ТЕХНИЧЕСКИЕ НАУКИ

TECHNICAL SCIENCES

УДК 004.8
ББК 32.813
Б 94

Бучацкая Виктория Викторовна
Доцент, кандидат технических наук, доцент кафедры прикладной математики, информационных технологий и информационной безопасности Адыгейского государственного университета, Майкоп, тел. (8772) 593904, e-mail: buch_vic@mail.ru
Бучацкий Павел Юрьевич
Доцент, кандидат технических наук, заведующий кафедрой автоматизированных систем обработки информации и управления Адыгейского государственного университета, Майкоп, тел. (8772) 593911, e-mail: buch@adygnet.ru
Лобанов Валерий Евгеньевич
Магистрант Адыгейского государственного университета, Майкоп, e-mail: valery2698@mail.ru

Анализ алгоритмов прогнозирования

(Рецензирована)

Аннотация. *Проведен анализ наиболее используемых на практике алгоритмов прогнозирования временных рядов. Представлена визуальная демонстрация прогноза различных моделей для различных временных рядов. Составлена сводная таблица метрик производительности моделей.*

Ключевые слова*: алгоритмы прогнозирования, временные ряды, авторегрессионные модели, градиентный бустинг над решающими деревьями.*

Buchatskaya Viktoriya Viktorovna
Associate Professor, Candidate of Technical Sciences, Associate Professor of Department of Applied Mathematics, Information Technologies and Information Security, Adyghe State University, Maikop, ph. (8772) 593904, e-mail: buch_vic@mail.ru
Buchatsky Pavel Yuryevich
Associate Professor, Candidate of Technical Sciences, Head of Department of Automated Systems of Processing Information and Control, Adyghe State University, Maikop, ph. (8772) 593911, e-mail: butch_p99@mail.ru
Lobanov Valeriy Evgenyevich
Magistrand of Adyghe State University, Maikop, e-mail: valery2698@mail.ru

Analysis of forecasting algorithms

Abstract. *The article analyzes the most commonly used time-series forecasting algorithms. The out-of-sample forecasting data of different models is provided. The list of metrics characterizing performance of models is summarized in the table.*

Keywords: *forecasting algorithms, time-series, autoregression models, gradient boosting on decision trees.*

В наши дни задача прогнозирования временных рядов встречается довольно часто. Происходит это потому, что многие явления, такие, как курс акций, статистика продаж, биржевые курсы валют, температура воздуха и, в общем, протекание различных процессов в течение определенного отрезка времени можно смоделировать как временной ряд.

Однако задача прогнозирования не всегда может быть решена одним стандартным набором методов. Основной проблемой прогнозирования временных рядов является корреляция наблюдений в различные промежутки времени и, соответственно, некоторые алгоритмы могут серьезно изменить или исказить порядок наблюдений, что делает полученный результат непригодным. Поэтому актуальной остается задача анализа работы различных моделей прогнозирования временных рядов.

На основе работ [1, 2] были изучены существующие подходы к решению задачи прогнозирования. Наиболее используемыми на сегодняшний день остаются статистические мо-

дели прогнозирования, а именно семейство авторегрессионных моделей [3]. Формально модель ARMA представима следующей формулой [4]:

$$ARMA(p,q):\ y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_{t-p} + \varepsilon_t + \vartheta_1 \varepsilon_{t-1} + \vartheta_2 \varepsilon_{t-2} + ... + \vartheta_q \varepsilon_{t-q},$$

где y_t – стационарный ряд с нулевым средним;

φ_1, ϑ_q – константы, неравные нулю;

ε_t – гауссов белый шум с нулевым средним и постоянной дисперсией τ_ε^2.

Основным преимуществом является гибкая настройка параметров, простота, прозрачность моделирования, единообразие анализа и проектирования [5]. Наибольший интерес в последние годы вызывают интеллектуальные подходы к решению задачи. Наиболее мощным инструментом можно назвать нейронные сети [2, 6], а именно архитектуру LTSM. Однако несмотря на превосходство точности прогноза в отдельных случаях и универсальность использования на различных данных, они имеют два существенных недостатка: невозможность анализа работы алгоритма и ресурсоемкость процесса обучения. В связи с этим был выбран набирающий популярность в практических задачах метод градиентного бустинга над решающими деревьями [5]. Модель представлена библиотекой XGBoost.

XGBoost строит композицию из K базовых алгоритмов b_k:

$$\hat{y}_i = \hat{y}_i^K = \sum_{k=1}^{K} b_k(x_i) = \hat{y}_i^{(K-1)} + b_k(x_i),$$

минимизируя следующий функционал:

$$Obj = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(b_k),$$

где N – размер обучающей выборки;

x_i, y_i, $\hat{y}_i$ – i-ый объект, правильный ответ и предсказание модели для него;

$\hat{y}_i^t$ – предсказание композиции из t уже обученных базовых алгоритмов для i-го объекта;

Ω – регуляризатор;

$(y_i, \hat{y}_i)$ – функция потерь.

Данное семейство моделей является серьезным улучшением классического статистического метода решающих деревьев, благодаря быстрым градиентным методам. К плюсам данного решения можно отнести встроенную регуляризацию, кросс-валидацию и обработку пропущенных данных, параллельность вычислений.

Таким образом, были проведен анализ двух семейств алгоритмов прогнозирования, ARIMA и XGBoost Regressor.

Для примера результата прогнозов были взяты два набора данных: ежемесячные продажи вина в Австралии в тысячах литров с января 1980 года по июль 1995 года и ежегодный размер поголовья рыси в период с 1821 года по 1934 год. Для авторегрессионной модели произведена предварительная обработка данных: устранены пропущенные значения, проверка данных на нормальное распределение и произведен тест Дики-Фуллера [7].

По найденным приближенным значениям тестов на автокорреляцию и частичную автокорреляцию для данных ежемесячных продаж были получена модель $SARIMAX(3,1,2)\times(1,0,2)_{12}$, а для данных ежегодной периодичности – модель $SARIMAX(2,1,1)\times(1,1,2)_5$.

С помощью перебора параметров были найдены наилучшие параметры из возможных, основываясь на тестах Акаике, Шварца и Ханна-Куина. Получены модели $SARIMAX(2,1,2)\times(0,1,1)_{12}$ и $SARIMAX(1,1,1)\times(1,1,2)_5$ для ежемесячных и годовых данных соответственно.

На рисунках 1 и 2 представлены графики прогноза моделей по тестовой выборке (выделено штриховыми линиями) и прогноз вне выборки на 48 месяцев и 10 лет вперед соответственно.

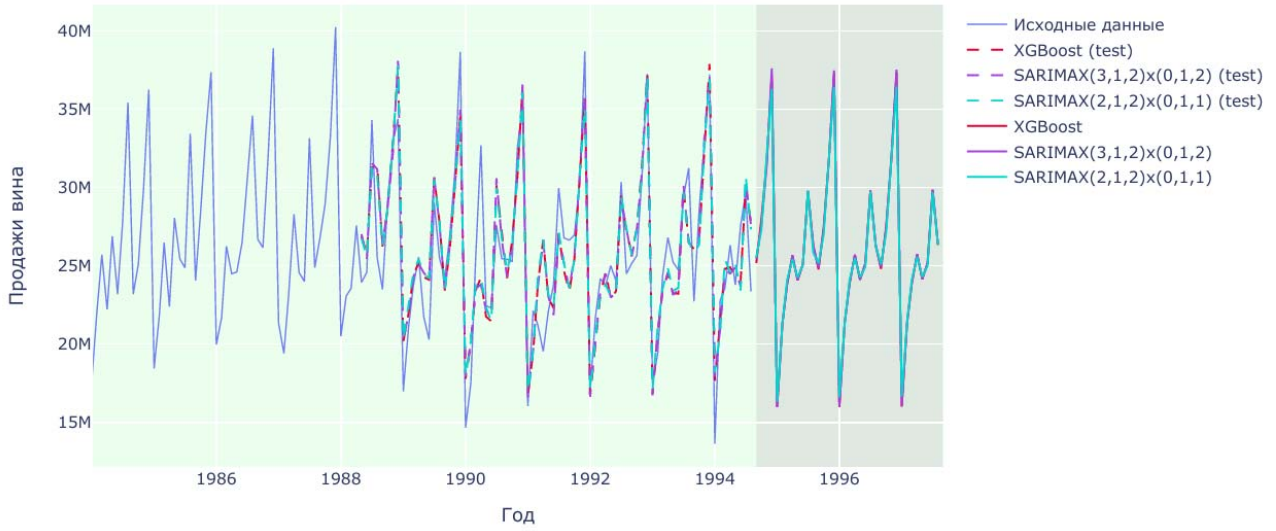

Рис. 1. Прогноз продаж на 48 месяцев

Как видно из рисунка 1, различия прогноза минимальны, модели и данные похожи друг на друга, прогнозы вне выборки выглядят адекватно и отражают годовую сезонность данных, и не выражают определенного тренда.

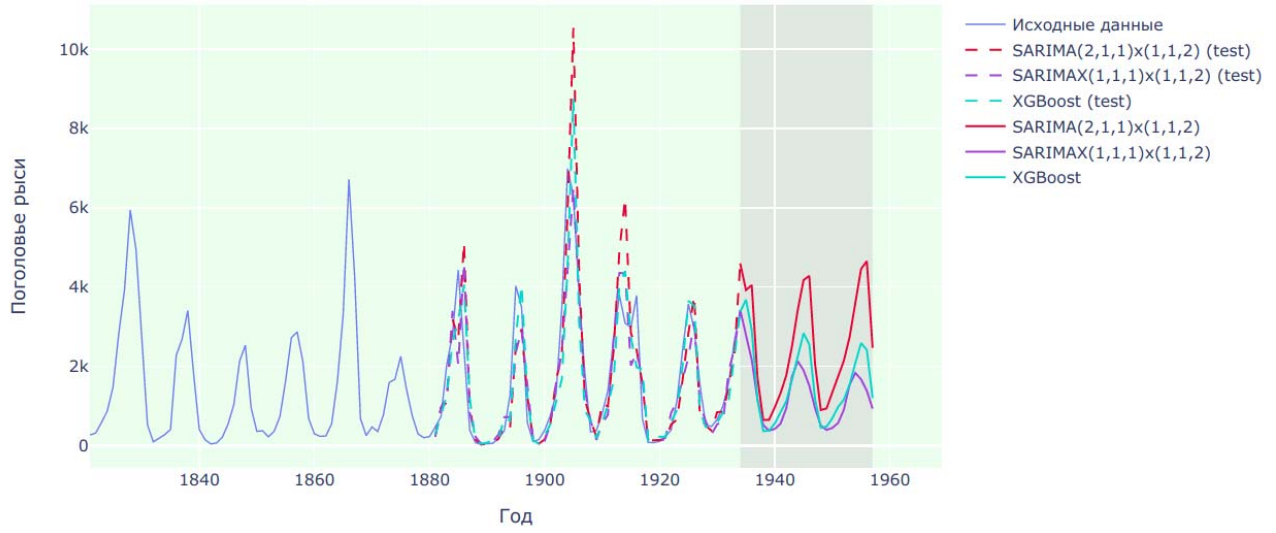

Рис. 2. Прогноз поголовья рыси на 10 лет

Несколько иная картина наблюдается при построении прогноза поголовья рыси на 10 лет (рис. 2). Базовая модель ARIMA явно имеет большую дисперсию и хоть сохраняет сезонность и характер ряда, кажется, отстает по точности от конкурентов.

Метрики по тестовой выборке данных продаж вина подтверждают состоятельность моделей. Все три алгоритма оказались очень схожи. Разброс ошибок для второй таблицы данных также небольшой (см. табл. 1).

По итогам работы были получены численные оценки двух семейств моделей прогнозирования временных рядов. Представленные визуальные характеристики прогнозных значений подтверждают состоятельность моделей.

Таблица 1

Сравнение метрик моделей для двух таблиц данных

Алгоритм	RMSE	AIC	BIC	MAPE	MAE	P(Q)
Ежемесячные продажи вина						
$SARIMAX(3,1,2)\times(1,0,2)_{12}$	2,779.6108	-296.8468	-269.0030	10.53	2,363.4	0.47
$SARIMAX(2,1,2)\times(0,1,1)_{12}$	2,639.8143	-295.794	-280.325	9.73	2,087.8	0.33
XGBoost (6)	2,347.6784	-	-	9.27	2,005.2	-
Ежегодное поголовье рыси						
$SARIMAX(2,1,1)\times(1,1,2)_{5}$	2,991.519	612.922	631.697	13.78	2,201,4	0.29
$SARIMAX(1,1,1)\times(1,1,2)_{5}$	2,789.477	626.482	645.574	9.31	2,099.2	0.12
XGBoost (10)	2,213.611	-	-	9.12	2,001.3	-

В дальнейшем планируется уделить внимание эвристикам временных рядов, на основе графиков декомпозиции составить план анализа остатков и в соответствии с этими данными выбрать наилучшие параметры дифференцирования для авторегрессионнных моделей.

Примечания:

1. Statistical forecasting: notes on regression and time series analysis. August 18, 2020. URL: https://people.duke.edu/~rnau/411home.htm
2. Симанков В.С., Бучацкая В.В. Обзор методов прогнозирования: деп. рукопись № 302-В2012 // ВИНИТИ РАН. 2012. № 7. 84 с.
3. Айвазян С.А. Основы эконометрики: учеб. пособие. Т. 2. Москва, 2001. С. 246–327.
4. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks / M.J. Kane, N. Price, M. Scotch [et al.] // BMC Bioinformatics. 2014. Vol. 15 (1). 276 p. DOI: 10.1186/1471-2105-15-276
5. Hyndman R.J., Athanasopoulos G. Forecasting: principles and practice // OTexts. 2015. URL: https://www.otexts.org/book/fpp
6. Симанков В.С., Бучацкая В.В. Выбор метода прогнозирования при исследовании сложных систем // Вестник Адыгейского государственного университета. Сер.: Естественно-математические и технические науки. 2012. Вып. 2 (101). С. 114–119. URL: URL: http://vestnik.adygnet.ru
7. Горелова Л.В., Мельникова Е.Н. Основы прогнозирования систем: учеб. пособие для инж.-экон. спец. ВУЗов. Москва: Высш. шк., 1986. 276 с.

References:

1. Statistical forecasting: notes on regression and time series analysis. August 18, 2020. URL: https://people.duke.edu/~rnau/411home.htm
2. Simankov V.S., Buchatskaya V.V. Review of forecasting methods: dep. manuscript No. 302-B2012 // VINITI RAS. 2012. No. 7. 84 p.
3. Ayvazyan S.A. Fundamentals of econometrics: a manual. Vol. 2. Moscow, 2001. P. 246–327.
4. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks / M.J. Kane, N. Price, M. Scotch [et al.] // BMC Bioinformatics. 2014. Vol. 15 (1). 276 p. DOI: 10.1186/1471-2105-15-276
5. Hyndman R.J., Athanasopoulos G. Forecasting: principles and practice // OTexts. 2015. URL: https://www.otexts.org/book/fpp
6. Simankov V.S., Buchatskaya V.V. Choice of methods of forecasting in researches of complicated systems // The Bulletin of the Adyghe State University. Ser.: Natural-Mathematical and Technical Sciences. 2012. Iss. 2 (101). P. 114–119. URL: URL: http://vestnik.adygnet.ru
7. Gorelova L.V., Melnikova E.N. Basics of forecasting systems: a manual for engineer and economic specialities of higher schools. Moscow: Vysshaya Shkola, 1986. 276 p.