# Прогнозирование котировок фьючерсов на индекс РТС на основе машинного обучения

Н. В. Воинов[1], М. К. Ворошилов, С. А. Молодяков,
П. Д. Дробинцев, О. В. Прокофьев, И. В. Зайцев
Санкт-Петербургский политехнический университет Петра Великого
[1]voinov@ics2.ecd.spbstu.ru

***Аннотация*. Работа посвящена исследованию методов машинного обучения для прогнозирования котировок финансовых инструментов. Проведен обзор наиболее распространенных методов и проанализированы результаты их применения. Разработан и реализован подход к краткосрочному прогнозированию котировок фьючерсов на индекс РТС с применением линейной и полиномиальной регрессий. Особенностью подхода является доступ модели к будущей стоимости фьючерса на этапе обучения. Проанализированы полученные результаты и сделаны выводы об эффективности разработанного подхода.**

***Ключевые слова: прогнозирование; линейная и полиномиальная регрессии; фьючерс на индекс РТС; финансовые инструменты***

## I. Введение

Анализ состояния фондового рынка с целью прогнозирования стоимости ценных бумаг и других финансовых инструментов всегда остается актуальной задачей. И хотя предсказать изменение котировок со стопроцентной вероятностью вряд ли когда-нибудь станет возможно, активно развиваются программные средства сбора и анализа больших объемов данных, влияющих на фондовый рынок, для повышения точности прогнозов. Современные программные решения в данной области часто основаны на методах машинного обучения. В первую очередь, это обусловлено тем, что фондовый рынок по сути своей очень динамичен, отсутствуют четкие гарантии на его поведение в будущем, это предоставляет хорошее экспериментальное поле для проверки различных алгоритмов машинного обучения. Закономерности развития в движении котировок финансовых инструментов постоянно изменяются, не имея чёткого закона, привнося большое количество рисков для финасновых игроков, а каждый подход в области машинного обучения пытается по-своему их минимизировать.

В данной статье рассмотрены наиболее известные методы машинного обучения для прогнозирования котировок финансовых инструментов, описан разработанный подход к краткосрочному прогнозированию котировок фьючерсов на индекс РТС с применением линейной и полиномиальной регрессий, проанализированы полученные результаты.

## II. Обзор существующих методов и подходов

Одним из наиболее распространенных алгоритмов машинного обучения применительно к прогнозированию фондового рынка является линейная регрессия [1, 2]. Это модель зависимости переменной от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости. Линейная регрессия относится к задаче определения «линии наилучшего соответствия» через набор точек данных. Если модель становится нелинейной комбинацией входных переменных, то такую регрессию называют полиномиальной. Линейная и полиномиальная регрессии применяются в прогнозировании числовых рядов, общее поведение которых можно представить с помощью линейной зависимости или функции-полинома.

Также широко применяемым подходом анализа является метод опорных векторов (SVM – Support Vector Machine) [3]. Это набор алгоритмов обучения с учителем, работающих по принципу разделения данных гиперплоскостью, которая максимизирует зазор между плоскостью и данными из обучающей выборки. Помимо SVM применяют деревья решений и случайные леса (ансамбли) решающих деревьев [2]. Классификация осуществляется с помощью голосования классификаторов, определяемых отдельными деревьями. Количество деревьев является параметром при использовании соответствующих классов или методов, реализованных в библиотеках [4].

Ещё одним распространённым подходом к прогнозированию цен на фондовом рынке является использование нейронных сетей [3]. Искусственная нейронная сеть – это математическая модель, а также ее программные или аппаратные реализации, построенная по образу сетей нервных клеток живого организма. Нейронные сети в области искусственного интеллекта являются упрощенным моделями биологических нейронных сетей. Простейший вид нейронной сети - персептрон. В его основе лежит математическая модель восприятия информации мозгом, состоящая из сенсоров, ассоциативных и реагирующих элементов.

В статье [5] в качестве способов прогнозирования цен рассматриваются линейная регрессия и регрессионная модель метода опорных векторов (SVR – Support Vector Regression). Результаты демонстрируют, что линейная регрессия может показываться себя лучше, чем полиномиальная модель на определённых интервалах из-за меньшей возможности переобучения линейной модели на тренировочных данных. В то же время полиномиальная модель более склонна к переобучению и показывает себя

лучше на данных из начала тестовой выборки, но её результаты на долгосрочном прогнозе начинают расходиться и их нельзя назвать точными. Недостатком данной работы можно считать отсутствие использование полиномиальной регрессии в чистом виде из класса, а применяется лишь SVR с полиномиальным типом ядра. В статье [6] используются два вида нейронных сетей и линейная регрессия для предсказания изменения цены на следующий день. Сравниваются многослойный персептрон (MLP – Multilayer Perceptron), рекуррентная нейронная сеть Элмана и линейная регрессия. Результаты работы демонстрируют, что MLP более точно предсказывает значение изменения цены в то время, как нейронная сеть Элмана и линейная регрессия лучше прогнозируют направление движения цены. В статье [7] для реализации модели используется комплекс из библиотеки TensorFlow и рекуррентной нейронной сети LSTM (Long Short-Term Memory), эффективность которого сопоставляется с популярной моделью для анализа временных рядов ARIMA. В качестве результатов точности предсказаний для краткосрочного прогноза приводятся значения в 94 % и 56 % для LSTM и ARIMA моделей соответственно.

Опираясь на исторические данные о движении фондового рынка, можно выявлять определённые шаблоны (паттерны) его поведения. Выделение подобных закономерностей часто применяется в профессиональных торговых инструментах и может положительно сказаться на инвестиционной стратегии. В статье [8] при анализе финансовых временных рядов удалось выделить более 17000 паттернов поведения. Для хранения и обработки такого объема данных использовали HDFS (Hadoop Distributed File System) и иерархическую кластеризацию, а в качестве инструмента для прогнозов – нейронную сеть, что позволило минимизировать среднеквадратичную ошибку до 500–2000 при учёте стоимости акций до 134500 у.е. и сделать точный прогноз на основе паттернов поведения цены в прошлом. Такое решение можно считать очень точным и эффективным, но его слабой стороной может оказаться неподготовленность алгоритмов к колебаниям, которых не было ранее среди исторических данных.

Предположения о будущем состоянии финансовых активов можно строить не только со стороны получения определённого значения стоимости ценных бумаг, но и с точки зрения задачи классификации. В статье [9] следуют именно такому подходу. Ставится задача предсказать направление изменения стоимости акций компании на 10% в сторону роста или падения на протяжении одного года. Среди множества протестированных алгоритмов случайный лес показал наибольшую точность в 76,5 %, при этом показатели SVM и логистической регрессии оказались ниже – 62,4 % и 62,5 % соответственно.

В результате обзора материалов по данной тематике были сделаны следующие выводы:

- в определённый момент времени движение котировок может не поддаваться описанию с помощью ранее выделенных паттернов;
- каждый алгоритм может показывать себя лучше в определенных ситуациях и на определенных наборах данных;
- нейронные сети имеют преимущества в виде способности адаптироваться к изменениям внешних факторов и решать задачи при неизвестных зависимостях между входными и выходными переменными;
- предиктивный анализ рынка возможен как с точки зрения задачи получения конкретных значений и проверки их на точность (регрессионные модели и нейросети), так и с точки зрения задачи классификации (метод опорных векторов, случайный лес и логистическая регрессия).

Было принято решение разработать подход к краткосрочному прогнозированию котировок фьючерсов на индекс РТС с применением линейной и полиномиальной регрессий, реализовав доступ модели к будущей стоимости фьючерса на этапе обучения. После анализа полученных результатов данный подход был доработан.

## III. Разработанный подход к прогнозированию котировок

### *A. Концепция*

Предлагается решение задачи регрессии для получения значений цены закрытия фьючерсного контракта в течение краткосрочного периода времени. Основная идея реализуемого подхода – обучить модель таким образом, чтобы во время выстраивания собственных зависимостей на тренировочной выборке в метках содержались значения прогнозируемой цены закрытия через определённое количество тиков (N) – изменений котировок. Иными словами, во время обучения модель будет иметь доступ к стоимости этого же фьючерса в будущем. Соответственно, при проверке модели на тестовой выборке ожидается, что регрессионные алгоритмы смогут применить выявленные тенденции развития финансовых временных рядов для получения будущих значений цены закрытия.

### *B. Базовый алгоритм*

Начальный этап – формирование датасета. Экспортированный с сервера брокера текстовый файл преобразуется в таблицу данных. Чтобы обеспечить доступ к будущей цене во время обучения модели, создается копия столбца признаков (Features) в столбце меток (Label) с предсказываемой ценой (рис. 1*а*). Далее столбец с метками сдвигается на N позиций вверх, где N – количество тиков, на которое производится краткосрочный прогноз (рис. 1*б*). В результате сдвига на N позиций вверх первые N ячеек в столбце меток удалятся. На место каждого текущего значения цены закрытия встанет соответствующее значение через N тиков. Последние N тиков приобретут пустые значения (NaN). Пример сдвига на два тика (N=2) показан на рис. 1*в*. Каждой цене закрытия из столбца признаков соответствует цена закрытия спустя 2 тика из столбца с метками.

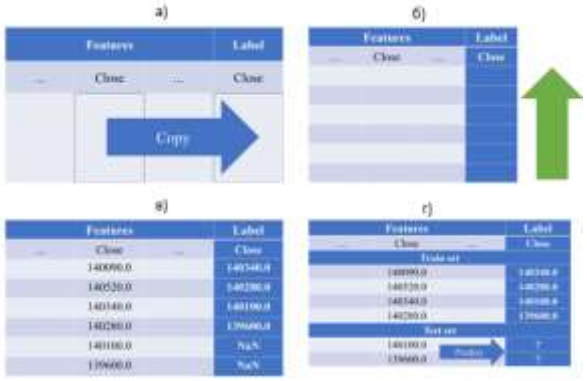

Рис. 1. Подготовка датасета

Далее выборка делится на обучающую и тестовую. Для каждого значения из тестовой выборки модель прогнозирует соответствующее значение через N тиков. На рис. 1*г* показан принцип получения будущих цен закрытия относительно текущих при разделении всего набора данных на обучающую и тестовую выборки.

После обучения производится проверка модели на тестовой выборке при различных N. Полученные в результате применения линейной и полиномиальной регрессий значения сравниваются с реальными и между собой, производится оценка точности. Далее строятся графики зависимости реальной и прогнозируемой цены от времени.

Программная реализация выполнена средствами языка Python с применением библиотек NumPy (работа с многомерными массивами), pandas (работа с датасетом), scikit-learn (алгоритмы машинного обучения), Matplotlib (построение графиков).

### *C. Анализ результатов*

Разработанный алгоритм был проверен на исторических данных по фьючерсу на индекс РТС с предсказанием цены закрытия. Данные были получены с Московской фондовой биржи, а котировки экспортировались с сайта брокера «Финам» с периодом торгов в один час. В качестве N были выбраны значения 1, 10 и 24. Критерия оценки точности стал метод score библиотеки scikit-learn. Чем полученное методом значение ближе к единице, тем более достоверным и точным считается результат.

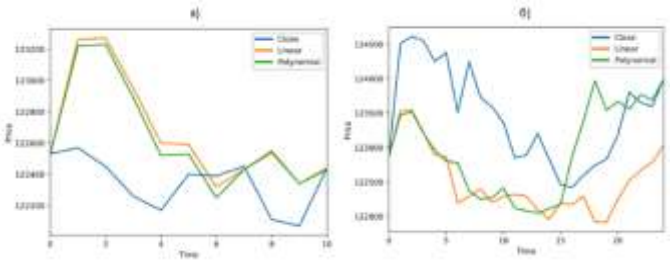

Рис. 2. График зависимости реальной (синий цвет) и прогнозируемой цены (зеленый цвет для полиномиальной регрессии, оранжевый - для линейной)

При предсказании следующего тика (N=1) показатель точности, полученный с помощью метода score, равен 0,995 как для линейной регрессии, так и для полиномиальной.

Прогноз на будущие десять тиков (N=10) при использовании линейной регрессии даёт показатель точности 0,9508, полиномиальной – 0,9510. Средний модуль отклонения составляет 175, среднеквадратичное отклонение – 220. В данном случае применение полиномиальной регрессии демонстрирует более точные результаты (рис. 2*а*).

Прогноз при N=24 происходит с меньшей точностью: 0,8875 для линейной регрессии и 0,8888 для полиномиальной. При этом средний модуль отклонения составил 374, а среднеквадратичное отклонение – 543 (рис. 2*б*).

Визуальное представление краткосрочного прогноза для полиномиальной и линейной регрессий может заметно отличаться, хотя при этом показатель точности метода score остаётся похожим. Причина в том, что оценка эффективности производится по всей тестовой выборке, учитывая множество краткосрочных прогнозов, а не по отдельному участку данных на графике.

По полученным результатам можно сделать вывод о том, что полиномиальная регрессия показывает себя более эффективно по сравнению с линейной в связи с общими колебаниями цены фьючерсного контракта за рассматриваемый период, который лучше поддаётся описанию через функцию-полином, чем через линейную зависимость. Также прогнозы полиномиальной регрессии пытаются сгладить риски, не так сильно отклоняясь в резкий рост или падение. Но нельзя однозначно говорить, что линейная регрессия всегда будет больше отклоняться от реальных значений. Если данные распределены более линейно, то её применение окажется эффективнее.

С увеличением количества прогнозируемых тиков (N) увеличивается и ошибка, но на краткосрочных прогнозах удается успешно спрогнозировать общее поведение цены в сторону роста или падения. Также необходимо учитывать, что показатель точности метода score (коэффициент детерминации) дополнительно принимает во внимание, насколько хорошо обучена модель. Такая оценка не гарантирует минимизацию ошибки в реальных условиях, поэтому дополнительно вычислялись значения модуля среднего отклонения и среднеквадратичного отклонения.

### *D. Доработка алгоритма*

Фактически описанный алгоритм не прогнозирует дальнейшую динамику цены, а повторяет её последние значения со сдвигом, равным количеству прогнозируемых тиков. Для усовершенствования подхода был применен метод скользящего окна. Под окном в данном случае понимается временной интервал, содержащий набор значений, которые используются для формирования обучающего примера (записи из обучающего набора данных). В процессе работы алгоритма окно смещается по временной последовательности на один тик, и каждое положение окна образует один пример.

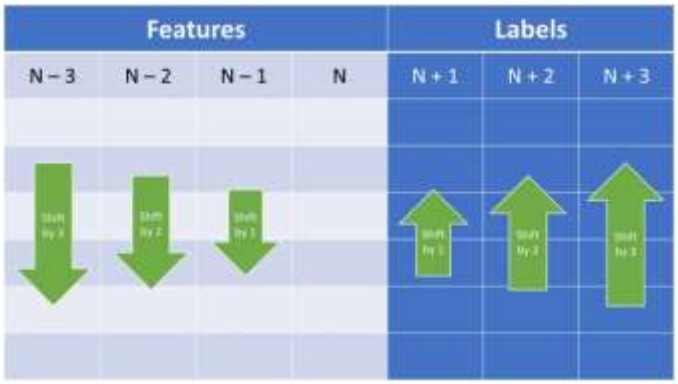


Рис. 3. Формирование датасета при использовании метода скользящего окна

После реализации метода скользящего окна раздел признаков (Features) в датасете выглядит следующим образом (рис. 3): слева от столбца с ценой закрытия в текущий день (N) находится столбец с ценой закрытия в предыдущий день (скопированный столбец с реальными значениями, но сдвинутый на одну позицию вниз). И таких признаков будет ровно N+1. Например, при прогнозе на 3 дня в датасете содержатся следующие признаки: цена сегодня (день N), цена в день N-1, в день N-2 и в день N-3.

Аналогичная ситуация с метками (Labels). Их будет ровно N: цена закрытия в день N+1 (завтра), в день N+2 и в день N+3. Только сдвиг этих столбцов будет производиться не вниз, а вверх на одну, две и три единицы соответственно относительно текущего дня (рис. 3).

В итоге модель будет содержать N линейных и N полиномиальных регрессий, и каждая из них, обучаясь на скользящем окне, будет пытаться получить свою собственную цену из столбца меток. Таким образом, каждая регрессия будет иметь свой собственных набор коэффициентов в отличие от базового алгоритма, где была только одна регрессия с одним и тем же набором коэффициентов для прогнозирования цены в день N+1, в день N+2 и т. д.

Доработанный алгоритм позволяет делать более точный прогноз. На рис. 4 приведено несколько графиков с реальной и прогнозируемой ценой различных ценных бумаг с прогнозом на 5 дней. По результатам можно сделать вывод о том, что алгоритм действительно реализует полноценный прогноз, пытаясь рассчитать будущую цену, а не просто повторяет последние участки исторических данных. При этом линейная регрессия предлагает консервативные прогнозы, выстраивая точки как результирующую аппроксимацию тех данных, на которых она обучалась. В то же время полиномиальная регрессия чаще рискует и старается, в первую очередь, угадать дальнейший тренд, что получается успешнее, чем у линейной модели. Показатель среднеквадратичной ошибки на рассмотренных наборах данных у полиномиальной регрессии также ниже, чем у линейной.

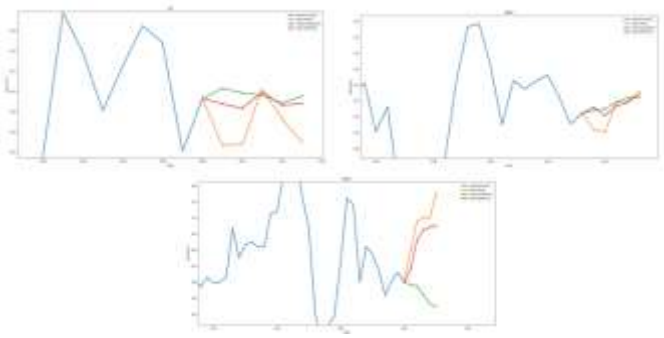

Рис. 4. График зависимости реальной (оранжевый цвет) и прогнозируемой цены (зеленый цвет для линейной регрессии, красный - для полиномиальной); синий цвет - исторические данные

## IV. Заключение

В рамках работы после изучения существующих методов и подходов к прогнозированию котировок финансовых инструментов был разработан подход, основанный на применении линейной и полиномиальной регрессий. Исходный алгоритм показал приемлемые результаты, но при этом обладал определенным недостатком, который был устранен путем усовершенствования алгоритма методом скользящего окна. Доработанный алгоритм продемонстрировал более точные результаты.

## Список литературы

[1] Emioma C.C., Edeki S.O. Stock price prediction using machine learning on least-squares linear regression basis. Journal of Physics: Conference Series. 2021, vol. 1734, no. 1. DOI: 10.1088/1742-6596/1734/1/012058

[2] Sun Z., Zhao S. Machine learning in stock price forecast. E3S Web of Conferences. 2020, vol. 214. DOI: 10.1051/e3sconf/202021402050

[3] Liu Y. Novel volatility forecasting using deep learning–Long Short Term Memory Recurrent Neural Networks. Expert Systems with Applications. 2019, vol. 132, pp. 99-109. DOI: 10.1016/j.eswa.2019.04.038

[4] Breiman L. Random Forests. Machine Learning. 2001, vol. 45, pp.5–32. DOI: 10.1023/A:1010933404324

[5] Nunno L. Stock Market Price Prediction Using Linear and Polynomial Regression Models: http://www.lucasnunno.com/assets/docs/ml_paper.pdf (дата обращения 09.03.2021)

[6] Naeini M.P., Taremian H., Hashemi H.B. Stock market value prediction using neural networks. 2010 International Conference on Computer Information Systems and Industrial Management Applications. 2010, pp. 132-136. DOI: 10.1109/CISIM.2010.5643675

[7] Manurung A.H., Budiharto W., Prabowo H. Algorithm and modeling of stock prices forecasting based on long short-term memory (LSTM). ICIC Express Letters. 2018, vol. 12, no. 12, pp. 1277-1283. DOI: 10.24507/icicel.12.12.1277

[8] Jeon S., Hong B., Kim J., Lee H.-J. Stock price prediction based on stock big data and pattern graph analysis. Proceedings of the International Conference on Internet of Things and Big Data. 2016, pp. 223-231. DOI: 10.5220/0005876102230231

[9] Milosevic N. Equity forecast: Predicting long term stock price movement using machine learning: https://arxiv.org/ftp/arxiv/papers/1603/1603.00751.pdf (дата обращения 09.03.21)