Литература

1. *Клименко И.С.* Теория систем и системный анализ: учебное пособие. М.: РосНОУ, 2014.
2. *Клименко И.С., Шарапова Л.В.* Общая задача принятия решения и феномен неопределенности // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ, управление». 2019. № 3. С. 44–58.
3. *Кох Р.* Принцип 80/20. М.: Эксмо, 2012.
4. *Куроуз Д., Росс К.* Компьютерные сети. Настольная книга системного администратора. М.: Эксмо, 2016.
5. *Лимончелли Т.А., Хоган К., Чейлап С.* Системное и сетевое администрирование. Практическое руководство. М.: Символ-Плюс, 2009.
6. *Олифер В., Олифер Н.* Компьютерные сети. СПб.: Питер, 2017.
7. *Фрайманн А.В.* Об особенностях применения принципа необходимого разнообразия для отображения функций системного администратора // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ, управление». 2019. № 2. С. 64–69.

Literatura

1. *Klimenko I.S.* Teoriya sistem i sistemnyj analiz: uchebnoe posobie. M.: RosNOU, 2014.
2. *Klimenko I.S., Sharapova L.V.* Obshchaya zadacha prinyatiya resheniya i fenomen neopredelennosti // Vestnik Rossijskogo novogo universiteta. Seriya “Slozhnye sistemy: modeli, analiz, upravlenie”. 2019. № 3. S. 44–58.
3. *Kokh R.* Printsip 80/20. M.: Eksmo, 2012.
4. *Kurouz D., Ross K.* Komp’yuternye seti. Nastol’naya kniga sistemnogo administratora. M.: Eksmo, 2016.
5. *Limonchelli T.A., Khogan K., Chejlap S.* Sistemnoe i setevoe administrirovanie. Prakticheskoe rukovodstvo. M.: Simvol-Plyus, 2009.
6. *Olifer V., Olifer N.* Komp’yuternye seti. SPb.: Piter, 2017.
7. *Freimann A.W.* Ob osobennostyakh primeneniya printsipa neobkhodimogo raznoobraziya dlya otobrazheniya funktsij sistemnogo administratora // Vestnik Rossijskogo novogo universiteta. Seriya “Slozhnye sistemy: modeli, analiz, upravlenie”. 2019. № 2. S. 64–69.

DOI: 10.25586/RNU.V9187.19.04.P.065 УДК 004.032.26

Д.М. Булычев

ПРОГНОЗИРОВАНИЕ РЕЗУЛЬТАТОВ ЭКСПЕРТНОГО ОЦЕНИВАНИЯ ТОЧЕК ПРОДАЖ С ПОМОЩЬЮ НЕЙРОННОЙ СЕТИ

Сформировано признаковое пространство экспертной оценки точек продаж на основе агрегированных данных. Введена гипотетическая формула оценки параметров и приведен пример отображения признакового пространства в пространство экспертных оценок специалистов. Разработан адаптивный алгоритм обработки больших данных, спроектирована и оптимизирована нейронная сеть, обучена модель, обеспечивающая прогноз с абсолютной средней ошибкой 0,448.

Ключевые слова: анализ данных, искусственные нейронные сети, геомаркетинг, полносвязная нейронная сеть, большие данные, Python.

D.M. Bulychev

PREDICTION THE RESULTS OF EXPERT EVALUATION OF POINTS OF SALE USING A NEURAL NETWORK

The features space of expert evaluation of sales points is formed on the basis of aggregated data. A hypothetical formula for parameter estimation is introduced and an example of mapping the feature space to the expert evaluation space is given. The adaptive algorithm of big data processing is developed, the neural network is designed and optimized, the model providing the prediction with an absolute average error of 0,448 is trained.
Keywords: data analysis, artificial neural networks, geomarketing, fully connected neural network, big data, Python.

Введение

В настоящее время точки продаж в большинстве случаев выбираются на основе опыта и интуиции экспертов, которые подкрепляют свой выбор абстрактными оценками, построенными на эвристиках, причем разные эксперты используют различные эвристики [3]. Представляется целесообразным, используя методы машинного обучения, стандартизировать и упростить получение оценки, обучив модель на оценках экспертов, использующих одинаковые или похожие способы оценивания. В настоящее время чаще всего геомаркетинговые компании продают данные и предоставляют услуги по прогнозированию определенных характеристик, таких как ожидаемая выручка, клиентский поток и др. Эти исследования проводятся экспертами вручную на основе статистических расчетов, а выводы делаются на небольшом наборе данных с привлечением эвристических оценок [2; 3].

Целью настоящей работы является разработка алгоритма машинного обучения, позволяющего предсказывать экспертные оценки коммерческого потенциала точек продаж на основе существующих результатов оценивания действующих экспертов. В качестве такого алгоритма была выбрана нейронная сеть. Данный выбор обусловлен способностью нейронных сетей к автоматическому отбору признаков и способностью дообучаться. Помимо этого, нейронные сети относительно легко масштабируются [5].

Автоматический отбор признаков позволит экспериментировать с различными данными без необходимости тщательного ручного подбора значимых признаков. То есть появляется устойчивость к информационному шуму.

Возможность дообучения – чрезвычайно важное свойство алгоритма в рамках рассматриваемой задачи. Это позволит модели при появлении новых данных сразу учитывать их, что открывает возможность постоянно улучшать качество прогноза и сглаживать сезонные эффекты.

Структура экспертной оценки

Исходя из практики маркетинговых исследований [4; 6; 7], можно выявить следующие существенные параметры, которые используются для оценивания точек продаж:

1) престижность района;

2) арендная плата;

3) стоимость строительства;
4) численность населения в окрестности;
5) доступность для личного транспорта;
6) близость к крупным торговым точкам другого типа;
7) видимость торговой точки;
8) потребительский потенциал;
9) суммарная площадь предприятий;
10) конкуренция;
11) размер офисов в окрестности;
12) автомобильный трафик;
13) пешеходный трафик.

Набор данных

В геомаркетинге данные обладают низкой доступностью, а их получение является сложной и дорогостоящей операцией, поэтому в открытом доступе находится мало полезной информации. Маркетологи для своих исследований покупают данные у геомаркетинговых сервисов или обеспечивают их поступление посредством организации собственных исследований. И то и другое требует значительных ресурсов. В связи с этим в настоящей работе будет рассматриваться довольно ограниченный набор данных, которые тем не менее было весьма непросто подготовить. Наборы данных, подготовленные для этой работы и актуальные на январь 2019 г., приведены в таблице 1. Указанные наборы ограничивают область исследований продуктовыми магазинами Москвы.

Таблица 1

Описание наборов данных, ед.

Набор данных	Количество записей
Магазины Москвы	10 107
Остановки наземного транспорта Москвы	11 414
Вестибюли метро Москвы	221
Объявления о сдаче жилой площади Москвы	1683
Бизнес-центры Москвы	1314
Торговые центры Москвы	669

Признаковое пространство

В таблице 2 указаны полезные признаки точек продаж, извлеченные из имеющихся данных (в скобках приведены переменные, которыми обозначаются параметры далее в формулах).

Таблица 2

Описание признаков

Признак модели	Связанные экспертные параметры оценки
Конкуренция в радиусе 500 м p_0	Конкуренция x_9
Конкуренция в радиусе 1000 м p_1	Конкуренция x_9
	Доступность для личного транспорта x_4
Конкуренция в радиусе 1500 м p_2	Конкуренция x_9
	Доступность для личного транспорта x_4

Окончание табл. 2

Признак модели	Связанные экспертные параметры оценки
Количество станций метро в радиусе 500 м p_3	Пешеходный трафик x_{12}
	Автомобильный трафик x_{11}
	Престижность района x_0
	Арендная плата x_1
	Стоимость строительства x_2
	Доступность для личного транспорта x_4
Количество маршрутов, проходящих через остановки в радиусе 300 м p_4	Численность населения в окрестности x_3
	Пешеходный трафик x_{12}
Средняя стоимость аренды одного квадратного метра жилья в радиусе 500 м p_5	Престижность района x_0
	Арендная плата x_1
	Стоимость строительства x_2
	Потребительский потенциал x_7
Количество офисов в радиусе 500 м p_6	Престижность района x_0
	Арендная плата x_1
	Стоимость строительства x_2
	Потребительский потенциал x_7
	Конкуренция x_9
	Размер офисов в окрестности x_{10}
	Пешеходный трафик x_{12}
	Автомобильный трафик x_{11}
	Доступность для личного транспорта x_4
Количество торговых центров в радиусе 300 м p_7	Арендная плата x_1
	Стоимость строительства x_2
	Потребительский потенциал x_7
	Конкуренция x_9
	Пешеходный трафик x_{12}
	Доступность для личного транспорта x_4
	Суммарная площадь предприятий x_8
	Близость к крупным торговым точкам другого типа x_5
	Видимость торговой точки x_6

Как видно, признаковое пространство, включающее 8 признаков, косвенно охватывает все абстрактные параметры экспертизы точек продаж из привлеченных маркетинговых исследований. Связь между ними установлена на основе фактов, приведенных в рассмотренных исследованиях, и не является окончательной, тем не менее она вполне достаточна для формирования целевого признака.

Выбор радиусов оценки был произведен согласно следующей эвристике:

- 300 м – радиус пешей доступности;
- 500 м – радиус пешей доступности при наличии точек интереса;
- 1000 м – радиус пешей доступности при наличии общественного транспорта и доступности для личного транспорта;
- 1500 м – радиус доступности для автомобиля.

Целевой признак

Прежде чем перейти к описанию целевого признака, необходимо преобразовать признаковое описание точек в параметры экспертного оценивания. Для этого, помимо установленной ранее связи между упомянутыми признаками и параметрами, нужно оценить их значимость. Весовые коэффициенты, соответствующие установленным связям, подобраны эвристически с учетом рассмотренных исследований:

$$
\begin{aligned}
&x_0 = p_3 + \frac{1{,}1 p_5}{750};\\
&x_1 = 1{,}5 p_3 + \frac{2 p_5}{750} + 1{,}4 p_6 + 2{,}1 p_7;\\
&x_2 = 0{,}8 x_1;\\
&x_3 = p_4;\\
&x_4 = 2 p_3 + 1{,}5 p_4;\\
&x_5 = p_7;\\
&x_6 = p_7;\\
&x_7 = \frac{1{,}5 p_5}{750} + 3 p_6;\\
&x_8 = p_7;\\
&x_9 = 2 p_0 + p_1 + 0{,}5 p_2 - \left(1 + 0{,}8 p_4 + \frac{0{,}4 p_5}{750} + 2 p_7 + 1{,}1 p_6\right);\\
&x_{10} = p_6;\\
&x_{11} = \frac{0{,}7 p_5}{750} + 0{,}5 p_6 + 0{,}6 p_7;\\
&x_{12} = p_7.
\end{aligned}
\qquad (1)
$$

Все признаки, кроме p_5 (средняя стоимость аренды одного квадратного метра жилья в радиусе 500 м), количественные, поэтому каждый коэффициент напрямую отображает значимость признака для параметра. Чтобы достичь такого же эффекта для признака p_5, нужно разделить его на близкое к среднему значение этого признака. Следовательно, можно рассматривать $\frac{p_5}{750}$ как степень отклонения среднего выборки стоимостей арендной платы квартир от среднего генеральной совокупности всех стоимостей. Таким образом, оценки экспертных параметров получаются как взвешенная сумма признаков модели.

В качестве целевого признака используется гипотетическая эвристическая оценка эксперта, которая складывается на основе оценки экспертных параметров:

$$
\begin{aligned}
y = \frac{1}{10}\big(&2x_0 - 1{,}5x_1 - 1{,}2x_2 + 1{,}3x_3 + 0{,}75x_4 + 0{,}55x_5 + 1{,}1x_6 +\\
&+ 2{,}5x_7 - 1{,}4x_8 - 2{,}5x_9 + 1{,}4x_{10} + 1{,}2x_{11} + 1{,}4x_{12}\big).
\end{aligned}
\qquad (2)
$$

Гипотетическая экспертная оценка моделируется так же, как и отдельные оценки параметров, как взвешенная сумма, где веса подбираются эвристически. Оценка получается

в диапазоне от 0 до 100 и для удобства восприятия делится на 10. Таким образом, нейронная сеть будет оценивать каждую точку по 10-балльной ранговой шкале.

Результирующая оценка охватывает все значения введенной шкалы. Для того чтобы оценка в большей степени соответствовала реальной ситуации, к вычисленной оценке добавляется гаусовский шум без смещения с $\sigma = 0{,}6$. Получившееся распределение оценок (рис. 1) похоже на распределение логорифмически нормальной случайной величины, что является совершенно естественным в рамках рассматриваемой задачи и косвенно указывает (но не доказывает) на адекватность выведенной оценки [1].

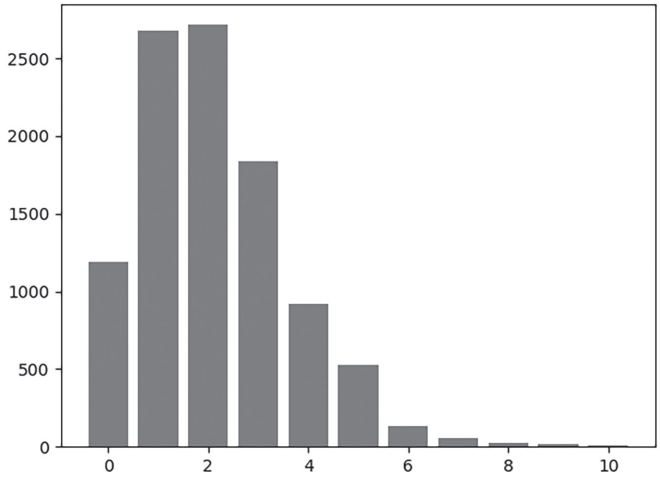

Рис. 1. Распределение значений оценки по шкале

Извлечение признаков

Вследствие большого объема данных, которые необходимо обработать для извлечения признаков, с учетом их разнородности задача обработки данных переходит в область Big Data. Помимо этого, все данные относятся к категории геоданных. Обработка геоданных – нетривиальная задача, требующая дополнительных вычислений в условиях отсутствия стандартных методов. Все это накладывает существенное ограничение на использование классических подходов, поэтому возникла необходимость создания собственного алгоритма обработки данных.

Для разработки был выбран язык Python, так как он обладает большим количеством библиотек как для машинного обучения, так и для обработки данных. Одной из таких библиотек является Dask. С ее помощью можно параллельно обрабатывать большие объемы данных, однако в рамках этой задачи, как показал опыт, использование стандартных средств Dask без существенной доработки оказывается неэффективным (см. табл. 3). Проблемы, возникающие при использовании класических подходов, указаны в таблице 3.

Для решения обозначенных проблем разработан адаптивный алгоритм на Dask (рис. 2). Алгоритм организует пакетную обработку данных Dask следующим образом: определяет количество доступной оперативной памяти, вычисляет размер пакета так, чтобы задействовать оптимальное количество оперативной памяти, и обрабатывает весь набор данных пакетами.

Таблица 3

Проблемы класических алгоритмов

Алгоритм	Время работы, ч	Проблема алгоритма
Сопоставление записей в цикле и обработка	13	Неприемлемо долго работает
Распараллеленное по процессам сопоставление записей в цикле и обработка	6	Все еще слишком долго работает
Параллельное сопоставление записей декартовым произведением, построчная обработка и группировка средствами Dask	12	Из-за декартова произведения требуется большой объем оперативной памяти (55 Гб) для обработки данных, что делает алгоритм слишком требовательным к аппаратному обеспечению

За счет оптимизации использования оперативной памяти удалось ускорить обработку данных до 400 с. Помимо этого, алгоритм стал менее требователен к аппаратному обеспечению и может работать как на 32-разрядных, так и на 64-разрядных системах.

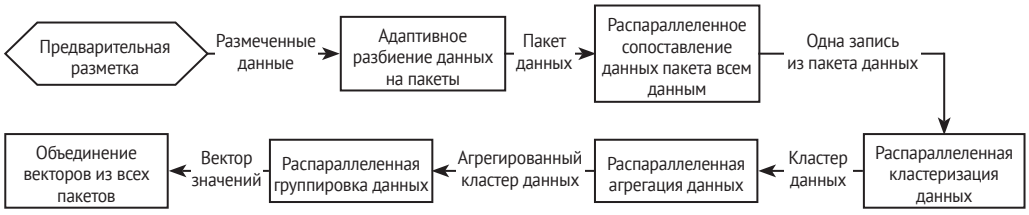

Рис. 2. Схема адаптивного алгоритма

Геоданные в наборах представлены широтой, долготой и адресом. Однако признаки привязаны к метрической системе, поэтому необходима конвертация. Переход от долготы и широты к метрам происходит в процессе вычисления расстояния между объектами по следующей формуле [8]:

$$\delta l = l_1 - l_2,$$

$$\delta\sigma = \arctan\frac{\sqrt{\left(\sin\phi_1\cos\phi_2\right)^2 + \left(\cos\phi_1\sin\phi_2\cos\delta l\right)^2}}{\sin\phi_1\sin\phi_2 + \cos\phi_1\cos\phi_2\cos\delta l}, \quad (3)$$

$$r = R\delta\sigma,$$

где l_1 и l_2 – долгота первой и второй точек соответственно; ϕ_1 и ϕ_2 – широта первой и второй точек соответственно; R – радиус Земли.

Данная формула позволяет получить точный результат даже на малых расстояниях. Если использовать внешние сервисы или дополнительные наборы данных, можно извлечь больше полезных признаков из геоданных. Однако в настоящей работе геоданные используются только для измерения расстояний.

Базовая модель нейронной сети

Предсказание оценки – задача регрессии, так как необходимо восстановить функцию получения оценки по входным данным.

Прежде всего следует определиться со структурой слоев нейронной сети.

Для различных задач и входных данных используются различные структуры слоев. Для решения задачи регрессии используется полносвязный персептрон. В качестве функции потерь – среднеквадратичная ошибка. Функция RMSprop выбрана как оптимизатор, так как она задает адаптивную скорость обучения и при этом устойчива к затуханию. На выходном слое активационная функция отсутствует, поскольку особых преобразований не требуется. На скрытых слоях используется активационная функция – гиперболический тангенс (tanh). Это обусловлено следующими причинами:

- функция симметрична, что обеспечивает быструю сходимость;
- функция имеет непрерывную первую производную;
- функция имеет простую производную, которая может быть вычислена через ее значение, что упрощает вычисления.

После извлечения признаков из данных формируется матрица признаков, где векторы-строки – это экземпляры записей о конкретных точках со всеми признаками. Перед тем как передавать матрицу в нейронную сеть, ее необходимо подготовить, а именно: признаки нужно нормализовать и центрировать. Для этого из каждого значения признака вычитается среднее всех его значений, а затем делится на стандартное отклонение. После такой обработки матрицу можно подавать на вход нейронной сети. За основу базовой модели было взято 4 скрытых слоя по 16 нейронов в каждом (рис. 3).

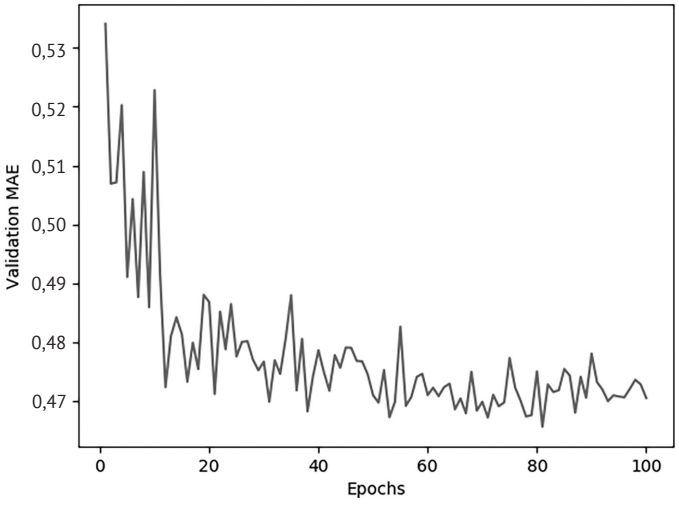

Рис. 3. Базовая модель нейронной сети

Чтобы использовать большее количество данных, тестирование модели проводится при помощи кросс-валидации по 4 блокам. Чтобы заведомо обнаружить момент переобучения, первый прогон будет длительностью 2000 эпох.

На рисунке 4 видно, что переобучение наступает примерно на 200-й эпохе. Для уточнения момента начала переобучения второй прогон был проведен на 250 эпохах (рис. 5). Как оказалось, переобучение наступает уже в окрестности 100-й эпохи.

После этого базовая сеть была обучена в 100 эпох и выдала прогноз со средней абсолютной ошибкой 0,459, что является приемлемым результатом. Тем не менее его можно улучшить посредством вариации количества нейронов и слоев (табл. 4).

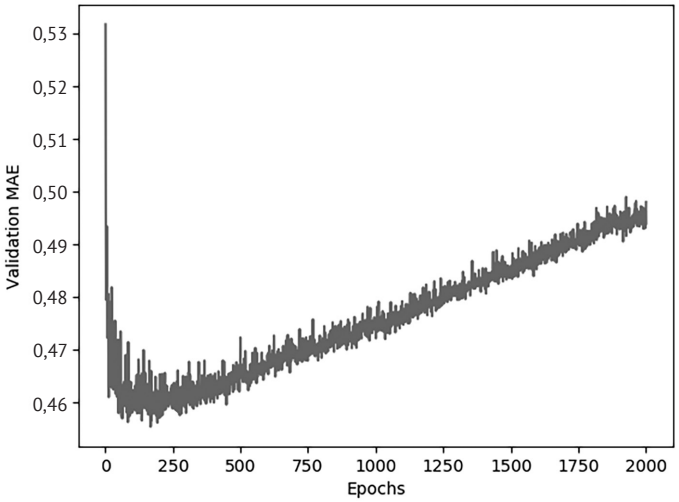

Рис. 4. Прогон базовой нейронной сети на 2000 эпохах

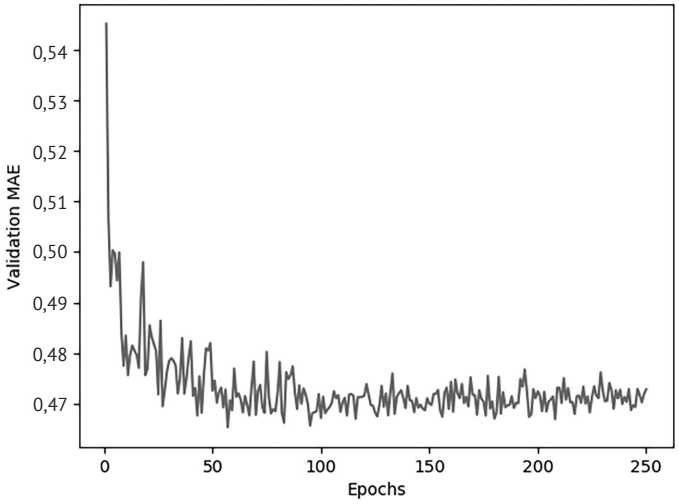

Рис. 5. Прогон базовой нейронной сети на 250 эпохах

Таблица 4

Вариация нейронов

Слои	Количество эпох	Абсолютная средняя ошибка
16-16-16-16	100	0,459
32-32-32-32	70	0,466
32-32-32-23	80	0,448
32-32-23-23	80	0,463
16-16-16-8	100	0,474
16-16-16-16	50	0,474
8-8-8-8-8	80	0,475

По данным таблицы 4 видно, что точность модели варьируется несущественно, а оптимальной является конфигурация 32-32-32-23 с абсолютной средней ошибкой 0,448.

Заключение

В результате проведенного исследования была разработана нейронная сеть, решающая поставленную цель – прогнозирование экспертной оценки. Для практического применения построенной модели необходимо обучить ее на реальных экспертных оценках. В зависимости от того, как строится эта оценка, может потребоваться модификация модели, так как ее структура зависит от данных, используемых при обучении. В частности, в рассмотренном случае линейная модель машинного обучения могла бы дать более точные прогнозы и обучалась бы быстрее, поскольку предложенная гипотетическая оценка изначально построена как линейная модель. Однако такой подход сделал бы модель гораздо менее гибкой: при появлении новых данных ее необходимо было бы обучать заново, а изменение структуры данных приведет к необходимости изменения модели, да и подготовка этих данных потребует гораздо больших усилий.

В перспективе для улучшения работы нейронной сети можно расширить признаковое пространство посредством добавления источников данных или извлечения большего количества признаков из имеющихся.

Автор благодарит профессора И.С. Клименко за полезное обсуждение.

Литература

1. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983. 487 с.
2. *Андрианов В., Леонов А., Бредюк К.* Геомаркетинг: на стыке маркетинга и географии // Маркетинг Менеджмент. 2010. № 7–8.
3. *Угаров А.С.* Методы выбора местоположения торговой точки // Маркетинг в России и за рубежом. 2005. № 6.
4. *Applebaum W.* Can Store Location Be a Science? // Economic Geography. 1965. № 41. P. 234–237.
5. *Chollet F.* Deep Learning with Python. N. Y.: Manning Publications, 2018.
6. *Kane B.J.* A Systematic Guide to Supermarket Location Analysis. N. Y.: Fanchild Publications, 1966.
7. *Nelson R.* The Selection of Retail Locations. N. Y.: F.W. Dodge Corp., 1958.
8. *Vincenty T.* Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations // Survey Review. 1975. № 23. P. 88–93.

Literatura

1. *Ajvazyan S.A., Enyukov I.S., Meshalkin L.D.* Prikladnaya statistika: Osnovy modelirovaniya i pervichnaya obrabotka dannykh. M.: Finansy i statistika, 1983. 487 s.
2. *Andrianov V., Leonov A., Bredyuk K.* Geomarketing: na styke marketinga i geografii // Marketing Menedzhment. 2010. № 7–8.
3. *Ugarov A.S.* Metody vybora mestopolozheniya torgovoj tochki // Marketing v Rossii i za rubezhom. 2005. № 6.
4. *Applebaum W.* Can Store Location Be a Science? // Economic Geography. 1965. № 41. P. 234–237.
5. *Chollet F.* Deep Learning with Python. N. Y.: Manning Publications, 2018.
6. *Kane B.J.* A Systematic Guide to Supermarket Location Analysis. N. Y.: Fanchild Publications, 1966.
7. *Nelson R.* The Selection of Retail Locations. N. Y.: F.W. Dodge Corp., 1958.
8. *Vincenty T.* Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations // Survey Review. 1975. № 23. P. 88–93.