WATER WELL-CONDITION ANALYSIS

Predicting and Understanding Well Functionality

BUSINESS PROBLEM

Content:

- **What is the Problem?**

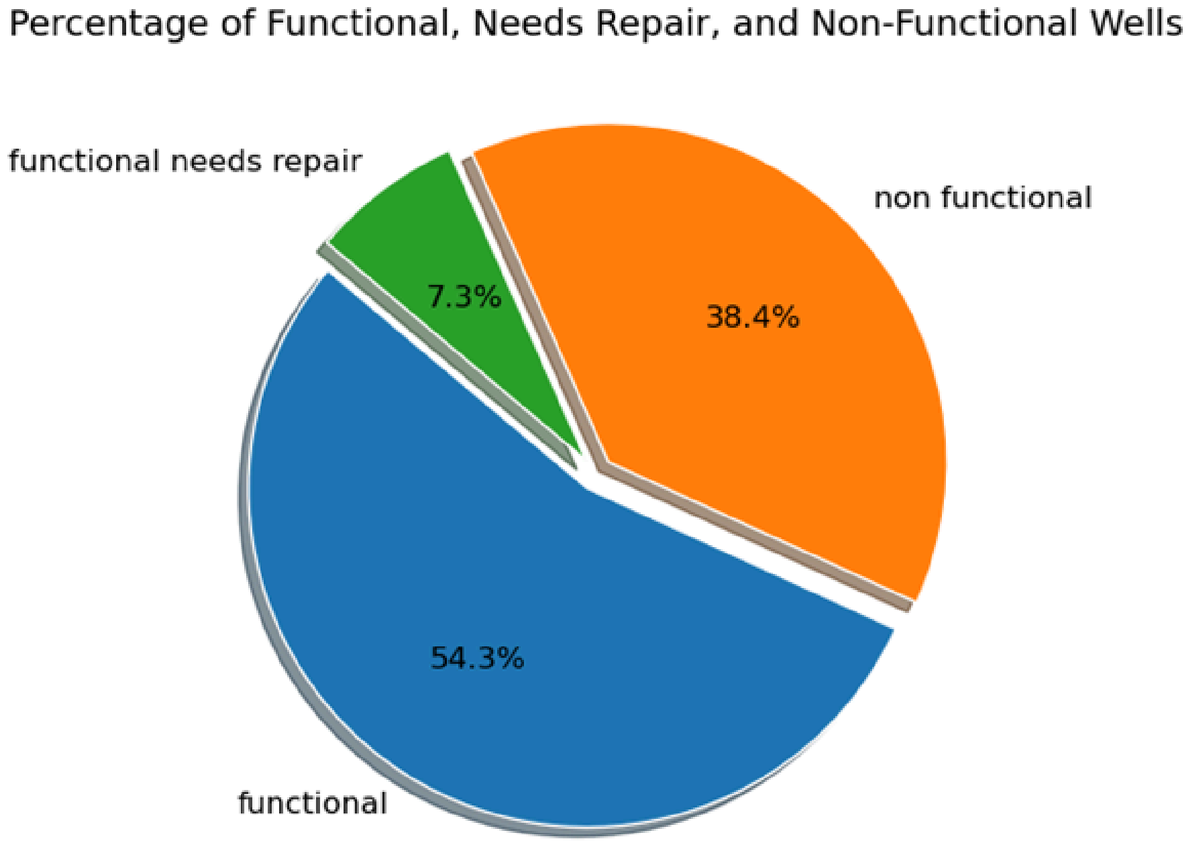
Many wells are non-functional or need repair, leading to water access challenges.

Why is this Important?

- Limited resources require targeted maintenance.
- Optimizing well performance ensures efficient resource utilization.

Our Goal:

- Predict which wells are functional, need repair, or are non-functional to prioritize maintenance.



PROJECT OBJECTIVES

- **Content:**
  - Build machine learning models to classify well conditions.
  - Identify the key factors that influence well functionality.
  - Compare multiple models to determine the best performer.
  - Provide actionable insights for decision-makers to allocate resources efficiently.

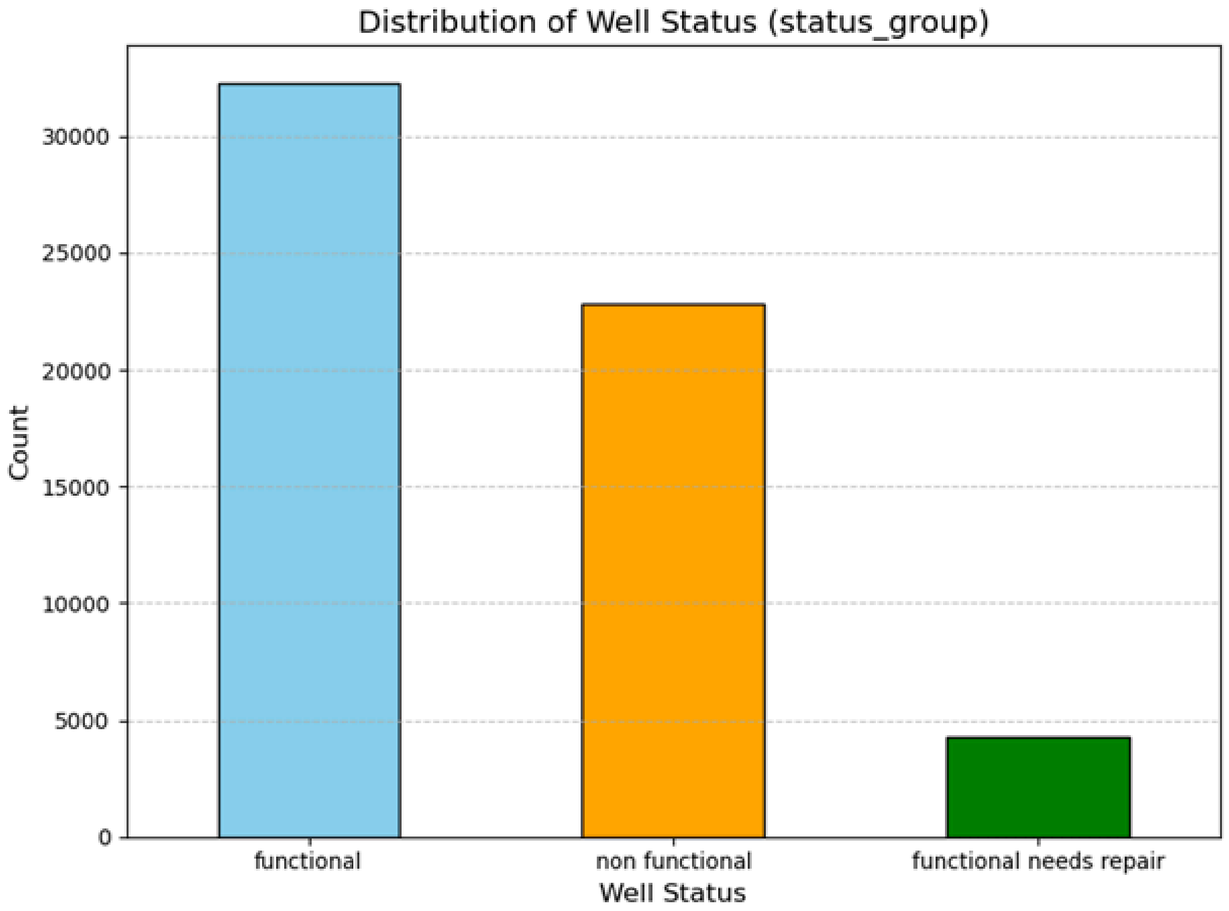
DATA OVERVIEW

Dataset Overview:

- Total Wells Analyzed: 59,400
- Features: 40 columns (e.g., gps_height, population, amount_tsh)
- Target Variable: status_group (functional, needs repair, non-functional).

Data Quality:

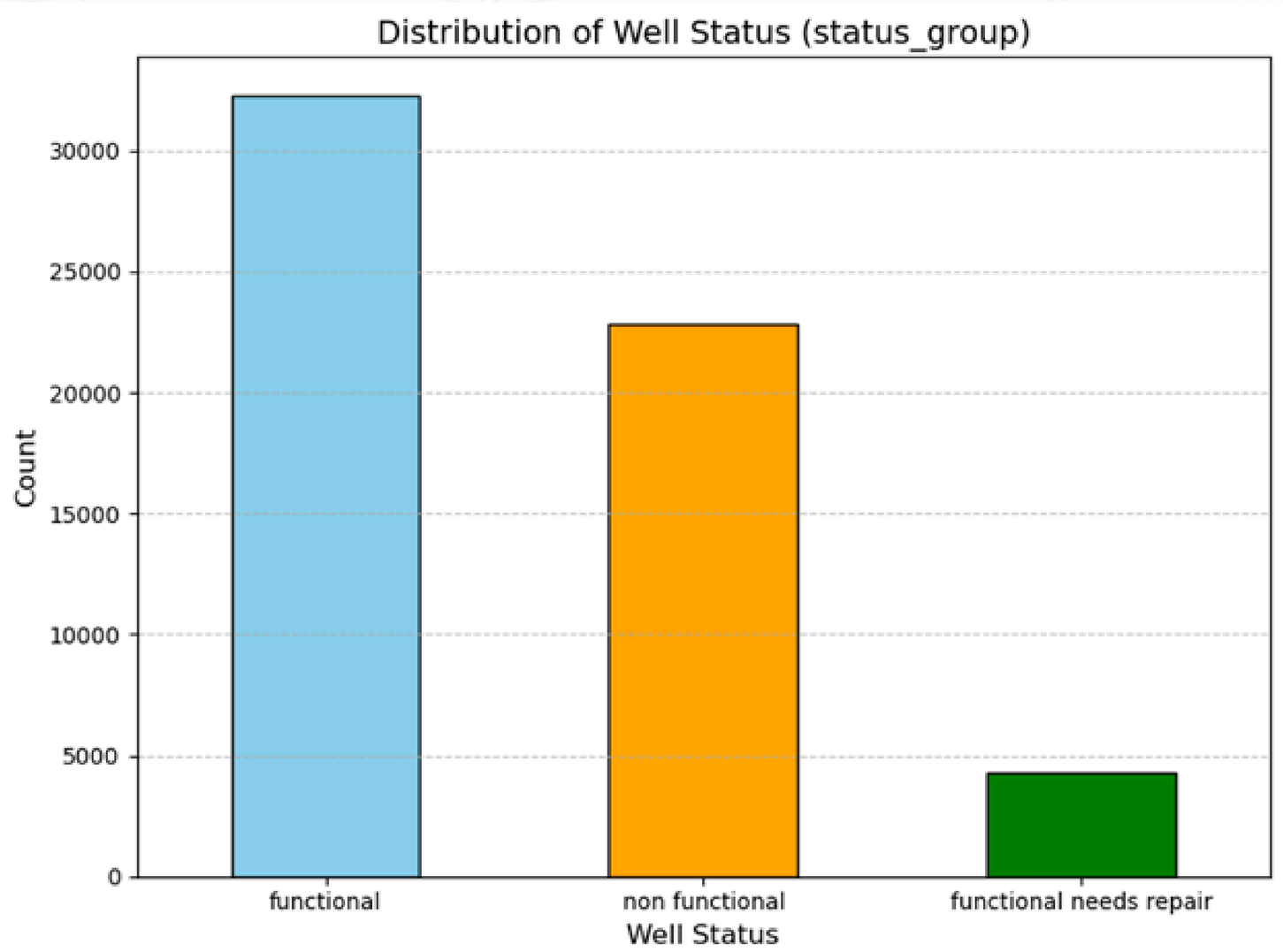
- Missing values were handled.
- Non-numeric data was cleaned or encoded.



EXPLORATORY DATA ANALYSIS

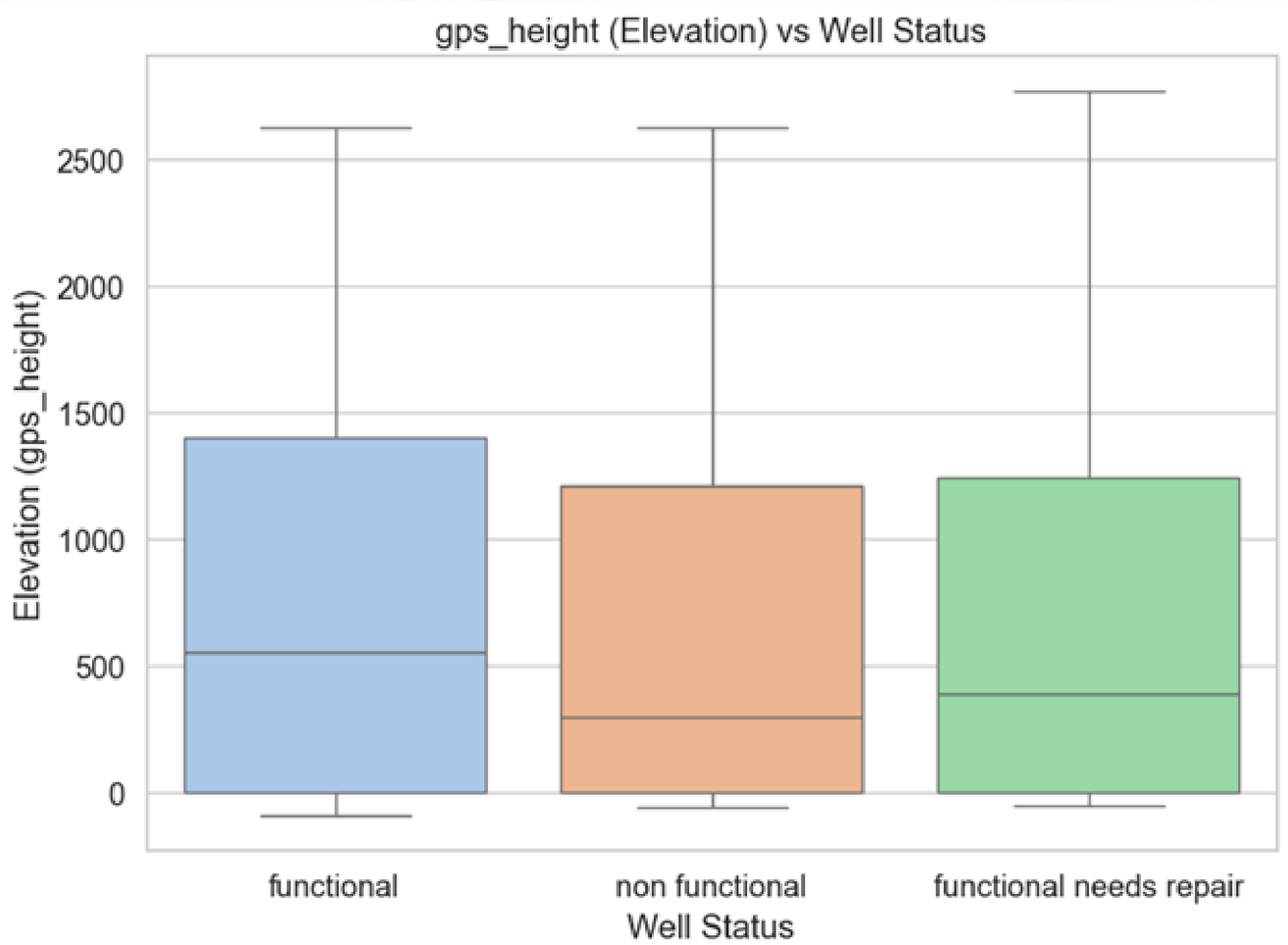
- Key Insights from EDA:
  - Wells at lower elevation (gps_height) are more likely to fail.
  - Older wells (construction_year) are at higher risk.
  - Areas with high population show increased wear and tear on wells.
  - Wells with low water output (amount_tsh) tend to fail frequently.

Bar Chart: Distribution of status_group.



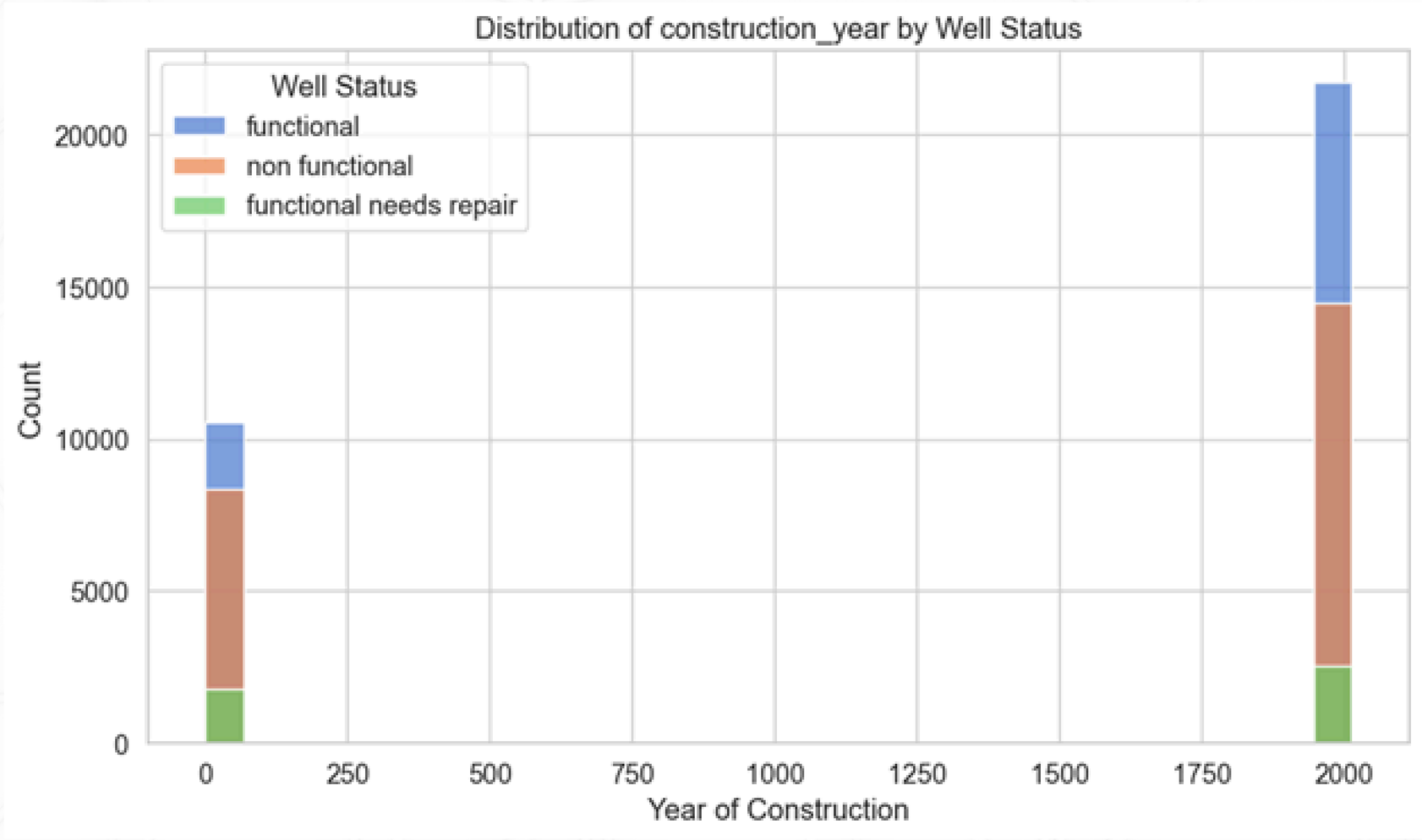
The graph shows that most wells are functional, followed by a significant number of non-functional wells, and very few require repairs.

Box Plot: gps_height vs. status_group.



The boxplot shows that wells at various elevations (gps_height) have similar distributions of functionality, non-functionality, and repair needs.

Histogram: Distribution of construction_year with well status overlay



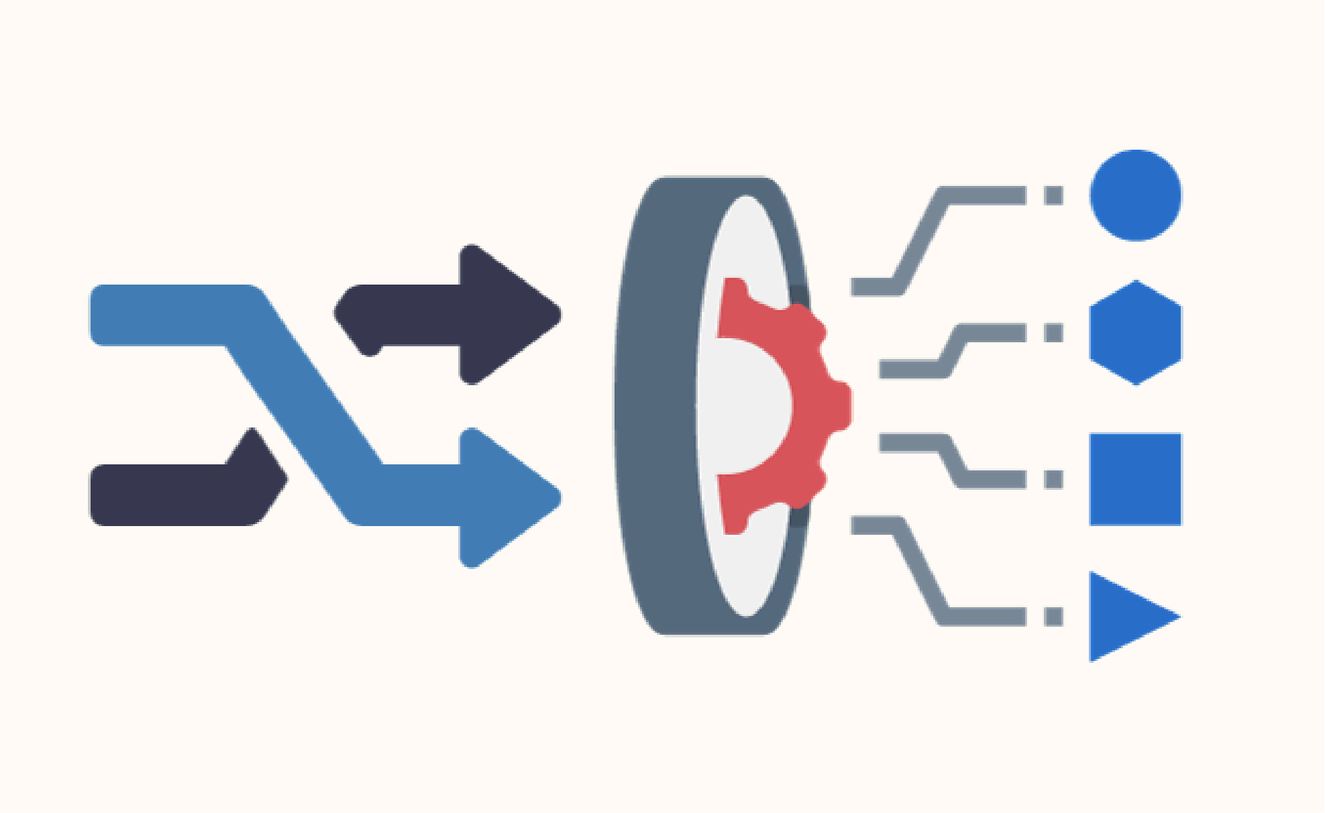
The graph shows that most wells with recorded construction years were built around 2000, while wells without a recorded year are distributed across all statuses.

FEATURE IMPORTANCE

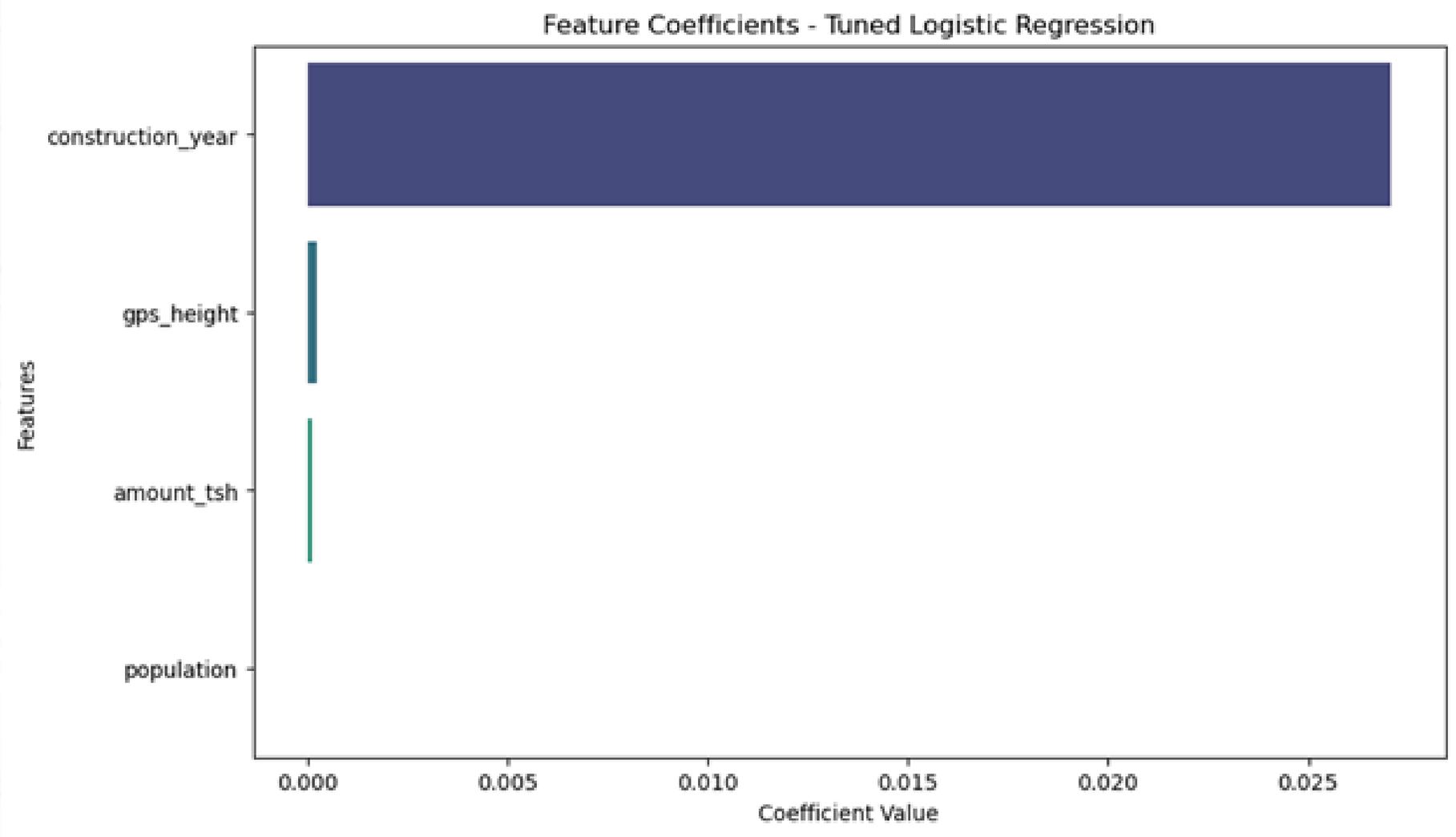
- **Content:**

What Did We Discover?

- The following features play a significant role in predicting well conditions:
  i. gps_height: Elevation of the well location.
  ii. construction_year: Year the well was constructed.
  iii. population: Population dependent on the well.
  iv. amount_tsh: Water output levels of the well.
- These features were derived from models like Random Forest and Logistic Regression.

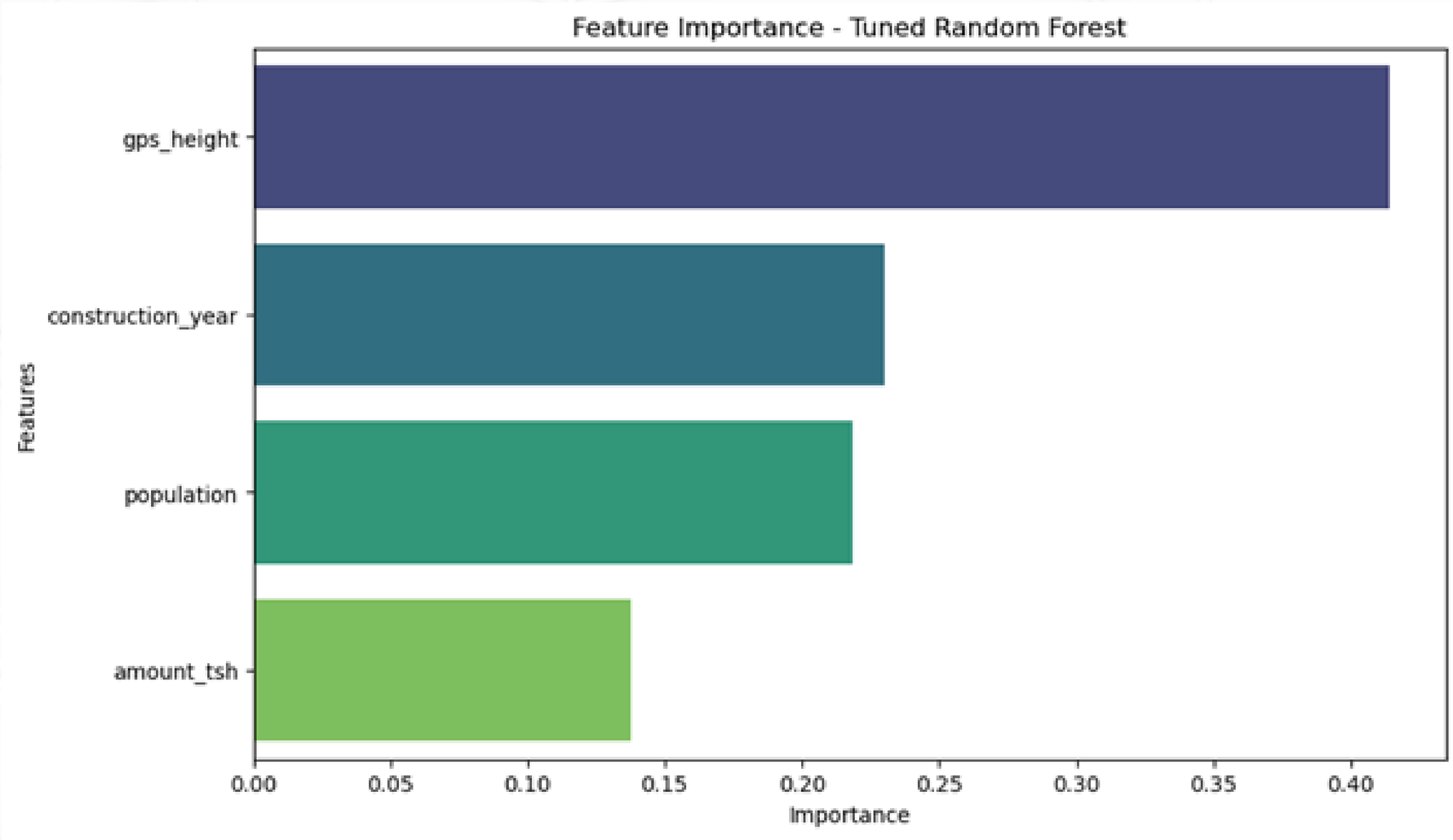


Feature Importance Plot (Bar Chart) from Tuned Logistic Regression.



The graph shows that construction_year has the most significant impact on the logistic regression model, while other features like gps_height, amount_tsh, and population have minimal influence.

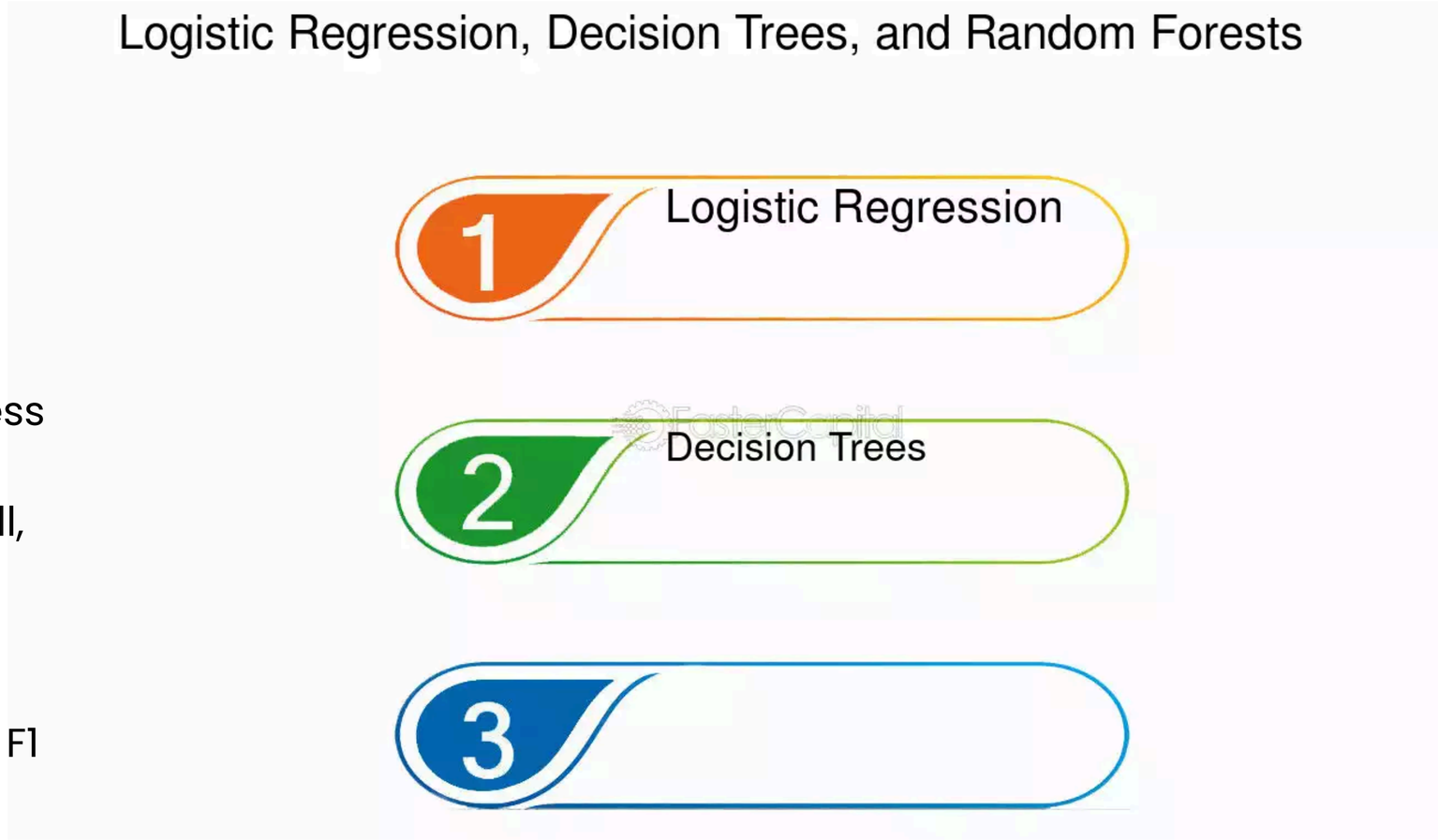
Feature Importance Plot (Bar Chart) from Tuned Random Forest



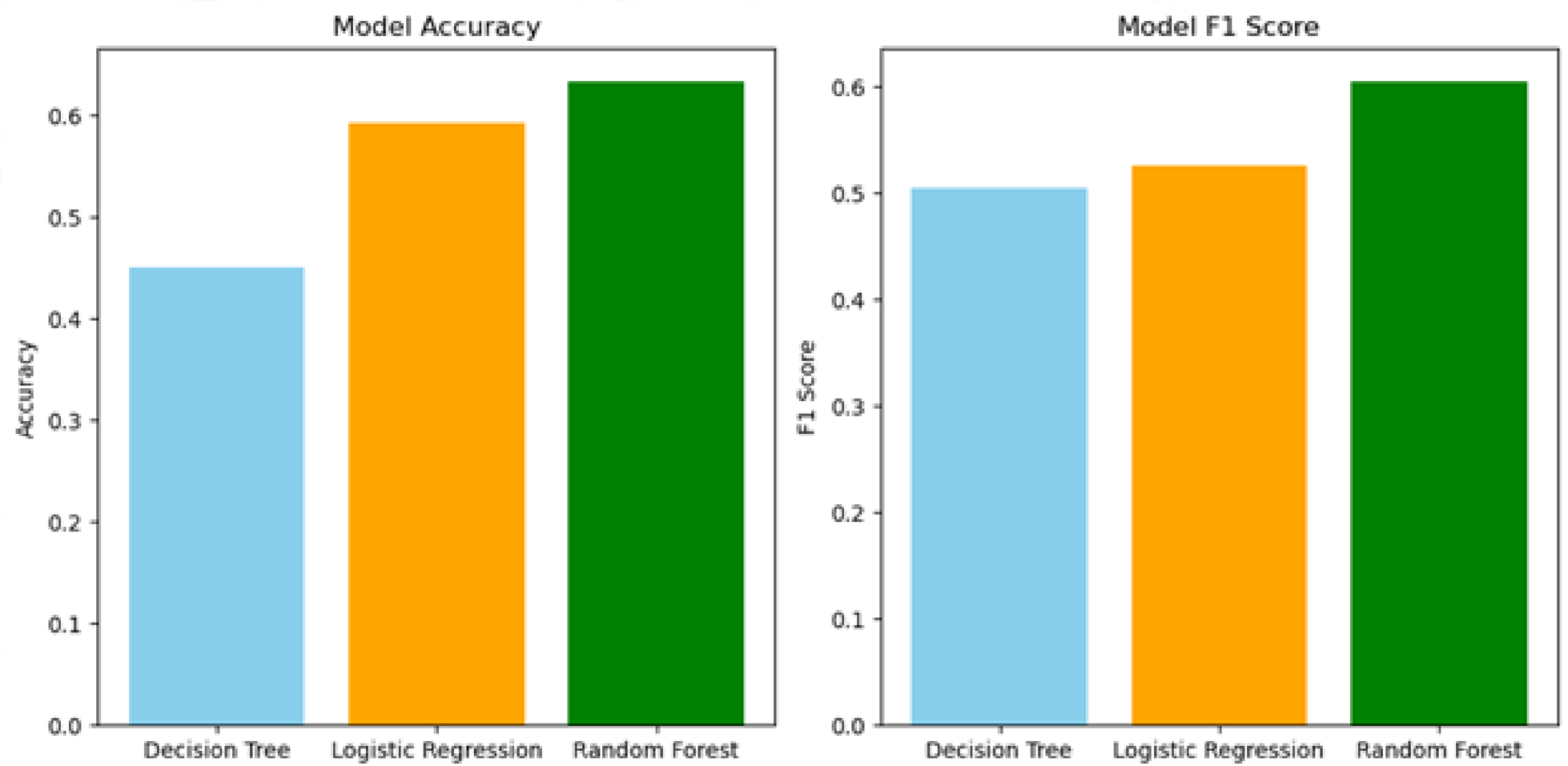
The graph shows that gps_height is the most important feature in the tuned random forest model, followed by construction_year, population, and amount_tsh.

MODEL COMPARISON

- **Content:**
  - **Models Tested:**
    - Decision Tree
    - Logistic Regression
    - Random Forest
  - **Evaluation Metrics:**
    - Accuracy: Measures overall correctness of the model.
    - F1 Score: Balances precision and recall, useful for imbalanced data.
  - **Results:**
  - Random Forest performed the best, achieving 63% accuracy and a weighted F1 score of 61%.
  - .



Bar Chart: Model Comparison (Accuracy and F1 Score).



The graphs show that Random Forest performs the best in both accuracy and F1 score, followed by Logistic Regression, while Decision Tree lags behind.

BEST PERFORMING MODEL

- **Content:**
  - **Model Picked:**
    - Random Forest
  - **Why it was selected:**
    - Highest Accuracy and F1 Score.
    - Robust to outliers and handles complex feature interactions well.
  - **Key Performance Metrics:**
    - Accuracy: 63%
    - F1 Score: 61%

Model	Accuracy	F1 Score
Decision Tree	0.58	0.56
Logistic regression	0.60	0.59
Random Forest	0.63	0.61

ACTIONABLE INSIGHTS

1. Wells in **low gps_height areas** are at a higher risk of failure. **Recommendation:** Prioritize maintenance for wells in lower elevations
2. **Older wells** (constructed before 2000) are more likely to fail. **Recommendation:** Replace or repair older wells.
3. **High population areas** contribute to well wear and tear. **Recommendation**: Allocate resources to high-demand regions.
4. Wells with low **amount_tsh** need monitoring. **Recommendation:** Schedule frequent checks for low-output wells

FINAL PREDICTIONS

- Content:
  - The Random Forest model was used to predict well statuses on new data.
  - Predictions have been saved to a CSV file for stakeholder review and use.

id	Predicted_Status
50785	functional
24806	non functional
8869	functional needs repair

THANK YOU!