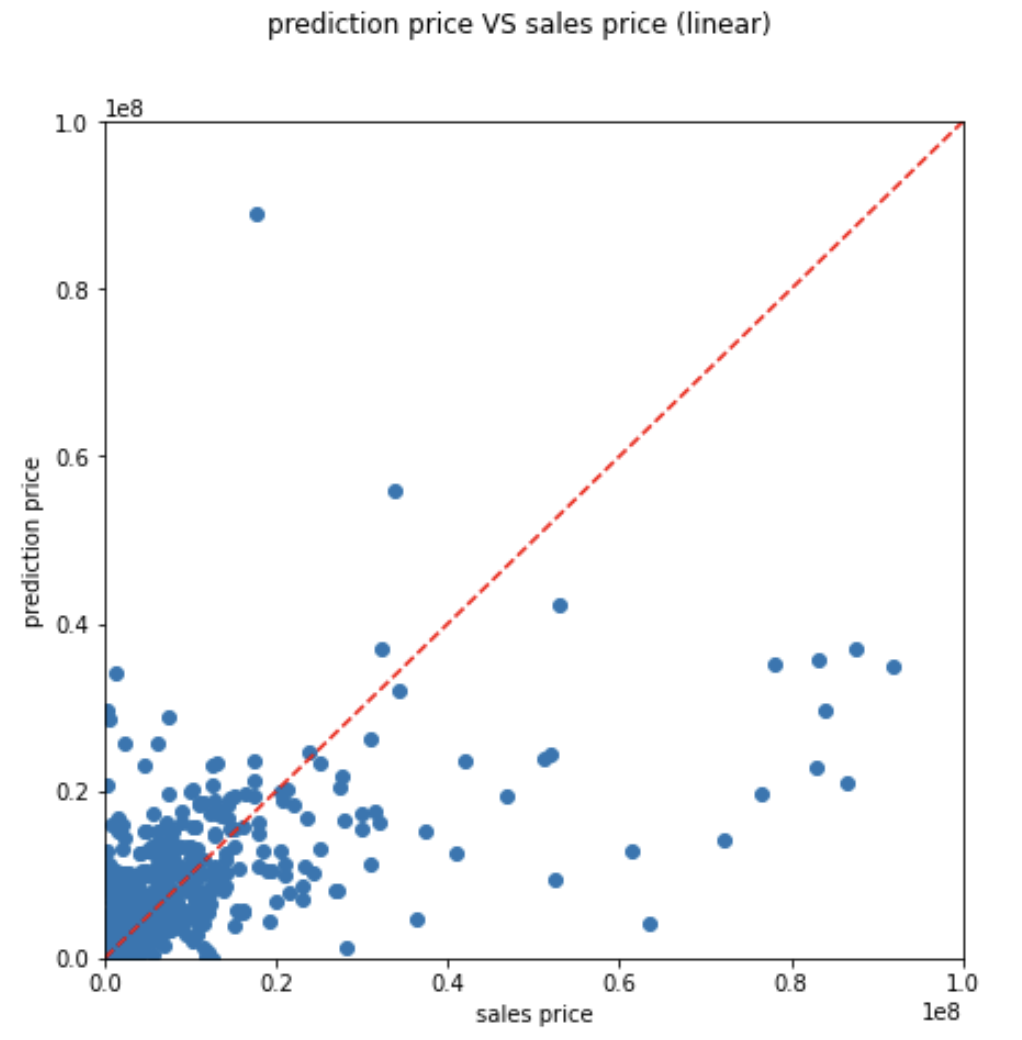
Model 1: Linear regression

This model is designed to predict property sales price based on the property information. it contains the columns, borough, neighborhood, building class category, residential units, commercial units, land square feet, gross square feet, building age as the predictors. We use both borough and neighborhood as the location information of property to make the location more detailed. As borough, neighborhood, building class category are all categorical columns and don’t have a specific order, we convert them into dummy variables. Considering the fact that unit property price varies from one location to another, we also multiply borough columns by land square feet and gross square feet columns and get new features borough\* gross square feet, borough\* land square feet so that the coefficient of gross square feet in linear model can be different for different borough.

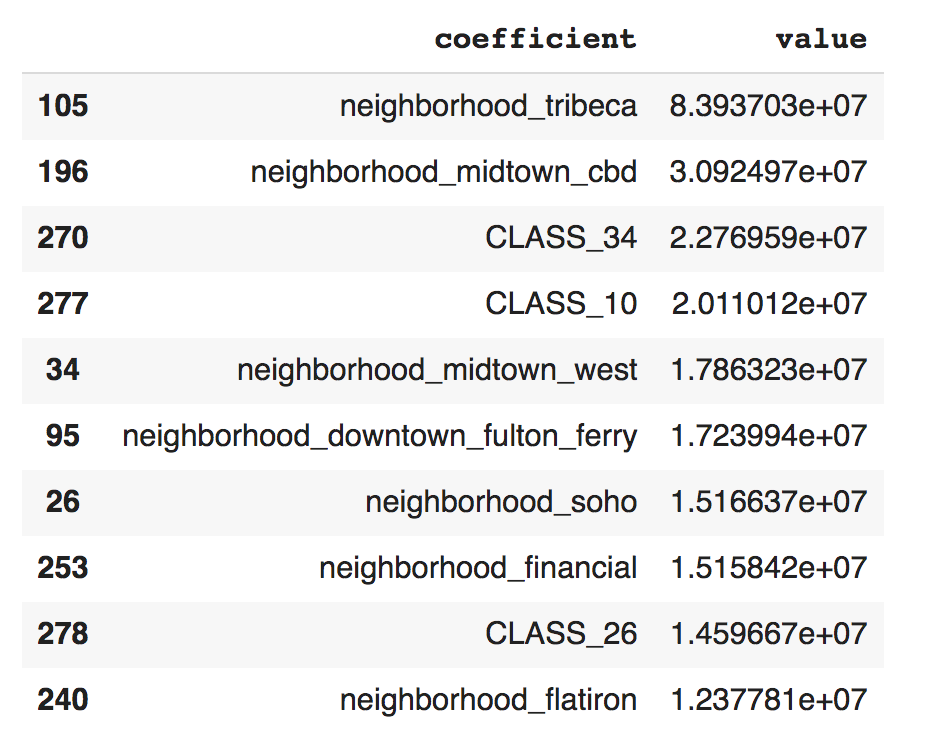
As we usually do in homework, we split the dataset into two parts, one part for training and the other for testing, and use MSE as the metric to evaluate the linear model.

The MSE score we get is 9.10489e+12, which means the mean estimation error on the price is more than one million dollars. The accuracy of the property price estimation is beyond our expectation. After trying to add more columns, the result doesn’t change a lot. We then check the original data in the CSV file and find that the sales price of property of larger square feet is sometimes lower than that of property of the same information other than square feet. We think some factors have more influence on sales price than the elements listed in the dataset like decoration. If one apartment is within the walking distance of some universities like NYU, it tends to have higher sales price compared to another apartment in the same borough far from NYU. The following graph shows the relation of our price prediction and the real price.



The closer the blue dots are to the red dotted line, the more accurate the linear model is.

Next, we use inference to determine the most important predictor order. We create a new pipeline which encapsulates a standard scalar and a linear regression object. After fitting the pipe, we create a pandas data frame with 2 columns named coefficient and value. The coefficient column contains the coefficient names and the value column contains the regression model coefficient absolute values.



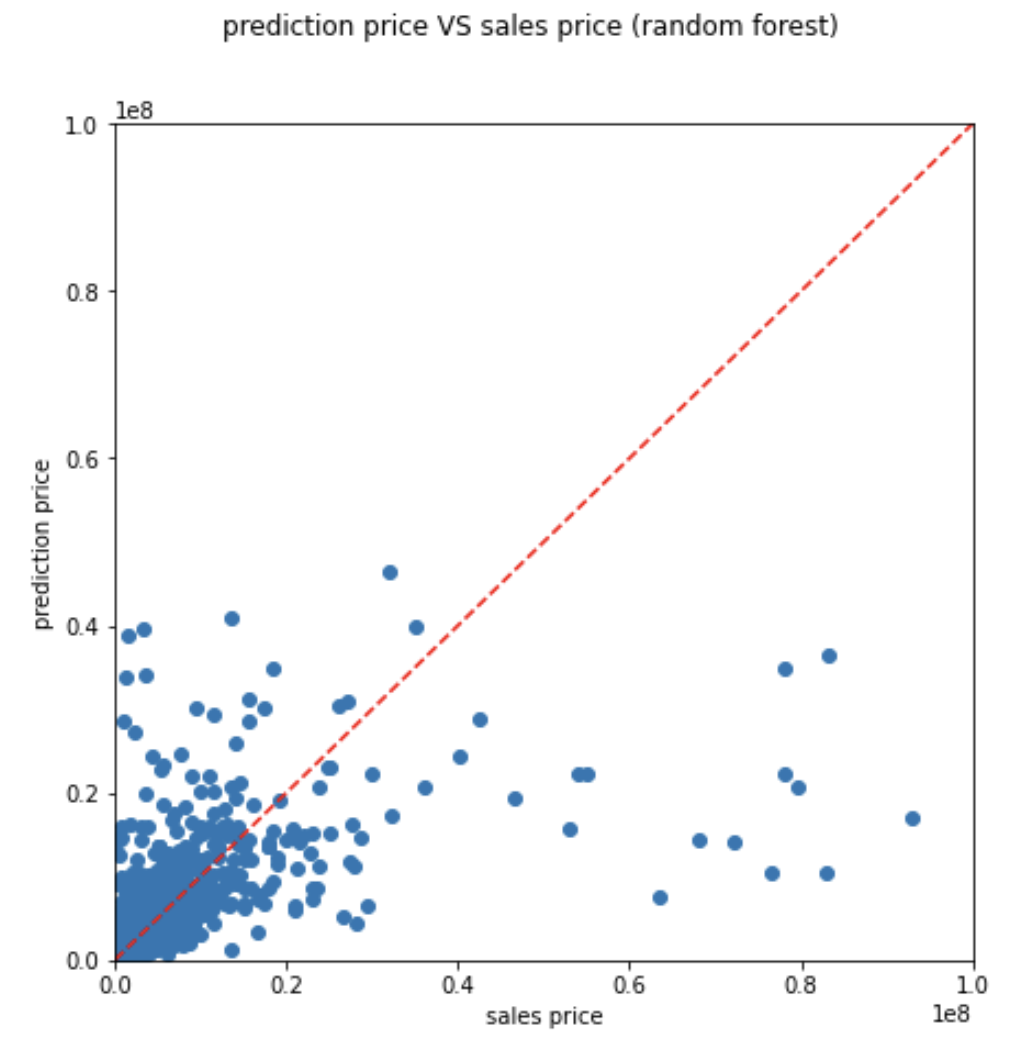
We can see the top ten predictors are all from neighborhood and building class category which is not consistent with our common sense.

Model 2: Random Forest

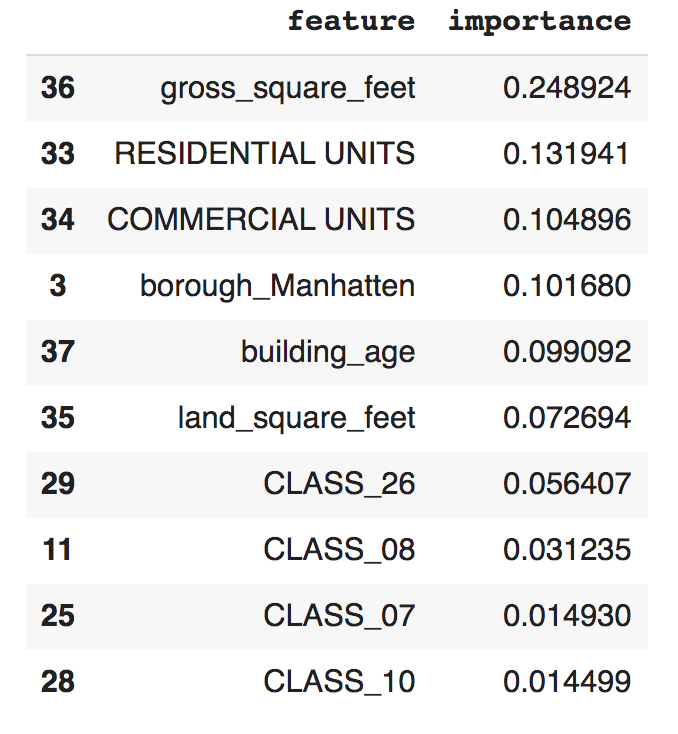
We use the same methods as linear model to transform categorical variables. To shorten the running time of the model, we only use borough as the location information of property.

The random forest model utilized a grid search to tune the regularization and elastic net parameters. The grid utilized num trees of 10, 30, 50 and max depth of 10 and 15. The trained model variations are evaluated by the mean square error (MSE) testing results after applying the model to the test dataset. The best model uses the num trees of 30 and max depth of 15.

The best random forest model was able to achieve a MSE score of 6.10829e+12 which is a little better than the result of linear regression model. The following graph shows the relation of our price prediction and the real price.



Below is the most important predictor in random forest model.



From the table, the model’s most important feature is the gross square feet of the property. The residential units of the property are the second feature people care about. This shows living area is people’s most important concern. It does make sense. Besides, the commercial units and the building age can only explain some buyers’ interest on properties. These two features are less preferred by the buyers.