МРНТИ 20.23.25
УДК 004.852

10.51889/2959-5894.2023.83.3.016

Ж.С. Есенгалиева [1], А.С. Есенгалиева [2], Р.Б. Биктимир [1], С.С. Есенғали [3]*

[1]Евразийский национальный университет им. Л.Н.Гумилева, г.Астана, Казахстан
[2]Казахстанский филиал Московского государственного университета им.М.В. Ломоносова, г.Астана, Казахстан
[3]Республиканская физико-математическая школа, г.Астана, Казахстан
[*]e-mail: jannayess@gmail.com

УПРАВЛЕНИЕ БОЛЬШИМИ ДАННЫМИ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Аннотация

В статье предложен ансамбль алгоритмов машинного обучения и программные результаты, включающие такие методы управления большими данными как регрессия, классификация и кластеризация. Предложенные методы в сравнении позволяют анализировать и интерпритировать полученные данные с реальными обстоятельствами на рынке недвижимости. В качестве данных рассматриваются сведения о недвижимости в столице Казахстана. Большие данные структурированы по таким полям как стоимость, классность, размер кухонного помещения, площадь и представляются в виде файла с расширением .csv, обрабатываются с помощью методов машинного обучения. В качестве среды программирования использован Python, при этом библиотеки numpy, pandas, matplotlib, Axes3D, LinearRegression, Scikit-learn, KMeans позволяют интерпретировать и визуализировать полученные данные. Проведенный вычислительный эксперимент наглядно демонстрирует классификацию данных, разделение на кластеры, а также формирует прогноз по стоимости в зависимости от заявленных признаков.

Ключевые слова: линейная регрессия, классификация, кластеризация, большие данные, дерево решений.

Аңдатпа

Ж.С. Есенгалиева [1], А.С. Есенгалиева [2], Р.Б.Биктимир [1], С.С. Есенғали [3]
[1]Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан
[2]М.В. Ломоносов атындағы Мәскеу мемлекеттік университетінің Қазақстан филиалы, Астана қ., Қазақстан
[3]Республикалық физика-математика мектебі, Астана қ., Қазақстан

МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІ АРҚЫЛЫ ҮЛКЕН ДЕРЕКТЕРДІ БАСҚАРУ

Мақалада үлкен деректерді басқару әдістерін қамтитын регрессия, классификация және кластерлеу сияқты машиналық оқыту алгоритмдерінің ансамблі және программалық нәтижелер ұсынылған. Ұсынылған әдістер салыстырмалы түрде алынған мәліметтерді жылжымайтын мүлік нарығындағы нақты жағдайлармен талдауға және түсіндіруге мүмкіндік береді. Деректер ретінде Қазақстан астанасындағы жылжымайтын мүлік туралы мәліметтер қарастырылады. Үлкен деректер құны, классы, ас үй кеңістігінің өлшемі, ауданы бойынша құрылымдалған және .csv кеңейтімі бар файл ретінде ұсынылған, машиналық оқыту әдістері арқылы өңделеді. Python бағдарламалау ортасы ретінде пайдаланылды, numpy, pandas, matplotlib, Axes3D, LinearRegression, Scikit-learn, KMeans кітапханалары арқылы алынған деректерді интерпретациялау және визуализациялау жасалды. Жүргізілген есептеу эксперименті деректерді классификациялауға, кластерлерге бөлуге, сондай-ақ мәлімделген белгілерге байланысты құны бойынша болжам жасауға мүмкіндік береді.

Түйін сөздер: сызықтық регрессия, классификация, кластерлеу, үлкен деректер, шешім ағашы.

Abstract

MANAGING BIG DATA WITH MACHINE LEARNING METHODS

Yessengaliyeva J.S.[1], Yessengaliyeva A.S.[2], Biktimir R.B. [1], Yessengali S.S.[3]
[1]Eurasian National University, Astana, Kazakhstan
[2] Kazakhstan branch of M. V. Lomonosov Moscow State University, Astana, Kazakhstan
[3]Republican Physics and Mathematics School, Astana, Kazakhstan

The article proposes an ensemble of machine learning algorithms and program results, including such big data management methods as regression, classification and clustering. The proposed methods in comparison allow to analyze and interpret the obtained data with the real circumstances in the real estate market. Information about real estate in the capital of Kazakhstan is considered as data. Big data is structured by such fields as cost, class, kitchen space size, area and is presented as a .csv file, processed using machine learning methods. Python was used as a programming environment, while the numpy, pandas, matplotlib, Axes3D, LinearRegression, Scikit-learn, KMeans libraries allow you

to interpret and visualize the received data. The conducted computational experiment clearly demonstrates the classification of data, division into clusters, and also forms a forecast for the cost depending on the declared features.

Keywords: linear regression, classification, clustering, big data, decision tree.

Введение

Современный мир уже невозможно представить без больших данных, их обработки, обучения с целью получения результатов прогноза, классификации объектов по группам, кластеризации. Рынок недвижимости в Казахстане и в мире в целом, ежедневно меняется, динамично развивается, пополняется данными и является актуальной темой для изучения в статистическом плане, прогноза стоимости, также затрагивает другие отрасли и явления как строительство, товарооборот, дизайн, демографический рост и др. Множество сайтов, редакций, социальных сетей предоставляют данные о недвижимости. Поток данных ежедневно обновляется, обрабатывается, увеличивается, при этом требует структурированной выходной информации для конечного пользователя. Извлекаемые большие данные при правильной обработке приносят значительный инвестиционный эффект, позволяющий более выгодно использовать имеющиеся ресурсы.

Однако, прежде чем приступить к обработке больших данных с целью получения достоверной своевременной и необходимой информации следует провести разведочный анализ данных, то есть наблюдаемые данные должны быть представлены в определенной форме, доступной для выявления каких-либо закономерностей, моделирования прогнозных решений. К примеру, в работе И. Керчева и др.[1] для разработки сверточной нейронной сети проведен разведочный анализ данных из полученных эталонных карт сегментации. Выделены признаки сегментов изображения, предложена модель нейронной сети для попиксельной классификации.

Объем данных продолжает расти, и как исследователям, так и компаниям требуются инструменты, помогающие анализировать и понимать эти данные, большая часть которых представлена в виде неструктурированных данных [2]. Правительственные учреждения и исследовательские группы стремятся обрабатывать и понимать данные, которые оказывают определенное влияние на развитие социальных, экономических, научно-технологических процессов.

Авторы в своих трудах [3] рассматривают многомерный анализ данных, основными особенностями которого являются возможность учитывать различные типы переменных (количественные или категориальные), различные типы структур данных (раздел переменных, иерархия переменных, раздел отдельных лиц) и наконец, дополнительная информация (дополнительные лица и переменные). Более того, параметры, полученные в результате различных анализов разведочных данных, могут автоматически описываться количественными и/или категориальными переменными. Предложенный многомерный анализ данных позволяет систематизировать неструктурированную информацию.

После этапа проведения раведочного анализа данных следует определить алгоритмы машинного обучения для дальнейшей обработки с целью получения искомой информации. Для поставленной цели предложены такие инструменты как регрессия, классификация и кластеризация. Методы машинного обучения позволяют осуществить прогноз, разделить извлеченные данные по класстерам и классам.

В статье З. Ахмад, М. Мансуровой [4] представлена модель машинного обучения для оценки значительной высоты океанской волны с целью прогнозирования состояния океана с помощью регрессии на основе метода опорных векторов. В предложенной модели вычислена среднеквадратическая ошибка, равная 0,044, приближенная к нулю, что свидетельствует об адекватности модели и проведенных вычислений.

В трудах А. Мифтаховой [5] описано применение метода дерева решений в задачах классификации и прогнозирования. Метод дерева решений, основанный на определенных решающих правилах структурирует большие данные по определенной иерархии.

М.Тиндова [6] описывает кластеризацию как набор точек для выявления неравномерностей в их распределении по пространству. Под «неравномерностями» понимаются сгущения точек и образование ими кластеров. Для решения этой традиционной задачи распознавания образов без учителя предложено множество подходов, базирующихся на теоретико-вероятностной модели либо на некоторой правдоподобной эвристике. В статье [7] предложены алгоритмы параметрической кластеризации на основе центроидов, такие как классические средние значения, алгоритм Линде-Бузо-Грея (LBG) и теоретико-информационная кластеризация, которые возникают в результате специального выбора дивергенции Брегмана. Алгоритмы сохраняют простоту и масштабируемость классического алгоритма kmeans, в то же время обобщая метод на большой класс функций потерь кластеризации.

С появлением больших данных в настоящее время во многих приложениях доступны базы данных, содержащие большое количество похожих временных рядов. Прогнозирование временных рядов в этих областях с помощью традиционных процедур одномерного предвидения оставляет неиспользованными большие возможности для получения точных прогнозов. Рекурентная нейронная сеть вместе с различными алгоритмами кластеризации, такими как kMeans, DBScan, Partition Around Medoids (PAM) и Snob позволяет достичь конкурентоспособных результатов при сравнительном анализе наборов данных в соответствии с процедурами оценки конкуренции [8].

На основе вышеописанных методов сформирован ансамбль алгоритмов машинного обучения, позволяющий наиболее точно определить прогноз, категориальность по имеющимся спарсенным данным.

Методы

Как описано выше, для запуска тех или иных моделей машинного обучения нужны данные. Обработка исходных данных является важным этапом в построении систем машинного обучения. Проведен разведочный анализ данных, при этом выявлены признаки, на которых будет обучаться модель. Сформированная датасет с помощью регрессии позволит получить прогнозную модель в рамках визуальной интерпретации, а также классифицировать и кластеризировать данные по признакам. Каждый представленный алгоритм имеет свою область применения. Ансамбли алгоритмов – наборы моделей для решения одной и той же задачи способствуют повышению точности модели.

Сначала рассмотрим регрессию с двумя признаками, затем с тремя признаками, по отдельности. В качестве признаков, так называемых фичей, определим переменные dim_1 – площадь квартиры, dim_2 – площадь кухонного помещения, price – цена. Следовательно, при использовании регрессии с двумя фичами получим зависимость согласно формуле 1:

$$price(dim\_1) = a^* dim\_1 + b. \quad (1)$$

В случае прогноза по трем признакам, имеем формулу 2:

$$price(dim\_1, dim\_2) = a + b1^* dim\_1 + b2^* dim\_2. \quad (2)$$

Далее, проведем классификацию. Точность классификации дерева решений (без ущерба для интерпретации и с сохранением структуры узлов и листьев) можно значительно улучшить модель с помощью адаптации параметров [7].

Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение [5]. Цель процесса построения дерева принятия решений – создать модель, по которой можно было бы классифицировать случаи и решать, какие значения может принимать целевая функция, имея на входе несколько переменных [5]. Построение модели основано на больших данных из сайта недвижимости по Республике Казахстан, а именно krysha.kz [9]. Зная параметры недвижимости, на основе входных данных можно выяснить его классность. Построим алгоритм классификации на основе дерева решений, где по определенным решающим правилам производится разделение на классы.

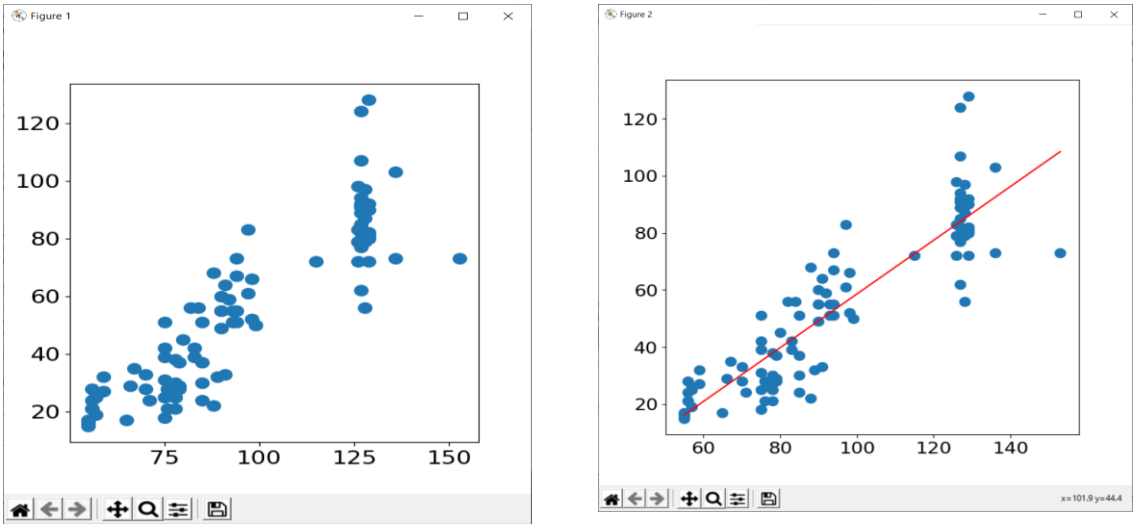
Какие бывают классы жилья? В РК официально существует четыре класса комфортности жилья. Но это вовсе не эконом, комфорт, бизнес и элит. В СНиПах они обозначаются римскими цифрами от I до IV. Сопоставить их можно так: первый класс – элит, второй – бизнес, третий – комфорт, четвёртый – эконом. По данным Krisha.kz [9], жильё IV класса комфортности обычно на 25–35 % дешевле III класса. Разница стоимости квадратного метра в эконом классе относительно II и I класса (бизнес и элит) достигает 50–80 % и более. Кроме того, особым спросом пользуются такие сопоставления как бизнес, комфорт+, комфорт. В связи с чем, за основу одного из фичей взяты данные категории.

Дерево, по сути, это вопрос. Один вопрос ведет к другому, пока не получим последний вопрос с искомым ответом. Библиотеки Python [10,11] позволяют произвести данный алгоритм. Здесь целевые признаки feature_0=> dim_1, a feature_1=> dim_2 соответствуют определенным фичам.

В предложенном ансамбле алгоритмов следующим является кластеризация. Здесь распределим данные по кластерам вышеупомянутого сопоставления. Разделение по кластерам позволяет сгруппировать данные по определенным критериям, является методом машинного обучения без учителя. На основе изложенных научных методов машинного обучения проведен вычислительный эксперимент, результаты которого представлены ниже.

Результаты и обсуждение

Проведен прогноз стоимости недвижимости по квадратуре и классности жилья. За образец взяты данные трехкомнатных квартир в г.Астана, Республика Казахстан. В качестве примера рассмотрен файл с расширением .csv и данными, извлеченными с сайта krysha.kz. Обучению подлежат непосредственно данные трехкомнатных квартир (г.Астана, Казахстан). На рис.1А) представлены обученные данные, на рис.1Б) высчитана регрессия, которая позволяет спрогнозировать стоимость недвижимости посредством регрессии (красной линии).

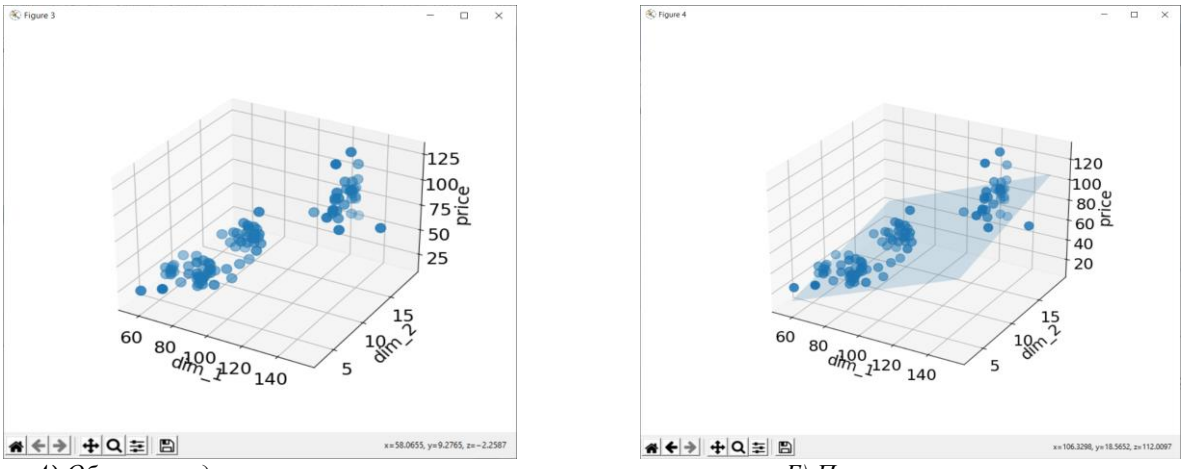


А) Обученные данные *Б) Линия регрессии*

Рисунок 1. Обучение с помощью регрессии

Здесь линия регрессии предполагает стоимость недвижимости по признакам площади и стоимости жилья. Полученные данные имеют общие признаки и коррелируют с математической моделью. При этом есть некоторые выбросы, в реальности такие экземпляры имеют место. По данной модели можно прогнозировать стоимость по площади.

Далее на рис. 2А) представлены обученные данные по трем признакам: dim_1 – площадь квартиры, dim_2 – площадь кухонного помещения, price – цена, на рис. 2Б) высчитана регрессия, которая позволяет спрогнозировать стоимость недвижимости посредством плоскости.



А) Обученные данные по трем признакам *Б) Плоскость регрессии*

Рисунок 2. Обучение с помощью регрессии по трем признакам

Плоскость регрессии предполагает стоимость недвижимости по признакам площади трехкомнатной квартиры, площади кухни и стоимости. Трехмерное пространство позволяет более наглядно

визуализировать полученные обработанные большие данные посредством методов машинного обучения в программной среде Python с использованием соответствующих библиотек.

Следующим результатом правильности суждений является классификация полученных данных по дереву решений. На рис.3 представлено распределение заявленных площадей по классности, а именно: бизнес, комфорт+, комфорт. Именно метод классификации является основополагающим в данном ансамбле алгоритмов. Посредством метода DecisionTreeClassifier в данном случае представлено дерево решений, которое являясь алгоритмом машинного обучения реализуется на основе правил образующих определенную иерархию. Данный алгоритм в вычислительном эксперименте основан на количественных признаках, взятых из .csv файла.

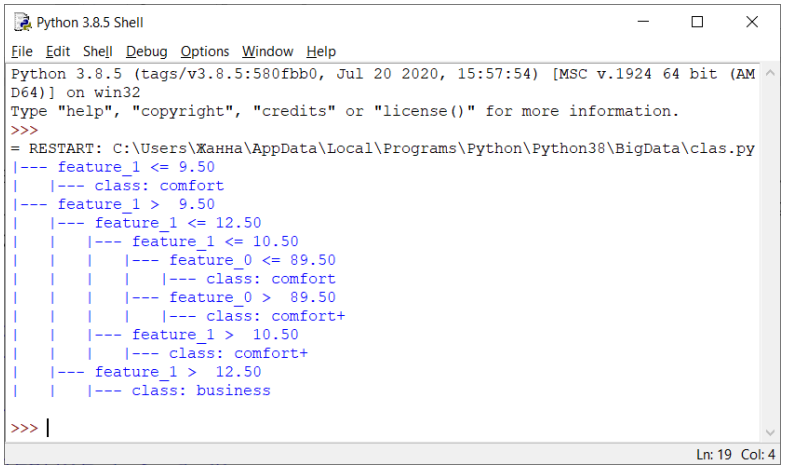
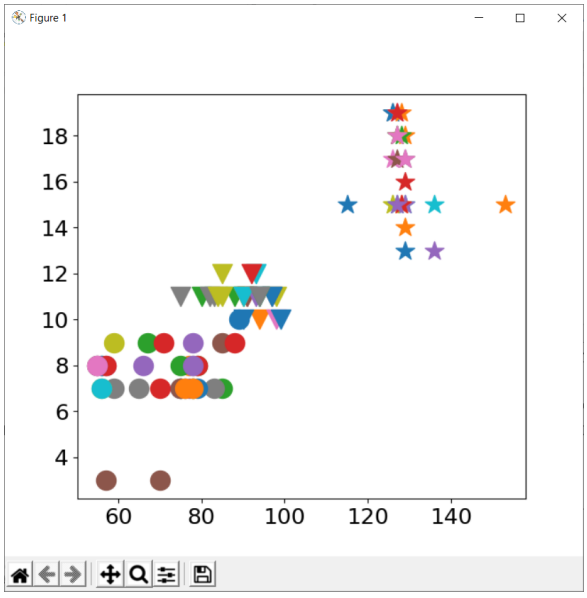


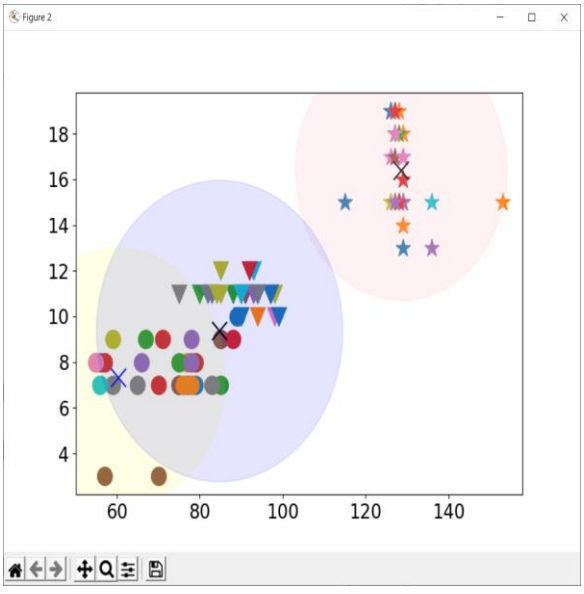
Рисунок 3. Машинное обучение. Классификация по дереву решений

В процессе исследования нами получены различные результаты множества дерева решений, где было сложно интерпретировать искомые данные. Подбирались различные фичи, и именно в разрезе трехкомнатных квартир с соответствующими площадями нами получено читабельное дерево решений, с определенными правилами по заявленным признакам.

В итоге, согласно метода машинного обучения, а именно полученного дерева решений на основе алгоритма DecisionTreeClassifier в среде Python проведен вычислительный эксперимент. Согласно результатам, если площадь кухни меньше или равна 9,5 кв.м., то недвижимость относится к классу комфорт. В диапазоне признаков 9<feature_1<=10.50, при feature_0<89,5 мы получаем также класс комфорт и т.д. Таким образом, при площади кухни больше 12,5 кв.м. мы получаем класс бизнес.



А) Кластеры



Б) Кластеры с выделенными областями

Рисунок 4. Машинное обучение. Кластеризация

В вычислительном эксперименте использована кластеризация полученных данных. На рис.4А) показаны сгруппированные кластеры: бизнес, комфорт+, комфорт. На рис.4Б) полученные кластеры объединены в соответствующие области по итогам проведенной кластеризации согласно критериям стоимости и площади кухонного помещения.

С помощью библиотеки Kmeans существует возможность сформировать определенные кластеры, а также центроиды. Данная процедура позволяет сгенерировать начальные центроиды кластера, на основе эмпирического распределения вероятностей вклада точек в общую инерцию с

В нашем случае мы имеем массив с параметрами (dim_1, dim_2) и получаем начальные центры.

Вызов библиотеки происходит по команде *from sklearn.cluster import Kmeans*. Здесь мы импортируем модуль, отвечающий за кластеризацию. Выгружаем признаки в отдельную переменную, затем создаем модель для кластеризации с помощью метода *fit:*

clust = KMeans(n_clusters=3).fit(X).

Метод fit(X[y, вес_выборки]) позволяет вычислить кластеризацию k-средних [11]. Для определения центра кластеров используем *clust.cluster_centers_*. Реализация областей позволяет визуально выделить кластеры, в нашем случае разделение происходит с помощью маркеров и библиотеки mathplotlib:

```
markers = {"comfort": "o", "comfort+": "v", "business": "*"}
plt.scatter(c1[0], c1[1], s=250, marker="x", c="blue")
plt.scatter(c1[0], c1[1], s=250 * 2e2, c="yellow", alpha=0.1).
```

Метрики качества машинного обучения для регрессии позволяют определить адекватность модели. Для оценки качества модели использовали коэффициент корреляции, также в научной литературе его называют коэффициентом детерминации. Кроме того, можно использовать среднюю квадратичную ошибку. Среда Python располагает соответсвующим функционалом по расчету данных метрик качества.

На рисунке 5 представлен файл с расширением .csv, данные с файла подгружаются в программу посредством библиотек с помощью команды *houses = pd.read_csv("houses.csv").*

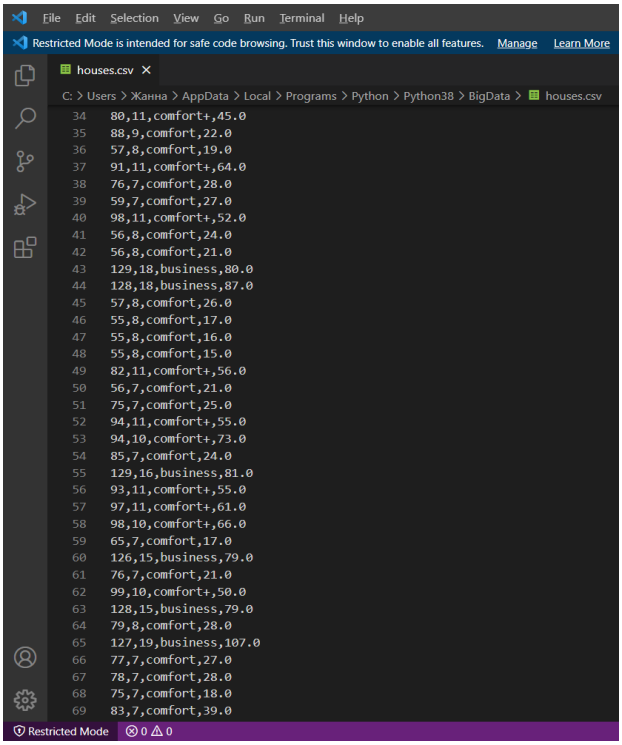


Рисунок 5. Фрагмент извлеченных данных

Таким образом, исследованы алгоритмы и методы машинного обучения. Проведен анализ полученных данных, который позволяет классифицировать, разделить на кластеры и прогнозировать большие данные, извлеченные с сайта посредством библиотеки BeautifulSoap с расширением .csv. Кроме того, разработано приложение в среде Python, которое осуществляет прогноз с применением ансамбля алгоритмов, базирующегося на регрессии, классификации, кластеризации.

Выводы

В статье проведен библиографический обзор в области методов машинного обучения. Рассмотрен разведочный анализ больших данных, освещены материалы по регрессии, классификации и кластеризации. Теоретически и практически изучены предложенные методы машинного обучения.

По полученным данным, на основе отобранных признаков выведены математические модели, разработаны программы в среде Python и объединены в единое приложение. Изучены библотеки среды программирования Python, такие как numpy, pandas, matplotlib, Axes3D, LinearRegression, Scikit-learn, KMeans позволяют интерпретировать и визуализировать полученные данные.

Проведены вычислительные эксперименты, результаты которых представлены в данной статье. Полученные результаты позволяют обрабатывать большие данные, моделировать рыночную ситуацию недвижимости в определенном сегменте и визуализировать в виде дерева решений, регрессии, областей кластеров. Представленный ансамбль алгоритмов машинного обучения наглядно демонстрирует адекватность моделей и результатов.

Ансамбль алгоритмов указывает на подход, при котором выполняется несколько методов, и их согласованная визуализация используется в качестве окончательной интерпретации. Ансамблевые решения лучше, чем простые модельные решения. Ансамбль считается лучшим ансамблем, если его члены являются действительными или качественными и если они участвуют в соответствии со своими качествами в построении консенсусной кластеризации, классификации, регрессии.

В статье предложена структура ансамбля алгоритмов, в которой используется алгоритм кластеризации, основанный на алгоритме кластеризации kmeans. Наш алгоритм кластеризации гарантирует, что обнаруженные кластеры действительны [13]. Структура ансамбля алгоритмов использует механизм для использования каждого обнаруженного кластера, класса, регрессии в соответствии с его качеством.

В перспективе планируется исследование полиномиальной регрессии с целью получения прогнозной модели оценки стоимости недвижимости. Также в качестве объектов исследования будут рассмотрены квартиры с различной комнатностью. Извлечение больших данных с веб-ресурсов играет важную роль для обработки с помощью методов машинного обучения. В этой связи планируется адаптация полученной модели для больших данных из других отраслей экономики, информационно-коммуникационных технологий.

Исследование представляет научный, экономический, маркетинговый интерес и эффективно для населения.

Список использованной литературы:

1 Kerchev I.A., Maslov K.A., Markov N.G., Tokareva O.S. Semantic segmentation of damaged fir trees in unmanned aerial vehicle images [Семантическая сегментация повреждённых деревьев пихты на снимках с беспилотных летательных аппаратов] (2021) Sovremennye Problemy Distantsionnogo Zondirovaniya Zemli iz Kosmosa, 18 (1), pp. 116 – 126. DOI: 10.21046/2070-7401-2021-18-1-116-126 Scopus

2 Edwards A., Sullivan M., Itkowsky E., Weinberg D. TextQ—A User Friendly Tool for Exploratory Text Analysis (2021) Information (Switzerland), 12 (12), art. no. 508 DOI: 10.3390/info12120508 Scopus

3 Lê S., Josse J., Husson F. FactoMineR: An R package for multivariate analysis (2008) Journal of Statistical Software, 25 (1), pp. 1 – 18 DOI: 10.18637/jss.v025.i01

4 Ahmad, Z., & Mansurova, M. (2021). Machine learning approach to predict significant wave height. Journal Of Mathematics, Mechanics And Computer Science, 110(2), 87-96. doi:10.26577/JMMCS.2021.v110.i2.08

5 Мифтахова, А. А. Применение метода дерева решений для решения задач классификации и прогнозирования / А. А. Мифтахова // Инфокоммуникационные технологии. – 2016. – Т. 14. – № 1. – С. 64-70. – DOI 10.18469/ikt.2016.14.1.10. – EDN WDCYVR.

6 Тиндова М. Г. Предварительная кластеризация многомерных объектов в интеллектуальном анализе данных // Вестник Саратовского государственного социально-экономического университета. 2008. – №. 4. – С. 137-138.

7 Абрамова, Т. В. Улучшение точности классификации методом обратного распространения параметра по иерархическому нечеткому дереву решений / Т. В. Абрамова // Актуальные проблемы современной науки, техники и образования. – 2020. – Т. 11. – № 1. – С. 126-127. – EDN CWUUCW.

8 Bandara K., Bergmeir C., Smyl S. Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach (2020) Expert Systems with Applications, 140, art. no. 112896

9 Электронный ресурс: https://krisha.kz/. Дата обращения: 14.11.2022

10 Сузи Р. А. Язык программирования Python //М.: Бином. Лаборатория знаний. – 2006.

11 Бэрри П. Изучаем программирование на Python. – Litres, 2019.

12 Электронный ресурс: scikit-learn: machine learning in Python — scikit-learn 1.3.1 documentation Дата обращения: 30.11.2022

13 Niu H., Khozouie N., Parvin H., Alinejad-Rokny H., Beheshti A., Mahmoudi M.R. An ensemble of locally reliable cluster solutions (2020) Applied Sciences (Switzerland), 10 (5), art. No. 1891

References:

1 Kerchev I.A., Maslov K.A., Markov N.G., Tokareva O.S. Semantic segmentation of damaged fir trees in unmanned aerial vehicle images [Semanticheskaya segmentaciya povrezhdyonnyh derev'ev pihty na snimkah s bespilotnyh letatel'nyh apparatov] (2021) Sovremennye Problemy Distantsionnogo Zondirovaniya Zemli iz Kosmosa, 18 (1), pp. 116 – 126. DOI: 10.21046/2070-7401-2021-18-1-116-126 Scopus (In Russian)

2 Edwards A., Sullivan M., Itkowsky E., Weinberg D. TextQ—A User Friendly Tool for Exploratory Text Analysis (2021) Information (Switzerland), 12 (12), art. No. 508 DOI: 10.3390/info12120508 Scopus

3 Lê S., Josse J., Husson F. FactoMineR: An R package for multivariate analysis (2008) Journal of Statistical Software, 25 (1), pp. 1 – 18 DOI: 10.18637/jss.v025.i01

4 Ahmad, Z., & Mansurova, M. (2021). Machine learning approach to predict significant wave height. Journal Of Mathematics, Mechanics and Computer Science, 110(2), 87-96. Doi:10.26577/JMMCS.2021.v110.i2.08

5 Miftahova, A.A. (2016) Primenenie metoda dereva reshenij dlja reshenija zadach klassifikacii i prognozirovanija [Application of the decision tree method to solve classification and forecasting problems]. A.A. Miftahova Infokommunikacionnye tehnologii. T. 14. № 1., 64-70. DOI 10.18469/ikt.2016.14.1.10. – EDN WDCYVR. (In Russian)

6 Tindova, M.G. (2008) Predvaritel'naya klasterizaciya mnogomernyh ob''ektov v intellektual'nom analize dannyh [Pre-clustering of multidimensional objects in data mining]. Vestnik Saratovskogo gosudarstvennogo social'no-ekonomicheskogo universiteta. № 4., 137-138. (In Russian)

7 Abramova, T.V. (2020) Uluchshenie tochnosti klassifikacii metodom obratnogo rasprostraneniya parametra po ierarhicheskomu nechetkomu derevu reshenij [Improving classification accuracy using the parameter backpropagation method over a hierarchical fuzzy decision tree]. Aktual'nye problem sovremennoj nauki, tekhniki i obrazovaniya. T. 11. № 1., 126-127. EDN CWUUCW. (In Russian)

8 Bandara, K., Bergmeir, C., Smyl, S. (2020) Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. Expert Systems with Applications, V.140, 112896.

9 Electronic resource: https://krisha.kz/. Date of the application: 14.11.2022

10 Suzi, R. A. (2006) Yazyk programmirovaniya Python [Python programming language]. M.: Binom. Laboratoriya znanij. (In Russian)

11 Berri, P. (2019) Izuchaem programmirovanie na Python [Learning Python Programming]. Litres. (In Russian)

12 Electronic resource: scikit-learn: machine learning in Python — scikit-learn 1.3.1 documentation Date of the application: 30.11.2022

13 Niu, H., Khozouie, N., Parvin, H., Alinejad-Rokny, H., Beheshti, A., Mahmoudi, M.R. (2020) An ensemble of locally reliable cluster solutions. Applied Sciences (Switzerland), 10 (5), 1891.