Prediction of Concrete Strength Using XGBoost

Gen Li *

Department of Financial Mathematics, Monash University, Australia

* Corresponding author email: glii0031@student.monash.edu

Abstract. This paper aims to test the extent to which different materials affect concrete strength and to use Machine Learning to test the accuracy of different mathematical algorithms in predicting concrete strength to select the most appropriate method to predict concrete strength through a given database. There is a strong need to predict concrete strength, as cement can only show its solidity after it has solidified and been put into use. So, it is not possible to propose some samples for strength testing before each use of cement to ensure safety. Although this is indeed a method, it is too expensive and cost-intensive. Thus, it is not cost-effective to implement. In this paper, some data was found through the academic forum and analyzed using Python, the degree of correlation between different parameters for concrete strength was summarized, and different algorithms were used to compare the accuracy. Finally, XGBoost was successfully used to predict concrete strength.

Keywords: Concrete Strength; Machine Learning; Regression.

1. Introduction

Concrete is one of the most important civil engineering materials today. It is an artificial stone made of cementitious materials, granular aggregates, water, as well as admixtures and admixtures in a certain proportion, and hardened by uniform mixing and maintenance. Concrete has many characteristics that make it the most popular building material, such as abundant raw materials, low prices, and simple production processes. At the same time, concrete also has many advantages, such as high compressive strength, good durability, and a wide range of strength grades. These characteristics make its use very wide and are involved in various fields. Concrete is widely used in many fields apart from construction. During the process of construction, checking the strength of the concrete to confirm that concrete has the required strength is always the first step [1]. After all, people cannot wait until the building collapses to know that the concrete strength is not enough, so it is necessary to know the approximate strength of the concrete according to the scale in the process of making concrete. The compressive strength of concrete determines its quality of concrete. However, the traditional cycle of obtaining concrete strength through traditional standard crushing experiments is at least 28 days [2]. Thus, how to predict if the concrete has sufficient strength has become a topic worth studying.

Machine learning (ML) is part of artificial intelligence (AI), a field of computational science that analyzes and interprets the patterns and structures of data for learning, reasoning, and decision-making without human interaction. At this point, machine learning can predict the strength of concrete through the number and proportion of raw materials to obtain the strength of concrete in a shorter time. This work seeks to explore the applicability of different classical ML models in concrete strength prediction, including Support Vector Machine (SVM), decision tree models such as decision tree and their successors, and several gradient boost methods. The experiment results show that the XGBoost method achieves the best performance of 0.882 accuracy.

2. XGBoost Model

XGBoost is a branch of boosting algorithms. The central theme of boosting algorithm is to combine many weak classifiers together so that forming a strong classifier [3]. Since XGBoost is a lifting tree model, it is a strong classifier that integrates many tree models together [4]. However, the specific XGBoost algorithm process is not the focus of this project, so XGBoost is only used here as a known tool without in-depth research.

As mentioned before, test data size accounts for 30%, and train data size accounts for 70%. Then, the best model parameters under this situation can be computed by Python code. Therefore, the predicted parameters and results can be expressed as [5,6]:

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (1)$$

Where $L(\phi)$ denotes the expression in linear space, i means the number of i th samples, k means the number of k th tree, and $\hat{y}_i$ means the prediction value of i th sample in x_i.

3. Experiments

Since there are many algorithms in machine learning, and different algorithms have their own characteristics, each algorithm's performance is different in different environments. So, in our experiment, author compare the accuracy of different algorithms and find that XGBoost has the best performance. Author implemented the algorithms in Python using corresponding libraries [7].

3.1 Dataset

For data science, the first step in machine learning is always to get a dataset, and Kaggle is a good choice to get a lot of data. This dataset contains 1030 samples, and each data contains 9 parameters. For each model to be tested, the dataset will be divided into test and train. The significance of this step is to test the performance of the model in the real environment. In this project, the test size accounts for 30%, and the training size accounts for the rest.

3.2 Result and Discussion

Table 1 shows the overall prediction of all the algorithms tested. The XGBoost has the best accuracy in predicting concrete strength at 0.882, followed by the Random Forest Model, which achieved 0.878 [8].

Table 1. Accuracy for different algorithm

Algorithm	Random Forest	Gradient Boost Regressor	Ada Boosting Regressor	KNN Regressor	Bagging Regressor	Support Vector Regressor	XGBoost	Decision Tree
Accuracy	0.877	0.761	0.767	0.754	0.861	0.655	0.882	0.779

Fig. 1 shows the prediction details produced by the XGBoost. The blue lines represent original data, and the orange lines represent predicted data. It actually shows the prediction of the strength of the concrete, and the two lines are generally very similar, so it means that the prediction model made with XGBoost is very accurate and works very well.

Apart from superb prediction accuracy, the advantage of the XGBoost model is that after boosted trees are constructed, the importance of each attribute for the dependent variable, concrete strength, can be retrieved [9].

A trained XGBoost model automatically calculates feature importance on this prediction model, and show the importance histogram in Fig. 2, where the x-axis represents the F score, and the y-axis represents the features [10]. Each feature corresponds to a different score, which represents the degree of influence of different materials on the strength of concrete in the composition of concrete. The 8 features shown by Y-axis, from f0 to f7, represent cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age, respectively. In reality, concrete may be more than these materials, but these 8 raw materials already cover most of the situation. The data was downloaded from Kaggle and had enough authenticity. It is easy to see in the table that f7 has the highest score, followed by f0 and f3, and the lowest score is f2. In other words, the strength of concrete has a great connection with the solidification time of concrete, and a longer period of time can make concrete more fully solidified. Thus, its strength has a positive correlation with age. Moreover, the

strength of the concrete is directly related to the water-ash ratio configured at the beginning. If the cement strength is relatively high, the strength of the concrete in the later period will also be enhanced. Since concrete strength is the result of intertwined factors, it is known from the experimental results that the best concrete strength prediction method may be the feature importance method mentioned above.

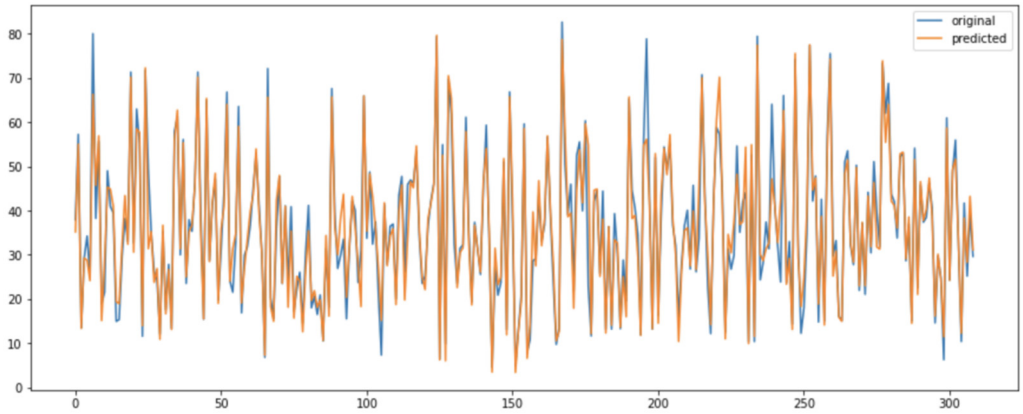


Fig 1. Concrete strength prediction graph

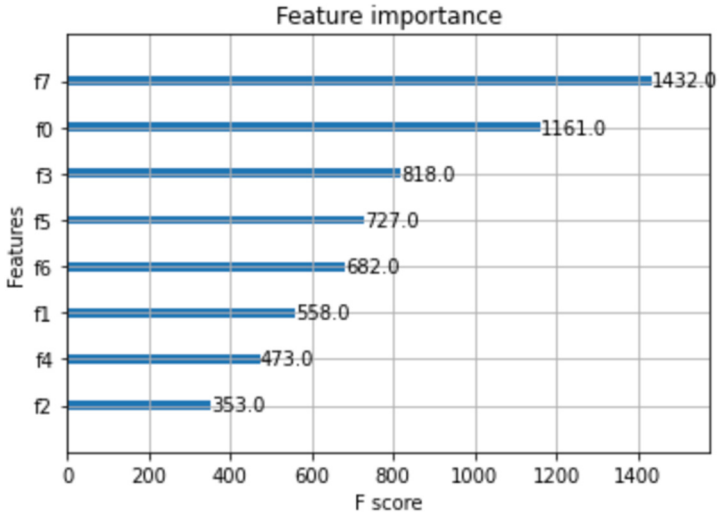


Fig 2. Feature Importance

4. Conclusion

Concrete strength is an utmost important factor for construction. However, traditional strength measuring methods take at least 28 days to get the result. This paper aims to explore the applicability of different classical ML models in concrete strength prediction and find the best suitable model. In the experiment, the XGBoost model outperforms other methods with an accuracy of 0.882. Furthermore, the model's prediction of concrete strength is not like other models that occur over time, such as the population growth model. There is no growth curve for the prediction of concrete strength, and the process of predicting the strength is based on the proportion of the components in the concrete. So, this is also one of the factors that chose XGBoost as the best choice for predicting the strength of concrete. The model also outputs feature importance, enabling users to analyze the quantity and proportion of components in the data to determine the level of concrete strength.

References

[1] D.-C. Feng, Z.-T. Liu, X.-D. Wang, Y. Chen, J.-Q. Chang, D.-F. Wei, and Z.-M. Jiang, "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach," Construction and Building Materials, vol. 230, p. 117000, 2020.

[2] H. Muliauwan, D. Prayogo, G. Gaby, and K. Harsono, "Prediction of concrete compressive strength using artificial intelligence methods," in Journal of Physics: Conference Series, vol. 1625, no. 1. IOP Publishing, 2020, p. 012018.

[3] Total Data Science. Machine Learning-Part 1-Predicting Compressive Strength Of Concrete | Feature Engineering. Sep 20, 2020.

[4] Total Data Science. Machine Learning-Part 2-Predicting Compressive Strength Of Concrete | Feature Engineering. Sep 20, 2020.

[5] Total Data Science. Machine Learning-Part 3-Predicting Compressive Strength Of Concrete | Feature Engineering. Sep 20, 2020.

[6] Young win, XGBoost principle and formula derivation. Jan 11, 2018.

[7] Qilin Liu, XGBoost principles, formula derivation, Python implementation, and applications. Jan 23, 2021.

[8] Jason Brownlee. Feature Importance and Feature Selection with XGBoost in Python. August 31, 2016 in XGBoost.

[9] Dr. Bright (PhD in Data Science). The Full Stack Data Scientist BootCamp, Stats| Python| SQL| Machine Learning & Cloud| Deep Learning| A.I | Computer Vision & NLP| Power BI | Internships & Jobs. 2022.

[10] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.