Russian Real Estate: Housing Price Prediction - A Case Study

Christina Karpagam
Student / VIT Chennai
19MIA1015

Kiran Benny
Student / VIT Chennai
19MIA1091

Aisha Samah Yahya
Student / VIT Chennai
19MIA1045

Abstract

The real estate sector takes up a significant portion of all the wealth on the planet. With real estate holdings, comes the feeling of power that accompanies that type of wealth. Various factors affect house pricing, such as type of house, type of building, number of rooms available. Predictive analytics will help customers to assess how profitable the purchase of a house or apartment will be for them, and also to predict the future price of houses and apartments in specific regions. This paper enumerates the various mathematical price predictive models for Russian real estate, and the process they follow in order to come up with an effective price forecasting model. This generalised predictive model can also be applied in the Indian context, on availability of relevant data.

1 Introduction

Predictive analytics is a technique derived out of machine learning and AI, that relies on various statistical models and regression analysis to study historical data points and understand trends, derive insights about them, and also predict future trends. Real estate is property consisting of land, and the buildings on the land, along with the natural resources of the specific area. According to Selim [*Selim, S. (2008). Determinants of house prices in Turkey: A hedonic regression model. Doğuş Üniversitesi Dergisi, 9(1)*], housing services are one of the most expensive household expenditures and changing housing prices have been of concern to both individuals and governments in that they influence socio-economic conditions and have a further impact on national economic conditions. Pricing for immovable objects is closely connected with the current state of the rate of national currency; in the economy of any state, the real estate is a basic unit, and pricing for it is secretly equated to the exchange auction [*Vladimirovna, V. E. (2016). Review of the Russian market of real estate. Science Time, (3 (27))*].

Another interesting fact is that, according to the study by Antipov, among the six floors that were analyzed in a building, i.e., 6th, 7th, 8th, 12th, 13th, and 14th, the sample proportion of sold apartments was the lowest for floor 13 (54.9%) and the highest for floor 7 (68.6%). [*Antipov, E. A., Pokryshevskaya, E. B. (2015). Are buyers of apartments superstitious? Evidence from the Russian real estate market. Judgment and Decision Making, 10(6)*]. The estimation of the worth of particular real estate entities is required to provide a quantitative measure of the benefit and liabilities, profits and losses, etc., occurring from the ownership and or buying or selling of the real estate. The economic and political situation in Russia over the past few years has been quite a hustle. With economic changes like the official ban on trade after their intervention in Crimea and Ukraine, the crash of the Russian currency, high inflation rates, and high interest rates many things have changed. These factors have supposedly put the country into a recession. As is bound to happen, this has affected housing prices across the country, including property prices in Moscow, their capital.

When dealing with such scenarios, a data-driven approach will prove to be more effective in dealing with forecasting as compared to conventional processes, and consistency is maintained over time. Insights that are derived from data, replace opinions with facts. A 2017 study by Imprev found that around 65% of top real estate executives are more likely to invest in predictive analytics, marketing automation, and Big Data by 2022. We are already here in 2022, and the growth of machine learning and AI techniques like predictive analytics has risen vastly.

Here, using the dataset, “Russia Real Estate 2018-2021”, we will be training the predictive models to

predict the price of housing in Russia and see the effect that different factors have on the price. The various models will be trained, and such a model will be found, which provides the highest accuracy while testing. This will be a generalised predictive model, which can be used in the future for a specific application in the Indian context.

2 Russian Real Estate

The Russian real estate market is still relatively new. All properties were owned by the state during the Soviet era. Citizens only had the right to utilize them if properties were assigned to them based on their place of employment. As a result, relocation possibilities were limited. However, with the fall of the Soviet Union, the Russian real estate market arose, allowing the natives of Moscow to privatize and then sell and buy houses. Until recently, no urban land had been privatized. The only exception was a small plot of land that had been allotted to families to live in, or for garages, or had been privatized by families who already lived in single-family residences.

Russian real estate is now booming. It provides numerous fascinating opportunities as well as significant returns on investment. Over the last two decades, the Russian real estate market (especially in the capital, Moscow) has been one of Europe's fastest expanding real estate segments. It was severely influenced by a variety of economic, political, demographic, and other relevant factors over the last decade. However, it has begun to recover and is now offering opportunities to both domestic and foreign investors who are considering adding a direct Russian real estate holding to their investment portfolio.

Moscow remains as Russia's most popular real estate investment destination. Moscow has the widest range of properties in terms of type and value, as well as the highest number of transactions. Therefore, it is wise to consider the Russian city and the options it offers for buyers.

3 Dataset

The dataset used for the development of the Russian real estate price predictive model was taken from kaggle, a subsidiary of Google LLC, which is an online community of data scientists and machine learning practitioners. It is an open, public database with GNU Affero General Public License made available on May 14, 2021.

The dataset has 13 fields. According to the publisher, the fields include the date of publication of the announcement and the time when the ad was published. The data also includes the geographical latitude and longitude of the apartment and its region in Russia. The building type, the apartment type, the apartment floor, number of storeys of the building, the number of living rooms, the total area of the apartment and the area of kitchen is also available in the dataset. The target variable price is given in rubles.

price	date	time	geo_lat	geo_lon	region	building_type	level	levels	rooms	area	kitchen_area	object_type
6050000	19-02-2018	20:00:21	59.8058084	30.376141	2661	1	8	10	3	82.6	10.8	1
8650000	27-02-2018	12:04:54	55.683807	37.297405	81	3	5	24	2	69.1	12	1
4000000	28-02-2018	15:44:00	56.29525	44.061637	2871	1	5	9	3	66	10	1
1850000	01-03-2018	11:24:52	44.996132	39.074783	2843	4	12	16	2	38	5	11
5450000	01-03-2018	17:42:43	55.918767	37.984642	81	3	13	14	2	60	10	1
3300000	02-03-2018	21:18:42	55.908253	37.726448	81	1	4	5	1	32	6	1
4704280	04-03-2018	12:35:25	55.6210965	37.4310016	3	2	1	25	1	31.7	6	11
3600000	04-03-2018	20:52:38	59.8755262	30.3954571	2661	1	2	5	1	31.1	6	1
3390000	05-03-2018	07:07:05	53.1950306	50.1069518	3106	2	4	24	2	64	13	11
2800000	06-03-2018	09:57:10	55.7369718	38.8464565	81	1	9	10	2	55	8	1
6909880	06-03-2018	18:34:48	55.9139498	37.7077118	81	1	9	14	3	76.1	8.8	11
4291950	06-03-2018	18:37:27	55.9139498	37.7077118	81	1	10	14	1	40.3	11	11
6675840	06-03-2018	18:37:28	55.9139498	37.7077118	81	1	25	25	3	73.2	12.4	11
6522650	06-03-2018	18:37:35	55.9139498	37.7077118	81	1	5	14	3	68.3	12.1	11
6522650	06-03-2018	18:37:40	55.9139498	37.7077118	81	1	7	14	3	68.3	12.1	11
4279770	06-03-2018	18:40:08	55.7817155	37.8566559	81	2	7	15	1	36.3	16.6	11
4550000	12-03-2018	12:37:08	55.738846	49.225437	2922	3	6	10	2	54.2	11.4	1
2880000	15-03-2018	14:38:45	55.7349712	52.3663848	2922	1	8	10	2	51	8	1
1450000	16-03-2018	14:51:58	45.069785	41.935019	2900	1	9	10	1	43	9	1
1650000	16-03-2018	16:21:54	44.9943012	41.1228103	2843	3	5	5	2	51	7	1
8000000	17-03-2018	06:46:32	55.738876	37.82537	3	1	5	9	2	45	6	1
2250000	17-03-2018	09:05:06	54.8155661	56.12556	2722	1	2	5	2	46	6	1
3843000	20-03-2018	14:44:11	56.3460273	43.8716477	2871	2	16	25	2	61	11	11
2697200	20-03-2018	14:44:11	56.3460273	43.8716477	2871	2	6	25	1	44	20	11
4214700	20-03-2018	14:44:11	56.3460273	43.8716477	2871	2	16	25	2	67	20	11
3773700	20-03-2018	14:44:11	56.3460273	43.8716477	2871	2	12	25	2	60	12	11
950000	25-03-2018	14:55:54	58.6930479	59.485351	6171	3	2	5	1	32	10	1
5300000	30-03-2018	12:08:09	54.88528	38.080196	81	3	2	5	3	75	9	1
4600000	02-04-2018	15:09:38	55.7078344	37.9596388	81	2	17	17	1	37	12.5	1

Figure 1: Dataset

The dataset comprises a variety of features and data types, encompassing a diverse range of information. The features incorporated within the dataset are rich and varied, encompassing a broad array of elements. These encompass a plethora of data types, including but not limited to numeric, categorical, textual, and temporal data. The dataset encompasses an assortment of attributes that capture distinct aspects of the data under consideration. These characteristics encapsulate a wide gamut of information, providing a comprehensive and multifaceted perspective of the dataset.

The features and data types in the dataset includes the following:

1. Categorical features:
    - Region
    - Building type
    - Object type
2. Numerical features:
    - Area
    - Kitchen area
    - Rooms
    - Level
    - Levels

3. Geospatial features:
    - Latitude
    - Longitude
4. Temporal features:
    - Date
    - Time

4 Methodology

The predictive data model was developed using the following steps:

4.1 Data Acquisition

The dataset; "Russia Real Estate 2018-2021" was taken from kaggle. It is an open, public database with GNU Affero General Public License which was uploaded in the context of a competition.

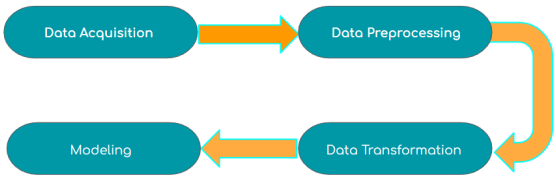

Figure 2: Working Modules

4.2 Data Preprocessing

The dataset consists of a total of 13 features; Date, Time, Geographical Latitude, Geographical Longitude, Region, Building Type, Object Type, Level, Levels, Rooms, Area, Kitchen Area and Price. The dataset is based on data from Russia during the years 2018 -2021. The price is in rubles. There are 85 subjects in the region of Russia. The facade type stretches for about 5 different styles including monolithic, brick and even wooden. The apartment type is either a secondary real estate market or a new building.
The stage involves filtering out the noisy data elements. Missing values were found to be null. Non-positive price values were identified and removed. Price values are divided by 100k for ease of data handling. Interquartile range (IQR) a concept from descriptive statistics is utilized to carry out outlier analysis.

4.2.1 Outlier Analysis

On application of IQR the lower limits were found to be empty as the necessary pre-processing was carried out. Beyond the upper limit there are around 4 lakh entries which is a huge number. To attend to the issue the average of the upper limit outliers was calculated and price values above the average were removed.

4.2.2 Anomaly Detection

Anomaly detection is the process of identifying patterns in data that are considered to be different from the normal or expected patterns. Anomalies, also known as outliers, can be caused by a variety of factors, including errors in data collection, sensor malfunctions, or unusual behavior. We have considered "Rooms", "Kitchen_Area", and "Area" for anomaly detection

Apartment Size: According to modern building codes and regulations, specifically SNiP - Stroitelnye Normy i Pravila (Russian Construction Codes and Regulations), the area of a room in a one-room apartment is required to be at least 14 sq.m., while kitchens should be a minimum of 5 sq.m. It is important to note that SNiP does not dictate the minimum area of an entire apartment, as this can vary based on the construction technology and dimensions of the initial structures such as bricks or reinforced concrete panels.
However, in recent years, the construction of small-sized apartments has become increasingly popular in Russia, resulting in a decrease in apartment size to 11.1 sq.m. for new buildings in Moscow in 2019. As of 2022, the minimum area for such micro-apartments is 10 sq.m. These micro-apartments are formally considered as apartments and belong to non-residential premises, allowing them to be built despite not meeting the minimum area requirement of 14 sq.m. for a room according to SNiP.

One-Room Apartment: Based on the presented database containing ads from 2018 until 2021, we can hypothesize that the minimum size for a one-room apartment is 10 sq.m. Given this assumption, any records indicating an area less than 10 sq.m. are considered irrelevant and will be deleted from the database.

Area Values: Upon analyzing the dataset, it has become evident that there are certain anomalies in

the area values of some apartments. In particular, for a few apartments with 1-, 2-, and 4-rooms, the areas are reported to be as high as 2, 3, 4, 5, and even 6 thousand square meters, which seems highly unlikely and suggests the presence of erroneous data.
On the other end of the spectrum, there are also instances of anomalies in the form of apartments with areas less than 10 sq.m. despite having multiple rooms, ranging from studios to 5-room apartments. This raises concerns about the accuracy of the data and the plausibility of such living spaces.Moreover, it is also worth noting that apartments with 6 to 9 rooms and an area of less than 50 sq.m. may also be anomalous and should be examined further.

Kitchen Area: According to the "softest" construction rules, the kitchen area should not be less than 5 sq.m. However, upon examining the database, it has come to light that there are instances of suspiciously large kitchen areas, even exceeding the total area of the apartment itself. Such anomalies are contradictory to common sense and suggest that the data may not be accurate. It is important to note that these anomalies are present in both the resale and new building markets. While micro-apartments, belonging to the "0" category, may have a kitchen area of just a couple of square meters, larger apartments with kitchen areas exceeding the total area of the apartment seem highly unlikely and warrant further investigation.

Identifying and addressing these anomalies is crucial in ensuring the reliability of the data for informed decision-making in the real estate market.

4.2.3 Data Normalization

Data normalization using StandardScalar from the sklearn library is applied to neutralize the heterogeneous sources. The features are standardized by removing the mean and scaling to unit variance. Temporal alignment and data formatting are also a part of this stage which is applied according to the requirement of the dataset.

4.3 Data Transformation

This stage is often called feature extraction and selection as the data is arranged, represented and the specific features are selected according to the requirements for the study. Data is altered to improve its organization. Humans and computers may find it easier to use transformed data. Data that has been properly prepared and validated enhances data quality and protects programmes against possible landmines like null values, duplication, erroneous indexing, and incompatible formats. Data transformation makes it easier for applications, systems, and data kinds to work together.
Data that is utilized for several purposes may require different transformations. Data type conversion and flattening of hierarchical data operations shape data to increase compatibility with the model. Correlations between the features were calculated and visualized. Date, Time, Geographical Latitude and Geographical Longitude were then removed from the dataframe as part of feature selection as the contribution of the labels were considerably less compared to the others.

4.4 Modeling

Various predictive models were designed, after correlation analysis, feature selection, and cross-validation. Anomaly detection were utilized to identify statistically deviant data. To identify the dependencies and correlations in the data, association rules are applied. Regression models are used to fit mathematical functions to data. Hyperparameter tuning is found to be crucial in the modeling process as they control the overall behavior of the model. The acquired data is then loaded in the model for prediction. The dataset is given in two parts, train and test. The train dataset will be passed into the models for the learning process. Following this, the test data will be used to test the accuracy of the model. Using this model, predictions of the future can be carried out.

5 The Model

The name "Gradient Boosting" comes from Friedman's paper Greedy Function Approximation: A Gradient Boosting Machine and the name XGBoost stands for "Extreme Gradient Boosting." The XGBoost algorithm was developed as a research project at the University of Washington by Tianqi Chen and Carlos Guestrin. Boosting as such, is an ensemble technique where new models are added with the motive of correcting the errors made by the existing models. Models are constantly added sequentially until there is no further improvement noticed.

Gradient boosting is an approach where new models are created, and these models predict the residuals (errors) of the previous models, and then they are added together to make the final prediction. It is known as gradient boosting, as it uses a gradient descent algorithm for the process of minimizing loss while new models are added. Gradient boosting supports both regression and classification predictive modeling problems. The working of gradient boosting is as follows:

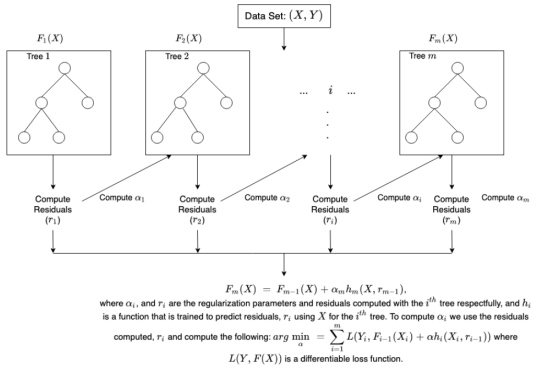

Figure 3: *docs.aws.amazon.com/sagemaker/xgboost*

The XGBoost library helps in developing fast and high performance gradient boosting tree models. In the recent past, XGBoost has been achieving the best performance on a wide variety of machine learning tasks. The XGBoost library implements the gradient boosting decision tree algorithm. This algorithm can also be called as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines.
XGBoost is an open-source implementation of the gradient boosted trees technique that is popular and efficient. Gradient boosting is a supervised learning approach that combines the estimates of a set of smaller, weaker models to attempt to accurately predict a target variable. XGBoost is implemented with a linear booster.
The booster parameter sets the type of learner. The weak learners in gradient boosting for regression are regression trees, and each regression tree transfers an input data point to one of its leafs containing a continuous score. There is a tree or a linear booster function. The linear booster will be a weighted sum of linear functions whereas in the case of trees, the model will consist of an ensemble of trees.
XGBoost combines a convex loss function (based on the difference between the anticipated and target outputs) with a penalty term for model complexity to minimize a regularized (L1 and L2) objective function (in other words, the regression tree functions). Iterative training is used to create new trees that forecast the residuals or errors of previous trees.
The objective determines the learning task and thereby, the type of the target variable. The options available for the object parameter include regression, logistic regression, binary and multi classification or rank. Using this, XGBoost can be implemented for a number of tasks. The default value is “reg:squarederror” , and the algorithm is implemented with the default for objective.

6 Results and Discussion

Regression analysis is a type of predictive modeling technique that looks into the relationship between a dependent (target) and independent (s) variable (predictor). It denotes the existence of significant correlations between the dependent and independent variables. It shows how strong many independent variables have on a dependent variable. The effects of variables assessed on different scales can be compared using regression analysis. For training and assessing a predictive model, we can first split the loaded dataset into input and output columns.The model is then fit to all available data and then by running predict() with a new row of data.
The evaluation metrics we implemented include:

1. **R2** : R squared: The r2 value works by measuring the amount of variance in the predictions obtained from the machine learning model. For our prediction, we got an R2 value of 0.996.

2. **RMSE**: Root mean squared error: This metric shows how far the predicted values fall, from the measured true values, using Euclidean distance. For our prediction, we got a RMSE value of 84053.68.

3. **MAE**: Mean Absolute Error: MAE measures the average magnitude of the errors in a set of predictions, without considering their direction; it takes the absolute value of the errors. For our prediction, we got a RMSE value of 21971.29.

4. **Explained Variance**: Explained variance is used to measure the proportion of the variability of the predicted values of a machine

learning model. For our prediction, we got an explained variance value of 0.996.

Training Accuracy	Testing Accuracy
0.999579	0.996456

Table 1: Training and Testing Accuracy

R2	Explained Variance
0.996	0.996

Table 2: R2 and Explained Variance

RMSE	MAE
84053.68	21971.29

Table 3: RMSE and MAE

7 Conclusion

Traditional approaches to real estate valuation can show error on the qualitative side, focusing more on intuition than logic. However, regression analysis can provide a reliable model for estimating property values based on previous transactions in a given area. In residential real estate, the comparable sales approach is most frequent, and it examines recent sales of similar properties to establish the value of a particular property. The sales prices are adjusted to account for variations between them and the subject property. Regression analysis is especially useful when dealing with enormous amounts of data. Although it would be hard to have a thorough understanding of every local real estate market in the country, regression modeling can aid in the search.

The cost of acquiring an identical piece of land and building a replica of the subject property is used to calculate value. The project's cost is then depreciated depending on the subject property's current state of obsolescence. Everything except Date, Time, Geographical Latitude, and Geographical Longitude affects the value of a home. This seemed to be more logical.

8 Future Works

In the course of this project, we have performed modeling and prediction, for house prices using the XGBoost machine learning boosting algorithm. In addition to this, time series forecasting can also be done, in order to predict the housing prices for the future, and analyze trends. Another implementation of such a model could be where we also integrate distance values from any map application, and predict the price of the house with respect to those attributes as well. For example, distance to the nearest metro station, bus station, or the closest mobile service provider tower, etc. These factors also influence the price of real estate and therefore they can be integrated into the analysis. Furthermore, using the latitude and longitude values given, the data can be segmented into different areas, and analysis and prediction can be done based on area. Big Data technologies like MapReduce can also be implemented, as the data is large, and we could not perform certain types of analysis owing to the lacking processing power of our GPU, and RAM space. These issues can be overcome by using Big Data frameworks.

9 References

1. Vladimirovna, V. E. (2016). Review of the Russian market of real estate. *Science Time, (3 (27)), 121-123*.

2. Nataliia Kharchenko, (2019) 6 use cases of big data AI in real estate.

3. Sharma, N., Arora, Y., Makkar, P., Sharma, V., Gupta, H. (2021). Real Estate Price's Forecasting through Predictive Modelling. In *Machine Learning for Predictive Analysis (pp. 589-597)*. Springer, Singapore.

4. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2, 1(4)*, 1-4.

5. Antipov, E. A., Pokryshevskaya, E. B. (2015). Are buyers of apartments superstitious? Evidence from the Russian real estate market. *Judgment and Decision Making, 10(6)*, 590.

6. Renaud, B. (1995). The real estate economy and the design of Russian housing reforms, Part I. *Urban Studies, 32(8)*, 1247-1264.

7. Peng, Z., Huang, Q., Han, Y. (2019, October). Model research on forecast of second-hand house price in Chengdu based on XGboost algorithm. In *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)* (pp. 168-172). IEEE.