CS 6375: ML Project

Names of students in your group:

Ankita Patil (asp160730)

Abhilash Gudasi (abg160130)

Shiva Ranga Chawla(sxc167630)

Project Name:

Pump it Up: Data Mining the Water Table

Source of Project:

DrivenData

Dataset Description

* The data is from Taarifa waterpoint dashboard which aggregates the data from the Tanzanian Ministry of Water
* Goal is to predict the operating condition of waterpoint. Basically, we will try to predict which water pumps are functional, which need some repairs and which don’t work at all.

From <https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/data/> we downloaded, the dataset.

Following table describes the information about dataset files

|  |  |
| --- | --- |
| **File** | **Description** |
| Training set values | The independent variables for the training set |
| Training set labels | The dependent variables(status\_group) for each rows in the Training set values |
| Test set values | The independent variables that need predictions |
| Submission format | The format for submitting predictions |

Following are the features in the dataset

|  |  |
| --- | --- |
| **Feature** | **Description** |
| amount\_tsh | Total static head (amount water available to waterpoint) |
| date\_recorded | The date the row/record was entered |
| funder | Who funded the well |
| gps\_height | Altitude of the well |
| installer | Organization that installed the well |
| longitude | GPS coordinate |
| latitude | GPS coordinate |
| wpt\_name | Name of the waterpoint if there is one |
| num\_private | - |
| basin | Geographic water basin |
| subvillage | Geographic location |
| region | Geographic location |
| region\_code | Geographic location (coded) |
| district\_code | Geographic location (coded) |
| lga | Geographic location |
| ward | Geographic location |
| population | Population around the well |
| public\_meeting | True/False |
| recorded\_by | Group entering this row of data |
| scheme\_management | Who operates the waterpoint |
| scheme\_name | Who operates the waterpoint |
| permit | If the waterpoint is permitted |
| construction\_year | Year the waterpoint was constructed |
| extraction\_type | The kind of extraction the waterpoint uses |
| extraction\_type\_group | The kind of extraction the waterpoint uses |
| extraction\_type\_class | The kind of extraction the waterpoint uses |
| management | How the waterpoint is managed |
| management\_group | How the waterpoint is managed |
| payment | What the water costs |
| payment\_type | What the water costs |
| water\_quality | The quality of the water |
| quality\_group | The quality of the water |
| quantity | The quantity of water |
| quantity\_group | The quantity of water |
| source | The source of the water |
| source\_type | The source of the water |
| source\_class | The source of the water |
| waterpoint\_type | The kind of waterpoint |
| waterpoint\_type\_group | The kind of waterpoint |

Following are the labels in the dataset

|  |  |
| --- | --- |
| **Labels** | **Description** |
| functional | the waterpoint is operational and there are no repairs needed |
| functional needs repair | the waterpoint is operational, but needs repairs |
| Non-functional | the waterpoint is not operational |

We explored the data on type of values each feature contained. The results are summarized in below table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **variable** | **dtypes** | **unique\_values** | **zeros %** |
| **0** | id | int | 59400 | 0 |
| **1** | amount\_tsh | float | 98 | 70.1 |
| **2** | date\_recorded | str | 356 | 0 |
| **3** | funder | str, float | 1898 | 0 |
| **4** | gps\_height | int | 2428 | 34.4 |
| **5** | installer | str, float | 2146 | 0 |
| **6** | longitude | float | 57516 | 3.1 |
| **7** | latitude | float | 57517 | 0 |
| **8** | wpt\_name | str | 37400 | 0 |
| **9** | num\_private | int | 65 | 98.7 |
| **10** | basin | str | 9 | 0 |
| **11** | subvillage | str, float | 19288 | 0 |
| **12** | region | str | 21 | 0 |
| **13** | region\_code | int | 27 | 0 |
| **14** | district\_code | int | 20 | 0 |
| **15** | lga | str | 125 | 0 |
| **16** | ward | str | 2092 | 0 |
| **17** | population | int | 1049 | 36 |
| **18** | public\_meeting | bool, float | 3 | 8.5 |
| **19** | recorded\_by | str | 1 | 0 |
| **20** | scheme\_management | str, float | 13 | 0 |
| **21** | scheme\_name | str, float | 2697 | 0 |
| **22** | permit | bool, float | 3 | 29.4 |
| **23** | construction\_year | int | 55 | 34.9 |
| **24** | extraction\_type | str | 18 | 0 |
| **25** | extraction\_type\_group | str | 13 | 0 |
| **26** | extraction\_type\_class | str | 7 | 0 |
| **27** | management | str | 12 | 0 |
| **28** | management\_group | str | 5 | 0 |
| **29** | payment | str | 7 | 0 |
| **30** | payment\_type | str | 7 | 0 |
| **31** | water\_quality | str | 8 | 0 |
| **32** | quality\_group | str | 6 | 0 |
| **33** | quantity | str | 5 | 0 |
| **34** | quantity\_group | str | 5 | 0 |
| **35** | source | str | 10 | 0 |
| **36** | source\_type | str | 7 | 0 |
| **37** | source\_class | str | 3 | 0 |
| **38** | waterpoint\_type | str | 7 | 0 |
| **39** | waterpoint\_type\_group | str | 6 | 0 |
| **40** | status\_group | str | 3 | 0 |

By looking at the dataset we observe that,

* Few features of the dataset contain numerical data whereas few features of the dataset contain categorical data.
* For few features, values are missing.
* Few features are not related or useful in our model prediction.

Hence, in preprocessing of the data we need to address the points which are stated above.

**Coding language used: Python 3.6.4**

**Preprocessing:**

* Even though we got train and test dataset from Data Driven for preprocessing we had to combine them, so we added one Boolean column to each train and test dataset using which we will split back again after the preprocessing is done.
* Missing Latitude and Longitude values are filled up using **geocoder.google("Place name")** package which will populate the latitude and longitude values based on the location and rounding them to 6 decimals.
* We are applying log functions for amount\_tsh and populations columns because few values are too small and some are too big so using log to make it a normal distribution.
* After populating missing values for latitude and longitude and applying log for amount\_tsh we are normalizing them using **MinMaxScaler** with a feature range (0,1).
* We are also dropping few columns like funder, installer, public\_meeting, recorded\_by, ward which are not useful in predicting the functionality of pump.
* As district\_code, construction\_year, region\_code, loc\_clust, num\_private are categorical values but are of integer type, we are converting them to string using str() function and then converting this categorical data into numeric using **LabelEncoder()**.
* We are converting the categorical data like basin, extraction\_type, management\_group etc. into numeric by using **LabelEncoder()** from sklearn.preprocessing.

**Techniques:**

* We are planning to use the models such as
  1. Logistic Regression,
  2. Deep learning,
  3. SVM
  4. ensemble learning methods such as
     1. Random forest and
     2. AdaBoosting.