Task 4 Group 6 README

Jack Cook

Govardhan Digumurthi

Thomas Okonkwo

Description of dataset

There are three input paths that are used in this task:

- the path to a random forest model from part 3 a)
- the path to a random forest model from part 3 b)
- the path to a testing dataset (20%) csv file

Work done

The following is a list of tasks done in this program:

- The spark configuration is setup
- Three file paths output from task 3 are read in from the command line
  - Path to random forest regression model a - read in as a random forest regression model
  - Path to random forest regression model b - read in as a random forest regression model
  - Testing dataset - read in as a pyspark sql dataframe
- Useful functions for accessing dataframes are defined
- Columns containing vectors for parts a and b are created (Note: this step is repeated in task 4 because storing vectors in an output file is no easy task, so the vectors are recreated here):
  - a) all features in the dataset
  - b) Only the feature with the highest correlation as determined in task 2 (specifically Bedroom)
- Both random forest regression models are used to make predictions on price
- Some of the predictions are shown from each dataframe
- The root mean squared error is computed for the all features and the bathroom model prediction, the rmse is shown in the console

Instructions to run the program

The random forest regression models are automatically stored in the hadoop file system when running `Task 3`. The testing dataset output in `Task 3` must be moved to the hadoop file system.

```
hdfs dfs -copyFromLocal Group6_Task_3_Output_Test.csv /user/cookjc
```

The program must now be submitted using `spark-submit`:

```
spark-submit Group6_Task_4_Code.py Group6_Task_3_Output_RF_A Group6_Task_3_Output_RF_B Group6_Task_3_Ou
```

Discussion of results

The root mean squared error values are horribly high for both predictions. The prediction using all of the features is lower than just using the bathroom.

The technique for finding the right combination of `tree depth` and `number of trees` is to provide very high numbers to get a very good fit, then reduce the size as small as is necessary to maintain a good fit. It did not matter how many trees I added for this model, the `rmse` would not go down.

The conclusion of these root mean squared error results is that random forest regression is the wrong machine learning network for this task.