GENERAL ASSEMBLY

DSI7-SF

Prepared by:
Manu Kalia

CAPSTONE PROJECT: CALIF. ELECTRICITY PRICE PREDICTION

Time-series Regressions Using Recurrent Neural Networks, ARIMA, & SARIMAX

PROBLEM STATEMENT

Electricity Hourly Price Prediction:

In the Calif. Electricity grid operator's footprint (CAISO), predict Day-ahead Market and Hourly Market prices for wholesale electricity, using system operator demand forecasts, weather observations, and reservoir water levels.

DATA SOURCES

Three different sources of publicly available data:

a) CAISO has an open access information platform called OASIS…
http://oasis.caiso.com/mrioasis/

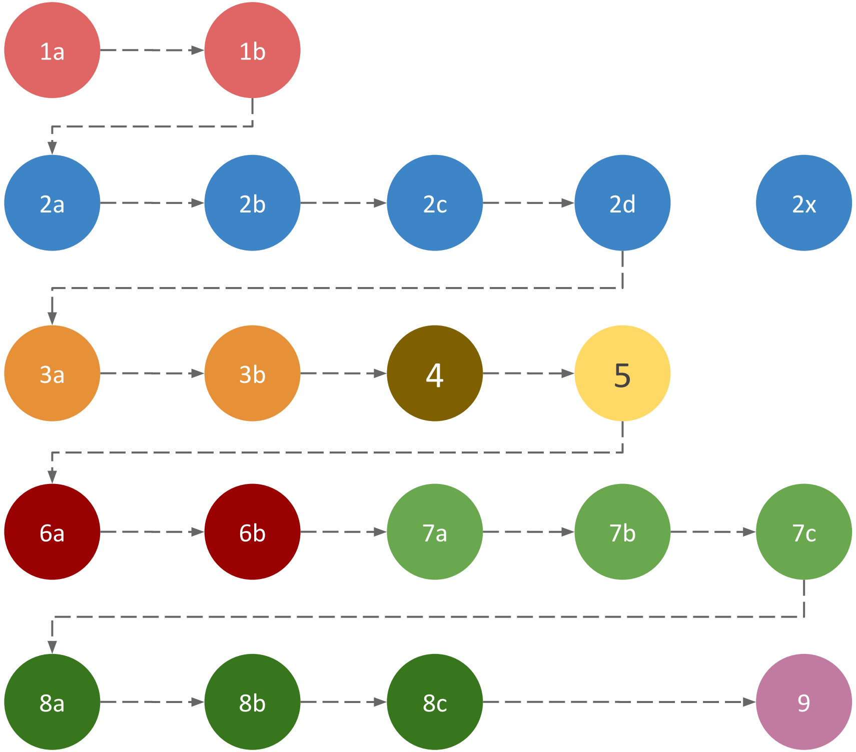
b) NOAA Global Weather Station Data Archive
https://www.ncei.noaa.gov/data/global-hourly/archive/csv/

c) CA Department of Water Resources
http://cdec.water.ca.gov/reportapp/javareports?name=RES

CODE NOTEBOOK FLOW

Set of 19 code notebooks that follow a sequential order, with interim saving of images, data files as .csv, and specific dataframes and fitted models as pickle byte files:

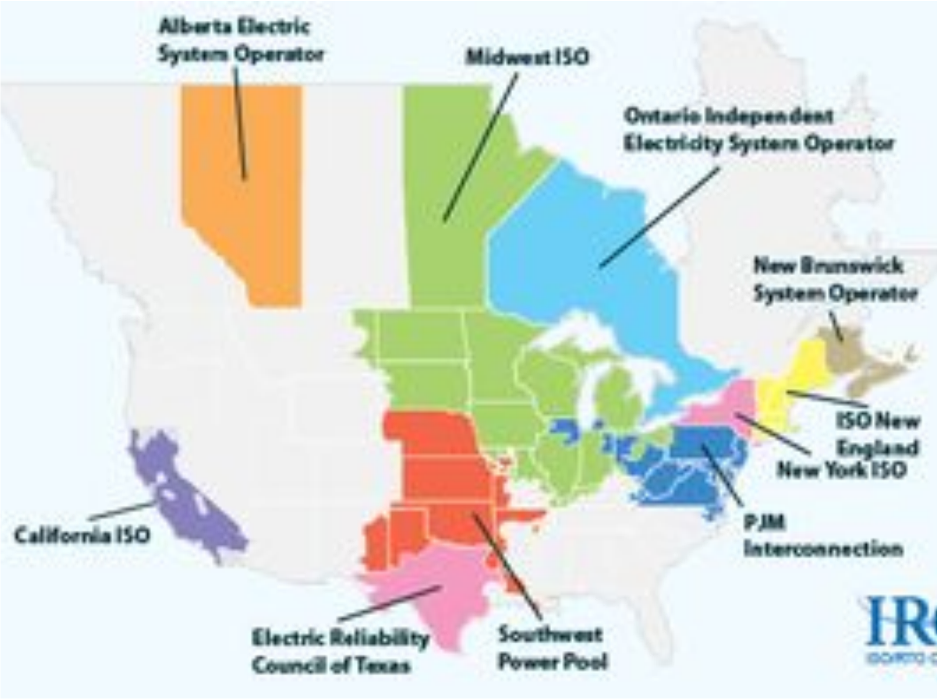
–	1a, 1b	… data acquisition
–	2a - 2d	… dataframe creation
–	3a, 3b	… dataframe merging
–	4	… pre-processing
–	5	… EDA / inspection
–	6a, 6b	… ARIMA, SARIMAX
–	7a, 7b, 7c	… RNN (dam) models
–	8a, 8b, 8c	… RNN (hasp) models
–	9	… predictions / eval



System Operator (ISO) Service Areas … Footprints

Certain parts of the country are organized by ISOs that are typically non-profit entities tasked with keeping a local electrical grid operating 24x7 with NO OUTAGES.

ISOs set up markets between counterparties, manage congestion pricing in the transmission and distribution networks.

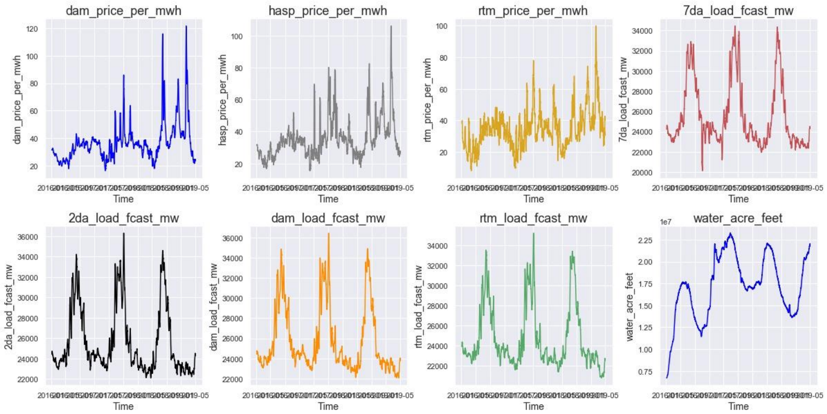

Alberta Electric System Operator
Midwest ISO
Ontario Independent Electricity System Operator
New Brunswick System Operator
ISO New England
New York ISO
PJM Interconnection
California ISO
Electric Reliability Council of Texas
Southwest Power Pool

CA ISO MARKET DATA (query construction for OASIS)

Each ISO typically has an API with different limits on how often you can hit it for data, but has a number of dimensions available for data:

a) Realtime prices (RTM LMP) … prices at particular "nodes" for the next period (5 min intervals, down-sampled to hourly)
b) DA (day ahead) market prices … set once a day for the next 24 hours (hourly)
c) HASP (hour ahead) prices … 15 minute intervals down-sampled to hourly
d) Load forecasts … ISO prediction for energy demand in MW at four different projection horizons: 7 days ahead, 2 days ahead, day-ahead (all 3 are hourly), and real-time (15 min freq down-sampled to hourly).

- Two target columns (DAM & HASP)
- 5 feature columns (realtime prices, and four load forecasts)

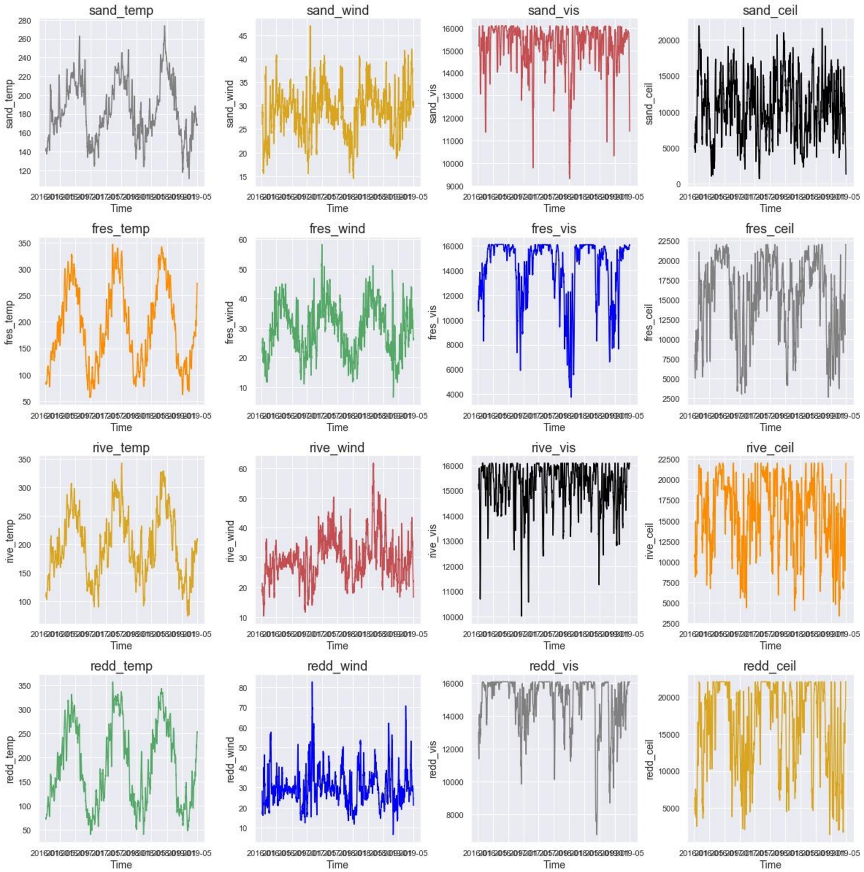


NOAA DATA

Two dimensions of data yield 16 features:

a) four variables: temp, wind speed, cloud ceiling, horizontal visibility

b) Selected four geographically spread out weather stations, so that a good cross-section of California is represented, all of which has an effect on the elec price. San Diego, Riverside, Fresno, and Redding.

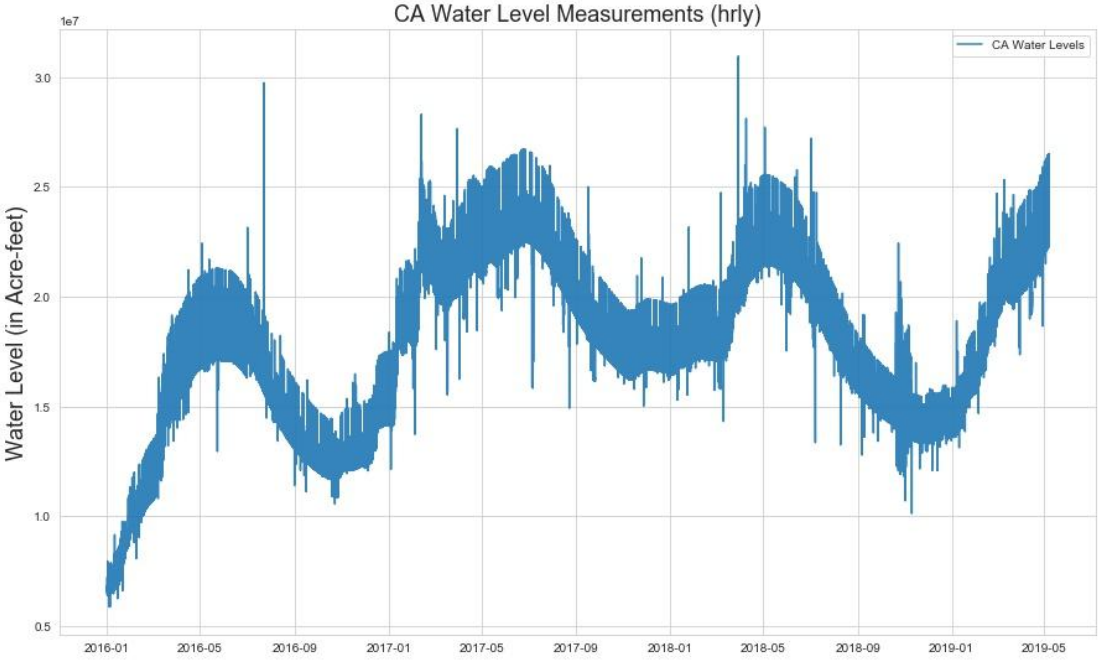


CA Dept. of Water Resources Data

Hourly sensor readings of the 47 major reservoirs in California.

Good indicator of the water capacity available for hydroelectric generation

Summed all these together each hour to get the CA Water Level feature



DATA ACQUISITION

Two APIs and one simple download site:

a) CAISO has an open access information platform called OASIS… http://oasis.caiso.com/mrioasis/ … hard limit of 31 days of data can be requested with any single API query. Daylight savings time plays havoc with this limit (automatic conversion to UTC, which has no DST).

b) CA Dept. of Water Resources has a relatively unrestricted API and easy-to-parse query structure.

c) http://noaa.gov/ allows one to download global weather station data by year. These are large zipfiles that unpack to 50GB folders of coded .csv files. Need to select specific files for weather stations of interest and discard the rest.

DUPLICATES AND GAPS

Methodology:

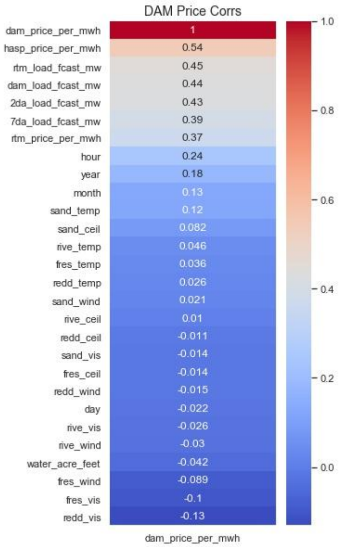
a) Allow gaps in CAISO data for DAM prices. Left join other dataframes to this.

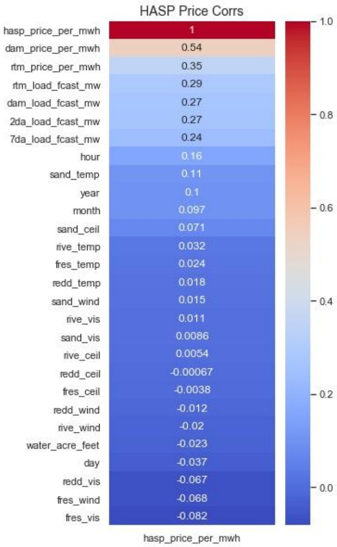
b) Convert various "Nan" numeric codes (e.g. "99999" in NOAA, "ART" and "---" in CADWR) to np.nan, then employ df['column'].fillna(method='ffill') to carry forward the last known observation. This mimics intuitive behavior when seeing gaps in time-series data for these datasets.

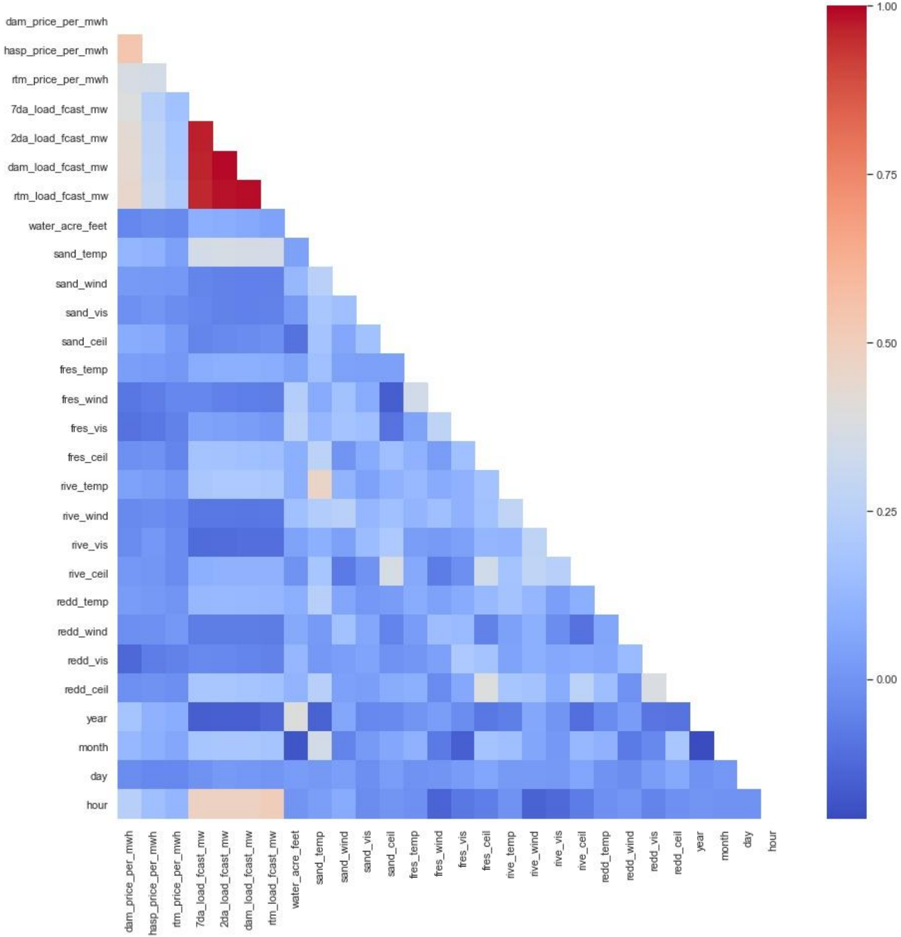
EDA / VISUALIZATIONS

29,000 hours captured, over 3.33 years...

- correlations are weak
- highly cyclical patterns







MODELING APPROACH

Three models will be used and compared for time series prediction on two different target variables: DAM (day-ahead market), and HASP (hour-ahead scheduling process)

a) ARIMA will serve as the baseline. Uses only the target variable's previous values to predict next value (auto-correlation).

b) SARIMAX used to inject seasonality (24-hour periodicity, as well as longer periods are relevant here), as well as the exogenous features, such as demand forecasts and weather.

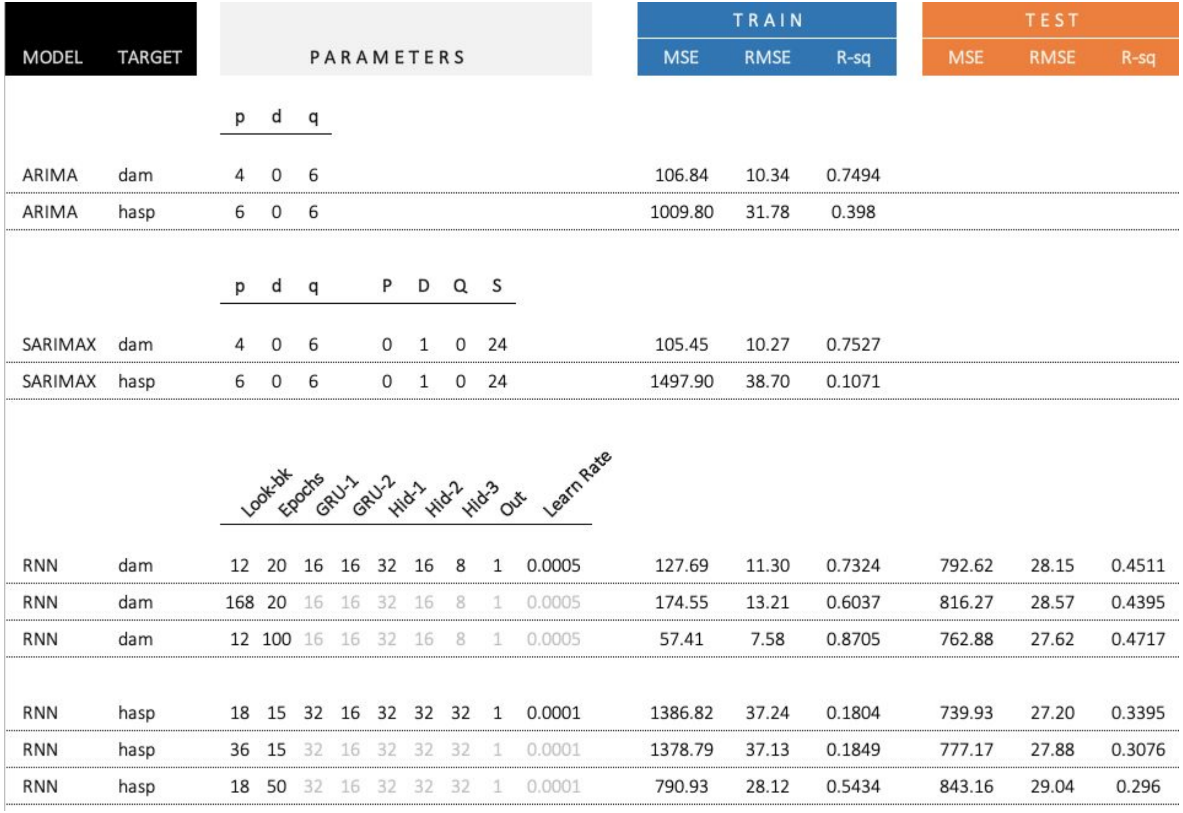
c) Recurrent Neural Networks will be the main area of exploration, since they have shown some promise in performing time-series regression with exogenous variables.

RESULTS SUMMARY

The hour ahead market proved to be more difficult for these methods to predict than the day ahead market.

ARIMA showed surprisingly good performance, matching that of the recurrent neural networks.

The RNNs were much faster to fit than ARIMA, and MUCH faster than SARIMAX, which was the worst performer overall.



											TRAIN			TEST		
MODEL	TARGET	PARAMETERS									MSE	RMSE	R-sq	MSE	RMSE	R-sq
		p	d	q												
ARIMA	dam	4	0	6							106.84	10.34	0.7494			
ARIMA	hasp	6	0	6							1009.80	31.78	0.398			
		p	d	q		P	D	Q	S							
SARIMAX	dam	4	0	6		0	1	0	24		105.45	10.27	0.7527			
SARIMAX	hasp	6	0	6		0	1	0	24		1497.90	38.70	0.1071			
		Look-bk	Epochs	GRU-1	GRU-2	Hid-1	Hid-2	Hid-3	Out	Learn Rate						
RNN	dam	12	20	16	16	32	16	8	1	0.0005	127.69	11.30	0.7324	792.62	28.15	0.4511
RNN	dam	168	20	16	16	32	16	8	1	0.0005	174.55	13.21	0.6037	816.27	28.57	0.4395
RNN	dam	12	100	16	16	32	16	8	1	0.0005	57.41	7.58	0.8705	762.88	27.62	0.4717
RNN	hasp	18	15	32	16	32	32	32	1	0.0001	1386.82	37.24	0.1804	739.93	27.20	0.3395
RNN	hasp	36	15	32	16	32	32	32	1	0.0001	1378.79	37.13	0.1849	777.17	27.88	0.3076
RNN	hasp	18	50	32	16	32	32	32	1	0.0001	790.93	28.12	0.5434	843.16	29.04	0.296

ARIMA & SARIMAX RESULTS

HIGHLIGHTS:

a) Performance. ARIMA showed a surprisingly strong performance (DA market), given that no exogenous features go into the prediction. Electricity usage is highly patterned, and costs do not change drastically, so this makes sense. SARIMAX was a worse performer, and given the addition of seasonality and exogenous features, this should not be the case. Both models (as well as RNNs) fared poorly for the HA market.

b) Deficiencies. The Statsmodel package is not intuitive, and difficult to set up for train/test splits to get predictions on a test set fitted on a train set. Extremely poor computational efficiencies, especially for SARIMAX.

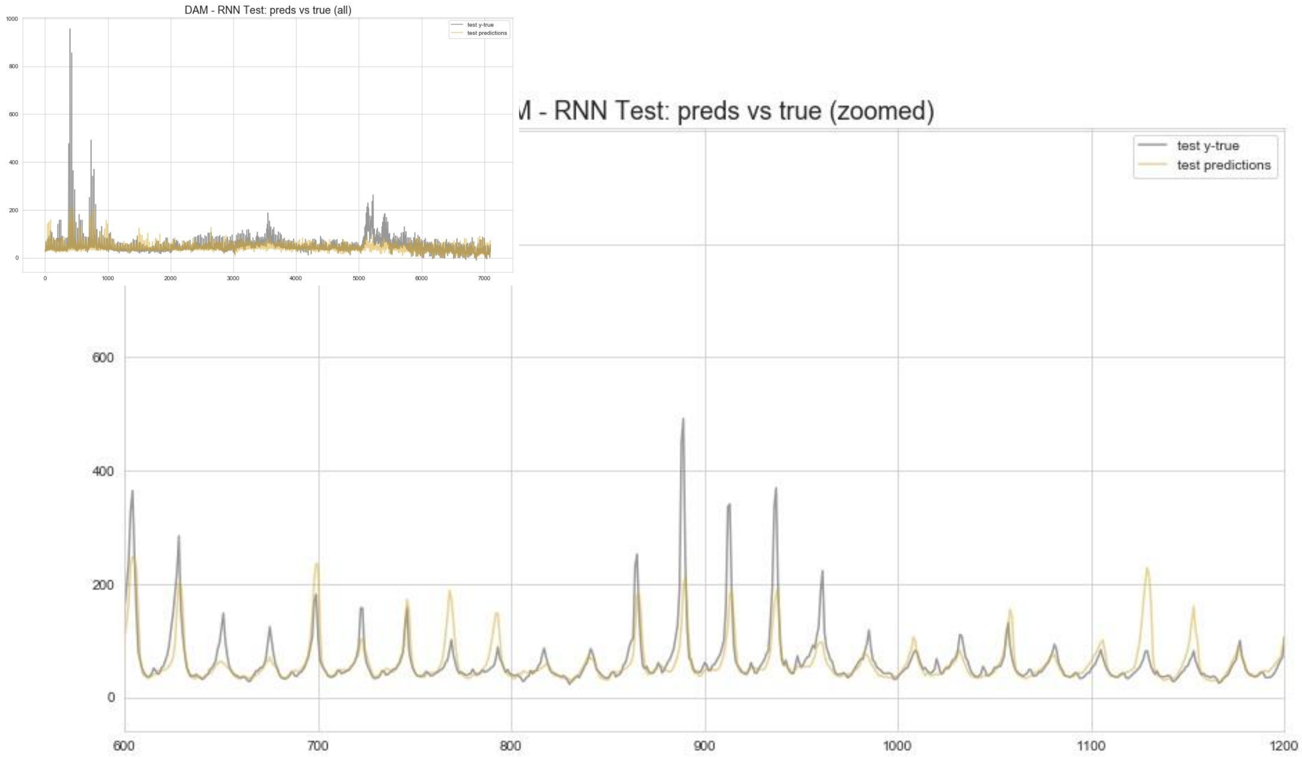
RNN (DAY-AHEAD MODEL) RESULTS

HIGHLIGHTS:

a) R2... 0.44 - 0.47

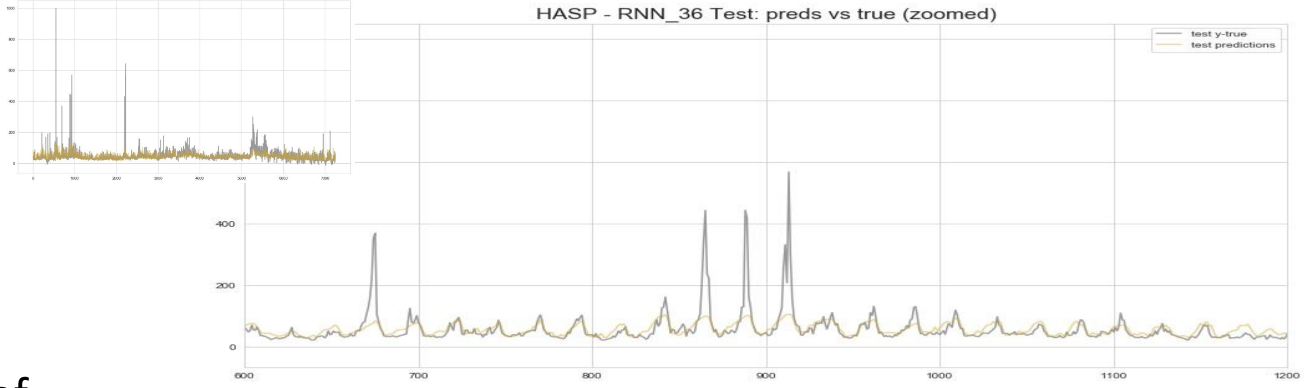
b) Sensitive to LR

c) Finicky to "tune"



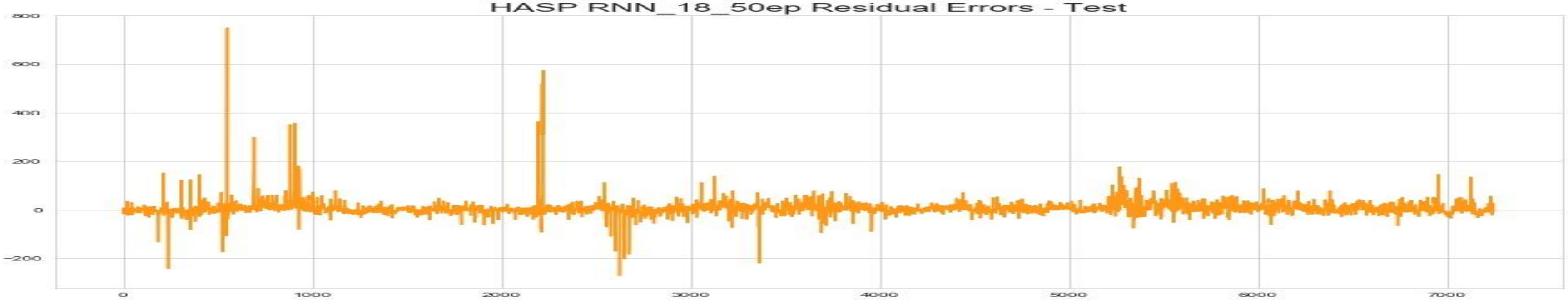
RNN (HOUR-AHEAD PRICES) RESULTS

HIGHLIGHTS:

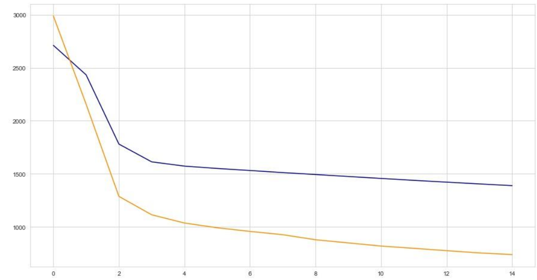


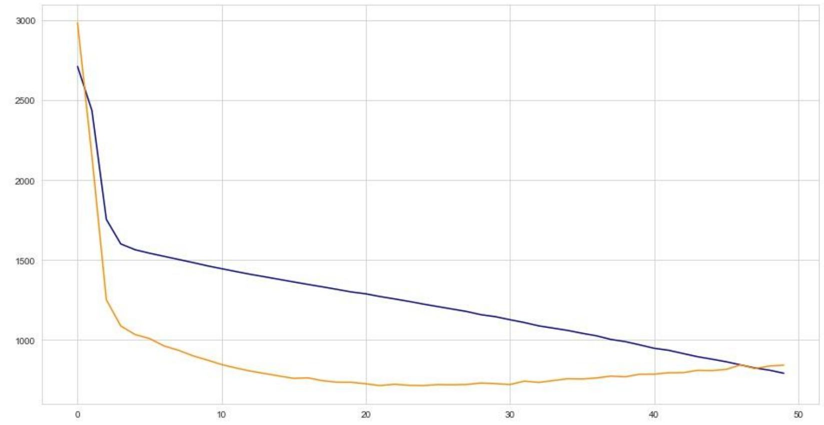
a) Poor prediction:
R2 ~ 0.30 - 0.34

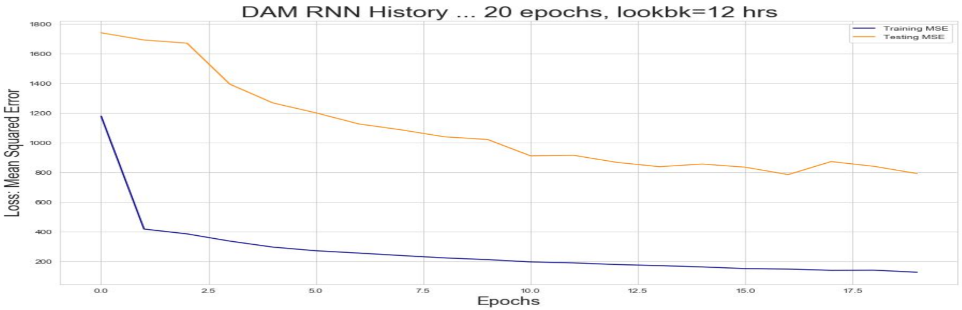
b) Strange behavior of test losses lower than train

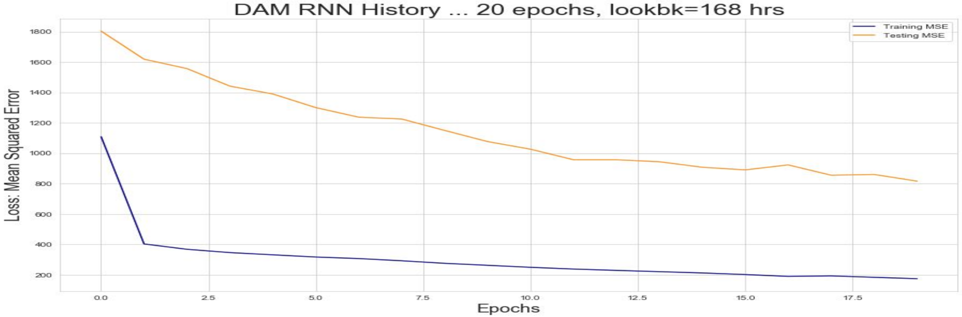


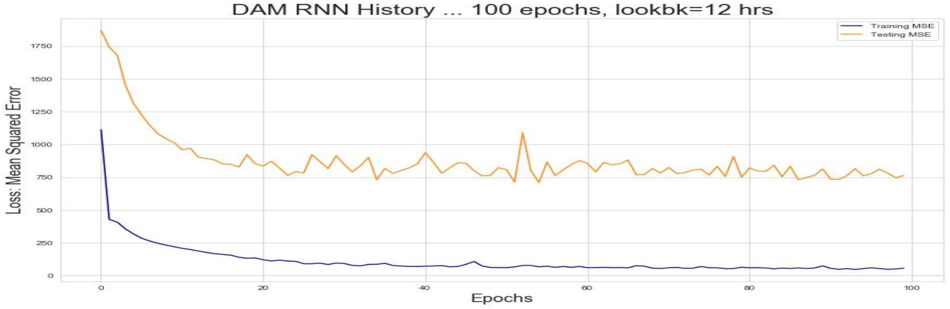
TUNING AN RNN











TAKE-AWAYS

The Day Ahead market appears to show promise for prediction, more so than the Hour Ahead market.

While SARIMAX should perform better than ARIMA, it fails to do so for the parameters attempted, and the long fit times make it unattractive.

RNNs show great promise, matching ARIMA results, but faster fit times, and more easily extractable predictions.

Overall, a successful first cut at this problem, with a number of rich options to pursue for future refinement...

FUTURE WORK ITEMS

Future work:

1. Expand data set back more years to get more train-test sets, fill in missing hours by going back to the CAISO API (skipped some days to come in under CAISO 31-day hard restriction per query)
2. Extend ARIMA grid-search, run SARIMAX grid search
3. Do a systematic exploration of all RNN node/layer/lr/look-bk combinations
4. Try the Facebook Prophet time-series ml tool
5. Cast the dataset into "tabular" format and remove the time index to allow for other estimators like random forest, etc.