# PREDICTING TEXAS ELECTRICAL LOADS

Texas WiDS Datathon 2021

Open-Source Excellence Award

Amanda J. Wright

The Challenge:

Forecast the hourly electrical load for each ERCOT region in Texas for the week of June 13, 2021.


http://www.ercot.com/content/wcm/landing_pages/89373/Map-weather.jpg

# Electrical load forecasting is a time-series problem

Features often used in load forecasting:

- electrical load history
- weather
- population growth, number of businesses and homes, price

Common approaches to time-series modeling:

- ARIMA or SARIMA models
- Machine learning regression algorithms
- More advanced algorithms such as deep learning (neural networks), rule-based methods, metaheuristic methods, and hybrid models

# Features came from 3 datasets

➢ Features used for model training also must be available for predicting

Datasets used:

- Historical electrical load data
- Historical weather data and predicted weather data
- Considered COVID-19 data, but rejected

# Feature selection/engineering

Weather features: tempC, windspeedKmph, precipMM, humidity, visibility, pressure, cloudcover, DewPointC, uvindex, weather_class

Date_time features: year, month, day_of_week, hour, weekend

Other: loadyear_lag

Preparation:

- Standard scaled the numerical features
- One-hot encoded the categorical features

# Evaluated many regression models

Linear Regression models:

- Linear Regression
- Stochastic Gradient Descent Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression

Tree-based models:

- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- Extreme Gradient Boosting Regressor

Other models:

- K-Nearest Neighbors Regressor
- Support Vector Machines

Null model:

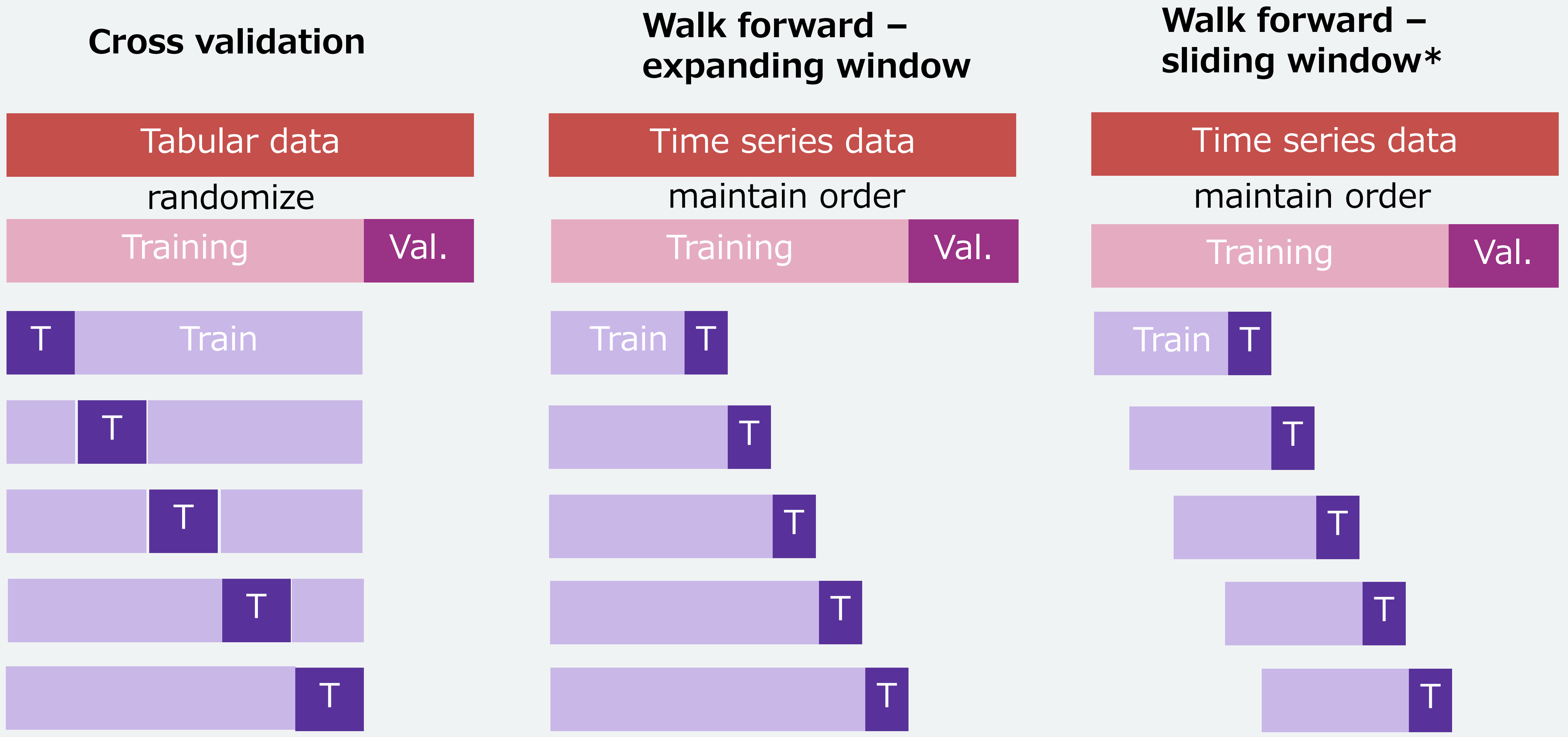
- naïve predictor that takes the last target value (y) of the training set and returns it as every predicted value

Validation metric:

- root mean square error (RMSE)

# Model validation schemes

- Cannot use cross validation as it is not appropriate to use future data to predict past or present values




* Scikit-learn's TimeSeriesSplit

# Random Forest, Gradient Boost, and Xtreme Gradient Boost models performed the best

Results from the ”West” region

| model | sliding_rmse | expand_rmse |
|---|---|---|
| null_naive | 399.3 | 399.2 |
| lin_reg | 8.52e+12 | 7.93e+12 |
| SDG_reg | 184.7 | 178.4 |
| ridge_reg | 157.6 | 164.0 |
| lasso_reg | 163.0 | 160.3 |
| EN_reg | 192.5 | 195.1 |
| DT_reg | 127.6 | 129.2 |
| RF_reg | 95.1 | 98.2 |
| GB_reg | 96.7 | 100.2 |
| KN_reg | 121.8 | 125.1 |
| XGB_reg | 90.8 | 91.3 |

# For 7 out of 8 regions, Xtreme Gradient Boost was the best model

Lowest RMSE = 3
2nd lowest RMSE = 2
3rd lowest RMSE = 1

| | RF_ slide | RF_ expand | GB_ slide | GB_ expand | XGB_ slide | XGB_ expand |
|---|---|---|---|---|---|---|
| West | | | 1 | | 3 | 2 |
| South | | | 1 | | 2 | 3 |
| North_Central | | 1 | | | 2 | 3 |
| Coast | 3 | | | | 2 | 1 |
| Far West | | | 2 | 1 | 3 | |
| South_Central | | 1 | | | 2 | 3 |
| East | | 2 | | | 1 | 3 |
| North | | | | 1 | 2 | 3 |
| Total | 3 | 4 | 4 | 2 | 15 | 18 |

# Model Optimization

Feature importance:

- TempC:  most important feature for 7/8 regions
- loadyear_lag:  most important feature for the Far West region
- ➢ Reran the models after removing features that were consistently in the bottom 30 for the 8 regional models, but model performance declined slightly so I used the larger feature set for the final model optimization.

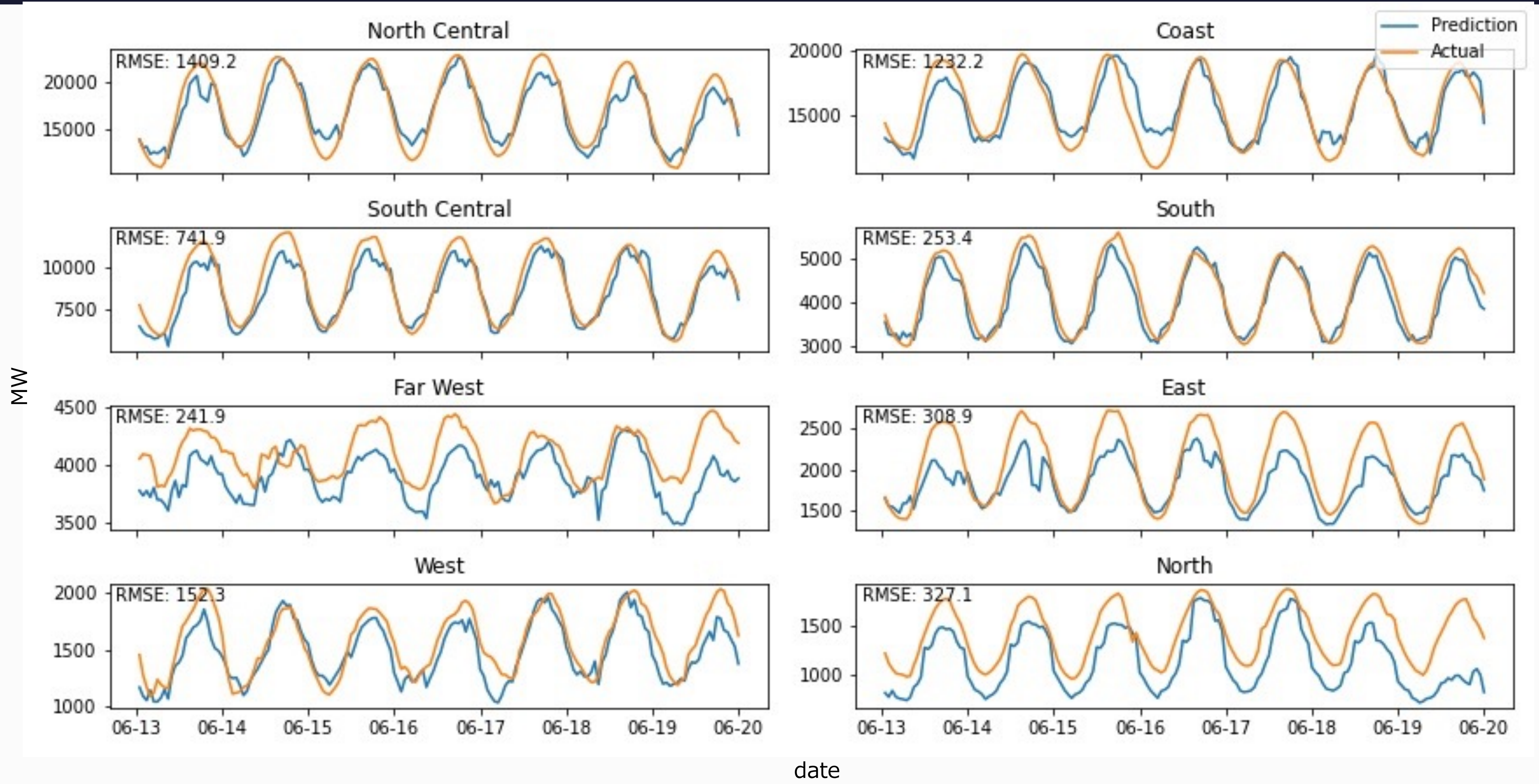
Hyperparameter optimization:

- Used an autoML tuner library, HyperOpt

# The Xtreme Gradient Boost models outperformed the null models for all regions

| | RMSE error | | |
|---|---|---|---|
| | **null_model** | **XGB_reg** | **times_better** |
| **West** | 187.430209 | 72.201993 | 2.595915 |
| **South** | 407.730529 | 191.473766 | 2.129433 |
| **North Central** | 3387.115596 | 1131.942963 | 2.992302 |
| **Coast** | 1246.488977 | 1099.359892 | 1.133832 |
| **Far West** | 192.997791 | 51.324080 | 3.760375 |
| **South Central** | 987.070433 | 586.539587 | 1.682871 |
| **East** | 332.539846 | 163.107765 | 2.038774 |
| **North** | 174.767288 | 59.821539 | 2.921478 |

# Actual electrical load versus my predicted values




Final RMSE was 741.883 (7th place).

# Acknowledgements

Thanks to WiDS Texas and their sponsors for hosting the datathon.

Thanks to Gia Gillis for project discussions.

To view my project:

https://www.kaggle.com/rhinophylla/wids-datathon-2021-forecasting-electrical-loads

# Connect with me! and/or Hire me!

**LinkedIn:**

https://www.linkedin.com/in/amandawrightphd/

**Portfolio:**

https://rhinophylla.github.io