International Journal of Software Innovation
Volume 11 • Issue 1

A Study on Prediction Performance Measurement of Automated Machine Learning:

Focusing on WiseProphet, a Korean Auto ML Service

Euntack Im, Soongsil University, South Korea

Jina Lee, Soongsil University, South Korea

Sungbyeong An, Soongsil University, South Korea

Gwangyong Gim, Soongsil University, South Korea*

ABSTRACT

In digital economics, where value creation using big data becomes important, the ability to analyze data using machine learning and deep learning technology is a key activity in corporate activities. Nevertheless, companies consider it difficult to introduce machine learning and artificial intelligence technologies because they need an understanding of the business as well as data and analysis algorithms. Accordingly, services such as automated machine learning have emerged for easy use of machine learning. In this study, the authors explored the automated machine learning service and compared the random forest and extreme gradient boosting analysis results using WiseProphet and Python. WiseProphet is used as a representative of automated machine learning solutions because it is a cloud-based service that anyone can easily access and can be used in various ways. It is contrasted with the model implemented by Python, which writes code with No coding. As a result of comparing the prediction performance, WiseProphet automatically outperformed the analysis result by parameter optimization.

KEYWORDS

Automated Machine Learning, Kaggle Data Set, Machine Learning, Wiseprophet, ML Performance Metrics

INTRODUCTION

The Fourth Industrial Revolution, which refers to digital transformation, created a digital economic system in which economic activities were carried out as a major factor in digital input such as digital technology, infrastructure, and data. As the COVID-19 pandemic promoted a non-face-to-face society centered on digital technology, the digital era came as digital technology was brought

DOI: 10.4018/IJSI.315656

*Corresponding Author

into a mainstream way of life, not an instrumental dimension for efficiency. Changes in the current economic environment require the use of strategic digital technology by companies (Jeong, 2019).

Through this, the corporate mindset is changing from product-oriented thinking to solving problems and inconveniences experienced by customers. Key accelerating factors include Connectivity and data collection (Lee et al, 2021). D.N.A (Data, Network, AI) is a key technology in the digital era and is rapidly growing around new businesses of companies using D.N.A. The existing resource and capital-oriented business model is not only changing to a data and artificial intelligence based business model, but also developing differentiated services using machine learning in existing industries such as finance and energy. In addition, D.N.A. is used in corporate activities in various ways such as reducing repetitive tasks or reducing existing costs by using it as a reference for important decisions(Jeong & Kim, 2021).

However, the Korea Institute for Information and Communication Policy suffers from a lack of quality data, difficulty in specifying tasks, understanding AI technology, and lack of internal talent in preparation, model development, and service operation after introduction. In a survey of 152 AI demand and supply companies in Korea, 34% of respondents chose a shortage of internal manpower as an obstacle to the introduction of AI technology (total %: 200%)(Lee&Kim, 2021). Also, Glue Coding, a simple task of putting multiple cords together, accounts for 90% of the total work in the stage of model development using machine learning. The version management of machine learning, the heterogeneity of the development environment, and the actual environment cause inefficiency and causes machine learning project failure(Valohai et al, 2021).

Automated Machine Learning means automating the steps from data preprocessing to model learning to build a model using machine learning. In automatic data processing, feature selection, machine learning algorithm selection, and parameter optimization, unnecessary repetitive tasks in place of human settings and coding can be reduced to enable efficient machine learning utilization projects(Simkek, 2019; Dawid, 2021).

Therefore, the introduction of Automated Machine Learning can contribute to the development of new business models through the universal data utilization of companies, as companies can reduce the probability of failure by reducing inefficiency in machine learning projects.

Thus, this study used python to evaluate the performance of automated machine learning that is being introduced. Random Forest and Extreme Gradient Boosting models were used to predict prediction accuracy. For comparison data, ‘the German Credit Prediction’ dataset(UCI machine learning, 2016), which is released on the big data platform Kaggle, was used.

THEORETICAL BACKGROUND

Machine Learning Project and Automated Machine Learning

The Machine Learning project goes through the process of Data collection, Data Cleaning, Feature selection and Engineering, Model selection, Hyperparameters Tuning, and Model validation and Exploration, as shown in Figure 1 (Simsek, 2019).

Data cleaning is not perfect, so machine learning models can learn properly by processing noise present in collected data with errors (Chai, 2020). Feature Selection is a task necessary to eliminate overfitting and shorten learning time from real data mixed with variables necessary and unnecessary to predict dependent variables (Jovic et al., 2015). If Feature selection is to find input variables, Feature Engineering is to transform input variables into ranges and shapes so that modeling can perform better (Zheng & Casari, 2018).

Hyper parameter refers to the value reflected in the model learning process. It refers to values such as a learning rate, a loss function, and a batch size that humans can adjust in advance before starting learning. Since it is a value reflected in the learning process, optimizing the Hyper-parameter has a great influence on machine learning performance. Hence, it is a very important process to explore the values that can improve learning outcomes the most using methods such as Bayesian.

Related Content