In this task, you are expected to build a regression model with the given dataset, which is sampled for various categories on different days. We dump data of the same day into files named `data.*i*.csv` in csv format. Each sample row should look like: C, Y, F1, F2, F3, ….

Meaning of each column:

1. C: is the categorical label, for example `C000`.

2. Y: is the variable to predict

3. N\_k: is the k-th numerical feature

We respect the time order of the data: the files are indexed according to their date order. For any given date *i* and category *c*, its samples are also ordered by time. Samples from different categories on the same day do not necessarily appear in time order. For the category label C, it is up to you whether to include it as your model input. Out sample data may contain both seen and unseen labels from the training data. The features are the same for in-sample and out-sample. Your goal is to minimize out-sample Root Mean Squared Error (RMSE).

1.     Out sample Root Mean Squared Error (RMSE)

2.     Merits in the report. It’s important to show your work (just like in a math proof).