4 ЛАБОРАТОРНАЯ РАБОТА «АНАЛИТИЧЕСКАЯ ПЛАТФОРМА DEDUCTOR. ЗАДАЧА ПРОГНОЗИРОВАНИЯ»

4.1 Теоретические сведения

Прогнозирование результата на определённое время вперёд, основываясь на данных за прошедшее время, – задача, встречающаяся довольно часто. К примеру, перед большинством торговых фирм стоит задача оптимизации складских запасов, для решения которой требуется знать, что и сколько должно быть продано через неделю и т.п., задача предсказания стоимости акций какого-нибудь предприятия через день и т.д. и другие подобные вопросы. Deductor Studio предлагает для этого инструмент «Прогнозирование».

Прогнозирование (регрессия) – установление зависимости непрерывных выходных переменных от входных. К этому же типу задач относится прогнозирование временного ряда на основе исторических данных.

Регрессия используется для установления зависимостей в факторах. Например, в задаче прогнозирования зависимой величиной является объемы продаж, а факторами, влияющими на эту величину, могут быть предыдущие объемы продаж, изменение курса валют, активность конкурентов и т.д. Или, например, при кредитовании физических лиц вероятность возврата кредита зависит от личных характеристик человека, сферы его деятельности, наличия имущества.

Прогнозирование появляется в списке Мастера обработки только после построения какой-либо модели прогноза: нейросети, линейной регрессии и т.д. Прогнозировать на несколько шагов вперед имеет смысл только временной ряд (к примеру, если есть данные по недельным суммам продаж за определенный период, можно спрогнозировать сумму продаж на две недели вперед). Поскольку при построении модели прогноза необходимо учитывать много факторов (зависимость результата от данных день, два, три, четыре назад), то методика имеет свои особенности.

4.2 Пример решения задачи прогнозирования

У аналитика имеются данные о помесячном количестве проданного товара за несколько лет. Ему необходимо, основываясь на этих данных, определить, какое количество товара будет продано через месяц и через два месяца. Исходные данные по продажам находятся в файле Trade.txt. Выполним импорт данных из файла, не забыв указать в Мастере, чтобы в качестве разделителя дробной и целой частей была точка, а не запятая (рисунок 1).

Таблица | Диаграмма

Дата (Год + Месяц)	Количество
2000-M01	462523,419
2000-M02	633208,196
2000-M03	660159,299
2000-M04	617455,3417
2000-M05	597354,4794
2000-M06	793517,4512
2000-M07	1015944,2862
2000-M08	1148052,2523
2000-M09	1156623,1715
2000-M10	1255021,9423
2000-M11	1410114,5606
2000-M12	1357230,3388
2001-M01	1003317,7317
2001-M02	1097048,6263
2001-M03	1498977,3427
2001-M04	1507696,4482
2001-M05	1520761,5589
2001-M06	1602674,5245
2001-M07	1685899,1625
2001-M08	1899255,945
2001-M09	1716804,1633
2001-M10	2069772,3982
2001-M11	2016227,4267
2001-M12	1817580,4566
2002-M01	1493788,5092
2002-M02	1449402,666
2002-M03	1734310,2126
2002-M04	2068457,1962
2002-M05	1899233,8657
2002-M06	2083559,1009
2002-M07	2157181,3292
2002-M08	2161689,8556
2002-M09	2315556,804
2002-M10	2367611,5581
2002-M11	2110853,3582
2002-M12	2297331,2201
2003-M01	1496683,6138
2003-M02	1629978,7451

Рисунок 1 – Фрагмент исходных данных

После импорта данных нужно воспользоваться диаграммой для их просмотра (установить флажок Диаграмма). Диаграмма данных представлена на рисунке 2. Если данные содержат аномальные значения (выбросы) и шумы, перед прогнозированием необходимо удалить аномальные значения и сгладить данные.

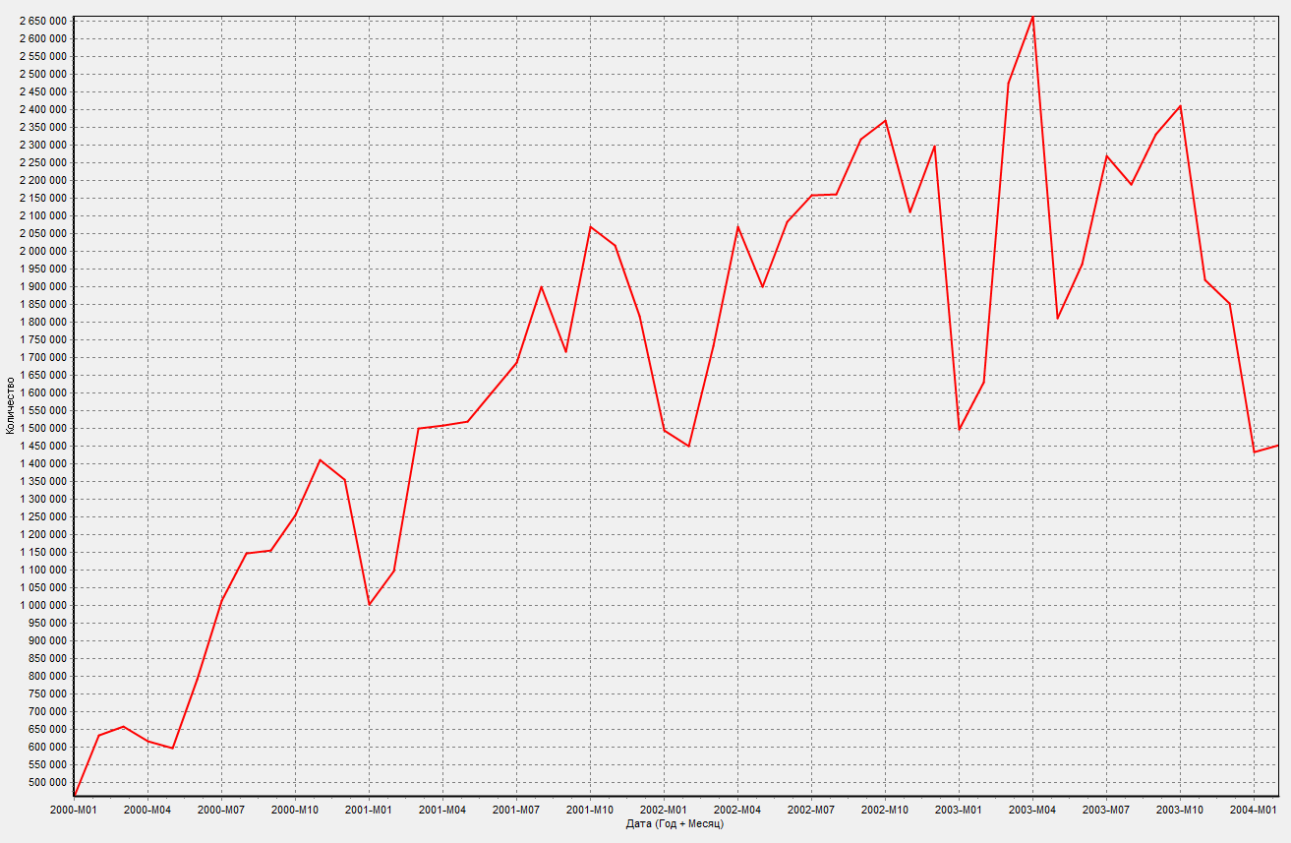

Рисунок 2 – Диаграмма

Очистить данные можно при помощи обработчиков «Оценка качества данных», «Заполнение пропусков», «Редактирование выбросов», «Спектральная обработка».

Подробное описание обработчиков можно найти, перейдя по ссылке: https://basegroup.ru/deductor/function/algorithm?algorytm_group%5B%5D=82.

Запустим обработчик «Спектральная обработка» в Мастере обработки и установить параметры, как указано на рисунке 3. Спектральная обработка предназначена для очистки от шумовой составляющей и сглаживания рядов данных. Сглаживание необходимо в том случае, когда ряд данных оказывается неравномерным, содержит большое количество мелких структур, препятствующих исследованию более значительных объектов и закономерностей. Спектральная обработка наиболее часто применяется для предварительной подготовки данных в задачах прогнозирования, т.к. позволяет сделать временной ряд более гладким, благодаря чему полученная прогнозная модель обладает высокими обобщающими качествами.

Результат можно просмотреть в виде таблицы и диаграммы (рисунок 4). Как видно из рисунка данные сгладились.

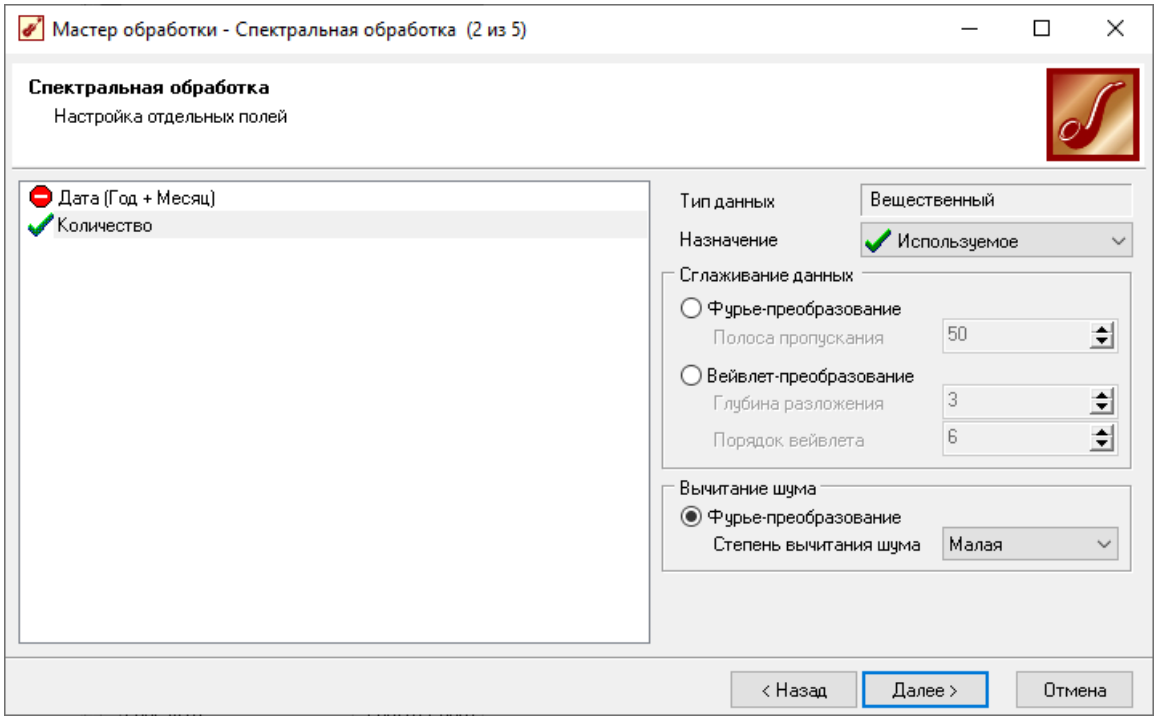

Рисунок 3 – Настройка параметров спектральной обработки

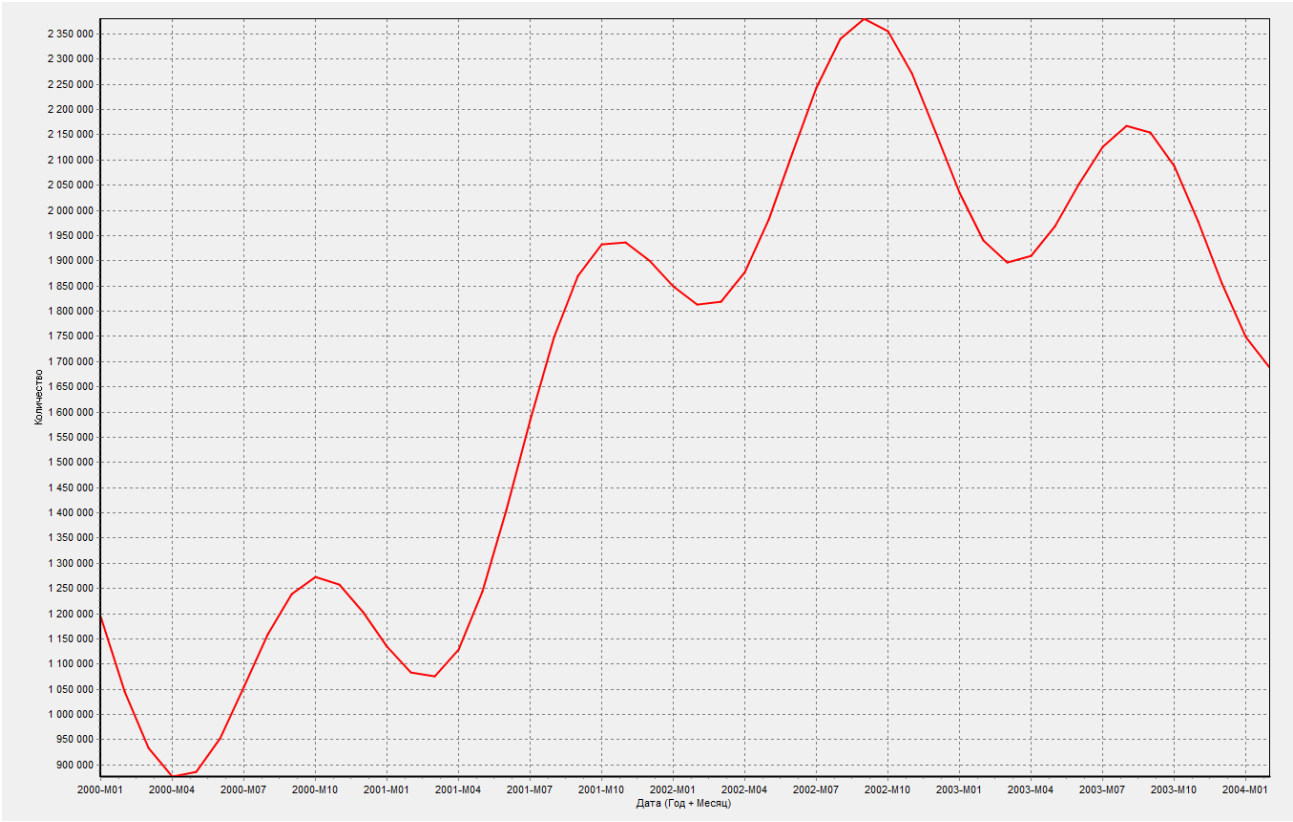

Рисунок 4 – Диаграмма после спектральной обработки

При решении задачи прогнозирования используются данные, в которых есть входные столбцы – факторы и есть выходные столбцы – результат. В данном случае набор данных (Trade.txt) содержит столбец с выходными значениями. Строить прогноз на будущее будем, основываясь на данных прошлых периодов, т. е. предполагая, что количество продаж на следующий месяц зависит от количества продаж за предыдущие месяцы. Это значит, что входными факторами для модели могут быть продажи за текущий месяц, продажи за месяц ранее и т.д., а результатом должны быть продажи за следующий месяц, т. е. здесь явно необходимо трансформировать данные к скользящему окну.

В случае наличия входных и выходных столбцов в наборе данных для решения задачи прогнозирования можно изучить пример, рассмотренный в ранее используемом пособии со станицы 107 по страницу 109. В пособии приведён пример построения нейросетевой модели прогнозирования стоимости недвижимости.

Обработчик «Скользящее окно» преобразует последовательность значений ряда в таблицу, где соседние записи представлены как соседние поля данных (окно – поскольку выделяется только некоторый непрерывный участок данных, скользящее – поскольку это окно «перемещается» по всему набору).

Потребность в такой таблице часто возникает при построении моделей, анализе и прогнозировании временных рядов, когда требуется подавать на вход модели значения нескольких смежных отсчётов из исходного набора данных (https://basegroup.ru/deductor/function/algorithm/sliding-window).

Запустим Мастер обработки и выберем Скользящее окно, затем укажем необходимые параметры (рисунок 5). Прогноз будем строить на месяц, поэтому требуется выбрать глубину погружения 12, назначив поле «Количество» используемым. Тогда данные трансформируются к скользящему окну так, что аналитику будут доступны все нужные факторы для построения прогноза. Например, если учитывать сезонность, то можно в качестве входных факторов использовать «Количество – 12», «Количество – 11» – данные по количеству 12 и 11 месяцев назад (относительно прогнозируемого месяца), а также «Количество – 2» и «Количество – 1» – данные за 2 предыдущих месяца. В качестве выходного поля укажем столбец «Количество».

Перейдём непосредственно к самому построению модели прогноза. Откроем Мастер обработки и выберем в нём узел Нейросеть. На втором шаге установим в качестве входных поля «Количество – 12», «Количество – 11», «Количество – 2» и «Количество – 1», а в качестве выходного – «Количество». Остальные поля сделаем информационными (рисунок 6).

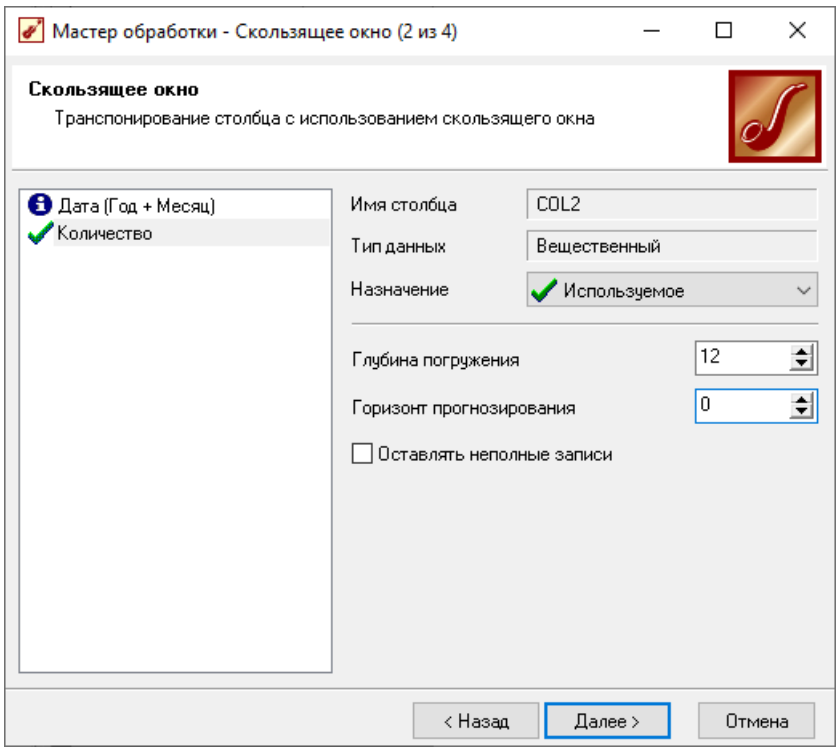

Рисунок 5 – Параметры обработки «Скользящее окно»

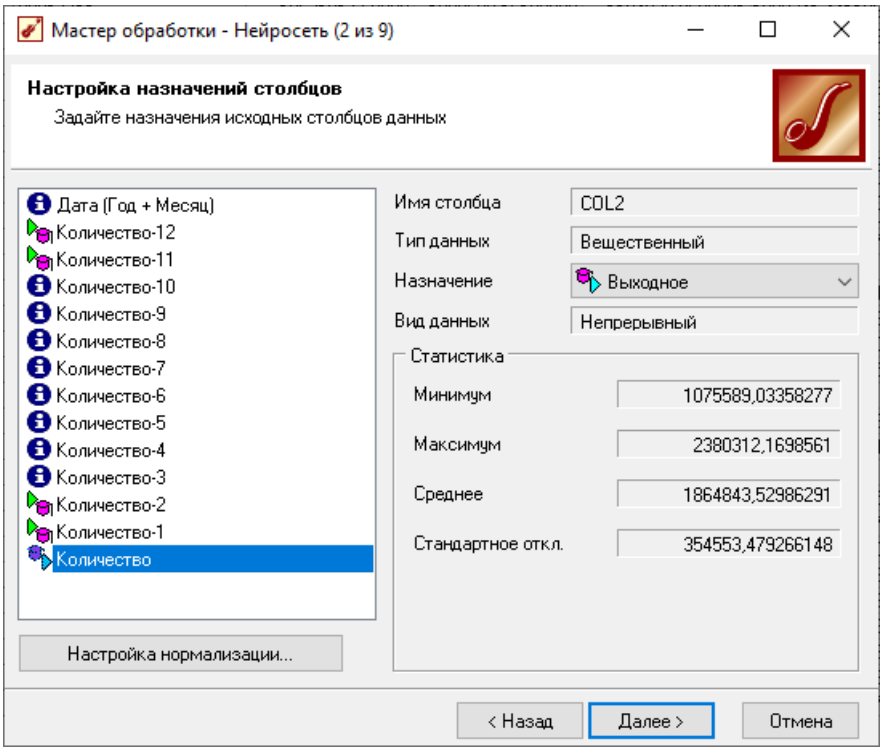

Рисунок 6 – Настройка назначений столбцов

Далее укажем по шагам все необходимые параметры обучения нейронной сети (см. лабораторную работу 4). После построения модели для просмотра качества обучения результаты представлены в визуализаторах «Граф нейросети» (рисунок 7), «Диаграмма рассеяния» (рисунок 8), «Что-если» (рисунок 9).

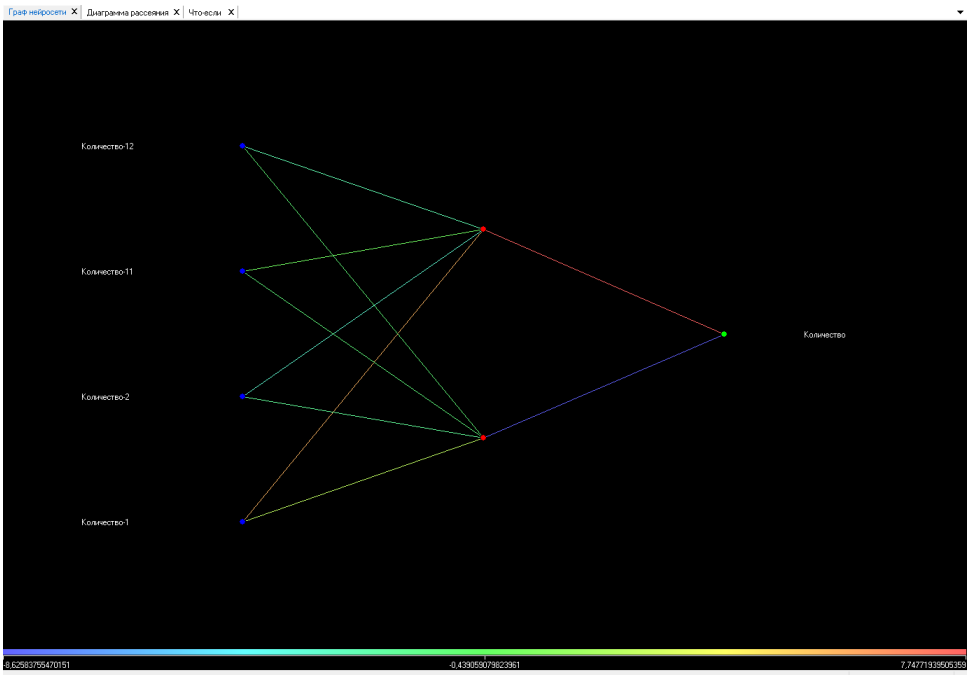

Рисунок 7 – Визуализатор «Граф нейросети»

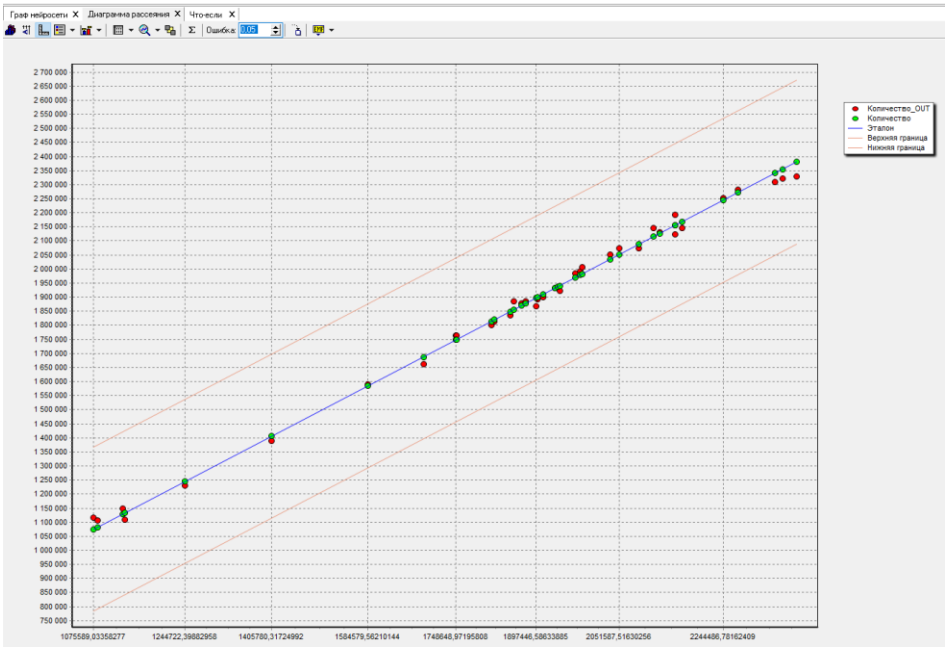

Рисунок 8 – Визуализатор «Диаграмма рассеяния»

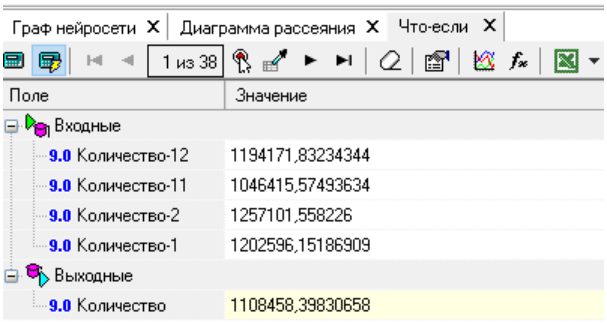

Рисунок 9 – Визуализатор «Что-если»

Диаграмма рассеяния более наглядно показывает качество обучения. На этой диаграмме отображается отклонение прогнозного значения величины от её истинного значения. Диаграмма рассеяния служит для наглядной оценки качества построенной модели с помощью результатов сравнения непрерывных значений выходного поля и непрерывных значений того же поля, но рассчитанных моделью. На диаграмме рассеяния отображаются выходные значения каждого для каждого из примеров обучающей выборки, координаты которых по оси X – это значение выхода на обучающей выборке (эталон), а по оси Y – значение выхода, рассчитанное обученной моделью на том же примере. Прямая диагональная линия представляет собой ориентир (линию идеальных значений). Чем ближе точка к этой линии, тем меньше ошибка модели.

Нейросеть обучена, осталось получить требуемый прогноз. Открываем Мастер обработки и выбираем появившийся обработчик «Прогнозирование».

На втором шаге Мастера предлагается настроить связи столбцов для прогнозирования временного ряда: откуда брать данные для столбца при очередном шаге прогноза. Мастер сам верно настроил все переходы, поэтому остаётся только указать горизонт прогноза (на сколько вперёд будем прогнозировать) равный трём, а также для наглядности следует добавить к прогнозу исходные данные, установив в Мастере соответствующий флажок (рисунок 10).

После этого необходимо в качестве визуализатора выбрать «Диаграмму прогноза», которая появляется только после прогнозирования временного ряда. В Мастере настройки столбцов диаграммы прогноза надо указать в качестве отображаемого столбец «Количество», а в качестве подписей по оси X указать столбец «Шаг прогноза».

Теперь аналитик может дать ответ на вопрос, какое количество товаров будет продано в следующем месяце и даже два месяца спустя (выделено жёлтым цветом на рисунке 11).

Данный пример показал, как с помощью Deductor Studio прогнозировать временной ряд.

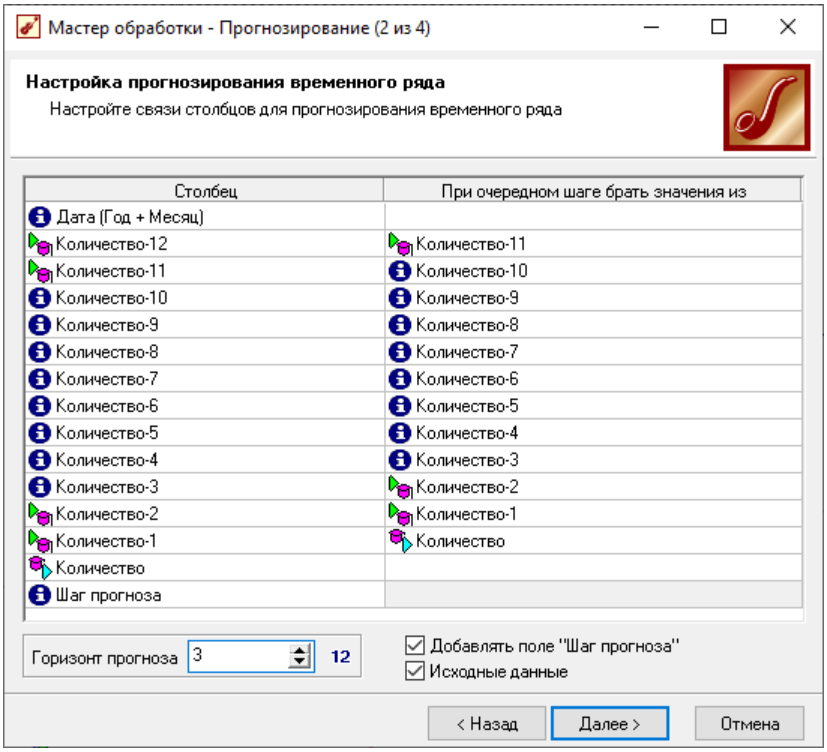

Рисунок 10 – Настройка прогнозирования временного ряда

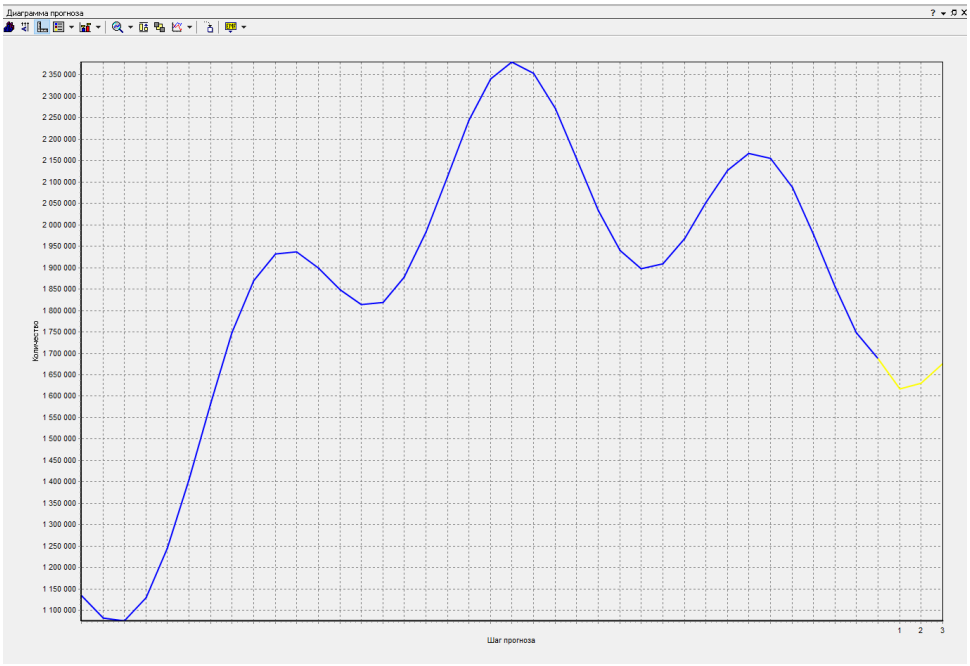

Рисунок 11 – Диаграмма прогноза

4.3 Задание и рекомендации

1. Изучить материал, представленный в лабораторной работе 4 и в пособии ранее указанном.
2. Для решения задачи прогнозирования найти набор данных (датасет) в сети Интернет. Набор данных может быть в формате *.csv. Для использования в Deductor нужно изменить расширение на *.txt. Можно использовать часть набора данных. В СДО есть примеры наборов данных.
3. Решить задачу прогнозирования в Deductor Studio Academic по исходным данным (набору данных) при помощи нейронной сети, получить прогноз. С помощью различных визуализаторов проанализировать полученные результаты, сделать выводы.
4. Подготовить отчёт и сдать преподавателю в электронной форме. Отчёт должен содержать краткое описание выполняемых действий, скриншоты, выводы. Наличие выводов является обязательным требованием.

4.4 Вопросы для защиты лабораторной работы

1. В чём суть задачи прогнозирования (регрессии)?
2. Для чего используется очистка данных? Какие обработчики в АП Deductor позволяют очистить данные?
3. Для чего используется обработчик «Скользящее окно»? В каких случаях применяется?
4. Для чего используются визуализаторы «Граф нейросети», «Диаграмма рассеяния» и «Что-если»?
5. Для чего используются обработчик «Прогнозирование»?