The best model is No2.gradient descent with l2 from cross-validation. However, the best RMSE and MAE of 20% test data are generated by model 3.gradient descent.

**1.gradient descent with l1**

Mean Root Mean Squared Error for Stochastic Gradient Descent in 10 Fold CV:3342.92783953

Root Mean Squared Error for Gradient Descent with l1: **3256.9017271**

Mean Absolute Error for Gradient Descent with l1: **947.5683125**

**2.gradient descent with l2**

Mean Root Mean Squared Error for Stochastic Gradient Descent in 10 Fold CV:3264.18019481

Root Mean Squared Error for Gradient Descent with l2: **3575.24800168**

Mean Absolute Error for Gradient Descent with l2: **975.92327334**

**3.gradient descent**

Mean Root Mean Squared Error for Stochastic Gradient Descent in 10 Fold CV:3442.69061474

Root Mean Squared Error for Gradient Descent: **2903.39092908**

Mean Absolute Error for Gradient Descent: **928.713940702**

**4.Stochastic gradient descent with l1**

Mean Root Mean Squared Error for Stochastic Gradient Descent in 10 Fold CV:3354.51249405

Root Mean Squared Error for Stochastic Gradient Descent with l1: **3315.34004725**

Mean Absolute Error for Stochastic Gradient Descent with l1: **960.213674871**

**5.Stochastic gradient descent with l2**

Mean Root Mean Squared Error for Stochastic Gradient Descent in 10 Fold CV:3329.42861221

Root Mean Squared Error for Stochastic Gradient Descent with l2: **3396.45371651**

Mean Absolute Error for Stochastic Gradient Descent with l2: **961.259818611**

**6.Stochastic gradient descent**

Mean Root Mean Squared Error for Stochastic Gradient Descent in 10 Fold CV:3370.48983568

Root Mean Squared Error for Stochastic Gradient Descent: **3415.12019827**

Mean Absolute Error for Stochastic Gradient Descent: **963.188306088**

Generally, SGD should run faster than GD, but in my actual program running, SGD models are slower than GD models. I checked my code, everything is correct. I think the reason is that models of SGD and SGD with l2 were run in one day, models of GD and GD with l1 were run in another day. However, running time of model of GD with l2 was longer than model of SGD with l2, which ran in the same day. Especially file size is not very big, so the time difference could be covered by machine performance.





