**ELECTRICITY SALES FORECASTING WITH EXTERNAL VARIABLES USING MACHINE LEARNING MODELS**

By Racheal Wanyoike

**ABSTRACT**

Forecasts on electricity consumption are becoming an essential component of sustainable energy management as the demand for electricity increases throughout the world. Traditional and machine learning methods, when applied, provide accurate predictions of future consumption. However, the presence of certain exogenous factors such as weather conditions and electricity prices complicate the prediction task and produce poor predictions. Models such as SARIMAX, Support Vector Machine, Random Forest have been known to capture relationships between variables and generalize well with data that contains exogenous variables. Making use of these models, electricity sales were forecasted using temperature, electricity price and unemployment rate as exogenous variables. The results show that SARIMAX has a prediction performance when temperature, price and unemployment rate are applied separately, with temperatures producing the best performance. When all variables are applied to the sales data, random forests had the best prediction performance.

**INTRODUCTION**

Over the years, the electricity industry has undergone dramatic changes, including the inclusion of renewable energy sources, demand management and changes in consumer behavior. All of these issues increase the importance of understanding electricity consumption, especially future trends. Accurate forecasting of electricity usage and demand is an important part of power system operation, ensuring effective management of utility operations at all stages of generation, transmission, and distribution grids. This forecasting ability plays an important role in various parts of power system operations, including day-to-day operations, production planning, economic dispatch, and unit commitment.

Daily purchase and sales forecasts provide an important input to operational supply and demand balancing, as well as financial management of the utility. Furthermore, they are an important input into the management of electricity losses - particularly non-technical losses (Minaar, M, & M, March,2023). Electricity sales and consumption are influenced by a variety of factors ranging from consumer usage to weather conditions to economic conditions. The types of benchmark models for electricity sales forecasting mainly include traditional econometric models and machine learning models with traditional econometric models being the first to be applied (Wen, et al., 2021)

The goal of this study is to investigate the impact of model complexity on predictive accuracy and generalization to different scenarios. The electricity sales will be forecasted for a short - medium horizon term, which is around one year in the future using external variables. Several approaches will be employed to time series Seasonal Autoregressive moving average with exogenous factors (SARIMAX), regularization hyperparameter tuning with Support Vector Regressor (SVR), random forest and XGBoost. The model performance will be evaluated based on evaluation metrics, RMSE, MSE and MAE. The models will incorporate external variables such as temperature and electricity price factors that have direct effect on electricity sale revenues and unemployment rate which will represent customer economic conditions.

**LITERATURE REVIEW**

The forecasting models and time horizons used vary with time horizon as follows: Very short-term (few minutes to an hour) for load and frequency control and economic dispatch functions, medium-term (week-year) for generator scheduling, and long-term (year-decade) for long term development plans (Hyndman & Fan, 2010). There are several feasible methods proposed by previous research proposed for electricity sales forecasting.

Early forecasting methods were based on traditional methods such as exponential smoothing(ES) and moving average(MA) which are simple and have proven to perform well in forecasting better than other sophisticated methods (Jiang, Fang, Spicher, Cheng, & B, 2019) In addition, autoregressive integrated moving average (ARIMA) model and the seasonal ARIMA are widely used for univariate time series forecasting. In short-term load forecasting, Cao et al. adopted a hybrid decision model based on ARIMA (autoregressive integrated moving average) and for daily load forecasting in households. The results showed that the hybrid method performed well compared to the individual basic methods (Cao, Dong, Wu, & Jing, 2015,October). The traditional methods  are simple to construct and fast to calculate, but it has significant limitations, such as the homogeneity of the variation in the regression model, changes in the season, climate, and other factors of the electricity consumption area that may cause the regression relationship to change, resulting in variability (Min, et al., 2022).

Combining two or more traditional methods is a favored approach as it can take advantage of the good parts of the various components. A study used SARIMA and STL decomposition to for a combination technique to forecast mid- and long-term monthly electricity consumption with seasonal characteristics (Zhang & Li, 2021) Multiple linear regression techniques are also detailed in the literature for residential electric energy demand forecasting. Sarduy et al applied linear and nonlinear models to estimate peak load at a University of Sao Paulo campus to select the best one for generalization (Humberto, Almendra, Cristhian, & Gabriel, 2017).

Traditional techniques have its advantages however machine learning algorithms and deep learning has shown stronger ability to capture complex patterns and work with large datasets. Wang et al. proposed a hybrid model integrating discrete wavelet transform and XGBoost for electricity consumption forecast (Wang, Shi, Lyu, & Deng, 2017) Rathore et al, focused on energy consumption prediction utilizing past charging data from EVs and modeled common machine learning (ML) algorithms such as Random Forest, XGBoost, linear Regression, ANN, and  Deep Neural Network (DNN). Random Forest and XGBoost outperformed other models including DNN (Rathore, Meena, & Jain, 2023). Guenoukpati et al. used three machine learning algorithms to forecast electricity generation in Togo by applying SVR, multilayer perceptron (MLP), and a long short-term memory (LSTM) recurrent neural network. According to their findings, machine learning methods outperform linear regression methods and are more suited for short-term predictions (Guenoupkati, Salami, Kodjo, & Napo, 2021)

Artificial network models are suitable for short-term predictions but are difficult to interpret, while multiple linear regression requires large amounts of data to predict mid- to long-term. Due to the limited data set and the existence of external variables, this article will utilize a model suitable for small data for predictive analysis, which can capture complex patterns with smaller errors and can accommodate external variables in predictions. Based on these factors and previous research, this study will apply STL decomposition with SARIMAX, SVR, XGBoost and Random Forest. Each of these models has shown a good ability to predict electricity usage, so study how they perform under weather and economic conditions.

**DATA**

Electricity demand can be studied by looking into the sales of electricity, which show the amount consumed. The dataset was sourced from multiple sources to acquire historical electricity sales consumption data, average temperatures, average electricity price, and unemployment rate for Pennsylvania from 2001 to 2022. The electricity data covers only the residential sector and was downloaded from the US Energy Information Administration. It provides monthly residential electricity sales data for the state of Pennsylvania.

The dataset consists of 264 samples with four features. The features in the dataset include:

1. Sales:  Electricity sales in million kilowatt hours (residential)
2. Ave\_temp: Average temperature of Pennsylvania (2001–2022) in Fahrenheit
3. PAUR: Unemployment Rate in Pennsylvania, Percent, Monthly, Seasonally Adjusted
4. Price: Average Price: Electricity per Kilowatt-Hour in Philadelphia-Camden-Wilmington, PA-NJ-DE-MD (CBSA), U.S. Dollars, Monthly, Not Seasonally Adjusted

Three factors were identified as potential predictors of electricity consumption: temperature, electricity price, and unemployment rate. Temperature data was obtained from a regional climate dataset, specifically targeting the Pennsylvania region. Electricity prices and unemployment rates were sourced from the FRED economic data repository, ensuring their relevance to the Pennsylvania context. The inclusion of the unemployment rate aims to capture the impact of economic conditions on the population's electricity consumption patterns.

**METHODOLOGY**

The dataset is small, and the target of the analysis is forecasting electricity sales to understand future electricity consumption in Pennsylvania including environmental and economic factors. In this study, statistical and machine learning techniques including regression analysis and SVR are applied to time series forecasting with external variables. These techniques have previously been used to predict electricity demand and sales with notable success with simple regression for simplicity and interpretability and artificial networks that capture. The reason for applying these techniques is because they can capture the influence of additional factors beyond the endogenous variable. Comparison table of statistical and AI techniques that contains pros and cons of each technique individually is indicated in Table 1.

Table 1: Comparison table of statistical and AI techniques

|  |  |  |
| --- | --- | --- |
|  | **Pros** | **Cons** |
| SARIMAX | Good for medium term forecasting.  Can include exogenous variables.  Ease to interpret.  Good for stationary time series data  Strong theoretical foundation. | Not ideal for very short- or long-term forecasting.  Requires careful selection of parameters.  May not capture complex patterns or nonlinear relationships. |
| SVR | Robust to outliers and noise.  Effective for both linear and non-linear relationships.  Can handle high-dimensional data.  Can incorporate kernel functions to capture complex patterns. | Need careful hyperparameters tuning.  Can be computationally expensive.  Not easy to interpret the model. |
| XGBoost | Can handle both continuous and categorical features.  Captures complex nonlinear relationships.  Effective for both regression tasks.  Robust to outliers and noise.  Can incorporate external variables to enhance predictive performance. | Computationally expensive.  Hyperparameter tuning is challenging.  Difficult to interpret the model. |
| Random Forest | Handles high-dimensional data.  Captures complex non-linear relationships.  Effective for both regression tasks.  Robust to outliers and noise. | Can be sensitive to correlated features.  May overfit the data.  Hyperparameter tuning can be challenging.  Difficult to interpret the model. |

1. **SARIMAX**

SARIMAX stands for Seasonal Autoregressive Integrated Moving Average with exogenous variables. It is seasonal ARIMA that includes exogenous variables. The (AR) part models the current value of the time series as a linear combination of its past values, while the (MA) part models the current value as a linear combination of past forecast errors. The (I) part handles non-stationarity in the time series by differencing the data, making it stationary and suitable for forecasting. The ARIMA model uses the differenced series (y't) instead of the original series (yt) because differencing can make the series stationary, a requirement for the ARIMA model to be valid. Differencing is subtracting a previous value of the series from the current value and can be performed multiple times as needed to achieve stationarity.

The mathematical expression of the ARIMA(p,d,q) process states that the present value of the differenced series y't is equal to the sum of a constant C, past values of the differenced series ϕpy't–p, the mean of the differenced series µ, past error terms θqϵt–q, and a current error term ϵt. The equation is as follows:

Equation 1

The order p determines how many lagged values of the series are included in the model, while the order q determines how many lagged error terms are included in the model. The order of integration is defined as d. Thus, the order of integration equals the number of times a series has been differenced to become stationary. If a series is differenced twice to become stationary, then d = 2.

When seasonal parameters are added, the model becomes SARIMA. The four new parameters in the model are P, D, Q, and m, where P, D, Q are basically seasonal counterparts of p,d,q. m is the number of frequencies observed per cycle. If data was recorded monthly, then m = 12. When exogenous variables are included, SARIMAX model is used, which is basically SARIMA with added effect of exogenous variables. SARIMAX model can be defined as:

Equation 2

And loosely as:

Equation 3

Therefore, for forecasting electricity sales, the lagged values of sales and lagged error terms will be included which will consider the impact of past values of sales observation. Also, the effect of past forecast error on the current observation. Seasonal patterns are also fitted, which is prevalent is electricity sales and the term will consider the effect of the included external variables.

1. **Support Vector Regressor (SVR)**

Support vector machine is a well-known classification algorithm; however, it can perform regression tasks as well. SVR is a non-linear regression technique that uses SVM to model relationship between the variables. It finds an appropriate line or hyperplane in a higher dimension to fit the data. SVR minimizes coefficients - the l2-norm of the coefficient vector. The error term is handled by using constraints called the maximum error, ϵ. This error tolerance allows some deviation of the data points from hyperplane while being counted as errors. The hyperplane is the best fit possible for data that falls with this maximum error.

The model parameters will be chosen based on best performance using gridsearchCV, which finds optimal hyper-parameters and hence improves the accuracy/prediction results. The SVR model will train on the electricity sales and external variable data, captures non-linear relationships present and the predicted values of sales will fit within the margin of error ϵ, inside the hyperplane.

1. **XGBoost**

XGBoost refers to extreme gradient boosting, is a fast, scalable gradient-boosting decision tree learning library. It provides parallel tree boosting learning and is popular for regression and classification problems. It applies ensemble learning where it combines multiple weak learners such as decision tree to make a stronger predictor. Gradient boosting sequentially adds new weak learners (decision trees) to the ensemble, each weak learner is trained to minimize the residual error, which is the difference between the predicted sales values and actual sales values. Each decision node leaf represents a prediction. This process will be repeated until desired accuracy is achieved.

XGBoost theory begins by identifying a training set with any number of factors and the target variable, y, along with a loss function and regularization term. The loss function compares the actual and predicted values of a target variable. The learning rate, which ranges between 0.1 and 0.3, describes how much new models learn from previous ones. This helps prevent overfitting. Next the XGBoost is initialized with constant value:

Equation 4: XGBoost

Argmin refers to the areas where the loss function is minimized, i.e., where the prediction error is the smallest (Le, Oktian, & Kim, 2022). θ is any unique value that acts as the first estimation value for the regression method. The estimation error with θ decreases with each additional iteration of the model.

The gradients and Hessian matrices for tree gradient boosting are then computed. The gradients demonstrate how the loss function changes with each unit change in feature value. The hessian determines how much the gradient changes, and thus how much the model will change. Each of them is required for the gradient descent procedure (Le, Oktian, & Kim, 2022). With these matrices, another tree is added by completing optimization problem in each iteration of the algorithm:

Equation 5: XGBoost (Le, Oktian, & Kim, 2022)

A Taylor approximation is used for optimization. That is, it will be able to calculate the model's current position in the gradient boosting process. If the gradient's rate of change is steep, the residuals are large, and the model requires significant change. If the rate of change is constant, the model is almost complete. The new tree added in each iteration is denoted by , and the learning rate determines how much the models must change. The model is refreshed with the new tree, and the process is repeated for each weak base learner. The final output of the XGBoost is expressed as the sum of each individual weak learning approach, *m* as weak learner:

Equation 6: XGBoost final (Le, Oktian, & Kim, 2022)

1. **Random forest Regressor**

Random forest is an ensemble of trees where each tree is slightly different from the others. There are two ways in which the trees in a random forest are randomized: by selecting the data points used to build a tree and by selecting the features in each split test**.** For regression tasks, the algorithm makes a prediction for every tree in the forest. The predicted values result is the mean of individual tree predictions. This is presented by the diagram below:

A diagram of a tree

Description automatically generatedFigure 1: Random Forest Structure

Random forest algorithm follows the following steps:

1. Randomly pick *k* data points from the training data set.
2. Build a decision tree based on these *k* data points.
3. Choose N, which is the number of trees to want to build and repeat step 1 and 2.
4. To make a prediction for a new data point, it passes through each tree in the forest. The individual predictions from all trees are then averaged to obtain the final prediction.

**Regression Performance Metrics**

When evaluating the performance of a regression model, the key concepts involved are the predicted outcome and the real outcome. The performance is reported as an error which is the difference between the predicted outcome and the real outcome. In regression analysis, the commonly used error metrics are: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE).

a) Mean Squared Error

MSE calculates the average of the errors and squares the results. This makes most MSE assume positive values. MSE has the same measurements as the square of the values being estimated. However, since it takes the average of the errors, it is prone to outliers. Therefore, outliers can affect the results greatly.

b) Root Mean Squared Error

RMSE calculates the square root of MSE. This helps provide an error measure that has the same measurement as the target variable. A perfect RMSE is usually zero however, that never is the case, therefore a good RMSE depends on the data.

c) Mean Absolute Error

MAE calculates the mean of the absolute error values. This helps distribute the values and no weight is given on larger values. The scores increase linearly with increase in error values. A perfect MAE is 0, meaning the predicted is equal to the expected outcomes.

For this analysis, a baseline for these metrics will be established with naïve regression model.

**DATA PREPROCESSING**

To prepare the data collected for time series analysis, pre-processing techniques were applied to the time series data. Data cleaning was performed to check for missing values, outliers and renaming of variables for better understanding. Data exploration was performed to gain understanding of the summary statistics and distribution of the time series, correlation relationships between electricity sales and the external variables. The relationship between residential electricity consumption and external factors is often not linear. Based on Pearson correlations, the external variables have a moderate linear relationship with electricity sales. There is presence of non-linear relationships therefore the SVR, XGBoost and Random Forest regressor will detect and account for the non-linear relationship.

1. **Time Series Components**

Stationarity is important to time series, it implies that the property of time series where its statistical properties, such as the mean, variance, autocorrelation, and covariance, do not change over time. This means that the time series exhibits a consistent pattern or behavior throughout its duration. STL decomposition and ADF tests can help identify stationarity. STL decomposes a time series into trend, seasonal and residual components. The long-term shift in the time series is represented by the trend component. This component oversees fluctuation in value over time. The periodic pattern in the time series is the seasonal component. It shows recurrent fluctuations that take place over an established length of time. In conclusion, the noise, also known as the residuals, represents any irregularity that cannot be accounted for by the seasonal component or the trend. STL decomposition and ADF tests to identify stationarity. If not present, time series were differenced and logged to achieve stationarity. In some instances, time series residuals were applied to modeling.

**a) Electricity consumption monthly sales.**

There is a slight presence of trend and seasonality in the sales series. To decompose the time series data into its components, STL (STL: Seasonal-Trend Decomposition Procedure Based on Loess) was employed. This method performs better than additive/multiplicative decomposition. It can handle any type of seasonality and can also control the rate of change of the seasonal component to better match the seasonal characteristics of our data.

ADF (Augmented Dickey-Fuller test) was performed to test for stationarity and the sales series was stationary. The ADF results showed a negative test statistic, small p-value, and stationarity was present since the test statistic is less than the critical values at various levels of significance.

The sales series was log transformed to normalize and stabilize the variance. Also, the residuals of the sales series were used in the analysis, to capture the noise and complex patterns excluding the effects of trend and seasonality. Residual modeling was applied to remove systemic patterns and capture unmodeled dynamics.

**b) Average price of electricity**

There is a slight trend and seasonality in the series. The price series is non-stationary based on the ADF test. Log-transformation and differentiation was performed but the ADF test still presented non-stationary state. Maintained original variable form, chosen model of use will have to consider non-stationary variables such as VAR. It is also linearly correlated with electricity sales.

**c) Average temperature**

Average temperature is linearly correlated to sales and its significant based on p-values. There is presence of trend and seasonality. Stationarity test was performed and confirmed stationarity. The test statistic of ADF test was less than the critical value (5%).

**d) Unemployment rate**

The rate of unemployment is not linearly correlated with sales of electricity; however, it is stationary. Hence it can be forecasted if included in predictive models. There is no trend or seasonality present.

**RESULTS AND DISCUSSION**

This paper aims to analyze the impact of external variables monthly average temperature, average electricity price and unemployment rate on forecasting future electricity sales of the state of Pennsylvania. The prediction relationships of these exogenous variables were observed by fitting the data to SARIMAX, SVR, XGBoost and Random Forest models and evaluate their predictive performance based on regression metrics.

The average monthly temperature has a moderate negative relationship with electricity sales. It's not an extremely strong negative correlation, but it does indicate a discernible tendency for the two variables to move in opposite directions. As the average temperature increases, electricity sales tend to decrease and vice versa. Based on the pairwise correlation coefficient, the p value < 0.05 suggests there is a significant correlation between electricity sales and average temperature. Average monthly electricity price has a moderate positive relationship with electricity sales. It a moderate positive correlation, and based on the p-value, the relationship is significant showing a relationship between electricity price and sale. As for the unemployment rate, it has a very small correlation, however the p-value shows that this small correlation is not statistically significant. However, Pearson correlation only accounts for linear relationship, therefore, it was included in the analysis to find any non-linear relationship that may affect electricity sales indirectly or directly.

1. **Monthly electricity sales forecast with average temperature**

The average monthly temperature has a moderate negative relationship with electricity sales. It is not an extremely strong negative correlation, but it does indicate a small tendency for the two variables to move in opposite directions. As the average temperature increases, electricity sales tend to decrease and vice versa. Based on the pairwise correlation coefficient, the p value < 0.05 suggests there is a significant correlation between electricity sales and average temperature. SARIMAX model performed best in the predictive performance accuracy based RMSE, MSE, MAE values. Below is a table with the predictive performance accuracy of the models applied:

Table 2: Electricity forecast with Average monthly temperature as external variable

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RMSE** | **MSE** | **MAE** |
| **SVR** | 0.0941 | 0.0732 | 0.0088 |
| **SARIMAX** | 0.0704 | 0.050 | 0.0089 |
| **XGBoost** | 0.0999 | 0.0782 | 0.01 |
| **Random Forest** | 0.099 | 0.0771 | 0.0098 |

Based on the SARIMAX results summary, the model effectively captures the seasonal and autoregressive patterns in the log of sales data. The negative coefficient for average temperature suggests that temperature plays a role in influencing sales, with warmer months associated with lower sales. The positive coefficient for autoregressive terms (AR) indicates that the electricity sales are positively correlated with their own value lagged by one month. There is also a seasonal pattern in the sales, with sales tending to be higher in the same month of the year compared to previous years. The positive coefficient for moving average terms (AM) indicates that a positive shock to the log of sales in one month is likely to be followed by a persistent positive effect on the log of sales in the following months.

The diagnostics checks; the Ljung-Box (LB) and Jarque-Bera (JB) tests assess the adequacy of the model. The LB p-value of 0.59 and the JB p-value of 0.84 suggest that there is no significant autocorrelation or non-normality in the residuals, indicating that the model fits the data well. The heteroskedasticity test, with a p-value of 0.63, indicates that there is no significant heteroskedasticity in the residuals, meaning that the variance of the error term is constant over time.

Overall, the model shows that an increase in temperature leads to a decrease in the log of sales. This could be due to lower energy consumption for heating during warmer months.

1. **Monthly electricity sales forecast with average electricity price**

Average monthly electricity price has a moderate positive relationship with electricity sales. It has a moderate positive correlation, and based on the p-value, the relationship is significant showing a relationship between electricity price and sales. SARIMAX model performed best in the predictive performance accuracy based RMSE, MSE, MAE values. Below is a table with the predictive performance accuracy of the models applied:

Table 3: Electricity forecast with Average monthly price as external variable

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RMSE** | **MSE** | **MAE** |
| **SVR** | 0.1855 | 0.0344 | 0.1626 |
| **SARIMAX** | 0.0715 | 0.0051 | 0.0579 |
| **XGBoost** | 0.176 | 0.0309 | 0.1517 |
| **Random Forest** | 0.1873 | 0.0351 | 0.1648 |

Based on the SARIMAX model results, there is a negative coefficient for average price differenced suggests that an increase in electricity price leads to a decrease in electricity sales. This could be due to price sensitivity among consumers. The model shows no presence of autocorrelation or non-normality in the residuals based on the Ljung-Box (LB) and Jarque-Bera (JB) tests that assess the adequacy of the model. This indicates that the model fits the data well. From the positive AR terms, the log of electricity sales is positive correlated with its own lagged value by one month. There is also a seasonal pattern accounted for in electricity sales with higher sales in the same month of the year. The negative coefficients of MA terms suggest positive sales in one month will likely be followed by a negative effect in sales in the same month the following year.

However, the p-value is > 0.05 suggesting price is not statistically significant as a coefficient in this model. SARIMAX assumes a linear relationship between price and sale, however, it seems that the relationship between price changes and sales may be nonlinear. For example, small price changes may have a small negligible impact on sales while large price changes have more impact on electricity sales.

When residual modeling is used, that is residuals for electricity sales and price, XGBoost model performed the best. This shows that XGBoost captures the non-linear relationship between sales and price and the noise in the data, improving the predictive accuracy.

1. **Monthly electricity sales forecast with unemployment rate**

The unemployment rate is used to assess possible economic conditions that may affect monthly electricity sales forecast. The SARIMAX model had the best performance predictive accuracy and effectively captures the seasonal, autoregressive, and moving average patterns in the log of sales data. The table below shows predictive performance accuracy of the various models applied.

Table 4: Electricity sale forecast with only unemployment rate as external variable

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RMSE** | **MSE** | **MAE** |
| **SVR** | 0.175 | 0.0306 | 0.1512 |
| **SARIMAX** | 0.0706 | 0.0306 | 0.058 |
| **XGBoost** | 0.176 | 0.0309 | 0.1517 |
| **Random Forest** | 0.1812 | 0.03285 | 0.1619 |

Based on the SARIMAX model results, the positive coefficient for unemployment rate suggests that the unemployment rate plays a role in influencing sales, with higher unemployment rates leading to lower sales. However, the P-value indicates insignificance of unemployment rate effect on sales however there is possible nonlinear relationship between the sales and unemployment rate. The positive coefficients for AR terms indicate the persistence of sales patterns over time and across seasons. The presence of moving average terms suggests that shocks to the log of sales have a persistent effect on future sales. The diagnostic checks on model adequacy confirm no significant autocorrelation in the residuals indicating model fits well with data.

The model suggests unemployment rate plays an indirect role in influencing electricity sales, for example, due to reduced consumer spending during economic downturns.

1. **Forecasting electricity sales using all the external variables combined.**

Average temperature, electricity price and unemployment rate were included as external variables to SARIMAX, SVR, XGBoost and Random Forest regressor models to forecast electricity sales. The SARIMAX model performed best when temperature and price were included in the model. Random forest regressor had the best predictive accuracy when unemployment rate and temperature were included same as when rate and electricity price were included. Random forest regressor also had the best predictive accuracy when all the external variables were included. Of all these variations, random forest regressor model with all external variables had the best overall with a RMSE, MSE and MAE: 0.0256, 0.0213, 0.0007.

Figure 2: Line chart of predicted values against the actual sales.

Random forest works well since it captures complex, non-linear relationships between predictor variables (price, temperature, unemployment rate) and the target variable (sales). It can account for interactions between variables. Based on feature importance analysis, temperature is the most important feature followed by electricity price and unemployment rate. It is important to note that the feature importance scores do not necessarily reflect the causal relationships between the external variables and the sales.

It is very clear that average temperature, electricity price and unemployment rate affect electricity sales forecast. We know that SVR, SARIMAX, Random Forest are known to be effective models to forecast monthly electricity sales from previous studies. These results show that Random Forest regressor works well on electricity sales forecast for medium horizon (1-2 years) when temperature, electricity price and unemployment rate are included. XGBoost regressor didn’t produce good results and that can be due to lack of proper fine tuning of the hyperparameters, and its complexity may not work well with the data.

**CONCLUSION**

The analysis of electricity sales forecasting with various external variables using different regression models provides valuable insights into the factors influencing sales patterns. The SARIMAX model effectively captures the seasonal and autoregressive patterns in sales data, revealing a negative correlation between temperature and sales. Warmer months are associated with lower sales, likely due to reduced energy consumption for heating. The SARIMAX model incorporating average electricity price demonstrates a negative correlation between price changes and electricity sales. The SARIMAX model with the unemployment rate as an external variable indicates a positive correlation, suggesting higher unemployment rates are associated with lower electricity sales.

However, a nonlinear relationship may not be fully captured by the model. The model outperforms other machine learning algorithms (SVR, XGBoost, Random Forest) in terms of predictive accuracy. When residual modeling is applied to the models, XGBoost and Random Forest have better performances. These models are good at capturing the noise in the data which can improve prediction accuracy.Random Forest regressor demonstrates superior predictive accuracy (RMSE, MSE, MAE) when all external variables are considered, emphasizing the importance of incorporating multiple factors in electricity sales forecasting. Future work should be done to explore and refine the models, especially in understanding nonlinear relationships that could enhance the accuracy of the forecasting models.

# **REFERENCES**

1. Cao, X., Dong, S., Wu, Z., & Jing, Y. (2015,October). A Data-Driven Hybrid Optimization Model for Short-Term Residential Load Forecasting. *IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 283–287.

2. Guenoupkati, A., Salami, A., Kodjo, M., & Napo, K. (2021). Short-Term Electricity Generation Forecasting Using Machine Learning Algorithms: A case study of the Benin Electricity Community (C.E.B). *TH Wildau Engineering and Natural Sciences Proceedings*.

3. Humberto, V., Almendra, A., Cristhian, B., & Gabriel, O. (2017). Statistic linear parametric techniques for residential electric energy demand forecasting. *A review and an implementation to Chile,*, Volume 24, Pages 512-521.

4. Hyndman, R., & Fan, S. (2010). Density Forecasting for Long-Term Peak Electricity Demand. *IEEE Trans. Power Syst*, 25, 1142–1153.

5. Jiang, H., Fang, D., Spicher, K., Cheng, F., & B, L. (2019). A New Period-Sequential Index Forecasting Algorithm for Time Series Data. *Applied Sciences*, 9(20), 4386.

6. Le, T.-T.-H., Oktian, Y. E., & Kim, H. (2022). XGBoost for Imbalanced Multiclass Classification-Based Industrial Internet of Things Intrusion Detection Systems. *MDPI, Sustainability*.

7. Min, C., Jinfeng, W., Xiaochen, S., Zhengmou, R., Haokai, C., Jie, Y., & Ning, L. (2022). hort-Term and Medium-Term Electricity Sales Forecasting Method Based on Deep Spatio-Temporal Residual Network. *Energies. 15. 8844. 10.3390/en15238844.* .

8. Minaar, U., M, V. Z., & M, H. (March,2023). Applied SARIMA Models for Forecasting Electricity Distribution Purchases and Sales. *CIGRE Science & Engineering*.

9. Rathore, H., Meena, H., & Jain, P. (2023). Prediction of EV Energy consumption Using Random Forest And XGBoost. *10.1109/ICPEE54198.2023.10060798.*, 1-6.

10. Wang, W., Shi, Y., Lyu, G., & Deng, W. (2017). Electricity Consumption Prediction Using XGBoost Based on Discrete Wavelet Transform. *DEStech Transactions on Computer Science and Engineering. 10.12783/dtcse/aiea2017/15003*.

11. Wen, M., Zhong, Y., Liao, J., Li, W., Zhou, H., & Xu, C. (2021). Multidimensional forecasting of electricity sales in Hunan Province based on decomposition-integration ideas. *E3S Web of Conferences* , 300, 01020.

12. Zhang, X., & Li, R. (2021). A Novel Decomposition and Combination Technique for Forecasting Monthly Electricity Consumption. *Frontiers in Energy Research. 9. 10.3389/fenrg.2021.792358*.