Wine Classifier Premium or Subprime

Ordinary chemical test results to predict quality.

Value Proposition

- Get from barrel to bottle faster
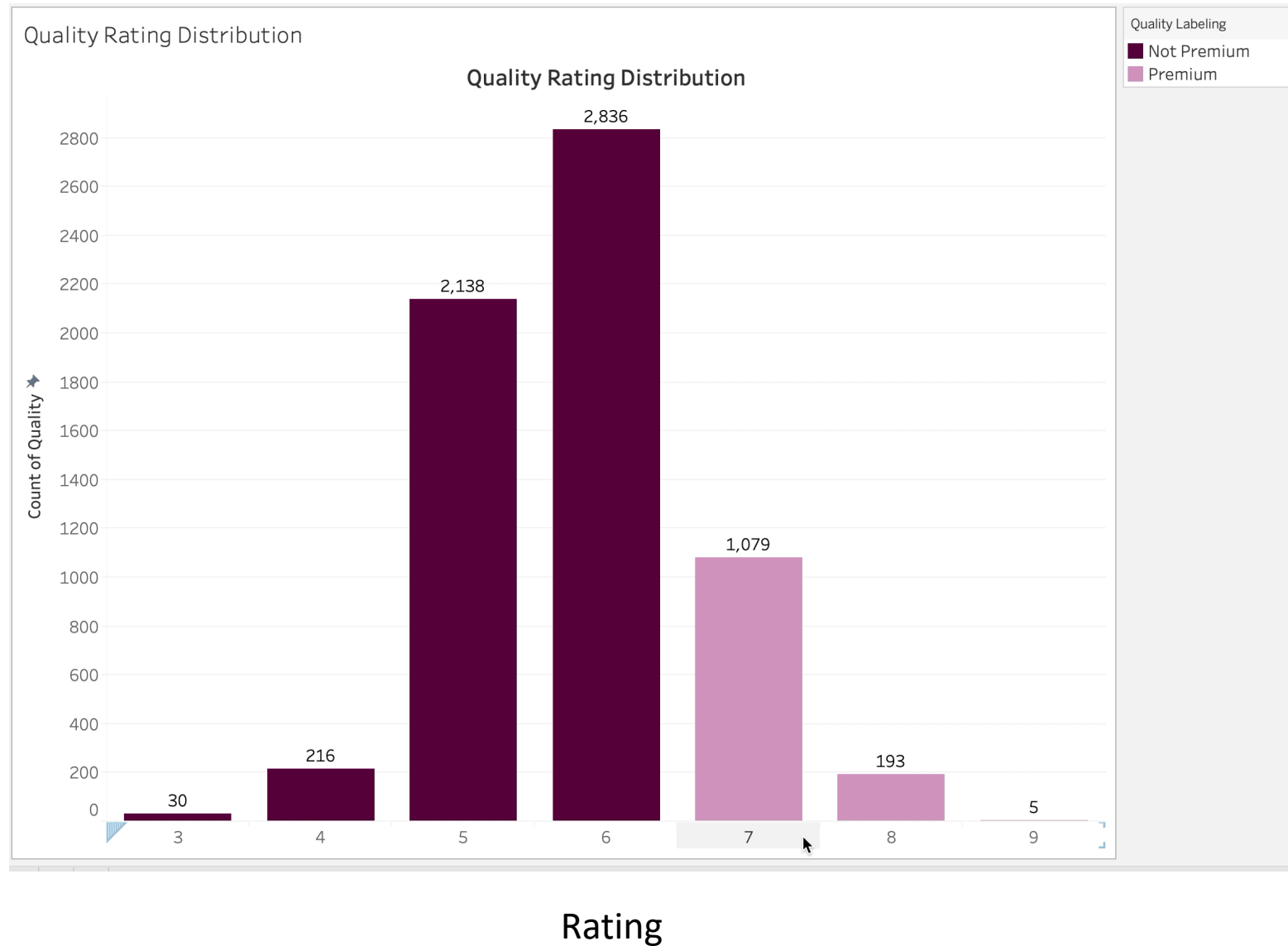- Improve cash flow

Which Metric?

Assumptions

- Niche vintner customer base

- Purchase based on Sales Forecast

Actual results	Subprime	True Negative True Subprime	False Positive False Premium	Potential $ Loss
	Premium	False Negative False Subprime	True Positive True Premium	
		Predicted Results		

Wine Classification

Count

Quality Rating Distribution

Quality Rating Distribution

Count of Quality

Rating	Count
3	30
4	216
5	2,138
6	2,836
7	1,079
8	193
9	5

Rating

20% Premium

80% Subprime

Commonly Tracked Attributes

- Alcohol
- Sulphates
- pH
- Density
- Total Sulphur Dioxide
- Free Sulphur Dioxide
- Chlorides
- Residual Sugar
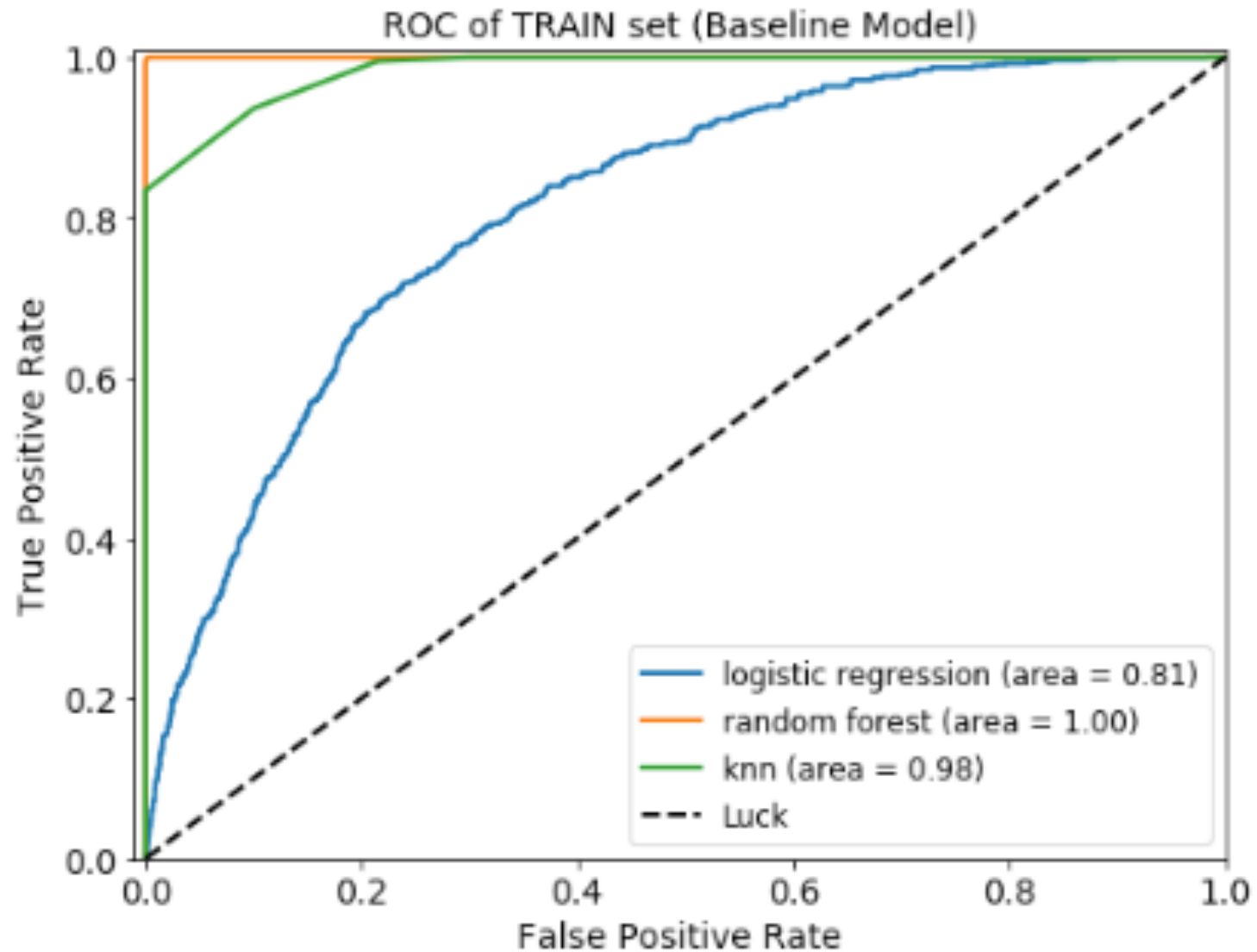- Volatile Acidity
- Fixed Acidity

Workflow

- Data acquisition
    postgreSQL

- Analysis
    Python, Numpy, Pandas, sklearn

- Visualizations
    Seaborn, Matplotlib, Tableau

- Precursor Application
    Flask

Methods

- 3 Models

  - K Nearest Neighbors
  - Random Forest
  - Logistic Regression

- Optimization & Component Analysis
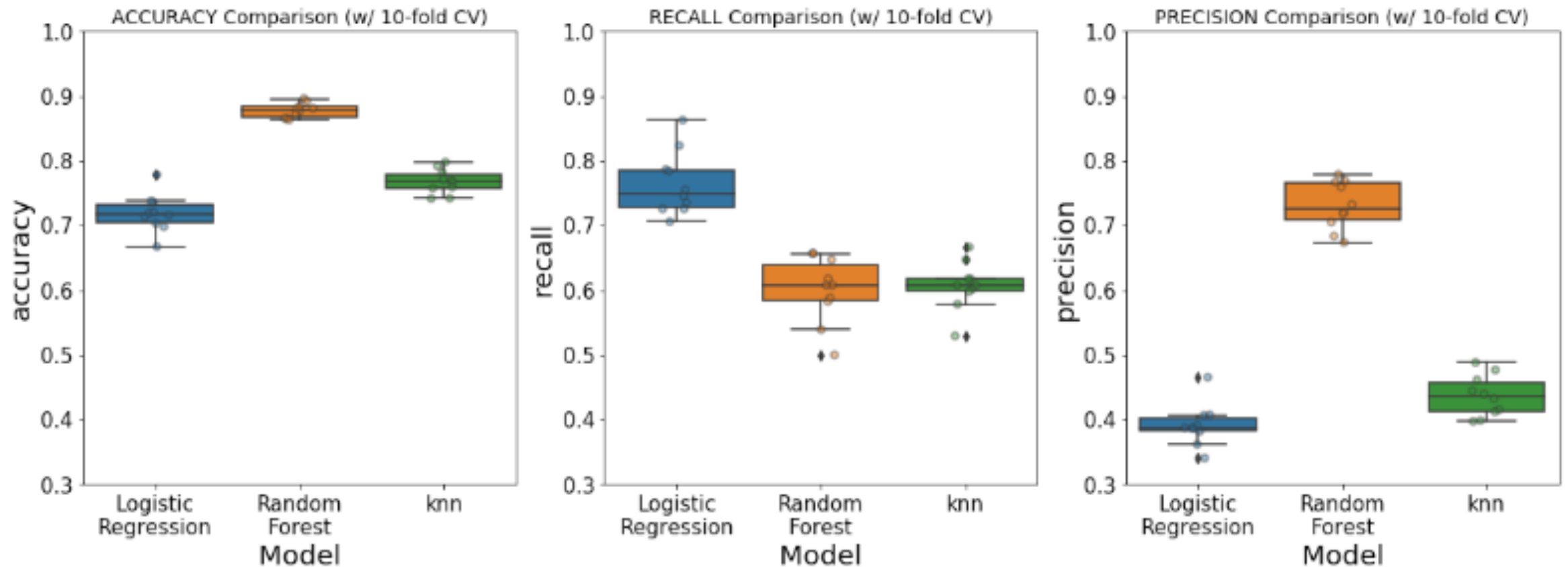
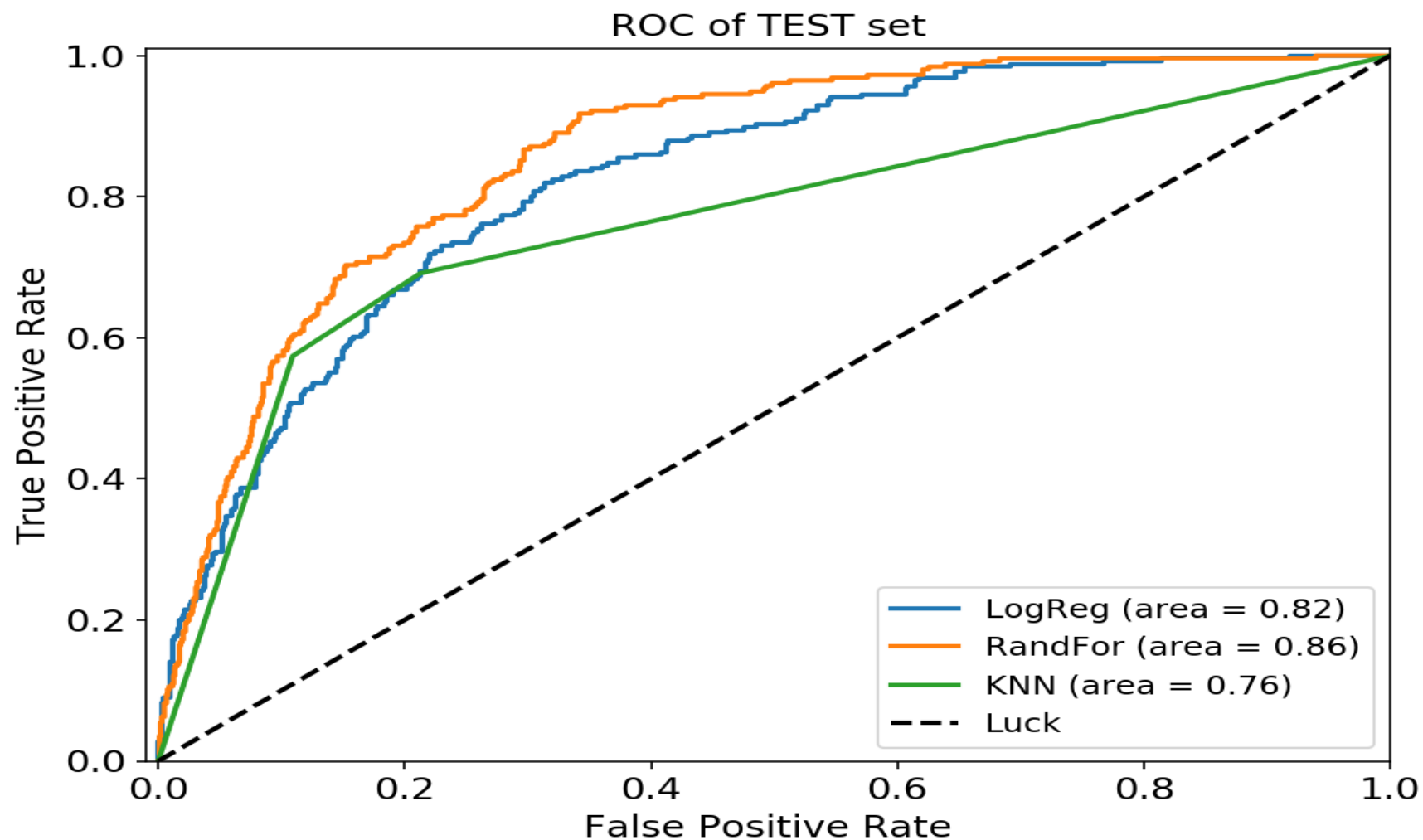- Evaluate & select best fit
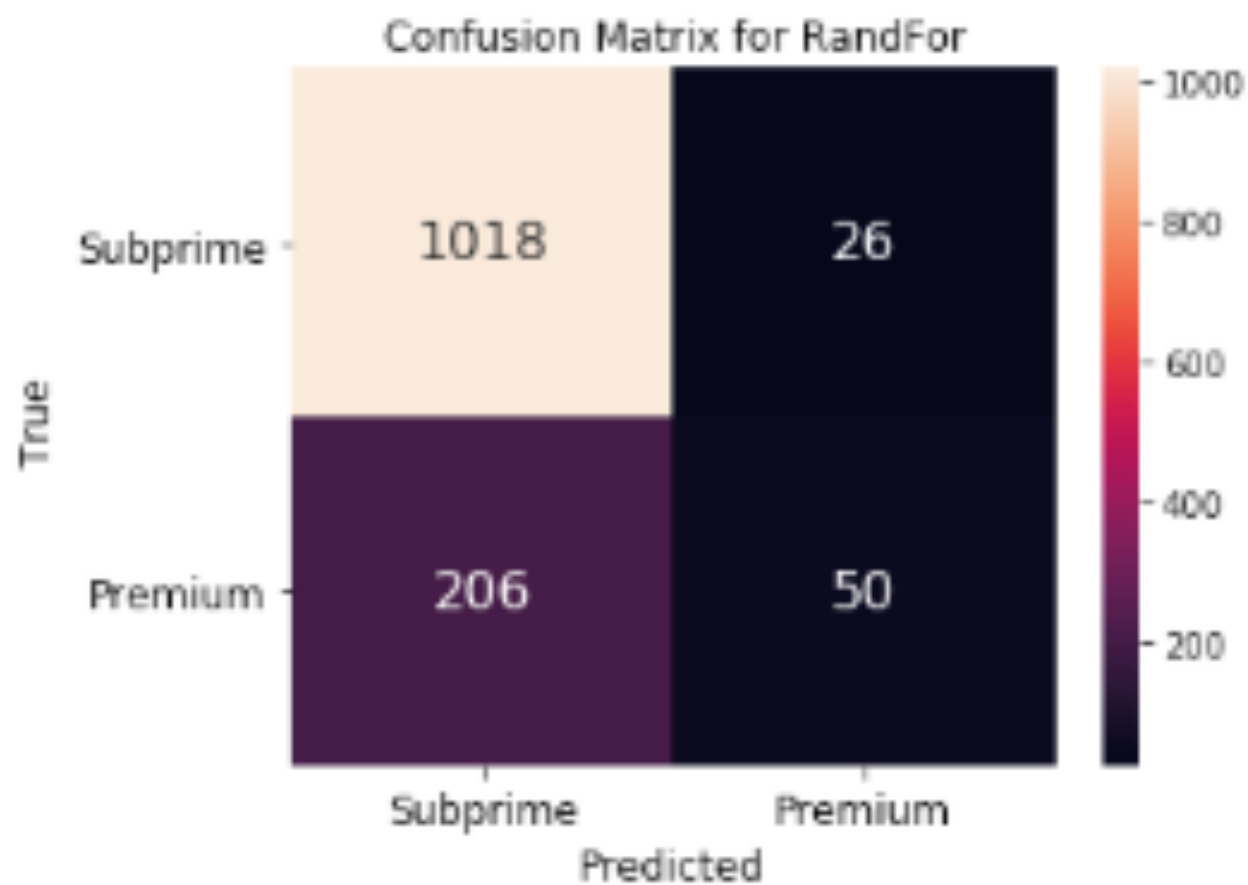
Which model is best fit?

ROC of TRAIN set (Baseline Model)

logistic regression (area = 0.81)
random forest (area = 1.00)
knn (area = 0.98)
Luck

Random f

Random Forest Reliability

METRICS Plots for 3 Crossvalidated Models

Out of Sample Results

Tuning the Threshold

Confusion Matrix for RandFor

Feature Importances in RF Model
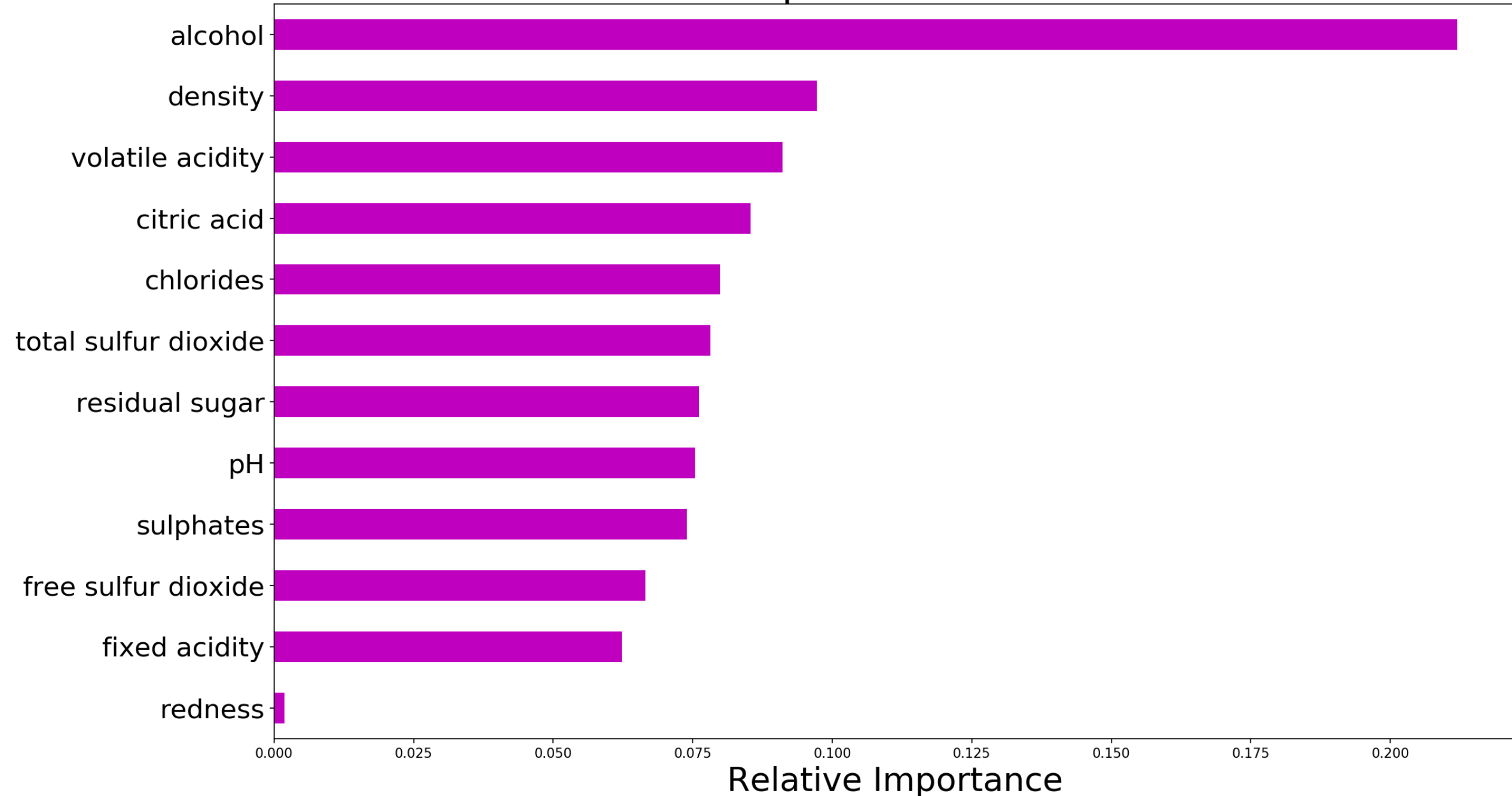
Link to Demonstration

Take Away

- Data Science can be applied to wine classification

- Recommend Random Forest

- Custom tune threshold

- Prototype Flask Application

Appendix

- **Citation:**

  P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

The Chemical Characteristics

Table 1. The physicochemical data statistics

Attribute (units)	Min	Max	Mean
fixed acidity (g(tartaric acid)$/dm^3$)	3.8	14.2	6.9
volatile acidity (g(acetic acid)$/dm^3$)	0.1	1.1	0.3
citric acid (g/dm^3)	0.0	1.0	0.3
residual sugar (g/dm^3)	0.6	65.8	6.4
chlorides (g(sodium chloride)$/dm^3$)	0.01	0.35	0.05
free sulfur dioxide (mg/dm^3)	2	260	35
total sulfur dioxide (mg/dm^3)	9	260	138
density (g/cm^3)	0.987	1.039	0.994
pH	2.7	3.8	3.1
sulphates (g(potassium sulphate)$/dm^3$)	0.2	1.1	0.5
alcohol (% vol.)	8.0	14.2	10.4