Water Well Depth Prediction

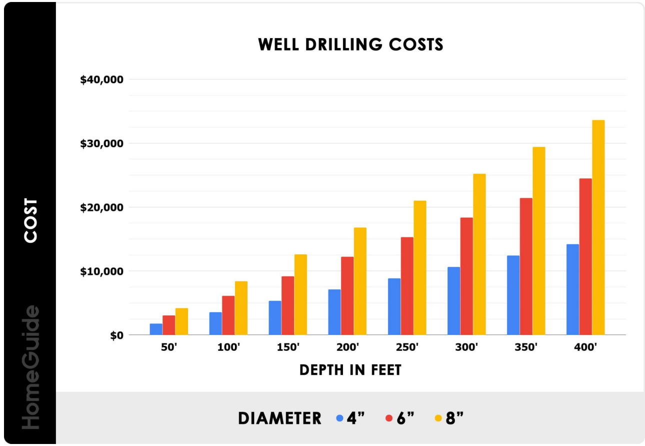
Capstone Project

2022SP_MSDS_498-DL_SEC61 Capstone Class: Data Engineering Capstone

Mark Stockwell
June 5, 2022
MarkStockwell2021@u.northwestern.edu

Water Scarcity is a Global Problem

Four billion people — almost two thirds of the world's population — experience severe water scarcity for at least one month each year. Over two billion people live in countries where water supply is inadequate. Half of the world's population could be living in areas facing water scarcity by as early as 2025. (https://www.unicef.org/wash/water-scarcity)



https://homeguide.com/costs/well-drilling-cost

Well drilling costs are high and rise linearly with well depth. Given limited resources, siting of wells in locations where water is near the surface and close to population centers is critical to solving the global water crisis.

This project aims to use Machine Learning tools and geographic datasets to determine the most cost effective locations to drill with highest probability of finding water.

Project Methodology

PRO

Google Cloud Storage

BigQuery ML

Google Earth Engine

Goal: Develop a regression model to predict water well depth for a given point on map.

Base Data: The US Geological Survey (USGS) is the nation's largest water, earth, and biological science and civilian mapping agency. It provides data on ~1M groundwater wells across the nation. This will be base labeled dataset.

Feature Selection: Precipitation, Lithology, Topography, and Elevation data are correlated with water well depth. Data are available via USGS and Google Earth Engine datasets.

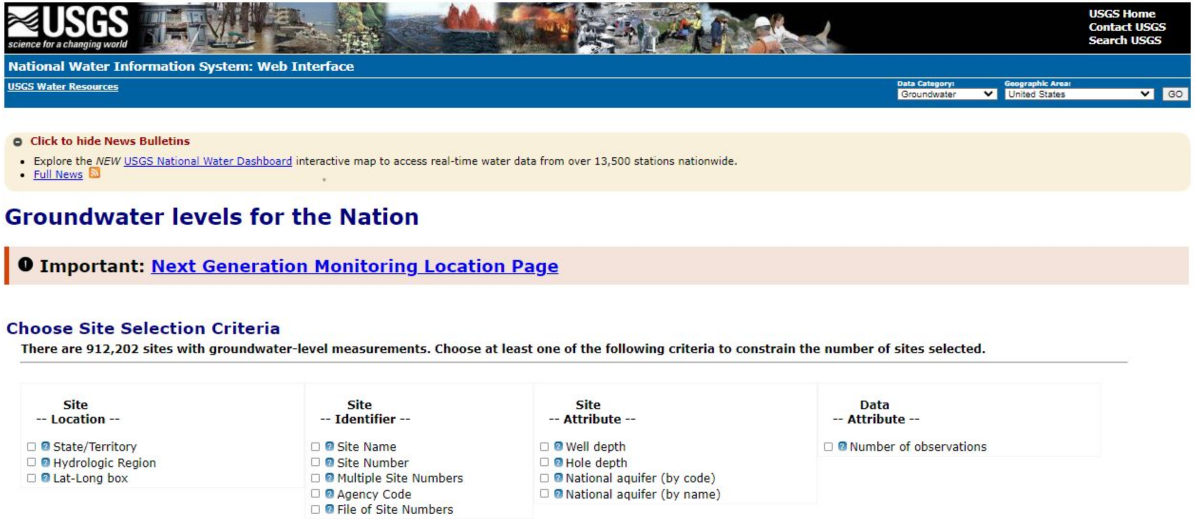
Data Collection: Colab Pro used to collect and process data from USGS and Earth Engine datasets into tabular relational format and stored in Google Cloud Storage buckets.

Data Preparation: BigQuery used to consolidate and link various data assets, including US Census public data.

Machine Learning: BigQuery ML used to build an XGBoost model based upon categorical and continuous variables linked to each ground water well.

User Interface: Google Earth Engine used as a display interface for exploratory data analysis and to link to predictions.

Datasets - Groundwater Wells


USGS
science for a changing world
USGS Home
Contact USGS
Search USGS
National Water Information System: Web Interface
USGS Water Resources
GO

Data Category: Groundwater

Geographic Area: United States

Click to hide News Bulletins

- Explore the *NEW* USGS National Water Dashboard interactive map to access real-time water data from over 13,500 stations nationwide.
- Full News

Groundwater levels for the Nation

Important: Next Generation Monitoring Location Page

Choose Site Selection Criteria

There are 912,202 sites with groundwater-level measurements. Choose at least one of the following criteria to constrain the number of sites selected.

Site -- Location --

- □ State/Territory
- □ Hydrologic Region
- □ Lat-Long box

Site -- Identifier --

- □ Site Name
- □ Site Number
- □ Multiple Site Numbers
- □ Agency Code
- □ File of Site Numbers

Site -- Attribute --

- □ Well depth
- □ Hole depth
- □ National aquifer (by code)
- □ National aquifer (by name)

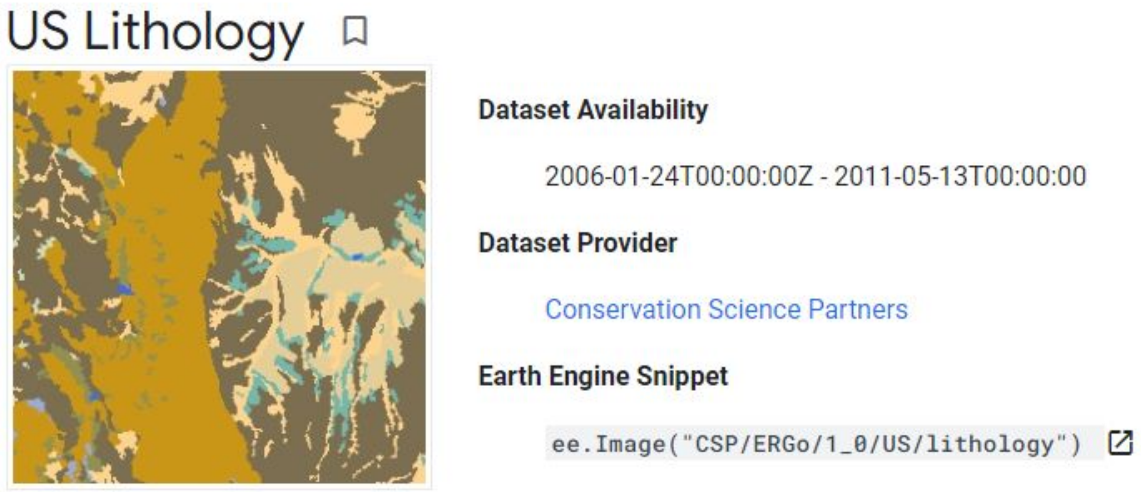
Data -- Attribute --

- □ Number of observations

The USGS National Water Information System (NWIS) contains extensive water data for the nation. The Groundwater database consists of more than 900,000 records of wells, springs, test holes, tunnels, drains, and excavations in the United States. Available site descriptive information includes well location information such as latitude and longitude, well depth, and aquifer.

Link: https://nwis.waterdata.usgs.gov/usa/nwis/gwlevels

Datasets - Lithology



US Lithology

Dataset Availability

2006-01-24T00:00:00Z - 2011-05-13T00:00:00

Dataset Provider

Conservation Science Partners

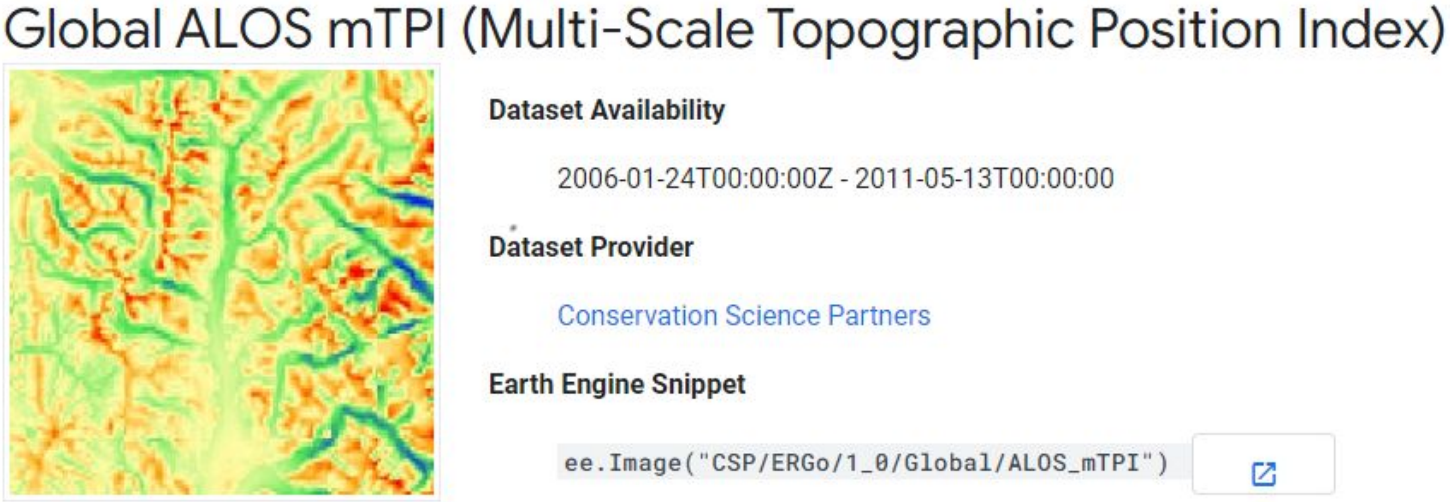
Earth Engine Snippet

```
ee.Image("CSP/ERGo/1_0/US/lithology")
```

The Lithology dataset provides classes of the general types of parent material of soil on the surface. The Conservation Science Partners (CSP) Ecologically Relevant Geomorphology (ERGo) Datasets, Landforms and Physiography contain detailed, multi-scale data on landforms and physiographic (aka land facet) patterns. The original purpose for these data was to develop an ecologically relevant classification and map of landforms and physiographic classes that are suitable for climate adaptation planning.

Link: https://developers.google.com/earth-engine/datasets/catalog/CSP_ERGo_1_0_US_lithology

Datasets - Topography



Global ALOS mTPI (Multi-Scale Topographic Position Index)

Dataset Availability

2006-01-24T00:00:00Z - 2011-05-13T00:00:00

Dataset Provider

Conservation Science Partners

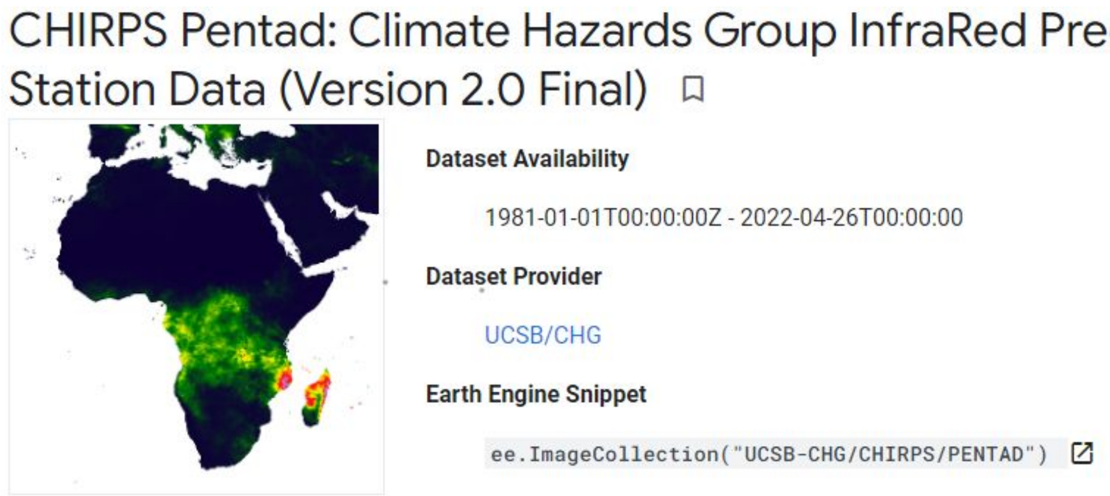
Earth Engine Snippet

```
ee.Image("CSP/ERGo/1_0/Global/ALOS_mTPI")
```

The mTPI distinguishes ridge from valley forms. It is calculated using elevation data for each location subtracted by the mean elevation within a neighborhood. mTPI uses moving windows of radius (km): 115.8, 89.9, 35.5, 13.1, 5.6, 2.8, and 1.2. It is based on the 30m "AVE" band of JAXA's ALOS DEM (available in EE as JAXA/ALOS/AW3D30_V1_1).

Link: https://developers.google.com/earth-engine/datasets/catalog/CSP_ERGo_1_0_Global_ALOS_mTPI

Datasets- Precipitation



CHIRPS Pentad: Climate Hazards Group InfraRed Precipitation With Station Data (Version 2.0 Final)

Dataset Availability

1981-01-01T00:00:00Z - 2022-04-26T00:00:00

Dataset Provider

UCSB/CHG

Earth Engine Snippet

```
ee.ImageCollection("UCSB-CHG/CHIRPS/PENTAD")
```

Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) is a 30+ year quasi-global rainfall dataset. CHIRPS incorporates 0.05° resolution satellite imagery with in-situ station data to create gridded rainfall time series for trend analysis and seasonal drought monitoring.

Link: https://developers.google.com/earth-engine/datasets/catalog/UCSB-CHG_CHIRPS_PENTAD

Data Prep - Base Groundwater Sites

- Colab used to loop through all states using USGS public data download URL
- Data for each state written in raw and tsv format, loaded to BQ
- BQ view used to consolidate data to single table
- State, County attributes appended from public BQ datasets

```python
import requests
import json
import pandas as pd
from google.cloud import bigquery
import urllib.request
import os

base url = "https://nwis.waterdata.usgs.gov/nwis/gwlevels?state_cd={state_cd}"
base url = base url +
"&group key=NONE&format=sitefile output&sitefile output format=rdb"
base url = base url + "rdb compression=file&list_of_search_criteria=state_cd"
column_list = [ 'agency_cd' , ... 'sv_count_nu' ]

for c in column list:
  base_url = base_url + "&column_name=" + c

for state cd in fips codes states:
  f = open("groundwater sites " + state cd + ".tmp", 'w')
  f2 = open("groundwater sites " + state cd + ".tsv", 'w')
  url = base_url.replace( "{state_cd}",state_cd.lower())
  payload={}
  headers = {}
  response = requests.request( "GET", url, headers=headers, data=payload)
  print(state cd,"Response code:" , response.status_code)
  if (response.status_code> 229):
    print("ERROR")
    break;
  else:
    f.writelines(response.text)
    f.close()
...
  f2.close()
```

Northwestern

Data Prep - Image Attribute Data

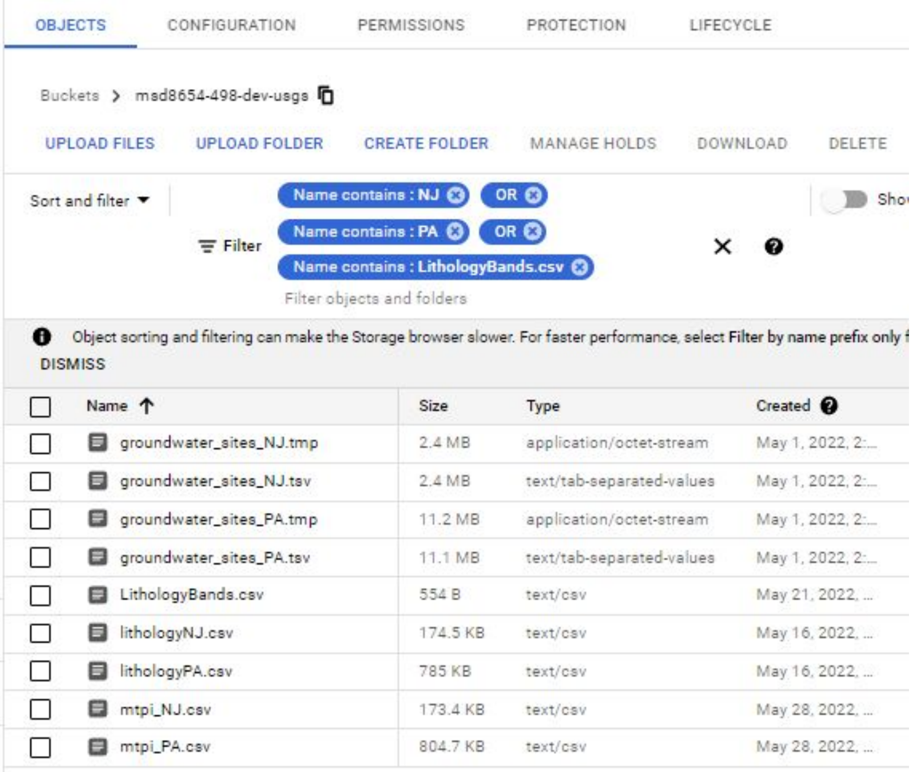
- Image data comes from publicly available datasets on Earth Engine Catalog
- Image data is composed of polygons with attributes storing information about the polygon, i.e. soil type, elevation, precipitation
- For each groundwater site, the location (lat, long) is passed to the image and the information retrieved.
- The site/attribute information is written to a file and uploaded to bucket.

```
from time import sleep
import pandas as pd

dem = ee.Image( 'CSP/ERGo/1_0/US/lithology' )

def get lith(long: float, lat: float):
  xy = ee.Geometry.Point([ long,lat])
  data = dem.sample(xy,  10).first().get( 'b1').getInfo()
  return data

for j, row in df state cds.iterrows():
  state postal abbreviation = row[ "state postal_abbreviation" ]
  print("state cd:", state postal abbreviation)
  df filtered = df[df.state cd==row[ "state fips code" ]]
  filename = f'lithology{state_postal_abbreviation} .csv'
  file = open(filename, 'w')
  for i, row in df_filtered.iterrows():
    try:
      val = get lith(row[ "dec long va" ], row["dec lat va" ])
      file.writelines( str(i) + "," + row["site_no" ] + "," + str(val) + '\n')
    except BaseException as err:
      print(f"   Unexpected {err}, {type(err)}")
      print('ERROR processed:' ,i,row.to_json(), val,  " "*10, datetime.datetime.now())
      continue

  file.close()
  upload_blob( 'msd8654-498-dev-usgs' ,filename, filename)

print(datetime.datetime.now(),  'END')
```

Northwestern

Data Prep - Cloud Storage Files

- After running data extraction process, the cloud storage bucket will contain all the files needed to load BigQuery.
- Naming conventions:
  - groundwater_sites_<state>.tsv - wide file with groundwater site no., lat/long, elevation, well depth, county/state.
  - <attribute>Bands.csv - lookup information for image attributes
  - lithology<state>.csv - data from images with site no., lithology category
  - mtpi<state>.csv - data from images with topographical index for each site.
  - precipitation<state>.csv - precip data for each site



Buckets > msd8654-498-dev-usgs

Name contains : NJ OR
Name contains : PA OR
Name contains : LithologyBands.csv

Filter objects and folders

Object sorting and filtering can make the Storage browser slower. For faster performance, select Filter by name prefix only f

□	Name ↑	Size	Type	Created
□	groundwater_sites_NJ.tmp	2.4 MB	application/octet-stream	May 1, 2022, 2:...
□	groundwater_sites_NJ.tsv	2.4 MB	text/tab-separated-values	May 1, 2022, 2:...
□	groundwater_sites_PA.tmp	11.2 MB	application/octet-stream	May 1, 2022, 2:...
□	groundwater_sites_PA.tsv	11.1 MB	text/tab-separated-values	May 1, 2022, 2:...
□	LithologyBands.csv	554 B	text/csv	May 21, 2022, ...
□	lithologyNJ.csv	174.5 KB	text/csv	May 16, 2022, ...
□	lithologyPA.csv	785 KB	text/csv	May 16, 2022, ...
□	mtpi_NJ.csv	173.4 KB	text/csv	May 28, 2022, ...
□	mtpi_PA.csv	804.7 KB	text/csv	May 28, 2022, ...

Northwestern

Data Prep - BigQuery Views

*.csv files are defined as external tables:

```
CREATE OR REPLACE EXTERNAL TABLE
  `msd8654-498-dev.usgs.mtpi_bands`
  ( id INT64,
    site_no STRING,
    mtpi_band STRING )
    OPTIONS ( format = 'CSV',
    uris =
['gs://msd8654-498-dev-usgs/mtpi*.csv'] )
```

```
SELECT
  gs.site no,
  gs.station nm,
  gs.dec lat va,
  gs.dec long va,
  gs.district cd,
  gs.state cd,
  gs.county cd,
  gs.country_cd,
  gs.alt va,
  lithology bands. Value AS lithology band,
  lithology bands. Description AS lithology_type,
  SAFE CAST(mtpi band  AS INT64) mtpi_band,
  gs.well_depth_va
FROM
  `usgs.groundwater_sites` gs
        INNER JOIN
          `usgs.lithology` lithology
        ON
          gs.site no =lithology.site_no
        INNER JOIN
          `usgs.lithology_bands` lithology_bands
        ON
          lithology.lithology_band  = lithology_bands. Value
        INNER JOIN
          `usgs.mtpi_bands` mtpi_bands
        ON
          gs.site_no =mtpi_bands.site_no  ...
...
```

Groundwater site base table is loaded with csv data and joined to multiple dimensions.

Northwestern

ML Model - Creation

BigQuery ML syntax used to create and train the model.

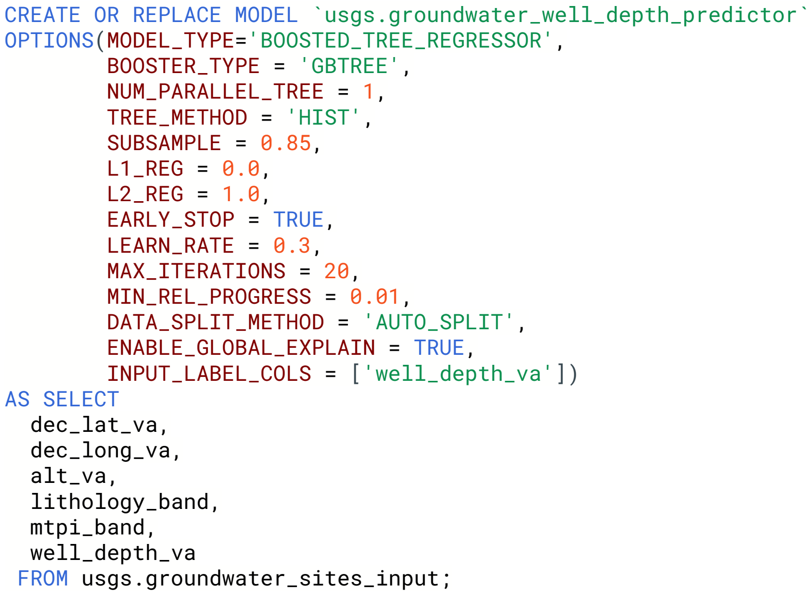
The dependent variable we are trying to predict is well depth.

The select statement contains all the features used in the model.

The BOOSTED_TREE_REGRESSOR model type leverages the XGBOOST library.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.

It was selected for efficiency and good performance with large amounts of tabular/categorical data.

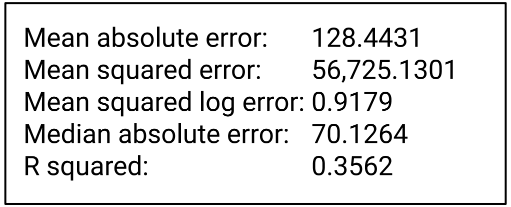


```
CREATE OR REPLACE MODEL `usgs.groundwater_well_depth_predictor`
OPTIONS(MODEL_TYPE='BOOSTED_TREE_REGRESSOR',
        BOOSTER_TYPE = 'GBTREE',
        NUM_PARALLEL_TREE = 1,
        TREE_METHOD = 'HIST',
        SUBSAMPLE = 0.85,
        L1_REG = 0.0,
        L2_REG = 1.0,
        EARLY_STOP = TRUE,
        LEARN_RATE = 0.3,
        MAX_ITERATIONS = 20,
        MIN_REL_PROGRESS = 0.01,
        DATA_SPLIT_METHOD = 'AUTO_SPLIT',
        ENABLE_GLOBAL_EXPLAIN = TRUE,
        INPUT_LABEL_COLS = ['well_depth_va'])
AS SELECT
  dec_lat_va,
  dec_long_va,
  alt_va,
  lithology_band,
  mtpi_band,
  well_depth_va
 FROM usgs.groundwater_sites_input;
```

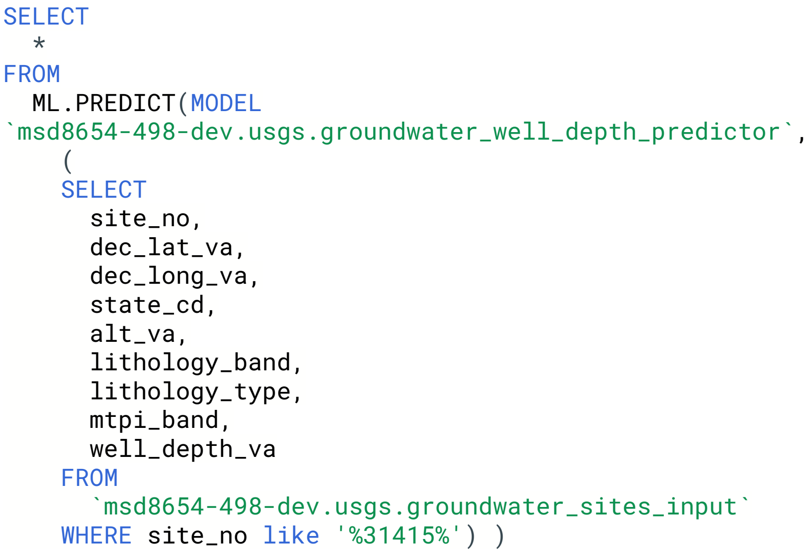
Link: The CREATE MODEL statement for boosted tree models using XGBoost | BigQuery ML | Google Cloud

ML Model - Usage and Evaluation

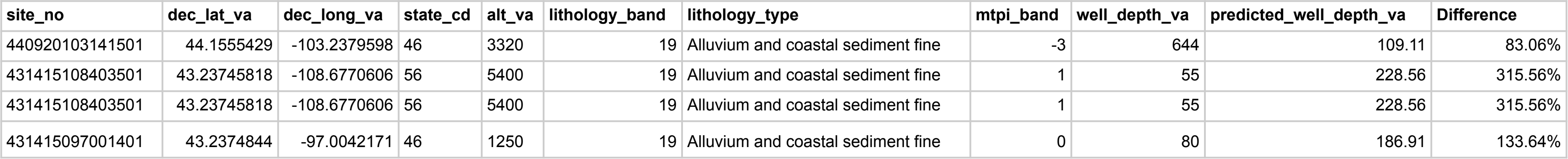
Initial evaluation indicates model performs poorly. Either additional features need to be added or alternatively different model type.



```
Mean absolute error:      128.4431
Mean squared error:       56,725.1301
Mean squared log error:   0.9179
Median absolute error:    70.1264
R squared:                0.3562
```

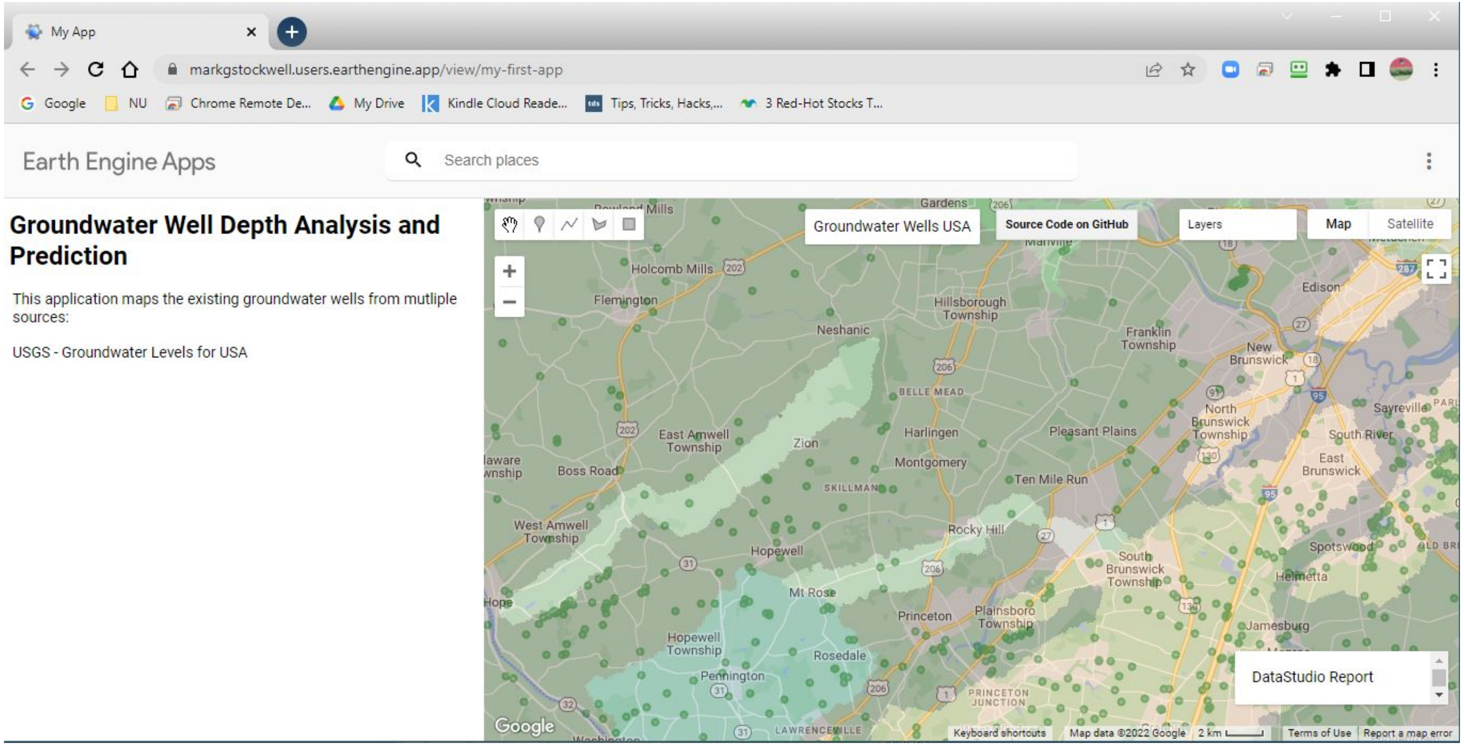


```
SELECT
  *
FROM
  ML.PREDICT(MODEL
`msd8654-498-dev.usgs.groundwater_well_depth_predictor`,
    (
    SELECT
      site_no,
      dec_lat_va,
      dec_long_va,
      state_cd,
      alt_va,
      lithology_band,
      lithology_type,
      mtpi_band,
      well_depth_va
    FROM
      `msd8654-498-dev.usgs.groundwater_sites_input`
    WHERE site_no like '%31415%') )
```



site_no	dec_lat_va	dec_long_va	state_cd	alt_va	lithology_band	lithology_type	mtpi_band	well_depth_va	predicted_well_depth_va	Difference
440920103141501	44.1555429	-103.2379598	46	3320	19	Alluvium and coastal sediment fine	-3	644	109.11	83.06%
431415108403501	43.23745818	-108.6770606	56	5400	19	Alluvium and coastal sediment fine	1	55	228.56	315.56%
431415108403501	43.23745818	-108.6770606	56	5400	19	Alluvium and coastal sediment fine	1	55	228.56	315.56%
431415097001401	43.2374844	-97.0042171	46	1250	19	Alluvium and coastal sediment fine	0	80	186.91	133.64%

User Interface - Earth Engine App



Groundwater Well Depth Analysis and Prediction

This application maps the existing groundwater wells from mutliple sources:

USGS - Groundwater Levels for USA

https://markgstockwell.users.earthengine.app/view/my-first-app

THANK YOU!