

DOI:

АНАЛИЗ ПРОГНОСТИЧЕСКОЙ СПОСОБНОСТИ ГИБРИДНЫХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ ДИНАМИКИ КОТИРОВОК АКЦИЙ

Оршанская Е.С.,

*Институт Национальный исследовательский университет «Высшая школа экономики»
(НИУ ВШЭ), г. Москва, ул. Мясницкая, д. 20
orshanskayaes@gmail.com*

Сизых Н.В.

*Институт Национальный исследовательский университет «Высшая школа экономики»
(НИУ ВШЭ), г. Москва, ул. Мясницкая, д. 20
sizykh_n@mail.ru*

Аннотация: В работе представлены результаты изучения прогностической способности гибридных моделей, основанных на объединении статистического подхода и методов машинного обучения, для прогнозирования динамики котировок акций. Данное исследование основано на анализе предсказательной силы двух гибридных моделей прогнозирования SARIMA/MLP и SARIMA/LSTM в сравнении с классической SARIMA как опорной точки на примере 30 компаний из двух сегментов рынка: автомобильном и информационных технологий.

Ключевые слова: динамика котировок акций, гибридные модели прогнозирования, прогнозирование временных рядов, прогностическая способность, качество прогнозирования

Введение

Согласно данным современных исследований в области прогнозирования динамики котировок акций точность прогнозирования можно оценить как умеренно низкую. Однако, некоторые аналитики указывают на возможность повышения точности прогноза при помощи применения методов, основанных на комбинации различных подходов к прогнозированию [1-4]. При этом данные практических исследований указывают на имеющуюся полезность учета финансовых ожиданий на рынках прогнозирования (особенно на интервалах прогнозирования до года) при целесообразности комбинирования современных методов с традиционными подходами к прогнозированию динамики финансовых рынков [5-7]. Некоторые аналитики указывают на возможность достичь достаточно высокую точность прогнозов, основанных на комбинации различных подходов к прогнозированию [1-4].

Учитывая нестационарность, присущую финансовым временным рядам, далеко не все разработанные модели прогнозирования временных рядов эффективны для предсказания курса акций. В последнее десятилетие в области прогнозирования курсов акций можно заметить смену концентрации внимания с традиционных статистических моделей на более сложные, в основе которых лежат алгоритмы машинного обучения и искусственного интеллекта. Данное положение подтверждается также существующими сервисами, предоставляющими услуги по автоматическому формированию прогноза для выбранной акции на необходимый промежуток времени. Базовыми моделями, лежащими в их основе, являются рекуррентные и сверточные сети, однако они не ограничиваются исследованием только одного финансового ряда, а используют дополнительные данные и показатели. Также одним из перспективных направлений развития современных методов и моделей прогнозирования является разработка и внедрение гибридных моделей, в основе которых может лежать как объединение традиционных статистических моделей и методов машинного обучения, так и техник, использующих более углубленный математический аппарат [8-12].

1 Гибридные модели прогнозирования динамики котировок акций

Существуют различные вариации гибридных моделей, представляющих из себя комбинацию статистических методов и методов машинного обучения. В основе большинства из них лежит идея, которая заключается в том, чтобы линейная и нелинейная часть прогноза формировались подходящей для этого моделью. Тем не менее, реализации сильно различаются.

В данной работе в качестве отправной точки для сравнения качества прогноза и проверки целесообразности использования гибридных техник взята статистическая модель SARIMA. Для анализа прогностической способности были выбраны две гибридные модели: SARIMA/MLP и SARIMA/LSTM, вдохновленные работой [13]. Авторами было предложено декомпозировать ряд котировок акций y_t при помощи простого фильтра скользящего среднего на две части: линейную с низкой волатильностью l_t , что является результатом сглаживания, и нелинейную часть с высокой

волатильностью g_t , где $g_t = y_t - l_t$. Фильтр подбирается таким образом, чтобы получившийся сглаженный ряд подчинялся нормальному закону, проверка происходит путем подсчета коэффициента эксцесса и его сравнения с критическим значением 3. Далее с помощью ARIMA с оптимально подобранными параметрами делается прогноз менее изменчивой части \hat{l}_t , а с помощью простой рекуррентной нейронной сети более изменчивую \hat{g}_t . После чего путем сложения результатов моделей получалось итоговое предсказание: $\hat{y}_t = \hat{l}_t + \hat{g}_t$.

Поскольку в данной работе представлено большое число компаний, данные которых значительно отличаются по своим характеристикам, было решено рассмотреть различные фильтры скользящего среднего на случай, если простого будет недостаточно для достижения нормального распределения сглаженного ряда. Таким образом, были выбраны следующие вариации:

1. *Просто скользящее среднее (SMA)* для периода t – это среднее арифметическое w последних наблюдений [14,15]:

$$SMA_t = \frac{1}{w} \sum_{j=t-w}^{t-1} x_j,$$

где x_j -значение ряда в момент времени j ,

w - окно скользящей средней (количество измерений, используемых для расчета скользящей средней)

2. *Экспоненциальное скользящее среднее (EMA)* [14,15]:

$$EMA_t = \alpha \cdot x_{t-1} + (1 - \alpha) \cdot EMA_{t-1},$$

где EMA_t - экспоненциальное скользящее среднее в момент времени t ,

α -коэффициент сглаживания. Принимает значение от 0 до 1.

Ее можно выбрать исходя из окна скользящей средней:

$$\alpha = \frac{2}{w + 1}$$

Взвешенное скользящее среднее (WMA) [15]:

$$WMA_t = \sum_{j=t-w}^{t-1} \alpha_j \cdot x_j,$$

где: WMA_t - взвешенное скользящее среднее в момент времени t ,

α_j -вес измерения в момент времени j .

4. *Адаптивная скользящая средняя Кауфмана (КАМА)* [16]:

$$КАМА_t = КАМА_{t-1} + SC_{t,w} \cdot (x_t - КАМА_{t-1}),$$

где: $КАМА_t$ – адаптивная скользящая средняя Кауфмана в момент времени t ,

SC (Сглаживающая константа в момент времени t для окна w) = $(ER_{t,w} * (FC - SC) + SSC)^2$,

FC (быстрый коэффициент сглаживания) = $\frac{2}{f+1}$,

SC (медленный коэффициент сглаживания) = $\frac{2}{s+1}$,

f, s - константы для быстрой и медленной скользящей средней;

w – окно скользящей средней,

$ER_{t,w}$ (коэффициент эффективности в момент времени t для окна w) = $\frac{|x_t - x_{t-w-1}|}{\sum_{i=0}^{w-1} |x_{t-i} - x_{t-i-1}|}$.

5. *Тройное экспоненциальное скользящее среднее (ТЕМА)* [17]:

$$ТЕМА = 3 \cdot EMA_1 - 3 \cdot EMA_2 + EMA_3,$$

где EMA_1 – простое экспоненциальное скользящее среднее,

EMA_2 – скользящее среднее от EMA_1 ,

EMA_3 – скользящее среднее от EMA_2 .

6. *Треугольное скользящее среднее (ТМА)* (25) [18]:

$$ТМА = \frac{1}{w} \sum_{j=1}^w SMA_j,$$

где: SMA_j – скользящее среднее от SMA_{j-1} для $j \in [2, w]$,

w - окно скользящей средней.

2 Прогностическая способность

Прогностическая способность методов предсказания динамики котировок акций зависит не только от выбранных методов и моделей, но и от множества факторов, среди которых можно выделить следующие: волатильность показателя; период прогнозирования; горизонт прогнозирования; субъект прогнозирования и т.п. [19]. Точность прогнозирования значительно различается в зависимости от периода прогнозирования (гораздо ниже при смене фазы деловых циклов), от прогнозируемого индикатора (гораздо ниже для финансовых показателей, а также ряда нефинансовых показателей, например, уровня запасов; выше для ВВП, инфляции, уровня безработицы) и от горизонта прогнозирования (падает с ростом горизонта).

При оценке точности прогноза учитывается и специфика прогнозирования акций, которая состоит в сложности финансовых рынков, влияние на их динамику огромного количества факторов, репутационные риски для субъектов прогнозирования при существенных отклонениях прогноза от действительности. С учетом имеющейся специфики аналитики выделяют следующие особенности прогнозирования динамики курсов акций:

- преобладание краткосрочных прогнозов (на несколько месяцев, год, иногда – до двух-трех лет);
- отсутствие точных значений в средне- и долгосрочных прогнозах;
- доступность во многом только текущих прогнозов и их закрытость вследствие коммерческой ценности.

С целью оценки качества выбранных методов были использованы безразмерные метрики MAPE, TIC и MDA, поскольку смысловая нагрузка не теряется при усреднении их значений; MAE, выраженная в исходных величинах, для финальной оценки прогноза и MPE для того, чтобы охарактеризовать выбранную оптимальную модель: свойственно ли ей завышать или занижать предсказание.

Исходные данные об исторических котировках акций компаний для проведения исследования были взяты с сервиса Investing.com посредством библиотеки InvestPy [20]. Для анализа были выбраны акции 30 компаний двух отраслей рынка: автомобильный сегмент и сегмент информационных технологий. Данное решение обуславливается тем, что указанные индустрии остаются относительно стабильными с течением времени и не претерпевали заметных кризисов за последнее десятилетие. Рассмотрение каждого сегмента в отдельности было реализовано в соответствии с предположением о его влиянии на данные о котировках акций компании.

Для сравнения прогностической способности моделей с разными параметрами и на разных периодах были выбраны следующие метрики [14]:

средняя абсолютная ошибка MAE

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

средняя абсолютная ошибка в процентах MAPE (относительная величина, не содержит единицу измерения и может быть использована для сравнения предсказаний по разным наборам данных)

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| * 100\%$$

средняя процентная ошибка MPE (позволяет определить смещение прогноза: свойственно ли модели завышение или занижение прогноза)

$$MPE = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \hat{y}_i)}{y_i}$$

коэффициент неравенства Тейла TIC (значение лежит в границах от 0 до 1, где нуль соответствует идеальной прогностической способности, а единица – ее полному отсутствию)

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

Средняя направленная точность MDA (подсчет числа правильного определения направления динамики курсов акций (возрастания или убывания))

$$MDA = \frac{1}{n} \sum_{t=2}^n z_t,$$

где: $z_t = 1$, если $(y_t - y_{t-1})(\hat{y}_t - y_{t-1}) > 0$,
 $z_t = 0$, если $(y_t - y_{t-1})(\hat{y}_t - y_{t-1}) < 0$,
 n -количество объектов в выборке,
 y_i - истинное значение,
 \hat{y}_i - прогнозное значение.

2 Анализ экспериментальных данных

Данные для исследования моделей прогнозирования были загружены с применением библиотеки InvestPy для языка программирования Python. Исторические данные были загружены с 01/06/2013 по 30/06/2019, где период с 01/06/2019 был выбран для формирования прогноза. Наиболее приемлемый исторический период для обучения был определен в ходе дальнейшего анализа. Архитектура нейронных сетей была выбрана в результате проведения экспериментов, которые заключались в усредненной оценке по отраслям качества прогноза с использованием метрик MAPE, TIC и MDA в зависимости от количества слоев и нейронов в слое и выборе компромиссного варианта. Были исследовано от одного до двух скрытых слоев; 128/256/512/1024/1500/2048 нейронов в слое для части MLP; 128/256/512/1024 нейронов в слое для LSTM. Лучшие результаты соответствуют однослойным архитектурам (за исключением SARIMA/LSTM ИТ-отрасли) с большим количеством нейронов (Таблица 1). Это может объясняться тем, что нейронные сети склонны к переобучению с увеличением количества слоев, поскольку за счет большого числа параметров модели легче подстроиться под обучающий набор данных, не выявляя при этом общие закономерности.

Таблица 1. Результаты эксперимента по подбору архитектуры сетей.

Отрасль	Модель	Архитектура сети
Автомобильная отрасль	SARIMA/MLP	Один слой, 1500 нейронов
	SARIMA/LSTM	Один слой, 1024 нейрона
Отрасль информационных технологий	SARIMA/MLP	Один слой, 1500 нейронов
	SARIMA/LSTM	Два слоя по 1024 нейрона в каждом

При выборе временного окна для составления выборки за начальную точку было взято значение 60. Однако, для улучшения дальнейшего прогноза необходимо провести проверку, действительно ли данное значение параметра хорошо работает на исследуемых данных. Для достижения данной цели был проведен анализ зависимости метрик качества от количества предыдущих данных котировок при составлении одного объекта выборки. Исследовались следующие значения параметра: 25, 40, 60 и 75.

Для модели SARIMA/MLP автомобильной отрасли, как видно исходя из верхней области рисунка 1, оптимальным вариантом временного окна является значение 60: можно заметить, что именно в данной точке MAPE и TIC достигают минимума, а MDA одного из максимальных значений. Для информационной отрасли лучшие результаты были получены при размерности выборки равном 75.

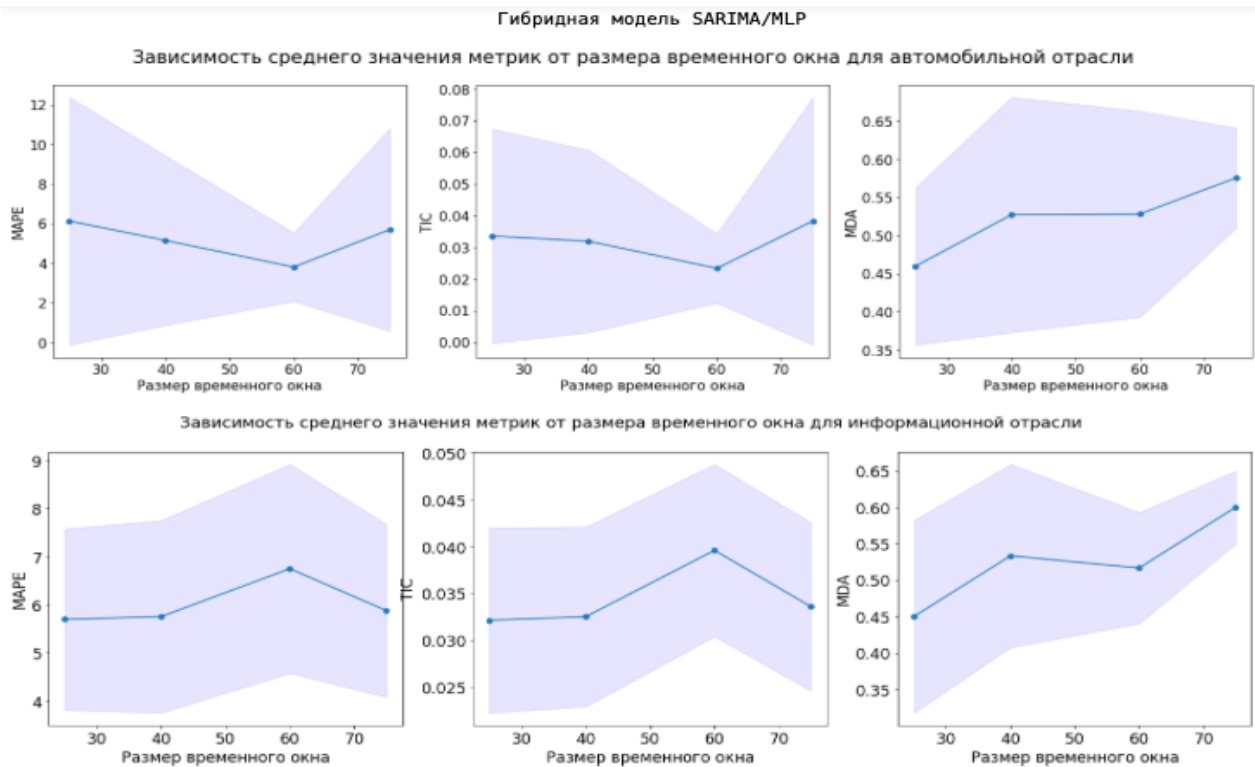


Рис. 1 Зависимость значения метрик качества от значения параметра временного окна для обеих отраслей модели SARIMA/MLP.

Модель SARIMA/LSTM, как видно из рисунка 2, для автомобильной отрасли показывает в среднем лучшие результаты при временном окне равном 75: MAPE и TIC достигают своего минимума, а MDA максимума. При этом, разброс значений относительно мал, в сравнении с другими вариантами.

Для отрасли информационных технологий ситуация противоположная: модель показывает лучшие результаты при минимальном количестве рассмотрений предыдущих значений цены акции - 25.

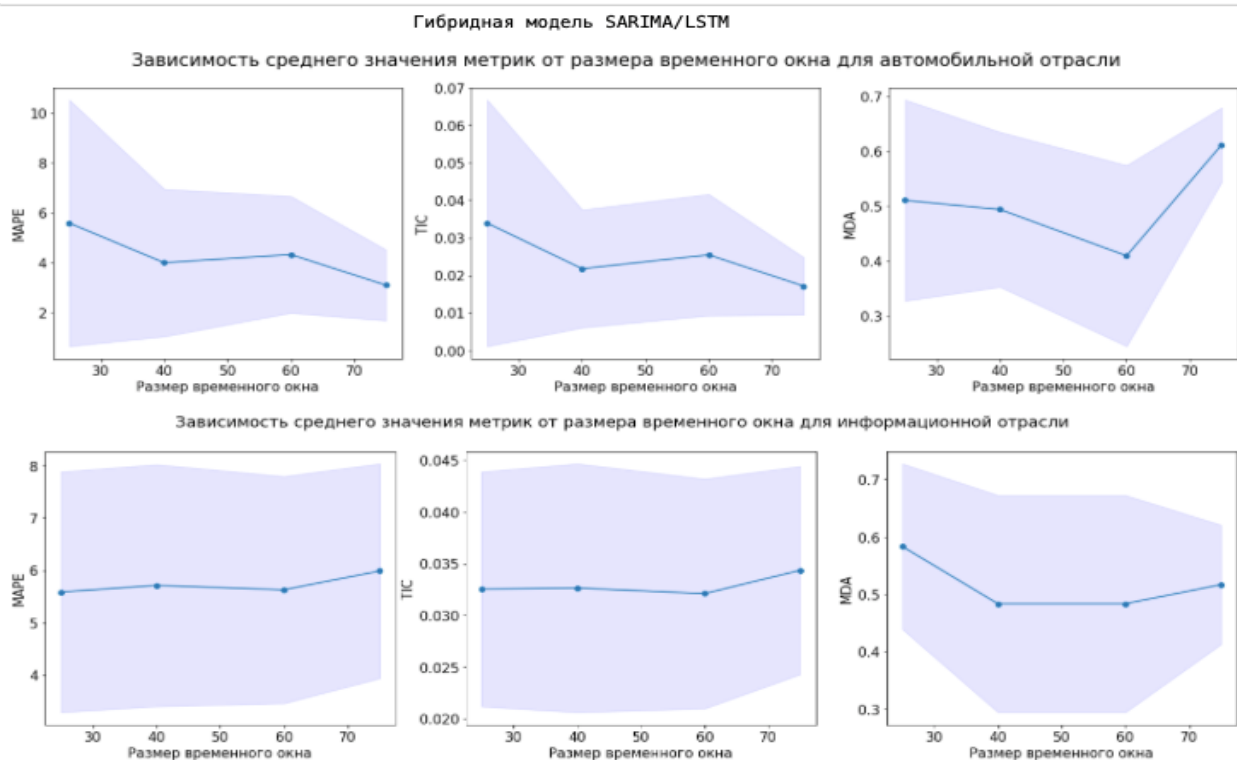


Рис. 2 Зависимость значения метрик качества от значения параметра временного окна для обеих отраслей гибридной модели SARIMA/LSTM.

В ряде работ по кластерам дается явная отсылка на цикл У-син (рис.6), например, у Г.Д. Боуш или неявно, в виде концепции «5И» М.П. Войнаренко, как необходимое условие создания кластеров.

Результаты проведенного эксперимента представлены в таблице 2.

Таблица 2. Результаты эксперимента для оптимизации размера временного окна при составлении выборки.

Отрасль	Модель	Размерность выборки
Автомобильная отрасль	SARIMA/MLP	60
	SARIMA/LSTM	75
Отрасль информационных технологий	SARIMA/MLP	75
	SARIMA/LSTM	25

Для определения оптимального исторического периода времени, на основе которого будет происходить обучение моделей и формироваться прогноз, был проведен эксперимент для каждой из отраслей в отдельности. Он заключался в анализе зависимости качества предсказания от размера обучающей выборки. В исследовании было рассмотрено 9 вариантов временных периодов от 2 лет до 6 лет с шагом в 6 месяцев.

В общем случае распознать основные тенденции для определения оптимального исторического периода времени достаточно сложно, однако метрика MAPE часто достигает своего локального и глобального минимума в точке, где значение исторического периода равно 3,5 годам и 6 годам (Рис.3). В тоже время, по показателю MDA, в данной точке модели также ведут себя достаточно стабильно и показывают хорошие результаты относительно других периодов. В связи с этим, прогноз был построен для обоих периодов. Тем не менее, результаты, полученные на основе данных за 6 лет, были в среднем хуже, чем за более короткий промежуток времени. Вероятно, это связано с усреднением прогноза статистических моделей с ростом обучающей выборки. В связи с этим, финальный прогноз был построен на основе временного горизонта в 3,5 года.

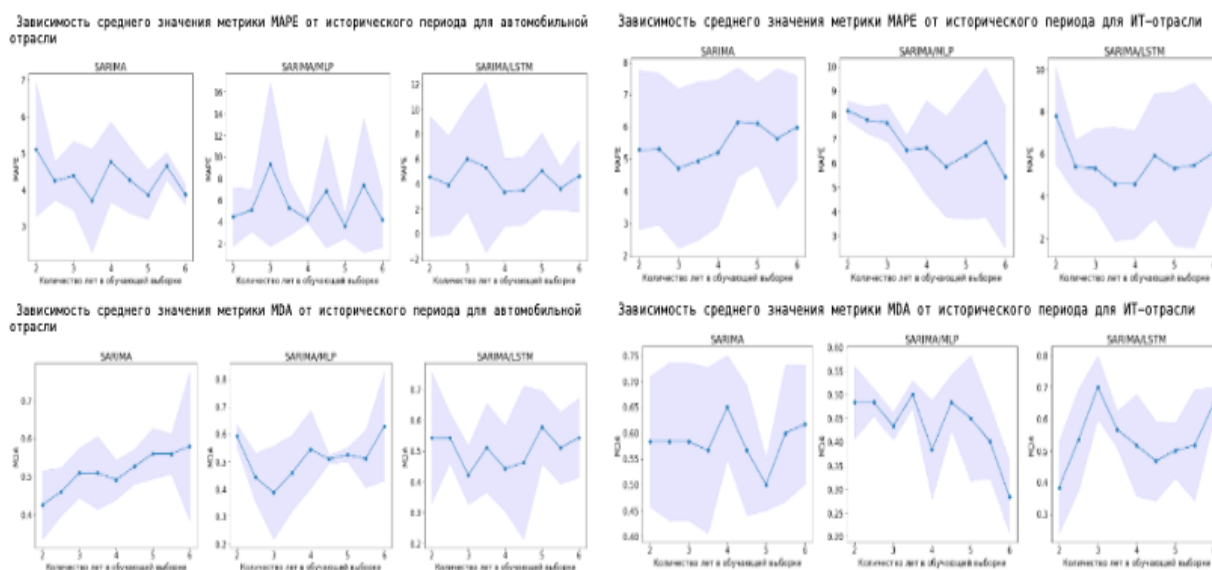


Рис. 3 Зависимость значения метрик качества от количества исторических данных для обучения для обеих отраслей.

Далее с помощью подготовленных моделей прогнозирования был сформирован прогноз для выбранных тридцати компаний автомобильной и информационной отраслей. Анализ полученных прогнозных данных показал, что использование гибридной модели, основанной на объединении статистическое модели SARIMA с рекуррентной нейронной сетью LSTM, позволило значительно повысить качество предсказания: приближенность к реальным значениям увеличилась и уменьшилась ошибочность распознавания направления движения котировок (Рис.4). Использование сети прямого распространения MLP в гибридной модели не показало целесообразности ее использования в сравнении с простой SARIMA, поскольку оно не ведет к значительному улучшению качества прогноза,

а в некоторых случаях ухудшает его. Оценка показателя MPE показала, что в среднем всем рассмотренным моделям свойственно занижение прогнозных значений.

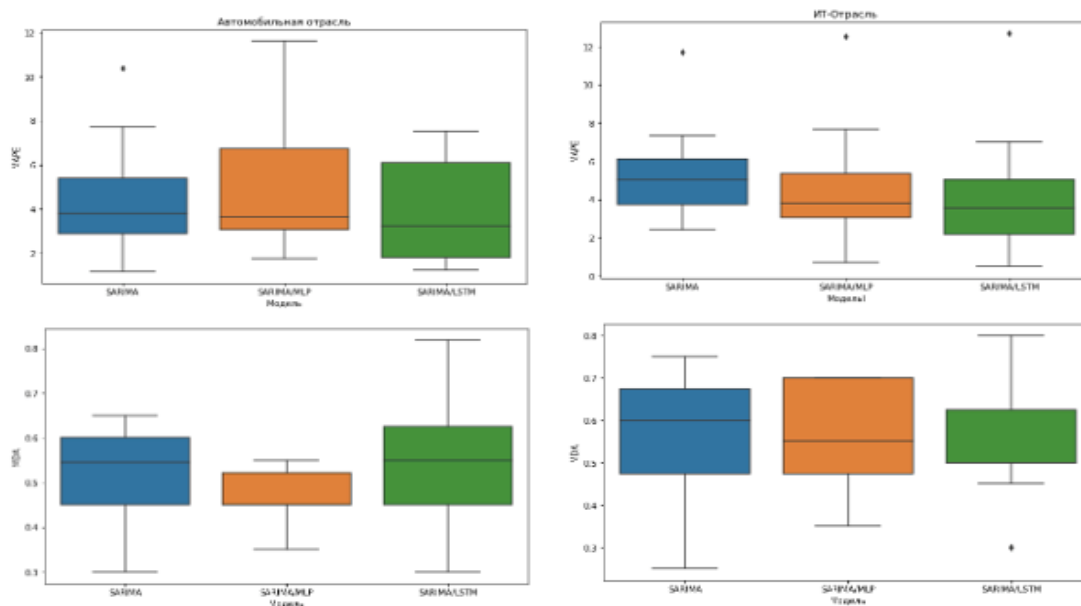


Рис. 4 Диаграммы распределения метрик качества прогноза MAPE и MDA, сгруппированных по моделям, для обеих отраслей

Пример прогноза для компании Peugeot для рассмотренных моделей приведен на рисунке 5.

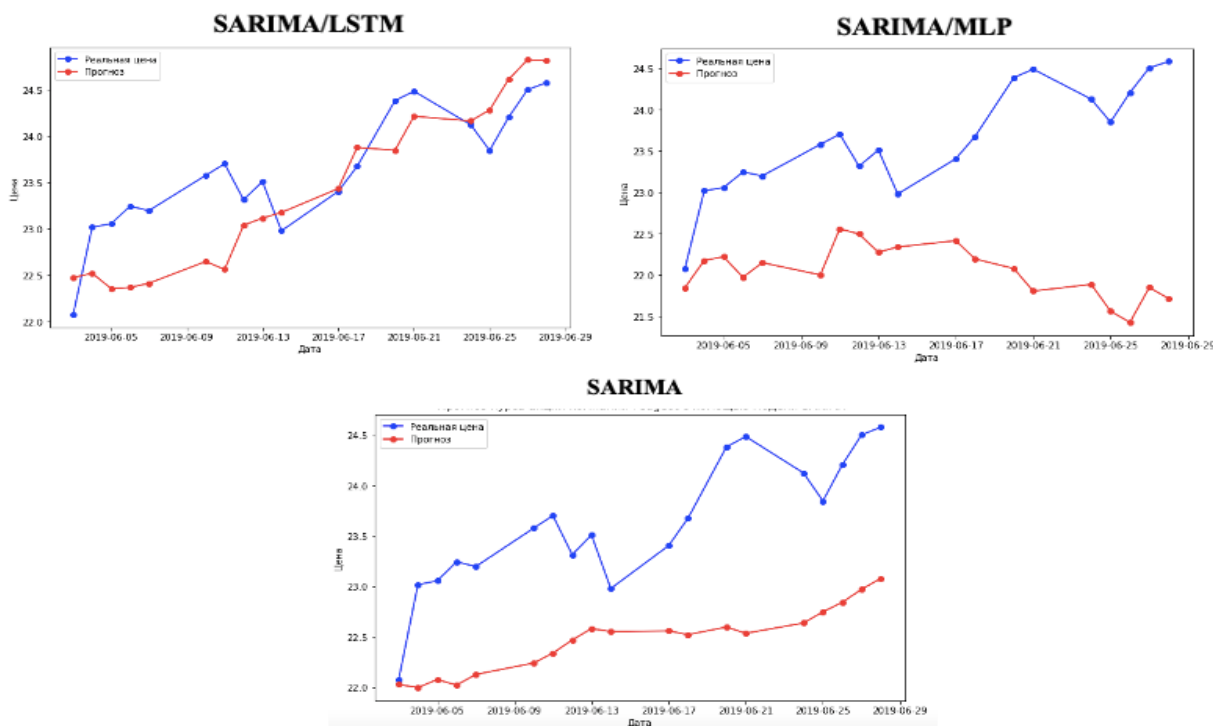


Рис. 5 Пример сформированного прогноза для компании Peugeot

Таким образом, анализ полученных прогнозных данных позволил получить исчерпывающее представление о прогностической способности каждого из методов и о том, как они соотносятся между собой. Оптимальным вариантом для обеих рассматриваемых отраслей является гибридная модель, объединяющая статистический метод и рекуррентную нейронную сеть, SARIMA/LSTM, которая позволила получить значительный прирост качества предсказания относительно простой статистической модели.

Заключение

Анализ прогностической способности рассматриваемых гибридных методов предсказания курса акций показал, что гибридная модель SARIMA/LSTM обладает большой силой предсказания как по показателям приближенности прогноза к реальной величине, так и по правильности определения направления движения котировок, в сравнении с классической статистической моделью SARIMA и гибридной SARIMA/MLP. Данный результат достигается за счет формирования прогноза из двух частей: стабильной и устойчивой линейной части, и добавления нелинейного компонента, отражающего высокую волатильность акций. Тем не менее, не стоит забывать, что прогнозирование курса акций только один из набора инструментов для эффективного оперирования на рынке ценных бумаг, поскольку акции являются высоко волатильным активом и требуют комплексного подхода.

Данная работа может быть полезна инвесторам и инвестиционным компаниям, поскольку она содержит выявление более точного метода прогнозирования котировок акций и для подготовки моделей. Кроме того, настоящее исследование может быть основой для дальнейшего анализа применимости и целесообразности использования результатов прогнозирования для составления и ребалансировки инвестиционного портфеля.

Литература

1. Hulbert on Markets: What's Working Now
2. Hulbert Financial Digest Monthly Newsletter
3. *Hulbert Mark*. The Past 17 Years and What They May Reveal About the Future // *AAII Journal*. June 1997. P. 31—33
4. *Blix Marten, Wadefjord Joachim, Wienecke Ulrika & Adahl Martin*. How Good Is the Forecasting Performance of Major Institutions // *Economic Review*. 2001. No. 3. P. 38—68
5. *Hushani P*. Using Autoregressive Modelling and Machine Learning for Stock Market Prediction and Trading // *Third International Congress on Information and Communication Technology*. – Springer, Singapore, 2019. – С. 767-774.
6. *Raoofi A., Zarra-Nezhad M., Bayani O*. Assessment and Comparison of linear and non-linear Methods for Forecasting Returns on Stock Market Index // *International Journal of Scientific Management and Development*. 2015. Vol.3 (10). p. 595-603.
7. *Ohanyan H*. Stock price forecast with deep learning LSTM and econometric ARIMA models : дис. – 2018.
8. *Hushani P*. Using Autoregressive Modelling and Machine Learning for Stock Market Prediction and Trading // *Third International Congress on Information and Communication Technology*. – Springer, Singapore, 2019. – С. 767-774.
9. *Ohanyan H*. Stock price forecast with deep learning LSTM and econometric ARIMA models : дис. – 2018.
10. *Jiang W*. Applications of deep learning in stock market prediction: recent progress // *arXiv preprint arXiv:2003.01859*. – 2020.