**Classical Methods versus Machine Learning: Forecasting Residential Electricity Consumption and Prices**

**Maram Salameh and Hongrui Liu\***

**Industrial and Systems Engineering, San Jose State University**

**One Washington Square, San Jose, CA 95192**

**\*Corresponding author: hongrui.liu@sjsu.edu**

**Abstract**

A typical California U.S. household uses more electricity to power numerous devices and equipment than ever before. For energy service providers to match the residential sector’s needs, it is crucial to accurately predict the demand of its end users. Assessing the cost and usage by customers enables the energy service providers to cut waste and losses that could incur due to excessive energy production and avoid blackouts due to underestimating energy consumption. This paper compares the application of classical forecasting methods against machine learning algorithms in predicting the electricity consumption and prices for California’s residential sector. Specifically, this paper investigates the application of five different algorithms; Holt-Winter’s Exponential Smoothing, Seasonal ARIMA (SARIMA), Multivariate Multiple Linear Regression, Random Forest Regressor, and Recurrent Neural Networks: Long-Short Term Memory (LSTM). Two metrics were used to evaluate the models; Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The analysis shows that Multivariate Multiple Linear Regression has the best accuracy out of the five implemented methods, this is because it is able to consider other variable factors when making predictions.

**Keywords**: Forecasting, Machine Learning, Exponential Smoothing, SARIMA, Multivariate Linear Regression, Random Forest, LSTM, Consumption Prediction, Price Prediction.

**Introduction**

The prediction of energy consumption and rates is essential in assisting energy supply companies to adjust to specific behaviors. Energy prediction allows companies to become familiar with the behavior of their customers to adapt their rates to consumption or know the periods in which it will anticipate peak demand for energy, thus having better management over their supply chains.Consumption and prices of energy services are liable to change and unpredictability due to seasonality, regional unique climate, and other factors. Residential electricity consumption is projected to grow due to increasing population and economic development, as a result, electricity prices will be adjusted to the projected increase. To support this anticipated trend and ensure an efficient energy management process, it is important to deploy accurate forecasting models.

Classical forecasting models use previous historical data to model the next step in the sequence as a linear function. Classical methods are best suitable for univariate time series with trend and or seasonality. Holt-Winter’s model is a popular method for forecasting time series data with seasonality [3]. Seasonal ARIMA is another widely used model in time series analysis, it is an extension of Autoregressive Integrated Moving Average (ARIMA). SARIMA extends the ARIMA model by adding a seasonal component. This component consists of three parameters that define the autoregression (AR), differencing (I), and moving average (MA) for the seasonal component, it also adds an additional parameter for the period of the seasonality.[3] Classical methods are often applied to univariate data; therefore, it does not take into consideration other variable factors that may impact the analysis

Machine Learning algorithms have become increasingly incorporated in time series forecasting; machine learning enables computers to make predictions through experience. The input data is used to train the machine to automate the model building. Machine learning algorithms can incorporate a higher number of variable factors than traditional forecasting methods, this allows for building models with higher accuracy. Supervised learning is a category of machine learning, it utilizes what is learned from past labeled data to make future predictions. Linear regression, Random forest, and LSTM are supervised algorithms hence they have the ability to compare the predicted values to the actual values.

**Literature Review**

The residential sector accounts for about 22% of total U.S. energy consumption, and electricity accounts for 41% of household end-use energy consumption [1]. In 2019, the typical U.S. residential customer’s annual electricity consumption was about 10,530 kilowatt-hours (kWh), for an average of about 877 kWh per month. In California, the average monthly residential electricity consumption is 577 kWh, and residential customers pay an average of 16 cents per kWh. The average Californian household electric bill is $119 in the summer, and $91 in the winter [2]. Energy-efficient tools are rapidly emerging; however, from 1990 to 2009, energy consumption in residential buildings increased by 24%. Residential energy consumption is projected to grow to 19% by 2040 [7].

Amasyali and El-Gohary conducted a review on building energy consumption prediction, they identified that 19% of models belong to residential buildings. 47% of the datasets focused only on the total energy consumption from electricity meters and 67% of the models used real data instead of modeled or public standard data. The most used machine learning algorithms were artificial neural networks and support vector machine, with 47% and 25% respectively. [8] Although there are many studies on energy prediction, most are focused on other areas such as commercial and industrial sectors. There are little studies on energy consumption at the residential level, yet due to the projected growth, it is important to build accurate models so that service providers can maintain and improve their service continuously.

As electricity generation became a liberalized market, determining electricity prices and rates has become a difficult task. Electricity trade requires a constant balance between production and consumption. Electricity is a non-storable commodity so any small changes in production greatly impact the price. [9] The increasing penetration of renewable energy sources also impacts electricity rates. Since the introduction of renewable energy, electricity generation has been dependent on and influenced by weather conditions. Price unpredictability increases with increasing renewable energy penetration. [10] Electricity price is vital for balancing electricity generation and consumption. Price prediction enables service providers to optimize their position on the market floor.

**Dataset**

The U.S. Energy Information Administration (EIA) is a federal agency. It is responsible for gathering and analyzing energy information for the purpose of developing efficient markets and promoting public understanding of energy and its impact on the economy and the environment. The EIA Monthly Electric Power Industry Report collects data from energy service companies that sell or supply electric power to end-users. The report includes monthly sales (consumption), revenue, prices & customers for all end-use sectors (residential, commercial, industrial, and transportation) statewide. For the analysis in this paper, only the residential sector data pertaining to the state of California was taken into consideration. The dataset includes monthly data from the year 1990 to the present, for the purpose of this study the dataset was restricted to data from January 2010 to August 2020 as shown in Figure 1. The testing dataset includes data from January 2010 to August 2019, while the training data set consists of data from September 2019 to August 2020.

|  |  |
| --- | --- |
| Chart  Description automatically generated | Chart, line chart  Description automatically generated |

**Figure 1: Plot of Residential Sales and Residential Price**

**Left: Plot line of Residential Electricity Consumption in Megawatt hours**

**Right: Plot line of Residential Electricity Price in Cents/kWh**

**Methodology**

The following Python modules were used to conduct the analysis: Pandas, NumPy, Matplotlib, statsmodels, sklearn, pmdarima, Keras, and TensorFlow.

**Classical Forecasting Methods:**

**Holt-Winter’s Method**: Also known as Triple Exponential Smoothing, it models three aspects of the time series: the average, the slope over time, and the seasonality. It is a technique for smoothing time series data by giving exponentially decreasing weights against historical data over time. In other words, more recent data is assigned a higher weight. Holt-Winter’s method has two parameters to its seasonal component: additive seasonality and multiplicative seasonality. The additive method is used when the seasonal variations are continuous, and the multiplicative method is applied when the seasonal variations are changing proportionally to the extent of the series [3]. In this analysis it was found that the multiplicative method resulted in a higher accuracy score in predicting electricity sales while the additive method had a better performance for predicting electricity prices.

The Multiplicative method is modeled as:

Text, letter

Description automatically generated

The Additive method is modeled as:

Text, letter

Description automatically generated

Where:

𝑙: level equation

𝑏: trend equation

𝑠: seasonal equation

**Seasonal ARIMA (SARIMA):** ARIMA is an acronym for Autoregressive Integrated Moving Average. This model can deal with non-stationary data because of its “integrate” component. The “integrate” component transforms a non-stationary time series into a stationary time series by differencing which eliminates or reduces the trend and seasonality. SARIMA adds a seasonal component to the ARIMA model [3, 4]. The model is expressed as follows:

Diagram

Description automatically generated

Where:

p: Trend autoregression order.

d: Trend difference order.

q: Trend moving average order.

P: Seasonal autoregressive order.

D: Seasonal difference order.

Q: Seasonal moving average order.

m: The number of time steps for a single seasonal period.

**Machine Learning Methods:**

**Multivariate Multiple Linear Regression:** It is an extension of the linear regression model used to assess the linear relation between independent and dependent variable sets. In an analysis where there are **n** observations, the number of dependent variables is **q**, and the number of independent variables is **p** [5]; the multivariate multiple linear regression model has the form:

Y = Χβ + ε

Where:

Y: The (n × q) response variable matrix

X: The n × (p + 1) independent variable matrix

β: The (p + 1) × q coefficient matrix

ε: The (n × q) error matrix

For the electricity sales prediction, the independent variables used are “Revenue”, “Customer count”, and “Price”. For the electricity price prediction, the independent variables used are “Revenue”, “Sales”, and “Customer count”

**Random Forest:** An ensemble supervised learning algorithm that utilizes bagging or bootstrap aggregation, it can perform both classification and regression tasks. Random Forests operate by building a large number of decorrelated trees then averaging them thus reducing the variance. Each tree in the model trains and learns from a random sample of the observations. Although each tree’s predictions may not be accurate, when combined and averaged they become close to the actual value. [6]

**Long Short-Term Memory:** A type of Recurrent Neural Networks (RNN), RNN is a kind of a neural network where the purpose is to make future predictions by learning from past sequential observations. In RNNs, the hidden layers operate as storage for information entered in earlier steps of reading sequential data. LSTM extends the typical RNN by adding features to memorize the sequence of data. The memorization is made possible through some gates along with a memory line incorporated in the LSTM. Each LSTM consists of three gates: (1) Forget gate: indicates whether information should be kept or discarded, (2) Memory gate: decides which new information need to be stored in the cell, and (3) Output gate: indicates the output of each cell based on the forget and memory gates. [4]

**Model Evaluation**

The accuracy of the models was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Root Mean Square Error (RMSE) is a measure of the standard deviation of the residuals, it measures how spread out these residuals are from the regression line.

Text, whiteboard

Description automatically generated

Mean Absolute Error (MAE) is the measure of the absolute value of the difference between the predicted values and actual values.

A picture containing schematic

Description automatically generated

Where:

n: The number of samples

p: The predicted values

a: The actual values

**Analysis Results**

**Electricity Sales Results:**

**Holt-Winter’s:** Using a model with multiplicative seasonality and no trend, the forecasted values are shown in Figure 2.

Chart, line chart

Description automatically generated

**Figure 2: Plot of Holt-Winter’s Electricity Sales Model**

**The figure depicts Holt-Winter’s predictions against actual values**

**Seasonal ARIMA:** Best model found using auto\_arima from the pmdarima module is ARIMA(0,0,0)(0,1,0)[12]. Figure 3 shows the SARIMA Forecasted values.

**Chart, line chart

Description automatically generated**

**Figure 3: Plot of SARIMA Electricity Sales Model**

**The figure depicts SARIMA predicted values against actual values**

**Multivariate Linear Regression:** The model parameters are as follows: Coefficients [ 5.58987966e+00, -1.45077666e-02, -3.94263828e+05], Intercept: [7232206.025894092], and Coefficient of determination: [0.977]. The forecasted values are shown in figure 4.

**Chart, line chart

Description automatically generated**

**Figure 4: Plot of Linear Regression Model for Electricity sales**

**The figure shows the Multivariate Linear Regression predictions against actual values**

**Random Forest Regressor:** Figure 5 shows the forecasted values using 200 trees before averaging the predictions.

Chart, line chart

Description automatically generated

**Figure 5: Plot of Random Forest Regressor for Electricity sales**

**The figure shows Random forest predictions against actual values.**

**Long Short-Term Memory:** Figure 6 shows the predicted values using LSTM model with the following parameters: 4 neurons, batch size of 1 and 1 epoch.

Chart, line chart

Description automatically generated

**Figure 6: LSTM model for Electricity Sales**

**The figure shows LSTM predictions against actual values**

**Residential Electricity Sales Accuracy Results:**

|  |  |  |
| --- | --- | --- |
| **Method** | **RMSE** | **MAE** |
| Holt-Winter’s | 702951 | 589194 |
| SARIMA | 876141 | 716207 |
| Multivariate Linear Regression | 240210 | 186133 |
| Random Forest | 608674 | 484645 |
| LSTM | 2007388 | 1583042 |

Table 1: Residential electricity sales model accuracy scores

**Electricity Prices Results:**

**Holt-Winter’s:** Using a model with additive seasonality and an additive trend, the forecasted values are shown in Figure 7.

**Chart, line chart

Description automatically generated**

**Figure 7:** **Plot of Holt-Winter’s Electricity Price Model**

**The figure depicts Holt-Winter’s predictions against actual values**

**Seasonal ARIMA:** Best model found using auto\_arima from the pmdarima module is ARIMA(0,0,0)(1,1,0)[12]. Figure 8 shows the predicted values against the actual values.

Chart, line chart

Description automatically generated

**Figure 8:** **Plot of SARIMA Electricity Price Model**

**The figure depicts SARIMA predicted values against actual values**

**Multivariate Linear Regression:** The model parameters are as follows: Coefficients [ 1.35713237e-05, -2.38566585e-06, 3.65098538e-08], Intercept: [17.003786976779438], and Coefficient of determination: [0.837]. The forecasted values are shown in figure 9.

**Chart, line chart

Description automatically generated**

**Figure 9: Plot of Linear Regression Model for Electricity Price**

**The figure shows the Multivariate Linear Regression predictions against actual values**

**Random Forest Regressor:** Figure 10 shows the forecasted values using 100 trees before averaging the predictions.

**Chart, line chart

Description automatically generated**

**Figure 10: Plot of Random Forest Regressor electricity price model**

**The figure shows Random forest predictions against actual values.**

**Long Short-Term Memory:** Figure 11 shows the predicted values using LSTM model with the following parameters: 4 neurons, batch size of 1 and 3 epoches.

**Chart, line chart

Description automatically generated**

**Figure 11: Plot of LSTM model for electricity price**

**The figure shows LSTM predictions against actual values.**

**Residential Electricity Prices Accuracy Summary Results:**

|  |  |  |
| --- | --- | --- |
| **Method** | **RMSE** | **MAE** |
| Holt-Winter’s | 1.15 | 0.92 |
| SARIMA | 1.27 | 1.06 |
| Multivariate Linear Regression | 0.66 | 0.50 |
| Random Forest | 1.48 | 1.25 |
| LSTM | 1.97 | 1.66 |

Table 2: Residential electricity prices model accuracy scores

**Conclusion**

This paper aims to evaluate the performance between classical forecasting methods and machine learning algorithms in predicting energy consumption and rates in the residential sector and compare their accuracy rate. As table 1 and 2 show Multivariate Linear regression performs better than classical forecasting methods, both in predicting electricity sales and prices. Random Forest and LSTM had a lower accuracy rate than classical methods, this is partly because it is difficult to tune their parameters to find the best fitting model.

**References**

1. U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. (2020, August 04). Retrieved December 2nd, 2020, from <https://www.eia.gov/energyexplained/use-of-energy/homes.php>
2. Rockzsfforde, R. R. (n.d.). Comparative Analysis of Utility Services & Rates in California (United States, California Public Utilities Commission, Policy & Planning Division). April 14, 2015.
3. Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on December 6th, 2020
4. S. Siami-Namini, N. Tavakoli and A. Siami Namin, *"A Comparison of ARIMA and LSTM in Forecasting Time Series," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, 2018, pp. 1394-1401, doi: 10.1109/ICMLA.2018.00227.
5. Topal, Mehmet. “*EXAMINATION OF MULTIVARIATE MULTIPLE LINEAR REGRESSION ANALYSIS 1”*. *International Research Journal of Mathematics, Engineering, and IT.* Erzurum, Turkey, 2016, Vol.3, Issue 6.
6. Hastie T., Tibshirani R., Friedman J. (2009) *“Random Forests. In: The Elements of Statistical Learning”.* Springer Series in Statistics. Springer, New York, NY. Pp. 587-604, <https://doi.org/10.1007/978-0-387-84858-7_15>
7. Repice, R., “Monthly Energy Review”, Energy Information Administration (EIA), 2020 (accessed 12.03.2020) <http://www.eia.gov/totalenergy/data/monthly/?src=Total-f1>.
8. Kadir Amasyali, Nora M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews, Volume 81, Part 1, 2018, pp. 1192-1205, ISSN 1364-0321, <https://doi.org/10.1016/j.rser.2017.04.095>.
9. Shahidehpour M, Yamin H, Li Z. *“Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management*”. New York (USA): John Wiley & Sons, Inc.; 2002. p. 1–20. http://dx.doi.org/10.1002/047122412X.ch1 [chapter 1]
10. Brancucci Martinez-Anido, Carlo & Brinkman, Greg & Hodge, Bri-Mathias. (2016). “*The impact of wind power on electricity prices. Renewable Energy*”. 94. 474-487. 10.1016/j.renene.2016.03.053.