УДК 330.47 **DOI: 10.26565/2311-2379-2020-99-12**

В.Д. Бобров
студент
Харківський національний університет імені В.Н. Каразіна
майдан. Свободи, 4, м. Харків, 61022, Україна
E-mail: bobrov.vld@gmail.com, ORCID: https://orcid.org/0000-0001-9042-8227
А.В. Кім
студент
Харківський національний університет імені В.Н. Каразіна
майдан. Свободи, 4, м. Харків, 61022, Україна
E-mail: andrykim05@gmail.com, ORCID: https://orcid.org/0000-0002-6879-2396

ПРОГНОЗУВАННЯ ЦІН НА РИНКУ ОРЕНДИ ЖИТЛА З ВИКОРИСТАННЯМ МЕТОДІВ МАШИННОГО НАВЧАННЯ

В роботі проведено дослідження факторів ціноутворення на ринку короткострокової оренди житла. У якості об'єкта дослідження обрано компанію Airbnb, що являє собою площадку для розміщення, пошуку та оренди житлових приміщень по всьому світу. На початок 2021 р. компанія налічує пропозиції 7 мільйонів житлових приміщень з більш ніж 220 країн світу. Чималу роль в успіху компанії відіграє використання методів Data Science. Одним з ключових алгоритмів компанії є алгоритм ціноутворення. З використанням функції «Рекомендації за цінами» власник житла може проаналізувати які дати швидше за все будуть заброньовані за поточною ціною, а які ні, та сформувати вигідну пропозицію. Система вираховує рекомендовану вартість житла на підставі сотень параметрів, деякі з яких легко розпізнати, однак є й менш очевидні фактори, що також можуть впливати на попит. В роботі запропоновано алгоритм виявлення неявних факторів ціноутворення на ринку короткострокової оренди з використанням методів машинного навчання, що включає: 1) збір та первинну обробку даних; 2) побудову та аналіз моделей лінійної регресії; 3) побудову та аналіз моделей нелінійної регресії. Дослідження проведено на прикладі об'яв з сайту Airbnb у штатах Вашингтон та Нью-Йорк з використанням програмних скриптів, які розроблено на Python. Побудовано та проаналізовано наступні моделі: однофакторна лінійна регресія, багатофакторна лінійна регресія, поліноміальна регресія, дерева рішень, випадковий ліс та бустинг. Результати дослідження показали, що найважливішими є фактори accommodates, cleaning_fee, room_type, bedrooms. Але виходячи з показників якості моделювання, отримані моделі не можна використовувати для впровадження: лінійні моделі мають невисоку якість, тоді як випадковий ліс, бустінг та дерева перенавчені. Але отримані результати можуть використовуватися при проведенні бізнес-аналізу.

Ключові слова: фактори ціноутворення, машинне навчання, регресія, Airbnb.
JEL Classification: C55, E37, L85.

Вступ. Airbnb – це унікальна площадка для розміщення, пошуку та оренди житлових приміщень по всьому світу. Компанія була створена у 2008 р. у Сан-Франциско Б. Чески, Д. Геббіа та Н. Блечарзиком. Стартап отримав інвестиції від фонду Y Combinator, та інших організацій (Businessofapps, 2020; Сайт компанії Airbnb, 2020). Зараз Airbnb займає близько чверті ринку бронювання житла (Airbnb Beat Expedia in Booked Room Nights, 2020). (рис. 1).

На кінець 2020 р. компанія налічує пропозиції 7 мільйонів житлових приміщень, близько 100 тисяч міст, більш ніж 220 країн по всьому світу, 150 мільйонів унікальних користувачів та 10 регіональних офісів (Businessofapps, 2020). У 2019 р. дохід компанії склав 4,7 мільярда доларів (Techcrunch, 2020), на 2020 р. планувалося (через коронавірус ця цифра зменшилася) близько 187 мільйонів бронювань (Доходи Airbnb у 2020 році, 2020). Але через світову кризу, сама компанія за рік подешевшала у два рази – з 38 мільярдів у 2019 р. до 18 мільярдів доларів у 2020 р. (Офіційний сайт статистики Airbnb, 2020)

Чималу роль в успіху компанії зіграло активне використання методів Data Science. Центральним елементом сайту Airbnb є функція пошуку, яка адаптує пошуковий досвід користувача в залежності від його характеристик. Крім того, алгоритми Airbnb використовують

інформацію про дії власника житла, включаючи відмови і прийняті запити, переваги бронювання, а також деталі поїздки для підрахунку ймовірності прийняття запиту на заселення (Charkov et al, 2013).

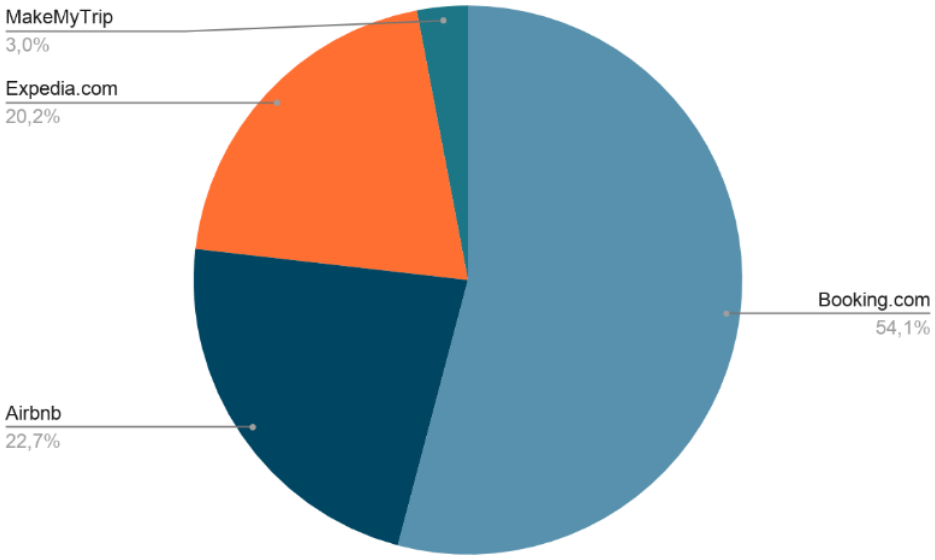


Рис.1. Розподіл ринку бронювання житла у 1 кварталі 2019 року

Джерело: (Yahoo, 2019)

Одним з ключових питань компанії є алгоритм ціноутворення, зокрема функція «Рекомендації за цінами». З її допомогою власник житла може проаналізувати які дати швидше за все будуть заброньовані за поточною ціною, а які ні, та сформувати вигідну пропозицію. Система вираховує рекомендовану вартість житла на підставі сотень параметрів. Деякі тенденції легко розпізнати, наприклад щорічні фестивалі та великі конференції можуть вплинути на підвищення цін по всьому місту. Однак є й менш очевидні фактори, що також можуть впливати на попит (McCarthy, 2018).

Таким чином, виявлення неявних факторів ціноутворення на ринку короткострокової оренди методами машинного навчання є актуальною задачею.

Для її вирішення в роботі було поставлено наступні завдання:

I. Збір та первинна обробка даних.
II. Побудова та аналіз моделей лінійної регресії.
III. Побудова та аналіз моделей нелінійної регресії.

Для проведення регресійного аналізу було обрано штати Вашингтон та Нью-Йорк через декілька причин:

1. Вашингтон – це столиця, тому тут є попит на житло, пов’язаний з адміністративними функціями міста.
2. Нью-Йорк – один з фінансових та туристичних центрів світу, тому вивчення міста може дати багато додаткової інформації щодо організації міжнародних економічних відносин (The World’s Cities, 2016).
3. Передмістя Вашингтону цікаві у порівнянні з містом та з огляду на те, що в США багато людей не бажають проводити багато часу у мегаполісах.

Дослідження проведено з використанням програмних скриптів, які розроблено на Python.

Підготовка даних. У роботі використано датасет Kaggle (Kaggle, 2020) у форматі таблиці, що описує вибірку житла з сайту Airbnb. Ці дані є відкритими для вивчення та не потребують сплати за використання у наукових цілях. Ключем є атрибут id, який однозначно визначає об’єкт у базі даних сайту. Перед очищенням даних ми мали 3013 спостережень (рядків) та 45

характеристик (стовпців). Зазначаємо, що більшість цих характеристик є нецікавими для дослідження та не підлягають інтерпретації у моделях, тому вони будуть виключені з вибірки.

Поглянемо на основні статистики вибраних нами змінних (рис.2).

	accommodates	bathrooms	bedrooms	beds	guests_included	minimum_nights	maximum_nights	number_of_reviews	review_scores_rating
count	3013	2988	3000	3003	3013	3013	3013	3013	2310
mean	3	1	1	2	2	2	716869	15	93
std	2	1	1	1	1	4	39123182	30	8
min	1	0	0	1	0	1	1	0	30
25%	2	1	1	1	1	1	120	1	90
50%	2	1	1	1	1	2	1125	4	95
75%	4	1	1	2	2	3	1125	16	100
max	16	8	10	16	16	180	2147483647	362	100

Рис.2. Основні статистики змінних

Джерело: власні розрахунки

Попередній аналіз показав, що потрібно позбутися пропущених значень, що можуть вплинути на якість моделювання (рис.3). Для заповнення пропущених значень використано KNN-Imputation. Цей алгоритм дозволяє зберегти кореляцію між змінними.

	Total	Percent
cleaning_fee	1127	0.374046
review_scores_checkin	707	0.234650
review_scores_cleanliness	707	0.234650
review_scores_accuracy	706	0.234318
review_scores_value	704	0.233654
review_scores_location	704	0.233654
review_scores_communication	704	0.233654
review_scores_rating	703	0.233322
reviews_per_month	672	0.223034
host_acceptance_rate	493	0.163624
host_response_rate	349	0.115831
bathrooms	25	0.008297
bedrooms	13	0.004315
beds	10	0.003319
property_type	1	0.000332
number_of_reviews	0	0.000000

Рис.3. Статистика пропущених значень

Джерело: власні розрахунки

Асиметричні числові змінні були прологарифмовані. Проблему аномальних спостережень було вирішено за допомогою правила трьох сигм.

На основі підготованих даних була побудована кореляційна матриця, яка дозволяє виявити найбільш значущі змінні (рис.4).

Моделювання. Для побудови моделей регресії методами машинного навчання початкова вибірка була поділена на навчальну та тестову у пропорції 80/20. Для аналізу було використано шість різних методів, кожен з яких буде розглянутий нижче.

Спочатку розглянемо лінійні моделі. Виходячи з рис.4, найбільше з ціною корелює змінна cleaning_fee, використаємо цей фактор для побудови однофакторної моделі лінійної регресії.

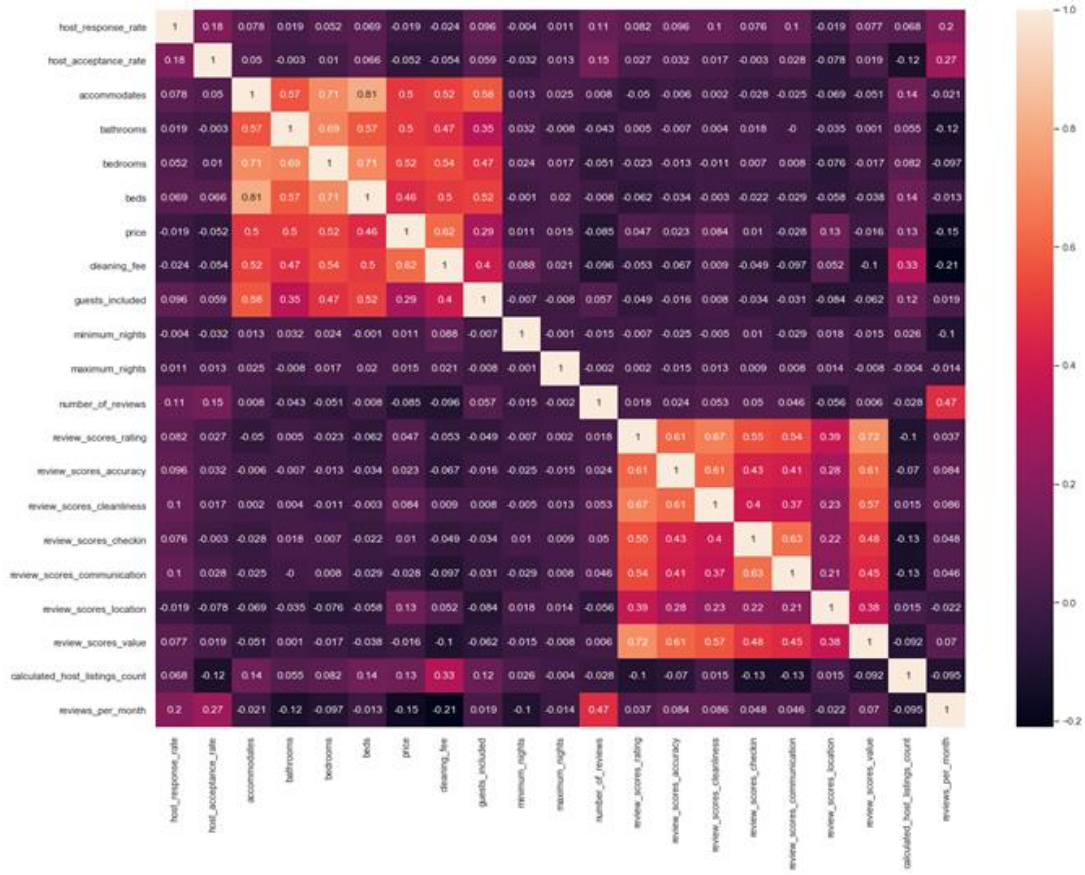


Рис.4. Кореляційна матриця

Джерело: власні розрахунки

Як можна побачити на графіку (рис.5 а), моделі вдалося знайти деяку тенденцію.

Для моделі множинної регресії були використані всі змінні. Це помітно відобразилося на якості моделі (рис.5 б).

Для моделі поліноміальної регресії ми взяли змінні з найбільшими коефіцієнтами кореляції – accommodates, bedrooms, cleaning_fee. Тут можна помітити, що прогнозні значення в середньому нижче, ніж реальні (рис.5 в).

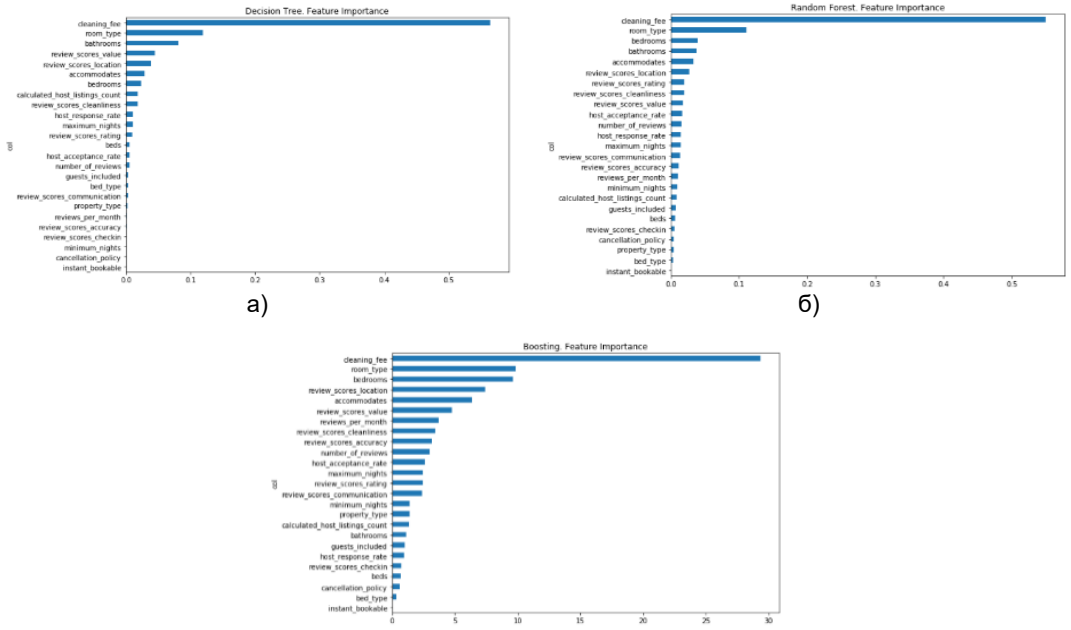
Для побудови нелінійних моделей (дерева рішень, випадкового лісу та бустингу) було обрано найбільш значущі змінні – cleaning_fee, room_type, bedrooms (рис.6).

Прогнозні значення моделей дерева рішень (рис.7 а) та бустингу (рис.7 в) дуже чутливі до нетипових спостережень, результати моделі випадкового лісу мають усереднений вигляд (рис.7 б).

а) б) в)

Рис.5. Лінійні моделі (результати моделювання на тестових вибірках)

Джерело: власні розрахунки



а) б)

в)

Рис.6. Значущість змінних: а) дерева рішень, б) випадковий ліс, в) бустинг

Джерело: власні розрахунки

а) б) в)

Рис.7. Нелінійні моделі (результати моделювання на тестових вибірках)

Джерело: власні розрахунки

Порівнюючи нелінійні моделі, можемо зафіксувати, що у всіх є ефект перенавчання, на це нам вказує метрика MSE (табл.1).

Таблиця 1

Метрики моделей

	Однофакторна лінійна регресія	Багатофакторна лінійна регресія	Поліноміальна регресія	Дерево рішень	Випадковий ліс	Бустінг
MSE (train)	8079.70	5360.79	5423.06	1722.53	1377.03	1366.44
MSE (test)	7209.08	5395.73	5005.85	3670.94	2456.20	2187.53
RMSE (train)	89.89	73.22	73.64	41.50	37.11	36.97
RMSE (test)	84.91	73.46	70.75	60.59	49.56	46.77
R^2 (train)	0.47	0.64	0.59	0.80	0.86	0.86
R^2 (test)	0.50	0.68	0.60	0.68	0.76	0.77

Джерело: власні розрахунки

Висновок. В роботі запропоновано алгоритм виявлення неявних факторів ціноутворення на ринку короткострокової оренди з використанням методів машинного навчання, що включає: 1) збір та первинну обробку даних; 2) побудову та аналіз моделей лінійної регресії; 3) побудову та аналіз моделей нелінійної регресії. Дослідження проведено на прикладі об'яв з сайту Airbnb у штатах Вашингтон та Нью-Йорк з використанням програмних скриптів, які розроблено на Python.

Побудовано та проаналізовано наступні моделі: однофакторна лінійна регресія, багатофакторна лінійна регресія, поліноміальна регресія, дерева рішень, випадковий ліс та бустинг. Результати дослідження показали, що найважливішими є фактори accommodates, cleaning_fee, room_type, bedrooms. Проте, виходячи з показників якості моделювання, отримані моделі не можна використовувати для впровадження: лінійні моделі мають невисоку якість, тоді як випадковий ліс, бустінг та дерева перенавчені. Але при проведенні бізнес-аналізу ці моделі можуть використовуватися для:

1. Оцінювання диференціальної ренти 1 та диференціальної ренти 2. А саме, скільки грошей заплатять за більш респектабельний район чи близькість до транспорту, а скільки платять за додаткову ванну кімнату.
2. Розподілення дорогих та недорогих апартаментів за районами та аналізу їх концентрації.
3. Обґрунтування рішень про те, яку інформацію треба вказувати орендодавцю для того, щоб підвищувати ціни.

Література

1. Airbnb Beat Expedia in Booked Room Nights. URL: https://uk.finance.yahoo.com/news/airbnb-beat-expedia-booked-room-180052599.html. (дата звернення: 13.12.2020).
2. Businessofapps. URL: https://www.businessofapps.com/data/airbnb-statistics/. (дата звернення: 13.12.2020).
3. Charkov M., Newman R., Overgoor J. Location Relevance at Airbnb. *Airbnb Engineering & Data Science.* 2013. URL: https://medium.com/airbnb-engineering/location-relevance-at-airbnb-12c004247b07#.vtj3t52mm. (дата звернення: 14.12.2020).
4. Kaggle. URL: https://www.kaggle.com/datasets. (дата звернення: 13.12.2020).
5. McCarthy N. Is Airbnb Really Cheaper Than A Hotel Room In The World's Major Cities? *Forbes*. 2018. URL: https://www.forbes.com/sites/niallmccarthy/2018/01/23/is-airbnb-really-cheaper-than-a-hotel-room-in-the-worlds-major-cities-infographic/?sh=1df2c5c978ac. (дата звернення: 15.12.2020).
6. Techcrunch. URL: https://techcrunch.com/2014/04/18/airbnb-has-closed-its-500m-round-of-funding-at-a-10b-valuation-led-by-tpg/. (дата звернення: 14.12.2020).

7. The World's Cities in 2016. URL: http://www.un.org/en/development/desa/population/publications/pdf/urbanization/the_worlds_cities_in_2016_data_booklet.pdf. (дата звернення: 15.12.2020).
8. Доходи Airbnb у 2020 році. . URL: https://fortune.com/2017/02/15/airbnb-profits/. (дата звернення: 12.12.2020)
9. Офіційний сайт статистики Airbnb. URL: http://airbnbstats.com/. (дата звернення: 11.12.2020).
10. Сайт компанії Airbnb. URL: https://news.airbnb.com/about-us/. (дата звернення: 15.12.2020).

References

1. Airbnb Beat Expedia in Booked Room Nights. Retrieved from https://uk.finance.yahoo.com/news/airbnb-beat-expedia-booked-room-180052599.html.
2. Businessofapps. Retrieved from https://www.businessofapps.com/data/airbnb-statistics/.
3. Charkov, M., Newman, R., Overgoor, J. (2013). Location Relevance at Airbnb. *Airbnb Engineering & Data Science.* Retrieved from: https://medium.com/airbnb-engineering/location-relevance-at-airbnb-12c004247b07#.vtj3t52mm
4. Kaggle. Retrieved from https://www.kaggle.com/datasets
5. McCarthy, N. (2018). Is Airbnb Really Cheaper Than A Hotel Room In The World's Major Cities? *Forbes*. Retrieved from https://www.forbes.com/sites/niallmccarthy/2018/01/23/is-airbnb-really-cheaper-than-a-hotel-room-in-the-worlds-major-cities-infographic/?sh=1df2c5c978ac.
6. Techcrunch. Retrieved from: https://techcrunch.com/2014/04/18/airbnb-has-closed-its-500m-round-of-funding-at-a-10b-valuation-led-by-tpg/.
7. The World's Cities in 2016. Retrieved from http://www.un.org/en/development/desa/population/publications/pdf/urbanization/the_worlds_cities_in_2016_data_booklet.pdf.
8. Airbnb revenues in 2020. Retrieved from https://fortune.com/2017/02/15/airbnb-profits/.
9. Official site of Airbnb statistics. Retrieved from http://airbnbstats.com/.
10. Airbnb. Retrieved from https://news.airbnb.com/about-us/.

Vladyslav Bobrov
Student
V.N. Karazin Kharkiv National University
4 Svobody Sq., 61022, Kharkiv, Ukraine
E-mail: bobrov.vld@gmail.com, ORCID: https://orcid.org/0000-0001-9042-8227
Andriy Kim
Student
V.N. Karazin Kharkiv National University
4 Svobody Sq., 61022, Kharkiv, Ukraine
E-mail: andrykim05@gmail.com, ORCID: https://orcid.org/0000-0002-6879-2396

FORECASTING PRICES IN THE RENTAL HOUSING MARKET WITH MACHINE LEARNING METHODS

The study of pricing factors in the market of the short-term rental has been done. Airbnb was chosen as the object of the study; it is a platform for accommodation, search, and rental around the world. At the beginning of 2021, the company offers 7 million homes from more than 220 countries. The Data Science methods play a significant role in the company's success. One of the key algorithms of the company is the pricing algorithm. Using the "Price Recommendations" feature, the homeowner can analyze which dates are most likely to be booked at the current price and which are not, it helps form a favorable offer. The system calculates the recommended cost of housing based on hundreds of parameters, some of which are easy to recognize, but there are less obvious factors that can also affect demand. The paper proposes an algorithm for identifying implicit pricing factors in the short-term rental market using machine learning methods, which includes: 1) data mining and data preparation; 2) building and analysis of linear regression models; 3) building and analysis of nonlinear regression models. The study was based on ads from the Airbnb site in Washington and New York using scripts developed in Python. The following models are built and analyzed: simple linear regression, multiple linear regression, polynomial regression, decision trees, random forest, and boosting. The results of the study showed that the most important factors are accommodates, cleaning_fee, room_type, bedrooms. But based on the model

evaluation criteria, they cannot be used for implementation: linear models are of low quality, while the random forest, boosting, and trees are overfitted. Still the results can be used in conducting business analysis.

Keywords: pricing factors, machine learning, regression, Airbnb.
JEL Classification: C55, E37, L85.

В.Д. Бобров
студент
Харьковский национальный университет имени В.Н. Каразина
пл. Свободы, 4, г. Харьков, 61022, Украина
E-mail: bobrov.vld@gmail.com, ORCID: https://orcid.org/0000-0001-9042-8227
А.В. Ким
студент
Харьковский национальный университет имени В.Н. Каразина
пл. Свободы, 4, г. Харьков, 61022, Украина
E-mail: andrykim05@gmail.com, ORCID: https://orcid.org/0000-0002-6879-2396

ПРОГНОЗИРОВАНИЕ ЦЕН НА РЫНКЕ АРЕНДЫ ЖИЛЬЯ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

В работе проведено исследование факторов ценообразования на рынке краткосрочной аренды жилья. В качестве объекта исследования выбрана компания Airbnb, которая представляет собой площадку для размещения, поиска и аренды жилых помещений по всему миру. На начало 2021 г. компания насчитывает 7 миллионов жилых помещений из более чем 220 стран мира. Немалую роль в успехе компании играет использование методов Data Science. Одним из ключевых алгоритмов компании является алгоритм ценообразования. С помощью функции «Рекомендации по ценам» владелец жилья может проанализировать какие даты скорее всего будут забронированы по текущей цене, а какие нет, и сформировать выгодное предложение. Система вычисляет рекомендованную стоимость жилья на основании сотен параметров, некоторые из которых легко распознать, но есть и менее очевидные, которые также могут влиять на спрос. В работе предложен алгоритм выявления неявных факторов ценообразования на рынке краткосрочной аренды с использованием методов машинного обучения, который включает: 1) сбор и первичную обработку данных; 2) построение и анализ моделей линейной регрессии; 3) построение и анализ моделей нелинейной регрессии. Исследование проведено на примере объявлений с сайта Airbnb в штатах Вашингтон и Нью-Йорк с использованием программных скриптов, разработанных на Python. Построены и проанализированы следующие модели: простая линейная регрессия, множественная линейная регрессия, полиномиальная регрессия, деревья решений, случайный лес и бустинг. Результаты исследования показали, что наиболее важными являются факторы accommodates, cleaning_fee, room_type, bedrooms. Но исходя из показателей качества моделирования, полученные модели нельзя использовать для внедрения: линейные модели имеют невысокое качество, тогда как случайный лес, бустинг и дерева переобучены. Но полученные результаты могут использоваться при проведении бизнес-анализа.

Ключевые слова: факторы ценообразования, машинное обучение, регрессия, Airbnb.
JEL Classification: C55, E37, L85.