Comparison of Machine Learning Techniques for Electricity Demand Forecasting

Abstract— **This paper aims to compare several machine learning techniques that can be used to predict the electricity demand in France. This paper implemented four machine learning algorithms including Multiple Linear Regression (MLR), Support Vector Machine (SVM), Autoregressive(AR), and K Neighbors Regressor. The importance of electrical demand forecasting has grown significantly as the electrical market has grown larger. In the previous research, seasonal algorithms are widely used in forecasting time series datasets. In this case, three regression analyses and 1 time-series analysis have been analyzed. According to the results, the xxx algorithm is optimal and has the highest accuracy when comparing these four algorithms.**

Keywords— **Support Vector Machine (SVM), Multiple Linear Regression (MLR), Autoregressive (AR), K Neighbors Regressor (KNN)**

1. Introduction

France is one of the largest European electricity markets [1], and its demand is a major factor in European electricity prices. During periods of high demand, France is likely to import electricity from neighboring countries, resulting in higher prices in both France and the rest of the European. And during periods of low demand, France is more likely to export, which results in lower electricity prices. As a result, market participants require a thorough understanding and forecast of this demand in France.

In this paper, three different machine learning algorithms and one deep learning algorithm have been applied to forecast the hourly electricity demand by using real-time electricity demand data from France in the period from 1st January 2017 to 8th Mar 2022. Each row of the data represents an hourly observation that has 16 features: Date, Time, Solar, Wind, Frozen, etc.

According to the previous studies, forecasting electricity demand can be categorized into three types: short-term forecasting: a couple of hours to a couple of days, mid-term forecasting: a couple of weeks to a couple of months, and long-term forecasting: a couple of years [2]. In this paper, the demand forecast is based on 48 hours ahead.

2. Data exploration and features selection

- Load and pre-process data

The dataset is provided by EnAppSys BV company. It has been imported into Jupyter (Python IDE) as a data frame type. It contains the datetime feature, in this paper, year, month, date, time, and time zone have been separated into columns for further analysis. After dropping 230 null and NA values, the total number becomes 45202, with no duplicated values.

- Pre-analysis data

Figure1 depicts the electricity demand in France over the last five years, along with weather data. Figure2 shows the yearly seasonality, with the highest value occurring during the winter months, from November to February. August is the lowest demand month of the year. Figure3 demonstrates that the demand is lower than the weekdays on the weekend.

To identify the correlation between demand and other features, a Pearson correlation heatmap has been shown in Figure4, the correlation coefficient of “air_tmp [Kelvin]” and “apparent_tmp [Kelvin]” are over 0.6 which can be regarded as a moderate positive relationship, which means the electricity demand of France related on weather data.

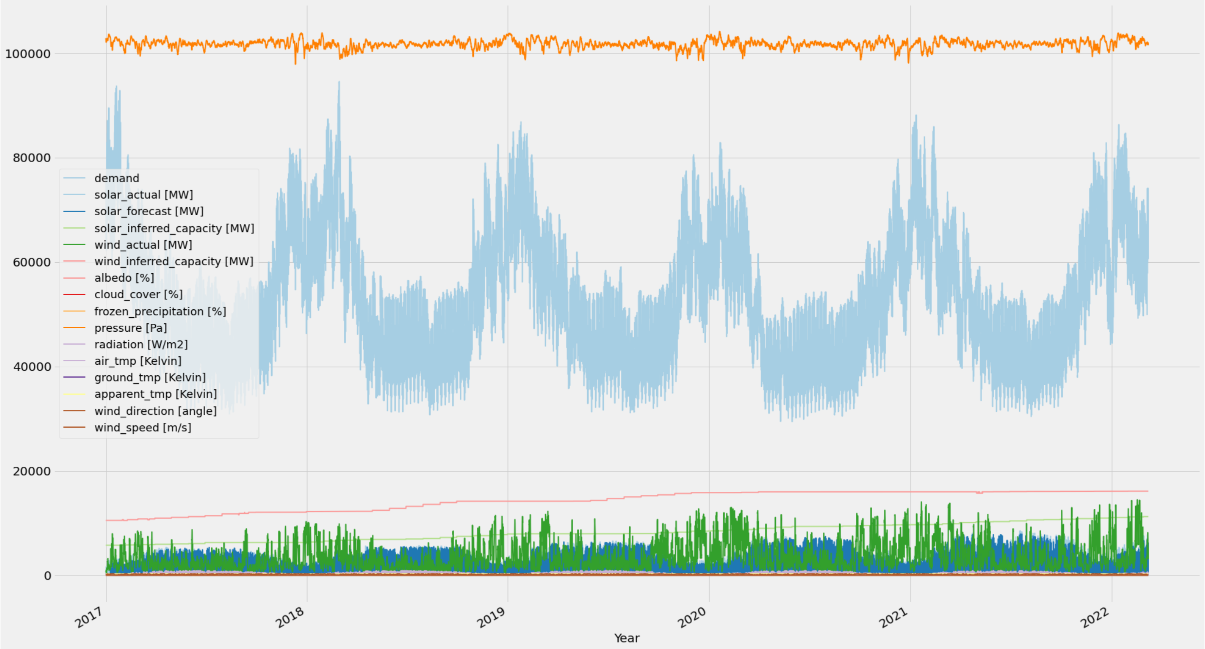


Figure1: Line chart of all the features in 2017–2022

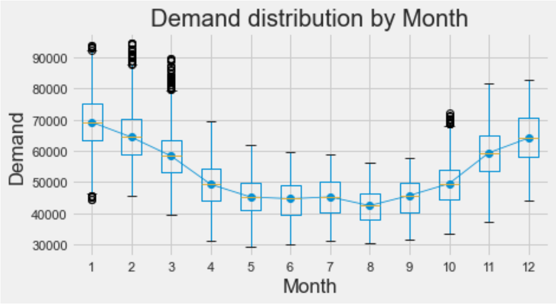


Figure2: Average monthly demand

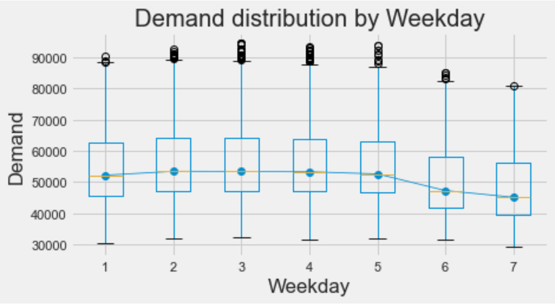


Figure3: Average weekly demand

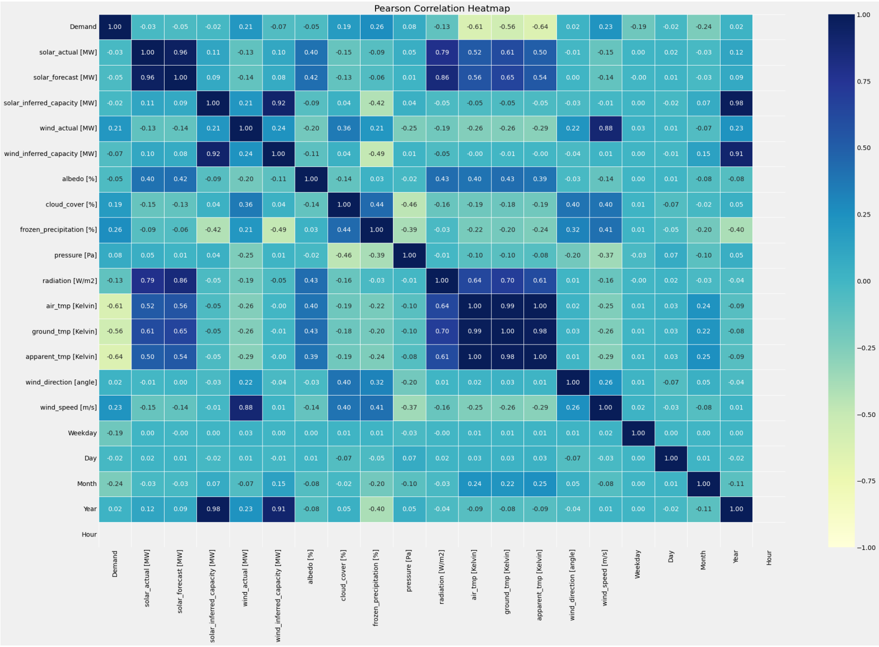


Figure4: Pearson Correlation Heatmap

- Split train data and test data

As this dataset is time series based, the train and test data are split by date rather than random in this paper. To control the ratio of train and test data at 8:2, it sorts date first, then makes the train sets with the first 80% of the data, and the test sets with the rest 20% of the data. The number of train sets is 36161, and the number of test sets is 9041.

- Experiments

Four different methods have been implemented to train and test the dataset. These are supervised machine learning algorithms and one of them is deep learning. k-fold cross-validation has been implemented, splitting data 5 times to train and validate.

2.1 Method

1) Multiple Linear Regression (MLR)

I. Theoretical background

The Multiple Linear Regression (MLR) model is a statistical algorithm that uses more than one independent variable to forecast output based on historical data[3].
The formula is：

$$y_i = b_0 + b_1 x_1 + b_2 x_2 + \mu_i$$

II. Implement

The Gradient boosting algorithm which can be used for regression only has been implemented to get the importance of the features [4]. As shown in Fig5, “apparent_tmp [Kelvin]”, “Month”, “ Hour”, “Weekday”, “radiation [W/m2]”, “air_tmp [Kelvin]”, “solar_actual [MW]” and “albedo [%]” have been selected to train the model.

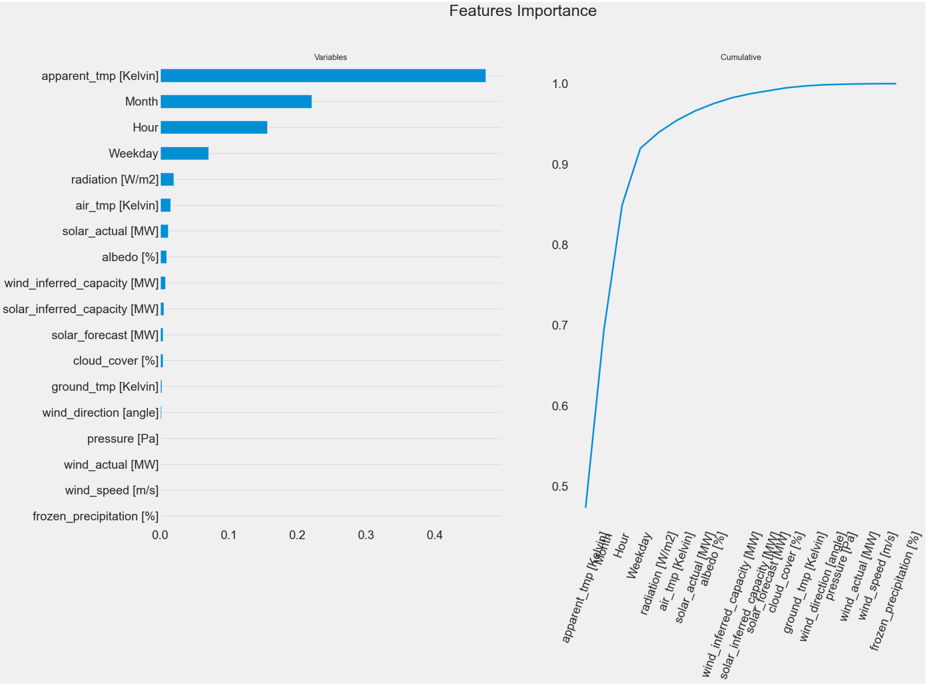


Figure5: The importance of the features

The K fold validation has been used in this algorithm. The R-squared (R2) value of each folder is in the range of 0.66-0.67.

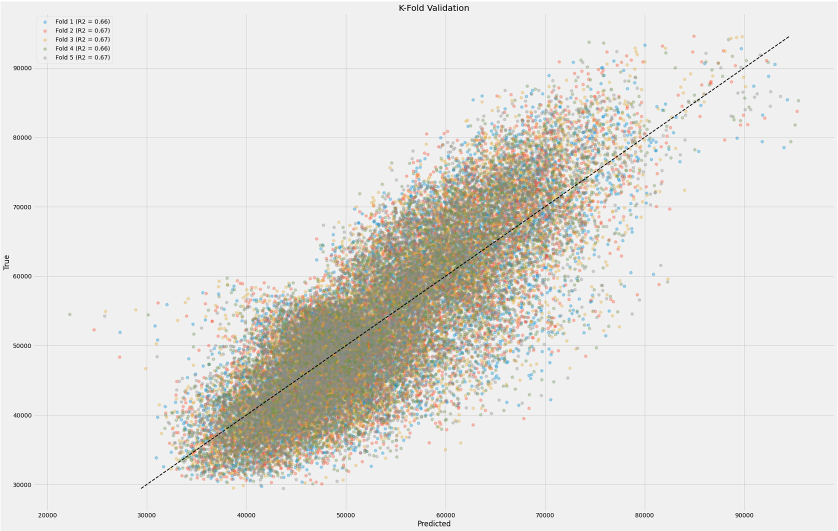


Figure6: The K-Fold validation and R-Squared value

2) Autoregressive (AR)

I. Theoretical background

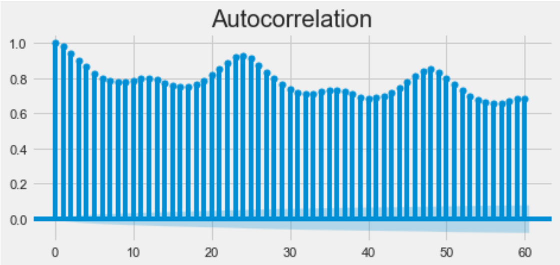
The Autoregressive (AR) model is a statistical algorithm that forecasts output based on historical data.
The formula is：

$$\begin{aligned} u(n) + a_1 u(n-1) + a_2 u(n-2) + \cdots \\ + a_M u(n-M) = \nu(n) \end{aligned}$$

II. Implement

In forecasting time series datasets, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are widely used to decide which models to use. PACF is a variable's final lag value with all intermediate lag values removed from the analysis, whereas ACF is a variable's correlation with its own lag values [5].

Because AR and MA processes behave differently, a combination of ACF and PACF plots can be used to find p and q for an ARMA process.



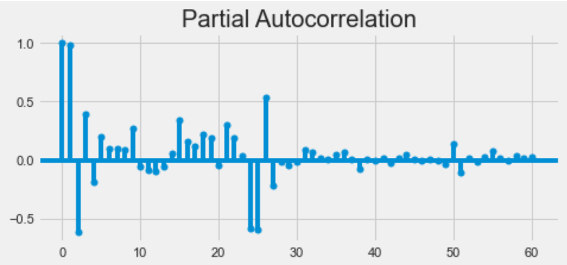


Figure7: Autocorrelation Function (ACF) Figure8: Partial Autocorrelation Function (PACF)

The above Figure7 clearly shows a Tails off, it reduces the autocorrelation value when the lags increase gradually. And in Figure8, the partial autocorrelation value shows a Cuts off instead, it jumps directly to 0 after lag p. It is a typical Autoregressive AR (p) model.

3) Support Vector Machine (SVM)

I. Theoretical background

The support vector machine (SVM) is a supervised machine learning model for classification, regression, and outliers' detection problems. It is able to reduce structural risk, which aims to find a balance between the model's complexity and the training error's complexity by lower the upper bound of the promotion error [6].

II. Implement

In this paper, epsilon stands for the width of the tube around the hyperplane, it has been set to 10, and regularization parameter C has been set to 1 as it is the penalty parameter of the error term.

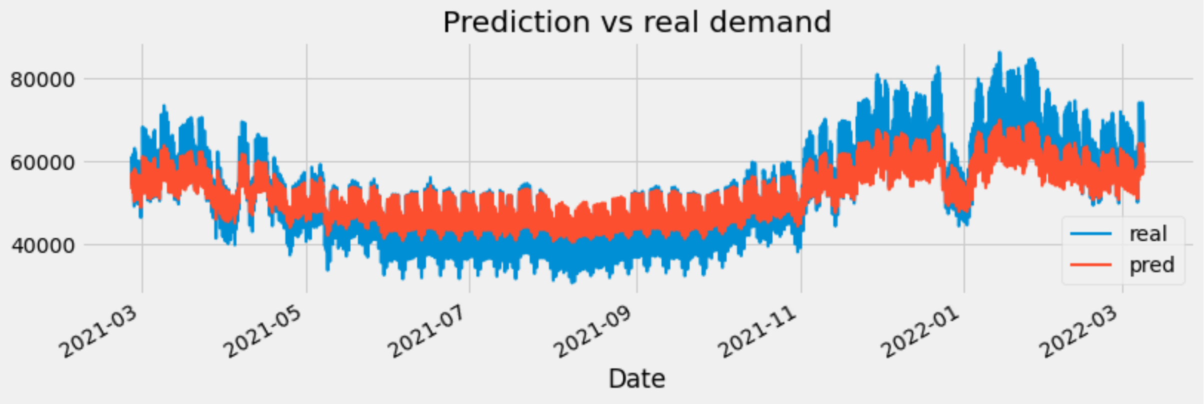


Figure8: SVM test set result

4) K-Nearest Neighbors Regressor (KNN)

I. Theoretical background

K-Nearest Neighbors Regressor uses a local interpolation of the targets associated with the training set's closest neighbors [7].

II. Implement

In this paper, k value has been defined by K-folder validation as 5.

3. Results and Discussion

The results of the comparison are shown in table1.

Models	Accuracy	R2 (Explained Variance)	Mean Absolute Error(MAE)	Mean Absolute Percentage Error(MAPE)	Mean Squared Error (MSE)
Multiple Linear Regression (MLR)	25	22	20		25
Autoregressive (**AR**)	9	10	10		28
Support Vector Machine (SVM)	0.65	0.65	5372.56		49121280.61
A* Search	9	10	10		9

Table 1: Results for different models result.

According to the data in Table 1,

The A* Search takes the least steps in both shortest path steps taken, and total steps taken, which means it can use fewer steps to find the shortest path. It is an optimal search algorithm in this case compared to the other search algorithms.

4. Conclusions

Mean absolute error: This is the average of absolute errors of all the data points in the given dataset.

Mean squared error: This is the average of the squares of the errors of all the data points in the given dataset. It is one of the most popular metrics out there!

Median absolute error: This is the median of all the errors in the given dataset. The main advantage of this metric is that it's robust to outliers. A single bad point in the test dataset wouldn't skew the entire error metric, as opposed to a mean error metric.

Explained variance score: This score measures how well our model can account for the variation in our dataset. A score of 1.0 indicates that our model is perfect.

R2 score: This is pronounced as R-squared, and this score refers to the coefficient of determination. This tells us how well the unknown samples will be predicted by our model. The best possible score is 1.0, but the score can be negative as well.

According to the result,

5. References

[1] A. Eybalin and M. Shahidehpour, ‘Electricity restructuring in France’, in *2003 IEEE Power Engineering Society General Meeting (IEEE Cat. No.03CH37491)*, Jul. 2003, vol. 1, pp. 330-335 Vol. 1. doi: 10.1109/PES.2003.1267192.

[2] A. Moradzadeh, H. Moayyed, K. Zare, and B. Mohammadi-Ivatloo, ‘Short-term electricity demand forecasting via variational autoencoders and batch training-based bidirectional long short-term memory’, *Sustain. Energy Technol. Assess.*, vol. 52, p. 102209, Aug. 2022, doi: 10.1016/j.seta.2022.102209.

[3] M. Flores-Sosa, E. León-Castro, J. M. Merigó, and R. R. Yager, ‘Forecasting the exchange rate with multiple linear regression and heavy ordered weighted average operators’, *Knowl.-Based Syst.*, p. 108863, Apr. 2022, doi: 10.1016/j.knosys.2022.108863.

[4] N. Dahiya, B. Saini, and H. D. Chalak, ‘Gradient boosting-based regression modelling for estimating the time period of the irregular precast concrete structural system with cross bracing’, *J. King Saud Univ. - Eng. Sci.*, Aug. 2021, doi: 10.1016/j.jksues.2021.08.004.

[5] J. Moon, M. B. Hossain, and K. H. Chon, ‘AR and ARMA model order selection for time-series modeling with ImageNet classification’, *Signal Process.*, vol. 183, p. 108026, Jun. 2021, doi: 10.1016/j.sigpro.2021.108026.

[6] H. Ma, F. Ding, and Y. Wang, ‘A novel multi-innovation gradient support vector machine regression method’, *ISA Trans.*, Mar. 2022, doi: 10.1016/j.isatra.2022.03.006.

[7] W. T. Ho and F. W. Yu, ‘Measurement and verification of energy performance for chiller system retrofit with k nearest neighbour regression’, *J. Build. Eng.*, vol. 46, p. 103845, Apr. 2022, doi: 10.1016/j.jobe.2021.103845.