Predicting Tanzanian Water Well Functionality: A Machine Learning Approach

Gregory Antony Mikuro

Business Understanding - The Challenge

- Tanzania faces a water crisis due to a high number of broken or non-functioning wells.
- This lack of access to clean water leads to disease, decreased productivity, and educational barriers.

Business Understanding: Our Goal

- Develop a machine learning model to accurately predict well functionality (functional, non-functional, or needs repair).
- Empower NGOs and the government to make data-driven decisions for well maintenance and construction.
- **Success Metric:**
- Target accuracy score of at least 80%.
- **Impact:**
- Improved water access for millions, leading to a healthier and more prosperous Tanzania.

Data Understanding: Data Source

- Data sourced from the DrivenData competition "Pump It Up: Data Mining the Water Table.“
- Includes information on well location, construction, funding, and functionality status.

- `amount_tsh`: Total static head (amount of water available to p
- `funder`, `installer`: Entities responsible for funding and insta
- `gps_height`: Altitude of the well.
- `longitude`, `latitude`: Geographic coordinates.
- `basin`: Geographic water basin.
- `population`: Population around the well.
- `public_meeting`: Indicator of a public meeting about the proj
- `scheme_management`: Entity managing the water supply schem
- `permit`: Indicator of a government construction permit.
- `construction_year`: Year of construction.
- `extraction_type_class`: Type of extraction technology.
- `management_group`: Management type of the well.
- `payment_type`: Water cost structure.
- `quality_group`: Water quality.
- `quantity`: Water quantity.
- `source_class`: General water source type.
- `waterpoint_type`: Type of well.
- `status_group`: Target variable (functional, non-functional, fun

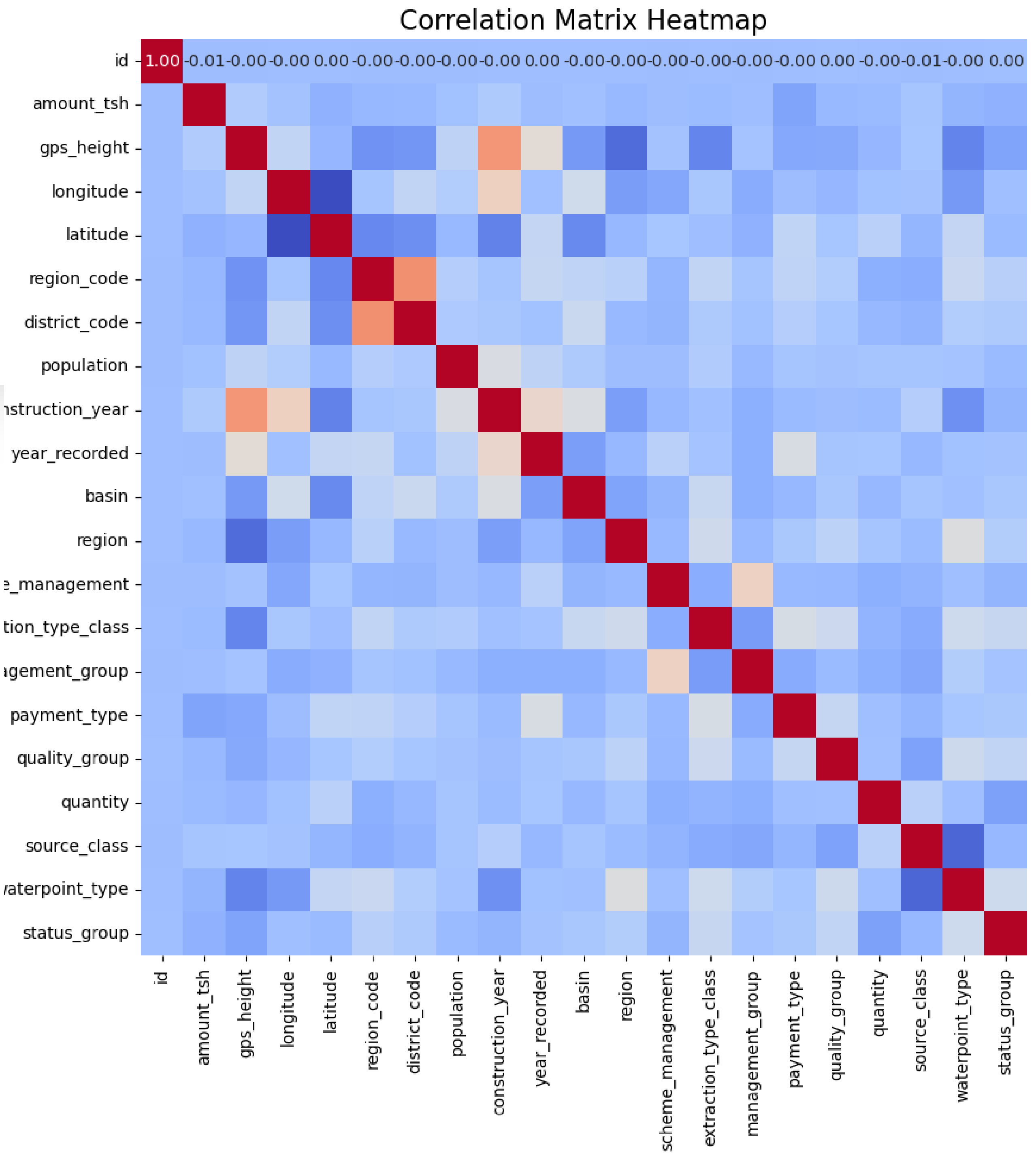
Data Cleaning

Cleaning Steps:

- Extracted year from date_recorded.
- Imputed missing values in categorical columns.
- Dropped irrelevant columns.

EDA

- Class imbalance in status_group: Most wells are "functional.“
- amount_tsh heavily skewed right with many zero values.
- gps_height has two distinct groups based on elevation.
- Weak correlations between most numerical and categorical features.



Modeling & Evaluation – Models Tested

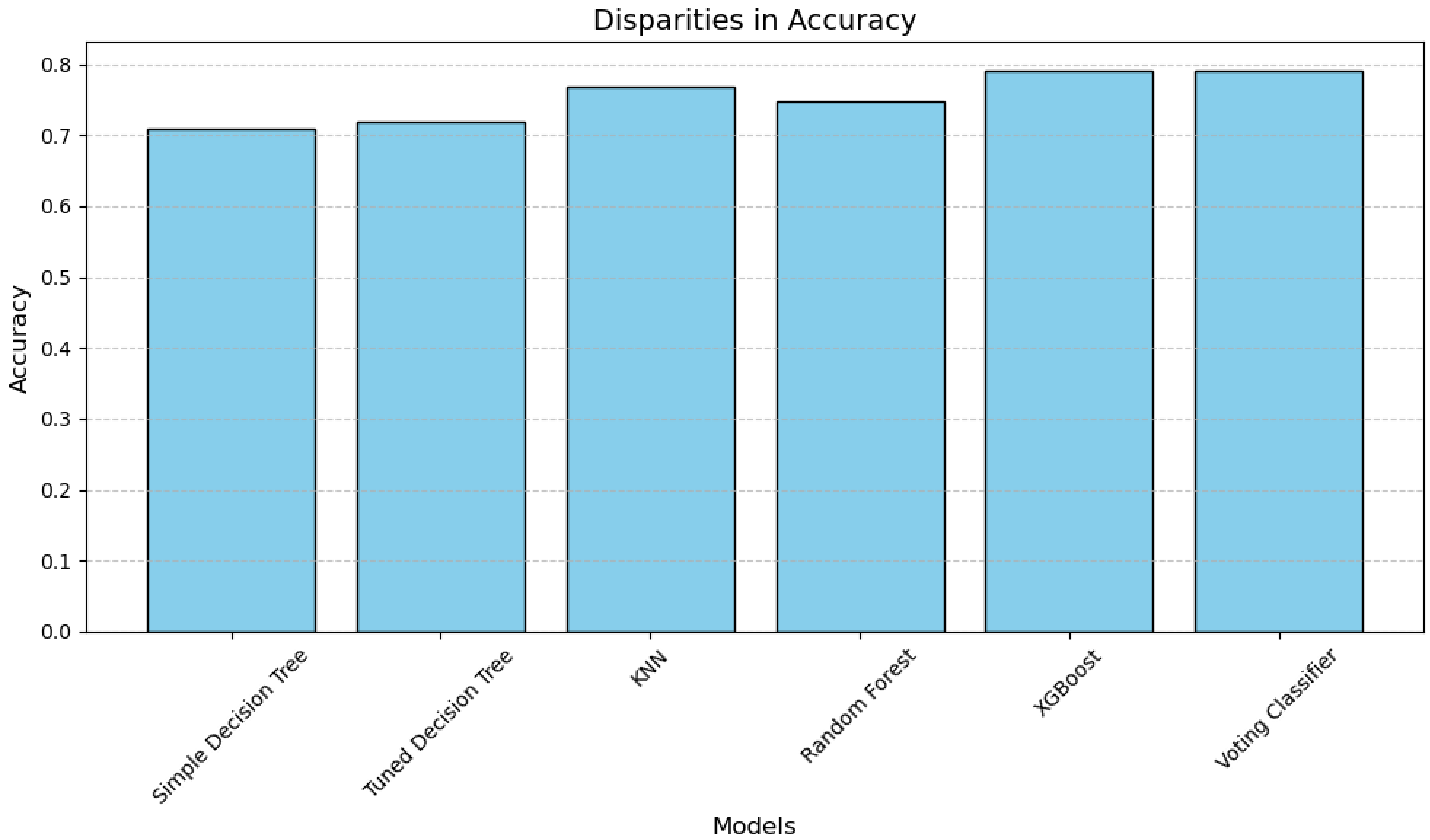
- Simple Decision Tree (Baseline)
- Tuned Decision Tree
- KNN
- Random Forest
- XGBoost
- Voting Classifier (Ensemble)

Performance

XGBoost and Voting Classifier performed best (around 79% accuracy).

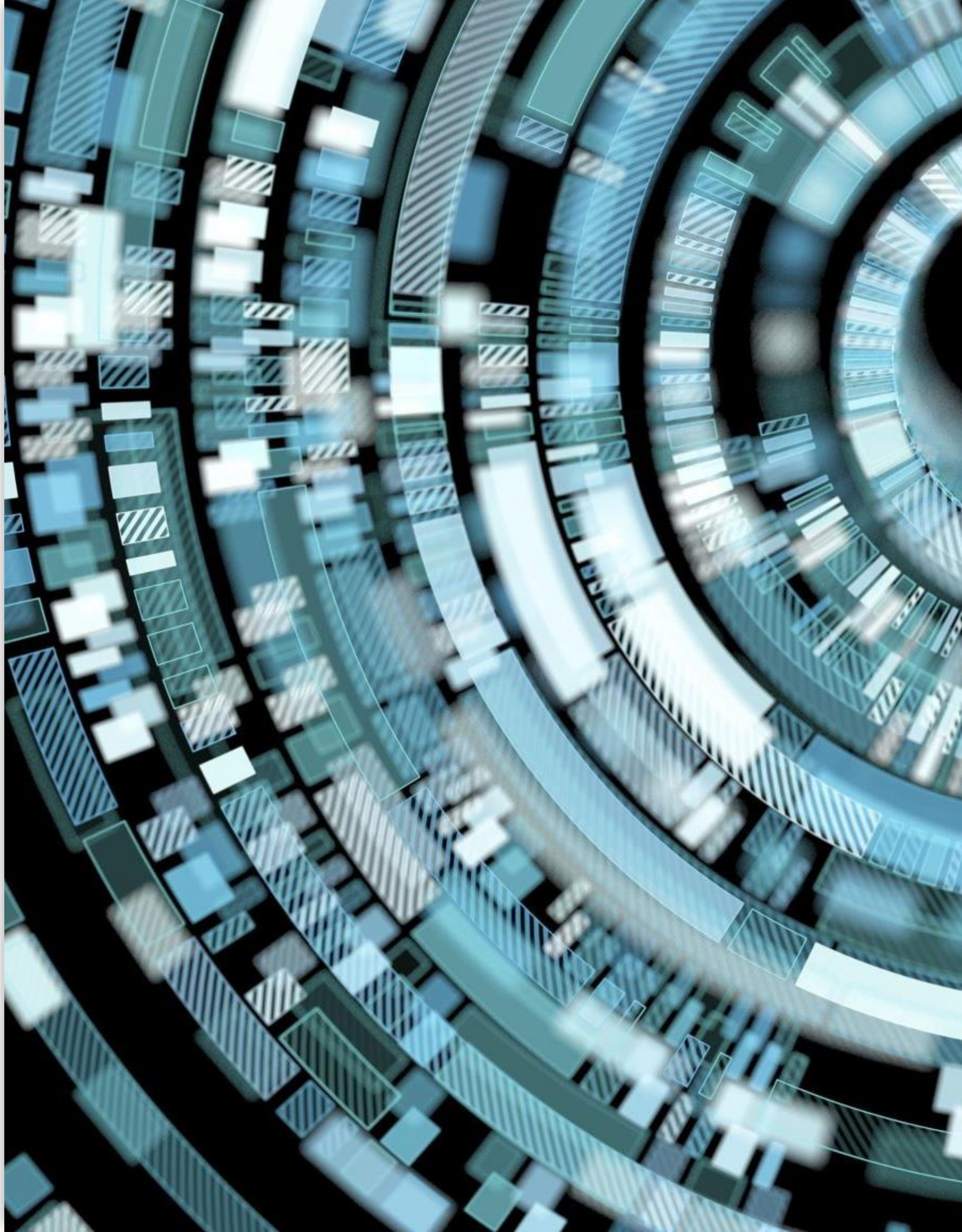
All models struggled with the "functional needs repair" class.

Note: Due to computational constraints, hyperparameter tuning was limited.


Disparities in Accuracy
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Simple Decision Tree
Tuned Decision Tree
KNN
Random Forest
XGBoost
Voting Classifier
Models

Deployment: Streamlit App

- Voting Classifier was chosen for deployment due to high accuracy
- User-friendly interface for inputting well features.

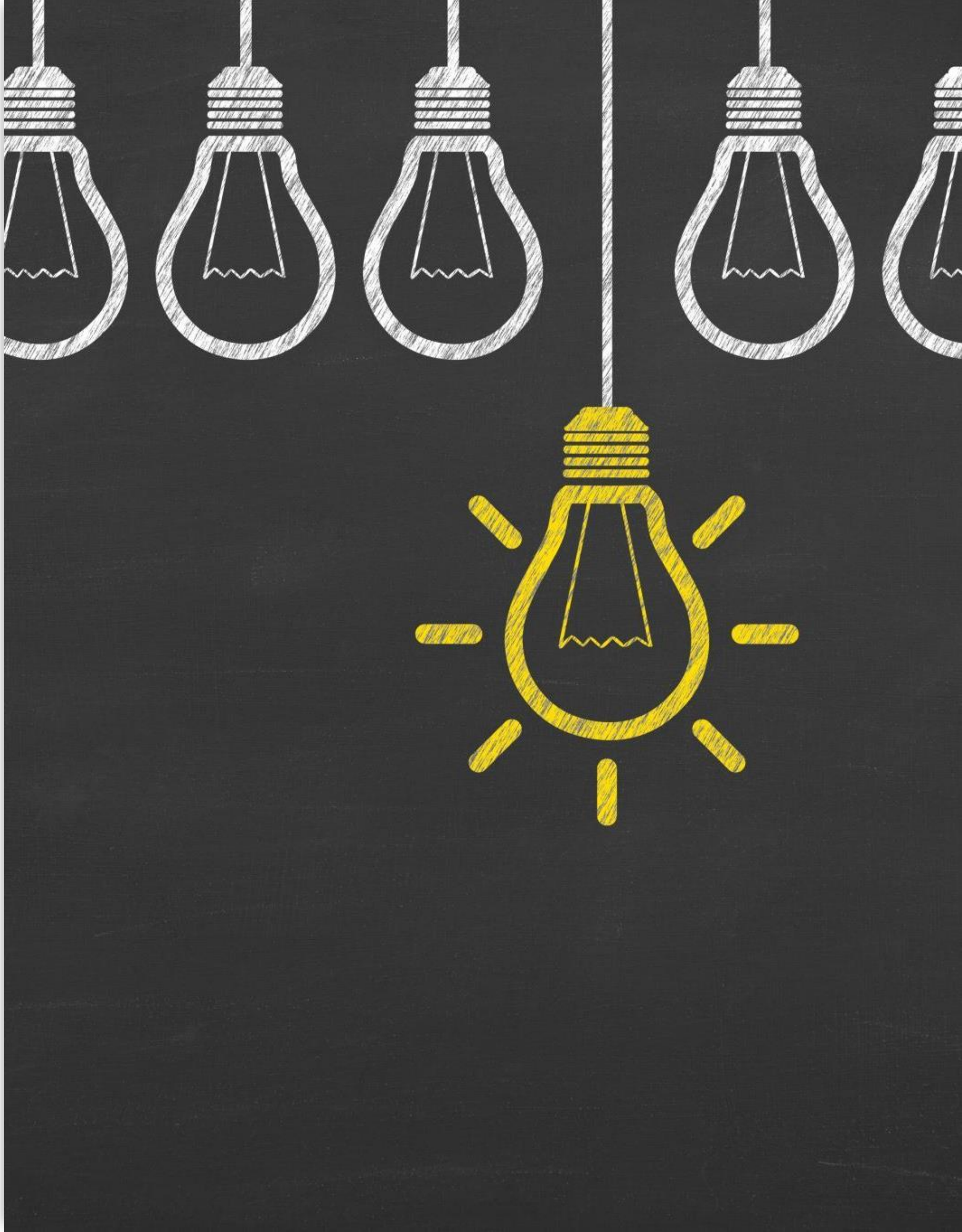


Conclusion & Recommendations

- Predictive models show promise, but improvements are needed for the "functional needs repair" class.
- Data limitations (class imbalance, zero imputation) may impact performance.

Recommendations:

- Gather more data, especially for the underrepresented class.
- Explore advanced feature engineering.Try different models (SVM, LightGBM, CatBoost).
- Implement cost-sensitive learning.
- Calibrate model probabilities.
- Deploy, monitor, and iterate on the model in production.



Thank you