Report ID5059 Practical 1 Student ID 200028066

I have described every step in the Jupyter Notebook I am submitting, hence I will outline the workflow in the report shortly. First I downloaded the data and then split it using GitBash console, to 86 pieces of 300 000 record lines each. Then I split the pieces in two folders, 'Train' with records up to 2016 and 'Test' with records 2017 and onward.

I am giving credit to StackOverflow website as I got some ideas for the sampling function from there, but the final working variant I used, took a lot of testing and improvement done by myself.

Loading and exploration of data: I created a function which takes a specified number of random samples from each piece and then adds all samples together in a pandas data-frame to work with, using the Glob library. I used this function to create workable data-frames from both Train and Test folders. Then I looked at the number of columns, naming them, the data types, etc

Preparing the data: I removed/dropped all unnecessary columns and left the ones specified in the task: house type, lease type and city. Then I used GetDummies method to transform the columns from categorical values to numerical ones, which also broke each feature to a number of columns equal to the number of variants it had. For instance if the house is detached it would have a 1 in the house_type_D column and 0 in all other house_type_ columns. For the City column, first I transformed it by naming all records that were in London as 1 and all others as 0, because that is the only thing we need to know as per the task. And then I used GetDummies to split it in two columns, as the others. Then I did some visualizations on the data and also looked at the linear correlations between my target (price) and the features.

Model attempts: I split the data into labels and features and trained as many models as I could, doing cross-validation and comparing the RMSE results. Then I tried fine-tuning the best models and again compared results.

Testing and conclusions: Finally, I used my sampling function to create a workable test data set from raw data pieces dated 2017 and onwards. I then used the best model so far (Decision Tree Regressor) to make predictions for the test data set. The resulting RMSE scores were very high and after many experiments and a lot of reading I could not improve that anymore. I believe the problem lies in the data itself and perhaps the way I prepared it before fitting the models, but I cannot know for certain.

Another hypothesis I have, which makes sense is that it might have something to do with the massive difference in the house prices between the recent years and the past years before 2017. The prices are much lower in the past, especially before 2000, hence a model trained on older pricing data cannot possibly make any good predictions when tested on more recent data where prices are much higher. Again, I cannot prove this as I do not know that much about machine learning yet, but I do hope I will be able to get to the bottom of this very soon.