Performance of Data Driven Algorithms to Predict Concrete Strength Using Production Raw Data

Arnaud Delaplace[1](✉), Ulli Olivetti Razinhas[1], Régis Bouchard[1], and Andreas Griesser[2]

[1] Holcim Innovation Center, 38291 Saint Quentin Fallavier, France
arnaud.delaplace@holcim.com
[2] Holcim, Zementweg 1, 5303 Würenlingen, Switzerland

Abstract. Predicting properties of concrete is a major issue for building sustainable structures. In the last decade, numerous publications have shown that machine learning algorithms can play a major role to predict these properties. The key factor is the availability of data to train the models. Collecting, cleaning and consolidating data can be challenging tasks, especially in a concrete industry in which the digitalization of the supply chain is still in progress. We propose in this study to use raw production data to evaluate the performance of a machine learning algorithm compared to an empirical model. The concrete strength value is predicted using both approaches and compared to the measured value. Even if machine learning algorithm shows good performance, no significant increase in the prediction accuracy is obtained.

Keywords: Strength prediction · mixdesign optimization · machine learning

1 Introduction

Designing a concrete becomes a more and more challenging task: the number of raw materials used in a concrete recipe is increasing, the available types of raw materials is expanding, the criteria and constraints of optimization is higher (CO2 footprint, cost, workability, strength, durability…). Complimentary to an experimental characterization, the possibility to predict concrete performance is needed. Numerous physical and phenomenological models have been developed, allowing an accurate prediction of properties. However some of these models become less accurate when dealing with cementitious materials (binary or ternary blended binders) or admixtures. As hydration of blended binders is a complex and strongly coupled process, developing new models can be challenging and time consuming.

In the last decades, data driven approaches have been proposed to predict concrete properties, with more or less efficiency. The key factor is the availability of consolidated and reliable data, with a sufficient amount to train the models. Unfortunately, such databases are still rare. As a result, a lot of studies are based on a collection of data

A. Jędrzejewska et al. (Eds.): SynerCrete 2023, RILEM Bookseries 44, pp. 695–702, 2023.
https://doi.org/10.1007/978-3-031-33187-9_64

issued from different sources, with different cements, different aggregates, limiting the relevance of the training. Without any doubt, the availability of large data set will be obtained in the next years, thanks to the connectivity of batching software, Quality control systems and IoT objects, with cloud-based storages.

Meanwhile, we propose in this study to evaluate the performance of machine learning algorithm compared to a phenomenological model, using production raw data, to predict concrete strength. The rationale is to evaluate the interest of substituting an existing model with a data-driven algorithm in existing quality control software. In a first part, the models used in this study are described. Then, a presentation of the data will be done. In a third part, the two algorithms will be tested and compared to the measured values.

2 Predictive Models

2.1 Empirical Models

Compressive strength is one of the most valuable properties of concrete. Numerous models have been developed to predict strength performance, based on the mix design (components dosage). The most known factor when looking at compressive strength is the water to cement ratio (*w*/*c*). Based on this consideration, the strength is predicted using a model derived from the well-known Bolomey's formula (see [1] for a review of most common strength models). It reads:

$$f_{c28} = K_g K_c \left(\frac{M_c + \sum_i M_{Ai} \times k_{Ai}}{W + A} - 0.5 \right) \tag{1}$$

where K_g is the aggregate coefficient, K_c is the cement coefficient, M_c is the cement dosage in kg/m^3, M_{Ai} is the dosage in kg/m^3 of the addition *i*, k_{Ai} is the activity coefficient of the addition *i*, *W* the effective water content in L/m^3, *A* the air content in L/m^3. Even if this empirical model is not the most accurate one, its main advantages are its simplicity and the limited number of properties needed to predict strength. It explains why the model is implemented in different production software in order to have an estimation of the concrete strength. It will be used as the reference model for this study.

Other models are more efficient (see for example models based on the Gel Space Ratio concept [2, 3], or on the packing density [4]), but are more complex to implement and require more material properties that are not always available.

2.2 Machine Learning Approach

Machine Learning is a branch of the Artificial Intelligence [5]. It's a set of statistical tools and algorithms that allow a model to learn from big data, in an iterative auto improvement process, to make predictions and classifications without needing any extra guidance. When we look at the Regression Analysis techniques, they differ by the fact that Machine Learning works without the use of predetermined equations. In general terms, Machine Learning can be seen as algorithms that learn by experience.

The general workflow can be summarized in the following steps:

1. Data are cleaned and filtered.

2. Data is divided into a train set and a test set.
3. The train set will be used as the input data to first build the model by refining the parameters of the prediction function, according to the statistical metric chosen.
4. Then the model is tested with the test set to check if it's able to be generalized for other input values.
5. If that is the case, then the model is validated and can be run with new data to predict new output. If not, additional data has to be considered to build the model again.

Many authors have worked on developing machine learning algorithms to predict some concrete indicators of performance, like durability, slump, compressive strength, elasticity modulus, etc. [6–8]. The objective of this study is not to develop a new algorithm, but to compare the performance of an algorithm available "off the shelf". Our choice fell on the Light Gradient Boosting Machine (LightGBM), based on decision tree algorithms, well known to be efficient and scalable [9]. A Python implementation will be used in this work.

The procedure to run the model is described next:

1. Standardization of the data, mandatory to be sure that the algorithm will learn from all properties even if their variances are small compared to others
2. Shuffle and split the dataset into test and train – apply K-Fold cross-validation
3. Define the model parameters. In our case, the LGBM gradient boosted decision trees method will be used. The main advantage of this method is its stability and reliability.

3 Production Data

Production data from three concrete plants, from 2019 until mid-2022, have been used. Data come from three different sources:

- Batch file: it contains the exact composition of each concrete produced, including the mass and the volume of each component, the amount of water in each component, and the mix design
- Quality control file: it contains information related to nominal and measured values for the concretes, including the air content, the nominal compressive strength, the 28-day measured compressive strength and the batch date and time.
- Technical sheets of raw materials: it includes the properties of the raw materials provided by the supplier.

The quality control and batch files are linked through a Delivery Note Number that is unique for each delivered concrete. A total of 81474 deliveries are recorded, corresponding to 390 different mix designs. From this collection, 1703 passed through a compressive strength test, covering 197 different mix designs. It is worth mentioning that even if the concrete volume produced globally is huge, just a very limited number of data are consolidated and usable today.

The raw materials used in the 1703 concretes are: 10 different binders, 16 different aggregates, 6 different additions, 31 different admixtures. These raw materials are not equally used in the concretes. For example, Fig. 1 shows the usage frequency of the different aggregates:

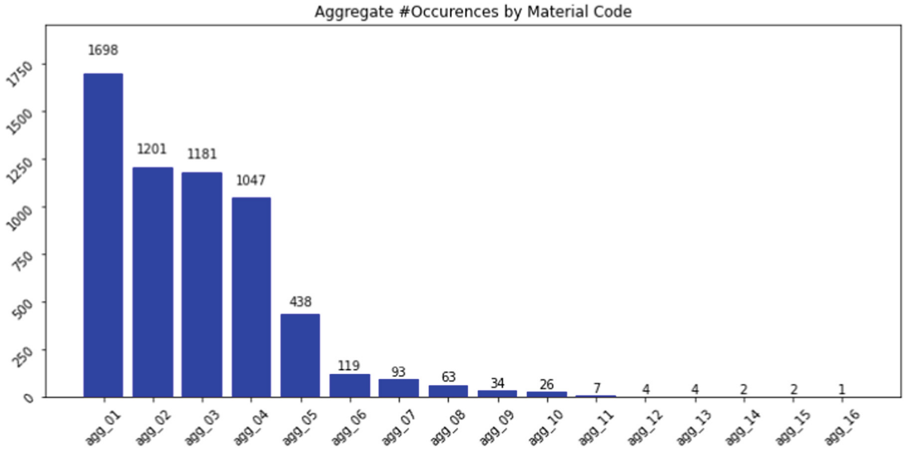

Fig. 1. Aggregates occurrence in the concrete mix designs

Just five of them are used extensively. We finally focus on concretes with raw materials present in more than 50 concretes, limiting the data to 4 different binders, aggregates classified in 4 different types, 1 addition, and admixtures classified by types (Superplasticizer, Air entraining agent, Retarder, Antifreeze, Shrinkage reducing agent, Viscosity modifying agent and Hardening accelerator). For concrete mixes with multiple aggregates or cements, a unique coefficient is computed based on the mass proportion of each single material coefficient.

4 Strength Prediction

4.1 Bolomey Model

In order to use the Bolomey model, we need to get some material coefficients:

- K_c: this coefficient is usually equal to the cement normalized strength at 28 days. This value is obtained from the cement technical sheet.
- K_g: the aggregate coefficients can be obtained by fitting the Bolomey's model for standard normalized concrete of different strength. We don't have these values from production data and we need to make some assumptions: the most important one is that we consider that it is an intrinsic property of the aggregate (in reality the quality of the paste has also an influence on this coefficient through the Interfacial Transition Zone). Note that some reference values, depending on the aggregate size and nature, are available.
- k_{Ai}: the addition coefficients can be obtained experimentally like for the aggregate coefficient. Again, without having this information, some assumptions have to be done. Note that some reference values can be found in the literature (0.25 for Limestone Filler, 0.40 for Fly Ash, 0.60 for Slag).

Different subsets of the database will be considered to fit the aggregate and addition coefficients: a first dataset is built with all concretes without additions, allowing fitting the aggregate parameters. A second dataset with concretes using the addition is used to get addition reactivity coefficient.

Table 1 gives the coefficient identified for the different materials, and Fig. 2 shows the predicted strength values *vs.* the measured ones. The Root Mean Square Error (RMSE) and the R-Squared measure (R^2) are used to estimate the accuracy of the prediction.

Table 1. Bolomey's coefficients and model performance

K_{agg1}	K_{agg2}	K_{agg3}	K_{agg4}	K_{a1}	RMSE	R^2
0.73	0.55	0.51	0.60	0.42	5.9	0.61

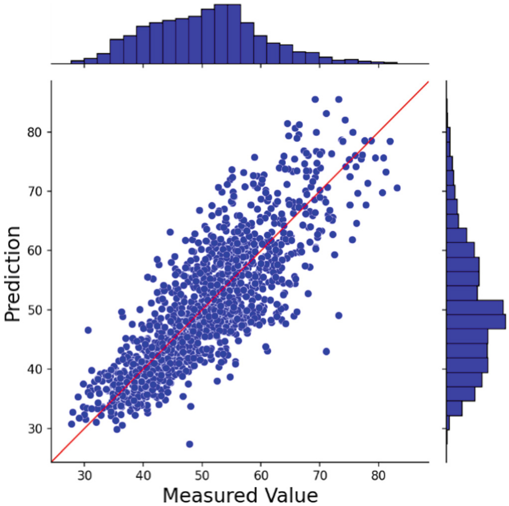

Fig. 2. Prediction of compressive strength in MPa using Bolomey's model vs. the measured values

We note that the strength prediction of high performance concretes is worse than for standard concretes. It's a well-known limitation of Bolomey's model.

4.2 Machine Learning Model

The selected features used for the ML model are:

- the dosage of each components,
- the total dosages of binders, aggregates and admixtures,
- the air content,
- the w/c ratio.

It corresponds to a total number of 20 features, directly linked to the concrete design. Different choices of features have been tested, such as reducing the total number by considering the total dosage of aggregates, or increasing it by considering the dosage of each different admixtures. No significant improvement in the model accuracy was observed, *i.e.* a decrease of the RMSE higher than 0.2. 20 features is a good balance for this dataset, as reducing their number speeds up computation time and avoids any overfitting, but with a lower accuracy, while increasing them does not improve the accuracy due to the limited number of data available.

For the model configuration [10], the learning rate is 0.005. In the K-Fold validation, K = 5 has been used. 75% of data are used for training, 25% for testing the model. The final results are given in Table 2 and Fig. 3. A better accuracy of the prediction is obtained with this model. In particular, the prediction of high performances is much better than Bolomey's model.

Table 2. ML model performance

RMSE	R^2
4.8	0.75

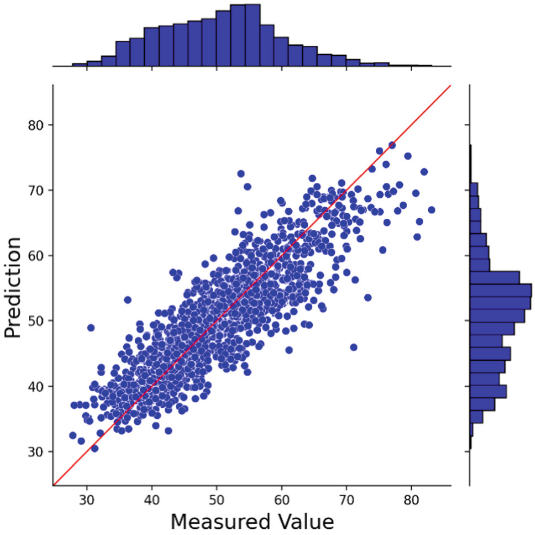

Fig. 3. Prediction of compressive strength in MPa using ML model vs. the measured values

The ML model gives also meaningful information regarding the importance of each feature in its learned model. Figure 4 shows the importance of each feature. Without any surprise, w/c ratio is the primary factor.

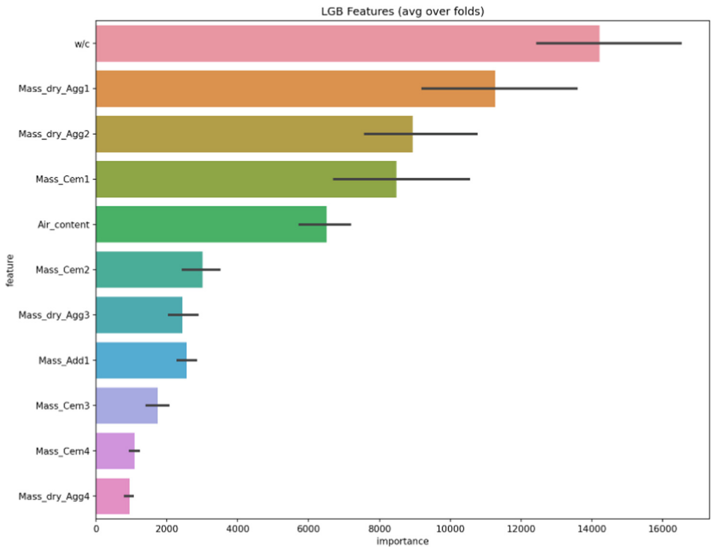

Fig. 4. Classification of feature importance obtained by the LightGBM model

5 Discussion

An empirical and a machine learning models have been used to predict 28d-strength of concretes, using raw data extracted from batching and Quality Control software. The objective is to evaluate if a significant improvement of prediction is obtained using the ML model. The conclusion of the analysis is just applicable to the used models and to the dataset available, but at least some comments could be drawn:

- Even if the volume of produced concrete is huge, the exploitable data remains small with the actual software used in production.
- A simple and old model like Bolomey's one performs well, excepted for strength higher than 65 MPa: the model doesn't take into account the quality of the aggregates packing for the aggregates or the interaction between the paste and aggregates. Its main advantage is its simplicity, the limited number of inputs and the ease of implementation in any software.
- A ML model provides more accurate results, but without a significant improvement for ordinary concretes for this dataset. Its implementation is less straightforward and it can be perceived as a black box.

6 Conclusion

The conclusion of this study could be that there is no rush to adopt ML models today. Nevertheless, we can predict that the number of available data will increase significantly in the coming years. The consequence is that ML models will perform better and will overpass the simplest empirical models. It is time to adapt the RMX industry to this new trend, by building consolidated databases, accessible and well structured. It's also important to develop standard and automated procedures for measurement, all along the RMX supply chain. This will allow as well having access to additional properties that are difficult to predict using physical or empirical models, due to the complexity of the phenomena.

References

1. Neville, A.M.: Properties of Concrete. 5^{th} edn. Pearson (2011)
2. Powers, T.C., Brownyard, T.L.: Studies of the physical properties of hardened Portland cement paste. Res. Lab. Portland Cem. Assoc. Bull. **22**, 101–992 (1948)
3. Acker, P.: Micromechanical analysis of creep and shrinkage mechanisms. In: Ulm, F.-J., Bažant, Z.P., Wittmann, F.H. (eds.), Creep, Shrinkage and Durability Mechanics of Concrete and Other Quasi-brittle Materials, 6th International Conference, Elsevier, Amsterdam, pp. 15–26 (2001)
4. De Larrard, F.: Concrete Mixture Proportioning: A Scientific Approach. CRC Press, Boca Raton (1999)
5. Copeland, B.J.: Artificial intelligence. https://www.britannica.com/technology/artificial-intelligence (2022)
6. Dao, D.V., et al.: A sensitivity and robustness analysis of GPR and ANN for high-performance concrete compressive strength prediction using a monte carlo simulation. Sustainability, **12**, 830 (2020)
7. Cui, L., Chen, P., Wang, L., Li, J., Ling, H.: Application of extreme gradient boosting based on grey relation analysis for prediction of compressive strength of concrete. Adv. Civ. Eng. **2021**(5), 1–14 (2021)
8. Ahmed, A., Jin, W., Ali, M.: Artificial intelligence models for predicting mechanical properties of recycled aggregate concrete. critical review. J. Ad. Concr. Technol. **20**, 404–429 (2022)
9. G. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
10. Bahmani, M.J.: Understanding lightGBM parameters (and how to tune them) (2022). https://neptune.ai/blog/lightgbm-parameters-guide