Nikita Traynin, nxt170002, 2021413175

9/19/2021

Linear Regression on the Russian Housing Market

**Introduction**

Our goal is to predict a property's sale price based on a collection of other known variables. We take a dataset from Kaggle and perform exploratory data analysis, followed by model application (linear regression in this case), and finally evaluating the model and re-tuning it.

We are using the Sberbank Russian Housing Market dataset. It is pre-split into two sets for us - train and test. These use an approximately 80/20 train/test split pattern. In total, there are 38,133 observations. The train observations are before 7/1/2015, and the test observations are after. The train/test split is not random, since it is based on timestamp, so I will avoid re-mixing the data and performing cross-validation.

There are 292 columns in this data. One is a superfluous id variable, and another is our target variable "price\_doc". So, we have a large 290 predictors. For this reason, we will implement a gradient descent, iterative algorithm for our linear regression instead of the closed-form solution.

**EDA**