



Tanzania Water Well Classification

Author: Chris Kucewicz

Introduction and Contents

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Exploratory Data Analysis
5. Modeling and Evaluation
6. Limitations
7. Recommendations
8. Next Steps



Business Understanding

Background

Goals

Success Criteria

- Tanzania faces challenges providing clean water
- Costly & complex:
 - **70k** water wells
 - **67 million** citizens
- NGO seeks efficient well-repair solutions

Business Understanding

Background

Goals

Success Criteria

- Assist NGO in identifying wells needing repair
- Prioritize minimizing false negatives for safety
- Ensure reliable access to clean water

Business Understanding

Background

Goals

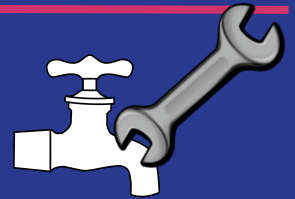
Success Criteria

Positive:

Needs Repair

+

Correctly identified



Business Understanding

Background

Goals

Success Criteria

Negative:

Repair not needed

+

Correctly identified

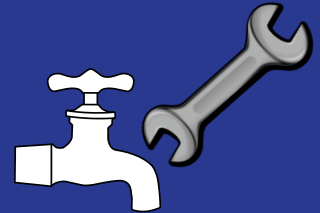


Positive:

Needs Repair

+

Correctly identified



Business Understanding

Background

Goals

Success Criteria

Negative:

Repair not needed

+

Correctly identified



False Positive:

Repair not needed

+

Wrongly flagged

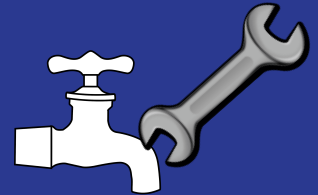


Positive:

Needs Repair

+

Correctly identified



Business Understanding

Background

Goals

Success Criteria

Negative:

Repair not needed

+

Correctly identified



False Positive:

Repair not needed

+

Wrongly flagged



False Negative:

Needs Repair

+

Missed

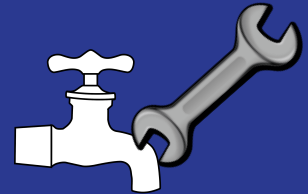


Positive:

Needs Repair

+

Correctly identified



Business Understanding

Background

Goals

Success Criteria

Negative:

Repair not needed

+

Correctly identified



False Positive:

Repair not needed

+

Wrongly flagged



False Negative:

Needs Repair

+

Missed

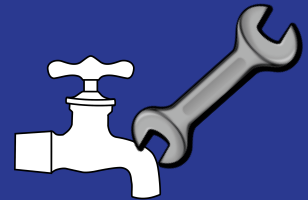


Positive:

Needs Repair

+

Correctly identified



Business Understanding

Background

Goals

Success Criteria

Key Metric: Recall

- Minimizes false negatives

False Negative:

Needs Repair

+

Missed

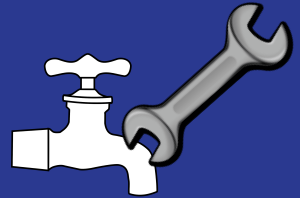


Positive:

Needs Repair

+

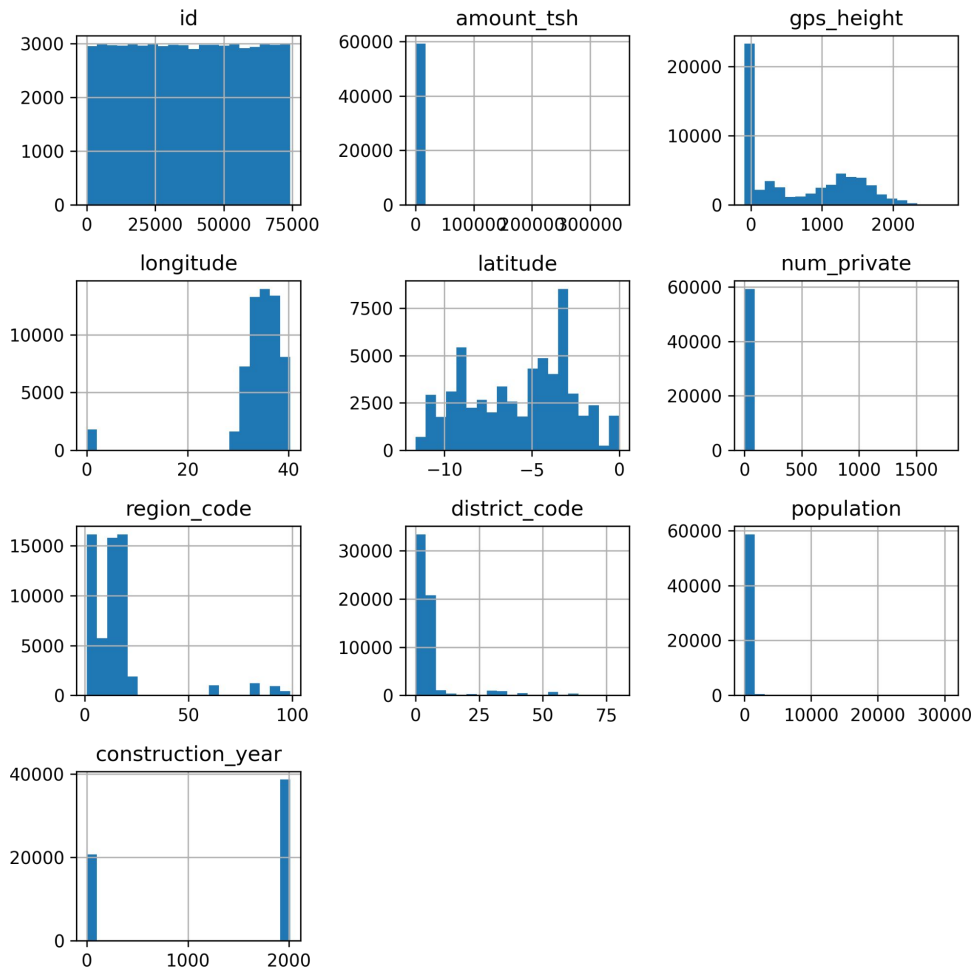
Correctly identified



Data Understanding

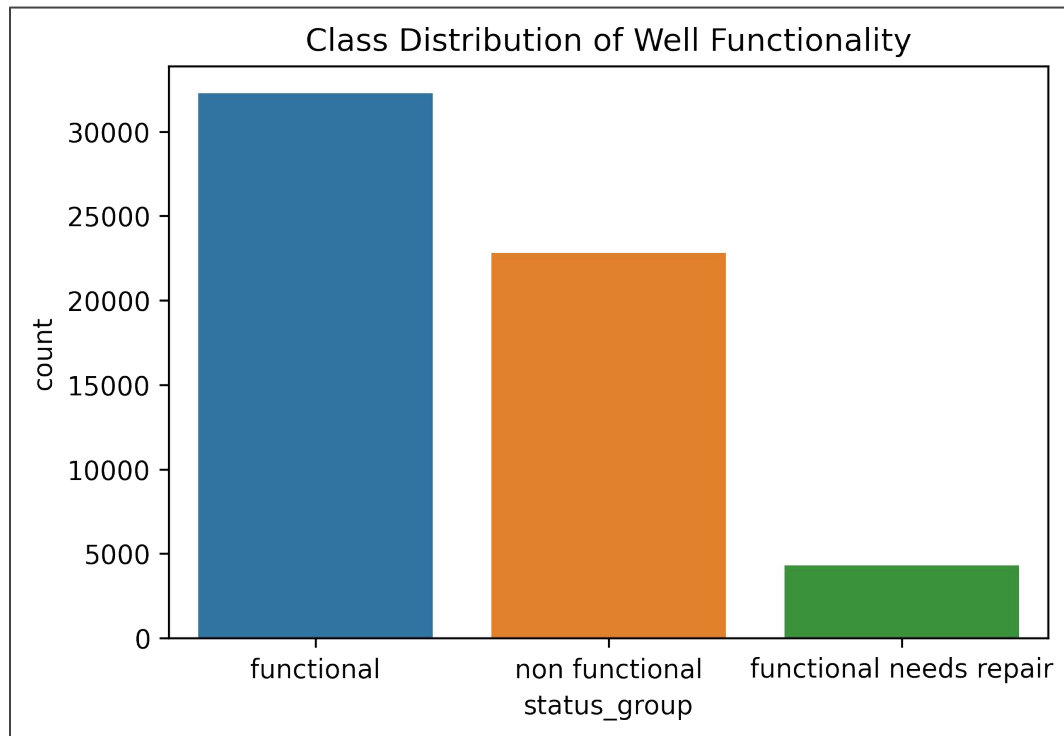
Data: **41** features, almost **60k** wells

Features: Included location, water source, installer



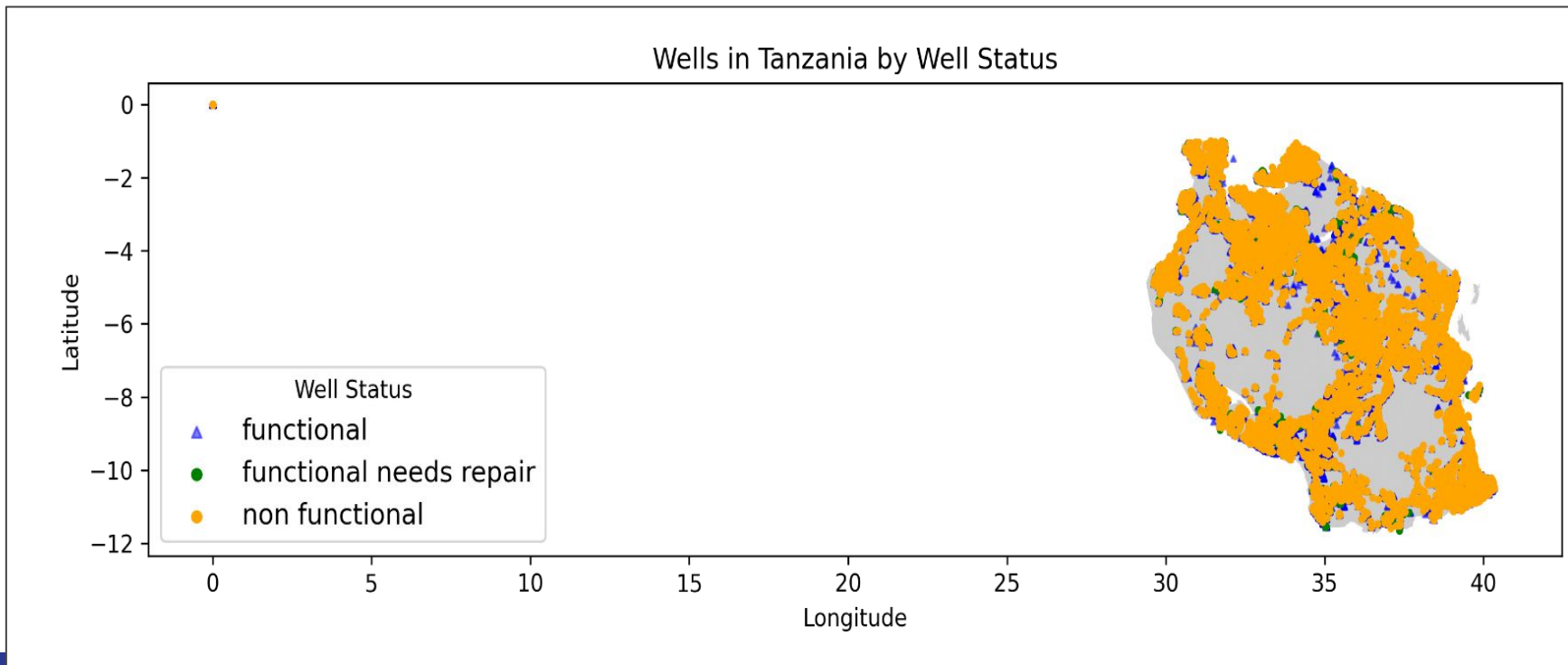
Data Understanding

Target: Well functionality with **3** classes



Data Understanding

Target: Well functionality with **3** classes

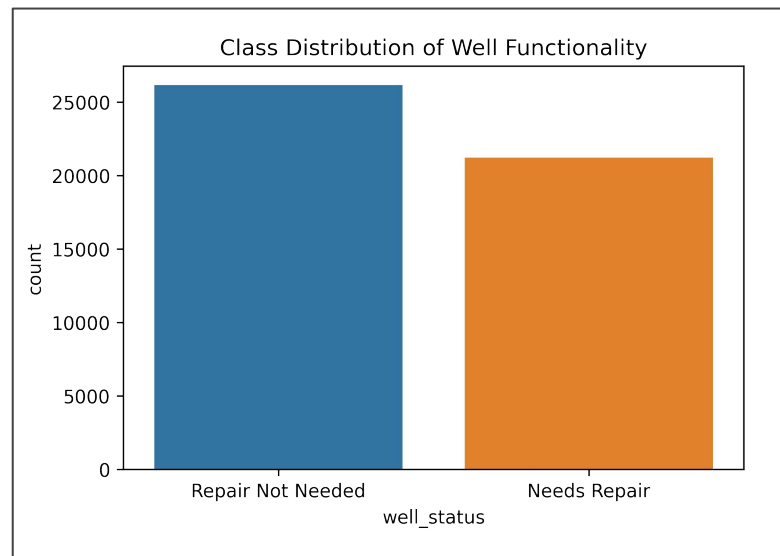
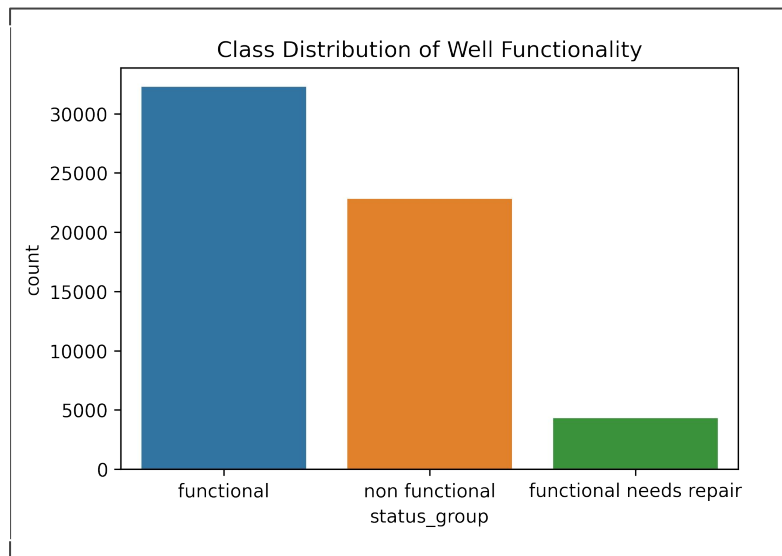


Data Preparation

1. Handled Duplicates
2. Processed features reducing cardinality:
3. Handled Null Values
4. Reclassified the Target Variable to Binary
5. Cleaned Dataset Overview:
 - Reduced to **19** features, **47k** rows



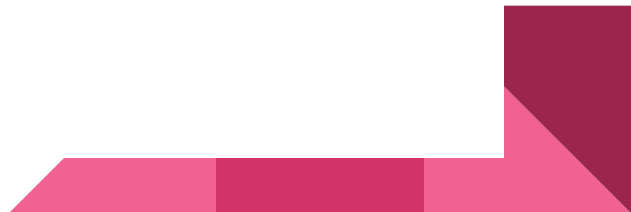
Data Preparation: Reclassified Target to Binary



Exploratory Data Analysis

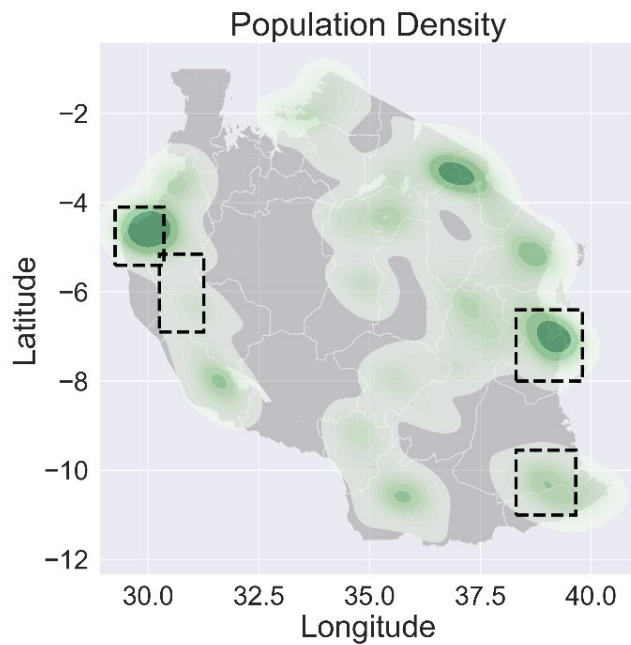
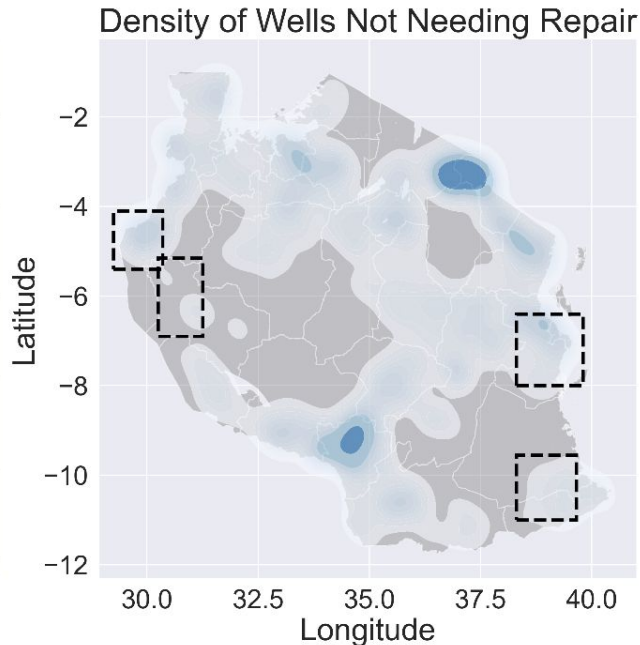
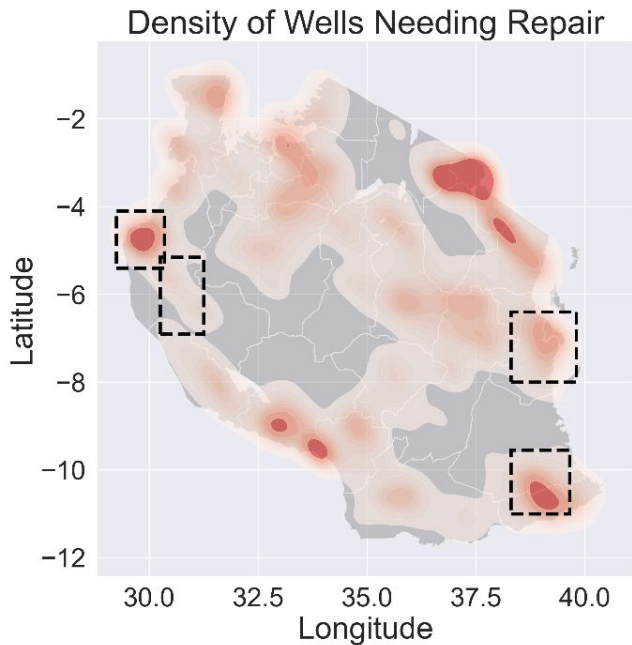
Findings

1. Government installer = higher non-functioning well rates
2. Ruvuma/Southern Coast and Lake Rukwa basins = higher non-functioning well rates
3. High-Priority Areas:
 - Areas with **high repair needs**, **low functional wells**, and **high population density**.



Exploratory Data Analysis: High-Priority Areas

Areas with **high repair needs**, **low functional wells**, and/or **high population density**.



Modeling and Evaluation

	Model Name	Training Recall (%)	Test/Val Recall (%)	Difference (Training - Test/Val Recall) (%)	False Negatives (%)
0	Decision Tree with RFE (val data)	100.00	78.60	21.40	9.70
1	Final Decision Tree (test data)	45.03	44.36	0.67	24.80
2	Final Random Forest (test data)	58.86	58.80	0.06	18.40
3	Initial Random Forest (test data)	99.90	77.45	22.45	10.10
4	Baseline Logistic Regression (test data)	66.85	66.00	0.85	15.20

Modeling and Evaluation

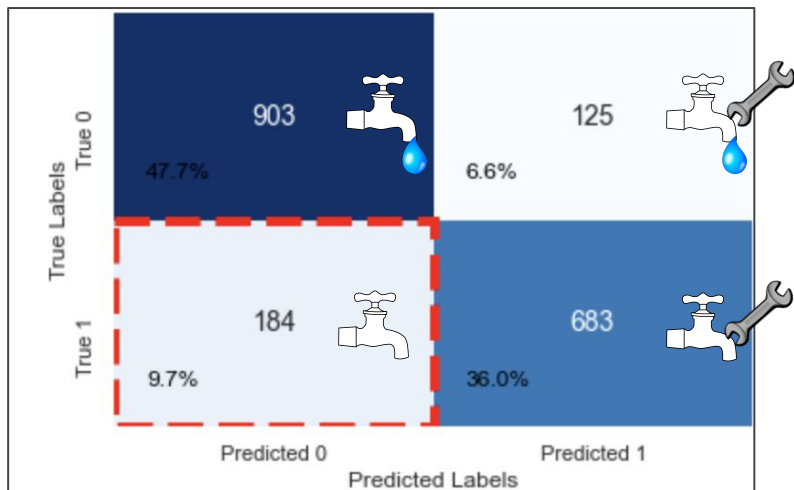
	Model Name	Training Recall (%)	Test/Val Recall (%)	Difference (Training - Test/Val Recall) (%)	False Negatives (%)
0	Decision Tree with RFE (val data)	100.00	78.60	21.40	9.70
1	Final Decision Tree (test data)	45.03	44.36	0.67	24.80
2	Final Random Forest (test data)	58.86	58.80	0.06	18.40
3	Initial Random Forest (test data)	99.90	77.45	22.45	10.10
4	Baseline Logistic Regression (test data)	66.85	66.00	0.85	15.20

Modeling and Evaluation: Confusion Matrices

Decision Tree Model with RFE

Train recall score: **100%**

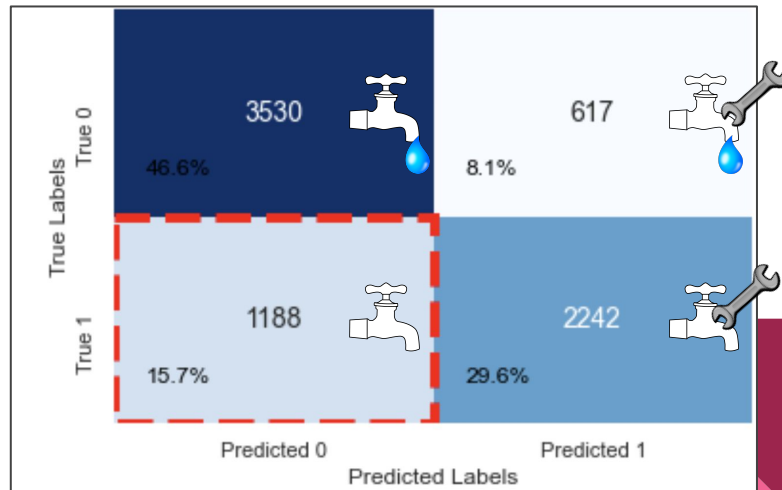
Test recall score: **78.6%**



Logistic Regression Model

Train recall score: **66.85%**

Test recall score: **66%**

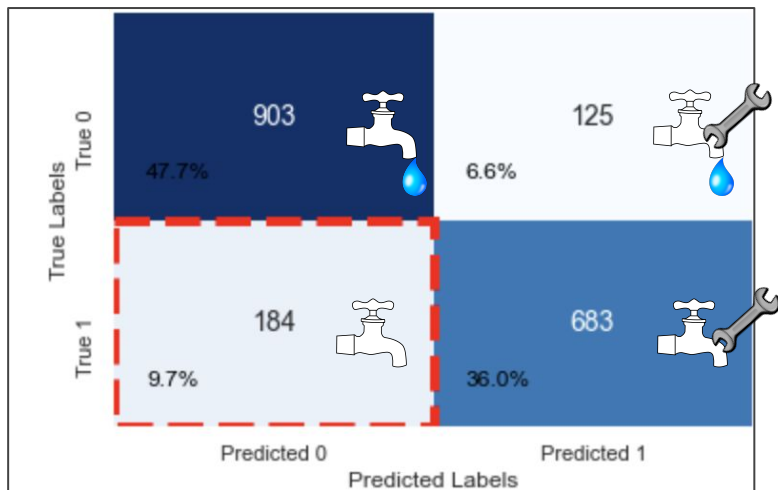
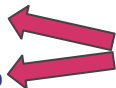


Modeling and Evaluation: Confusion Matrices

Decision Tree Model with RFE

Train recall score: **100%**

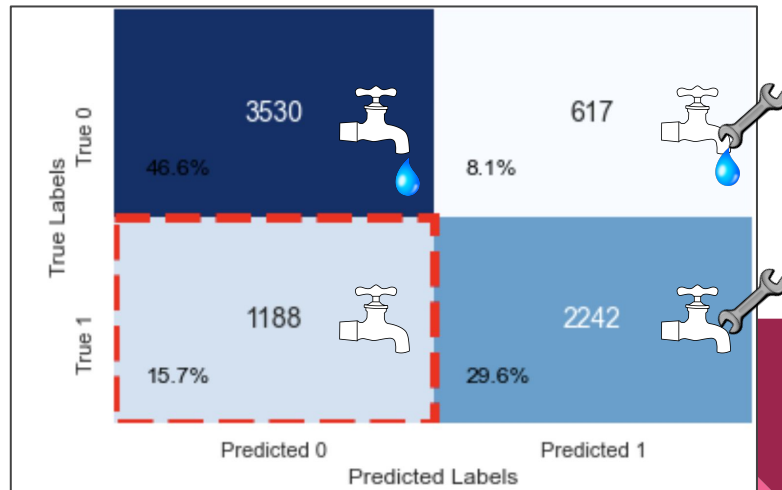
Validation recall score: **78.6%**



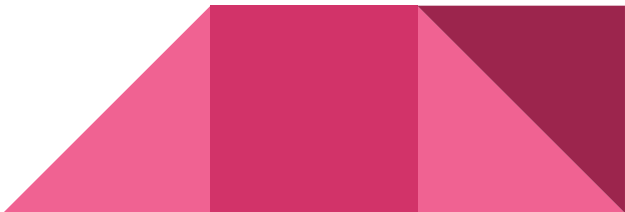
Logistic Regression Model

Train recall score: **66.85%**

Test recall score: **66%**



Limitations

- **Data Quality:**
 - Model is only as good as its data
 - **Computing Constraints:**
 - Restricted parameter grid exploration
 - **Time Constraints:**
 - Impacted model optimization
 - **Domain Knowledge:**
 - Lack of local expertise affected feature selection
- 

Recommendations

Recommendation 1: High-Demand Areas

Focus repairs in following regions:

- **Northwest (Kigoma)**
- **Southeast (Dar Es Salaam, Mtwara)**

Recommendation 2: Logistic Model

Choose **logistic regression model** for unseen data

Recommendation 3: Reduce Model Complexity

To reduce overfitting

- Improve **feature selection**
- Apply **cross-validation**

Recommendations

Recommendation 1: High-Demand Areas

Focus repairs in following regions:

- **Northwest (Kigoma)**
- **Southeast (Dar Es Salaam, Mtwara)**

Recommendation 2: Logistic Model

Choose **logistic regression model** for unseen data

Recommendation 3: Reduce Model Complexity

To reduce overfitting

- Improve **feature selection**
- Apply **cross-validation**

Recommendations

Recommendation 1: High-Demand Areas

Focus repairs in following regions:

- **Northwest (Kigoma)**
- **Southeast (Dar Es Salaam, Mtwara)**

Recommendation 2: Logistic Model

Choose **logistic regression model** for unseen data

Recommendation 3: Reduce Model Complexity

To reduce overfitting

- Improve **feature selection**
- Apply **cross-validation**

Next Steps

1. Enhance Data Collection
2. Reduce Overfitting in Models
3. Explore Advanced Algorithms



Thank you!



Github Repository:

https://github.com/ckucewicz/water_well_classification

Contact Chris Kucewicz at

cfkucewicz@gmail.com with additional questions