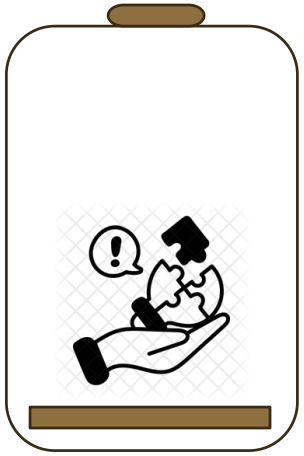
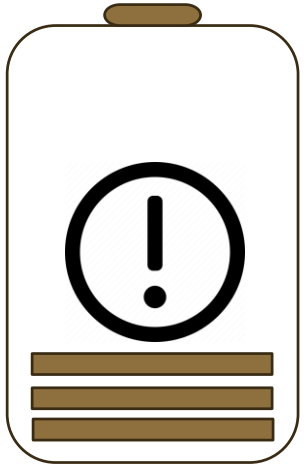
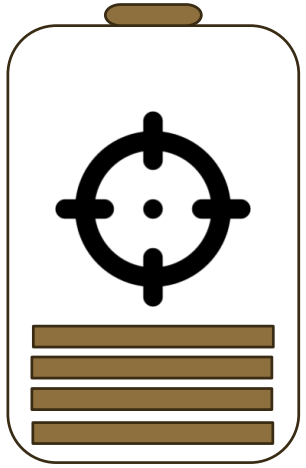
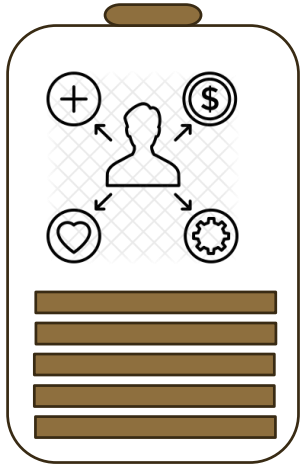
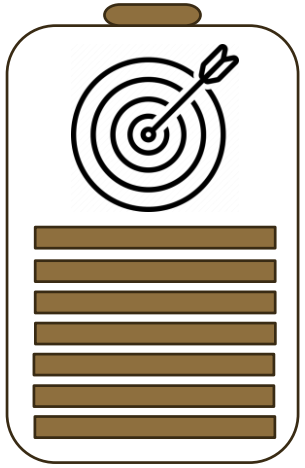
ENERGY MARKET TIME SERIES FORECASTING

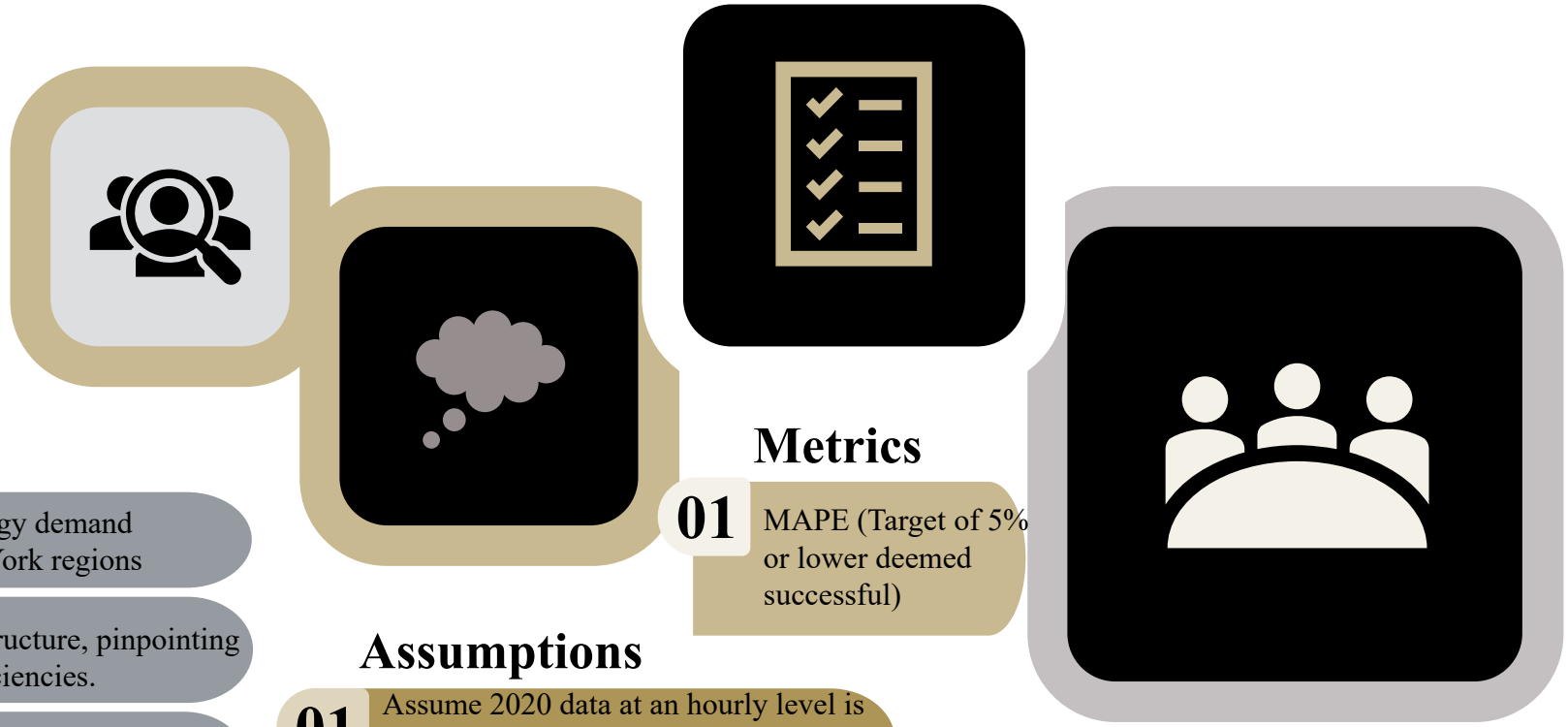
Amrutha Gabbita, Ridvan Kanca, Sahana Reddy, Alexander Scarcelli

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Business Problem Framing

BUSINESS PROBLEM	PURPOSE	IMPORTANCE	CONTEXT	POTENTIAL BUSINESS BENEFITS	STAKEHOLDERS	END GOAL
						
Determining optimal regions for grid-scale battery deployment to enhance energy distribution efficiency in New York for a major US-based company with offerings in a wide variety of industries including grid-scale commercial energy storage.	The analytics serve a dual purpose: primarily aiding a commercial, utility-scale B2B product with potential future applications in national security, focusing on regional volatility assessments to inform market positioning and gap analysis.	Strategic prioritization enhances the firm's market competitiveness by identifying generation or storage gaps, guiding efforts to optimize battery cycling in grid-scale systems, which is vital in the rapidly evolving energy market.	Currently focused on non-military, commercial markets, utilizing proof-of- concept analytics for initial regional volatility assessments.	Guides strategic decisions, enhancing operational efficiencies and influencing profit margins.	Targets commercial sectors with potential future applications in national security, addressing both B2B and potentially B2G stakeholders.	Optimize battery charge and discharge cycles for installed grid-scale storage systems, leading to improved market positioning and cost efficiencies and consumers potentially enjoy lower energy pricing.

Business to Analytics Problem Framing



Objectives

01 Analyze hourly energy demand and supply in New York regions

02 Evaluate grid infrastructure, pinpointing bottlenecks or inefficiencies.

03 Assess renewable energy variability and predictability

04 Identify regions for optimal grid-scale battery deployment, prioritizing efficiency, reliability, & cost-effectiveness.

Metrics

01 MAPE (Target of 5% or lower deemed successful)

Assumptions

01 Assume 2020 data at an hourly level is representative of seasonal hourly trends for 2024 and beyond

Strategy

01 Leverage open source algorithms like SARIMAX, ARMA, and relevant neural network architectures along with K-Means clustering to forecast hourly energy

Business to Analytics Problem Framing

Task 1: Data Pre-processing and Exploratory Analysis

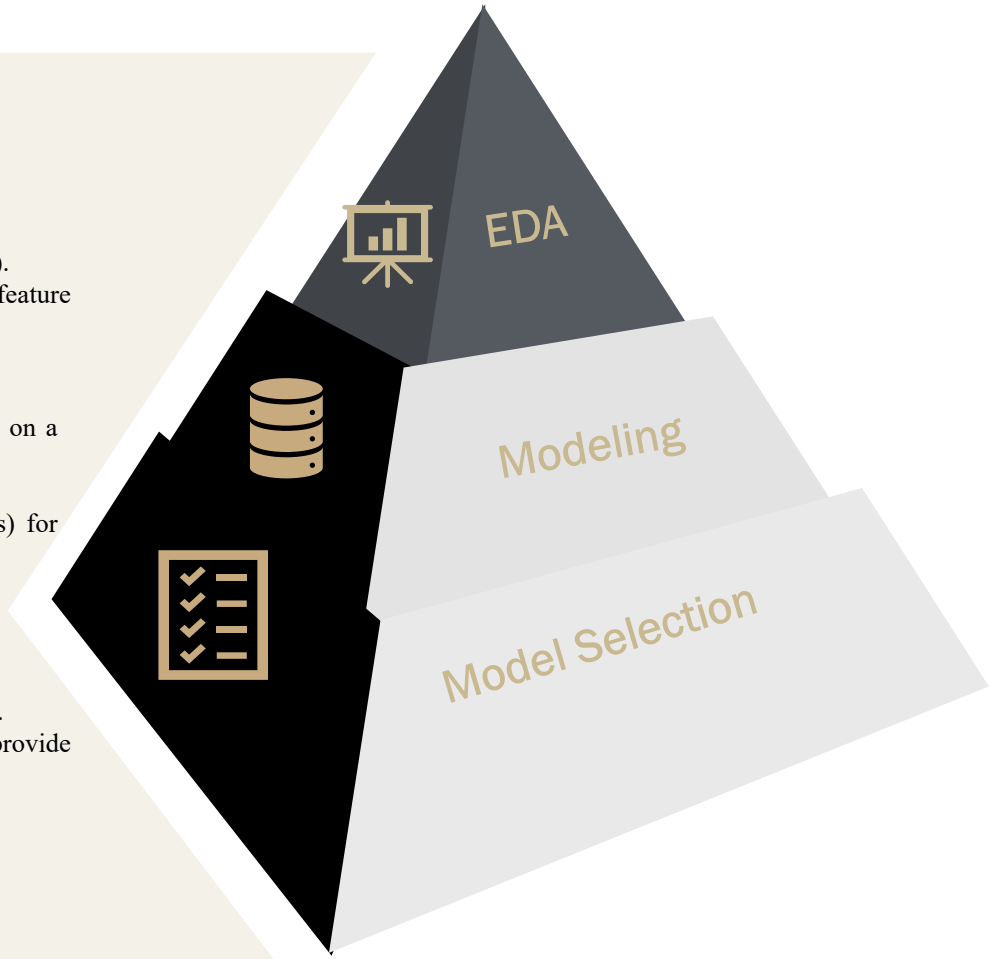
- Assess data quality, completeness, and consistency.
- Perform statistical analysis to understand the distribution of key features.
- Identify and analyze relationships between features and the target variable (energy grid load).
- Determine the most significant features for load forecasting using correlation matrices or feature importance algorithms.

Task 2: Algorithm Selection and Modeling

- Compare different time series forecasting methods (like ARIMA, XGBoost, Prophet, etc.) on a subset of the data.
- Evaluate performance metrics for different models to select the most promising approach.
- Experiment with different configurations of neural networks (RNN, LSTM, Transformers) for deep learning-based forecasts.

Task 3: Algorithm Selection and Modeling

- Refine the chosen models with hyperparameter tuning and cross-validation.
- Implement a method for geographic segmentation of the energy grid data (e.g., by zip code).
- Develop a prototype that integrates model forecasts with geographic segmentation to provide focused regional load predictions.



Literature Review

STUDY	BEST MODEL	MAPE	NOVELTY
Jacob N. Silver, 2016	ARIMAX	4.14	Focuses on finding the best combination of exogenous variables from past literature
Varshney, Sharma, & Kumar 2016	SSA-NN	4.73	Combination of singular spectrum analysis and artificial neural network reduces the time required to train and performs better compared to other NN models
Patil, Deshmukh, & Agrawal, 2017	ARIMA with Clustering	4.1	Classifies electric price data using K-mean and k-NN data mining techniques rather than by using a calendar
Neupane, Lee Woon, Aung, 2017	VWM	3.94	Propose two different strategies for selecting each hour's expert algorithm from the set of participating algorithms/ Final ensemble shows better results over ARIMA
Zahid et al, 2019	ESVR	4.2	Utilizes unique machine learning techniques at different forecasting stages, outperforming benchmark schemes
Our Study	**Similiarities**: Forecasts LBMP using best found models and exogenous variables similar to previous literature **Novelty**: Focus on hourly price prediction through feature experimentation and different aggregations		

Additional Research

Fuels

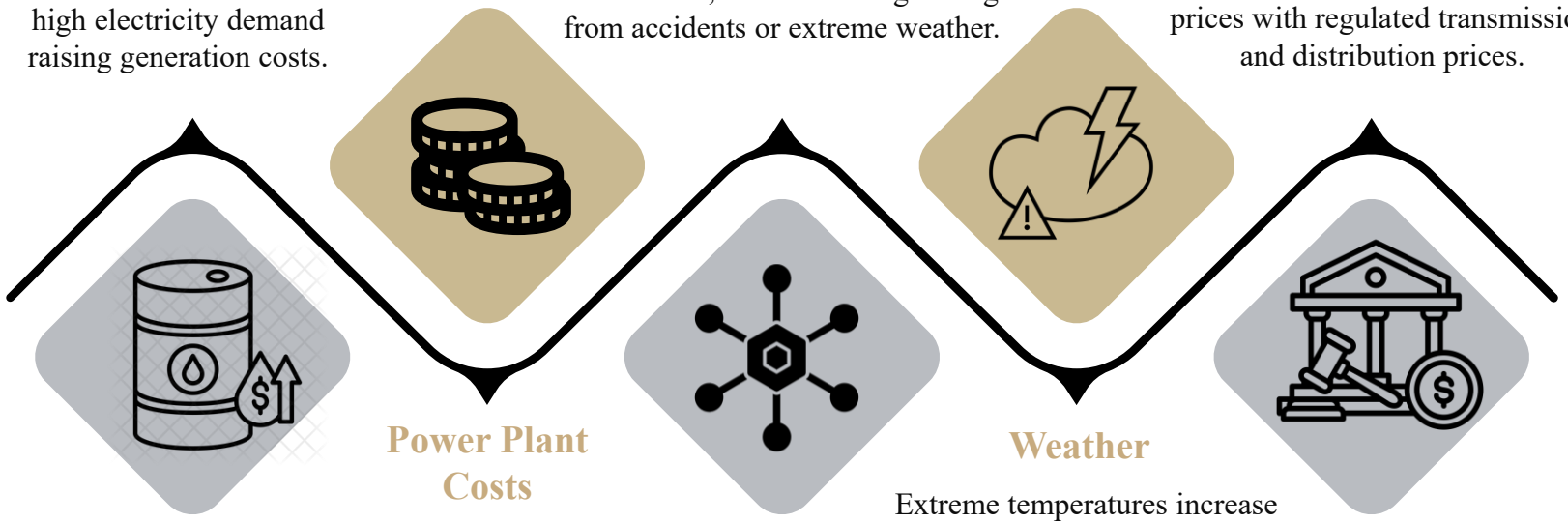
Fluctuating prices of natural gas and petroleum can spike during high electricity demand raising generation costs.

Transmission and Distribution Systems

The infrastructure connecting power plants to consumers involves costs for construction, operation, maintenance, and addressing damage from accidents or extreme weather.

Regulations

Pricing regulations vary across states, with some having fully regulated prices by public service commissions, while others combine unregulated generator prices with regulated transmission and distribution prices.



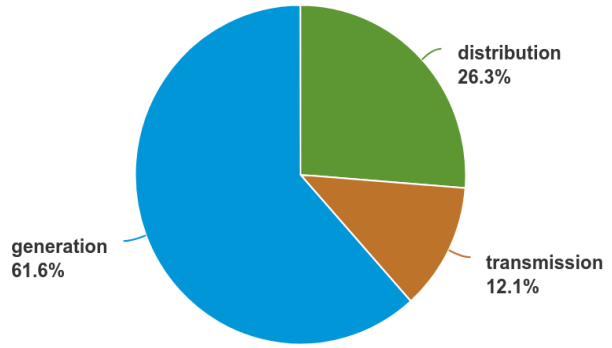
Power Plant Costs

Each power plant incurs expenses for financing, constructio n, maintenance, and operation.

Weather

Extreme temperatures increase heating and cooling demands, affecting fuel and electricity prices. Rain and snow support low-cost hydropower, while favorable wind speeds enable cost-effective wind power.

Major components of the U.S. average price of electricity, 2022



eia Data source: U.S. Energy information Administration, *Annual Energy Outlook 2023* , Reference case, Table 8, March 2023

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Data Sources and Data Basics

DATA SOURCE

- Primary Data Source: NYISO Open source data
- Secondary Data Source: Energy Information Administration and NYISO

PREPROCESSING

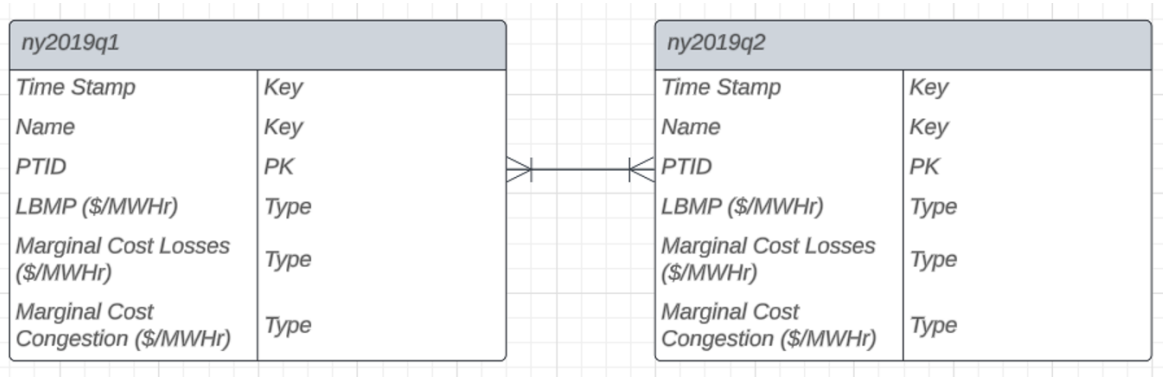
- Aggregate Timestamp column and filter and impute any null values with forward fill

- January - March 2019 & January - December 2020

Data Basics

Column Name	Description	Type	Example
Time Stamp	This column records the date and time of a record	String	1/1/2019 0:00
Name	Street address of the place of interest.	String	59TH STREET_GT_1
PTID	This column contains unique identifiers assigned to each address.	Int	25648
LBMP ($/MWHr)	“Locational Based Marginal Pricing”. The price of electricity at a specific location on the grid, calculated per megawatt-hour	Float	25.57
Marginal Cost Losses ($/MWHr)	least bit marginal price. The lowest price at which electricity can be sold in the wholesale market at a given time	Float	1.07
Marginal Cost Congestion ($/MWHr)	The costs associated with congestion in the power grid.	Float	-14.97

Datasets are broken into several csv files for easy sharing, but technically we only have one dataset, the only difference is the "time".



Exploratory Data Analysis

Checking the Dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
import seaborn as sns
df=pd.read_csv('ny2019q1.csv')

# Checked the dataset, very clean, no NaN values,only contains 6 columns
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 6 columns):
 #   Column                            Non-Null Count    Dtype
---  ------                            --------------    -----
 0   Time Stamp                        1048575 non-null  object
 1   Name                              1048575 non-null  object
 2   PTID                              1048575 non-null  int64
 3   LBMP ($/MWHr)                     1048575 non-null  float64
 4   Marginal Cost Losses ($/MWHr)     1048575 non-null  float64
 5   Marginal Cost Congestion ($/MWHr) 1048575 non-null  float64
dtypes: float64(3), int64(1), object(2)
memory usage: 48.0+ MB
```

Summary Statistics (Numerical Data)

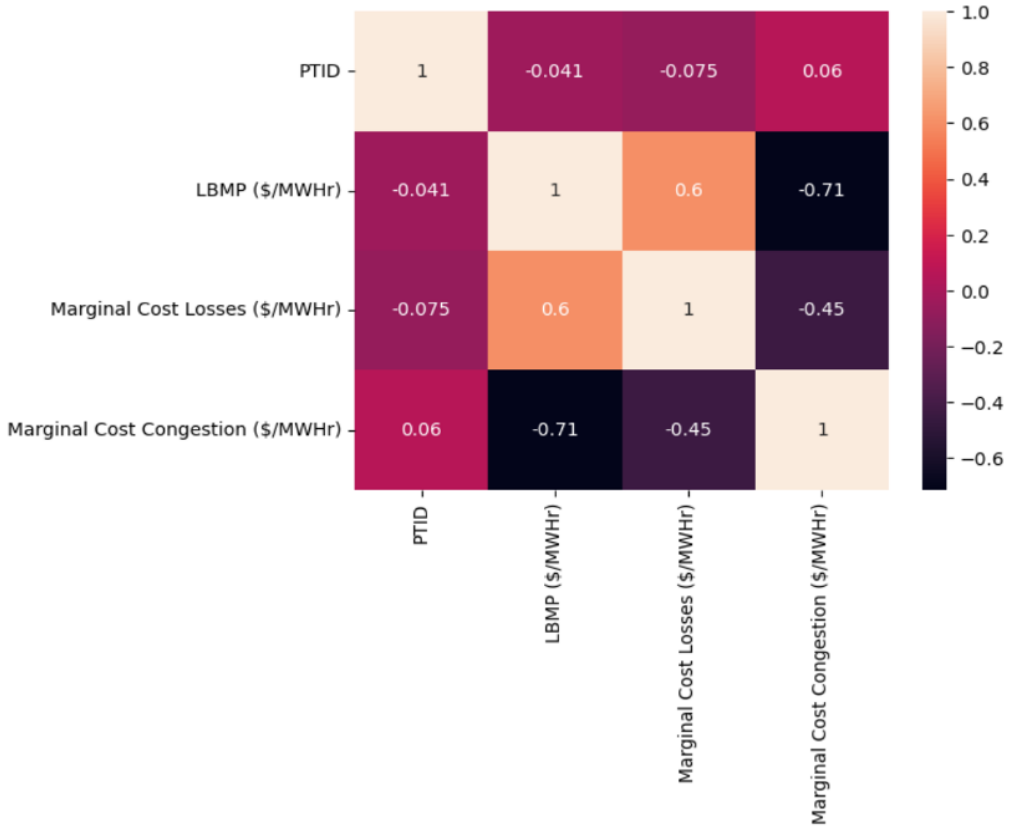
	PTID	LBMP ($/MWHr)	Marginal Cost Losses ($/MWHr)	Marginal Cost Congestion ($/MWHr)
count	1048575.0	1048575.0	1048575.0	1048575.0
mean	100888.2	37.0	1.9	-6.8
std	132602.2	22.1	2.4	11.9
min	23512.0	0.1	-18.4	-208.6
25%	23712.0	25.0	0.3	-8.9
50%	24102.0	30.5	2.0	-2.2
75%	323558.0	40.6	3.1	0.0
max	345034.0	252.7	35.1	54.2

Summary Statistics (Non-Numeric)

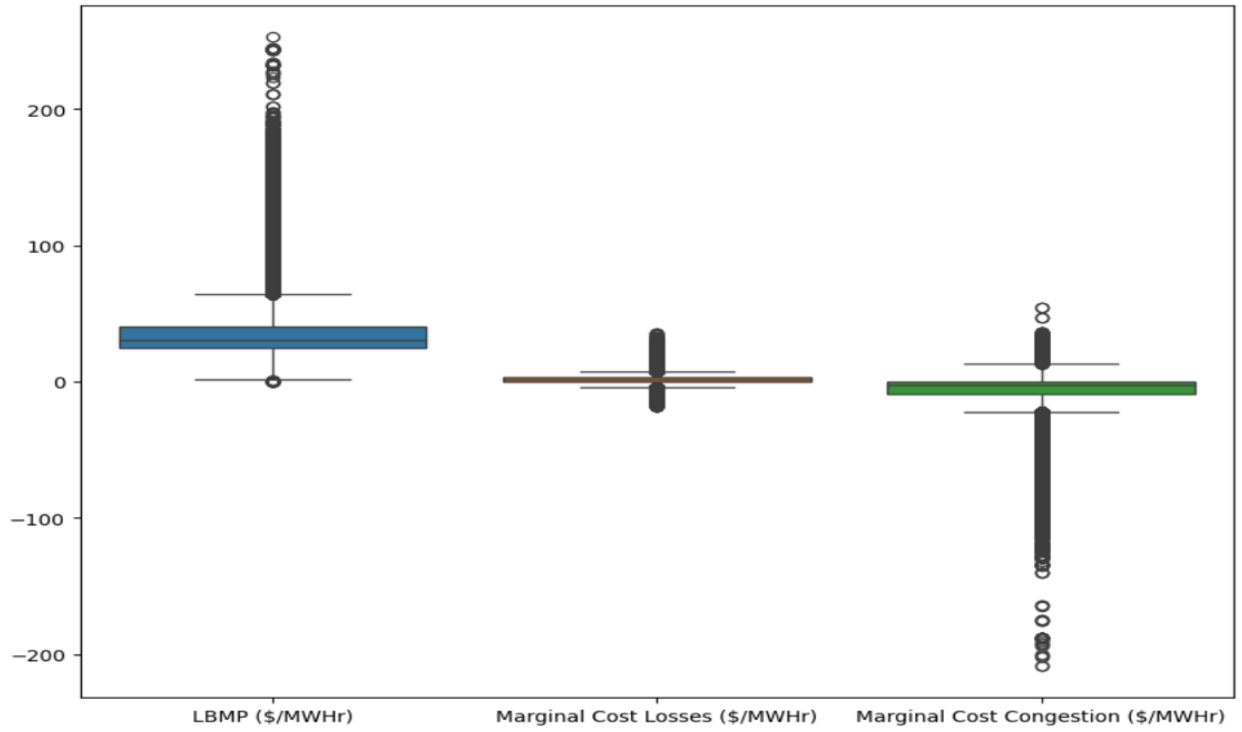
	Time Stamp	Name
count	1048575	1048575
unique	1890	556
top	2019-03-15 10:00:00	59TH STREET_GT_1
freq	556	1890
first	2019-01-01 00:00:00	NaN
last	2019-03-20 18:00:00	NaN

Exploratory Data Analysis

Correlation Matrix



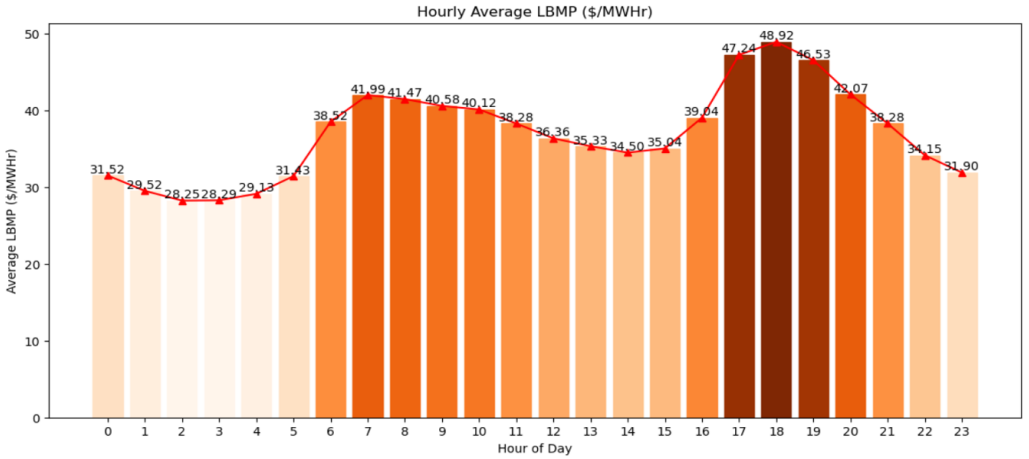
Outliers*

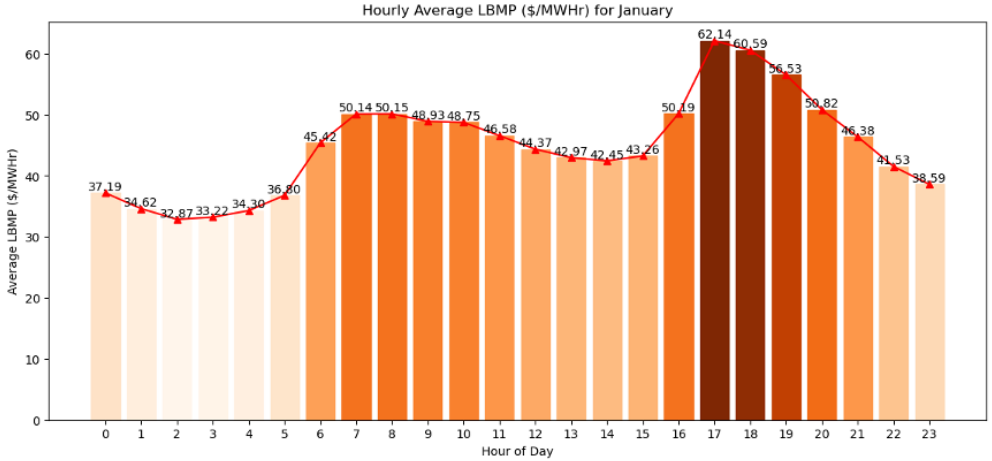


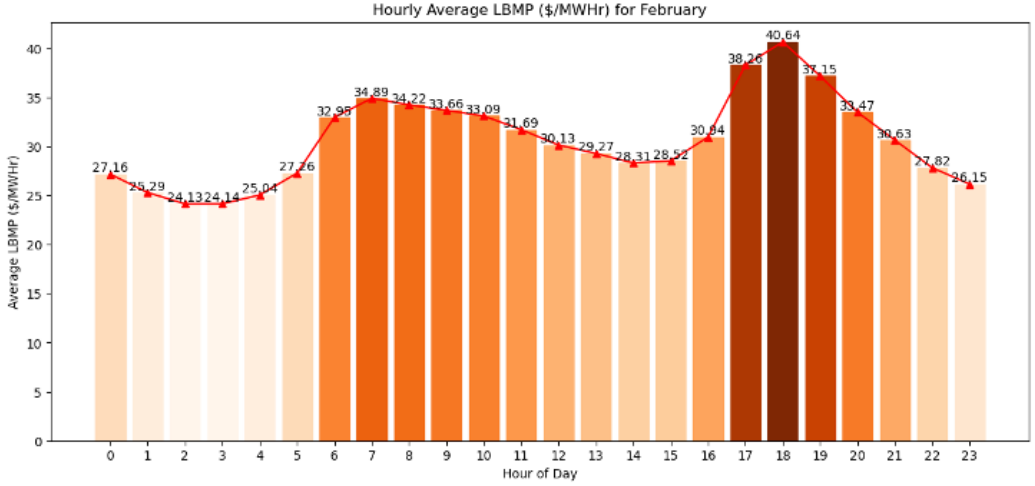
* Outliers were not removed, intentionally.

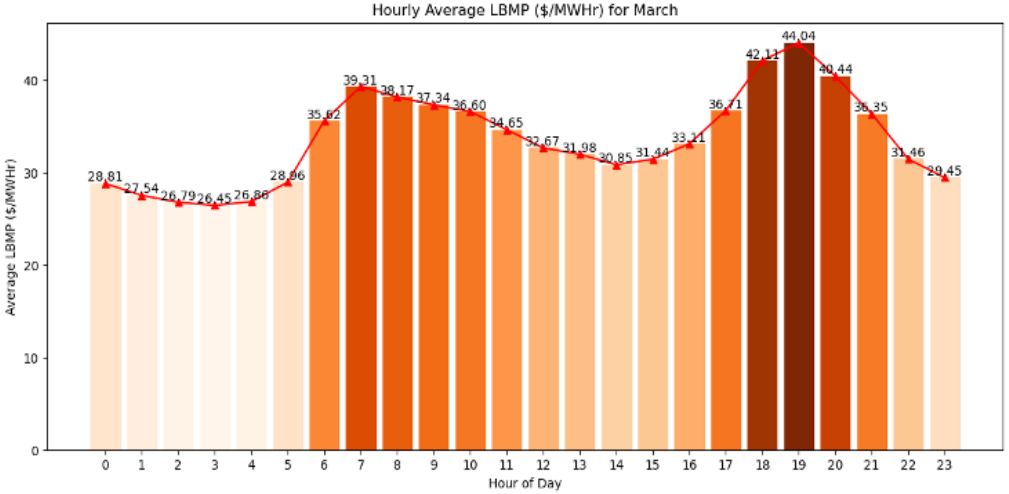
PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Exploratory Data Analysis – Average LBMP distribution by each hour and month









PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Assumptions and Constraints

Assumptions:

- Unfamiliarity with NYISO = keep models simple
- Focus on hour ahead forecasting
- Aggregate at simple levels (hourly mean, median, std)
- Runtime should be under 1 hour for hour ahead forecasting
- Time series forecasting models to focus on: ARIMA and SARIMA

Constraints:

- Repetition of Time Steps across generators
- Size of data set
- Data leakage in feature generation
- Computing power
- Forecasting more than 1 hour ahead

Aggregation

Mean, Median, and STD

- Mean: Measure central tendency while retaining information about the overall trend of the data
- Median: Provides a more robust measure of central tendency in the presence of extreme values/ spikes
- Std: Identifies deviations from central trend, providing a more robust forecast in terms of spikes and seasonal patters

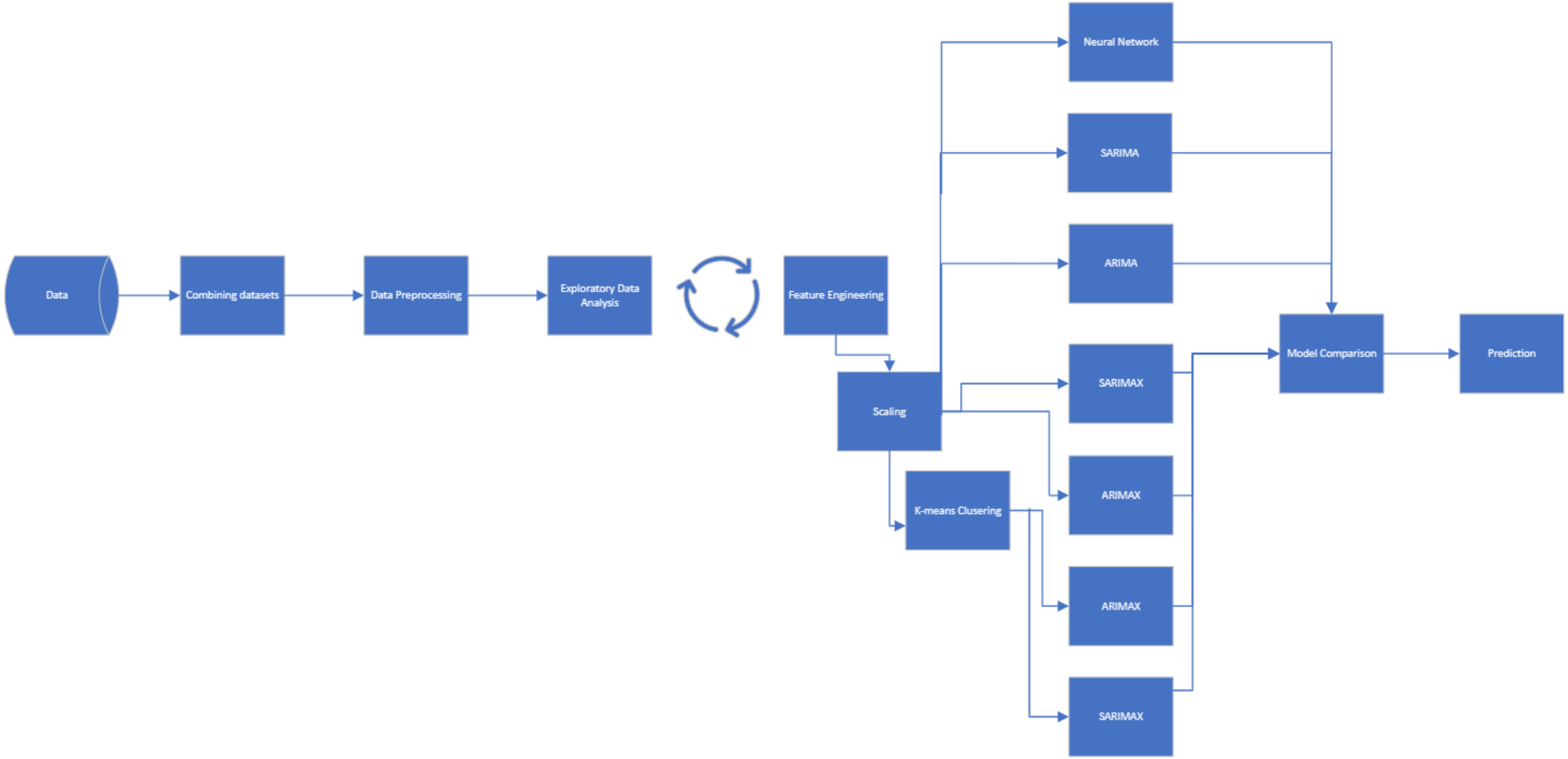
Feature Generation and Selection

- Features are generated and combined before aggregation as a representation of pre-aggregated data
- Generated/ Added Features EIA and NYISO:
  - Fuel Prices: EIA
  - Power Plant Costs: EIA
  - Transmission and Distribution Costs: NYISO
  - Weather Conditions: Iowa State ASOS Network
  - Regulations: NYISO
  - Self Generated Features: Lags for 1-168 hours, rolling means of 3, 12, 24 hours, seasonal, monthly, weekly, and hourly features

Modeling Steps

1. EDA and Correlation Analysis
2. Feature Generation
3. Aggregation
4. Recursive Feature Selection
5. Grid Search For Best Parameters
6. Graph Results
7. Evaluate Results

Methodological Workflow Sketch



AR Best Models

Mean Agg.

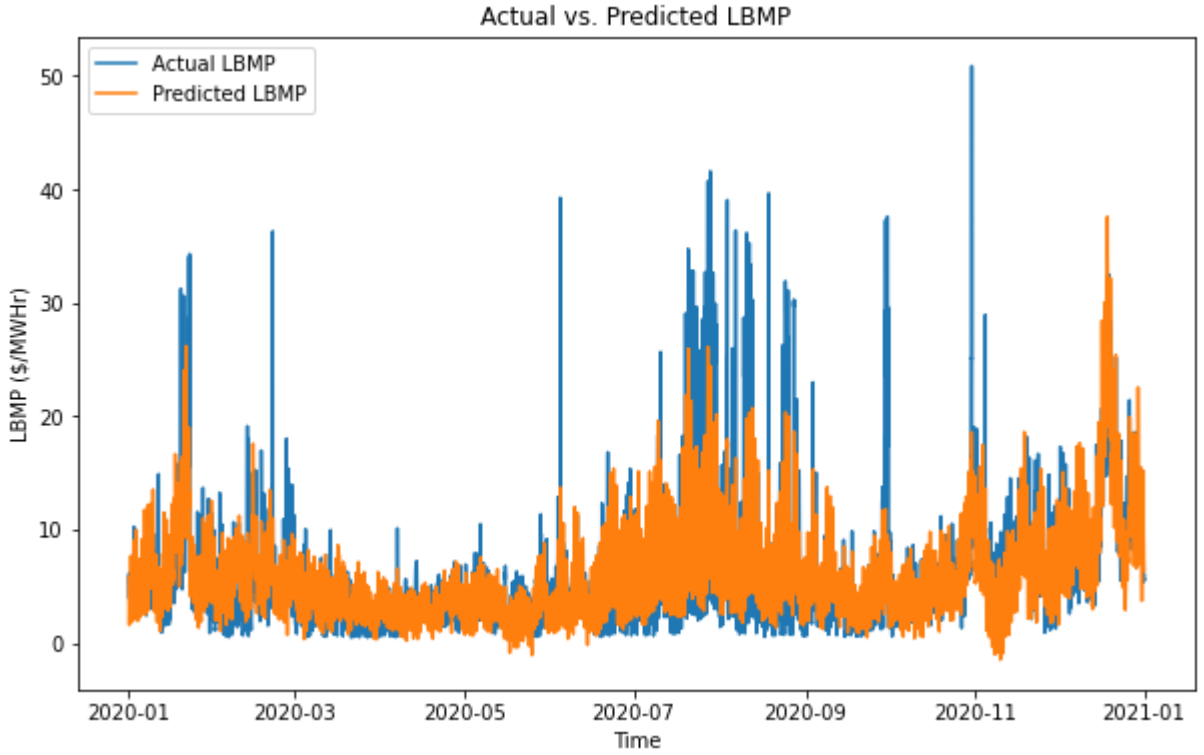
Runtime	<1 Second
RMSE	0.129
MSE	0.017
MAPE	0.051

Median Agg.

Runtime	<1 Second
RMSE	0.253
MSE	0.064
MAPE	0.411

STD Agg.

Runtime	<1 Second
RMSE	2.637
MSE	6.952
MAPE	34.813



ARIMA Best Models

Mean Agg.

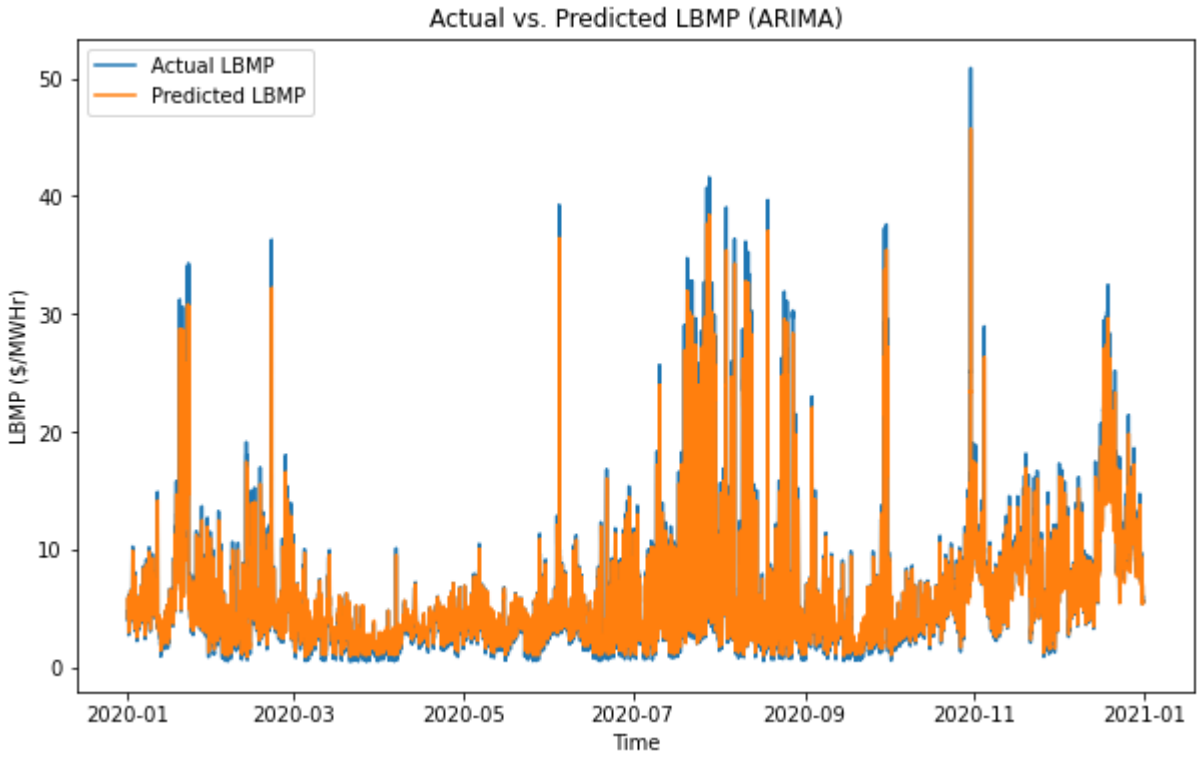
Runtime	1.3 Minutes
RMSE	0.108
MSE	0.012
MAPE	0.127

Median Agg.

Runtime	36 Seconds
RMSE	0.237
MSE	0.056
MAPE	0.393

STD Agg.

Runtime	17 Seconds
RMSE	1.463
MSE	2.142
MAPE	17.414



SARIMA Best Models

Mean Agg.

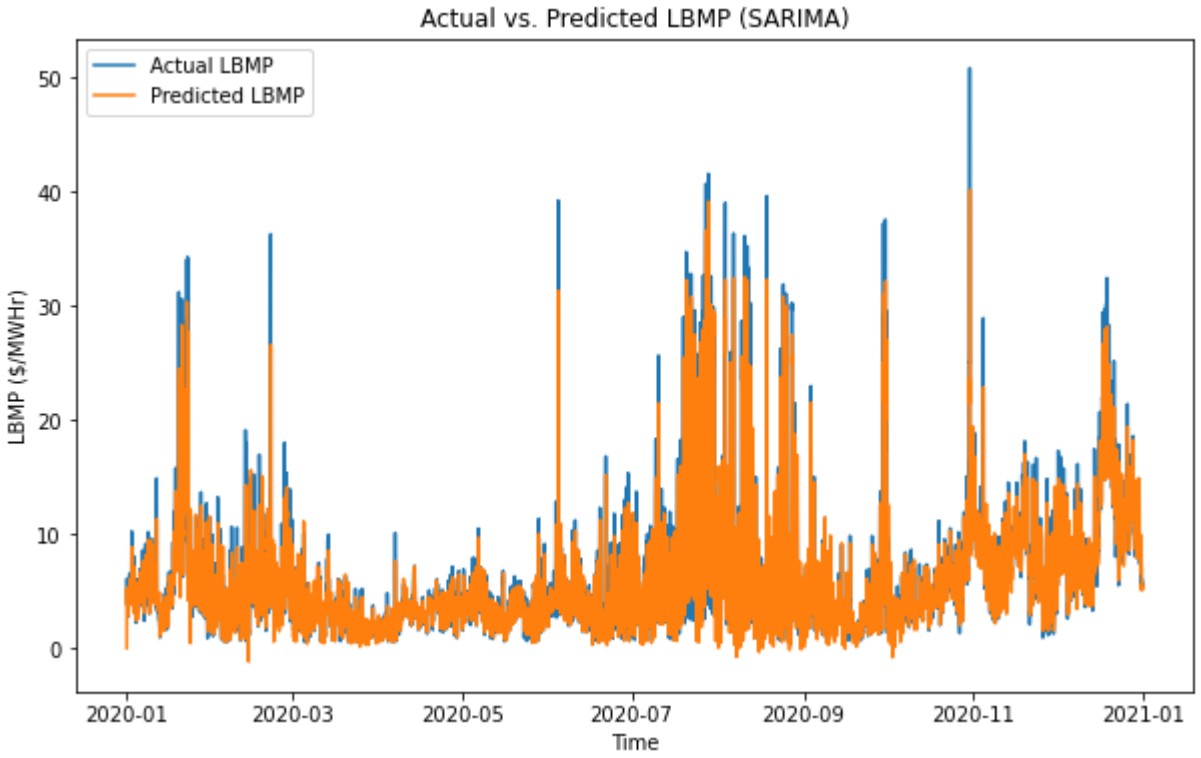
Runtime	25 Minutes
RMSE	0.111
MSE	0.012
MAPE	0.048

Median Agg.

Runtime	6 Minutes
RMSE	0.236
MSE	0.056
MAPE	0.396

STD Agg.

Runtime	14 Minutes
RMSE	1.321
MSE	1.744
MAPE	17.317



Attempts to Improve STD Forecast

Attempted Models:

1. AR
2. ARIMAX
3. SARIMAX
4. GARCH
5. ARIMA GARCH
6. SES
7. DES
8. Random Forest
9. Gradient Boosting and XGBoost
10. SVM
11. Prophet
12. LSTM

STL and ARIMAX Ensemble

Our Method

1. Deseasoning LBMP by extracting seasonal component of STL decomposition
2. Point Forecasting by running ARIMAX model on adjusted data
3. Reseasonalizing by adding seasonal component back for final forecast results

Optimization

- Selected Features: 'Marginal Cost Congestion ($/MWHr)', 'LBMP_lag_1h', 'LBMP_lag_2h', 'LBMP_lag_3h', 'LBMP_lag_22h', 'LBMP_lag_23h', 'LBMP_lag_24h', 'LBMP_lag_2d', 'rolling_mean_3h', 'rolling_mean_12h'
- Decomposed hourly (Period = 1)
- Estimation Method: hannan rissanen- allows for convergence
- order=(2, 0, 0), trend='n'
- Potential Overfitting: Need more data to test on

STL and ARIMAX Ensemble

Mean Agg.

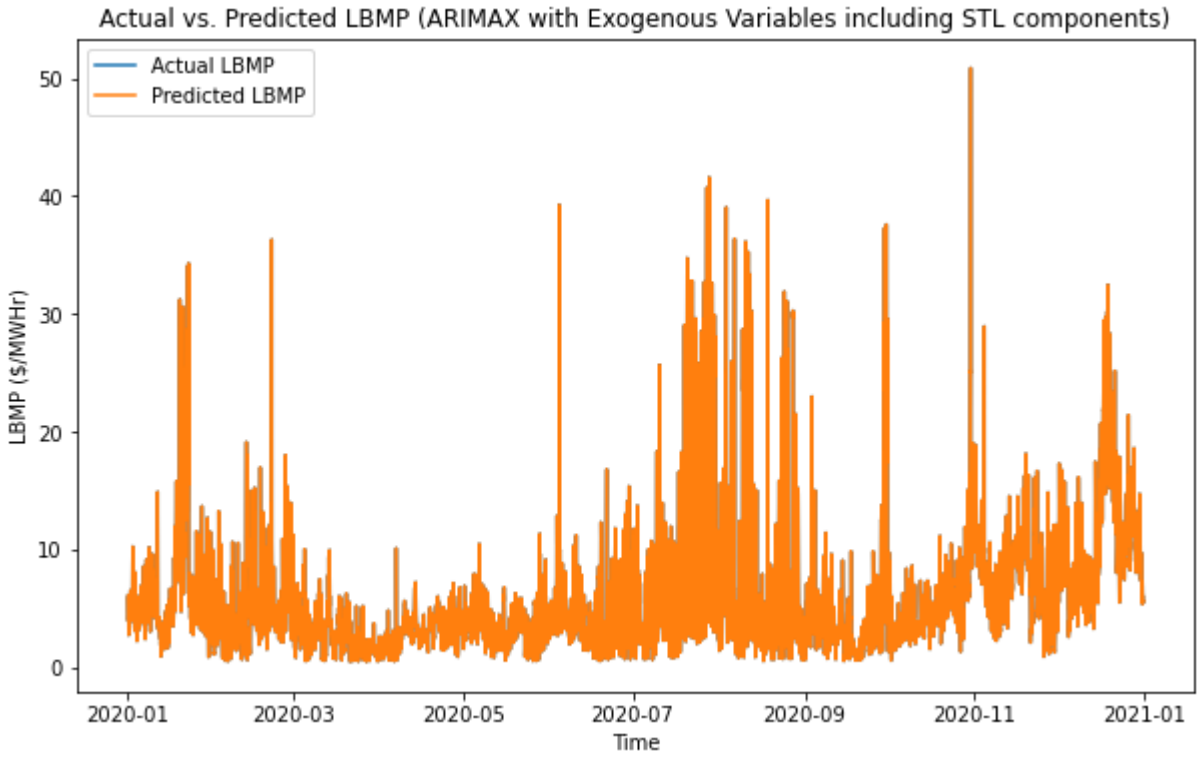
Runtime	<1 Second
RMSE	3.1428E-14
MSE	9.877E-28
MAPE	1.3004E-13

Median Agg.

Runtime	<1 Second
RMSE	3.1428E-14
MSE	9.877E-28
MAPE	1.3004E-13

STD Agg.

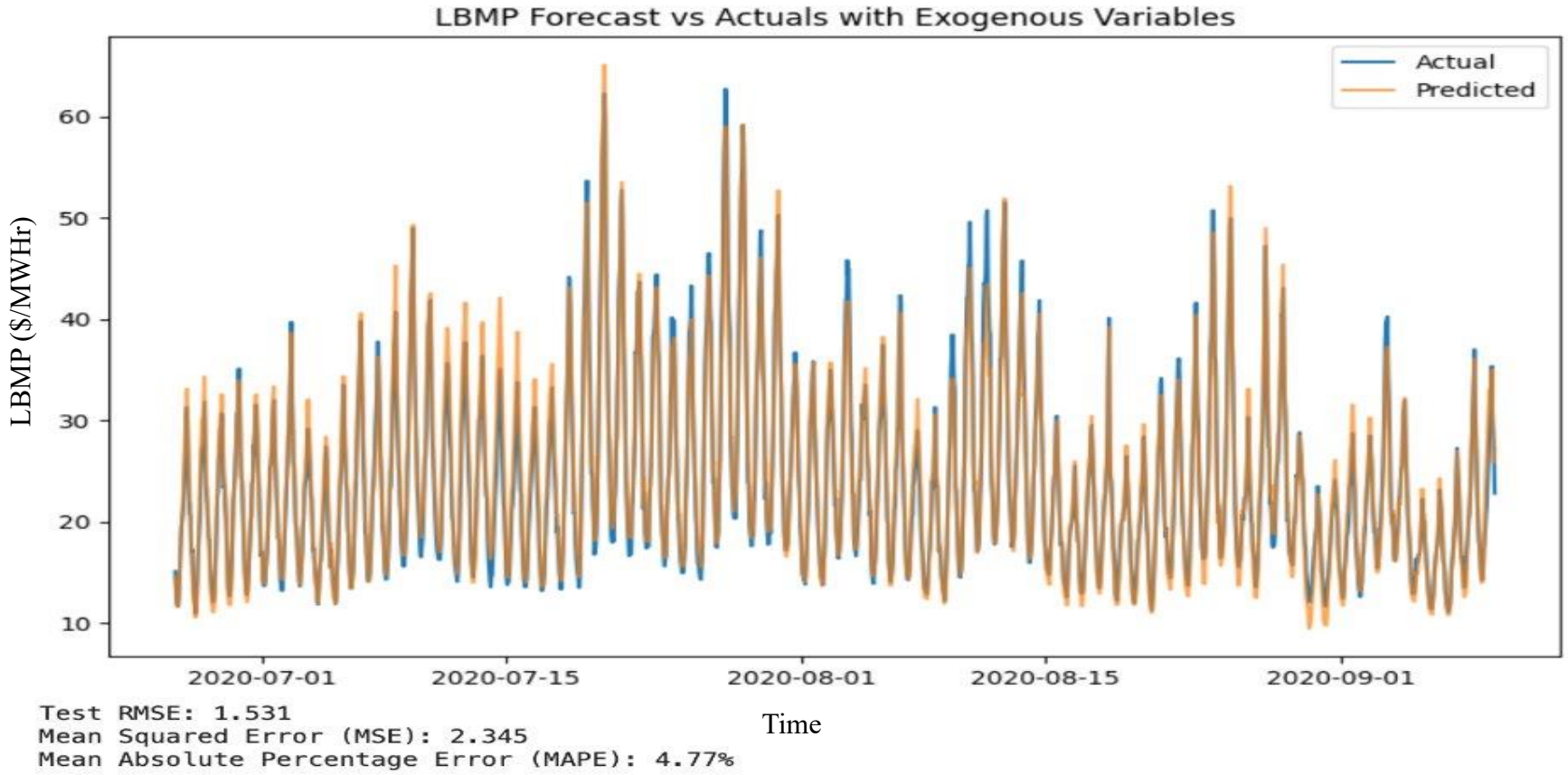
Runtime	<1 Second
RMSE	3.3529E-15
MSE	1.1242E-29
MAPE	5.2932E-14



SARIMAX: No Rolling Window

- **Exogenous Variables: MCC, MCL**

- **100% of 2020 q1-4 data**
- **Preprocessing frequency of time to aggregate hourly**
- **Preprocessing for NaN values of time**
- **70/30 data split**

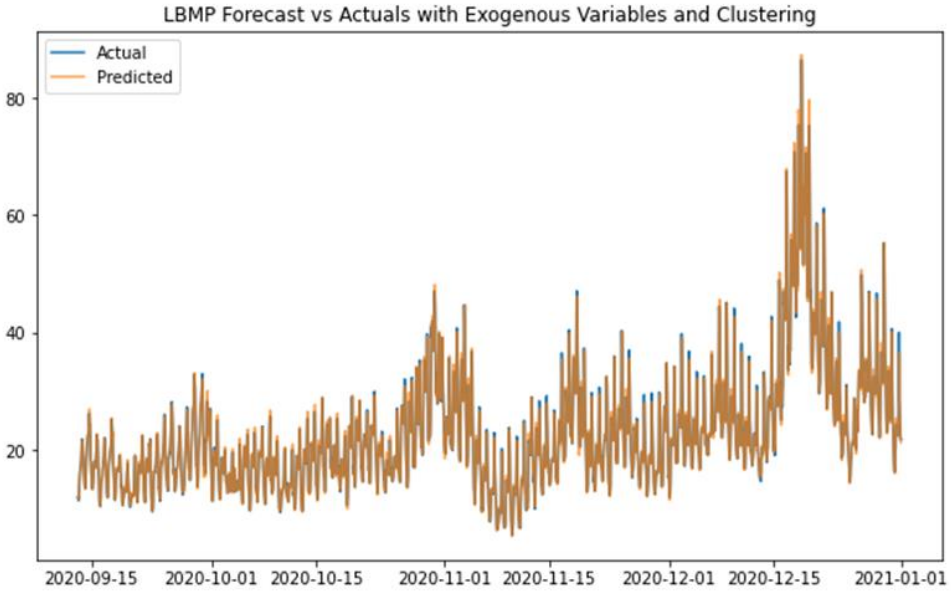


SARIMAX: No Rolling Window

- **Exogenous Variables: MCC, MCL, rolling_mean_3h, LBMP_lag_2h**

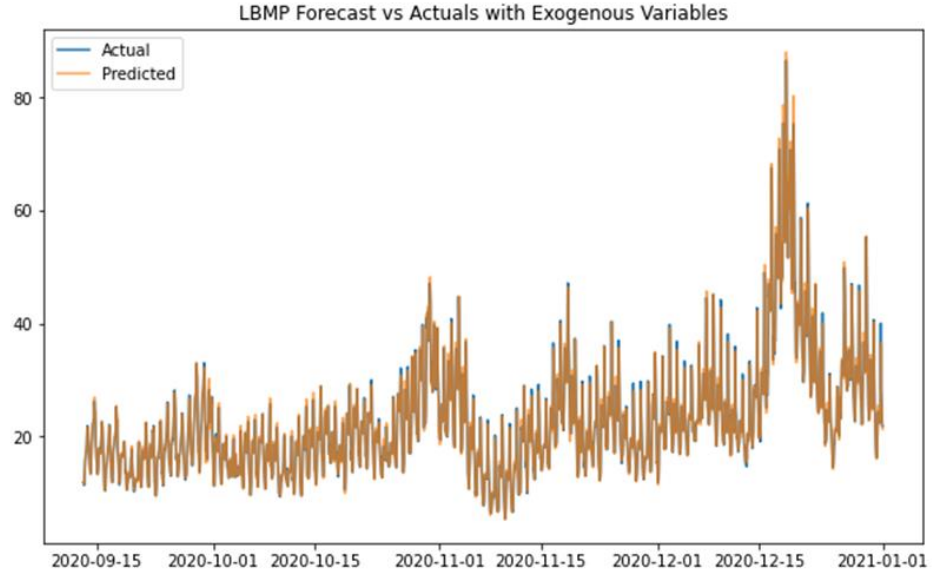
K-Means

Runtime	44.97 seconds
RMSE	0.904
MSE	0.817
MAPE	2.35%



Without K-Means

Runtime	41.71 seconds
RMSE	0.914
MSE	0.835
MAPE	2.36%



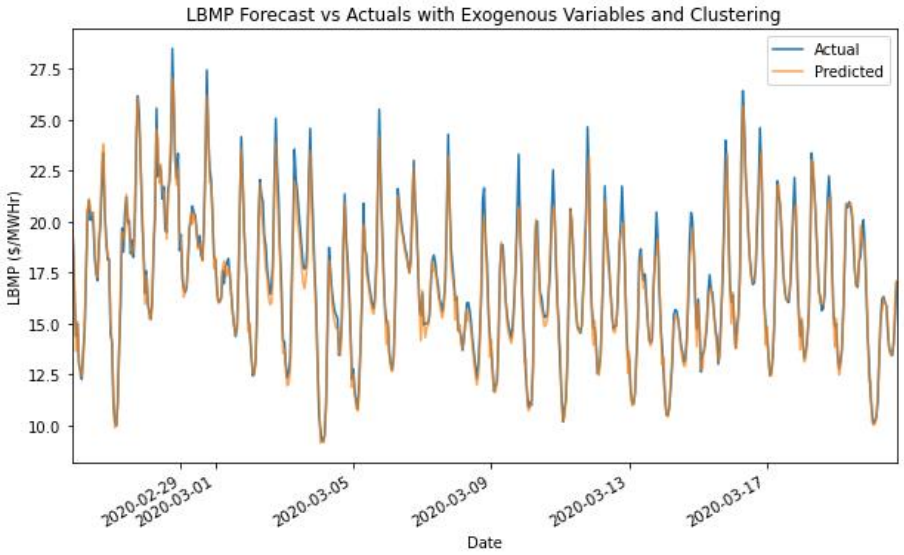
PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

SARIMAX: Rolling Window 168h, Additional Exogenous Variables

- **Exogenous Variables: MCC, MCL, rolling_mean_3h, LBMP_lag_2h**

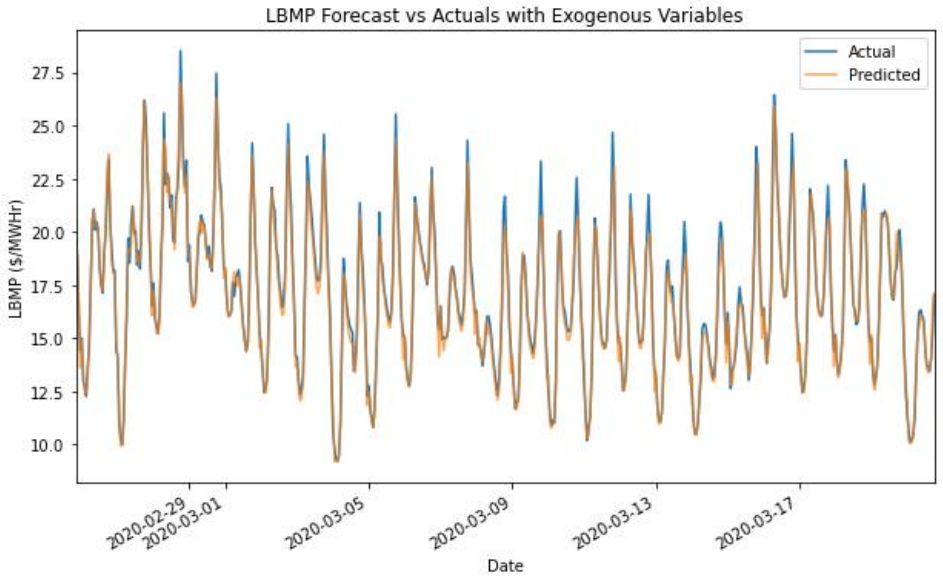
K-Means Clustering

Runtime	48.16 seconds
RMSE	0.581
MSE	0.338
MAPE	2.41%



Without K-Means Clustering

Runtime	22.03 seconds
RMSE	0.565
MSE	0.319
MAPE	2.28%

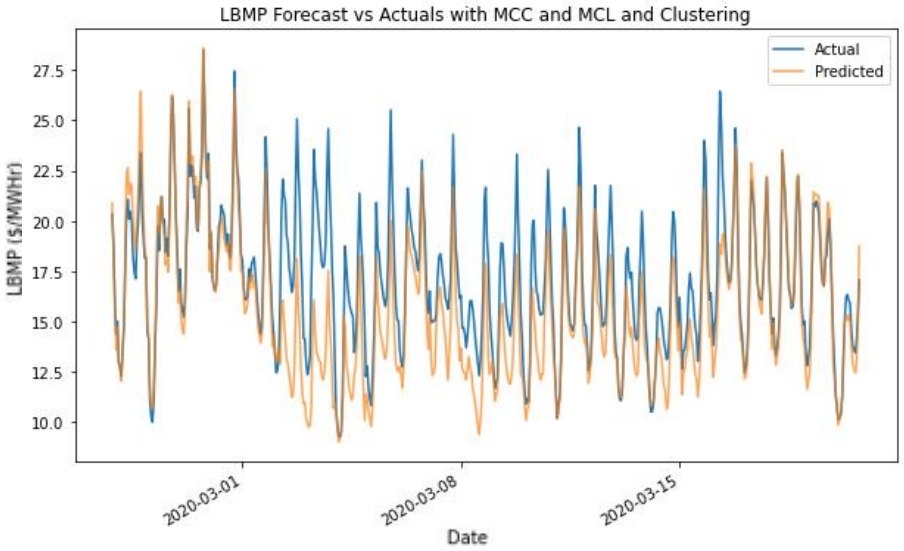


PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

SARIMAX: Rolling Window 168h, Baseline Exogenous Variables

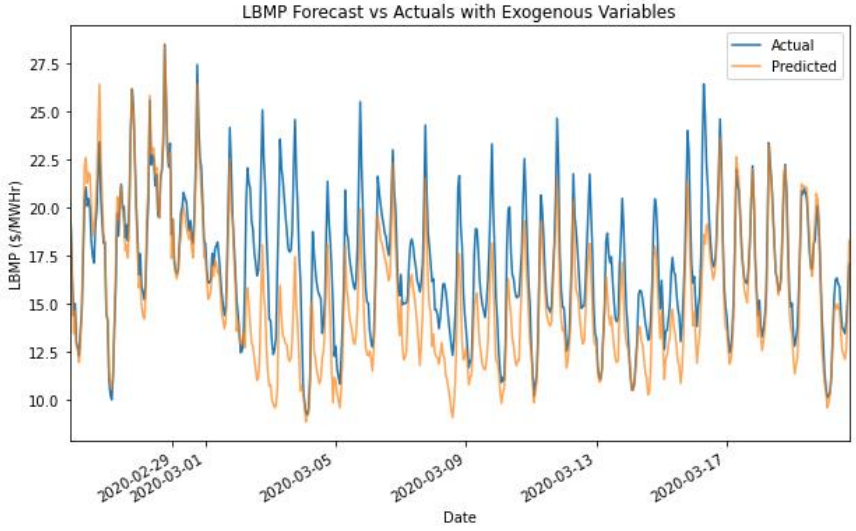
K-Means Clustering

Runtime	33.4 seconds
RMSE	2.348
MSE	5.511
MAPE	9.65%



Without K-Means Clustering

Runtime	16.15 seconds
RMSE	2.501
MSE	6.254
MAPE	10.73%

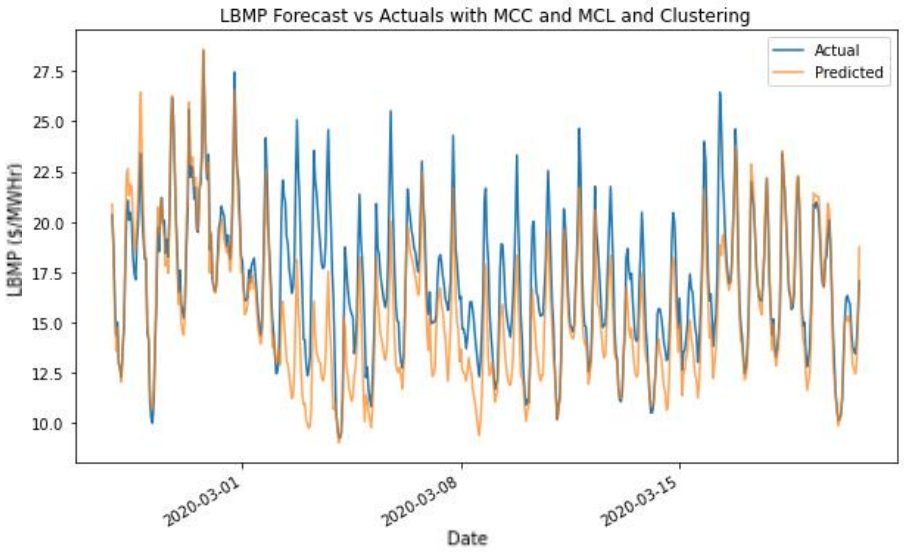


PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

SARIMAX: Rolling Window 168h, Baseline Exogenous Variables

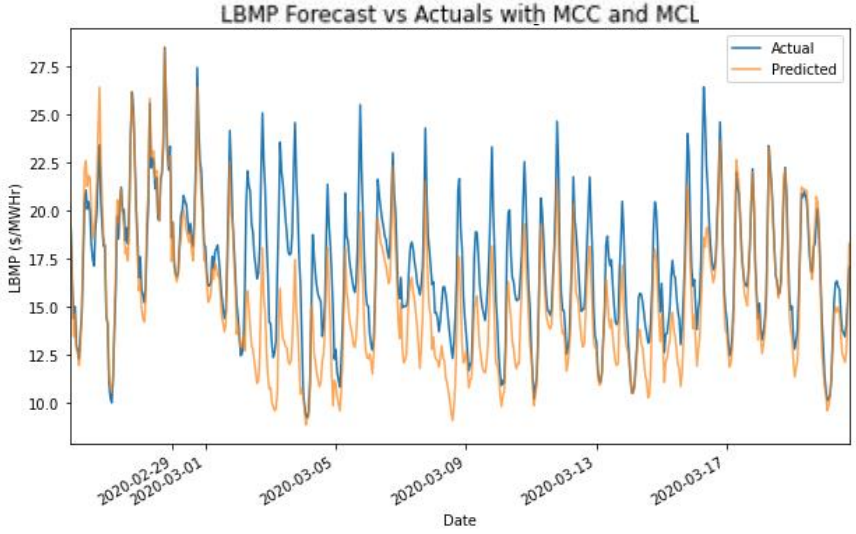
K-Means Clustering

Runtime	33.4 seconds
RMSE	2.348
MSE	5.511
MAPE	9.65%



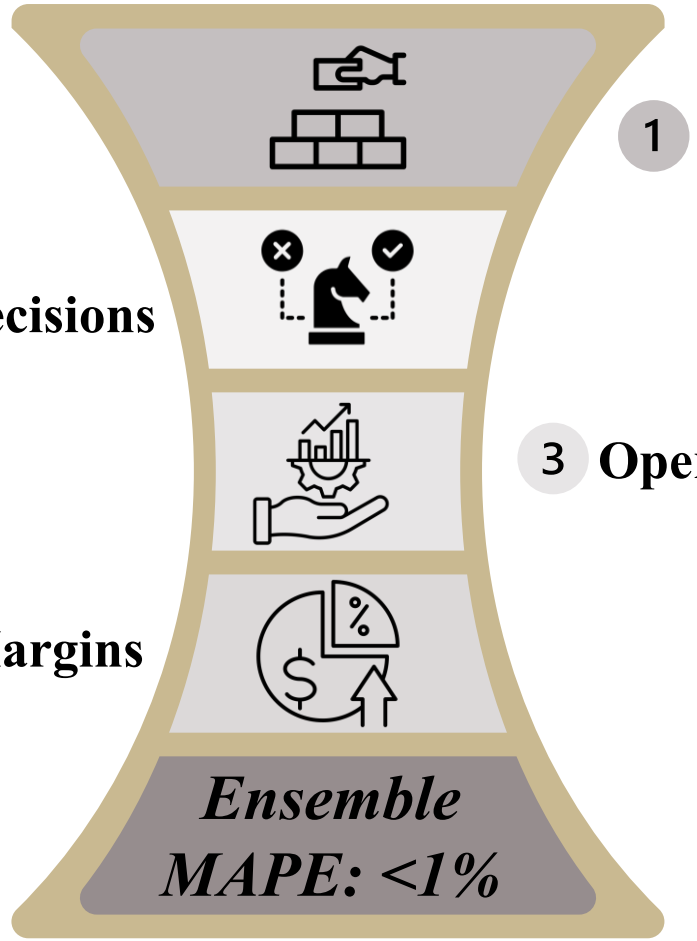
Without K-Means Clustering

Runtime	16.15 seconds
RMSE	2.501
MSE	6.254
MAPE	10.73%



PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Conclusion


1 Solid Foundation
2 Strategic Decisions
3 Operational Efficiencies
4 Profit Margins
Ensemble
MAPE: <1%

Thank You

We would like to answer your questions if you have.