Forecasting US Electricity Prices Using Time Series Models

Stanford STATS 207 Project

Ngoc Vo
Department of Statistics
Stanford University
hnngocvo@stanford.edu

Kevin Liu
Department of Statistics
Stanford University
liuxk@stanford.edu

Yihan Wang
Department of Statistics
Stanford University
yihan330@stanford.edu

Abstract

Electricity price forecasting plays a critical role in the decision-making processes of deregulated energy markets, where accurate predictions can mitigate risks arising from volatile market conditions. This study investigates the efficacy of classical and deep learning time series models for forecasting U.S. electricity prices using a dataset spanning 2001 to 2024. We compare the performances of a baseline seasonal autoregressive integrated moving average (SARIMA) model, a modified SARIMAX model that incorporates temperature as an exogenous variable, and a Long Short-Term Memory (LSTM) network, emphasizing their ability to capture price trends, seasonality, and complex temporal dependencies. While the SARIMAX model exhibits better forecasting (especially in the short run) and model quality than the SARIMA model, the LSTM model outperforms both in terms of predictive accuracy, highlighting its capability to handle non-linear relationships and long-term dependencies.

1 Introduction

Time series analysis of electricity prices is essential for navigating the complexities of today's competitive and deregulated energy market. Accurate electricity price forecasting has become crucial to decision-making in the energy industry, enabling stakeholders to make informed choices in a volatile environment shaped by diverse factors, including weather patterns, evolving energy policies, the rise of renewable energy sources, and increasing competition. Despite its significance, electricity price forecasting remains a challenging area, with considerable opportunities for enhancing model accuracy and predictive power.

Our research adopts a comparative approach to further this body of knowledge, applying both classical time series models (SARIMA(X)) and deep learning models (LSTM) within the context of the U.S. electricity market. By exploring these newer and more adaptive techniques, we aim to improve forecasting accuracy, capturing complex price dynamics, volatility, seasonality, and trends that traditional models may find challenging to model effectively.

2 Related Work

A review of the literature reveals that electricity price forecasting is primarily approached through two methods: time series models and simulation-based models (Singh and Mohanty, 2015). Amongst these, time series models are more commonly applied, particularly for day-ahead forecasting. In defining the scope of our study, we will focus on time series forecasting.

One of the critical challenges in forecasting electricity prices, as previously noted, is the high degree of price variability over time. To address this,Wang et al. (2022) proposed a SARIMAX model, incorporating exogenous variables such as gas, coal, and carbon prices to enhance predictions of electricity price fluctuations. Building on this, Lehna et al. (2022) conducted an analysis of German electricity prices, comparing both classical models (like SARIMAX) and advanced deep learning approaches (such as LSTM, CNN, and a two-stage VAR model). Their study included exogenous factors like consumer load, fuel prices, carbon emissions, and weather data to forecast prices across three time horizons, ultimately finding that hybrid models combining multiple approaches often outperformed single models. Amongst the single models, LSTM broadly performed the best; however, the VAR model performs exceptionally well with short-term predictions. Similarly, Wagner et al. (2022) examined deep learning methods, finding that embedding seasonal layers into neural network models led to more accurate results, even outperforming LSTM models under certain conditions.

We will follow a time series approach similar to that of Lehna et al. (2022), constructing SARIMA(X) and LSTM models but applying them to US electricity prices from 2001 to 2024. We place less emphasis on exogenous variables (but discuss this later as a next step), although we do incorporate temperature as an exogenous variable in the SARIMAX model.

3 Data

We use the U.S. electricity price dataset from Kaggle King (2024), which spans from 2001 to 2024. This dataset provides monthly average electricity prices per kilowatt-hour (kWh) for residential, commercial, and industrial sectors, with state-level disaggregation. Additionally,

it includes total electricity sales, revenue, and customer counts. This dataset is well-suited for analyzing trends, seasonality, and price dynamics in the U.S. electricity market.

To construct a SARIMAX model, we also consider the average monthly temperature in the U.S. as an exogenous variable, since studies have shown that weather effects have a significant impact on electricity demand, e.g., Hong et al. (2013). We use monthly temperature data collected by the National Centers for Environmental Information (NCEI) tem (2024).

Figure 1 shows a time series plot of the average electricity prices (blue) and temperatures (red) from 2001 to 2024, highlighting seasonal fluctuations and long-term trends (for price). Note that electricity prices experience a seasonal peak during the summer months when temperature is highest (and sometimes a smaller peak during the winter months), indicating some relationship between price and temperature.

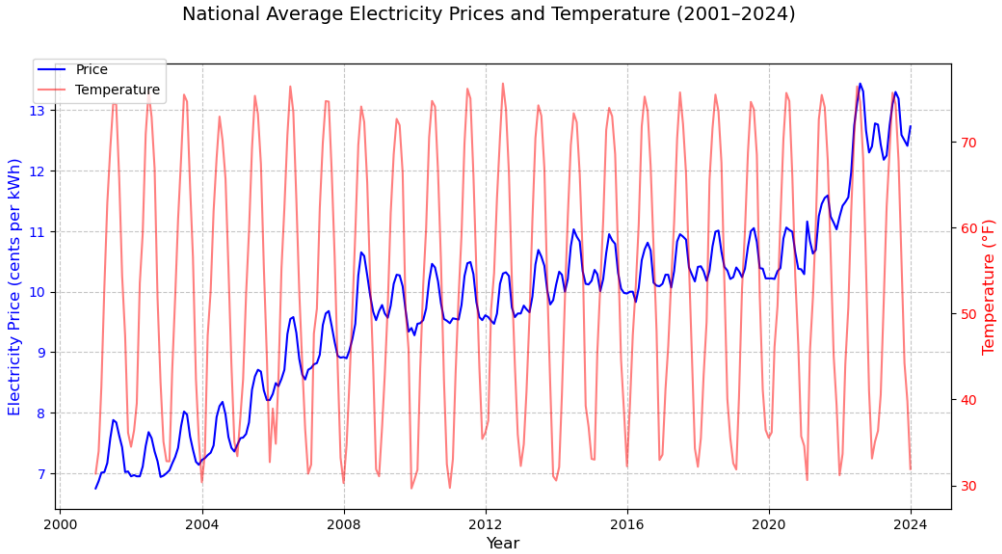
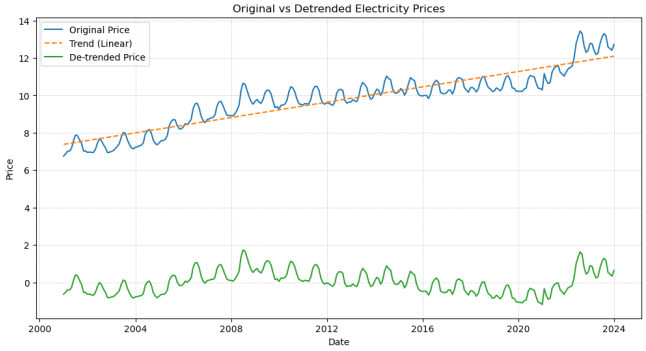
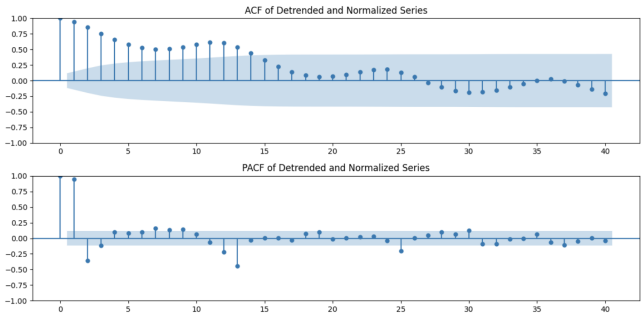


Figure 1: Time series plot of U.S. electricity prices (2001–2024) and average temperatures, showing seasonal patterns and price trends.

Given the presence of a clear trend in the original data, we applied two approaches to ensure the data satisfies time-series modeling assumptions. First, we detrended the data by fitting and removing a linear trend. Second, we achieved stationarity by applying first differencing. Following these adjustments, we conducted diagnostic tests. The augmented Dickey-Fuller test verified stationarity, autocorrelation plots were used to detect residual patterns, the Ljung-Box-Pierce test confirmed residual independence, and Normal Q-Q plots visually assessed residual normality. These steps ensured the data met model assumptions.



(a) Original and detrended data with a linear trend removed.



(b) ACF and PACF plots of the detrended data.

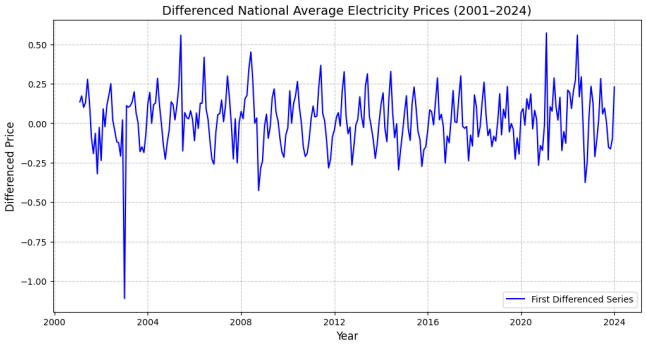
Figure 2: Visualization of the detrended data and its autocorrelation structure.

4 Methods

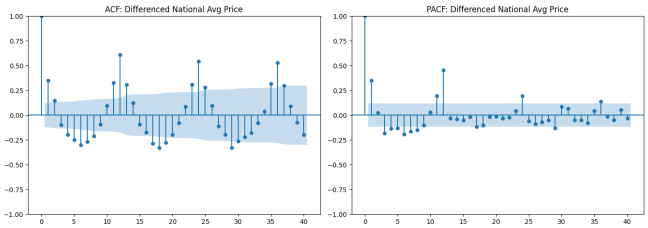
4.1 SARIMA(X) Models

We used a seasonal autoregressive integrated moving average (SARIMA) model as our baseline model. In addition to the autoregressive (AR) and moving average (MA) terms offered by the standard ARMA model, the SARIMA model allows us to capture the seasonal (S) and non-stationary (I) aspects of our price data.

To determine appropriate parameter ranges, we analyzed the ACF and PACF plots. Using these ranges, we found the optimal parametrization through grid-search cross-validation, using the average root mean-squared error (RMSE) as our performance metric. In addition



(a) First-differenced data.



(b) ACF and PACF plots of the first-differenced data.

Figure 3: Visualization of the first-differenced data and its autocorrelation structure.

to this cross-validation RMSE score (which measures predictive power), we also evaluated the SARIMA model through its Akaike information criterion (AIC), which considers both goodness of fit and model simplicity. Finally, we conducted a Ljung-Box test on the residuals to determine whether the residuals had any significant autocorrelations.

We originally trained a SARIMA model on the untransformed price data but found that it failed the Ljung-Box test. To improve from this, we trained the model on the detrended data (obtained by fitting a linear trend and removing it from the original data). The resulting SARIMA model, passed the Ljung-Box test and was used as our official baseline model.

4.1.1 SARIMAX: Temperature as an Exogenous Variable

To explore the potential of exogenous variables within this SARIMA framework, we modified our original SARIMA model to include the national monthly average temperature as an exogenous variable, resulting in a SARIMAX model. As in the original case, we chose to train the SARIMAX model on the detrended data, since we found that it failed the Ljung-Box test when trained on the untransformed data.

The SARIMA hyperparameters $(p, d, q) \times (P, D, Q)_s$ consist of:

- The ARIMA hyperparameters (p, d, q), corresponding to the autoregressive, differencing, and moving average orders, respectively
- The seasonal hyperparameters $(P, D, Q)_s$ for seasonality of period s, which are analogous to the ARIMA hyperparameters

In general, the SARIMAX model extends the SARIMA model by incorporating exogenous variables X_t. Given observed data Y_t, noise terms Z_t, and exogenous variables $X_t^{(i)}$, the SARIMAX model is given by Equation 1.

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d Y_t = \delta + \Theta(B^s)\theta(B)Z_t + \sum_{i=1}^{n}\beta_i X_t^{(i)} \tag{1}$$

where $\nabla^d = (1-B)^d$ is the differencing operator, $\nabla_s^D = (1-B^s)^D$ is the seasonal differencing operator (for period s), and $\beta_1, \ldots, \beta_n$ are the coefficients for the endogenous variables. In this case, we have $n = 1$ (only temperature as an endogenous variable).
The equation for the SARIMA model is similar to Equation 1. The SARIMAX model captures both the seasonal patterns in the time series and the influence of exogenous variables.

4.2 LSTM

LSTM (Long Short-Term Memory) networks are an extension of Recurrent Neural Networks (RNNs) designed to address the limitations of RNNs in handling long-term dependencies. Traditional RNNs struggle to retain information over long sequences due to vanishing or exploding gradients during training. LSTMs overcome this issue by introducing a cell state as a direct information pathway, enabling better preservation of long-term dependencies.

An LSTM module is composed of four interacting neural networks that collectively manage the flow of information. As illustrated in Figure 4, the LSTM updates the cell state C_t and hidden state h_t using the previous cell state C_{t-1}, previous hidden state h_{t-1}, and the current input x_t. The architecture of an LSTM relies on three primary gates: forget, input, and output gates, which regulate the addition, removal, and modification of information in the cell state.

- **Forget Gate**: The forget gate determines which information from the previous cell state (C_{t-1}) should be retained or discarded. It generates a vector f_t, with values ranging from 0 (forget) to 1 (retain), based on the sigmoid activation function. The filtered cell state is calculated by performing element-wise multiplication of f_t with C_{t-1}, as described in Equation 2.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

- **Input Gate**: The input gate determines which new information should be added to the cell state. It computes candidate values $\tilde{C}_t$ using the $\tanh$ activation function (Equation 3) and uses a sigmoid function to identify the relevant parts (Equation 4). These values are combined to update the cell state, as shown in Equation 5.

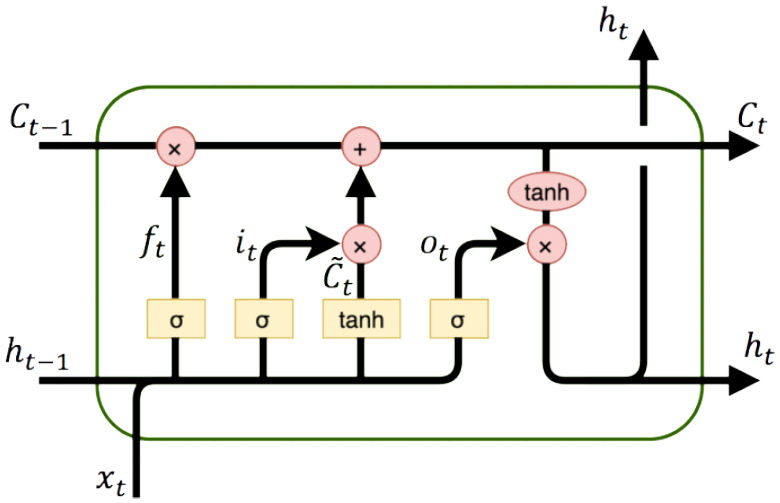


Figure 4: LSTM Module Architecture (adapted from Hedengren (n.d.)).

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{5}$$

- **Output Gate**: The output gate determines the proportion of the cell state (C_t) that contributes to the hidden state (h_t). This process is controlled by the sigmoid activation function (Equation 6), and the final hidden state is computed as shown in Equation 7.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \odot \tanh(C_t) \tag{7}$$

These mechanisms allow LSTMs to effectively manage long-term dependencies by systematically preserving, discarding, or updating information in the cell state. This makes LSTMs particularly well-suited for tasks such as time-series forecasting, including electricity price prediction, where long-term patterns and dependencies are critical for accurate predictions.

5 Experiments / Results / Discussion

To compare the performance of the SARIMA(X) models with the LSTM model, we compute the average root mean squared error (RMSE) across 5-fold cross-validation to tune and evaluate the best model.

5.1 SARIMA(X) Models

5.1.1 Hyperparameter Tuning

For the SARIMA(X) models, given the yearly seasonality observed in the price data, we set $s = 12$ and used grid-search cross-validation to optimize the remaining hyperparameters. The resulting hyperparameter sets for both the SARIMA and SARIMAX models are listed below in the general model summary (see Table 1).

We see that the best SARIMA model does not incorporate differencing and emphasizes the seasonal parameters. On the other hand, the best SARIMAX model incorporates first-order differencing but not any ARMA parameters. This suggests that the inclusion of an exogenous variable does influence the hyperparameter tuning process to some extent.

5.1.2 Results

After fitting the SARIMA and SARIMAX models to the data, we evaluated their forecasting ability by computing their average root mean squared error (RMSE) during 5-fold cross-validation (CV) on the original price data (adding the trend back to the models' predictions). We also evaluated the models' quality by comparing their Akaike information criteria (AICs). These results, along with the optimal hyperparameter setup, are summarized in Table 1.

We conclude from these metrics that the SARIMAX model slightly outperforms the SARIMA model, since the former achieves both a lower RMSE and a more negative AIC. However, we also recognize that this handful of metrics may not tell the whole story behind these models' performances, which motivates us to look for additional approaches to evaluation, one of which we will now discuss.

Model	*SARIMA*	*SARIMAX*
Hyperparameters	$(2,0,0) \times (1,1,2)_{12}$	$(0,1,0) \times (0,1,2)_{12}$
Average CV score (RMSE)	1.217	0.933
AIC	-320.31	-323.07

Table 1: Comparison of the results from the SARIMA and SARIMAX models.

5.1.3 Historical Analysis

Besides comparing individual metrics, it is also worth visually examining the nature of each model's forecast, especially during the period around 2022, when the average price increased more than usual, deviating from the linear trend (see Figure 2a). For example, we fit both models to the data during 2001-2016 and compared their forecasts for 2016-2024, shown in Figure 5.

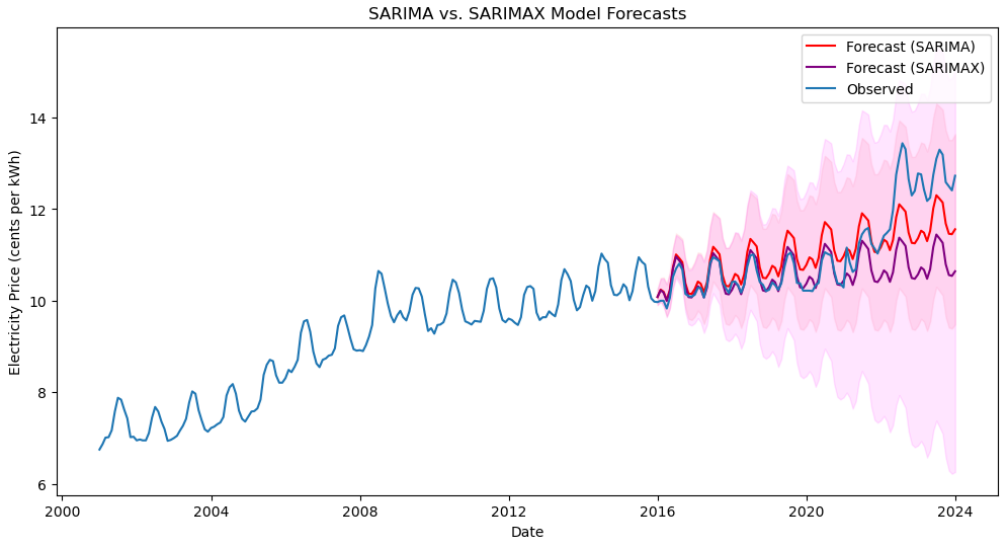


Figure 5: 2016-2024 forecasts given by the SARIMA (red) and SARIMAX (purple) models.

Both models struggle to capture the 2022 increase, although the SARIMA model (despite having worse metrics, as noted above) does slightly better in this task. It resembles the estimated linear trend (see Figure 2a), so it slightly overestimates the 2016-2022 period and underestimates the 2022-2024 period. On the other hand, the SARIMAX model is more accurate for the 2016-2022 period (short-term) at the expense of underperforming after the 2022 increase. It also exhibits more uncertainty (i.e., wider confidence intervals) than the SARIMA model.

These findings suggest that the SARIMAX model may be powerful for forecasting in the short run or in the absence of deviations from the linear trend, but otherwise it may not be as robust as the SARIMA model.

5.2 LSTM

5.2.1 Hyperparameter Tuning

The hyperparameter tuning process for the LSTM model was conducted using grid search combined with 5-fold cross-validation to identify the optimal combination of parameters. The following hyperparameters were tuned, and their potential impact on RMSE is discussed below:

- **Units**: The number of LSTM units in each layer determines the model's capacity to learn complex temporal patterns. A lower number of units may result in underfitting, leading to higher RMSE, while an excessively large number of units can overfit the training data, negatively impacting generalization and increasing RMSE on validation sets.
- **Dropout Rate**: Dropout is applied to prevent overfitting by randomly deactivating a fraction of neurons during training. A low dropout rate may fail to regularize the model, increasing the likelihood of overfitting and leading to higher RMSE on validation data. Conversely, a high dropout rate can result in underfitting, as the model may struggle to capture important patterns.
- **Learning Rate**: The learning rate controls how quickly the model updates its weights during training. A learning rate that is too high may cause the model to converge to suboptimal solutions, leading to higher RMSE. On the other hand, a very low learning rate can slow convergence, potentially leaving the model undertrained within the given number of epochs.
- **Batch Size**: The batch size influences the stability and efficiency of the training process. Smaller batch sizes can lead to noisier gradient updates, potentially increasing RMSE, while larger batch sizes may smooth gradients but could miss finer details in the data, also leading to suboptimal performance.

- **Epochs**: The number of epochs determines how long the model is trained. An insufficient number of epochs can lead to underfitting, with higher RMSE, as the model does not have enough time to learn the underlying patterns. However, training for too many epochs may result in overfitting, where the model performs well on the training data but poorly on validation data, increasing RMSE.

By systematically exploring combinations of these hyperparameters and evaluating their impact on RMSE, the tuning process aims to strike a balance between underfitting and overfitting, ensuring optimal predictive performance for the LSTM model. We include the range of RMSE from each parameter configuration in Figure 6.

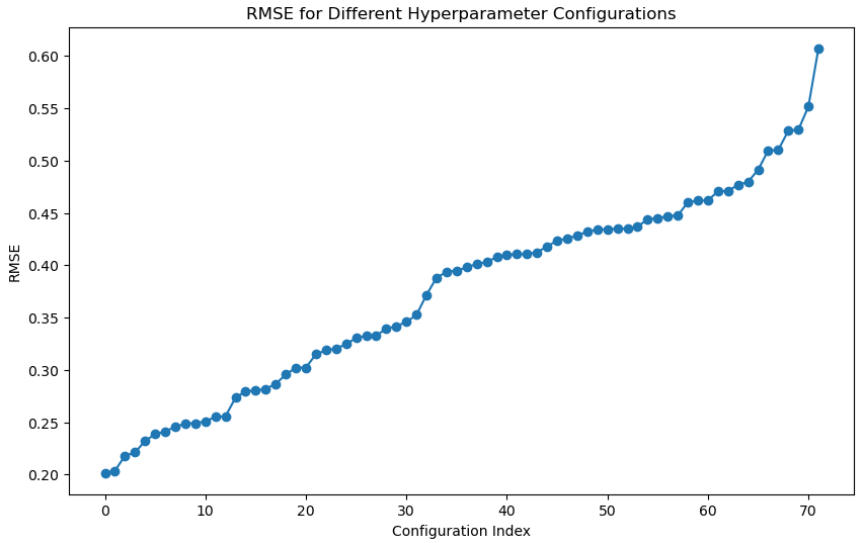


Figure 6: RMSE for Different Hyperparameter Configurations

5.2.2 Results

The LSTM model was evaluated using 5-fold cross-validation, and the results demonstrated its effectiveness in forecasting electricity prices.

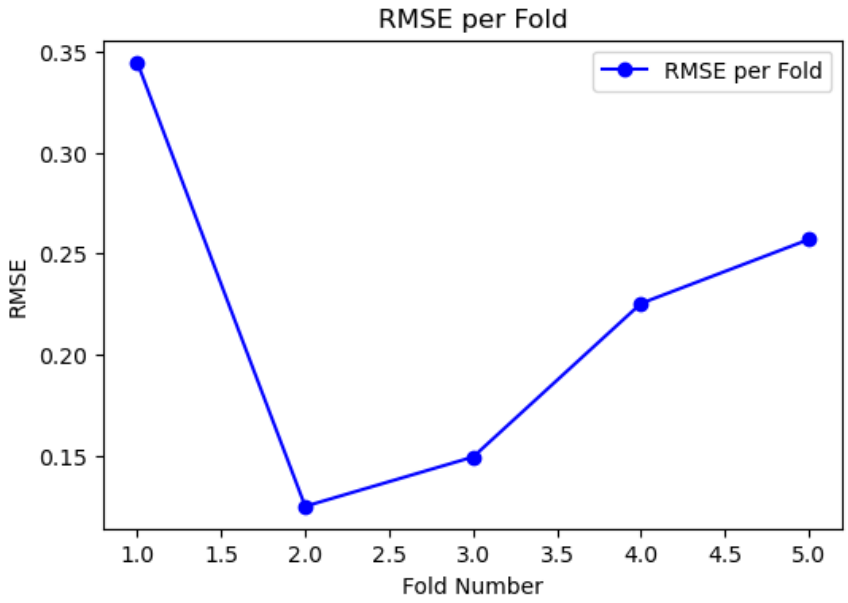


Figure 7: RMSE per Cross-Validationn Fold

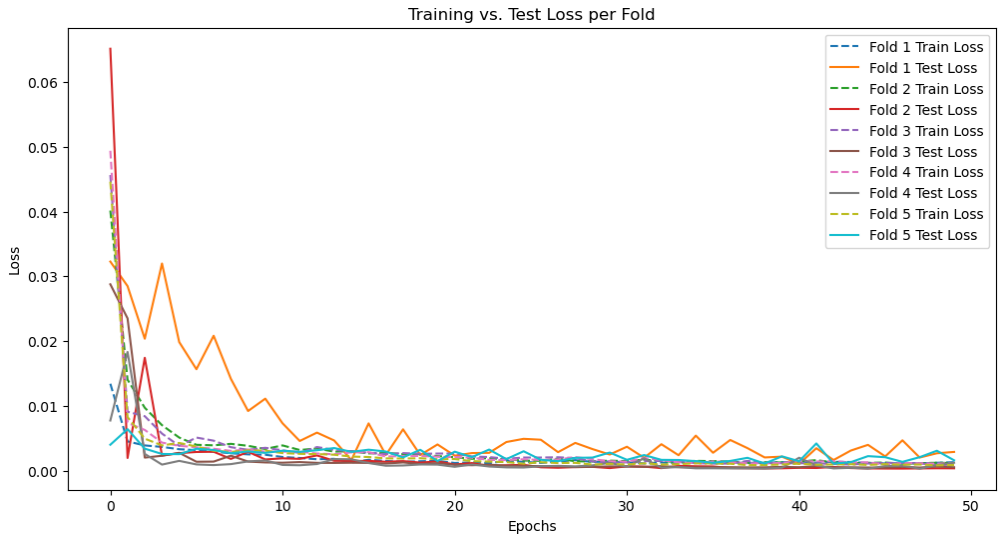
Figure 7 shows the RMSE of each fold. The average root mean square error (RMSE) across the folds was calculated to be **0.2203**, indicating a high level of accuracy in capturing the temporal patterns of the data. Notably, this RMSE indicates a large improvement over the RMSEs achieved by the SARIMA and SARIMAX models (see Table 1).

The best hyperparameters, identified through grid search, are as follows:

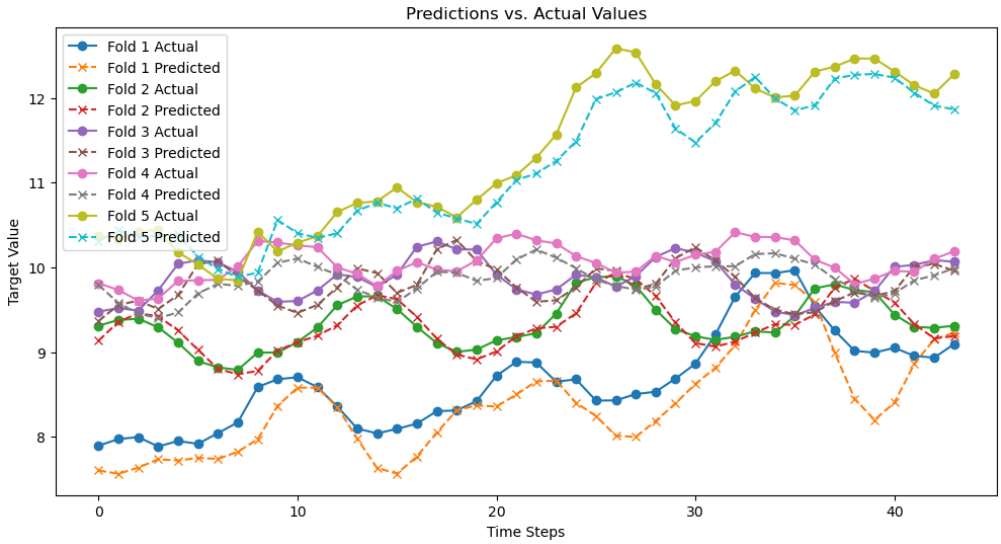
- **Number of LSTM Units:** 64

- **Dropout Rate:** 0.1
- **Learning Rate:** 0.01
- **Batch Size:** 16
- **Number of Epochs:** 50

The inclusion of a dropout rate as a regularization technique was crucial in preventing overfitting of the model. This ensured that the model did not rely excessively on specific patterns in the training data, improving its generalization to unseen test data. The selected hyperparameters was able to balance the model's complexity and generalization.



(a) Training and Test Loss per Fold



(b) Predictions vs Actual Values

Figure 8: Evaluating LSTM Loss and Values

From Figure 8a, we observe a steady decrease in training loss as the model progressively learns from the training data. The LSTM model, leveraging its ability to capture temporal dependencies, achieves a significant reduction in training loss over time, ultimately stabilizing at a low value towards the end of the training phase. Initially, there were minor fluctuations in the training loss due to the smaller batch size used, but these diminished as training continued. The test loss was consistently higher than the training loss at the start, but it decreased steadily over time, causing the gap between the training and test losses to narrow. This trend is also evident in Figure 8b, where the predicted values closely track the actual values. However, deviations are observed in instances of higher volatility or unexpected patterns in the data, which the model struggles to capture entirely.

6 Conclusion / Future Work

In our search for a strong model that fits the electricity price data, we first considered the traditional SARIMA model as a baseline and then evaluated the viability of an exogenous variable (temperature) by comparing it with a modified SARIMAX version. The improvement

in forecasting accuracy (RMSE) and model quality (AIC) for the SARIMAX model suggests that some exogenous variables may be useful in electricity price forecasting, and it supports the idea that the weather plays a role in influencing electricity demand. However, historical analysis suggests that this forecasting power of the SARIMAX model may be limited to the short run or to situations in which the trend does not experience any sort of deviation.

In addition, the LSTM model demonstrated superior predictive performance compared to the both SARIMA(X) models in forecasting electricity pricing. The key advantage of the LSTM lies in its ability to capture complex temporal dependencies in the data without being constrained by assumptions like linearity and white noise residuals, which is inherent in the SARIMA model. This flexibility allows the LSTM to model non-linear, long-range dependencies in electricity pricing data, where such assumptions may not hold, leading to improved predictive accuracy.

In contrast, SARIMA's reliance on these assumptions often requires preprocessing steps, such as detrending, which can obscure important long-term trends in electricity prices. LSTM, on the other hand, is able to learn from the raw data directly, maintaining these trends and capturing both short-term fluctuations and long-term shifts more effectively. The LSTM's capacity to adapt to complex patterns in electricity prices gives it an edge over SARIMA, especially when dealing with unpredictable market dynamics, seasonal variations, and price spikes driven by external factors.

However, the LSTM model is also less interpretable, whereas the SARIMA(X) models offer us a way to directly quantify the effects of seasonality, exogenous variables, etc., through the coefficient estimates. As another example, while we were able to evaluate the forecasting accuracy for all models through the average RMSE score, the SARIMA(X) models also yield metrics like the AIC, which can reveal other aspects of model quality (e.g., simplicity) beyond predictive performance.

To further enhance these models and their application to electricity price forecasting, the following next steps can be further explored:

- **New Data (Integration of Exogenous Variables)**: We have incorporated a single exogenous variable in our SARIMAX model and seen its influence on model quality. On the other hand, currently, the LSTM model is univariate, relying only on historical price data for predictions. In both cases, we can incorporate more exogenous variables, such as weather patterns (e.g., precipitation, humidity), energy demand, fuel prices, and policy changes, all of which could significantly improve the model's predictive power. These external factors directly influence electricity pricing, and by including them, the models can capture a broader range of influences, leading to more accurate forecasts, especially during periods of high price volatility or sudden price changes driven by external shocks.
- **New Models (Hybrid SARIMA)**: While the LSTM outperformed SARIMA, hybrid approaches combining SARIMA with machine learning models like LSTM could be explored. A potential hybrid approach could involve using SARIMA to model the linear and seasonal components of electricity pricing, while the residuals from SARIMA are fed into an LSTM to capture any remaining non-linear relationships. This approach could leverage the strengths of both models, offering a more robust and flexible solution to electricity price forecasting.

As electricity markets become increasingly complex and volatile, improvements in forecasting models have the potential to play a crucial role in forecasting price movements, enabling better decision-making and more efficient management of electricity resources.

Appendix: Project Code

The code (and other relevant files) used for this project can be found at the following GitHub repo:

`https://github.com/hnngocvo/STATS-207-Final-Project`

References

2024. National average temperature dataset. Accessed: 2024-12-12.

John D. Hedengren. n.d. Long short term memory. `https://apmonitor.com/pds/index.php/Main/LongShortTermMemory`. Accessed: 2024-12-05.

Tianzhen Hong, Wen-Kuei Chang, and Hung-Wen Lin. 2013. A fresh look at weather impact on peak electricity demand and energy use of buildings using 30-year actual weather data. *Applied Energy*, 111:333–350.

Alistair King. 2024. U.s. electricity prices dataset. Accessed: 2024-12-11.

Malte Lehna, Fabian Scheller, and Helmut Herwartz. 2022. Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account. *Energy Economics*, 106:105742.

N. Singh and S. Mohanty. 2015. A review of price forecasting problem and techniques in deregulated electricity markets. *Journal of Power and Energy Engineering*, 3:1–19.

Andreas Wagner, Enislay Ramentol, Florian Schirra, and Hendrik Michaeli. 2022. Short- and long-term forecasting of electricity prices using embedding of calendar information in neural networks. *Journal of Commodity Markets*, 28:100246.

D. Wang, I. Gryshova, M. Kyzym, T. Salashenko, V. Khaustova, and M. Shcherbata. 2022. Electricity price instability over time: Time series analysis and forecasting. *Sustainability*, 14(15):9081.