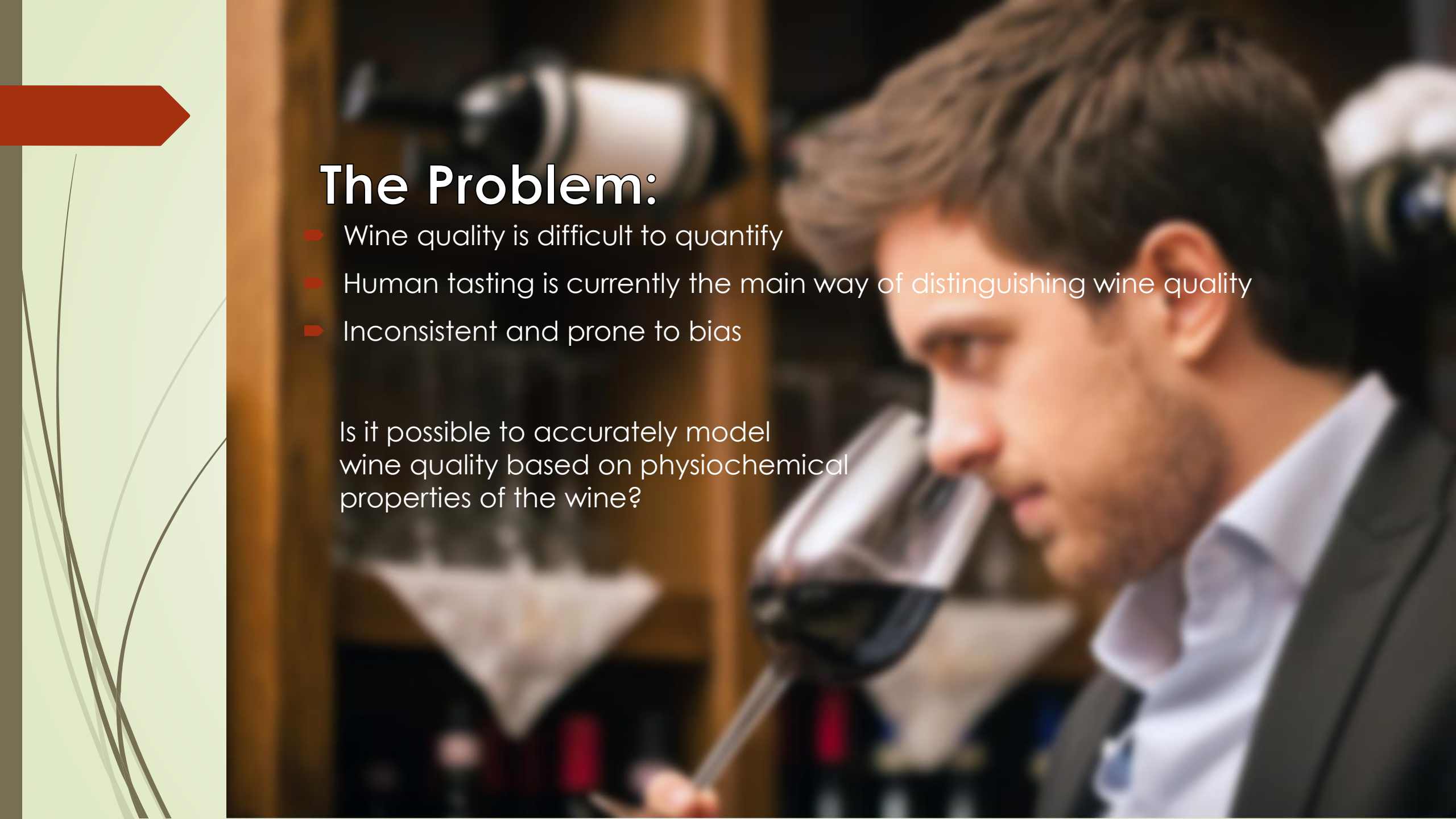


Modeling Wine Quality Using Physiochemical Properties

A Data Mining
Analysis
By Clement
Chen



A man in a dark suit and light blue shirt is shown in profile, holding a glass of red wine up to his nose and inhaling. He is in a dimly lit room with wooden shelves in the background, which appear to be a wine cellar or a tasting room. The lighting is warm and focused on the man and the wine glass. On the left side of the image, there is a vertical green bar with a red arrow pointing right and some faint, stylized line art.

The Problem:


- Wine quality is difficult to quantify
- Human tasting is currently the main way of distinguishing wine quality
- Inconsistent and prone to bias

Is it possible to accurately model wine quality based on physiochemical properties of the wine?



Business Case I: Small Family-Owned Winery

- A small business like this receives most of their revenue from individuals willing to pay top dollar for high quality wines.
- In order to retain their clientele, this business must make certain that the wines they sell are always of top quality. One bad wine could mean losing years of future profits.
- A business like this would be interested in a model that has high precision, that is to ensure that of all the wines labeled high quality, an overwhelming majority are truly high quality thus minimizing the chance that a loyal customer would receive a low-quality wine.



Business Case II: Large Winery

- A large winery would supply wines in boxes of 6 to retailers such as Costco or Walmart. As such, they are okay if not all the wines in the boxes are of the highest quality. From internal focus group tests, they have found out that consumers buying wine in bulk are okay with some of the wines being of lower quality because of the already cheap prices.
- In order to not waste lower quality wines, they purposely package some lower quality wines with higher quality wines instead of getting rid of all low-quality inventory.
- A business like this would be okay with a model that has lower precision, that is to ensure that of all the wines labeled high quality, a moderate fraction are truly high quality allowing less production loss due to rejected products.

Data Information

- Source University of Minho, Viticulture Commission of the Vinho Verde region (CVRVV)
- Each row contains a wine sample
- Contains two data sets
 - Red wine 1599 samples
 - White wine 4898 samples
- 11 feature columns 1 class label column
- Class labels are a quality rating from 0-10 based off the median value of at least 3 wine experts
- CSV format

The physicochemical data statistics per wine type

Attribute (units)	Red wine			White wine		
	Min	Max	Mean	Min	Max	Mean
fixed acidity ($g(\text{tartaric acid})/dm^3$)	4.6	15.9	8.3	3.8	14.2	6.9
volatile acidity ($g(\text{acetic acid})/dm^3$)	0.1	1.6	0.5	0.1	1.1	0.3
citric acid (g/dm^3)	0.0	1.0	0.3	0.0	1.7	0.3
residual sugar (g/dm^3)	0.9	15.5	2.5	0.6	65.8	6.4
chlorides ($g(\text{sodium chloride})/dm^3$)	0.01	0.61	0.08	0.01	0.35	0.05
free sulfur dioxide (mg/dm^3)	1	72	14	2	289	35
total sulfur dioxide (mg/dm^3)	6	289	46	9	440	138
density (g/cm^3)	0.990	1.004	0.996	0.987	1.039	0.994
pH	2.7	4.0	3.3	2.7	3.8	3.1
sulphates ($g(\text{potassium sulphate})/dm^3$)	0.3	2.0	0.7	0.2	1.1	0.5
alcohol (% vol.)	8.4	14.9	10.4	8.0	14.2	10.4

Data Wrangling

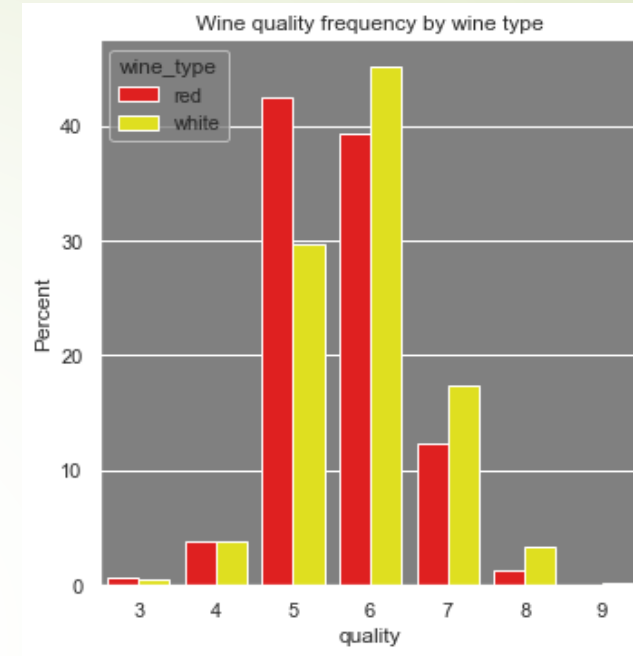
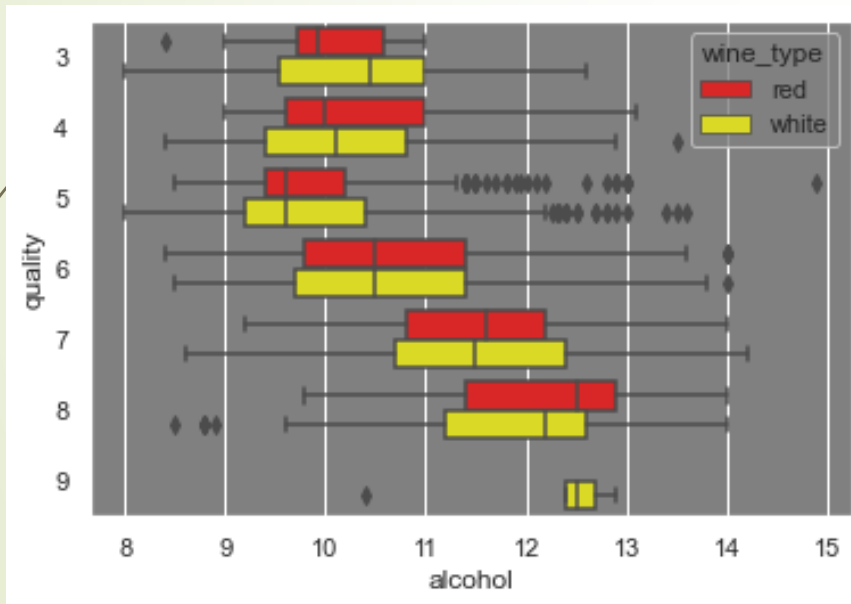
- Data were very clean
- Many duplicate rows
- Based on the unlikely chance that feature values were coincidentally the same values, and metadata stating wine quality is the median value of at least 3 wine experts, we can conclude that these are true duplicates

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	wine_type	size
460	7.0	0.150	0.28	14.70	0.051	29.0	149.0	0.99792	2.96	0.39	9.0	7	white	7
622	7.3	0.190	0.27	13.90	0.057	45.0	155.0	0.99807	2.94	0.41	8.8	8	white	7
661	7.4	0.160	0.30	13.70	0.056	33.0	168.0	0.99825	2.90	0.44	8.7	7	white	6
360	6.8	0.180	0.30	12.80	0.062	19.0	171.0	0.99808	3.00	0.52	9.0	7	white	6
728	7.6	0.200	0.30	14.20	0.056	53.0	212.5	0.99900	3.14	0.46	8.9	8	white	5
660	7.4	0.160	0.27	15.50	0.050	25.0	135.0	0.99840	2.90	0.43	8.7	7	white	5
664	7.4	0.190	0.30	12.80	0.053	48.5	229.0	0.99860	3.14	0.49	9.1	7	white	5
665	7.4	0.190	0.31	14.50	0.045	39.0	193.0	0.99860	3.10	0.50	9.2	6	white	5
684	7.4	0.330	0.26	15.60	0.049	67.0	210.0	0.99907	3.06	0.68	9.5	5	white	4
118	6.2	0.230	0.36	17.20	0.039	37.0	130.0	0.99946	3.23	0.43	8.8	6	white	4
267	6.6	0.220	0.23	17.30	0.047	37.0	118.0	0.99906	3.08	0.46	8.8	6	white	4

Table of some of the duplicate values removed.
The 'size' column represents the number of the duplicate rows found.

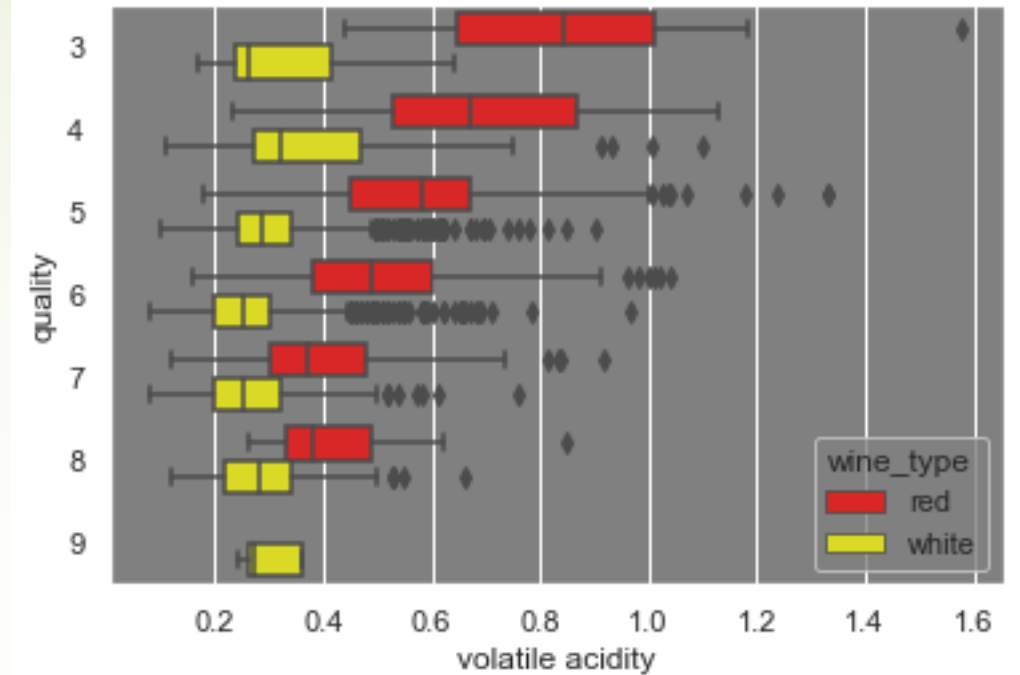
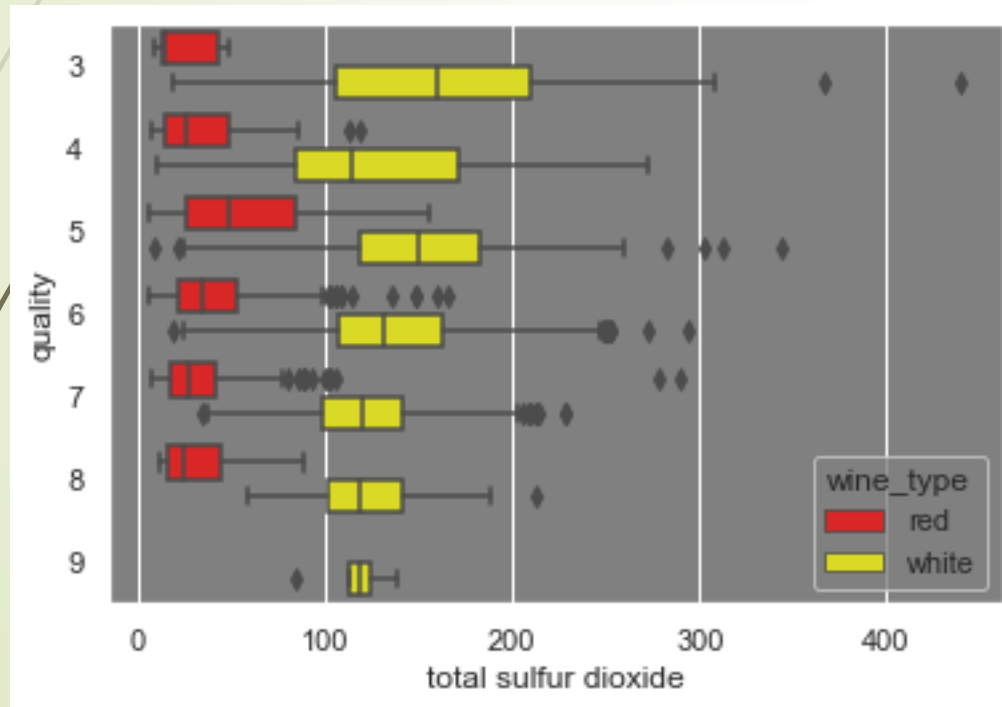
EDA

- White wines tend to have higher quality than red wines
- Most wines are in the 5-6 quality range



- Alcohol volume seems to be an important factor in determining wine quality for both red and white wines
- Higher alcohol content can be the result of
 - Higher sugar content from the grapes
 - More sugar being turned into alcohol
 - Length of fermentation

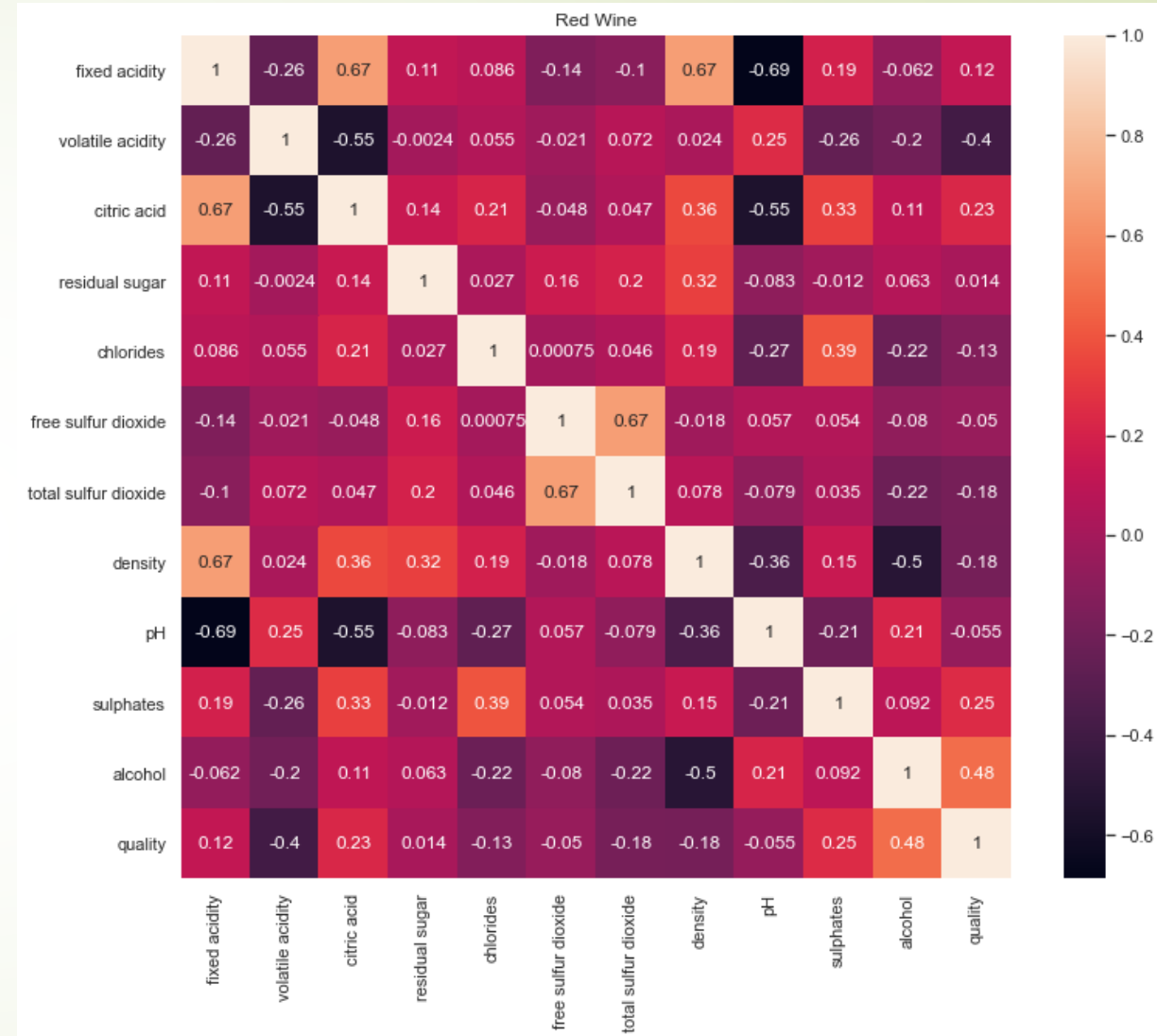
- Volatile acidity (acetic acid) is a significant indicator of quality in red wines
- Acetic acid content is the result of naturally occurring bacteria processing alcohol and oxygen into acetic acid



- Less total sulfur dioxide is an indicator for higher quality in white wines
- Sulfurs are added into wines for stability and to protect against oxidation but in this case, a tradeoff was made.

Red Wine Correlated Features

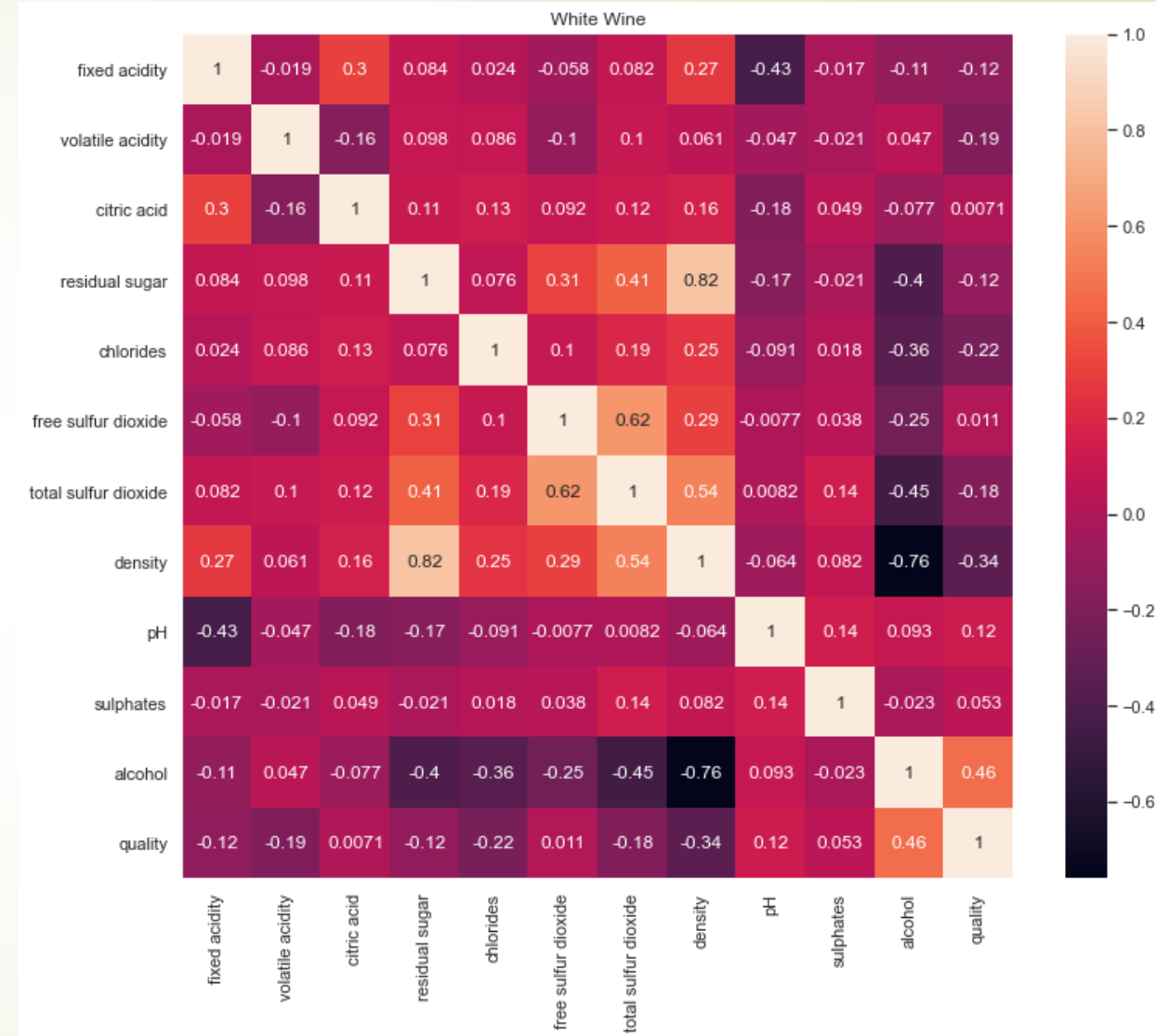
Feature 1	Feature 2	Correlation
Fixed Acidity	pH	-0.69
Fixed Acidity	Density	0.67
Fixed Acidity	Citric Acid	0.67
Total Sulfur Dioxide	Free Sulfur Dioxide	0.67



White Wine

Correlated Features

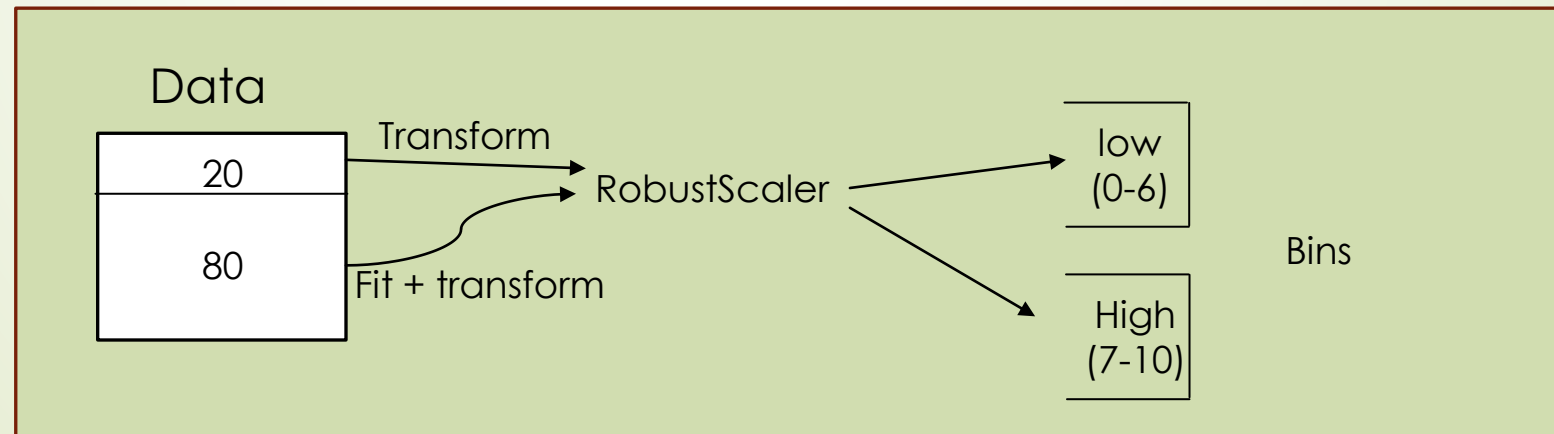
Feature 1	Feature 2	Correlation
Residual Sugars	Density	0.82
Density	Total Sulfur Dioxide	0.67
Alcohol	Density	-0.76
Total Sulfur Dioxide	Free Sulfur Dioxide	0.62



Preprocessing

Data Preprocessing:

1. Train-test split of ratio 80-20
2. Training data was fitted with RobustScaler
3. Test data was scaled with the fitted RobustScaler to prevent data leakage
4. Data was binned into low (0-6) and high (7-10) quality to create a binary classification task



Modeling

Classification Algorithms Used

Logistic Regression
Logistic Regression with PCA
Random Forest
Gradient Boosted Trees

Hyperparameter Tuning

5-fold CV with GridSearchCV and
RandomizedSearchCV

Scoring

ROC_AUC was used for scoring models

Results

Red Wine: Small Business

Model Name	f1 Score	Test Accuracy	ROC AUC	Precision	Recall	Winner
Logistic Regression	0.333	0.897	0.868	0.778	0.212	✗
PCA Logistic Regression	0.300	0.897	0.874	0.857	0.182	✓
Random Forest	0.375	0.890	0.874	0.600	0.273	✗
Gradient Boosted Tree	0.279	0.886	0.872	0.600	0.182	✗

Red Wine: Large Winery

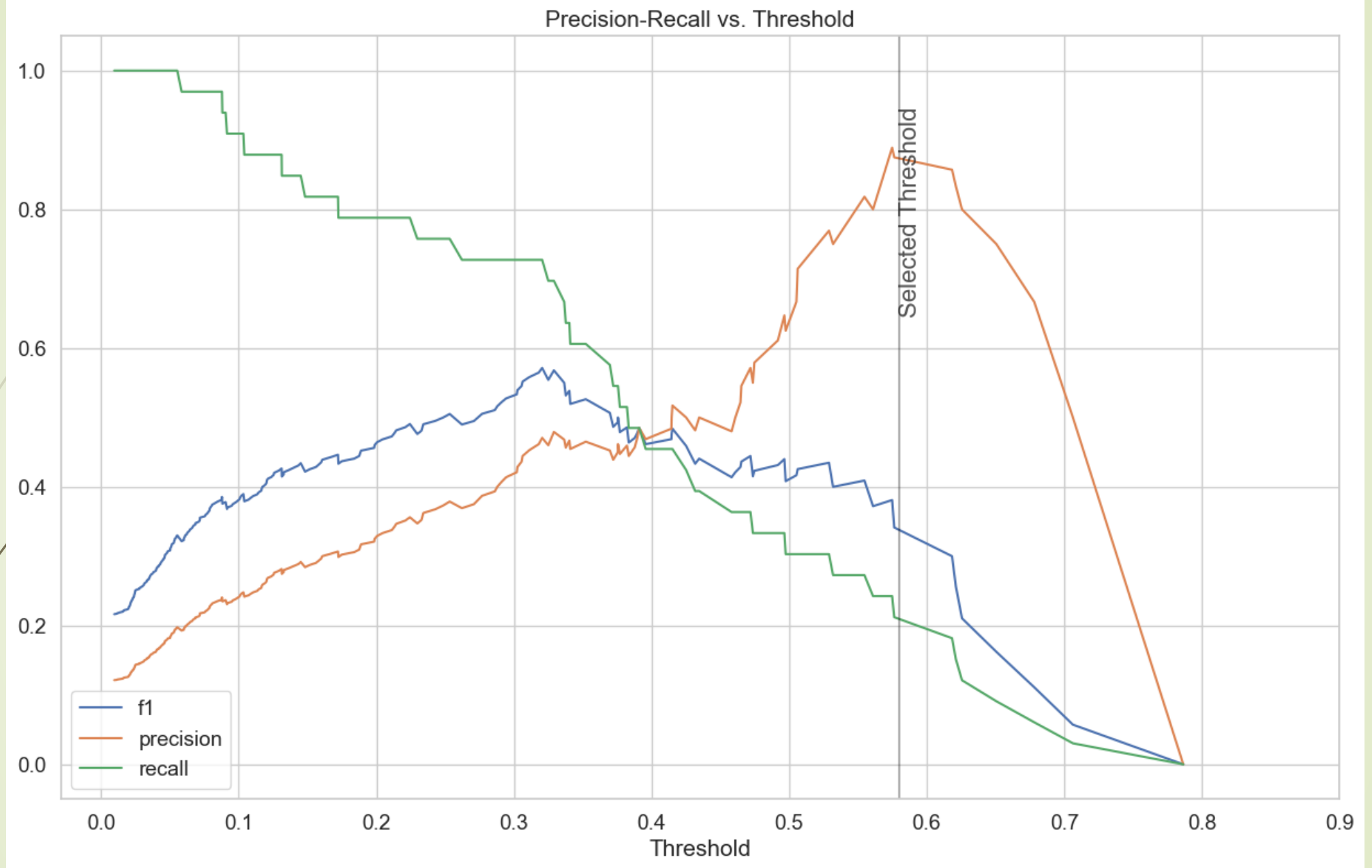
Model Name	f1 Score	Test Accuracy	ROC AUC	Precision	Recall	Winner
Logistic Regression	0.361	0.857	0.868	0.393	0.333	✗
PCA Logistic Regression	0.414	0.875	0.874	0.480	0.364	✗
Random Forest	0.495	0.820	0.874	0.375	0.727	✗
Gradient Boosted Tree	0.537	0.860	0.872	0.449	0.667	✓

White Wine: Small Business

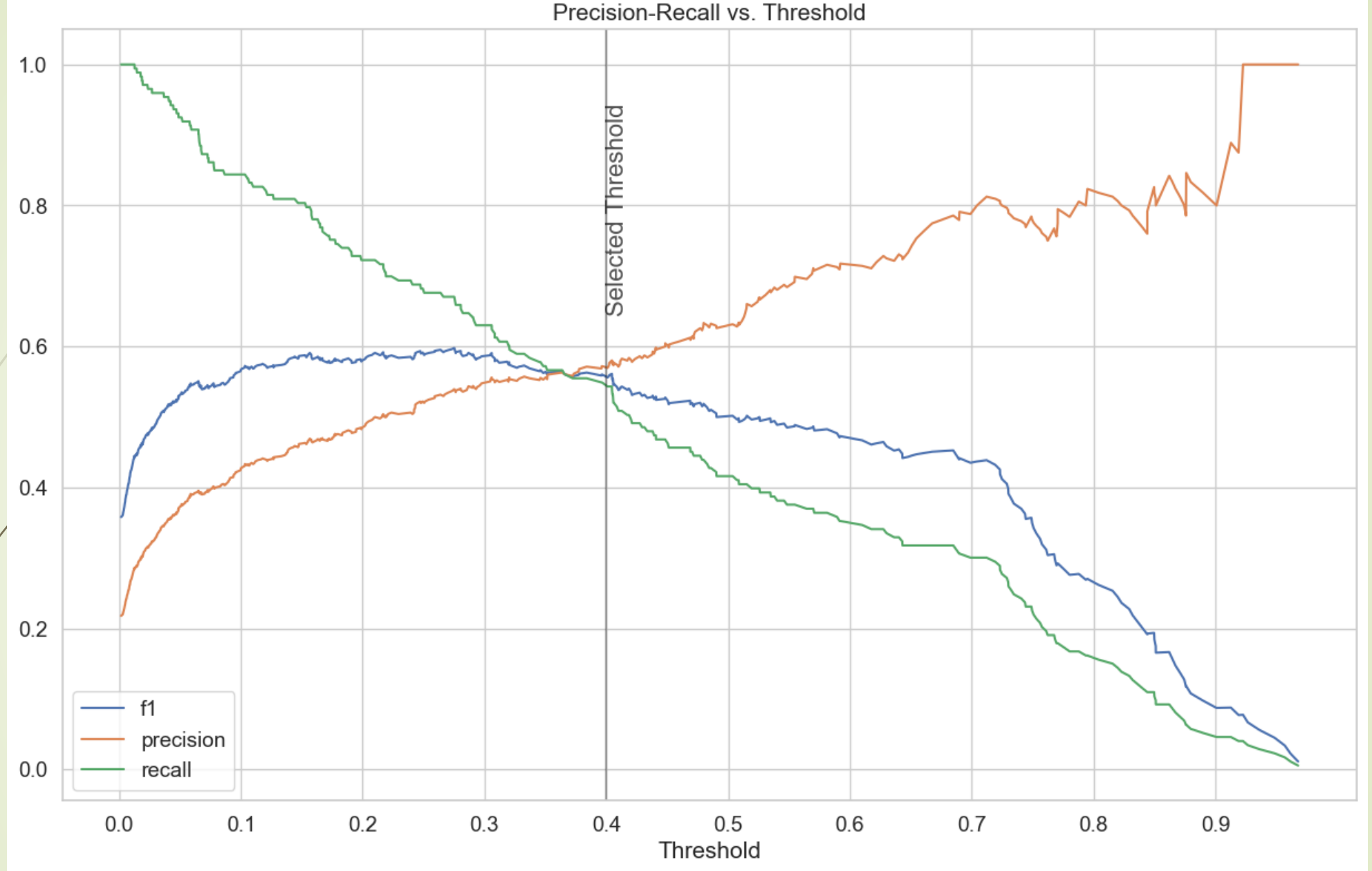
Model Name	f1 Score	Test Accuracy	ROC AUC	Precision	Recall	Winner
Logistic Regression	0.203	0.802	0.832	0.833	0.116	✓ tied
PCA Logistic Regression	0.203	0.802	0.832	0.833	0.116	✓ tied
Random Forest	0.494	0.835	0.865	0.744	0.370	✗
Gradient Boosted Tree	0.540	0.832	0.846	0.672	0.451	✗

White Wine: Large Winery

Model Name	f1 Score	Test Accuracy	ROC AUC	Precision	Recall	Winner
Logistic Regression	0.398	0.821	0.832	0.746	0.272	✗
PCA Logistic Regression	0.398	0.817	0.832	0.706	0.277	✗
Random Forest	0.552	0.687	0.865	0.402	0.884	✗
Gradient Boosted Tree	0.546	0.813	0.846	0.582	0.514	✓



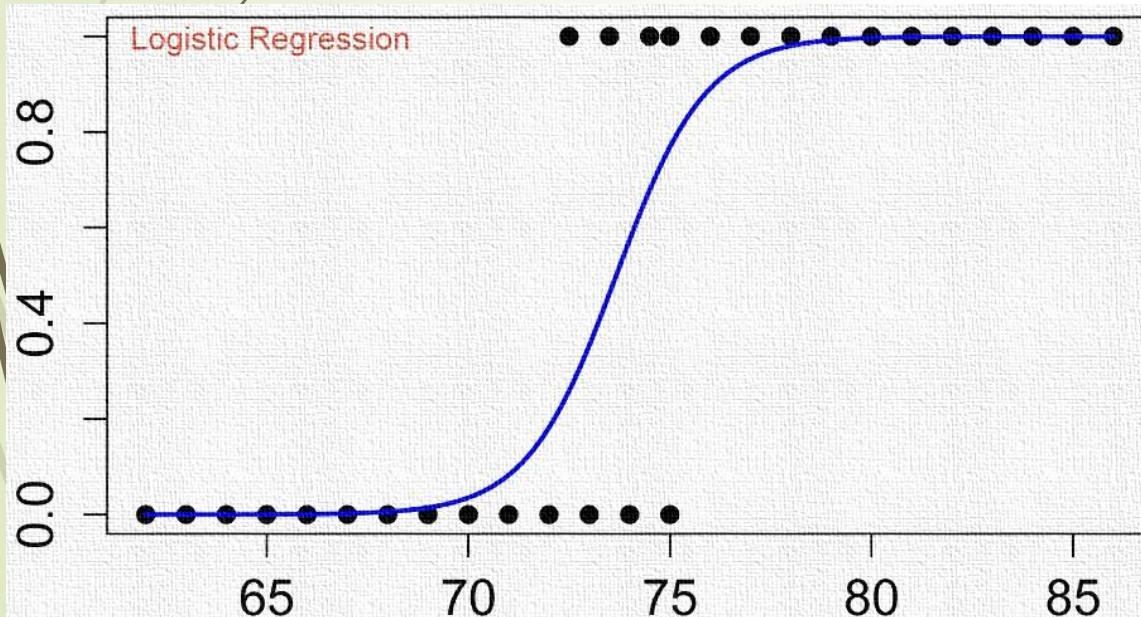
Threshold chart on red wine test data from winning PCA Model for small family-owned business



Threshold chart on white wine test data from winning GBT Model for large winery

Conclusion

- The dataset is made of many duplicates which have been classified as true duplicates and removed prior to modeling.
- The red wines tend to have lower quality than the white wines.
- Logistic Regression-based Models won for the small business thresholding while Gradient Boosted Tree Models won with large wineries.
- The results of this analysis would be useful for owners of small family-owned wine shops looking to retain loyal high paying customers or R&D departments at large wineries looking to maximize sales by reducing product recall.



Ideas for Future Research

Apply New Machine Learning Algorithms

- Neural Nets
- Support Vector Machines

Acquired More Features

- Wine Color
- Wine Price
- Wine Name
- Carbonic Acid Levels (carbonation)
- Cask or Bottle Fermented

Adjust Optimization Metric

Target regression on wine price instead of classification of wine quality

More Robust Labeling of Sensory Data

Wine quality pulled from the median of more than 3 tasters, possibly from different regions or from markets with different palettes

Focus on Improving Red Wines

White wine had proportionally 30% more high quality wines than the red wine data. Efforts could be directed to improving red wine performance.

Cheers!

