E

УДК 681.518.25

Т. В. Ермоленко, Д. В. Попадин, В. Н. Котенко
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Донецкий государственный университет»
283001, Донецкая Народная Республика, г. Донецк, ул. Университетская, 24

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ В ПРОГНОЗИРОВАНИИ ФОНДОВОГО РЫНКА

T. V. Yermolenko, D. V. Popadin, V. N. Kotenko
Federal State Budgetary Educational Institution of Higher Education "Donetsk State University"
283001, Donetsk People's Republic, Donetsk, University st, 24

APPLICATION OF MACHINE LEARNING IN STOCK MARKET FORECASTING

В статье описаны особенности временных рядов финансовых данных, основные трудности их прогнозирования и методы их обработки и структурирования для получения прогноза. Основная проблема – проблема стационарности и наличия памяти: преобразования, приводящие к стационарности ряда, удаляют долгосрочную память. Для ее решения применяют дробное дифференцирование. Можно выделить три класса ценовых трендов: восходящий, нисходящий и боковой. С учетом высокой волатильности рынка трансформированные данные маркируют по этим классам с помощью динамических порогов доходности. При принятии решения о ставке авторами предлагается использовать две модели машинного обучения: одна для маркировки движения цены по методу трех барьеров, вторая вычисляет коэффициент уверенности для проставленных маркеров двухступенчатым методом Лопеса Де Прадо.

Ключевые слова: финансовый временной ряд, бары, японские свечи, тиковый график, дробное дифференцирование, метод тройного барьера, двухступенчатая маркировка Лопеса де Прадо.

The article describes the features of time series of financial data, the main difficulties in their forecasting and methods for processing and structuring them to obtain a forecast. The main problem is the problem of stationarity and the availability of memory: transformations leading to stationarity of the series remove long-term memory. To solve it, fractional differentiation is used. Three classes of price trends can be distinguished: ascending, descending and sideways. Given the high market volatility, the transformed data is labeled for these three classes using dynamic yield thresholds. When making a decision on a rate, the authors propose to use two machine learning models: one for marking the price movement using the three-barrier method, the second calculates the confidence coefficient for the marked markers using the Lopez De Prado two-stage method.

Key words: financial time series, bars, Japanese candlesticks, tick chart, fractional differentiation, triple barrier method, Lopez de Prado two-stage labeling.

Введение

Задача прогнозирования фондового рынка состоит в определении будущей стоимости акций компании или другого товара, торгуемого на бирже. Вполне очевидно, что качественное прогнозирование будущей цены акций способствует получению значительной прибыли. Необходимость анализа фондового рынка у профессиональных инвестиционных институтов рынка сегодня не вызывает сомнений: удачно предсказанная восходящая тенденция изменений цены может в итоге принести прибыль компании, а нисходящая – спасти от убытков. В связи с чем практически в любой инвестиционной компании и коммерческом банке есть своя аналитическая служба.

Задача прогнозирования роста цен и акций на рынке, решаемая аналитиками, является одной из самых сложных в экономике в силу наличия значительного количества факторов, способных повлиять на изменение рыночной курсовой стоимости ценных бумаг [1]. В процессе ее решения менеджеры вынуждены учитывать большое количество противоречивых и неоднозначных данных, важность которых невозможно оценить объективно. Во всех подобных случаях принимаемые решения носят весьма субъективный характер и, как все интуитивные решения, не могут быть корректно объяснены [2], [3].

Машинное и глубокое обучение стали новой эффективной стратегией трейдинга, которую для увеличения доходов используют многие инвестиционные компании. Изменения в сфере финансов происходят нелинейно [4], и иногда может показаться, что цены на акции формируются совершенно случайным образом, поэтому очевидно преимущество нейронных сетей, показавших высокую эффективность в решении задачи обобщения и выделения скрытых зависимостей между данными, а также способности на их основе прогнозировать или классифицировать новые данные [5].

Однако появляется новая проблема, так как данные должны быть специально структурированы определённым образом. Именно это и является одной из главнейших текущих проблем, которые влияют на успешность прогнозов, полученных с помощью моделей машинного обучения [6], [7]. В данной работе проводится анализ особенностей данных фондового рынка и определения необходимых рекомендаций к их трансформации и структурированию, которые позволят эффективно применить методы машинного обучения для получения прогноза.

Проблемы, возникающие при прогнозировании фондового рынка

Для прогнозирования цен, спроса или предложения чаще всего используется определённый временной ряд, затем строится модель распределения и описываются данные. Одной из распространенных и часто применяемых в контексте прогнозирования временных рядов является модель авторегрессионной скользящей средней (ARMA) и ее обобщение – модель авторегрессионной интегрированной скользящей средней (ARIMA) [8-10]. Эффективность таких методов требует того, чтобы ряд был предварительно обработан и приведён к стационарности, т.е. его среднее значение не должно изменяться, чего нельзя гарантировать для реальных данных рынка, в которых присутствуют длинные истории, смещающие смещают среднюю цену во времени. Финансовые временные ряды обладают высокой волатильностью, которая является показателем, отражающим активность рынка – динамику изменения цены.

В таких рядах периоды колебаний чередуются с периодами относительного спокойствия, поэтому для анализа финансовых рядов используется обобщенная модель авторегрессионной условной гетероскедастичности (GARCH) [11], в которой дисперсия ошибки моделируется как функция от времени. Помимо наличия нестационарности, есть и другие проблемы ценовых диаграмм: данные на рынках изначально не структурированы, очень низкое отношение сигнала к шуму, удаление памяти временных рядов [12].

Финансовые данные чрезвычайно зашумлены и подвержены разрежённости и выбросам, что вызвано большим количеством участников на ликвидных рынках, работающих с различными инвестиционными горизонтами. Однако существующие модели не включают методы устранения шума. В результате большинство эконометрических исследований приходят к ложным выводам, подкрепленным шумом, а не сигналом.

Предсказательные и классифицирующие модели особенно чувствительны к наличию выбросов [13], [14]. Даже небольшой процент выбросов может вызвать очень большой процент ошибок модели: покупки, которые должны быть продажами (ложноположительные результаты), и продажи, которые должны быть покупками (ложноотрицательные результаты).

Корреляция является полезной мерой линейной взаимозависимости, однако в случае нелинейных взаимозависимостей зависимостей, каковые наблюдаются в финансовых данных, она бесполезна. Кроме того, на значение корреляции сильно влияют выбросы, что делает использование корреляции для выявления закономерностей в данных весьма ограниченным.

Статистические тесты эффективности прогнозирования модели обычно проводятся путем разделения заданного набора данных на обучающую выборку, используемую для выбора модели и оценки её гиперпараметров, и тестовую выборку, на данных которой модель не обучалась, и которая используется для оценки эффективности обобщающей способности модели. Модель, полученная в результате обучения на данных, в которых имеется корреляция и выбросы, имеет низкое качество прогнозирования.

Еще одной проблемой при исследовании финансовых рядов является проблема стационарности и наличия памяти [15], [16]. В ценовых рядах каждое значение зависит от длинной истории ценовых уровней, следовательно, они имеют память. Исследователи в большинстве своем анализируют изменение цены, т.е. работают с приращениями цен (или логарифмами приращений), изменениями доходностей или волатильности. Подобные трансформации временного ряда делают временной ряд стационарным, но при этом удаляют всю память из ценовых последовательностей, потому что строятся на ограниченной длине окна. Долгосрочная зависимость (долгая память или долгосрочная персистентность) является основой для предсказательных свойств моделей, поэтому модель должна содержать некоторое количество памяти для оценки того, как далеко цена удалилась от своего ожидаемого значения.

Таким образом, для действительно успешного применения различных моделей машинного обучения в задаче анализа фондового рынка нужно в первую очередь уделить внимание разведочному анализу данных и их структурированию, а также трансформации временного ряда таким образом, чтобы сделать его стационарным, но сохранить максимально возможное количество памяти. Именно качество данных и их преобразование в конечном итоге определяет качество результата [17].

Методы визуализации и анализа динамики цен в задаче прогнозирования фондового рынка

Для визуализации и анализа динамики цен используется такой инструмент биржевой торговли как технический анализ – метод прогнозирования изменения цены с помощью анализа ее графика за предыдущий период времени.

Все изменения рыночной цены отображаются в виде графиков, к основным видам которых относятся: линейный график, график баров, свечной график, тиковый график.

Самым простейшим видом графика является линейный (зонный), который чаще всего строится по ценам закрытия каждого периода и представляет собой сплошную ломаную линию. Он показывает общее направление движения цены (тренд).

Графики баров строятся по четырём одинаковым точкам интереса внутри рассматриваемого временного отрезка (времени экспирации). Бары иногда называют OHLC-графиком, так как он строится по четырём точкам интереса:

– открытие *(open*) – значение цены в начале тайм-фрейма, которая отображается в виде горизонтальной черты слева от бара;
– максимум (*high*) – максимальное ценовое значение, которое было достигнуто внутри тайм-фрейма;
– минимум (*low*) – минимальное ценовое значение, которое было достигнуто внутри тайм-фрейма;
– закрытие (*close*) – значение цены в конце тайм-фрейма, которое отображается в виде горизонтальной черты справа от бара.

Бары оставались самым популярным типом отображения графиков цен вплоть до 90-х гг., им на смену пришли японские свечи, с помощью которых было удобнее отслеживать трендовые движения.

По тому же интервальному принципу, что и бары, работают японские свечи. Каждая свеча отображает нам следующую информацию: цена открытия, максимум свечи, минимум свечи, цена закрытия, используя для этого графические символы в форме японских свечей, каждая из которых дает краткий обзор результатов торговой деятельности за определенный период времени.

На рис. 1 изображены схемы свечи и бара.

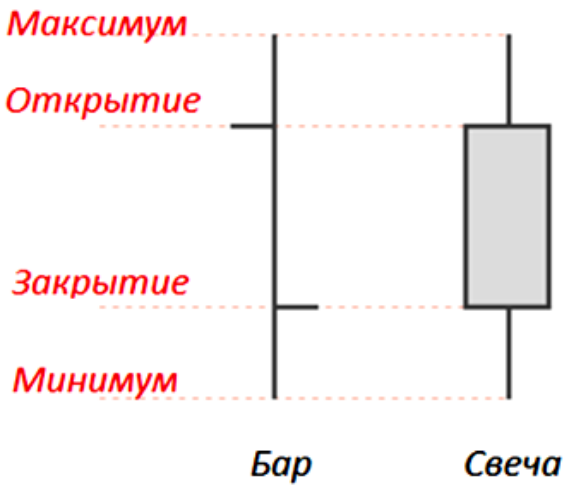

Рисунок 1 – Схема элемента OHLC-графика (бара) и свечного графика (японской свечи)

Свечной график строится наподобие графика баров, только результат закрытия периода выделяется цветом и называется «телом» свечи. За счет выделения цветом «тел» свечей он получается более наглядным.

Существуют множество самых разных вариаций свечных паттернов, однако алгоритм их анализа сводится к ответам на два вопроса: как закрылась свеча относительно своего диапазона и какой размер данной свечи относительно предыдущих свечей на графике. Если цена закрылась на максимуме диапазона, то это говорит нам о том, что покупатели контролируют ситуацию. Размер же свечи говорит нам о том, какой силой она обладает. Трендовое движение происходит на больших свечах. Откаты – это слабость в текущем тренде, они происходят на маленьких свечах, которые движутся против тренда.

Здоровому тренду характерно трендовое движение, за которым периодически следуют откаты. Однако, когда тренд ослабевает, у коррекционного движения уже не маленькие свечи, а большие, что свидетельствует о большой вероятности разворота цены. Техника анализа свечного графика позволяет предсказать поворотные моменты на рынке и находить торговые возможности с низким соотношением риска к прибыли.

Таким образом происходит усреднение линейного графика в первом приближении, при этом свечи и бары выполняют функцию индикаторов. Однако конструктивно они показывают одни и те же ценовые значения, отсеивая лишние для анализа ценовые флюктуации.

Свечной график с шагом с тайм-фреймом в 2 часа для фьючерсов *E-mini Nasdaq* 100 приведён на рис. 2.

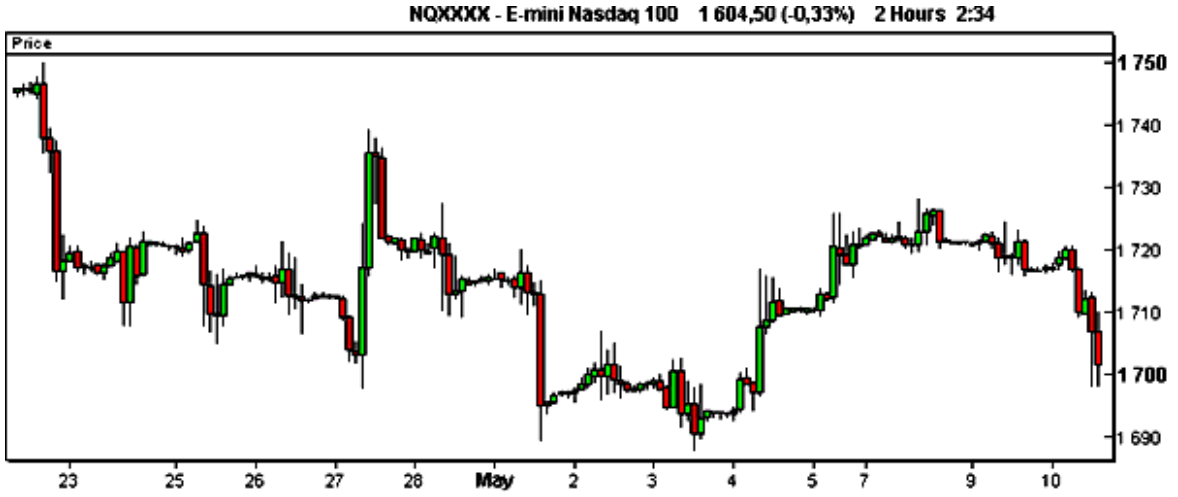

Рисунок 2 – Свечной график фьючерсов *E-mini Nasdaq* 100

Преимущества использования барового и свечного графика состоят в следующем:

– позволяет определить наличие ценовых разрывов (гэпов);
– можно быстро оценить ситуацию внутри торгового периода – наличие значительных подъемов или падений цен (максимумов/минимумов).

К недостаткам можно отнести то, что невозможно определить характер движения цены внутри рассматриваемого периода, для этого следует переключаться на более низкий таймфрейм.

Еще одной проблемой в этом подходе является то, что рынок не следует временному правилу. Люди не совершают сделки каждый фиксированный период вре-

мени (каждые N минут/часов). Альтернативой этому подходу может быть тиковый график, который выстраивается не по времени, а по количеству тиков, т.е. не привязан к временным периодам. Тиковый график отображает каждое изменение рыночной цены, на нем тики отображаются в одинаковом количестве, независимо от того, в течение какого периода времени они возникли.

Тиковый график с шагом в 3000 тиков для тех же фьючерсов *E-mini Nasdaq* 100 приведён на рис. 3.

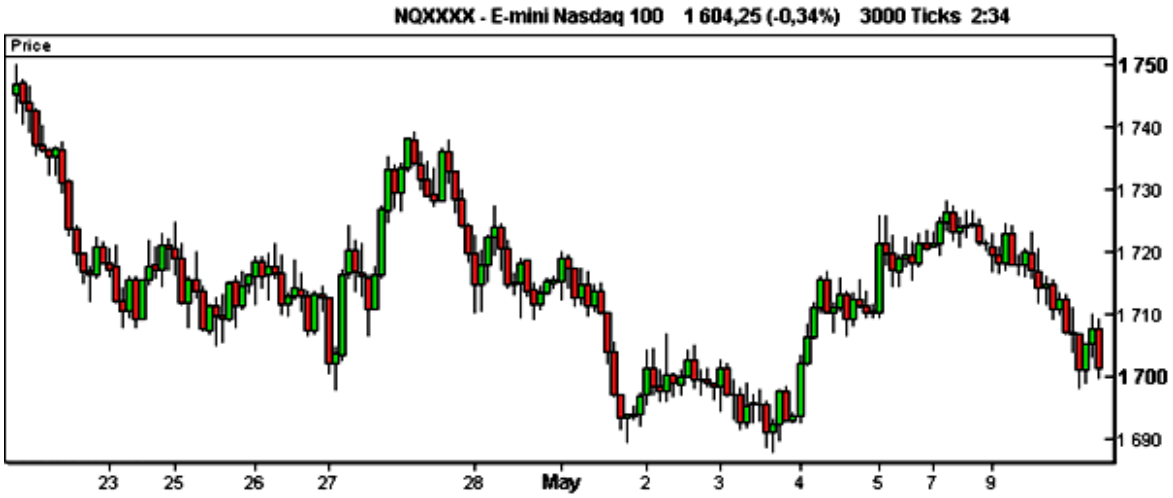

Рисунок 3 – Тиковый график фьючерсов *E-mini Nasdaq* 100

Применение тиковых графиков может иметь некоторые преимущества перед классическим представлением на основе внутридневного времени: более четкий анализ, нет привязки по времени, подтверждение прорыва линии тренда, более четкие признаки выхода из рынка, корреляция между объемом и изменением цены.

Сравнение графиков по времени и по тикам приведено на рис. 4.

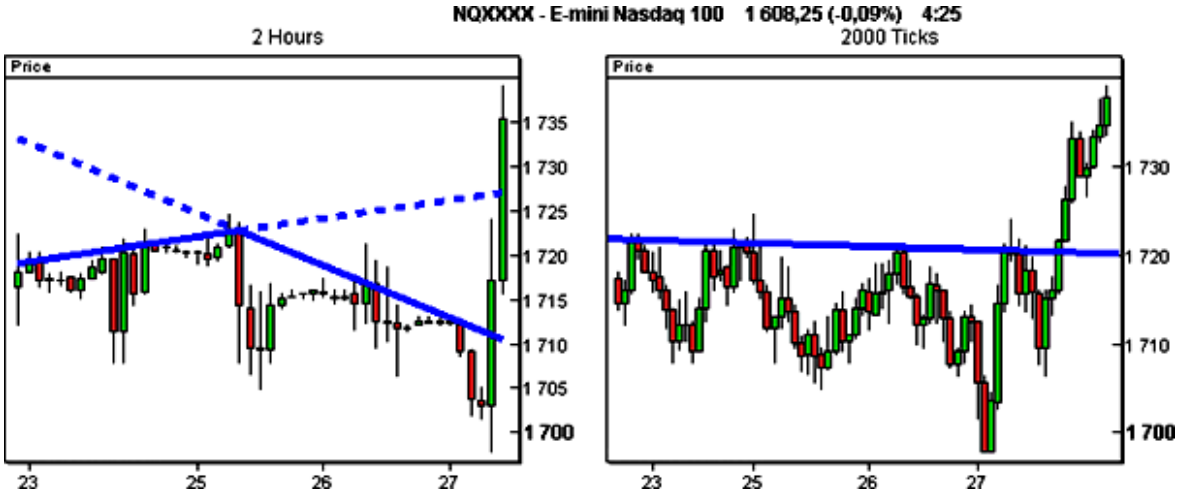

Рисунок 4 – Графики по времени и по тикам

Можно заметить, что на 2-часовом графике отображаются несколько маленьких свечей (обычных ночью) и несколько длинных свечей (обычных днем). Линия тренда может быть проведена между 23-м и 25-м числом, но не будет указывать на изменение тренда, которое происходит 25-го и 27-го числа. На самом деле нужно было бы провести две линии, что еще больше усложняет решение, куда может пойти

тренд. В представлении с тиками, длинные свечи разделены на более мелкие бары по 2000 тиков. С 23-го по 27-е можно провести сильную линию сопротивления, что позволит лучше понять тренд в эти дни.

Помимо тикового также можно использовать график объёмов, который использует количество реальных сделок для анализа рыночного движения. Оба этих подхода характерны тем, что в каждой отдельной свече одинаковое количество информации внутри, а тиковые свечи имеют одинаковое количество изменений цен, но при этом объём может сильно отличаться, поэтому в некоторых ситуациях лучше будет использовать график объёмов.

Поскольку сделок на рынке совершается очень много, то даже за одну минуту получается очень большое количество изменений цены, что делает тиковый график слишком хаотичным для анализа за длительный период, поэтому он часто используется трейдерами для внутридневной торговли.

В результате анализа графиков изменений цены выделяют 3 вида тренда – восходящий, нисходящий и боковой. Восходящий и нисходящий тренды характеризуются тенденциями повышения и понижения цен соответственно, в то время как боковой тренд отображает колебание цены в боковом диапазоне, когда не происходит ни явного роста, ни заметного снижения.

Маркировка и дифференцирование данных фондового рынка для использования прогностических моделей машинного обучения

При проведении анализа изменения цен алгоритмами машинного обучения, как правило, подают им на вход дифференцированный ряд. Как было сказано выше, возникает проблема: приращения цен стационарны, но не содержат памяти о прошлом, тогда как ценовой ряд содержит весь объем доступной памяти, но нестационарен. Следовательно, необходимо дифференцировать временной ряд, таким образом, чтобы он стал стационарным, при этом содержал максимально возможное количество памяти. Это достигается с помощью дробного дифференцирования, введенного Хоскингом в [18]. Операция дробного дифференцирования ряда $\{X_t\}$ с показателем d сводится к скалярному произведению:

$$\hat{X}_t = \sum_{k=0}^{\infty} w_k X_{t-k},$$

с весами w

$$w = \left\{1, -d, \frac{d(d-1)}{2!}, -\frac{d(d-1)(d-2)}{3!}, \ldots, (-1)^k \prod_{i=0}^{k-1} \frac{d-1}{k!}\right\}$$

и значениями X

$$X = \{X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-k}, \ldots\}.$$

Пример дробного дифференцирования показан на рис. 5.

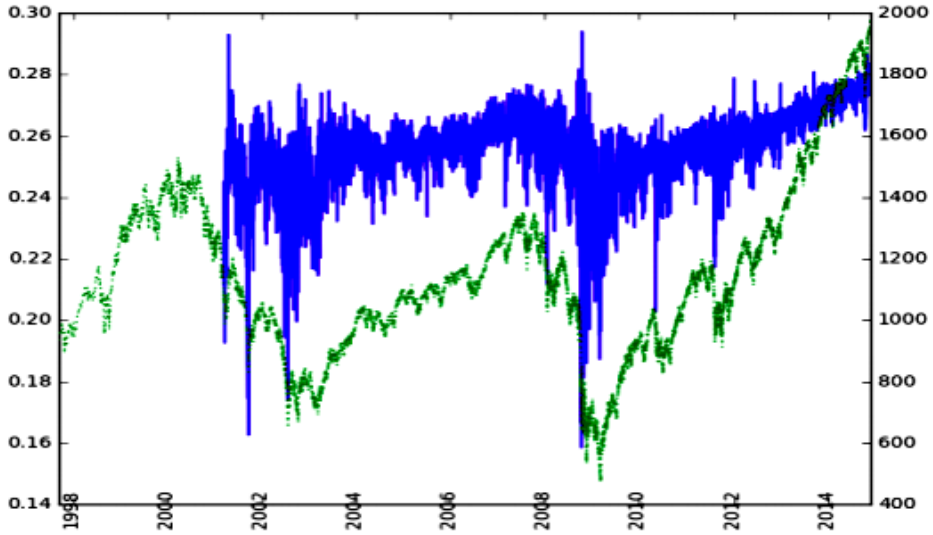

Рисунок 5 – График, отображающий дробное дифференцирование на фьючерсах *E-mini S&P 500*

Зелёная линия отображает фьючерсы *E-mini S&P 500*, а синяя – результат дробного дифференцирования. На длинной дистанции это напоминает уровень цен. Использование дробного дифференцирования поможет сохранить информацию о движении цены, поэтому будет использован именно этот вид дифференцирования.

Дробное дифференцирование имеет большую вычислительную сложность и отрицательное смещение трансформированного ряда, поскольку применяется для всей его последовательности. Маркос Лопес Де Прадо в [19] предложил метод фиксированного окна, в котором последовательность коэффициентов w отбрасывается, когда их модуль становится меньше заданного порогового значения τ. Эта процедура дает преимущество перед классическим методом расширяющегося окна, поскольку позволяет иметь одинаковые веса для любой последовательности исходного ряда, снижает сложность вычислений и избавляет от отрицательного смещения.

В тот момент, когда трейдер открывает позицию после какого-то сигнала, он помнит, какие были *take-profit* (приемлемая прибыль) и *stop-loss* (максимальные потери). Это значит, что более важно не то, как поменяется цена через прогнозируемое время, а то, как она будет вести себя на протяжении этого времени. Также цели могут меняться со временем из-за волатильности рынка, сумма ставки не менее важна в данном случае. Все эти требования к маркировке данных стоит учесть.

Для того чтобы, адаптироваться к изменчивой волатильности рынка, логичнее будет использовать динамические пороги вместо статических. Для чего используется следующая техника. Статический порог делится на 3 класса: если доходность текущей цены и будущего больше некоторого порога T; меньше, чем -T; все остальные в диапазоне от T до -T. Пример гистограмм распределения по этим классам приведен на рисунке 6. Можно зафиксировать порог T для всего набора данных или вычислять его адаптивно, используя стандартную вариацию доходности.

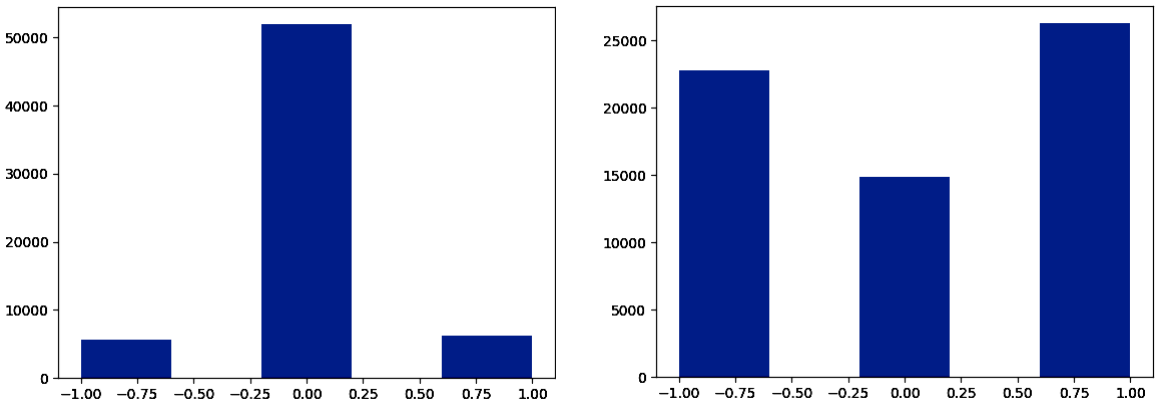

Рисунок 6 – Гистограммы статического порога и меток, основанных на волатильности

Для того чтобы учесть появление ситуации *stop-loss* в ближайшее время, или зафиксировать прибыль в нужный момент, используют метод тройного барьера. По этому методу нужно определить 3 барьера: два горизонтальных, которые представляют собой *stop-loss* и *take-profit* и одни вертикальный, который обозначает конечный горизонт (статический горизонт) [20]. На рис. 7 приведён график, который изображает метод тройного барьера.

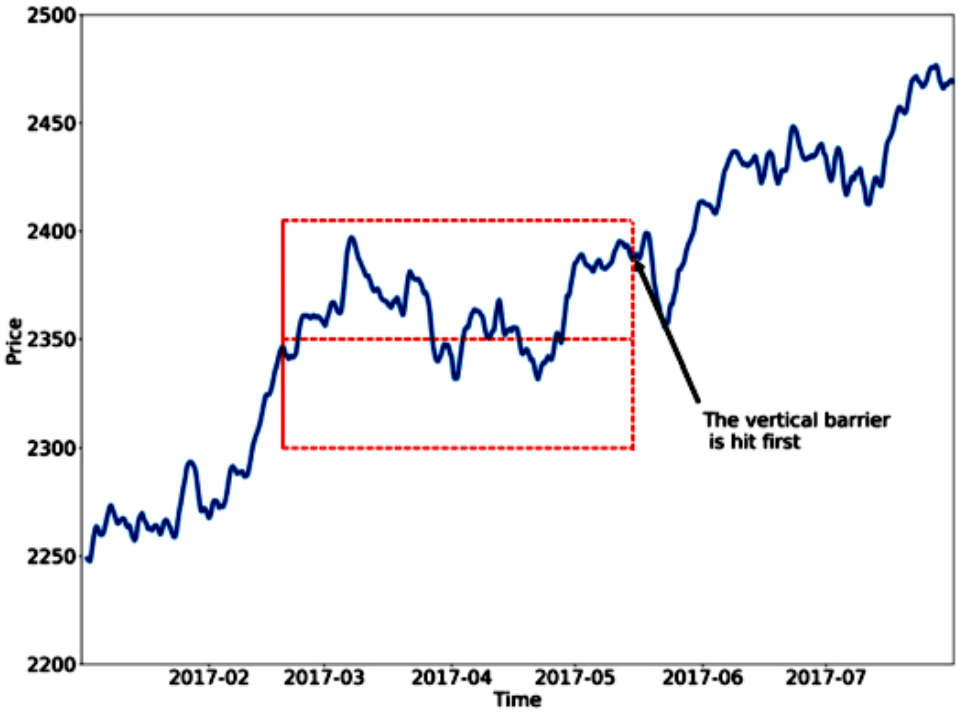

Рисунок 7 – График с применением тройного барьера

Если первым коснется верхний барьер, мы пометим наблюдение как 1. Если первым коснется нижний барьер, мы пометим наблюдение как -1. Если сначала коснулись вертикального барьера, мы помечаем наблюдение как 0. В ситуации, изображенной на графике видно, что первым из барьеров был задет правый вертикальный барьер.

Для решения вопроса о размере ставки предлагается использовать двухступенчатую маркировку Лопеса де Прадо. В этой маркировке одна метка отвечает за

направление, а вторая – за достоверность ставки и соответственно за её размер. Для первоначальной разметки (вверх или вниз) используют динамический порог в качестве предварительного *stop-loss* и барьера *take-profit*. Решение делать ставку или нет основывается на соответствии между направлением от первой метки и *stop-loss* или *take-profit*. Значение второй метки устанавливается как 1, если:

– первая метка имеет значение «вверх» и есть подтверждение того, что *take-profit* достигнет цели;
– первая метка имеет значение «вниз» и может быть попадание в *stop-loss*.

В случае, когда нет соответствия между направлением первой метки и *stop-loss* или *take-profit*, второй метке присваивается значение 0.

Применение моделей машинного обучения для прогнозирования фондового рынка

После того как рассмотрены структуры, встречающиеся во временных рядах изменения цены, проведена их маркировка, необходимо построить модели машинного обучения, позволяющие прогнозировать данные фондового рынка согласно классификации по введенной маркировке. Для решения этой задачи в работе предлагается обучить две модели, которые будут решать задачу классификации: результатом работы первой модели является маркировка движения цены по методу трех барьеров, а второй – коэффициент уверенности для маркеров, проставленных первой моделью, полученный по методу Лопеса Де Прадо.

Для получения прогноза и принятия решения о ставках с помощью методов машинного обучения предлагается осуществить следующе:

1) провести трансформацию данных тикового графика стандартизированных цен закрытия, объемов и доходности с помощью дробного дифференцирования методом фиксированного окна с целью приведения рядов к стационарности;
2) разделить трансформированный ряд на окна с некоторым количеством тиков и провести маркировку каждого окна по методу тройного барьера с динамическим порогом на три класса (цена движется вверх, вниз или колеблется);
3) оценить достоверность проведенной маркировки двухступенчатым методом Лопеса Де Прадо путём проверки достижения заданного *take-profit* при росте цены или *stop-loss* при падении цены.

Данные, маркированные методом тройного барьера, являются обучающими для модели, которая должна предсказывать, какой барьер цена пройдёт первой, т.е. проводить классификацию движения цены: вверх, вниз или стагнация. На вход модели поступает отрезок временного ряда, соответствующий окну, а выходом является маркер движения цены, поставленный для этого окна. После чего обучается вторая модель, на вход которой поступают данные изменения цен и результат предсказания первой модели. Выходом второй модели является показатель достоверности (от 0 до 1), предсказывающий, достигнет ли ставка *take-profit* или же *stop-loss*. Для обучения этой модели данные маркируются двухступенчатым методом Лопеса Де Прадо, где на первом этапе маркировка выполняется методом тройного барьера.

Таким образом первая модель предсказывает один из трех классов направления движения цены, а вторая – эффективность ставки. Предлагаемая технология анализа финансовых рядов позволит принять решения: чем ближе к единице будет предсказанное значение второй модели, тем больше и уверенней стоит делать ставку. В случае низкой достоверности (менее 0.5), следует воздержаться от ставки несмотря на то, что первая модель предсказала рост цены.

Выводы

Сложность решения задачи прогнозирования роста цен и акций на рынке обусловлена наличием значительного количества факторов, влияющих на стоимость ценных бумаг и приводящих к высокой волатильности и зашумлению данных. Для приведения к стационарности ряд дифференцируют, т.е. аналитики работают с приращениями или логарифмами приращений цен. Это позволяет применить модели, основанные на автокорреляционной функции (ARMA, ARIMA). Однако дифференцирование ряда удаляет всю память из ценовых последовательностей.

Для того чтобы ряд был стационарным и при этом содержал максимально возможное количество памяти, используют дробное дифференцирование.

Для отображения колебаний котировок цен используют интервальные графики в виде баров или свечей, а также тиковый график, который не привязан к временным периодам, а строится по количеству тиков. В графиках изменений цены можно выделить три класса тренда: восходящий, нисходящий и боковой (стагнация).

Для решения задачи прогнозирования финансовых временных рядов представляется перспективным использовать методы машинного обучения, в частности, нейронные сети, поскольку они обладают тем преимуществом, что способны обобщить и выделить скрытые зависимости между данными, на основе чего прогнозировать или проводить классификацию.

В работе предлагается проводить по имеющимся данным временного ряда как классификацию (маркировку) движения цены, так и вычислить показатель достоверности этой маркировки. Достигается это с помощью двух моделей. Первая по поступившему на вход отрезку трансформированного с помощью дробного дифференцирования ряда ставит ему в соответствие маркер движения цены. Вторая, в свою очередь, определяет показатель достоверности маркировки от 0 до 1. Обучение второй модели предлагается проводить двухступенчатым методом Лопеса Де Прадо. Для краткосрочного инвестирования рациональным представляется использовать тики вместо временного ряда цен.

Предлагаемый подход может быть полезен при принятии решений о ставках и инвестировании: чем ближе к единице будет предсказанное значение второй модели, тем больше и уверенней стоит делать ставку. Если первая модель предсказала рост цены, а вторая – низкую достоверность (менее 0.5), то от ставки следует воздержаться.

Список литературы

1. Пшекоп, В. Ю. Математические модели прироста цены финансовых инструментов на основе симметричного и асимметричного распределения Лапласа // Проблемы искусственного интеллекта. 2019. № 2(13). С. 87-92.
2. Павлыш, В. Н., Миньковская, М. В. Компьютерные средства анализа рисков в условиях рыночной деятельности предприятия // Проблемы искусственного интеллекта. 2016. № 2(3). С. 55-65.
3. Павлюк, Е. Н., Криводубский, О. А. Разработка математической модели тактического прогноза деятельности строймаркета // Проблемы искусственного интеллекта. 2022. № 2(25). С. 16-28.
4. Жмыхова, Т. В., Чудина, Е. Ю. Вероятность разорения страховой компании, оперирующей на биномиальном финансовом рынке, определяемая на основе полиномов Лагерра // Проблемы искусственного интеллекта. 2022. № 4(27). С. 16-23.
5. Ермоленко, Т. В., Котенко, В. Н., Винник, А. О. Исследование эффективности предсказательных моделей для системы анализа и мониторинга энергопотребления на предприятиях угольной промышленности // Проблемы искусственного интеллекта. 2022. № 4(27). С. 25-34.
6. Лабусов, М.В. Нейронные сети долгой краткосрочной памяти и их использование для моделирования финансовых временных рядов // Инновации и инвестиции. 2020. № 3. С. 167-171.

7. Лабусов, М.В. Обзор моделей анализа и прогнозирования высокочастотных финансовых временных рядов // Экономика и предпринимательство. 2019. № 6 (107). С. 1256-1258.
8. Хайндман, Р. Дж., Атанасопулос Дж. Прогнозирование: принципы и практика. 3-е изд. Мельбурн: OTexts, 2021. 382 с.
9. Щербинина, А. В., Алжеев, А. В. Сравнительный анализ качества прогнозирования классической статистической модели и модели машинного обучения на данных российского фондового рынка // Ученые записки Российской академии предпринимательства. 2021. № 20(3). С. 52-63.
10. Кратович, П. В. Нейронные сети и модели ARIMA для прогнозирования котировок // Программные продукты и системы. – 2011. – №1. – С. 95-98.
11. Лобанов, А. А., Чугунов, А. В. Энциклопедия финансового риск-менеджмента. Москва: Альпина Паблишер, 2009. 878 с.
12. Применение машинного обучения в сфере финансовой математики [Электронный ресурс]. URL: https://doicode.ru/doifile/lj/68/lj-12-2020-221.pdf (дата обращения: 16.08.2023).
13. Беспалова, С .В. Построение предсказательных моделей параметров давления воды в водораспределительных сетях с помощью методов машинного обучения / С. В. Беспалова, С. М. Романчук, Т. В. Ермоленко, В. И. Бондаренко // Проблемы искусственного интеллекта. 2019. №2(13). С. 24-38.
14. Беспалова, С. В. Построение регрессионных моделей режимов работы водораспределительных сетей с помощью методов регуляризации и анализа главных компонент / С. В. Беспалова, С. М. Романчук, Т. В. Ермоленко, В. И. Бондаренко // Информатика и кибернетика. 2019. № 2(16). С. 35-49.
15. Мэрфи, Д.Дж. Технический анализ фьючерсных рынков. Теория и практика; перевод с английского О. Новицкой и В. Сидорова. Москва: Альпина Паблишер, 2015. 610 с. ISBN 978-5-9614-5332-4.
16. Трегуб, И.В. Технический анализ финансовых рынков: учебник. Москва : Финансовый университет, 2013. 224 с. ISBN 978-5-7942-0993-8.
17. Machine Learning Applied to Real World Quant Strategies [Электронный ресурс]. URL: https://www.quantstart.com/advanced-algorithmic-trading-ebook/ (дата обращения: 23.08.2023).
18. JHosking. Fractional diffencing // Biometrika, Vol. 68, № 1, 1981. P. 165-176. [Электронный ресурс]. – URL: https://www.ma.imperial.ac.uk/~ejm/M3S8/Problems/hosking81.pdf (дата обращения: 28.08.2022).
19. Прадо М. Л. Машинное обучение: алгоритмы для бизнеса. Санкт-Петербург: издательский дом «Питер», 2019. 432 с.
20. Yves J. H. Python for Finance Mastering Data-Driven Finance. Sebastopol: O'Reilly Media, 2018. 720 с.

References

1. Pshekop, V. YU. Matematicheskie modeli prirosta ceny finansovyh instrumentov na osnove simmetrichnogo i asimmetrichnogo raspredeleniya Laplasa. *Problemy iskusstvennogo intellekta*. 2019. № 2(13). S. 87-92.
2. Pavlysh, V. N. Komp'yuternye sredstva analiza riskov v usloviyah rynochnoj deyatel'nosti predpriyatiya / V. N. Pavlysh, M. V. Min'kovskaya. *Problemy iskusstvennogo intellekta*. 2016. № 2(3). S. 55-65.
3. Pavlyuk, E. N. Razrabotka matematicheskoj modeli taktricheskogo prognoza deyatel'nosti strojmarketa / E. N. Pavlyuk, O. A. Krivodubskij. *Problemy iskusstvennogo intellekta*. 2022. № 2(25). S. 16-28.
4. Zhmyhova, T. V. Veroyatnost' razoreniya strahovoj kompanii, operiruyushchej na binomial'nom finansovom rynke, opredelyaemaya na osnove polinomov Lagerra / T. V. ZHmyhova, E. YU. CHudina. *Problemy iskusstvennogo intellekta*. 2022. № 4(27). S. 16-23.
5. Yermolenko, T. V. Issledovanie effektivnosti predskazatel'nyh modelej dlya sistemy analiza i monitoringa energoporebleniya na predpriyatiyah ugol'noj promyshlennosti / T. V. Ermolenko, V. N. Kotenko, A. O. Vinnik. *Problemy iskusstvennogo intellekta*. 2022. № 4(27). S. 25-34.
6. Labusov, M.V. Nejronnye seti dolgoj kratkosrochnoj pamyati i ih ispol'zovanie dlya modelirovaniya finansovyh vremennyh ryadov / M.V. Labusov. *Innovacii i investicii*. 2020. № 3. S. 167-171.
7. Labusov, M.V. Obzor modelej analiza i prognozirovaniya vysokochastotnyh finansovyh vremennyh ryadov. *Ekonomika i predprinimatel'stvo*. 2019.№ 6 (107). S. 1256-1258. ISSN 1999-2300.
8. Hyndman R. J., Athanasopoulos G. *Forecasting: Principles and Practice*. 3rd edition. Melbourne: OTexts, 2021. 382.
9. ShcHerbinina, A.V., Alzheev A. V.Sravnitel'nyj analiz kachestva prognozirovaniya klassicheskoj statisticheskoj modeli i modeli mashinnogo obucheniya na dannyh rossijskogo fondovogo rynka. *Uchenye zapiski Rossijskoj akademii predprinimatel'stva*. 2021. № 20(3). S. 52-63.

10. Kratovich P. V. Nejronnye seti i modeli ARIMA dlya prognozirovaniya kotirovok. *Programmnye produkty i sistemy*. 2011. №1. S. 95-98.
11. Lobanov A. A. *Encyclopedia of financial risk management*. Moscow, Alpina Publisher, 2009. 878.
12. *Application Machine Learning to financial mathematics* [Electronic resource]. URL: https://doicode.ru/doifile/lj/68/lj-12-2020-221.pdf (date of the application: 16.08.2022).
13. Bespalova, S .V. Postroenie predskazatel'nyh modelej parametrov davleniya vody v vodoraspredelitel'nyh setyah s pomoshch'yu metodov mashinnogo obucheniya / S. V. Bespalova, S. M. Romanchuk, T. V. Yermolenko, V. I. Bondarenko. *Problemy iskusstvennogo intellekta*. 2019. №2(13). S. 24-38.
14. Bespalova, S. V. Postroenie regressionnyh modelej rezhimov raboty vodoraspredelitel'nyh setej s pomoshch'yu metodov regulyarizacii i analiza glavnyh komponent / S. V. Bespalova, S. M. Romanchuk, T. V. Yermolenko, V. I. Bondarenko. *Informatika i kibernetika*. 2019. № 2(16). S. 35-49.
15. Merfi, D.Dzh. *Tekhnicheskij analiz f'yuchersnyh rynkov. Teoriya i praktika* / Dzh.D. Merfi ; perevod s anglijskogo O. Novickoj i V. Sidorova. Moskva: Al'pina Pablisher, 2015. 610 s.
16. Tregub, I.V. *Tekhnicheskij analiz finansovyh rynkov: uchebnik* / I.V. Tregub. Moskva : Finansovyj universitet, 2013 224 s. ISBN 978-5-7942-0993-8.
17. *Machine Learning Applied to Real World Quant Strategies* [Electronic resource]. URL: https://www.quantstart.com/advanced-algorithmic-trading-ebook/ (date of the application: 23.08.2022).
18. JHosking. Fractional diffencing. *Biometrika*, Vol. 68, № 1, 1981. P. 165-176. [Electronic resource]. URL: https://www.ma.imperial.ac.uk/~ejm/M3S8/Problems/hosking81.pdf (date of the application: 28.08.2022).
19. Prado M. L. *Machine learning: algorithms for business*. St. Petersburg: publishing house «Piter», 2019. 432.
20. Yves J. H. *Python for Finance Mastering Data-Driven Finance*. Sebastopol: O'Reilly Media, 2018. 720.

RESUME

T. V. Yermolenko, D. V. Popadin, V. N. Kotenko

Application of machine learning in stock market forecasting

The task of predicting the growth of prices and stocks in the market is complicated by a large amount of conflicting and ambiguous data that affect price formation, non-linear changes in the financial sector, and high volatility that financial series have. This leads to noisy series, the presence of outliers, which affects the quality of forecasts obtained using machine learning models.

For effective forecasting, it is necessary to bring the time series to stationarity by transforming the data, as well as to identify structures in it that allow analysis, despite volatility.

To study price dynamics, such structures as bars, candles and ticks are used. Unlike bars and candles, ticks are not tied to a fixed time period, the chart displays ticks in the same number, regardless of the period of time they occurred, which is an undeniable advantage of this structure. However, the tick chart is chaotic for analysis over a long period, so along with it, a volume chart is used, that is, the number of real transactions, as well as profitability.

To bring the series to stationarity, a transformation called fractional differentiation is used, which allows you to save information about the price movement, in contrast to ordinary differentiation. This solves the problem of stationarity and memory.

To analyze financial series and make an investment decision, it is proposed to use two models. For their training, the data is divided into windows with a certain number of ticks and marked depending on the price movement using the triple barrier method with a dynamic threshold into three classes (up, down or stagnation). To assess the reliability of the classification, a two-stage labeling of Lopez de Prado is used. The first model predicts one of the three classes of price movement direction, and the second predicts the effectiveness of the rate.

The implementation of the proposed approach to the analysis of the stock market, which takes into account not only the direction of price changes, but also its reliability, will improve the efficiency of trading, reducing investment risks.

РЕЗЮМЕ

Т. В. Ермоленко, Д. В. Попадин, В. Н. Котенко

Применение машинного обучения в прогнозировании фондового рынка

Задача прогнозирования роста цен и акций на рынке осложняется большим количеством противоречивых и неоднозначных данных, которые влияют на формирование цены, нелинейными изменениями в сфере финансов, высокой волатильностью, которой обладают финансовые ряды. Это приводит к зашумлению рядов, наличию выбросов, что влияет на качество прогнозов, полученных с помощью моделей машинного обучения.

Для эффективного прогнозирования необходимо привести временной ряд к стационарности путем трансформации данных, а также выделить в нем структуры, позволяющие проводить анализ, несмотря на волатильность.

Для исследования динамики цен используются такие структуры как бары, свечи и тики. В отличие от баров и свечей, тики не привязаны к фиксированному временному периоду, график отображает тики в одинаковом количестве, независимо от того, в течение какого периода времени они возникли, что является неоспоримым преимуществом этой структуры. Однако, тиковый график хаотичен для анализа за длительный период, поэтому наряду с ним используется график объёмов, т.е. количества реальных сделок, а также доходности.

Для приведения ряда к стационарности используется преобразование, называемое дробном дифференцированием, которое позволяет сохранить информацию о движении цены, в отличие от обычного дифференцирования. Таким образом решается проблема стационарности и памяти.

Для анализа финансовых рядов и принятия решения об инвестировании предлагается использовать две модели. Для их обучения данные делятся на окна с некоторым количеством тиков и маркируются в зависимости от движения цены с помощью метода тройного барьера с динамическим порогом на три класса (вверх, вниз или стагнация). Для оценки достоверности классификации используется двухступенчатая маркировка Лопеса де Прадо. Первая модель предсказывает один из трех классов направления движения цены, а вторая – эффективность ставки.

Реализация предложенного подхода к анализу фондового рынка, учитывающего не только направление изменения цены, но и его достоверность, позволит повысить эффективность трейдинга, снизив риски инвестирования.

Статья поступила в редакцию 20.03.2023