PART 3: SUMMARY REPORT

Xingchen (Estella) Ye
Department of Computer Science
Columbia University
New York, NY 10027, USA
xy2527@columbia.edu

1 PROBLEM FORMULATION

The objective of this analysis is to create an Exploratory Data Analysis (EDA) and forecast model to predict the Real-Time Locational Marginal Price (RTLMP) for ERCOT North hub. The data includes the following hourly time series:

- **RTLoad**: ERCOT real-time hourly actual load.
- **WIND_RTI**: ERCOT real-time hourly wind generation.
- **GENERATION_SOLAR_RT**: ERCOT real-time solar generation.
- **RTLMP**: ERCOT North hub real-time price, which is the target variable dependent on the other three variables.

The dataset includes `DATETIME` and `HOURENDING` in HE (Hour Ending) format. Additionally, there are three peak types associated with the datetime:

- **WEPEAK**: HE7 to HE22 (6:00 AM to 10:00 PM) on weekdays
- **WDPEAK**: HE7 to HE22 (6:00 AM to 10:00 PM) on weekends.
- **OFFPEAK**: Times other than WEPEAK and WDPEAK.

2 ANOMALY DETECTION AND EDA

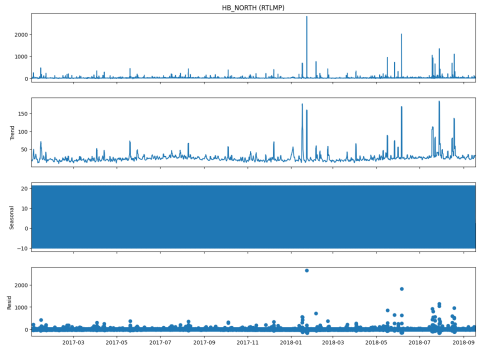
2.1 ANOMALY DETECTION

Anomalies in the dataset are identified by checking if the peak type matched the timestamp. We find 160 data points where non-WEPEAK hours are incorrectly classified as WEPEAK. These anomalies are removed to improve the quality of the EDA and model training.

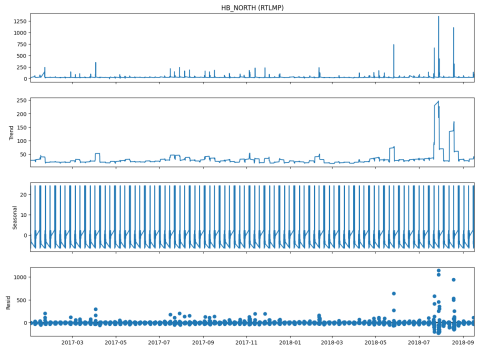
2.2 SEASONAL DECOMPOSITION OF RTLMP BY PEAK TYPE

Seasonal decomposition is performed on the RTLMP data segmented by peak type (WDPEAK, WEPEAK, and OFFPEAK).

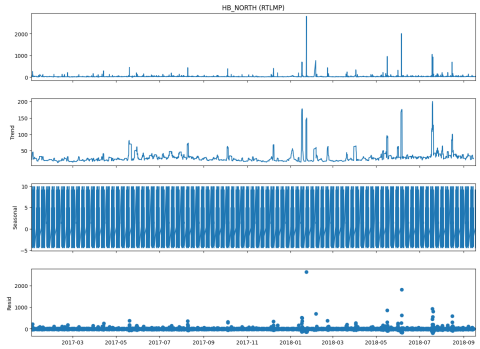
In Figure 1. (a), the seasonal decomposition reveals increasing RTLMP values with infrequent spikes. The general trend remains steady around 100, with peaks showing significantly high RTLMP values to around 3000. When the graph is broken down by the three peak types, it is evident that (b) WEPEAK and (c) WDPEAK account for those peaks. The WEPEAK type shows RTLMP peaks that are higher than those of the other two types, which are around 1200 for WEPEAK and around 400 for WDPEAK.



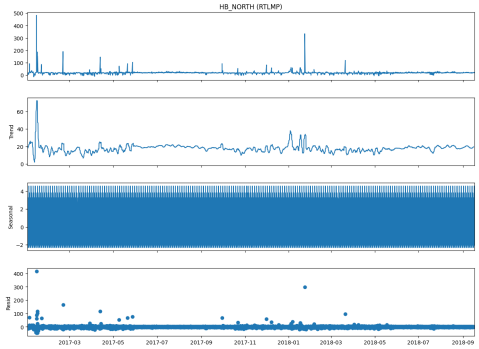
(a) Seasonal Decomposition of RTLMP (All 3 Peaktypes)



(b) Seasonal Decomposition of RTLMP (WEPEAK)



(c) Seasonal Decomposition of RTLMP (WDPEAK)



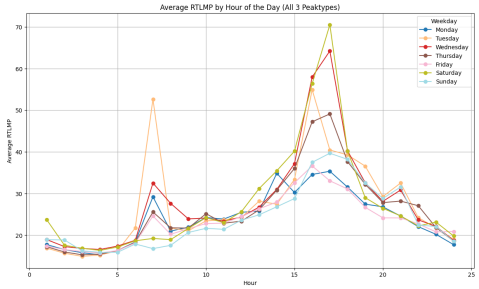
(d) Seasonal Decomposition of RTLMP (OFFPEAK)

Figure 1: Seasonal Decomposition of RTLMP for Different Peaktypes

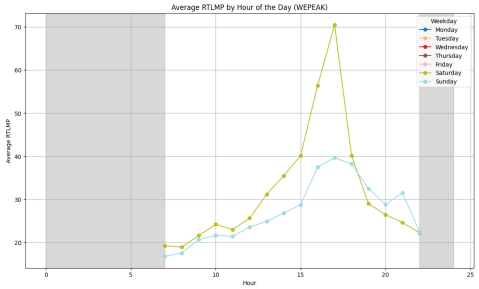
2.3 On Hourly and Daily Basis

Then, we closely examine the trend of RTLMP on hourly and daily basis.

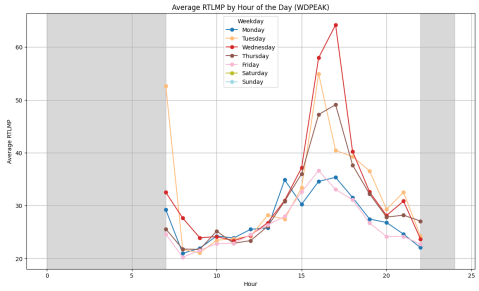
2.3.1 On Hourly Basis



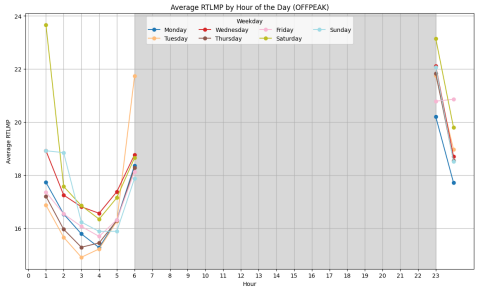
(a) Average RTLMP by Hour of the Day (All 3 Peaktypes)



(b) Average RTLMP by Hour of the Day (WEPEAK)



(c) Average RTLMP by Hour of the Day (WDPEAK)

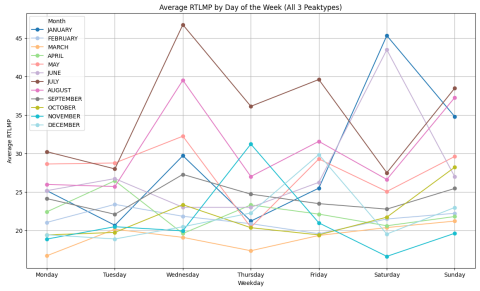


(d) Average RTLMP by Hour of the Day (OFFPEAK)

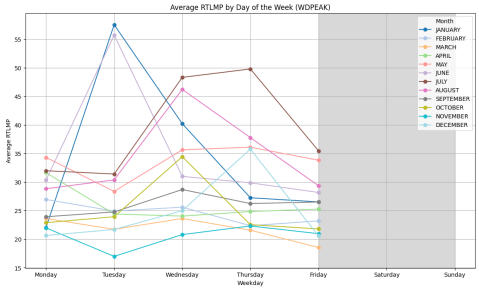
Figure 2: Average RTLMP by Hour of the Day for Different Peaktypes

RTLMP peaks at HE6 and HE16, marking the transition from OFFPEAK to ONPEAK (WEPEAK and WDPEAK). The general hourly trend remains consistent regardless of whether it is a weekday or weekend. The average RTLMP during OFFPEAK is significantly lower than that during WEPEAK and WDPEAK.

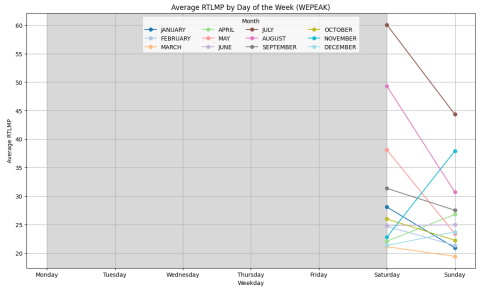
2.3.2 On Daily Basis



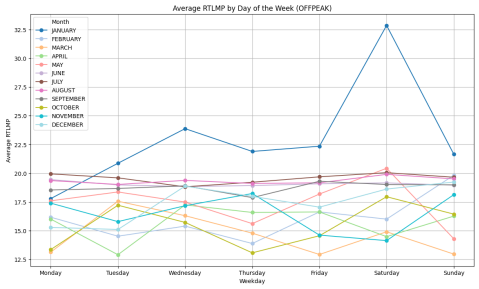
(a) Average RTLMP by Day of the Week (All 3 Peaktypes)



(b) Average RTLMP by Day of the Week (WEPEAK)



(c) Average RTLMP by Day of the Week (WDPEAK)



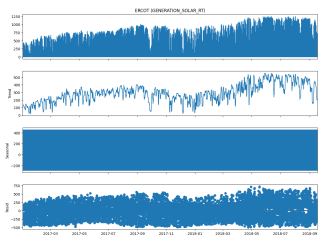
(d) Average RTLMP by Day of the Week (OFFPEAK)

Figure 3: Average RTLMP by Day of the Week for Different Peaktypes

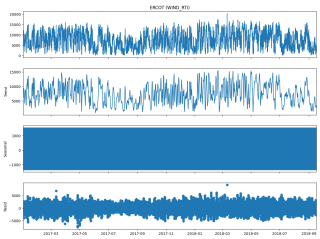
Similarly, we analyze the average RTLMP on a daily basis for different months. There is no evident general trend by day of the week. The peaks may be due to outliers on certain dates, resulting in peaks of RTLMP in a particular day only in some month.

2.4 Seasonal Decomposition of Independent Variables

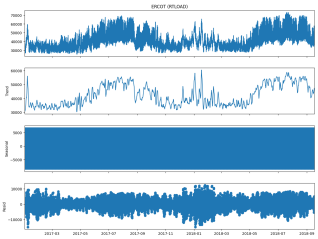
Seasonal decomposition was also performed on the independent variables (GENERATION_SOLAR_RT, WIND_RTI, and RTLOAD). This analysis helps to identify the seasonal patterns and trends in these variables, which are critical for predicting RTLMP.



(a) Seasonal Decomposition of GENERATION_SOLAR_RT



(b) Seasonal Decomposition of WIND_RTI



(c) Seasonal Decomposition of RTLOAD

Figure 4: Seasonal Decomposition of Independent Variables

The observations are as follows.

1. The decomposition (a) reveals an increasing trend in solar generation.
2. The residual component in (b) shows noise, indicating the unpredictable nature of wind generation. There are some notable outliers around the time period where RTLMP peaks

more frequently. We could hypothesize that an unusual event in wind generation might account for the price peak.

3. The residual component in (c) shows substantial variability around January 2018, which might be due to the same event that caused the outlier in the residual component of WIND_RTI.

3 FORECASTING

3.1 PREPROCESSING

1. **Data Loading**
   - The dataset is loaded from the `timeseries_data.xlsx` file. Convert `DATETIME` and `MARKETDAY` columns to datetime format.
2. **Anomalies Correction**
   - Correct anomalies in the `PEAKTYPE` column for records labeled as `WEPEAK` on weekdays, changing them to `WDPEAK`.
3. **Feature Engineering**
   - **Lag Features**: Shift the columns `ERCOT (WIND_RTI)`, `ERCOT (GENERATION_SOLAR_RT)`, and `ERCOT (RTLOAD)` by one time step to create lag features. This is because when forecasting the price at time step t, we only have access to the data until time step $t-1$.
   - **Rolling Statistics**: Calculate rolling mean and standard deviation with a window of 24 hours.
   - **Time-Based Features**: Create time-based features, such as the day of the week, to capture the cyclical patterns demonstrated in EDA.
4. **One-Hot Encoding**
   - **One-Hot Encoding**: Apply one-hot encoding to categorical features, including `PEAKTYPE`, `MONTH`, `WEEKDAY`, `YEAR`, and `HOURENDING`.
5. **Normalization**
   - **Normalization**: Use `MinMaxScaler` to scale the features and target variable to a range of 0 to 1 for better model training.

3.2 MODEL CONSTRUCTION

- **Data Preparation**
  - **Train/Val/Test Split**: Split the data into training and validation (80%) and test (20%) sets, and further split the training and validation data into training (80%) and validation (20%) sets.
  - **Sequence Creation**: Create sequences of data with a specified time step (24 hours), as the data exhibit daily cycles.
- **LSTM (Long Short Term Memory)** The LSTM model captures long-term dependencies and temporal patterns in sequential data, making it suitable for predicting time-series data. The model architecture includes 4 LSTM layers, each with 200 units, with a dropout rate of 0.3 applied after each layer. A dense layer with a single unit is used to output the predicted RTLMP. The Adam optimizer is employed, and MAE (mean absolute error) is chosen as the loss function because it does not place heavy weights on extreme values, such as peak values. The model is trained for 50 epochs with a batch size of 32.

  As suggested by the literature, the evaluation metrics used are MAPE (mean absolute percentage error) and RMSE (root mean square error).

3.3 RESULTS

The LSTM model's performance on the test set shows an RMSE of 73.000 and a MAPE of 0.358. The figures below compare the predicted and actual values. While the model struggles to capture infrequent peaks, it successfully identifies the overall upward trend and cyclical patterns (Figure 7).

We also implement the LSTM model with various configurations: LSTM units set to 100, 200, and 500; batch sizes of 1 and 32; and sequence lengths of 1, 24, and 24x7. Additionally, we test an LSTM model with an attention block. The LSTM configuration described in section 3.2 yields the best performance, achieving the lowest RMSE and MAPE scores.

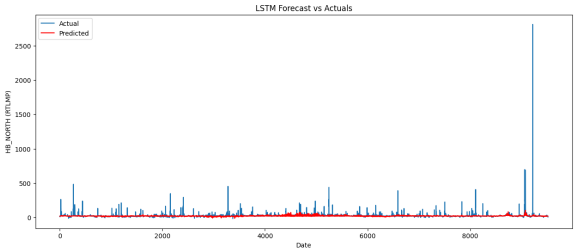


Figure 5: LSTM Forecast vs Actual on Training Set

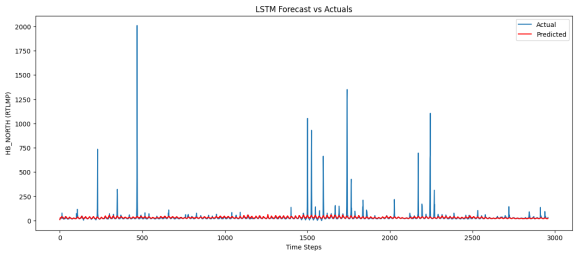


Figure 6: LSTM Forecast vs Actual on Test Set

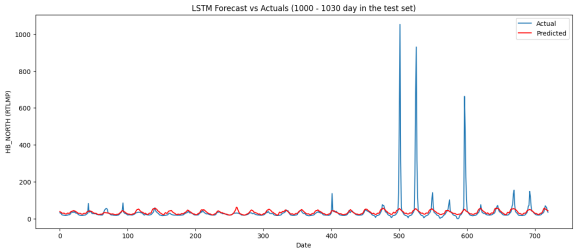


Figure 7: LSTM Forecast vs Actual on Test Set in 30-day Period

We also implement Random Forest and XGBoost models for this prediction task. The feature engineering process remains unchanged, expect that validation stage is not used for these models. Both models are configured with 1,000 estimators.

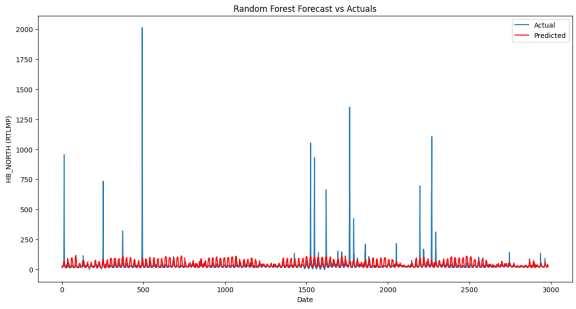


Figure 8: Random Forest Forecast vs Actual on Test Set

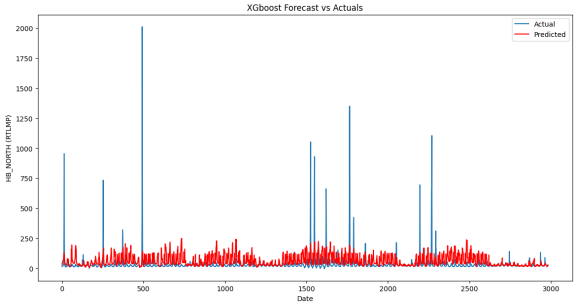


Figure 9: XGBoost Forecast vs Actual on Test Set

Table 1: Comparison of Models

Model	**RMSE**	**MAPE**
LSTM	73.000	0.358
RF	75.496	0.734
XGBoost	81.229	1.532

It is evident that the LSTM model outperforms the regression models, highlighting LSTM's effectiveness in capturing the cyclic behavior in this time-series prediction task.

3.4 VISION

For future work, additional feature engineering can be incorporated into this task. Specifically, identifying peaks as outliers would be productive in predicting peak prices. Separate predictions can be made for the cyclic daily fluctuations and the peak prices. The former is a time-series prediction task, while the latter, due to the nature of market behaviors, could be approached as a regression task that depends on many factors, such as market data, events, and other prices. Incorporating additional categories of information would enable us to increase the model's complexity and improve forecasting accuracy.