DOI: 10.25728/mlsd.2022.0303

ПРОГНОСТИЧЕСКАЯ СПОСОБНОСТЬ ГИБРИДНЫХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ КОТИРОВОК АКЦИЙ НА ПРИМЕРЕ *ARIMA/LSTM*

Сизых Н.В., Оршанская Е.С., Сизых Д.С.
Институт Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ),
Россия, г. Москва, ул. Мясницкая, д. 20
sizykh_n@mail.ru, orshanskayaelena@gmail.com, D.Sizykh@gmail.com

Аннотация: В статье представлены результаты исследования прогностической способности моделей прогнозирования котировок акций с помощью гибридной модели ARIMA/LSTM. Данное исследование основано на анализе прогнозирующей способности с использованием выборки, включающей 30 компаний из трех секторов: энергетический, финансовый и технологический.

Ключевые слова: методы и модели прогнозирования, прогнозирование временных рядов, прогнозирующая способность, котировки акций.

Введение

В последнее десятилетие в области прогнозирования курсов акций происходит смена концентрации внимания с традиционных статистических моделей на более сложные, в основе которых лежат алгоритмы машинного обучения и искусственного интеллекта. Одним из перспективных направлений развития современных методов 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. Данное положение поддерживается также существующими сервисами, предоставляющими услуги по автоматическому формированию прогноза для выбранной акции на необходимый промежуток времени.

Прогностическая способность методов прогнозирования 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. [12]. Точность прогнозирования значительно различается в зависимости от периода прогнозирования (гораздо ниже при смене фазы деловых циклов), от прогнозируемого индикатора и от горизонта прогнозирования (падает с ростом горизонта).

При оценке точности прогноза учитывается и специфика прогнозирования акций, которая состоит в сложность финансовых рынков, влияние на их динамику огромного количества факторов, репутационные риски для субъектов прогнозирования при существенных отклонениях прогноза от действительности. В настоящее время выделяют следующие особенности прогнозирования динамики биржевых индексов 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.

В проведенном исследовании было проанализировано влияние различных параметров модели выбранной гибридной модели прогнозирования на точность прогноза.

1 Гибридные модели прогнозирования

Рассмотрим особенности гибридных моделей прогнозирования и, в частности, выбранной модели *ARIMA/LSTM*, как одной из наиболее оптимальных для прогнозирования курсов акций, что было показано в ряде исследований за последние несколько лет [19].

Котировки акций представляют собой сложный нестационарный временной ряд, подверженный влиянию большого количества разнообразных факторов. В настоящее время существует множество различных методов, которые применяются для анализа и прогнозирования будущих значений цен акций. Исследователи Шах и другие [19] провели анализ наиболее распространенных методов, используемых для определения прогнозирования курса акций и предложили собственную классификацию. Авторы данной работы выделил следующие группы методов: статистические, распознавание образов, машинное обучение, анализ тональности текста и гибридные методы. В свою очередь, методы машинного обучения можно разделить на две подгруппы: алгоритмы классического

машинного обучения и нейронные сети.

Статистические методы основаны на статистических гипотезах о временных рядах, включающих в себя стационарность, линейность, соответствие нормальному распределению, которые могут не соблюдаться для такого волатильного ряда как цены акций. Эта группа методов позволяет определять относительно простые зависимости в данных, таких как сезонность или линейная зависимость от предыдущих значений и широко используется из-за своей стабильности и высокой интерпретируемости полученных результатов, пусть они и не являются самыми точными. Наиболее популярной среди статистических методов является модель авторегрессионного интегрированного скользящего среднего (ARIMA), предложенная Боксом и Дженкинсом в 1972 [17] и сочетающая в себе авторегрессионный подход (зависимость искомой переменной от отдельных ее предыдущих значений) и метод скользящего среднего (зависимость переменной от средних по наборам предыдущих значений). Однако ряд исследований на финансовых данных (Алжеев и Кочкаров, 2020 [4]; Адебийи и другие [2], Хашани [8]) доказывает значительное превосходство точности нелинейных моделей. Добавление к ARIMA сезонного тренда (SARIMA), позволяет несколько повысить точность (Бодриа [6] на данных индекса DZAIRINDEX), однако все еще не помогает выявлять сложные нелинейные зависимости, характерные для финансовых показателей.

Наиболее точно выявлять неочевидные и нелинейные закономерности в данных позволяют модели нейронных сетей, построенные по принципу работы нервной системы человека и отличающиеся от моделей классического машинного обучения наличием нелинейных функций перехода. Так как предсказание будущих значений временного ряда строится на его же предыдущих точках, наибольшее распространение получили рекуррентные нейронные сети (RNN), в частности нейронная сеть с долгой краткосрочной памятью (LSTM). Архитектура LSTM была предложена Хохрайтером и Шмидхубером в 1997 году [22], но данная модель остается актуальной и в настоящее время. Это подтверждается, например, исследованием эффективности LSTM для прогнозирования цен акций индийских IT-компаний Infosys и Zensar (февраль 2021: Кумар, Гандмал [14]).

Среди других гибридных моделей наиболее распространено сочетание статистического метода и нейронной сети. Используя этот подход и данные о курсах акций банка, Абдулрахман и др. [1] заявили, что комбинация ARIMA и LSTM может уменьшить ошибку прогноза в 95 раз по сравнению с просто LSTM. Такая модель не только точнее прогнозирует цены акций, но и чаще угадывает направление динамики переменной, что наиболее важно для принятия инвестиционных решений.

В общем случае гибридные модели прогнозирования объединяют в себе различные группы методов. Такой подход позволяет отдельно предсказывать части переменной с различными распределениями или разделять прогнозирование на несколько этапов для обнаружения на каждом из них новых зависимостей в данных. Помимо этого, качество прогноза в значительной степени зависит от настроек модели, поэтому дополнительные алгоритмы могут использоваться для оптимизации параметров предсказывающей модели.

2 Модель ARIMA/LSTM

Для анализа влияния временных факторов на качество прогноза котировок акций была выбрана модель ARIMA/LSTM [9]. Авторами было предложено декомпозировать ряд котировок акций, при помощи подходящего фильтра скользящего среднего, на две части и предсказывать их соответствующей моделью. Первая часть l_t с меньшей волатильностью получается в результате сглаживания фильтром скользящего среднего, который подбирается таким образом, чтобы получившийся после сглаживания ряд подчинялся нормальному закону.

Для получения второй части g_t с более высокой волатильностью из исходного ряда вычитается сглаженный: $g_t = y_t - l_t$. Далее с помощью ARIMA с оптимально подобранными параметрами делается прогноз менее изменчивой части $\widehat{l_t}$, а с помощью простой рекуррентной нейронной сети более изменчивой $\widehat{g_t}$. Итоговое предсказание является результатом сложения полученных данных моделей $\widehat{y_t} = \widehat{l_t} + \widehat{g_t}$.

Метрики качества. Для оценки качества прогноза были выбраны метрики MAPE и MDA. Первая для оценки близости прогноза к реальности, а вторая для оценки способности моделей распознавать направление движения котировок акций.

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

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

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

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

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

3 Исходные данные и подготовка моделей для прогнозирования

Для проведения анализа было выбрано 30 компаний, представленных на Нью-Йоркской фондовой бирже, каждая из которых оперирует в одном из трех секторов: энергетический, финансовый и технологический. Для исследования был выбран отрезок времени с 01/12/2014 по 01/01/2020, где декабрь 2019 года используется для формирования прогнозов.

В качестве временных показателей для исследования их влияния на метрики качества были выбраны такие параметры, как исторический период исходных данных и временное скользящее окно. Оценка проводилась на разных периодах прогнозирования. Таким образом, в работе были исследованы следующие значения параметров:

- скользящее окно: 10 (2 недели), 20 (около месяца), 40 (около 2 месяцев), 60 (около 3 месяцев) и 80 (около 4 месяцев);
- Исторический период: от 2 до 5 лет с шагом в полгода;
- Период прогнозирования: 1 день, 1 неделю, 2 недели и 1 месяц.

Подготовка моделей. Так как качество прогнозов в значительной степени зависит от входных данных, необходимо также подобрать оптимальный набор данных для обучения и тестирования.

Обучающая выборка не должна быть менее года, чтобы учесть сезонные изменения цен акций и не должна быть слишком большой, чтобы избежать переобучения моделей и чрезмерных затрат временных и вычислительных ресурсов. Кроме того, влияние на качество оказывает размер движущегося окна, соответствующий количеству предыдущих записей, необходимых для расчета прогнозируемой цены закрытия.

Фильтр скользящего среднего и соответствующее окно сглаживания выбирались автоматически для каждого временного ряда путем сравнения коэффициента эксцесса получившегося ряда с критическим значением 3. В связи с тем, что 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. В исследовании были использованы фильтры SMA, EMA, WMA, KAMA, TEMA, TMA с окном от 4 до 100.

Выборка для обучения ARIMA была предварительно логарифмирована. Параметры для моделей были подобраны автоматически путем минимизации RMSE. Для нейронной сети LSTM была выбрана архитектура с одним скрытым слоем и 1024 нейронами на основе опыта предыдущего исследования. С целью уменьшения проблемы переобучения также был добавлен слой DropOut. Обучение модели проходило на 100 эпохах, входные данные были предварительно нормированы. Формирование прогноза части LSTM было реализовано с использованием рекурсивной стратегии, т.е. модели пошагово дообучались на каждом своем последующем предсказании. Перед формированием конечного прогноза выводы моделей были приведены к исходному масштабу.

4 Практические результаты оценки влияния временных параметров на качество прогноза

Оценка качества прогноза и его зависимости от рассматриваемых временных показателей была проведена на основе корреляционного анализа (Таблица 1) и анализа распределения значений в целом с помощью диаграмм типа Box Plot и Bar Plot (для исследования показателя MDA на периоде прогноза

в один день) (см. Рисунок 1-5). Секторы были исследованы в отдельности и совокупности. Для более компактного визуального представления переменная “Исторический период” был приведен к порядковой шкале с рангами “Меньшие периоды” (2-3 года), “Средний период” (3,5 года) и “Большие периоды” (4-5 лет). Корреляционный анализ был проведен для исходной шкалы.

При совместном рассмотрении прогнозных данных по акциям всех секторов выбранных компаний для исследуемых сроков прогнозов наблюдается обратная зависимость значения метрики MAPE и длины исторического периода, значимые отрицательные коэффициенты корреляции это подтверждают. В разрезе временных окон заметен рост ошибки с увеличением значения параметра, между показателями наблюдается прямая зависимость. По метрике MDA визуально заметна слабая прямая связь с длиной исторического периода, коэффициенты корреляции значимы. Большим историческим периодам чаще свойственна более высокая точность и меньший размах значений для всех сроков прогнозов. В разрезе временных окон худшие результаты соответствуют наибольшему окну 80, разница у остальных не велика. Значимые коэффициенты корреляции в данном случае не превосходят 0.1 по абсолютному значению, явная зависимость не наблюдается.

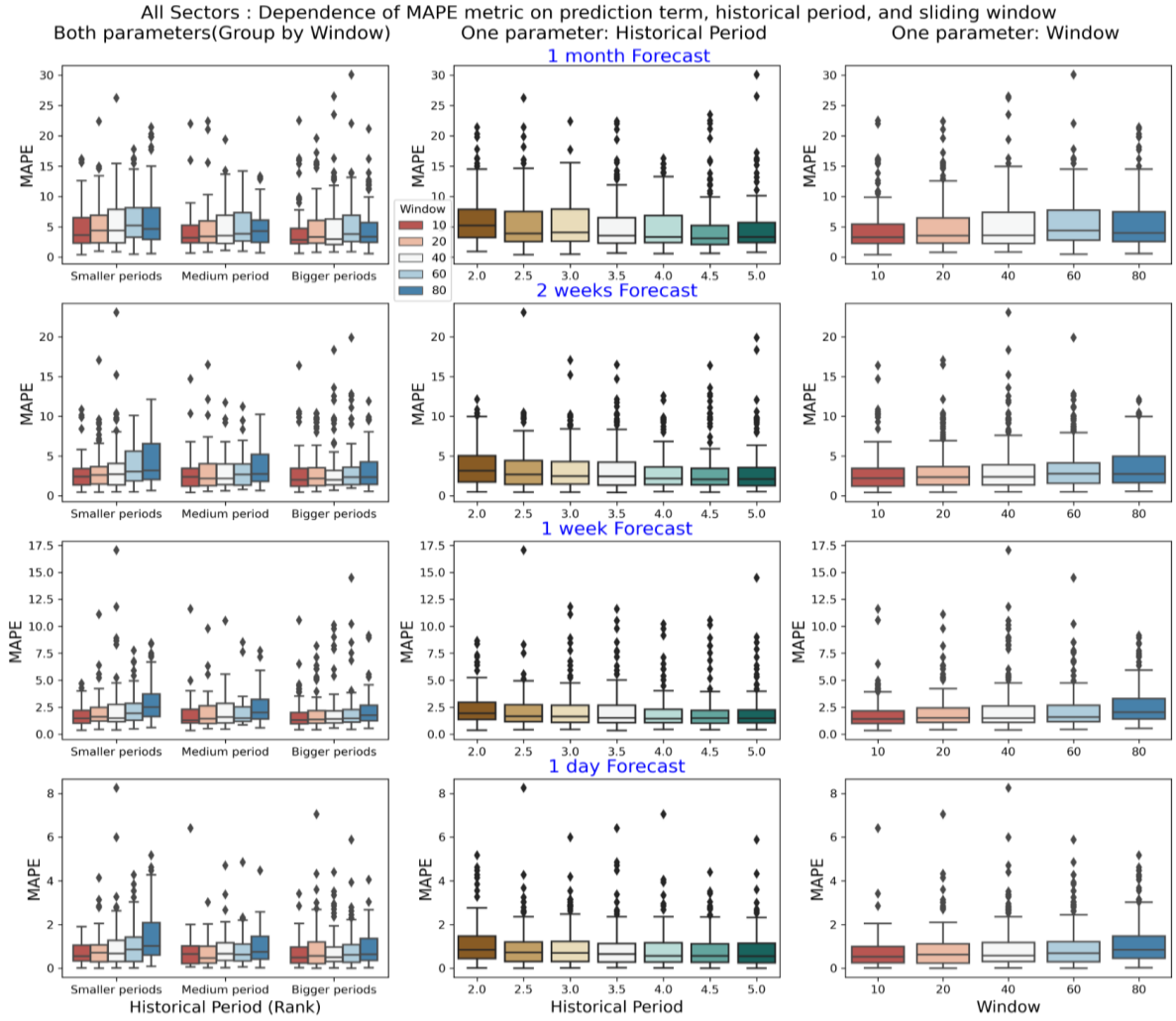
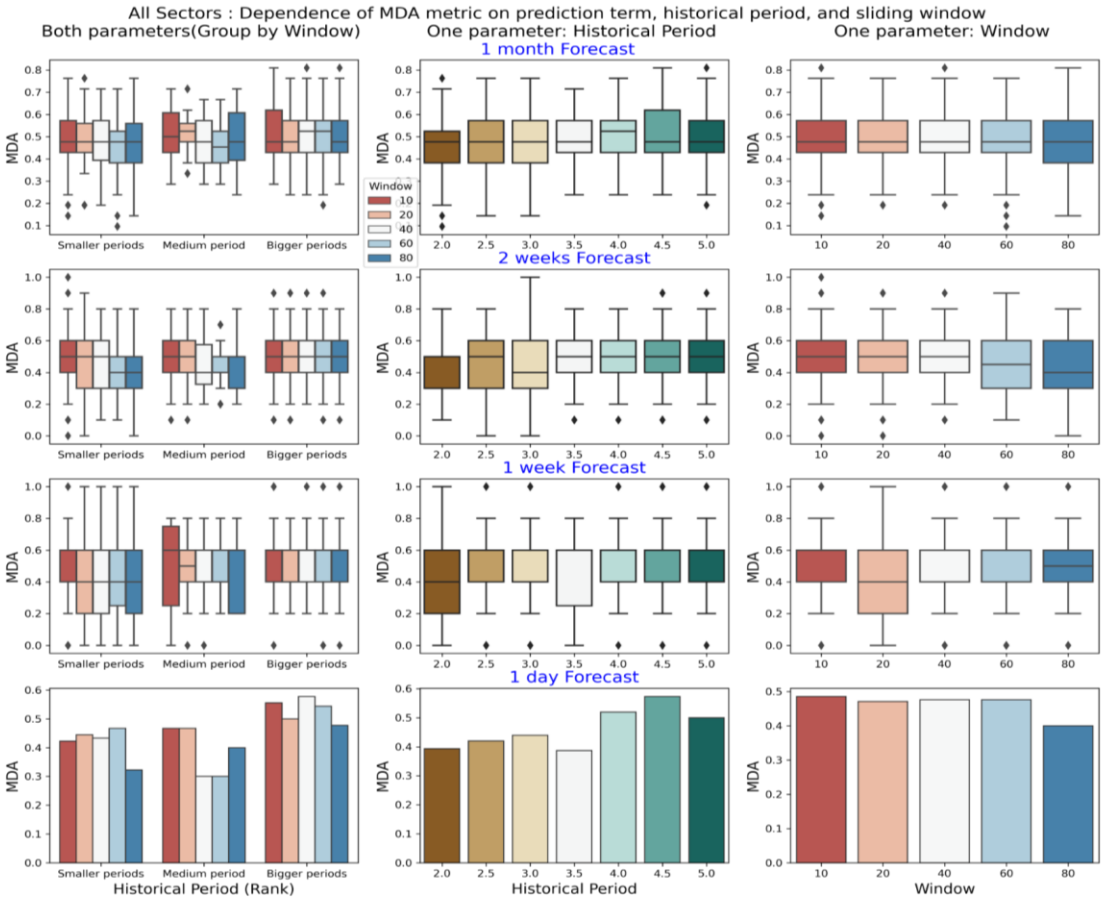


Рис. 1. Совместное рассмотрение секторов. Зависимость значения MAPE от временных показателей



.

Рис. 2. Совместное рассмотрение секторов. Зависимость значения MDA от временных показателей

В энергетическом секторе в разрезе исторических периодов также визуально наблюдается обратная зависимость значения метрики MAPE и исторического периода, что и подтверждают коэффициенты корреляции. Визуально при рассмотрении временных окон наблюдается явная прямая зависимость между размером параметра и значением процентной ошибки на всех периодах прогнозирования, однако статистически это подтверждается только для прогнозов в одну и две недели. По метрике MDA просматривается прямая связь с длиной исторического периода, во временных же окнах наблюдается изменчивая ситуация за исключением окна 40 - оно показало стабильный высокий результат на всех сроках прогноза. Тем не менее, при совместном рассмотрении параметров исторического периода и временного окна для прогнозов на месяц и две недели заметно, что меньшее окно 10 ведет себя лучше на больших временных промежутках.

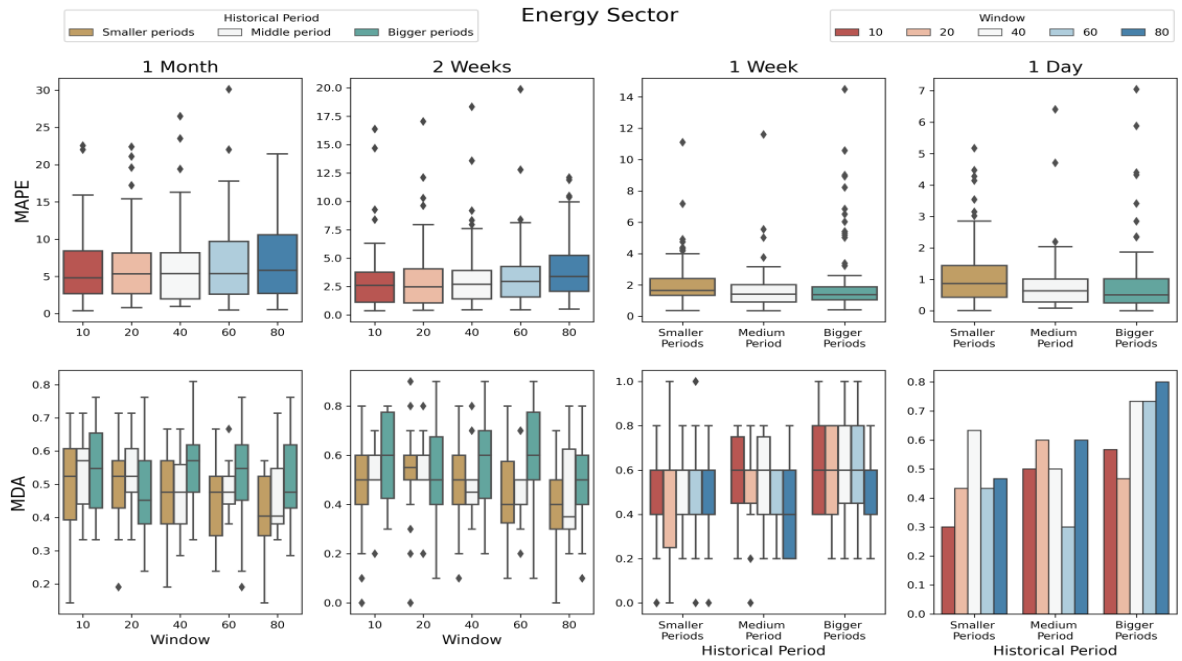


Рис. 3. Энергетический сектор. Зависимость значения метрик MAPE и MDA от временных показателей

В финансовом секторе для всех сроков прогнозов наблюдается обратная зависимость метрики MAPE и длины исторического периода, а с размером временного окна - прямая взаимосвязь. Это подтверждается статистически для всех случаев кроме временного окна при прогнозе на месяц. По метрике MDA более высокие показатели соответствуют меньшим временным окнам (10-20) для всех сроков прогноза, за исключением недельного, где значительно отличалось от других окно размером 80. Линейная связь между показателями отсутствует.

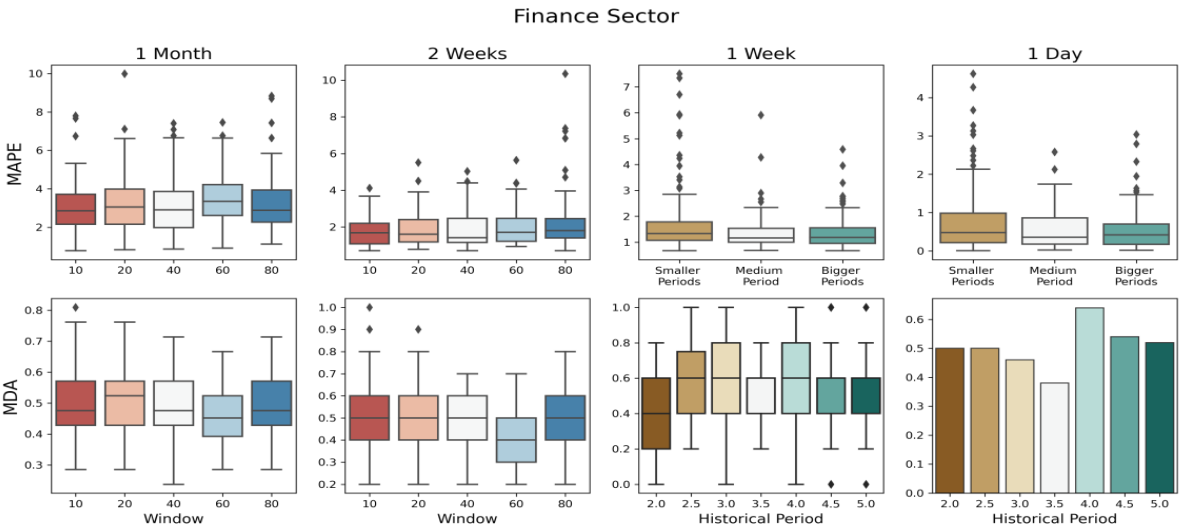


Рис. 4. Финансовый сектор. Зависимость значения метрик MAPE и MDA от временных показателей

В технологическом секторе для всех сроков прогноза по метрике MAPE наблюдается прямая зависимость с размером временного окна, наименьшему соответствует значительно меньшая ошибка, все коэффициенты корреляции значимы. В разрезе исторических периодов не наблюдается линейная зависимость показателей, хорошее качество наблюдается при меньших исторических периодах с использованием меньших окон. Возможно, это связано с достаточно быстрым развитием технологического сектора. Большее влияние на котировки акций соответствующих компаний оказывают наиболее близкие по времени события, а слишком давние данные избыточны и не соответствуют текущим тенденциям. Для однодневного прогноза по показателю MDA в разрезе временных окон лучший результат также соответствует наименьшему окну, а в исторических периодах

– 4,5. Для прогнозов в две недели и один месяц наблюдается прямая связь между показателями, а меньшим историческим промежуткам свойственен больший разброс значений.

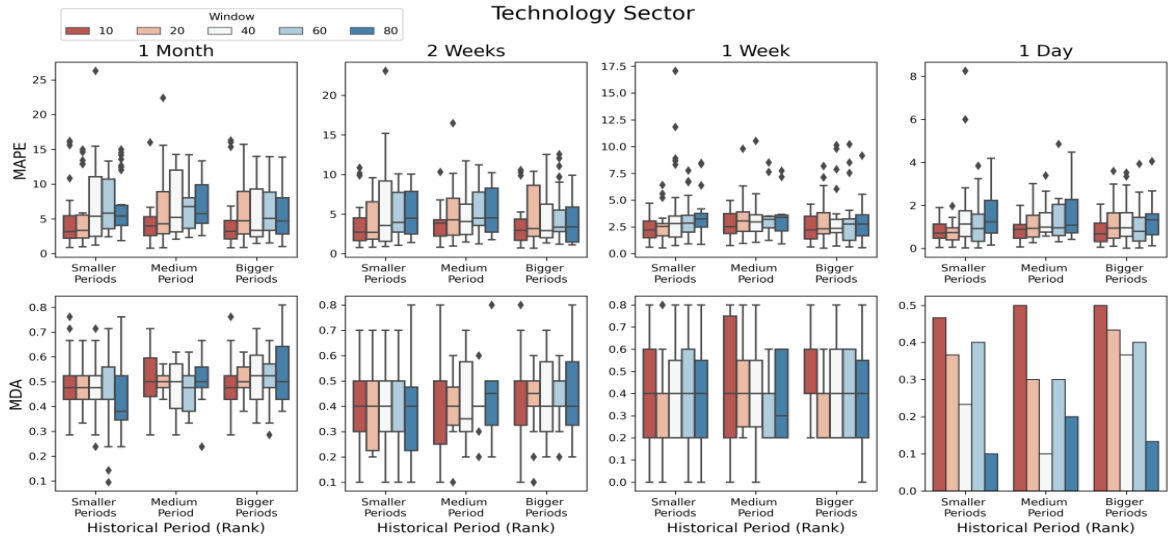


Рис. 5. Технологический сектор. Зависимость значения метрик MAPE и MDA от временных показателей

Таблица 1. Показатели качества прогноза в зависимости от исторического периода и размера окна на периоды: 1 день и 1 неделя

Forecast Period	Parameter		Sector							
			All Sectors		Energy Sector		Finance Sector		Technology Sector	
			Mean value of metric							
			MAPE	MDA	MAPE	MDA	MAPE	MDA	MAPE	MDA
1 Week	Window	10	1,796	0,503	1,655	0,534	1,271	0,549	2,461	0,426
		20	2,059	0,452	1,800	0,520	1,401	0,474	2,975	0,363
		40	2,244	0,489	1,817	0,566	1,426	0,503	3,489	0,397
		60	2,199	0,487	1,996	0,537	1,447	0,514	3,153	0,409
		80	2,546	0,490	2,302	0,486	2,151	0,606	3,186	0,377
	Historical Period	2	2,407	0,425	2,339	0,472	1,873	0,412	3,008	0,392
		2.5	2,154	0,485	1,743	0,500	1,743	0,568	2,977	0,388
		3	2,228	0,459	1,911	0,464	1,569	0,544	3,205	0,368
		3.5	2,254	0,480	1,802	0,496	1,488	0,544	3,471	0,400
		4	2,006	0,524	1,706	0,580	1,291	0,592	3,022	0,400
		4.5	2,053	0,521	1,999	0,616	1,418	0,540	2,742	0,408
		5	2,078	0,493	1,897	0,572	1,393	0,504	2,944	0,404
1 Day	Window	10	0,711	0,486	0,847	0,443	0,470	0,529	0,815	0,486
		20	0,838	0,471	0,957	0,471	0,562	0,557	0,995	0,386
		40	0,918	0,476	0,772	0,657	0,582	0,500	1,401	0,271

		60	0,944	0,476	0,976	0,543	0,663	0,500	1,193	0,386
		80	1,141	0,400	1,061	0,629	0,942	0,443	1,421	0,129
	Historical Period	2	1,101	0,393	1,294	0,320	0,928	0,500	1,079	0,360
		2.5	0,939	0,420	1,008	0,480	0,682	0,500	1,127	0,280
		3	0,955	0,440	0,893	0,560	0,693	0,460	1,280	0,300
		3.5	0,921	0,387	0,882	0,500	0,571	0,380	1,310	0,280
		4	0,847	0,520	0,848	0,580	0,505	0,640	1,187	0,340
		4.5	0,815	0,573	0,778	0,760	0,611	0,540	1,056	0,420
		5	0,795	0,500	0,754	0,640	0,516	0,520	1,114	0,340

Помимо объема выборки, значительную роль играет структура данных: предсказание цены закрытия акций компании может основываться как исключительно на ценах закрытия предыдущих дней, так и на дополнительных данных, таких как цена открытия, максимальная и минимальная за день, а также объем проданных за день акций. Было установлено, что в случае гибридной модели дополнительные признаки позволяют значительно повысить качество прогнозирования.

Заключение

Проведенный анализ позволил выявить тренд к усложнению методов прогнозирования и популярности исследований, посвященных новым видам гибридных моделей. Выявлена недостаточность исследований, в которых сравниваются между собой различные типы гибридных моделей и анализируются их оптимальные параметры.

Посредством проведения экспериментов подобраны оптимальные общие параметры и архитектура модели. Таким образом, исходя из проведенного анализа можно рекомендовать следующее: для более точного предсказания с помощью гибридной модели ARIMA/LSTM в случае относительно стабильных рынков преимущественно использовать временной период не менее 3-3,5 лет. Если же рассматриваемый сектор подвержен частым сильным изменениям, как технологический, то стоит рассматривать вариант прогнозирования с меньшими длинами исторического периода, например, в 2 года. Исследование временных окон показало, что для достижения хороших результатов достаточно использовать небольшой размер окна, который содержит 10 предыдущих значений (2 недели). Его сильное увеличение может оказать негативное влияние на точность прогноза.

Не было установлено существование зависимости между сложностью гибридной модели (количеством составляющих модулей) и точностью ее прогноза. Но было установлено, что подобранные оптимальные показатели исторического периода, движущего окна позволяют снизить среднюю по рассмотренных компаниям ошибку прогнозирования MAPE на 15%.

Данное исследование может быть полезно как для управляющих компаний, так и для частных инвесторов. Учитывая наличие в настоящее время различных программных сервисов для прогнозирования, в частности, для котировок акций, процесс инвестиционных решений становится более доступным для инвесторов любого уровня.

Литература

1. *Abdulrahman1, U. F. I., Ussiph, N., & Hayfron-Acquah, H.* A Hybrid Arima-Lstm Model for Stock Price Prediction. International Journal of Computer Engineering and Information Technology. – 2020. 12, 48–51.
2. *Adebiyi A. A., Adewumi A. O., Ayo C. K.* Comparison of ARIMA and artificial neural networks models for stock price prediction //Journal of Applied Mathematics. – 2014. – Т. 2014.
3. *Agrawal R. K., Adhikari R.* An introductory study on time series modeling and forecasting //Nova York: CoRR. – 2013.
4. *Alzheev A. V., Kochkarov R. A.* Comparative analysis of ARIMA and lSTM predictive models: Evidence from Russian stocks //Finance: Theory & Practice. – 2020. – Т. 24. – №. 1. – С. 14-23.
5. *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
6. *Boudrioua M. S.* Modeling and forecasting the Algiers Stock Exchange returns using the Box-Jenkins methodology. – 2019.

7. *Hank D.* Business Forecasting 7th (seventh) Edition. – 2003.
8. *Hochreiter S., Schmidhuber J.* Long short-term memory //Neural computation. – 1997. – Т. 9. – №. 8. – С. 1735-1780.
9. *Hulbert Mark* The Past 17 Years and What They May Reveal About the Future // AAII Journal. June 1997. P. 31—33
10. *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.
11. Investpy Documentation [E-resource] // Investpy. URL: https://investpy.readthedocs.io/ (application date: 01.04.2020)
12. *Jiang W.* Applications of deep learning in stock market prediction: recent progress //arXiv preprint arXiv:2003.01859. – 2020.
13. *Kambil Ajit* CFO insights: Social analytics: Tapping prediction markets for foresight. Deloitte Development LLC. 2010.
14. *Kumar, K., & P. Gandhmal, D.* (2021). An intelligent indian stock market forecasting system using LSTM deep learning. Indonesian Journal of Electrical Engineering and Computer Science, 21(2), 1082.
15. *Marten Blix, Joachim Wadefjord, Ulrika Wienecke & Martin Adahl* How Good Is the Forecasting Performance of Major Institutions // Economic Review. 2001. No. 3. P. 38—68
16. *Ohanyan H.* Stock price forecast with deep learning LSTM and econometric ARIMA models: dis. – 2018.
17. *Rao J. N. K.* Time Series Analysis Forecasting and Control. – 1972.
18. *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.
19. *Shah D., Isah H., Zulkernine F.* Stock Market Analysis: A Review and Taxonomy of Prediction Techniques //International Journal of Financial Studies. – 2019. – Т. 7. – №. 2. – С. 26.
20. *Shui-Ling Y. U., Li Z.* Stock Price Prediction Based on ARIMA-RNN Combined Model // 4th International Conference on Social Science (ICSS 2017). 2017
21. *Sizykh, N., & Orshanskaya, E.* (2020, September). Analysis of the Prognostic Ability of Methods for Forecasting Share Quotations Dynamics. In 2020 13th International Conference" Management of large-scale system development"(MLSD) (pp. 1-3). IEEE.
22. *Snowberg Erik, Wolfers Justin & Zitzewitz Eric* Prediction Markets for Economic Forecasting. NBER, Working Paper No. 18222. July 2012.
23. *Spann Martin & Skiera Bernd* Internet-Based Virtual Stock Markets for Business Forecasting // Management Science. Vol. 49. Issue 10. October 2003. P. 1310—1326.