ORIE 4741: Mid-term Project Report

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

Data Description

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. It 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 lists the sales price of 30472 houses in various sub-areas of Moscow Metropolitan based on 390 features which can be broadly classified into three 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) Neighbourhood 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

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

1. Missing values: Many features had missing values and corrupted data. We removed some of these features having more than 50% missing values. We were able to delete those features because they were found to be highly correlated with other features in the data set. Also, we removed observations which had missing values for more than 50 features.

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 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. For a value like 20052009, it is obvious that this value was double entered so we picked randomly from 2005 and 2009. Also we found the floor of the building in which the apartment was located was more than the maximum floors of the building. In these cases we updated the max floors of the building to be equal to the floor of the apartment.

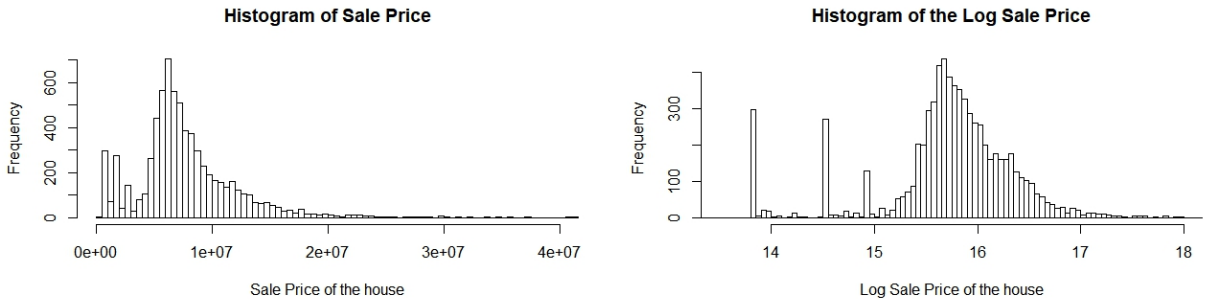
After data cleaning, we had a total of 10250 observations with 374 features on which we performed feature engineering.

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. 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 and Modeling

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 overfitting will be reduced; practically, it enables less computation operations and shortens training time.

In our project, we used two feature selection methods–Sequential Forward Selection (SFS) and Least Absolute Shrinkage and Selection Operator (LASSO) and we will compare their predicting power.

We randomly split our data set into two train set and test set by 7:3. The train set contains 7175 observations and the test set has 3075 observations.

Sequential Forward Selection (SFS)

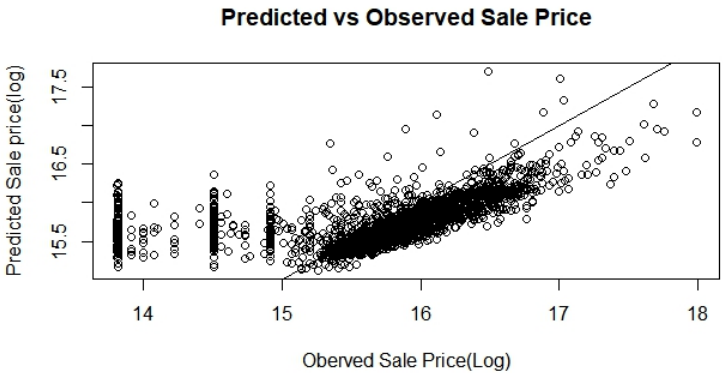
We chose to use Sequential Forward Selection (SFS), also called Stepwise Forward Selection, because it is the simplest greedy search algorithm and proved to performs well when the optimal subset is small. Basically, this algorithm starts from an empty set, predictors are added one at a time beginning with the predictor with the highest correlation with the dependent variable, until Akaike information criterion (AIC) is no longer improved.

Our final model based on SFS has 362 predictors including dummies of the original features. We conducted 10-fold Cross Validation on the train set and got the Root Mean Squared Error (RMSE) as **0.5685**.

Lasso Regression (Least Absolute Shrinkage and Selection Operator)

The idea here is to calculate the least square error by penalizing the coefficients of the features.The mathematical expression for the model is as given below, where y_i is the observed value or the predictor variable, w is coefficient vector , x is the covariate matrix and λ is penaliser. minimize $\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{n} |w_i|$

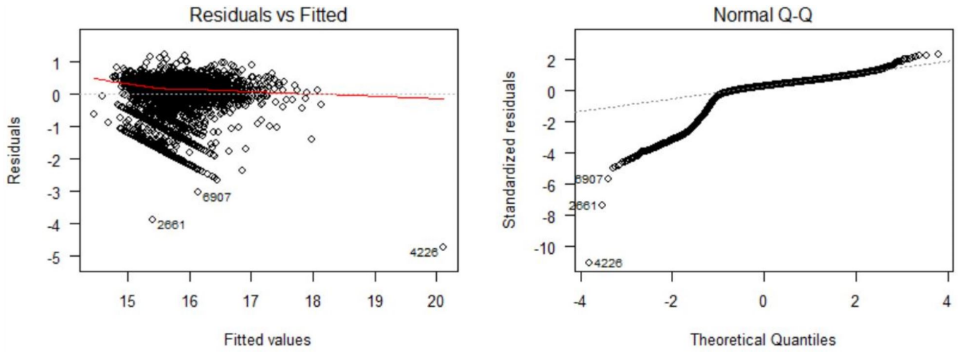
We used 10-fold cross-validation to find the lambda which minimised the entire function. λ min was calculated to be 0.00527. As per the Lasso model , 110 features were considered significant to make the prediction. The plot of observed value of sale price vs predicted value of sale price is shown in the figure below:



The RMSE calculated using lasso model on the test data was **0.1744**, much smaller than that of SFS. So we choose to use the LASSO model.

Future Work:

Diagnostic Figures



We performed diagnostic analysis based on the lasso model and the results are shown in the above figure. We can see there are three straight line segments in the Fitted value vs Residuals figure, which is not desirable and seems to be abnormal. However, if we compare it to the histogram of log(sales price) and the Predicted vs Observed Sale Price figure, we found some commonality of abnormality. The three vertical lines in the Predicted vs Observed Sale Price figure and the three downward sloping line segments correspond to the three lonely standing bins in the histogram of log(sales price) below 15. The departure from normal distribution in the Q-Q plot may also be caused by those data points. Further exploration is needed to tackle these data. Our hypothesis is that these data comes from another population and this is why they do not fit well. Another finding from the diagnostics figure is that there are some outliers which are worthy of further discussion.

Also for the future work we plan to create separate models for different size of houses and different sub-areas and explore how the predictions vary. Alternatively, we will also look for smooth coefficients model to check if the prediction improves. We will also test the sensitivity of the model on the test set data by bootstrapping samples from the test set and testing the performance. We will also look at few nonlinear models like random forest and neural network to check how the prediction changes.