УДК 004.94, 519.2

И. Х. Утакаева
Финансовый университет при Правительстве РФ, Москва, e-mail: utakaev@yandex.ru

ПРИМЕНЕНИЕ ПАКЕТА СТАТИСТИЧЕСКОГО АНАЛИЗА PYTHON ДЛЯ АНАЛИЗА ДАННЫХ АВТОМОБИЛЬНОГО РЫНКА

Ключевые слова: анализ данных, эконометрика, модель, python, программирование.

Автомобильная отрасль занимает одно из ключевых мест в экономике многих стран. В российской экономике эта отрасль также остается на ведущих позициях. Сегодня компаниям отрасли приходится развивать и поддерживать бизнес, адаптируясь к нестабильной экономической ситуации и новым правилам регулирования. От характера развития автомобильной отрасли зависят уровни развития смежных отраслей.

В условиях современной рыночной экономики возникает необходимость получения более точной и достоверной информации о состоянии и тенденциях рынка автомобилей, чтобы иметь возможность оперативно реагировать на изменения и управлять рисками. Динамика продаж и цен влечет за собой определенные риски для производителей, работников, дилеров и государства в целом. Поэтому важно детально исследовать данный феномен и предложить инструмент для моделирования экономической ситуации. Важное значение для производителей и дилеров имеет адекватная оценка потенциала рынка.

Одной из мер, которая может помочь в управлении рисками автомобильной отрасли, является составление эконометрической модели прогнозирования. Таким образом, в условиях рыночной неопределенности производители и дилеры смогут вырабатывать механизмы, необходимые для выживания и стабильного функционирования. Государство сможет принимать решение о необходимости поддержки отрасли посредством реализации различных мер, способствующих обеспечению занятости населения в автомобильной отрасли, повышения доступности автомобилей для населения и развития производства автомобилей.

Введение

Специфической особенностью деятельности экономиста является работа в условиях недостатка информации и неполноты исходных данных для моделирования. Анализ такой информации требует использования специальных методов, которые составляют один из аспектов эконометрики. Одной из центральных проблем эконометрики является построение эконометрической модели и определение возможности ее использования для описания, анализа, прогнозирования реальных экономических процессов. Сегодня, применению эконометрических методов может мешать высокая стоимость коммерческих пакетов статистического анализа. Выходом из сложившейся ситуации может быть использование открытого программного обеспечения, удачным примером которого является пакет сатистического анализа языка программирования Python.

Цель исследования: демонстрация новых возможностей эконометрического моделирования с использованием современного языка программирования Python. В статье предлагается использовать возможности языка Python, как инструмента обеспечивающего высокую производительность и точность при использовании эконометрических и статистических методов анализа данных.

Материал и методы исследования

Библиотека Pandas языка Python позволяет с легкостью манипулировать исходными данными и анализировать их. В принципе, библиотека Pandas построена на еще одной замечательной библиотеки в python 3 – Numpy. Использование пакета Pandas дает широкие возможности при работе с электронными таблицами.

В качестве примера выбраны автомобили марок Ford Focus и Opel Astra. В результате анализа выделены факторы, которые наиболее ощутимо влияют на стоимость автомобиля, разработана многофакторная математическая модель описывающая процесс ценообразования на вторичном рынке автомобилей, получены уравнение регрессии и матрицы корреляции переменных, построены графики влияния исследуемых факторов на стоимости автомобилей. В работе рассматриваются широкие возможности открытого и свободного программного обеспечения – FLOSS (Free\Libre and

Open Sourse Software). Исследуются особенности ценообразования [3].

Для проведения исследования, необходимо выбрать базовый набор данных DataSet. Разнообразные наборы данных можно скачать прямо с сайта, который содержит такую информацию. DataSet, как правило, представляет собой файл с таблицей данных в формате json или csv. **Цель работы** – показать простоту обработки достаточно большого объема данных средствами Python.

Язык программирования Python в последнее время активно используется для анализа данных в различных социально значимых сферах. Это один из наиболее популярных современных языков программирования, который широко используется в анализе данных. Связано это прежде всего, с простотой языка, а также доступностью и разнообразием современных библиотек. В статье приведен пример исследования и классификации неструктурированных данных, а также построения эконометрической модели прогнозирования стоимости автомобиля с использованием возможностей и инструментов языка Python.

Python – это современный язык программирования, востребованный и популярный в мировой научной среде. В настоящей работе демонстрируются новые возможности ценообразования на вторичном рынке автомобилей. Язык программирования Python – это мощный высокоуровневый кроссплатформенный язык. Он поддерживает объектно-ориентированное программирование, и в последнее время стал серьезной альтернативой таким языкам программирования как C++. В отличие от MATLAB, язык Python изначально не заточен под научные вычисления.

Объектом исследования является рынок подержанных автомобилей, целью – выявление критериев оценки и уровня их влияния на цену подержанного автомобиля. Данные для проведения исследования получены с web-сайта avito.ru – крупнейший в Европе сайт частных объявлений с посещаемостью более 25 000 000 пользователей ежемесячно. Выбор сайта avito.ru объясняется, во-первых, тем, что сайт имеет достаточно большую базу предложений, во-вторых, по каждому продаваемому автомобилю в базе имеется подробная информация о его характеристиках. О каждом автомобиле в извлеченной выборке имеется следующая информация: марка автомобиля, модель автомобиля, тип кузова, год выпуска, пробег, коробка передач, объем двигателя, тип двигателя, привод, подробная информация о комплектации автомобиля. Для построения модели использованы такие модули как: pandas, библиотека для визуализации данных в statsmodels.formula.api, бибилиотека двумерной графики matplotlib. pyplot [2].

По теме исследования опубликованы работы, в которых не представлены современные эконометрические модели, которые могут включать не только числовые параметры.

Для исследования в качестве примера выбраны автомобили марок Ford Focus и Opel. После удаления из полученной выборки недостоверной и противоречивой информации, объем выборки составил 618509 автомобилей для автомобилей марки Ford Focus и 100955 для автомобилей Opel .

Введем следующие обозначения: Year – возраст, Mileage – пробег, объем двигателя – Capacity, мощность двигателя – Power, стоимость Price.

Получены следующая статистическая информация о выборке автомобилей Ford Focus.

Проведен корреляционный анализ извлеченной из базы выборки, результаты приведены в таблице.

Рассмотрим следующее уравнение регрессии:

$$Y = S + \sum_{i=1}^{n} a_i x_i ,$$

где x_i – это независимые переменные, характеризующие автомобиль (Year, Mileage, Capacity, Power); S – некоторая фиксированная величина, зависящая от модели авто; a_i – коэффициенты, отражающие степень влияния, соответствующего параметра x_i на цену автомобиля; Y – зависимая переменная – стоимость автомобиля (Price).

Регрессионное уравнение можно использовать для построения модели расчета стоимости автомобиля определенной модели. Воспользуемся методом наименьших квадратов [3].

```
In [3]: #Первые 5 записей
path = '/Users/alex/Documents/PYTHON/AvitoDatasetFord.txt'
#path = '/Users/alex/Documents/PYTHON/ID.txt'
data = pd.read_csv(path, names=['Year', 'Mileage', 'Capacity', 'Power', 'Price'])
data.head()
```

Out[3]:

	Year	Mileage	Capacity	Power	Price
0	6	75000	1.8	125	357000
1	4	66000	2.0	150	660000
2	9	166000	1.8	125	330000
3	11	210000	2.0	145	235000
4	14	104000	2.0	130	170000

```
#Информация о наборе данных
data.describe()
```

	Year	Mileage	Capacity	Power	Price
count	1716.000000	1716.000000	1716.000000	1716.000000	1.716000e+03
mean	7.455128	111607.250000	1.722786	119.375291	3.890623e+05
std	3.252229	55634.700013	0.178034	19.008764	1.388090e+05
min	1.000000	2000.000000	1.000000	70.000000	8.000000e+04
25%	5.000000	74800.000000	1.600000	101.000000	2.850000e+05
50%	7.000000	100009.000000	1.600000	125.000000	3.750000e+05
75%	10.000000	140998.500000	1.800000	125.000000	4.850000e+05
max	17.000000	430000.000000	2.500000	250.000000	1.100000e+06

Рис. 1. Первые 5 автомобилей Ford Focus и следующая статистическая информация о выборке автомобилей Ford Focus

```
#Матрица корелляции переменных
data.corr()
```

	Year	Mileage	Capacity	Power	Price
Year	1.000000	0.631289	0.202119	-0.138462	-0.870056
Mileage	0.631289	1.000000	0.075783	-0.146194	-0.630609
Capacity	0.202119	0.075783	1.000000	0.799377	-0.039703
Power	-0.138462	-0.146194	0.799377	1.000000	0.314947
Price	-0.870056	-0.630609	-0.039703	0.314947	1.000000

Рис. 2. Матрица корреляции переменных для автомобилей Ford Focus

Результаты исследования и их обсуждение

Проведенный анализ показывает, что для автомобилей Ford Focus с увеличением возраста на 1 год, стоимость автомобиля снижается на 31 386 рубля, а с увеличением пробега на 10 000 километров, стоимость снижается на 2918 рублей, с увеличением мощности двигателя на 100 лошадиных сил, стоимость увеличится на 206 064 рубля.

Для автомобилей Opel с увеличением возраста на 1 год, стоимость автомобиля снижается на 31 354 рубля, а с увеличением пробега на 10 000 километров, стоимость снижается на 2176 рублей, с увеличением мощности двигателя на 100 лошадиных сил, стоимость увеличится на 96 548 рублей. Подобные расчеты возможно произвести для авто любых марок.

```
#Уравнение регрессии и коэффициенты переменных
lm = smf.ols(formula = 'Price ~ Year + Mileage + Capacity + Power', data=data).fit()
print lm.params
```

```
Intercept    554405.329814
Year         -31386.186722
Mileage          -0.291816
Capacity     -84036.020544
Power          2060.643268
dtype: float64
```

```
#Уравнение регрессии и коэффициенты переменных
lm = smf.ols(formula = 'Price ~ Year + Mileage + Capacity + Power', data=data).fit()
print lm.params
```

```
Intercept    764101.181277
Year         -31353.783205
Mileage          -0.217647
Capacity    -131900.230257
Power           965.481407
dtype: float64
```

Рис. 3. Результаты расчета параметров модели автомобилей Ford Focus и Opel Astra

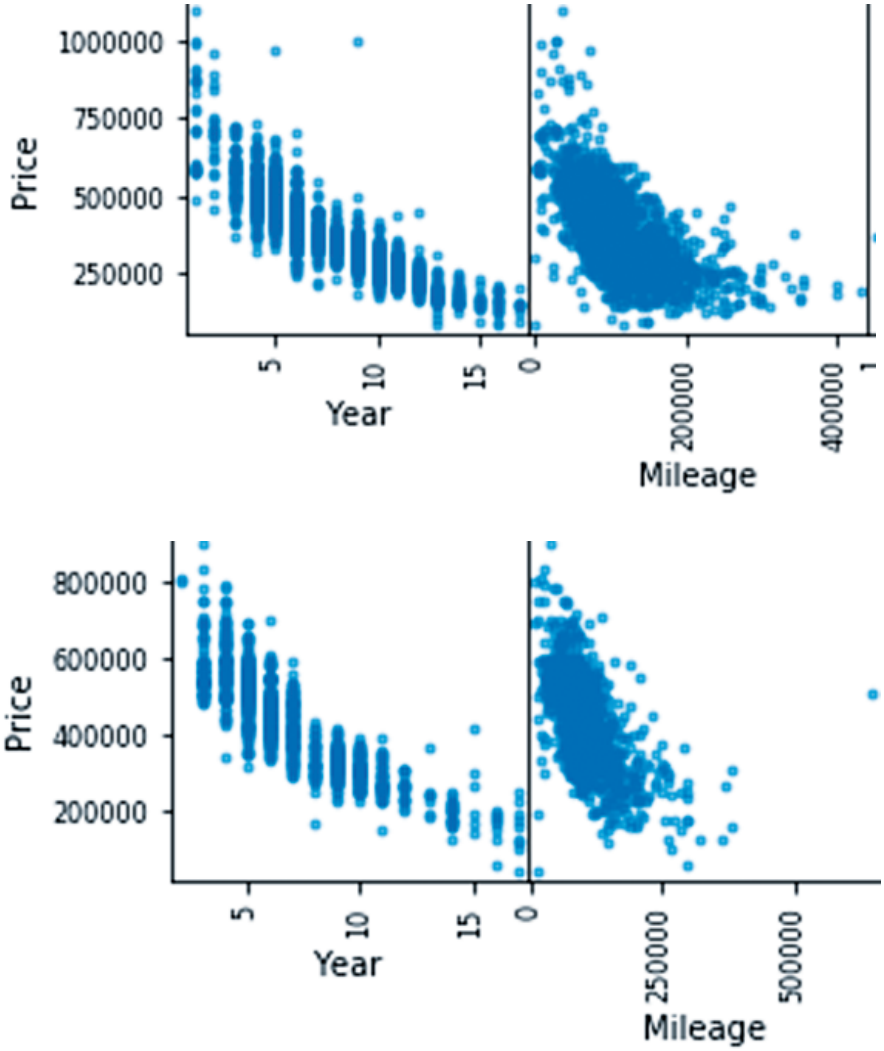

Рис. 4. Влияние возраста и пробега на стоимость автомобилей Ford Focus и Opel

```
In [375]: #Вводим нужные значения и предсказываем зависимую переменную
          model = pd.DataFrame({'Year': [4], 'Mileage': [50000], 'Capacity': [1.6], 'Power': [150]})
          lm.predict(model)
Out[375]: array([ 588908.6182675])

In [376]: model = pd.DataFrame({'Year': [7], 'Mileage': [50000], 'Capacity': [1.6], 'Power': [150]})
          lm.predict(model)
Out[376]: array([ 494750.05810091])

In [381]: model = pd.DataFrame({'Year': [4], 'Mileage': [200000], 'Capacity': [1.6], 'Power': [150]})
          lm.predict(model)
Out[381]: array([ 545136.15232695])

In [385]: model = pd.DataFrame({'Year': [4], 'Mileage': [50000], 'Capacity': [1.6], 'Power': [90]})
          lm.predict(model)
Out[385]: array([ 465270.02219018])

In [8]:  model = pd.DataFrame({'Year': [4], 'Mileage': [50000], 'Capacity': [1.8], 'Power': [150]})
         lm.predict(model)
Out[8]:  array([ 572101.41415863])

In [388]: model = pd.DataFrame({'Year': [10], 'Mileage': [150000], 'Capacity': [1.6], 'Power': [125]})
          lm.predict(model)
Out[388]: array([ 319893.77227508])
```

```
#Вводим нужные значения и предсказываем зависимую переменную
model = pd.DataFrame({'Year': [5], 'Mileage': [75000], 'Capacity': [1.6], 'Power': [115]})
lm.predict(model)
```

```
array([ 490998.69854596])
```

```
model = pd.DataFrame({'Year': [11], 'Mileage': [75000], 'Capacity': [1.6], 'Power': [115]})
lm.predict(model)
```

```
array([ 302875.99931455])
```

```
model = pd.DataFrame({'Year': [5], 'Mileage': [20000], 'Capacity': [1.6], 'Power': [115]})
lm.predict(model)
```

```
array([ 502969.30930443])
```

```
model = pd.DataFrame({'Year': [5], 'Mileage': [75000], 'Capacity': [2.0], 'Power': [115]})
lm.predict(model)
```

```
array([ 438238.60644329])
```

```
model = pd.DataFrame({'Year': [5], 'Mileage': [75000], 'Capacity': [1.6], 'Power': [140]})
lm.predict(model)
```

```
array([ 515135.73372652])
```

```
model = pd.DataFrame({'Year': [3], 'Mileage': [15000], 'Capacity': [1.8], 'Power': [140]})
lm.predict(model)
```

```
array([ 564522.1021858])
```

```
#Итоговые показатели модели
lm.summary()
```

Рис. 5. Результаты тестирования модели для автомобилей Ford Focus и Opel

Выводы

Представленная модель может использоваться для прогнозирования цен на автомобили производителями, дилерами и государством в целях управления рисками в автомобильной отрасли. Государственные органы, основываясь на прогнозных значениях, могут вырабатывать стимулирующие меры для поддержки рынка автомобилей.

В целом, несмотря на качество и адекватность построенной модели анализа и прогнозирования, ее можно усовершенствовать, добавив иные факторы, например, ставку по автокредитованию или стоимость обслуживания автомобиля.

Также данная модель не учитывает факторы, которые могли бы повлиять на стоимость, которые сложно количественно оценить. К ним можно отнести кризисное состояние экономики, появление новых марок автомобилей, государственные меры поддержки, такие как субсидирование процентной ставки по автокредитованию, программа утилизации автомобилей. В таком случае, можно использовать фиктивные переменные.

Следует отметить, что проведенные исследования демонстрируют эффективность использования возможностей языка Python для анализа больших данных, получены результаты, которые заслуживают дальнейшего изучения с помощью предложенного инструмента.

Библиографический список

1. https://www.avito.ru/.
2. Бывшев, В.А. Эконометрика: учеб. пособие. – М.: Финансы и статистика, 2008. – 480 с.
3. Войтковская Е.И., Шабельникова Е.В. Анализ отечественной автомобильной отрасли с применением эконометрического моделирования // Устойчивое развитие науки и образования. – 2017.
4. Костромин, А.В., Кундакчян, Р.М. Эконометрика: учебное пособие. – М.: КНОРУС, 2015. – 228 с.
5. Савин, А.В. Перспективы развития автомобильной отрасли в России // Международный журнал прикладных и фундаментальных исследований. – 2015. – № 7–2. – С. 311–316.
6. Ассоциация европейского бизнеса [Электронный ресурс]. – Режим доступа: http://www.aebrus.ru/ru/ (дата обращения: 10.12.2016).
7. Бюллетень социально-экономического кризиса России [Электронный ресурс] Аналитический центр при Правительстве Российской Федерации. – Режим доступа: http://ac.gov.ru/files/publication/a/9154.pdf (дата обращения: 11.12.2016).
8. Мировой рынок легковых автомобилей в 2015-2016 годах [Электронный ресурс] EREPORT.RU. – Режим доступа: http://www.ereport.ru/articles/commod/auto.htm (дата обращения: 11.12.2016).
9. Прошлые данные – Нефть Brent [Электронный ресурс] Investing.com. – Режим доступа: https://ru.investing.com/commodities/brent-oil-historical-data (Дата обращения: 10.02.2016).
10. Российский авторынок в 2015 году опустился на пятое место в Европе [Электронный ресурс] Аналитическое агентство АВТОСТАТ. – Режим доступа: https://www.autostat.ru/news/24492/ (Дата обращения: 11.12.2016).
11. Федеральная служба государственной статистики [Электронный ресурс]. – Режим доступа: http://www.gks.ru (дата обращения: 10.12.2016).
12. У. Маккинли. Python и анализ данных. – 2015.
13. Утакаева И.Х., Хмелевская К.А. Опыт эконометрического моделирования с использованием пакета статистического анализа Python//Международный научный журнал. – 2017. – № 5. – С. 67.