Federal State Institution of Higher Professional Education

Saint Petersburg State University

Graduate School of Management

«MACHINE LEARNING METHODS APPLICATION FOR REAL ESTATE PRICE PREDICTION IN THE RUSSIAN MARKET»

Bachelor’s Thesis by the 4th year student

Concentration – Information Management

RAU Vladislava L.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Signature)

Thesis Supervisor:

Senior Lecturer

GOROVOY Vladimir A.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Signature)

«MEET THE REQUIREMENTS»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Thesis Supervisor Signature)

«\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2017.

Saint Petersburg

2017

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Я, Рау Владислава Леонидовна, студентка 4 курса направления 080500 «Менеджмент» (профиль подготовки – Информационный менеджмент), заявляю, что в моей выпускной квалификационной работе на тему «Применение методов машинного обучения для предсказания цены объекта недвижимости на российском рынке», представленной в службу обеспечения программ бакалавриата для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата.

Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее курсовых и выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

Мне известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за предоставление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».

(Подпись студента)

(Дата)

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF THE BACHELOR THESIS

I, Vladislava Rau, fourth year bachelor student, Bachelor of Information Management program, state that my master thesis on the topic «Machine Learning Methods Application for Real Estate Price Prediction in the Russian Market», which is presented to the Bachelor Office to be submitted to the Official Defense Committee for the public defense, does not contain any elements of plagiarism.

All direct borrowings from printed and electronic sources, as well as from term papers and bachelor and master theses, PhD and doctorate theses which were defended earlier, have appropriate references.

I am aware that according to paragraph 9.7.1. of Guidelines for instruction in major curriculum programs of higher and secondary professional education at St. Petersburg University «A bachelor thesis must be completed by each of the degree candidates individually under the supervision of his or her advisor», and according to paragraph 51 of Charter of the Federal State Institution of Higher Professional Education Saint-Petersburg State University «a student can be expelled from St. Petersburg University for submitting of the course or graduation qualification work developed by other person (persons)».

(Student’s signature)

(Date)

# Аннотация

|  |  |
| --- | --- |
| Автор | Владислава Рау |
| Название выпускной квалификационной работы | Применение методов машинного обучения для предсказания цены объекта недвижимости на российском рынке |
| Факультет | Высшая Школа Менеджмента |
| Направление подготовки | Информационный менеджмент |
| Год | 2017 |
| Научный руководитель | Владимир Андреевич Горовой |
| Описание цели, задачи и основных результатов | В данной выпускной квалификационной работе была разработана модель машинного обучения, которая прогнозирует цен на недвижимость и может быть использована в качестве первой оценки владельцев недвижимости и агентов по недвижимости.  Для разработки модели использовался язык программирования Python и некоторые из библиотек машинного обучения, такие как scikit-learn.  В качестве датасета был массив из 50000 объявлений, размещенных на сайте Яндекс.Недвижимости, и главной задачей было создать модель прогнозирования цен с высочайшим уровнем точности.  При проведении анализа оценки эффективности моделей были использованы реальные цены объявлений, методы оценки агентов по недвижимости, модели конкурентов Яндекс.Недвижимости, а также разработанные модели, использующие методы машинного обучения.  Лучший результат показала xgboost модель, дающая 75% точность. |
| Ключевые слова | предсказание цены на недвижимость, оценка недвижимости, машинное обучение, машинное обучение в недвижимости |

# Abstract

|  |  |
| --- | --- |
| Bachelor Student’s Name | Vladislava Rau |
| Bachelor Thesis Title | Machine Learning Methods Application for Real Estate Price Prediction in the Russian Market |
| Faculty | Graduate School of Management |
| Main field of study | Information Management |
| Year | 2017 |
| Thesis Supervisor’s Name | Vladimir A. Gorovoy |
| Description of the goal, tasks and main results | In this thesis, I designed a machine learning model which predicts real estate price that could be used as the first estimation by real estate owners or real estate agents.  The source programming language is Python and some of the machine learning libraries such as scikit-learn.  I was given a 50000 items dataset and the main task was to create a price prediction model with the highest level of accuracy.  In comparison analysis, I evaluate the effectiveness of real estate agents price estimation, Yandex competitors’ models, only machine learning based models and combined models containing algorithmic and machine learning part.  The best price prediction model was xgboost model with the 75% accuracy |
| Keywords | real estate price prediction, machine learning |

# Table of Content

[Introduction 9](#_Toc483252006)

[Background 9](#_Toc483252007)

[Research problem 11](#_Toc483252008)

[Organization of the study 12](#_Toc483252009)

[Chapter 1. Price prediction models overview 14](#_Toc483252010)

[1.1 Real estate price prediction concept used by professional realtors 14](#_Toc483252011)

[1.2 Real estate price prediction concepts used by online real estate aggregators 16](#_Toc483252012)

[1.2.1 Cian.ru 17](#_Toc483252013)

[1.2.2 IRN.RU 19](#_Toc483252014)

[1.2.3 Domofond.ru 20](#_Toc483252015)

[1.2.4 Ocenchik.ru 21](#_Toc483252016)

[1.2.5 Flatorial 22](#_Toc483252017)

[1.2.5 Comparison results 22](#_Toc483252018)

[1.3 Proposed machine learning based models 23](#_Toc483252019)

[1.3.1 Kaggle competitions and other approaches 23](#_Toc483252020)

[1.3.2 Scientific papers published in Google Scholar 24](#_Toc483252021)

[Chapter 2. Model development 26](#_Toc483252022)

[2.1 Dataset description 26](#_Toc483252023)

[2.2 Creating models with clear data 30](#_Toc483252024)

[2.3 Adding new features to dataset 31](#_Toc483252025)

[2.4 Anomalies detection 33](#_Toc483252026)

[2.5 Re-running previous models on updated dataset 36](#_Toc483252027)

[2.6 Implementing grid search and XGBoost models 37](#_Toc483252028)

[Chapter 3. Findings and conclusion 39](#_Toc483252029)

[Chapter 4. Further research and development 41](#_Toc483252030)

[4.1 Decreasing subjectivity 41](#_Toc483252031)

[4.2 Adding dynamic features 41](#_Toc483252032)

[4.3 Dealing with small dataset 41](#_Toc483252033)

[4.4 Using advanced machine learning models and techniques 42](#_Toc483252034)

[List of references 43](#_Toc483252035)

[Appendix A 46](#_Toc483252036)

[Appendix B 55](#_Toc483252037)

[Appendix C 57](#_Toc483252038)

# Introduction

## Background

There are different ways to know the approximate market price of the apartment. However, even if one invites several professional appraisers or experienced realtors, each of them will call their own number, and all of them will be different (in some cases very different) for sure. Furthermore, the real sales price, most likely, will not coincide with any of them. Real estate prices in Moscow depend on many different factors, until the mental attitude of the seller or/ and the buyer at the time of bargaining. That is why for seller it is important to know the price corridor of his/her apartment and have quick apartment analysis to understand the appraisal process and make it more transparent for him/her. Here online price prediction service come.

In the Russian market there are already several services providing online real estate price estimation, but the difference between their results are volatile and need extra attention from real estate agency to reconsider the object and give estimation that is more precise. In addition, there is no apartment analysis given, so customer does not understand on which criteria the price depends on and what its weight in final price formula. This research is an attempt to develop an objective price prediction model with high level of accuracy using cutting-edge technologies and techniques that could help increase truthfulness of given estimation.

Nowadays customers’ expectations are constantly increasing, and one of the main organization’s goals is to create such benefit, which will help to meet these expectations. In terms of real estate market, several factors are vital for positive customer experience: declaration of sale authenticity and information about the object truthfulness, transparent real estate object assessment criteria. The main goal for real estate owner is to rent or sell apartment at maximum possible price, but this price should be realistic.

Before machine learning wide spreading, there was attempts to predict prices using different math formulas, but one should take into account many subjective criteria that cannot be understand from data without any special processing.

Many management decisions are turning into data-driven one relying on data science methods and advanced business analytics. Quite the same story is with using machine learning to solve business-related issues: nowadays there is a huge demand to machine learning (especially neural networks and deep learning) solutions to a wide variety of issues.

One of outstanding examples in cracking this kind of cases is Yandex Data Factory. It has already created solutions for such sectors like manufacturing, retail, telecommunication, banking and finance.

Key definitions:

Real estate price prediction – the process of estimation the price regarding the data about real estate given using straight-forward algorithm to weight every feature definition.

Machine learning – a broad subfield of artificial intelligence that studies methods of building algorithms that are able to learn. One of its main sections - supervised learning, which was designed to solve the following task: there are a lot of objects (X) and a lot of possible responses (Y), also there is some dependence between objects and responses, and however, it is unknown. The goal of machine learning is to understand this dependence and to create a model which could give the most accurate answers to given objects.

Machine learning model – model, which contains machine-learning methods and solves machine-related problem (classification, prediction, clustering etc.)

Machine learning classification – a set of machine learning models where we need to understand what the object is related to. [[1]](#footnote-1)

Machine learning clustering – a set of machine learning models where we need to split the data into the groups according to some unknown rule.

Cross-validation testing – a testing method where the dataset is cut down into smaller datasets independently and the models are running of each dataset, as a result we have an average error without using extra data for testing phase.

Linear regression – a statistical method for drawing a straight line in accordance with a number of objects and responses.

Neural Network – the machine learning model, consisting of connected and interacting perceptrons - artificial neurons. The perceptrons send each other the signals and learn based on them.

Decision Tree – the tree, the leaves of which are attributes, and in other nodes there are the signs, which are classified the object.

Random Forest – an ensemble of decision trees, in whose nodes the random signs are located. The tree predictions are combined to obtain the most accurate result.

KNN (K-nearest neighbor) – the machine learning model, in which the object's class is determined by influence of the k nearest objects of the training dataset.

Gradient boosting – the ensemble of decision trees, in which each of the following the tree is added with the calculation of the gradient to correct the error of the existing ensemble.

Business-to-consumer (B2C) business model – business model where value chain ends with the customer without any additional agents between business and direct customer

Customer-generated content – content, which is generating by the customer. Customer provides all necessary data and information to the service that could help in creating the best solution to the problem.

## Research problem

Despite the fact that organizations do their best to implement machine-learning solutions and there are many researches covering this topic, it is not clear how to solve this issue in commercial way with high level of accuracy. Data is different and researcher could not even predict which features and which metrics would represent the business issue in the right way.

The topic of machine learning solutions in real estate Russian market seems very attractive and has room for organizational improvement and change. It is important to mention, that real estate market is different from other selling products industries. Nowadays all price prediction model work with less than 60 per cent accuracy and machine-learning methods implementation could increase this characteristic to more than 90 per cent (but the model accuracy depends on the dataset).

The aim of this research is to identify the best suitable machine learning application to price prediction issue, identify the opportunities and value of machine learning implementation for customers and organization itself.

The main research question: What is the best machine learning method solving price prediction real estate problem in Russian market with the highest accuracy?

Sub-questions:

* What are the possible positive effects of this opportunity implementation?
* How to solve this issue in cities without huge dataset and data in general?
* How to decrease the subjectivity of real estate evaluation process in terms of price and set the fair and transparent deal?

## Organization of the study

The research contains four main parts. In the first chapter, there are key theoretical background that could help for better understanding of the issue. For instance, how real estate agent evaluates an object and sets the price. What techniques and algorithms are already exist in the market or going to be implemented or improved by online real estate aggregators like Cian and Domofund. In addition, finally yet importantly, in this chapter all key machine learning fundamentals that I found useful.

In the second chapter, the key part is the entire way of machine learning development with all problems that I have coped with. Here there is full description of all features that were added to dataset and metrics that were a key indexes of model accuracy.

The third part contains all key findings and conclusion including the last comparison analysis with existing competitors and realtors-based approach. As part of the research, it is important to understand to what extent the model developed could be used as business related model which can bring significant profit to the company. And if not so at that time, how it could be improvement in very near perspective.

In the very last part, all key conclusions would be covering and here there is a research containing ideas for model improvement during the further research. In addition, answers to sub-questions of this research will be given. While the price estimation is demanded in all regions of Russia (especially in other city with millions people lived), one should have a strict algorithm how to deal with small datasets and inappropriate data that could be extracted from the open sources.

# 

# Chapter 1. Price prediction models overview

To understand the accuracy of these calculators let appraise one apartment and then compare the results by Yandex. The apartment characteristics will be following:

Table 1. The apartment characteristics for appraisal

|  |  |
| --- | --- |
| Address | Varshavskoye shosse, 16 |
| District | Nagatino-Sadovniki |
| Closest subway station | Nagatino |
| Distance to subway station (on foot) | 5 minutes |
| Total/ living/ kitchen area | 55/32/7 meters sq feet |
| Rooms amount | 2 |
| Apartment floor/ floors | 3/9 |
| Repair condition | good |
| House type | brick |
| Windows orientation | into the yard |

## 1.1 Real estate price prediction concept used by professional realtors

Professional realtors and appraisers have developed several publicly available methods of evaluating real estate objects. The two main criteria considered in price prediction methods in most cases are house category and house location (for Moscow it is municipal district).

In Moscow the municipal district is the most important feature in price estimation process because two apartments with the same meter sq feet, house category, repair etc. in Krylatskoe (the most prestigious residential district) and Kapotnya (the least prestigious residential district) could be different in 1.6-1.8 times. To solve this kind of issue realtors use the scale of prestige of all 120 of the municipal districts (see Appendix A): the less the corresponding rating is, the higher will be the average cost per square meter of total area of apartment in one type of house.

As for house category, this feature is also important, but not as the district one. Usually difference in meter sq feet price in Stalin-era buildings and Khrushchev-era buildings is about 1.5 times. In the Russian market, 95 per cent of real estate can be divided into three main categories: business, economy and khrushchevka. Let us consider all of them more precisely and give the characteristics of each category.

First category is business-class apartments. This category includes modern brick, built on an individual project, super Stalin-era buildings or just Stalin-era buildings, and profitable houses of the early twentieth century, then constructivist buildings (20-30’s years) and Stalin-era five-floor buildings with high ceilings. To go deeper, the classification will be:

* 1A2
* 1A1
* 1
* 1B1
* 1B2

Second category is economy-class apartments. This includes a modern panel and block multi-floor buildings (12, 14, 16, 17 b 22 floors), the Brezhnev-era multi-entrance buildings built in the late 60's - early 80-ies. Also for the price in this class are included 2-5 floor building going under renovation (“workers” and “German” towns). Classification that is more detailed is following:

* 2A2
* 2A1
* 2
* 2B

Third category is Khrushchev-era apartments. In this category, there are brick, block, prefabricated five-floor buildings built during era of mass industrial housing. Among them, one can distinguish "improved" apartments: with a relatively large kitchens 6-8 sqm or separate rooms, or brick. Important to mention, that khrushchevka or emergency condition building is those that soon will be demolished, and could cost possible higher because of government house-replacement program. In the third category there are three sub-classes:

* 3A
* 3
* 3B

Furthermore, there are two classes out of list: “out of list luxe” and “out of list”. If “out of list luxe” is special class that has its own appraisal algorithm, “out of list” is old house stock that is going to be reconstructed or even demolished. In addition, apartments in this type are not subject to sale and privatization, so from customers’ point of view they are out of interest.

Last but not least, to appraise the apartment as accurately as possible, one should takes into account the average price for district, average price for house type, and as a percentage of average cost, all existing advantages and disadvantages of considering apartment.

In Appendix B there is a table containing detailed description of percentage change according to concrete apartment features. Also in Appendix C one could find an example of apartment appraisal using the professional realtor’s methodology.

## 1.2 Real estate price prediction concepts used by online real estate aggregators

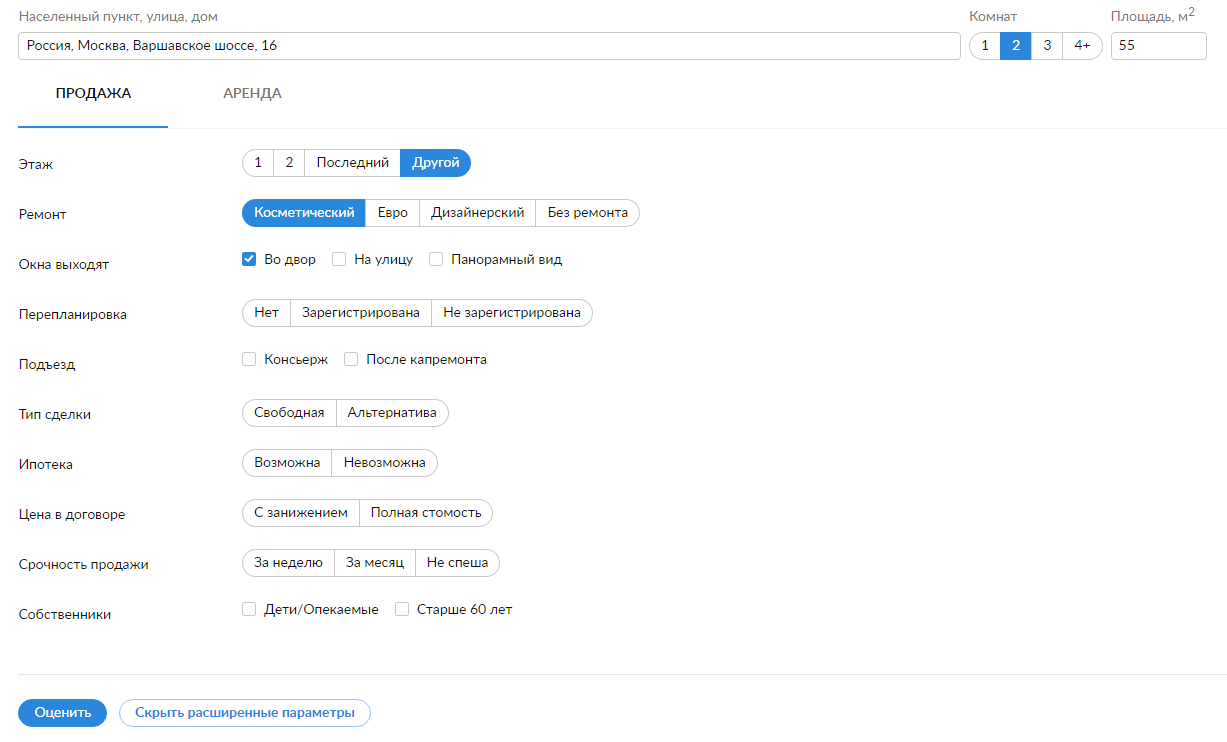
In the Russian market there are many services that provides quick online real estate price estimation, but only few of them seems relevant in terms of accuracy. In this subchapter, all popular online calculators will be considered with determining their potential disadvantages.

To determine the cost of an apartment the customer needs to specify apartment main parameters, which significantly affect the price of housing. Less important factors such as the presence or absence of a balcony, phone, floor covering, ceiling height, separate or combined bathroom and other such parameters have an impact on the price of the apartment at 1% -2% and are within the error estimation method. In most cases, this error amounts to 3% -5%, which is quite acceptable even in the professional assessment with the participation of a live expert. Error of estimate can be more expensive for luxury homes, apartments in the city center, atypical apartments, and apartments with renovated and other non-standard options. Rather moody for the automatic evaluation are representatives of the other extreme – a small one-bedroom apartments and studio apartments, as the cost per square meter of apartments may significantly stand out from the average level of housing prices.

### 1.2.1 Cian.ru

Cian price prediction methodology is one of the benchmark because it contains machine learning methods (k nearest neighbor classifier) and gives percentage of appraisal accuracy to the customer.

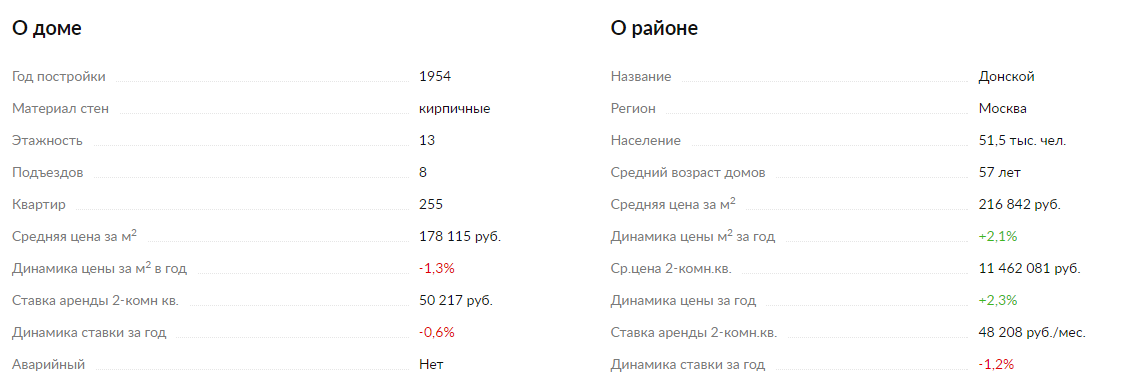
Let us consider the model more precisely: it compares the apartment to customer’s location with many similar ads from the Cian database for the last 5 years. The smaller differences at the apartment in the ad is, the higher the weight of this equivalent in the calculation will be. In determining the price, the most important prices for apartments in customer’s house with the same number of rooms, then all the other apartments, the apartments in similar houses in customer’s area. For the similar houses selection, there is a study on the typology of the housing stock in Moscow and MO. All the houses in the Moscow region are divided into 15 groups according to their number of floors, period of construction, design features. Relevant ads have a higher weight compared to archived and removed from sale. To bring the archive to the current price system is used correction factors. The calculations do not use duplicate ads apartments, which have received valid complaints, suggestions on the first and last floor, ad sales with non-market conditions. In the end, to calculate the cost per square meter in the house we use 5-7 counterparts (as standard), and hundreds and thousands of similar ads that have passed not only through strict moderation, but also through numerous filters for the mass appraisal model. 15 listings with the greatest weight are displayed on the page of the calculator in the section "history of the announcement to this house."



1. Apartment appraisal Cian.ru interface.[[2]](#footnote-2)

Cian provides not only average price including the price corridor, but also the accuracy percentage and detailed description of house and district. Moreover, the customer sees what apartments he/she can buy or rent in another different locations with the same average price.

This information seems useful for customer and because of a lot of extra information given customer starts thinking about how transparent and increase loyalty to service.



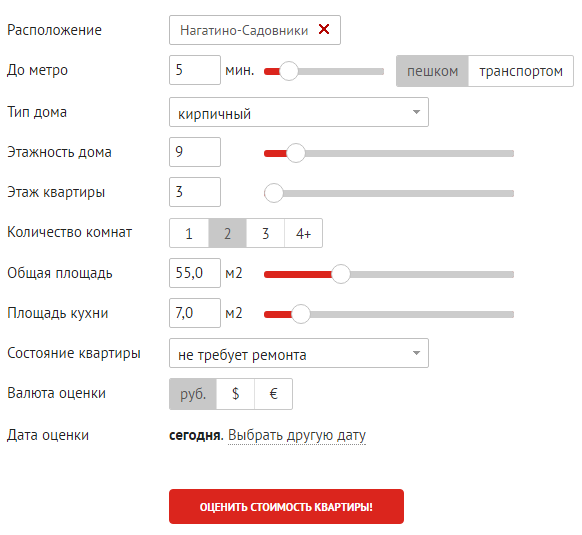
1. Additional information about the house and the district.

Despite the high accuracy in most cases, Cian can’t predict some sort of apartment because of the lack of relevant data to build KNN near the estimated apartment. This could be considered as one of the weaknesses of this model and decrease average model accuracy.

### 1.2.2 IRN.RU

The IRN.RU online calculator was developed 11 years ago and during this time more than 9.7 million of apartment were appraised online.

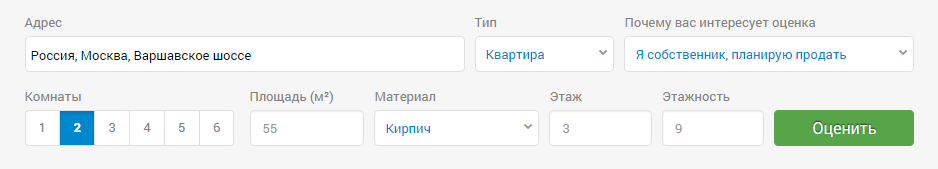
The calculator features needed are following: location (subway station or district), distance to subway (on foot or by transport), house type, floors, apartment’s floor, number of rooms, total area, kitchen area, repair condition. Methods of evaluation are based on multidimensional matrices indexes of the cost of housing in Moscow and the Moscow region, as well as matrices of valuation adjustments. Real estate appraisal in such a way is an extension of the method of comparative analysis of sales. [[3]](#footnote-3)



1. Apartment appraisal IRN.ru interface

### 1.2.3 Domofond.ru

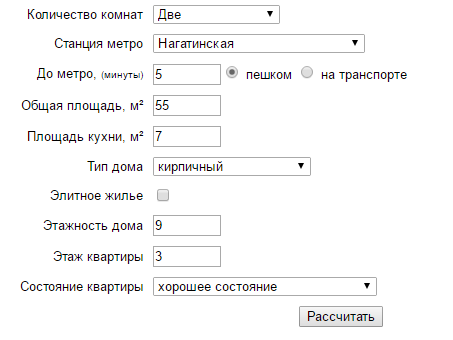
Domofond online calculator is based on prices of ads posted in Domofond.ru and Avito.ru. The average price is calculated taking into account the apartment and apartments in the neighborhood. The closer apartment is, the higher its weight in the formula will be. The service provide only qualitative estimation of price appraisal accuracy like “very good”, “good”, “need more information”, so one cannot say how well this model can predict. Also one of the outstanding features for this service is the customer’s role: owner (sales or rent purpose), realtor or just curious Internet user. To have a price appraisal, one should input the email address and automatically subscribe to Domofond customers database, so it could be considered as one of the ways to engage new customers and collect information about their possible interests.[[4]](#footnote-4)



1. Apartment appraisal Domofond.ru interface

### 1.2.4 Ocenchik.ru

The Ocenchik.ru online calculator has only beta-version and its features are following: location (subway station), distance to subway (on foot or by transport), house type, floors, apartment’s floor, rooms amount, total area, kitchen area, repair condition and check “is it elite real estate”.[[5]](#footnote-5)

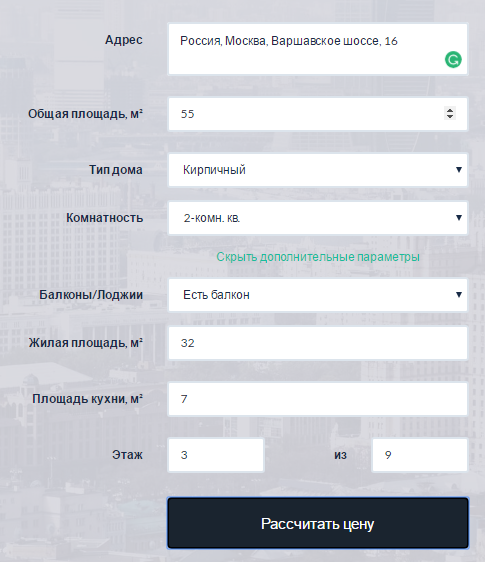


1. Apartment appraisal Ocenchik.ru interface

The model features are quite standard, but does not seems relevant because customer by his- or herself inputs the check “is it luxe real estate?’. In addition, there is the second point for discussion: the repair condition as feature: it does not have clear difference between bad, good repair or renovation. So it absolutely subjective feature.

### 1.2.5 Flatorial

Besides the services provided by huge real estate companies, in Russia there is a startup named Flatorial, the main goal of which is real estate price prediction using cutting-edge machine leaning methods.[[6]](#footnote-6)



1. Apartment appraisal Flatorial.ru interface

### 1.2.5 Comparison results

As one can see the difference between the highest and lowest appraisals is ~2 million rubles (or ~33 thousands euro) that is really huge possible price corridor. Even if one compares only Cian, IRN and Domofond, 1 million rubles also matters.

Table 2. Comparison of online real estate calculator appraisals

|  |  |  |
| --- | --- | --- |
| **Online calculator name** | **Apartment price (rub)** | **Price corridor (rub)** |
| Cian.ru | 10 019 295  10 219 681\* | 9 215 030 - 10 708 060  9 399 331 - 10 922 221\* |
| IRN.RU | 10 220 047 | 9 709 000 - 10 731 000 |
| Domofond.ru | 9 354 962 | 7 722 793 - 10 987 130 |
| Ocenchik.ru | 8 231 098 | 7 901 854 - 8 560 342 |
| Flatorial | 10 894 000 | 9 484 000 – 12 303 000 |

\*after input advanced parameters.

As one possible ways to make this comparison more illustrative, one can estimate the real estate price using professional realtor methodology and by apartment use one of the apartments that was as ad in Yandex Real Estate to make all side comparison. But the thing is there is no transparent apartment online price prediction service: every service gives quite different predictions without any appropriate analysis or just explanation how this prices were reached. In this terms, the customer should 100% trust the service and don’t even go deep into details.

## 1.3 Proposed machine learning based models

In this subchapter, several machine learning based models are considering for their ability to be implemented to thesis research problem. Firstly, the main priority is given to Kaggle competitions solutions because Kaggle is the biggest data science community where in every competition one can find interesting and outstanding examples of machine learning implementation to business issue. However, in R&D data science world many papers were published, so other goal is to process the most interesting of them and make a summary of methods used.

### 1.3.1 Kaggle competitions and other approaches

There were several Kaggle competitions regarding to real estate price prediction. Especially in Russia there is increasing demand of such price prediction models: during the thesis writing two data science competitions from the biggest Russian bank, Sberbank, have been launched. The first one was held in the end of 2016 and the winner used some kind of combination of neural networks[[7]](#footnote-7). The second competition is currently running on the Kaggle platform[[8]](#footnote-8)

Before that the Kaggle data science community had created another competition about advanced regression techniques[[9]](#footnote-9). During the model development I consider several solutions to this competition as resource to understand the things right and kind of benchmark whit I need to follow. [[10]](#footnote-10)[[11]](#footnote-11)[[12]](#footnote-12)[[13]](#footnote-13)

### 1.3.2 Scientific papers published in Google Scholar

As machine learning approach popularity is constantly growing, there are many research papers covering different areas of implementation. In Google Scholar, the hugest search engine for R&D practices. I searched for papers and found some catching-eye papers that one should definitely have a look[[14]](#footnote-14).

One of the papers that one should be aware of is “Application of Fuzzy Neural Network for Real Estate Prediction”.[[15]](#footnote-15) As for advanced machine learning methods, “Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal” covering three ML approaches that seemed cutting-edge and “solution for everything” during last several years.

In Russia there is no cases of real estate pricing model that uses machine learning in any state, but there is one example in New Zealand[[16]](#footnote-16). Moreover, I paid attention to such papers like “A dynamic mode of housing price determination”[[17]](#footnote-17), “discovering the hidden structure of house prices with a non-parametric latent manifold model”[[18]](#footnote-18) and “The research on price prediction of second-hand houses based on KNN and stimulated annealing algorithm”[[19]](#footnote-19)

Moreover, I searched for the benchmark solutions on professional forum discussions, massive open online courses etc. and gain quite of expertise which model are the most suitable to try.

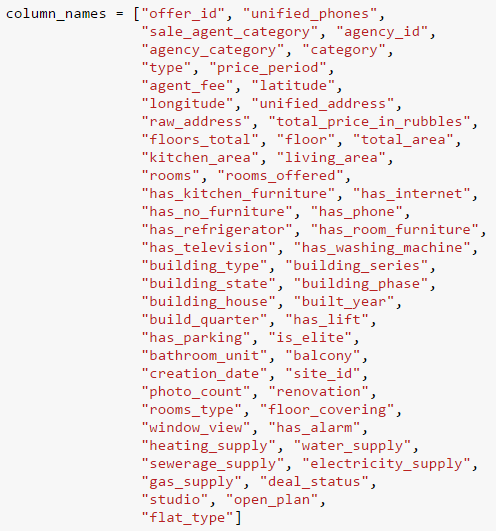
To sum up all scientific papers related to this topic, I can say following: there is no universal problem-solving approach for this type of issue. The results are strongly depends on the dataset volume and sometimes the combination of different machine learning models can bring the better outcome.

# Chapter 2. Model development

## 2.1 Dataset description

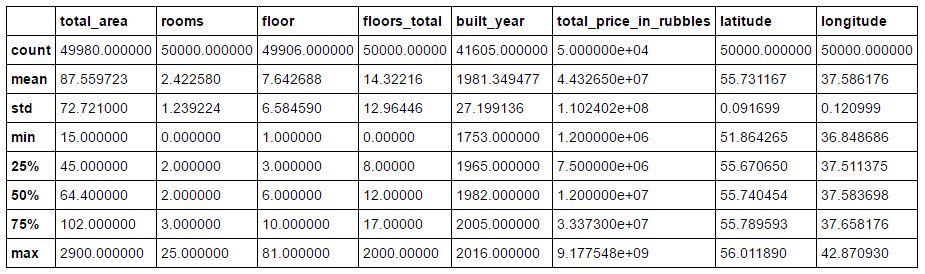
I was given the cast of real-time data from Yandex Real Estate database dated October 2016. First research phase (clear data processing, adding new static features to the dataset and implement a standard machine learning models to estimate how accuracy clear data could possibly be) was on October dataset and then in the second phase (comparison analysis in real-time data, deleting anomalies and implementing advanced features to dataset using different API and external information).

The dataset contains 50000 ads with all information inside. From ad to ad it depends on what features concrete ad will have (because every ad was customer-generated content) and in most cases some features left blank. That is why not every machine learning models is it possible to implement (for instance, for some regression one needs more information about the apartment or one could use less features that causes less model quality). The start-pack features were following:



1. Dataset start features

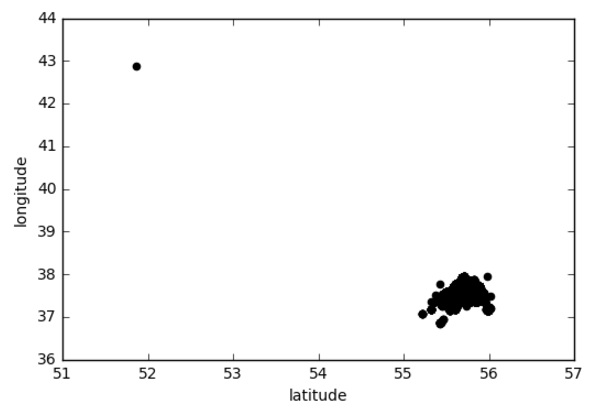
After analyzing the features that exist in dataset, the data processing phase begins. Firstly, let’s run the describe function to understand the data distribution (Fig. 8). As one can see, the average apartment in this dataset has 2 or 3 rooms with 87 m sq feet and was built in early 80s previous century.



1. Dataset key quick statistics description

Secondly, as main features in dataset are «total area» and «total price in rubbles», all ads which does not have this information (for any reason) were deleted from dataset (20 ads in summary). As final dataset, there were 49980 ads to process and to create machine learning models.

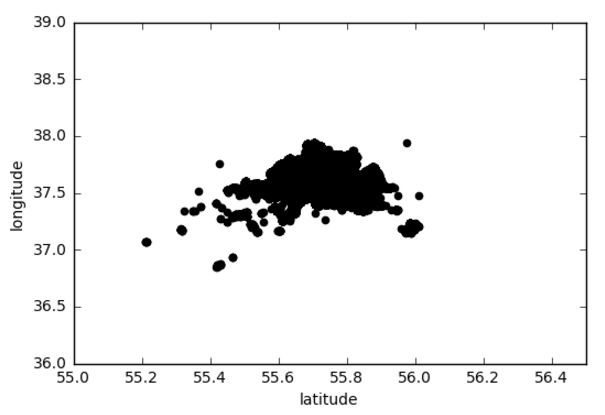
In Moscow there is high sensitivity to district, as it was mention before in realtor methodology description, I draw plot to understand is there concentrated data or it distributed throw districts.



1. Latitude-longitude data distribution

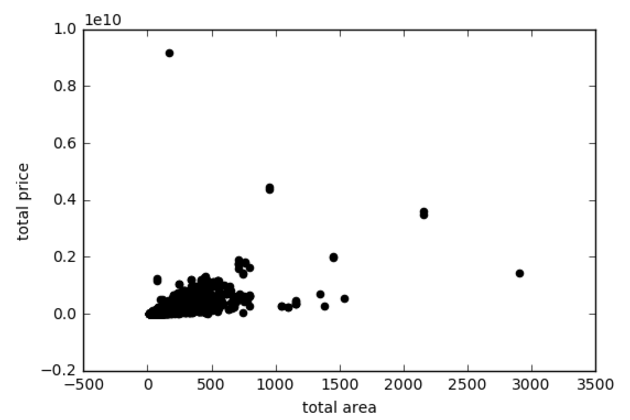
As one can see, there is a statistics anomaly in up-left corner. However, there could be a part of district, so it is really early (before real model development), establish that it is the anomaly and it should be deleted from the dataset.

Then I had a closer look at concentrated data and understood that the data distribution is uniform and models should be working for every district because as one of machine learning advantages, quite few data need to teach the model right, and if one have enough data for every district, it would work with high accuracy (see Fig. 10).



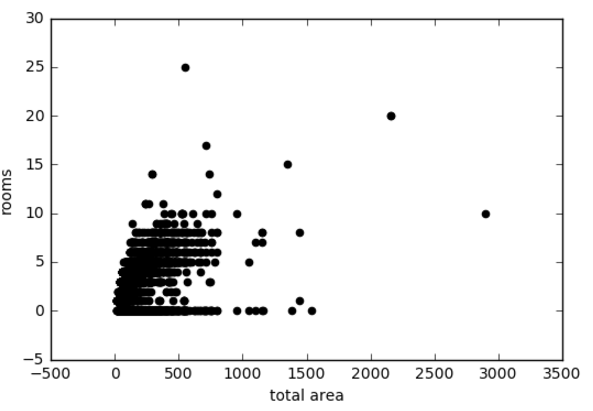
1. Latitude-longitude concentrated data distribution

Next thing that I was interesting in is the total area vs. total price distribution (apartments of what category in dataset there is more than the others categories) (see Fig. 11). Fortunately, there is no extra ordinary anomalies that should be deleted immediately and most apartments are in second and third house categories (as one of research hypothesis).



1. Total area – total price data distribution

In addition, the last distribution that seemed interesting for further research was the total area vs rooms (see Fig. 12). As one can see, in 90 per cent cases in dataset is studio or 2-rooms apartment which is quite relevant to customers’ profiles. Nevertheless, there are extremely big apartments (maybe they are irrelevant because data could be wrong due to user-base content generation). These apartments could be useful to create approaches to appraise such type of apartments, but from another point of view, this data with high probability could be the “data noise”.



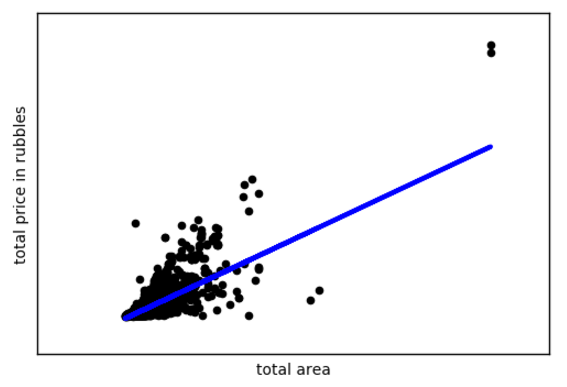
1. Total area – rooms data distribution

## 2.2 Creating models with clear data

After data processing, I developed few linear regression models as one of the standard models used for price prediction. The default optimization parameter is score (the same as accuracy) which calculated using r2\_score function. It is very similar to mean\_squarred\_error (MSE), but it displays the answer regarding the scope and is a universal compared with a naive algorithm, which would predict only average (it would have scored r2\_score = 0). If our method is worse, score < 0, then we need to change something.

For training and testing we will use the cross validation. Cross validation is the method to understand the accuracy involving all data in training and testing sessions. For instance, one divides data to 5 datasets and for training one uses 3 of them, and for testing other 2. Combining the all five in different variants, one has the most fulfil model which covering all data.

As first model I developed simple linear regression with one feature (meter sq feet). The CV score was 0.5528, which means that with 55.28% chance model estimates the price right. See Fig. 13



1. Linear regression with meter sq feet as a feature

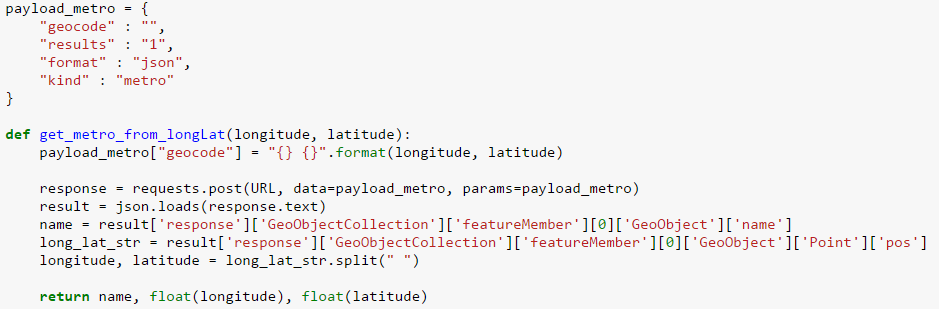
Then I added as a feature the number of rooms. The CV score was 0.5329, so the model became worse. After adding the floor with floors total and year of construction, the model became worse once again, and CV score became 0.4587.

So, the results was quite strange for machine learning models accuracy: as a standard practice, after adding new features and data, the model has to increase its accuracy, but it does not. The next step was the features normalization. And after this the CV score became quite higher 0.4983, but it does not seem like this model could be considered as efficient one to solve business-related problem.

## 2.3 Adding new features to dataset

As new feature, I decided to add a district affiliation. To make it in right and efficient way, I downloaded a JSON file with all districts’ coordinates in polygon format and then developed an algorithm which returns Boolean parameter is this house affiliates this district or not. This feature should increase the accuracy because the appraisal process is really sensitive to districts and their prestige rate.

Another feature that was successfully added was the closest subway station with its coordinates using Yandex Geocoder API. The main extraction function is shown on Fig. 14.



1. Closest subway station extraction using Yandex Geocoder API.

After adding these highly important geographic features, I ran all models from machine learning map that seemed relevant to the task[[20]](#footnote-20): Linear Regression, Lasso, Elastic Net, SVR with linear kernel and rbf kernel, NN 10 (neural network with MLP Regressor), Decision Tree (including Decision Tree Regressor) and Random Forest (including Random Forest Regressor).

As key features that could not be blank, I chose “total area”, “rooms”, “floor”, “floors\_total”, “built\_year”, “has lift” and “kitchen\_area”. After deleting irrelevant data items, dataset size decreased to 35435 ads that it is still enough to create a machine learning model.

Table 3. Multi-model run results

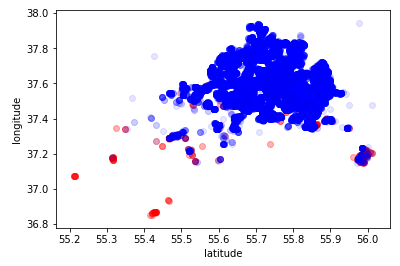
|  |  |
| --- | --- |
| **Model name** | **CV score** |
| Linear Regression | 0.5083063089345885 +- 0.2110353369021043 |
| Lasso | 0.5083073578756937 +- 0.2110357674226658 |
| Elastic Net | -0.007835631280501177 +- 0.007830536707496285 |
| SVR (linear kernel) | 0.23516213894707363 +- 0.10619937629053769 |
| SVR (rbf kernel) | -0.0606058262572569 +- 0.022290465892775623 |
| NN 10 | -0.034969973219004614 +- 0.017037993371217017 |
| Decision Tree | -1.4898002788496403 +- 2.596128495343518 |
| Random Forest | 0.2324390079237034 +- 0.3215922453400611 |

So the results did not look optimistic yet. The next step is to detect anomalies in the dataset and understand the nature of such a big model mistake.

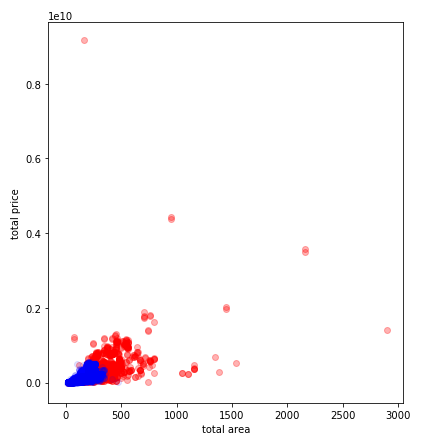
## 2.4 Anomalies detection

In this dataset there are anomalies that are necessary to detect and delete from the dataset to get more precise model in terms of accuracy.

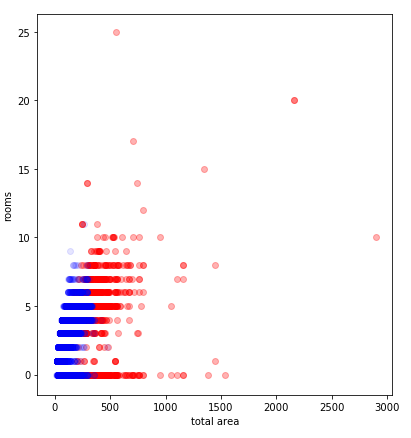
Firstly, let’s delete from the dataset the ads where floor\_total is null because it weird and can bring a problems in the future. For anomalies detection using machine learning methods we use IsolationForest from sklearn.ensemble. Let’s have a look on data distribution graphics that we considered previously.



1. Latitude vs. longitude dataset anomalies detection

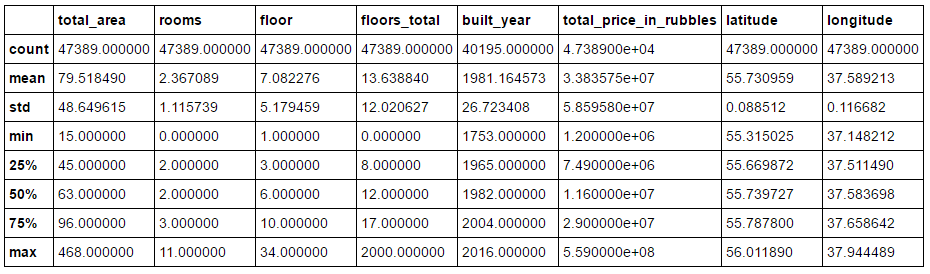


1. Total area vs. total price dataset anomalies detection



1. Total area vs rooms dataset anomalies detection.

As one can see, the anomalies are pretty likely to be detected. Now it is the time to have a look on updated statistics description (see Fig. 18).



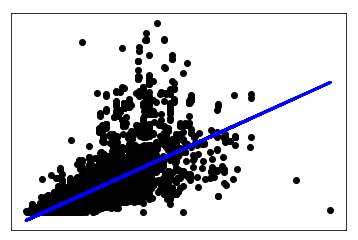
1. Updated dataset statistics description.

## 2.5 Re-running previous models on updated dataset

Now it is the time to re-run the models and understand to what extent the accuracy has increased (see Table 3)

Table 4. Comparison Linear regressions on old and updated datasets.

|  |  |  |
| --- | --- | --- |
| Linear model features | Original dataset | Processed dataset |
| Meter sq feet | 0.5528 | 0.5840 |
| Meter sq feet + rooms | 0.5329 | 0.5928 |
| Meter sq feet + rooms + floor | NA | 0.6235 |
| Meter sq feet + rooms + floor + year | 0.4587 | 0.6227 |
| Meter sq feet + rooms + floor + year (normalized) | 0.4985 | 0.6150 |
| Meter sq feet + rooms + floor + year (normalized + relative floor) | NA | 0.6147 |



1. Total area vs total price linear regression model

As one can noticed, the regression models have been improved quite well. Now let’s have a look on features importance:

total\_area: 0.136724789255

rooms: 1.0

floor: 0.0146520322805

floors\_total: 0.180951231351

built\_year: 0.225325349857

relative\_floor: 0.207599811711

After closer consideration, we can notice that the number of the rooms has the highest weight and the lowest priority has the floor number. This could be interpreted as when it comes to price estimation, the number of rooms sets the apartment category, while the floor number does not actually 100% correspond with future possible price.

Let’s re-run the multi-models and compare the results with original data before anomalies detection (see Table 5)

Table 5. Multi-model re-run results

|  |  |
| --- | --- |
| **Model name** | **CV score** |
| Linear Regression | 0.6291 +- 0.0354 |
| Lasso | 0.6291 +- 0.0354 |
| Elastic Net | -0.0058 +- 0.0050 |
| SVR (linear kernel) | 0.2807 +- 0.0354 |
| SVR (rbf kernel) | -0.1318 +- 0.0229 |
| NN 10 | 0.6652 +- 0.0338 |
| Decision Tree | 0.5034 +- 0.0386 |
| Random Forest | 0.6555 +- 0.0468 |

## 2.6 Implementing grid search and XGBoost models

After models re-running, I tried grid search methods (it automatically search the ideal model’s parameters giving the highest accuracy) and XGBboost as one of the most powerful model in machine learning nowadays. As it was predicted, the grid search using the MLPRegressor gives the 0.6591 accuracy that it the highest accuracy at this point.

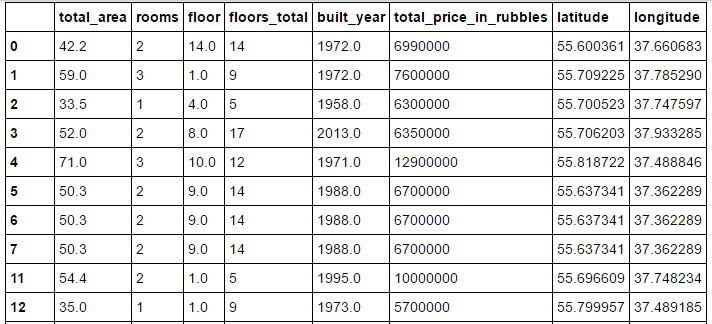
Let’s try the XGBoost with five main features like “total area”, “rooms”, “floor”, “floors total” and “built year”. It gave us 0.7487 score. At this time this model is the best one, but we need to take into account the dataset subjectivity: the apartment cost was settled by customers and it brings a subjectivity to the model. Also this score looks fine because it describe the whole model, not unique estimation as the competitors do. Later on, in the Chapter 4 I’ll consider ways to increase the model accuracy with further model development.

# Chapter 3. Findings and conclusion

During this research I set for me goal not only machine learning development, but also understanding the key principles of machine learning and its necessity to the business. As the first part of the research, I paid a lot of attention to existing machine learning methods and working principles, so now I’m able to solve this type of issues and continue my professional degree as data scientist or even business analyst.

As for model, I ran several models: different linear and non-linear regressions, multiple models such as lasso, elastic net, SVR with different kernels, random forest, decision tree, neural networks, use grid search method to choose the right model parameters and ran xgboost, and currently it is the best model in terms of accuracy. Also now I know how one can efficiently clear data from anomalies which can bring significant error during the research. It was predicted before, that the highest accuracy score will be models like neural networks, random forests and xgboost. The room for model improvement is considered in the next chapter.

To be more precise in model accuracy, let’s compare random 10 apartments with their values in original dataset, model value and competitors’ responses. This comparison can give us the background of model accuracy not only in terms of model development in IPython Notebook, but also from market perspective: to what extent this model seems competitive.



1. Randomly chosen apartment for further comparison

Table 6. Apartments’ prices comparison through different models (RUB)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Original price | Model price | Cian | IRN | Domofond | Flatorial |
| 1 | 6 990 000 | 6 990 452 | 7 001 806 | 6 960 945 | 6 705 985 | 7 105 835 |
| 2 | 7 600 000 | 7 500 340 | 7 630 780 | 7 891 034 | 8 100 678 | 7 906 893 |
| 3 | 6 300 000 | 6 700 010 | 6 089 000 | 6 466 801 | NA | 6 502 829 |
| 4 | 6 350 000 | 6 250 090 | 6 370 059 | 6 406 891 | 6 201 984 | 6 209 985 |
| 5 | 12 900 000 | 11 900 567 | 13 104 901 | 12 750 945 | NA | 13 206 945 |
| 6 | 6 700 000 | 6 802 458 | 7 004 982 | 7 105 340 | 6 90410 | 6 850 013 |
| 7 | 6 700 000 | 6 700 340 | 6 510 842 | 6 490 397 | 6 903 000 | 6 578 923 |
| 8 | 6 700 000 | 6 700 000 | 6 620 934 | 6 800 892 | 6 569 450 | 6 598 249 |
| 9 | 10 000 000 | 9 640 060 | NA | 10 256 960 | NA | 10 106 052 |
| 10 | 5 700 000 | 5 670 080 | 6 005 950 | 5 800 397 | 6 106 400 | 5 904 0060 |

As one can noticed, the model developed gives us pretty the same price range as competitors do. At this point, the model is not worse than existing one, but of course if one would like to improve this model with deep analysis provided to the customer and dynamic features consideration that would definitely increase the model accuracy because it will include not only parametric given values, but also meta information collected from the open sources.

Currently, model has 0.75 accuracy, and this is significant result if we will remember that this accuracy is true for all dataset, not only for concrete apartment. Another point is that due to different machine learning models implemented and different internal datasets, some competitors can’t predict prices for concrete apartments because of lack of data.

In the first chapter there I provided the theoretical background such as existing competitors’ models and scientific papers which covers price prediction in the real estate. Second chapter was dedicated to practical model development: with original clear data, then anomalies detection, then re-running models and choosing optimal parameters to increase the accuracy. In the last, fourth chapter, further recommendations regarding the model improvement provided.

# Chapter 4. Further research and development

## 4.1 Decreasing subjectivity

Even if due to machine learning approach model could decrease subjectivity from realtor side, it cannot eliminate it from customer side: when one inputs apartments’ characteristics (for instance, distance to subway station on foot or repair condition), one make subjective estimation. And this type of issue can be solved by adding new features that would be automatically counted (distance to subway station on foot, by car and by bus using Yandex Maps services) and implementing apartment images analysis using image recognition and machine learning to understand the repair condition (is it renovation or just very good repair).

Another point that important to mention is the proximity of the houses with quite opposite price categories: for instance, if near luxury properties there is old housing stock, it could decrease the apartment price because of neighborhood that does not correspond to luxury category. These criteria can be processed by adding advanced district analysis containing not only municipal district name, but also the full description of neighborhood (what is main house type, how many new houses have been built during last ten years etc.).

## 4.2 Adding dynamic features

Adding this type of features are relevant even if there is little data density in neighborhood. For instance, if the considered apartment is located in new district where there is no enough (~20 similar ads) data in next buildings, one needs extra external data regarding to the district quality (is there gardens and shopping malls, how far they are from the building) and governmental programs (will be in near future in this district new subway station or high-speed road, or maybe this building will be reconstructed or even demolished soon). But if there is enough similar ads all this information will be in the model as meta-information from the ads.

## 4.3 Dealing with small dataset

If one want to implement the similar model in different city or town in Russia, with high probability, there will be a problem related to small dataset which is vital to machine learning. When it comes to that question, one of possible solutions could be implementation of mixed approach: the algorithmic (from professional realtors’ indexes) and machine learning in some concrete parts and/ or features. Also as another approach, one could compare the real estate prices in the city X with the prices in the city Y (where there is enough data and transparent model) using economic indicators of generated profits, GPD per person etc.

## 4.4 Using advanced machine learning models and techniques

In this thesis, I used the basic machine learning methods without implementing my own machine learning libraries or tensor flow opportunities at least. Nowadays machine learning techniques are evolving every day, so I’m pretty sure that in one year (or maybe less) there will be a huge variety of methods and libraries that will be appropriate to use for solve this price prediction issue.

Also nowadays the idea of machine learning ensembles increasing in its popularity. For instance, one could try random forest and xgboost or neural networks with decision trees. Everything that could match the dataset and the final research goal.

# List of references

* Machine Learning Algorithms for Classification [Electronic resource]/ Princeton University [United States]. URL: <http://www.cs.princeton.edu/~schapire/talks/picasso-minicourse.pdf>
* Cian online price estimation [Electronic resource]/ Cian.ru site. URL: <https://goo.gl/Wwymj7> (date of access 24.04.2017)
* IRN online price estimation [Electronic resource]/ IRN.ru site. URL: <https://goo.gl/1kW9qP> (date of access: 24.04.2017)
* Domofond online price estimation [Electronic resource]/ Domofond.ru site. URL: <http://www.domofond.ru/ozenka> (date of access: 24.04.2017)
* Ocenchik online price estimation [Electronic resource]/ Ocenchik.ru site. URL: [www.ocenchik.ru/ocenka.kvartiry/1492985132/](http://www.ocenchik.ru/ocenka.kvartiry/1492985132/) (date of access 24.04.2017)
* Flatorial online price estimation [Electronic resource]/ Flatorial.ru site. URL: <http://flatorial.ru/calc/> (date of access 01.05.2017)

Data science competition announcement [Electronic resource]/ Boosters.pro [Russia, 2016]. URL: <https://boosters.pro/news/2> (date of access 01.05.2017)

Data science competition description [Electronic resource]/ Kaggle.com [United States, 2017]. URL: <https://www.kaggle.com/c/sberbank-russian-housing-market> (date of access 10.05.2017)

* Data science completion “Advanced regression techniques in the house prices” [Electronic resource]/ Kaggle.com [United States, 2016]. URL: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques> (date of access 10.10.2016)
* Solution to the Kaggle competition [Electronic resource]/ Github.com. URL: <https://github.com/duttashi/House-Price-Prediction> (date of access 15.10.2016)
* Solution to the Kaggle competition [Electronic resource]/ Github.com. URL: <https://github.com/rehassachdeva/House-Prices-Advanced-Regression-Techniques---Kaggle-Competition> (date of access 15.10.2016)
* Solution to the Kaggle competition [Electronic resource]/ Github.com. URL: <https://github.com/NischalBhatewara/Ames-housing-prices> (date of access 15.10.2016)
* Solution to the Kaggle competition [Electronic resource]/ Github.com. URL: <https://github.com/Shitao/Kaggle-House-Prices-Advanced-Regression-Techniques> (date of access 15.10.2016)
* Google Scholar Search for Scientific papers related to the topic [Electronic resource]/ Scholar.google.com [United States], URL: <https://scholar.google.ru/scholar?q=price+prediction+real+estate+machine+learning&btnG=&hl=ru&as_sdt=0%2C5&oq=price+prediction+real+estate+machine+lea> (date of access 19.12.2016)
* Jian-Guo Liu, Xiao-Li Zhang, Wei-Ping Wu “Application of Fuzzy Neural Networks for Real Estate Prediction” [Electronic resource]/ International Symposium of Neural Networks. URL: <https://link.springer.com/chapter/10.1007%2F11760191_173?LI=true> (date of access 05.01.2017)
* “House prediction: hedonic price model vs. artificial neural network” [Electronic resource]/ Lincoln University [New Zealand]. URL: <https://researcharchive.lincoln.ac.nz/handle/10182/5198> (date of access: 20.12.2016)
* Engle Robert F, Lilien David M, Watson Mark. A dynamic model of   
  housing price determination // Journal of Econometrics. –– 1985. –– Vol. 28, no. 3. –– P. 307–326.
* Discovering the hidden structure of house prices with a non-parametric latent manifold model / Sumit Chopra, Trivikraman Thampy, John Leahy et al. // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007. –– 2007. –– P. 173– 182. –– URL: <http://doi.acm.org/10.1145/1281192.1281214>. (date of access: 03.05.2017)
* Zhao Weikun, Sun Cao, Wang Ji. The Research on Price Prediction of Second-hand houses based on KNN and Stimulated Annealing Algorithm // International Journal of Smart Home. –– 2014. –– Vol. 8,  no. 2. –– P. 191–200
* Scikit-learn. –– 2017. –– May. –– URL: <http://scikit-learn.org/stable/>
* Scipy. –– 2017. –– May. –– URL: <https://www.scipy.org/>.
* Shabalin V. – «Real estate deals on first- and second-hand markets»// Omega-L – 2012 – 768p.
* Cheng Li «A Gentle Introduction to Gradient Boosting»/ College of Computer and Information Science Northeastern University [United States, 2016]. URL: <http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf> (date of access: 25.04.2017)

# Appendix A

As advanced feature, I’ve added to every apartment the district affiliation with prestige rate. Moscow is really sensitive to its municipal districts and the price could extremely volatile from one district to another although the apartments’ features could be the same or pretend to be the same. This district prestige rating was published in Shabalin book “Real estate deals in the Russian market” and could be considered as the source of truth because of author’s expertise.

The scale of prestige of 10 administrative districts in Moscow:

1. Tsentral'nyy
2. Yugo-Zapadnyy
3. Zapadnyy
4. Severo-Zapadnyy
5. Severo-Vostochnyy
6. Severnyy
7. Vostochnyy
8. Yuzhnyy
9. Yugo-Vostochnyy
10. Zelenogradskiy

The scale of prestige of all 120 of the municipal districts in Moscow.

Table 7. The scale of prestige for municipal districts

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Place | Municipal (RU) | Administrative (RU) | Municipal (ENG) | Administrative (ENG) |
| 1 | Арбат | Центральный | Arbat | Tsentral'nyy |
| 2 | Хамовники | Центральный | Khamovniki | Tsentral'nyy |
| 3 | Тверской | Центральный | Tverskoy | Tsentral'nyy |
| 4 | Пресненский | Центральный | Presnenskiy | Tsentral'nyy |
| 5 | Якиманка | Центральный | Yakimanka | Tsentral'nyy |
| 6 | Замоскоречье | Центральный | Zamosedch'ye | Tsentral'nyy |
| 7 | Басманный | Центральный | Basmannyy | Tsentral'nyy |
| 8 | Мещанский | Центральный | Meshchanskiy | Tsentral'nyy |
| 9 | Таганский | Центральный | Taganskiy yazyk | Tsentral'nyy |
| 10 | Дорогомилово | Западный | Dorogomilovo | Zapadnyy |
| 11 | Гагаринский | Юго-Западный | Gagarinskiy | Yugo-Zapadnyy |
| 12 | Красносельский | Центральный | Krasnosel'skiy | Tsentral'nyy |
| 13 | Сокол | Северный | Sokol | Severnyy |
| 14 | Академический | Юго-Западный | Akademicheskiy | Yugo-Zapadnyy |
| 15 | Ломоносовский | Юго-Западный | Lomonosovskiy | Yugo-Zapadnyy |
| 16 | Крылатское | Западный | Krylatskoye | Zapadnyy |
| 17 | Аэропорт | Северный | Aeroport | Severnyy |
| 18 | Донской | Южный | Donskoy | Yuzhnyy |
| 19 | Черемушки | Юго-Западный | Cheremushki | Yugo-Zapadnyy |
| 20 | Беговой | Северный | Begovoy | Severnyy |
| 21 | Кунцево | Западный | Kuntsevo | Zapadnyy |
| 22 | Раменки | Западный | Ramenki | Zapadnyy |
| 23 | Фили-Давыдково | Западный | Fili-Davydkovo | Zapadnyy |
| 24 | Сокольники | Восточный | Sokol'niki | Vostochnyy |
| 25 | Даниловский | Южный | Danilovskiy | Yuzhnyy |
| 26 | Строгино | Северо-Западный | Strogino | Severo-Zapadnyy |
| 27 | Пр-т Вернадского | Западный | Pr-t Vernadskogo | Zapadnyy |
| 28 | Хорошевский | Северный | Khoroshevskiy | Severnyy |
| 29 | Филевский парк | Западный | Filevskiy park | Zapadnyy |
| 30 | Савеловский | Северный | Savelovskiy | Severnyy |
| 31 | Коньково | Юго-Западный | Kon'kovo | Yugo-Zapadnyy |
| 32 | Обручевский | Юго-Западный | Obruchevskiy | Yugo-Zapadnyy |
| 33 | Ясенево | Юго-Западный | Yasenevo | Yugo-Zapadnyy |
| 34 | Преображенское | Восточный | Preobrazhenskoye | Vostochnyy |
| 35 | Котловка | Юго-Западный | Kotlovka | Yugo-Zapadnyy |
| 36 | Можайский | Западный | Mozhayskiy | Zapadnyy |
| 37 | Алексеевский | Северо-Восточный | Alekseyevskiy | Severo-Vostochnyy |
| 38 | Соколиная гора | Восточный | Sokolinaya gora | Vostochnyy |
| 39 | Останкинский | Северо-Восточный | Ostankinskiy | Severo-Vostochnyy |
| 40 | Очаково-Матвеевское | Западный | Ochakovo-Matveyevskoye | Zapadnyy |
| 41 | Тропарево-Никулино | Западный | Troparevo-Nikulino | Zapadnyy |
| 42 | Теплый Стан | Юго-Западный | Teplyy Stan | Yugo-Zapadnyy |
| 43 | Зюзино | Юго-Западный | Zyuzino | Yugo-Zapadnyy |
| 44 | Чертаново Северное | Южный | Chertanovo Severnoye | Yuzhnyy |
| 45 | Лефортово | Юго-Восточный | Lefortovo | Yugo-Vostochnyy |
| 46 | Марьина роща | Северо-Восточный | Mar'ina roshcha | Severo-Vostochnyy |
| 47 | Северное Измайлово | Восточный | Severnoye Izmaylovo | Vostochnyy |
| 48 | Хорошево-Мневники | Северо-Западный | Khoroshevo-Mnevniki | Severo-Zapadnyy |
| 49 | Щукино | Северо-Западный | Shchukino | Severo-Zapadnyy |
| 50 | Измайлово | Восточный | Izmaylovo | Vostochnyy |
| 51 | Войковский | Северный | Voykovskiy | Severnyy |
| 52 | Свиблово | Северо-Восточный | Sviblovo | Severo-Vostochnyy |
| 53 | Покровское-Стрешнево | Северо-Западный | Pokrovskoye-Streshnevo | Severo-Zapadnyy |
| 54 | Чертаново Центральное | Южный | Chertanovo Tsentral'noye | Yuzhnyy |
| 55 | Восточное Измайлово | Восточный | Vostochnoye Izmaylovo | Vostochnyy |
| 56 | Богородское | Восточный | Bogorodskoye | Vostochnyy |
| 57 | Бутырский | Северо-Восточный | Butyrskiy | Severo-Vostochnyy |
| 58 | Тимирязевский | Северный | Timiryazevskiy | Severnyy |
| 59 | Марфино | Северо-Восточный | Marfin | Severo-Vostochnyy |
| 60 | Бабушкинский | Северо-Восточный | Babushkinskiy | Severo-Vostochnyy |
| 61 | Нижегородский | Юго-Восточный | Nizhegorodskiy | Yugo-Vostochnyy |
| 62 | Головинский | Северный | Golovinskiy | Severnyy |
| 63 | Отрадное | Северо-Восточный | Otradnoye | Severo-Vostochnyy |
| 64 | Ростокино | Северо-Восточный | Rostokino | Severo-Vostochnyy |
| 65 | Новогиреево | Восточный | Novogireyevo | Vostochnyy |
| 66 | Лерово | Восточный | Lerovo | Vostochnyy |
| 67 | Коптево | Северный | Koptevo | Severnyy |
| 68 | Южнопортовый | Юго-Восточный | Yuzhnoportovyy | Yugo-Vostochnyy |
| 69 | Левобережный | Северный | Levoberezhnyy | Severnyy |
| 70 | Чертаново Южное | Южный | Chertanovo Yuzhnoye | Yuzhnyy |
| 71 | Гольяново | Восточный | Gol'yanovo | Vostochnyy |
| 72 | Ивановское | Восточный | Ivanovskoye | Vostochnyy |
| 73 | Северное Тушино | Северо-Западный | Severnoye Tushino | Severo-Zapadnyy |
| 74 | Южное Щукино | Северо-Западный | Yuzhnoye Shchukino | Severo-Zapadnyy |
| 75 | Южное Медведково | Северо-Восточный | Yuzhnoye Medvedkovo | Severo-Vostochnyy |
| 76 | Алтуфьевский | Северо-Восточный | Altuf'yevskiy | Severo-Vostochnyy |
| 77 | Царицыно | Южный | Tsaritsyno | Yuzhnyy |
| 78 | Вешняки | Восточный | Veshnyaki | Vostochnyy |
| 79 | Нагатино-Сабурово | Южный | Nagatino-Saburovo | Yuzhnyy |
| 80 | Метрогородок | Восточный | Metrogorodok | Vostochnyy |
| 81 | Москворечье-Сабурово | Южный | Moskvorech'ye-Saburovo | Yuzhnyy |
| 82 | Ховрино | Северный | Khovrino | Severnyy |
| 83 | Бибирево | Северо-Восточный | Bibirevo | Severo-Vostochnyy |
| 84 | Северное Медведково | Северо-Восточный | Severnoye Medvedkovo | Severo-Vostochnyy |
| 85 | Лосиноостровский | Северо-Восточный | Losinoostrovskiy | Severo-Vostochnyy |
| 86 | Митино | Северо-Западный | Mitino | Severo-Zapadnyy |
| 87 | Северное Бутово | Юго-Западный | Severnoye Butovo | Yugo-Zapadnyy |
| 88 | Кузьминки | Юго-Восточный | Kuz'minki | Yugo-Vostochnyy |
| 89 | Лианозово | Северо-Восточный | Lianozovo | Severo-Vostochnyy |
| 90 | Бескудниковский | Северный | Beskudnikovskiy | Severnyy |
| 91 | Куркино | Северо-Западный | Kurkino | Severo-zapadnyy |
| 92 | Ярославский | Северо-Восточный | Yaroslavskiy | Severo-Vostochnyy |
| 93 | Нагатинский Затон | Южный | Nagatinskiy Zaton | Yuzhnyy |
| 94 | Выхино-Жулебино | Юго-Восточный | Vykhino-Zhulebino | Yugo-Vostochnyy |
| 95 | Солнцево | Западный | Solntsevo | Zapadnyy |
| 96 | Дмитровский | Северный | Dmitrovskiy | Severnyy |
| 97 | Западное Дегунино | Северный | Zapadnoye Degunino | Severnyy |
| 98 | Печатники | Юго-Восточный | Pechatniki | Yugo-Vostochnyy |
| 99 | Рязанский | Юго-Восточный | Ryazanskiy | Yugo-Vostochnyy |
| 100 | Новопеределкино | Западный | Novoperedelkino | Zapadnyy |
| 101 | Орехово-Борисово Северное | Южный | Orekhovo-Borisovo Severnoye | Yuzhnyy |
| 102 | Южное Бутово | Юго-Западный | Yuzhnoye Butovo | Yugo-Zapadnyy |
| 103 | Новокосино | Восточный | Novokosino | Vostochnyy |
| 104 | Бирюлево-Западное | Южный | Biryulevo-Zapadnoye | Yuzhnyy |
| 105 | Орехово-Борисово Южное | Южный | Orekhovo-Borisovo Yuzhnoye | Yuzhnyy |
| 106 | Текстильщики | Юго-Восточный | Tekstil'shchiki | Yugo-Vostochnyy |
| 107 | Восточное Дегунино | Северный | Vostochnoye Degunino | Severnyy |
| 108 | Забликово | Южный | Zablikovo | Yuzhnyy |
| 109 | Выхино-Жулебино | Юго-Восточный | Vykhino-Zhulebino | Yugo-Vostochnyy |
| 110 | Люблино | Юго-Восточный | Lyublino | Yugo-Vostochnyy |
| 111 | Внуково | Западный | Vnukovo | Zapadnyy |
| 112 | Косино-Ухтомский | Восточный | Kosino-Ukhtomskiy | Vostochnyy |
| 113 | Бирюлево-Восточное | Южный | Biryulevo-Vostochnoye | Yuzhnyy |
| 114 | Марьино | Юго-Восточный | Mar'ino | Yugo-Vostochnyy |
| 115 | Молжаниновский | Северный | Molzhaninovskiy | Severnyy |
| 116 | Поселок Восточный | Восточный | Poselok Vostochnyy | Vostochnyy |
| 117 | Северный пос. | Северо-Восточный | Severnyy pos. | Severo-Vostochnyy |
| 118 | Братеево | Южный | Brateyevo | Yuzhnyy |
| 119 | Поселок Некрасовка | Юго-Восточный | Poselok Nekrasovka | Yugo-Vostochnyy |
| 120 | Капотня | Юго-Восточный | Kapotnya | Yugo-Vostochnyy |

# Appendix B

Professional real estate price estimation contains several parameters due to which the apartment category can be determined. Here there is a table with all surcharges and discounts provided according to parameters values.

Table 8. Apartment features evaluation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameters | | Surcharge (%) | | Discount (%) | |
| 1 | Floor | 11 – 4-7 with elevator  12 | +0+1 | 011 – first  012 – second  013 – last, with elevator  014 – 4 without elevator  015 – 5 without elevator | -8+13  -0+1  -1+3  -5+7  -10+15 |
| 2 | Balcony  Loggia  Bay window | 21 - balcony  22 - loggia  23 – b + l  24 – bay window  25 | +0+2  +1+2  +2+3  +3+5 | 021 |  |
| 3 | Phone  Signalization | 31 – second number  32 - signalization  33 – media channels  34 | +3+5  +0+1  +0+1 | 031 – without phone  032 – doesn’t work  033 - shared  034 | -2+3  -0+1  -0+1 |
| 4 | Rooms | 41 - > 22 m sq feet  42 – for host  43 – not standard form  44 – two windows  45 | +1+2  +2+3  +0+1  +0+1 | 041 - shared  042 | -2+8 |
| 5 | Plumbing,  Equipment | 51 – foreign plumbing  52 – bathroom >= 4 m sq feet  53 – 2 bathroom units  54 | +1+2  +2+4  +3+5 | 051 – gas plumbing  052 – shared bathroom  053 | -0+1  -2+4 |
| 6 | Windows | 61 – all in the yard  62 – panoramic view  63 | +1+3  +2+5 | 061 – north side  062 – less ½ noise  063 – more ½ noise  064 – low level of isolation | -0+1  -3+5  -5+10  -1+3 |
| 7 | Distance to the subway | 71 - < 1 km (~10 min)  72 | +1+3 | 071 - > 2-3 km (3-5 bus stops)  072 – 6-10 bus stops  073 - > 10 bus stops  074 | -2+4  -5+7  -8+10 |
| 8 | Repair condition | 81 – after total repair  82 – after minor repair  83 – after euro repair  84 | +2+3  +5+8  +6+10 | 081 – total repair needed  082 – minor repair needed  083 – leakage  084 - unregistered redevelopment | -0+1  -3+5  -1+2  -3+10 |
| 9 | Yard condition | 91 – super clean  92 - watchman  93 – video recording  94 – parking zone  95 | +1+2  +1+2  +0+1  +2+5 | 091 – old elevator  092 – without minor repair  093 – super dirty  094 – many autos  095 | -0+1  -2+4  -0+1  -0+1 |

# Appendix C

In this appendix there is an example of the evaluation list used by professional realtors during real estate price evaluation.

Table 9. Apartment evaluation list

|  |  |
| --- | --- |
| **Basic apartment characteristics** | |
| The apartment address |  |
| Number of rooms |  |
| Total area |  |
| Living area |  |
| **Basic meter sq feet characteristics:** | |
| Floor |  |
| Floors total |  |
| House material (type) |  |
| Elevator presence |  |
| Garbage disposal presence |  |
| Kitchen area |  |
| Ceiling height |  |
| Interfloor overlapping |  |
| Interflat vestibule |  |
| Year built |  |
| Municipal district |  |
| **House category** |  |
| **Meter sq feet / USD** |  |
| **Basic quality characteristics:** | |
| Parameter | Surcharge / discount (%) |
| Floor |  |
| Balcony  Bay window |  |
| Phone  Signalization |  |
| Rooms |  |
| Plumbing,  Equipment |  |
| Windows |  |
| Distance to the subway |  |
| Repair condition |  |
| Yard condition |  |
| **Final meter sq feet / USD** |  |

1. Machine Learning Algorithms for Classification [Electronic resource]/ Princeton University [United States]. URL: <http://www.cs.princeton.edu/~schapire/talks/picasso-minicourse.pdf> [↑](#footnote-ref-1)
2. Cian online price estimation [Electronic resource]/ Cian.ru site. URL: <https://goo.gl/Wwymj7> (date of access 24.04.2017) [↑](#footnote-ref-2)
3. IRN online price estimation [Electronic resource]/ IRN.ru site. URL: <https://goo.gl/1kW9qP> (date of access: 24.04.2017) [↑](#footnote-ref-3)
4. Domofond online price estimation [Electronic resource]/ Domofond.ru site. URL: <http://www.domofond.ru/ozenka> (date of access: 24.04.2017) [↑](#footnote-ref-4)
5. Ocenchik online price estimation [Electronic resource]/ Ocenchik.ru site. URL: [www.ocenchik.ru/ocenka.kvartiry/1492985132/](http://www.ocenchik.ru/ocenka.kvartiry/1492985132/) (date of access 24.04.2017) [↑](#footnote-ref-5)
6. Flatorial online price estimation [Electronic resource]/ Flatorial.ru site. URL: <http://flatorial.ru/calc/> (date of access 01.05.2017) [↑](#footnote-ref-6)
7. Data science competition announcement [Electronic resource]/ Boosters.pro [Russia, 2016]. URL: <https://boosters.pro/news/2> (date of access 01.05.2017) [↑](#footnote-ref-7)
8. Data science competition description [Electronic resource]/ Kaggle.com [United States, 2017]. URL: <https://www.kaggle.com/c/sberbank-russian-housing-market> (date of access 10.05.2017) [↑](#footnote-ref-8)
9. Data science completion “Advanced regression techniques in the house prices” [Electronic resource]/ Kaggle.com [United States, 2016]. URL: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques> (date of access 10.10.2016) [↑](#footnote-ref-9)
10. Solution to the Kaggle competition [Electronic resource]/ Github.com. URL: <https://github.com/duttashi/House-Price-Prediction> (date of access 15.10.2016) [↑](#footnote-ref-10)
11. Solution to the Kaggle competition [Electronic resource]/ Github.com. URL: <https://github.com/rehassachdeva/House-Prices-Advanced-Regression-Techniques---Kaggle-Competition> (date of access 15.10.2016) [↑](#footnote-ref-11)
12. Solution to the Kaggle competition [Electronic resource]/ Github.com. URL: <https://github.com/NischalBhatewara/Ames-housing-prices> (date of access 15.10.2016) [↑](#footnote-ref-12)
13. Solution to the Kaggle competition [Electronic resource]/ Github.com. URL: <https://github.com/Shitao/Kaggle-House-Prices-Advanced-Regression-Techniques> (date of access 15.10.2016) [↑](#footnote-ref-13)
14. Google Scholar Search for Scientific papers related to the topic [Electronic resource]/ Scholar.google.com [United States], URL: <https://scholar.google.ru/scholar?q=price+prediction+real+estate+machine+learning&btnG=&hl=ru&as_sdt=0%2C5&oq=price+prediction+real+estate+machine+lea> (date of access 19.12.2016) [↑](#footnote-ref-14)
15. Jian-Guo Liu, Xiao-Li Zhang, Wei-Ping Wu “Application of Fuzzy Neural Networks for Real Estate Prediction” [Electronic resource]/ International Symposium of Neural Networks. URL: <https://link.springer.com/chapter/10.1007%2F11760191_173?LI=true> (date of access 05.01.2017) [↑](#footnote-ref-15)
16. “House prediction: hedonic price model vs. artificial neural network” [Electronic resource]/ Lincoln University [New Zealand]. URL: <https://researcharchive.lincoln.ac.nz/handle/10182/5198> (date of access: 20.12.2016) [↑](#footnote-ref-16)
17. Engle Robert F, Lilien David M, Watson Mark. A dynamic model of   
    housing price determination // Journal of Econometrics. –– 1985. –– Vol. 28, no. 3. –– P. 307–326. [↑](#footnote-ref-17)
18. Discovering the hidden structure of house prices with a non-parametric latent manifold model / Sumit Chopra, Trivikraman Thampy, John Leahy et al. // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007. –– 2007. –– P. 173– 182. –– URL: <http://doi.acm.org/10.1145/1281192.1281214>. (date of access: 03.05.2017) [↑](#footnote-ref-18)
19. Zhao Weikun, Sun Cao, Wang Ji. The Research on Price Prediction of Second-hand houses based on KNN and Stimulated Annealing Algorithm // International Journal of Smart Home. –– 2014. –– Vol. 8,  no. 2. –– P. 191–200 [↑](#footnote-ref-19)
20. Machine learning map [Electronic resource]/ Scikit-learn manual. URL: <http://scikit-learn.org/stable/tutorial/machine_learning_map/> (date of access: 05.02.2017) [↑](#footnote-ref-20)