УДК 004

МЕТОДЫ ПРОГНОЗИРОВАНИЯ ПРОДАЖ АПТЕЧНОЙ СЕТИ

Егунова Алла Ивановна
к.и.н, доцент
Афонин Виктор Васильевич
к.т.н, доцент
Бабин Андрей Павлович
ФГБОУ ВО "МГУ им. Н. П. Огарёва"

Аннотация: В статье проводится сравнительный анализ методов оценки розничных продаж аптечной сети для прогнозирования плана закупок. Оценка выполнялась по среднему модулю отклонения и среднеквадратической ошибке модели по временным рядам. Результаты исследования показали, что алгоритм ARIMA обеспечивает более точные результаты для экстраполирования по временным рядам.

Ключевые слова: розничная торговля, прогнозирование, регрессионная модель, модель ARIMA,алгоритм XGBoost, градиентный бустинг, SVM – метод опорных векторов.

DRUG STORE SALES PREDICTION METHODS

Egunova Alla Ivanovna
Afonin Viktor Vasilevich
Babin Andrey Pavlovich

Abstract: The article provides a comparative analysis of methods for evaluating the retail sales of a pharmacy chain to forecast a procurement plan. The estimation was performed by the mean deviation modulus and the standard error of the model in time series. The results of the study showed that the ARIMA algorithm provides more accurate results for extrapolating over time series.

Keywords: retail, forecasting, regression model, ARIMA model, XGBoost algorithm, gradient boosting, SVM - support vector method.

При организации любого вида розничной торговли важной задачей является определение количества заказываемого у поставщика товара. В идеальном случае необходимо заказать такое количество какого-либо товара, чтобы оно успело реализоваться к моменту получения новой партии товаров от поставщика. Также сам процесс заказа товаров должен быть организован таким образом чтобы получить на выходе оптимальное соотношение затрат на транспортировку и хранение товаров.

Для аптечных сетей оптимальной и наиболее распространённой является практика, когда интервал между поставками составляет порядка 7-10 дней. Исходя из этого, стоит формировать заказ именно таким образом, чтобы он покрывал этот период.

Исходя из того что для формирования заказа необходимо определить количество заказываемого товара то основной задачей является успешное прогнозирования продаж на основе данных по предыдущим периодам.

Данные или же статистика продаж по предыдущим периодам являются по своей сути временными рядами и таким образом к ним могут быть применены все те методы анализа и прогнозирования, которые существуют для временных рядов.

Для осуществления предсказаний существует обширный перечень регрессионных методов, некоторые из которых и будут рассмотрены.

В общем виде регрессионная модель записывается следующим образом:

$$y = f(x, b) + \varepsilon,$$

где ε – случайная ошибка модели, $f(x, b)$ – функция регрессии, b – параметры модели.

Перейдем непосредственно к рассмотрению различных моделей регрессии.

Наиболее простой моделью является линейная регрессия она позволяет моделировать зависимость между некоторой независимой последовательностью и одной или множеством зависимых переменных. Функция линейной регрессии:

$$f(x, b) = \sum_{i=1}^{n} b_i x_i.$$

Данная модель может быть применена для прогнозирования продаж, но лишь в том случае если прослеживается линейная зависимость.

Еще одной моделью является полиномиальная регрессия. Данная модель в отличие от предыдущей позволяет предсказывать нелинейную зависимость. Функция полиномиальной регрессии:

$$f(x, b) = \sum_{i=1}^{n} \sum_{j=1}^{k} b_j x_i^j,$$

где j – настраиваемый параметр модели.

Модель является более гибкой в отличие от линейной регрессии, но в тоже время стоит отметить, что она также является более чувствительной к настройке.

ARIMA – модель и методология предназначенная для анализа нестационарных временных рядов. Является одной из самых распространенных моделей для построения краткосрочных прогнозов. Модель имеет следующий вид:

$$\triangle^d X^t = c + \sum_{i=1}^{p} a_i \triangle^d X_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t,$$

где c,a_i,b_j – параметры модели.

$$\triangle^d = (1 - L)^d,$$

L – лаговый оператор.

Три основные параметра модели: p,q,d. Параметр d задает число взятия последовательной разности для приведения ряда к стационарному. Параметры p и q, это авторегрессия и скользящее среднее соответственно [2].

XGBoost – является наиболее популярной реализацией алгоритма градиентного бустинга. Данный метод применяется для решения задач классификации, регрессии и ранжирования. Градиентный бустинг строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Обучение ансамбля проводится последовательно. На каждой итерации вычисляются отклонения предсказаний уже обученного ансамбля на обучающей выборке. Следующая модель, которая будет добавлена в ансамбль будет предсказывать эти отклонения. Новые деревья добавляются в ансамбль до тех пор, пока ошибка уменьшается, либо пока не выполняется одно из правил "ранней остановки" [3].

Функция оптимизации градиентного бустинга выглядит следующим образом:

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t),$$

где l – функция потерь, y_i – значение элемента обучающей выборки, $\hat{y}_i^{(t-1)}$ – сумма предсказаний первых t деревьев, f_t – обучаемая функция, x_i – набор признаков элемента обучающей выборки.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2,$$

где T – кол-во вершин в дереве, w – весовые коэффициенты в листьях, γ, λ – параметры регуляризации.

SVM – метод опорных векторов. Используется при решении задач классификации или регрессии. Идея метода заключается в переводе исходных векторов в пространство более высокой размерности и нахождение разделяющей гиперплоскости с максимальным зазором в этом пространстве. Затем строятся две параллельные гиперплоскости по обеим сторонам от разделяющей гиперплоскости. Предполагается, что ошибка классификации обратно пропорциональна расстоянию между двумя параллельными гиперплоскостями.

Для реализации нелинейной классификации используется переход от скалярных произведений к произвольным ядрам. Таким образом, линейная классификация в новом пространстве, становится эквивалентной нелинейной классификации в исходном пространстве. Данная версия метода называется SVR(метод опорных векторов для регрессии) и применим для прогнозирования временных рядов[4].

В SVM на вход подается исходный вектор x, который сопоставляется с признаковым пространством высокой размерности для улучшения линейной отделимости. Если обучающие данные линейно разделимы, после отображения в пространство признаков, то функция принятия решения будет иметь следующий вид:

$$f(x_i) = (w * x_i) + b,$$

где w – вектор весовых коэффициентов, b – параметр смещения.

Для оценки и сравнения различных регрессионных методов и моделей использовались следующие характеристики: средний модуль отклонения и корень среднеквадратической ошибки модели.

Средний модуль отклонения модели (Mean absolute error) – это величина, которая показывает разницу между двумя временными рядами. Формула среднего модуля отклонения модели:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n},$$

где y_i – наблюдаемое значение, $\hat{y}_i$ – предсказанное значение.

Корень среднеквадратической ошибки модели также отражает различие между исходным и спрогнозированным временными рядами, но в отличие от среднего модуля отклонения грубые ошибки становятся более заметными за счет возведения в квадрат. Формула расчета корня среднеквадратической ошибки модели:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}.$$

Сравнение осуществлялось следующим образом бралась выборка данных длинною 52 элемента, которая содержала в себе данные о продажах в течение 52 недель соответственно. Затем выборка разбивалась на две части тренировочную – длинной 34 элемента, и тестовую – длинной 18 элементов. После чего производилось обучение модели на основе тренировочной выборки, а затем предсказание будущих 18-ти значений. Получившиеся значения сравнивались с тестовой выборкой, путем расчета среднего модуля отклонения и среднеквадратической ошибки модели.

Первым тестировалась модель полиномиальной регрессии, результаты тестирования представлены в таблице 1. На рисунке 1 изображен график, на котором сопоставлены функция предсказания, получившаяся в результате использования метода полиномиальной регрессии и исходные данные тестовой выборки.

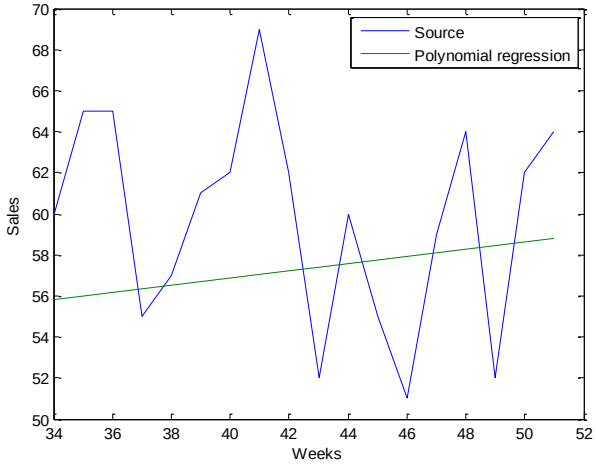

Рис. 1 Полиномиальная регрессия с параметром 1

Таблица 1

Показатели полиномиального распределения

Значение параметра	Средний модуль отклонения	Корень среднеквадратической ошибки
1	4,9477	5,7370
2	12,5032	15,2662
3	6,5126	8,3821

Характеристики модели ARIMA с различными параметрами представлено в таблице 2. На рисунке 2 приведено сравнение прогнозов выполненных при помощи модели ARIMA(5,1,0) с тестовой выборкой.

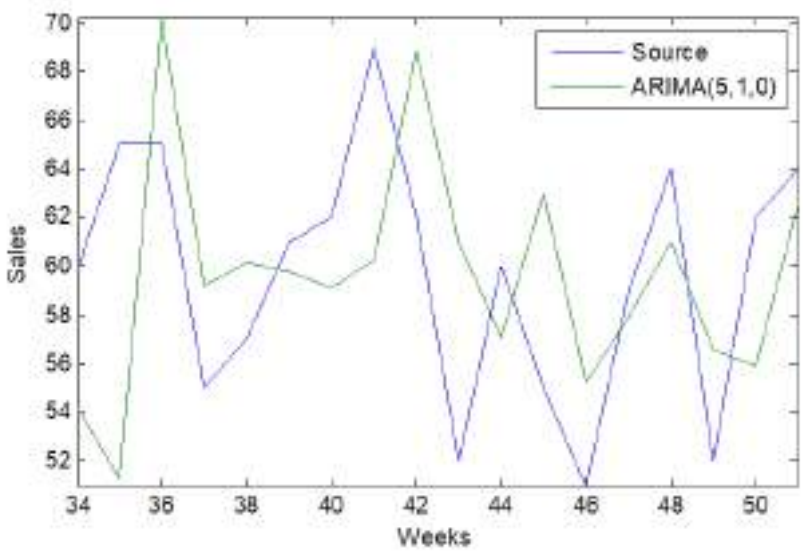

Рис. 2 Модель ARIMA(5,1,0)

Таблица 2

Показатели модели ARIMA

Значения параметров	Средний модуль отклонения	Корень среднеквадратической ошибки
ARIMA(2,0,1)	4.240	4.906
ARIMA(3,0,0)	3.886	4.754
ARIMA(4,0,0)	3.671	4.445
ARIMA(4,1,0)	3.941	4.802
ARIMA(5,0,0)	3.653	4.348
ARIMA(5,1,0)	3.875	4.210
ARIMA(5,2,0)	4.088	4.811

Для модели XGBoost в результате настройки были получены следующие значения: MAE – 4.946, RMSE – 5.663. Результаты сравнения с тестовой выборкой представлены на рисунке 3.

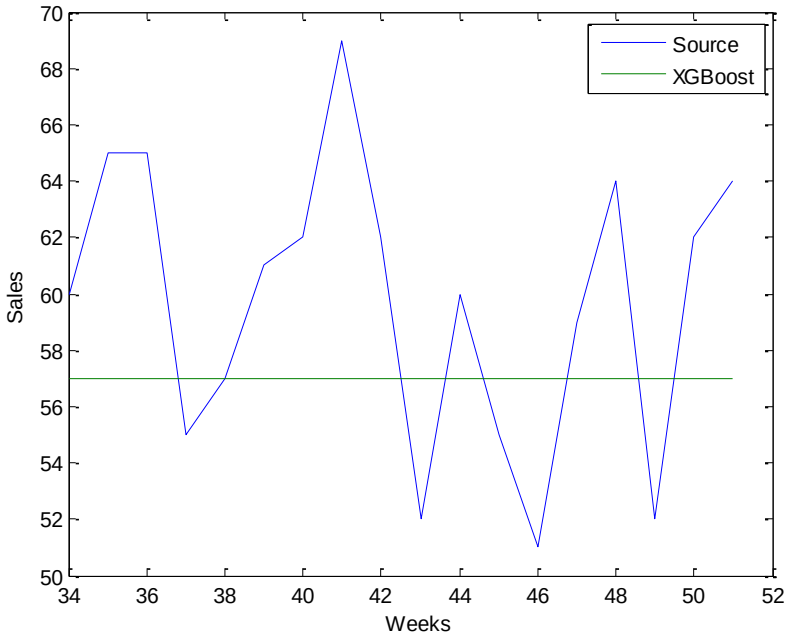

Рис. 3 Модель XGBoost

В результате тестирования модели SVR были получены следующие значения MAE – 7.439, RMSE – 8.689. Сравнение с результатами тестовой выборки представлено на рисунке 4.

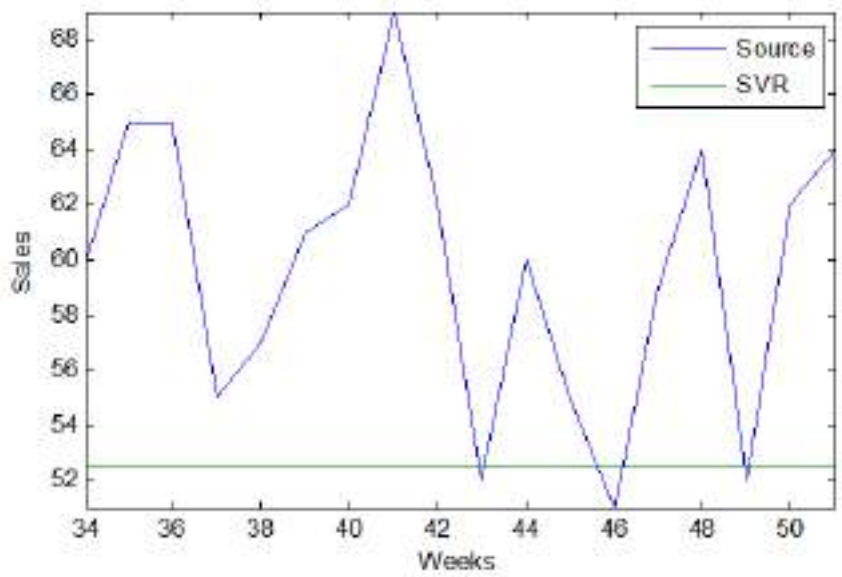

Рис. 4 Модель SVR

После проведения испытания лучшие результаты каждого из методов были сведены в таблицу 4.

Таблица 4

Сравнительная таблица методов

Модель	Средний модуль отклонения	Корень среднеквадратической ошибки
Poly(1)	4,9477	5,7370
ARIMA(5,1,0)	3.875	4.210
XGBoost	4.946	5.663
SVR	7.4397	8.6895

В результате произведенного тестирования описанных выше методов было выявлено, что модель ARIMA наилучшим образом подходит для работы с временными рядами в задаче прогнозирования продаж. Что подтверждается полученными графиками и расчетными характеристиками. Остальные же методы в силу своих структурных особенностей плохо подходят для задач экстраполяции.

Список литературы

1. Афонин В.В. Моделирование систем: учебно-практическое пособие / В.В. Афонин, С.А. Федосин. – М.: Интуит, – 2016. – 23 с.

2. ARIMA – Википедия [Электронный ресурс]. — Режим доступа: https://ru.wikipedia.org/wiki/ARIMA

3. XGBoost – Викиконспекты [Электронный ресурс]. — Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=XGBoost

4. Метод опорных векторов [Электронный ресурс]. — Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Метод_опорных_векторов