Predictive Modeling of Tanzanian Water Well Functionality

Nancy Ho

Summary

- Use of machine learning models to provide method of predicting functionality of Tanzanian wells based on descriptive information about various wells within the region

Outline

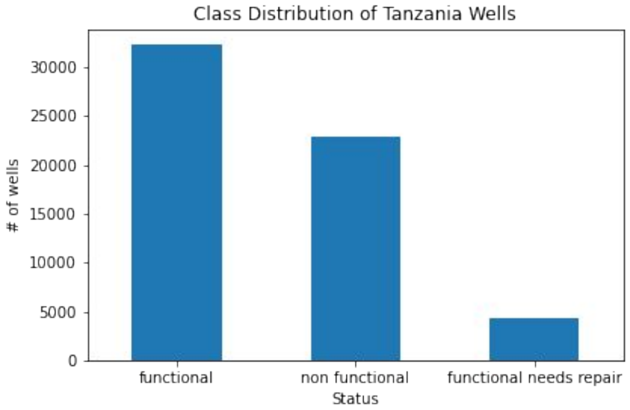
- Business Problem
- Data Understanding
- Model Creation
- Model Evaluation
- Next Steps

Business Problem

- Water crisis in Tanzania, many non-profit organizations have focused on drilling wells to provide clean water to Tanzanian villages
- While developing new wells is important, we must also pay attention to condition of existing water wells
    - Need to ensure Tanzanian villages have consistent, sustainable access to clean water
- Goal: create a model that can predict condition of wells in Tanzania based on descriptive information
    - Can assist in allocating resources towards well maintenance and upkeep

Data Understanding

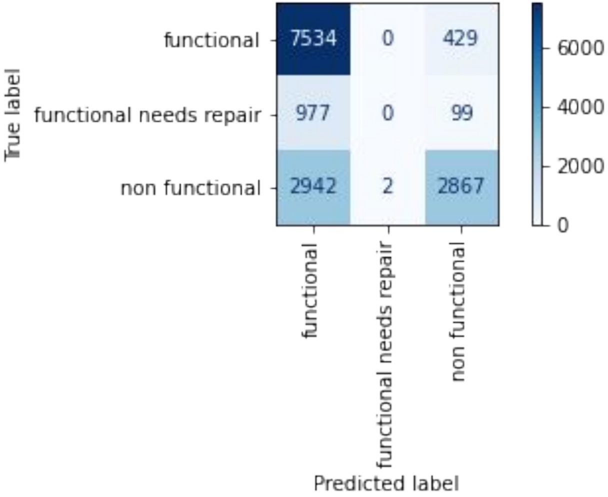
- Information about water wells in Tanzania
  - Provided by DrivenData, derived from Taarifa software and Tanzania Ministry of Water
  - Contains large amount of information about each well -- will only use information most relevant to predicting water well condition (e.g. extraction type, water quality)
  - Class imbalance - may lead to error later, but data is left as is for identification purposes



Model Evaluation

- Logistic regression
  - Measures the difference in probabilities among each class
- F1 score (average of precision and recall): 0.70
  - Metric based on how well model minimizes misclassifying wells as false positive or false negative
- Shows more bias towards identifying functional wells
  - May be result of imbalance in data

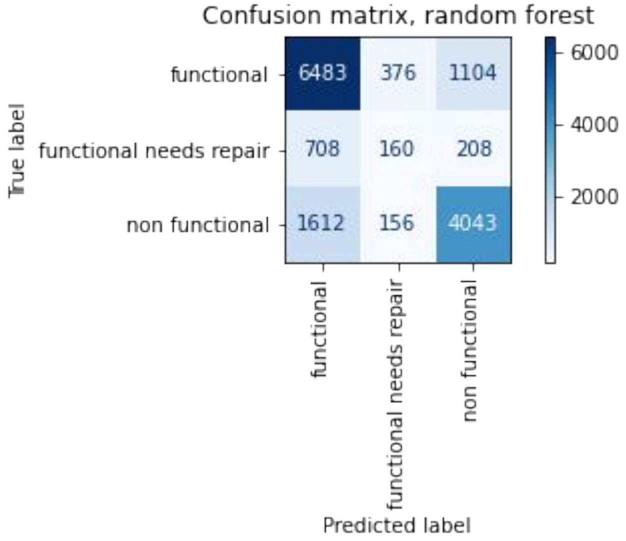
Confusion matrix, logistic regression



	functional	functional needs repair	non functional
functional	7534	0	429
functional needs repair	977	0	99
non functional	2942	2	2867

Model Evaluation

- Random forest
  - Aggregates multiple decision tree classifiers to improve overall accuracy of model
- F1 score: 0.72
- Shows greater balance in identifying non-functional/functional needs repair wells

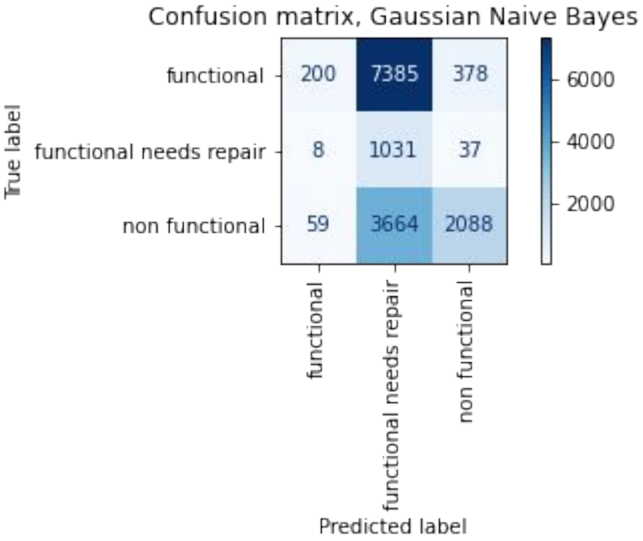


Confusion matrix, random forest

True label \ Predicted label	functional	functional needs repair	non functional
functional	6483	376	1104
functional needs repair	708	160	208
non functional	1612	156	4043

Model Evaluation

- Gaussian Naive Bayes
  - Assumes all features are independent, estimates overall probability given conditional probabilities of each feature
- F1 score: 0.22
- Shows significant bias towards marking most wells as “needs repair”

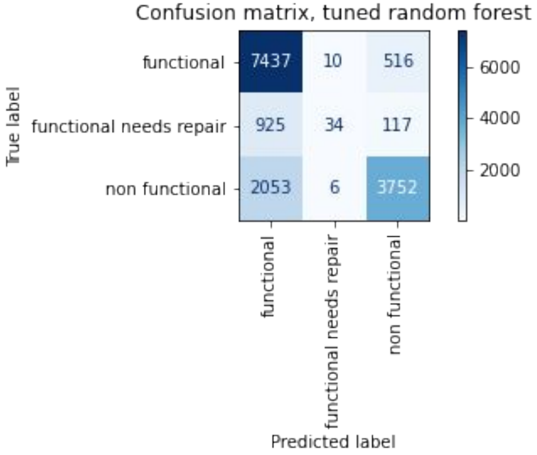


Confusion matrix, Gaussian Naive Bayes

	functional	functional needs repair	non functional
functional	200	7385	378
functional needs repair	8	1031	37
non functional	59	3664	2088

Model Evaluation and Conclusion

- Random forest classifier predicts well condition most accurately, performs even better with optimal parameters
- Still shows some bias towards functional wells due to categorical imbalance
- While model does not perform exceptionally well, it still provides a satisfactory baseline model to perform predictions on a well's condition



Next Steps

- Attempt to re-train models on balanced data
- Create model that can process more specific attributes of water wells rather than subcategories of attributes
- Doing more research on potential geological and ecological factors that impact well condition and water quality
- Possibly use model to assist other well maintenance efforts across Africa?

Thank you!

Email: nancyho83@yahoo.com
GitHub: @nancyho83
LinkedIn: linkedin.com/in/nho3/