Concrete Truth: Predicting concrete strength and carbon footprint
Executive Summary

Project
Concrete is the most important building material and has a significant carbon footprint, mostly due to its cement content. Our goal is to eventually use a model to find mixtures that have a prescribed strength but minimize CO_2 emissions. To this end, we compare regressors to best predict the strength of concrete from its ingredients and age.
While the optimization is beyond the scope of our project, we pave the way for a future study.

Data Description
Our dataset is available at UCI's Machine Learning Repository and kaggle. The data consists of 1030 instances with 9 columns, consisting of 7 ingredients, age, and strength.
There are no missing entries, but we found outliers in the instances with age greater than 120 days. Indeed, these data points comprised only 5% of the total set and had value 0 for a few other features. Whether or not this is due to missing information, we did not find it representative of the rest of the data.

Models
We considered 6 models for regression: LinearRegression, RandomForestRegressor, XGBoostRegressor, HistGradientBoostingRegressor, MLPRegressor, and keras. The linear model served as a base, and we used the exponential relationship between concrete strength and age to improve it. Initial success with random forest led us to try more effective ensemble methods. We also tried some neural networks but had some trouble getting them to converge. After some initial tuning using grid searches, we compared the models, using the metric mean squared error. Under the same cross-validation, the XGBoost model performed best.

Results
After further refining the XGBoost model, we tested it on the test data and found that it fit as well as our cross-validations had indicated. Compared to the base linear model's mean squared error of 49.07, the best model had an error of 18.88, less than 40% of the base.

Applications
Our best model works as an effective tool for modeling concrete strength in any context. Any concrete manufacturer wanting to make concrete with prescribed strength will find it useful. One can enter amounts of ingredients to find the predicted strength until they find a mixture one likes.
But we can also develop a type of inverse function that recommends mixtures based on desired strength. To answer our initial question about CO_2 emissions, we can also factor in emissions as well to recommend mixtures that minimize carbon footprint.