## BGD301x – Big Data: Assignment 2

The idea of finding the best model is training the Random Forest Regressor with different parameter sets (numTrees, maxDepth, maxBins). Evaluate will based on the following item:

* **Root Mean Squared Error (RMSE)**: This is a measure of the differences between the values predicted by the model and the actual values. Lower values of RMSE indicate better fit.
* **Mean Squared Error (MSE)**: This is similar to RMSE but without taking the square root.
* **R-squared**: This is a statistical measure that represents the proportion of the variance for a dependent variable that’s explained by an independent variable or variables in a regression model. Higher values (closer to 1) are better.
* **Mean Absolute Error (MAE)**: This is the average of the absolute differences between the predicted and actual values. It measures the average magnitude of errors in a set of predictions, without considering their direction. Lower values are better.
* **Explained Variance**: This measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set. Higher values are better.

The table below shows the parameters sets and the corresponding result when testing on the validation data (30% data of train.csv):

|  |  |  |  |
| --- | --- | --- | --- |
| **numTrees** | **maxDepth** | **maxBins** | **Accuracy/Result** |
| 10 | 4 | 32 | * Root Mean Squared Error (RMSE) = 2296.5633006758185 * MSE = 5274202.99401101 * RMSE = 2296.5633006758185 * R-squared = 0.3780186003664311 * MAE = 1454.8678000676305 * Explained variance = 2455157.8613324366 |
| 3 | 4 | 32 | * Root Mean Squared Error (RMSE) = 2240.199896086758 * Validation data MSE = 5018495.574427121 * Validation data RMSE = 2240.199896086758 - Validation data R-squared = 0.38127144897412935 * Validation data MAE = 1478.7279755396378 * Validation data Explained variance = 2466765.4558192524 |
| 15 | 5 | 32 | * Root Mean Squared Error (RMSE) = 2236.8482425794077 * MSE = 5003490.060330586 * RMSE = 2236.8482425794077 * R-squared = 0.4053454757458218 * MAE = 1425.5202631763589 * Explained variance = 2736020.998644885 |
| 20 | 7 | 32 | * Root Mean Squared Error (RMSE) = 2203.432733950088 * MSE = 4855115.813042758 * RMSE = 2203.432733950088 * R-squared = 0.45726522185584695 * MAE = 1355.3167200906312 * Explained variance = 3091417.2843252826 |
| 40 | 9 | 32 | * Root Mean Squared Error (RMSE) = 2086.3156419133456 * MSE = 4352712.957692295 * RMSE = 2086.3156419133456 * R-squared = 0.5015171443124632 * MAE = 1318.9620863455677 * Explained variance = 3502721.033020182 |
| 60 | 10 | 48 | Training time is so long and out of memory error occurred |

**Explanation**:

* First, test with a random value: the *numTrees = 10, maxDepth = 4 and maxBins = 32* and log the result.
* Next reduce numTrees from 10 to 3 and run the test to compare with the first parameters set. Root Mean Squared Error reduce, this means this is better than the first set.
* The second parameter set is better, however the numTrees is the lowest value. So, try with *numTrees = 20, maxDepth = 7 and maxBins = 32*. The result is better than the second parameter set.
* Keep increasing the numTrees and maxDepth to see if the result will be improved.
* With *numTrees = 40, maxDepth = 9 and maxBins = 32,* the result seems good.
* With *numTrees = 60, maxDepth = 10 and maxBins = 48,* the training time is so long and an out of memory error occurred.

So, I chose *numTrees = 40, maxDepth = 9 and maxBins = 32* and ran on test data.

(For detail of parameters set results, please look at the *\*.ipynb* file attached along with source code).