ПРИМЕНЕНИЕ МЕТОДОВ ФАКТОРНОГО АНАЛИЗА И НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ЦЕНЫ ЗАКАЗА СЛУЖБЫ ТАКСИ[52]

Андриянов Н.А.

Финансовый университет при Правительстве Российской Федерации,
Россия, г. Москва Ленинградский пр-т, д.49
naandriyanov@fa.ru

Аннотация: В работе проведено исследование многомерных данных о работе службы заказа такси с целью прогнозирования стоимости заявки. Показано, что можно выделить главные компоненты, на основе которых возможно достаточно точное прогнозирование относительно исходного набора данных.

Ключевые слова: прогнозирование, анализ данных, метод главных компонент, служба заказа такси

Введение

Искусственный интеллект в настоящее время может быть успешно применен в самых различных сферах деятельности человека [1-5]. Одним из актуальных направлений является применение методов машинного обучения для прогнозирования динамики COVID-19 [1,2]. Рассматривается также его вклад в сферу техники и бизнеса [3]. Задача прогнозирования величин на основе множества факторов значительно усложняется при большом количестве параметров и исторических данных. В этом случае возникает задача обучения нейронных сетей. Так, в работе [4] была обучена модель для прогнозирования преступности. Наконец, работа [5] посвящена исследованию результатов и способов внедрения искусственных нейронных сетей в процессы управления социальными и экономическими системами. Действительно, применение интеллектуальных систем для управления бизнесом сегодня всё больше и больше находит применение в его различных сферах. Одним из направления бизнеса, генерирующим огромные объемы информации, является работа службы заказа такси. Ряд работ [6-8] посвящен оценке и прогнозированию количества заказов за определенный период времени. Работа [8] также дает детальные модели для описания потоков трафика такси. Тем не менее, помимо предсказания самого числа заказов, важно предсказать также его стоимость. Это довольно сложная задача, поскольку в настоящее время на стоимость заказа влияет множество факторов. Вместе с тем, в области обработки данных службы заказа такси практически отсутствуют исследования по сокращению размерности. Однако, используя методы сокращения размерности, например, метод главных компонент и его различные модификации [9], можно добиться значительного уменьшения объема обрабатываемых данных при приемлемой эффективности их обработки.

С другой стороны, в некоторых менее интеллектуальных службах заказа такси, сегодня наоборот используется изначально ограниченное число факторов, которое не позволяет настраивать качественную тарификацию, а учитывает, например, лишь дальность или время поездки. В данной работе совместно с менеджерами службы заказа такси были выбраны некоторые факторы и сформированные цены заказов в соответствии с данными факторами. Вместо процесса нормализации данных были выполнены процессы эквалайзинга и масштабирования для ряда данных, поскольку информация в абсолютных величинах составляет коммерческую тайну службы заказа такси. На основе этих данных строятся нейросетевые модели прогнозирования цены заказа по входным параметрам заказа, которые более подробно рассмотрены далее. Кроме того, применяется корреляционный анализ данных и метод главных компонент для сокращения размерности. Нейронные сети переобучаются по новым данным и сравнивается результат прогнозирования в таком случае.

Мерой оценки эффективности прогноза выступает относительная дисперсия ошибки прогноза для цены.

Первый раздел статьи посвящен описанию исходных данных о работе службы заказа такси. Во втором разделе рассматривается применение метода главных компонент, а третий раздел приводит результаты экспериментов по обучению нейронных сетей и прогнозированию данных с их помощью.

[52] *Работа выполнена при поддержке РФФИ, Проект №19-47-730011*

1 Описание исходных данных и их преобразования

После обсуждения с экспертами в области менеджмента в сфере услуг такси, были отобраны некоторые факторы из имеющихся в базах данных службы. Поскольку некоторые параметры описываются категориальными переменными, дата и время имеет соответствующий тип, а большинство факторов всё же имеют числовые значения, было принято решение для простоты анализа все факторы представить в числовом виде. Итак, в качестве входных данных для первичного обучения были выбраны следующие факторы:

- время, в которое был осуществлен заказ (поскольку теоретически час-пик может быть использован для повышения цен). Данный фактор подвергался эквалайзингу в интервале от 0 до 1, где 0 – начало дня, 1 –конец дня;
- корректировки стоимости). Данный фактор был преобразован в оценку экспертов по шкале от «-1» до «1», где «-1» соответствует такой погоде, когда требуется понизить цену, а «+1» – когда требуется повысить цену заказа;
- минимальная стоимость поездки (величина, также называемая «стоимость посадки»), может зависеть от конкурентных цен, от политики самой службы заказа такси, и естественно влияет на суммарную стоимость заказа. Данный фактор подвергался масштабированию с целью сохранения коммерческой тайны службы заказа такси;
- количество свободных машин в районе заказа (поскольку при необходимости посадки и последующей транспортировки клиента может понадобиться перегонять машину из района в район, то цена должна соответствовать такому перемещению). Данный фактор нормировался по величине 10 машин, т.е. если была одна машина в районе, то значение фактора было 0.1.
- дальность поездки (поскольку это один из главных факторов, определяющих начисляемую стоимость заказа за расстояние). Данный фактор не изменялся и использовался в км.

На основе выбранных факторов рассчитывалась стандартная цена поездки в соответствии с простыми алгоритмами на базе расстояния. Однако данная цена корректировалась менеджерами заказа такси на основе формируемого ими управляющего воздействия. Управляющее воздействие могло повышать или понижать рассчитанную программой стоимость на некоторый процент. Для простоты анализа рассмотрено влияние перечисленных условий на минимальную стоимость поездки.

Всего по данным параметрам было проанализировано 104 заказа. При этом многомерный набор данных имел вид, представленный в таблице 1 (для первых 10 заказов).

Таблица 1. Пример исходных для анализа данных

Время	Погода	Посадка	Свободные машины	Расстояние	Управляющее воздействие	Цена
0.833333	-0.58	0.97	0.2	3	-0.03	0.726087
0.28125	1	1.25	0.2	4.3	0.25	1.130435
0.020833	-0.46	0.98	0.7	2.2	-0.02	0.647826
0.385417	-0.92	0.95	0.1	2.7	-0.05	0.678261
0.895833	-0.6	0.98	0.3	4.4	-0.02	0.878261
0.020833	-0.36	1	0.1	0.6	0	0.5
0.354167	-1	0.96	0.4	5.2	-0.04	0.943478
0.989583	0.04	0.97	0.3	3.2	-0.03	0.743478
0.802083	-0.94	1.01	0.4	2.4	0.01	0.695652
0.96875	-0.18	1	0.1	2.7	0	0.717391

Анализ данных в таблице 1 показывает, что для заказов 2 и 5, близких по расстоянию, корректировки приводят к достаточно большой разнице по цене. В первую очередь, это объясняется ранним временем второго заказа (пиковая нагрузка), а также плохими погодными условиями, что привело к 25% корректировке стоимости посадки, что в свою очередь, сильно повлияло на итоговую стоимость заказа. Тем не менее, в остальных представленных случаях можно отметить незначительное управляющее воздействие, и явную зависимость стоимости заказа от расстояния.

Рассмотрим влияние иных факторов на стоимость заказа. На рис. 1 представлены точечные графики, где по одной из осей откладывается исследуемый фактор, а по другой – итоговая стоимость поездки. В частности, зависимость итоговой стоимости заказа от времени его поступления представлена на рис. 1а (следует отметить, что время распределено в диапазоне от 0 до 1). В то же время на рис. 1б представлена зависимость итоговой стоимости заказа при различной погоде (характеризуется числовыми значениями от «-1» до «1»). На рис. 1в показано, как на цену заказа

влияет количество свободных машин в районе заказа (нормировано с коэффициентом 0,1). Наконец, на рис. 1г показана связь между главным фактором («расстояние») и стоимостью поездки. Важно отметить, что стоимость, как объясняемая переменная, всегда отложена по оси Y, а факторы, как объясняющие переменные, всегда отложены по оси X. При этом окончательная цена заказа масштабирована по некоторому среднему значению.

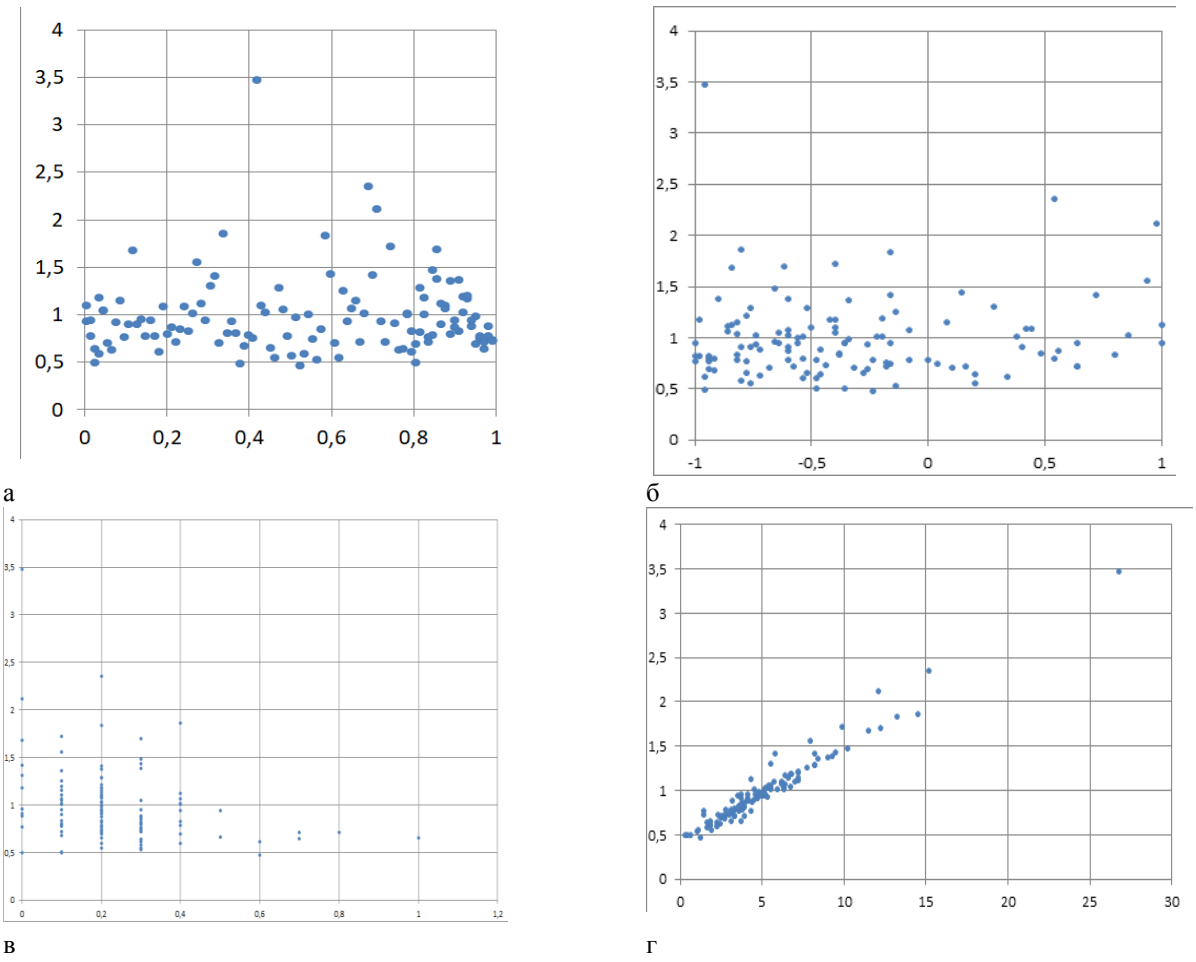


Рис. 1. Соотношение цены заказа и остальных факторов

Анализ двумерных отображений данных, представленных на рис. 1, показывает, что линейная зависимость и сильная корреляция между стоимостью поездки и другими параметрами наблюдается только для расстояния. Это объясняется тем, что базовая модель службы заказа такси действительно настроена так, что оперирует только расстоянием. Однако для ведения эффективного бизнеса необходимо построить гораздо более интеллектуальную систему, способную учитывать и другие факторы. Эти факторы учитываются в управляющем воздействии, которое направлено на изменение цены в различных условиях заказа.

На рис. 2 показаны аналогичные отображения на двумерной плоскости для различных факторов и управляющего воздействия. В частности, зависимость управляющего воздействия от времени поступления заказа представлена на рис. 2а (следует отметить, что время распределено в диапазоне от 0 до 1). В то же время на рис. 2б представлена зависимость управляющего воздействия на стоимость заказа при различной погоде (характеризуется числовыми значениями от «-1» до «1»). На рис. 2в показано, как на управляющее воздействия на цену заказа влияет количество свободных машин в районе заказа (нормировано с коэффициентом 0,1). Наконец, на рис. 2г показана связь расстоянием и управляющим воздействием. Важно отметить, что управляющее воздействие, как объясняемая переменная, всегда отложено по оси Y, а факторы, как объясняющие переменные, всегда отложены по оси X.

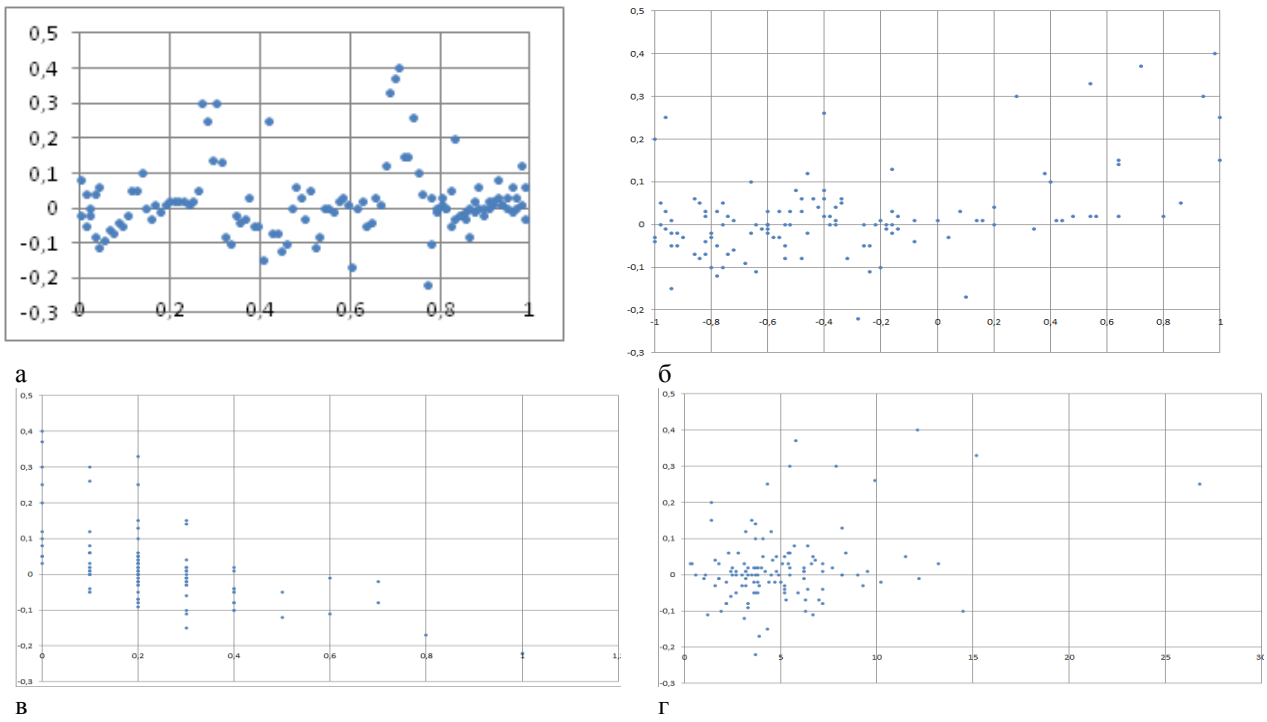


Рис. 2. Соотношение управляющего воздействия на стоимость заказа и остальных факторов

Анализ представленных зависимостей между переменными показывает, что управляющее воздействие имеет сложные нелинейные связи с исследуемыми факторами. Особенно это сильно заметно по изменению рис. 2г по сравнению с рис. 1г. Для понимания такой связи можно использовать нейронные сети с обратным распространением ошибки, поскольку у нас имеется выходная переменная – управляющее воздействие, а входными переменными являются все исследуемые факторы.

Таким образом, будет рассмотрена задача обучения и прогнозирования как для управляющего воздействия, так и для итоговой стоимости заказа.

2 Сокращение размерности данных

Исходный набор данных содержит 7 столбцов, среди которых 2 объясняемых (Y1,Y2) и 5 объясняющих (X1,X2,X3,X4,X5) переменных, а также 104 строки, соответствующих определенному заказу. Поскольку для применения метода главных компонент необходимо вычислить собственные значения корреляционной матрицы, построим ее для исследуемых данных. Важно отметить, что выбранный метод обучения нейронных сетей подразумевает наличие обучающей и тестовой выборки, то данные о первых 86 заказов будут составлять обучающую выборку, а оставшиеся 18 заказов будут использоваться для тестирования. Корреляционная матрица рассчитана для обучающего набора данных и представлена в таблице 2.

Таблица 2. Корреляционная матрица

Факторы	**X1**	**X2**	**X3**	**X4**	**X5**	**Y1**	**Y2**
X1	1	-0.044	0.119	-0.034	0.007	0.119	0.036
X2	-0.044	1	0.504	0.007	0.096	0.504	0.225
X3	0.119	0.504	1	-0.514	0.293	1	0.530
X4	-0.033	0.007	-0.514	1	-0.199	-0.514	-0.309
X5	0.007	0.096	0.293	-0.199	1	0.293	0.966
Y1	0.119	0.504	1	-0.514	0.293	1	0.530
Y2	0.036	0.225	0.530	-0.309	0.966	0.530	1

Анализ данных в таблице 2 показывает, что имеется линейная связь между факторами Y1 (управляющее воздействие) и X3 (цена посадки). Это объясняется тем, что изначально цена посадки имеет фиксированное значение, а управляющее воздействие направлено именно на ее изменение. Следовательно, данный фактор можно исключить из анализа (Таблица 3).

Таблица 3. Сокращение размерности корреляционной функции

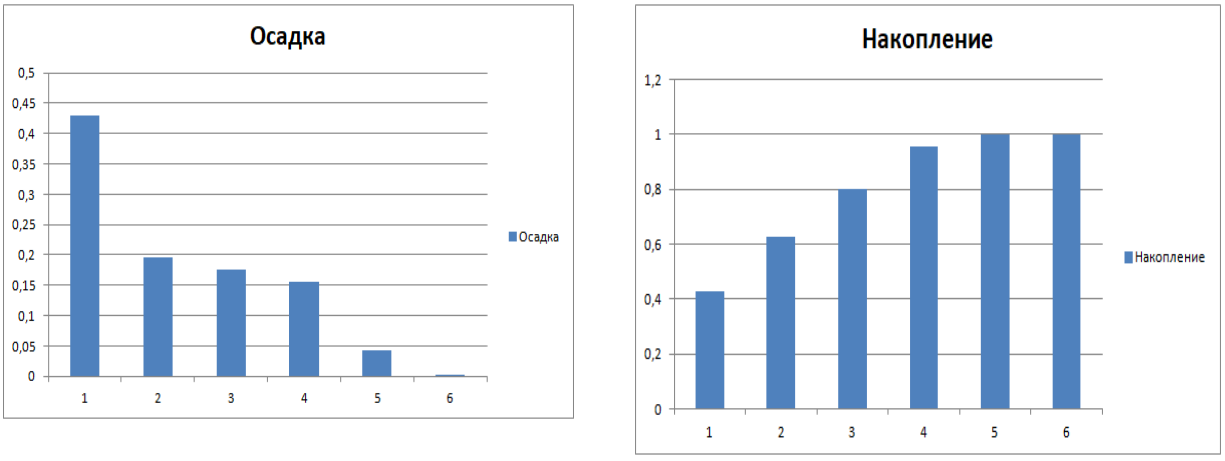
Факторы	X1	X2	X4	X5	Y1	Y2
X1	1	-0.044	-0.034	0.007	0.119	0.036
X2	-0.044	1	0.007	0.096	0.504	0.225
X4	-0.033	0.007	1	-0.199	-0.514	-0.309
X5	0.007	0.096	-0.199	1	0.293	0.966
Y1	0.119	0.504	-0.514	0.293	1	0.530
Y2	0.036	0.225	-0.309	0.966	0.530	1

На основе таблицы 3 рассчитаем собственные значения и собственные векторы для корреляционной матрицы. Собственные значения описываются вектором из шести величин: $\lambda = (2.5744 \quad 1.1743 \quad 1.0535 \quad 0.9356 \quad 0.2620 \quad 0.0002)$. Собственные векторы для новых компонент представлены в таблице 4.

Таблица 4. Собственные векторы

V1	V2	V3	V4	W1	W2
0.051449437	-0.113495786	0.754958642	0.631938658	-0.123117672	0.002163894
0.264535128	-0.595412263	-0.470822021	0.338743304	-0.488963052	-0.000367124
0.48409108	-0.466268311	0.075796329	-0.075792495	0.704335883	-0.201670687
-0.33398733	0.187214799	-0.431660031	0.668061132	0.469805452	-0.00706123
0.501179329	0.518528522	-0.1.02200368	0.145339978	-0.160684793	-0.650045556
0.574718793	0.33357997	-0.076444046	0.112838907	0.055958646	0.732608375

На рис. 3 построен график осадки для собственных значений. При этом рис. 3а показывает относительную описываемую долю изменчивости дисперсии по компонентам, а рис. 3б – кумулятивное накопление описываемой доли изменчивости дисперсии по компонентам.



а б

Рис. 3. График осадки

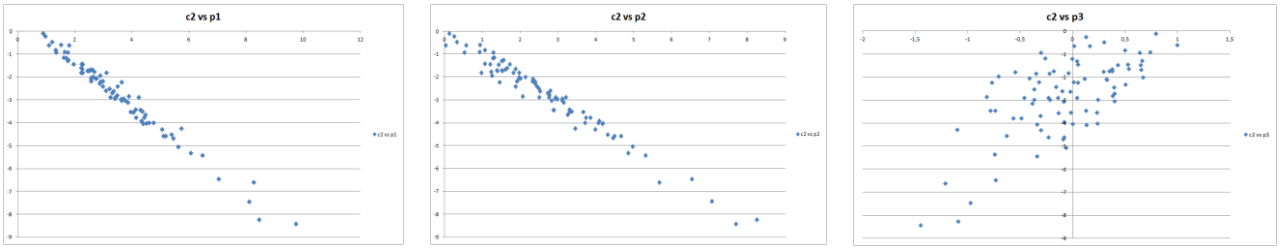
Анализ представленных графиков показывает, что в новом разложении, наибольшее количество информации содержится в нескольких первых компонентах. Поскольку собственные значения первых трёх компонент больше 1, то будет выполнять дальнейший анализ исходя из необходимости учета трёх новых факторов.

В таблице 5 представлены новые данные для описания работы службы заказа такси на основе трёх компонент, а также прогнозируемая 6-я компонента для первых десяти заказов.

Таблица 5. Сокращение размерности данных

p1	p2	p3	c2
2.21304885	1.63371562	0.52729328	-1.61321592
3.6220749	1.43403766	-0.77595309	-2.22028903
1.59491618	1.30249712	-0.26993844	-1.15786358
1.9459384	1.70608395	0.42518041	-1.44934493
2.97152418	2.42929128	0.38677329	-2.41437605
0.94459856	0.24234419	0.11831231	-0.22592271
3.23318625	3.19355795	0.03473846	-2.88433166
2.71677141	1.99648033	0.10574998	-2.0523066
3.87824064	2.04488086	-0.82492914	-2.84143567
2.56748254	1.95326584	-0.4177632	-2.02612386

Теперь рассмотрим описание зависимостей компоненты c2 от главных компонент p1,p2,p3. Для этого также представим точечные отображения на плоскости. Рис. 4 показывает связи между новыми данными.



а б в

Рис. 4. Связи между главными компонентами

Анализ распределений, представленных на рис. 4, показывает, что во многом прогнозируемую переменную можно описывать с помощью первой или второй компоненты. В частности, индекс детерминации в линейной модели на основе данных рис. 4а составляет R=0.9765, а для данных на рис. 4б – R=0.9682.

Таким образом, прогнозирование возможно без обучения нейронных сетей на основе моделей линейной регрессии. Однако задача требует получения на основе собственных векторов новых компонент, прогнозирование новой переменной и затем обратное преобразование к исходным размерностям данных.

3 Обучение нейронных сетей и прогнозирование на основе линейной регрессии главных компонент

Исходный датасет без выделения главных компонент содержал 104 вектора с 5 входными и 2 выходными параметрами. Среди 104 векторов 86 было выбрано для обучения, остальные 18 пошли на тестовую выборку. Поскольку большинство входных переменных не имело сильной корреляции с выходной переменной были выбраны нейросетевые модели. Использовалась обычная структура многослойного персептрона с полносвязным слоем на выходе. Обучение производилось методом обратного распространения ошибки в течение 100 эпох. Также рассмотрено обучение с помощью сетей Элмана – рекуррентных нейронных сетей [10]. И самый простой прогноз делается на основе модели общей регрессии.

В результате прогноза 18 тестовых параметров Y1 (управляющее воздействие), Y2 (итоговая стоимость) может быть посчитано среднее отклонение прогноза, как

$$\sigma_{er} = \frac{\sigma_{Y2}\sqrt{\sum_{i=1}^{18}\left(Y1_{pr(i)} - Y1_i\right)^2} + \sigma_{Y1}\sqrt{\sum_{i=1}^{18}\left(Y2_{pr(i)} - Y2_i\right)^2}}{36\sigma_{Y1}\sigma_{Y2}}, \quad (1)$$

где σ_{Y1} и σ_{Y2} – среднеквадратичные отклонения в наборе данных по управляющему воздействию и по цене; $Y1_{pr}$ и $Y2_{pr}$ – прогнозируемые значения управляющего воздействия и цены соответственно.

Результаты относительной среднеквадратичной ошибки представлены в таблице 6 для прогнозов управляющего воздействия и стоимости по обучающей выборке и тестовой при различном числе нейронов в сети. В первом столбце указаны названия моделей обучения и количество нейронов во

внутреннем слое, а во втором и третьем – дисперсии ошибок прогнозирования, рассчитанные как квадрат из выражения (1).

Таблица 6. Ошибка прогнозирования без сокращения размерности

Сеть	Дисперсия ошибки на обучающей выборке	Дисперсия ошибки на тестовой выборке
ОРО 1	0.05853497	0.0411185
ОРО 3	0.00315664	0.00409658
ОРО 5	0.01757205	0.0073322
ОРО 10	0.01094354	0.00549245
ОРО 50	0.13521128	0.43645541
Элман 10	0.01240906	0.01364026
ОР	0.00186543	0.40803577

В таблице 6 использованы следующие обозначения: многослойный персептрон – нейронная сеть с обратным распространением ошибки (ОРО), рекуррентная нейронная сеть (Элман), модель общей регрессии (ОР). Числовое значение после названия модели говорит о количестве нейронов во внутреннем слое.

Анализ полученных ошибок, представленных в таблице 6, показывает, что при прогнозировании тестовой цены и управляющего воздействия наиболее близкие к реальным значениям прогнозы обеспечивает сеть с обратным распространением ошибки, у которой задействовано 3 нейрона во внутреннем слое. Возможно, это связано с количеством входных параметров, которых всего 5, однако некоторые из которых оказывают маленькое влияние на итоговую стоимость. Модель общей регрессии на обучающей выборке выявила зависимость между минимальной стоимостью и управляющим воздействием, что привело к минимальной ошибке. Однако на тестовой выборке данная модель показала один из худших результатов. Также можно видеть, что насыщение сети до 50 нейронов приводит к значительному снижению качества модели, поскольку обрабатываемые данные могут описываться гораздо более простыми связями.

Для главных компонент прогнозировался только параметр c2, что значительно упростило модель и вычисление ошибки, которое для измененных данных выполняется в соответствии с выражением:

$$\sigma^{*}_{er} = \frac{\sqrt{\sum_{i=1}^{18}\left(c1_{pr(i)} - c1_i\right)^2}}{18\sigma_{c1}}, \qquad (2)$$

где σ_{c1} – среднеквадратичное отклонения в наборе данных по компоненте c1; $c1_{pr}$ – прогнозируемые значения компоненты c1.

Результаты расчетов представлены в таблице 7. Следует отметить, что использовалась только модель регрессии.

Таблица 7. Ошибка прогнозирования при сокращении размерности

Сеть	Дисперсия ошибки на обучающей выборке	Дисперсия ошибки на тестовой выборке
ОР	0.00008	0.00012

Анализ полученных результатов показывает, что в пространстве новых компонент выполняется более точный прогноз. Однако не следует забывать, что при возврате к обычным данным погрешность прогноза снова может возрасти.

Заключение

Проведен факторный и интеллектуальный анализ данных о работе службы заказа такси. Для прогнозирования цены от различных условий формирования заказа применены модели машинного обучения, включая сети с обратным распространением ошибки, рекуррентные сети и модели множественной регрессии. Пожелания менеджеров по воздействию на итоговую стоимость поездки в пространстве исходных данных лучше всего учитываются моделью с обратным распространением ошибок с 3 нейронами во внутреннем слое. Это связано с небольшой размерностью данных и связями внутри переменных. Наличие таких связей позволило перейти к новым компонентам с помощью метода главных компонент. В это пространстве удалось уменьшить дисперсию ошибки в 3,5 раза по сравнению с наилучшей моделью в исходном пространстве признаков. Дальнейшие исследования могут быть связаны с обратным преобразованием данных и большим уточнением моделей для

пространства главных компонент. Разработанные модели могут быть учтены при формировании цен на поездку в такси в различных условиях, включая время заезда, погодные условия и количество близстоящих машин.

Литература

1. *Niazkar H.R., Niazkar M.* Application of artificial neural networks to predict the COVID-19 outbreak // Glob health res policy. Vol. 50. 2020, №5. – P. 123-128. doi: 10.1186/s41256-020-00175-y.
2. *Farooq J., Abid B.M.* A deep learning algorithm for modeling and forecasting of COVID-19 in five worst affected states of India // Alexandria Engineering Journal. Vol. 60. 2021, №1. – P. 587-596.
3. *Иванов А.А., Рожкова Л.В.* Искусственный интеллект как основа инновационных преобразований в технике, экономике, бизнесе // Известия СПбГЭУ. 2018, №3. – С. 111-114.
4. *Суходолов А.П., Бычкова А.М.* Искусственный интеллект в противодействии преступности, ее прогнозировании, предупреждении и эволюции // Всероссийский криминологический журнал. 2018, №6. – С. 753-766
5. *Тетерин Д.А., Хабибулин Р.Ш., Гудин С.В.* Обзор применения искусственных нейронных сетей в управлении социальными и экономическими системами // Экономика и Информатика. Vol. 45. 2018, №3. – P. 574-583.
6. *Andriyanov N A., Sonin V.A.* Using mathematical modeling of time series for forecasting taxi service orders amount // CEUR Workshop Proceedings. Vol. 2258. 2018. – P. 462–472.
7. *Andriyanov N.A.* Development of Prediction Methods for Taxi Order Service on the Basis of Intellectual Data Analysis // Advances in Intelligent Systems and Computing. Vol. 1230 AISC. 2020. – P. 652–664. doi: 10.1007/978-3-030-52243-8_49
8. *Zhizhen L., Hong Ch., Yan L., and Qi Zh.* Taxi Demand Prediction Based on a Combination Forecasting Model in Hotspots // Journal of Advanced Transportation. Vol. 2020. 2020. Article ID 1302586. – P. 1-13.
9. *Поляк Б.Т., Хлебников М.В.* Метод главных компонент: робастные версии. // Автомат. и телемех., 2017, № 3. – С. 130–148.
10. *Лила В.Б., Пучков Е.В.* Методология обучения рекуррентной искусственной нейронной сети с динамической стековой памятью // Программные продукты и системы. 2014, №14. – С. 132-135.