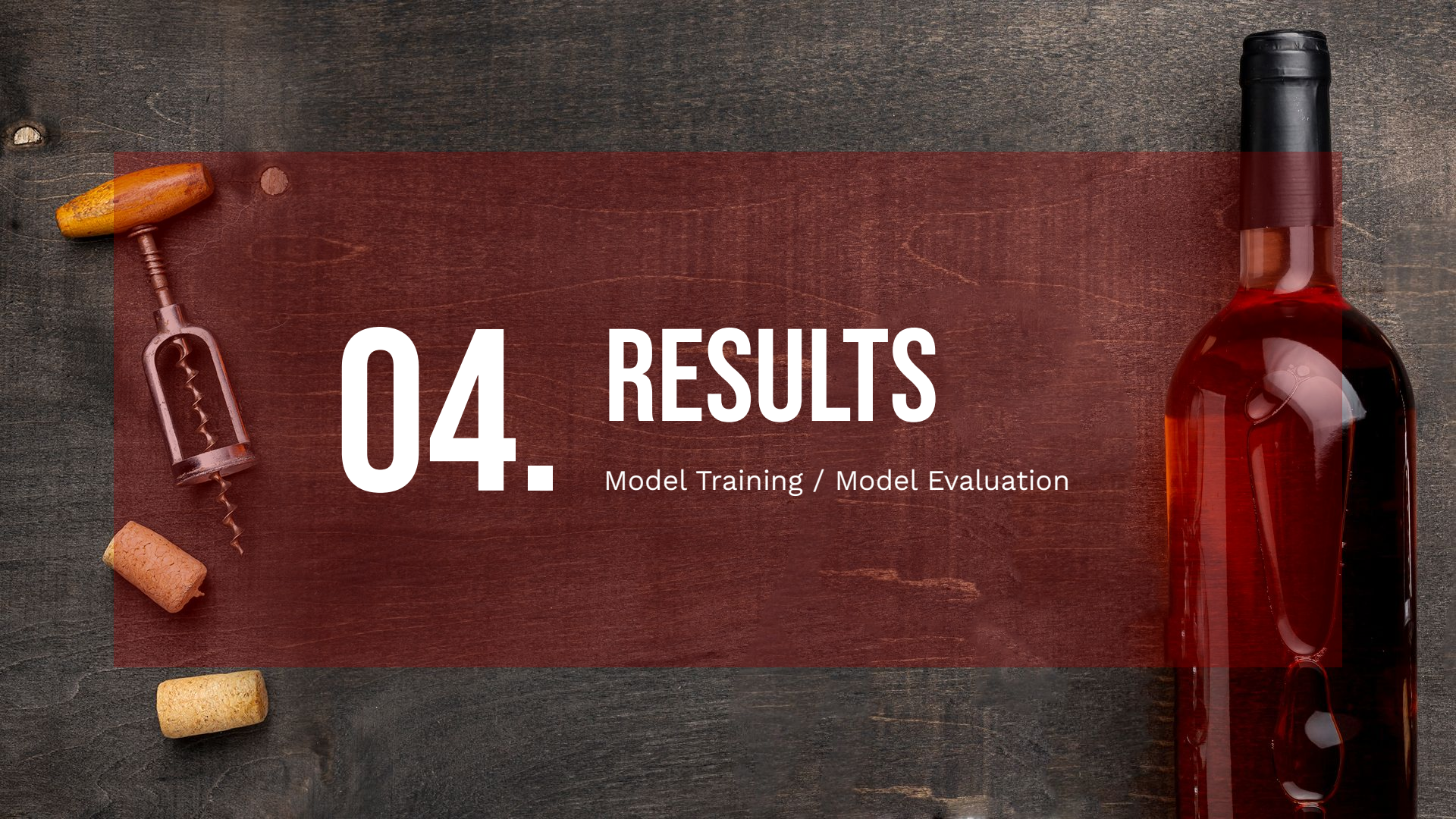
HISTOGRAM



- Outliers are founded in some features and normalization has been applied
- Quality data are ordered and not balanced, in real life there are many more normal wines than excellent or poor ones
- Quality rating have been classified into either high or low, for retail or consumer recommendations, the distinction is more relevant than exact score

04. RESULTS

Model Training / Model Evaluation



HYPER-PARAMETER TUNING – GRID SEARCH

DECISION TREE

1. **Max. Depth:** None
2. **Min. Samples Split:** 2
3. **Max. Leaf Nodes:** 40

LOGISTIC REGRESSION

1. **Max. Iteration:** 50

RANDOM FOREST

1. **Max. Depth:** None
2. **No. of trees in the forest (n_estimators):** 50

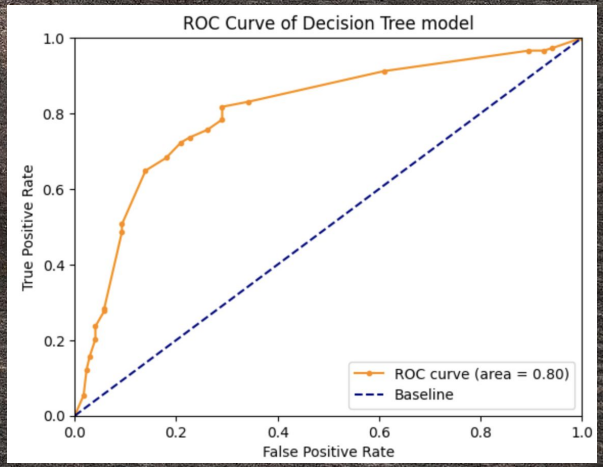
ML MODEL RESULTS – CLASSIFICATION REPORT

	DECISION TREE	LOGISTIC REGRESSION	RANDOM FOREST
ACCURACY	0.756	0.747	0.822
PRECISION	0.773	0.769	0.836
RECALL	0.773	0.756	0.831
F1 SCORE	0.773	0.762	0.834

ML MODEL RESULTS – ROC CURVE

01

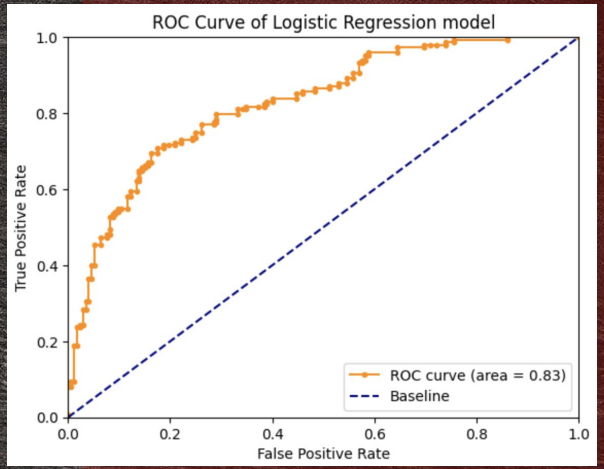
DECISION TREE



AUC: 0.80

02

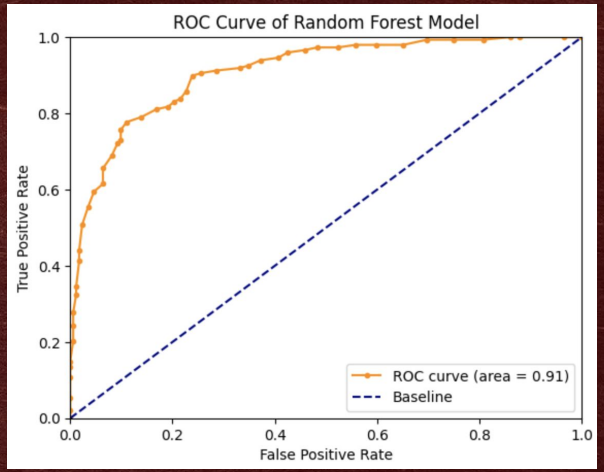
LOGISTIC REGREssION



AUC: 0.83

03

RANDOM FOREST

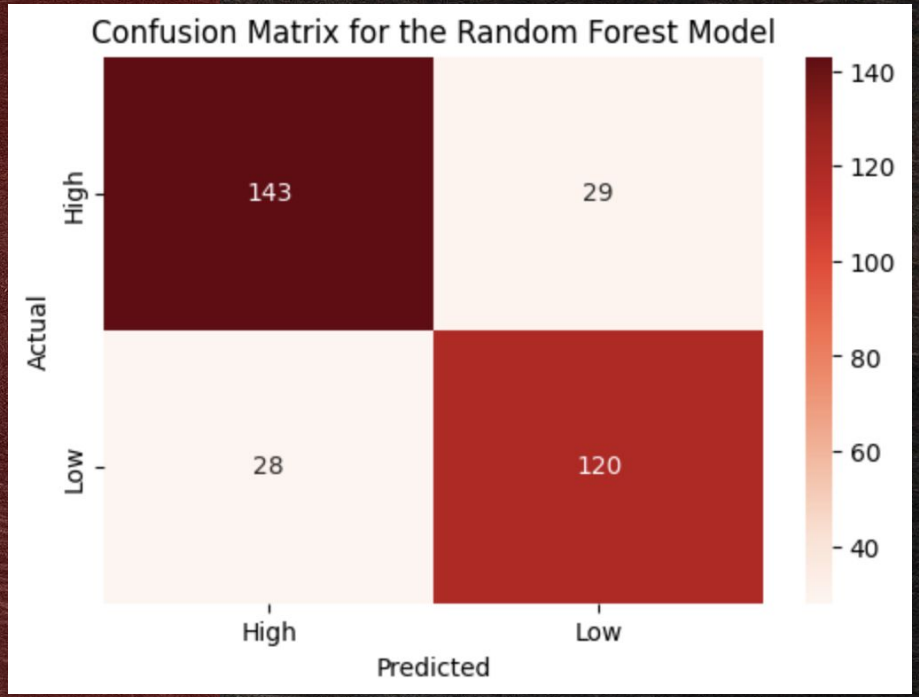


AUC: 0.91

BEST ML CLASSIFIER

Random Forest

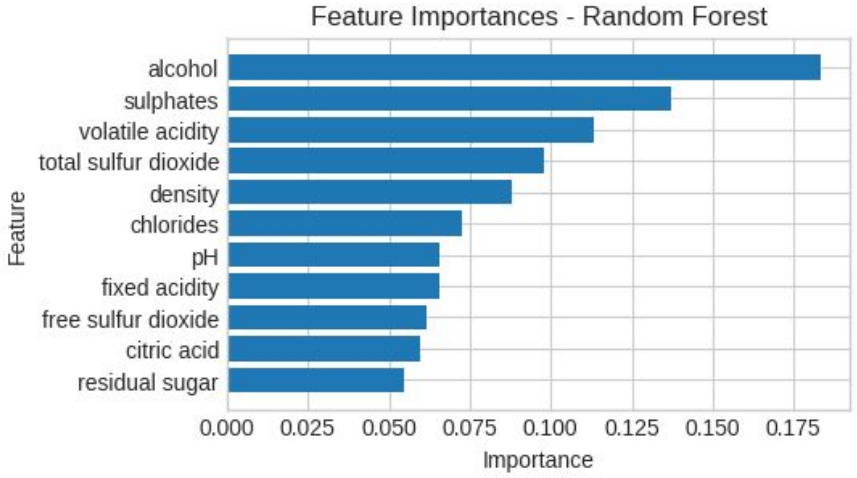
- Highest accuracy, precision, recall, F1-score, and AUC score
- Low False Positives and False Negatives
- Best predictive model for high-quality wines



05. EXPLAINABILITY

FEATURE IMPORTANCE/ICE/PDP

FEATURE IMPORTANCE



Feature Importances:

	Feature	Importance
10	alcohol	0.183361
9	sulphates	0.137153
1	volatile acidity	0.113269
6	total sulfur dioxide	0.098153
7	density	0.088177
4	chlorides	0.072543
8	pH	0.065732
0	fixed acidity	0.065538
5	free sulfur dioxide	0.061832
2	citric acid	0.059473
3	residual sugar	0.054770

Goal: find important features.

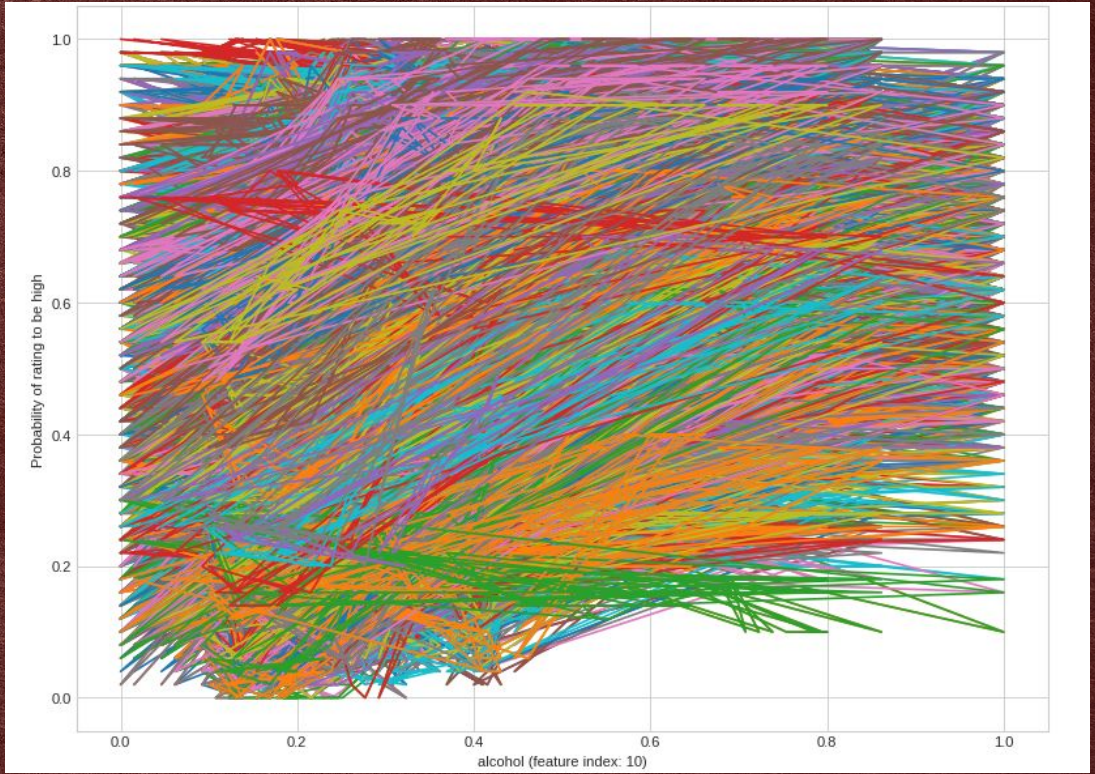
Top 3 features

- Alcohol: 0.1834
- Sulphates: 0.1372
- Volatile Acidity: 0.1133

The bigger the number is, the more important it is.

There are no explicit coefficients (like Logistics Regression) in Random Forest, so we cannot know the direction between the feature and the target.

INDIVIDUAL CONDITIONAL EXPECTATION (ICE)



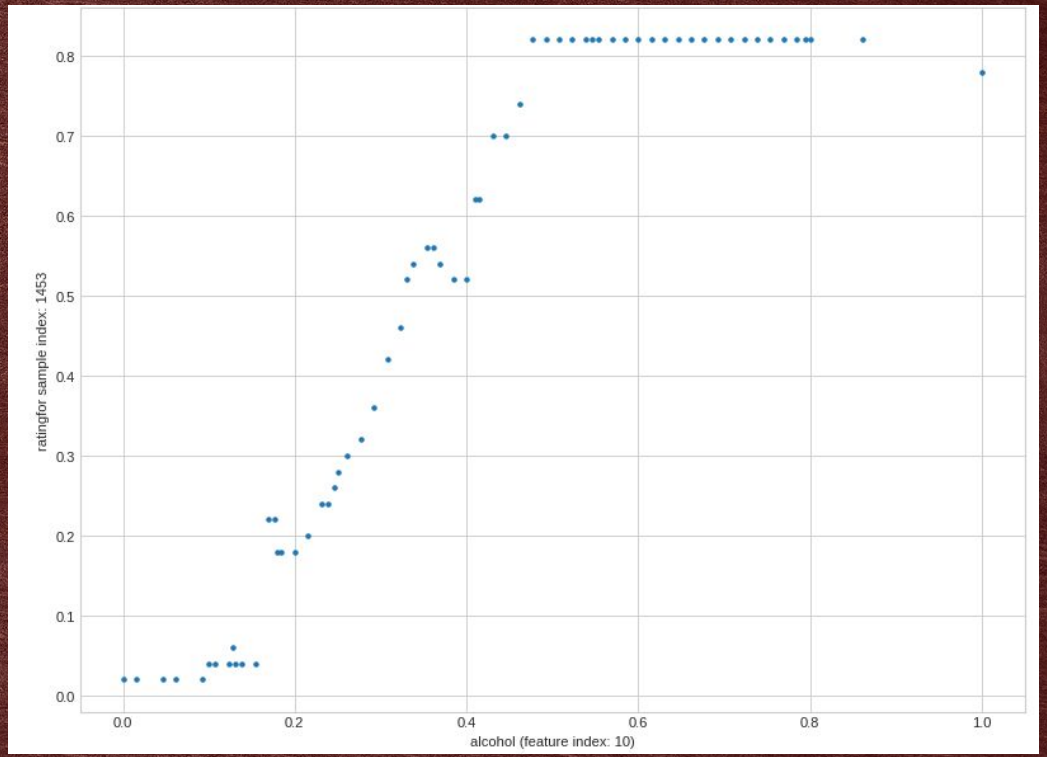
Goal: explore the impact of the selected feature on a specific instance.

Each line is an individual instance with N augmented records.

N is the number of unique values of the selected feature (e.g. alcohol) in all records.

Except for the selected feature, other features' values remain constant.

INDIVIDUAL CONDITIONAL EXPECTATION (ICE)



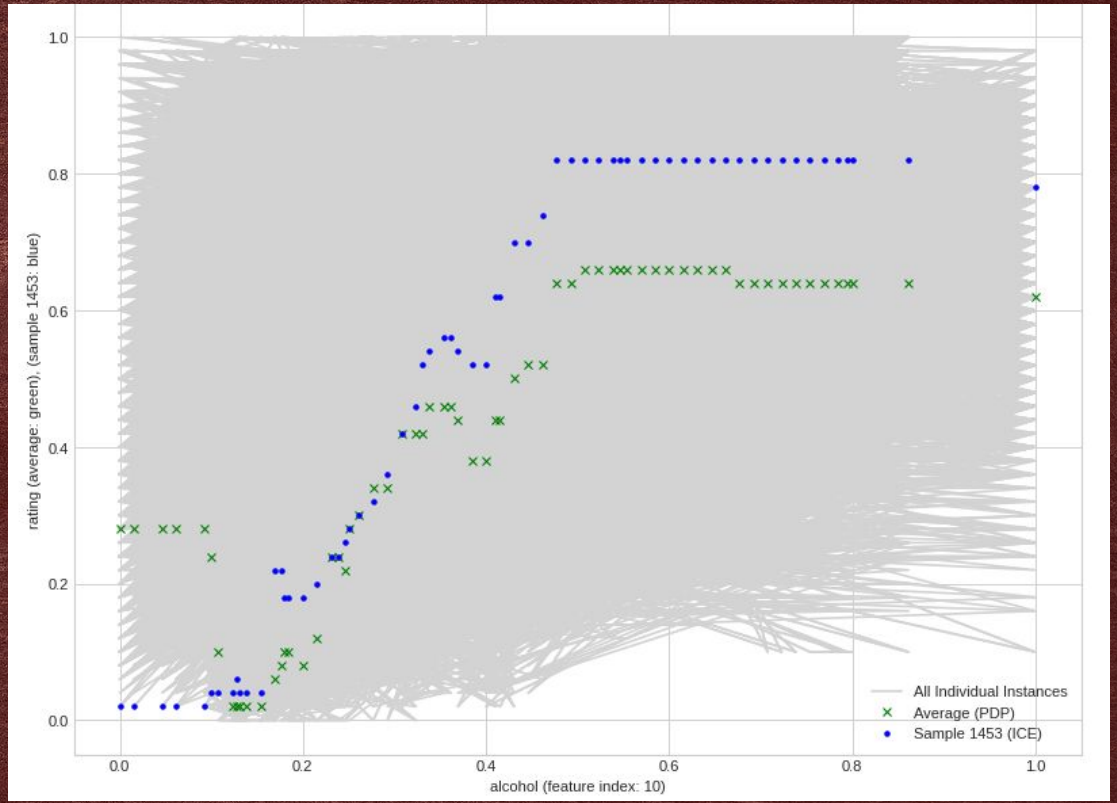
Instance Index: 1453.

As the alcohol increases, the probability of this instance to be rated as high goes up.

After 0.45 (normalized value), the rating keeps stable.

The rating even goes down after 0.8.

PARTIAL DEPENDENCY PLOTS (PDP)



Goal: explore the overall average impact of the selected feature on all instances.

May be different with a specific instance.

06. USE AND LIMITATION

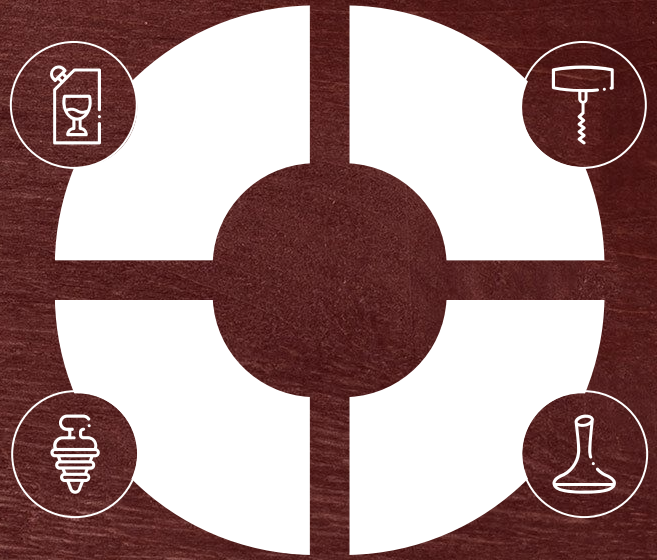
MODEL USE IN REAL-WORLD

GOAL 1

Automated
Preliminary Ratings

GOAL 2

Consistency in
Ratings



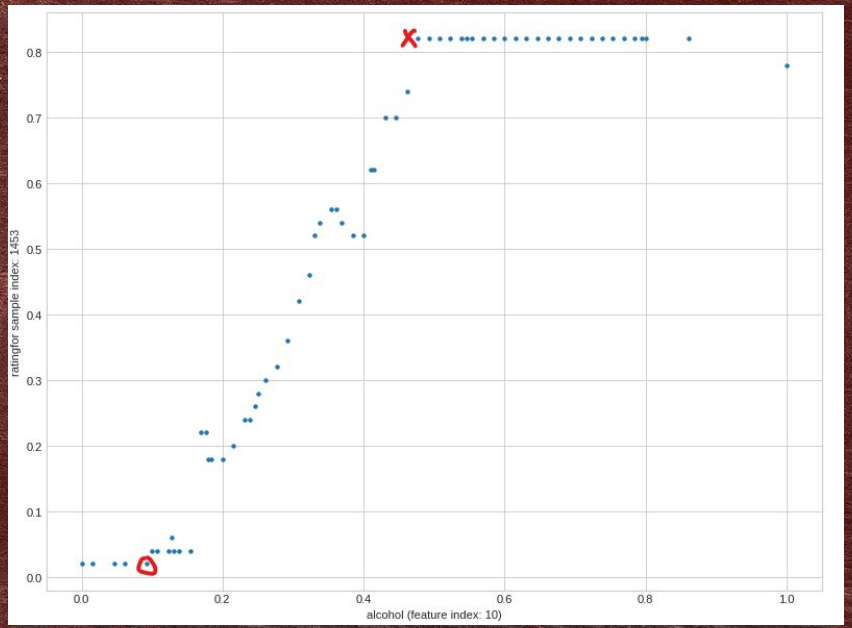
GOAL 3

Understandable
Index for Customers

GOAL 4

Product
Improvement
Consultations

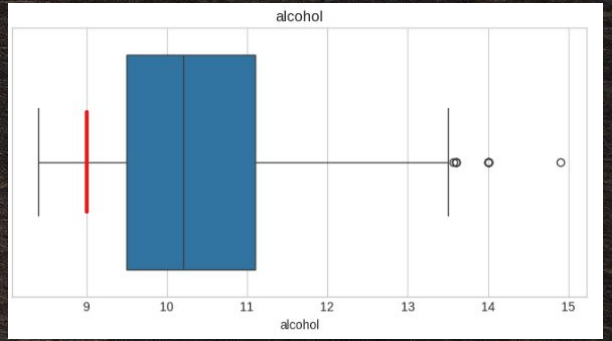
IMPROVE RATING BY ALCOHOL




```
1 df.iloc[1453]
```

fixed acidity	7.6
volatile acidity	0.49
citric acid	0.33
residual sugar	1.9
chlorides	0.074
free sulfur dioxide	27.0
total sulfur dioxide	85.0
density	0.99706
pH	3.41
sulphates	0.58
alcohol	9.0
quality	5
rating	Low

Name: 1453, dtype: object




```
1 print(X[1453])
```

```
[0.26548673 0.25342466 0.33       0.06849315 0.10350584 0.36619718
 0.27915194 0.51321586 0.52755906 0.1497006  0.09230769]
```

```
1 df.iloc[:,10].describe()
```

count	1599.000000
mean	10.422983
std	1.065668
min	8.400000
25%	9.500000
50%	10.200000
75%	11.100000
max	14.900000

Name: alcohol, dtype: float64

Increase the alcohol to improve the rating.
Pay attention to the threshold.
(14.9-8.4)*0.45+8.4 ≈ 11.3

LIMITATIONS

SUBJECTIVE

Rating is an artificial variable created by wine experts rather than a physicochemical index.

MULTICOLLINEARITY

It is difficult to hold other features remain constant when change the selected feature.

TOO SIMPLE

Only rating cannot showcase a kind of wine comprehensively.


THANK YOU