Сравнение моделей машинного обучения для прогнозирования курса акций

Comparison of Machine Learning Models for Stock Price Forecasting

DOI: 10.12737/2306-627X-2024-13-3-46-51 Получено: 30 мая 2024 г. / Одобрено: 11 июня 2024 г. / Опубликовано: 25 сентября 2024 г.

Борцова Д.Э.

Канд. экон. наук, доцент кафедры безопасности и информационных технологий, Национальный исследовательский университет «МЭИ», г. Москва
e-mail: BortsovaDinE@mpei.ru

Алёшечкина Т.А.

Студентка 3-го курса группы ИЭ-65-21, Национальный исследовательский университет «МЭИ», г. Москва
e-mail: taisiaaleshechkina@mail.ru

Bortsova D.E.

Candidate of Economic Sciences, Associate Professor, Department of Security and Information Technology, National Research University "Moscow Power Engineering Institute", Moscow
e-mail: BortsovaDinE@mpei.ru

Alyoshechkina T.A.

3rd year Student of the IE-65-21 Group, National Research University "Moscow Power Engineering Institute", Moscow
e-mail: taisiaaleshechkina@mail.ru

Аннотация

В статье представлен сравнительный анализ моделей машинного обучения для прогнозирования курса акций. Охарактеризован процесс алгоритмического трейдинга. Рассмотрено использование искусственного интеллекта на фондовом рынке, преимущества и недостатки его применения. Выбраны модели машинного обучения: линейная регрессия и случайный лес, дана их характеристика. Определены метрики для оценки качества прогнозов и представлено их математическое описание. Выполнено обучение и тестирование моделей, получены прогнозируемые значения, найдены необходимые метрики. Все расчеты, анализ, машинное обучение выполнены в среде программирования Python с подключение библиотек Pandas, Numpy, Matplotlib, Sklearn. В результате модель случайного леса оказалась наиболее надежной с учетом высокой точности и минимизации ошибок, для модели линейной регрессии среднеквадратическая ошибка и средняя абсолютная ошибка больше почти на 90%.

Ключевые слова: линейная регрессия, случайный лес, машинное обучение, прогнозирование котировок, алгоритмический трейдинг.

Abstract

The article presents a comparative analysis of machine learning models for stock price forecasting. The process of algorithmic trading is characterized. The use of artificial intelligence in the stock market, the advantages and disadvantages of its application are considered. The models of the machine learning model are selected: linear regression and random forest, and their characteristics are given. Metrics for assessing the quality of forecasts are defined and their mathematical description is presented. Training and testing of models were performed, predicted values were obtained, and the necessary metrics were found. All calculations, analysis, and machine learning are performed in the Python programming environment using the Pandas, Numpy, Matplotlib, and Sklearn libraries. As a result, the random forest model turned out to be the most reliable, taking into account high accuracy and error minimization, for the linear regression model, the standard error and the average absolute error are almost 90% greater.

Keywords: linear regression, random forest, machine learning, quote forecasting, algorithmic trading.

1. ВВЕДЕНИЕ

Прогнозирование временных рядов — это процесс определения будущих значений временного ряда на основе прошлых и текущих данных. Оно заключается в анализе исторических данных, выявлении закономерностей и использовании математических моделей для предсказания будущих значений временного ряда.

Временной ряд представляет собой последовательность упорядоченных во времени числовых показателей, характеризующих уровень состояния и изменения изучаемого явления. Прогнозирование временных рядов используется в экономике, бизнесе и других сферах для предсказания будущих тенденций и принятия обоснованных решений.

Цель научной статьи — провести сравнительный анализ моделей прогноза для исследования стоимости курса акций и определить наиболее подходящую модель для прогнозирования динамики рынка акций.

Задачи научной статьи:

- провести обзор регрессионных моделей прогноза стоимости акций, используемых в машинном обучении;
- охарактеризовать рынок акций;
- выбрать датасет, проанализировать его, выполнить предварительную обработку;
- построить прогнозы с помощью моделей;
- сравнить результаты;
- оценить перспективы развития рынка акций;

Актуальность научной статьи обусловлена сложностью прогнозирования и оценки динамики рынка акций.

Распространённые модели для прогнозирования цен на акции включают линейную регрессию и случайный лес.

Линейная регрессия определяет зависимость между целевой переменной (ценой акции) и одним или несколькими факторами (индексами, финансовыми показателями и т. д.). Она широко используется в финансовых моделях.

Случайный лес — это метод, объединяющий несколько деревьев решений в единую модель. Каждое дерево обучается на подмножестве данных и выборе случайного подмножества факторов. Конечный прогноз определяется усреднением прогнозов всех деревьев.

В данном исследовании сравниваются модели линейной регрессии и случайного леса. Критериями для сравнения являются значения ошибок прогноза.

2. МЕТОДЫ ИССЛЕДОВАНИЯ

Для проведения исследования в области определения наиболее подходящей модели для прогнозирования стоимости акций использованы методы систематизации данных (системный метод), сравнительного анализа на основе полученной точности и достоверности результатов, применения математических моделей для предсказания будущих значений временного ряда, анализ исторических данных для выявления закономерностей и зависимостей, а также использование машинного обучения.

3. РЕЗУЛЬТАТЫ

Алгоритмическая торговля или алгоритмический трейдинг — процесс совершения торговых операций на финансовых рынках по заданному алгоритму с использованием специализированных компьютерных систем — торговых роботов [4, с. 18].

В последние несколько лет динамику фондового рынка можно назвать нацеленную на цифровизацию процесса совершения сделок с ценными бумагами, при этом инвесторам приходится «отрабатывать» все более мелкие рыночные колебания, использовать в своей работе все более короткие временные интервалы (в том числе внутридневные), т.е. следует переходить от пассивного инвестирования к активному трейдингу.

Системная торговля предполагает осуществление операций в соответствии с некоторым набором правил для входа и выхода из позиции. Если правила торговой системы четко сформулированы, то в 90% случаев такую систему можно автоматизировать. Достоинства и недостатки применения торгового робота представлены ниже (табл. 1).

Таблица 1

Достоинства и недостатки применения торгового робота

Достоинства	Недостатки
Экономия времени	Зависимость от алгоритма
Снижение эмоционального влияния	Невозможность использования фундаментального анализа
Помощь новичкам	Риск технических сбоев
Высокая скорость и непрерывность работы	Необходимость системного контроля

Источник: составлено авторами на основе источников [4; 6].

По данным Московской биржи, в настоящее время, алгоритмический торговый оборот на фондовом рынке превышает 50%. В то же время на срочном рынке Московской биржи доля роботизированных операций в объеме торгов составляет порядка 60% (из них высокочастотных — около 45%). На валютной секции наблюдается схожая ситуация: *HFT* обеспечивали порядка 65% от объема торгов. Также стоит отметить, что многие российские банки — Сбербанк, ВТБ, Альфа-Банк, Росбанк и др. — предоставляют услугу покупки ценных бумаг и имеют интеграцию с роботами-советниками.

Алгоритмическая торговля стала неотъемлемой частью современного финансового рынка, обеспечивая высокую скорость и эффективность проведения операций. Она позволяет институциональным инвесторам и крупным клиентам брокеров совершать сделки большого объёма без риска потерь. Алгоритмические стратегии, такие как *TWAP* (*Time Weighted Average Price*), *VWAP* (*Volume Weighted Average Price*) и *Iceberg*, помогают равномерно исполнять заявки и минимизировать влияние на рынок.

Машинное обучение — это процесс, когда компьютеры, анализируя данные, учатся самостоятельно формировать прогнозы и выполнять задачи, обычно требующие человеческого участия.

На фондовых рынках машинное обучение используется для анализа временных рядов, прогнозирования цен на акции, оптимизации управления портфелем и обнаружения аномалий [6, с. 8].

Основные методы машинного обучения, применяемые на фондовых рынках, следующие:

1) прогнозирование стоимости ценных бумаг: алгоритмы анализируют исторические данные ценных бумаг и строят модели для предсказания будущих цен;
2) портфельное управление: алгоритмы помогают инвесторам выбирать комбинации активов для максимизации ожидаемой доходности при заданном уровне риска;
3) обнаружение аномалий: машинное обучение помогает обнаруживать аномалии в финансовых данных, такие как мошенничество с кредитными картами или непредсказуемые изменения рынка.

Преимущества машинного обучения на фондовых рынках заключаются в обработке больших объёмов данных, автоматизации процессов и улучшении точности прогнозирования. Однако финансовые рынки могут быть подвержены непредсказуемым событиям, которые могут исказить результаты моделей. Основными недостатками применения искусственного интеллекта в данной сфере являются возможность возникновения сильной волатильности цен на рынке, если все трейдеры будут использовать исключительно искусственный интеллект, сложности в предотвращении взломов платформ и их восстанов-

лении после сбоев, высокая стоимость разработки качественных алгоритмов, ботов и платформ, что может ограничивать их доступность для многих трейдеров [8].

На сегодняшний день выделяют следующие основные этапы и подходы машинного обучения для прогнозирования стоимости акций:

1) *сбор и подготовка данных.* Собираются исторические данные о ценах акций, объёмах торгов, макроэкономических показателях и других факторах, влияющих на стоимость акций. Данные очищаются от ошибок и нормализуются;
2) *выбор алгоритма машинного обучения*. Используются различные методы, такие как линейная регрессия, деревья решений, случайный лес, градиентный бустинг, нейронные сети и другие. Выбор алгоритма зависит от специфики задачи и предпочтений аналитика;
3) *обучение модели.* На основе собранных данных и выбранного алгоритма создаётся модель, которая будет прогнозировать стоимость акций. Модель обучается на исторических данных, чтобы учесть взаимосвязь между различными факторами и стоимостью акций;
4) *тестирование и оптимизация модели*. После обучения модели проводится тестирование на новых данных, чтобы проверить её качество и точность. Если необходимо, модель оптимизируется путём настройки гиперпараметров, добавления или удаления переменных и других методов;
5) *применение модели для прогнозирования*. После успешной проверки и оптимизации модель готова к использованию для прогнозирования стоимости акций. Аналитик может использовать эту модель для определения потенциальных точек входа и выхода из рынка, а также для формирования инвестиционных стратегий;
6) *мониторинг и обновление модели*. Со временем рыночная ситуация может меняться, поэтому модель необходимо регулярно обновлять и адаптировать к новым условиям. Это позволит сохранять её актуальность и повышать точность прогнозов.

Машинное обучение играет важную роль в прогнозировании стоимости акций на фондовых рынках, позволяя аналитикам и инвесторам принимать обоснованные решения на основе объективных данных и статистических моделей.

Модель линейной регрессии в машинном обучении — это математическая модель, которая описывает связь между несколькими переменными.

Линейная регрессия выражается уравнением вида

$$f(x) = b + m * x,$$

где m — наклон линии; b — смещение по оси Y.

Изменение коэффициентов m и b влияет на расположение прямой на графике, а оптимальное значение определяется с помощью функции потерь, которая минимизирует расстояние между объектами и прямой [2, с. 51–56].

В машинном обучении линейная регрессия используется для решения задач классификации и регрессии, а также для создания искусственных нейронных сетей и глубокого обучения.

Модель линейной регрессии может быть использована для прогнозирования стоимости акций с учётом различных факторов, влияющих на их стоимость. Однако следует отметить, что результаты прогнозирования могут быть неточными из-за сложности финансовых рынков и влияния множества факторов на стоимость акций.

Модель случайного леса в машинном обучении — это ансамблевая модель, основанная на методе бэггинга — (метаалгоритм, предназначенный для улучшения стабильности и точности алгоритмов машинного обучения, уменьшающий дисперсию и помогающий избежать переобучения). Модель использует множество решающих деревьев для классификации или регрессии данных. Дерево решений — это средство поддержки принятия решений, используемое в машинном обучении, анализе данных и статистике. Оно представляет собой структуру, состоящую из «листьев» и «веток». На рёбрах дерева записаны признаки, от которых зависит целевая функция, а в «листах» записаны значения этой функции. Каждый лист представляет собой значение целевой переменной, изменённое в ходе движения от корня по рёбрам дерева до листа [7].

При построении дерева решений используются различные алгоритмы, такие как *ID3, C4.5* и *CART*. *ID3* основан на информационной энтропии, *C4.5* улучшает предыдущий метод, позволяя работать с числовыми атрибутами, а *CART* строит бинарные деревья решений. Эти алгоритмы выбирают признаки для разделения на основе прироста информации или нормализованного прироста информации, что позволяет создать оптимальное решающее дерево [3, с. 45–58].

Для достижения цели и определения наиболее точной модели прогнозирования будем использовать датасет, содержащий 1826 данных с 23.11.2015 по 20.11.2020. Датасет разделен на 7 семь параметров: дата, максимальная стоимость, минимальная стоимость, стоимость при открытии, стоимость при закрытии, объём, скорректированные значения. Про-

гнозирование будет выполнено для параметра скорректированные значения стоимости акции. Набор исходных данных разделён на обучающую и тестовые выборки в соотношении 80% к 20%. Исследование будет выполнятсявыполняться с использованием языка программирования *Python*, подключены библиотеки *Pandas, Numpy, Matplotlib, Sklearn*.

Библиотека *Pandas* используется для анализа и обработки данных, особенно в контексте работы с табличными данными. Применяется для загрузки, очистки, преобразования и хранения данных [1, с. 36].

Numpy — библиотека для работы с многомерными массивами и матрицами. Используется для выполнения математических операций над большими наборами данных, таких как линейная алгебра, случайные числа и быстрое преобразование Фурье.

Matplotlib — библиотека для создания визуализаций данных, таких как графики, диаграммы и изображения. Позволяет создавать статические и интерактивные графики, а также настраивать их внешний вид и параметры отображения.

Библиотека *Scikit-learn* применяется для машинного обучения, которая она предоставляет инструменты для классификации, регрессии, кластеризации и уменьшения размерности данных. В и включает в себя различные алгоритмы и методы машинного обучения, а также инструменты для оценки производительности моделей.

Для оценивания точности получившихся прогнозов будут найдены *RMSE, MAE, MAPE,* R^2.

RMSE (среднеквадратическая ошибка) — это метрика, которая измеряет среднее расстояние между прогнозами модели и фактическими значениями. Она используется для оценки качества моделей регрессии и имеет преимущество перед *MSE* (средней квадратической ошибкой) в том, что её значение легче интерпретировать.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2},$$

где n — количество наблюдений по которым строится модель и количество прогнозов, y_i — фактические значение зависимой переменной для i-го наблюдения, $\hat{y}_i$ — значение зависимой переменной, предсказанное моделью.

MAE (средняя абсолютная ошибка) — это метрика, которая вычисляется как среднее абсолютных разностей между наблюдаемыми и предсказанными значениями. *MAE* используется для оценки качества моделей регрессии и является линейной оценкой, что означает, что все ошибки в среднем взвешены одинаково.

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

MAPE (средняя абсолютная процентная ошибка) — это метрика, которая измеряет отклонение прогнозов от фактических значений в процентах. Она используется для оценки качества моделей регрессии.

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

R^2 (коэффициент детерминации) — это метрика, которая измеряет долю вариации зависимой переменной, объясняемую независимыми переменными в модели. Она используется для оценки адекватности модели и для сравнения моделей с одинаковыми данными [5, с. 239–241].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(\bar{y}_i - y_i\right)^2}.$$

Приступим к созданию и обучению моделей (рис. 1).

Далее выполним прогнозирование. Для наглядности прогноз будет строиться как на обучающей выборке, так и на тестовой (рис. 3, 4).

```
'''Создание и обучение необходимых моделей'''
'''Линейная регрессия'''
LinearRegression_model = LinearRegression()
LinearRegression_model.fit(X_train, Y_train)
'''Случайный лес'''
RandomForestRegressor_model = RandomForestRegressor()
RandomForestRegressor_model.fit(X_train, Y_train)

# Прогнозирование на тестовой выборке
dict_models_data["LinearRegression"]["Y_predict"] = LinearRegression_model.predict(predict_data)
dict_models_data["RandomForestRegressor"]["Y_predict"] = RandomForestRegressor_model.predict(predict_data)
```

Рис. 1. Отрывок программного кода

Источник: составлено авторами.

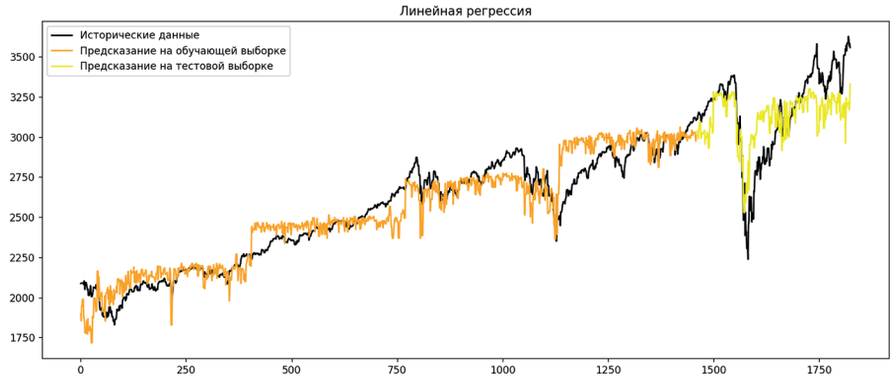

Рис. 3. Прогнозирование с помощью модели линейной регрессии

Источник: составлено авторами.

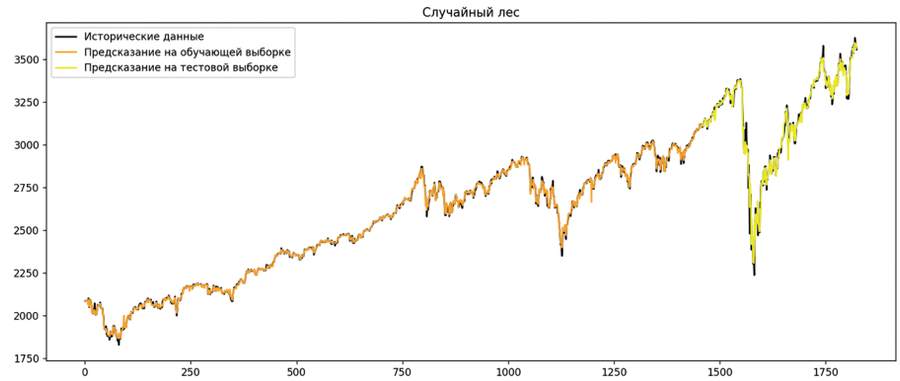

Рис. 4. Прогнозирование с помощью модели случайного леса

Источник: составлено авторами.

По результатам прогнозирования, изображенных изображенным на графиках, можно сделать вывод, что наилучший прогноз получился при использовании модели случайного леса.

Определим точность полученного прогноза. Для этого вычислим метрики для оценки качества модели (рис. 5).

```
Модель Линейная регрессия:
        Оценка R2 = 0.44099562278394744
        Оценка MAE = 165.34274243165368
        Оценка RMSE = 203.71521917776374
        Оценка MAPE = 0.05387616338118937


Модель Случайный лес:
        Оценка R2 = 0.9915760419384695
        Оценка MAE = 14.752357602472165
        Оценка RMSE = 25.007712580963553
        Оценка MAPE = 0.004860541272782921
```

Рис. 5. Вычисление метрик точности полученных моделей

Источник: составлено авторами.

Исходя из полученных оценок, можно сделать вывод, что наиболее высокий результат показала модель случайного леса.

4. ОБСУЖДЕНИЕ И ЗАКЛЮЧЕНИЕ

Модель случайного леса предсказывает стоимость акций точнее, чем линейная регрессия, потому что она формирует множество независимых алгоритмов, которые охватывают различные возможные исходы для каждого входного вектора. Это позволяет деревьям принимать разнообразные решения и описывать разные исходы для входных векторов. При усреднении результатов эффект переобучения естественным образом нивелируется, и итоговое выходное значение оказывается достаточно точным и устойчивым к отдельным выбросам.

Литература

1. *Богатырев С.Ю.* [и др.]. Машинное обучение в финансах: учебник [Электронный ресурс]. — М.: Прометей, 2024. — 224 с.
2. *Кремер Н.Ш.* Эконометрика [Текст] / Н.Ш. Кремер, Б.А. Путко. — М.: ЮНИТИ-ДАНА, 2010. — 328 с.
3. *Лимановская О.В.* Основы машинного обучения [Текст] / О.В. Лимановская, Т.И. Алферьева. — Екатеринбург: Изд-во Урал. ун-та, 2020. — 88 с.
4. *Малыхин Е.М.* Алгоритмический трейдинг для профессионалов [Текст] / Е.М. Малыхин. — СПб.: БХВ-Петербург, 2021. — 176 с.
5. *Харрисон М.* Машинное обучение: карманный справочник «Краткое руководство по методам структурированного машинного обучения на Python» [Текст] / М. Харрисон. — СПб.: Диалектика, 2020. — 320 с.
6. *Янсен С.* Машинное обучение для алгоритмической торговли на финансовых рынках. Практикум [Текст] / С. Янсен. — СПб.: БХВ-Петербург, 2019. — 560 с.
7. *Krauss C., Do X.A., Huck N.* Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research (2016), 259(2), 689–702. DOI: 10.1016/j.ejor.2016.10.031/ URL: https://www.sci-hub.ru/10.1016/j.ejor.2016.10.031?ysclid=lwtkxebont933507928 (accessed 28 May 2024).
8. *Hong K.* Modelling Intervalling Effect of High Frequency Trading on Portfolio Volatility. Theoretical Economics Letters, October 2019, vol. 9, no. 7. URL: https://www.scirp.org/journal/paperinformation?paperid=95423 (accessed 28 May 2024).

References

1. Bogatyrev S.Y., Pomulev A.A., Zatevakhina A.V., Kruglova I.A., Barabanova M.I., Tegin A.V., Solodovnikov M.A., Shashina I.A., Matrosov S. V. Mashinnoe obuchenie v finansah [Machine learning in finance]. [Electronic resource]: textbook. M.: Prometheus Publ., 2024. 224 p.
2. Kremer N.S., Putko B.A. Econometrica [Econometrica]. Moscow: UNITY-DANA Publ., 2010. 328 p.
3. Limanovskaya O.V., Alferyeva T.I. Osnovy mashinnogo obucheniya [Fundamentals of machine learning]. Yekaterinburg: Ural Publishing House. unita Publ., 2020. 88 p.
4. Malykhin E.M. Algoritmicheskij trejding dlya professionalov [Algorithmic trading for professionals]. St. Petersburg: BHV-Petersburg Publ., 2021. 176 p.
5. Harrison M. Mashinnoe obuchenie: karmannyj spravochnik Kratkoe rukovodstvo po metodam strukturirovannogo mashinnogo obuchenii na Python [Machine learning: a pocket guide A brief guide to methods of structured machine learning in Python]. St. Petersburg: Dialectics LLC, 2020. 320 p.
6. Jansen S. Hands-On Machine Learning for Algorithmic Trading BIRMINGHAM – MUMBAI: Packt, 2018. 484 p. (Russ. ed.: Jansen S. Mashinnoe obuchenie dlya algoritmicheskoj torgovli na finansovyh rynkah. Praktikum.. St. Petersburg: BHV-Petersburg Publ., 2019. 560 p.).
7. Krauss C., Do X. A., Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research (2016), 259(2), 689–702. DOI: 10.1016/j.ejor.2016.10.031/ URL: https://www.sci-hub.ru/10.1016/j.ejor.2016.10.031?ysclid=lwtkxebont933507928 (accessed 28 May 2024).
8. Hong K. Modeling Intervalling Effect of High Frequency Trading on Portfolio Volatility. Theoretical Economics Letters, October 2019, vol. 9, no. 7. URL: https://www.scirp.org/journal/paperinformation?paperid=95423 (accessed 28 May 2024).