COMP 4030 Coursework2 Interim Report

Hong Li psxhl10@nottingham.ac.uk
Wenxuan Zhu alywz30@nottingham.ac.uk

***Title*—The prediction and characteristics of operating condition of waterpoints in Tanzania by using data modelling, analysis and machine learning**

I. Introduction to the Data Set

A. *Introduction to the data set*

The data set comes from the waterpoints dashboard of an open source web app named "Taarifa" which aggregates data from the Tanzania Ministry of Water. The data set of the training part has 59.4k rows and 41 columns including the label that save in the a separate file, in the case of testing data has 14.85k rows. The target variable is status_group that has three possible values functional, non-functional and functional needs repair.

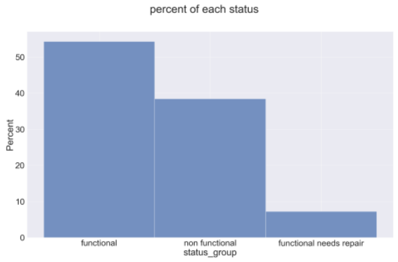

Figure_1

It can be clearly seen that functional and non-functional accounts for the majority, of which functional accounts for around 54.5% and non-functional accounts for about 38.5%, while functional needs repair only accounts for only a small part, about 7%.

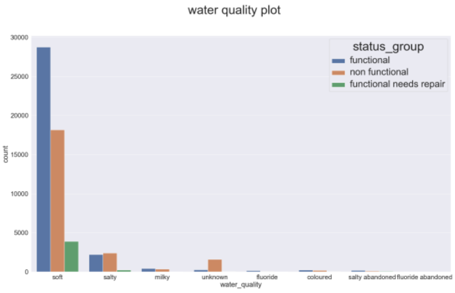

Figure_2

It shows that wells with soft water quality have a really high probability being functional, while if the water is salty then there are almost equal probability between functional and non-functional. But roughly 18,000 wells with soft water quality are currently non-functional.

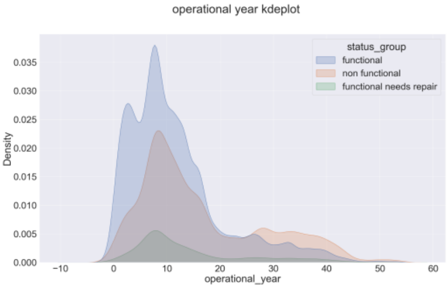

Figure_3

From the figure_3, the operational year of most wells is less than 45 years. And it could be found that the functional well still accounts for the majority when the operational year of the well is under around 26 years, on the other hand, the proportion of the non-functional well is not low. But after the operational year exceeds 26 years, the situation starts to be different. The proportion of the non-functional well has exceeded that of the functional well, which indicates that the operational year of the well has a great impact on the operating condition of the well.

B. *Initial Research Questions*

Question 1:

How to predict the operating condition of a water-point for each record in the data set. (For improving maintenance operations and ensure that clean)

- How to make sure data is accurate, sufficient, and well distributed.
- The correlation of the influence of each feature on the prediction result and how to use them to predict.
- How to choose the right machine learning model for prediction.
- How to judge the quality of forecast results.

Question 2:

Whether natural factors (Like water quality, source type and so on) have more influence on the condition of a water-point or human factors (Like management). (For key decision maker to know they need to pay more attention on natural factors or human factors in the future)

- Which features are natural factors.
- Which features are human factors.
- How to reflect the relationship between a class of eigenvalues and condition of a waterpoint.

Question 3:
Whether the type of water pump has a relatively large relationship with the operating condition of a water-point. (For key decision maker to decide what type of water-point to be used in the future)

II. Summary of Required Data Wrangling and Pre-Processing Approaches

A. *Data Wrangling*

- Deal with the missing values, outliers and bad data.
- Deal with duplicates

B. *pre-processing*

- Normalisation.
- Analysis, compare and select one from the similar features.
- Deal with the data imbalance.
- Add a new feature operational_year which represents the number of years from which the waterpoint was operational.