PUMP IT UP! Predicting the faulty Water pumps

In this Model , we'll be predicting the faulty water pipes that need repair and replacement and the ones which don't need anything i.e. perfectly working.

The motivation behind this Project is to create a model for the welfare of the society and unprivileged areas and also to put up my Data analytics skill sets to a real life problem.

Data Discription :

Total number of attributes: 39

Number of training instances: 59400

Number of testing instances: 14850

Class labels: Functional, Functional but needs repair and Non-functional.

The following are the 39 attributes used to build a model:

1. amount_tsh: Amount of water available to the waterpoint.
2. date_recorded: Date of entry.
3. funder: Person who funded the well.
4. gps_height: Altitude of the well.
5. installer: Organization that installed the well.
6. longitude: GPS Longitudinal coordinate.
7. latitude: GPS Latitudinal coordinate.
8. wpt_name: Name of the waterpoint.
9. num_private: Private number.
10. basin: Geographic water basin.
11. subvillage: Geographic location.
12. region: Geographic location.
13. region_code: Geographic location code.
14. district_code: Geographic location code.
15. lga: Geographic location.
16. ward: Geographic location.
17. population: Population around the well.
18. public_meeting: True/False.
19. recorded_by: Person entering the data.
20. scheme_management: One who operates the waterpoint.

21. scheme_name: Who operates the waterpoint.
22. permit: If a permit exists for the waterpoint or not.
23. construction_year: Year of construction of the waterpoint.
24. extraction_type: Kind of extraction that the waterpoint uses.
25. extraction_type_group: Kind of extraction that the waterpoint uses.
26. extraction_type_class: Kind of extraction that the waterpoint uses.
27. management: How the waterpoint is managed.
28. management_group: How the waterpoint is managed.
29. payment: What the water costs.
30. payment_type: What the water costs. 31. water_quality: Quality of the water.
32. quality_group: The quality of the water.
33. quantity: The quantity of water. 34. quantity_group: The quantity of water.
35. source: Source of water. 4
36. source_type: The source of water.
37. source_class: The source of water.
38. waterpoint_type: The kind of waterpoint.
39. waterpoint_type_group: The kind of waterpoint.

Pre-Processing Techniques

Step 1: Remove redundant attributes

At the outset, we decided to take a close look at the 39 attributes provided. Our first impression upon taking a closer look was that a few attributes seemed redundant. That is, it seemed repetitive to have an Attribute B as part of the model when another Attribute A was already a part of it. For example, waterpoint_type and waterpoint_type_group. Another such example is quantity and quantity group.

Step 2: Remove attributes that have just one category

We found attributes that just had one category in all. Since such an attribute would be in no way useful for predicting the final class label, we decided to remove this attribute from our model. Example: recorded_by had just one category – GeoData Consultants Ltd.

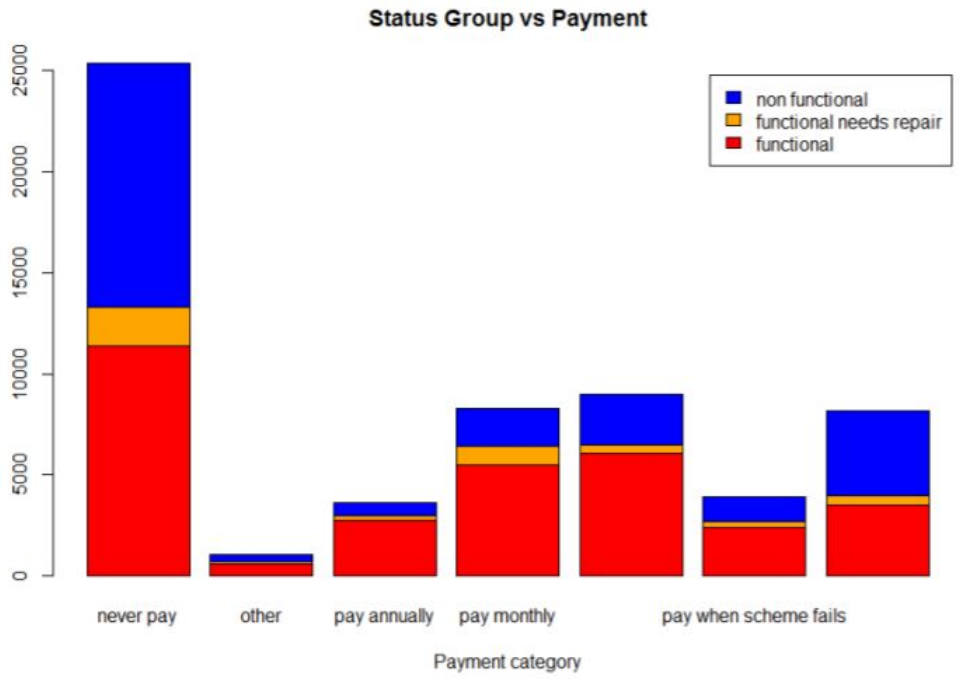
Step 3: Combining attributes to achieve a more meaningful attribute

We figured that the age of a well would strongly have to do with whether it is functional or not. So, we decided to track that in our model. To achieve this, we took attributes date_recorded and construction_year. We took the **difference** in values of these two attributes as the age of the well, as of the time of recording.

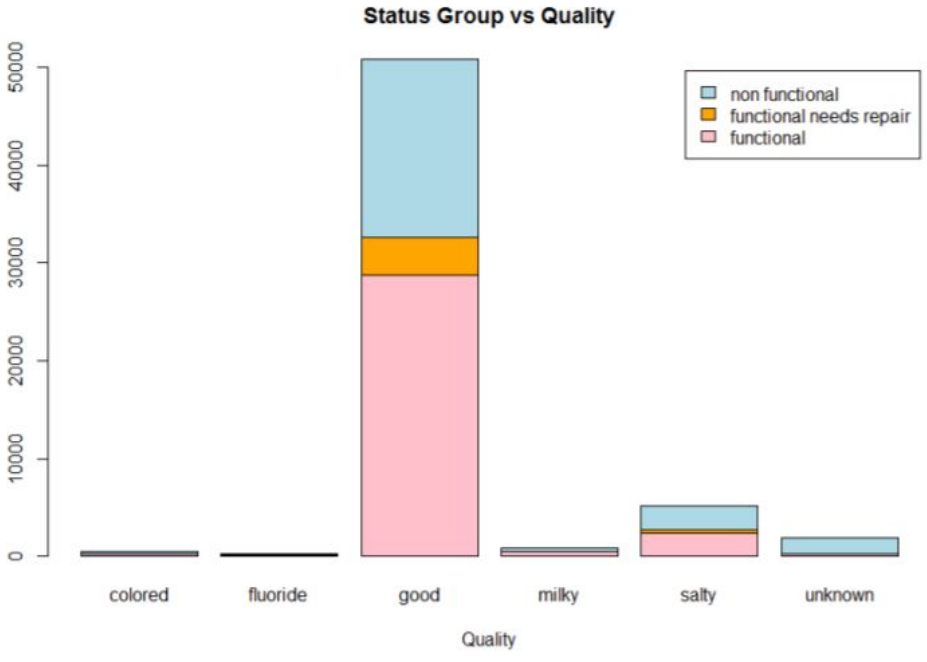
Step 4: Remove rows containing missing or 0 data as required

There were quite a few rows that had missing data (0 or blank) for some of the attributes that we were considering in the model. We decided to not consider the partial data available in these rows and hence, these rows were not fed to the classifier during model building.

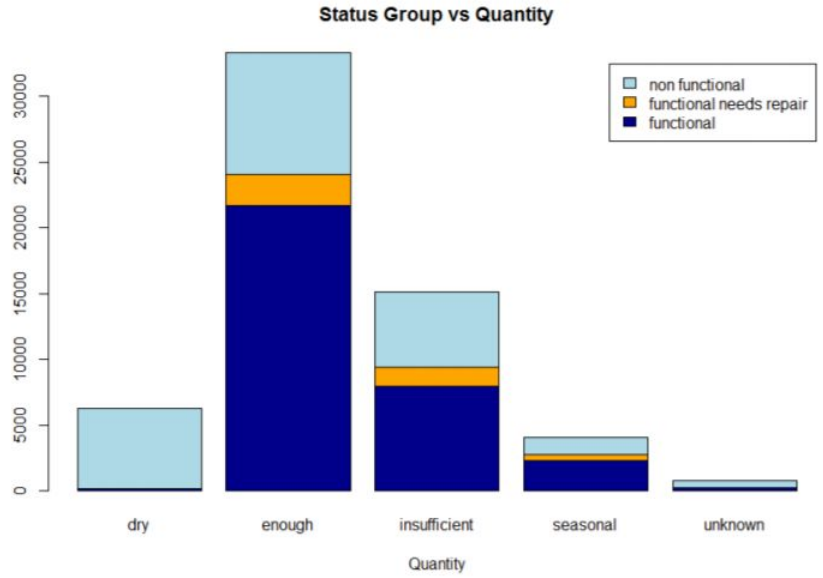
Step 5: Analyze the class labels of the training data



Status group vs Quality



Status group vs Quantity



Approach to Solve :

1) **Logistic Regression :**

Logistic Regression

Error Tolerance	Time taken	Accuracy
0.00001	~ 2 minutes	73.67
0.01	~ 2 minutes	73.03
0.05	~ 2 minutes	57.08
0.1	~ 2 minutes	52.67

2) **Random Forest :**

Random Forest

No. of trees	Max. features	Time taken	Mean Accuracy
120	25	~ 3 minutes	81.18
100	15	~ 2 minutes	80.98
70	10	~ 2 minutes	81.11
50	45	~ 2 minutes	80.09