DSCI 5240-Final report

Predictive Modeling of Water Pump Functionality

Instructor name

Javier Rubio Herrero

University of North Texas

Group-

Team Members list

1.Shivani Nalanagula

2.Veda Samskruthi Kancherla

3.Vineela Kandunuru Muni

4. Diya Reddy Katpally

**Introduction**

**Project Objective:**

This project's primary goal is to use various modeling approaches to forecast the state of water pumps' functionality. It is suggested to classify a collection of water pumps as functional or non-functional by utilizing a variety of machine learning algorithms and taking into account multiple assets in the data set. Therefore, forecasting future water conditions is essential for stable water delivery systems and effective management of water resources.

**Executive Summary:**

This paper presents the findings of the project assigned to the authors, which required employing machine learning to forecast the functionality status of water pumps. Water is an essential commodity and in rural or regions few can access such regions and water points are very important for the welfare of those people. The focus of this project was to develop the best classifier that could well distinguish between water pumps that are functional, and non-functional, and require repairs based on several characteristics of the water pumps.

Key objectives of the project included:

**Data Exploration and Preparation:** Early engagement with the dataset, removal of outliers, and other discrepancies that are not important in determining the data’s pattern and characteristics before data visualization.

**Model Development:** Model fitting and selecting several machine learning models to select the one that could predict the best of the pump function accurately.

**Feature Engineering:** This entails the activities of data cleaning, normalization, transformation such as feature extraction, and encoding the categorical data which are enabled in the model.

**Model Evaluation and Comparison**: To achieve this, the efforts and performance of each model will be assessed and ranked based on the level of accuracy, precision, recall, and F1-score of each model to find out which one is best suited for the application.

Included in the dataset were characteristics of the water pumps such as the year of installation, funding agency, source type, quality as well as the distance to the nearest established town. We applied three different machine learning models: Among these algorithms, the used algorithms are Logistic Regression, Random Forests, and Gradient boosting classifier. Hence, after careful consideration, the Gradient boosting classifier model was said to have provided the highest accuracy in the prediction of pump functionality.

Out of the data analyzed we established the following basic observations; Age of the pump, water source type, and distance to the nearest facility affected pump functionality status. Among these, the Gradient boosting classifier model revealed these variables as the most critical predictors based on their feature importance.

These findings can help governments and organizations to better allocate funds for maintenance, decide how to invest in new infrastructure, and evaluate the optimization of water distribution. Another beneficial implication of our proposed model is to deploy a mode to guarantee the functionality of water pumps, thus positively contributing to the availability of clean water in devastated areas.

Thus, our project showcases the use and necessity of machine learning algorithms to process and analyze a real-world problem and come up with tangible and dependable insights and a prediction model for water pump maintenance.

**Data Description**

**Dataset Overview:**

The dataset used for this research includes information on water pumps located throughout a certain nation. A single pump is represented by each row of the table, and the data set includes multiple fields that detail the characteristics of the water pump. Predicting each water pump's "Functioning Status" in relation to these attributes is hence the difficult task at hand.

The dependent variable in the dataset provided is:

Functioning Status: We employ this as the dependent variable in each of our three prediction model algorithms. By labeling a water pump as "Functioning" or "Not Functioning," accordingly, it indicates whether the pump is operating or not.

**Features of the dataset:**

Some of the trivial details of the dataset are its purpose is to identify key features of the dataset:

* Water Pump ID: They maintain an identity, something associated with each water pump. It is done to make sure that each record needs to have a unique identifier so that it can be easily identified.
* Water Source Type: It denotes the kind of water source that the pump is tied to; whether it is a well, borehole, river, or lake. Aspects like source reliability and quality can easily be understood with the aid of this feature.
* Water Quality: Quantifies the water quality that is available under two different conditions namely clean and contaminated source. It is crucial to determine the health concerns arising out of the water that is supplied.
* Distance to Nearest Town: This is the distance in kilometers between the water pump site and the nearest town or community where people commonly reside. This is helpful in getting an appreciation of how easily manageable the pump is and the ease of getting spare parts and other logistical factors needed in its operation.
* Population Served: The number of people benefited from the water pump means the total of population that uses water from the pump. This feature is the most crucial one needed to see how the functioning of the pump affects the inhabitants.
* Installation Year: The year the water pump was installed & it is employed in the determination of the age of the pump. Older devices may have some weaknesses which make them fail in the process of pumping.
* Funder: The organization/entity that has funded the installation of water pump. The quality and standards which are followed or required to be maintained can be different in different funds.
* Payment Type: Specifies the following aspects as pertaining to payment of water usage; free, per use, per gallon, etc. This may play negative roles on how often the machines will be maintained and the consequent operations upkeep.
* Water Pump Age: Generated from the name of the installation years meaning the number of actual years that the pump has been used. This assists in determining the. However, it helps in knowing the end user utilization and consequently the wear and tear of the pump.
* Pump Type: Establishes whether it is a hand pump, motorized pump, or a solar pump depending on the kind of water pumping station in the intervention site. There are tremendous varieties of pump types on the market today, and these wetted types vary in terms of their basic maintenance requirements and failure frequency.
* GPS Coordinates: Geographical data in the form of the latitude and longitude of the water pump. This can be useful in identifying the pumps and understanding the geographical locations of the pumps as well as the functionality of the pumps.
* Functioning Status: The target variable which shows if the quantity of water pumped is correct, if the water pump is broken and does not pump the right amount of water, or if the water pump is broken but its components are replaceable. This is what the project is trying to forecast, so it is the main dependent variable expected to be influenced by the independent variables of the study.

These features give a rich description of every pump, allowing for critical examination, as well as the development of strong forecast models to help in water management.

**Data Preparation**

Here’s a summary of the data preparation process focusing on handling missing values and outliers:

**Handling Missing Values:**

* **Categorical Features:**

Features: The various independent variables that were identified include water source type, water quality, payment type, functioning status of an equipped pump, funder, and pump type.

Method: In case of missing Continuous data, use the mode which maintains the most frequent category.

* **Numerical Features:**

Features: Installation Year, Water Pump Age.

Method: When one value is involved, the other can be determined by using the current year.

When both are missing then fill them using median to keep it in a harmony and which gives consistency.

* **String to Numeric Conversion:**

Features: Distance to Nearest Town, Population Served, Water Pump Age

Method: To avoid distorting the data, missing values in string format enclosed by brackets are converted to the median of the corresponding variable.

* **ID Column:**

Feature: Water Pump ID.

Method: Use missing IDs for purposes of data integrity.

* **GPS Coordinates:**

Method: When converting the tables, remove the rows where the GPS coordinates are absent to ensure no compromised data is used.

**Handling Outliers:**

* **Identifying Outliers:**

Features: Distance to the nearest town, the population served, and the age of the water pump.

Method: Interquartile Approach is more effective when it comes to identifying the outliers.

* **Visualization:**

Method: To identify patterns of values and outliers in the columns, express the data in terms of box plots.

* **Removing Outliers:**

Method: Several outliers discovered for the height measurements; exclude these rows so as to clean up the data.

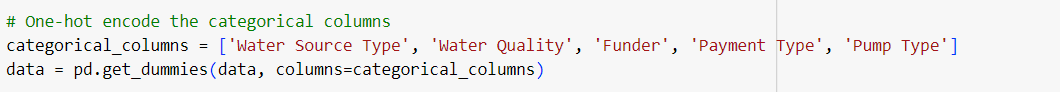
**Post-Cleaning Statistics:**

Method: Perform initial summaries of the cleaning result data by using mean and standard deviation tests on the samples after removing them from the outliers.

**Feature engineering:**

**Encoding Categorical Variables:**

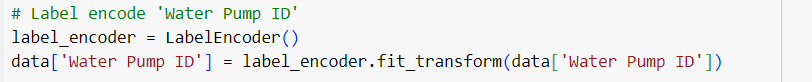
* One-Hot Encoding: Performed on the Water Source Type, Payment Type, Funder and Pump Type variables since these categorical inputs need to be pre-processed for use by a machine learning program.



Explanation: One hot encoding can be regarded as a type of feature scaling since it transforms a categorical feature into a binary matrix pattern. Every of the values of the category is turned into a new column, and the observation is associated with a binary value of 0 or 1 depending on the presence of the category in the observation.

Purpose: This step converts categorical features into an understandable format for the machine learning algorithms since they accept only numerical inputs.

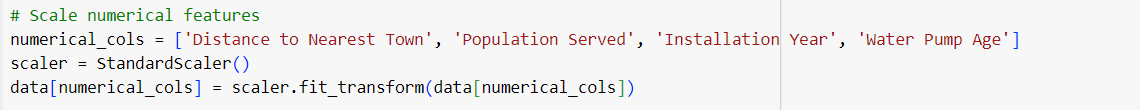
* Label Encoding: Implemented when dealing with binary categorical features and in case one of the classes was represented by the number 0 and the second one – by 1.



Explanation: Label encoding converts categorical values into numeric labels. Here, it is used for the 'Water Pump ID' column, which contains unique identifiers for each water pump.

Purpose: Although 'Water Pump ID' is not a categorical feature with meaningful categories, converting it to numerical format ensures it is processed correctly by the model without being treated as a purely categorical feature.

* Scaling: It is worthy of note that distance and population served were forms of continual data hence Min-Max was used when scaling the data to fit for several algorithms that have sensitivity to this info.



Explanation: Standard Scaler also deals with feature normalization by decomposing the data to remove the mean and dividing by standard deviation. This means that each of the numerical features will have a mean which will be zero and the standard deviation which will be one.

Purpose: When we extract numerical attributes, we need to make sure that the corresponding attributes of different types are on a similar scale, which is essential for many of the machine learning algorithms that use distance metrics (for example Logistic Regression, Gradient Boosting). For instance, if the value range of certain features is big, they’re likely to have huge impacts on the model in case they are not standardized.

**Splitting of dataset (70/30 Split):**

A close-up of text

Description automatically generated

* Firstly we split the dataset into features and target. Here we need to predict the functionality of the pump so it is the target column alone. The features columns include all other columns.
* The feature columns are all columns in the dataset excluding the target column which is Functioning Status, GPS Coordinates and Population Served. The GPS coordinates is not directly useful as features are in the raw form, and Population Served was excluded we felt it may not contribute significantly to the model.
* Then we encoded the target using its two classes into a binary format which makes functioning class as 1 and not functioning as 0.
* To train the model, splitting it into feature and target variables is essential. So we initialized a variable X for input and Y for output target to predict.

A computer screen shot of a computer code

Description automatically generated

* The train\_test\_split function splits the data in the ratio of 70:30 where 70% of the data is used to train the model and the remaining 30% is used for testing the model and it has set the random\_state to 42.
* Then displayed the first few rows of the training and testing data using head function.

**Prediction Algorithms:**

1. **Logistic regression:**

Logistic regression is a statistical technique for analyzing binary data and estimating the probability of occurring one event over another. Instead, it applies the sigmoid function to convert the inputs to a probability within the range of (0,1). The model describes the how the predictor variables are related to the outcome via the equation of a straight line.

A screen shot of a computer code

Description automatically generated

* Logistic Regression is perfect for our model cause it’s easy to interpret, it will explain how each of our features contributes to the probability of having a functioning water pump. It is very easy to apply and requires little computational power, so is best used as a first-pass analysis. Developed for binary classification, it returns a probability, which aids in choosing pumps to inspect according to their failure risk. It is efficient with linearly separable data and fine with any other data as long as the data is preprocessed properly for features. In summary, the method of Logistic Regression provides clear results, simplicity, and information relevance to predict the future condition of the water pump in our project.

A screenshot of a computer

Description automatically generated

* The Logistic Regression model for predicting water pump functionality helped in achieving a training accuracy of 0.54. The precision was 0.52, and the recall was 0.54, for an F1-score of 0.50. The confusion matrix revealed 609 TNs, 157 TPs, 156 FPs, and 496 FNs. According to the classification report, the model predicted non-functional pumps more well (precision of 0.55, recall of 0.80) than functional pumps (precision of 0.50, recall of 0.24). All in all, the accuracy of the model was rather low therefore more detailed feature extraction or using a more complex model might be required.
* The confusion matrix, ROC curve and precision recall curve are shown below:

A blue squares with white text

Description automatically generated A graph of a curve

Description automatically generated

A graph of a graph

Description automatically generated

Fig 1: Confusion matrix, ROC Curve, Precision-Recall curve of Logistic Regression Model

1. **Random Forest Classifier:**

Random Forest is one of the methods of ensemble learning, where during the training process, it builds different decision trees and in the final stage of classification, the most frequent value among trees’ outputs is used, while in the case of regression, the mean value is used instead. It minimizes the problem of overfitting and increases the reliability of results since the outcome is obtained by making averages from various trees.

A screen shot of a computer code

Description automatically generated

* Randomness is helpful in Random Forests since it forms the basis of constructing numerous decision trees that are different; therefore, there is more reliability when the predictions are pooled together than when individual trees are used.

A screenshot of a computer

Description automatically generated

* The Random Forest classifier has a high tendency of overfitting the training data set as depicted by a training accuracy of 1.00 but a mediocre test accuracy of 0.53. Precision and recall are both roughly equal to 0.52 meaning that there is only 52% accuracy in positive prediction while 53% accuracy in positive cases. The confusion matrix shows values such as 464 of the true negatives, 301 of the false positives, 370 of the false negatives, and 283 of the true positives. The classification report gives separate scores for each class, which is helpful for comparison, along with the macro and weighted averages. This implies that although the model may have a very good fit with the training data, the ability to transfer this model to unseen data is poor.

A blue squares with white text

Description automatically generated A graph of a curve

Description automatically generated

A graph of a graph

Description automatically generated

Fig 2: Confusion matrix, ROC Curve, Precision-Recall curve of Random Forest Model

1. **Gradient Boosting Classifier:**

The Gradient Boosting Classifier is an ensemble learning technique that produces an increased number of decision trees; each subsequent tree tries to reduce the error of the previous tree. It uses these trees to build a powerful predictive model and, therefore, increases the model’s efficiency by eliminating bias.

**A computer screen shot of a code

Description automatically generated**

* It is used because it least biases and variances in the given predictive models with high accuracy by learning from the errors of a previous model. This characteristic makes it flexible to solve different types of data and different complexities of problems, making it useful for regression and classification**.**

**A screenshot of a computer

Description automatically generated**

* The result of the Gradient Boosting Classifier has moderate accuracy with the accuracy of 0.66, and a test accuracy of 0.56, which shows that the current model is quite fitting to the training data but somewhat less applicable to the test data set. The precision and recall are 0.55. The accuracy of positive predictions was at 55% while the ability to detect actual positive cases was observed to be 56%. The F1-score is 0.64 which represent the values of precision and recall in their balanced state. The confusion matrix reveals 563 TN values, 202 FP values, 424 FN values, and 229 TP values. The classification report offers metrics for each class and the macro and weighted averages conclude the overall performance.

**A blue and white chart

Description automatically generated** **A graph of a curve

Description automatically generated**

**A graph of a line

Description automatically generated**

Fig 3: Confusion matrix, ROC Curve, Precision-Recall curve of Gradient Boost Classifier Model

**Comparison of models:**

* Accuracy: Comparing the two we observe that Gradient Boosting is slightly better than Random Forest with a higher accuracy of 56%.
* Precision: The performance comparison in terms of precision of Gradient Boosting and Random Forest indicates the former outperforms the latter.
* Recall: What is more, looking at the results, Gradient Boosting perform better than Random Forest in the aspect of recall.
* F1-Score: Regarding the evaluation metric, which is the F1-score, we observe that Gradient Boosting outperforms Random Forest.

Its accuracy, precision, recall and the F1 score are higher than those of the Random Forest and Logistic Regression, which proves that Gradient Boosting is a better model when a classification problem is in question.