ОБРАБОТКА БОЛЬШИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ СРЕДСТВ ЯЗЫКА PYTHON

PROCESSING BIG DATA USING THE PYTHON LANGUAGE TOOLS

Карпов Даниил Константинович, бакалавр, Московский Государственный Технический Университет имени Н. Э. Баумана, г. Москва

Karpov D.K. dkkarpov@gmail.com

Аннотация

Постановка проблемы. Большие данные, неумолимо внедряются практически во все сферы жизни и пользователь повседневно сталкивается с необходимостью длительного анализа получаемой информации. Поэтому в последнее время активно развиваются способы упрощения возможностей добычи и анализа данных: RapidMiner, язык программирования R, методы *Data Mining*. Однако все вышеперечисленные способы довольно сложны для обычного пользователя. Автором данной статьи предлагается использовать небольшие парсеры для упрощения работы пользователя и обработки объемных сайтов с данными.

Цель. Расмотреть возможность использования инструментов языка python для обработки и анализа больших данных.

Практическая значимость. Предложенная методика может быть использована широким кругом пользователей, которым не обязательно быть программистами или инженерами. Такие скрипты могут быть легко настраиваемыми, динамично изменяющимися и простыми в обращении

Annotation

Problem statement. Big data is inexorably introduced into almost all spheres of life, and the user is daily faced with the need for a long analysis of the information received. Therefore, recently, ways to simplify data mining and analysis capabilities have been actively developed: RapidMiner, the R programming language, and Data Mining methods. However, all of the above methods are quite difficult for the average user. The author of this article suggests using small parsers to simplify the user's work and the processing of large sites with data.

Goal. To consider the possibility of using python tools for processing and analyzing big data.

Practical significance. The proposed method can be used by a wide range of users who do not have to be programmers or engineers. Such scripts can be easily configurable, dynamically changing, and easy to handle.

Ключевые слова: большие данные, Data Moning, парсинг, python разработка, анализ данных, машинное обучение.

Keywords: big data, Data Moning, parsing, python development, data analysis, machine learning.

1. Введение

Big Data, или «большие данные» по-русски - термин, появившийся совсем недавно - всего шесть лет назад. Но это не означает, что одновременно возникло и само явление. Большими данными принято называть большие объемы информации со сложной неоднородной и / или неопределенной структурой. Иногда о больших данных говорят как о неструктурированной информации, но это неверно - большие данные всегда имеют структуру, они могут быть сложными из-за того, что данные поступают из разных источников и содержат совершенно разную информацию или совершенно неизвестны. То есть, как правило, собрать эту стопку в одну таблицу не удается.

Большие данные (Big Data) - это структурированные и неструктурированные данные огромных объемов и разнообразия, а также методы их обработки, которые позволяют распределенно анализировать информацию.

Для аналитической обработки Больших Данных используется широкий спектр методов и алгоритмов. Это методы классов Data Mining (поиск ассоциативных правил, классификация, кластеризация и др.) и Machine Learning, искусственные нейронные сети и распознавание образов, имитационное моделирование, статический анализ и др.

Стоит заметить, что в России под термином «Big Data» подразумевают также технологии обработки, а в мире — лишь сам объект исследования.

2. Актуальность

Актуальность работы обусловлена тем, что количество источников данных стремительно растёт, а значит технологии их обработки становятся всё более востребованными.

Стоит отметить, что Большие данные, неумолимо внедряются практически во все сферы жизни. Не иметь подобающих инструментов для анализа и обработки этих данных, значит подтверждать свою беспомощность и несостоятельность в информационном обществе.

3. Функции и задачи больших данных

Когда говорят о Big Data, упоминают правило VVV — три признака или свойства, которыми большие данные должны обладать:

- Volume (объем)— данные измеряются по величине физического объема документов.
- Velocity (быстрота) — данные регулярно обновляются, что требует их постоянной обработки.

- Variety (разнообразие) — разнообразные данные могут иметь неоднородные форматы, быть неструктурированными или структурированными частично.

4. Методики анализа больших данных

Существует множество разнообразных методик анализа массивов данных, в основе которых лежит инструментарий, заимствованный из статистики и информатики (например, машинное обучение). Вот некоторые из них:

- ***методы класса Data Mining:*** изучение правил ассоциации (англ. association rule learning), классификация (методы классификации новых данных на основе принципов, ранее применявшихся к существующим данным), кластерный анализ, регрессионный анализ;

  Знания, обнаруженные в процессе Data Mining, должны быть нетривиальными и ранее неизвестными. етривиальность означает, что такие знания не могут быть обнаружены простым визуальным анализом. Они должны описывать взаимосвязи между свойствами коммерческих объектов, прогнозировать значения одних функций на основе других и т.д. Полученные знания должны применяться к новым объектам.
- ***краудсорсинг*** — категоризация и обогащение данных широким, неопределенным кругом лиц, работающих на основе публичной оферты без вступления в трудовые отношения.
- ***смешение и интеграция данных*** (англ. data fusion and integration) — набор методов, которые объединяют разнородные данные из разных источников для проведения углубленного анализа. Примеры таких методов, составляющих этот класс методов, включают в себя цифровую обработку сигналов и обработку естественного языка (включая тональный анализ);

- ***машинное обучение, включая обучение с учителем и без учителя, а также Ensemble learning*** (англ.) — использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей (англ. constituent models, ср. состатистическим ансамблем в статистической механике);

  Стоит отметить, что принято выделять 2 типа машинного обучения:

  - Индуктивное или по прецедентам, которое основано на выявлении эмпирических закономерностей во входных данных;
  - Дедуктивное, которое предполагает формализацию знаний экспертов и их перенос в цифровую форму в виде базы знаний.
- ***искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы***;
- ***распознавание образов;***
- ***прогнозная аналитика;***
- ***имитационное моделирование***;
- ***пространственный анализ*** (англ. Spatial analysis) — класс методов, использующих топологическую, геометрическую и географическую информацию в данных;
- ***статистический анализ***, в качестве примеров методов приводятся A/B-тестирование и анализ временных рядов;
- ***визуализация аналитических данных*** — представление информации в виде изображений, диаграмм с использованием интерактивных функций и анимации, как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

В рамках данной работы рассмотрим поподробнее класс Data mining.

5. Data Mining.

Data Mining – это процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных, практически полезных и интерпретируемых знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining является одним из шагов Knowledge Discovery in Databases.

К методам и алгоритмам Data Mining относятся:

- искусственные нейронные сети
- деревья решений, символьные правила
- методы ближайшего соседа и k-ближайшего соседа
- метод опорных векторов
- байесовские сети
- линейная регрессия
- корреляционно-регрессионный анализ
- иерархические методы кластерного анализа
- неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы
- эволюционное программирование и генетические алгоритмы
- метод ограниченного перебора
- эволюционное программирование и генетические алгоритмы
- разнообразные методы визуализации данных и множество других методов.

Большинство аналитических методов, используемые в технологии Data Mining – это известные математические алгоритмы и методы.

Важно понимать, что информация, полученная в процессе применения методов сбора данных, должна быть нетривиальной и ранее неизвестной, например, средние продажи не являются таковыми. Знания должны описывать новые отношения между свойствами, предсказывать значения одних признаков на основе других и т.д. Найденные знания должны быть применимы и на новых данных с некоторой степенью достоверности. Полезность

заключается в том, что эти знания могут приносить определенную выгоду при их применении. Знания должны быть в форме, понятной пользователю, а не в математическом виде.

Задачи, решаемые методами Data Mining:

1. **Классификация** – это отнесение объектов (наблюдений, событий) к одному из заранее известных классов.
2. **Регрессия**, в том числе задачи прогнозирования. Установление зависимости непрерывных выходных от входных переменных.
3. **Кластеризация** – это группировка объектов (наблюдений, событий) на основе данных (свойств), описывающих сущность этих объектов. Объекты в кластере должны быть "похожими" друг на друга и отличаться от объектов, включенных в другие кластеры. Чем больше похожих объектов внутри кластера и чем больше различных кластеров, тем точнее кластеризация.
4. **Ассоциация** – определение закономерности между связанными событиями. Примером такого шаблона является правило, указывающее, что событие X следует за событием Y. Такие правила называются ассоциативными правилами. Изначально эта задача была разработана для поиска типичных моделей покупок, сделанных в супермаркетах, поэтому ее иногда также называют анализом рыночной корзины (market basket analysis).
5. **Последовательные шаблоны** – определение закономерностей между связанными во времени событиями, т.е. обнаружение такой зависимости, что если произойдет событие X, то спустя заданное время произойдет событие Y.
6. **Анализ отклонений** – выявление наиболее нехарактерных шаблонов.

6. Практическая часть.

Бывает, что для выполнения конкретной задачи нужно найти сотни и тысячи номеров, адресов страниц в социальных сетях на сотнях сайтов при

определенных условиях и запросах, но управлять таким объемом информации вручную невозможно. Некоторая информация также может быть скрыта от глаз пользователя, но она содержится в коде веб-сайта. Тогда к нам на помощь и приходят парсеры.

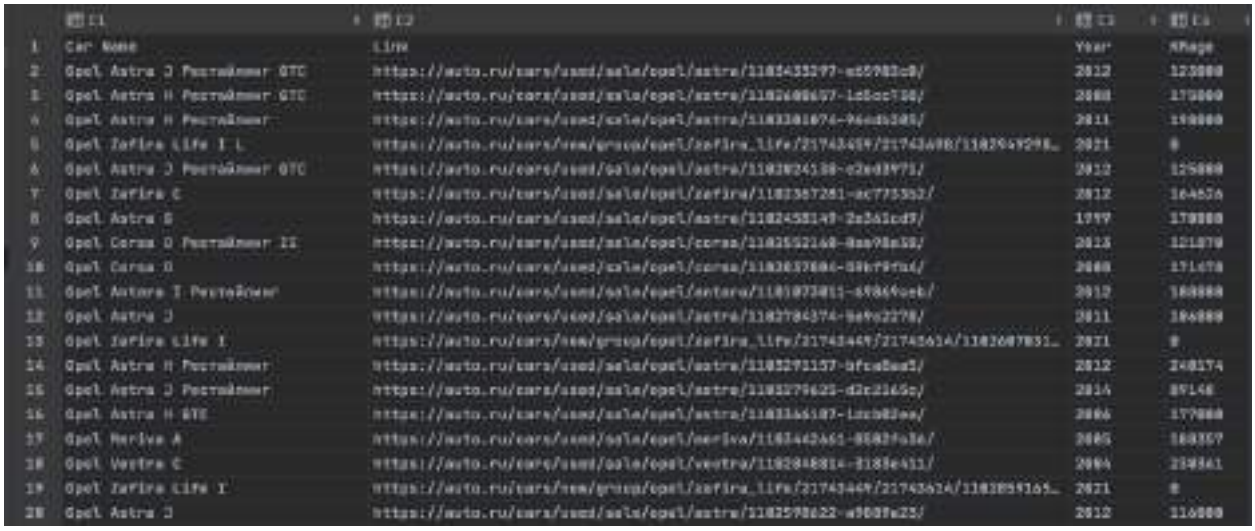
Специальные программы анализируют код страницы с помощью различных алгоритмов от совсем простых до сложнейших статистических моделей с использованием теории хаоса и нейронных сетей.

Мы разработаем парсер сайтов на языке Python, который будет обрабатывать информацию с сайтов, продающих автомобили. Получив, обработанные данные, проверим их на адекватность, удалим ошибочные записи (при наличии таковых), при необходимости декодируем. Затем сможем оценить эффективность работы парсера путем оценки качества, полученных данных. Если полученные данные будут в удобоваримом виде, будем считать, что парсеры – простой и современный способ обработки больших данных.

Код парсера можно посмотреть в репозитории GitHub по ссылке: https://github.com/mistergahan/BigData

Для примера попробуем работать с относительно небольшими объемами данных. Пропарсим сайт Auto.ru, а точнее соберем все автомобили марки Opel на рынке.

Получим датасет (.csv файл) со следующими данными (см. *Рис. 1*).



	列1	列2	列3	列4
1	Car Name	Link	Year	KMage
2	Opel Astra J Рестайлинг GTC	https://auto.ru/cars/used/sale/opel/astra/1105433297-e65983c0/	2012	123000
3	Opel Astra H Рестайлинг GTC	https://auto.ru/cars/used/sale/opel/astra/1102498697-1d6cc738/	2008	170000
4	Opel Astra H Рестайлинг	https://auto.ru/cars/used/sale/opel/astra/1103381874-94eab305/	2011	190000
5	Opel Zafira Life I L	https://auto.ru/cars/new/group/opel/zafira_life/21743459/21743498/1102649298…	2021	0
6	Opel Astra J Рестайлинг GTC	https://auto.ru/cars/used/sale/opel/astra/1103024138-e2e43973/	2012	125000
7	Opel Zafira C	https://auto.ru/cars/used/sale/opel/zafira/1103367281-ac775302/	2012	164576
8	Opel Astra G	https://auto.ru/cars/used/sale/opel/astra/1102455149-2a361cd9/	1999	170000
9	Opel Corsa D Рестайлинг II	https://auto.ru/cars/used/sale/opel/corsa/1102552168-8aa98a58/	2013	121870
10	Opel Corsa D	https://auto.ru/cars/used/sale/opel/corsa/1103037886-59b79fbd/	2008	171478
11	Opel Antara I Рестайлинг	https://auto.ru/cars/used/sale/opel/antara/1101973851-69864ueb/	2012	160000
12	Opel Astra J	https://auto.ru/cars/used/sale/opel/astra/1102784374-9e9c2278/	2011	186000
13	Opel Zafira Life I	https://auto.ru/cars/new/group/opel/zafira_life/21743449/21743614/1102607851…	2021	0
14	Opel Astra H Рестайлинг	https://auto.ru/cars/used/sale/opel/astra/1103279157-9fca8ea5/	2012	240174
15	Opel Astra J Рестайлинг	https://auto.ru/cars/used/sale/opel/astra/1103279625-d2c2165c/	2014	89148
16	Opel Astra H GTC	https://auto.ru/cars/used/sale/opel/astra/1103366107-1acb02ee/	2006	177000
17	Opel Meriva A	https://auto.ru/cars/used/sale/opel/meriva/1103442461-8582fa3e/	2005	160357
18	Opel Vectra C	https://auto.ru/cars/used/sale/opel/vectra/1102984814-8185e411/	2004	230361
19	Opel Zafira Life I	https://auto.ru/cars/new/group/opel/zafira_life/21743449/21743614/1103859165…	2021	0
20	Opel Astra J	https://auto.ru/cars/used/sale/opel/astra/1103578622-a9089e25/	2012	116000

Рис. 1. Первые 20 строк полученного файла

Данные читаемы, пропусков нет, можем перейти непосредственно к анализу полученного датастета.

Для анализа полученных данных будем использовать Anaconda Jupiter Notebook. Посмотрим, как коррелируют между собой и распределяются год выпуска и пробег автомобиля.

Выведем первые 5 строк датасета (см. *Рис. 2*).

```
# Первые 5 строк датасета
data.head()
```

	Car Name	Link	Year	KMage
0	Opel Astra J Рестайлинг GTC	https://auto.ru/cars/used/sale/opel/astra/1103...	2012	123000
1	Opel Astra H Рестайлинг GTC	https://auto.ru/cars/used/sale/opel/astra/1102...	2008	175000
2	Opel Astra H Рестайлинг	https://auto.ru/cars/used/sale/opel/astra/1103...	2011	198000
3	Opel Zafira Life I L	https://auto.ru/cars/new/group/opel/zafira_lif...	2021	0
4	Opel Astra J Рестайлинг GTC	https://auto.ru/cars/used/sale/opel/astra/1102...	2012	125000

Рис. 2. Первые 5 строк полученного файла в Jupiter Notebook

Можем убедиться, что все поля отображаются корректно и без пропусков.

Выведем размер датасета и типы данных в колонках (см. *Рис. 3*).

```
# Размер датасета
data.shape
```

```
(2334, 4)
```

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 2334
```

Рис. 3. Размер полученного файла

Просмотрим типы данных чтобы понимать, как их анализировать (см. *Рис. 4*).

```
# Основные статистические характеристки набора данных
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2334 entries, 0 to 2333
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Car Name  2334 non-null   object
 1   Link      2334 non-null   object
 2   Year      2334 non-null   int64 
 3   KMage     2334 non-null   int64 
dtypes: int64(2), object(2)
memory usage: 73.1+ KB
```

Рис. 4. Список колонок и типы представленных данных

Используем метод seaborn, чтобы оценить плотность распределения пробега автомобилей марки Opel (см. *Рис. 5*).

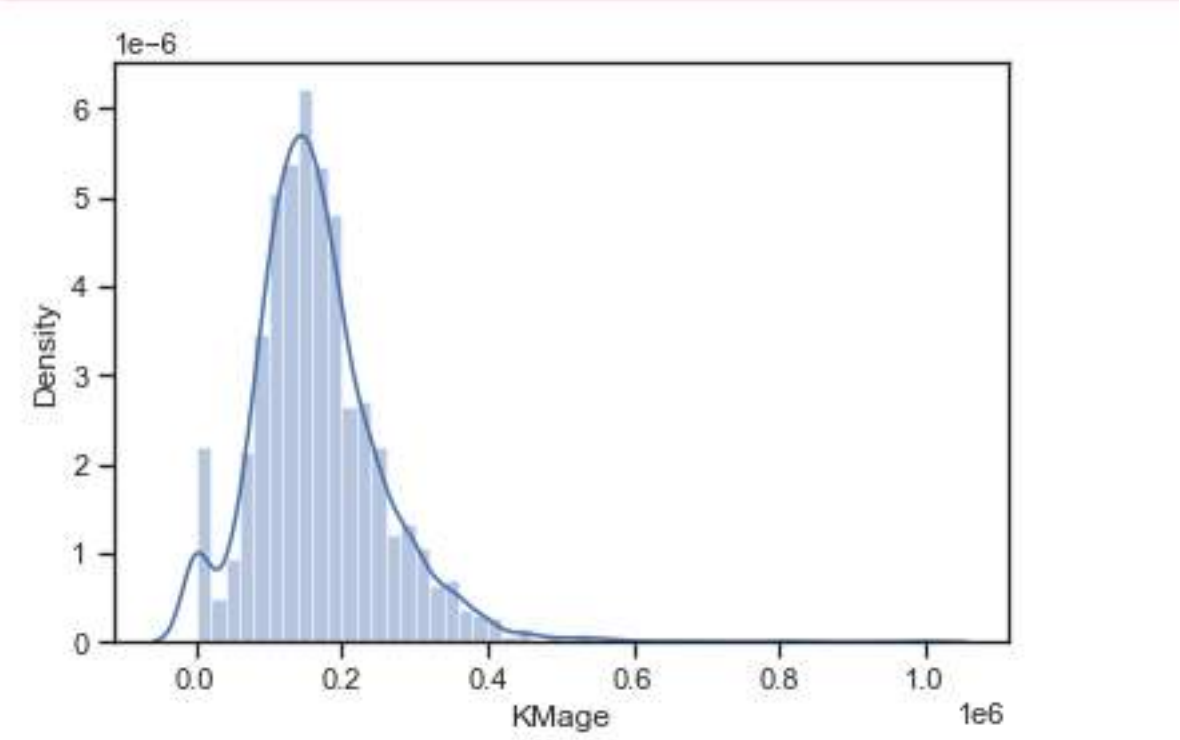


Рис. 5. Распределение пробега автомобилей марки Opel

1 Теперь построим обычную гистограмму, показывающую частотное распределение Года выпуска и Пробега авто (см. *Рис. 6*).

```
Out[33]: <AxesSubplot:xlabel='Year', ylabel='KMage'>
```

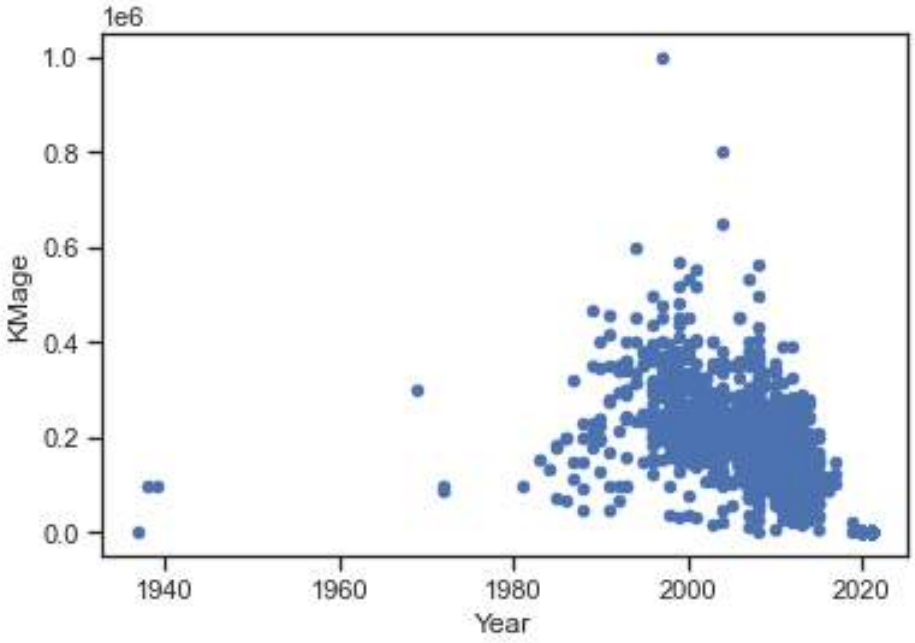


Рис. 6. Частотное распределение Года выпуска и пробега автомобилей марки Opel

Добавим линию линейной регрессии, чтобы наглядно увидеть отклонение от моды (см. Рис. 7).

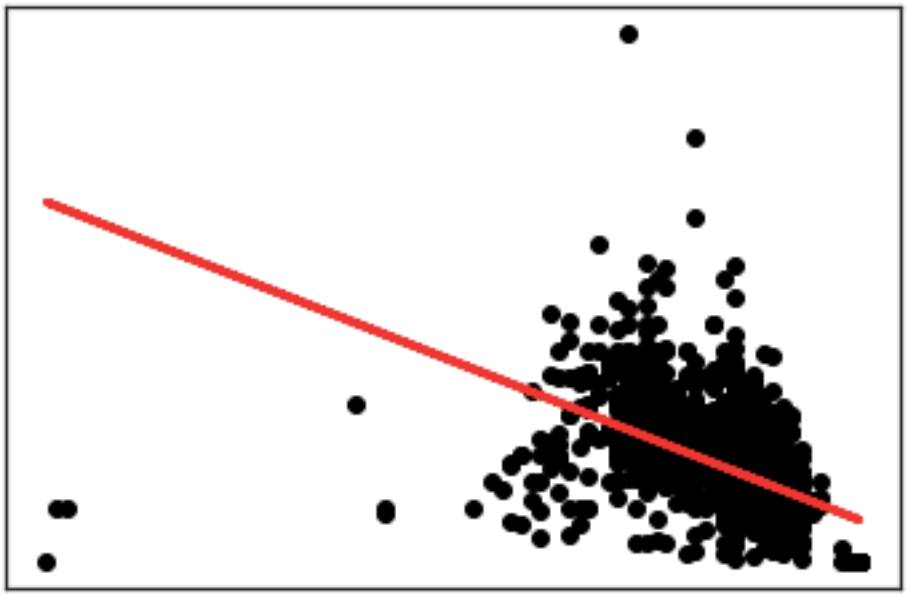


Рис. 7. Частотное распределение Года выпуска и пробега автомобилей марки Opel с Линейной регрессией (выделена красным)

Из полученного графика мы видим, что распределение Года выпуска и Пробега авто далеко не линейно, как мы могли предполагать заранее. В то время как у части автомобилей на рынке распределение действительно происходит линейно, у большой доли соотношение год/пробег значительно выбивается из моды.

Из последнего графика явно видно, что присутствуют даже явно заметные аномалии, когда у автомобиля с возрастом сохраняется малое число пройденных километров. Если посмотреть, спаршенный нами ранее датасет, то мы можем без труда выявить подобные примеры (см Рис. 8 - 11).

Рис. 8. Запись в датасете с аномально низким значением пробега для года выпуска авто.

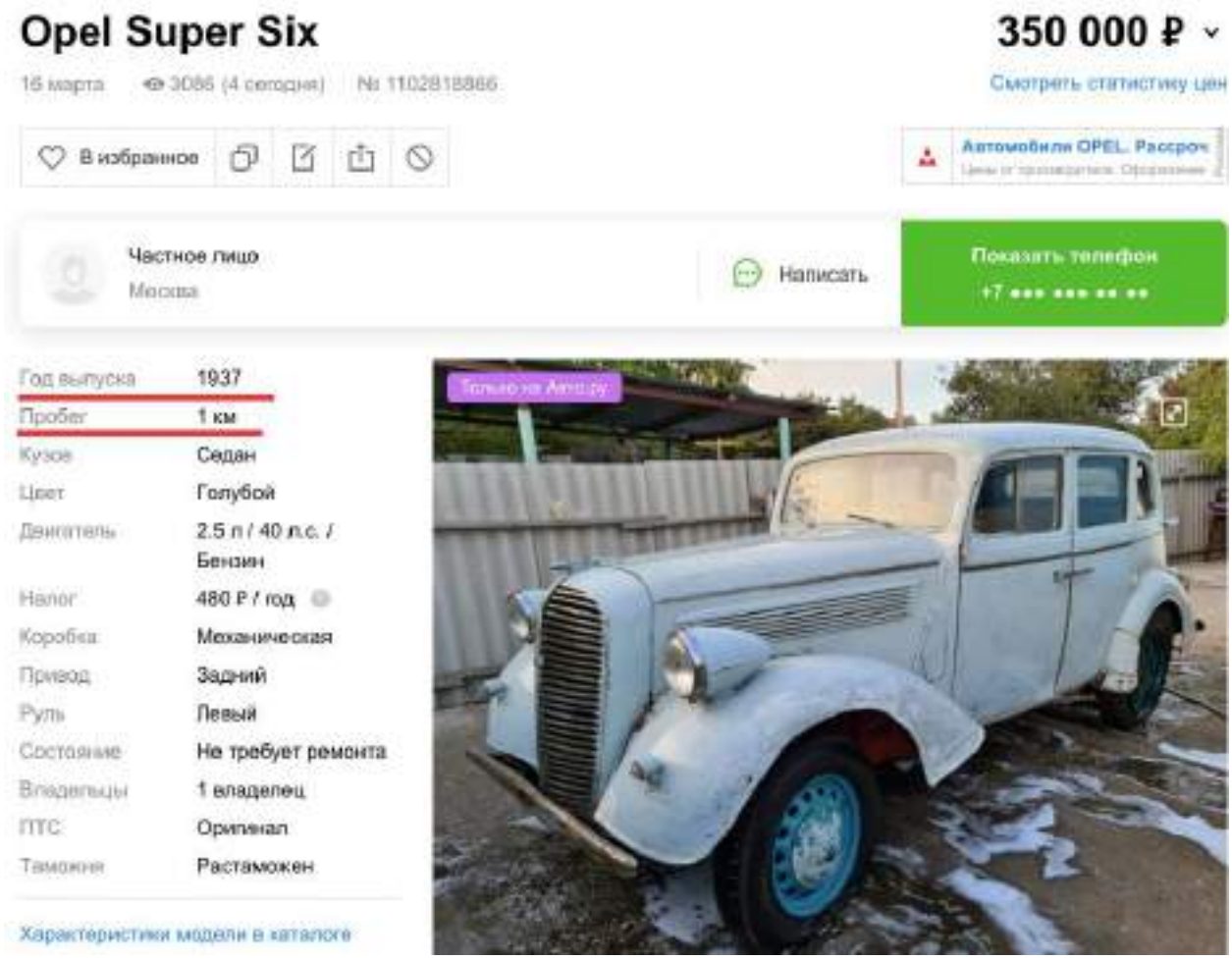


Рис. 9. Карточка автомобиля с аномально низким значением пробега для года выпуска на сайте auto.ru.

Рис. 10. Запись в датасете с аномально высоким значением пробега для года выпуска авто

Рис. 11. Карточка автомобиля с аномально высоким значением пробега для года выпуска на сайте auto.ru.

Заключение

В данной статье автор работы рассмотрел понятие Big Data и убедился в актуальности данной тематики. В частности, был затронут метод Data Mining и разработан свой скрипт, который позволил оперативно добывать данные с сайта Auto.ru в удобном для анализа формате. Проанализировав полученную выборку при использовании методов машинного обучения, было обнаружено, что распределение пробега автомобилей относительно года их выпуска происходит не по линейному закону и нашли обоснование данному феномену. Дальнейшими шагами по развитию данной тематики может быть расширение функциональности скрипта с целью его универсификации.

Литература

1. Хабрахабр. Аналитический обзор рынка Big Data // Электронный ресурс – https://habrahabr.ru/company/moex/blog/256747/

2. Шаньгин В. Ф. Защита информации в компьютерных системах и сетях. // ДМК Пресс. 2017 г.
3. Егоров А.А., Чернышова А.В., Губенко Н.Е. Анализ средств защиты больших данных в распределенных системах // Первая международная научно-практическая конференция Программная инженерия: методы и технологии разработки информационно-вычислительных систем (ПИИВС-2016). Донецк, 2016 г. – Сборник научных трудов. – ДонНТУ, Том 2, с. 28-33.
4. Егоров А.А., Чернышова А.В. Исследование инструментов распределенной системы Hadoop // Конференция Современные информационные технологии в образовании и научных исследованиях (СИТОНИ-2017). Донецк, 2017 г.
5. Хабр. Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce // Электронный ресурс – https://habr.com/ru/company/dca/blog/267361/
6. Михаил Цымблер. Какие методы и технологии используются для обработки Больших Данных //Электронный ресурс - https://mzym.susu.ru/papers/Zymbler_Supercomputers-14b.pdf
7. Loginom. Добыча данных (Data Mining) // Электронный ресурс – https://wiki.loginom.ru/articles/data-mining.html
8. Анна Вичугова Machine Learning // Электронный ресурс – https://www.bigdataschool.ru/wiki/machine-learning
9. АСУ-Аналитика ОБЗОР МЕТОДОВ DATA MINING // Электронный ресурс – http://asu-analitika.ru/obzor-metodov-data-mining
10. Хабрахабр. Визуализация данных с Python // Электронный ресурс – https://habr.com/ru/company/ods/blog/323210/
11. Пятифан. Информационные технологии. Формы и способы представления данных // Электронный ресурс - http://5fan.ru/wievjob.php?id=39531

12. Латышева А. М. Big data. Актуальность и перспективы использования // Электронный журнал Молодежный Научно-Технический Вестник ISSN 2307-0609 - http://sntbul.bmstu.ru/doc/724143.html

Literature

1. Habrahabr. Analytical review of the Big Data market / / Electronic resource – https://habrahabr.ru/company/moex/blog/256747/
2. Shangin V. F. Information protection in computer systems and networks. // DMK Press. 2017
3. Egorov A. A., Chernyshova A.V., Gubenko N. E. Analysis of big data protection tools in distributed systems / / First International scientific and practical conference Software Engineering: methods and technologies for developing information and computing systems (PIIVS-2016). Donetsk, 2016- Collection of scientific works. - DonNTU, Volume 2, pp. 28-33.
4. Egorov A. A., Chernyshova A.V. Research of tools of the distributed Hadoop system / / Conference Modern information technologies in education and scientific research (SITONI-2017). Donetsk, 2017
5. Habr. Big Data from A to Z. Part 1: Principles of working with big data, the MapReduce paradigm // Electronic resource – https://habr.com/ru/company/dca/blog/267361/
6. Mikhail Tsymbler. What methods and technologies are used to process Big Data //Electronic resource - https://mzym.susu.ru/papers/Zymbler_Supercomputers-14b.pdf
7. Loginom. Data Mining / / Electronic resource – https://wiki.loginom.ru/articles/data-mining.html
8. Anna Vichugova Machine Learning // Electronic resource – https://www.bigdataschool.ru/wiki/machine-learning
9. AUTOMATED CONTROL SYSTEM-Analytics OVERVIEW OF DATA MINING METHODS // Electronic resource – http://asu-analitika.ru/obzor-metodov-data-mining

10. Habrahabr. Data visualization in Python / / Electronic resource – https://habr.com/ru/company/ods/blog/323210/

Ссылка: Карпов Д.К. Обработка больших данных с использованием средств языка python // StudNet - 2021 г. - №6 - С. 1397-1412 - Ссылка: https://cyberleninka.ru/article/n/obrabotka-bolshih-dannyh-s-ispolzovaniem-sredstv-yazyka-python