**ECE 537 Data Mining, Winter 2025  
Final Project Report**

**Project Title:** Predicting Day-Ahead MISO’s Locational Marginal Prices Using Data Mining Techniques and Publicly Available Data.

**Students in the project group**: LaRico Andres (in-class) and Michael Acquah (online)

**Roles**: **LaRico** – Data Collection and Preprocessing, Background research, Literature Review **Michael** – Model Development, Project Planning, Coordination, Visualization

**Both** - Presenting and Reporting, Documentation, Communication, Feature Engineering

**Introduction**

In this project we apply data mining techniquessuch as data preprocessing- transformations, correlation, discretization, and normalization, time variant data warehousing, and various statistical analytical processes to analyze and predict day-ahead locational marginal prices (LMP) in the MISO (Midcontinent Independent System Operator) wholesale electricity market.LMP is a pricing mechanism used in electricity markets to reflect the cost of delivering power to specific locations, or nodes, within a transmission network. It accounts for the cost of electric power generation, the cost of delivering that power, and the physical limitations of the transmission system. LMP is crucial in managed wholesale markets, providing real-time pricing signals that help balance supply and demand while considering factors like congestion and load patterns.

Publicly available datasets from the U.S. Energy Information Administration (EIA), MISO and Open-Meteo.com were leveraged to extract meaningful patterns, identify key relationships, and derive actionable insights from historical electric market data. These insights were then used to develop three predictive models capable of forecasting future electricity prices each to varying positive degrees of confidence with **each step using data mining techniques learned in ECE 537**. This methodology is novel and can be quite valuable given more development as we have no access to proprietary information and unique algorithms used in industry. The results demonstrate the efficacy of data-driven approaches using publicly available data in enhancing decision-making and operational efficiency in energy markets.The literature on electricity market forecasting reveals several established techniques and emerging trends, which are summarized below.

The research paper by Gołębiewska et al [1]compares four data mining methods – MLP neural networks, SVM, MARSplines, and regression – for forecasting Polish electricity production one and seven days ahead. Using five years of data, MLP networks, especially regression-based ones, proved most accurate, showing lower forecast errors and higher correlation coefficients than other methods. MARSplines, while robust and handling outliers well, yielded the lowest accuracy. The study highlights the trade-off between the high accuracy of MLPs and the interpretability challenges they present, contrasting them with the robustness but lower accuracy of MARSplines. The best method ultimately depends on prioritizing accuracy versus model interpretability. We want to prioritize accuracy.

The next significant contribution comes from Rosano & Nerves, who used Artificial Neural Networks (ANNs) to forecast locational marginal prices (LMPs) in the Philippine electricity market, improving accuracy by including data from adjacent generator nodes. A feedforward ANN with backpropagation training achieves 6.8-6.9% Mean Average Percentage Error (MAPE), outperforming ARIMA models. Strengths include improved accuracy and handling of non-linear price behavior. Weaknesses are limited data (three months), model complexity, and assumptions about generator behavior. The method offers practical value for market participants but requires further research with more data and broader geographic scope[2].

In this study[3], Martínez-Álvarez et al examines data mining techniques for electricity-related time series forecasting. It covers linear methods (AR, MA, ARMA, ARIMA, ARCH, GARCH, VAR) and non-linear methods (neural networks, SVMs, genetic programming, k-NN, rule-based, wavelets, ensemble methods, and others). The paper's strength lies in its comprehensive overview and showcasing the superiority of data mining over classical statistical methods for complex electricity data. However, it lacks depth in individual method descriptions and lacks original empirical comparisons. The focus on specific markets might also introduce bias. Computational aspects and data preprocessing are not thoroughly addressed.

However, none of the authors mentioned above incorporated weather variability in their models or explicitly used previous hour (R2 = .80+) and previous 24 hour (R2 = .15+) LMP values which our feature importance data mining techniques proved to have high predictive importance.

**Methodology**

This section outlines the comprehensive methodology employed to predict day-ahead LMP in the MISO electricity market. The approach encompasses data collection, preprocessing, feature engineering, model development, and evaluation.

The primary datasets used in this project were sourced from publicly available platforms, including the EIA for LMP data, and the Open-Meteo.com Application Programming Interface (API) for weather forecasts. The data was either downloaded as CSV, preprocessed and merged into our dataframe or downloaded directly from API into our python environment and then preprocessed. This dataset ultimately encompassed a time span from February 2021 to March 2025, focusing on hourly timestamps.

For **data preprocessing**, rows with missing values were dropped to ensure data integrity. To maintain the quality of the data, outliers were detected using the **Interquartile Range (IQR) method.** This involved calculating the first quartile (Q1) and the third quartile (Q3) to determine the bounds for acceptable values. Data points falling outside these bounds were flagged as outliers. **Log transformations** were also applied to the LMP values to address the **positively-skewedness and normalize the distribution**. Other transformations (**square root, cube root, and Box-Cox**) were also explored but ultimately not chosen because their results were inferior to the log transformation as seen in the presentation.

For **feature engineering**, several new features were created after linear regression models were run to access baseline performance using the Pandas library to enhance the model's predictive capabilities and to inject temporal dynamics into the feature set. Previous hour (LMP\_lag\_1) and same hour from the previous day (LMP\_lag\_24) values of LMP were included to capture the temporal dependencies. Combined variables like the total wind strength and a temperature-humidity index were computed to better represent the impact of weather on electricity prices. Sinusoidal transformations of the hour of the day were applied to account for daily seasonality patterns as well.

In **Exploratory Data Analysis** (EDA), **summary statistics**, **uniqueness, consecutive and null rule** data mining techniques were utilized on key variables for better data understanding and to aid in data integration to create the most robust model possible. Also, **Pearson Correlation Maps** were integral in helping to confirm key drivers of LMP variability. Then, **K-means clustering**, an unsupervised machine learning algorithm used to group data into clusters based on similarity was applied to group similar hours/days/weather conditions based on weather variables, time of day, and LMP behavior, revealing four distinct clusters. Cluster 0 represents moderate conditions with medium LMP pricing, Cluster 1 represents off-peak conditions with low LMP pricing, Cluster 2 represents high weather variability such as snow/high winds with a wide range of LMP pricing, and finally Cluster 4 was the most interesting in warm peak hours and high LMP prices. This allowed our model to classify individual points in each cluster for predictive analysis later.

To build our final model, advanced models like Random Forest and XGBoost were also employed for their ability to manage non-linear relationships and feature importance. The most influential features (e.g., lagged LMP values, actual load, temperature) were selected based on importance scores from Random Forest and XGBoost.

The dataset was divided into training (80%) and testing (20%) sets using the train\_test\_split method from Scikit-learn. Models were trained on the training set and evaluated on the test set using metrics such as Root Mean Squared Error (RMSE) and R² scores. To ensure the model's generalizability, 5-fold cross-validation was employed. This method allowed for the assessment of model performance across different subsets of the dataset, providing a more reliable estimate of its predictive power.

**Experiments**

We conducted experiments to evaluate the effectiveness of different models and feature sets in predicting LMP. The following experiments were performed: The first experiment involved establishing a baseline using a Linear Regression model(**RMSE of $39.18/MWh and an R² score of 0.1002**). Just by using the log transformation of LMP the model improved to **RMSE of $32.38/MWh and an R² score of 0.2082**.

We also used feature-reduced models to improve model efficiency by selecting the most impactful features. A forecast – only model used features like temperature, dew point, wind total, lagged LMP values, and actual load, achieving an RMSE of $26.61/MWh and R² score of 0.5689. The second model (full reconstruction) included congestion and loss components in addition to the features from the forecast – only model, achieving a lower RMSE of $20.28/MWh and a higher R² score of 0.7500.

For the third and most advanced models, XGBoost best highlighted the importance of lagged LMP values, actual load, and weather variables suggesting XGBoost picks up on seasonal and weather trends more sharply, these variables are critical to our novel approach.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **YEAR** | **MAE** | **RMSE** | **MAPE** | **R2** |
| 2023 | 3.37 | 4.43 | 13.18 | 0.4837 |
| 2024 | 10.97 | 17.21 | 27.12 | 0.3673 |
| 2025 | 7.71 | 10.51 | 17.87 | 0.3264 |

These experiments demonstrated that advanced models like Random Forest and XGBoost, combined with carefully engineered features, significantly outperform baseline methods in predicting LMP, with XGBoost performing best overall with the lowest RMSE metrics.

The table presents performance metrics for predicted LMP across April 1 of 2023, 2024, and 2025 using a machine learning forecasting model. The model performed best in 2023, achieving the lowest Mean Absolute Error (MAE) of $3.37/MWh and a RMSE of $4.43/MWh, with a relatively strong R² value of 0.4837, indicating it captured nearly half of the variance in actual LMP values. However, performance deteriorated in 2024, with errors more than doubling (MAE = $10.97, RMSE = $17.21), and a significantly higher mean absolute percentage error (MAPE) of 27.12%, suggesting the model struggled with increased price volatility or unusual market conditions that year. In 2025, the model showed moderate performance (MAE = $7.71, RMSE = $10.51), but the R² remained low (0.3264), implying limited explanatory power. Overall, the model exhibits varying accuracy year to year, performing best under stable price conditions (2023), while requiring improvement for capturing volatility in years like 2024.  
**Conclusion**

This project successfully applied advanced data mining techniques and predictive modeling to analyze and forecast day-ahead locational marginal prices (LMP) and actual load in the MISO electricity market. By leveraging publicly available data from the U.S. Energy Information Administration (EIA) and weather variables we employed a comprehensive datamining approaches including data preprocessing, time-series analysis, and machine learning to uncover meaningful patterns and enhance predictive accuracy.

Future work could explore hybrid modeling approaches, real-time data integration, and expanded regional analyses to further refine predictive performance and operational insights.

**References**

[1] B. Gołębiewska and J. Trajer, “Analysis of energy market using data mining methods.” [Online]. Available: [www.cire.pl](http://www.cire.pl)

[2] K. R. Jay Rosano and A. C. Nerves, “Give to AgEcon Search Forecasting Locational Marginal Prices in Electricity Markets by Using Artificial Neural Networks.” [Online]. Available: <http://ageconsearch.umn.edu>

[3] Francisco Martínez-Álvarez, Alicia Troncoso, “A Survey on Data Mining Techniques Applied to Electricity-Related Time Series Forecasting,” *Energies (Basel)*, vol. 8, no. 11, pp. 13096–13111, 2015, doi: 10.3390/en81112361.