Classifying Safe Water Sources

Potential Potables

By: William Grennan

Earth’s Most Valuable Resource

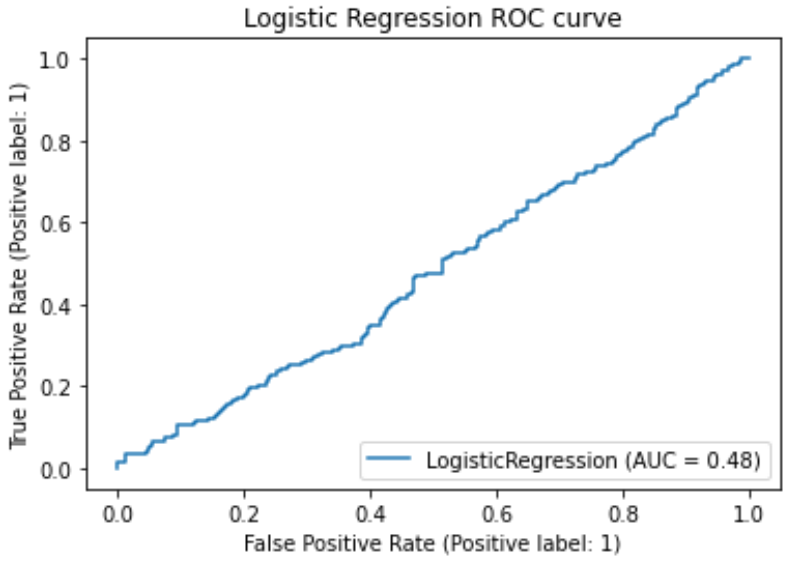
- Fresh Water accounts for only ~2.5% of all water on earth
- But not all fresh water is safe to drink
- Water Pollution takes many forms
  - pH imbalance
  - Clarity
  - Toxic substances
  - Etc.

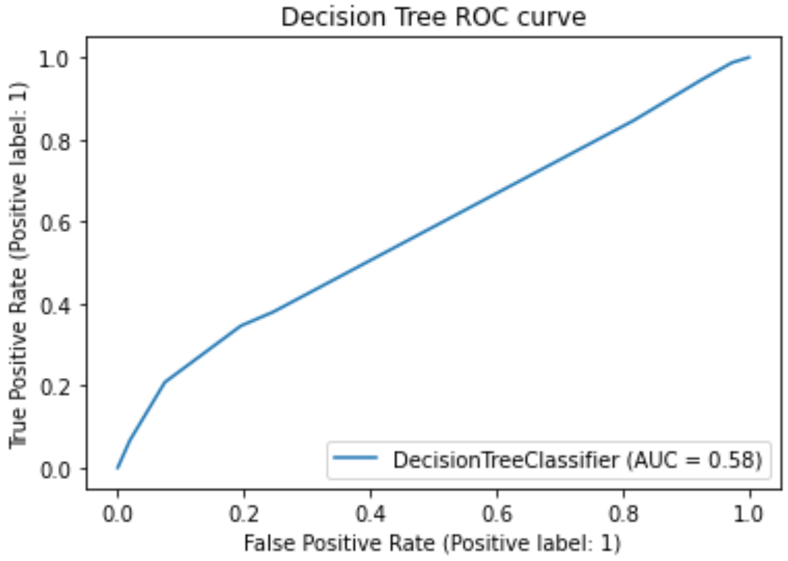
How Can Data Science Help?

- 9 key measurements can be taken from the water sources quickly
  - pH, Hardness, Dissolved Solids, Chloramines, Dissolved Sulfates, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity
- Using these measurements, we can narrow the search
- 2 Objectives depending on need
  - **Precision – Quickly find water most likely to be safe**
  - Recall – Identify sources that were previously overlooked

What Successes Can We Find?

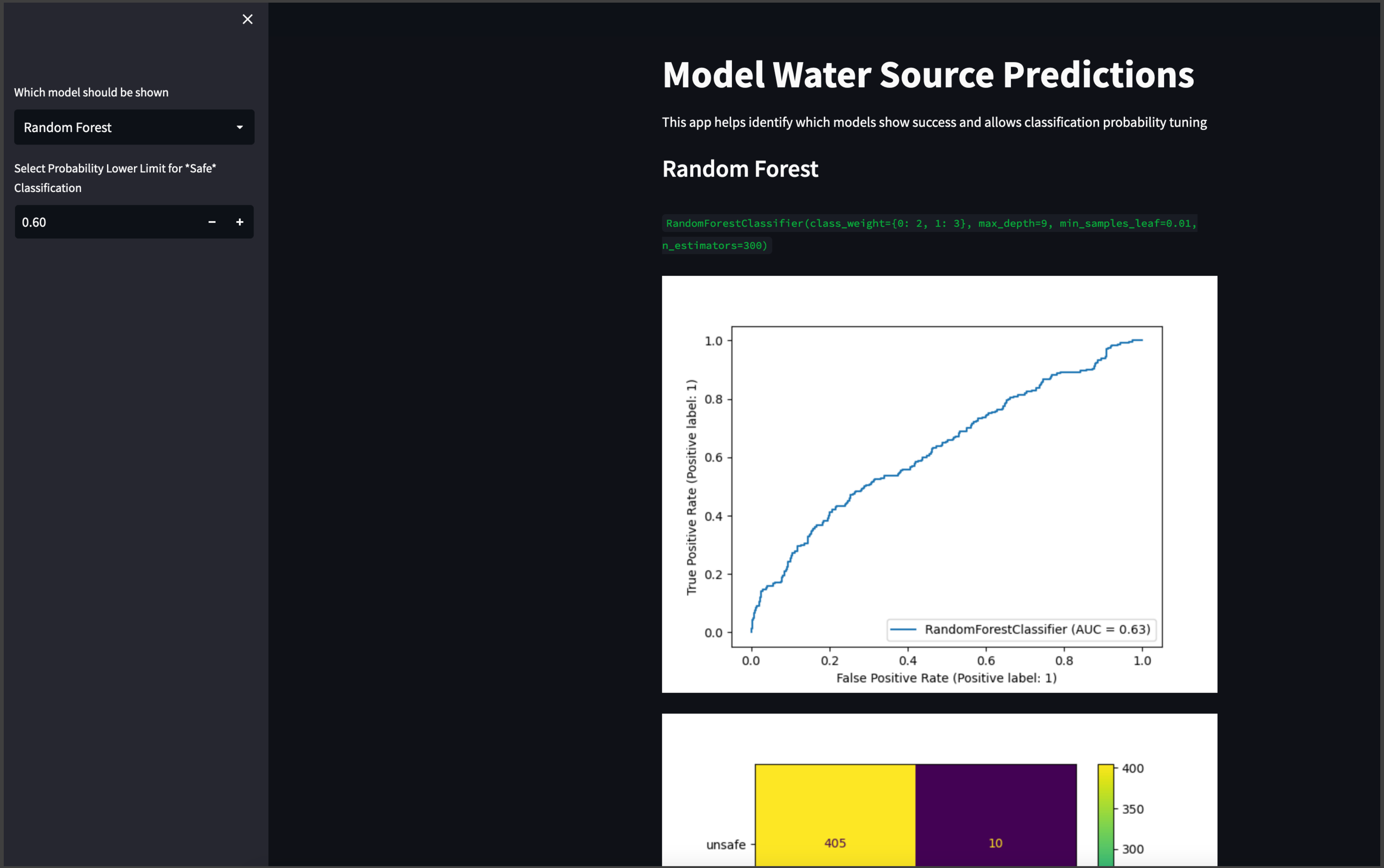
- Data modeling can be tricky
- Lots of measurement overlap between the classes
- Decision Trees and Forests tend to perform better than Logistic Regression


Logistic Regression ROC curve


Decision Tree ROC curve

Directly Compare the Models in Streamlit

- Selectable Model
- View Model Hyperparameters
- Viewable ROC Chart
- Adjustable Confusion Matrix



Which model should be shown

Random Forest

Select Probability Lower Limit for *Safe* Classification

0.60

Model Water Source Predictions

This app helps identify which models show success and allows classification probability tuning

Random Forest

```
RandomForestClassifier(class_weight={0: 2, 1: 3}, max_depth=9, min_samples_leaf=0.01,
n_estimators=300)
```

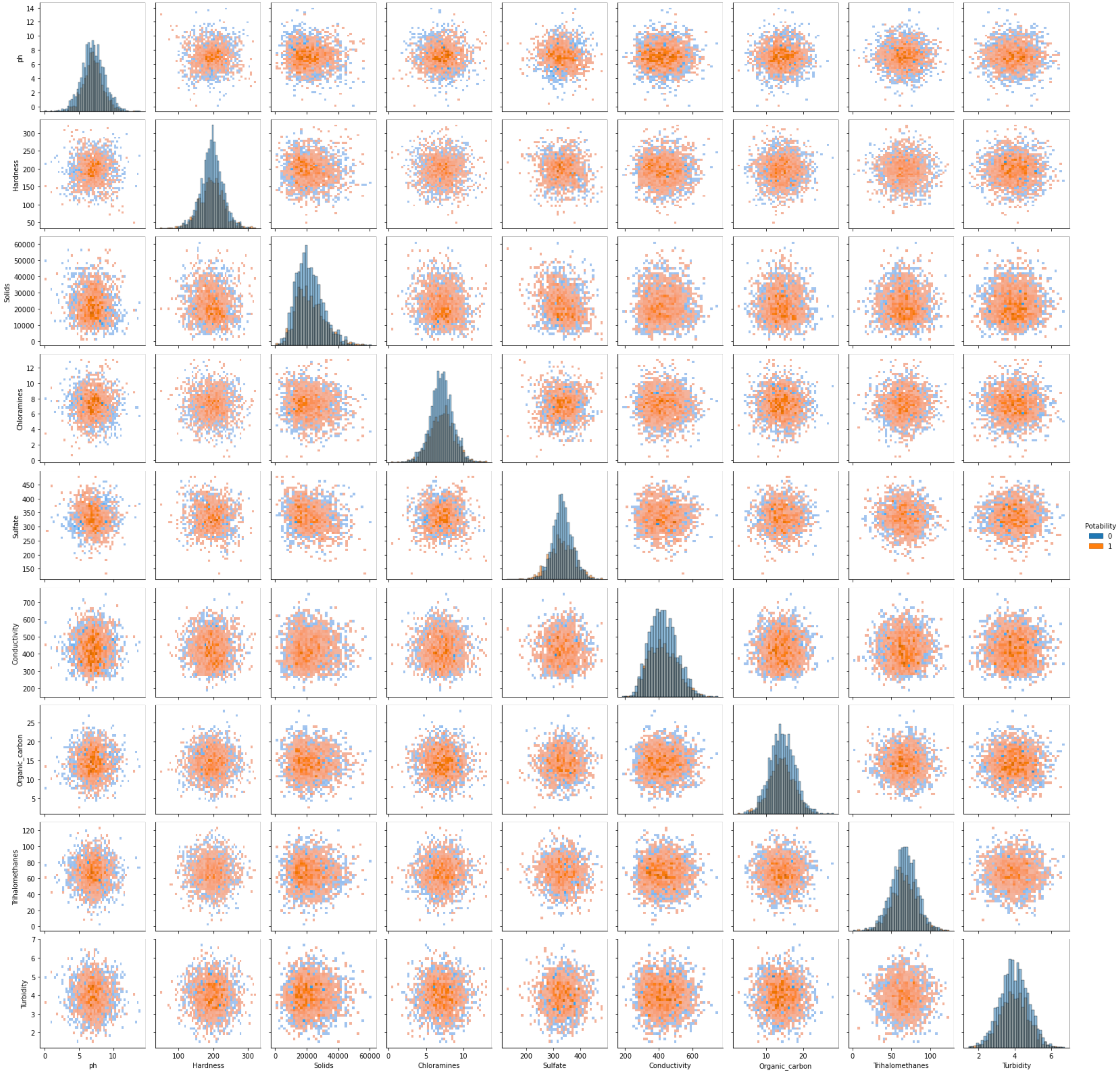
×
−
+
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)
RandomForestClassifier (AUC = 0.63)
unsafe
405
10
400
350
300

Conclusions

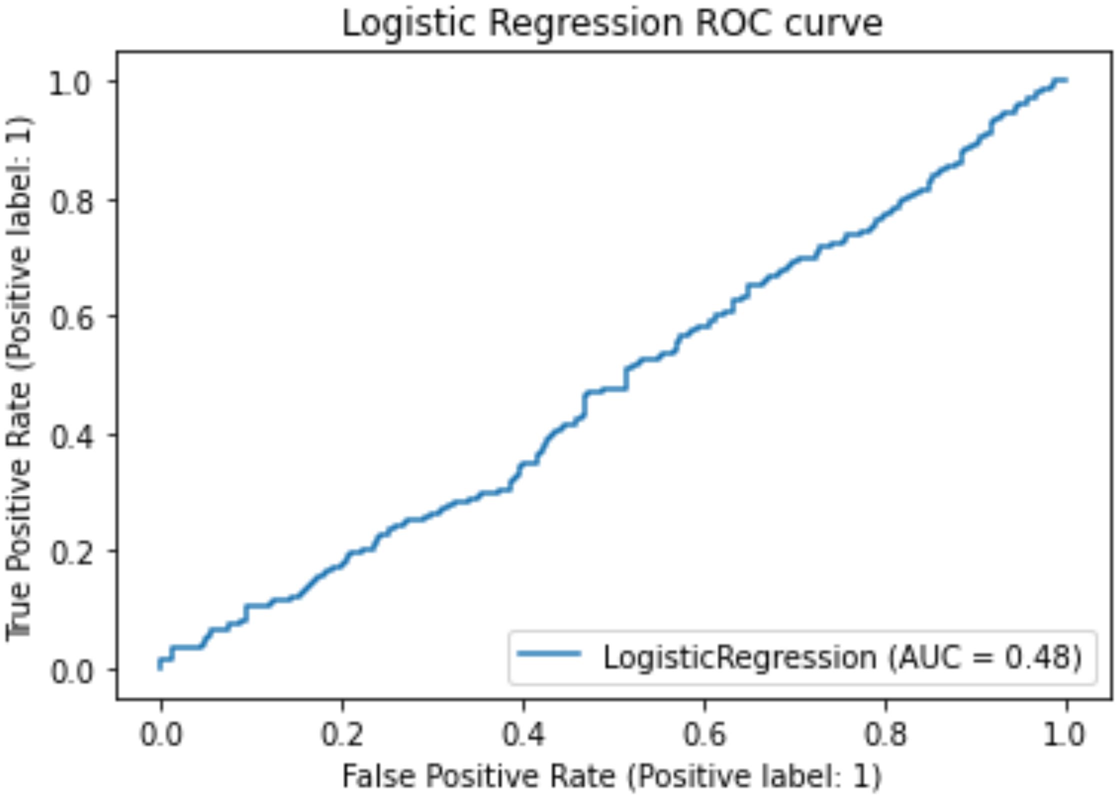
- XGBoost and Random Forests perform the best on test data
- Current modeling is too risky for a definitive answer
- Further scientific study may find more reliable measurements

Testing Data	Accuracy	ROC AUC
XGBoost	61.28 %	57.25 %
KN Neighbor	60.82 %	56.26 %
Random Forest	64.02 %	60.82 %

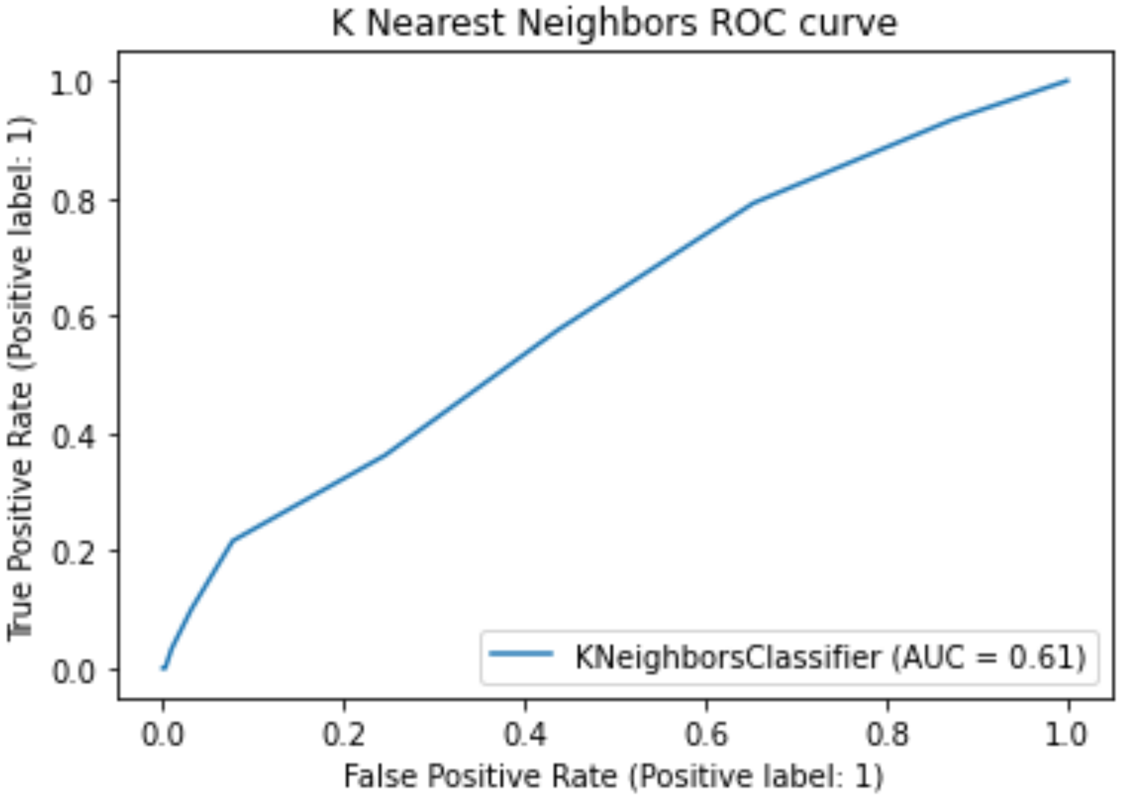
Appendix (Pair Plot)



Appendix (ROC Curves)

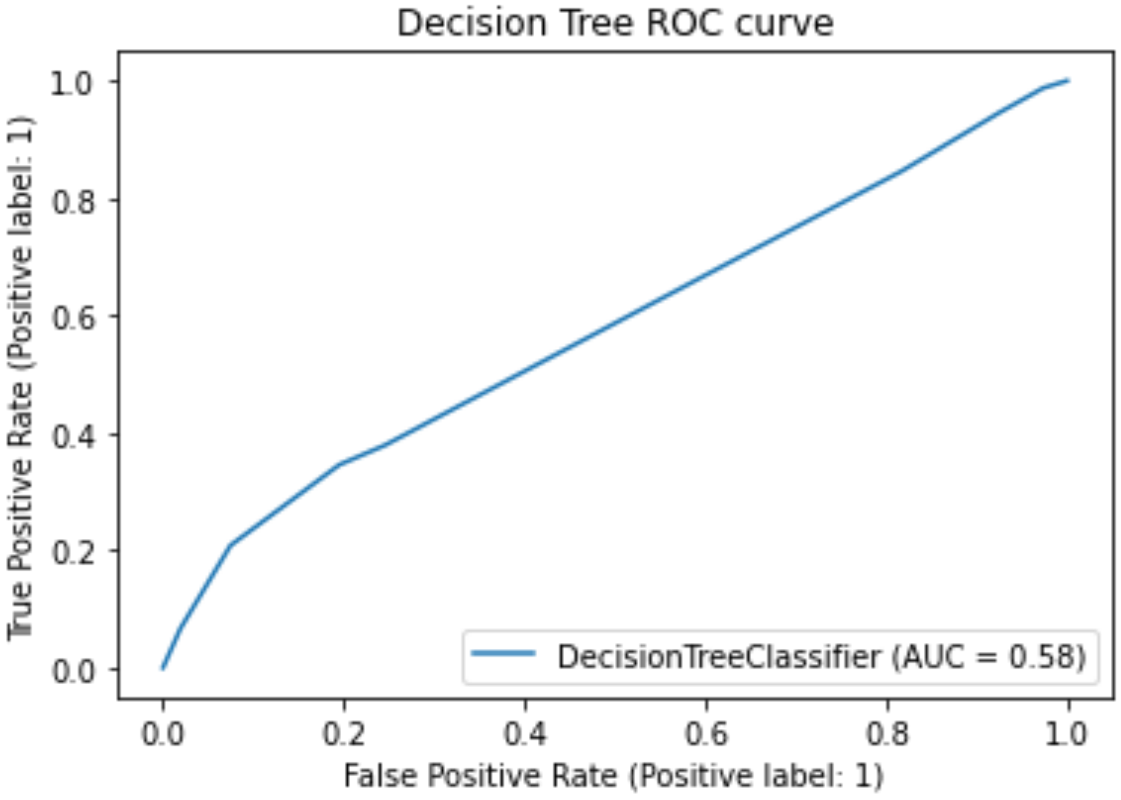


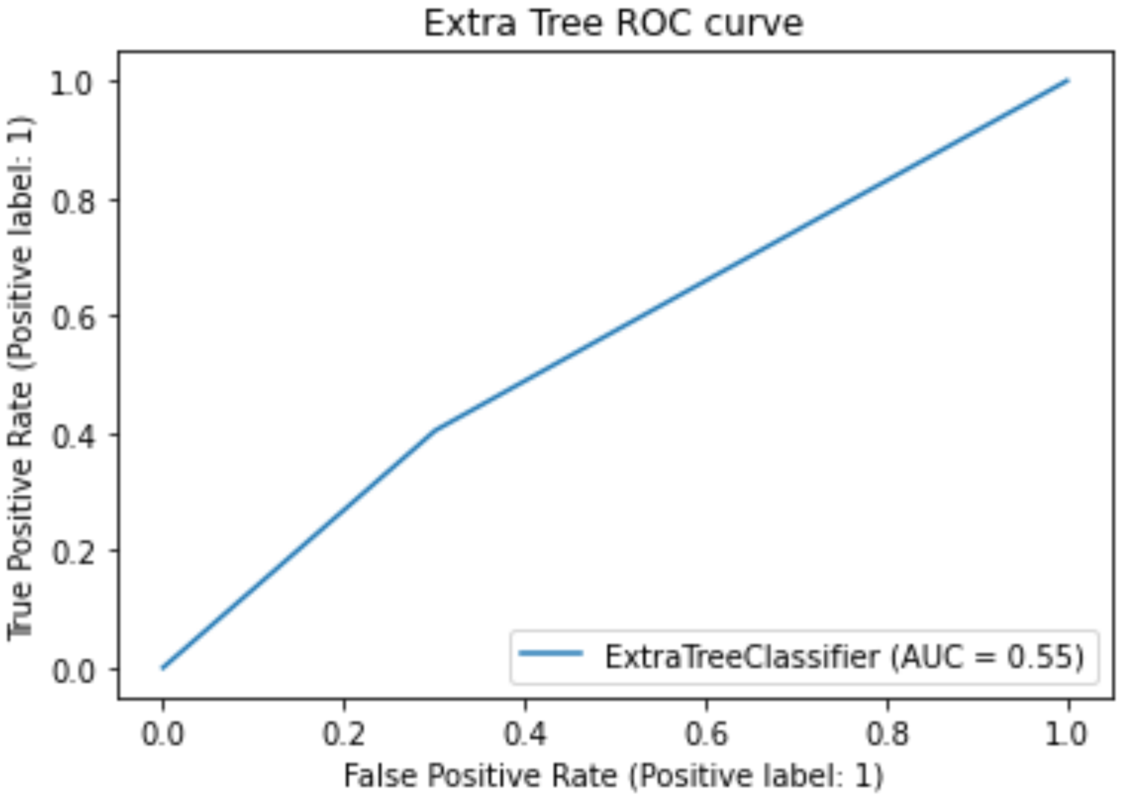
Logistic Regression ROC curve



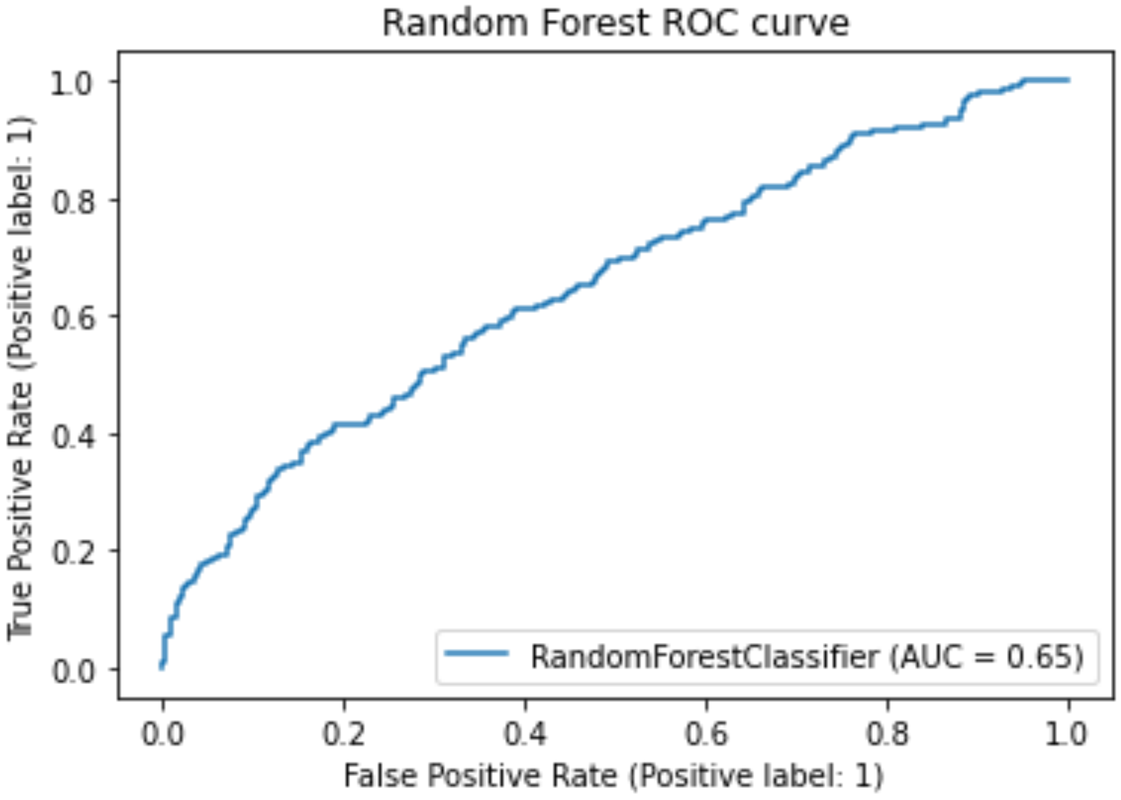
K Nearest Neighbors ROC curve

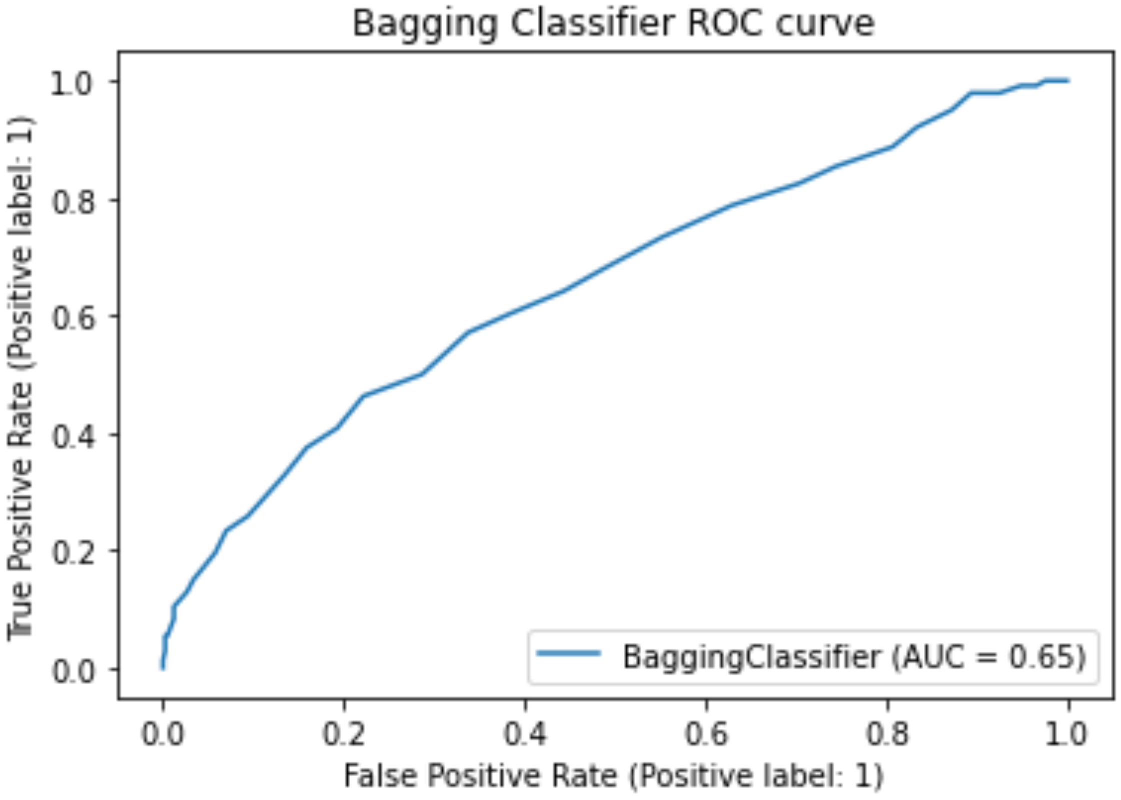
Appendix (ROC Curves)


Decision Tree ROC curve

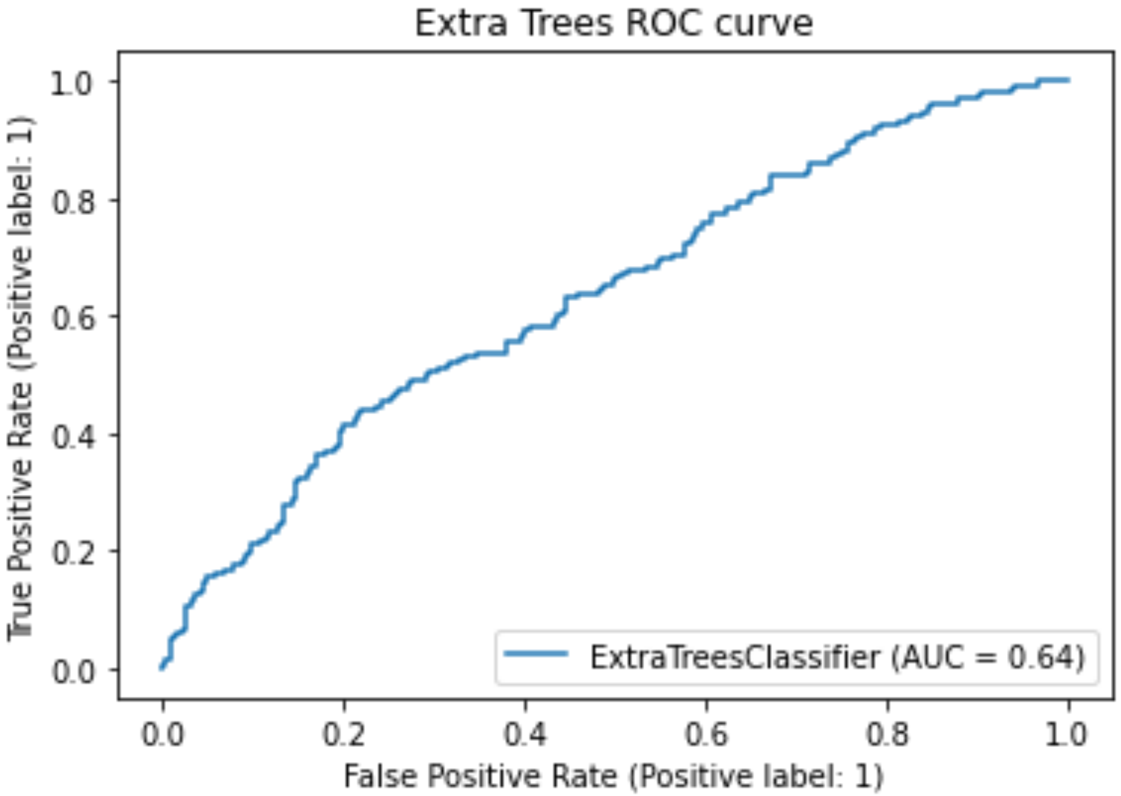

Extra Tree ROC curve

Appendix (ROC Curves)

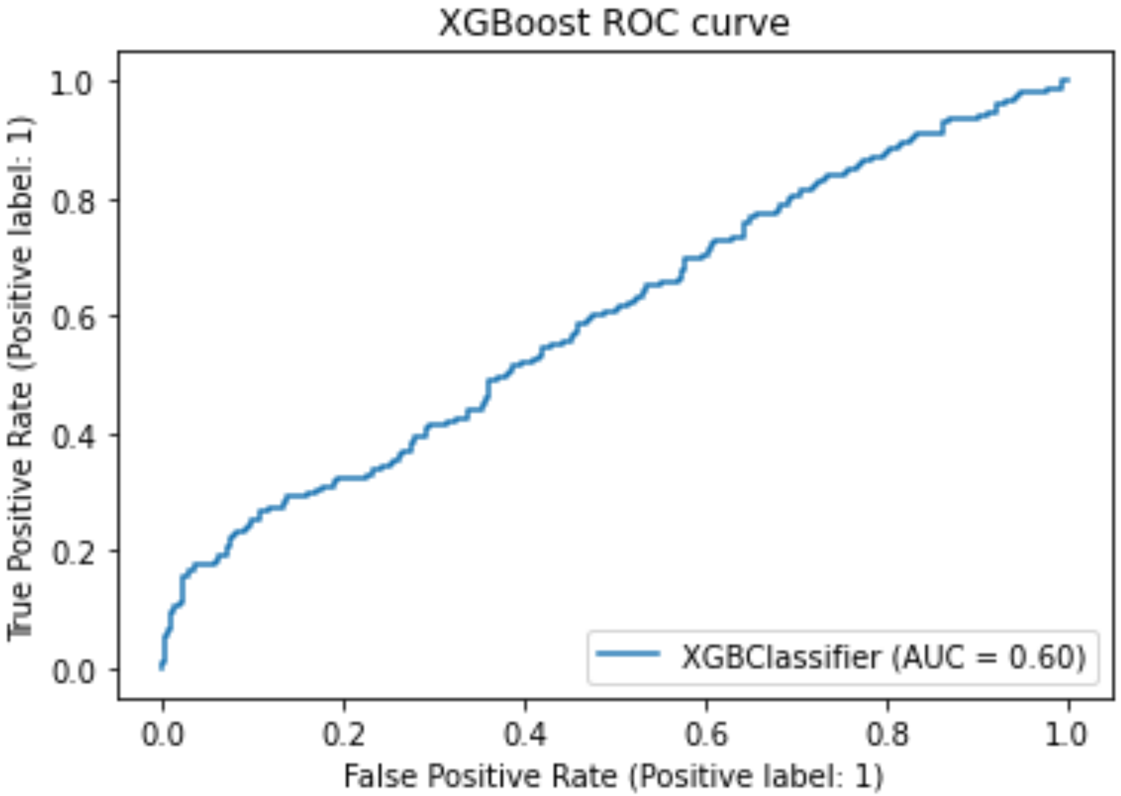




Appendix (ROC Curves)



Extra Trees ROC curve



XGBoost ROC curve