Andrea Schmidt
Sales Engineer
SparkCognition

UV Yadav
Data Scientist
SparkCognition

Akshata Mohan
Data Scientist
Cloudflare

Hack The Trail

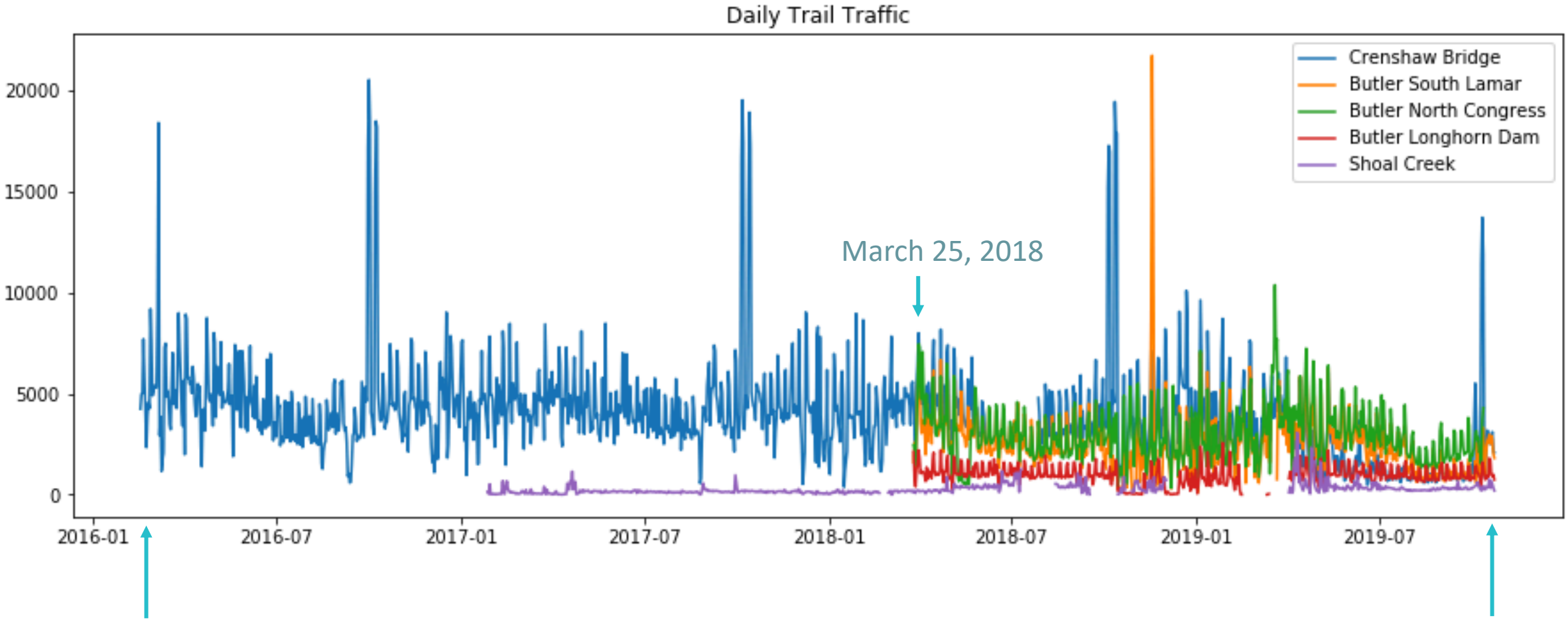
Team Calavera

Challenge #1

Outdoor Voices

Walmart

The Trail Foundation



Path Traffic

Weather

Local Events

Feature Engineering

- **Eventful API** extract events in Austin around downtown
- Weighted score to capture events around trail

Time	Crenshaw Bridge	Butler South Lamar	Butler North Congress	Butler Longhorn Dam	Sh[illegible]reek
2019-10-20	2510.0	2510.0	NaN	1699.0	329.0
2019-10-21	2839.0	2911.0	NaN	961.0	674.0
2019-10-22	3101.0				
2019-10-23	2107.0				
2019-10-24	2076.0				

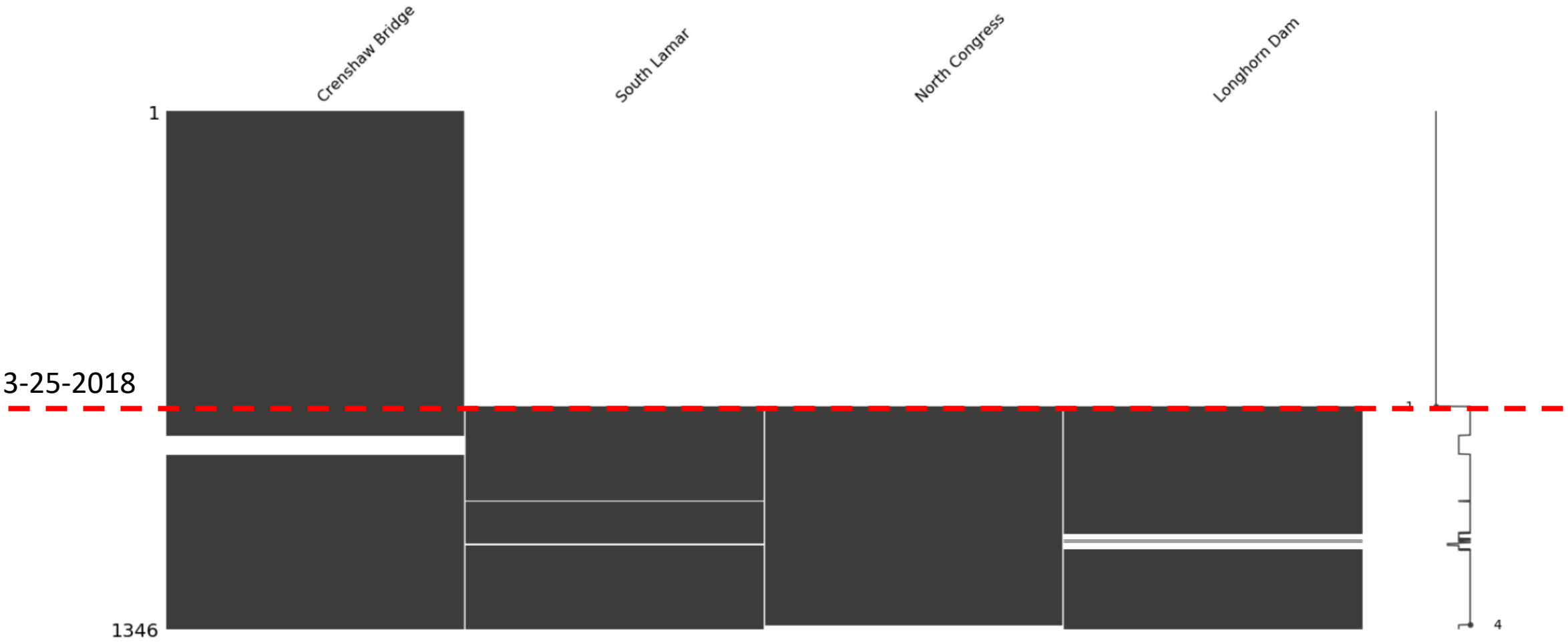
datetime	PRCP	TMAX	TMIN	day	month	woy
2019-10-18	0.0	74	47	4	10	42
2019-10-19	0.0	82	48	5	10	42
						42
						43
						43

	title	venue_name	start_time	stop_time
0	SAFe® Advanced Scrum Master with SASM Certific...	Sheraton Austin Hotel @ the Capitol	2019-10-19 09:00:00	None
1	Guide a Bike Tour!	Postal code 78741, US	2019-10-19 11:53:00	None
2	Early Bird Brunch	Cafe Blue	2019-10-19 09:00:00	None
3	Guide a Bike Tour!	Postal code 78741, US	2019-10-19 11:53:00	None
4	Energetic S@x Practicum & Playshop w/ Reid Mih...	Austin	2019-10-19 10:30:00	None

	Butler Longhorn Dam	Butler North Congress	Butler South Lamar	Crenshaw Bridge	PRCP	TMAX	TMIN	Time	day	event_weight	month	woy
571	899.0	3017.872624	2786.0	2549.0	0.0	74.0	47.0	2019-10-18	4	12	10	42
572	1804.0	4872.386092	2959.0	2707.0	0.0	82.0	48.0	2019-10-19	5	26	10	42
573	1699.0	3876.792360	2510.0	2510.0	0.0	92.0	61.0	2019-10-20	6	23	10	42
574	961.0	2940.687694	2911.0	2839.0	0.4	90.0	59.0	2019-10-21	0	24	10	43
575	918.0	3063.454755	2753.0	3101.0	0.0	76.0	48.0	2019-10-22	1	21	10	43

```
nulls:   49      | Crenshaw Bridge
nulls:   7       | South Lamar
nulls:   11      | North Congress
nulls:   39      | Longhorn Dam
```

```
df2 = df.copy().drop("Shoal Creek",axis=1)
msno.matrix(df2)
plt.show()
```

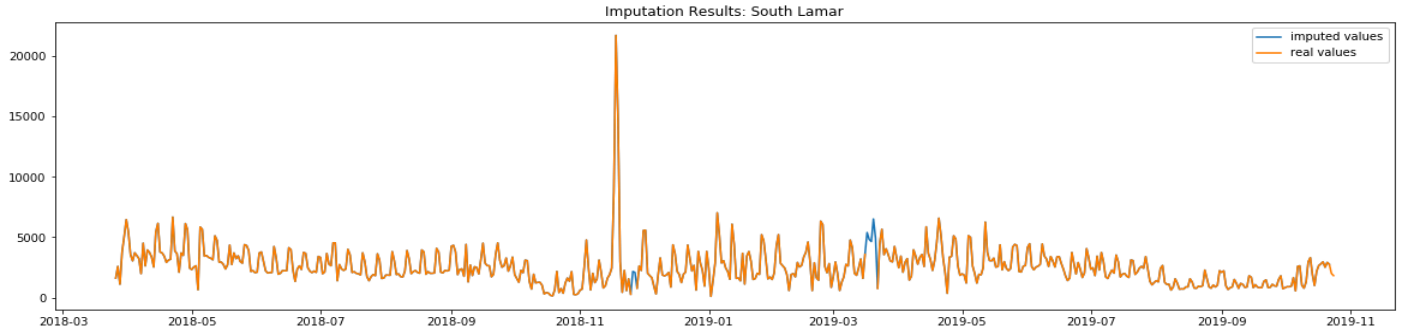


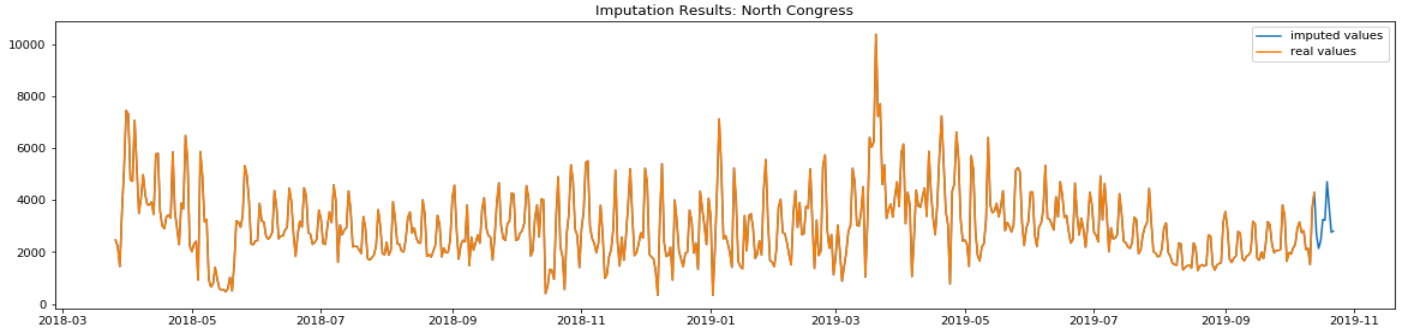
3-25-2018

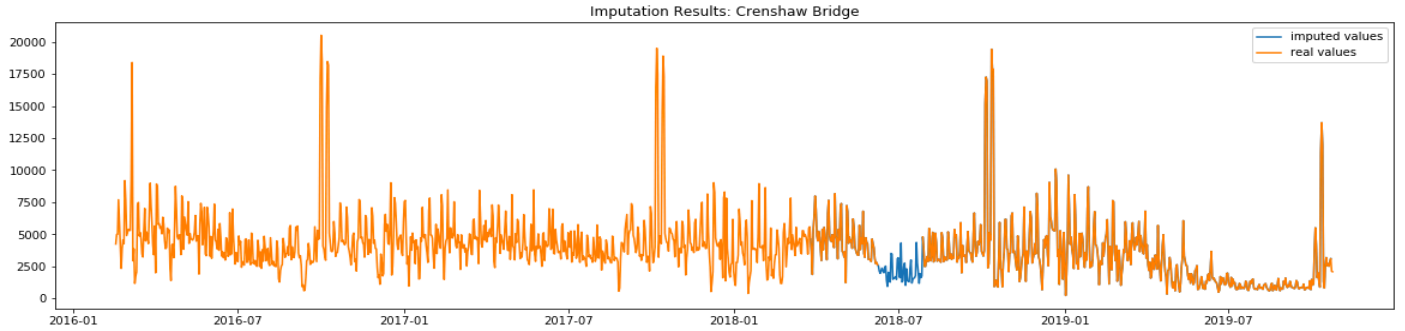
Imputation using AutoML platform

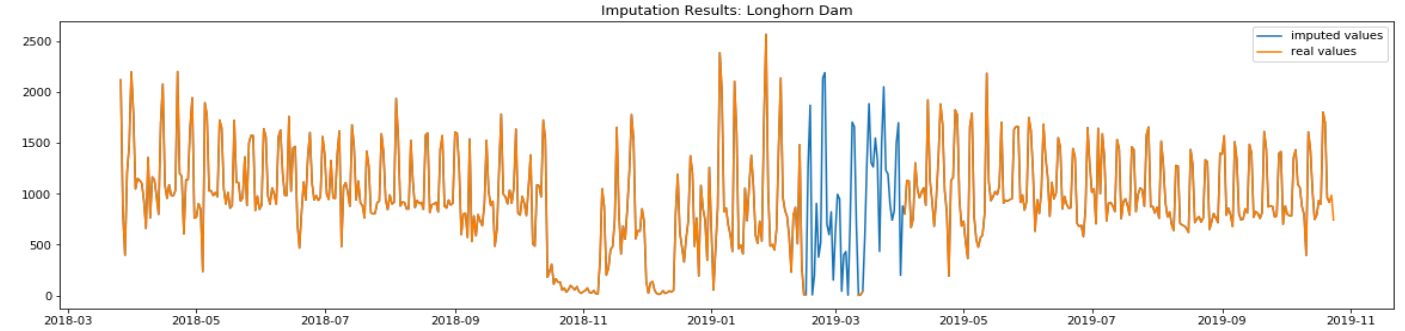
DARWIN™

- ***South Lamar*** (DeepNet : $R^2 = 0.41$)
  - week of the year – 38%
  - North Congress – 16%
  - Precipitation – 13%
- ***North Congress*** (DeepNet: $R^2 = 0.70$)
  - week of the year – 37%
  - South Lamar– 14%
  - Temperature low – 14%
- ***Crenshaw Bridge*** (RandomForest: $R^2 = 0.92$)
  - week of the year – 38%
  - North Congress – 16%
  - Precipitation – 13%
- ***Longhorn Dam*** (SkLearn: $R^2 = 0.84$)
  - week of the year – 32%
  - Precipitation – 11%
  - The other 3 sensors contributed about 10% each









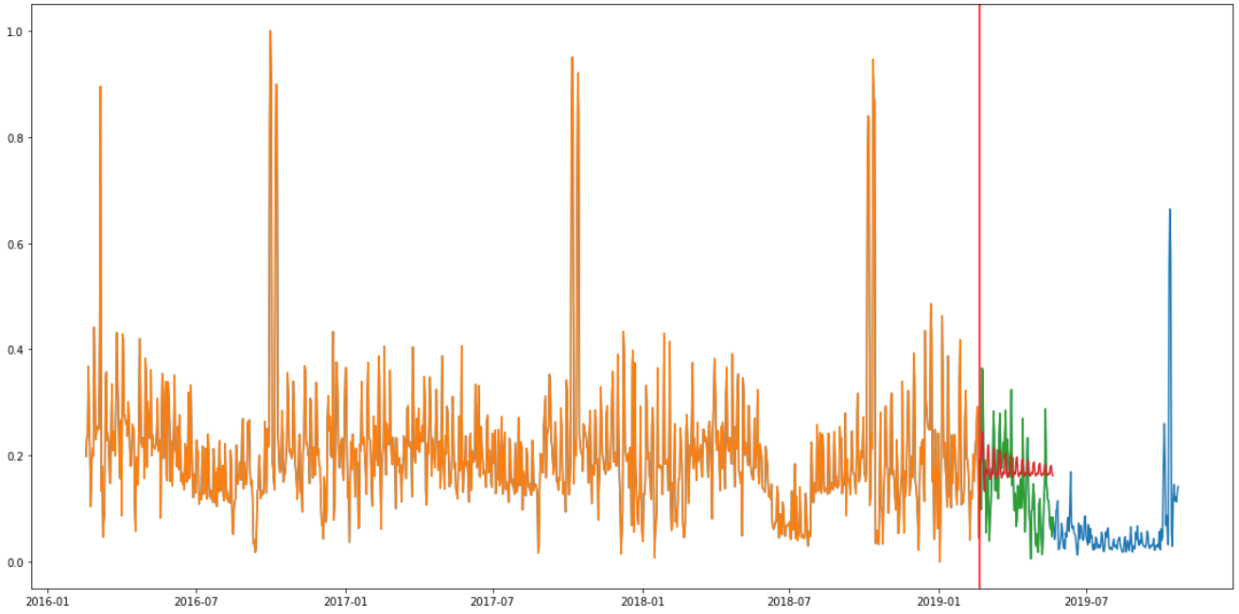
Forecasting Results

SARIMA | SARIMAX | Kalman Filter

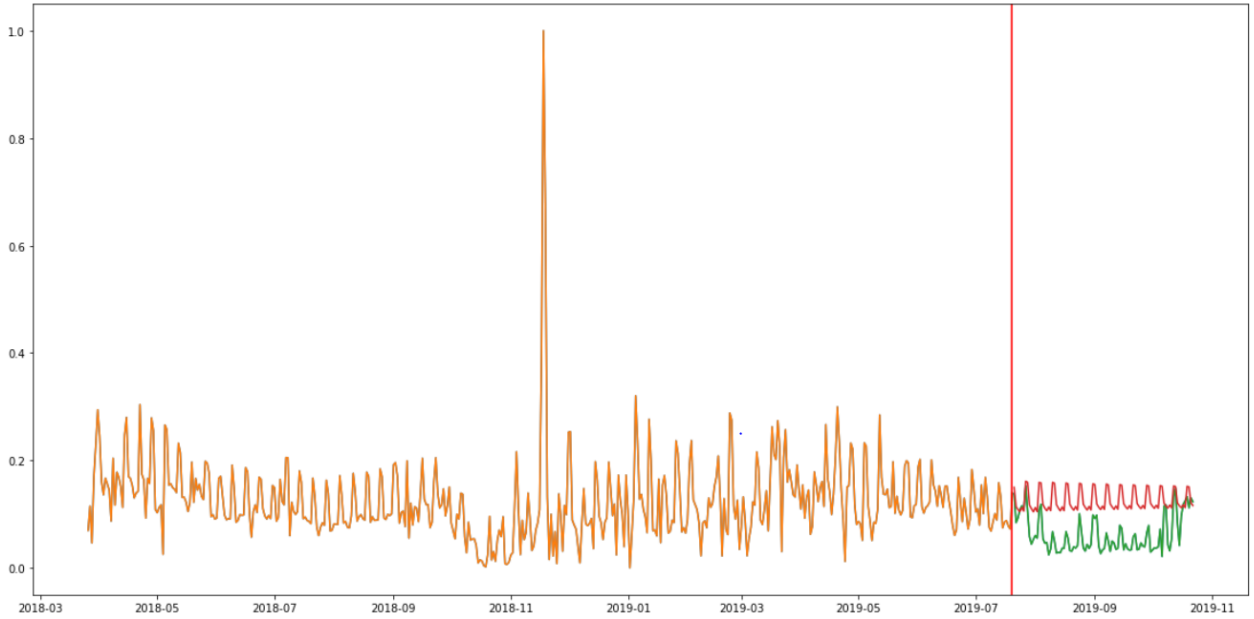
Crenshaw Bridge :	0.08276	MSE
South Lamar:	0.1027	MSE
North Congress:	N/A	
Longhorn Dam:	N/A	

SARIMA

Crenshaw Bridge (MSE: 0.08276)



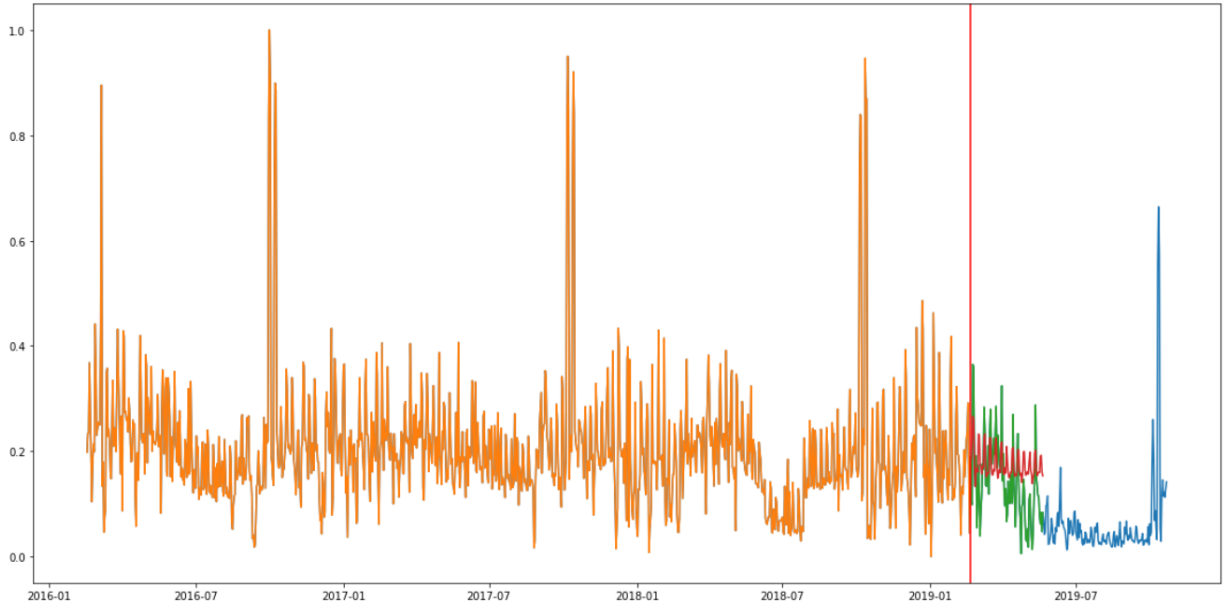
South Lamar (MSE: 0.10271)



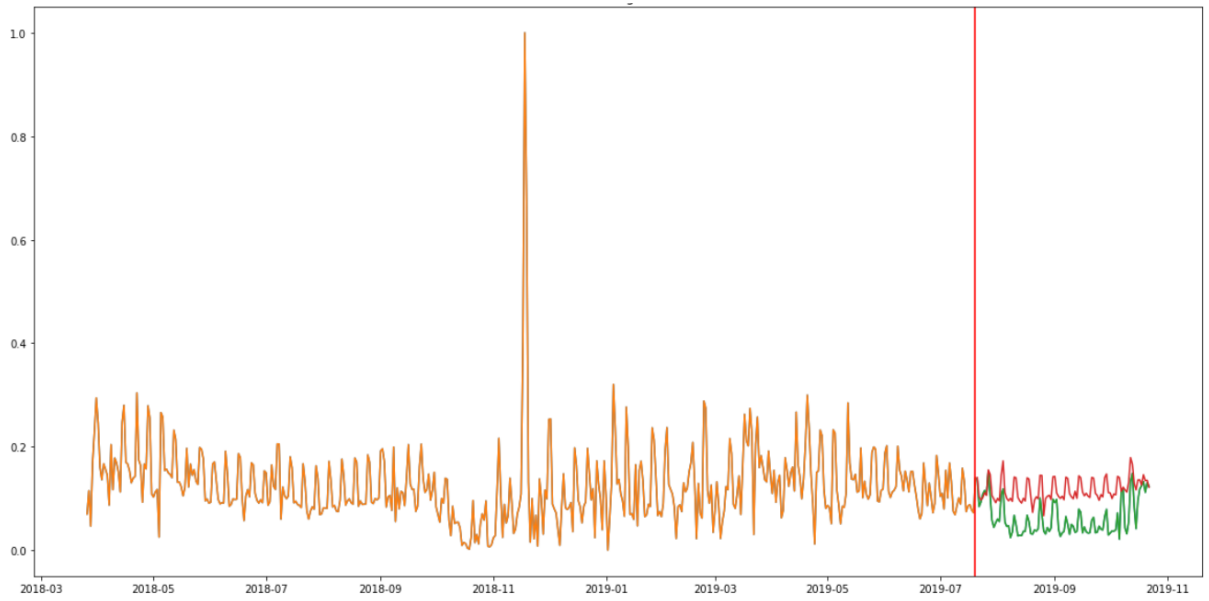
Crenshaw Bridge :	0.08276	MSE
South Lamar:	0.1027	MSE
North Congress:	N/A	
Longhorn Dam:	N/A	

SARIMAX

Crenshaw Bridge (MSE: 0.03861)



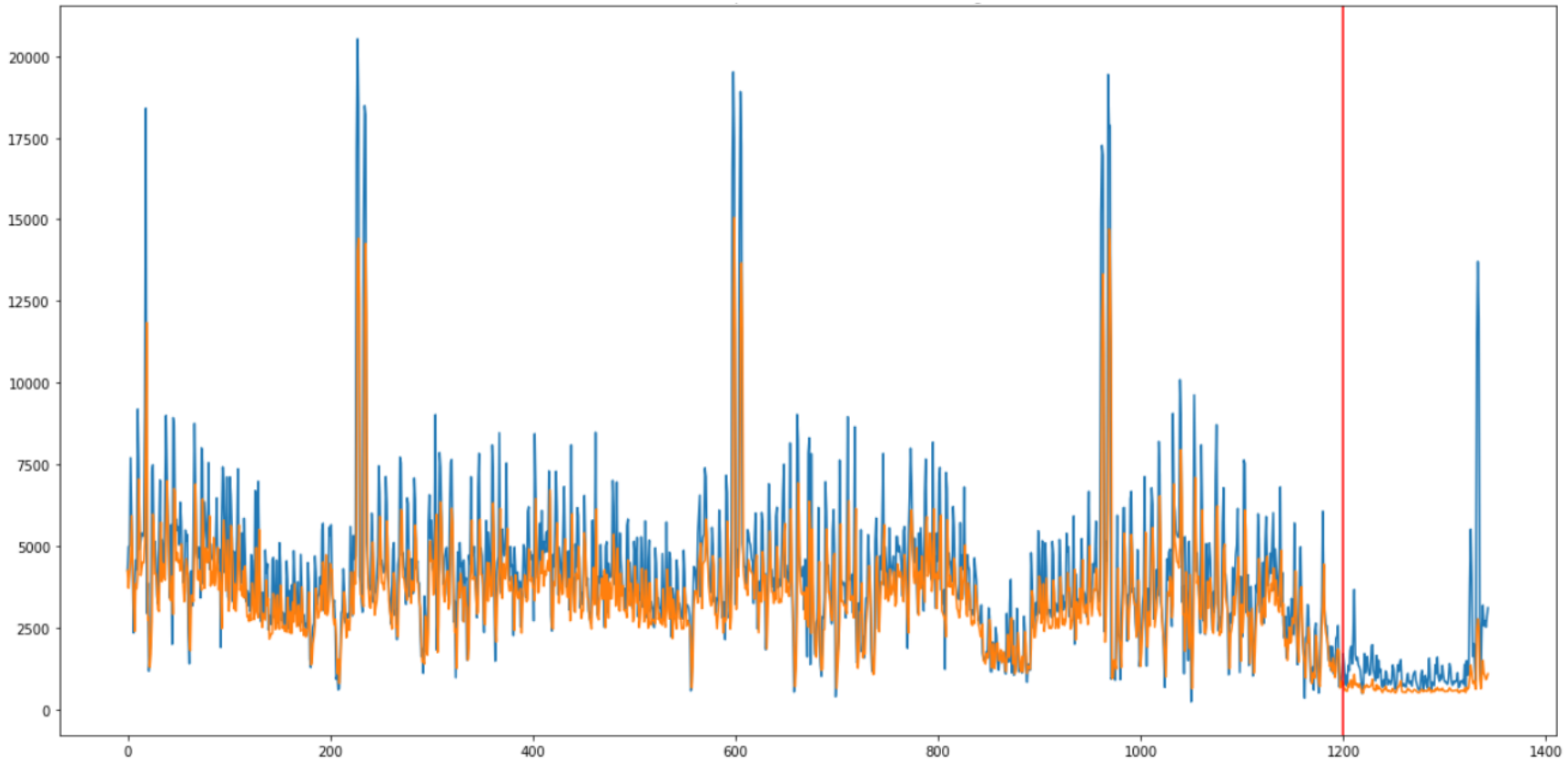
South Lamar (MSE: 0.10328)



Crenshaw Bridge :	0.5475	MSE
South Lamar:	N/A	
North Congress:	N/A	
Longhorn Dam:	N/A	

Kalman Filter

Crenshaw Bridge (MSE =0.5475)

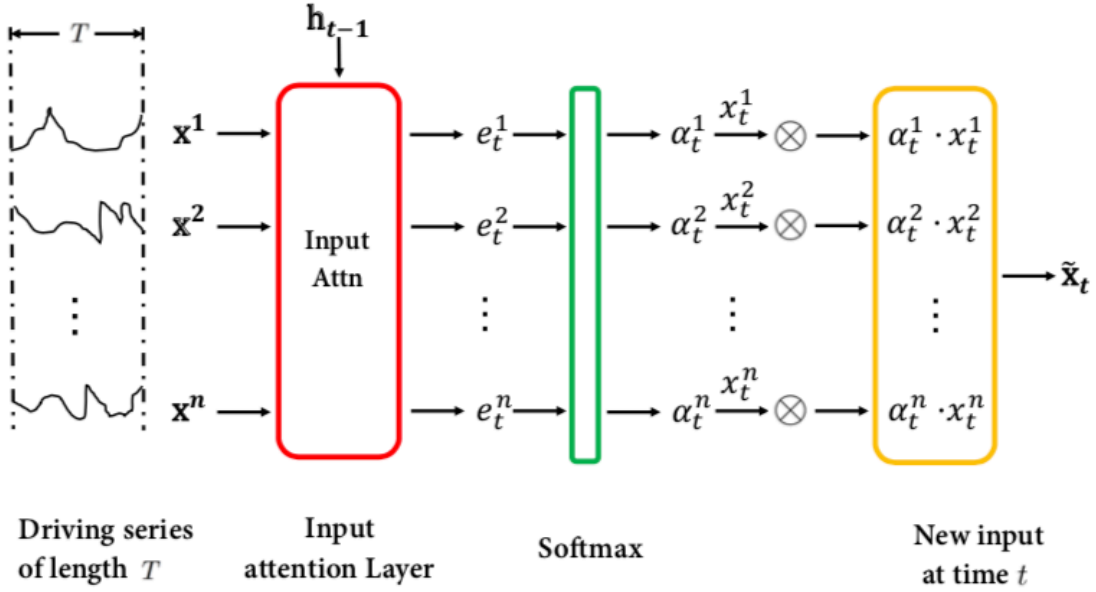


Forecasting Results

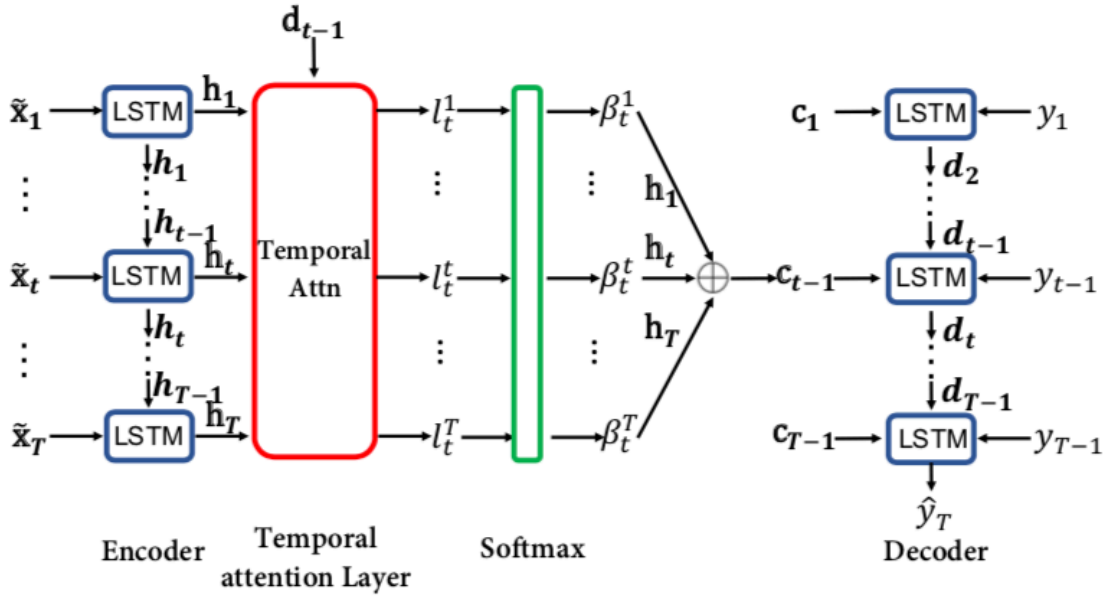
RNNs with dual-Attention| Self-Attention + TCNs

RNNs with dual-attention

$y_1, y_2, \ldots y_n$ + Exog (weather & events)



(a) Input Attention Mechanism

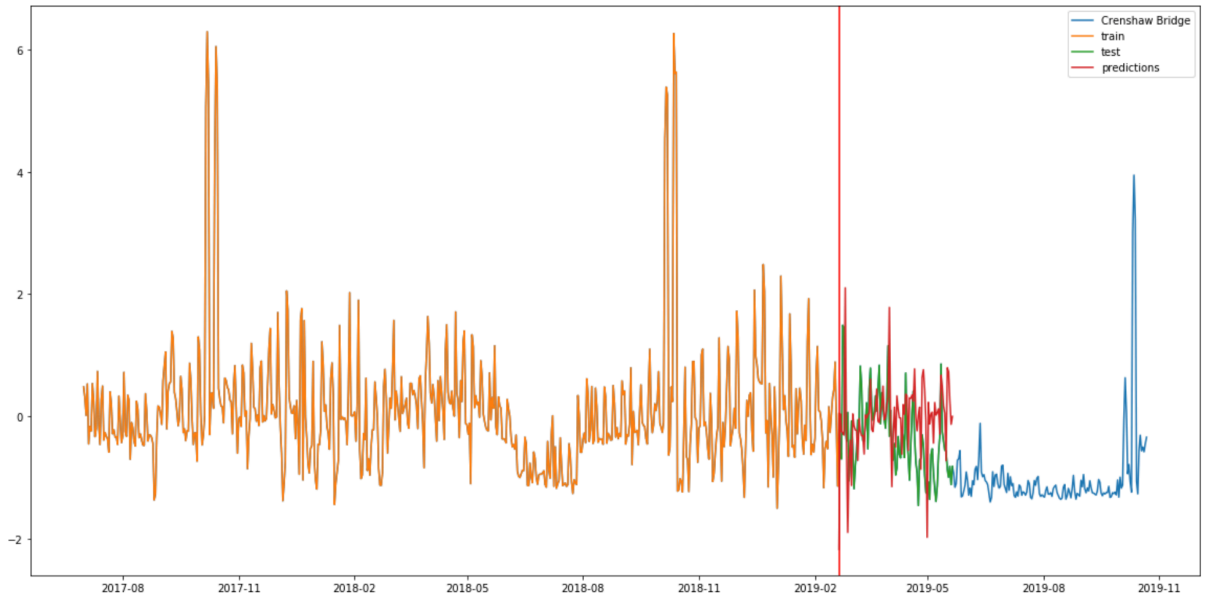


(b) Temporal Attention Mechanism

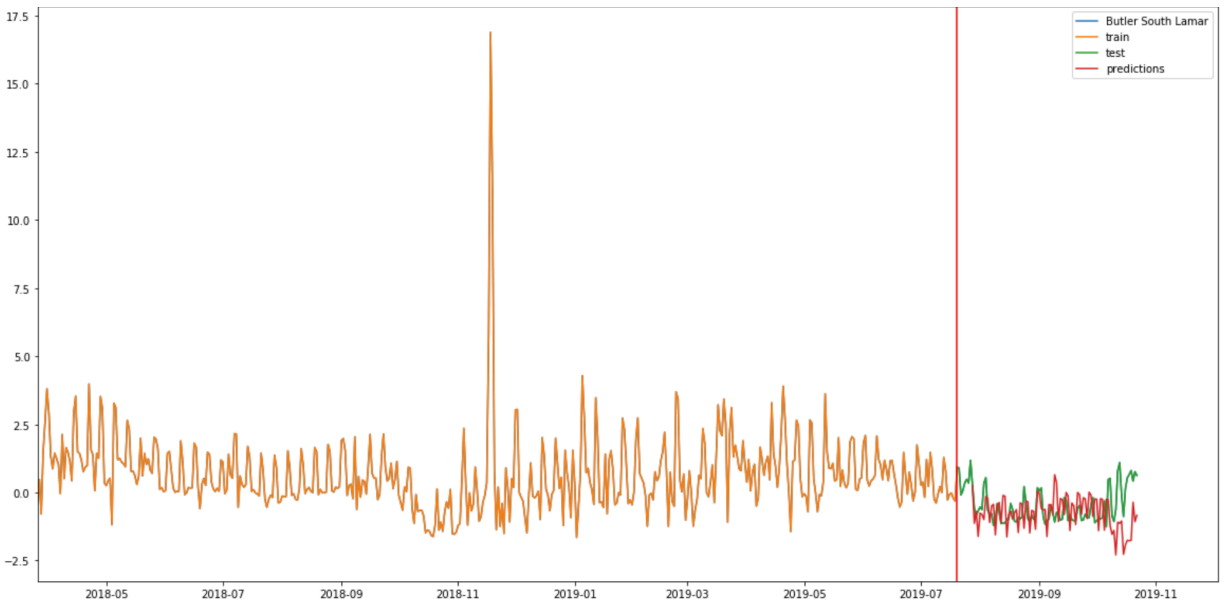
Crenshaw Bridge :	0.798	MSE
South Lamar:	0.865	MSE
North Congress:	0.7281	MSE
Longhorn Dam:	0.2350	MSE

RNNs with DA

Crenshaw Bridge

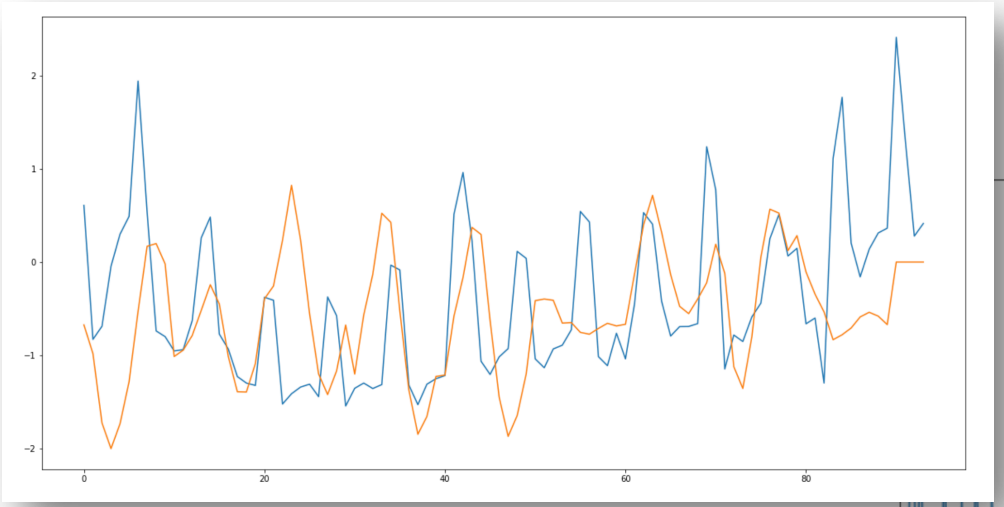


South Lamar

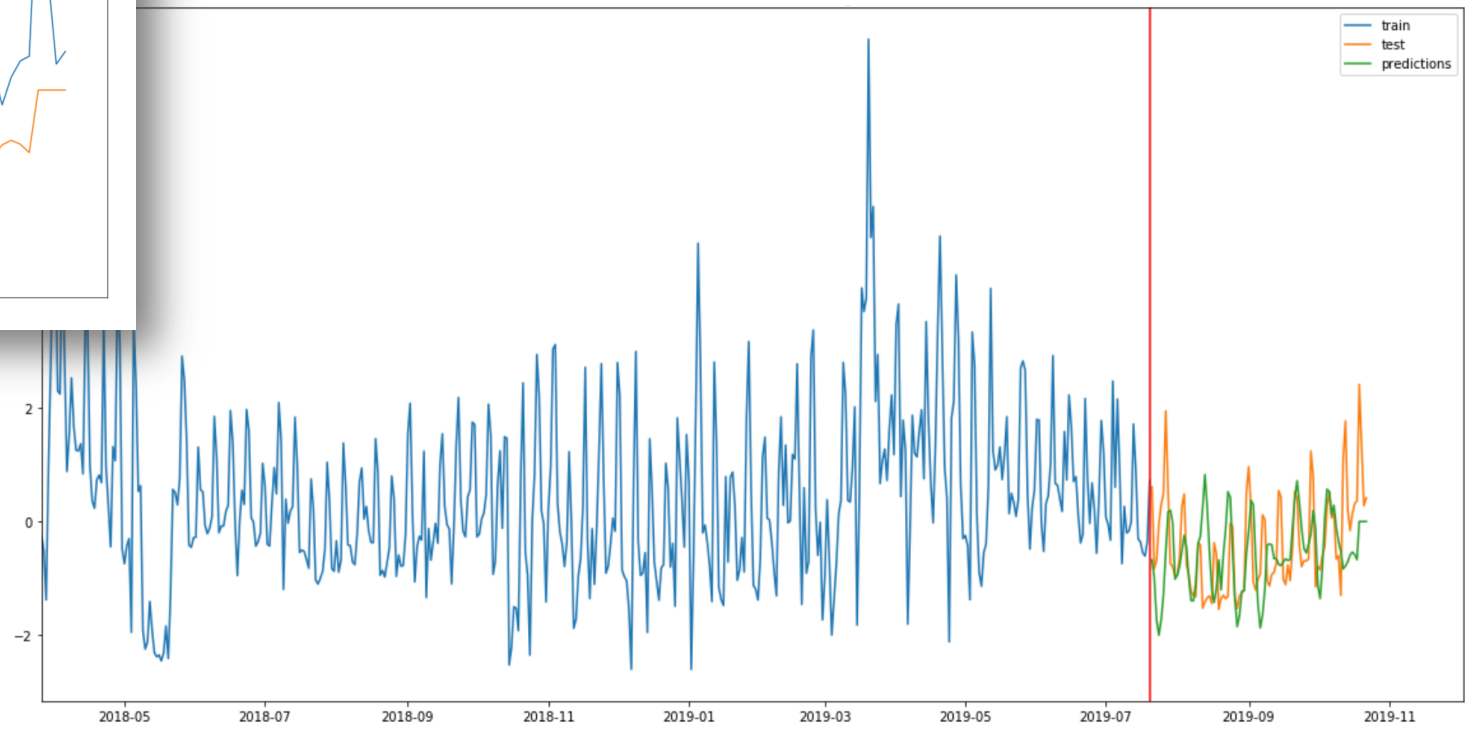


Crenshaw Bridge :	0.798	MSE
South Lamar:	0.865	MSE
North Congress:	0.7281	MSE
Longhorn Dam:	0.2350	MSE

RNNs with DA

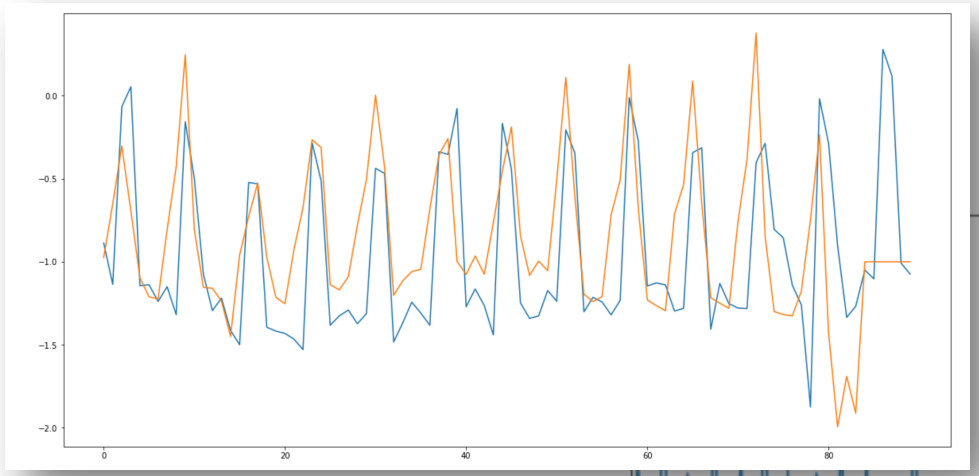


North Congress

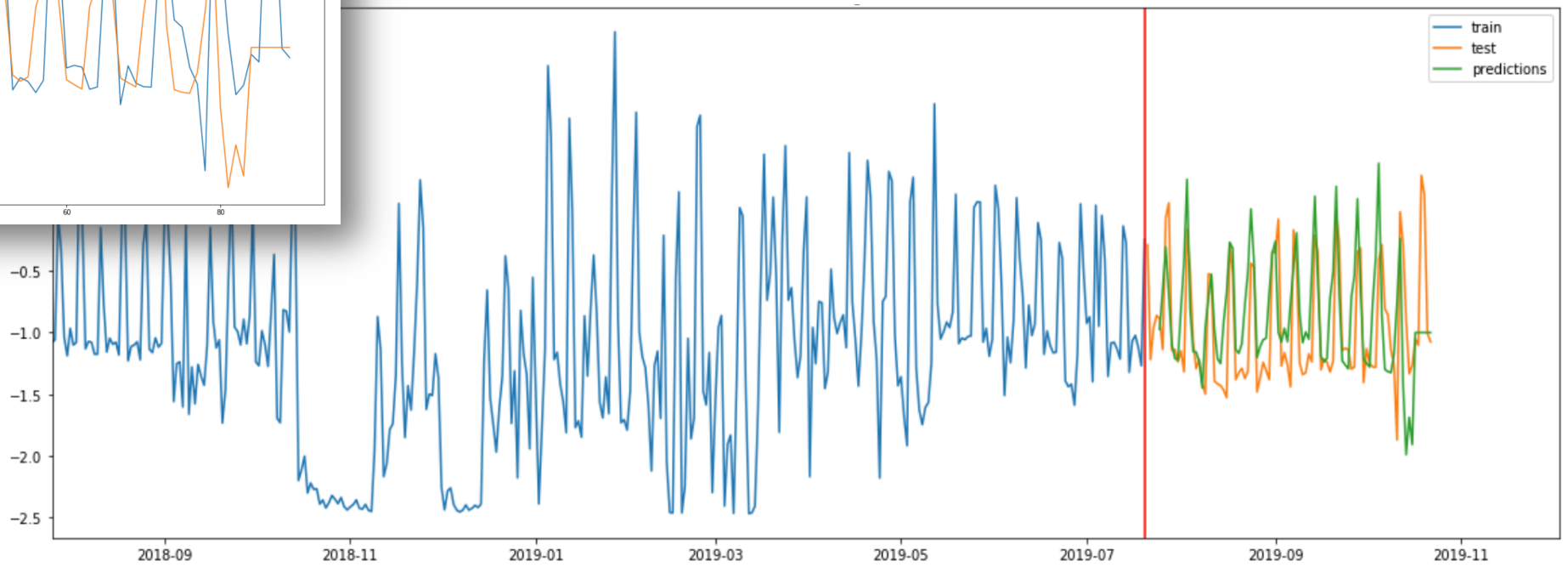


Crenshaw Bridge :	0.798	MSE
South Lamar:	0.865	MSE
North Congress:	0.7281	MSE
Longhorn Dam:	0.2350	MSE

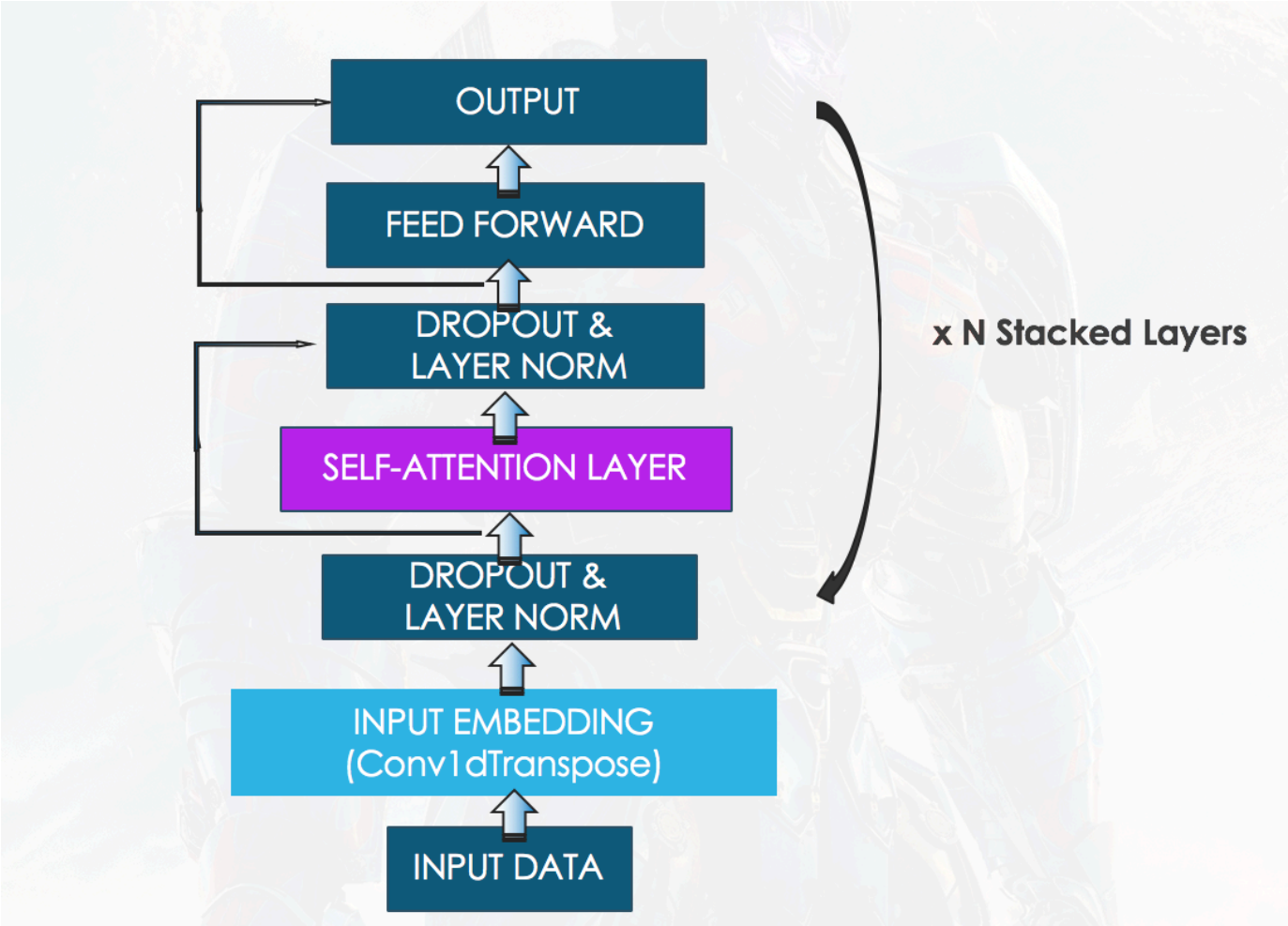
RNNs with DA



Longhorn Dam



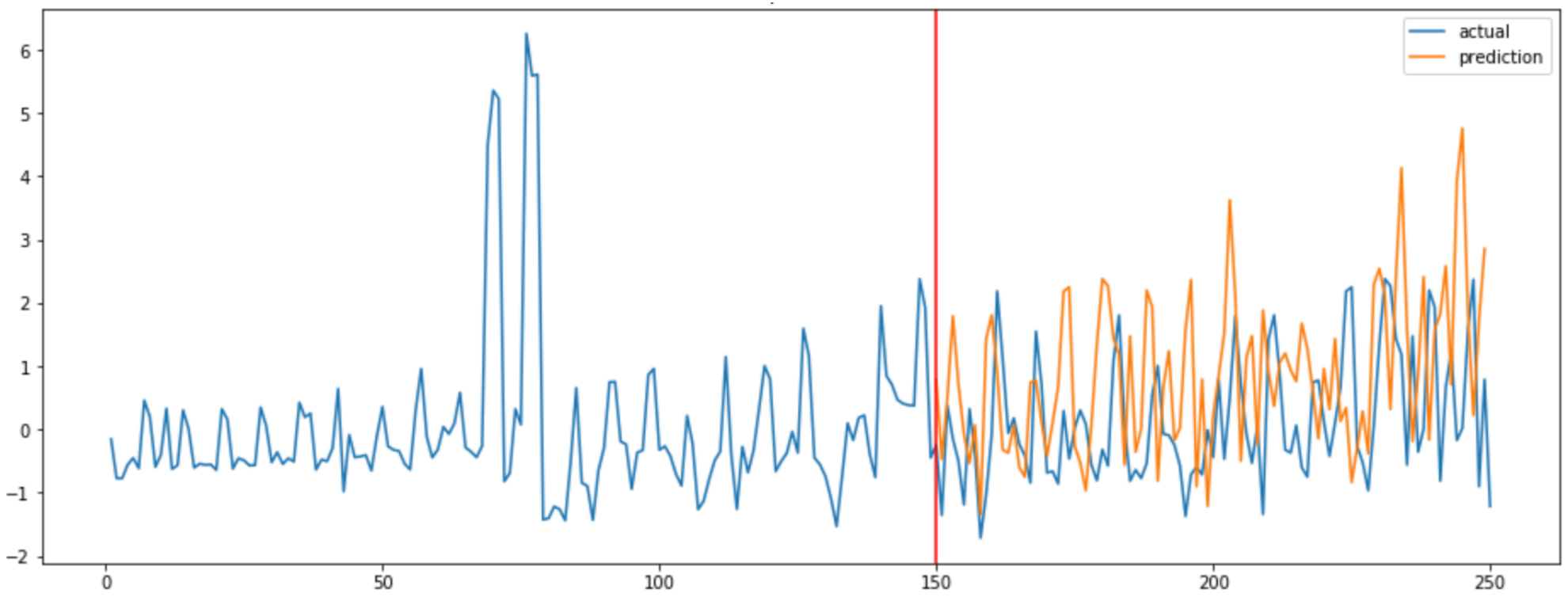
TCNs with Self-attention Architecture



Crenshaw Bridge :	0.340	MSE
South Lamar:	0.983	MSE
North Congress:	0.560	MSE
Longhorn Dam:	0.435	MSE

TCNs with SA

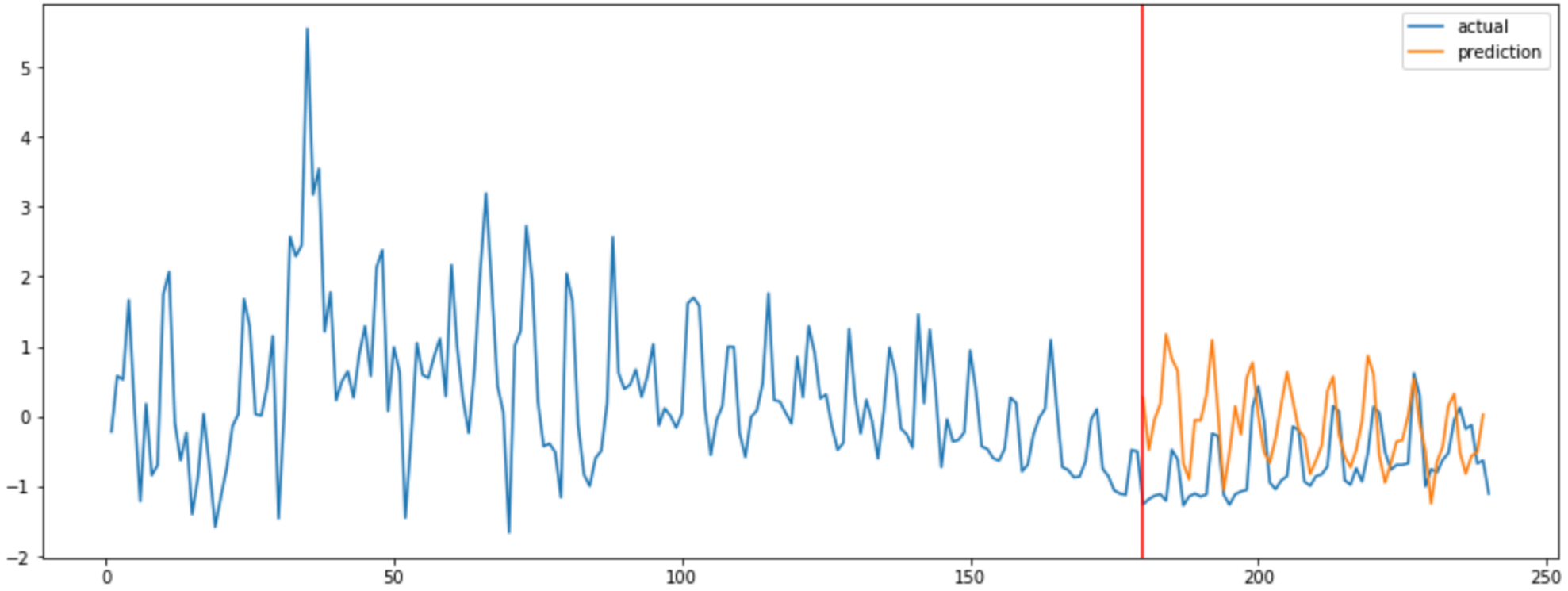
Crenshaw Bridge – 90 Day Forecast



Crenshaw Bridge :	0.340	MSE
South Lamar:	0.983	MSE
North Congress:	0.560	MSE
Longhorn Dam:	0.435	MSE

TCNs with SA

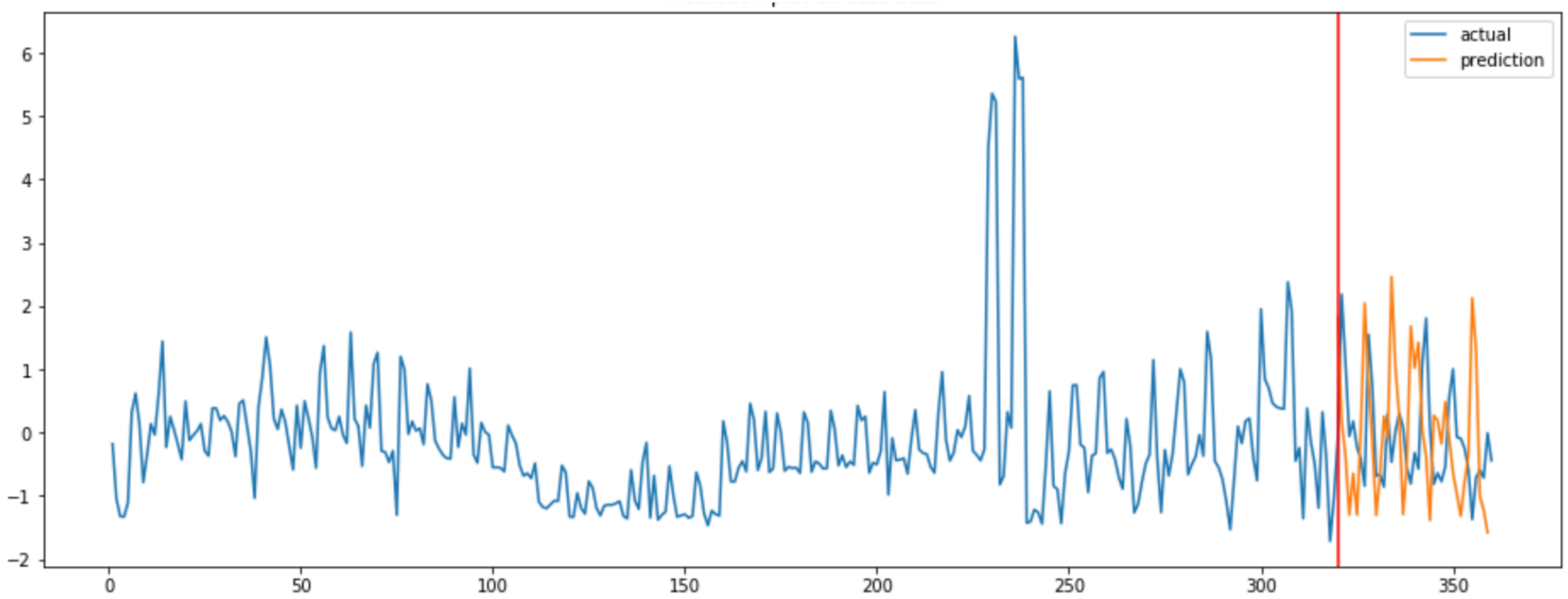
North Congress – 60 Day Forecast



Crenshaw Bridge :	0.340	MSE
South Lamar:	0.983	MSE
North Congress:	0.560	MSE
Longhorn Dam:	0.435	MSE

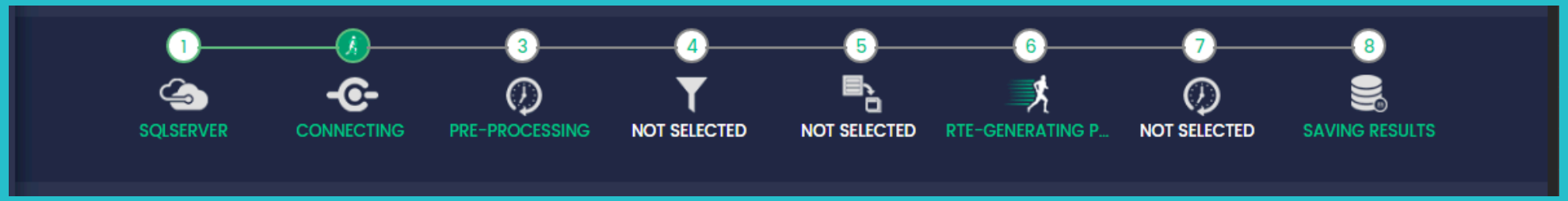
TCNs with SA

Crenshaw Bridge – 40 Day Forecast



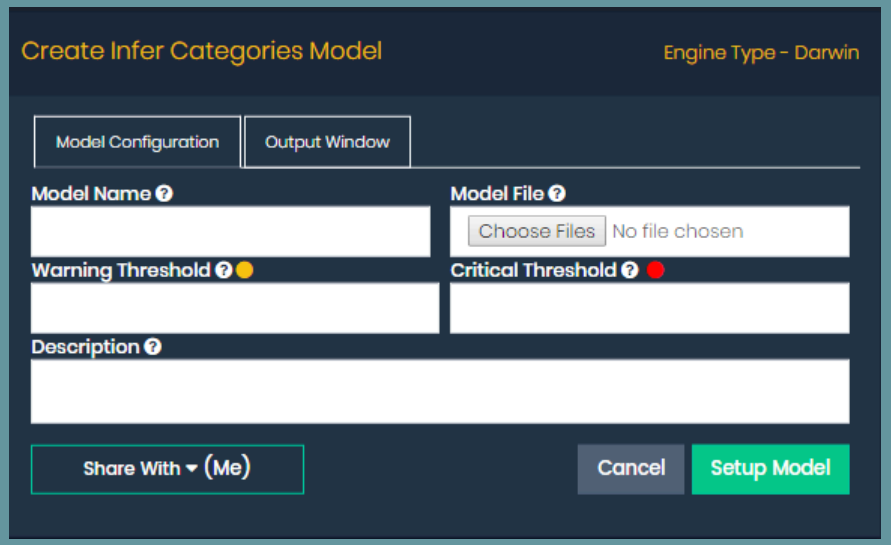
Summary

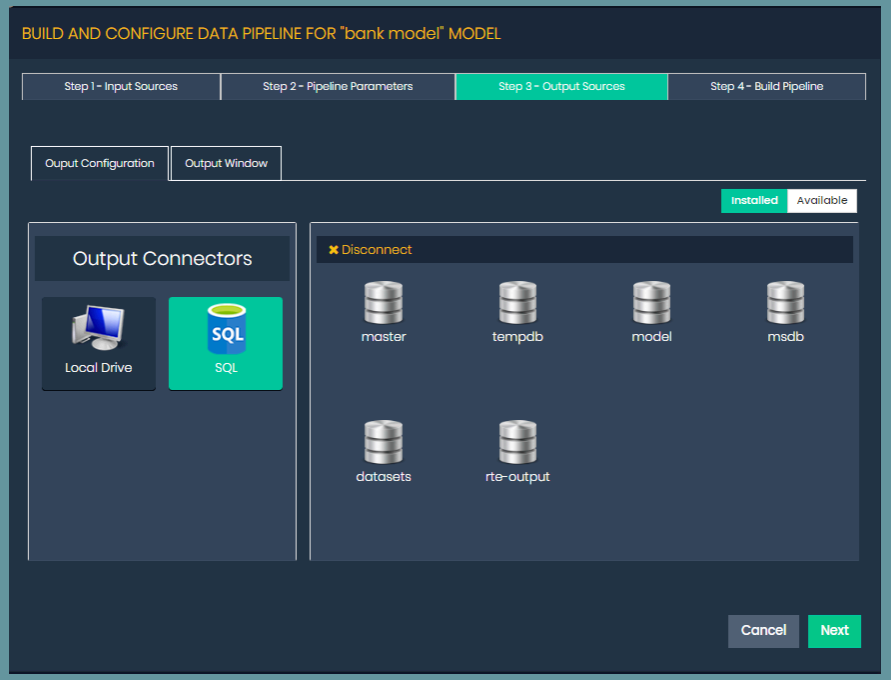
- *Events weight provided more explainability to the models*
- *Events weights value calculation could be improved by including the events*
- *SARIMAX gave lower MSE than SARMIA because of improved predictions due to exogenous variables like weather and events*
- *Kalman filters work best when observed data is accurately captured and less noisy*
- *TCNs with Self-attention is good in learning long term dependencies (like 50 to 90 days in the future)*
- *TCNs with Self – attention is comparatively faster in training as compared to RNNs*
- *Multi-variate forecasting can be possible with TCNs + SA*

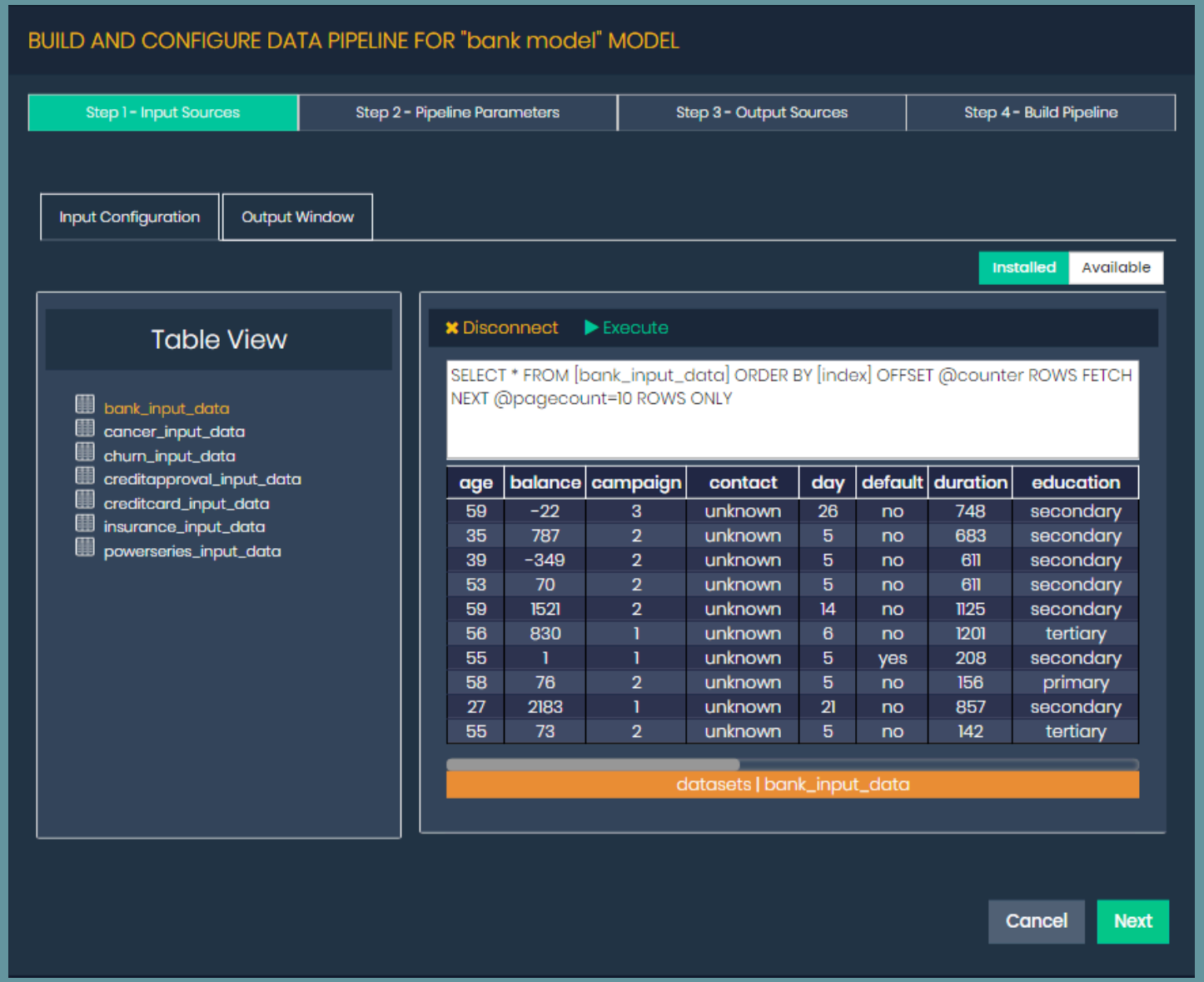


Solution Deployment

DARWIN™


Create Infer Categories Model
Engine Type - Darwin
Model Configuration
Output Window
Model Name
Model File
Choose Files
No file chosen
Warning Threshold
Critical Threshold
Description
Share With (Me)
Cancel
Setup Model


BUILD AND CONFIGURE DATA PIPELINE FOR "bank model" MODEL
Step 1 - Input Sources
Step 2 - Pipeline Parameters
Step 3 - Output Sources
Step 4 - Build Pipeline
Ouput Configuration
Output Window
Installed
Available
Output Connectors
Local Drive
SQL
Disconnect
master
tempdb
model
msdb
datasets
rte-output
Cancel
Next


BUILD AND CONFIGURE DATA PIPELINE FOR "bank model" MODEL
Step 1 - Input Sources
Step 2 - Pipeline Parameters
Step 3 - Output Sources
Step 4 - Build Pipeline
Input Configuration
Output Window
Installed
Available
Table View
bank_input_data
cancer_input_data
churn_input_data
creditapproval_input_data
creditcard_input_data
insurance_input_data
powerseries_input_data
Disconnect
Execute
SELECT * FROM [bank_input_data] ORDER BY [index] OFFSET @counter ROWS FETCH NEXT @pagecount=10 ROWS ONLY
age balance campaign contact day default duration education
59 -22 3 unknown 26 no 748 secondary
35 787 2 unknown 5 no 683 secondary
39 -349 2 unknown 5 no 611 secondary
53 70 2 unknown 5 no 611 secondary
59 1521 2 unknown 14 no 1125 secondary
56 830 1 unknown 6 no 1201 tertiary
55 1 1 unknown 5 yes 208 secondary
58 76 2 unknown 5 no 156 primary
27 2183 1 unknown 21 no 857 secondary
55 73 2 unknown 5 no 142 tertiary
datasets | bank_input_data
Cancel
Next

Run_1 Cycles: 3
Oct 04 2019 08:36AM

Description

Update Model Beta | View Logs

Confidence Level		Predictions					
Cycle	Average	Current	Total				
90 %	86.5 %	10	30	Auto	Manual	3	00:00:45	38 s

No. of Cycles | Refresh Interval | Last Cycle Time

1	2	3	4	5	6	7	8
SQLSERVER	CONNECTING	PRE-PROCESSING	NOT SELECTED	NOT SELECTED	RTE-GENERATING P...	NOT SELECTED	SAVING RESULTS

Monitor | Results | Model Detail | Data Insights (Beta)

Predicted_Results	Prediction_prob	age	balance	campaign	contact	day	default	duration	education	housing	index	job	level0	loan	marital	m
no	85.54	59	-22	3	unknown	26	no	748	secondary	yes	0	services	0	no	married	
no	76.2	35	787	2	unknown	5	no	683	secondary	yes	1	technician	1	no	divorced	
no	87.13	39	-349	2	unknown	5	no	611	secondary	yes	2	admin.	2	no	divorced	
no	94.31	53	70	2	unknown	5	no	611	secondary	yes	3	blue-collar	3	no	married	
yes	54.84	59	1521	2	unknown	14	no	1125	secondary	yes	4	retired	4	yes	married	
yes	54.77	56	830	1	unknown	6	no	1201	tertiary	yes	5	management	5	yes	married	
no	99.97	55	1	1	unknown	5	yes	208	secondary	yes	6	services	6	no	divorced	
no	100	58	76	2	unknown	5	no	156	primary	yes	7	blue-collar	7	no	married	
no	75.78	27	2183	1	unknown	21	no	857	secondary	yes	8	technician	8	yes	married	
no	100	55	73	2	unknown	5	no	142	tertiary	yes	9	management	9	no	married	

Appendix A.

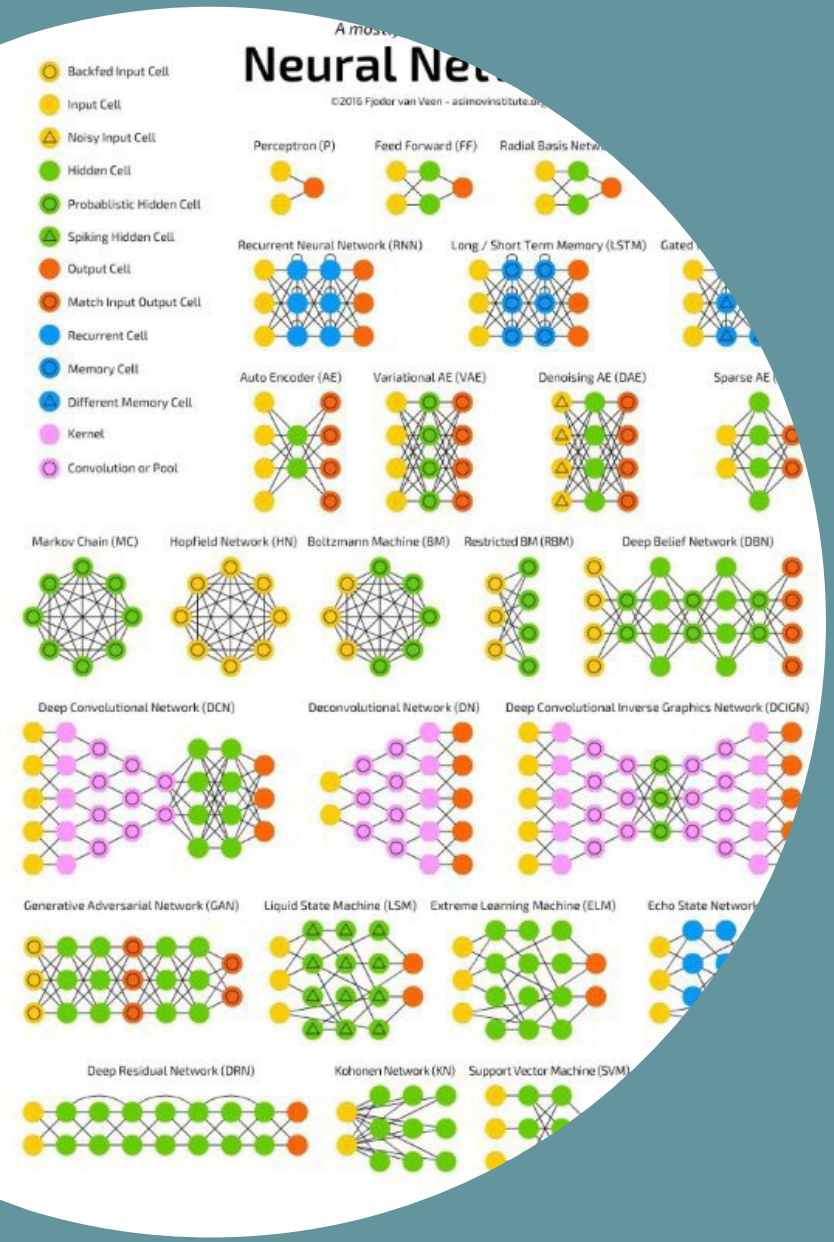
Training Results for Imputation Models

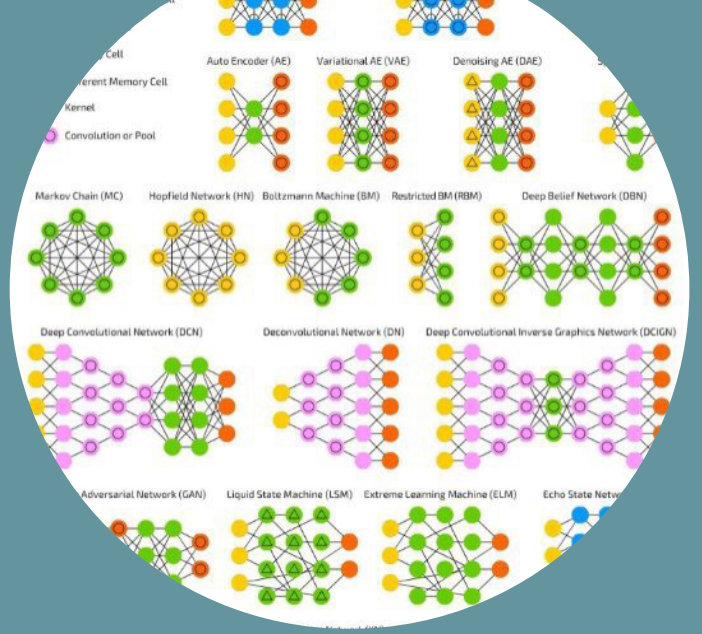

← Back to Models
Models > SouthLamarModel
Actions
TRAINING
PREDICTIONS
TRAINING TIME
10 minutes 18 seconds
ALGORITHM
DeepNet
MSE
1560713.7248
R²
0.4095
Initial Setup and Training Sessions
Training Results
Feature Importance
Model Architecture Diagram
Training Sets
1
FILENAME
R²
SL_TRAIN.csv
0.4095
Results
Export
Actual
Predicted
South Lamar
8k
6k
4k
2k
0
5
10
15
20
25
30
35
40
45
50
Visible Points:
50
100
250
1
2
3
4
...
11
12

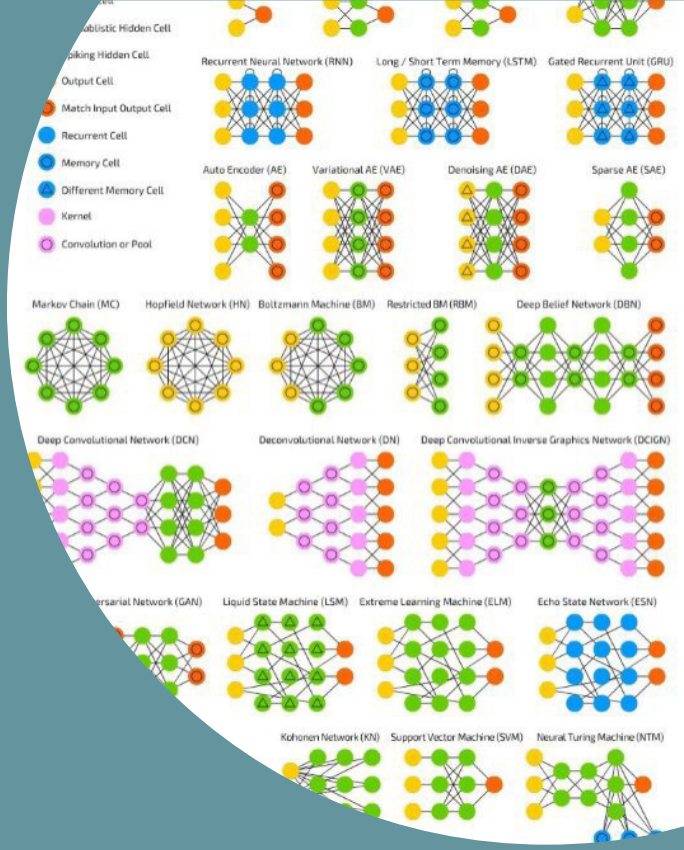

← Back to Models
Models > NorthCongressModel
Actions
TRAINING
PREDICTIONS
TRAINING TIME
10 minutes 12 seconds
ALGORITHM
DeepNet
MSE
534702.4967
R²
0.6988
Initial Setup and Training Sessions
Training Results
Feature Importance
Model Architecture Diagram
Training Sets
1
FILENAME
R²
North Congress_TRAIN.csv
0.6988
Results
Export
Actual
Predicted
North Congress
8k
6k
4k
2k
0
5
10
15
20
25
30
35
40
45
50
Visible Points:
50
100
250
1
2
3
4
...
11
12


← Back to Models
Models > CrenshawBridgeModel
Actions
TRAINING
PREDICTIONS
TRAINING TIME
10 minutes 18 seconds
ALGORITHM
RandomForestRegressor
MSE
514922.2043
R²
0.9189
Initial Setup and Training Sessions
Training Results
Feature Importance
Training Sets
1
FILENAME
R²
Crenshaw Bridge_TRAIN.csv
0.9189
Results
Export
Actual
Predicted
Crenshaw Bridge
10k
8k
6k
4k
2k
0
5
10
15
20
25
30
35
40
45
50
Visible Points:
50
100
250
1
2
3
4
...
10
11


← Back to Models
Models > LonghornDamModel
Actions
TRAINING
PREDICTIONS
TRAINING TIME
10 minutes 19 seconds
MSE
38212.5430
R²
0.8362
Initial Setup and Training Sessions
Training Results
Feature Importance
Training Sets
1
FILENAME
R²
Longhorn Dam_TRAIN.csv
0.8362
Results
Export
Actual
Predicted
Longhorn Dam
2500
2000
1500
1000
500
0
5
10
15
20
25
30
35
40
45
50
Visible Points:
50
100
250
1
2
3
4
...
9
10



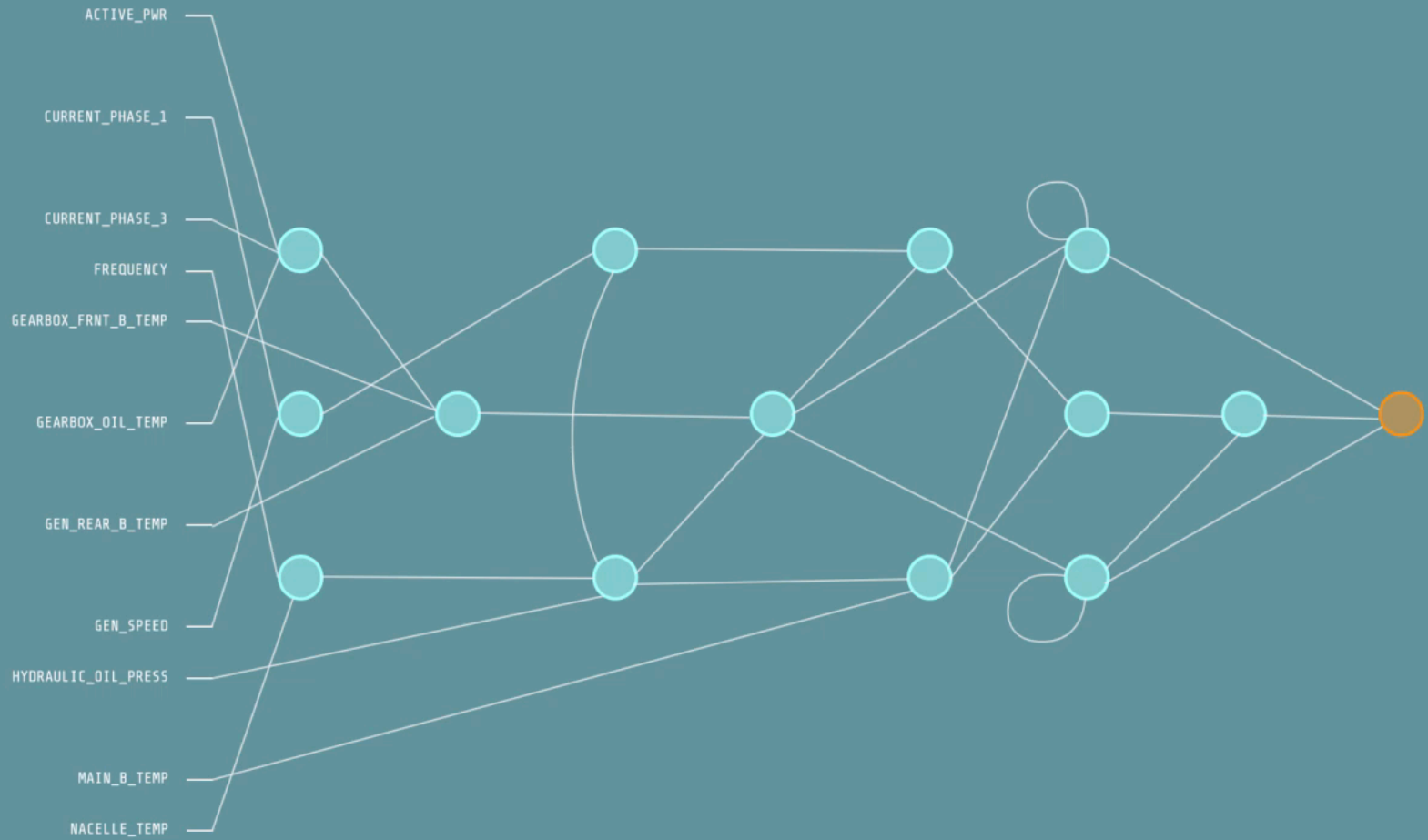




sparkcognition™

DARWIN™

Imputing with AutoML



sparkcognition™

Forecasting Results

DARWIN™

Darwin Forecasting Crenshaw Bridge

Models > Forecast_CrenshawBridge

TRAINING | PREDICTIONS

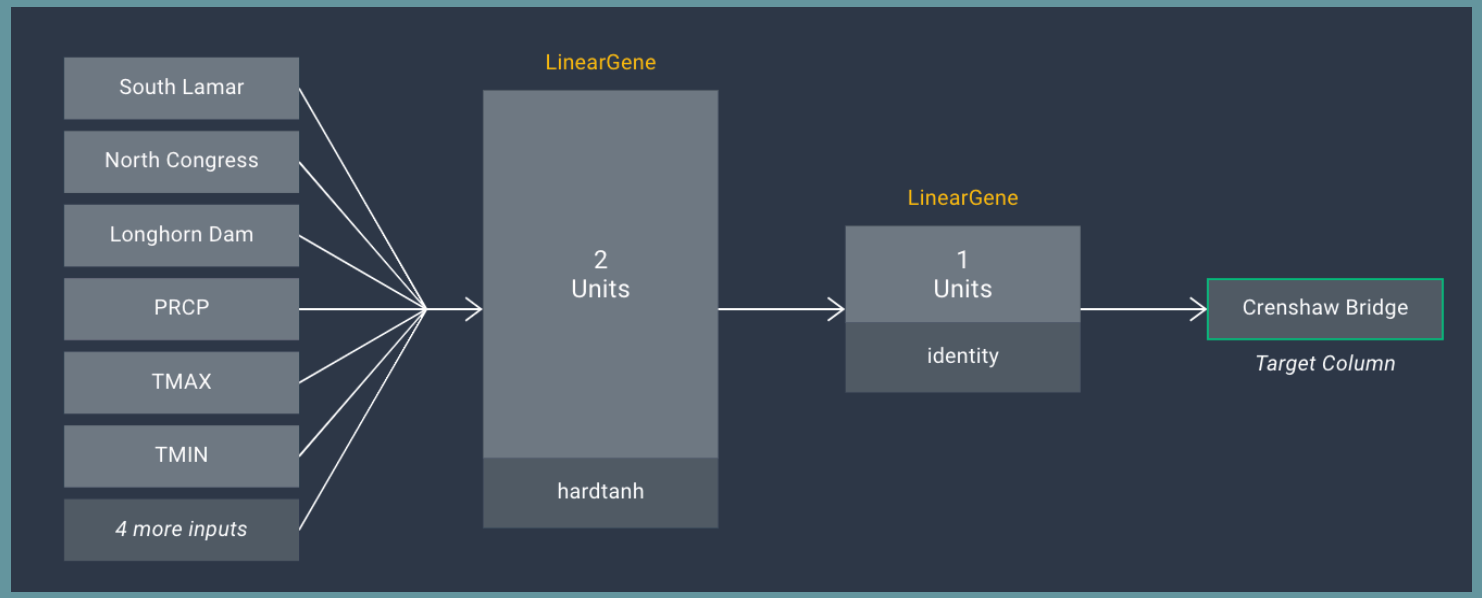
TRAINING TIME	ALGORITHM	MSE	R^2
40 minutes 27 seconds	DeepNet	5383178.7732	0.0402

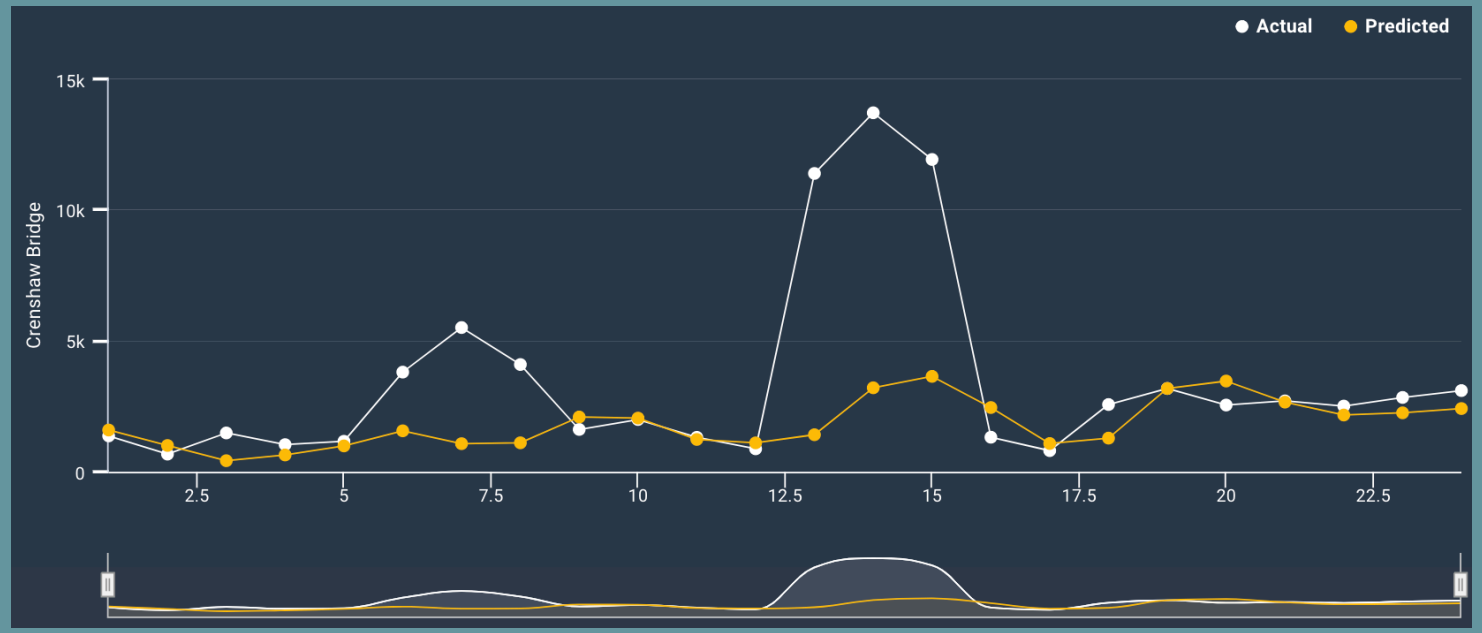
Initial Setup and Training Sessions | Training Results | **Feature Importance**

Feature Importance

Overall

FEATURE	VALUE
woy	0.3636
event_weight	0.1288
South Lamar	0.1126
PRCP	0.0987
Longhorn Dam	0.0902
North Congress	0.0684





Darwin Forecasting South Lamar

Models > Forecast_SouthLamar

TRAINING | PREDICTIONS

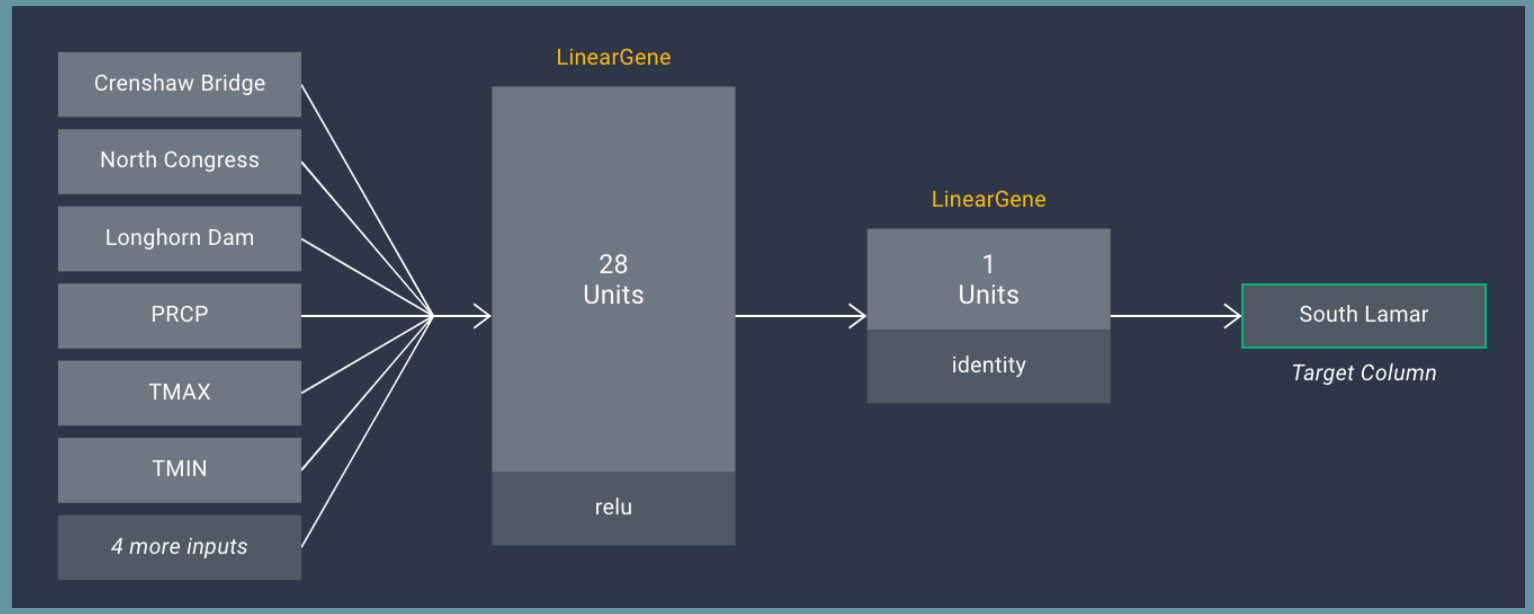
TRAINING TIME	ALGORITHM	MSE	R^2
40 minutes 16 seconds	DeepNet	1655095.9664	0.3867

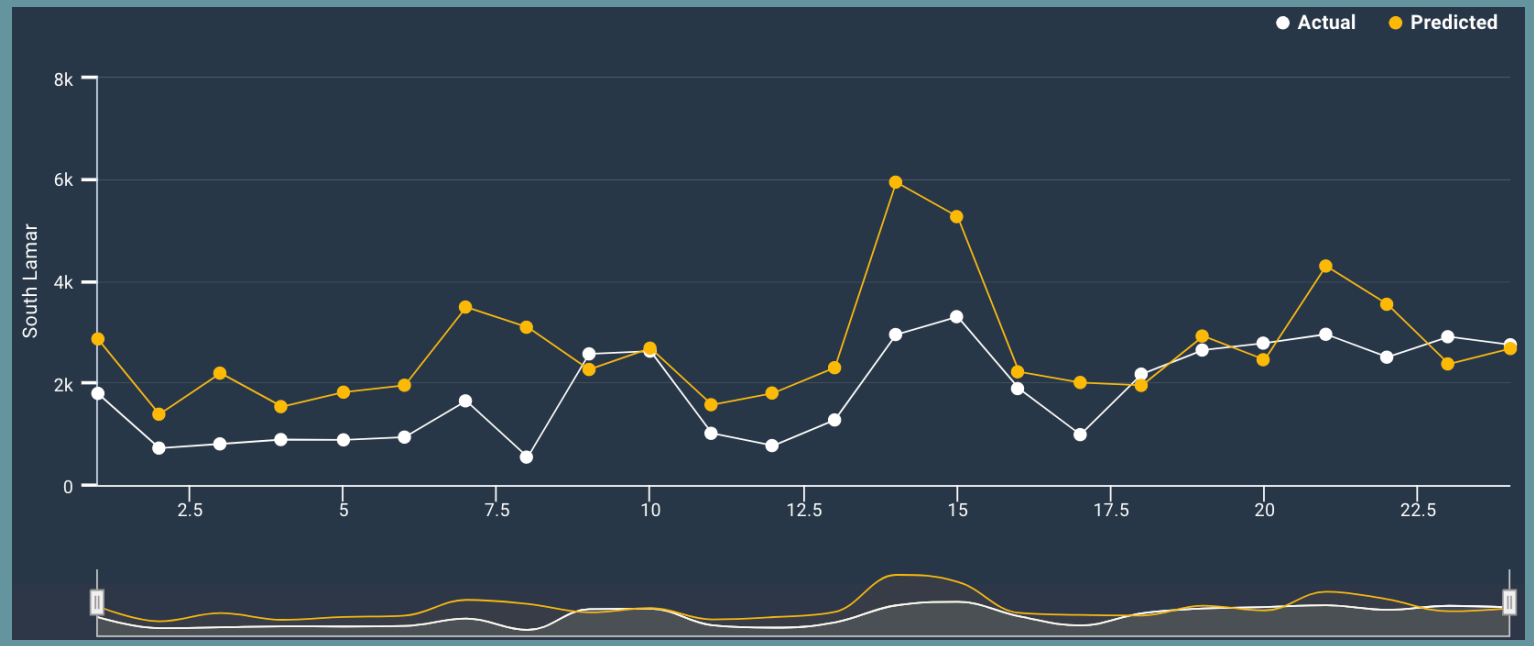
Initial Setup and Training Sessions | Training Results | **Feature Importance**

Feature Importance

Overall

FEATURE	VALUE
woy	0.3572
Longhorn Dam	0.1215
event_weight	0.1029
PRCP	0.0981
Crenshaw Bridge	0.0797
North Congress	0.0765





Darwin Forecasting North Congress

Models > Forecast_NorthCongress

TRAINING | PREDICTIONS

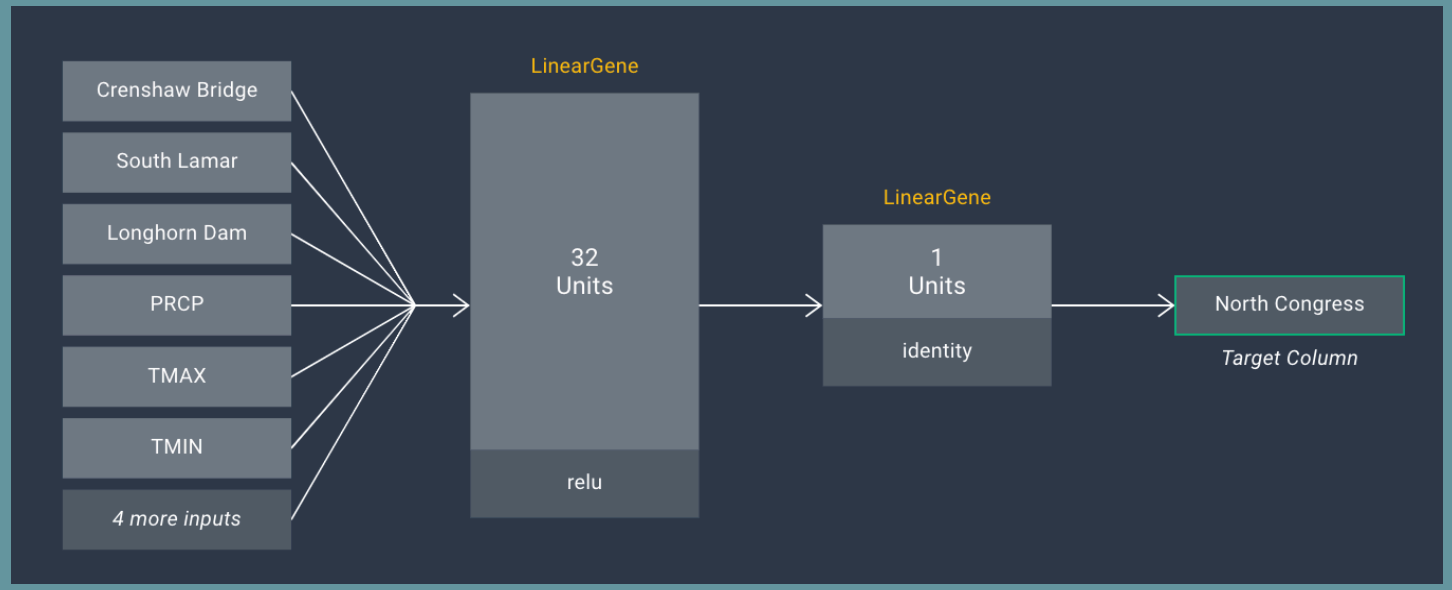
TRAINING TIME	ALGORITHM	MSE	R^2
40 minutes 16 seconds	DeepNet	1281824.4078	0.2891

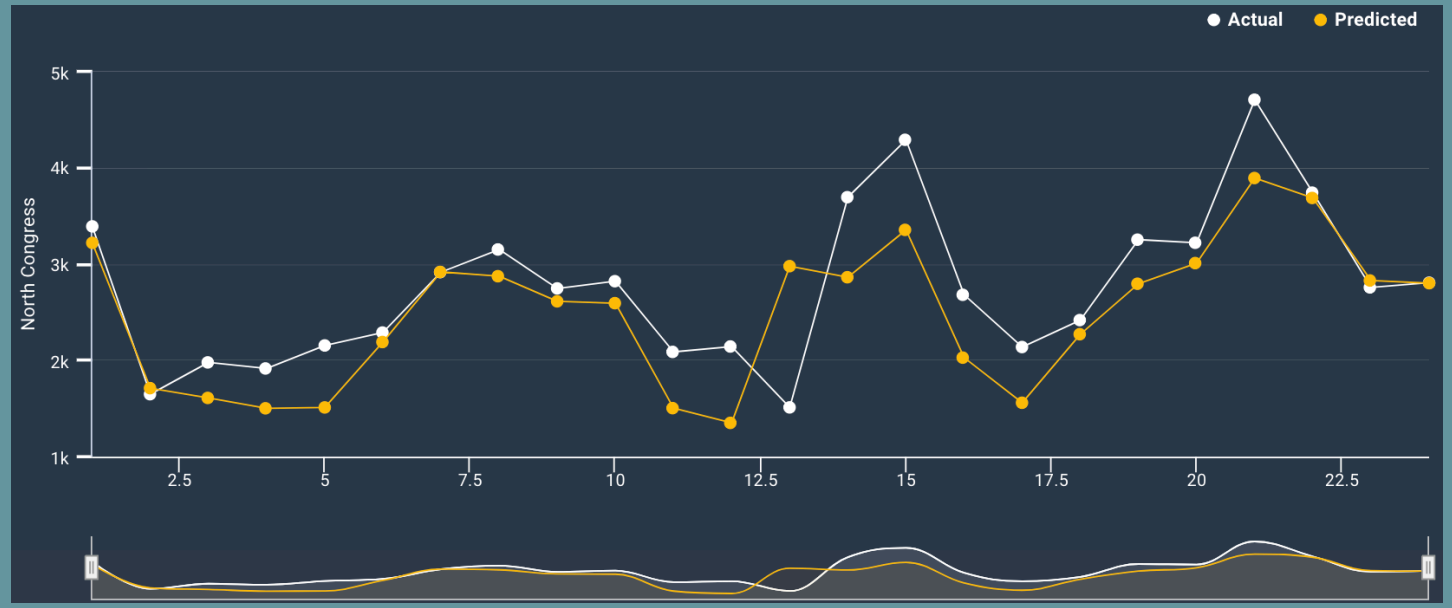
Initial Setup and Training Sessions | Training Results | **Feature Importance**

Feature Importance

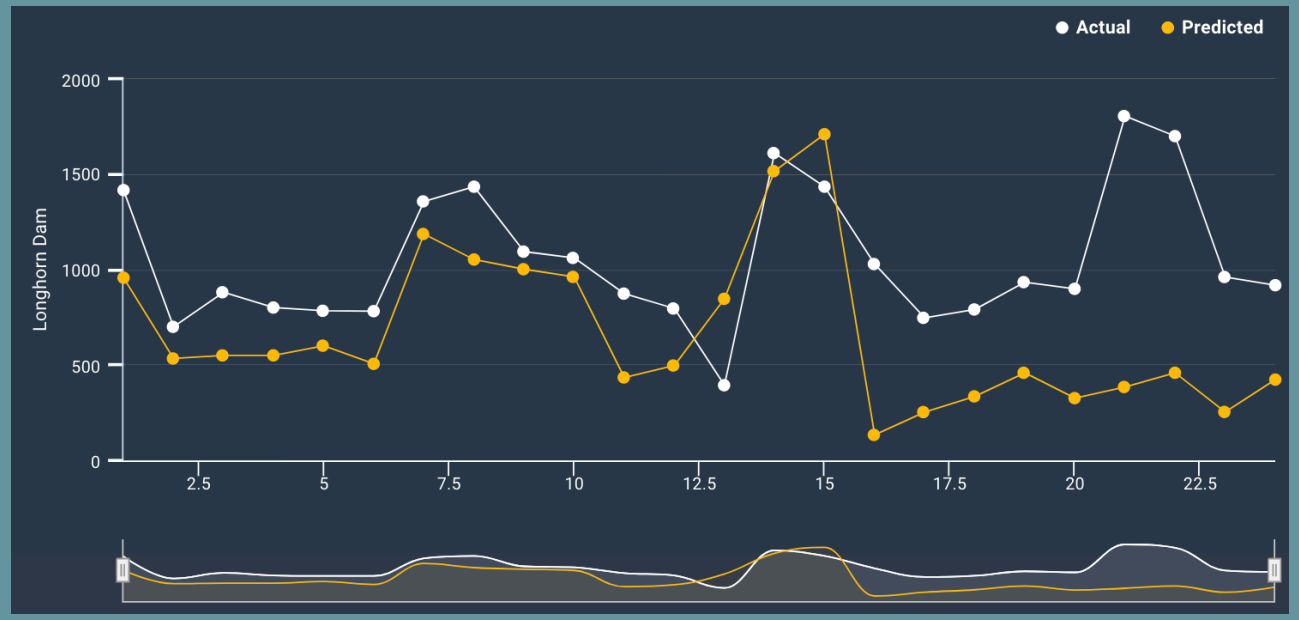
Overall

FEATURE	VALUE
woy	0.3445
event_weight	0.1234
PRCP	0.1044
Longhorn Dam	0.1001
South Lamar	0.0942
Crenshaw Bridge	0.0686





Darwin Forecasting Longhorn Dam



Models > Forecast_LonghornDam

TRAINING | PREDICTIONS

TRAINING TIME	ALGORITHM	MSE	R^2
40 minutes 24 seconds	XGBRegressor	12443.5143	0.9466

Initial Setup and Training Sessions | Training Results | **Feature Importance**

Feature Importance

Overall

FEATURE	VALUE
woy	0.3235
event_weight	0.1233
South Lamar	0.1129
North Congress	0.1053
PRCP	0.0921
Crenshaw Bridge	0.0771
month	0.0435