WINE PREDICTIONS
MACHINE LEARNING PROJECT

Akash Barathan
Andrew White
Ethan Wong
Maria-Laura Peña
Agnitra Das

CONTENTS

INTRO

What we are predicting and why

DATA

Quick overview of our data

DATA PREP

Cleaning & Transformation

MODEL

Selection, Tuning, and Implementation

EVALUATION

Measures of model performance

FINDINGS

Key insights and conclusion

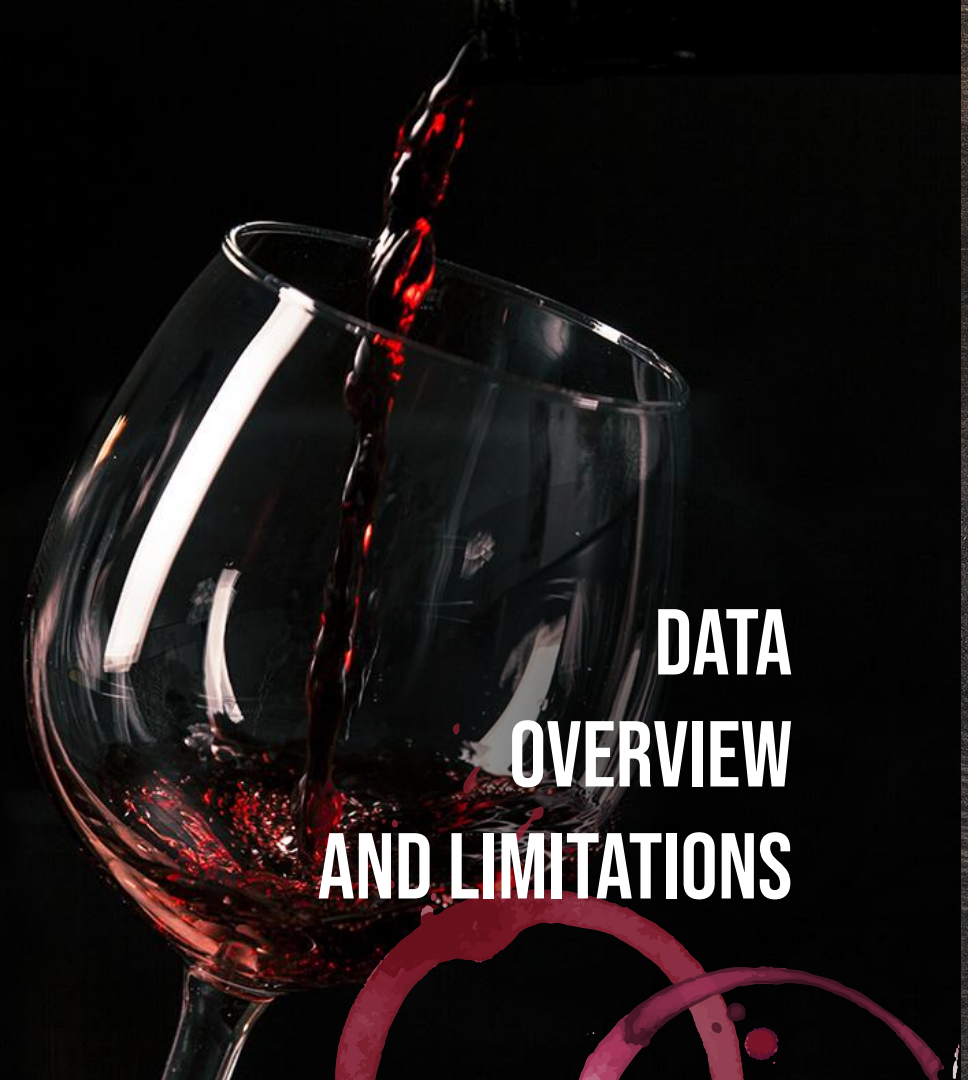
01.
INTRO

WINE!

The goal: Classify wines as either red or white based on various features.

POTENTIAL USE CASES

02.
DATA



DATA OVERVIEW AND LIMITATIONS

Original

Kaggle wine quality dataset

Extension

UCI machine learning repository (wine color)

Size

Original =
1143 rows, 13 columns

With Extension =
6497 rows, 14 columns

VARIABLES

Fixed Acidity
Stable acids affecting wine's taste

Volatile Acidity
Acids contributing to wine's sourness

Citric Acid
Adds a citrus flavor to the wine

Residual Sugar
Sugar left after fermentation, affects sweetness

Chlorides
Salt content, influences taste and mouthfeel

Free Sulfur Dioxide
Preserves wine, prevents spoilage

Density
Mass per volume, affects wine's body

pH
Acidity level, impacts taste and stability

Total Sulfur Dioxide
Overall sulfur content, used as preservative

Sulphates
Enhance flavor and stability, impact taste

Alcohol
Ethanol content, affects body and flavor

Color_White
TARGET VARIABLE

03.
DATA PREP

Cleaning & Transformation

Duplicates & Null

Dropped duplicate rows (1177) and checked for null values (found none)

Drops

We dropped the quality and ID variables (reducing columns)

New shape:

5320 rows, 12 columns

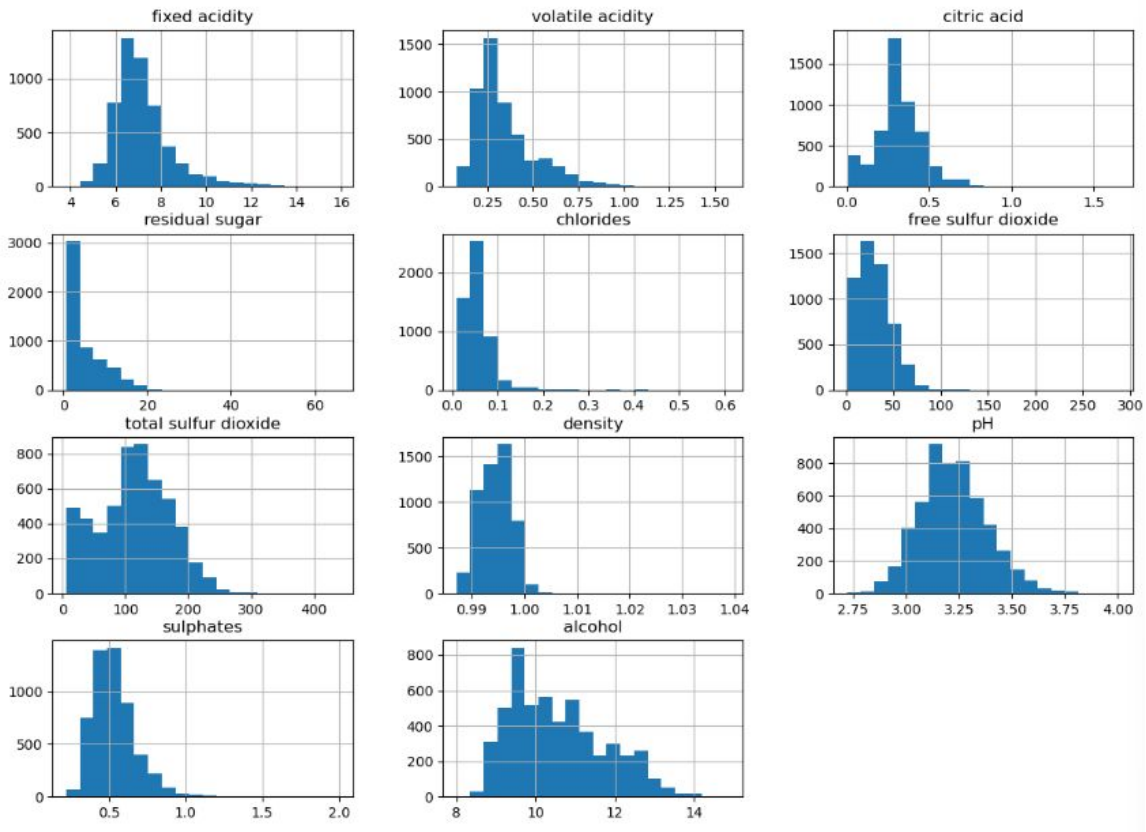
Scaling

For logistic regression and K-NN to make variables comparable

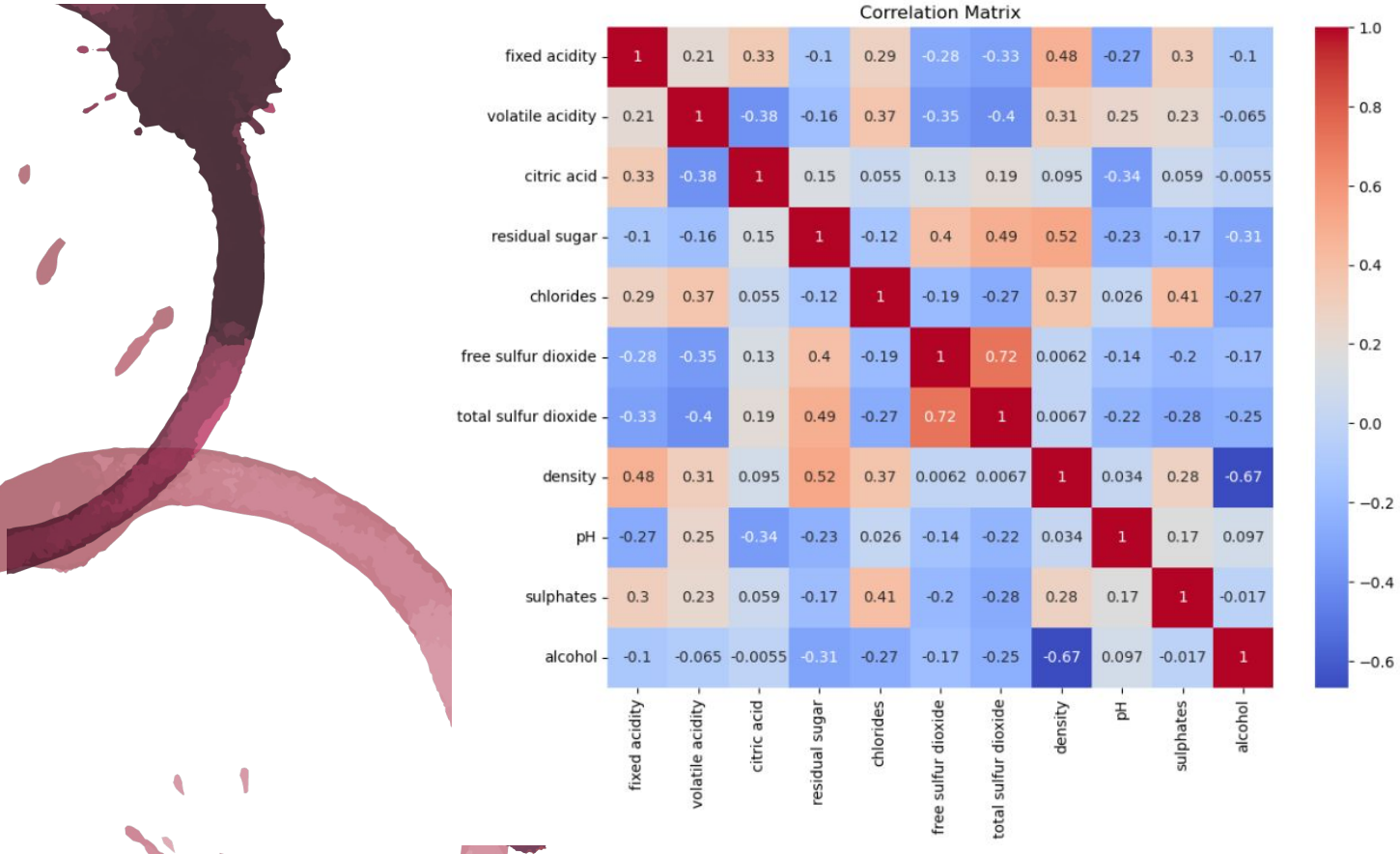
Binary

Made the target variable (Red or White) a binary variable

Data Distribution



Correlation Matrix of Predictors

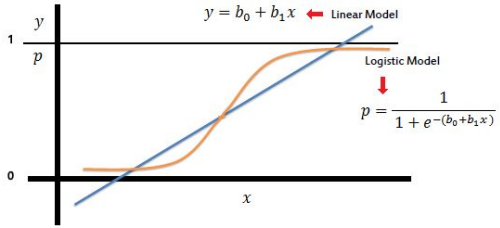


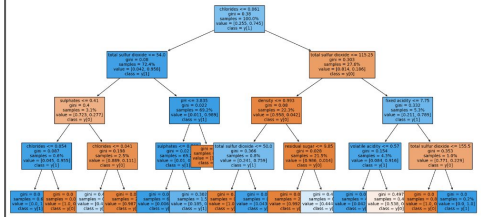
04.
MODEL

Some models we tried

Logistic Regression

Predicts a binary outcome with probabilities. Simple to interpret, efficient to run, and robust to noise.



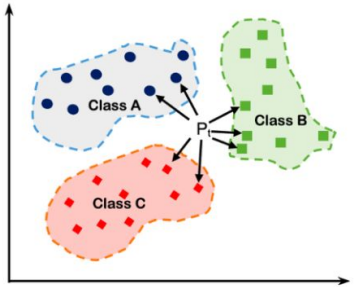


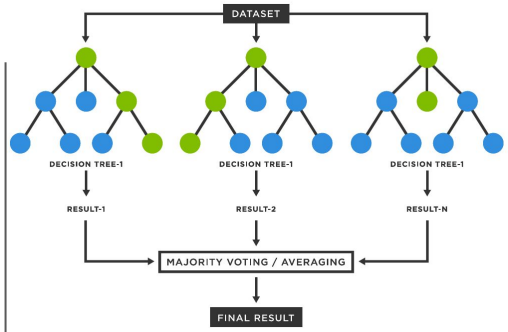
Decision Tree

Simple classification model that splits into more nodes at each depth until maximum depth is reached

K-NN

Classifies based on nearest neighbors in feature space.





Random Forest

Averages predictions from an ensemble of trees, using subsets of predictors to reduce variance

TUNING

Applied Bagging and Boosting techniques to improve performance

Conducted 10-fold Stratified Cross-Validation to optimize results

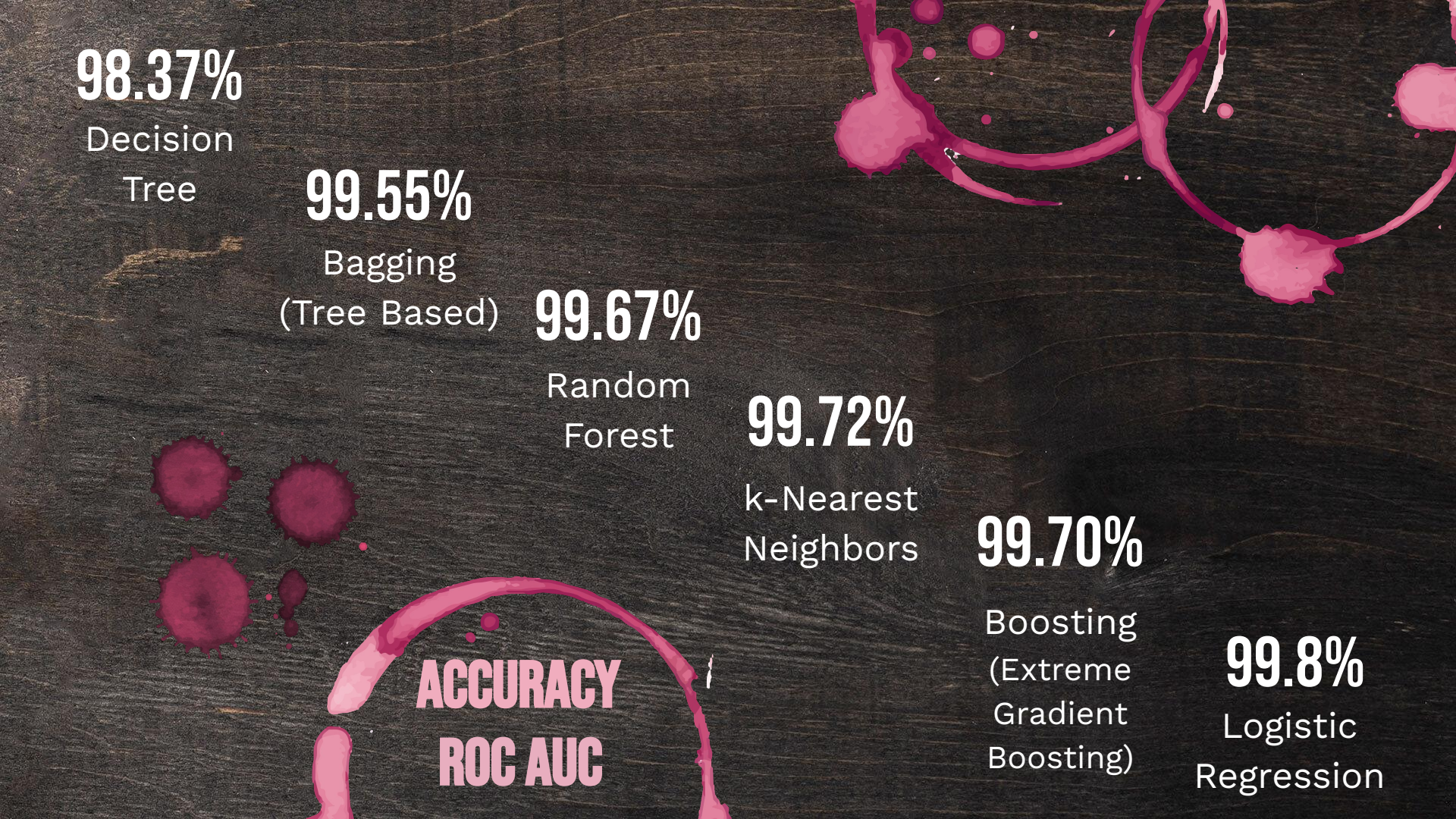
Stratified Cross-Validation to handle imbalanced data

Optimized for the highest ROC AUC

Evaluated various configurations:

- Number of estimators in ensemble methods
- Learning rates
- Max depth in tree-based methods
- Number of nearest neighbors

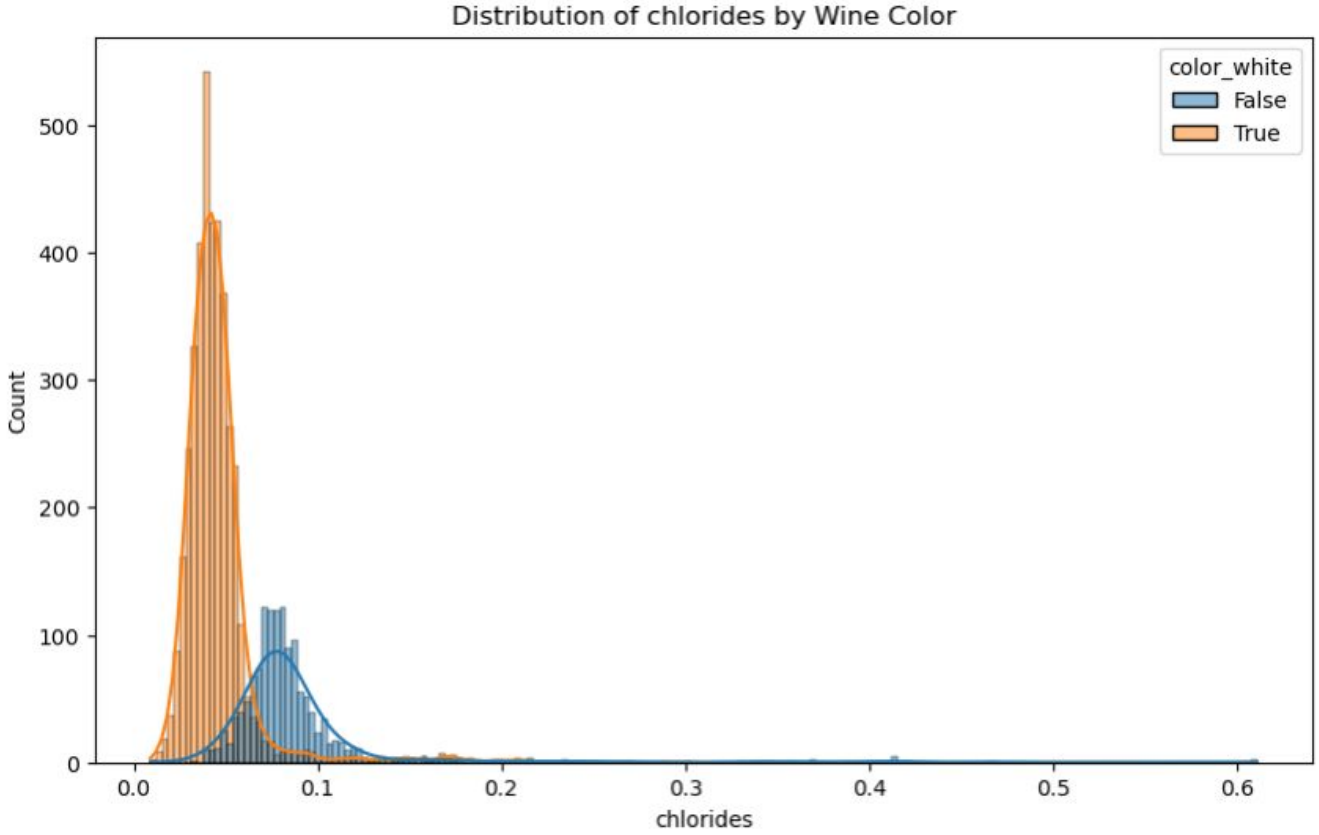
05.
EVALUATION


98.37%
Decision Tree
99.55%
Bagging (Tree Based)
99.67%
Random Forest
99.72%
k-Nearest Neighbors
99.70%
Boosting (Extreme Gradient Boosting)
99.8%
Logistic Regression
ACCURACY
ROC AUC

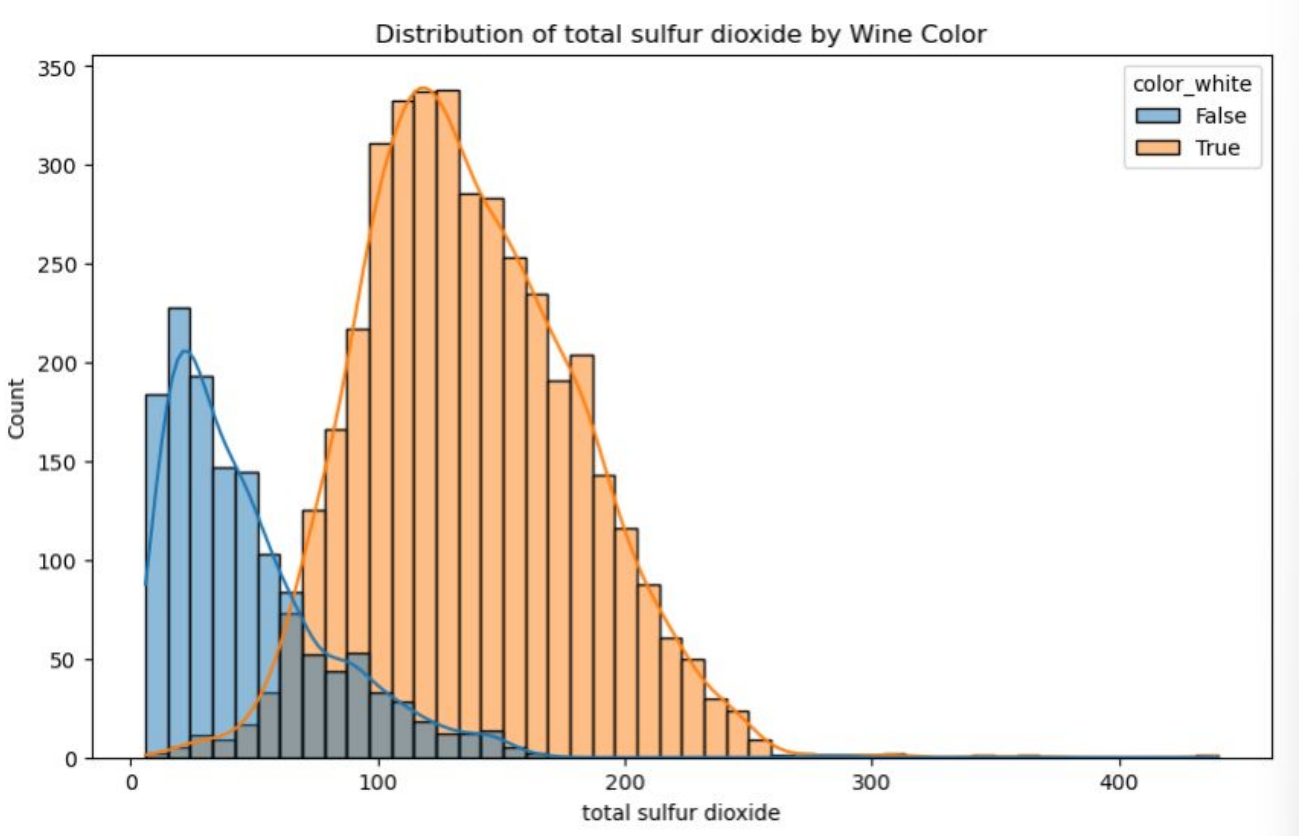
COMPARING MODELS

Model	Accuracy	ROC AUC	Cohen's Kappa	Log Loss	TPR	TNR
Logistic Regression	0.9937	0.9981	0.9836	0.0303	0.9946	0.9911
Decision Tree	0.9829	0.9837	0.9549	0.0733	0.9908	0.9599
Bagging	0.9943	0.9955	0.985	0.0892	0.9969	0.9866
Random Forest	0.9949	0.9967	0.9865	0.0691	0.9977	0.9866
KNN	0.992	0.9972	0.9791	0.0651	0.9939	0.9866
XGBoost	0.9954	0.997	0.988	0.0217	0.9962	0.9933

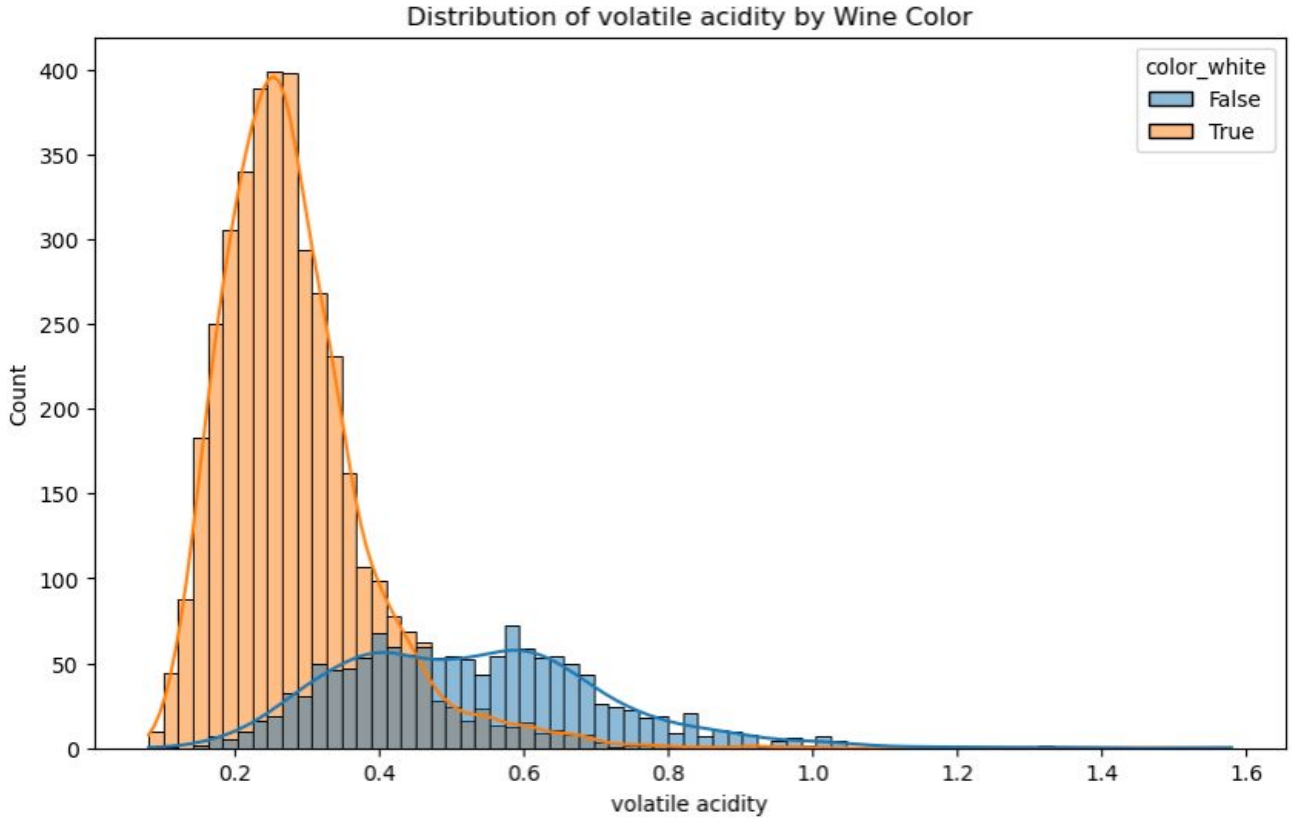
Distribution Plots



Distribution Plots



Distribution Plots



06. FINDINGS

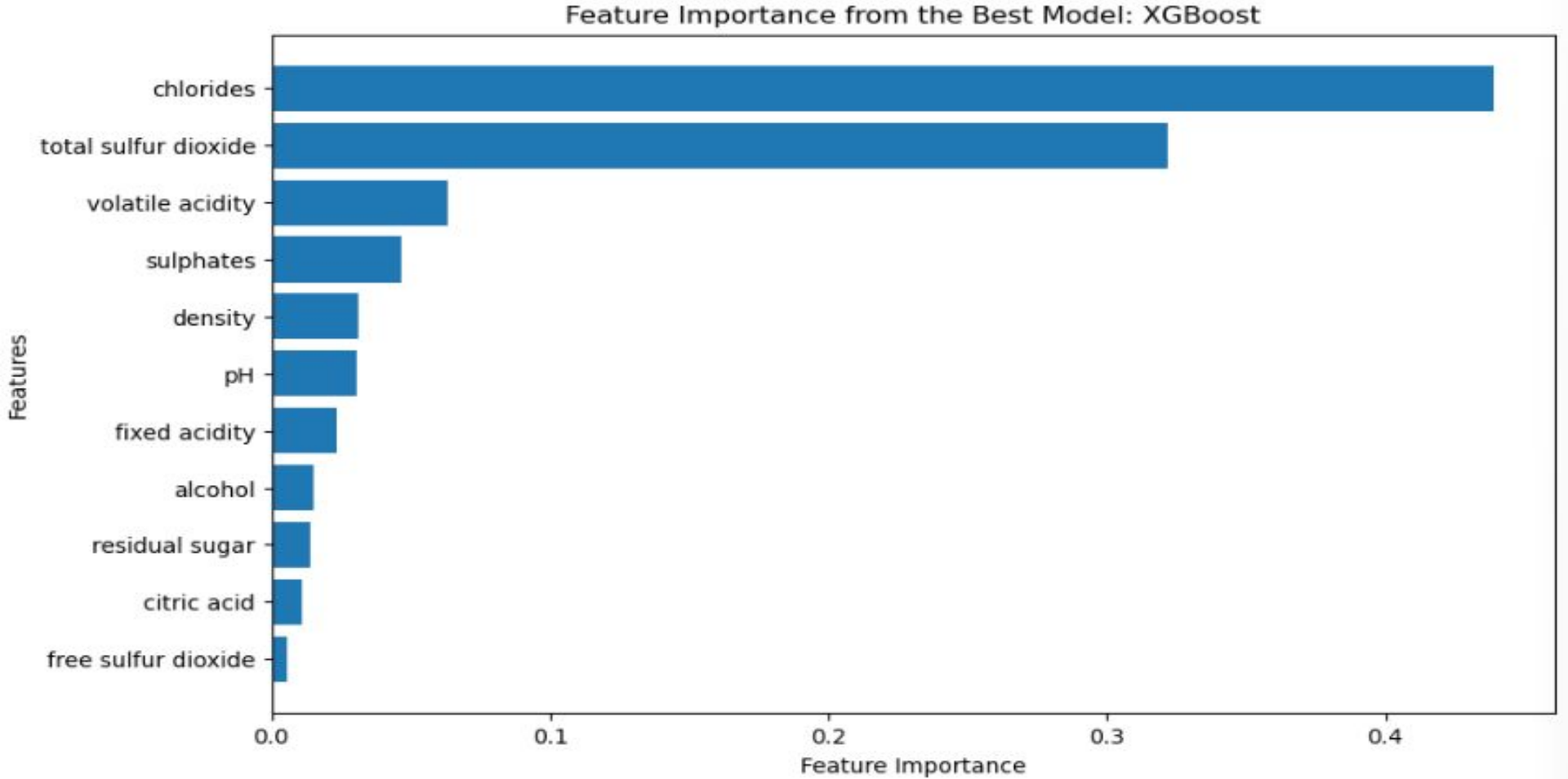
BEST MODEL: XGBOOST

Highest Accuracy

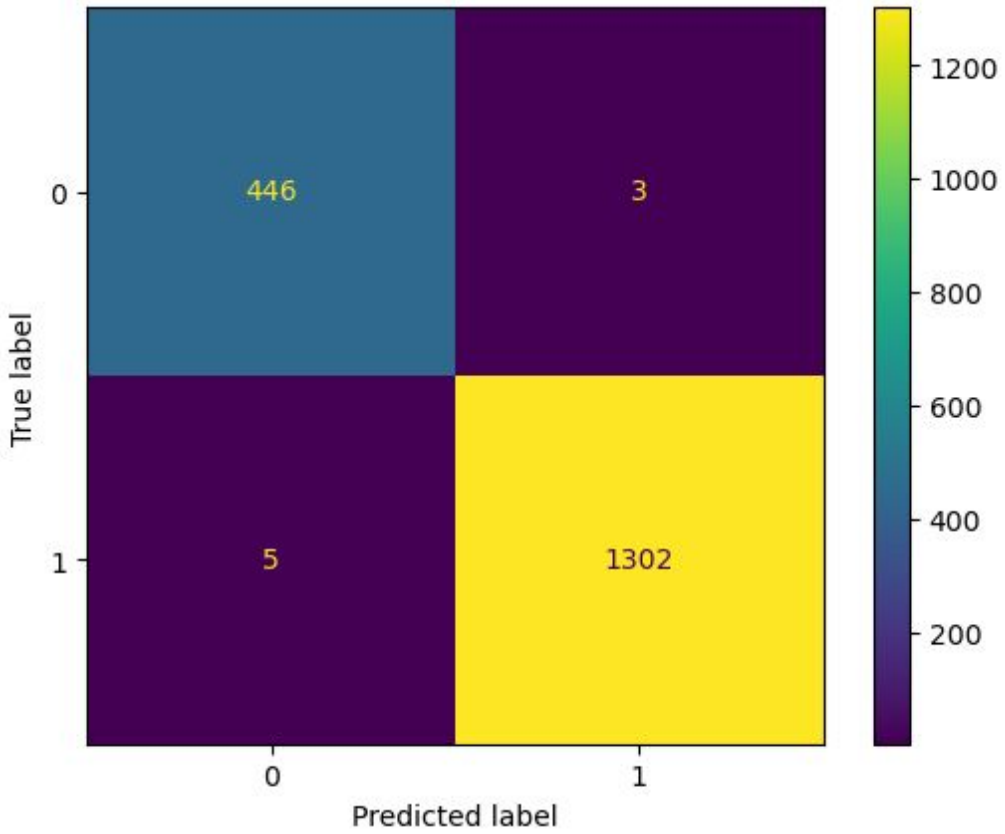
Lowest Log Loss

Comparable ROC AUC, TPR, and TNR

FEATURE IMPORTANCE



Confusion Matrix



IMPLICATIONS & TAKEAWAYS

Extension

To regions across the world

Artificial wine

Perfect imitations of flavor

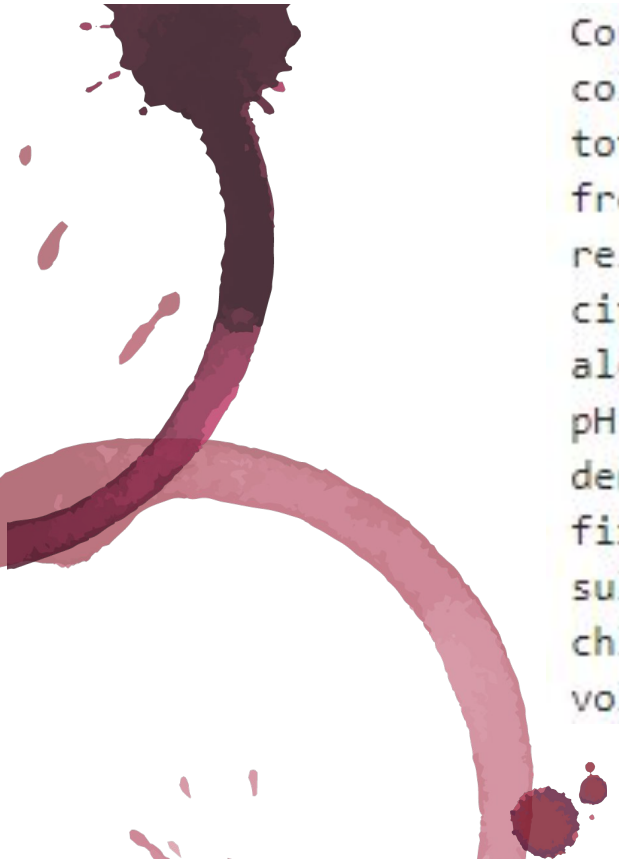
Attributes

Attributes of white vs. red wine


QUESTIONS?

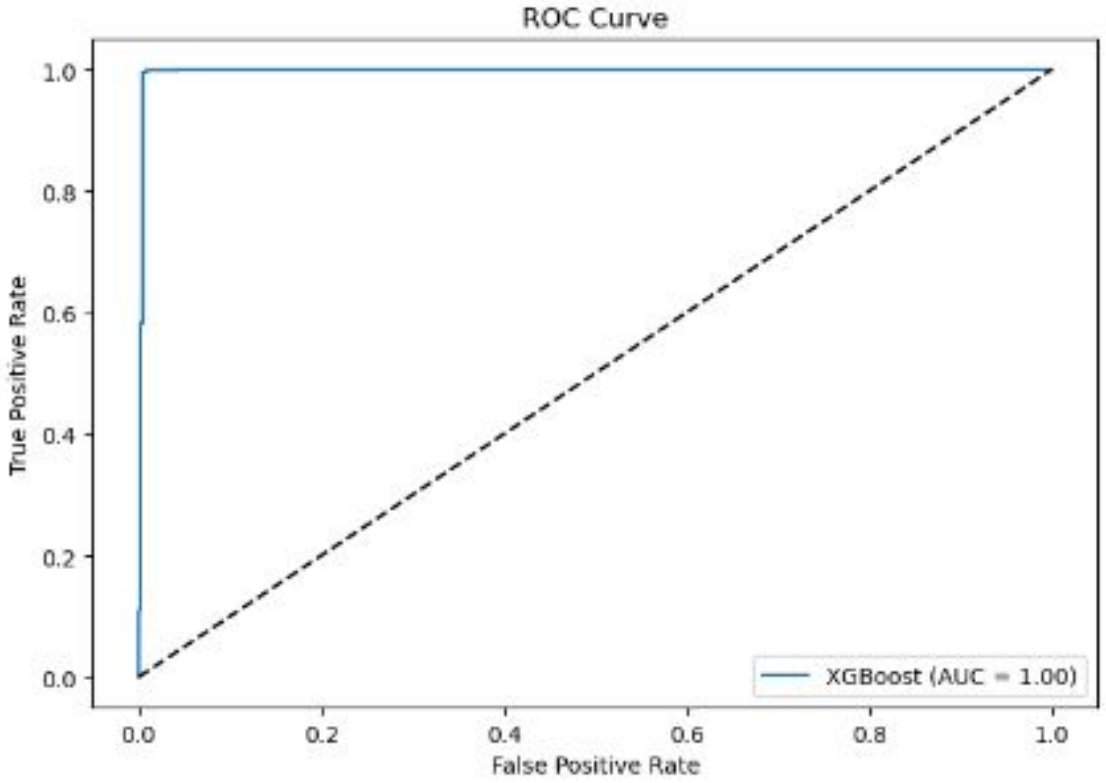

APPENDIX

Target Variable Correlations

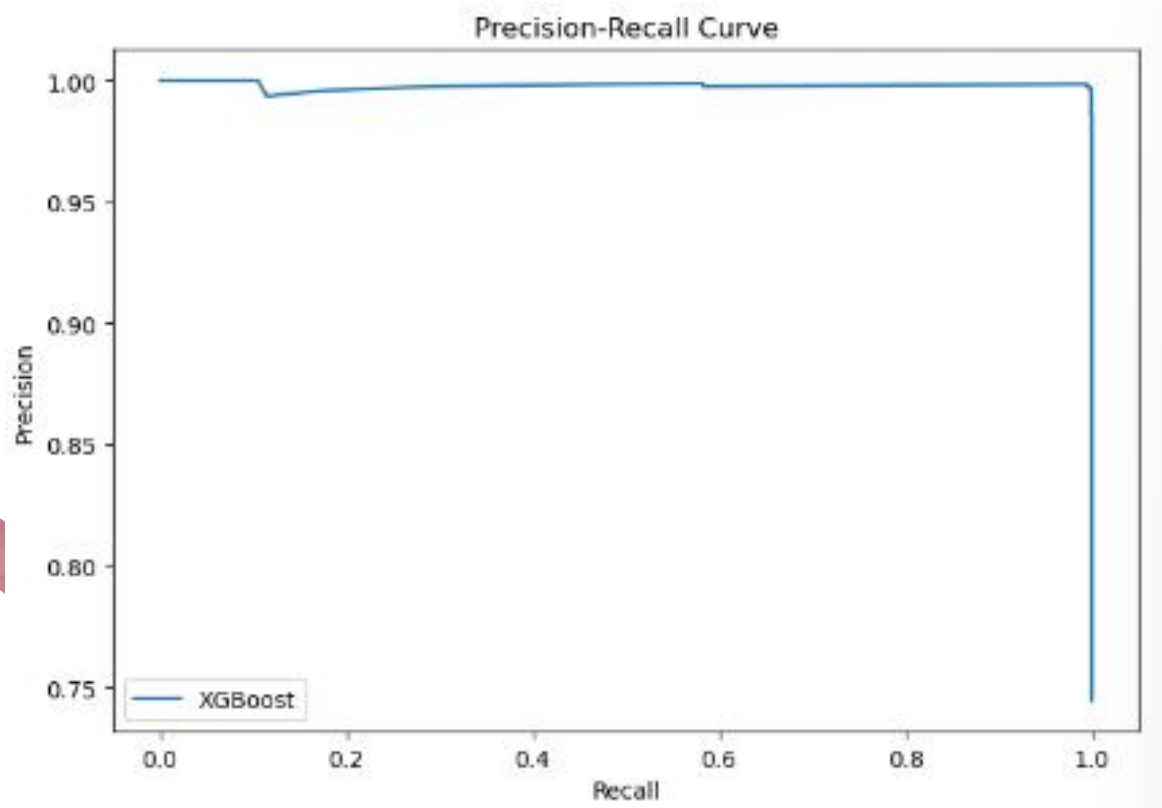


Correlation with Target Variable:	
color_white	1.000000
total sulfur dioxide	0.694229
free sulfur dioxide	0.465326
residual sugar	0.328695
citric acid	0.183759
alcohol	0.057756
pH	-0.310919
density	-0.429377
fixed acidity	-0.486253
sulphates	-0.490364
chlorides	-0.499517
volatile acidity	-0.645335

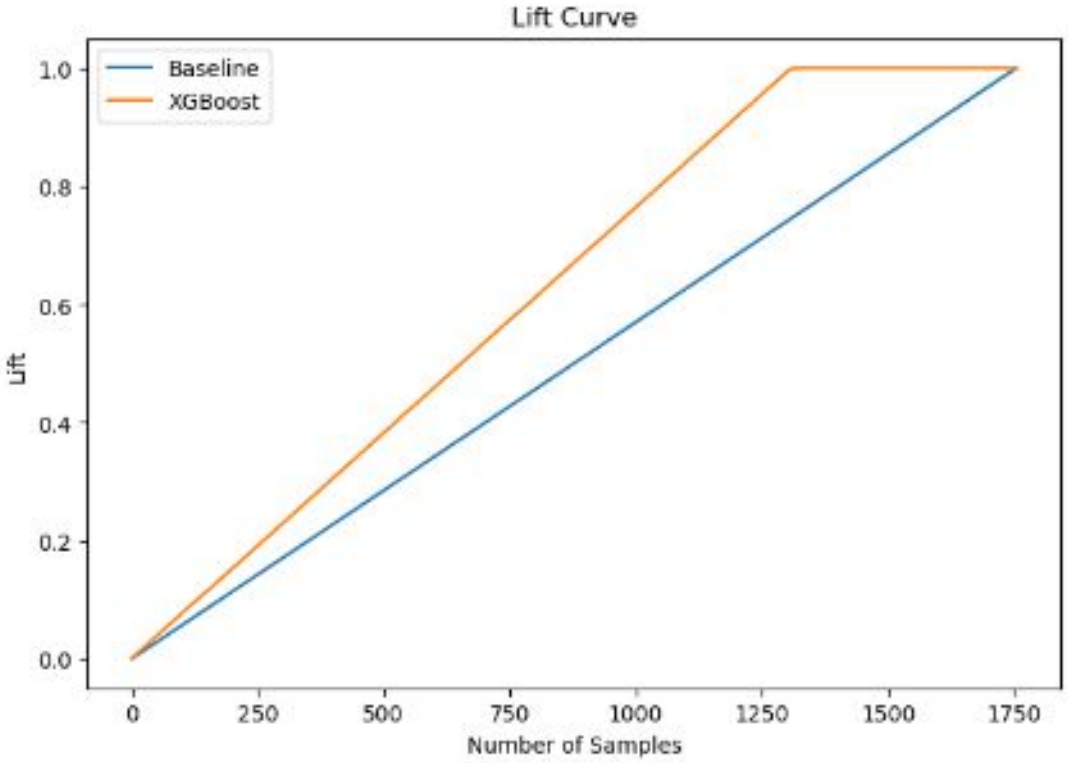
ROC Curve -XGBoost



Precis. Recall Curve - XGBoost



Lift Curve - XGBoost



PCA Distribution