Table of Contents

Problem Statement

Predict Day-ahead Market and Hourly Market pricing for wholesale power in the California Electricity Grid Operator's footprint (CAISO) using system operator demand projections, weather data, and reservoir water levels.

Features

There are 16 weather features, one water level feature, four DateTime features, four power demand prediction features, one real-time spot settlement price feature, and two-goal variables: DAM (Day-Ahead Market) and HASP (Horizonary Average Spot Price) (Hour Ahead Scheduling Process). All regressions will be two-model regressions. As a feature, each target variable will take the other target.

Altogether, there are 16 weather features, one water level feature, four datetime features, four electricity demand forecast features, one real-time spot settlement price feature, and two target variables: DAM (Day-Ahead Market) and HASP (Hour Ahead Scheduling Process). All regressions will be pairs of models. Each target variable will take the other target as a feature.

Data Dictionary

Item	Description
Index	Datetime in Pacific time zone
Train set	First 75% of data
Test set	Last 25% of data
dam_price_per_mw	Day-Ahead Market elec. price
hasp_price_per_mw	Hour-Ahead Market elec. price
rtm_price_per_mw	Realtime Market elec. price
7da_load_fcast_mw	7-day load fcast
2da_load_fcast_mw	2-day load fcast
dam_load_fcast_mw	Day-ahead load fcast
rtm_load_fcast_mw	Hour-ahead load fcast
water_acre_feet	CA water level
sand_temp	San Diego weather stn Temp
sand_wind	San Diego weather stn Wind speed
sand_vis	San Diego weather stn Visibility
sand_ceil	San Diego weather stn Celling
rive_temp	Riverside weather stn Temp
rive_wind	Riverside weather stn Wind speed
rive_vis	Riverside weather stn Visibility
rive_ceil	Riverside weather stn Celling
redd_temp	Redding weather stn Temp
redd_wind	Redding weather stn Wind speed
redd_vis	Redding weather stn Visibility
redd_ceil	Redding weather stn Celling
fres_temp	Fresno weather stn Temp
fres_wind	Fresno weather stn Wind speed
fres_vis	Fresno weather stn Visibility
fres_ceil	Fresno weather stn Celling

Data Collection

1. Electricity wholesale prices. I created an automated loop to get information from CAISO on both pricing and load (demand) forecasts... automated generation of the right string format url + query syntax, as well as a 5-second delay between monthly data requests extending back 40 months to January 1, 2016. Depending on the datum in issue, data is collected at hourly, five-minute, and fifteen-minute intervals.

2. The California Department of Water Resources (CA DWR) provides a statewide database of the reservoir, inflow/outflow, water levels, snowfall water content, and other variables. I added an hourly aggregate of the 47 reservoir's water content, measured in acre-feet, as one feature. The API queries to CADWR have no date range restrictions.

The California Department of Water Resources (CA DWR) maintains a statewide database that includes information on reservoirs, inflow/outflow, water levels, snowfall water content, and other factors. As one feature, I included an hourly aggregate of the 47 reservoir's water content, measured in acre-feet. There are no date range constraints for CADWR API queries.

3. Hourly weather station data from the National Oceanic and Atmospheric Administration (NOAA). Downloads of very big zip tar files are available by year. These files unzip into 50 GB folders containing a slew of.csv files, each titled after an 11-digit numerical number that identifies a specific meteorological station or buoy. Four specific weather stations in California were chosen for inclusion of four weather-related statistics: San Diego, Riverside, Redding, and Fresno. The four locations were chosen for their data quality and geographic dispersion throughout California.

   Surface temperature, wind speed, cloud ceiling height, and horizontal visibility are the four statistics retrieved. All of these variables have an impact on either the demand for or supply of power. Temperature encourages air cooling and heating with electricity, wind speed suggests wind farm output, and cloud ceiling + visibility may indicate the amount of potential solar generation available.

Data Inspection and EDA

Missing value. CAISO's pricing and load prediction data set is fairly thorough, with no missing numbers, however, I had to remove one day every 31-day month to avoid the API from rejecting my calls during daylight savings time, when the additional hour (relative to UTC) causes the 31-day limit to trip by one hour. These were left as blanks in the datetime index. Data was down-sampled to hourly averages from 5- and 15-minute intervals. Because the majority of the data was already in hourly granularity, it made the most sense to retain it that way.

Missing entries (encoded with '9999'), duplicated datetime index values, and many duplicates abound in the NOAA dataset. After slicing off the necessary information from lengthy tuples of comma-separated string values, I left linked the four locations' dataframes together. I used a forward fill approach to fill Nan values on a case-by-case basis, based on the reasoning that a missing observation will most likely be the same as the prior observed value. All columns were graphed to look for anomalies and outliers/errors.

Correlations. To see how predictive the traits are, we used a color heatmap of Pearson correlations. I didn't notice any predictive power, therefore this will most likely be a challenging challenge to solve.

Two Regression Modeling Estimators

1. ARIMA Autoregressive integrated moving average
2. SARIMAX ARIMA with Seasonal effect, and eXogenous variables

Conclusion

The hour ahead market proved to be more difficult to anticipate than the day-ahead market for these strategies. ARIMA matched the recurrent neural networks' performance, which was unexpected.

SARIMAX and ARIMA

a) Performance. Given that no exogenous factors are included in the forecast, ARIMA performed fairly well (DA market). This makes sense because electricity demand is highly structured and expenses do not fluctuate much. SARIMAX performed worse, which should not have been the case given the addition of seasonality and exogenous factors. In the HA market, both models performed poorly.

b) Weaknesses. The Stats model package is difficult to set up for train/test splits to generate predictions on a test set fitted on a train set since it is not obvious. Extremely inefficient computational efficiency, particularly with SARIMAX.

Future works

1. Use other methods such as RNN and LSTM
2. Extend the data set back a few years to collect additional train-test sets, then go back to the CAISO API to fill in the gaps (I skipped a few days to stay under the CAISO 31-day hard limit per query).
3. Run SARIMAX grid search after extending ARIMA grid search.
4. Use the Prophet time-series ml tool on Facebook.
5. Convert the dataset to "tabular" format and remove the time index so that alternative estimators, such as random forest, may be used.