Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA10.20250285
Published: June 17, 2025

Comparison of Prediction Accuracy of Logistic Regression and Random Forest in Weather Data Time Series Analysis

Yejun Hwang and Dongho Shin
Student and Professor, My Paul School
12-11, Dowontongmi-gil, Cheongcheon-myeon, Goesan-gun
Chungcheongbuk-do, Republic of Korea
eavatar@hanmail.net

Abstract

The purpose of this study is to compare the performance of logistic regression model and random forest model in meteorological time series data analysis and analyze the difference when applying them to precipitation probability prediction. Using logistic linear regression model and random forest algorithm, we examined how major weather variables (Temperature, Dew Point, Humidity, etc.) contribute to precipitation occurrence prediction.In the experiment, the logistic regression accurately predicted 5,545 out of 6,812 samples, achieving a prediction accuracy of 81.4%. In contrast, the random forest model, an ensemble technique that can learn nonlinear interactions, utilizes an ensemble learning technique that combines multiple decision trees to improve prediction performance. In the experiment, the random forest model achieved a prediction accuracy of 90.61% on the test data, showing about 9% higher performance than the logistic regression model.In summary, in time series data analysis, it is important to select an appropriate model considering the linearity and complexity of the data. The logistic regression model is suitable for situations that require a simple structure and quick interpretation, while the random forest model excels in problems that include nonlinear data and complex interactions.

Keywords

Logistic regression, Time series data, Random forest, Weather prediction and Machine learning

1. Introduction

Time series data are collected chronologically and play an important role in various fields including economy, environment, meteorology, and medicine (Park et al. 2018). In particular, weather data is closely related to our daily lives, and precipitation is an important factor that affects decision-making in the agriculture, transportation, and energy use sectors (Kim and Seo 2014).Recent advances in data science and artificial intelligence have expanded the possibilities of time series data analysis and prediction (Jeon et al. 2024). For example, precipitation prediction technology using weather data can contribute to a more accurate and efficient decision-making process (Park 2006).
In this paper, we aim to determine which model is better among linear regression models and nonlinear regression models in order to improve the accuracy of precipitation prediction. To this end, we will compare and analyze the performance of the logistic regression model and the random forest model. The objective of this study is to analyze the advantages of logistic regression, one of the linear regression models, and random forest, one of the nonlinear regression models, in precipitation prediction, and which model shows better performance.

2. Body

The multiple linear regression model is a technique that predicts changes in the dependent variable based on changes in independent variables using a linear function, similar to simple linear regression. However, the key difference between multiple linear regression and simple linear regression lies in the number of independent variables. A linear regression model with two or more independent variables is referred to as a multiple linear regression model. As

mentioned earlier, the multiple linear regression model expresses changes in the dependent variable as a linear function of the independent variables. It is an artificial intelligence model that predicts the value of a dependent variable by taking multiple independent variables as input (LaValley 2008).

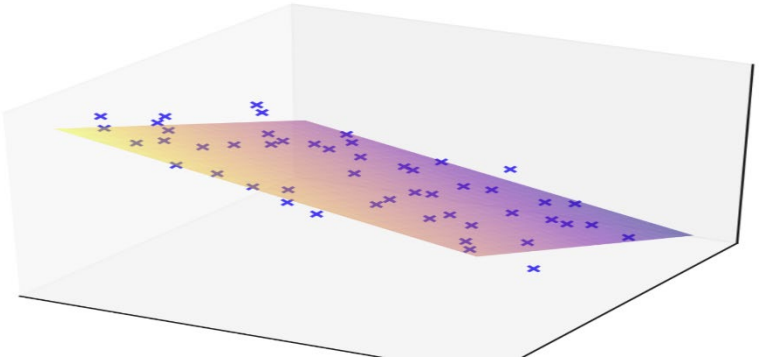

Figure 1. Logistic regression graph

In order to obtain the predicted value of a linear function like this, first, the weights (w) and bias (b) of the model are randomly initialized, and the learning rate (α) that controls the learning speed is set. After that, a linear model is calculated based on the given input data (X), and this is converted into a probability value using the sigmoid function (Sperandei 2014).

$$\text{linear\_model} = Xw + b$$

$$y_{\text{pred}} = \frac{1}{1 + e^{-\text{linear\_model}}}$$

Next, the error is measured by calculating the difference between the actual value (y) and the predicted value (y_perd). In this study, the mean squared error (MSE) was used as the loss function (Kwak 2002).

$$\text{error} = y - y_{\text{pred}}$$

Using gradient descent, the gradients of weights and biases are calculated to minimize the loss. The change in weights is derived from the inner product of the input data and the error, and the change in bias is calculated based on the average of the errors (Dreiseitl 2002).

$$w_{\text{gradient}} = \left(-\frac{2}{n}\right) X^T \cdot \text{error}$$

$$b_{\text{gradient}} = \left(-\frac{2}{n}\right) \sum \text{error}$$

Finally, the update is performed by subtracting the gradient values, multiplied by the learning rate, from the existing weights and biases (Hellevik 2009).

$$w = w - \alpha \cdot w_{\text{gradient}}$$

$$b = b - \alpha \cdot b_{\text{gradient}}$$

By repeating the above process multiple times, the optimal weights and biases are obtained and used to predict new values (Biau and Scornet 2016).

The random forest model is an algorithm based on decision trees. The modeling approach that combines multiple algorithms to produce results is called an ensemble method. There are various ensemble techniques, with the most representative being voting, bagging, and boosting (Speiser et al. 2019).
The voting method determines the final prediction by aggregating results from different algorithms through majority voting or averaging. The bagging method employs the same algorithm, randomly extracting subsets of data to create multiple sub-datasets, which are then trained in parallel. The final prediction is determined by majority voting or averaging. The boosting method, like bagging, utilizes the same algorithm, but instead of training in parallel, it sequentially learns from previous results, gradually reducing errors (Probst et al. 2019).
The random forest model is an algorithm constructed using the bagging technique. It leverages the bootstrap method to generate diverse sub-datasets, and multiple decision trees are trained on these datasets. By aggregating their predictions, the model mitigates the overfitting issue commonly associated with a single decision tree. This sampling process reduces variance and enhances the model's stability, enabling more reliable predictions across various data scenarios (Oshiro et al. 2012).

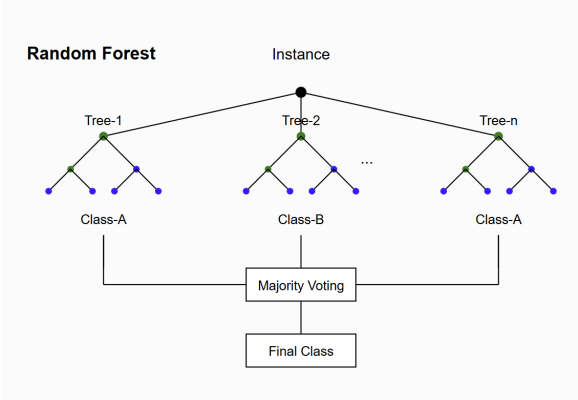

Figure 2. Random Forest Diagram

In this study, precipitation probability was predicted using both linear and nonlinear techniques by leveraging multiple variables. A total of 20 variables were utilized for prediction, including Max Temperature, Mean Temperature, Min Temperature, Dew Point, Mean Dew Point, Min Dew Point, Max Humidity, Mean Humidity, Min Humidity, Max Sea Level Pressure (hPa), Mean Sea Level Pressure (hPa), Min Sea Level Pressure (hPa), Max Visibility (km), Mean Visibility (km), Min Visibility (km), Max Wind Speed (km/h), Mean Wind Speed (km/h), Max Gust Speed (km/h), Precipitation (mm), and Cloud Cover (Farnaaz and Jabbar 2016).

In the linear approach using logistic regression, an overfitting issue was observed. When predicting precipitation probability based on multiple variables, certain variables exhibited extreme convergence toward 0 or 1. This indicates that the model may be overly sensitive to the training data, potentially leading to poor generalization performance (Jeon et al. 2024).

Below is a visualization of precipitation probability predictions using logistic regression, showing the probability distribution for each individual variable as well as for all variables combined (Park 2018).

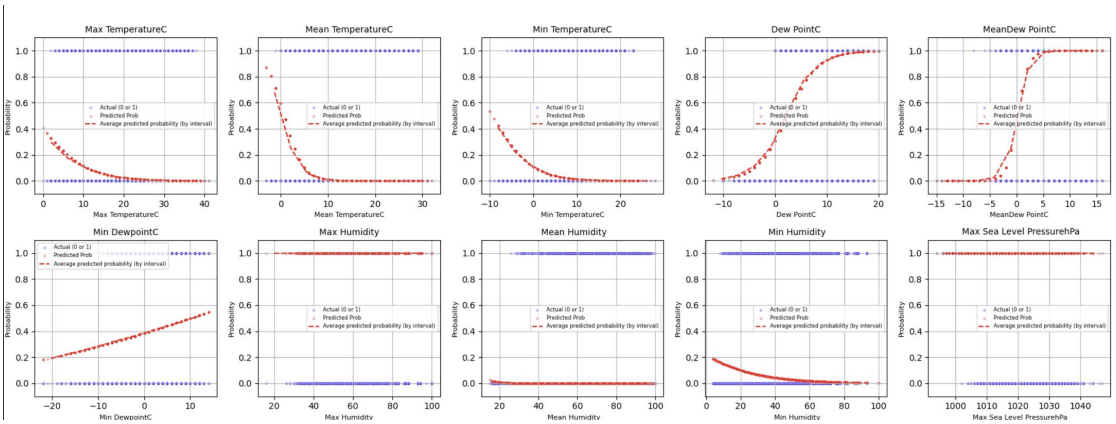

Figure 3. Logistic regression graph for each variable-1

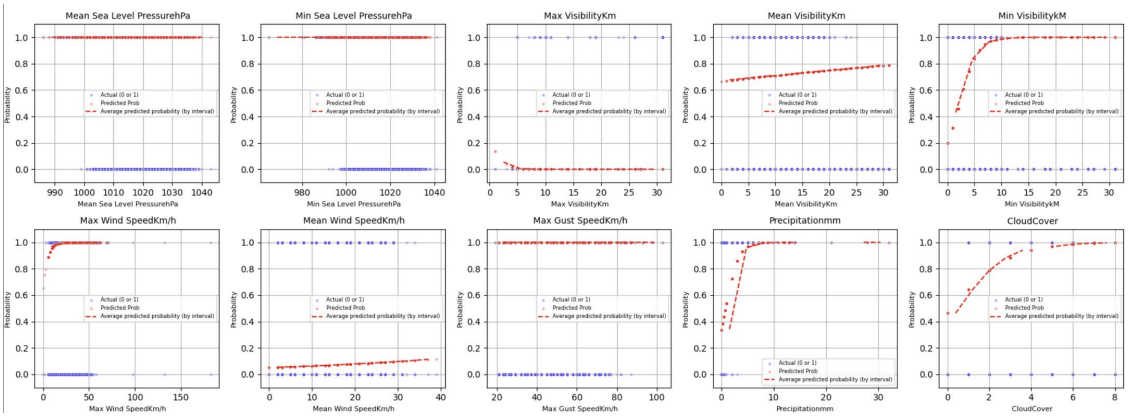

Figure 4. Logistic regression graph for each variable-2

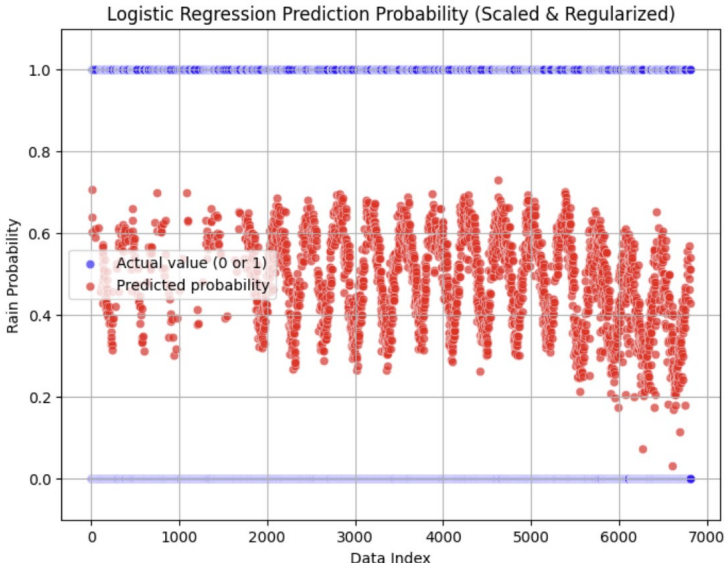

Figure 5. Logistic regression graph of all variables

The graph below illustrates the precipitation probability predictions made by the random forest model, considering each variable individually as well as all variables combined. Unlike logistic regression, which is a linear model and produces an S-shaped curve, the random forest model captures nonlinear and complex patterns, resulting in more intricate graph shapes (Park 2006).

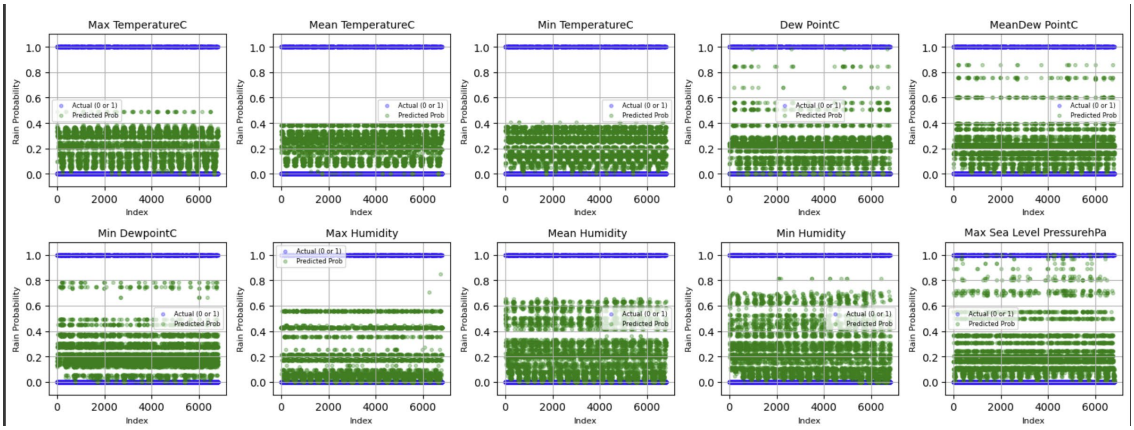

Figure 6. Random forest graph for each variable-1

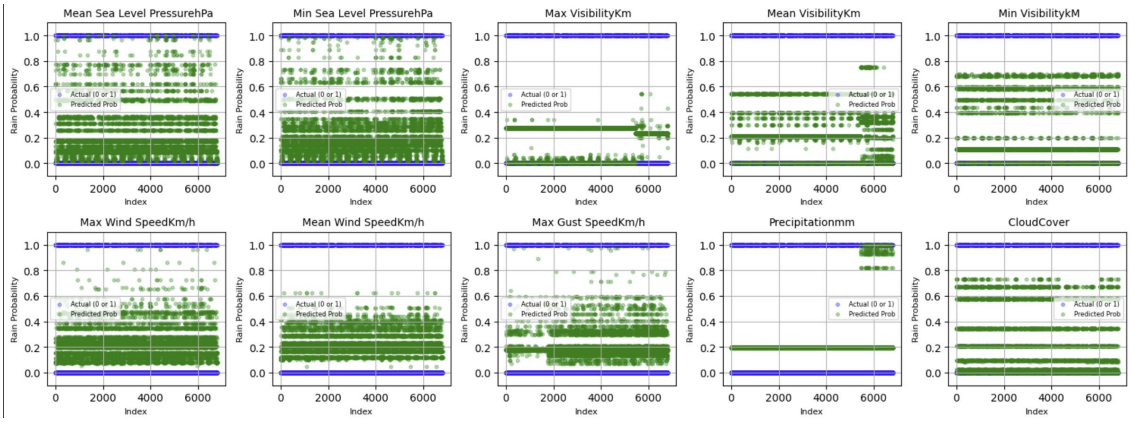

Figure 7. Random forest graph for each variable-2

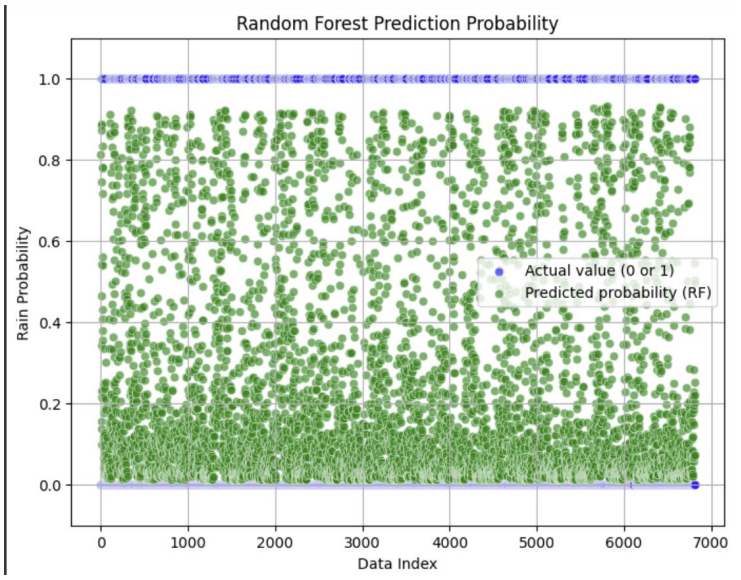

Figure 8. Random forest graph of all variables

In this study, a total of 6812 training data points and 6812 prediction data points were prepared. Among the 6812 data points, the multiple linear regression algorithm accurately predicted 5545 data points, achieving an accuracy of 81.40%. The random forest algorithm accurately predicted 6173 data points out of 6812, achieving an accuracy of 96.02% (rounded to two decimal places) (Kim and Seo 2014).

3. Conclusion

The random forest model effectively captures nonlinear patterns, and in cases where there are sudden changes in certain variable values, it may exhibit a step-like pattern. In contrast, logistic regression, being a linear model, follows an S-shaped prediction curve. Weather data often involves complex, nonlinear interactions between various factors, and the random forest model, which can more flexibly adapt to these characteristics, demonstrated superior prediction performance (Kang 2007).

References

LaValley, M., Logistic regression. *Circulation*, *117*(18), 2395-2399, 2008.

Sperandei, S., Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12-18, 2014.

Kwak, C., & Clayton-Matthews, A., Multinomial logistic regression. *Nursing research*, *51*(6), 404-410, 2002.

Dreiseitl, S., & Ohno-Machado, L., Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5-6), 352-359, 2002.

Hellevik, O., Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & quantity*, *43*, 59-74, 2009.

Biau, G., & Scornet, E., A random forest guided tour. *Test*, *25*(2), 197-227, 2016.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E., A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93-101, 2019.

Probst, P., Wright, M. N., & Boulesteix, A. L., Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301, 2019.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A., How many trees in a random forest?. In *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8* (pp. 154-168). Springer Berlin Heidelberg, 2012.

Farnaaz, N., & Jabbar, M. A., Random forest modeling for network intrusion detection system. *Procedia Computer Science*, *89*, 213-217, 2016.

Jeon, S. R., Lee, J. H., Kim, S. E., & Park, J. H., Optimization of Growth Environments Based on Meteorological and Environmental Sensor Data. *Journal of Sensor Science and Technology*, *33*(4), 230-236, 2024.

Park, C., Moon, J. Y., Cha, E. J., Yun, W. T. & Choi Y. E., Recent Changes in Summer Precipitation Characteristics over South Korea. *Journal of the Korean Geographical Society*, 43(3), 324-336, 2018.

Park, J. S., Precipitation probability prediction using the Markov logistic regression model. *Proceedings of the Korean Data & Information Science Society Conference*, 345-352, 2006.

Kim, J. Y., & Seo, K. H., The Development of Ensemble Statistical Prediction Model for Changma Precipitation. *Atmosphere, 24*(4), 533-540, 2014.

Kang, B. S., Rieu, S. Y. & Ko, I. H.,. Long-term Probabilistic Streamflow Prediction using Weather Outlook Weighted Ensemble Streamflow Prediction. *Journal of the Korean society of civil engineers*. B, *27*(2B), 183-191, 2007.

Biographies

Yejun Hwang is student in MY PAUL SCHOOL. He is interested in block chains, artificial intelligence, deep learning, cryptography, robots, autonomous vehicles, etc., and is conducting related research.

Dongho Shin is Professor and Teacher in MY PAUL SCHOOL. He obtained his Ph.D. in semiconductor physics in 2000. He is interested in artificial intelligence, deep learning, cryptography, robots, block chains, drones, autonomous vehicles, mechanical engineering, the Internet of Things, metaverse, virtual reality, and space science, and is conducting related research.