MACHINE LEARNING APPROACH FOR THE PREDICTION OF THE STATUS OF TANZANIAN WELLS [COMP4030 CW2 - Data Science and Machine Learning]

*Thomas Cotter
Computer Science
University of Nottingham
Nottingham, England
psytc8@nottingham.ac.uk

*Loo Yang Shen Jason
Computer Science
University of Nottingham
Nottingham, England
hfyyl5@nottingham.ac.uk

***Abstract*—This paper details our approaches and results for the COMP4030 CW2. We have used a number of machine learning algorithms in order to predict the status ('Functional', 'Functional - Needs Repair' & 'Non-Functional') of water wells in Tanzania. This is important as not only does knowing the status of the well help keep the total percentage of functioning wells higher, but it also allows for more effective spending from the government as they no longer would have to send workers out to check the status of the well. Our results suggested that the most important features were ADD RESULTS, and the best classification model as ADD RESULTS**

***Index Terms*—Machine Learning, Data Science, Classification**

I. Introduction

This section will provide an introduction to the dataset and the research questions we have attempted to solve.

A. Research Questions

We have decided that our research question will be:

What factors are most important for determining the status of a well, and how accurately can we classify wells based on these features?.

We choose this question because we are interested in the factors that determine the status of a well, and using ML to try to classify these wells into 1 of 3 classes: Functional, Non-Functional & Functional Needs Repair. From this question, we can think about some follow-up questions. These could include:

- How does the accuracy of the classification model vary with different feature sets and classification algorithms?
- Could we use our results to ensure that wells are built and repaired so that fewer wells are non-functional?

B. Dataset

We will be using the dataset from the Tanzanian Ministry of Water, which contains information on the status of wells in Tanzania to answer our research question. This dataset has 59400 rows, with 40 different features. These 40 features could be broken down into three subgroups which hold information regarding: a) Geographic Location of the Wells. b) Management of the wells. and c) Water Condition of the wells. The dataset is originally split into 2 different files, one for labels and one for the actual data. These can be merged easily with pandas through left join on the "ID" column.

We have done a range of statistical analysis on the dataset, including the value counts and graphical representations of each feature. This can be seen here https://github.com/tomcotter7/dma-project/blob/main/Data%20Understanding%20%26%20Statistical%20Analysis.ipynb. This will also be discussed further in the Results section.

Fig. 1 shows the distribution of the target variable, which is the status of the well. We can see from this that we might need to oversample the 'functional needs repair' class.

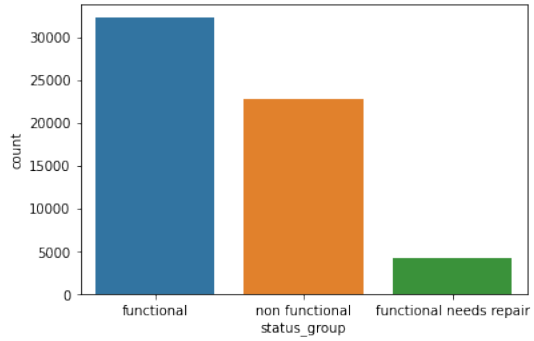

Fig. 1: Distribution of the target variable

C. Management Structure

Talk about the christmas tree structure of our approach to solving this problem.

II. Literature Review

In this section, we review the relevant literature on predicting machine failure, with a focus on studies that have used similar datasets to Pump It Up. Prediciting machine failure is an important area of reasearch, as it has the potential to reduce downtime and maintenance costs. Some machines may be part of a critical infrastructure, so preventing them from failure is of upmost importance.

One study by Pathak et al. (2023) [5] compared the performance of TabNet, a sequential attentive classification architecutre designed for tabular data, and tree-based approaches such as XBBoost. They found that TabNet outpeformed XGBoost, boasting an 83% accuracy compared to XGBoosts 78%. TabNet makes use of Transformers, a machine learning algorithm which uses self-attention to differentially weight the significance of each part of the input. A point of note is that TabNet does not require feature engineering to perform at these standards. While TabNet is an interesting solution, we have not used it in our report, as our primary goal was to showcase an end-to-end machine learning solution, and this includes feature enginering.

Jabeur et al [3] conducted a detailed review of different algorithms for predicting financial distress, including, SVMs, Neural Networks, RandomForest, XGBoost & CatBoost. They found that the ensemble methods worked the most effectively (RandomForest, XGBoost & CatBoost). Ensemble methods combine multiple models in order to improve accuracy.

Similary, Mahabub [4] investigated the effectiveness of ensemble methods to detect fake news. They found that combining different classification methods into a Ensemble Voting Classifier produced the best results.

Finally, Celikmih et al [2] used classification models to predict the failure of aircraft equipment and employed the ReliefF feature selection algorithm. This algorithm estimates feature weights iteratively, according to their ability to make a distinction between neighboring models. They used this to select the best subset of features to feed into their classification model, resulting in the best performance.

III. Methodology

A. Data Analysis / Preprocessing

To perform data analysis, we primarily used the python library Seaborn to create a number of graphs to visualise the data [6]. This library creates much more aesthetically pleasing graphs than matplotlib, and in our opinion is simpler to use. We used this in conjuction with Pandas [1], a python library for advanced data handling. We used the result of our initial data analysis to make decisions for the pre-processsing section (See Results Chapter IV for detailed results).

In the pre-processing section, we followed our christmas tree planning approach, working seperatly, and then combining our results. We each tried different techniques for pre-processing to find the best ones. To perform pre-processing, we both used pandas techniques to format the data in the format we wanted. Pandas is fast data manipulation tool, which makes it perfect for this problem.

The final step in our pre-processing step was feature engineering. Again we used the christmas tree approach, each working separately to build new features and coming back together with our ideas. We also used Pandas for this, for it's speed and ease of use.

B. Data Classification

We used a number of different classification algorithms to classify the data. Initially, we wanted to test the performance of different algorithms to gain an idea of which ones to focus on. The results from this test can be seen in Table. I.

IV. Results

A. Data Analysis

B. Data Preprocessing

C. Classification

V. Discussion

VI. Conclusion

References

[1] Pandas: Python data analysis library. https://pandas.pydata.org/. Accessed: 2023-04-24.

[2] Kadir Celikmih, Onur Inan, and Harun Uguz. Failure prediction of aircraft equipment using machine learning with a hybrid data preparation method. 2020.

[3] Sami Ben Jabeur, Cheima Gharib, Salma Mefteh-Wali, and Wissal Ben Arfi. Catboost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166:120658, 2021.

[4] Atik Mahabub. A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers. *SN Applied Sciences*, 2020.

[5] Karan Pathak and L Shalini. Pump it up: Predict water pump status using attentive tabular learning, 2023.

[6] Micheal Waskom. Seaborn: statistical data visualization. https://seaborn.pydata.org/. Accessed: 2023-04-24.

TABLE I: Initial Classification Results

Algorithm	**Accuracy**	**Precision**	**Recall**
XGB	0.797811	0.751084	0.632675
CatBoost	0.794865	0.745754	0.633800
Bagging	0.793434	0.704477	0.658841
HistGradientBoosting	0.790909	0.743381	0.623847
DT	0.756902	0.643165	0.644585
KNN	0.708754	0.623125	0.563458