Cross Validation:-

Cross validation is a technique in which we train our model using the subset of the data set and then evaluate using the complementary subset of dataset.

* While using traintest split we split our data into train-test split. Means the data~~set~~ will be randomly splitted. we have a major drawback in this which is if i split 70% training data and 30% testing data, if some important data is with testing data then our accuracy will fall, because those important records need to be trained.

* The random_state parameter decides the spliting of data and accuracy, if random_state changes then accuracy will also change.

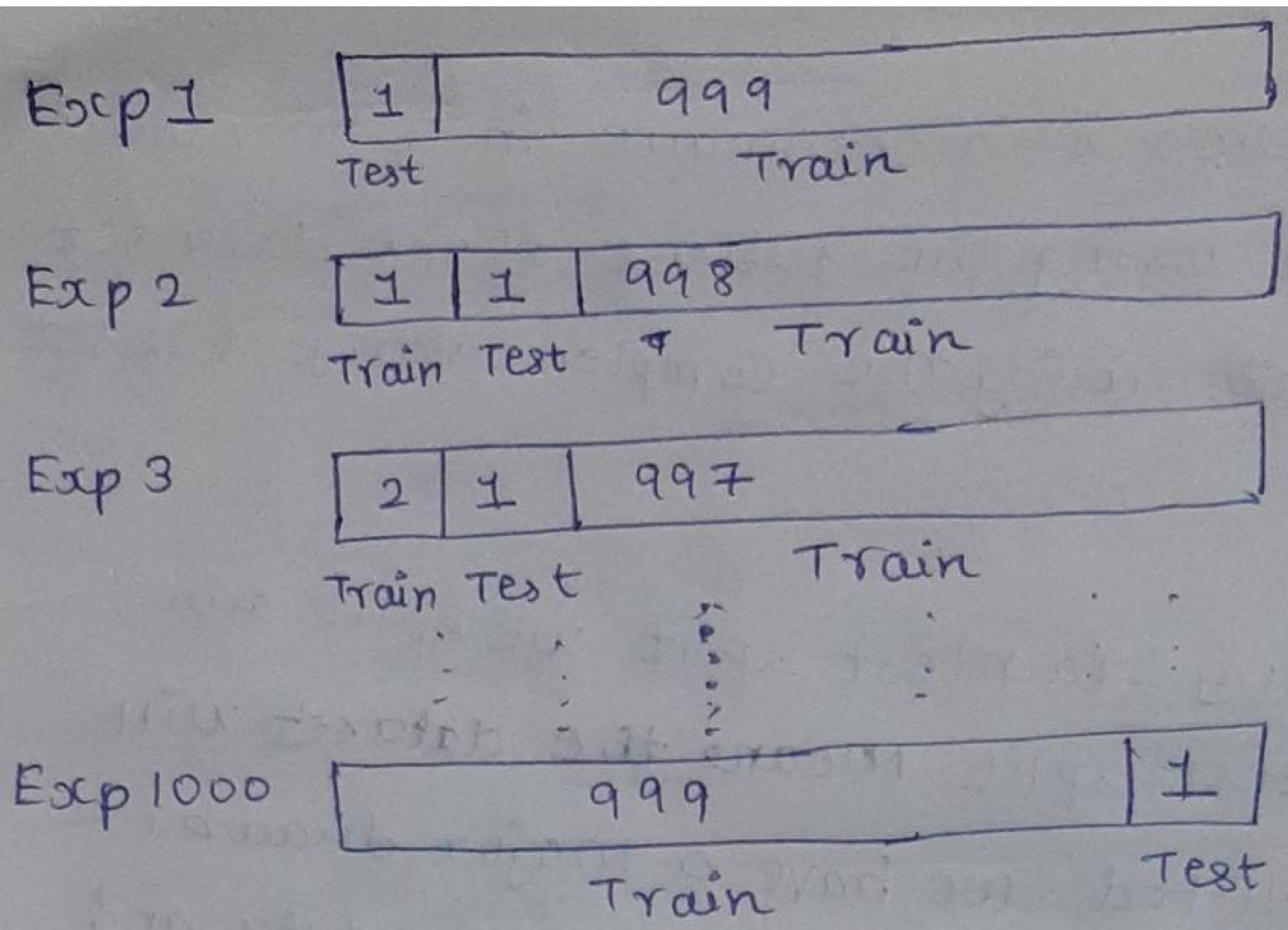
* So to overcome this difficulty the cross validation is done. There are different cross validation techniques.

- Leave one out cross Validation (LOOCV)

~~* lets take 1000~~

* Suppose we have 1000 data pts in the dataset

* In this method, we perform training on whole dataset but leaves only one data point as the testing. This iteration ~~co~~will continue till the ~~las~~ last datapoint.



* Like this LOOCV will be implemented, Basic idea is testing data will be 1 datapoint, that 1 will be keep on iterated to last data point.

* advantage of using this method is that we make use of all data points and hence its low bias.

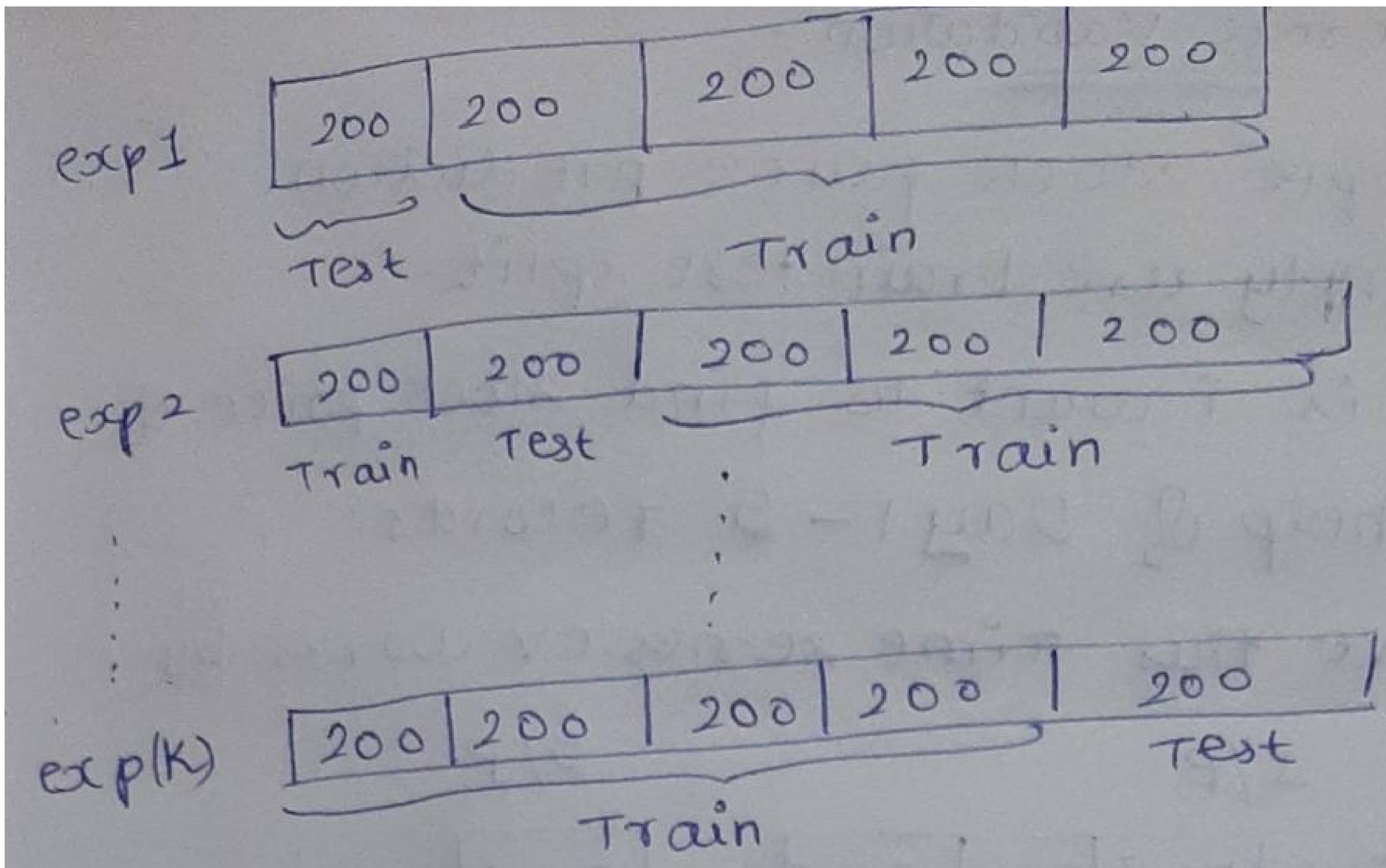
* drawback is it has high variance and it takes lot of computational time.

K-Fold cross validation:-

* In this method we will split the dataset into K subsets and we will use K-1 Part for training and remaining 1 subset for testing.

* Suppose we have 1000 records and K = 5

$$\frac{1000}{K} = \frac{1000}{5} = 200$$



* advantage is it runs faster than LOOCV and has limited K iterations

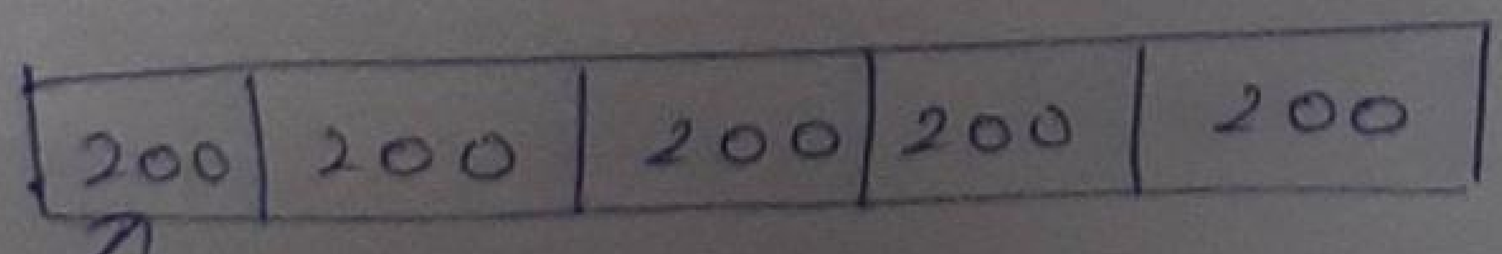
* disadvantage is while splitting into subsets the proportion of importance ~~will be~~ may be biased, like if it is a classification problem the first 200 data pts will have entry as 1.

* So it will be biased.

Stratified K-fold cross validation

* It is same as K-fold, but the disadvantage of K-fold is solved.

* Stratified K-fold CV ensures ~~that~~ the proper proportion of data in ~~th~~ each



each set will correct proportion of data.

* This technique should be used if dataset is imbalanced even after balancing in feature engineerin

Time series cross Validation:-

* For example stock price prediction we cannot simply use train test split.

* The case is i want to find stock price of Day 6 with help of Day 1-5 records.

* To achieve this Time series CV works by

	I/P					O/P
exp1	Day1	Day2	Day 3	Day4	Day5	Day6
exp2	Day2	Day3	Day4	Day5	Day 6	Day 7

* It's like cumulative addition of data for the successive prediction of ~~next~~ next O/P requirement

*