Electricity Price Forecasting for Freight Electrification: A Spatial and Temporal Model for BET Charging along California's I-5

Juan C. Lopez, Dae Hyeun Cheong, Hee Won Ahn, Thommas Phan

1. INTRODUCTION

Freight transportation significantly contributes to pollutant and greenhouse gas (GHG) emissions, accounting for about 10% of global GHG and nearly 65% of all global pollution [1–4]. Medium- (MDT) and Heavy-Duty Trucks (HDT) have lagged behind Light-Duty Vehicles (LDVs) in decarbonization efforts, necessitating more ambitious goals to reduce their emissions [5]. Battery electric trucks (BETs) have been considered a vital decarbonization option in freight transportation. However, a significant limitation to electrifying the long-haul segment is the need for large batteries to complete long routes, reducing cargo capacity and increasing ownership costs [6,7]. Electrification decisions are also influenced by factors such as electricity pricing, investments in and availability of charging infrastructure, vehicle characteristics (e.g., vehicle type, range), and driving patterns (daily mileage, trip timing, dwell time) [7,8]. These factors lead to varying charging costs and increased total cost uncertainty, which can hamper fleets' willingness to adopt BETs.

Accurately predicting electricity prices across different regions is critical to support the widespread adoption of BETs in California, aligning with the State's climate and air-quality objectives to decarbonize the freight sector. By providing high-frequency (e.g., hourly or even more granular) electricity price forecasts, large fleet operators can make informed decisions to optimize charging schedules, thereby reducing costs and enhancing the economic viability of BETs. The California Independent System Operator (CAISO) [9] currently provides hourly and day-ahead electricity price forecasts for the entire state. However, the hourly forecasts lack spatial resolution, while the more spatially focused predictions are only available on a day-ahead basis. This creates a gap in the information that fleet operators need to spatially pinpoint optimal charging locations along predetermined routes based on real-time electricity prices.

In addition, the volatility of wholesale electricity prices poses a significant challenge for large fleet operators, as they purchase electricity at wholesale rates. Fluctuations in price can occur due to demand-supply imbalances, renewable energy availability, and grid conditions [10]. Predicting periods of low electricity prices allows fleet operators to reduce charging costs. Therefore, this project aims to develop a model to forecast wholesale electricity prices in regions along California's I-5 Highway, where a large share of the State's truck flows and freight facilities are concentrated [11]. Given the real-time variations in electricity prices, we will employ machine learning algorithms to identify regions with the lowest prices at different times of the day. This approach will support strategic planning for BET charging scheduling along major freight corridors and address the following research questions: How can we use machine learning

models to better predict spatial and temporal variations in California's wholesale electricity prices along the I-5 highway?

The remaining portion of this report is organized as follows: Section 2 presents a literature review on predictive models for electricity pricing, highlighting the use of Long-Short-Term Memory (LSTM) networks as key models to accurately predict the short-term price of electricity markets. Section 3 presents the methodology and the data acquisition process. Section 4 presents the exploratory data analysis. Section 5 describes the model training process for the SARIMA and LSTM models. Finally, Sections 6 and 7 present the results with discussion and conclusions with limitations.

2. LITERATURE REVIEW

Predicting electricity prices has been a focus of research for several studies. For instance, Yousefi, Sianaki, and Sharafi [10], developed a forecasting model for California's average monthly retail electricity prices over five years. The authors aimed to improve price prediction accuracy by identifying key factors influencing electricity prices and using machine learning techniques and data analytics to offer valuable insights for investors and energy stakeholders. Their dataset used data from the U.S. Energy Information Administration (IEA), covered January 2001 to 2017, and included variables such as natural gas consumption (for electricity and industrial sector), coal electricity generation, net electricity generation and imports, GDP, and renewable energy generation.

The authors' methodology involved data collection and cleaning, feature correlation analysis to identify variables correlated to retail electricity prices, and time series prediction using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. They developed a feature-based machine learning algorithm, training several models to minimize the mean squared error. The study assumed that the selected features significantly impacted electricity prices and that the models would generalize well to new data. Results showed that machine learning models could predict retail electricity prices accurately, bridging the gap between traditional methods and advanced analytics. The authors recommended further exploration of feature correlations and the inclusion of real-time data and domain knowledge to improve forecasting accuracy and adaptability.

The literature on electricity price forecasting (EPF) has advanced significantly, focusing on improving accuracy through various models and price determinants. This paper [12] reviews 62 EPF studies published between 2012 and 2022, examining their data preprocessing techniques, forecasting methods, model optimization, and evaluation criteria. It also considers key variables such as electricity supply and demand, fossil fuel prices, climate, imports, exports, and carbon prices, which have been shown to improve forecasting performance. The review categorizes EPF methods into three types: traditional econometric models, artificial intelligence models, and hybrid models. While conventional models like ARMA and GARCH have been widely applied to linear time series, newer ones like neural networks (CNN, DNN, LSTM) and support vector

machines (SVM) often offer greater flexibility in handling nonlinear and nonstationary data. The paper underscores the importance of model optimization and evaluation, noting that evaluation metrics MAPE, MAP, and RMSE are widely used but highlighting the absence of a standard performance evaluation system. Referencing this review can guide our choice of variables and appropriate models for our forecasting work.

Based on the existing literature, we anticipated that Long Short-Term Memory (LSTM) networks, enhanced with exogenous variables and refined through feature selection, would yield the most accurate short-term electricity price forecasting results. This approach is especially advantageous in highly volatile electricity markets with cyclical and seasonal pricing. LSTM models are particularly suited to capture time-dependent patterns due to their sequential memory capabilities, allowing them to learn from complex dependencies across time [10]. While Seasonal Autoregressive Integrated Moving Average (SARIMA) is commonly used for time series forecasting when seasonality and trend components are prominent, it is less effective in short-term contexts where price fluctuations are irregular and do not strictly follow seasonal trends. In contrast, the literature indicates that LSTM consistently outperforms SARIMA in short-term electricity price prediction, especially in capturing rapid, nonlinear price changes [10].

Integrating exogenous variables, such as temperature, load demand, and renewable generation output, further enhances the model's predictive power. Empirical results show that these factors have high correlations with electricity prices, and incorporating them has led to a 47.3% reduction in error for LSTM models [10]. Additionally, Random Forest is identified as a practical feature selection method that prioritizes the most impactful variables for LSTM input, thus enhancing model accuracy. The Random Forest algorithm ranks previous DAM and RTM prices, solar irradiance, and weather indicators among the top influential factors, allowing the model to focus on the primary drivers of price volatility without noise from less relevant data [12].

Furthermore, recent research advocates for predicting the gap between day-ahead (DAM) and real-time (RTM) prices directly rather than forecasting each separately. This approach mitigates the compounding of errors and captures critical market adjustments more accurately. The literature highlights that learning the DAM-RTM gap enables the model to account for short-term fluctuations driven by rapid demand changes, generation variability, and sudden grid condition shifts, which are central to accurate short-term forecasting. Additionally, modeling the DAM-RTM gap provides insights into price adjustments within an hour, aligning well to predict finer intervals such as 5-10 minutes [13]. Consequently, an LSTM model that integrates exogenous factors, utilizes Random Forest for feature selection, and directly predicts the DAM-RTM gap is expected to deliver highly accurate short-term price forecasts in volatile electricity markets.

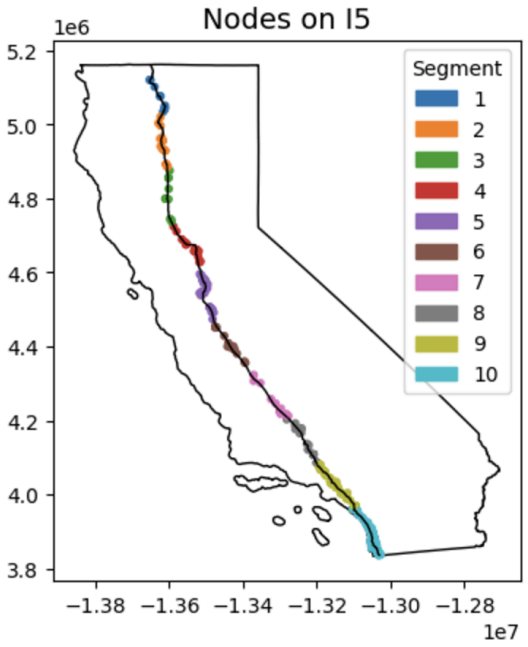
3. METHODOLOGY

3.1. Data acquisition

The research team built the data set by collecting data from different sources, such as GridStatus API [14], MeteoStat API [15], Caltrans [16], and California Open Data Portal [17], as shown in Appendix I.

3.2. Data Integration and Preprocessing Strategy

We employed a systematic procedure to merge datasets, including steps for data cleaning, aligning formats, handling missing values, and ensuring consistency across critical variables. Appendix II presents a detailed description of the data merging process.



	closest_segment	node_count
0	1	14
1	2	23
2	3	11
3	4	25
4	5	56
5	6	21
6	7	16
7	8	36
8	9	79
9	10	124

Figure 1. a. Distribution across the I5 highway, b. Node distribution per segment.

After gathering and preprocessing the data, we narrowed our analysis to segments 6 through 9 of the I-5 highway, corresponding to the section between Los Angeles and San Francisco (see Figure 1). We then conducted exploratory data analysis (EDA) to examine key interactions between variables, identify outliers, and detect highly correlated features. EDA also played a critical role in modeling SARIMA, mainly using ACF and PACF plots to assess autocorrelations and seasonal patterns in the outcome variables.

Following EDA, we applied elementary variable selection using the Variance Inflation Factor (VIF) method, excluding variables with a VIF higher than 10 to reduce multicollinearity. Using the selected variables, we developed SARIMA and LSTM models with the training dataset and made predictions on the test dataset. Then, the analysis was evaluated using graphical representations and RMSE.

3.3. SARIMA

LMP (Locational Marginal Price), representing electricity prices, is inherently highly volatile, driven by real-time market dynamics and various external factors such as weather conditions, demand fluctuations, and regulatory changes. Despite this volatility, electricity prices also exhibit seasonal patterns, with recurring high- and low-price periods throughout the day due to predictable shifts in supply and demand, influenced by external factors such as temperature and sunlight availability. Given these characteristics, SARIMA emerged as a compelling choice for electricity price forecasting. Its ability to model both trend and seasonality makes it suitable for this application, even when volatility is present.

SARIMA extends the ARIMA model by incorporating seasonal components. While ARIMA is effective for time series with trends but no seasonality, SARIMA introduces terms that capture periodic patterns, making it robust for forecasting in domains like economics, energy markets, and weather prediction. SARIMA's flexibility allows it to handle time series data with stationary and seasonal characteristics, making it particularly relevant for electricity price prediction. The SARIMA model is represented as:

$$SARIMA\ (p,\ d,\ q)(P,\ D,\ Q,\ s)$$

Where *p, d,* and *q* are the traditional ARIMA coefficients representing non-seasonal components: autoregression, differencing, and moving average, respectively. *P, D, Q* represent their seasonal counterparts. The seasonal term *s* indicates the length of the seasonal cycle, which is typically determined through autocorrelation and partial autocorrelation plots and fine-tuned using the training and test datasets. This flexible parameterization allows SARIMA to adapt to various seasonal patterns observed in time series data. The general SARIMA model equation is expressed as:

$$\Phi_p(B^s)(1 - B^s)^D \phi_p(B)(1 - B)^d Y_t = \theta_Q(B^s)\theta_q(B)\varepsilon_t$$

Where the terms correspond to seasonal and non-seasonal AR/MA polynomials and a white noise error term, expressed using the backward shift operator. The equation defines a comprehensive model that accounts for seasonal trends, short-term dynamics, and random fluctuations in the data. A more detailed breakdown of this equation is provided in Appendix III.

SARIMA operates under two key assumptions that must be verified before applying the model. First, the time series must be expressible as a linear combination of trend, seasonality, and noise. Second, the seasonal patterns must remain constant over time. After conducting EDA and verifying the stationarity of the dataset, SARIMA was selected as a baseline forecasting model. Despite its somewhat restrictive assumptions, its structured approach to handling seasonality and trends provides a strong benchmark for evaluating more advanced machine learning models, such as LSTM.

3.4. LSTM

While SARIMA is a widely used predictive time-series model, it has a significant limitation: it only considers time and LMP variables for generating predictions. According to the literature, electricity prices are highly volatile due to multiple external factors such as weather conditions, demand spikes, and regulatory changes. Therefore, a more advanced time-series model capable of incorporating these external influences is required to improve prediction accuracy.

LSTM is a specialized Recurrent Neural Network (RNN) designed for processing and predicting time series, sequential data, and temporal dependencies. Unlike SARIMA, which focuses on a limited number of variables, LSTM can integrate multiple features and adjust its learning based on historical and contextual information. This flexibility makes LSTM suitable for modeling highly dynamic and multi-variable datasets like electricity prices.

As SARIMA extends the traditional ARIMA model, LSTM enhances standard RNNs by solving RNN's inherent limitation: the inability to capture long-term dependencies. Traditional RNNs struggle with the vanishing gradient problem, where repeated backpropagation causes gradients to shrink exponentially, making it difficult for the network to learn long-term relationships. This issue arises from repeatedly multiplying small gradient values during backpropagation, limiting the model's ability to retain information across long sequences.

The key structural difference between RNN and LSTM lies in their memory management approach. Standard RNNs process sequences by passing the hidden state from one time step to the next, using only a simple recursive loop. This process treats all previous information equally, making RNNs prone to forgetting older contexts due to the vanishing gradient problem. LSTM, by contrast, introduces a memory-based architecture centered on a specialized component called the cell state. This cell state acts as a long-term memory storage, retaining relevant information while discarding irrelevant data through three core gating mechanisms:

1. **Forget Gate:** Determines which past information should be discarded.
    - In RNNs: No such mechanism exists—older information is gradually lost.
    - In LSTMs: The forget gate adjusts memory retention by learning what to keep and what to discard.
2. **Input Gate:** Decides which new information should be added to the memory.
    - In RNNs: New input simply overwrites the hidden state.
    - In LSTMs: The input gate regulates how much of the new input updates the cell state.
3. **Output Gate:** Controls what information to output as the next hidden state.
    - In RNNs: The hidden state from the previous time step is directly used as output.
    - In LSTMs: The output gate selectively passes relevant information from the memory for prediction while retaining critical long-term context.

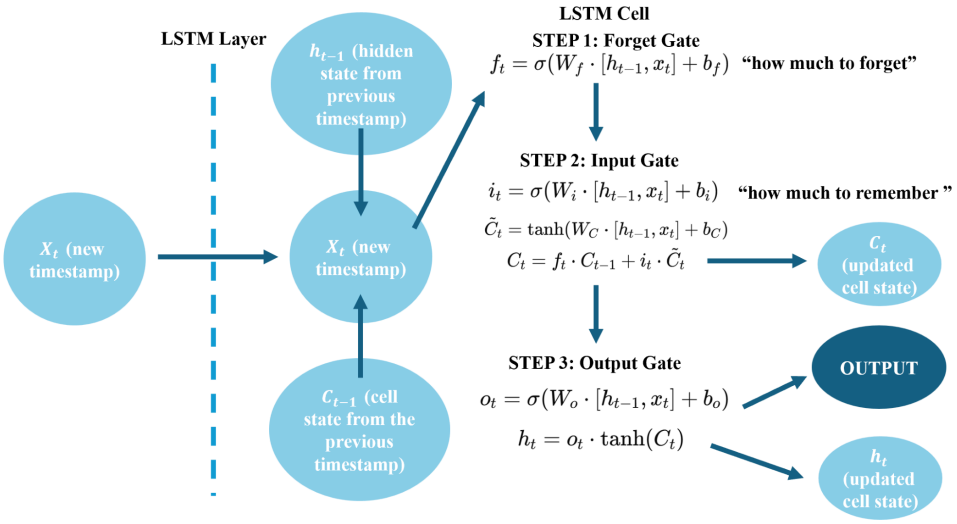


Figure 2. Visualization of LSTM Data Processing Procedure

These gates rely on sigmoid and tanh activation functions to regulate the flow of information effectively, ensuring that the model learns both short-term and long-term dependencies while mitigating the vanishing gradient problem. By dynamically managing memory, LSTMs excel at time-series forecasting and sequence-based prediction, making them one of the most potent tools in deep learning for sequential data. (See Figure 2 for more details of the LSTM model).

Another key advantage of using LSTM in this analysis is its less restrictive assumptions than SARIMA. LSTM requires the dataset to have temporal structure, long-term dependencies, and sufficient data points, all of which are satisfied in this study. Unlike SARIMA, LSTM does not require the data to be stationary or follow fixed seasonal patterns, making it more adaptable to dynamic and volatile electricity prices. By integrating external factors and utilizing LSTM's memory-based structure, we expect better performance than SARIMA, offering a more robust framework for electricity price forecasting.

4. Exploratory Data Analysis

We plotted the dataset to better understand the dynamics of LMP over time. Figure 3 illustrates the LMP trends, capturing the temporal variations and highlighting critical patterns. This visualization reveals notable price fluctuations driven by market volatility, which aligns with demand-supply imbalances and renewable energy variability. Two significant gaps were identified in the sequential data during the data collection and merging phase. These gaps were attributed to missing records for specific nodes during certain periods.

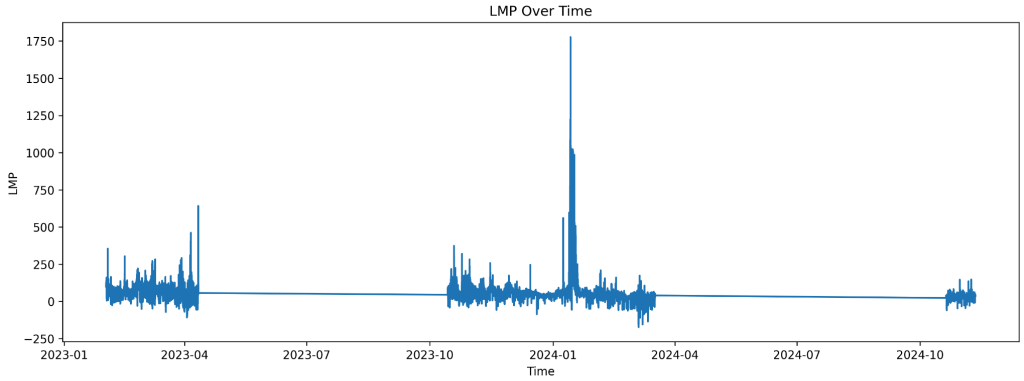


Figure 3. LMP data gathered.

For model training and testing purposes, we selected two specific time frames:

a. Training Data: A six-month sequence spanning October 2023 to April 2024, chosen for its completeness and representation of seasonal variations in LMP.

b. Testing Data: A three-week sequence from October to November 2024 reflects recent market conditions and allows for model evaluation against up-to-date scenarios.

Two initial questions were raised from Figure 3: Do our segments of interest (6, 7, 8, 9) trend similarly or differently? What is the cause of the large peak in LMP in January 2024? We analyzed LMP trends by segment along the I-5 corridor to better understand regional electricity price dynamics. Figure 4 presents LMP for segments 6 through 9 over time, capturing spatial variations across these segments.

The LMP trends for these segments appear highly similar, with only small differences observed in hourly price fluctuations. This similarity suggests that, within this portion of the I-5 corridor, LMP variability is more influenced by temporal factors (e.g., demand, supply, and grid conditions) than geographic location. Other features, however, remain highly similar due to nearby geographic locations, such as wind speed and temperature, as illustrated in Appendix IV.

A notable spike in January LMP persists across all segments, though its magnitude is considerably lower than the aggregate LMP trends shown in Figure 3. Upon further investigation, we attribute the exaggerated peak in Figure 3 to the rise in gas prices experienced during that time (see Figure A.2 in Appendix IV) and to LMP variation in Segment 1, located in the northernmost part of California. This region experiences higher LMP variance, likely driven by its sparse population, harsh weather conditions, and reduced grid reliability. These factors collectively lead to more significant price volatility in Segment 1 compared to the more populated and stable grid regions near Los Angeles and San Francisco.

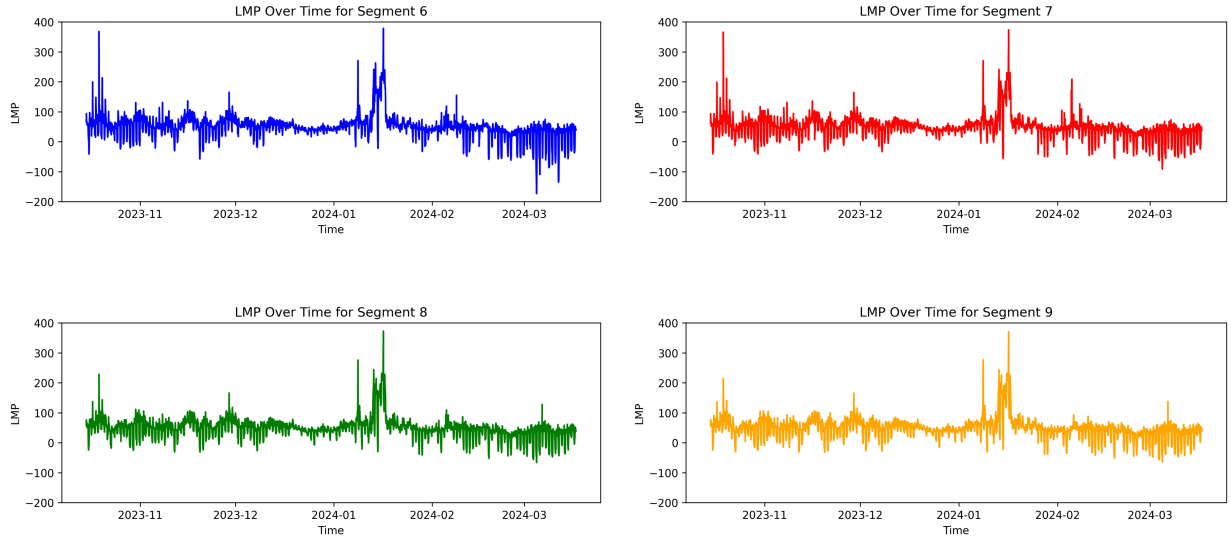


Figure 4. LMP across the segments of interest.

4.1. Trend analysis

In preparation for SARIMA and LSTM modeling, we looked into both the trend and seasonal analysis (Figure 5) since checking temporal and seasonal dependency is crucial in modeling both models. Figure 5 provides valuable insights into the observed LMP values, highlighting the trend, seasonality, and residual components. As mentioned earlier, the trend shows a significant upward spike in early January 2024, reflecting a temporary but sharp increase in electricity prices. This spike likely corresponds to heightened energy demand during colder months or other seasonal grid stress factors. After this peak, the trend returns to relatively expected values. The seasonal component demonstrates a clear daily periodic pattern, which implies strong seasonality with the long-term and temporal dependency of the dataset.

The ACF and PACF plots in Figure 6 further reveal the underlying structure of the time series and guide parameter selection in the SARIMA model. The ACF plot shows significant peaks at multiple lags, indicating the presence of moving average (MA) components in the time series. Notably, recurring peaks at 24 and 48 hours lag confirm a strong seasonal pattern aligned with daily periodicity. The ACF plot also shows significant peaks at lags 1, 2, and others. This suggests moving average (MA) components in the time series. The PACF plot exhibits a strong spike at lag 1, indicating the inclusion of an autoregressive (AR) term. Similar to the ACF plot, seasonal peaks around lag 24 support the inclusion of seasonal autoregressive terms in the model.

Figure 5. LMP Trend and Seasonality Plots.

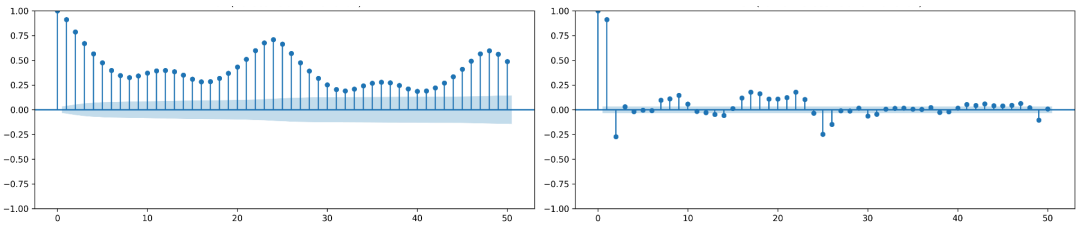


Figure 6. LMPs Autocorrelation Plots. (Left) ACF (Right) PACF.

4.2. Feature Selection and Multicollinearity

Multicollinearity was a critical consideration in the feature selection process for the predictive models, as highly correlated features can distort the coefficients of machine learning models and lead to overfitting. Figure 7 presents the correlation matrix heatmap before feature selection, showing the relationships between Locational Marginal Pricing (LMP) and various features, including weather variables, renewable generation sources, and grid operations. Several variables demonstrated high pairwise correlations, such as solar generation with imports and batteries and natural gas with LMP. These relationships suggested potential redundancies in the dataset.

In addition to mitigating multicollinearity, we sought to reduce model complexity due to the size of the initial merged dataset, which exceeded 350 MB, included 22 features, and comprised over 2 million rows. A high-dimensional dataset of this scale can strain computational resources and prolong model training and testing times. To address this, we employed the Variance Inflation Factor (VIF) method, iteratively removing features with VIF values above the threshold of 10. This approach ensured that only variables contributing independent information were retained in the dataset. For instance, features such as geothermal, biomass, and biogas, which exhibited high VIF values and low direct correlation with LMP, were excluded.

Figure 7 illustrates the correlation matrix before and after feature selection. The refined dataset demonstrates reduced multicollinearity, with fewer features showing strong pairwise

correlations. Key retained variables include natural gas price, temperature, wind speed, solar generation, large hydro, batteries, and imports. These features collectively capture the major drivers of LMP variability, such as weather-dependent renewable generation, grid storage usage, and external market conditions.

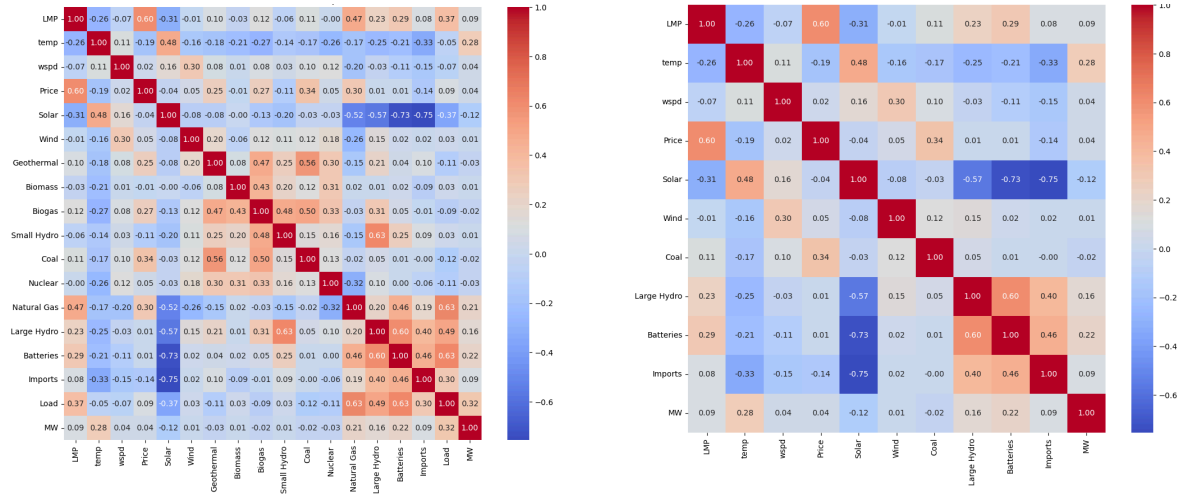


Figure 7. Correlation matrix heatmap. (Left) All Variables (Right) VIF Selected Variables.

5. Model Training

5.1. SARIMA Model Training

The exploratory data analysis (EDA) revealed the presence of AR/MA components and strong seasonality in the LMP time series. Trend analysis confirmed recurring daily patterns, suggesting the dataset follows a seasonal time-series structure. The Augmented Dickey-Fuller (ADF) Test also produced a p-value significantly lower than 0.05, indicating that the LMP time series is stationary at a 5% significance level. Given this result, we determined that the differencing term $d = 0$ is appropriate for our SARIMA model, as no further differencing is needed to achieve stationarity.

Table 1. Hyperparameter tuning results for SARIMA and LSTM

Segment	SARIMA	LSTM
Segment 6	(2, 0, 2)(1, 1, 1, 24)	LSTM unit: 32 dropout: 0.2 dense unit: 256 LR: 0.001 # layers: 1
Segment 7	(2, 0, 2)(1, 1, 1, 24)	LSTM unit: 64 dropout: 0.2 dense unit: 256 LR: 0.001 # layers: 1
Segment 8	(2, 0, 2)(1, 1, 1, 24)	LSTM unit: 64 dropout: 0.2 dense unit: 256 LR: 0.001 # layers: 1
Segment 9	(2, 0, 2)(1, 1, 1, 24)	LSTM unit: 64 dropout: 0.2 dense unit: 256 LR: 0.001 # layers: 1

The EDA findings also suggested the inclusion of autoregressive (p) and moving average (q) components, along with seasonal terms (P, D, Q). Given the observed 24-hour cyclic pattern in the LMP time series, we set the seasonal lag to 24 hours to capture the daily periodicity effectively. Considering this baseline model structure, we divided the dataset into training and test sets and conducted a grid search to fine-tune the SARIMA hyperparameters. See Table 1 for the tuning result for each segment of the highway.

The results confirmed our initial assumption that the LMP trends across segments exhibit similar patterns, with only minor hourly price fluctuations observed during the EDA. This consistency supports the choice of the same SARIMA configuration across all four segments, providing a unified and robust model for predicting electricity prices in the target area.

5.2. LSTM Model Training

Implementing the LSTM model involves a specific preprocessing method, model framework construction, and parameter tuning to optimize performance. First, we selected variables based on a prior variable selection process with VIF and included only those variables as columns in the dataset. Then, we applied a sliding window approach and scaling method to transform the dataset into the input structure required for the LSTM model. We began with a Bidirectional LSTM layer for the model framework to capture variations in past and future data points. Next, we added a fully connected dense layer as a hidden layer with a nonlinear activation function, followed by a final thick layer as the output layer. Dropout layers were incorporated between these layers to prevent overfitting. We used the Adam optimizer with a learning rate of 0.001, a commonly chosen value that balances convergence speed and stability while minimizing the risk of overshooting the optimal solution.

To ensure robust performance, we used a time series split to maintain the chronological order of the data while separating the training set and validation set. This method allowed us to systematically tune parameters such as:

- **LSTM units**: number of cells in each LSTM layer. Increasing the number of LSTM units allows the model to capture more diver patterns, but it can also increase the risk of overfitting.
- **The dropout rate**: the probability of units being dropped during training. This parameter helps the model prevent overfitting.
- **Dense units**: number of neurons in each dense layer. Higher dense units improve the capacity to learn complex relationships but may increase unnecessary model complexity.
- The **number of LSTM layers**, where we limited the structure to one or two layers. A single LSTM layer is efficient for simple patterns, while two layers capture long-term dependencies but require more data to prevent overfitting.

We built the same framework for four different models and tuned parameters for each segment from 6 through 9 to find the best model for the unique characteristics of each dataset. See Table 1 for the tuning result. After selecting the optimal parameters, we trained the final model using the entire training dataset to build the final model and tested it with each test dataset.

6. Results and Discussion

According to Figure 8, both the SARIMA and LSTM models perform reasonably well in predicting the actual electricity prices on the test set by capturing general seasonal patterns and the underlying trend. However, a closer inspection reveals that the SARIMA model struggles to capture the non-linear behavior in electricity prices caused by various external factors. This limitation is evident in its frequent failure to predict sudden peaks and periods of high volatility in the time series. Moreover, SARIMA's predictions appear overly constant and static across time intervals, resulting in over-smoothed forecasts that fail to reflect the real-world volatility in the LMP dataset.

The primary cause of SARIMA's limitations is its reliance on restrictive assumptions about the time series. SARIMA assumes that the dataset can be modeled as a linear combination of trend, seasonality, and noise components. It also expects seasonal patterns to remain static and constant over time, making the model less adaptable to datasets with dynamic seasonality and frequent fluctuations. Since the LMP dataset is highly volatile, containing many nonlinear behaviors influenced by external factors such as market supply and demand, weather conditions, and electricity generation levels, SARIMA's assumptions are violated, leading to suboptimal performance.

In contrast, the LSTM model successfully addresses these challenges by incorporating external factors directly into its predictive framework. As a machine learning-based RNN model, LSTM can handle a broader set of input features, including variables such as electricity demand and supply, weather conditions, and generation capacity. This enables it to capture the dataset's short-term irregularities and long-term dependencies effectively. This flexibility allows LSTM to adapt to electricity prices' non-linear and dynamic nature. Figure 8 demonstrates that LSTM predictions closely follow price trends, including volatile spikes.

Furthermore, comparing the final prediction RMSE values in Table 2 reveals that LSTM consistently outperforms SARIMA, reinforcing its superiority over traditional time-series models. This performance advantage is due to LSTM’s ability to incorporate multiple variables and its flexibility in handling real-life, volatile datasets with minimal assumptions. While SARIMA operates under rigid assumptions about stationarity and static seasonality, LSTM's adaptive architecture allows it to learn complex patterns from data, making it a more suitable model for electricity price forecasting in a dynamic environment. Appendix IV presents the results of other segments.

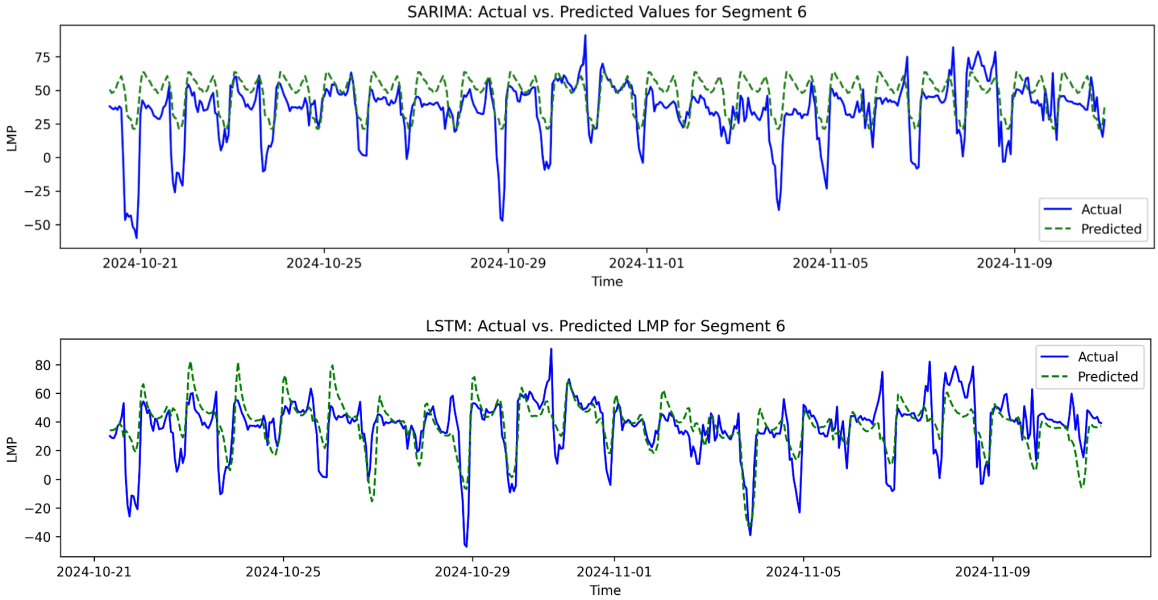


Figure 8. Graph of Actual vs Prediction of Segment 6 (T) SARIMA (B) LSTM

Table 2. RMSE of SARIMA and LSTM Prediction

Segment	**SARIMA RMSE**	**LSTM RMSE**
6	20.560034	15.096201
7	18.900282	14.977289
8	15.644757	13.078254
9	16.215970	15.958361

7. CONCLUSIONS AND LIMITATIONS

This study investigated the application of time-series models for predicting electricity prices along California's I-5 Highway, focusing on comparing the performance of SARIMA and LSTM models. Given the critical role of electricity price forecasting in optimizing battery electric truck (BET) charging schedules, accurate and high-frequency predictions are essential for reducing costs and supporting California's climate goals. Our findings demonstrated that while both models captured general seasonal patterns and long-term trends, LSTM outperformed SARIMA due to its ability to incorporate external variables and adapt to the highly volatile nature of electricity prices.

The SARIMA model effectively modeled the dataset's seasonality and trend, as our exploratory data analysis (EDA) and statistical tests suggested. However, its rigid assumptions hindered its performance, including stationarity, linearity, and static seasonality. These assumptions limited its ability to capture nonlinear behaviors and unexpected spikes in electricity prices driven by external factors like market demand and weather conditions. As expected, SARIMA produced over-smooth and static forecasts, particularly during volatile periods, resulting in higher root-mean-square error (RMSE) scores than LSTM.

Conversely, the LSTM model addressed these challenges by incorporating multiple external factors, including supply and demand variables, weather data, and electricity generation capacity. Its adaptive memory-based architecture enabled it to capture short-term irregularities and long-term dependencies without assuming stationarity or fixed seasonality. As demonstrated by the RMSE results in Table 2, LSTM consistently outperformed SARIMA across all highway segments, providing more accurate and responsive forecasts. These results validate the effectiveness of deep learning-based models in predicting electricity prices in a dynamic, real-world environment.

In addition to predicting electricity prices, we explored a secondary research question: "Can the LSTM model identify the cheapest segment along the I-5 highway at any given hour?" Although this analysis was included in Appendix V due to space constraints, its findings are notable. While electricity prices across segments followed similar patterns, the LSTM model successfully identified the cheapest segment at various times, demonstrating the potential for real-time electricity procurement. Specifically, the model’s average predicted price for the lowest-cost segment was $30.12 per MWh, compared to the average of $31.83 across all segments—an approximate 5.4% cost reduction. This result highlights the model's practical value for reducing electricity expenses in real-time applications. We direct readers to Appendix V for additional details on our analysis of the secondary research question, including corresponding visualizations (Figures A5 and A6).

Despite these promising results, several limitations should be acknowledged. First, while LSTM can incorporate many external variables, our dataset may have missed key factors such as global market conditions, political events, and energy policy changes, potentially affecting the model's accuracy. Expanding external features through a broader literature review could reduce omitted-variable bias. Second, computational limitations restricted our ability to perform full hyperparameter tuning, preventing the use of a complete grid search or deeper LSTM architectures with more than two layers. Future research could benefit from powerful computing environments and advanced tuning methods such as Bayesian optimization. Lastly, the dataset covered only five months for training and 20 days for testing, limiting the model’s ability to learn long-term seasonal patterns and diverse price fluctuations. Expanding the dataset to include multiple years would enhance the model's ability to capture complex temporal dynamics, making its forecasts more robust and accurate.

REFERENCES

[1] ITF. Transport Outlook 2019. OECD; 2019. https://doi.org/10.1787/TRANSP_OUTLOOK-EN-2019-EN.

[2] IEA. Transport 2023. https://www.iea.org/topics/transport (accessed June 8, 2023).

[3] IEA. Trucks and Buses 2023. https://www.iea.org/energy-system/transport/trucks-and-buses (accessed July 27, 2023).

[4] IEA. Global energy-related CO2 emissions by sector 2020. https://www.iea.org/data-and-statistics/charts/global-energy-related-co2-emissions-by-sector (accessed July 27, 2023).

[5] IEA. Transport 2022. https://www.iea.org/reports/transport (accessed April 28, 2023).

[6] Rohith G, Devika KB, Menon PP, Subramanian SC. Sustainable Heavy Goods Vehicle Electrification Strategies for Long-Haul Road Freight Transportation. IEEE Access 2023;11:26459–70. https://doi.org/10.1109/ACCESS.2023.3257431.

[7] Zhao HP, Wang Q, Fulton LP, Jaller MP, Burke AP. A Comparison of Zero-Emission Highway Trucking Technologies. 2018. https://doi.org/10.7922/G2FQ9TS7.

[8] Li B, Jing D, Zhong H, He G, Ma Z, Ruan G, et al. Centralized charging station planning for battery electric trucks considering the impacts on electricity distribution systems. Energy Reports 2023;9:346–57. https://doi.org/10.1016/J.EGYR.2023.04.090.

[9] California ISO. Prices, Today's Outlook 2024. https://www.caiso.com/TodaysOutlook/Pages/prices.html (accessed May 27, 2024).

[10] Yousefi A, Sianaki OA, Sharafi D. Long-Term Electricity Price Forecast Using Machine Learning Techniques. 2019 IEEE PES Innovative Smart Grid Technologies Asia, ISGT 2019 2019:2909–13. https://doi.org/10.1109/ISGT-ASIA.2019.8881604.

[11] Caltrans. 2023 California Freight Mobility Plan. 2023.

[12] Chai S, Li Q, Abedin MZ, Lucey BM. Forecasting electricity prices from the state-of-the-art modeling technology and the price determinant perspectives. Res Int Bus Finance 2024;67:102132. https://doi.org/10.1016/J.RIBAF.2023.102132.

[13] Nizharadze N, Soofi AF, Manshadi SD. Learning the Gap in the Day-Ahead and Real-Time Locational Marginal Prices in the Electricity Market, 2020

[14] Kanter M. GridStatus API 2023. https://docs.gridstatus.io/en/latest/index.html (accessed December 4, 2024).

[15] Meteostat. Meteostat Developers 2024. https://dev.meteostat.net/ (accessed December 4, 2024).

[16] Caltrans. Caltrans GIS Data 2024. https://gisdata-caltrans.opendata.arcgis.com/ (accessed December 4, 2024).

[17] California Open Data. California Open Data 2024. https://data.ca.gov/ (accessed December 4, 2024).

APPENDIX I: Data Collection Source and Explanation of Variable Considered

Dataset	Variables	Description and variables	Source
Locational Marginal Pricing (LMP)	Date Hour (every 15 minutes) LMP (every 15 minutes)	LMP is the real-time wholesale price of electricity at specific locations, named 'nodes'. In California, buyers and sellers in electricity markets use LMP prices as benchmark signals.	GridStatus API [14]
Node's information	Node_id, TAC Latitude and longitude	The nodes and the geographical locations. TAC denotes the transmission area across California.	GridStatus API [14]
Grid demand	Date, TW, TAC Hour (every hour)	Denotes the MW of electricity demand at a specific time and day in each transmission area.	GridStatus API [14]
Grid mix	Date Hour (every hour) Solar, Wind, Geothermal, Biomass, Biogas, Small Hydro, Coal, Nuclear, Natural Gas, Large Hydro, Batteries, Imports	Solar to Imports represents the electricity each source produces during each hour. Batteries denote the total amount of energy used from storage, while imports represent the amount imported across California from external markets.	GridStatus API [14]
Gas prices	Date, Hour, Price	Price of natural gas used for electricity generation.	GridStatus API [14]
Grid load	Date, Load Hour (every 5 min)	Load represents the flow of electricity on high-voltage power lines.	GridStatus API [14]
Storage	Date Hour (every 5 min) Stand-alone Batteries	Stand-alone Batteries represent the amount of electricity stored in batteries in California.	GridStatus API [14]
Weather stations	Node_id Weather station	Identifies the closest weather station given the longitude and latitude of a node.	MeteoStat API [15]
Weather data	Date, Hour, Temp, Wspd	Weather data for the closest weather stations to the nodes in California. Temp denotes the average temperature of that hour in degrees Celsius, and Wspd denotes the wind speed in m/s.	MeteoStat API [15]
California's Shapefile	Geometry	Geographical definition of the State	California Open Data Portal [17]
I-5 Shapefile	Geometry	Geographical shape of the I5 highway in California	Caltrans [16]

APPENDIX II: Data Merging Procedure

The data merging procedure is graphically represented in Figure A.1 and described below:

1. **California Nodes**: The locations of nodes were plotted alongside California's shapefile. Since some nodes were outside California, the analysis was restricted to nodes within the state, resulting in 3,184 identified nodes.
2. **I-5 Segments**: The I-5 shapefile was divided into 10 equal-length segments to represent potential charging regions.
3. **I-5 Nodes:** California nodes were mapped with I-5 segments, identifying nodes within 5 miles of the highway. These nodes were assigned a segment_id, representing potential charging locations. 404 nodes were identified along I-5 (see Figure 1 for distribution by segment).
4. **Weather Data:** Using the MeteoStat API [ref], the closest weather station was assigned to each node identified in Step 3 (station_id). Weather data for these stations was collected from 01-01-2022 to 11-11-2024.
5. **LMP Data:** LMP data was gathered for each node (from Step 3), including segment_id, for the same period (01-01-2022 to 11-11-2024).
6. **Grid Mix, Gas Prices, Grid Load, and Storage Data:** These datasets were collected statewide from 01-01-2022 to 11-11-2024.
7. **Grid Demand Data:** For the same period, data on grid demand was collected for each Transmission Access Charge (TAC) area in California.
8. **LMP and Weather Data Merge:** LMP and weather data were combined based on each node's closest weather station (station_id), date, and hour. Data gaps were identified, with full coverage only between 10-14-2023 to 03-17-2024 and 10-20-2024 to 11-11-2024.
9. **LMP, Weather, and Demand Data Merge:** The merged LMP_weather data was combined with grid demand data using each node's TAC, date, and time, resulting in 354 nodes matching the TAC data.
10. **Final Merge (LMP_Full):** The remaining datasets, independent of location, were merged based on time and date.
11. **Segment-level Grouping:** Finally, all node-level information was grouped by segment_id to compute average values for variables of interest within each potential charging area.

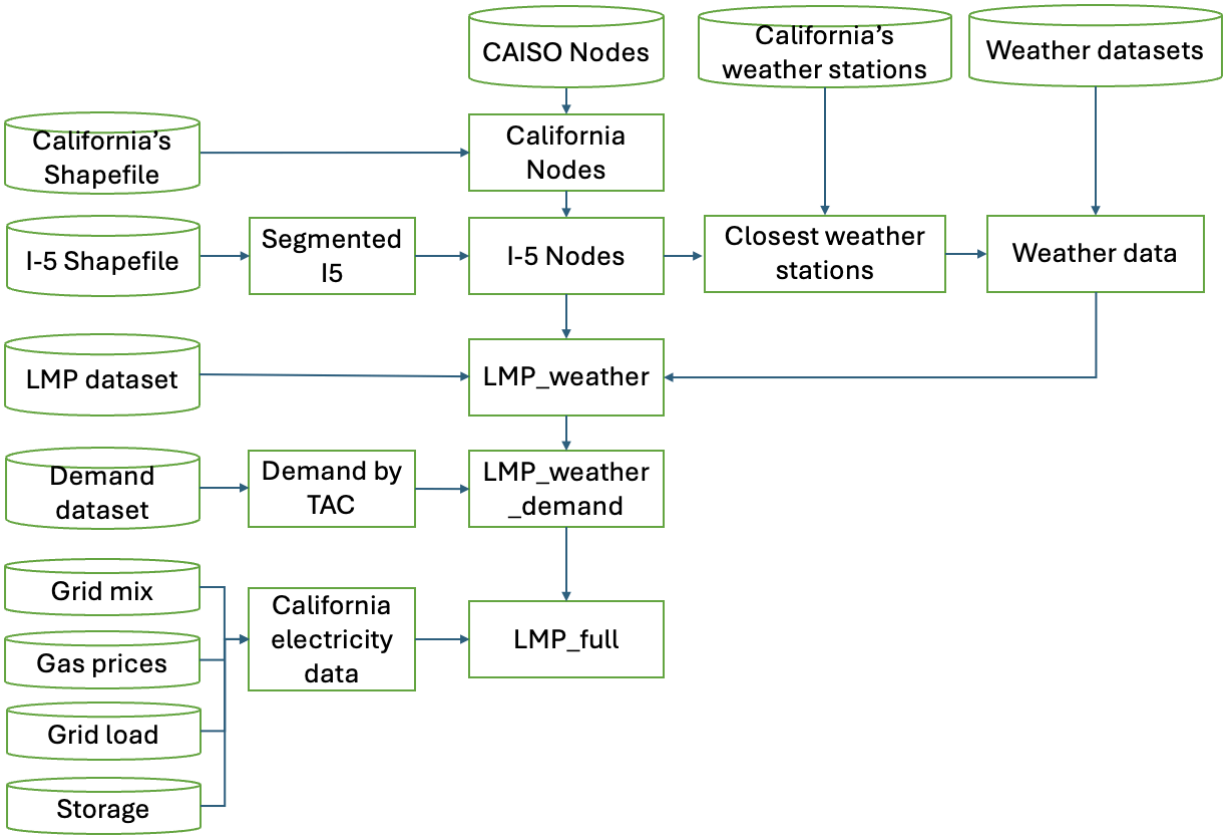


Figure A.1 - Merging data process

APPENDIX III: Additional Explanation of SARIMA:

The SARIMA model is represented as

$$SARIMA(p,\ d,\ q)(P,\ D,\ Q,\ s)$$

Non-Seasonal Components:

- *p*: Order of autoregression (number of past observations to use)
- *d:* Degree of differencing (number of times to subtract past values)
- *q*: Order of moving average (number of past forecast errors)

Seasonal Components:

- *P*: Seasonal autoregressive order (similar to p, but at seasonal lags)
- *D*: Degree of seasonal differencing (number of seasonal differences)
- *Q*: Seasonal moving average order (similar to q, but for seasonal periods)
- ***s*:** Length of the seasonal cycle (e.g., 12 for monthly data with yearly seasonality)

The general SARIMA model can be expressed as:

$$\Phi_p(B^s)(1 - B^s)^D \phi_p(B)(1 - B)^d Y_t = \theta_Q(B^s)\theta_q(B)\varepsilon_t$$

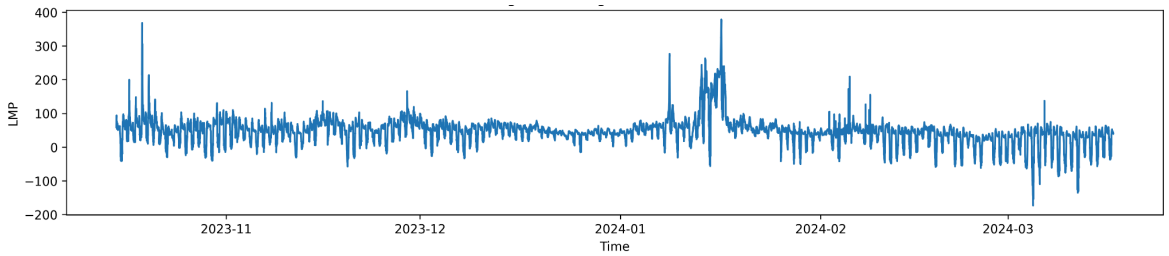
- Y_t: Time series value at time t.
- B: Backward shift operator ($BY_t = Y_{t-1}$)
- ε_t: White noise error term
- $\phi_p(B)$: Non-seasonal AR polynomial
- $\theta_q(B)$: Non-seasonal MA polynomial
- $\Phi_P(B^s)$: Seasonal AR polynomial
- $\Theta_Q(B^s)$: Seasonal MA polynomial

Where each term of the equation representing:

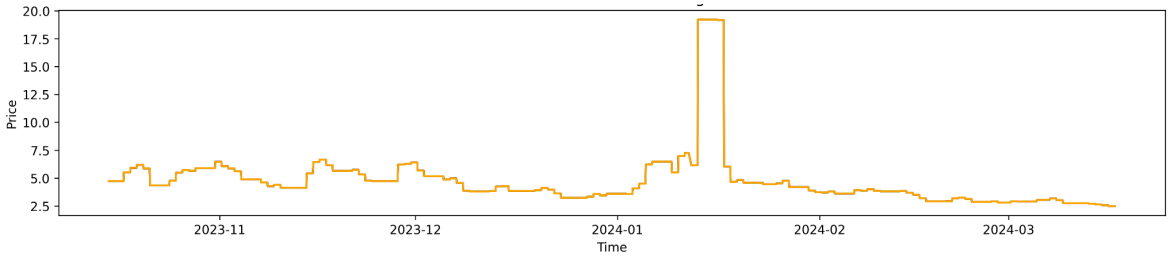
- $\phi_p(B)(1 - B)^d Y_t$: Captures the linear trend in the data
- $\theta_q(B)$: Captures short-term dependencies.
- $\Phi_P(B^s)(1 - B^s)Y_t$: Models seasonal patterns.
- $\Theta_Q(B^s)$: Captures seasonal error corrections.

APPENDIX IV: COMPLEMENTARY ANALYSIS OF THE EDA

Figure A2 shows that the peak in January 2024 correlated to the price of natural gas in California during the same time span. It is reasonable to expect such a large increase in LMP when the Natural Gas Price rises since California is reliant on renewable energy to generate electricity, one of the primary sources being natural gas.

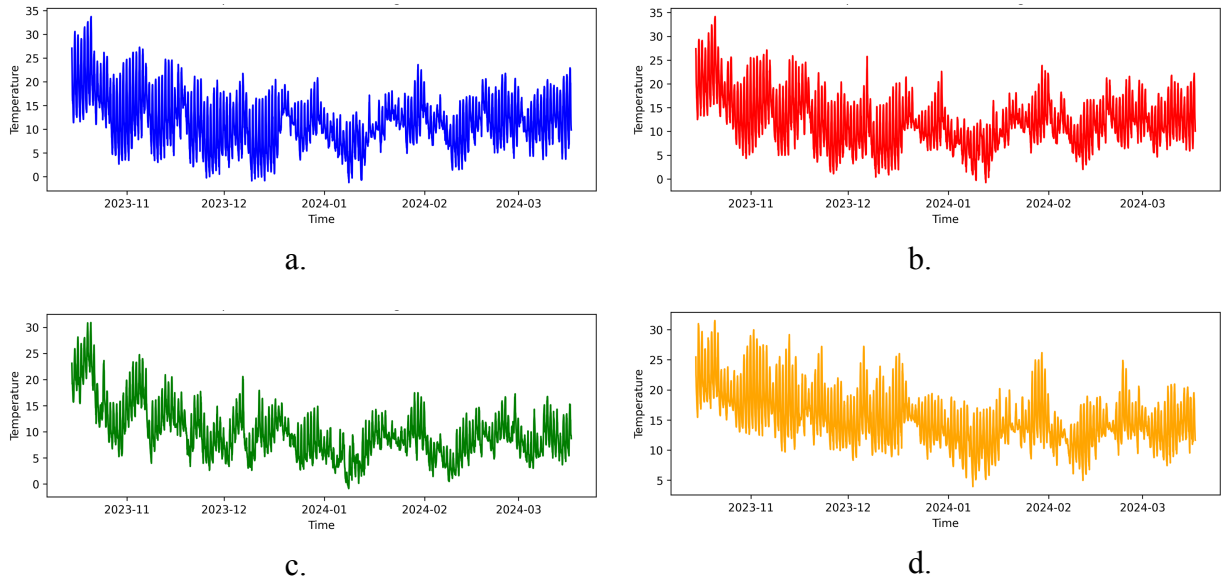


a.



b.

Figure A.2. Prices between 10-2023 and 04-2025 for all segments. a. LMP, b. Natural gas



a. b. c. d.

Figure A.3. Temperature over time. a. Segment 6, b. Segment 7, c. Segment 8., d Segment 9.

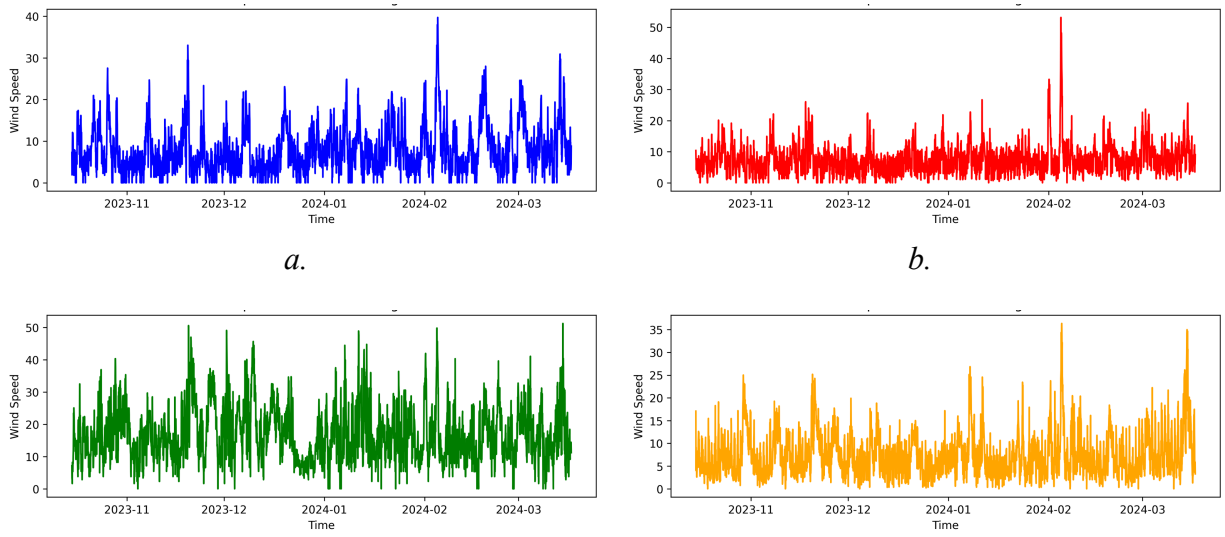


Figure A.4.Wind speed over time. a. Segment 6, b. Segment 7, c. Segment 8., d Segment 9.

APPENDIX V - SECONDARY RESEARCH QUESTION

The primary question of interest in this work is: How can we use machine learning models to predict better spatial and temporal variations in California's wholesale electricity prices along the I-5 highway? According to the RMSE of the LSTM model, it appears the model can reliably predict variation in California's electricity price based on the features selected in EDA. However, a secondary question arises: *If the LSTM model can predict spatial and temporal variations in California's wholesale electricity prices by segments along the I-5 highway, how can the model also predict where the cheapest segment is by hour?*

We evaluated the model's ability to predict the cheapest segment by plotting the LMP values by segment (Figure A5). We found the expected result from EDA: LMP values move very similarly among the four regions. While they are not the same, there is a slight variation between the four segments at any given time, making predicting the cheapest one a difficult task.

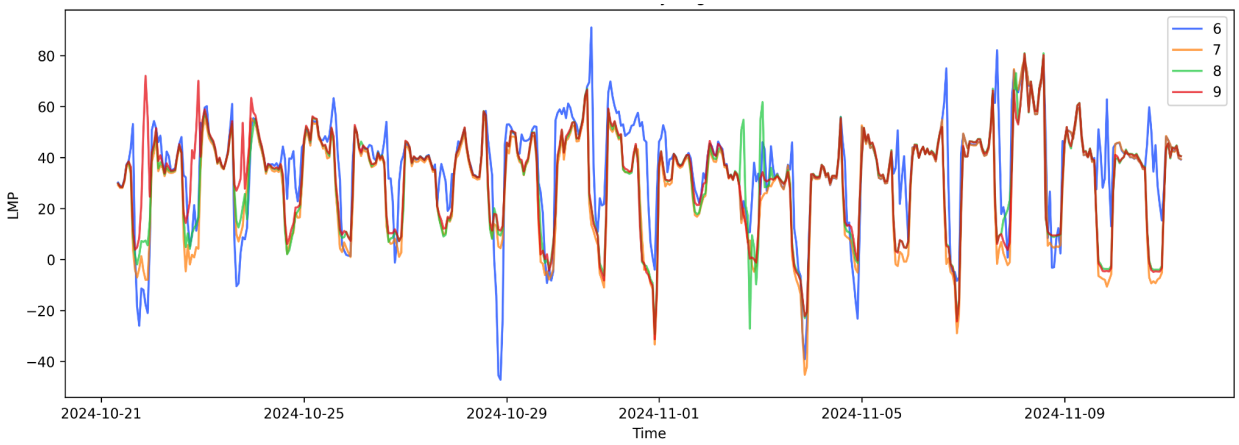


Figure A.5. LMP by segment

We chose to compare the model's "pick" (lowest predicted value of LMP by segment) at each hour and then collected the actual LMP value of the chosen segment and compared it to the mean LMP of all four segments at each hour, as illustrated by Figure A6. We see that the model mainly selected a segment cheaper than the LMP average for most of the 3-week testing dataset. The average of the model selection's choices is $30.12, while the total timespan average for all four segments is $31.83, resulting in ~5.4% less than the average. We conclude that the model can consistently select a "cheaper-than-average" region.

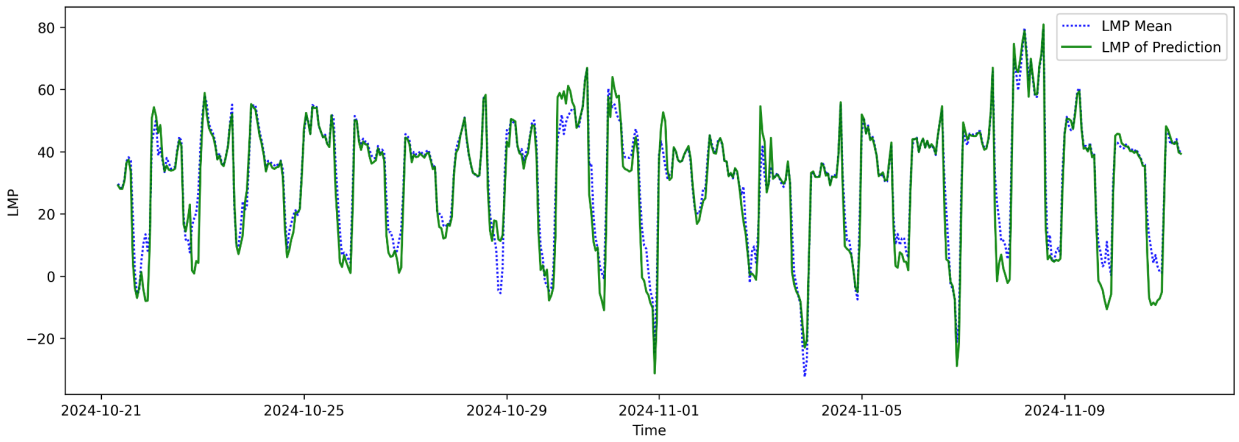


Figure A.6. Mean LMP vs. predicted LMP.

APPENDIX VI - GITHUB REPOSITORY

For further reference and additional resources related to this research, please visit GitHub repository at the following link: LMP_project

This repository includes:

- **LMP_Raw_Collection**: Raw datasets collected from various sources.
- **LMP_Data_Merge**: Steps for merging datasets into the final analysis-ready format.
- **LMP_EDA**: Exploratory Data Analysis (EDA) with visualizations and descriptive statistics.
- **LMP_LSTM_SARIMA**: Implementation of LSTM and SARIMA models for electricity price prediction.