# **California Ground Water Dataset : Prediction of Groundwater depth**

NOTE: GSE\_GWE is term that will refer to the depth of ground water. GSE is ground surface elevation and GWE is ground water elevation. The difference between the two is the comparable groundwater depth for all wells.

INTRODUCTION

{Project Proposal}

DATASET

Groundwater, which is found in aquifers below the surface of the earth, is one of our most important natural resources. Groundwater provides drinking water for a large portion of California, nay, the nation's population. It also supplies business and industries, and is used extensively for irrigation. California depends on groundwater for a major portion of its annual water supply, particularly during times of drought. This reliance on groundwater has resulted in overdraft and unsustainable groundwater usage in many of California’s basins, particularly so in the San Joaquin River basin.

![What is groundwater](../images/groundwater.png)

**What is groundwater?**

Groundwater is water that exists underground in saturated zones beneath the land surface. The upper surface of the saturated zone is called the water table. Groundwater is a part of the natural water cycle. Some part of the precipitation that lands on the ground surface infiltrates into the subsurface. The part that continues downward through the soil until it reaches rock material that is saturated is groundwater recharge. Water in the saturated groundwater system moves slowly and may eventually discharge into streams, lakes, and oceans. An aquifer is a body of rock and/or sediment that holds this groundwater.

The water level in an aquifer that supplies water to a well does not always remain the same. Factors affecting groundwater levels that are studied in this project include:

1. Droughts  
2. Seasonal variations in precipitation  
3. Reservoir levels  
4. Pumping for human needs such as domestic, agriculture and industrial

If water is pumped at a faster rate than an aquifer is recharged by precipitation or other sources  
of recharge, water levels drop. This can happen during drought, due to the extreme deficit of rain.

The analysis is performed against the backdrop of the [Sustainable Groundwater Management Act](https://water.ca.gov/programs/groundwater-management/sgma-groundwater-management) that was passed in 2014 in California. SGMA requires locals agencies to form groundwater sustainability agencies (GSAs)for the high and medium priority basins.  
  
Each dataset is acquired at the available granularity and transformed to features at a township-range and year granularity.

Graphical user interface, text

Description automatically generated

ACQUIRING DATA AND CLEANING

Provide links to each dataset ETL documentation.

FEATURE PREPROCESSING USING SKLEARN PIPELINE

Graphical user interface, text, application, Word, email

Description automatically generated

Chart, histogram

Description automatically generated

*We have several variables that are null as seen in the chart above. A pipeline was created using SKLearn to perform the following imputations and transformations.*

**Vegetation dataset  
Cause**We do not have data for the above vegetation after 2014 for any of the TownshipRanges. That is, it is absent for the years [2015, 2016, 2017, 2018, 2019, 2020, 2021]. **Imputation**  
We will be making an assumption that the vegetation landscape of a TownshipRange does not radically differ from year to year. The vegetation values will carry over from 2014 for these vegetation types for each of the township ranges. The imputation is carried out by finding the mean of the data for all Townships. This will be the only available data in the year 2016. This is then filled into the rest of the years for each TownshipRange using a merge.

**Crops dataset  
Cause**We only data for 3 years over the analysis survey, 2014, 2016, 2018 **Imputation**   
We assume little year-to-year variation in crop farming and extended the data for the missing years (2015, 2017, 2019) with the data from the previous years (i.e. we assume that 2015 crops = 2014 crops). We understand that farming practices like crop rotation would challenge these assumptions. The imputation is carried out by sorting the dataframe by year and then township and then forward filling the data over to the years following. **Soils dataset  
Cause**  
We have data for soils for 2016 and not for other years. **Imputation**We will be making an assumption that the basic nature of the soil of a TownshipRange does not radically differ from year to year. As we do not expect the soil type to change from year, the 2016 soil data are used for all the other years. The imputation is carried out by finding the mean of the data for all Townships. This will be the only available data in the year 2016. This is then filled into the rest of the years for each TownshipRange using a merge.

**Well Completion dataset  
Cause**  
Absence of well completion reports filed in township range in certain, or all years causes values to be null.  
**Imputation**We impute the GROUNDSURFACEELEVATION\_AVG for a TownshipRange as the mean of the values over all years for that TownshipRange. This value that is stored in the data is the average of the elevation of the ground where wells were constructed in that TownshipRange and hence elevation was known.

But this does not cover all the NaNs since 76 TRs do not have any value at all and they should be filled with the average elevation of the entire area.

For the other variables since they are the average of the well yield, static water level, top of perforated interval, bottom of perforated interval and total completed depth of the wells constructed in each TownshipRange and so if there are no wells newly constructed, in a township range in a certain year, this value will be set to 0 as these variables are well specific measurements.

**Reservoir dataset   
Cause**Reservoir data is available from the California reporting stations from 2018, we are missing the data from 2014 through 2017.  
**Imputation**  
Since in these years, California was stricken with drought, we will impute the values to be the minimum of the PCT\_OF\_CAPACITY for that TownshipRange.

**Population dataset  
Cause**Population density is missing for the final year of 2021 for all townships. **Imputation**  
We will be using the previous years(2020's) trend over the year 2019 and add to the previous years (2020) value for each TownshipRanges for the missing values in year 2021.

**Challenges faced :** Pipeline steps are executed serially, where the output from the first step is passed to the second step, and so on. ColumnTransformers are different in that each step is executed separately, and the transformed features are concatenated at the end. By default, any columns you pass into the ColumnTransformer that aren’t specified to be transformed will be dropped (remainder='drop'). If you have columns that you want to include but do not need to be transformed, specify remainder='passthrough'.

CHOICE OF EVALUATION SCORE METRIC

Baseline

Since this is a regression problem with continuous variables, the evaluation scores find the difference (error) between the predicted value and the observed value. Some of the choices are **M**ean **S**quare **E**rror, **R**oot of the **M**ean **S**quare **E**rror and **R**oot **M**ean **S**quare **E**rror. We can also look into the mean magnitude of the errors in a set of predictions (**M**ean **A**bsolute **E**rror). These are negatively oriented scores which means lower scores are better. Also, the scores where the errors are squared give a higher weight to large errors and used when large errors are undesired. They ignore the direction of error.

R2 (R squared), MSE (Mean Squared Error) or RMSE (Root Mean Squared Error)

For the initial comparison of algorithms, before hyper tuning and for generating a list of models for PyCaret to compare, R squared gave us a quick indication of the relative performance and was relied on. R2 provides the proportion of the variance for the target that’s explained by selected features in the model. It is independent of the scale of the features and ranges from 0 to 1. A negative value implies worse than mean model. When evaluating an algorithm, it is prudent to look into multiple regression scores and not just R-squared since the acceptable threshold of the error will depend on the distribution of the target value itself. To keep the groundwater depth prediction difference from the mean in check we additionally looked into MSE and RMSE.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **R2 Score** | **MSE** | **RMSE** |
| Dummy Regressor | -0.0571 | 17442.601 | 132.070 |
| Extra Tree | 0.4340 | 6499.055 | 68.117 |
| XGBoost |  |  |  |
| SVM |  |  |  |

SCALING

Many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularization of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

**IMPACT OF SCALING WITH MINMAX OR STANDARD SCALER**

1.There is a difference in clustering when the features are scaled using MinMaxScaler (no negative values for the range set) versus when they are scaled using StandardScaler (required for algorithms such as SVM) . Features scaled using MinMaxScaling do not show the large area of incorrectly clustered data points.

Chart

Description automatically generated

Figure 1: Clustering with features scaled using StandardScaler

Chart, scatter chart

Description automatically generated

Figure 2: Clustering with features scaled using MinMaxScaler

We are considering a region within a state that is a known agricultural area. It is very likely that the features in the area that contribute to the land being fertile and suitable for agriculture (among others) is resulting in one major cluster, with the other not as significant.  
2. Major difference in number of components that explain 70% of the variation when difference is observed scaling of the feature is performed to be within the [MinMaxScaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html) ( range of [0, 1] ) as compared to features scaled with [StandardScaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html) (remove the mean and have unit variance). 3.The variations in the evaluation scores and PCA number of components explaining 70% of the variation is pronouncedly different and so the two scaling methods were evaluated in different notebooks.   
4. With Standard Scaling, the Biplot is difficult to read, but the components make clear sense.  
5. The Silhouette score in clustering is better with StandardScaler.  
6. The evaluation metrics (R2, MSE, RMSE) for testing improve with StandardScaler.  
7.With MinMaxScaling, the Silhouette Score does not show as much negative area for the first three clusters.

## **EVALUATING VARIOUS ALGORITHMS AND CHALLENGES FACED IN PREDICTIONS**

1. Normalizing the target.  
 SVM was the first linear algorithm to provide non-negative test evaluation scores, but to improve the training scores, in SVM, we employed SKLearn’s TransformedTargetRegressor that can be used to wrap any SKLearn ML algorithm class and has the following advantages:

a. You do not have to call inverse\_transform every time after prediction.  
b. You can specify the transformation and inverse transformation function in the initialization  
c. The wrapped regressor can be used in cross validation and parameters can be set using the wrapping classes’s parameters deduced by calling model.get\_params().keys().   
d. The wrapper does not have feature importance property though a wrapper class was created (for TransformedTargetRegressor) that extracts the feature\_importance\_ from the parent.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **R2 Score** | **MSE** | **RMSE** |
| Unnormalized SVM | 0.462 | 8879.232 | 94.230 |
| Normalized | 0.526 | 7817.253 | 88.415 |

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure: Normalized Target

Figure: Target Normalized

2. High training score and very low test score was obtaining after tuning XGBoost an industrial strength algorithm that returned extremely high training scores .  
3. Increasing number of folds in Pycaret upto exactly 51, improved R-squared, MSE and RMSE.

Other strategies we tried out with no effect or led to reduction in scores:  
1. Added cluster label as a feature after applying unsupervised learning to create clusters with highest Silhouette Score.  
2. Reduced features using PCA.

SUPPORT VECTOR MARGIN REGRESSION

Hyperplane

Support vectors

Margin

XGBOOST

It is E**X**treme **G**radient **Boost**ing that is best suited for large complicated datasets.

Understanding Boosting

A family of models with production level implications. When we learn a final predition function F(x), it will be a sum of  **underpowered sequential weak learners**  (these weak learners could be any linear  
regression/svm/tree).

STEP 0

    1. Define your loss function. True label and y-hat (prediction). This function should be differentiable.

STEP 1

    1. Start with extremely weak learner (F1(x)). Initially, say it is the mean of the y targets.  
    2. This is the initial prediction.  
    3. The residuals (sum of difference of each data point from the mean, predicted by the current model  
 line, define how  good or bad this initial prediction is.

STEP 2    This is essentially gradient descent.

    At every iteration we add the previous weak learner to a eta\*(new learner).  
 That is, we move in the direction of the decreasing loss by a factor of a learning rate (eta).  
 Every week learner descends down the gradient by reducing the loss.

* Start with all residuals in a leaf node
* Calculate the Similarity Score for this leaf which is the Square of the Sum of the Residuals
* divided by (number of Residuals + lambda). Lambda is the regularization parameter that minimizes dependence on any one feature.
* Note the residuals are not sqaured before summing.Some of the residuals might cancel each other out.
* Now the question is : Can we do a better job if we cluster resdiuals together in some manner.(create a line separating the clusters)
* We split the observations into two parts based on say an average of a feature.
* Calculate the similarity score of each leaf node with residuals clustered into the two nodes.
* The Gain is the sum of the similarity scores of these two leaf nodes minus the similarity score of the "root" node.
* We keep the clusters that result in the largest Gain.
* Pruning trees. Set a threshold: Tree Complexity Parameter (gamma) to compare Gains against. If the Gain is greater than gamma, then the tree is not pruned. We start with the lowest tree.
* Lambda, the regularization parameter when set to more than 0, also can cause tree pruning, since it reduces the Similarity Scores and hence the Gain (which is compared to the threshold Gamma)
* Lambda has a greater effect on tree pruning than Gamma.
* After the trees have been created, we calculate the Outcome.Outcome is the Sum of Residuals/(number of residuals + Lambda)
* Once we have the Outcomes we can make new predictions. Learning Rate = ETA

= original prediction (typically mean) + Learning Rate x Output

[EXTREME TREE ALGORITHM](https://www.youtube.com/watch?v=ULFDltLnR6U)

**Basic Principle of Random Forests that Extreme Tree algorithm is similar to:** Bootstrapping + Aggregating (Bagging) to create random trees characterized by repeated sampling with replacement from the dataset and randomly using a subset of columns in each tree and finally aggregating the results of all trees. Evaluation is carried out on the out-of-bag samples (those that did not go through a tree) and evaluating prediction errors.

Difference in Extra Tree from random Forest:

1. Splits in Random Forest are evaluated based on purity pf samples and GINI Index, where as in ET, splits are entirely random. Leads to better speed.  
2. Whole samples considered for each tree (no bootstrapping).  
3. Faster runtime encourages use with datasets with large number of attributes and noisy features.  
4. Randomly splitting by attributes can lead to smaller sample size available for a tree and hence bias.  
5. Randomness also generally leads to less interpretability.

**How much does each of the features contribute to the variance and to the prediction?**

CORRELATION BETWEEN FEATURES

Graphical user interface, chart, scatter chart

Description automatically generated

We are considering features with absolute value of 0.7 as being highly correlated.  
Well construction specific correlations that can be considered as expected since each value describes the well's dimensions and capacity.

To the right of the heatmap, we see TOPOFPERFORATEDINTERVAL and BOTTOMOFPERFORATEDINTERVAL highly correlated.

BOTTOMOFPERFORATEDINTERVAL and TOPOFPERFORATEDINTERVAL are heavily correlated with TOTALCOMPLETEDDEPTH, while TOTALCOMPLETEDDEPTH is correlated with STATICWATERLEVEL

**Well Count**

All well counts, further categorized as Domestic, Agriculture and Public form highly correlated categories (WELL\_COUNT --> (WELL\_COUNT\_DOMESTIC, WELL\_COUNT\_AGRICULTURE))

**Crops**

We observe some few other correlations between some type of crops (e.g. 0.57 between CROP\_D16 plums, prunes or apricots and CROP\_D5 which is peaches/nectarines and between some crops and soils (e.g. 0.61 between CROP\_T31 potatoes and sweet potatoes and SOIL\_ENTISOLS\_A soil of hydrographic group A corresponding to coarse texture soils with high saturated hydraulic conductivity. This correlation also makes sense since potatoes tend to not grow well in soils retaining too much water.

**Vegetation**

We also see some level of correlation between the different type of forest vegetation. For example VEGETATION\_CANYON\_LIVE\_OAK, VEGETATION\_HARD and VEGETATION\_KNOBCONE\_PINE show correlation values between 0.46 and 0.6. This might be explained by the fact that some type of vegetation tends to found frequently together.

VARIANCE USING PCA BIPLOT

After observing the feature correlation in the heatmap, we decided to check for latent features in the dataset that will effectively combine the correlated features, and create a new set of features that are a weighted linear combination of original features, using PCA. The number of components selected explain **70%** (a threshold we selected based on the Scree plot), of the original variance of the dataset.

The first component contains all features describing a well in an area.  
The second component includes a linear combination of well counts and population in an area.  
The third component is largely about precipitation and reservoir capacity, along with ground surface elevation.

Top 4 most important features in each component

===============================================

Component 0: ['TOTALCOMPLETEDDEPTH\_AVG', 'BOTTOMOFPERFORATEDINTERVAL\_AVG', 'TOPOFPERFORATEDINTERVAL\_AVG', 'STATICWATERLEVEL\_AVG']

Component 1: ['WELL\_COUNT\_DOMESTIC', 'WELL\_COUNT\_PUBLIC', 'WELL\_COUNT\_AGRICULTURE', 'POPULATION\_DENSITY']

Component 2: ['PCT\_OF\_CAPACITY', 'AVERAGE\_YEARLY\_PRECIPITATION', 'GROUNDSURFACEELEVATION\_AVG', 'WELL\_COUNT\_AGRICULTURE']

Component 3: ['GROUNDSURFACEELEVATION\_AVG', 'WELL\_COUNT\_INDUSTRIAL', 'AVERAGE\_YEARLY\_PRECIPITATION', 'POPULATION\_DENSITY']

Component 4: ['WELL\_COUNT\_INDUSTRIAL', 'AVERAGE\_YEARLY\_PRECIPITATION', 'POPULATION\_DENSITY', 'WELL\_COUNT\_AGRICULTURE']

Component 5: ['POPULATION\_DENSITY', 'GROUNDSURFACEELEVATION\_AVG', 'TOTALDRILLDEPTH\_AVG', 'WELL\_COUNT\_AGRICULTURE']

Component 6: ['TOTALDRILLDEPTH\_AVG', 'WELLYIELD\_AVG', 'STATICWATERLEVEL\_AVG', 'GROUNDSURFACEELEVATION\_AVG']

Projections IN THE NEW PCA SUBSPACE

Graphical user interface, chart, scatter chart

Description automatically generated

Figure 4 PCA Biplot

The PCA biplot emphasizes the collinearity we saw in the heatmap (see in notebook) in features related to a well such as completed depth, top and bottom of perforations (for filter).

Crop D12 (almonds) and Alfisols\_D soil are correlated (small angle between them), where as Crop\_D12 “pistachios” are quite different as a feature with a greater than 90 degrees angle between pistachios and almonds. As per [USDA soil taxonomy](https://dbpedia.org/page/Alfisol), because of their productivity and abundance, the Alfisols represent one of the more important soil orders for food and fiber production. They are widely used both in agriculture and forestry,

FEATURE IMPORTANCE IN PREDICTION

**SHAPELY VALUES (Shapely Additive Explanations) : Model Agnostic**

Graphical user interface, application

Description automatically generated

Figure 5: Shapley Values

Shapely values are a commonly used machine learning technique to explain a model. It is **model agnostic.** In Pycaret, we use the **interpret\_model(best)** thatprovides the above plot.  
[Shapely works by changing one feature at a time](https://www.youtube.com/watch?v=NBg7YirBTN8&t=301s), taking into account the probability of creating improbable samples and isolated effects. The target is predicted and compared as a difference to the mean of all original predictions. The expected value of these differences is the Shapely value of this feature.

Ground Surface Elevation of the area is the more important indicator followed by depth of wells and current depth of water in the wells in the area and some soils such as the “alfisols” which is the fertile soil that supports San Joaquin valley’s reason for being an agricultural belt, Entisols and Mollisols, which are soils of grassland ecosystems. Among crops, Pistachios (“D14”) and alfalfa and alfalfa mixtures ("P1") carrying areas, influence predictions.

[The SHAP plot](https://www.youtube.com/watch?v=ZkIxZ5xlMuI) also shows the positive and negative relationships of the predictors with the target variable with 0.0 as the midline in the graph. The less (more blue) the groundsurface elevation and afisols, the more the depth. The more (more red) the depth of the wells, the more the ground water depth (kind of intuitive). More pistachios, more depth, but more alfalfa, predicts, less depth.

**FEATURE IMPORTANCE FROM THE BEST MODEL : Model Specific**

Chart, scatter chart

Description automatically generated

Figure 6: Feature Importance for ET regressor

Feature importance in the random tree boosting algorithm Extra Trees also afforda Groundsurface Elevation the most importance along with well specific feature of the depth of the well and current level of water in the wells in the area. In addition to this, Alfisol\_D (the fertile soil) and Pistachios also feature prominently in the prediction. Reservoir capacity is considered an important feature by this algorithm.

ML FLOW IN DEEPNOTE