

Применение регрессионного анализа и нейронных сетей для построения моделей рынка подержанных автомобилей Toyota Prius

Козич Виталий Геннадьевич

*Приамурский государственный университет имени Шолом-Алейхема
Студент*

Бондаренко Владислав Витальевич

*Приамурский государственный университет имени Шолом-Алейхема
Студент*

Баженов Руслан Иванович

*Приамурский государственный университет им.Шолом-Алейхема
к.п.н., доцент, зав.кафедрой информационных систем, математики и
методик обучения*

Аннотация

В данной статье рассматриваются методы регрессионного анализа в среде Gretl и построения нейронной сети в Neural Network Wizard. Исследование производится на основе данных рынка подержанных автомобилей Toyota Prius. Экспериментальные данные были взяты с сайта drom.ru. В результате исследования получены модели, позволяющие спрогнозировать стоимость покупки либо продажи подержанного автомобиля в зависимости от определенных параметров, которые считаются важными.

Ключевые слова: регрессионный анализ, нейронная сеть, рынок подержанных автомобилей

Using regression analysis and neural networks to build models of the market for second-hand cars Toyota Prius

Kozich Vitaliy Gennadievich

*Sholom-Aleichem Priamursky State University
Student*

Bondarenko Vladislav Vitalievich

*Sholom-Aleichem Priamursky State University
Student*

Bazhenov Ruslan Ivanovich

*Sholom-Aleichem Priamursky State University
Candidate of pedagogical sciences, associate professor, Head of the Department of Information Systems, Mathematics and teaching methods*

Abstract

The article deals with methods of regression analysis in the environment Gretl and methods of neural network of Neural Network Wizard. The research is based on data from the market of used the Toyota Prius car model. Experimental data were taken from the site drom.ru. The study resulted in models allowing to predict the cost of buying or selling a second-hand car depending on certain parameters, which are considered important.

Keywords: regression analysis, neural network, the market for used cars

В настоящее время почти каждый нуждается в автомобиле. Для большинства граждан предпочтительнее купить поддержанный автомобиль, который экономичен и имеет низкий уровень загрязнения, такими качества обладает Toyota Prius. Какую сумму правильно будет заплатить за этот бывший в употреблении автомобиль? Чтобы правильно дать ответ, необходимо учесть определенные параметры: год выпуска, объем двигателя, мощность, пробег. На основе данных по рынку предложений можно построить различные модели для прогнозирования цены. Для этого хорошо подходят регрессионный анализ и искусственные нейронные сети.

Ряд ученых изучает проблемы и методы построения различных регрессионных моделей и нейронных сетей. Линейный множественный регрессионный анализ в статистическом машинном эксперименте рассматривал С.В.Недошивин [1]. Корреляционно-регрессионный анализ влияния инвестиционной активности на рост регионального валового продукта изучал А.С.Коротченко [2]. Сущность, преимущества и недостатки кластерного анализа показал С.А.Суслов [3, 4]. Р.И.Остапенко [5, 6]. рассмотрел особенности моделирования латентных изменений. А.А.Серов исследовал скоринг экспериментальных данных с применением прогнозных моделей [7]. Р.И.Баженов и др. изучали возможности применения различных видов анализа [8-14]. Применением нейронных сетей для аппроксимации данных занималась Т.В. Филатова [15]. А.В. Чернявский и В.Г. Спицын [16] разработали эволюционирующую нейронную сеть для улучшения качества изображений. В.Ю.Осипов показал применение рекуррентной нейронной сети с двумя сигнальными системами [17]. В.Г.Манжула и др. [18] изучали нейронные сети Кохонена и нечеткие нейронные сети в интеллектуальном анализе данных. О качестве восстановления образов искусственной нейронной сетью Хопфилда писали И.В.Колбасина и др. [19]. В.А.Седов и Н.А.Седова моделировали расхождения судов в зоне чрезмерного сближения нейронными сетями [20, 21]. Использованием нейронных сетей для оценки уровня заболоченности территории на основе данных дистанционного зондирования занимались А.Д.Варламов и Р.В.Шарапов [22, 23]. Е.А.Шумков и И.К.Чистик [24] использовали генетические алгоритмы для обучения нейронных сетей. Гимаров В.А. и др. разработали нейро-нечеткий метод классификации объектов с разнотипными признаками [25]. Р.И.Баженов и др. [26-29] применили нейронные сети в распознавании

образов, прогнозировании. Изучением методов регрессионного анализа и нейронных сетей занимаются также зарубежные ученые [30-32].

Рассмотрим разработку регрессионной модели в среде Gretl и нейронную сеть в Neural Network Wizard для нахождения зависимости стоимости автомобиля от данных параметров.

По данным сайта www.drom.ru за октябрь, ноябрь 2015 г. был осуществлен сбор экспериментальных наблюдений. Рассматривались объявления о продаже автомобилей Toyota Prius по всей России. Были выбраны следующие критерии: год, объем двигателя, мощность, пробег, стоимость.

Создадим таблицу данных в MS Excel. Используя данные 100 объявлений (рис. 1).

	A	B	C	D	E
1	Год	Объем двигателя, л	Мощность, л.с.	Пробег, км	Стоимость, руб.
2	2004	1,5	77	94000	375000
3	2009	1,8	99	108000	650000
4	2004	1,5	76	200000	410000
5	2001	1,5	72	250000	247000
6	2009	1,5	76	100000	560000
7	2010	1,8	99	48000	635000
8	2007	1,5	76	80000	490000
9	2005	1,5	76	179000	380000
10	2008	1,5	76	117000	430000
11	2009	1,8	99	80000	630000
12	2009	1,8	99	117000	600000
13	2005	1,5	77	200000	410000
14	2011	1,8	99	80000	770000
15	2010	1,8	99	75000	710000
16	2011	1,8	99	99000	735000
17	2009	1,8	99	68000	670000
18	2007	1,5	76	120000	450000
19	2006	1,5	77	120000	420000
20	2008	1,5	76	160000	470000
21	2010	1,8	99	79000	780000

Рисунок 1 - Фрагмент данных в Excel

Введем обозначения переменных: год – God, объем двигателя – V, мощность – P, пробег – S, стоимость – price (рис. 2).

	A	B	C	D	E
1	God	V	P	S	price
2	2004	1,5	77	94000	375000
3	2009	1,8	99	108000	650000
4	2004	1,5	76	200000	410000
5	2001	1,5	72	250000	247000

Рисунок 2 - Обозначение переменных

Следующий шаг – импортировать данные таблицы MS Excel в Gretl. В окне программы Gretl появляются переменные, которые необходимы, чтобы построить регрессионную модель (рис. 3).

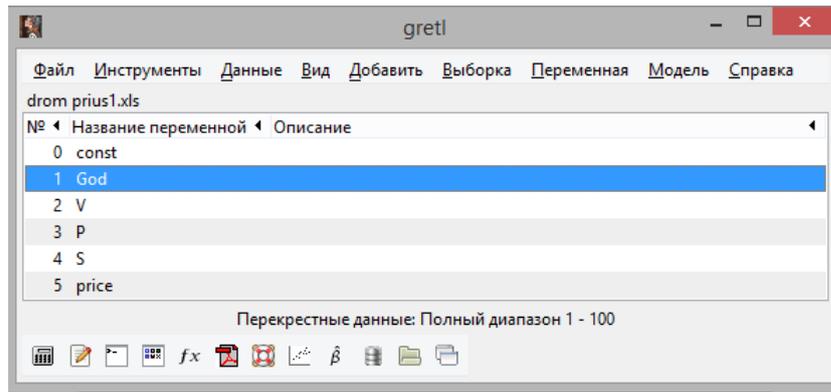


Рисунок 3_- Рабочее окно программы Gretl

Необходимо просмотреть получившуюся таблицу (рис. 4-5).

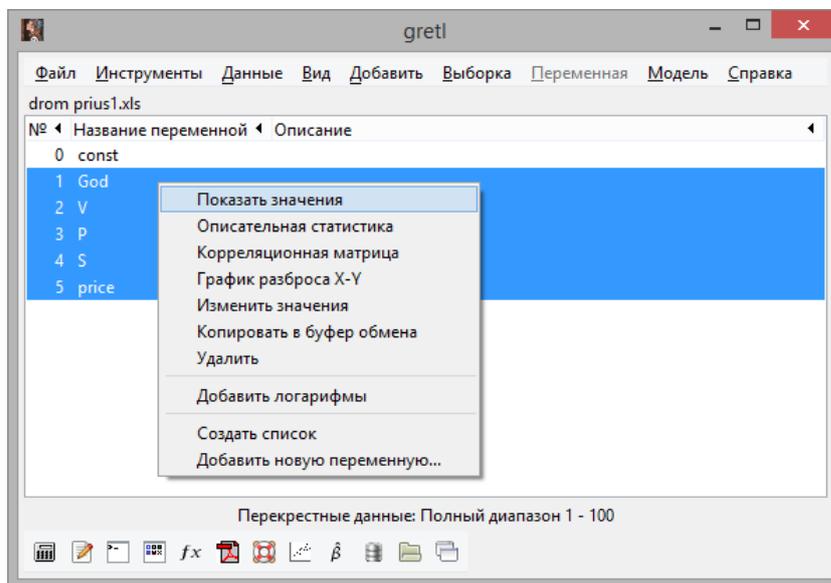


Рисунок 4_- Контекстное меню выделенных переменных

The screenshot shows a window titled 'gretl: показать данные'. It displays a data table with the following columns: 'God', 'V', 'P', 'S', and 'price'. The rows represent individual observations, numbered 1 through 17. The data is as follows:

	God	V	P	S	price
1	2004	1,5	77	94000	375000
2	2009	1,8	99	108000	650000
3	2004	1,5	76	200000	410000
4	2001	1,5	72	250000	247000
5	2009	1,5	76	100000	560000
6	2010	1,8	99	48000	635000
7	2007	1,5	76	80000	490000
8	2005	1,5	76	179000	380000
9	2008	1,5	76	117000	430000
10	2009	1,8	99	80000	630000
11	2009	1,8	99	117000	600000
12	2005	1,5	77	200000	410000
13	2011	1,8	99	80000	770000
14	2010	1,8	99	75000	710000
15	2011	1,8	99	99000	735000
16	2009	1,8	99	68000	670000
17	2007	1,5	76	120000	450000

Рисунок 5_- Просмотр таблицы данных

Перейдем к построению уравнения модели $y = a * x_1 + b * x_2 + c * x_3 + d * x_4$. Для решения нашей задачи найдем регрессионную модель, используя метод наименьших квадратов (рис. 6, 7).

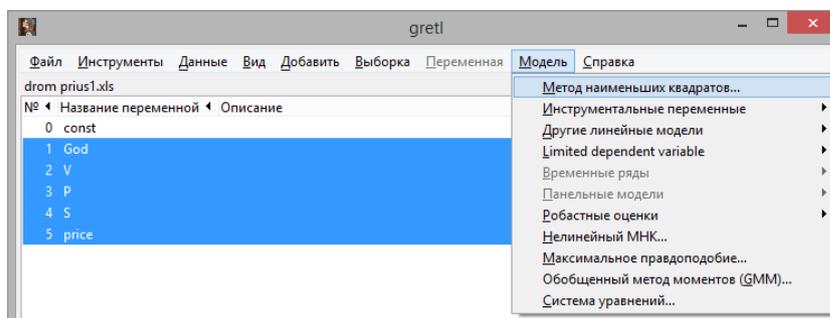


Рисунок 6 - Меню Модель

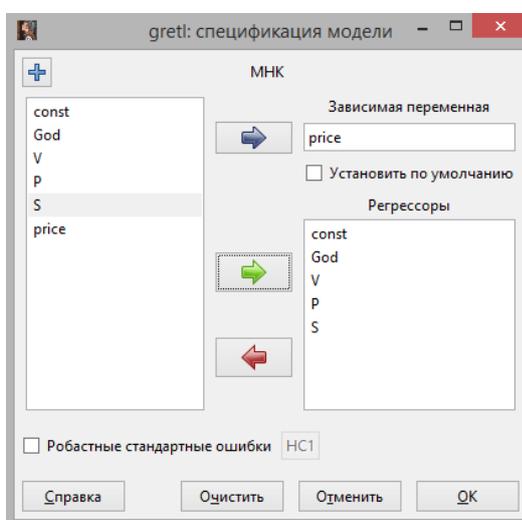


Рисунок 7 - Окно спецификации модели

Получившаяся модель и ее описательные статистики показаны на рис. 8.

	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	-7,25821e+07	6,33972e+06	-11,45	1,39e-019 ***
God	35970,4	3159,37	11,39	1,89e-019 ***
V	620254	97789,6	6,343	7,60e-09 ***
P	-1424,98	1324,44	-1,076	0,2847
S	0,176913	0,151660	1,167	0,2463

Среднее зав. перемен	541550,0	Ст. откл. зав. перемен	159226,7
Сумма кв. остатков	3,19e+11	Ст. ошибка модели	57960,09
R-квадрат	0,872850	Испр. R-квадрат	0,867497
F(4, 95)	163,0379	P-значение (F)	1,21e-41
Лог. правдоподобие	-1236,080	Крит. Акаике	2482,160
Крит. Шварца	2495,186	Крит. Хеннана-Куинна	2487,432

Исключая константу, наибольшее p-значение получено для переменной 3 (P)

Рисунок 8 - Регрессионная модель

По значению исправленного R-квадрата можно судить о доле вариации результативного признака с учетом воздействия изучаемых факторов. В данной модели 86,7497% вариации переменной price зависит от влияния включенных факторов. Если коэффициент выше 80%, то модель считается достаточно хорошей.

Найдем зависимость $\ln(\text{price})$ от $\ln(\text{God})$, V, P, $\ln(S)$ (рис.9).

	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	-1396,54	97,6524	-14,30	2,01e-025 ***
l_God	185,175	12,8293	14,43	1,10e-025 ***
V	0,881885	0,212502	4,150	7,25e-05 ***
P	-0,00229602	0,00288075	-0,7970	0,4274
l_S	0,0204768	0,0267241	0,7662	0,4454

Среднее зав. перемен	13,15019	Ст. откл. зав. перемен	0,346443
Сумма кв. остатков	1,510541	Ст. ошибка модели	0,126097
R-квадрат	0,872874	Испр. R-квадрат	0,867522
F(4, 95)	163,0729	P-значение (F)	1,20e-41
Лог. правдоподобие	67,74126	Крит. Акаике	-125,4825
Крит. Шварца	-112,4567	Крит. Хеннана-Куинна	-120,2107

Исключая константу, наибольшее p-значение получено для переменной 7 (l_S)

Рисунок 9 - Новая регрессионная модель

В этой модели исправленный R-квадрат=86,7522% у нас чуть больше. Значит, модель более точная, чем предыдущая.

Перейдем к решению уравнения

$$\ln(\text{price}) = a \cdot \ln(\text{God}) + b \cdot V + c \cdot P + d \cdot \ln(S) + \text{const}$$

Введем в Excel данные const, l_God, V, P, l_S. В ячейку, окрашенную в зеленый цвет, вводим формулу расчета (рис.10). В таблицу Excel вводим необходимые параметры нашего автомобиля, который мы хотим купить или продать, в ячейки, окрашенные в голубой цвет.

	A	B	C	D	E	F	G
1	const	-1396,54					
2	l_God	185,175		Год	2004		
3	V	0,881885		Объем двигателя, л	1,5		
4	P	-0,00229602		Мощность, л.с.	77		
5	l_S	0,0204768		Пробег, км	94000		
6							
7	l_Price	13					
8							

Рисунок 10 - Ввод данных

В ячейку, окрашенную в желтый цвет, вводим формулу расчета стоимости автомобиля. После этого в ячейке отобразится стоимость данного автомобиля (рис.11).

	A	B	C	D	E	F
1	const	-1396,54				
2	I_God	185,175		Год	2004	
3	V	0,881885		Объем двигателя, л	1,5	
4	P	-0,00229602		Мощность, л.с.	77	
5	I_S	0,0204768		Пробег, км	94000	
6						
7	I_Price	13				
8						
9	Цена	330250,888				
10						

Рисунок 11 - Расчет стоимости автомобиля по заданным параметрам

Если мы хотим узнать стоимость другого автомобиля Toyota Prius, то нужно просто в ячейки, окрашенные в голубой цвет ввести другие параметры.

Далее, создадим нейронную сеть в программе Neural Network Wizard. Для этого, предварительно, создадим текстовый файл с данными об автомобилях (рис. 12).

God	V	P	S	price
2004	1,5	77	94000	375000
2009	1,8	99	108000	650000
2004	1,5	76	200000	410000
2001	1,5	72	250000	247000
2009	1,5	76	100000	560000
2010	1,8	99	48000	635000
2007	1.5	76	80000	490000

Рисунок 12 – Фрагмент данных в текстовом документе

Откроем его в программе (рис. 13).

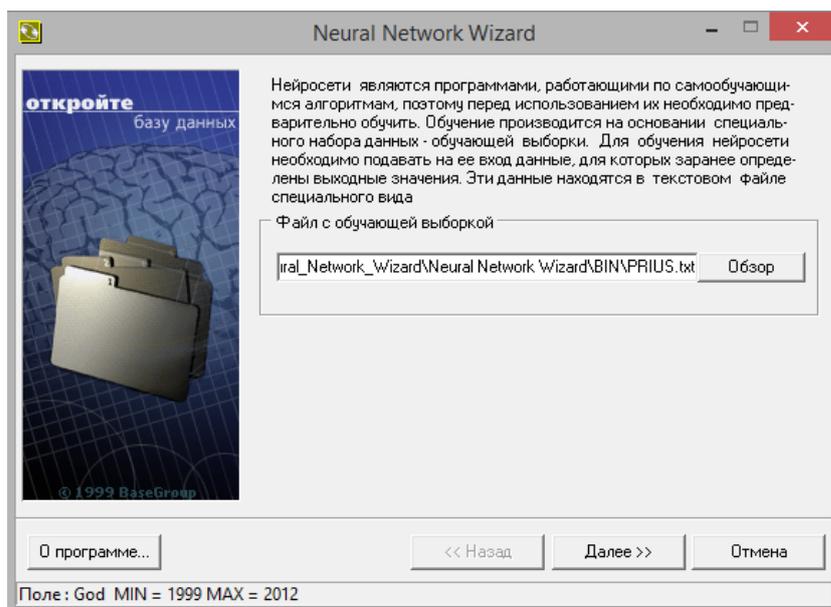


Рисунок 13 - Программа Neural Network Wizard

Далее, нужно указать, что поле `price` — целевое. То есть, нейронная сеть будет пытаться определить каким образом значения полей `God`, `V`, `P` и `S` влияют на поле `price` (рис. 14).

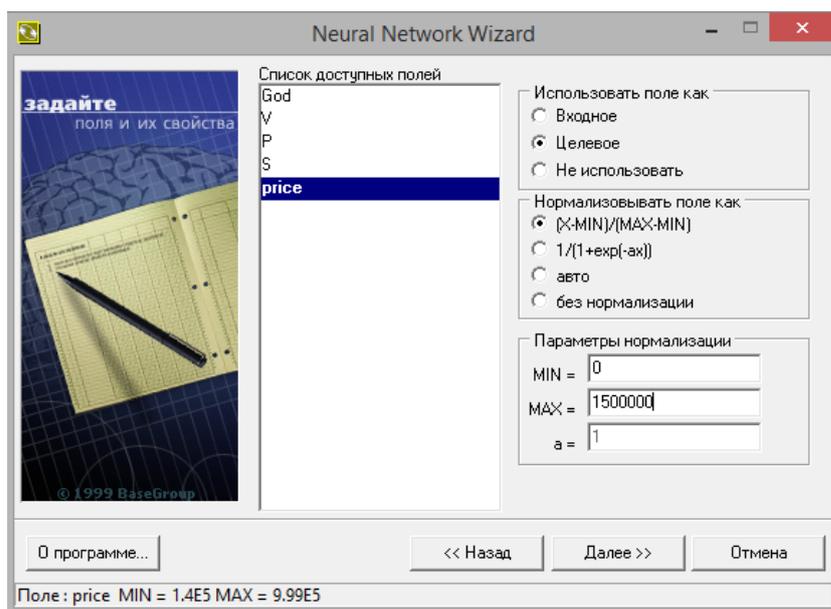


Рисунок 14 – Окно программы со списком полей

Определим конфигурацию нейронной сети. Зададим количество скрытых слоев – 1. Количество элементов в первом слое – 4. Параметр сигмоиды зададим 0,5 (рис. 15).

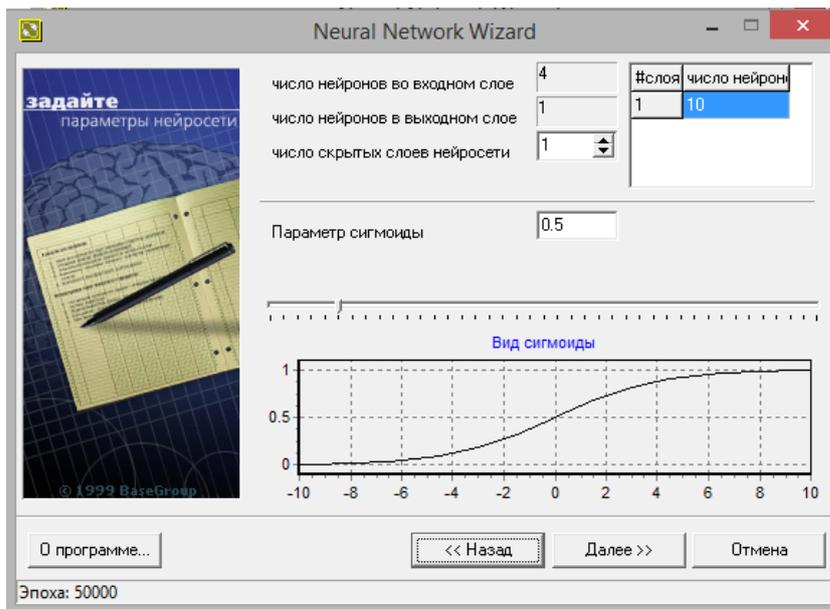


Рисунок 15 - Окно конфигурации нейронной сети

Определим параметры обучения — остановить обучение по прошествии 50000 эпох (рис. 16).

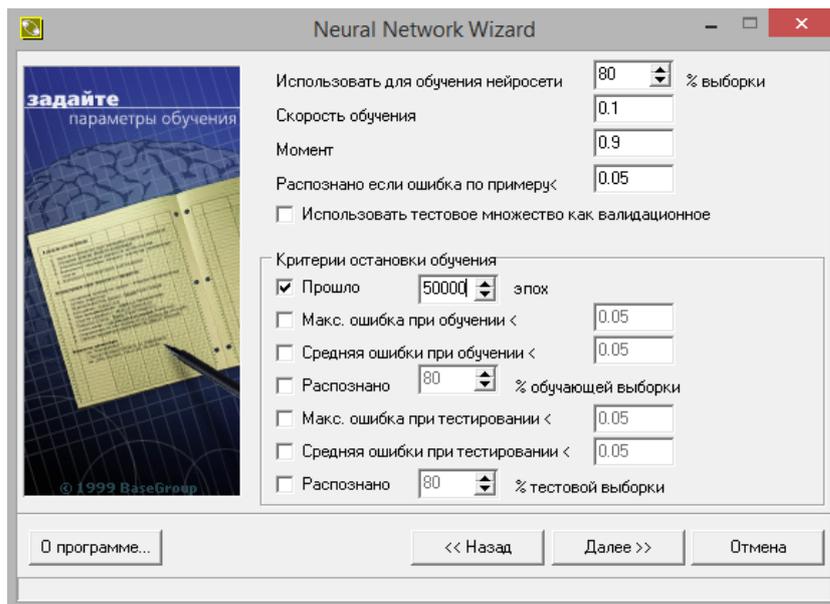


Рисунок 16 - Окно параметров обучения

Проверим правильность конфигурации сети и параметров обучения (рис. 17).

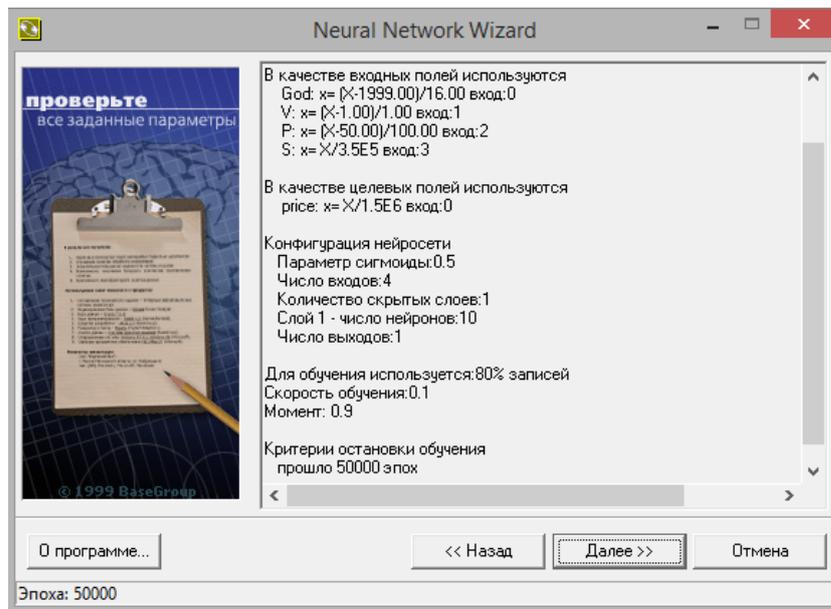


Рисунок 17 - Проверка конфигурации сети и параметров обучения

Запустим систему на обучение (рис. 18).

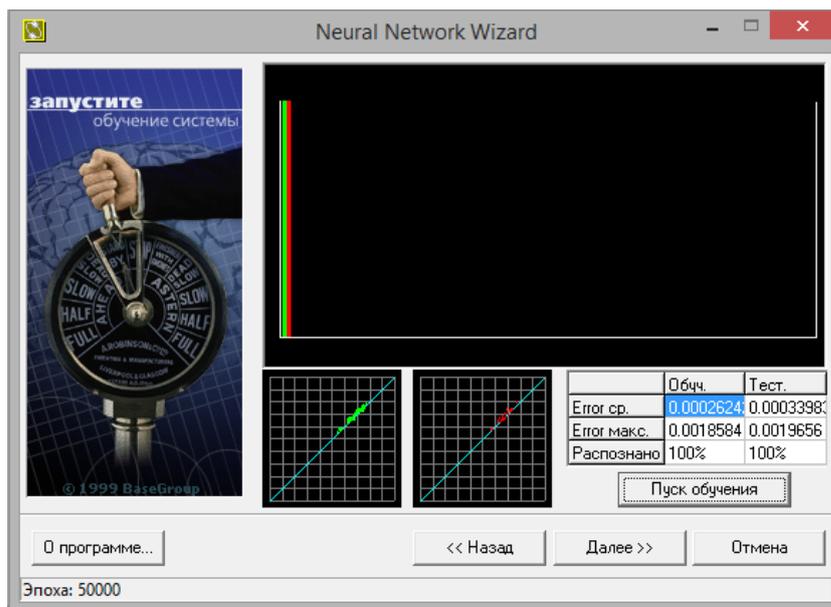


Рисунок 18 - Окно запуска обучения

По окончании обучения протестируем полученную модель. Обозначения входных параметров: год – God, объем двигателя – V, мощность – P, пробег – S. Введем начальные параметры и рассчитаем стоимость автомобиля – price (рис. 19).

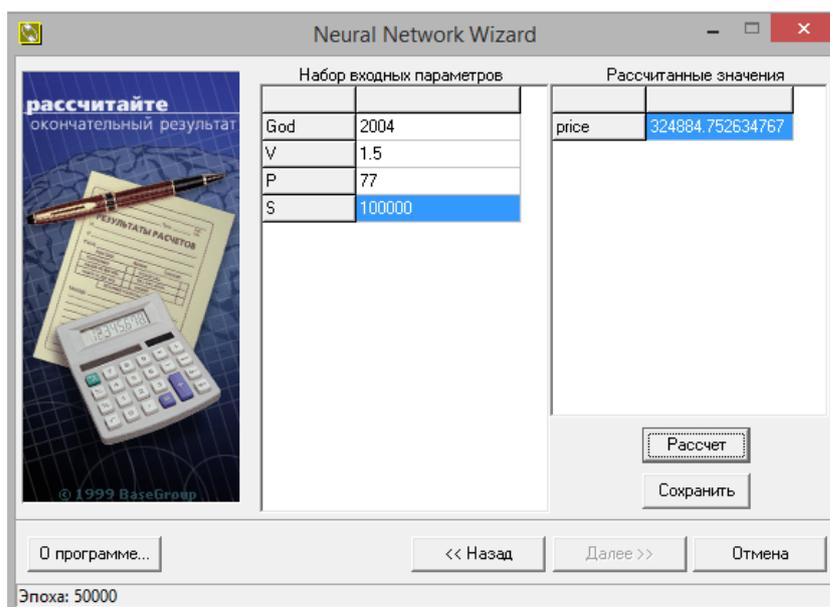


Рисунок 19 - Окно тестирования нейронной сети

Теперь сравним методы регрессионного анализа и нейронных сетей. В таблицу внесем стоимости автомобилей по одинаковым параметрам в столбцы «Регрессионный анализ» и «Нейронная сеть». Покажем разницу стоимости автомобилей.

Таблица 1 – Сравнение данных экспериментов

Регрессионный анализ	Нейронная сеть	Абсолютная разница	Относительная разница, %
330300	320900	9400	3
476100	490700	14600	3
429900	447900	18000	4
223800	216100	7700	3
720500	706300	14200	2

По данным табл.1 видно, что оба метода отлично подходят для установления стоимости автомобиля. Полученные регрессионную модель и нейронную сеть можно использовать для покупки и продажи авто модели Toyota Prius. Проведенное исследование может быть использовано в при обучении методам интеллектуального анализа студентов различных направлений [33-35].

Библиографический список

1. Недошивин С.В. Линейный множественный регрессионный анализ в статистическом машинном эксперименте // Известия Тульского государственного университета. Технические науки. 2014. №7. С. 38-52
2. Коротченко А.С. Корреляционно-регрессионный анализ влияния инвестиционной активности на рост регионального валового продукта //

- Вестник Балтийского федерального университета им. И. Канта. 2010. №3. С. 108-115.
3. Суслов С.А. Кластерный анализ: сущность, преимущества и недостатки // Вестник НГИЭИ. 2010. Т. 1. № 1. С. 51-57.
 4. Демин Л.А., Суслов С.А., Экономико-математические методы моделирования экономических систем: учебно-методическое пособие. Княгинино: Нижегородский государственный инженерно экономический институт, 2006. 56 с.
 5. Остапенко Р.И. Особенности моделирования латентных изменений с помощью AMOS SPSS // Перспективы науки и образования. 2014. № 1 (7). С. 89-95.
 6. Остапенко Р.И. Особенности анализа лонгитюдных данных в психолого-педагогических исследованиях с помощью AMOS SPSS // Современные научные исследования и инновации. 2014. № 4 (36). С. 74
 7. Серов А.А. Скоринг экспериментальных данных с применением прогнозных моделей в среде пакета SPSS // Традиции и новации в профессиональной подготовке и деятельности педагога. Материалы Всероссийской научно-практической конференции преподавателей и студентов. Ответственный редактор В.П.Анисимов. Тверь: Тверской государственный университет, 2013. С. 53-54.
 8. Пивенко К.А., Баженов Р.И. Построение регрессионной модели в среде Gretl на примере рынка поддержанных автомобилей г. Биробиджана и г. Хабаровска // Экономика и менеджмент инновационных технологий. 2015. № 4-1 (43). С. 72-80.
 9. Пронина О.Ю., Баженов Р.И. Исследование методов регрессионного анализа программной среды EVIEWS // Nauka-Rastudent.ru. 2015. № 1 (13). С. 45.
 10. Лагунова А.А., Баженов Р.И. Разработка в среде gretl регрессионной модели рынка вторичного жилья г. Биробиджана // Nauka-Rastudent.ru. 2015. № 1 (13). С. 40.
 11. Муллинов Д.О., Баженов Р.И. Разработка в среде EVIEWS регрессионной модели рынка гаражных помещений г. Биробиджана // Nauka-Rastudent.ru. 2015. № 1 (13). С. 43.
 12. Эм А.А., Баженов Р.И. Разработка в среде EVIEWS регрессионной модели реализации продукции компании по производству резинометаллических изделий // Экономика и менеджмент инновационных технологий. 2015. № 4-2 (43). С. 58-709.
 13. Широкова Н.А., Баженов Р.И. Применение корреляционного анализа для исследования данных спортивных показателей студентов в среде SPSS // Nauka-Rastudent.ru. 2015. № 6 (18). С. 25.
 14. Муллинов Д.О., Винокуров А.С., Баженов Р.И. Разработка в среде SPSS регрессионной модели рынка автомобилей // Nauka-Rastudent.ru. 2015. № 6 (18). С. 24.
 15. Филатова Т.В. Применение нейронных сетей для аппроксимации данных

- // Вестник Томского государственного университета. 2004. №284. С. 121-125.
16. Чернявский А.В., Спицын В.Г. Применение эволюционирующей нейронной сети для улучшения качества изображений // Известия Томского политехнического университета. 2006. №7. С. 26-31.
 17. Осипов В.Ю. Рекуррентная нейронная сеть с двумя сигнальными системами // Информационно-управляющие системы. 2013. № 4 (65). С. 8-15.
 18. Манжула В.Г., Федяшов Д.С. Нейронные сети Кохонена и нечеткие нейронные сети в интеллектуальном анализе данных // Фундаментальные исследования. 2011. № 4. С. 108-114.
 19. Колбасина И.В., Старовойт Е.Д., Бежитский С.С. О качестве восстановления образов искусственной нейронной сетью Хопфилда // Решетневские чтения. 2013. Т. 2. № 17. С. 26-27.
 20. Седов В.А., Седова Н.А. Моделирование расхождения судов в зоне чрезмерного сближения нейронными сетями // Научные проблемы транспорта Сибири и Дальнего Востока. 2014. № 3. С. 102-105.
 21. Седов В.А., Седова Н.А. Гибридная система управления судном // Научно-технический вестник Поволжья. 2013. № 2. С. 204-207.
 22. Варламов А.Д., Шарапов Р.В. Использование нейронных сетей в задачах мониторинга экзогенных процессов дистанционными методами // Геоинформатика. 2014. № 4. С. 62-68.
 23. Варламов А.Д., Шарапов Р.В. Использование нейронных сетей для оценки уровня заболоченности территории на основе данных дистанционного зондирования // Нейрокомпьютеры: разработка, применение. 2015. № 3. С. 29-33.
 24. Шумков Е.А., Чистик И.К. Использование генетических алгоритмов для обучения нейронных сетей // Политематический сетевой электронный научный журнал КубГАУ. 2013. №7. URL: <http://ej.1gb.ru/2013/07/pdf/78.pdf>
 25. Гимаров В.А., Дли М.И., Битюцкий С.Я. Нейро-нечеткий метод классификации объектов с разнотипными признаками // Системы управления и информационные технологии. 2004. Т. 16. № 4. С. 13-18.
 26. Ступников А.В., Баженов Р.И. Прогнозирование цены легковых автомобилей с помощью нейронных сетей в среде Neural Network Wizard // Современная техника и технологии. 2015. № 7 (47). С. 3-10.
 27. Николаев С.В., Баженов Р.И. Распознавание образов с помощью нейронных сетей в среде MATLABR2009B // Nauka-Rastudent.ru. 2015. № 1 (13). С. 44.
 28. Николаев С.В., Пронина О.Ю., Баженов Р.И. Исследование методов интеллектуального анализа для формирования краткосрочного прогноза в программной среде Statistica // Экономика и менеджмент инновационных технологий. 2015. № 7 (46). С. 89-102.
 29. Баженов Р.И. Информационная безопасность и защита информации:

- практикум. Биробиджан: Изд-во ГОУВПО «ДВГСГА», 2011. 140 с.
30. Kanyongo G. Y., Certo J., Launcelot B. I. Using Regression Analysis to Establish the Relationship between Home Environment and Reading Achievement: A Case of Zimbabwe // International Education Journal. 2006. Т. 7. №. 4. С. 632-641.
31. Fechete F., Nedelcu A. Analysis of the economic performance of a organization using multiple regression // Scientific Research & Education in the Air Force - AFASES . 2014. Т. 2, С. 411-416.
32. Dede G., Sazli M. H. Speech recognition with artificial neural networks // Digital Signal Processing. 2010. Т. 20. №. 3. С. 763-768.
33. Баженов Р.И. О применении балльно-рейтинговой системы для оценивания курсовых работ по дисциплине «Интеллектуальные системы и технологии» // Приволжский научный вестник. 2014. № 5 (33). С. 135-138.
34. Баженов Р.И. Проектирование методики обучения дисциплины «Интеллектуальные системы и технологии» // Современные научные исследования и инновации. 2014. № 5-2 (37). С. 48.
35. Баженов Р.И. Проектирование методики обучения дисциплины «Информационные технологии в менеджменте» // Современная педагогика. 2014. № 8 (21). С. 24-31.