Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра системного анализа

Лабораторная работа

Прогнозирование спроса

Студент 415 группы
К. Ю. Егоров

Научный руководитель
к.ф.-м.н., доцент П. А. Точилин

Москва, 2020

Содержание

1 Цель работы **3**

2 Постановка задачи **3**

3 Общие соображения **4**

4 Модель SARIMA **5**

5 Линейная регрессия **6**

6 Заключение **8**

1 Цель работы

Основной целью выполнения данного задания является получение базовых навыков обработки и анализа больших объемов информации, используя современные прикладное программное обеспечение. В рамках лабораторной работы ставится задача разработать адекватный алгоритм прогнозирования спроса на представленную группу товаров по данным покупкам за предыдущий период. Выбор методов и программных средств прогнозирования оставлены на выбор автора.

2 Постановка задачи

Для исследования нам даны данные о покупках товаров одной категории за 2017–начало 2018 годы. Данные хранятся в отношении с заголовком:

(Код товара, Дата покупки, Количество),

причем в одну дату может быть быть совершено более одной покупки.

Мы будем исследовать совокупный спрос на все товары данной группы. Будем пользоваться СУБД *MySQL* для быстрого доступа к данным и средствами модулей *Python3* для их обработки, построения моделей и построения графиков. Визуализация совокупных покупок всех товаров, представленных в 2017 и 2018 годах, доступные для разработки алгоритма представлены на рисунке Рис. 1.

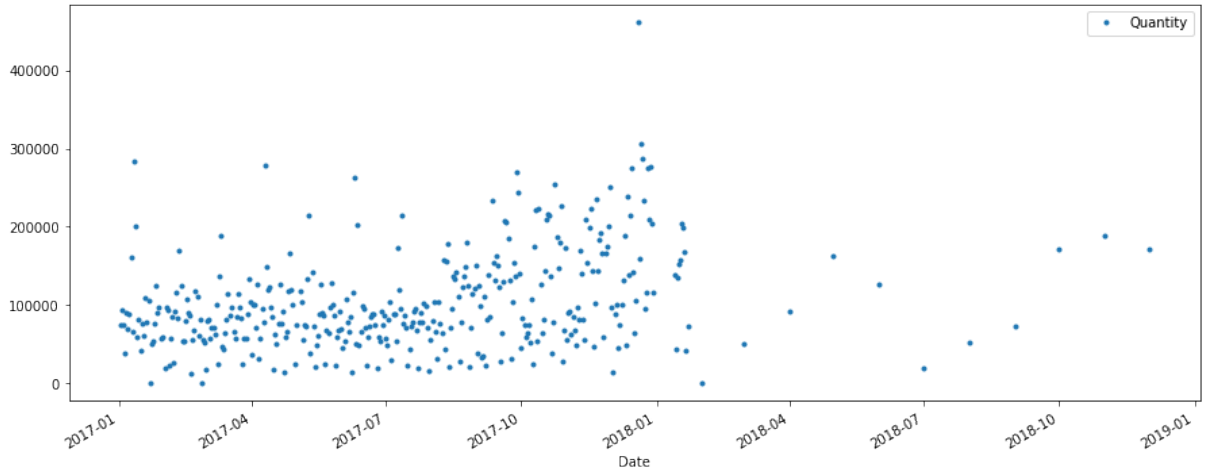

Рис. 1: Совокупные покупки по дням.

Из рисунка видно, что данные за 2018 год не полные, поэтому мы сформулируем задачу следующим образом: *Будем предсказывать дневной спрос в декабре 2017 года по считающемуся известным спросу за предыдущие месяцы.*

3 Общие соображения

Посмотрим на график Рис. 1 подробнее. Видно, что наблюдается восходящий тренд: каждый следующий месяц в среднем покупают большее количество товара, чем в предыдущий. Это может быть связано как с общим трендом на увеличение покупок: товар становится более востребован, так и с годичной сезонностью: так, например, известно, что в декабре существенно возрастает количество покупок большего числа товаров, так как это время перед новогодними (рождественскими) праздниками. Имея данные только за один год и не зная, какая именно группа товаров исследуется, мы не можем конкретно полагать, о каком из двух случаев идет речь. Чего нельзя отрицать, так это восходящего тренда в рамках года. Для предсказания на конец года нам должно быть достаточно этого знания.

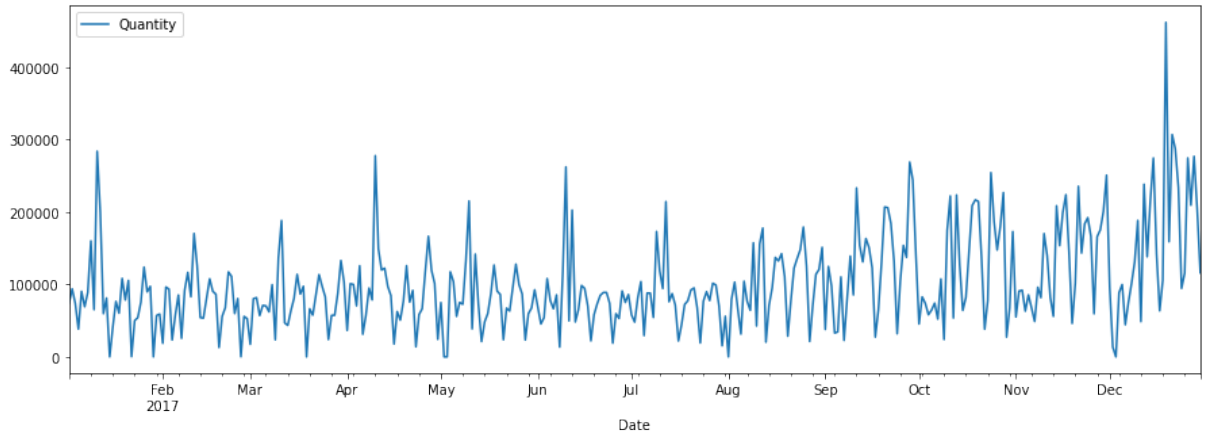

Рис. 2: Данные за 2017 год.

Модуль `statmodels` предоставляет возможность визуально разделить временной ряд на три компоненты: тренд, сезонность и шум — для заданного периода. Проведя серию экспериментов с разной величиной периода, было выбрано значение с визуально наименьшей выборочной дисперсией шума — 28 дней. На рисунке Рис. 3 можно посмотреть на соответствующую деком-

позицию. Также напрашивается вывод, что месячная периодичность более выражена, чем любая другая, например, недельная.

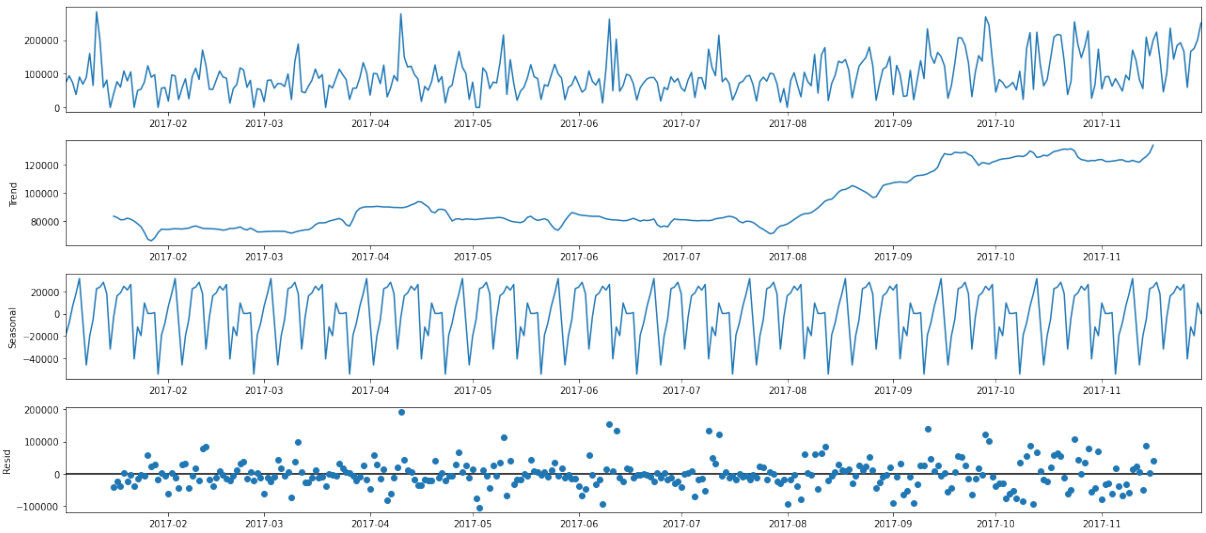

Рис. 3: Декомпозиция временного ряда на компоненты: тренд, сезонность и шум — для периода в 28 дней.

4 Модель SARIMA

Одной из самых популярных моделей для предсказания временного ряда на сегодняшний день является модель *SARIMA*. Она также реализована в модуле `statmodels` и представляет собой учитывающее сезонность улучшение модели авторегрессионного интегрированного скользящего среднего. Эта модель предполагает набор параметров для каждой из частей модели: авторегрессионной, интегрированной, скользящего среднего, период сезонности, а также дополнительные термины сезонности. Сначала необходимо провести перебор по этим параметрам, чтобы добиться наиболее стационарного ряда: в таком случае мы сможем предполагать, что будущие характеристики ряда не будут отличаться от нынешних.

После перебора параметров модели мы выбрали те, при которых достигает минимума *информационный критерий Акаике* для данной модели. Этот критерий не только вознаграждает за качество приближения, но и штрафует за излишнее использование параметров модели. Таким образом мы пытаемся избежать переобучения модели. Характеристики получившегося стационарного ряда можно посмотреть на рисунке Рис. 4.

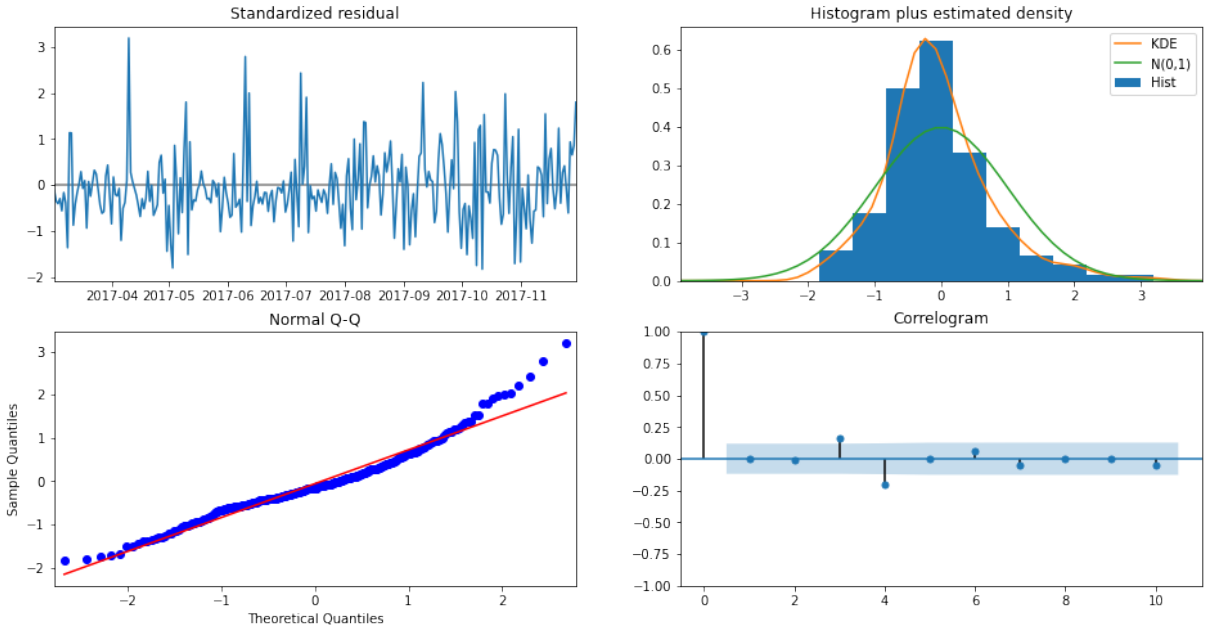

Рис. 4: Характеристики модели SARIMA с наиболее подходящими параметрами.

На графике Рис. 5 можно визуально оценить точность приближения второго полугодия построенной модели SARIMA, обученной на первом полугодии. На графике Рис. 6 видно предсказание на декабрь и дальше. Средняя ошибка за декабрь составляет 72 340,11.

5 Линейная регрессия

Предыдущий подход часто называют эконометрическим: он учитывает такие присущие товарам свойства как тренд и сезонность, используя статистические методы. Сейчас попробуем дать предсказание, используя базовый численный метод — линейную регрессию.

Напомним, что модель линейной регрессии представляет собой один нейрон с n входами (признаками) $x \in \mathbb{R}^n$, некоторой активационной функцией и одним выходом:

$$\hat{y} = f(w_0 + \langle x, w \rangle),$$

где $w_0 \in \mathbb{R}$, $w \in \mathbb{R}^n$ — параметры модели (веса), подбираемые методом градиентного спуска, минимизирующего некоторый функционал цены.

Мы возьмем в качестве признаком данные за 28 дней перед предсказанием и день недели. (День недели кодируется в виде семи признаков, принимаю-

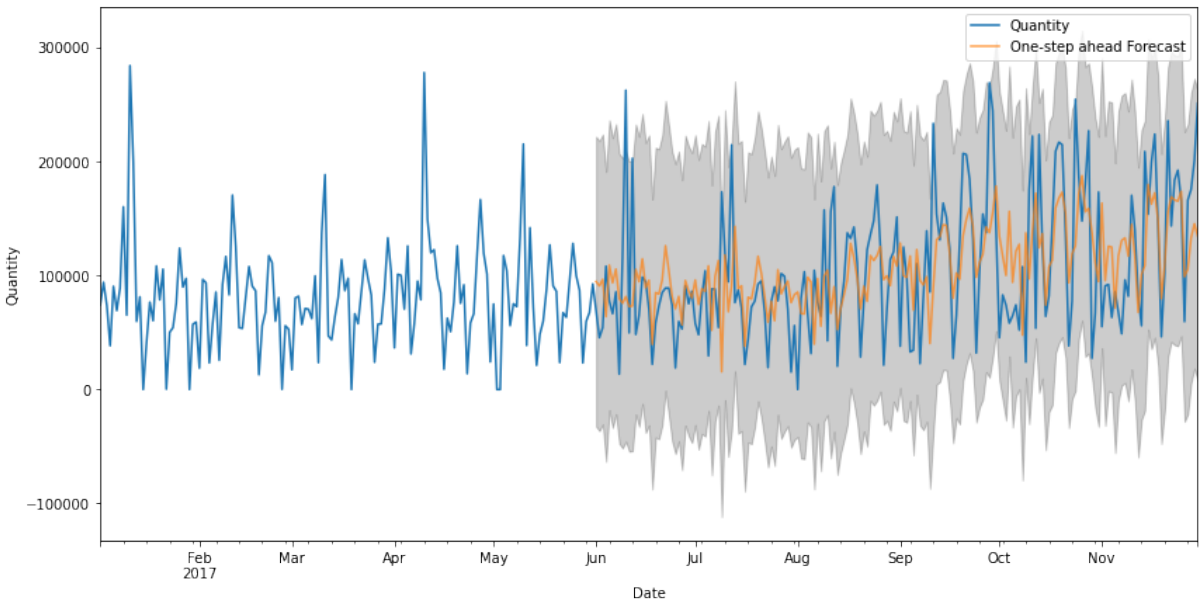

Рис. 5: Предсказание второго полугодия по первому. Здесь синий — истинное значение, желтый — предсказание, серым закрашен доверительный интервал предсказания с уровнем доверия 0,05.

щих единичное значение, если предсказывается данный день недели, иначе ноль. Такой способ кодирования позволяет задать различные веса дня каждого дня недели и не дает какому-либо дню недели априорного преимущества перед другими.) Активационную функцию возьмем линейную $f(z) = z$. Будем минимизировать средне-квадратичную ошибку $J = \frac{1}{N}\sum_{k=1}^{N}(y^k - \hat{y}^k)^2$, где N — размер обучающей выборки. Обучение произведем на первых 11 месяцах.

Наша надежда в том, что, имея данные о 28 днях и о дне недели, модель сможет распознать оба вида сезонности: месячную и недельную, при этом тренд должен быть распознан за счет большого количества весов, больших единицы. В случае, когда мы хотим предсказывать данные на больший срок, чем один день, мы можем использовать значение, сгенерированное на предыдущих итерациях как признак: таким образом работают наибольшее число реккурентных моделей. Использование всего одного нейрона должно уберечь нас от переобучения модели.

На графике Рис. 7 можно посмотреть качество приближения на обучающей выборке. На графике Рис. 8 можно посмотреть качество предсказания на проверочной выборке (декабрь). Средняя ошибка в декабре составила 64 334,37.

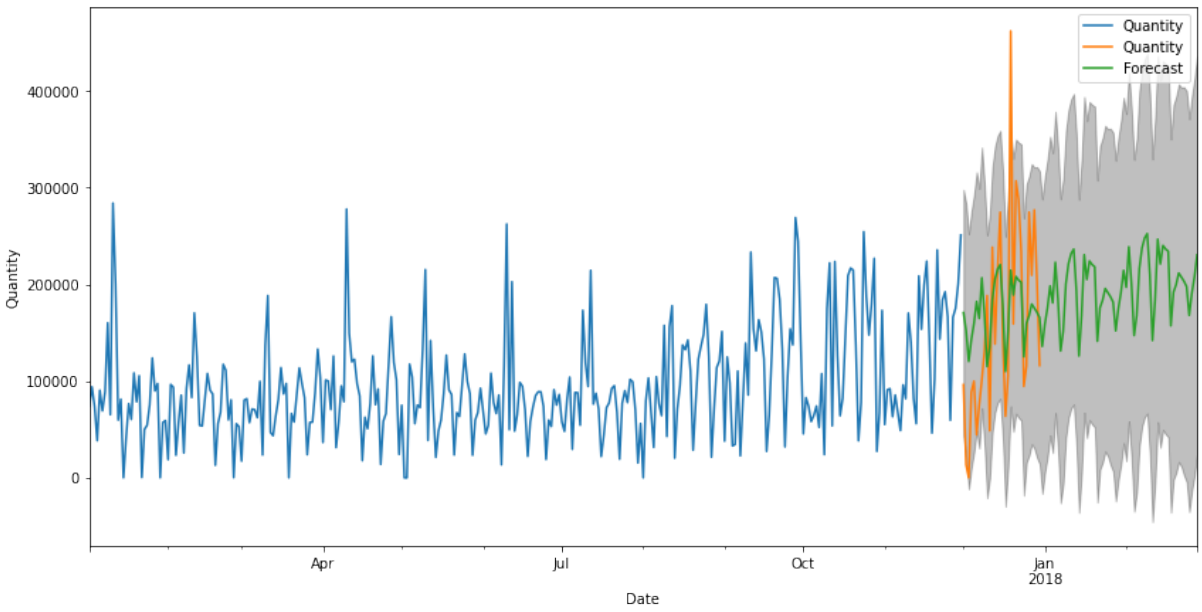

Рис. 6: Предсказание на 100 дней по 11 месяцам. Здесь синий — данные, на которых обучалась модель, зеленый — предсказание модели, желтым — истинное значение за декабрь (для визуального сравнения), серым закрашен доверительный интервал предсказания с уровнем доверия 0,05.

6 Заключение

Мы предложили два метода предсказания временного ряда для спроса некоторой группы товаров используя статистический и численный методы. Посчитали некоторый критерий предсказания, используя проверочные данные для одного месяца. Построенные методы были опробованы на больших данных, с которыми мы научились работать.

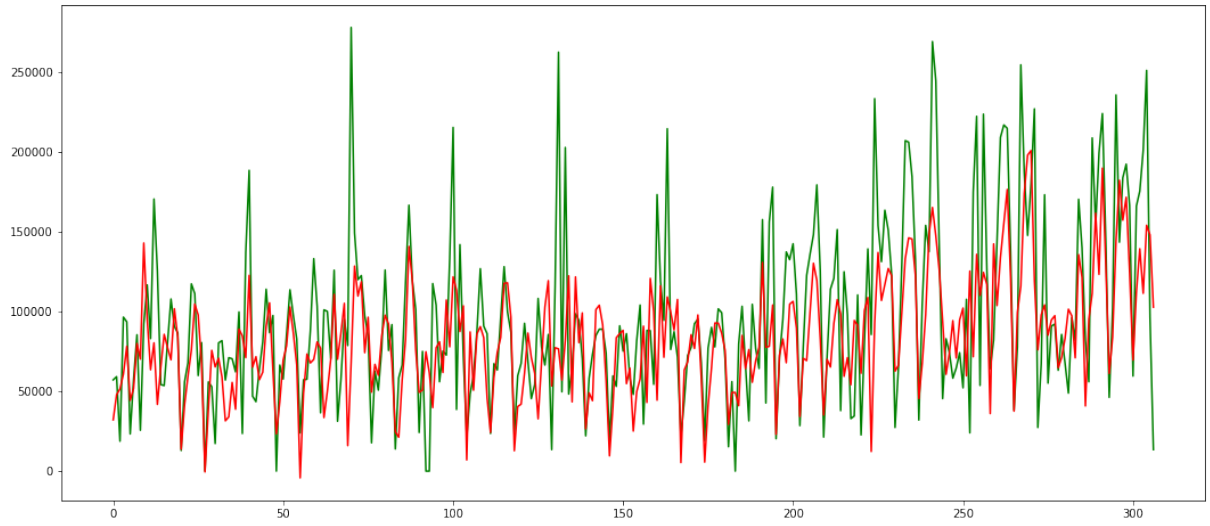

Рис. 7: Предсказание модели на обучающей выборке. Здесь красным — предсказание, зеленым — истинное значение.

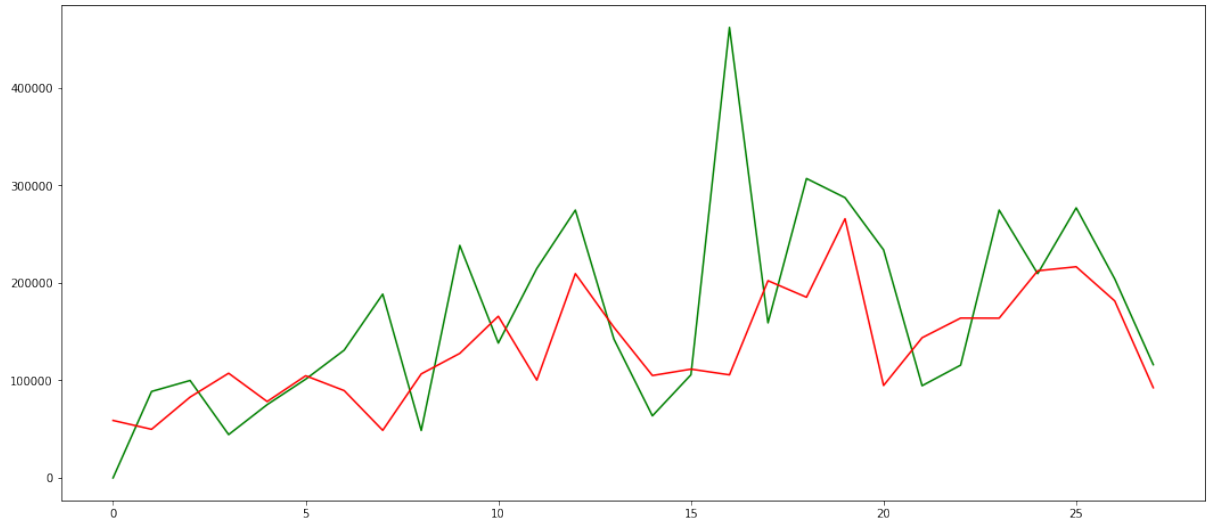

Рис. 8: Предсказание модели на проверочной выборке (за декабрь). Здесь красным — предсказание, зеленым — истинное значение.

Список литературы

[1] David A. Freedman. *Statistical Models: Theory and Practice.* Cambridge University Press, 2009.

[2] Mills, Terence C. *Time Series Techniques for Economists.* Cambridge University Press, 1990.

[3] Robert S. Pindyck and Daniel L. Rubinfeld. *Econometric Models and Economic Forecasts*, ch. 1, 1998.

[4] Yan, Xin. *Linear Regression Analysis: Theory and Computing.* World Scientific, 2009.