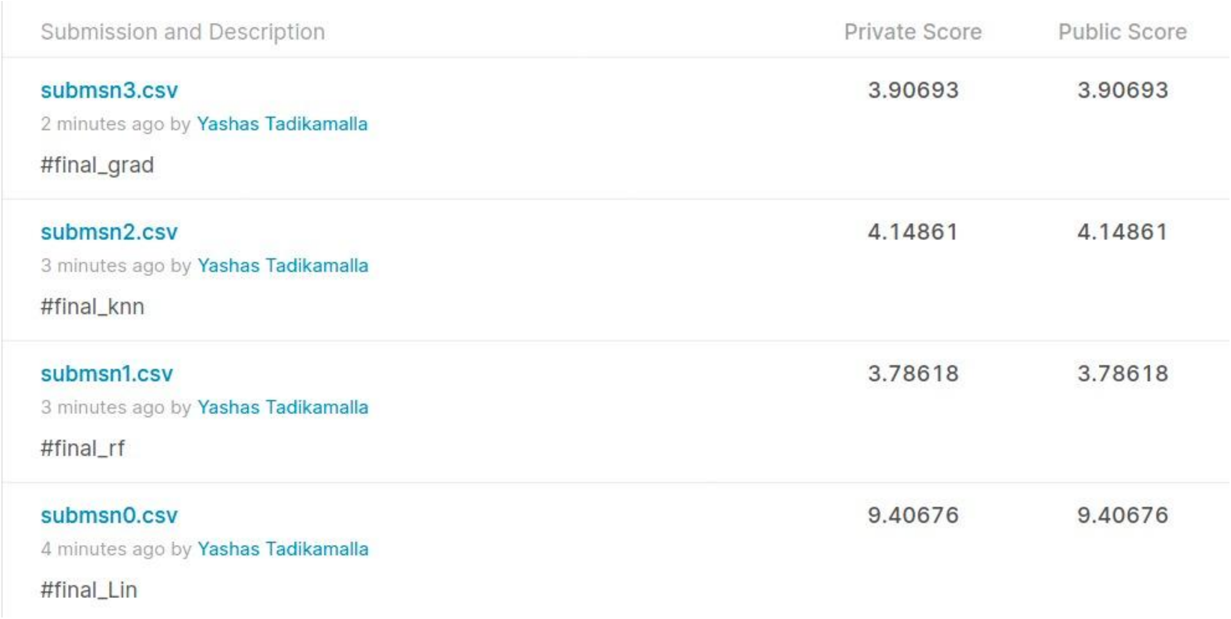
Assignment 4

Q6 Report



Submission and Description	Private Score	Public Score
submsn3.csv 2 minutes ago by Yashas Tadikamalla #final_grad	3.90693	3.90693
submsn2.csv 3 minutes ago by Yashas Tadikamalla #final_knn	4.14861	4.14861
submsn1.csv 3 minutes ago by Yashas Tadikamalla #final_rf	3.78618	3.78618
submsn0.csv 4 minutes ago by Yashas Tadikamalla #final_Lin	9.40676	9.40676

final_Lin: Linear Regressor
LinearRegressor()

final_rf: Random Forest Regressor (best)
RandomForestRegressor(max_features='sqrt', n_estimators=1000)
Ensemble model-**Bagging**.
Trains 1000 trees, employs bootstrapping, uses only sqrt(features) for training the trees.

final_knn: K Neighbors Regressor
KNeighborsRegressor()

final_grad: Gradient Boosting Regressor (next best, after Random Forest)
GradientBoostingRegressor(max_features='sqrt', n_estimators=1000)
Ensemble model-**Boosting**.
Trains 1000 trees, employs bootstrapping, uses only sqrt(features) for training the trees.

For comparison sake, all models have been trained on the same dataset size(1e5).

Linear Regression is underfitting. It assumes a linear trend, and tries to fit, which isn’t true with the data in this case. Hence, it doesn't perform well. It runs very quickly, but will underfit.

K Nearest Neighbors performs better than Linear regression, but it still gives an RMSE>4. Its performance will increase on training with more data. But with the amount of data used for comparison, it is not accurate enough. It runs very quickly, but needs a lot of data to produce better results.

Gradient Boosting and Random Forest being ensemble methods, in general perform better than models like Linear regression and KNN.

Random Forests train multiple decision trees and reduce variance by majority voting. Due to max_features=sqrt(n_features), it's robust to noise in data. It takes time to run, even for small datasets. If we increase dataset size or n_estimators, accuracy will increase, but so will time taken to fit the model.

Gradient boosting trains multiple weak learners, and makes predictions out of them. It takes time to run, even for small datasets, but can guarantee reasonably good accuracy. If we increase dataset size or n_estimators, accuracy will increase, but so will time taken to fit the model.