Predicting Snow Water Equivalent in regions in Western United States

Estimating snow water equivalent (SWE) at a high spatiotemporal resolution over the Western U.S.

Author: Hanis Zulmuthi

May 2022

Source: Reddit.com

Overview

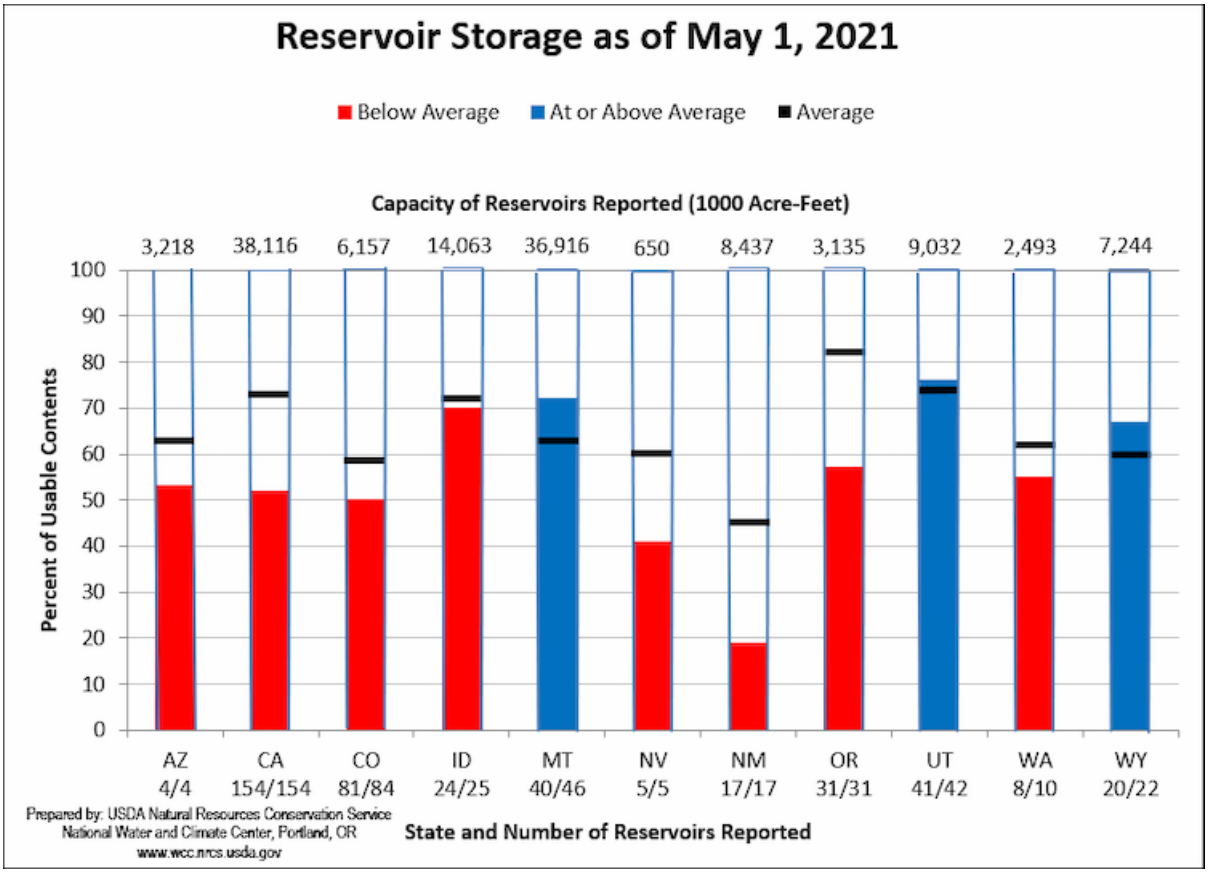
This project budded from a competition titled Snowcast Showdown on Driven Data. The goal of the project is to develop a predictive model to estimate the distribution of Snow Water Equivalent (SWE) at a high spatiotemporal resolution over the Western U.S. This predictive model will assist NOAA in their National Integrated Drought Information System (NIDIS), an initiative to monitor snow drought in the wester United States.

Introduction

Snow Water Equivalent (SWE) is a common snowpack measurement used by hydrologists and water managers to gage amount of liquid water contained within snowpack. It is equal to the amount of water contained within the snowpack when it melts. It can be thought of as the depth of water that would theoretically result if you melted the entire snowpack instantaneously [1].

Water in a snow pack is determined by depth, density, type of snow, changes in the pack, previous freeze/thaw cycles, recent rainfall events, etc. Available water is the amount of water that would be released if the snow pack melted. SWE is an important measure of availability of water resources, since it relates to the runoff of rivers and variations in groundwater levels, so knowing how much water is available in the snow pack is valuable for those managing reservoirs and flood forecasting [2][3].

Reports by NOAA (through their National Integrated Drought Information System (NIDIS) program) on the intensifying snow drought over western U.S raises the alarm on the importance of predicting SWE as accurately possible, especially for remote, high elevation areas where manual ground measure measurements are not feasible. It was reported that the loww snowpack, rapid and early snow melts and poor runoffs had resulted in a significant drop in water supply in the summer of 2021 (fig 1).



Source: NIDIS, Drought.gov

Data Understanding

Historical Ground Measures data: Ground measures help provide regularly collected, highly accurate point estimates of SWE at designated stations. Ground measures data range from 2013-2019 and 2020-2021 was provided in ground_measures_train_features.csv and ground_measures_test_features.csv. The ground measures data are from Snow Telemetry (SNOTEL) and California Data Exchange Center (CDEC). The dataset used from these sources is available in this repo here.

SNOTEL: The Snow Telemetry (SNOTEL) program consists of automated and semi-automated data collection sites across the Western U.S.

CDEC: The California Data Exchange Center (CDEC) facilitates the collection, storage, and exchange of hydrologic and climate information to support real-time flood management and water supply needs in California. CDEC operates data collection sites similar California.

Ground-based sites from SNOTEL and CDEC are used both as an input data source and in ground truth labels for our predictive model. ***Note that, sites that we are predicting SWE for, are entirely distinct from those in the features data.***

MODIS Satellite Imagery: The MODIS satellite images consist of MODIS/Terra and MODIS/Aqua Snow Cover Daily L3 Global 500m SIN Grid. Terra's orbit around the Earth is timed so that it passes from north to south across the equator in the morning, while Aqua passes south to north over the equator in the afternoon. Snow-covered land typically has very high reflectance in visible bands and very low reflectance in shortwave infrared bands. The Normalized Difference Snow Index (NDSI) reveals the magnitude of this difference. The snow cover algorithm calculates NDSI for all land and inland water pixels in daylight using MODIS band 4 (visible green) and band 6 (shortwave near-infrared).

The satellite imageries from MODIS were not used for modelling due to contraints in computing power and memory. We did however, pull down the satellite images from their Azure blob and saved it as numpy arrays of pixels. This process was done in this notebook that was executed in Google Colab.

Notebooks in this repo

MODIS-DEM-Preprocessing_colab(01.1).ipynb - This notebook details the process of pulling down MODIS satellite imageries from Azure blob and save them as numpy arrays of pixels.

MODIS-Preprocessing(01.2).ipynb - This notebook details the same process as the above notebook but for local machines and using conda environment. The environment to run this notebook is provided in the repo here. **However,** this notebook wasn't executed to completion due to restraints on computational memory on my local machine.

Data-Preprocessing-EDA(02).ipynb - This notebook is where the data processing of ground measures data into model features was done. It then saves the resulting dataframe for modeling. The environment to run this notebook is provided in the repo here.

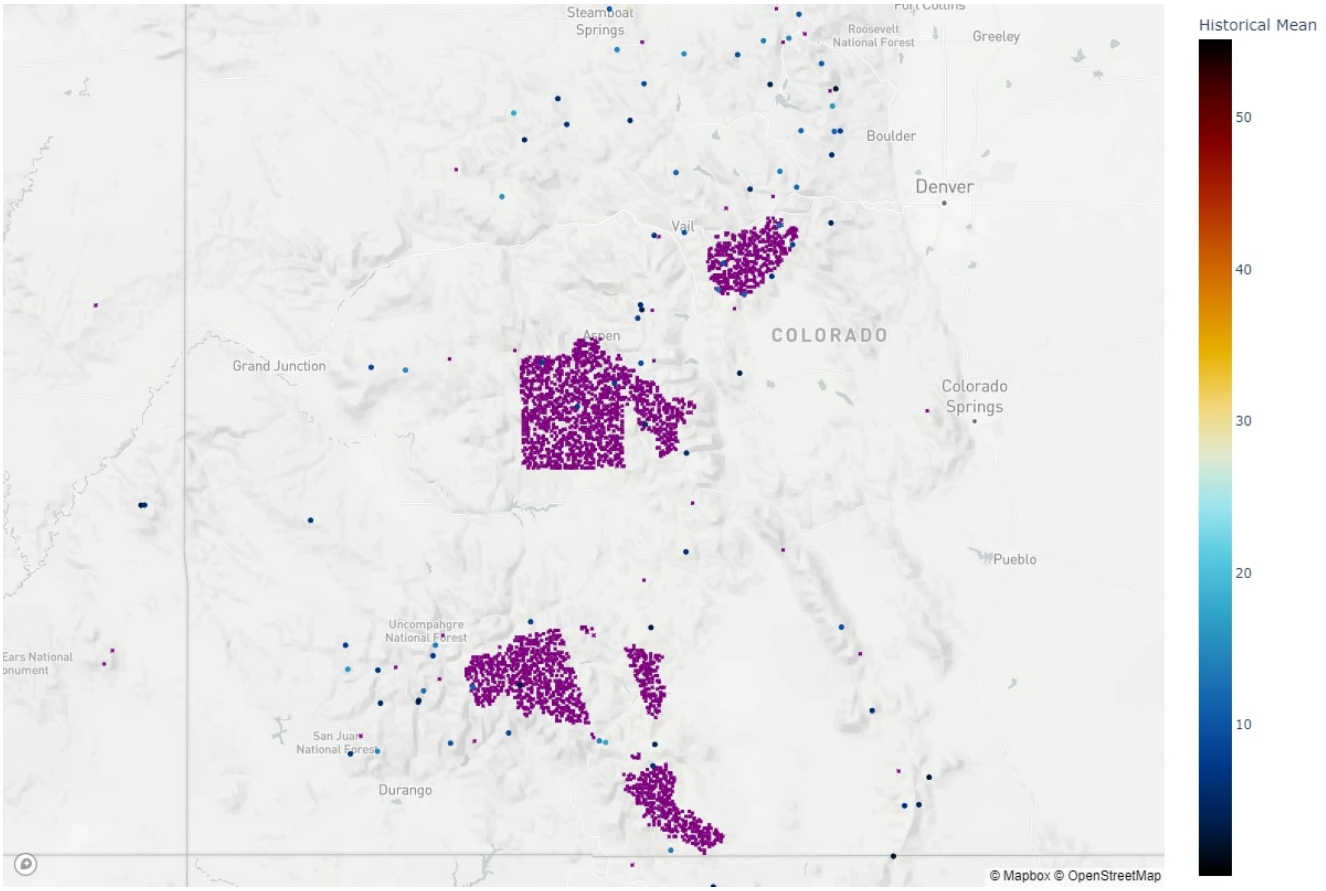
Time-Series-EDA(03).ipynb - This notebook contains time series data exploration of the ground measure SNOTEC and CDEC stations.

Modeling(04).ipynb - This notebook contains dummy model, linear regression and 3 different types of Gradient Boosting models trained on the data saved at the end of Data-preprocessing notebook. This notebook also contains model evaluations and comparisons. The environment to run this notebook is provided in the repo here. It is the same environment used for Data Processing notebook.

Modeling & Results

Model Target (predicted)

Our model is predicting snow water equivalent (SWE) measures for 2013-2021 for 1km x 1km grid cells. For model training, we SWE values for 2013-2019 were provided in train_labels.csv. For model evalutation, data for 2020-2021 is in labels_2020_2021.csv.

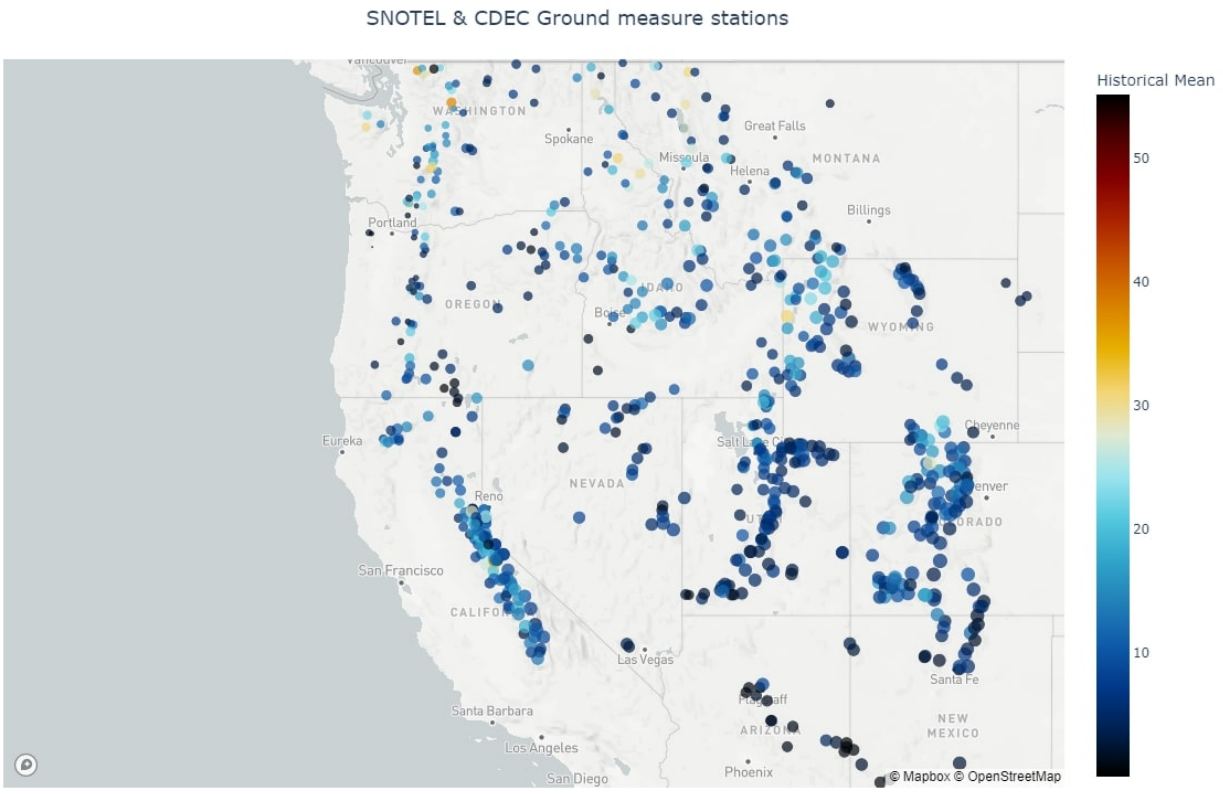


The purple grids are a sample of 1km x 1km grid cells in mountains of Colorado, US. The scatter markers are SNOTEC and CDEC stations colored by the historical mean of their SWE.

Model Features (predictors)

1. Ground measures Data (SNOTEL and CDEC)

SWE values of the top 20 nearest (by distance) ground measure stations to the grid cell were used as model features.



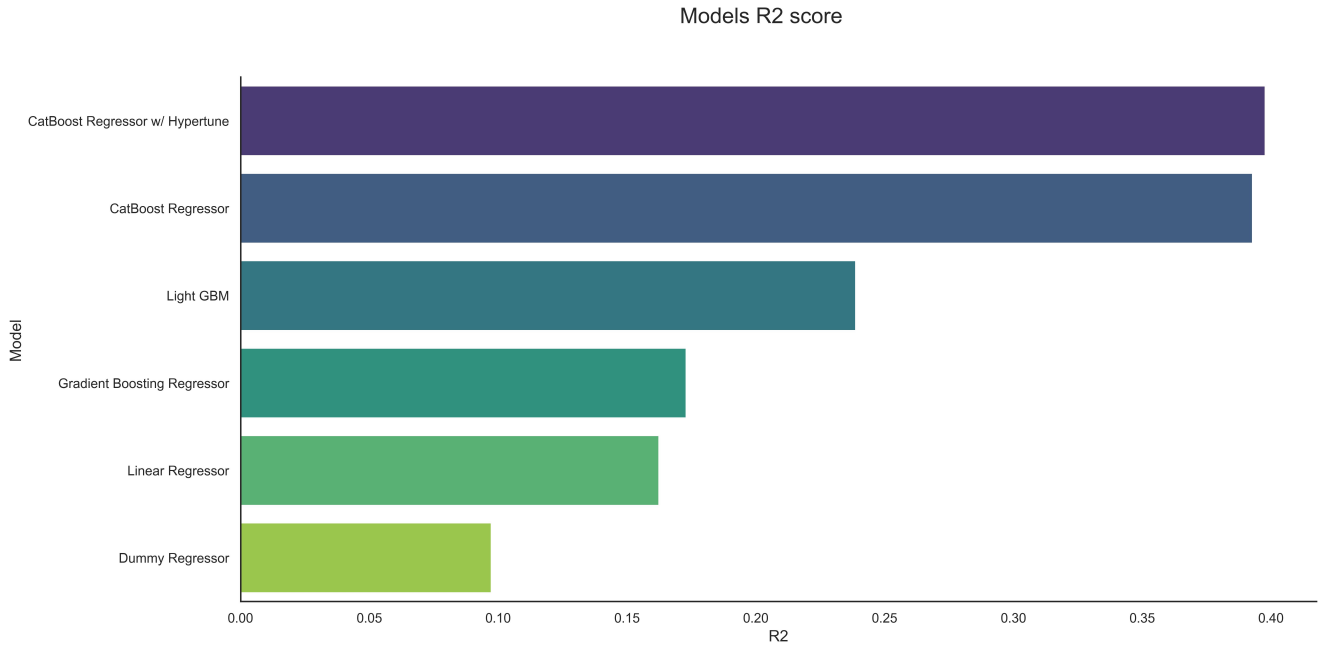
Scatter plot of SNOTEC and CDEC stations colored by the historical mean of their SWE. Size of markers represent elevation of those stations.

2. Geo-spatial features

- Latitude
- Longitude
- Region type

Model Comparison

We started off with our first simple model that returned an R2 of 0.09730. This will be our baseline, and yes, it's a pretty low baseline. Our subsequent models manage to beat the R2 of our first simple models as follows:



All models except CatBoost Regressor with the parameters Hypertune were ran using default parameters. The R2 is the mean of cross-validations R2. As seen in the figure above, our best model is a Catboost Regressor. The hypertuned CatBoost only improved by a mere 0.05. With all the models being around the same R2 range, it tells that the way to improve our R2 is to include more features that are more descriptive of the geolocation of the grid cell regions themselves and features that have influence on the yearly snow packs. And these features could be temperature, precipitation, elevation and satellite images.

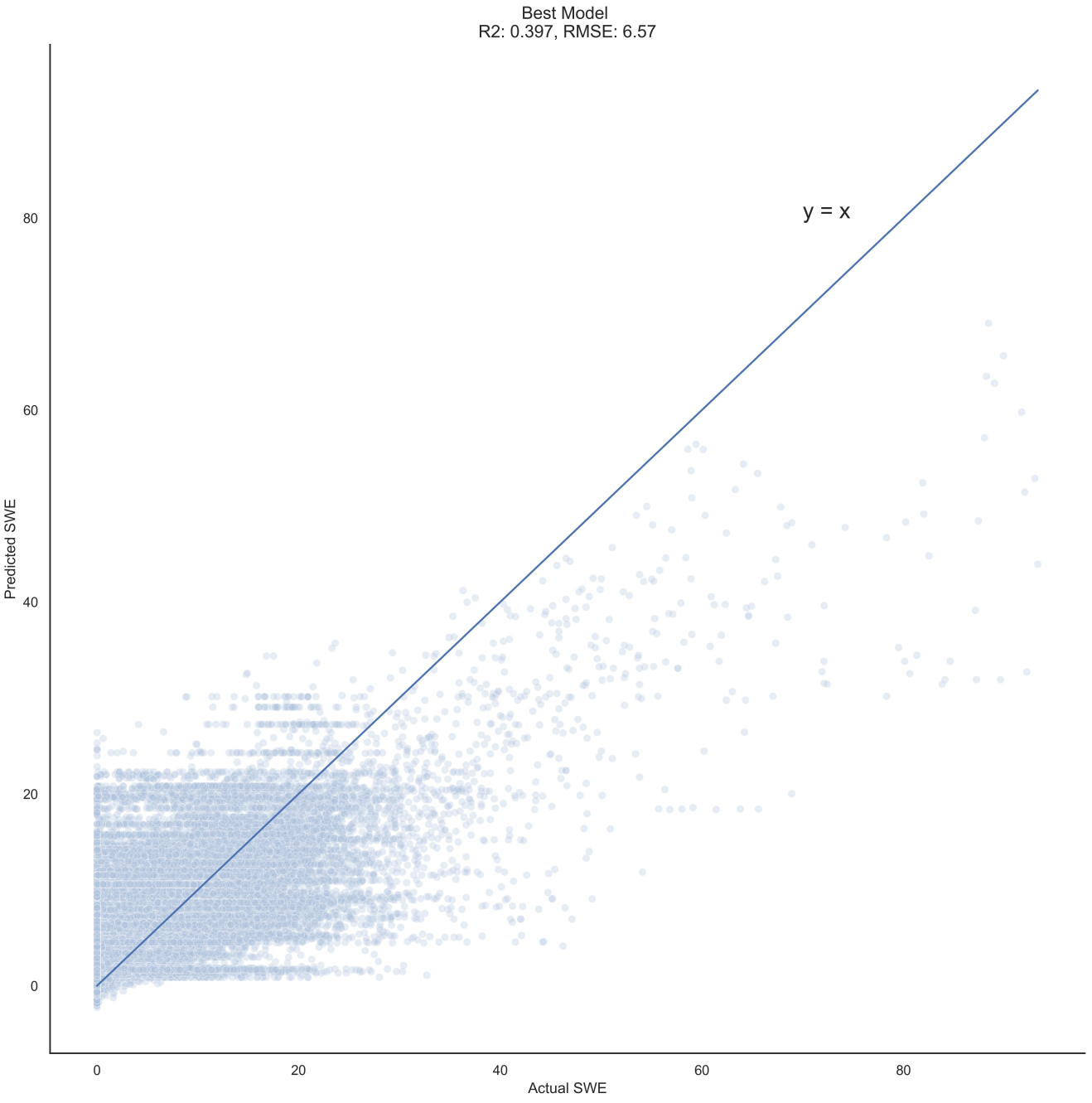
Best Model

CatBoost Regressor with hypertuned parameters

To reiterate, our best model is our CatBoost Regressor with some hypertuning. It gave a R2 value of 0.3977, the highest out of 5 and an RMSE of 6.57. For reference, an ideal RMSE would be in the range of 1-3. The best parameter setting that resulted in this R2 value is:

```
{'catboost__depth': 10,
 'catboost__learning_rate': 0.1,
 'catboost__min_data_in_leaf': 10,
 'catboost__n_estimators': 100}
```

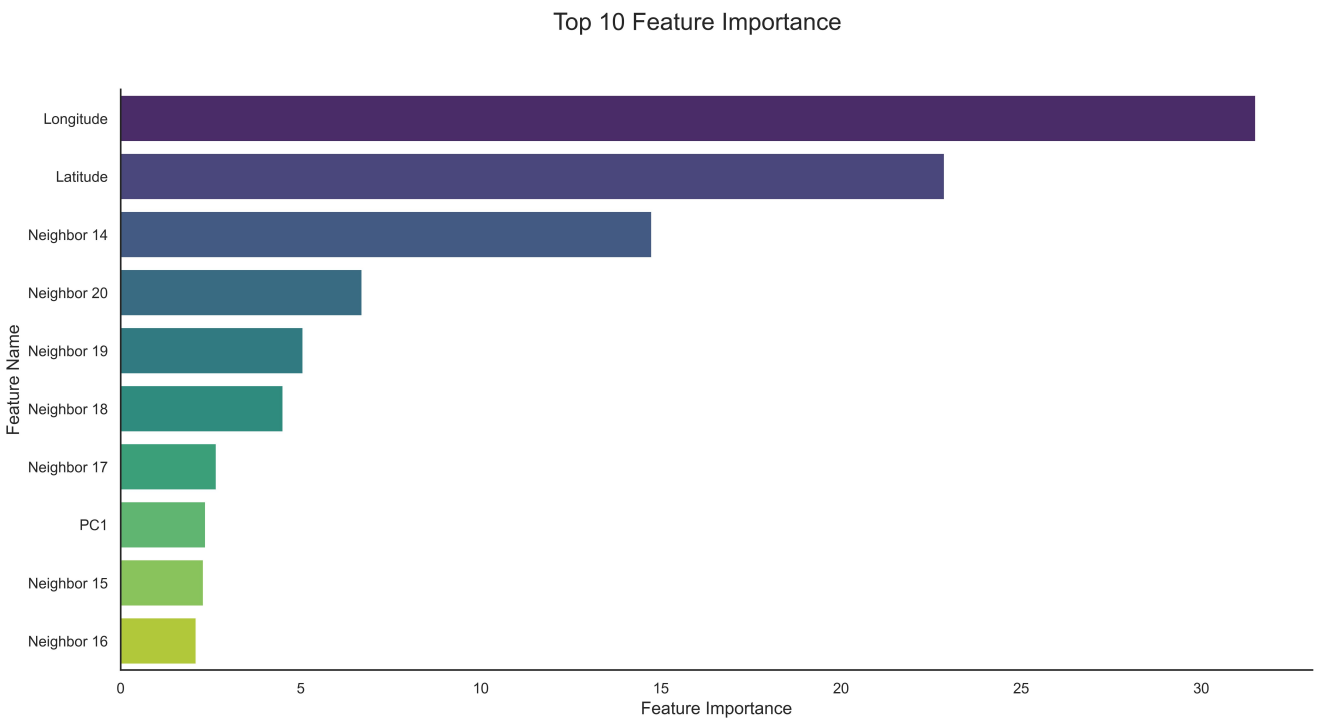
The figure below is a visualization of our best model's performance.



We can see that our model with the input features wasn't able to capture the variation of SWE.

Feature Importance

Tehe figure below shows which of our Model features are the most important.



It is no surprise that the most important features is latitude and longitude since these two features are the only features that are **direct** geospatial features of our grid cells. This further emphasizes how beneficial it would be to incorporate satellite imagery, elevation and climate data as part of our model features to improve our R2 and RMSE since the thickness of snow packs at any given time depend on these climatological and geospatial features at those grid cells.

Conclusion

Future Work

1. Explore Time-Series to forecast SWE at SNOTEL and CDEC stations

On top of predicting SWE at the location of grid cells, it would be highly valuable to conduct a time series analysis of SWE at SNOTEL and CDEC (ground measure) stations and furthermore, to be able to forecast SWE at these locations.

2. Exploration of feature engineering

As observed from our model performance, our model could do better with more features that are engineered. Some possibilities include:

- Using historical mean of SWE and SWE relative to historical mean. This could possibly capture long term trends in the region.
- Calculate the snow day at which the observation was measured. This could help capture seasonal trends in our data.

3. Incorporate data from satellite imageries & remote sensing data

- Satellite imageries (MODIS Terra/Aqua Data)
- Climate data
- DEM The model could've beneffited just from the mean and variance of pixel values over an entire grid cell for the satellite imageris (MODIS and DEM).

4. Use near Real-time data to predict SWE Make a dashboard of predictions and forecast with streams of near real-time data

Repository Structure

```
├── data
├── figures
├── models
├── src
│       ├── catboost_info
│       ├── Data-Preprocessing.ipynb
│       ├── EDA.ipynb
│       ├── MODIS-DEM_Preprocessing_colab.ipynb
│       ├── MODIS-Preprocessing.ipynb
│       └── Modeling.ipynb
│
├── .gitignore
├── geo_env.yml
├── LICENSE
├── README.md
└── modis.yml
```

References

[1] What is SWE? (Natural Resources Conservation Service Nevada, USDA)

[2] Snow Water Equivalent (SWE) Measurement Systems (Campbell Scientific)

[3] Climatology of snow cover and snow water equivalent (Eumetrain.org)

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%