Water Potability

Team 9: Mild Trakarnsakdikul, Sylvie Zhou,
Jason Hamilton, Faarid Sanaan

Team 9 Introduction

Mild Trakarnsakdikul

Jason Hamilton

Sylvie Zhou

Faarid Sanaan

Water is considered “the most important resource for sustaining ecosystems, which provide life-supporting services for people, animals, and plants.”

—The Centers for Disease Control and Prevention (CDC)

Business Understanding: Clean Water

Why clean water is a necessity, according to the UN:

- Sustainable development
- Socio Economic development
- Energy and food production
- Health and survival
- Healthy ecosystems

Business Understanding: Measurement

POTABLE WATER = CLEAN WATER

- The United Nation adopted Goal 6: Access to Clean Water Supply and Adequate Water Sanitation in their “Sustainable Development Goals”
  - Testing for water potability using coliform bacteria test
  - Improve water quality by:
    - Reducing pollution
    - Eliminating dumping
    - Increase water recycling

Business Problem: Predicting Potability

Objective: Build a predictive model to determine whether or not the water is potable

- Potable water or drinking water comes from surfaces and ground sources, then is treated to safe levels

Data Understanding: Water Features

Our 9 features were presented as numeric values:

- **pH of water:** the acidity or basic of water
  - 0 (Basic) to 14 (acidic)
- **Hardness:** capacity of water to precipitate soap (mg/L)
- **Solids:** total dissolved solids (ppm)
- **Chloramines:** Amount of Chloramines (ppm)
- **Sulfate:** Amount of Sulfates dissolved (mg/L)
- **Conductivity:** Electrical conductivity of water (µS/c)
- **Organic Carbon:** Amount of organic carbon (ppm)
- **Trihalomethanes:** Amount of Trihalomethanes (µg/L)
- **Turbidity:** Measure of light emitting property of water (NTU)

Data Understanding: Dataset

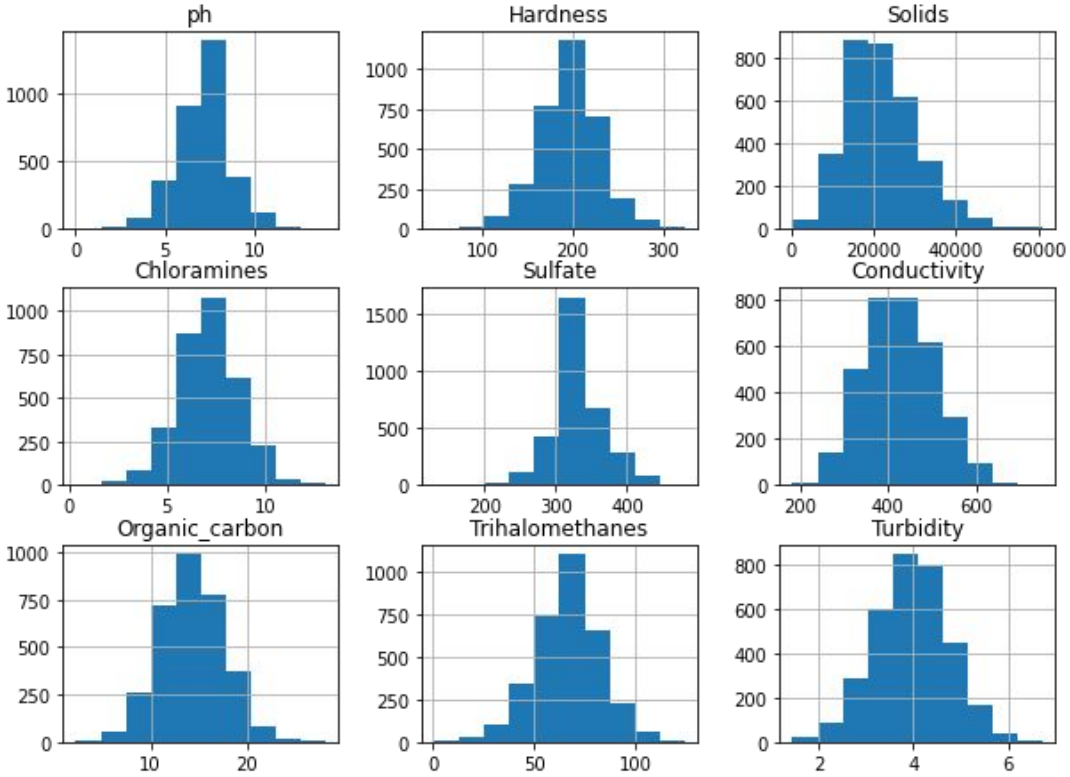
The data contain 3,276 samples of different water bodies:

- Target variable: **potability**
  - 1,998 water bodies are not potable
  - 1,278 water bodies are potable
  - Ratio is 5:3 to the majority class
- Missing values filled using the data mean
  - pH: 491 missing values
  - Sulfate: 781 missing values
  - Trihalomethanes: 162 missing values

Exploratory Data Analysis

Features Distribution

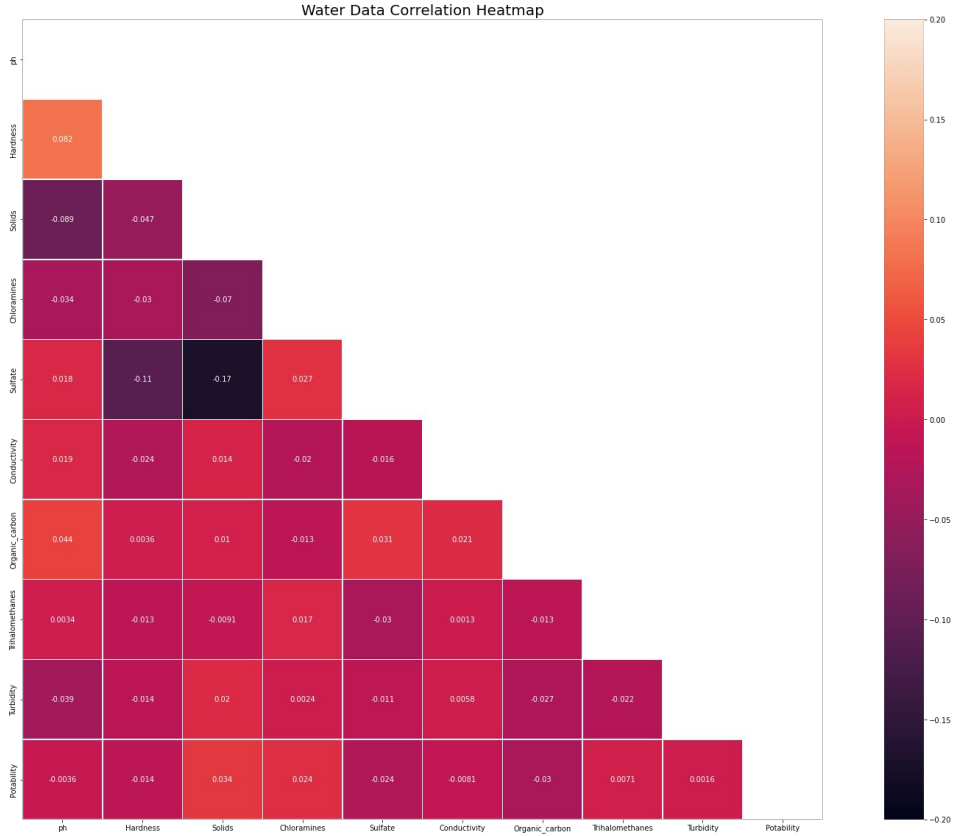
- Normally distributed
- No transformations required



Exploratory Data Analysis

Our Correlation Heatmap:

- Little to no correlation
- Effect on model performance



Data Preparation: Feature Engineering

Used polynomial features to generate new features to uncover nonlinear trends:

- Quadratic features
- Cubic features

Modeling

Four classification models:

- Logistic Regression, K-NN, Decision Tree, Ridge Classifier

Utilized cross validation and feature engineering

- 5-fold cross validation
- Balanced accuracy, Recall and F1 Score

Grid search for hyperparameter tuning

3,575,000

people die each year from diseases stemming from dirty water

Model Prioritization: Maximizing Balanced Accuracy

Unsafe water causes water-borne illnesses

- Escherichia coli-induced diarrhea, cholera, typhoid fever, giardia, Hepatitis A, and dysentery

Scoring by balanced accuracy

- Average of true positives and true negatives
- Correctly identify water

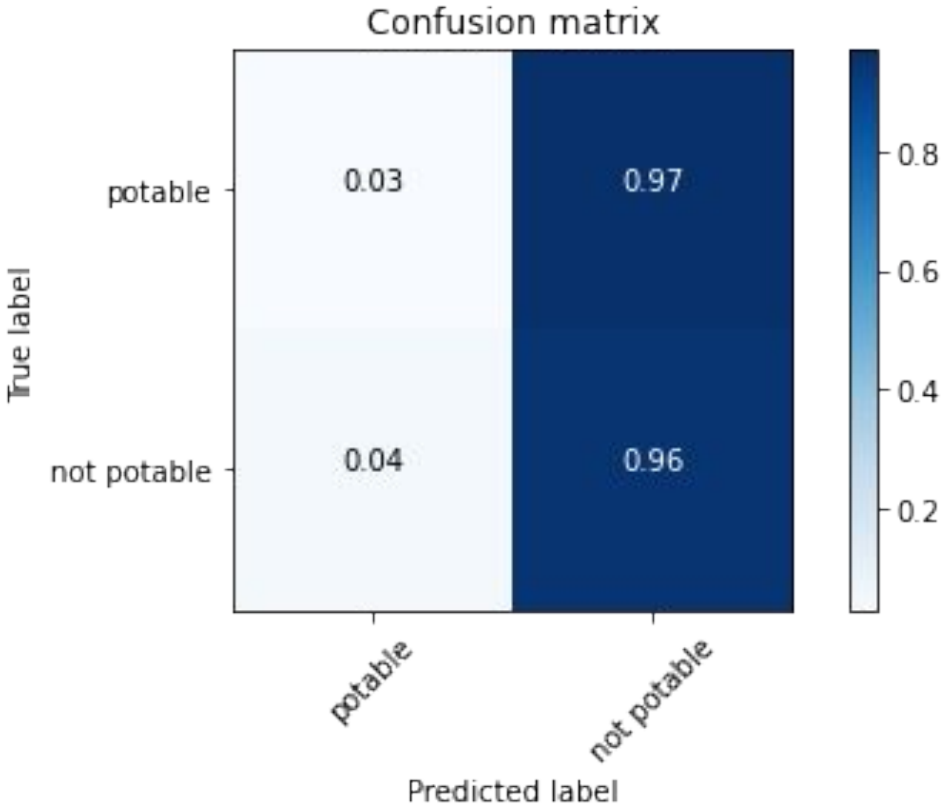
Model Result

	Logistic Regression	k-NN	Decision Tree	Ridge Classifier
Best Parameters	C = 0.01 max_iter = 1000 penalty = none	metric = euclidean n_neighbors = 11 p = 1 weight = distance	criterion = entropy max_depth = 8 max_features = 5 min_samples+leaf = 15 min_samples_split = 10	Alpha = 1 Max_iter = 50
Best Features	Quadratic	No difference	Quadratic	Cubic

Model Result: Confusion Matrix

Logistic Regression:

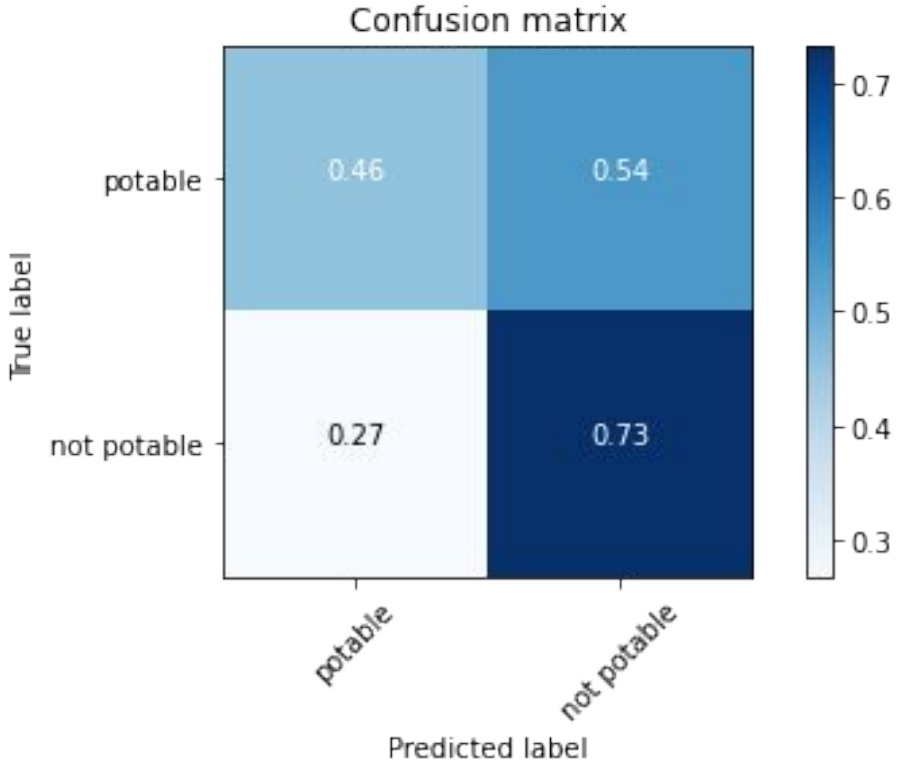
- Highest true positive rate: 96%
- Lowest true negative: 3%
- Smallest false negative rate: 4%
- False positive rate of 97%
    - Model is wasteful



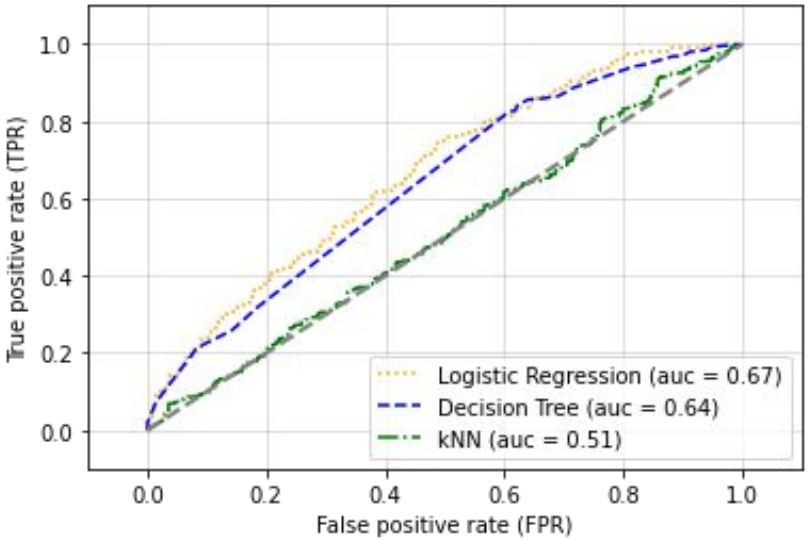
Model Result: Confusion Matrix

Decision Tree

- High true positive rate: 73%
- Relatively high true negative rate: 46%
- Relatively low false positive rate: 54%



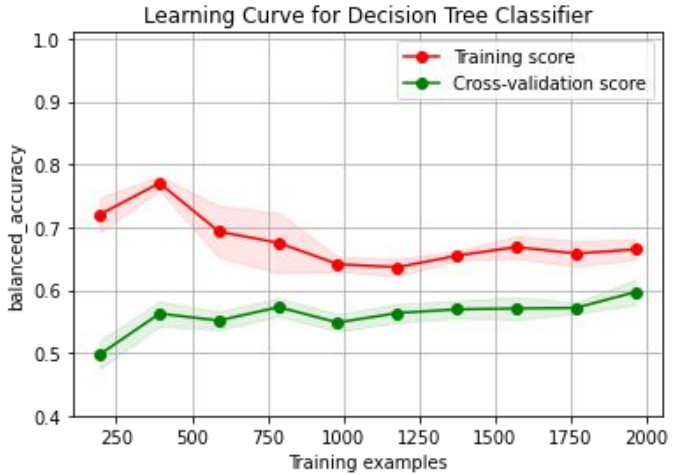
Evaluation: True Positive



- Logistic Regression maximize true positives
  - Highest AUC of 0.67
  - Use with quadratic features and parameter tuning

Evaluation: Balanced Accuracy

	CV Balanced Accuracy	AUC
Logistic Regression	0.56 +/- 0.02	0.67
Decision Tree	0.57 +/- 0.01	0.64
k-NN	0.54 +/- 0.02	0.51
Ridge Classifier	0.57 +/- 0.02	



- Best Decision Tree Classifier
  - Maximized the true positives and true negatives, avoided false negatives
  - Highest Balanced Accuracy at 57%
- Best features
  - Sulfate, Hardness, Conductivity, Chloramines, Trihalomethanes
- Potential Scaling Problem

Deployment

The predictive model can be used to determine if the body of water is potable or not.

- Entities like the CDC, UN or governments can apply this model to:
  - Deliver and test safe water in their community
  - Improve population health
  - Increase economic stability
  - Grow food and natural resources
- Other stakeholders to considers:
  - NGOs
  - Private water bottling companies

Conclusion

Improved water supply can boost economic growth and reduce poverty.

- This model has applications to varying entities, so we expect
  - Scaling the model to be a challenge with the balance accuracy score
  - FIX: Focus on other scoring such as precision or F1 to also mitigate false negatives
- Costs and benefits
  - Cost of drinking non-potable water outweighs the cost of misclassifying potable water
  - Graphing profit curve and exploring ways to minimize false negatives

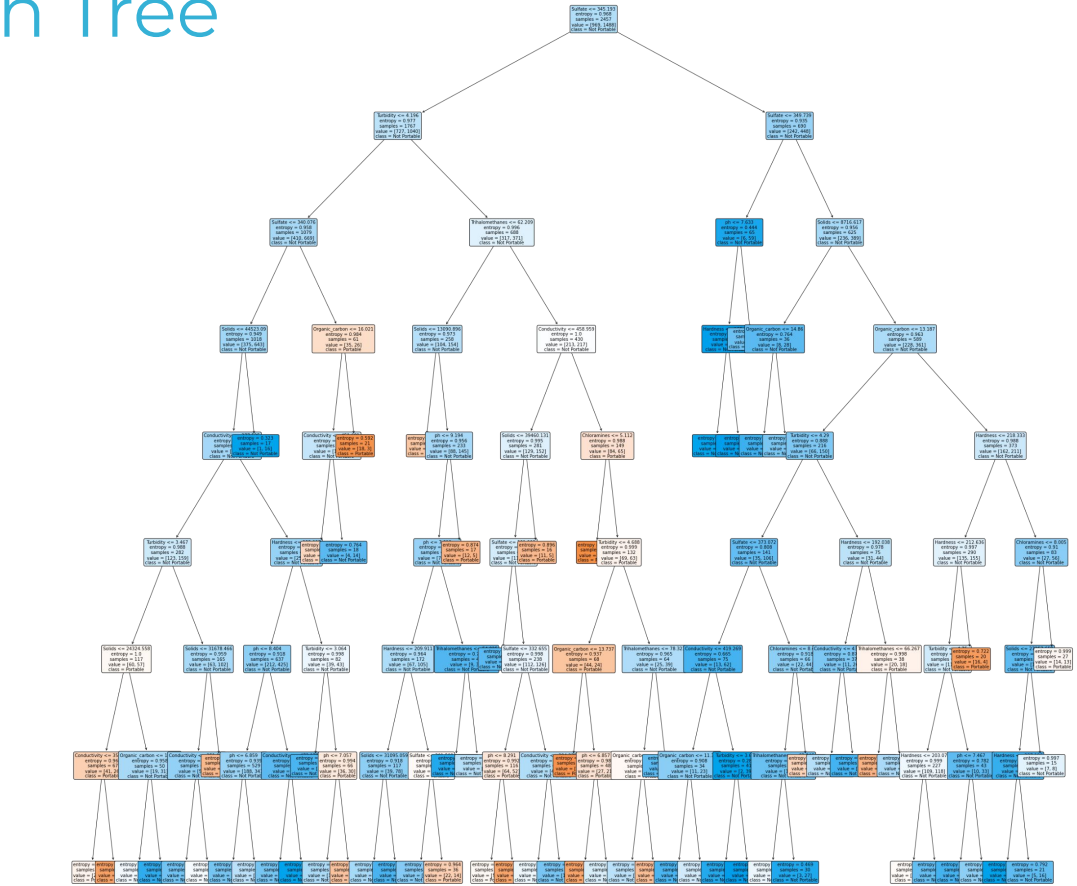
Thank you!
Any Questions?

Citations

- https://caltestlabs.com/analytical-services/regulated-drinking-water-homeowners/wateranalyses/#:~:text=All%20drinking%20waters%20should%20be,potable%20with%20respect%20to%20bacteria.
- https://www.un.org/sustainabledevelopment/water-and-sanitation/
- https://healingwaters.org/why-do-communities-need-clean-water/#:~:text=Unsafe%20water%20causes%20water%2Dborne,%2C%20and%20hygiene%20(WASH).
- https://sdgs.un.org/goals/goal6
- https://www.who.int/news-room/fact-sheets/detail/drinking-water#:~:text=Safe%20and%20readily%20available%20water,contribute%20greatly%20to%20poverty%20reduction.

Appendix: Decision Tree

Decision Tree Mapping

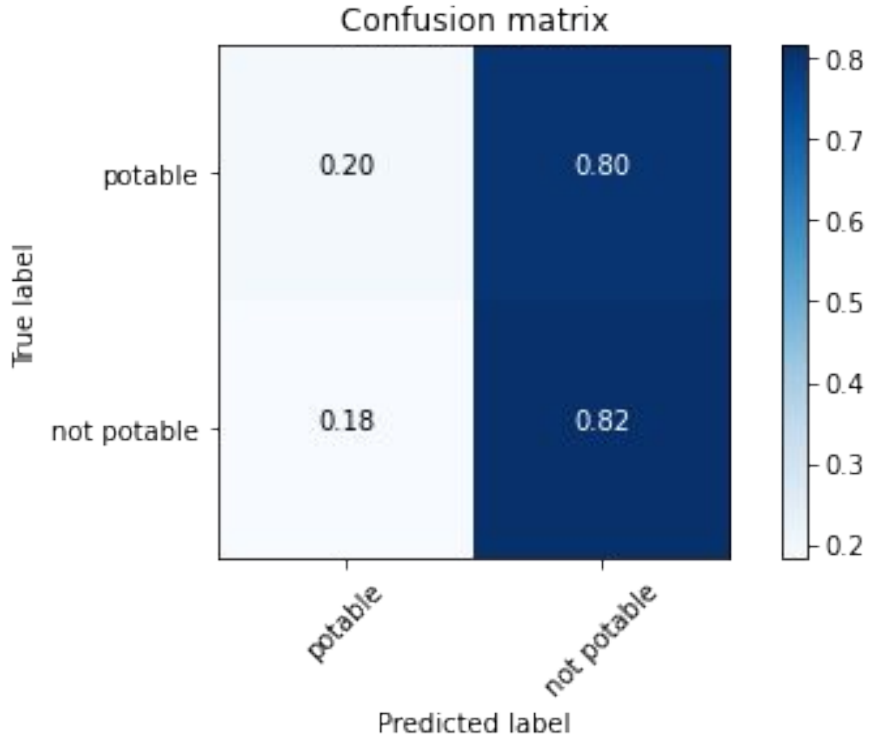


Top 5 factors of water quality:

Sulfate, Hardness, Conductivity, Chloramines, Trihalomethanes

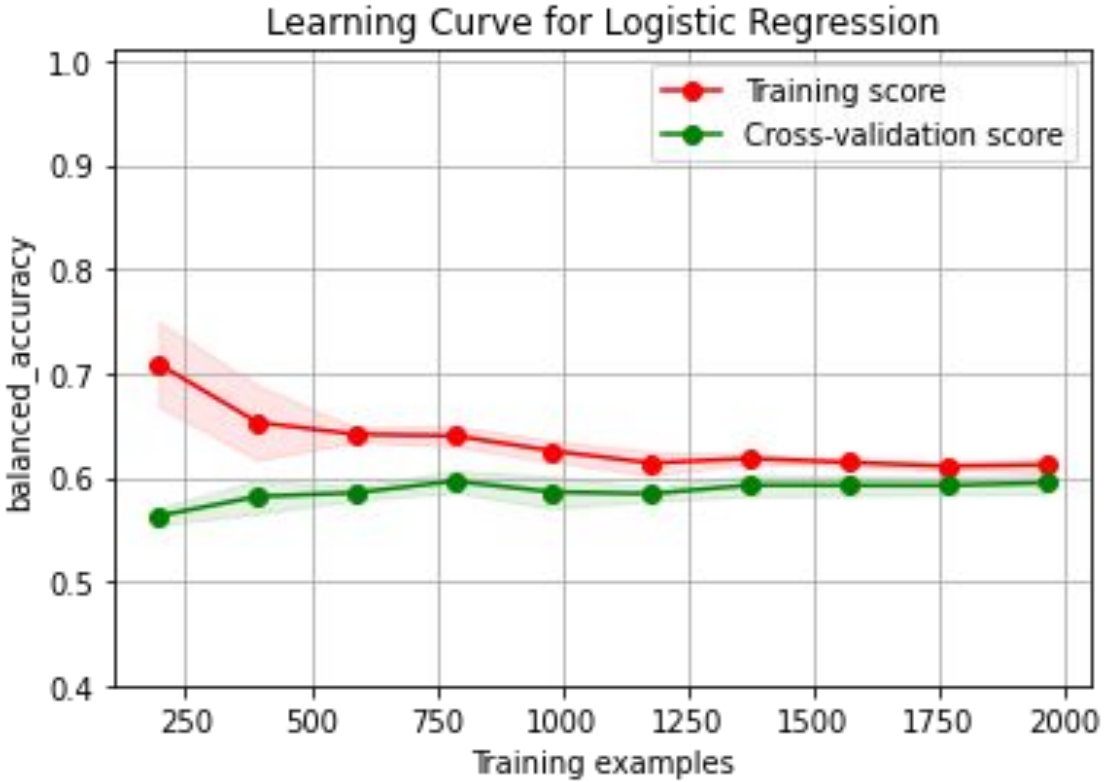
Appendix: Confusion Matrix

Ridge Classifier



Appendix: Learning Curve

Logistic



Appendix: Learning Curve

Ridge Classifier