# Housing Price Prediction

Pooja Choudhary (pc676), Ning Li (nl488) and Sreenu Chandran (sc2886)

December 4, 2017

**Abstract**

In our project, we try to predict house prices in Moscow Metropolitan Area in Russia, which consists of the city of Moscow and parts of the surrounding Moscow Oblast. Our dataset comprises of the sales price of 30472 houses in various sub-areas of Moscow Metropolitan based on 391 features like House Details, Neighborhood features and Macroeconomic conditions. During pre-processing, we found that there were 92 features which had missing values. So, we applied Matrix Completion by iterative low-rank SVD Decomposition to impute the missing values in the data set. We used the training data set to formulate five different prediction models -Simple Linear Regression, Lasso Regression, Huber Regression, Principle Component Analysis and Random Forest. We then tested these models on the test data set and find that Random Forests and Lasso give the lowest error on the test set data.

## Introduction

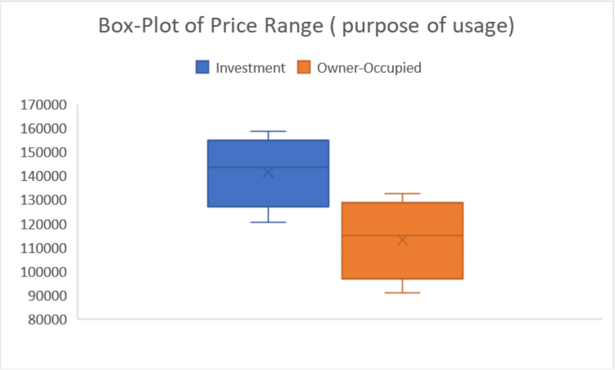
Moscow Metropolitan Area is the largest metropolitan area in Russia with a population of around 16,170,000 and total area of 46,811 square kilometers. Its economy is primarily dependent on its various industries spread across fields like space, metallurgy, chemical, and energy production. Moscow also houses the Kremlin and various other government offices. It also has various green zones like parks and forests which cover about 40% of the region.
There is a huge demand for houses in the region and the region has been witnessing constant rise in new houses and apartment complexes.

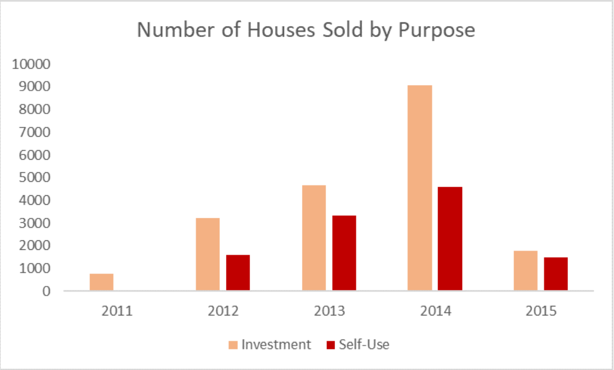
Our dataset comprises of the sales price of 30472 houses in various sub-areas of Moscow Metropolitan based on 391 features which can be broadly classified into three important categories - 1)House Details like built year, material used for construction, number of rooms, area of the rooms and kitchen, number of floors etc. 2)Neighborhood features like the proximity of the house to schools, universities, transportation system, health-care facilities and various sports, recreational, cultural and religious landmarks 3)Macroeconomic conditions prevalent at the time of sale of the house like GDP of the country, CPI , mortgage rate, deposit rate, etc.

## Data Pre-processing

Data exploration is one of the most important steps of data pre-processing. It helps in forming the preliminary idea about model selection.

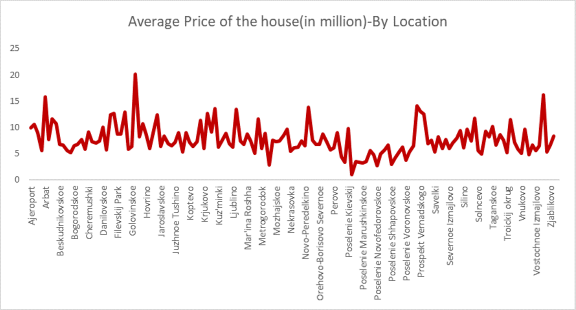
Our data-set comprised of both owner occupied houses and houses purchased for investment purpose. We first check if there are any inherent price differences in these two types using box plots. We also plotted number of houses that were sold in these types to identify any pattern.

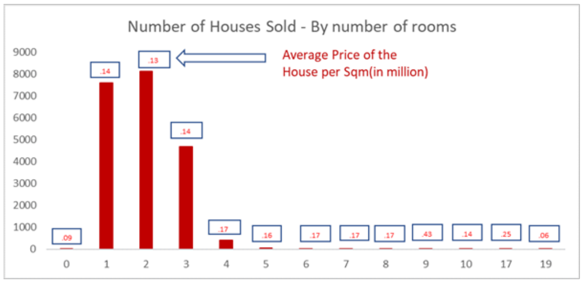

Box-Plot of Price Range ( purpose of usage)


Number of Houses Sold by Purpose

As can be seen from the graphs above, more houses were sold for investment purpose than for self-use. Further, from the box plots it is clear that the price range for investment purpose houses are higher than for self-use. So, we will incorporate this information while making the future prediction.

The data comprises of houses from 128 sub-areas. To check if the price of the houses are varying by areas , we plotted average price of house by area. We also plotted number of houses sold by number of rooms. More than 95% of the houses had 1-3 rooms. We also checked the average price of the house per square meter for different number of rooms and found that it was almost consistent for different number of rooms. Below are the plots for the same:


Average Price of the house(in million)-By Location


Number of Houses Sold - By number of rooms

Since there is not much variation in the average prices we conclude that we do not need to fit separate models for different sub-areas. A single model is sufficient.

On analysing the data, we observed that our data was messy in the following ways:

1. Missing values: Out of 391 features, 92 of them had some missing values. Instead of deleting data and possibly losing valuable information, we decided to use a Matrix completion method by iterative low-rank SVD decomposition to impute the missing values. When we ran a simple linear regression on the data sets with imputed values and with deleted values, we got lower RMSE for the former set. Therefore we decided to go ahead with the imputed data set for all further analysis.

2. Mixed data type: Approximately one fourth of the features are categorical. For example, material of the building or house has type 1, 2, 3. We also have text variables (eg.locality names) as well as numeric variables (continuous variables like GDP growth rate, discrete variables like number of rooms).

3. Noisy data: There are some mis-entered or inaccurate data. For example, variable $\text{build}_{year}$ is

the year the apartment/house was built. But in the values, we found 4965, which we suspected that it should be 1965. There were other data-point having values like 20052009, where it was obvious that this value was double entered. Also, we found the floor of the building in which the apartment was located was more than the maximum floors of the building. In all these cases, we imputed these values using matrix completion approach.

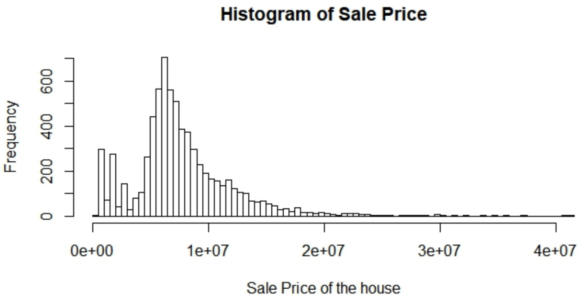
### Feature Engineering

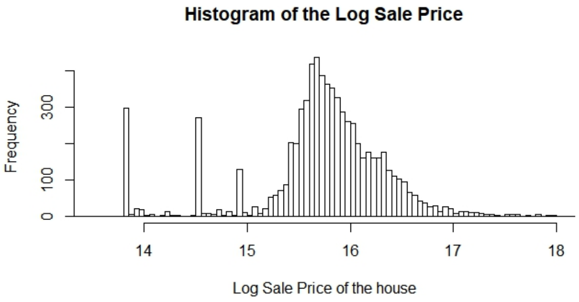
1. One-hot Encoding Transformation: We used one-hot encoding transformation to convert all the categorical features into binary variables.

2. Generating New Variable: We also generated new variables which we think may catch the variance in sales price better. For example, Instead of using build year and sale year separately, we created a new variable, age, where age=sale_year-build_year.

3. Delete predictors that have zero variance:zero-variance predictors lead to biased estimation and may cause all kinds of problems. We found out those predictors that have near zero variance and deleted them (140 features are deleted).

3. Log-Transformation: The sale price was heavily skewed to the right with skewness = 3.47. So, to normalise the sale price, we used the log transformation of sale price to make better prediction. This made the distribution close to symmetric with a skewness of -0.9. Histogram of both the sale price and log of sale price is shown below:





### Feature Selection

Feature selection is crucial in predictive modeling for several reasons: it simplifies models so that interpretation becomes easier; generalization will be enhanced if only a subset of the features are used in the model and the problem of over-fitting will be reduced; practically, it enables less computation operations and shortens training time.
So, we incorporated various feature selection into our predictive modeling using different regulariser like Lasso, Principle Component Analysis etc.

### Predictive Modeling

In this section we use different models (different loss functions and regularizers) to fit the data and we will show a summary of all these models, including visulizations in the conclusion part.

Simple Linear Regression (SLR)

This is one of the simplest linear models for continuous variables. Here we predict the value of dependent variable by finding the best fitting hyperplane between the dependent variable and the covariates(independent variable). This is done by minimising the sum of squares of the residuals. i.e., find

$$h(x) = W^T X \text{ by minimising } \sum(y_i - w^T x_i)$$

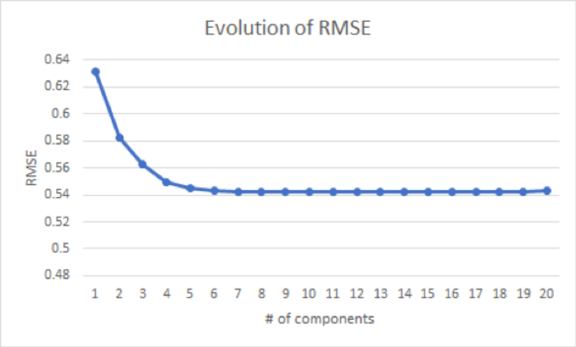
We used the training set to fit a linear model and got RMSE of 0.65. We used this result to make the prediction on the test set and got a Root Mean Squared Error(RMSE) of 0.76 on the test set.

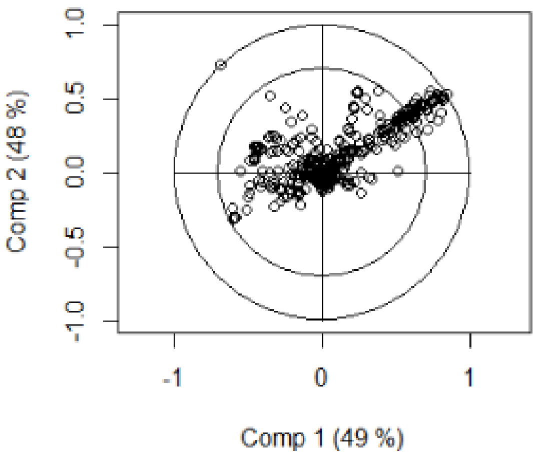
Random Forest (RF)

Random Forest model using many decision trees and give mean prediction by aggregating the prediction from each tree. At every node, it randomly selects some features find the best split for the node. Since we are not sure about whether the house price we want to predict has a linear structure in all these predictors, plus we have a decent amount of dummy variables after fearture engineering, there is a good reason to try random forest regression for the prediction.
We used random forest regressor under sklearn package in python to predict the sales price using all the predictors. We got a $R^2$ of 0.68 for the test set and the RMSE value for the test set data was 0.535.

Partial Least Squares (PLS)

Like Principle Component Analysis (PCA), PLS utilizes linear combinations of predictors (Stone & Brooks, 1990). However, instead of maximally summarizing predictor space variability, the PLS chooses linear combinations to maximally summarize covariance with the dependent variable. Because of this, PLS can be seen as a supervised counterpart of PCA. But it is especially designed for data sets where predictors are highly correlated, which is suitable for our data. The below figure (left) shows how RMSE changes with number of components, and we can find that using 8 components generates the lowest RMSE of 0.54. The right figure shows the correlations of predictors to components. Many predictors crowded in the center of the circle, indicating that many predictors have very small weights (nearly zero) in constructing components 1 and 2.



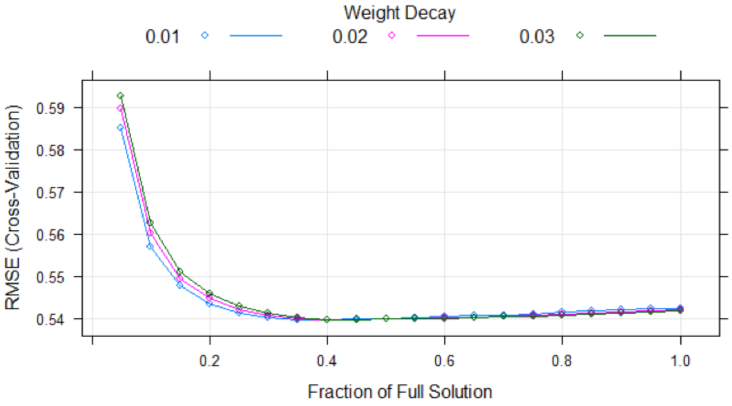


Lasso Regression

Here we use a generalized lasso model elastic net (Zou & Hastie, 2005).

$$SSE_{enet} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda_1 \sum_{j=1}^{d} |w_j|^2 + \lambda_2 \sum_{j=1}^{d} |w_j|$$

This model is advantageous by incorporating the ridge-type penalty and the feature selection function of lasso penalty. It was suggested that this model is more efficient with highly correlated predictors in presence. When lambda is 0, the model boils down to a pure Lasso model. We tuned the model with three different values 0, 0.01, 0.02 for $\lambda_1$, which we call lambda, and 20 values for $\lambda_2$ (0.05 to 1), which we call fraction. The following figure shows the evolution of RMSE (Cross-Validation) with different combinations of lambda and fraction. We can see that RMSE drops in the beginning and then after a certain value of fraction, it slightly goes up.The lowest RMSE of y, 0.54, happens at lambda=0.02, and fraction=0.40, which generates R$^2$ value of 0.32.



Huber Approach

From above analysis, we learnt that there are some outliers in the data set and we already cleaned out some. Since our data set is huge and messy, maybe Huber approach is a good one to try to avoid the impacts of outliers on coefficients by using a quadratic loss function.

$$SSE_{Huber} = \sum_{i=1}^{n} \mathbf{huber}(y_i - \hat{y_i}) + r(w)$$

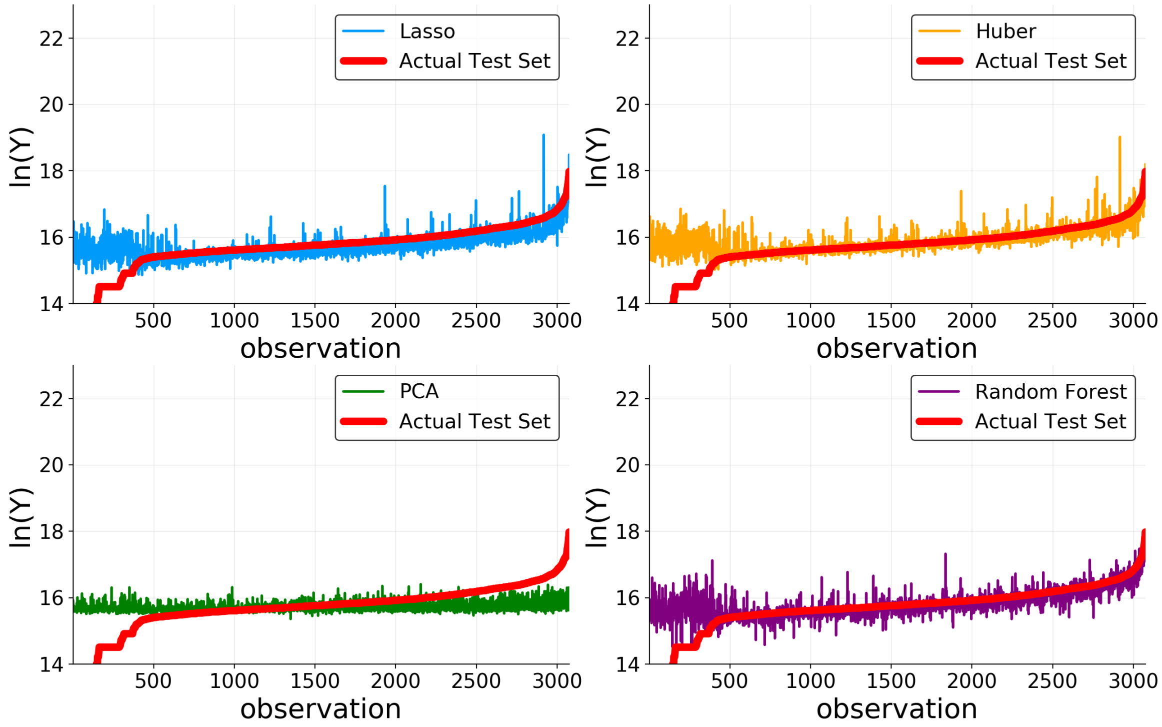
$$\mathbf{huber}(z) = \begin{cases} \frac{1}{2}z^2 & |z| \leq 1 \\ |z| - \frac{1}{2} & |z| \geq 1 \end{cases}$$

We got a RMSE of 0.565 for the training set and the RMSE value for the test set data was 0.579.

## Conclusion

We used five different models including linear and non-linear models- to predict the price of houses using multiple covariates. We divided the entire data-set randomly into training and test set in 70:30 ratio and formulated all the models on the training set using 10-fold cross validation. We then, tested our models on the test set which was not exposed for training the models. The results of different models are as shown below:

| **Models** | **SLR** | **RF** | **Lasso** | **Huber** | **PLS** |
|---|---|---|---|---|---|
| RMSE (CV on train set) | 0.650 | 0.524 | 0.540 | 0.565 | 0.542 |
| RMSE (on Test Set) | 0.760 | 0.535 | 0.553 | 0.579 | 0.643 |



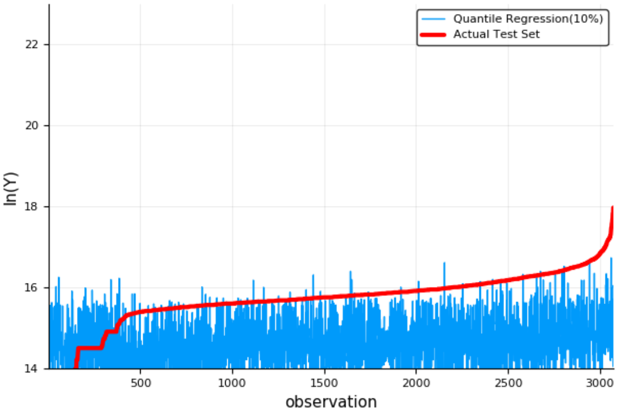
As evident from the above table, all the models are generalising well except for simple linear regression and PLS which is giving lower root mean squared error for the training set but higher error when it was used to fit the test set. It means that they are over-fitting the model and shouldn't be used to future prediction. Among all other models, Lasso and Random forest is giving almost similar training set and test set error and can be used for future prediction with confidence.

One of the heaviest item on the balance sheet of the real-estate companies is the money spent in evaluating the price of houses. Having a predicted price using the models can save a lot of money for the real-estate consultant give them competitive edge. The models built by us is based on a comprehensive data-set of more than 30,000 observation and 390 features, covering all aspects of property valuation. Thus, we are confident that the Lasso and Random Forest model will help the enterprise in making correct prediction in future.

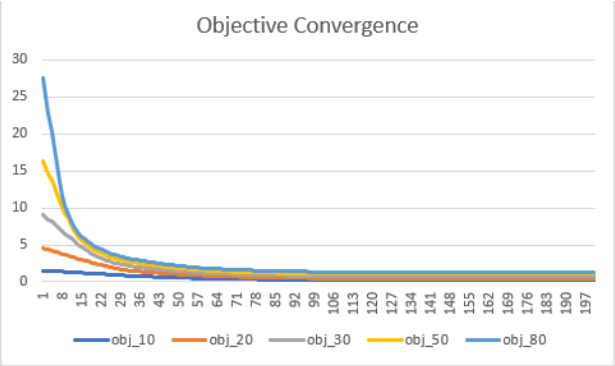
## Discussions

### Quantile Regression

As discussed above, although our predictive models did a good job in predicting most of the data points in the test set, they all have a weak performance on those 500 observations with the lowest price levels (the data was sorted before we drew the prediction graphs). The reasons could be, these houses are underpriced, or they belong to another population, etc. What is sure is that using mean or medians of our models to estimate those houses is not a good idea. Maybe we can try a lower quantile regression to fit those data points. The following graph shows the result of a 10% quantile regression.



Out of curiosity, we also did quantile regressions using other quantiles (10%, 20%,...90%.) (the following graph illustrates how the objectives converge) and calculated the standrdized coefficients. Their coefficients display different values for different quantiles. The most important ones in common include interest rate, mortgage value, footage of the house, squre of offices within a 5 km range, etc. An interesting finding is that mortgage value plays a way more important role in 90% quantile regression than that in middle quantiles. Interest rate exhibits a similar property. In the fututre, we will have a closer look at those 500 observations to see what characteristics they bear and come up with a better predictive model for them.



## References

Stone, M., & Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 237–269.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.