AI200 A4 Q6-Report

Vijay Varma - AI20BTECH11012

November 18, 2021

Question 6

The given training data is preprocessed to remove null values, add new features ('year','month','day','hour') and some features ('pickup-datetime','key') are removed.

And also, the latitudes and longitudes not in range of New York City are removed from the data, passengers more than 6 are removed etc.

Model 1 is trained using Light Gradient Boosting Machine (LGBM) Regressor and took 10000000 rows of training data for Training.

It's Kaggle Score is **3.57481** (RMSE of the test data).

Model 2 is trained using Random Forest Regressor (RFR) and took 1000000 rows of training data for Training.

It's Kaggle Score is **3.39883** (RMSE of the test data).

The First Model (LGBM) trains very quickly on the training data and does not overfit the data. It gives a pretty good score (3.57481).

The Second Model (RFR) trains slowly compared to Model 1 and it does not overfit on the data. It gives an increased score (3.39883).

Other Models are tried but they take too much too time for training and also does not give decent scores (RMSE).