Machine Learning Practise
Playground Series - Season 3, Episode 9

Ben Colquhoun

7/3/23

Outline of the Problem

The task here is to predict the strength of concrete, given an instance of the concrete. This task was found as a Kaggle competition -"Playground Series - Season 3, Episode 9", joined 7 days into the 2 week duration.

Exploratory Data Analysis

Initial Features

Inital investigation of the data provided the following features from the the training set, with expanded explanations with reference to a Kaggle discussion of Phong Nyugen [1]. There are 5407 instances in the training dataset.

- id
  - Id of the instance
  - Integer
  - Unimportant feature for comparing strength of concrete
- CementComponent
  - Amount of cement added to concrete
  - Float
  - More cement improves strength of the concrete, though too much may cause it to be brittle. Important feature.
- BlastFurnaceSlag
  - Amount of slag added to concrete

  - Float
  - Slag can be used as a cement substitute in creating concrete. Important feature.

- FlyAshComponent
  - Amount of fly ash added to concrete
  - Float
  - Fly ash can be used as a cement substitute in creating concrete. Important feature.

- WaterComponent
  - Amount of water added to concrete
  - Float
  - Water is the binding agent of the other components. Important feature.

- SuperplasticizerComponent
  - Amount of superplasticizer added to concrete
  - Float
  - Chemical additive to improve workability without changing water content. Reasonably important feature.

- CoarseAggregateComponent
  - Amount of coarse aggregate added to concrete
  - Float
  - Gravel/crushed stone added to cement for structure. Reasonably important feature.

- FineAggregateComponent
  - Amount of fine aggregate added to concrete
  - Float
  - Sand added to cement for structure. Reasonably important feature.

- AgeInDays
  - Days since concrete poured
  - Integer
  - Longer drying times increases the strength up to a point dependant on other factors. Important feature when combined with other factors.

All of the above features have maximum and minimum values that make sense for the context that they are in.

Missing Values and Duplicates

Investigation into the duplicated values showed that there were no full instance duplications (with and without the ID), however discussions about duplicated data revealed that identical features could result in a wide variation of strength values [2]. This is evidenced by the following distributions of strength values in Figure 1.

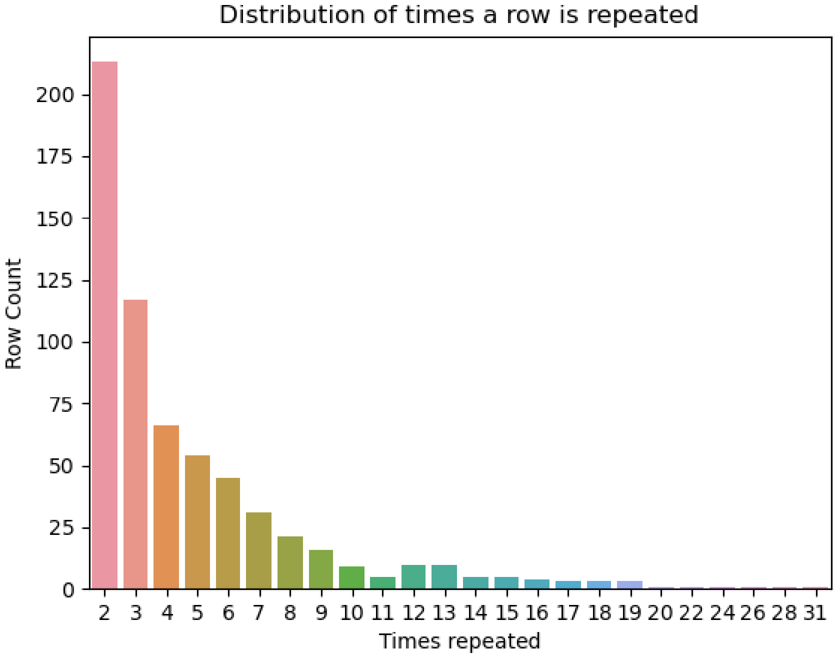

Figure 1: Duplicated Row Counts

Dealing with this may require the mean strength for the duplicated row to be calculated and then input. This is especially the case as the range between the given strength values of these duplicated rows can vary wildly, as seen in Figure 2.

Feature Engineering

Possible feature additions have been considered and are listed below. Some of the entries are from Phong Nyugen's discussion [1]. ChatGPT was also used to provide assistance in contextual knowledge and combinations of features.

- TotalComponentWeight = CementComponent + BlastFurnaceSlag + FlyAshComponent + WaterComponent + SuperplasticizerComponent + CoarseAggregateComponent + FineAggregateComponent

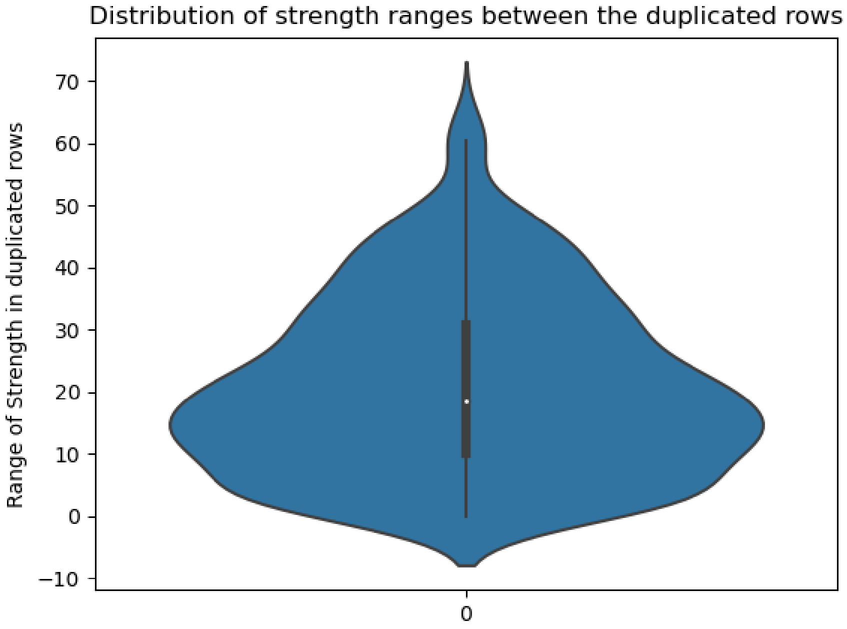

Figure 2: Strength Value Range over Duplicates

- Water/Cement Ratio = WaterComponent / CementComponent
- Aggregate Ratio = (CoarseAggregateComponent + FineAggregateComponent) / CementComponent
- Water/Binding Ratio = WaterComponent / (CementComponent + BlastFurnaceSlag + FlyAshComponent)
- Superplasticizer/Binder Ratio = SuperplasticizerComponent / (CementComponent + BlastFurnaceSlag + FlyAshComponent)
- Cement-Age Relation = CementComponent * AgeInDays
- Age Factor = Age ** n, with n some factor

References

[1] Phong Nyugen. (2023, March). *Detailed feature description and feature engineering by ChatGPT*. Retrieved March 7, 2023, from `https://www.kaggle.com/competitions/playground-series-s3e9/discussion/391066`

[2] Matt OP. (2023, March). *The great duplicate saga*. Retrieved March 7, 2023, from `https://www.kaggle.com/competitions/playground-series-s3e9/discussion/391011`