МРНТИ 004.852
УДК 28.23.25

https://doi.org/10.51889/2981.2022.83.29.026

С.З. Сапакова[1]

[1]*Халықаралық Ақпараттық технологиялар университеті, Алматы қ., Қазақстан*
e-mail: sapakovasz@gmail.com

АЛМАТЫ ҚАЛАСЫНЫҢ ЖЫЛЖЫМАЙТЫН МҮЛІК НАРЫҒЫНДА МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНУ

Аңдатпа

Бұл жұмыста Алматы қаласының Әуезов және Бостандық аудандары бойынша жылжымайтын мүлік нарығы қарастырылды. Бұл нарық әлемдік экономиканың маңызды саласы екені белгілі. Қазақстанда тұрғын үй жылжымайтын мүлік нарығы көптеген ерекшеліктерімен сипатталатын жүз мыңдаған пәтерлерден тұратын күрделі құрылым болып табылады. Бұл ретте нарықтағы кез келген өзгерістер алыпсатарлыққа және жылжымайтын мүлік бағасының әдейі көтерілуіне себеп болуы мүмкін. Сондықтан пәтердің нақты құны қандай және қай жерде қымбаттырақ екенін түсіну маңызды. Зерттеу барысында машиналық оқыту әдістерін пайдалана отырып, жылжымайтын мүліктің құнын болжау моделін құру қарастырылды. Жұмыста машиналық оқытудың бірнеше алгоритмі: Linear Regression, Lasso, Ridge, Decision Tree Regression, Random Forest Regression, SVM (Gaussian kernel) қолданылды, олардың жұмыс нәтижелері мен дәлдіктері көрсетілді.

Түйін сөздер: машиналық оқыту, жылжымайтын мүлік, мәліметтерді өңдеу, регрессиялық талдау, алгоритм.

Аннотация

С.З. Сапакова[1]

[1]*Международный университет информационных технологий, г.Алматы, Казахстан*

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ НА РЫНКЕ НЕДВИЖИМОСТИ АЛМАТЫ

В данной работе был рассмотрен рынок недвижимости в Ауэзовском и Бостандыкском районах города Алматы. Известно, что этот рынок является важным сектором мировой экономики. Рынок жилой недвижимости Казахстана представляет собой сложную структуру, состоящую из сотен тысяч квартир, характеризующихся многими особенностями. В то же время любые изменения на рынке могут стать причиной спекуляций и преднамеренного повышения цен на недвижимость. Поэтому важно понимать, какова реальная стоимость квартиры и где она дороже. В ходе исследования было рассмотрено создание модели прогнозирования стоимости недвижимости с использованием методов машинного обучения. В работе использовались несколько алгоритмов машинного обучения: Linear Regression, Lasso, Ridge, Decision Tree Regression, Random Forest Regression, SVM (Gaussian kernel), показаны результаты их работы и точность.

Ключевые слова: машинное обучение, недвижимость, обработка данных, регрессионный анализ, алгоритм.

Abstract

APPLICATION OF MACHINE LEARNING METHODS IN THE ALMATY REAL ESTATE MARKET

Sapakova S.Z.[1]

[1]*International University of Information Technologies, Almaty, Kazakhstan*

In this work, the real estate market in Auezov and Bostandyk districts of Almaty city was considered. It is known that this market is an important sector of the world economy. The residential real estate market in Kazakhstan is a complex structure consisting of hundreds of thousands of apartments characterized by many features. At the same time, any changes in the market can cause speculation and deliberate increase in real estate prices. Therefore, it is important to understand what the real cost of the apartment is and where it is more expensive. In the course of the study, the creation of a real estate value prediction model was considered using machine learning methods. Several algorithms of machine learning: Linear Regression, Lasso, Ridge, Decision Tree Regression, Random Forest Regression, SVM (Gaussian kernel) were used in the work, their work results and accuracy were shown.

Keywords: machine learning, real estate, data processing, regression analysis, algorithm.

Кіріспе

Қазіргі уақыттағы дүниежүзіндегі саяси және экономикалық жағдайларды ескере отырып, көптеген әлеуметтік маңызы бар салалардың қай бағытта дамитынын болжау оңай емес. Қазақстанның көптеген маңызды секторлары мұнай бағасынан тәуелді және еліміздің батыс аймағында сол саланың ірі өндіру компаниялары шоғырланған. Бұл компаниялардағы көптеген жұмысшылар вахталық әдіспен қызмет атқарып, Қазақстанның түкпір-түкпірінен келеді. Соған байланысты, көптеген ірі өнеркәсіптік кәсіпорындарда жұмысшыларды жұмысқа қабылдаудың, сондай-ақ оларды көтермелеудің мотивациялық ынталандыруларының бірі тұрғын жылжымайтын мүлікті уақытша немесе тұрақты пайдалануға беру болып табылады. Бұл нарықтағы әртүрлі ұсыныстардың ішінен жылжымайтын мүліктің ең жақсы нұсқасын таңдау мәселесін тудырады. Көп жағдайда өзіндік құнын бағалау мүмкін емес, сырттан мамандар тартылады, бұл ақша мен уақыттың шығынына әкеледі. Жылжымайтын мүліктің құнына әсер ететін әртүрлі параметрлері бар. Баға белгілеу процесінің өзі күрделі, сондықтан пәтердің оңтайлы құнын түсіну әрқашан мүмкін емес. Пәтердің шамамен алынған құнын білу сатуды тездетеді және тұрғын үйді сатып алу кезінде шығындарды азайтады. Қазіргі уақытта пәтер құнын бағалаудың бірнеше әдістері ұсынылған. Мысалы, [1] мақалада авторлар жасанды нейрондық желіні пайдалана отырып, гедоникалық модельді болжау дәлдігін салыстыру мақсатын қойды. Авторлар Жаңа Зеландиядағы Крайстчерч қаласындағы 200 үйді кездейсоқ таңдаған. Құрылыс үлгілері үшін үйдің өлшемі, үйдің жылы, үй түрі, жатын бөлмелердің саны, ванна бөлмесінің саны, гараждар саны, үй айналасындағы қолайлы орналасуы және географиялық орны ескерілген. Гедоникалық модель болжам бойынша сұралатын баға үшін шешуші болатын үйдің атрибуттарымен байқалатын оның бағасының регрессиясын қамтиды. Бұл келесі мақалада авторлар жартылай журнал моделін пайдаланды, өйткені баға өте сезімтал және тұрақсыз компонент [2-4] болып табылады. Нейрондық желінің моделі гедоникалық модельді құруда қолданылатын процеске ұқсас [5-8]. Жасанды нейрондық желінің оңтайлы моделін анықтау үшін сынақ және қателік әдісі қолданылады [9-11]. Нәтижесінде нейрондық желі моделі гедонистік модельге қарағанда жақсырақ болды. Дегенмен, екі модель де айнымалы орындардың тұрғын үй бағасының маңызды рөл атқаратынын көрсетті. Қарастырылған [1] пен [3] айырмашылығы, мақала [3] орналасу мен жалға алу арасындағы сызықтық қатынасты зерттеді. Авторлар Орегон штатындағы Портленд қаласының пәтерлерін зерттей отырып, өз зерттеулерін жүргізді. Модельді құрастырмас бұрын авторлар 600 астам пәтерлерге бақылау жүргізген. Нәтижелер көрсеткендей, қала орталығынан 10 км қашықтықта орналасқан сайын жалға алу құны төмендейді. Алайда, содан кейін 7 км жалға алу бағасы көтеріледі. Бұл қала маңындағы тұрғындарды, қала орталығын айналып өтетін айналма жолға көшірумен байланысты. Содан кейін пәтерлер айналма жолдан және қала орталығынан алыстаған сайын жалдау құны тағы да өседі. Пәтерлер бойынша жиналған деректерге сәйкес зерттеушілер үлгіні құрастырды:

$$R_i = \beta X_i + \gamma_1 DCC_i + \gamma_2 DCC_i{}^2 + \gamma_3 DCC_i{}^3 + \gamma_4 DH_i + \gamma_5 DH_i{}^2 + \gamma_6 DI_i + \gamma_7 DI_i{}^2 \qquad (1)$$

мұндағы R_i – ай сайынғы пәтерді жалдау, X_i – пәтер атрибуттарының векторы, β – осы пәтер атрибуттары үшін жасырын баға шектеулерінің векторы, DC_i – қала орталығынан пәтерге дейінгі қашықтық, DH_i – жақын маңдағы тас жолдан пәтерге дейінгі қашықтық, DI_i – екі тас жолдың ең жақын қиылысынан пәтерге дейінгі қашықтық, e – стохастикалық қате. Алынған модель (1) орталықтан қашықтығы мен Портленд қаласындағы пәтердің жалдау құны арасындағы байланысты жақсы сипаттады. Сондай-ақ болжамның әртүрлі әдістерін қолдана отырып, жылжымайтын мүлік құнының бағасын болжауға болады. Мысалы, [8-15] мақалаларында авторлар логистикалық регрессия, SVM, Lasso регрессиясы, шешім ағашының регрессиясы, кездейсоқ орман регрессиясы және нейрондық желілер сияқты алгоритмдерді пайдаланады. Осы зерттеулер үшін жылжымайтын мүлікті сату бағасы үйдің орналасқан жері, үйдің материалы, пәтердің ауданы, үйдің жылы (жасы), бөлмелер саны және т.б. сияқты факторлармен анықталды. Осылайша, пәтер құнына оның ауданы немесе бөлмелер саны сияқты ішкі параметрлері ғана емес, сонымен қатар пәтердің орналасуын сипаттайтын сыртқы параметрлер де әсер етеді.

Мәліметтер мен әдістер. Қарастырылатын мәселе

Жұмыстың мақсаты Алматы қаласындағы жылжымайтын мүліктердің сатылым бағасына болжамдық үлгі алу. Үлгілерге қажетті мәліметтер krisha.kz веб-сайтындағы пәтерді сату туралы хабарландырулар болып табылады. Қарастырылған жұмыста Әуезов және Бостандық аудандары

бойынша пәтерлер туралы мәліметтер толық талданып, жинақталды. Сатылымдағы пәтерлердің артық бағаланбағанын анықтау мақсатында олардың сметалық құнын үлгі арқылы құрастыру қажет. Бұл баға көптеген компанияларға пәтерлердің нарықтағы анық бағасын көрсетеді.

Деректер

Зерттеу барысындағы деректер жоғарыда айтылғандай krisha.kz және olx.kz сайттарынан қолданылды (1-суретте көрсетілгендей). Алынған мәліметтердің сапасын арттыру үшін жылжымайтын мүлік құнына шектеу енгізілді - 100 миллион тенгеден аспайды. Ең минимальді баға 1 млн және максимальді баға 100 млн. Бұл шектеу қымбатырақ пәтерлердің бағасы сәл өзгеше ережелер мен тәуелділіктерге бағынады деген болжамға негізделген.

	title	price	Город	Дом	Этаж	Площадь	Состояние	Санузел	Балкон	Балкон остеклён	...	Интернет	М
0	\n 1-комнатная квартира, 32...	23000000 ₸	\nАлматы, Бостандыкский р-н\nпоказать на карте\n	кирпичный, 1982 г.п.	4 из 5	32 м²	хорошее	совмещенный	балкон	да	...	ADSL	чаг мебли
1	\n 3-комнатная квартира, 93...	71000000 ₸	\nАлматы, Бостандыкский р-н\nпоказать на карте\n	монолитный, 2021 г.п.	11 из 19	93 м²	требует ремонта	2 с/у и более	NaN	NaN	...	NaN	
2	\n 3-комнатная квартира, 12...	65000000 ₸	\nАлматы, Бостандыкский р-н\nпоказать на карте\n	кирпичный, 2009 г.п.	8 из 12	123 м², кухня — 15.6 м²	хорошее	2 с/у и более	NaN	NaN	...	NaN	
3	\n 5-комнатная квартира, 28...	от 181097631 ₸	\nАлматы, Бостандыкский р-н\nпоказать на карте\n	монолитный, 2016 г.п.	NaN	289.9 м²	NaN	2 с/у и более	NaN	NaN	...	NaN	
4	\n 3-комнатная квартира, 88...	75000000 ₸	\nАлматы, Бостандыкский р-н\nпоказать на карте\n	монолитный, 2015 г.п.	14 из 14	88 м²	хорошее	2 с/у и более	несколько балконов или лоджий	NaN	...	оптика	
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Сурет 1. Өңделмеген мәліметтер жиыны

Қарастырылатын параметрлер: бөлмелер саны, студия, жалпы алаңы, қабаты, үйдегі қабаттар саны, тұрақ, жөндеу түрі, балкондар саны, ванна түрі, лифттер саны, салынған жылы, үйдің апаттылығы, едендердің түрі, қабырға материалы, шаршысы, бөлмелер саны. Алынған деректер қорының жалпы көлемі 3882 жазбаны құрайды және 23 бағаннан тұрады.

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   title             3645 non-null   object
 1   price             3645 non-null   int32
 2   Состояние         3645 non-null   int32
 3   Санузел           3645 non-null   int32
 4   Балкон            3645 non-null   int32
 5   Балкон остеклён   3645 non-null   int32
 6   Дверь             3645 non-null   int32
 7   Телефон           3645 non-null   int32
 8   Интернет          3645 non-null   int32
 9   Мебель            3645 non-null   int32
 10  Пол               3645 non-null   int32
 11  Безопасность      3645 non-null   int32
 12  В прив. общежитии 3645 non-null   int32
 13  description       3645 non-null   int32
 14  Жилой комплекс    3645 non-null   int32
 15  Парковка          3645 non-null   int32
 16  Возможен обмен    3645 non-null   int32
 17  district          3645 non-null   int32
 18  year              3645 non-null   int32
 19  type              3645 non-null   int32
 20  real_floor        3645 non-null   int32
 21  from_floor        3645 non-null   int32
 22  area              3645 non-null   float64
 23  ceiling           3645 non-null   float64
dtypes: float64(2), int32(21), object(1)
memory usage: 412.9+ KB
```

Сурет 2. Жылжымайтын мүлік бағасына әсер ететін факторлар

Жалпы нәтижеге елеусіз мәндерді қосатын факторлар моделдеу барысында алынып тасталады және қабылданбайды. Деректер жинағы екі бөлікке бөлінеді - оқу жинағы және сынақ жинағы. Машинамен оқытудың әртүрлі үлгілері оқу жинағының көмегімен оқытылады. Содан кейін сынақ жинағы машиналық оқытудың барлық үлгілерінің өнімділігін тексеру үшін пайдаланылады. Дәлдігі есептеледі және барлық үлгілердің орташа квадраттық қатесі есептеледі. Соңғы қадамда үй бағасын болжау үшін ең жоғары дәлдік ұпайы және ең аз RMSE (Root Mean Square Error) мәні бар модель пайдаланылады.

Алматы қаласы бойынша пәтерлер бағасының өзгеру динамикасын зерттеуде, сайттан тікелей алынған мәліметтерден байқайтынымыз, хабарландыруларда жылжымайтын мүлікке қатысты көптеген параметрлер көрсетілген. Көріп отырғанымыздай бұл мәліметтер сандық және сапалық категориялардан болып келеді. Ары қарай мәліметтерді өңдеу кезеңінде барлық сипаттамалардың мәні сандық форматқа келтіріледі. Жалпы бағаның құраушылар саны -23, енді осылардың ішінен бағаға ықпал ететін маңызды факторларды анықтауда корреляциялық матрица есептеледі, оның нәтижесі 3-суретте көрсетілді.

	price	Состояние	Санузел	Балкон	Балкон остеклён	Дверь	Телефон	Интернет	Мебель	Пол	...	Жилой комплекс	Парковка
price	1.000000	0.221730	-0.527768	0.141388	0.035776	-0.071543	0.051320	0.036315	-0.037936	0.031784	...	-0.448827	-0.331374
Состояние	0.221730	1.000000	0.028789	0.223104	0.179126	0.088353	0.177957	0.180292	0.111717	0.143500	...	-0.136608	0.062555
Санузел	-0.527768	0.028789	1.000000	0.087313	0.124311	0.192776	0.085908	0.131490	0.168507	0.124857	...	0.196075	0.223865
Балкон	0.141388	0.223104	0.087313	1.000000	0.632696	0.260000	0.152759	0.264745	0.217277	0.387116	...	-0.091127	0.119352
Балкон остеклён	0.035776	0.179126	0.124311	0.632696	1.000000	0.363932	0.186267	0.336133	0.161314	0.381137	...	-0.062165	0.182100
Дверь	-0.071543	0.088353	0.192776	0.260000	0.363932	1.000000	0.275215	0.380632	0.240325	0.394572	...	0.018578	0.325904
Телефон	0.051320	0.177957	0.085908	0.152759	0.186267	0.275215	1.000000	0.519618	0.181616	0.224823	...	-0.093277	0.184395
Интернет	0.036315	0.180292	0.131490	0.264745	0.336133	0.380632	0.519618	1.000000	0.220425	0.383341	...	-0.099818	0.242926
Мебель	-0.037936	0.111717	0.168507	0.217277	0.161314	0.240325	0.181616	0.220425	1.000000	0.346258	...	-0.047757	0.148675
Пол	0.031784	0.143500	0.124857	0.387116	0.381137	0.394572	0.224823	0.383341	0.346258	1.000000	...	-0.154616	0.230483
Безопасность	0.129415	0.218839	0.076493	0.306994	0.342046	0.356496	0.349977	0.355760	0.154075	0.283388	...	-0.102048	0.204185
В прив. общежитии	0.278996	0.250998	-0.176580	0.095095	0.077782	0.130853	0.087719	0.116741	0.072741	0.152006	...	-0.252591	0.093263
description	-0.059592	-0.091501	-0.033302	-0.176581	-0.148953	-0.077597	-0.002695	-0.088876	-0.055144	-0.161407	...	0.016985	-0.058972
Жилой комплекс	-0.448827	-0.136608	0.196075	-0.091127	-0.062165	0.018578	-0.093277	-0.099818	-0.047757	-0.154616	...	1.000000	0.355036
Парковка	-0.331374	0.062555	0.223865	0.119352	0.182100	0.325904	0.184395	0.242926	0.148675	0.230483	...	0.355036	1.000000
Возможен обмен	0.039875	-0.002999	-0.019874	0.073073	0.156365	0.125613	0.026036	0.086413	-0.220808	0.137885	...	-0.044134	0.076179
district	0.406669	0.101709	-0.148933	0.159318	0.106620	0.030509	0.069806	0.110590	0.035183	0.151419	...	-0.367713	-0.137784
year	0.568147	0.210257	-0.280240	0.177562	0.087062	-0.035297	0.091769	0.112013	0.030303	0.145499	...	-0.643684	-0.373098
type	-0.233529	-0.074786	0.137121	-0.049098	-0.034738	0.010021	-0.048325	-0.050040	0.032070	-0.025996	...	0.199689	0.141589
real_floor	0.302367	0.013765	-0.170055	0.020461	-0.040141	-0.085797	0.003843	0.002293	-0.054502	-0.003496	...	-0.237945	-0.340651
from_floor	0.433502	0.005194	-0.196822	0.048186	-0.032661	-0.129228	0.001407	-0.006768	-0.073769	-0.005395	...	-0.379726	-0.460278
area	0.846378	0.157910	-0.535190	0.088792	0.008066	-0.040140	0.047107	0.010173	-0.013351	0.023618	...	-0.242113	-0.232077
ceiling	0.288372	0.079969	-0.193381	-0.054010	-0.085816	-0.139607	-0.068041	-0.066438	-0.051515	-0.118096	...	-0.398736	-0.335056

Сурет 3. Пирсон әдісімен есептелген корреляциялық матрица

Нәтижені талдап артық факторларды алып тастап, машиналық оқыту алгоритмдерін қолданамыз. Сонымен 4-суретте баға тәуелді факторымен ең жоғары коррелциялық коэфициентке ие сипаттамалар тізіміне шолу жасалған.

```
high_corr_var
```

```
[('price', 'Санузел'),
 ('price', 'Жилой комплекс'),
 ('price', 'district'),
 ('price', 'year'),
 ('price', 'from_floor'),
 ('price', 'area'),
 ('Санузел', 'area'),
 ('Балкон', 'Балкон остеклён'),
 ('Телефон', 'Интернет'),
 ('Жилой комплекс', 'year'),
 ('Парковка', 'from_floor'),
 ('year', 'from_floor'),
 ('year', 'area'),
 ('real_floor', 'from_floor')]
```

Сурет 4. Баға тәуелді факторымен ең жоғары коррелциялық коэфициентке ие сипаттамалар

Сонымен, бізде келесі

```
: X = data.loc[:, ['Санузел', 'Жилой комплекс', 'district', 'year', 'from_floor', 'area']]
  y = data.iloc[:, 1]
  print(X.shape)
  print(y.shape)

  (3645, 6)
  (3645,)
```

Сурет 5. Корреляциялық талдаудан кейін қалған факторлар

Жылжымайтын мүлік бағасына болжамдар алу барысында осы 5-суреттегі факторларды ғана қолданамыз, байқағанымыздай 23 сипаттаушы хабарландыру сайтында сатып алушыларға толық ақпарат беру мақсатында қолданылғанымен, белгілі бір алгоритмдер кезінде ықпалы жоқ болғандықтан оларды датафреймнен алып тастаймыз.

Әдістер

Функцияларды таңдау мүмкіндіктерді рекурсивті жою әдісі арқылы орындалды, мұнда регрессия мәселесін шешу кездейсоқ орман машинасын оқыту алгоритмі арқылы жүзеге асырылды және жоталық және сызықтық регрессия әдістері. Үздік үлгіні анықтау критерийі RMSE (орташа квадрат қатесінің түбірі) формуламен анықталатын кросс-валидация болды:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{m}\left(y_{ij} - \widehat{y_{ij}}\right)^2}{n}} \quad (2)$$

мұндағы y_{ij} - i-ші үлгінің валидация жиынының j-ші нүктесіндегі жауаптың мәні; $\hat{y}_{ij}$ - j-ші нүктедегі i-ші үлгінің шығуы; k - қарсы тексеру блоктарының саны (10 блок); m - элементтер саны валидация үлгісінің; n – бастапқы үлгінің өлшемі. Тордағы іздеу арқылы оңтайлы гиперпараметрлерді таңдағаннан кейін элементтік қарсы тексерудің MAE (орташа абсолютті қате) мәні қатені анықтау үшін есептелді, өлшем бірлік теңгемен.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|, \quad (3)$$

мұнда y_i, - i-ші модель үшін басқару деректер жиынының жауап мәні; ŷi - - i-ші үлгінің шығысы бақылау сынама нүктесі; n- бастапқы үлгінің өлшемі.

1. Қарапайым сызықтық регрессия

Қарапайым сызықтық регрессияда біз бір айнымалының бағалауын екінші айнымалының бағалауы арқылы болжаймыз. Регрессия сызығының формуласы

$$Y' = bX + A,$$

мұндағы, Y' - болжамды бағалау, яғни тәуелді айнымалы, x - тәуелсіз айнымалы, b - түзудің еңісі және A - Y-пен қиылысу нүктесі. Бір ғана x предикторлы айнымалысы болғанда болжау әдісі қарапайым сызықтық регрессия деп аталады.

Ең кіші квадраттар (OLS) регрессиясы жиі тұрақсыз болуы мүмкін, яғни жаттығу деректеріне қатты тәуелді болады, бұл әдетте шамадан қайта оқуға деген тенденцияны көрсетеді. Регуляризация мұндай артық орнатуды болдырмауға көмектеседі - қалаған параметрлерге қосымша шектеулер енгізуден тұратын жалпы әдіс, бұл модельдің шамадан тыс күрделілігіне жол бермейді. Процедураның мағынасы b коэффициенттерінің векторын баптау барысында олар абсолютті мәнде ең кіші квадраттарды оңтайландыруға қарағанда біршама кішірек болатындай етіп «келтіру» болып табылады.

2. Жоталық регрессия (Ridge Regression)

Жотаның регрессиясы немесе жоталық регрессия өлшемділікті тексеру әдістерінің бірі болып табылады. Оны тәуелсіз айнымалылар бір-бірімен корреляциялану нәтижесінде көп айнымалы

сызықтық регрессия коэффициенттерінің бағалауларын тұрақсыз ететін артық деректерді азайтуда қолданылады (яғни мультиколлинеарлық орын алғанда).

$$\sum_{i=1}^{M}(y_i - \widehat{y_i})^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2. \quad (4)$$

Қарапайым тілмен айтқанда,

$$\text{Ridge R} = \text{loss} + \lambda||w||^2,$$

мұнда λ тұрақты,

$$||w||^2 = w_1{}^2 + w_2{}^2 + w_3{}^2 \ldots,$$

мұнда w – коэффициент векторы. Сонымен, ридж-регрессиясы коэффициенттерге (w) шектеулер қояды. Айыппұл термині (лямбда) коэффициенттерді реттейді, егер коэффициенттер үлкен мән қабылдаса, айыппұл салынады. Осы сызықтар бойымен жотаның қайталануы коэффициенттерді қысқартады және бұл модельдің болжаусыздығы мен көп коллинеарлылығына ықпал етеді.

3. Лассо регрессия әдісі (LASSO, ең аз абсолютті қысқарту және таңдау операторы) модельді оңтайландыру функционалдығына қосымша реттеу қосылғышын енгізуден тұрады, бұл көбінесе орнықты шешім алуға мүмкіндік береді. Предикторлардың параметрлерін бағалаудағы квадраттық қатені азайту шарты келесі формуламен өрнектеледі:

$$\sum_{i=1}^{M}(y_i - \widehat{y_i})^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j|. \quad (5)$$

Қарапайым тілмен айтқанда,

$$\text{Lasso} = \text{loss} + \lambda||w||.$$

Көріп отырғанымыздай Лассо регрессия коэффициенттері Ridge сияқты шектеулерге бағынады.

Бұл жағдайда регрессия қатесі мен |w| коэффициенттерінің абсолютті мәндерінің қосындысы ретінде көрсетілген пайдаланылатын мүмкіндіктер кеңістігінің өлшемі арасында белгілі бір ымыраға қол жеткізіледі. Минимизациялау барысында кейбір коэффициенттер нөлге тең болады, бұл шын мәніндегі ақпараттық белгілердің таңдалғандығын көрсетеді. Регуляризация параметрінің мәні λ=0 болғанда, лассо регрессиясы кәдімгі ең кіші квадраттар әдісіне келтіріледі, ал λ өскен сайын генерацияланған модель нөлдік модельге айналғанша көбірек «ықшамдала» береді. λ оңтайлы мәні кросс-валидация көмегімен табылады, яғни ол модельдің $\widehat{\boldsymbol{y_i}}$ өзін құруға қатыспаған бақылаулар бойынша минималды болжам қатесіне сәйкес келеді.

4. SVM векторлық регрессияны қолдау

SVM – классификациялау алгоритмдері үшін кеңінен қолданылатын бақыланатын оқыту алгоритмі. SVM сызықты түрде бөлінетін деректер үшін ғана қолданылады. Сызықты емес деректерде ядроның функциялары қолданылады. SVM мәліметтерді екі класқа «гипержазықтық» көмегімен жіктейді. Гипержазықтық берілген деректерді класқа бөлу үшін жоғары өлшемді кеңістіктегі ең үлкен мәнге ие болуы керек. Екі класс арасындағы айырмашылық олардағы ең жақын деректер нүктелері арасындағы ең үлкен қашықтықты білдіреді. Алгоритм осы параллель гипержазықтықтар арасындағы айырмашылық немесе қашықтық неғұрлым көп болса, классификатордың орташа қателігі соғұрлым аз болады деген болжамға негізделген. (1) теңдеу келесідей көрсетілген (SVR туралы егжей-тегжейлі талқылау үшін Basak et al., 2007; Mu et al., 2014).

$$min\frac{1}{2}w^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i{}^*).$$

$$\begin{cases} y_i - f(x_i, w) \le \varepsilon + \xi_i^* \\ f(x_i, w) - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, i = 1, 2, \dots, n \end{cases} \quad (6)$$

мұндағы w^2 – модель күрделілігі; ε – сезімтал емес жоғалту функциясы; ξ_i– і деректер нүктесінің қате жіктелу дәрежесін өлшейтін бос айнымалы; C – мақсат функциясындағы шығын параметрі. ξ_i – нөл емес және ол C шығын параметріне көбейтіледі.

Оңтайландыруды қосарлы есептерге айналдыруға болады және шешімдер (2) теңдеу ретінде көрсетіледі.

$$f(x) = \sum_{i=1}^{n_{SV}} (\alpha_i + \alpha_i^*) K(x_i, x) s.t. \begin{cases} 0 \le \alpha_i^* \le C \\ 0 \le \alpha_i \le C \end{cases} \quad (7)$$

мұндағы n_{SV} – тірек векторларының саны; $K(x_i, x)$ – (3) теңдеуде көрсетілгендей ядро функциясы.

$$K(x_i, x) = \sum_{j=1}^{m} g_j(x)\, g_j(x_i) \quad (8)$$

Біз SVM-ді бағалағанда, біздің деректер жинағымыз k= 5 тең ішкі жиындарға бөлінеді, мұнда әрбір қатпар белгілі бір уақытта сынақ жинағы ретінде пайдаланылады. Әрбір итерация R^2 бағасын береді, содан кейін біз моделіміздің жалпы дәлдігін анықтау үшін өтінімде олардың орташа мәнін есептей аламыз.

5. Decision tree (Шешім ағашы) – адамның болжау есептерін шешуіне ұқсастыра отырып негізделген бақыланатын оқу мәселесінің алгоритмі. Жалпы жағдайда, бұл жапырақты емес шыңдардағы (түйіндер) шешім ережелері және жапырақ төбелеріндегі мақсаттық функция туралы кейбір қорытынды (болжау) бар k-өлшемді ағаш. Шешім ережесі – қарастырылып отырған нысанды еншілес шыңдардың қайсысына орналастыру керектігін анықтауға мүмкіндік беретін нысанның қандайда бір функциясы. Жапырақ төбелерінде әртүрлі объектілерді орналастыруға болады: сол жерге жеткен объектіге тағайындалуы керек класс (жіктеу мәселесінде), класс ықтималдықтары (жіктеу мәселесінде), мақсат функциясының тікелей мәні (регрессия мәселесі). Кездейсоқ орман регрессиясы айнымалыларды кездейсоқ таңдау негізінде көптеген шешімдер ағашын дамытады. Ол көптеген ағаштарға негізделген тәуелді айнымалылар класын береді.

1. Деректерді кездейсоқ таңдау:- original data= subset 1+subset 2+subset 3+.....

Ағаштар деректерді, сондай-ақ айнымалыларды кездейсоқ таңдауға негізделгендіктен, бұл кездейсоқ ағашты құрайды. Осындай көптеген кездейсоқ ағаштар кездейсоқ орманға әкеледі.

Модельдер және нәтижелер

Регрессиялық модельдер үшін біз келесі мәселені шешуге тырысамыз: үйді сипаттайтын факторлардың өңделген тізімін ескере отырып, біз оның ықтимал сатылу бағасын болжағымыз келеді. Сызықтық регрессия – регрессия есептері негізгі модельдің табиғи таңдауы болып келеді. Сондықтан біз алдымен 23 факторы (сипаттамасы) және 3882 жаттығу үлгілерін пайдаланып, барлық мүмкіндіктерді қамтитын сызықтық регрессияны іске қосамыз. Одан кейін моделді біздің сынақ деректеріміздегі мүмкіндіктер бойынша үйлердің сатылу бағасын болжауда пайдаланып және оның нәтижесін сынақ деректер жинағында берілген үйлердің нақты сату бағасымен салыстырамыз.

Модель өнімділігі болжамды нәтижелер мен нақты нәтижелердің дәлдік ұпайымен өлшенді. Біздің базалық үлгіміз детерминация коэффициенті 83,72 % құрады.

```
from sklearn.linear_model import LinearRegression, Ridge, BayesianRidge

model = DecisionTreeRegressor(random_state=1)
model.fit(X_train, y_train)
prediction = model.predict(X_test)
```

```
print('Train accuracy: ', model.score(X_train, y_train))
print('Test accuracy: ', model.score(X_test, y_test))
```

```
Train accuracy:  0.9984660574563022
Test accuracy:  0.8546135389248579
```

Сурет 6. Decision Tree Regression алгоритмінің дәлдігі

```
test_pred = lin_reg.predict(X_test)
train_pred = lin_reg.predict(X_train)

print('Test set evaluation:\n_____________________________________')
print_evaluate(y_test, test_pred)
print('Train set evaluation:\n_____________________________________')
print_evaluate(y_train, train_pred)

results_df = pd.DataFrame(data=[["Linear Regression", *evaluate(y_test, test_pred) , cross_val(LinearRegression())]],
                          columns=['Model', 'MAE', 'MSE', 'RMSE', 'R2 Square', "Cross Validation"])
```

Сурет 7. Сызықты регрессияның метрикаларын есептеу

```
Test set evaluation:
_____________________________________
MAE: 5295765.736173671
MSE: 58752946753314.62
RMSE: 7665047.080958773
R2 Square 0.8372620206452704
_____________________________________
Train set evaluation:
_____________________________________
MAE: 5349868.883522503
MSE: 62854764054601.945
RMSE: 7928099.649638742
R2 Square 0.8081434416634332
_____________________________________
```

Сурет 8. Сызықты регрессияның метрикаларының берілген мәлеметтер бойынша нәтижесі

Базалық үлгі ретінде сызықтық регрессия үлгісін пайдаланғаннан кейін, артық сәйкестікті азайту үшін сызықтық регрессия үлгілеріне қосымша реттеу параметрлерін қостық.

	true	prediction
0	98310000	99744400.0
1	95000000	80000000.0
2	92000000	91500000.0
3	92000000	91500000.0
4	91989990	91854917.0
...	...	...
724	10700000	10700000.0
725	10000000	11100000.0
726	9500000	10500000.0
727	9500000	11300000.0
728	9300000	10300000.0

729 rows × 2 columns

Сурет 9. Бағаның болжамды және эмпирикалық мәндері

Жалпы жоғарыдағы 8-суреттен R^2 детерминация коэффициенттерінің мәні мақсатты айнымалыдағы дисперсияның қаншалықты біздің үлгімен түсіндірілетінін көрсетеді. Lasso алгоритімінің детерминация коэффициенті 82,34 % құрады, бұл біздің базалық үлгіміздің төмен. Лассо реттегішінен басқа Ridge-де 83,68 дәлдік алынды. Ал, сызықты регрессия мен Elastic net регрессияларының детерминация коэффициенттері шамалас болып тұр. Гаусс ядросы бар қолдау векторлық регрессия (SVM) да қарастырылып отырған факторларға бағытталды, бірақ өте төменгі нәтижелерді көрсетті.

	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	5.295766e+06	5.875295e+13	7.665047e+06	0.837262
1	Lasso Regression	5.568230e+06	6.375224e+13	7.984500e+06	0.823415
2	Elastic Net Regression	5.282822e+06	5.874988e+13	7.664847e+06	0.837271
3	Random Forest Regressor	3.356879e+06	3.084995e+13	5.554273e+06	0.914550
4	SVM Regressor	1.291767e+07	3.387686e+14	1.840567e+07	0.061655
5	Ridge Regression	5.261640e+06	5.888893e+13	7.673912e+06	0.836885
6	Decision Tree	3.970189e+06	5.248856e+13	7.244899e+06	0.854614

Сурет 10. Қолданылған алгоритмдер бойынша қорытынды нәтижелер

Біздің кездейсоқ орман регрессия моделінің коэффициенті 91,45 % дәлдікке ие, бұл біздің базалық үлгіден де жақсырақ екенін көреміз. Соңында біз 85,46 дәлдік ұпайын берген деректер жиынына шешім ағашының жіктеуішін қолдандық. Тұтастай алғанда, кездейсоқ орман классификаторы мен шешімдер ағашы модельдері негізгі сызықтық регрессия үлгісінен жақсырақ жұмыс істеді. Ең жоғары дәлдік көрсеткішін кездейсоқ орман классификаторы көрсетті. Сонымен, таңдалған факторлар негізінде құрылған тұрғын үй бағасын болжауға арналған үлгі алынған нәтижелер бойынша толықтай тәуелді және болашақта басқа аудандар немесе қалалар үшін қолдана аламыз.

Қорытынды

Ұсынылған жұмыс барысында құрастырылған модельдер жылжымайтын мүліктің ұсыныс бағасының аталған факторларға тәуелділігі Алматы қаласының Әуезов және Бостандық аудандандарындағы пәтер нарығының ағымдағы жағдайын сипаттайды. Пәтер бағасын болжауға арналған модельдер оның параметрлеріне байланысты жақсы статистикалық сипаттамаларға ие және алдағы уақытта кез келген аудан немесе Қазақстан қалалары үшін тұрғын үй құнының болжамды бағалауында пайдаланыла алады. Қолданылған моделдерің ерекшелігі болжамның дәлдігін әртүрлі алгоритмдерді қолдана отырып арттыру болып табылады, сонымен қатар сатушының пәтер құнын асыра бағалау кезінде және жарияланған хабарландырулардағы басқа да анық емес ақпараттар кездескенде мәліметтерді алдын-ала өңдеу кезінде ыңғайлы реттеп алуға болады. Бұл жұмыс барысында жылжымайтын мүлік нарығындағы тек пәтерге құнына қатысты ішкі факторларды қарастырдық, ал алдағы уақытта бағаға байланысты жергілікті жердің сыртқы факторларында ескере отырып жаңа моделдер құруға болады.

Пайдаланылған әдебиеттер тізімі:

1 *De Aquino Afonso, B. K., Melo, L. C., de Oliveira, W. D. G., Da Silva Sousa, S. B., & Berton, L., (2020). Housing prices prediction with a deep learning and random forest ensemble [Unpublished manuscript]. Anais do Encontro Nacional de Inteligencia Artificial e Computacion.*

2 *Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. Applied Sciences, 8(11), 1–24. https://doi.org/10.3390/app8112321*

3 *Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption,". Energy and Buildings, 147(2386), 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038*

4 *Einav, L., & Levin, J. (2014). Economics in the age of big data. Science, 346(6210), 715–721. https:// doi.org/10.1126/science.1243089*

5 *Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10), 906–914. https://doi.org/10.1093/bioinformatics/16.10. 906*

6 *Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439), 531–537. https://doi.org/10.1126/science.286.5439.531*

7 *Mukhlishin, M. F., Saputra, R., & Wibowo, A., (2017) Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor. 2017 1st International Conference on Informatics and Computational Sciences, 171–176.*

8 Muralidharan, S., Phiri, K., Sinha, S. K., & Kim, B. (2018). Analysis and prediction of real estate prices: A case of the Boston housing market. Issues in Information Systems, 19(2), 109–118. http://www.iacis/org/iis/2018/2_iis_2018_109-118.pdf

9 Limsombunchai, V., Gan, C., & Lee, M. (2004). House price prediction: Hedonic price model vs. artificial neural network. American Journal of Applied Sciences, 1(3), 193–201. https://doi.org/ 10.3844/ajassp.2004.193.201

10 Elaine M. Worzala, Margarita Lenk, Ana Silva (1995). An Exploration of Neural Networks and Its Application to Real Estate Valuation // Journal of Real Estate Research; American Real Estate Society. Vol. 10(2). P. 185–202.

11 Nils Kok, Eija-Leena Koponen, Carmen Adriana Martinez-Barbosa (2017). Big Data in Real Estate From Manual Appraisal to Automated Valuation» // The Journal of Portfolio Management. 43(6). P. 202–211.

12 GeoPhy: [сайт]. URL: https://geophy. com/ (дата обращения: 24 мая 2019 года).

13 Ясницкий В. Л. Нейросетевое моделирование в задаче массовой оценки жилой недвижимости города Перми // Фундаментальные исследования. 2015. № 10-3. С. 650– 653. URL: http://www.fundamental-research. ru/ru/article/view?id=39274.

14 Сурков Ф. А., Петкова Н. В., Суховский С. Ф. Нейросетевые методы анализа данных в оценке недвижимости» // Известия вузов. Северо-Кавказский регион. Технические науки. 2016. № 3. С. 38-45.

15 Арефьева Е. А., Костяев Д. С. Использование нейронных сетей для оценки рыночной стоимости недвижимости // Известия Тульского государственного университета. Технические науки. 2017. Вып. 10. С. 177–184

References:

1 De Aquino Afonso, B. K., Melo, L. C., de Oliveira, W. D. G., Da Silva Sousa, S. B., & Berton, L., (2020). Housing prices prediction with a deep learning and random forest ensemble [Unpublished manuscript]. Anais do Encontro Nacional de Inteligencia Artificial e Computacion.

2 Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. Applied Sciences, 8(11), 1–24. https://doi.org/10.3390/app8112321

3 Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption,". Energy and Buildings, 147(2386), 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

4 Einav, L., & Levin, J. (2014). Economics in the age of big data. Science, 346(6210), 715–721. https:// doi.org/10.1126/science.1243089

5 Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10), 906–914. https://doi.org/10.1093/bioinformatics/16.10. 906

6 Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

16 Mukhlishin, M. F., Saputra, R., & Wibowo, A., (2017) Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor. 2017 1st International Conference on Informatics and Computational Sciences, 171–176.

7 Muralidharan, S., Phiri, K., Sinha, S. K., & Kim, B. (2018). Analysis and prediction of real estate prices: A case of the Boston housing market. Issues in Information Systems, 19(2), 109–118. http://www.iacis/org/iis/2018/2_iis_2018_109-118.pdf

8 Limsombunchai, V., Gan, C., & Lee, M. (2004). House price prediction: Hedonic price model vs. artificial neural network. American Journal of Applied Sciences, 1(3), 193–201. https://doi.org/ 10.3844/ajassp.2004.193.201

9 Elaine M. Worzala, Margarita Lenk, Ana Silva (1995). An Exploration of Neural Networks and Its Application to Real Estate Valuation // Journal of Real Estate Research; American Real Estate Society. Vol. 10(2). P. 185–202.

10 Nils Kok, Eija-Leena Koponen, Carmen Adriana Martinez-Barbosa (2017). Big Data in Real Estate From Manual Appraisal to Automated Valuation» // The Journal of Portfolio Management. 43(6). P. 202–211.

11 GeoPhy: [site]. URL: https://geophy. com/ (accessed 24 May 2019).

12 Yasnitsky V. L. Neural network modeling in the problem of mass assessment of residential real estate in the city of Perm // Fundamental research. 2015. No. 10-3. pp. 650–653. URL: http://www.fundamental-research. ru/ru/article/view?id=39274.

13 Surkov F. A., Petkova N. V., Sukhovskiy S. F. Neural network methods of data analysis in real estate valuation. Izvestiya vuzov. North Caucasian region. Technical science. 2016. No. 3. C. 38-45.

14 Arefieva E. A., Kostyaev D. S. The use of neural networks to assess the market value of real estate // News of the Tula State University. Technical science. 2017. Issue. 10. S. 177–184