Machine Learning Engineer Nano degree
Capstone Project Proposal
Predicting the sustainability of Hand Pump
NAGI REDDY SEELAM
February 14th, 2019

Proposal:-
Predicting the Sustainability of hand pump.

Domain Background:-

My project is a tool for development agencies and governments to understand the state of water resource infrastructures in underdeveloped and vulnerable regions of the world. In 2015, an estimated 184 million people living in Africa (Sub-Saharan) relied on hand pumps for their water supply and more than 300 million people lacked access to an improved water source.

Historically, development agencies have been supporting those populations by providing infrastructures, such as hand pumps, but very little attention had been directed to their sustainability and their maintenance ,. Functionality rate of hand pumps in selected Sub-Saharan countries was 36% in 2009, and is respectively 15% and 25% one year and two years after construction in 2016.

My goal is to apply machine learning techniques to evaluate the sustainability of a water scheme using data that is already being collected by managing agencies. We look at different aspects of sustainability, including whether a water point is functional or not, the quantity of water it outputs, and its water quality.

These predictions can shorten the time required for agencies to provide support and organize maintenance operations. Ideally, this project can inform the water sector and help improve the lives of those that rely on such hand pumps for daily tasks and basic human needs.

I take this problem from a challenge by Africa government organization. Already a paper has been published named “smart hand pump” by applying Machine Learning algorithm of Support Vector Machine .Link to this paper is: https://pdfs.semanticscholar.org/3b76/6769a289d345559444d2372eda88f4fc5dbe.pdf

I take this paper as consideration and I want to develop more efficient algorithm to efficiently predict the Sustainability of hand pump.

PERSONAL MOTIVATION:-

As from the starting of this Nano degree program I very much interested in the topics like Neural Networks , Linear Regression, Random Forest algorithm .I want to know the best among them. But as we already know these algorithm results changes from problem to problem .Even though I choose this problem so I can apply number of algorithms on this dataset and choose the best among them(related to this problem).

Problem Statement:-

The aim of this project is to predict the functionality of a hand pump as well as the quantity and the quality of water it outputs based on a minimum of data collected on the field. (Predicting those characteristic of a hand pump at a given point in time can help shorten the time required for managing agencies to provide support and plan targeted maintenance operations of hand pumps in Remote areas).

Datasets and Inputs:-

Data used in this comes from *DrivenData*, an online web platform for data science practice competitions aimed at tackling social challenges. The datasets used are a compilation of data from *Taarifa* and the *Tanzanian Ministry of Water*.

The data has been downloaded from:

'https://s3.amazonaws.com/drivendata/data/7/public/4910797b-ee55-40a7-8668-10efd5c1b960.csv'

This dataset contains data for 59,400 hand pumps, each with 40 features. Some of the features are binary/categorical, and some numerical. These include the location of the water pump, water source type, date of construction, the population it serves, and whether there were public meetings for the point.

Out of 40 features, we use only 24 features which are required for our problem. We transform the categorical features into binary variables through a One Hot Encoding (OHE) process.

Categorical features are:

Funder, installer, subvillage, public_meeeting, scheme_management, scheme_name, permit, construction_year

(These features need to be transform into binary values using OHE)

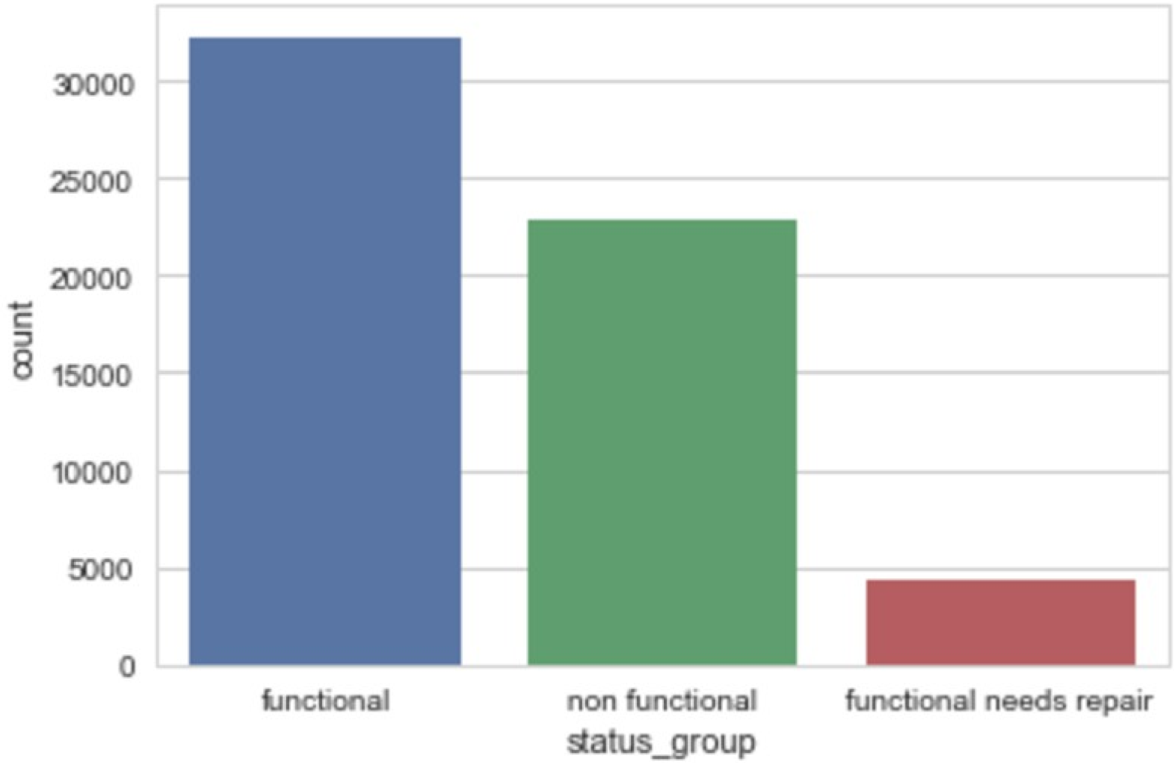
Numerical features are:

amount_tsh, gps_height, population, longitude, latitude, num_private

The water pumps are any of these three categories

1. functional,
2. non-functional, or
3. Functional but need repair.

Now, let us check the count of each fuctional type of the pumps in status_group so that we will understand the functional scenario of the pump status.



By the above result, we can say that there are 54.31% of Functional Pumps, 38.42% of non-functional pumps and 7.27% of functional but which needs to be repaired.

By the above figures, we can roughly estimate that there is 54.31% chance that if we take a random pump in the data to be a functional one.

Data Cleaning and Analysis

Looking at the data, some of the features that seemed discriminative based on human intuition. amount_tsh (amount of water available to water point), gps_height, basin, installer, population, scheme_management, construction year, extraction_type, management_group, water_quality, payment type, source, and waterpoint_type seemed like they could be extremely important in identifying the pump status.

There are null values in our features which are needed to be updated for better training of our model.

In process of cleaning the null values we found out some of the features have high arity. so for features with high arity, let us keep the top 10 values, based on frequency and assign all the remaining values to 11th synthetic value as “others”.

And after all the cleaning process is done we have I split the data in the form of

Trainig set-75% of data
Testing set-25% of data

Solution Statement:-

I will apply some classification algorithms Like Logistic Regression, Random Forest and Neural Networks. And then I will choose the best one among them.

Models were optimized using a grid search with CV to fine tune hyper parameters. Final results were obtained using 5-fold CV with a 75%-25% train –test split. Algorithms were evaluated and optimized based on the F1 score.

Hyper parameters that were optimized are:
number of trees, max depth, max number of features per split, minimum number of samples by leaf, minimum number of sample by split

Benchmark Model:-

The Benchmark model I chosen is the gradient Boosting classifier. Because some previous works have been done on this project using Gradient Boosting Classifier and it gives test accuracy score of 79.37 and train accuracy score of
92.38.

Link to the previous project is: https://towardsdatascience.com/predicting-the-functional-status-of-pumps-in-tanzania-355c9269d0c2

Evaluation Metrics:-

The micro average F1 is used as an evaluation metric. And also accuracy

Precision = TP/TP+FP

TP=total positives

FP=false positives
FN=false negatives

Recall = TP/TP+FN

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Accuracy =Number of correct predictions/Total number of predictions.

Project Design:-

The project is composed of different steps as follows:-

Preprocessing:

- We first performed a feature screening and decided to use only 24 of the 40 features. Our screening process excluded 16 features for the following reasons:
- Irrelevance: some of the features were deemed irrelevant to our project and we decided to exclude them to reduce the computational cost of our algorithm.
- Redundancy: some of the categorical features had exact or almost exact duplicates and we decided to only keep one out of the two or three identical features. In these cases, we kept the most granular feature. In particular, this reduced the number of geographical features.
- We then transformed most of the remaining categorical features into binary variables through a One Hot Encoding (OHE) process.
- Finally, we imputed values where data was missing, and replaced those
- Data points with the mean (continuous) or mode (categorical/binary) of the feature that was missing. This

allowed us to keep over 24,000 data points that were missing at least one feature.

<u>Preprocessing</u>:

- First remove the duplicate features like…..

(extraction_type, extraction_type_group, extraction_type_class), (payment, payment_type), (water_quality, quality_group), (source, source_class), (subvillage, region, region_code, district_code, lga, ward), and (waterpoint_type, waterpoint_type_group)

So these features has to be dropped

- **There are null values in our features which are needed to be updated for better training of our model.**

- Models were optimized using grid search cross-validation (5-fold) to fine tune hyper parameters and final results were obtained using 5-fold cross validation on the test set. Algorithms were evaluated and optimized based on the micro F1 score

- For Neural Networks:
  I use MLP classifier

1)First Step in Training:-

First I want to choose the Benchmark model that will gives a low F1 score

2)Second Step in Training:-

Then I want to choose the different classification algorithms like

Linear Regression
Random Forests
Neural Networks

3)Final Step in Training:-

And finally I choose the best one among the above three which gives high F1 score for all the classes (Functionality, Quality, and Quantity).