Ludwig-Maximilians-Universität München

Faculty of Economics

ROLE OF MARKET POTENTIAL IN PREDICTING PRICES FOR REAL ESTATE IN MOSCOW

Iana Fedorchuk

Munich
2020

Contents

Introduction

Many people are facing the issue of buying or selling real estate, and an important criterion here is not to buy more expensive or not to sell cheaper relative to other, comparable options. The simplest way is a comparative one, focusing on the average price of a meter in a particular place and expertly adding or lowering percentages of the cost for the advantages and disadvantages of a particular apartment.

But this approach is time-consuming, inaccurate and will not allow us to take into account the whole variety of differences of apartments from each other. Therefore, we need to build a model for predicting a price by using data analysis and machine learning. This paper describes the main stages of such an analysis.

In large cities prices for the real estates increase and decrease at different speeds, prices depend on location, affiliation with a particular area, surrounding infrastructure, distance to public transportation, state of the property and many other factors. In order to predict property prices, a larger number of relevant factors should be used.

For the measure of infrastructure we use so called market potential in this paper for each apartment and for different groups of infrastructure.

The term ”market potential”, was proposed by Colin Clark. This is an abstract indicator of the intensity of possible contact with the markets[8]. The idea ultimately stems from physics, in which similar formulas are used to determine the strength of a field, whether electric, magnetic, or gravitational.

In this paper, we will consider the impact of market potential on the real estate price prediction. Market potential is a concept used in the scientific literature often in relation to the New Economic Geography (NEG)[7], which studies how agglomeration forces act on the development of cities. New Economic Geography dates back to 1991. Therefore, all papers on this topic and including NEG models are relatively novel. thus, the aim of this work is to improve real estate price predictions by adding a market potential factor. In this article, we will assess the impact of adding the accumulated market potential to the data set, and the market potential collected by groups.

In addition to market potential, various categorical variables are given in this work as explanatory variables: affiliation to a particular district, building material, building’s gas and heating supply type, house overlap type.

The data on the real estate involved in the empirical study were taken from cian[3] database. Location data was taken from Yandex.Maps[6]. Market potential was calculated on their basis.

Chapter 1 provides a theoretical basis for the following analysis and an overview of assigned empirical work that used google search data for the analysis and provides a critical review of assigned paper.

Chapter 2 discusses the models used, describes the data for analysis, and the method of calculating market potential.

Chapter 3 provides the results of empirical research.

1 The Impact of The Unemployment Insurance on Job Search: Evidence from Google Search Data

1.1 Summary

The article examines the impact of unemployment insurance on the job search. The authors use Google search data. Job search is a key variable that is used in labor market theories but it's very difficult to measure, so the authors of the article decided to use an index calculated on Google data which they called Google Job Search Index (GJSI)[2].

Since Google Trends provides search information for each term only in integers in the range from 0 to 1, and information about the absolute value of the number of searches is kept confidential. To get raw data related to job search searches and to make sure that their number is large enough authors decided to use other Google features such as Adwords showed that there were about 68 million monthly requests for the term work during the year from April 2012 to April 2013. Thus, the authors of the article confirmed that the number of searches for the term work is very large.

While the authors did their research they noticed that google job search index has some advantages over other databases related to job search by citizens before previous job search tracking methods. This is a report which is a questionnaire that aims to assess how much people spend searching for jobs. This is a report that is issued annually. This questionnaire interviews approximately 26,400 households using a telephone every year, but it is not targeted at the unemployed, so it gets very few unemployment observations at about 5 observations per month in every state in America.

Search engine for more specific queries that include words such as industry company name or country and city in which the person is looking for work, therefore, thanks to the specificity of the search engine Google, researchers can get job search data for specific industries, for specific cities, countries, states or companies.

Other online platforms such as Career Builder provide data that can be used by researchers in their work, but they are not always available and have fewer observations compared to Google trends, although it would be useful to look at the online work platform data so there catch some other specific patterns.

Authors decided to pay attention to databases in which there may be a smaller number of observations, in order to better estimate the labor market, they decided to look at other databases such as comScore which includes the habits of finding 100,000 Americans the authors found that the index is a good proxy for the effort of finding a job on the Internet.

The authors also compared their results with the American Time Use Survey. They found

that Google searches fluctuate equally to the results of the questionnaire. The authors showed that higher unemployment rates correspond to higher indexes.

Results of empirical analysis showed that in areas where the average unemployed citizen had unemployment insurance for less than 10 weeks, the index showed 66 more search activity than in nearby regions, where people had unemployment insurance for 10-20 weeks and 108 more than in areas where people had 30 weeks of unemployment insurance.

1.2 Critical Discussion

Job search activity may also depend on external factors not related to unemployment insurance. The authors could use the financial series as an explanatory variable responsible for external factors. This could be a control variable on the state of the economy. For example, if an economic agent notices that economic instability is growing, he or she may start spending more time looking for work. Because of the instability in the economy increases, the pressure on the unemployed person increases in the direction of job search.

As another explanatory variable, an indicator of population sentiment can be used. The method for obtaining this variable can be as follows:

- applying text analysis to newspaper news
- compilation of a list of words found in articles related to a specific topic
- compilation of a categorical variable (for example, dummy breakdown of variables of different news categories)
- adding such a variable to the data set can control an external factor such as the mood of the population.

Depending on the topic and sentiment of articles and information that economic agents consume, they form their subsequent activity on job search.

As a benchmark authors could use random walk for calculating error term.

2 Data and Method

2.1 Data

Explanatory variable - market potential will be calculated as follows:

$$mp_i = \sum_{j \neq i} \frac{1}{D_{ij}}, \tag{2.1}$$

where potential of i flat is a sum of fractions, where D_{ij} is a geodesic distance between i flat and j unit of infrastructure.

The program for calculating market potential is presented in the Appendix A.

It is generally accepted that the Earth has the shape of a ball with an average radius of 6371 km for computational problems. Therefore, in this work, we will use the following formula to calculate the distance between apartments and infrastructure units:

$$D_{ij} = arccos(sin(\varphi_i) \cdot sin(\varphi_j) + cos(\varphi_i) \cdot cos(\varphi_j) \cdot cos(\lambda_i - \lambda_j)) \cdot R, \tag{2.2}$$

where φ_i и φ_j - the latitude values of these apartments and infrastructure units, λ_i и λ_j- the longitude values of these apartments and infrastructure units, R - the radius of the Earth.

For calculating market potential variable were collected latitude and longitude for the following types if infrastructure:

1) bars and restaurants;
2) coffee shops;
3) grocery stores;
4) markets;
5) shopping malls;
6) theatres;
7) cinemas;
8) parks;
9) universities;
10) hotels;
11) office centers;
12) and sum of the stated market potentials as overall market potential.

Market potential grouped by types of infrastructure and overall market potential are used in different specifications of model.

Geographical coordinates were collected by using Yandex maps API and Python library ”yandex geocoder”.

Code for collecting the address line from page using JavaScript

```
var address_list = [];
var address_line = document.querySelectorAll("div.search-business-snippet-
    view__address");
for (var i = 0; i<address_line.length; i++){
      address_list.push(address_line[i].textContent);
}
address_list
```

The following variables are taken from Cian database. Cian.ru is the largest Internet service for buyers and tenants of housing in Russia. The project was launched in 2001. Today CIAN is a specialized portal with a large database of urban, suburban and commercial real estate in the Moscow region. The total number f observations in data set is 37418.

The following categorical variables are given in this work as explanatory variables:

1) district (affiliation to a particular district);
   - eastern administrative district, western administrative district, northern administrative district, southern administrative district, central administrative district, south-eastern administrative district, south-western administrative district, north-eastern administrative district, north-western administrative district, zelenogradsky administrative district, novomoskovsky administrative district, troiysky administrative district
2) building material:
   - block, brick, monolith, panel, wood
3) building's gas supply type:
   - autonomous, central, unknown, without
4) building's heating supply type:
   - autonomous boiler, boiler, central, itp, stove, without
5) building's overlap type:
   - concrete, mixed, other, unknown, wood
6) the closest underground line
7) user trust level:
   - danger, excluded, involved, new, not involved

Data set includes the following dummy variables:

1) new (1, if flat is new, 0, if flat is old);
2) Top3 (1, if the offer is in the top 3 list of the search results);
3) demolishedInMoscowProgramm (1, if the building takes place in the housing renovation program);
4) mortgageAllowed (1, if the mortgage allowed);

5) UndergroundUnderConstruction (1, if there are new underground station is under construction nearby).

The following variables are given as explanatory variables:

1) average daily number of visits for the ad page;
2) total number of visits;
3) number of agent's offers;
4) number of floors in the building;
5) number of flats in the building;
6) offer's floor number;
7) travel time to public transport;
   - byFoot, byCar
8) number of rooms;
9) number of similar offers.

Descriptive characteristics of variables

	price	st total	agent offers	build floors
count	37418	37418	37418	37418
mean	18569640	948.17	948.67	16.58
std	22422070	1696.31	2729.52	10.88
min	950000	0	0	1
25%	8700000	154	7	9
50%	12500000	456	47	15
75%	20000000	1134	498	21
max	1900000000	104540	15599	97
	building total area	floor number	rooms	similar offers
count	37418	37418	37418	37418
mean	67.74	8.46	2.14	0.5
std	32.41	7.85	0.77	2.49
min	11.5	1	1	0
25%	45	3	2	0
50%	61	6	2	0
75%	80	11	3	0
max	1971	84	4	124

2.2 The expected effect of explanatory variables on the dependent variable

How real estate pricing works

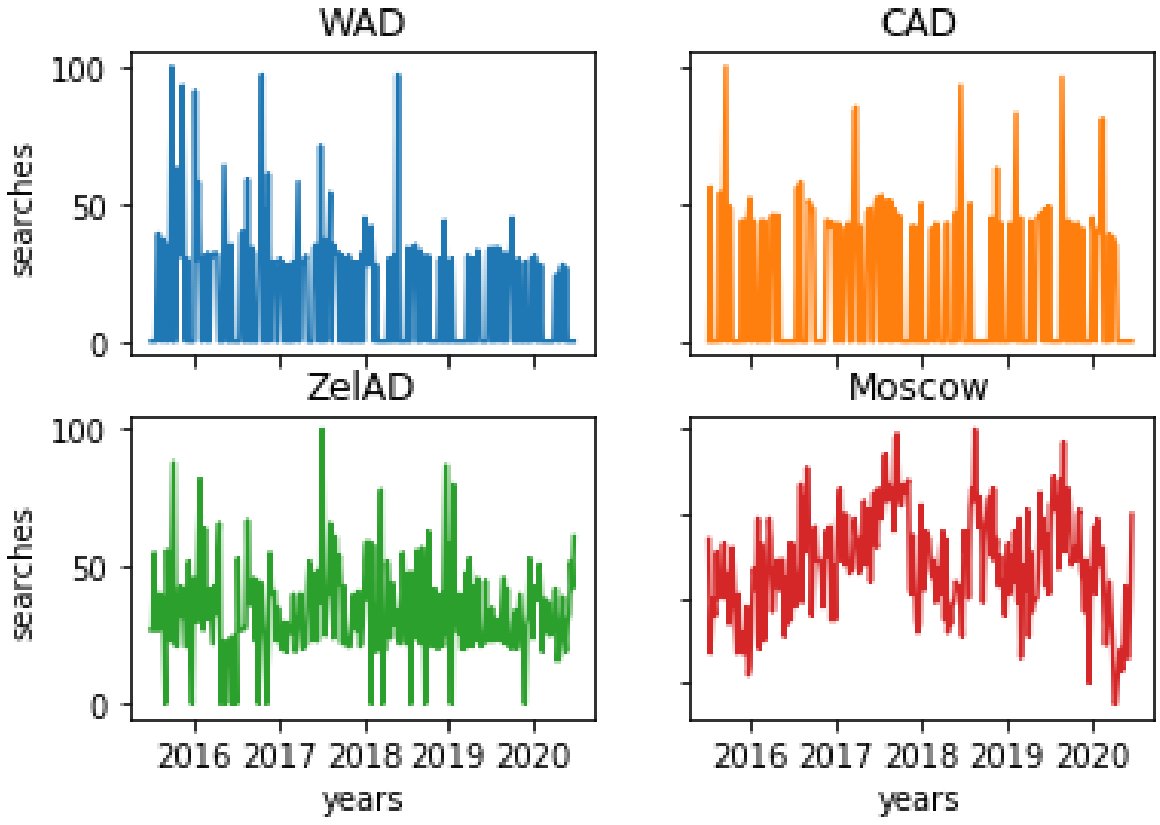
First, the price of real estate appears like for any other product when there is a need, the utility of the product for people, the rarity of it, that is, the value of the product.

Second, the real estate price depends on the construction costs and the future profit.

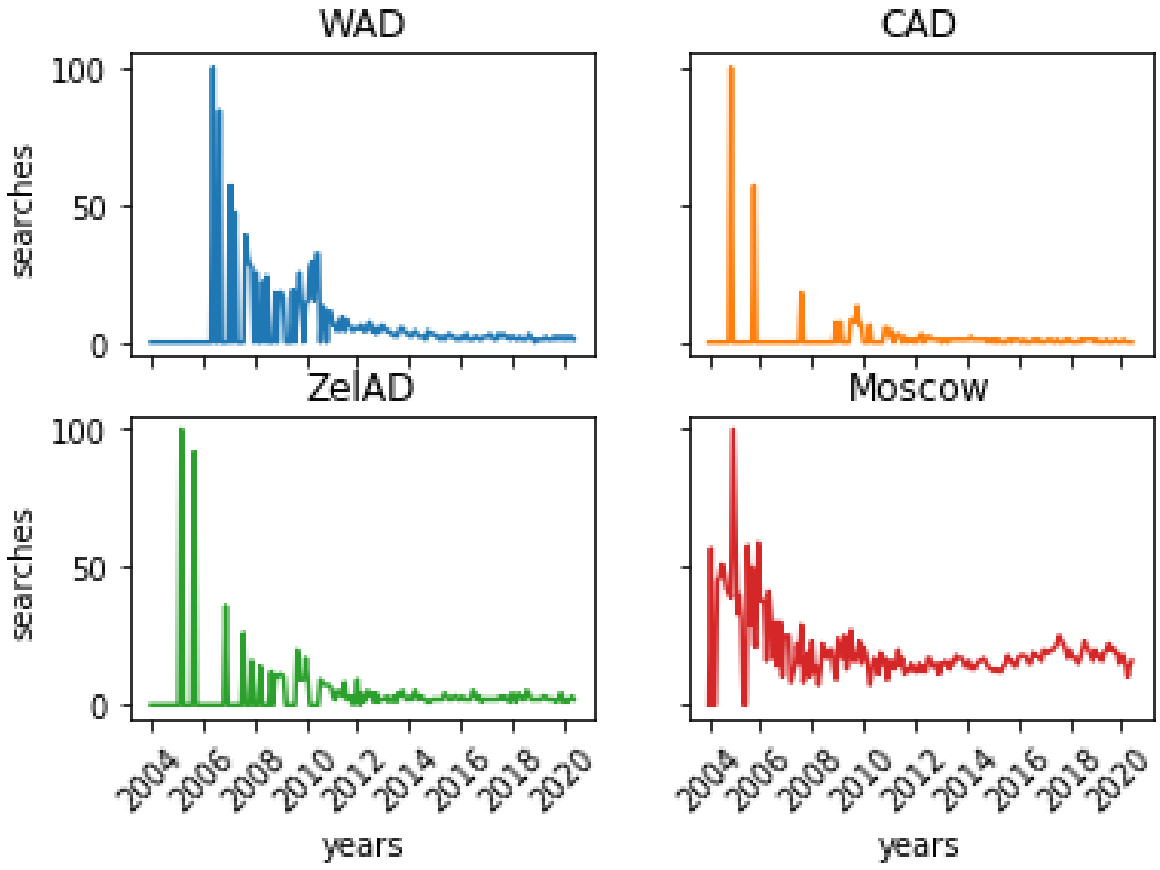
Third, the real estate price is formed by supply and demand, which ultimately determine the equilibrium price. The above factors cannot be considered separately as they all affect the value of real estate at the same time[5].

In this paper, we don't include the demand variable into the data set, because we are concentrated more on the material variables, but it's important to mention the economical part of the real estate pricing.

To see how the demand in the real estate market in Moscow has changed, we look at Google trends. We construct a plot for the change in the requests number for Moscow and its districts (western administrative district, central administrative district, Zelenogradsky administrative district, Moscow in total (WAD, CAD, ZelAD, Moscow) for time intervals (2015-2020, 2004-2020). See Plot 2.1 and Plot 2.2 respectively.



Plot 2.1 — Change in the number of searches given in per cents 2015-2020



Plot 2.2 — Change in the number of searches given in per cents 2004-2020

As a general rule, the basis of real estate pricing depends on construction pricing, i.e. from the profit that the construction company is planning to yield and costs that are laid in a particular real estate object. These costs, taking into account the minimum profit of further prices.

When real estate is put into operation and then it is being placed in the secondary market, where various factors make a big contribution to the change in the future price over time. Newly built apartments have higher prices than apartments in the secondary market. Condition of apartments in the secondary market can vary widely. What also affects the price of apartments in the secondary market.

Material factors

In addition to the factors described above, there are material factors that affect prices in the real estate market. Supply and demand are greatly influenced by the location factor, i.e. two similar apartments can cost differently if they are located in different areas. Because every coordinate is characterized by the development of infrastructure and the main specialization of the district (the housing neighborhood, office center, city center with a lot of entertainment properties). Location is a unique factor, because it cannot be changed for a real estate unit. For

residential real estate, it is important that infrastructure is highly developed, closeness to grocery stores, fitness centers, parks, cafes, low noise level.

For the measure of infrastructure we use so called market potential in this paper for each apartment and for different groups of infrastructure as stated in previous sub-chapter.

The more infrastructure units located near the apartment, the more attractive it is in the real estate market, therefore the price of such an apartment is higher.

The next pricing material factor in real estate is the floor on which the apartment is located. In residential real estate the ground floor is the cheapest and with an increase in the floor, the price rises and then falls again. Because living in the ground floor is connected to some insecurities (noise levels are higher on the ground floor) as well as living on the top floor (the temperature is higher during summer, leaking roof).

The next material factor that affects the price of a property is the building material. The most common materials that are used in residential construction are monolith, panel, brick, wood, and concrete. For the most part, the area of 1 square meter in a panel and monolithic houses is cheaper than in brick houses. Brick houses have a strong foundation and thick walls, which entails high costs for construction materials, labor, and equipment. And also people prefer to live in brick houses, because the quality of living in such a building is better, because of good thermal and noise insulation.

For prefabricated houses, special factories manufacture the panels, so the construction is faster compared to brick houses. Monolithic construction is developing at a high pace. Thus, panel buildings are cheaper in terms of construction. So the basis of pricing for monolithic buildings is lower.

The next material factor that affects the price of the apartment is the total area of the property. In general, real estate with a small area is in great demand, it is easier to sell than a property with a large area, so small areas have a high price per 1 square meter in relation to large areas. In Russian Federation, housing with small areas has received great development, which is highly demanded and, accordingly, has a high cost per 1 square meter.

2.3 Method

Machine learning is a subsection of artificial intelligence that studies algorithms that can learn without directly programming what needs to be learned.

Due to machine learning, programmers are not required to write instructions that take into account all possible problems and contain all the solutions. Instead, an algorithm for the independent finding of solutions is put into a particular program by the joint use of statistical data, from which the patterns are carried out and on the basis of which forecasts are being made subsequently.

The technology of machine learning based on data analysis dates back to 1950. Machine learning was not used that broad in the beginning, because computing power did not allow to solve complex problems quickly enough. In the past years, devices with a high computing power became relatively cheaper and accessible to almost everyone. Thus, the range of tasks and problems solved using machine learning has increased. Functions and libraries for solving problems are increasing due to the fact that the open-source phenomenon has been adopted in the data science community. Therefore, machine learning is the most effective and modern way of solving problems, for example, such as regression and classification.

All tasks solved using ML belong to one of the following categories:

1) The task of regression is a forecast based on a sample of objects with various explanatory variables (features). The output of the model should be a real number, for example, the price of a diamond, the price of a wine bottle, the price of the apartment, etc.
2) The task of classification is to answer a question to what class belong an object. It can be binary or multi-class. The most common example of the classification problem is a problem when computer needs to distinguish photos of cats and dogs from each other and give an answer whether the cat or dog is in the picture. Another popular use of classification problem is to predict whether a patient has a risk of a particular disease.
3) The task of clustering is when we need to distribute data objects into groups: the separation of articles of newspaper by subject (politics, economics, etc), the classification of vehicles in one category or another, the classification of job offers by industries (IT, Human Resources, Sales, Marketing).
4) The task of detecting anomalies is to separate the anomalies from standard cases.
5) Dimensionality reduction task is to reduce a large number of variables to a smaller number for easier visualization (Reducing features from 18 to 3 to get a 3-dimensional plot).

The main types of machine learning

The bulk of the tasks solved using machine learning methods relates to two different types: learning with a teacher “supervised learning” or without a teacher “unsupervised learning”. However, this teacher is not necessarily the programmer himself or herself, who stands above the computer and controls every action in the program. “Teacher” in terms of machine learning is the intervention of a person in the processing information process. For better understanding the difference between the two types, their examples are in the following[1].

Machine learning with a teacher

1) We have information about fifty thousand diamonds: carat, depth, cut, color, clarity, price, etc. We need to create a model that predicts the market value of diamond by its parameters. This is an example of machine learning with a teacher: we have the initial data (the number of diamonds and their parameters, which are called features) and a ready answer for each of the diamonds is its cost. The program has to solve the problem of regression.

2) Other examples: to confirm or deny the presence of heart disease for a patient, taking into account his medical indicators. Find out if the applicant's CV is relevant for a particular job offer.

Machine learning without a teacher

1) Training without a teacher is when ready-made "right answers" are not provided to the computer. For example, we have information about foot length and width of number of children, and we need to divide this data into ten groups, to determine 10 sizes of shoes that we want to produce. This is a clustering task.
2) If we take a different situation, we have in our sample has more than a hundred of different features for each object, then the main difficulty will be the graphical display of such a sample. Therefore, the number of features is reduced to 2 or 3. This is the task of reducing dimension.

We have information 37418 Moscow apartments: area, floor, district, distance from the metro, apartment price, etc. We need to create a model that predicts the market value of an apartment by its parameters. We have the initial data (the number of apartments and their features). Thus, in this paper, we need to solve the regression problem. This is machine learning with the teacher. We have the "right" answers - prices of the apartments.

For the empirical analysis this paper uses Random Forest Machine Learning Method[4].

Features of Random Forest:

- Random Forest - Bagging on Decisive Trees
- Build independent deep trees and average them
- One of the most powerful machine learning methods
- Works well with diverse features, is not whimsical, is not sensitive to outliers, does not require scaling, does not overfit

2.4 Implementation description

In the analytical research were used three specifications of the model.

1) Base model: includes features that were extracted from Cian database. Data were cleaned and dummies were extracted for the categorical features.
2) Model with accumulated market potential. Difference between second specification and base model is one feature that is a sum of market potential (not divided by categories).
3) Model with market potential divided by categories. Difference between third specification an d base model is 11 features that are market potential divided by the following categories:
   - bars and restaurants;
   - coffee shops;
   - grocery stores;

- markets;
- shopping malls;
- theatres;
- cinemas;
- parks;
- universities;
- hotels;
- office centers.

Idea of these specifications were to check the following hypothesis:

1) Adding market potential feature as an infrastructure measure improves predictions. Because the more infrastructure units located near the apartment, the more attractive it is in the real estate market, therefore the price of such an apartment is higher.
2) Adding market potential by categories as an infrastructure measure improves predictions. Because prices of apartments may change depending on the types of infrastructure units that are nearby.

For choosing the best parameters for Random Forest Regressor were used GridSearchCV function in Python.

The search was performed for the following hyper-parameters:

- bootstrap: [True, False],
- max depth: [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
- max features: [’auto’, ’sqrt’],
- min samples leaf: [1, 2, 4],
- min samples split: [2, 5, 10],
- n estimators: [200, 400, 600, 800, 1000, 1200, 1400, 1600].

Function for evaluating GridSearchCV and basic model (MAPE)

```
def evaluate(model, X_test, y_test):
    predictions = model.predict(X_test)
    errors = abs(predictions - y_test)
    mape = 100 * np.mean(errors / y_test)
    accuracy = 100 - mape
    print('Model Performance')
    print('Average Error: {:0.4f} degrees.'.format(np.mean(errors)))
    print('Accuracy = {:0.2f}%.'.format(accuracy))

    return accuracy
```

3 Results

Results of the three specifications model fitting presented in the table.

Results of every model specification compared together with a base Random forest model and with best parameters of GridSearchCV function.

	base RF model	GridSearchCV	Improvement
base specification	41.58%	43.69%	5.06%
accumulated mp	41.53%	44.95%	7.39%
mp with categories	43.44%	46.42%	6.86%

Improvement compared to base specification	base RF model	GridSearchCV
accumulated mp	-1.00%	2.88%
mp with categories	4.47%	6.25%

Improvement compared to accumulated mp	base RF model	GridSearchCV
mp with categories	4.60%	3.27%

The hypothesis of the research are not rejected by the results of the empirical analyses. Market potential impacts positively on the apartment price forecasting.

Better accuracy of predictions received when market potential divided by categories were added.

Best parameters of Grid Search for base model specification:

- bootstrap: False,
- max depth: 30,
- max features: sqrt,
- min samples leaf: 1,
- min samples split: 2,
- n estimators: 800.

Best parameters of Grid Search for mp model (without categories):

- bootstrap: False,
- max depth: 30,
- max features: sqrt,
- min samples leaf: 1,
- min samples split: 2,

– n estimators: 200.

Best parameters of Grid Search for categorial mp model:

– bootstrap: False,
– max depth: None,
– max features: sqrt,
– min samples leaf: 1,
– min samples split: 2,
– n estimators: 1400.

Best hyper-parameters of models' specifications in our research vary only in max depth of each tree and number of estimators.

Conclusion

In this research we looked at the paper that used Google search data for the formation of the index that were used as a variable.

For the empirical analysis in this paper were used Cian database that includes a large number of features for the apartments. We used web-scraping to collect geo data and calculate a measure of infrastructure that we called market potential. Our goal was to increase accuracy of apartment prices prediction by adding market potential as a feature.

Hypothesis of the paper:

1) Adding market potential feature as an infrastructure measure improves predictions. Because the more infrastructure units located near the apartment, the more attractive it is in the real estate market, therefore the price of such an apartment is higher.
2) Adding market potential by categories as an infrastructure measure improves predictions. Because prices of apartments may change depending on the types of infrastructure units that are nearby.

The following conclusions were made regarding the results obtained:

The hypothesis of the research are not rejected by the results of the empirical analyses. Market potential impacts positively on the apartment price forecasting.

Better accuracy of predictions received when market potential divided by categories were added.

Literature

1. Alpaydin E. Introduction to machine learning. — MIT press, 2020.
2. Baker S. R., Fradkin A. The impact of unemployment insurance on job search: Evidence from Google search data // Review of Economics and Statistics. — 2017. — Vol. 99, № 5. — P. 756–768.
3. Cian. — 2020. — URL: https://www.cian.ru/.
4. Classification and regression by randomForest / A. Liaw, M. Wiener, [et al.] // R news. — 2002. — Vol. 2, № 3. — P. 18–22.
5. Lusht K. M. The real estate pricing puzzle // Real Estate Economics. — 1988. — Vol. 16, № 2. — P. 95–104.
6. Yandex.Maps. — 2020. — URL: https://yandex.com/maps/.
7. Hassink R., Gong H. New economic geography // The Wiley Blackwell Encyclopedia of Urban and Regional Studies. — 2016. — P. 513.
8. Krugman P. What's new about the new economic geography? // Oxford review of economic policy. — 1998. — Vol. 14, № 2. — P. 7–17.

Appendices

A Program for calculating market potential (Python)

Program and Data can be found with the following URL:
https://github.com/yanafedorchuk/DataScience_Seminar
Program was written with Python programming language.

```
def distance(lat1,lon1,lat2,lon2):

    radius = 6371  # km

    dlat = math.radians(lat2 - lat1)
    dlon = math.radians(lon2 - lon1)
    a = (math.sin(dlat / 2) * math.sin(dlat / 2) +
         math.cos(math.radians(lat1)) * math.cos(math.radians(lat2)) *
         math.sin(dlon / 2) * math.sin(dlon / 2))
    c = 2 * math.atan2(math.sqrt(a), math.sqrt(1 - a))
    d = radius * c

    return d
def mp_cal(lat1,lon1,lat2,lon2):
    if distance(lat1,lon1,lat2,lon2)==0:
        first_step=0
    else:
        first_step=1/distance(lat1,lon1,lat2,lon2)
        return first_step

column_np=np.array([])
for lat1,lon1 in zip(df_train1.offer_geo_coordinates_lat, df_train1.
    offer_geo_coordinates_lng):
    Pls=np.array([])
    for lat2,lon2 in zip(df_br.latitude, df_br.longitude):
        Pls=np.append(Pls,mp_cal(lat1,lon1,lat2,lon2))

        e = np.array([])
    for i in Pls:
        if i is None:
            e=np.append(e,0)
        else:
            e=np.append(e,i)
    column_np=np.append(column_np,sum(e))
new=list(column_np)
df_train1['mp_br']=new
```