

Отбор признаков и их влияние на модели машинного обучения для прогнозирования цен акций.

Введение

Прогнозирование цен акций – задача, усложненная высоким уровнем шума, волатильности и множеством факторов влияния ([Predicting Economic Trends and Stock Market Prices with Deep Learning and Advanced Machine Learning Techniques](#)). Для улучшения точности моделей машинного обучения (ML) исследователи привлекают различные группы признаков: *технические индикаторы* (информацию, извлекаемую из динамики цен и объемов), *макроэкономические показатели* (характеризующие состояние экономики и рынков), *фундаментальные данные* компаний (финансовые коэффициенты и отчетность) и *новостной фон/тональность* (настроения инвесторов по данным СМИ и соцсетей). Однако избыток признаков может приводить к переобучению, поэтому важны методы отбора наиболее информативных признаков. В данном обзоре последовательно рассмотрены популярные типы признаков, способы их формирования для подачи в модель, методы feature selection и оценка вклада разных факторов, а также примеры исследований по прогнозированию фондового рынка. Мы проанализируем, какие признаки доказали свою эффективность (и на каких горизонтах прогнозирования), а также сравним результаты различных работ, включая случаи, где добавление сложных данных (например, новостей) не приносило пользы. Обсудим метрики, которыми оценивают качество моделей и вклад признаков, и укажем ограничения подобных исследований.

Технические индикаторы

Понятие и популярные индикаторы. Технические индикаторы (ТИ) вычисляются на основе исторических цен (и иногда объемов) и служат для выявления паттернов – трендов, моментов перекупленности/перепроданности, волатильности и т.д. Наиболее распространены осцилляторы и скользящие средние: например, RSI (индекс относительной силы), стохастический осциллятор, скользящие средние (простая SMA и экспоненциальная EMA) ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Широко применяются индикаторы тренда и инерции, такие как MACD (конвергенция/дивергенция скользящих средних) и полосы Боллинджера (диапазон волатильности), а также индикаторы объема (OBV) и другие. В частности, в работах отмечается, что стохастический осциллятор (KDJ), MACD, Bollinger Bands и RSI оказываются «очень важными» для краткосрочного прогнозирования цен акций

[\(\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach\)](#).

Эти индикаторы популярны, поскольку отражают повторяющиеся поведенческие эффекты на рынке – импульсы, перекосы спроса/предложения и т.п., – и могут служить готовыми признаками для алгоритмов ML.

Использование и эффективность. В краткосрочном трейдинге (интрадей, горизонтом в дни или недели) технические индикаторы зачастую улучшают результаты модели по сравнению с использованием только «голых» цен и объемов. Они предоставляют модели сводную информацию о направлении тренда или силе движения. Например, добавление стандартных индикаторов к историческим данным повышает точность классификации направления движения акций – одна из работ показала, что модель на фундаментальных показателях давала 64,4% правильных прогнозов направления против ~62,8% у модели только на ценовых трендах ([Paper Title \(use style: paper title\)](#)) (в данном случае техническая модель была немного слабее фундаментальной, см. ниже). В целом, комбинация ценовых рядов с техническими индикаторами – базовый подход в прогнозирующих моделях ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Тем не менее эффективность индикаторов зависит от частоты данных и сложности модели. Так, при прогнозировании на минутных интервалах (high-frequency trading) было обнаружено, что включение множества индикаторов может приводить к переобучению: хотя в обучающей выборке риск-доходность улучшалась, на тестовых данных модель теряла обобщающую способность ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)). В исследовании на минутных данных S&P500 (SPY) с 13 техническими индикаторами (полосы Боллинджера, ЕМА, уровни Фибоначчи и др.) улучшались некоторые метрики риска, но *feature importance*-анализ показал, что простые ценовые признаки (например, последние изменения цены) переигрывают сложные индикаторы, т.е. вклад технических индикаторов в высокочастотной торговле оказался ограниченным ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)). Эти результаты намекают, что для очень коротких интервалов модель сама может извлечь нужные шаблоны из сырого ценового ряда, а добавление индикаторов иногда лишь дублирует информацию и увеличивает риск переобучения.

Ограничения и отбор технических признаков. Существует огромное число технических индикаторов, и не все из них независимы. Многие сильно коррелируют между собой (напр., различные типы скользящих средних). Исследования

указывают, что избыток индикаторов может ухудшить результаты модели. Alsubaie et al. (2019) отмечают: использование более ~30 технических индикаторов начинает снижать точность классификации на их данных ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Поэтому важен отбор ключевых индикаторов. Например, метод отбора *Relief* (фильтр, основанный на оценке важности признаков через ближайших соседей) был признан одним из лучших для выбора подмножества индикаторов из 50 возможных ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Другой подход – рекурсивное исключение признаков (RFE): Weng et al. (2018) применили RFE на технических индикаторах в задаче внутридневного прогноза и обнаружили скрытый оптимальный набор показателей, повысив точность модели ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). В китайском рынке (Shen & Shafiq, 2020) сравнили три метода (логистическая регрессия, SVM и собственный алгоритм) для отбора из 44 индикаторов – все три подхода выбрали одинаковые 20 лучших признаков, и включение именно этих индикаторов дало максимальную точность классификации направления цен ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Таким образом, добавление технических индикаторов оправдано, но с соблюдением меры – из множества возможно избыточных индикаторов следует выбирать информативные. Стоит отметить и современный тренд: используются методы преобразования исходного ценового ряда для усиления сигналов индикаторов. Например, применение вейвлет-преобразования для разложения финансового ряда с последующей оптимизацией признаков (метод DWT-CSO) позволило добиться значительного роста точности. В эксперименте на индексах S&P500, DJI, NIFTY50 и BSE предложенная модель с обработкой (DWT + отбор оптимальных технических признаков) показала прирост точности прогноза на **~15–19%** по сравнению с базовой моделью ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Эти результаты подтверждают, что грамотно сконструированные технические признаки заметно повышают качество предсказания. В то же время, ценность технического анализа проявляется в краткосрочной перспективе. Для более длительных прогнозов (месяцы, годы) одних технических сигналов недостаточно – рыночные тенденции на больших масштабах определяются фундаментальными факторами, что требует привлечения других видов данных ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)).

Макроэкономические факторы

Основные макропоказатели и их связь с акциями. Акции компаний (особенно крупных, индексов) реагируют на изменения макроэкономической среды: состояние экономики, денежно-кредитную политику, цены на ресурсы и прочие внешние факторы ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). К типичным макропоказателям, используемым в моделях, относятся: процентные ставки (в частности, ключевая ставка центрального банка), инфляция, показатели деловой активности (ВВП, PMI), безработица, курсы основных валют, цены сырьевых товаров – нефти, газа, золота и др., а также индексы фондового рынка и волатильности (например, индекс VIX). Эти переменные влияют на прибыли компаний и настроения инвесторов. Например, рост процентных ставок обычно давит на фондовый рынок (повышаются издержки для бизнеса, становятся привлекательнее облигации), а рост цен на нефть поддерживает акции нефтегазового сектора, но может негативно сказаться на секторах-потребителях топлива. Поэтому добавление макроэкономических признаков может дать модели более широкое понимание контекста, особенно для прогнозирования индексов и долгосрочных тенденций.

Применение в моделях и эффект. Макрофакторы нередко включаются в модели прогнозирования индексов или корзин акций. Так, Latif et al. (2025) интегрировали в модель прогнозирования S&P 500 сразу несколько макроэкономических индикаторов, характеризующих неопределенность и риск: мировые цены на золото и нефть, индекс волатильности VIX, индекс экономической политической неопределенности (EPU), индекс финансового стресса (FSI), индекс геополитических рисков (GPR) и *shadow* ставку (SSR), наряду с десятью техническими индикаторами ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)) ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Используя гибридную нейросеть (CNN для выделения признаков + GRU для учета временной динамики), они добились заметного повышения точности прогноза по сравнению с моделями на одних только технических показателях ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)) ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Более того, анализ важности признаков методом Шэпли (SHAP) показал, что *самыми значимыми предикторами* оказались именно макроиндикаторы – индекс волатильности VIX, индекс финансового стресса

и индекс неопределенности политики ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Это подчеркивает огромную роль внешних факторов неопределенности: в современный турбулентный период они существенно влияют на фондовый рынок, и модель, учитывающая такие переменные, лучше справляется с прогнозом ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)) ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)).

Другой пример – включение ожиданий по будущей макрополитике. **Haque et al. (2023)** предложили улучшать прогноз акций за счет учета *ожидаемых* изменений экономической политики ([\[2311.06278\] Boosting Stock Price Prediction with Anticipated Macro Policy Changes](#)). Их идея: инвесторы часто заранее переоценивают акции, предвосхищая, скажем, будущие решения регуляторов или макроусловия, поэтому модель должна знать не только текущие макроданные, но и прогнозы/ожидания. Авторы добавили в признаки информацию о грядущих вероятных изменениях (например, ожидание повышения ставки) и показали, что такая модель превосходит обычный подход. **Во всех протестированных ML-моделях** включение макропоказателей улучшило результаты; лучшая модель дала RMSE=1,61 против 1,75 у базовой (без учета ожидаемых макроизменений) ([\[2311.06278\] Boosting Stock Price Prediction with Anticipated Macro Policy Changes](#)), то есть ошибка снизилась примерно на 8%. Эти результаты статистически значимы и подтверждают пользу макроэкономической информации для прогнозирования цен акций ([\[2311.06278\] Boosting Stock Price Prediction with Anticipated Macro Policy Changes](#)). Особенно актуально это для периодов, когда инвесторы кардинально меняют ожидания по экономике – модель, “чувствующая” эти ожидания, улавливает соответствующие движения цен.

Помимо глобальных индикаторов, могут учитываться специфические макропеременные, связанные с деятельностью компании. Например, курс национальной валюты важен для экспортеров (укрепление валюты может снизить выручку в пересчете), а цены на металлы – для металлургических компаний. В моделях для отдельных отраслей добавляют профильные индексы: в прогнозировании акций авиакомпаний могут включаться цены на керосин, в прогнозах банков – показатели денежной массы и учетной ставки и т.д. Макропоказатели зачастую обновляются реже (месяц, квартал) и публикуются с задержкой, поэтому при их использовании следует соблюдать хронологическую

корректность (например, не использовать показатель ВВП Q2 для прогноза цен в середине Q2, пока он не опубликован). В моделях же “скользящего” прогноза обычно просто предполагается, что последнее известное значение макропоказателя актуально вплоть до выхода нового.

Влияние макрофакторов на точность. В исследованиях в целом присутствует консенсус, что макроэкономический фон значим прежде всего для среднесрочных и долгосрочных прогнозов и для широкого рынка. Добавление макропеременных чаще всего увеличивает объясненную дисперсию модели. Так, интеграция макроиндикаторов в нейросетевую модель позволила *увеличить точность прогноза доходностей индекса* по сравнению с моделью на одних технических данных ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). В кросс-страновом исследовании (Chen et al., 2021) отмечается, что учет глобальных макропеременных (цен сырья, мировых индексов) улучшает прогнозирование локальных рынков, особенно в периоды внешних шоков. Тем не менее бывают ситуации, когда макрофакторы не дают выигрыша – например, для краткосрочных прогнозов отдельных акций, сильно зависящих от внутренних новостей компании, либо в стабильных макроусловиях (когда ключевые показатели почти не меняются, их влияние уже “учтено” рынком). В отсутствие крупных макросдвигов основными драйверами цены будут фирменные события и рыночный сентимент. В целом же современные работы показывают, что **игнорировать макроэкономический фон нельзя**: он создает общий “ветер” для рынка. В частности, неопределенность (волатильность, политический риск) часто становится главным фактором движений – так, метод **SHAP** выявил, что примерно половина влияния на прогноз **S&P500** в 2020-е годы приходилась на показатели неопределенности (**VIX, EPU, FSI**) ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Поэтому включение макроданных обычно повышает стабильность и реалистичность моделей фондового рынка.

Фундаментальные данные компаний

Сущность и примеры фундаментальных признаков. Фундаментальные данные – это финансовые показатели компании, отражающие ее бизнес и эффективность. Их источник – бухгалтерская отчетность (баланс, отчет о прибылях и убытках и др.), корпоративные события и показатели корпоративных действий. В модели ML такие данные обычно входят в виде финансовых коэффициентов и индикаторов, например: отношение цены акции к прибыли на акцию (P/E), к

балансовой стоимости (P/B), рентабельность капитала (ROE) и активов (ROA), маржа прибыли, темпы роста выручки или прибыли, размер и рост EBITDA, долговая нагрузка (Debt/Equity), текущий коэффициент ликвидности, размер и доходность дивидендов, *free float* (доля акций в свободном обращении) и т.д. Эти показатели помогают оценить “внутреннюю стоимость” компании и финансовое здоровье, давая модели информацию о том, недооценена или переоценена акция относительно фундаментальных ориентиров.

Фундаментальный анализ vs технический: горизонты применения. В классических финансах фундаментальный анализ применяется для средне- и долгосрочных прогнозов – на горизонтах кварталов и лет ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Он фокусируется на оценке справедливой стоимости акций на основе бизнес-показателей и предполагает, что рынок в перспективе придет к этой стоимости. Поэтому фундаментальные признаки наиболее полезны для прогнозирования трендов цен на интервале от одного квартала до нескольких лет. В краткосрочном же периоде (дни-недели) влияние фундаментальных метрик невелико: отчетность выходит раз в квартал, и между релизами показатели неизменны, а ежедневные колебания вызваны в основном рыночным сентиментом и техническими факторами. Это подтверждается обзором: последние исследования по ML-прогнозированию в основном концентрируются на технических индикаторах и настройках, тогда как **финансовые коэффициенты компаний используются реже** ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)) – видимо, потому что многие такие работы решают задачу краткосрочного трейдинга. Однако для более длительных прогнозов интеграция фундаментальных данных может существенно повысить точность. Khairi et al. (2019) отмечают, что **технические индикаторы хорошо работают на коротких отрезках, но “поймать” долгосрочные колебания цены по одним только техническим данным очень трудно**; чтобы учесть долгосрочные движения, нужно привлекать “core” данные компаний (т.е. фундаментальные) ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)). Объединив технические и фундаментальные признаки, авторы смогли построить модель на алгоритме решающих деревьев (J48 с бэггингом), которая дала более высокую точность прогнозов по акциям по сравнению с моделями на одном техническом анализе ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)) ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)).

Методы включения фундаментальных признаков. В практических моделях фундаментальные показатели обычно кодируются на тот же временной шкале, что и данные цен. Например, можно добавить переменные “текущий P/E” или “долг/EBITDA” ко входам модели каждый день – при этом их значение обновляется лишь когда выходит новый отчет (раз в квартал). Таким образом, модель получает ступенчатый сигнал, постоянный между отчетными датами. Альтернативно, прогнозируют не ежедневную цену, а, скажем, доходность акции за квартал вперед на основе фундаментальных данных текущего квартала – этот подход ближе к традиционному фундаментальному анализу акций (выбор недооцененных акций). Некоторые исследования преобразуют фундаментальные данные для удобства модели: например, нормируют показатели относительно отраслевых средних (чтобы модель понимала, насколько компания лучше/хуже рынка), или применяют логарифмические преобразования к финансовым величинам (капитализации, прибыли) для снятия масштабных эффектов. Также практикуется дискретизация – Namdari & Li (2018) для прогноза тренда акций high-tech сектора **дискретизировали 12 финансовых коэффициентов** для подачи в нейросеть ([Paper Title \(use style: paper title\)](#)).

Результаты: фундаментальные данные в ML-моделях. Ряд работ демонстрирует успешность использования фундаментальных признаков. Huang et al. (2022) собрали **22 года квартальных финансовых данных** по американским компаниям (по всей видимости, данные S&P 500 с ~1995 по 2017) и протестировали три алгоритма ML – полносвязную нейросеть (FNN), случайный лес (RF) и нечеткую систему ANFIS – для прогнозирования доходности акций на основе исключительно фундаментальных показателей ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Они применили RF-отбор признаков и бэггинг ансамблей, добившись улучшения качества моделей. Лучший результат показал Random Forest, за ним – ансамбль моделей. Важно, что их **комбинированная модель на фундаментальных данных переиграла бенчмарк рынка (индекс Dow Jones)** по доходности на тестовом периоде ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Авторы делают вывод: модели ML вполне могут помочь фундаментальным аналитикам в отборе акций, выявляя бумаги с перспективой опережающего роста ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)).

Интересное сравнение фундаментального и технического подхода приведено Namdari & Li (2018). Они обучили одну нейросетевую модель предсказывать тренд акций только по **12 финансовым коэффициентам**, а другую

– только по историческим ценам (технический подход) за тот же период и набор компаний (NASDAQ tech-сектор, 2012–2017). Результат: модель на фундаментальных данных показала *долю правильных прогнозов 64,38%*, тогда как чисто техническая – 62,84% ([Paper Title \(use style: paper title\)](#)). То есть фундаментальные признаки дали небольшой, но ощутимый выигрыш (~1,5 п.п. точности) в предсказании направления движения акций за период квартала. В рамках долгосрочного инвестирования такой прирост может быть важным.

Конечно, эффективность фундаментальных факторов проявляется не во всех ситуациях. В краткосроке (скажем, прогноз на завтра) их влияние близко к нулю – цена за день не успеет среагировать на изменение P/B, которого и не было за день. Фундаментальные данные приносят пользу, когда прогнозный горизонт соизмерим со временем, за которое рынок переоценит акцию под влиянием новых финансовых данных. Например, при прогнозе доходности на год вперед фундаментальные метрики (особенно оценочные коэффициенты – P/E, P/B) оказываются сильными предикторами: существует обширная литература по факторным стратегиям (value, quality), где акции с низким P/E или высоким ROE статистически опережают рынок на горизонте 6–12 месяцев. ML-модели могут автоматизировать учет этих факторов и улучшить их комбинацию. В одном из недавних исследований **Bohn (2018)** попытался объединить технический, фундаментальный и новостной подходы: на выборке ~1500 акций из S&P500 (2002–2016) он построил модели, ранжирующие акции по прогнозной доходности, и оценивал их по корреляции Спирмена между предсказанным ранком и фактическим ([Paper Title \(use style: paper title\)](#)). Модель на сочетании всех типов признаков (с фильтрацией признаков и нейросетевой архитектурой) показала качество ранжирования, сравнимое с моделью, созданной *экспертами инвестиционной фирмы* ([Paper Title \(use style: paper title\)](#)). Иными словами, ML с техническими, фундаментальными и sentiment-факторами сумел не хуже человека отобрать перспективные акции. Yu et al. (2019) также подтвердили, что **комбинация технических и фундаментальных данных** может генерировать прибыль, опережающую рынок: их алгоритм на данных Шанхайской биржи 2005–2012 создал портфель акций, существенно обыгравший бенчмарк ([Paper Title \(use style: paper title\)](#)).

Таким образом, фундаментальные признаки в моделях особенно эффективны для средне- и долгосрочных прогнозов и при отборе акций (stock picking). Для внутридневных и краткосрочных задач они обычно не используются из-

за отсутствия влияния на таких интервалах. Впрочем, даже в краткосроке косвенные фундаментальные индикаторы могут пригодиться – например, признак “было ли недавно объявлено о дивидендах или байбэке” может повлиять на настроение инвесторов в ближайшие дни. К сложным фундаментальным данным можно отнести и показатели цепочки поставок, внутренних операций компании (производство, продажи), но их редко можно достать оперативно. В эпоху больших данных некоторые фонды пытаются учитывать *альтернативные фундаментальные данные* – например, показатели продаж из сторонних источников (сканирование чеков, трафик на сайт компании, количество активных пользователей и т.п.), однако в академических работах подобное почти не встречается из-за труднодоступности таких данных.

Новостной фон и сентимент инвесторов

Роль информационного поля. Новости, слухи и публичные настроения имеют непосредственное влияние на цены акций, особенно в краткосрочной перспективе. Позитивная новость (например, о росте прибыли или заключении выгодного контракта) способна вызвать скачок цены, а негативная (скандал, убытки, отставка CEO) – падение. Даже общий новостной фон (экономические новости, геополитика) влияет на рынок через настроения участников. Поэтому *новостные данные* и *сентимент-аналитика* (sentiment analysis) все чаще используются как признаки в моделях прогнозирования. Сырые тексты новостей и сообщений соцсетей не подходят напрямую для ML-моделей, поэтому их превращают в числовые показатели – тональность (отрицательная, нейтральная, положительная), упоминания определенных событий или ключевых слов, индексы “буйности” информационного потока и т.д.

Методы обработки новостей и соцмедиа. Для анализа текстов применяются технологии NLP (Natural Language Processing). Ранние подходы опирались на словари тональности: например, словарь финансовых терминов Loughran–McDonald, где каждому слову присвоена метка “позитивное” или “негативное”. Подсчитав количество позитивных и негативных слов в статье, получают оценку тональности. Более современные методы – машинное обучение на текстах. Классический пример: обучить классификатор (например, Naïve Bayes или SVM) определять тональность заголовков новостей по размеченным данным. Теперь все большую популярность имеют модели на базе трансформеров, такие как **BERT**. Специально для финансового текста была разработана модель **FinBERT**, обученная на тональности финансовых новостей. Преимущество

трансформеров – понимание контекста, идиом, полисемии; например, для общего словаря слово “*bullish*” (бычий) может быть нейтральным или негативным, а в финансовом контексте означает позитивный настрой. FinBERT учитывает такие особенности. В одном исследовании новостные заголовки по акциям пропускали через предобученную модель FinBERT для оценки тональности, и включение этих **оценок тональности как признаков значительно улучшило результаты LSTM-сети** по прогнозированию цен в рамках дня ([Stock Price Prediction Using Sentiment Analysis - IEEE Xplore](#)). Причем показатели FinBERT оказались лучше, чем если использовать стандартный BERT, – доменная специфичность дала эффект ([Stock Price Prediction Using Sentiment Analysis - IEEE Xplore](#)).

Структурирование новостного потока зачастую требует агрегации. Популярный подход – расчет *ежедневного индекса сентимента*: например, берутся все новости по компании за день, оцениваются по тональности, затем усредняются (или берется разность доли положительных и отрицательных). Такой признак “сентимент новостей за день” затем сопоставляется с динамикой цены на следующий день. Аналогично для твитов: можно считать долю положительных твитов о данной акции за последний час. Кроме тональности, выделяют *упоминания важных событий*: например, если в тексте новости обнаружены слова “merger” (слияние) или “lawsuit” (судебный иск), можно завести бинарный признак, сигнализирующий о соответствующем событии. Более сложный вариант – тематическое моделирование: кластеризация текстов на темы (LDA и др.) и отслеживание, какие темы обсуждаются (например, “финансовые результаты”, “скандалы”, “инновации” и т.п.).

Социальные медиа и альтернативные источники мнений. Помимо новостей, огромный пласт информации – это посты и обсуждения в соцсетях (Twitter, Reddit, форумы инвесторов, блоги). Они отражают настроения розничных инвесторов и могут предвосхитить ценовые движения, особенно на отдельных акциях с сильным комьюнити. Классический пример – **Twitter-сентимент**. В 2013–2015 гг. ряд исследований (в т.ч. Bollen et al., Ranco et al.) показали наличие корреляции между общим настроением твитов и фондовым рынком. Так, Ranco et al. (2015) анализировали 15-месячный поток твитов о 30 компаниях Dow Jones. В целом корреляция тональности твитов с доходностью акций была невысока, и причинно-следственные связи по всей выборке слабы ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)). Однако при более детальном анализе обнаружилось, что **в периоды всплеска твиттер-активности** (так называемые

event peaks, когда внезапно выходит очень много твитов о компании – обычно вокруг квартальной отчетности или важной новости) наблюдается статистически значимая зависимость: преобладающая тональность твитов в такие моменты влияет на *накопленную аномальную доходность* акций после события ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)). Иными словами, когда резко возрастает внимание аудитории, **полярность мнений в соцсети указывает, в какую сторону пойдет цена**. По оценкам, если в момент события тональность в Twitter была положительной, то совокупная аномальная доходность акции за несколько дней после события составит порядка +1–2%, и наоборот для негативной тональности ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)). Хотя 1–2% – относительно небольшой эффект, он статистически значим и держится несколько дней ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)), что можно использовать в торговой стратегии. Вне таких пиковых ситуаций твиттер-поток, видимо, уже “встроен” в рыночные цены, поэтому сигнал слабый. Интересно, что это исследование применило *event study*-методологию: автоматически определяя моменты резко возросшего объема твитов, они фактически находили важные события и изучали реакцию рынка на них с учетом сентимента твитов ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)).

Другой пример влияния соцсетей – форумы типа Reddit (подфорум WallStreetBets) или поисковые тренды Google. В 2021 году массовое обсуждение на Reddit привело к взрывному росту акций GameStop и AMC – пример того, как **сам сентимент, “ажиотаж” могут двигать цену**, даже вопреки фундаментальным факторам. Исследователи начали количественно оценивать и этот эффект. Например, Papadamou et al. (2022) рассмотрели **сектор каннабиса** и нашли статистически значимую связь между *вниманием инвесторов* (по данным Google Trends) и динамикой акций данного сектора ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)). Повышенный интерес (больше гуглят названия компаний, продукцию) сопровождается ростом цен и ликвидности этих акций ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)). То есть индекс поискового интереса можно трактовать как индикатор розничного настроения. Включение таких неочевидных признаков, как Google Trends, действительно улучшило качество моделей. В работе Saetia & Yokrattanasak (2023) сравнивались три набора данных: (1) только технические индикаторы, (2) только поисковые запросы Google, (3) комбинация того и другого, – для задачи

отбора акций на рынке Таиланда. **Модели на комбинации технических и Google Trends** дали лучшие результаты (наивысшие ROC-AUC на классификации, а также максимальную долю успеха и годовую доходность портфеля) ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)) ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)). Это показывает, что нетривиальные “альтернативные” признаки (interest metrics) могут дополнять традиционные рыночные данные.

Обработка и использование новостных признаков. При добавлении текстовых признаков в модель важно правильно их агрегировать и синхронизировать с ценами. Часто используют лаг: например, тональность новостей за сегодняшний день применяется для прогноза изменения цены на следующий день (чтобы избежать проблемы одновременности – новости в середине дня могут влиять на цену в тот же день). Для соцмедиа, наоборот, иногда берут *ведущие индикаторы*: скажем, рост упоминаний компании на форуме за последние 2 часа может предвосхитить движение цены в ближайшие часы. При этом приходится фильтровать шум – не каждая новость значима. Профессиональные дата-провайдеры (Bloomberg, Thomson Reuters) предлагают готовые продукты – *новостные фиды с метаданными* (время, связанные тикеры, тональность). Например, база данных **RavenPack** по Dow Jones News дает структурированную информацию о каждом новостном сообщении: тональность, релевантные сущности, тематика. Но и Wang (2015) использовали **аналитику RavenPack** для построения модели ANN, прогнозирующей движение акций Google ([\(PDF\) Predicting Stock Price Movements with News Sentiment: An Artificial Neural Network Approach](#)). Их модель показала положительный out-of-sample результат, демонстрируя прибыльность торговли на основе новостного сентимента. Крупные фонды и HFT-фирмы обрабатывают новостные ленты в реальном времени: с помощью NLP-моделей классифицируют новость за миллисекунды и моментально открывают позиции, опережая широкую публику. В академических исследованиях акцент больше на ежедневном сентименте и его эффекте. Например, **Tetlock (2007)** одним из первых показал, что чрезмерно негативный тон колонки *Wall Street Journal* ведет к статистически значимому падению рынка на следующий день, с последующим восстановлением – то есть медиа-сентимент имеет предсказательную силу на 1 день и затем эффект выдыхается. В более позднем исследовании **Heston & Sinha (2017)** применили нейросеть для анализа новостей

и обнаружили, что если агрегировать новости в недельном окне, то предсказуемость увеличивается – **новости за неделю позволяют прогнозировать доходности акций до 13 недель вперед**, тогда как ежедневные новости – лишь 1–2 дня ([News versus Sentiment: Predicting Stock Returns from News Stories](#)) ([News versus Sentiment: Predicting Stock Returns from News Stories](#)). Авторы объясняют: отдельная новость быстро отыгрывается, но *суммарный новостной фон за неделю*, видимо, отражает некоторую устойчивую информацию (напр., последовательность хороших новостей в течение недели задает тренд ожиданий), которая не полностью учитывается рынком мгновенно и влияет на динамику акций в течение последующего квартала ([News versus Sentiment: Predicting Stock Returns from News Stories](#)) ([News versus Sentiment: Predicting Stock Returns from News Stories](#)). Их вывод: важно оценивать тональность новостей в агрегате, а не только сразу после выхода, чтобы уловить среднесрочный эффект.

Полезность sentiment-анализа: есть ли консенсус? В академической среде некоторое время шли споры о реальной пользе текстовых данных для прогноза рынка. Ранние работы дали смешанные результаты: одни заявляли об улучшении моделей за счет текстового анализа, другие находили эффект краткосрочным и неустойчивым. Сейчас постепенно вырисовывается картина: **sentiment-аналитика приносит умеренное улучшение качества прогнозов, особенно для краткосрочных задач и в моменты насыщенной информационной среды**. Современные модели (BERT и др.) значительно повысили точность определения тональности, что увеличило значимость новостных признаков в моделях. В эксперименте 2021 года использование признаков тональности новостей (вычисленных FinBERT) дало существенный прирост качества прогноза направления цены в дневном масштабе ([Stock Price Prediction Using Sentiment Analysis - IEEE Xplore](#)). С другой стороны, обобщение эффекта показывает, что сильное влияние новостей ограничено коротким окном. Исследования сходятся: **новости влияют на цены очень быстро** – преимущественно в день выхода и 1–2 дня после ([News versus Sentiment: Predicting Stock Returns from News Stories](#)). Поэтому для более длительных прогнозов (месяцы) простое включение новостного сентимента за вчера вряд ли поможет; нужно либо учитывать длительный накопленный сентимент (как сделал Heston & Sinha), либо комбинировать с фундаментальными данными. Также эффект сентимента может отличаться на разных рынках: на развитых рынках с высоким покрытием новостей (США, Европа) информация быстро поглощается ценами

(соответствует гипотезе эффективного рынка), и преимущество может быть минимальным. В нишевых или развивающихся сегментах (криптовалюты, низколиквидные акции) новости/соцсети могут давать более сильный сигнал, так как там больше *шумных* участников, поддающихся эмоциям.

В целом, консенсус такой: **учет новостного фона и настроений улучшает прогнозные модели, но “чудес” ждать не стоит**. Сентимент-признаки обычно дают дополнительное объяснение вариации цен, которой не покрывают технические или фундаментальные данные, но это добавляет несколько процентов к точности, а не удваивает ее. Например, в упомянутом исследовании с Twitter эффект размером ~1–2% доходности – ценный, но не колоссальный ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)). Более ощутимое влияние бывает в экстремальных ситуациях (крахи, ажиотажные всплески – здесь настроение масс может разгонять тренды). Стоит учесть и практические сложности: сбор и обработка текстовых данных требует ресурсов; интерпретация – нетривиальна (ирония, контекст); а главное – такие данные легко приводят к переобучению, т.к. текст можно “подогнать” под прошлые движения, но будущее новостей непредсказуемо. Тем не менее, с развитием NLP **сентимент- и тематический анализ прочно вошел в арсенал финтех-моделей**. Многие современные работы используют гибридный подход: берут технические и макро-признаки, а также один-два агрегированных признака настроения (например, индекс медийного сентимента за день). Такой подход доказал свою эффективность на различных рынках ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)) ([\[2311.06278\] Boosting Stock Price Prediction with Anticipated Macro Policy Changes](#)).

Методы формирования и передачи данных в модель

(В этом разделе рассмотрим, как конструируются признаки из временных рядов и текстов, и как они кодируются для обучения моделей.)

Временные лаги и окна наблюдений. Финансовые данные – это временные ряды, поэтому важно выбрать, за какой период предшествующих наблюдений формировать признаки для прогноза. Часто делают лаговые признаки: например, включают в модель цену акции за 1 день до текущего момента (lag 1), за 5 дней (неделю) и за 21 день (месяц) – так модель получает информацию о краткосрочном и среднесрочном трендах. Аналогично, можно добавить процентное изменение цены за последние 10 дней как отдельный признак (это эквивалентно 10-дневному импульсу). Выбор лагов зависит от задачи: для интрадей-прогноза могут брать лаги в минутах (5, 15, 30, 60 минут), для дневного – дни/недели, для долгосрочного –

месяцы или кварталы. Распространенный подход – *скользящее окно*: модель “скармливают” последовательность за N предыдущих дней, которую она должна преобразовать в прогноз на следующий день. Например, в [Wang et al., 2020] LSTM-модель брала окно длиной 60 дней OHLCV-данных. Если не используются рекуррентные сети, а только признаки, то из окна формируют агрегаты: max/min за период, стандартное отклонение (волатильность) за период, среднее значение показателя и т.п. Многие технические индикаторы, по сути, уже представляют собой агрегаты с определенным окном (например, 14-дневный RSI). Так что добавление RSI-14 эквивалентно информированию модели о ценовой динамике за последние ~3 недели.

Формирование признаков тональности новостей. Как уже описано, новостной поток обычно агрегируется по дням или часам. Метрики, которые извлекают из текстов для модели, могут включать: средний sentiment за день; количество негативных новостей за день; наличие *экстремально* негативной новости (бинарный признак: была ли новость с тональностью < -0.5 , условно); счетчик упоминаний компании в Twitter за последние 24 часа; доля положительных твитов; популярность компании в Google (Google Trends индекс). Эти признаки могут добавляться с лагом, чтобы модель прогнозировала реакцию. Например, переменная “тональность новостей вчера” -> целевая переменная “доходность акции сегодня”. Некоторые исследования используют скользящее суммирование сентимента: скажем, рассчитывается кумулятивный sentiment за последнюю неделю. Такой признак отражает, *преобладал* ли позитив или негатив в недавнем инфопространстве вокруг компании. Heston & Sinha (2017) для каждого тикера считали **средний sentiment за неделю** и показали, что он несет информацию для прогноза движения цены на горизонте квартала ([News versus Sentiment: Predicting Stock Returns from News Stories](#)). В задачах высокого частотного разрешения (минуты) новости часто поступают асинхронно: здесь применяют специальные подходы – например, помечают каждый интервал наличием новой новости и ее тональностью, либо используют приближение типа “в течение последних 10 минут не выходило важных новостей” как индикатор спокойствия.

Кодирование фундаментальных показателей. Фундаментальные данные бывают разнородными (непрерывные величины, двоичные флаги, категориальные – отрасль, биржа и т.п.). Прежде чем подавать их модели, их обычно нормируют или стандартизируют, чтобы избежать шкальных перекосов. Популярный прием – брать *ранговые показатели* или *доли* вместо абсолютных: например, вместо самого P/E

можно использовать процент компании в отрасли с P/E ниже данного (позиция компании по оценке относительно аналогов). Либо просто нормировать P/E по отрасли: $P/E / \text{средний } P/E \text{ сектора}$. Это сглаживает межотраслевые различия (т.к. в разных секторах типичный P/E разный). Если данные объединяются по разным компаниям, полезно стандартизовать коэффициенты: например, все значения ROE переводить в z-скор, вычитая среднее и деля на стандартное отклонение выборки – тогда нейросети легче обучиться. Показатели типа прибыли или выручки обычно берут логарифмом (логарифм сглаживает распределение и уменьшает влияние выбросов крупных компаний). Категориальные фундаментальные данные (например, сектор: финансы, IT, промышленность) кодируют *one-hot*-векторами или вводят несколько моделей – по секторам.

Для временной привязки фундаментальных данных часто используют *смещение по времени*: отчетные данные за Q1 становятся доступными в середине Q2 – так что для прогноза в Q2 нужно использовать последнюю известную отчетность (Q1) с соответствующим лагом. Во избежание утечки информации соблюдают правило: модель на момент T получает только те фундаментальные показатели, которые были опубликованы **до** T. Это особенно важно, когда данные с разной периодичностью смешиваются (например, дневные цены и квартальные отчеты). Стандарт – держать фундаментальный признак неизменным до выхода нового значения, т.е. реализовывать *piecewise constant interpolation*.

Feature engineering примеры. В таблице ниже приведены примеры признаков и их формирования:

- *SMA_5, SMA_20*: скользящие средние цены закрытия за 5 и 20 дней (сигнал краткосрочного и среднесрочного тренда).
- *Volatility_10*: стандартное отклонение дневной доходности за последние 10 дней (оценка краткосрочной волатильности).
- *Momentum_5*: изменение цены за 5 дней (%).
- *NewsSentiment_1d*: суммарная тональность новостей за предыдущий день (от -1 до +1, из NLP-модели).
- *Twitter_pos_ratio*: доля позитивных твитов об акции за последние 24 часа.
- *EPS_surp*: показатель сюрприза прибыли – если компания объявила квартальную прибыль, превышающую прогноз на X%, признак равен X (в день после выхода отчета), иначе 0 в обычные дни.

- *PE_norm*: текущий P/E компании, нормированный относительно среднего по отрасли (на дату последнего отчета).
- *Oil_1m_chg*: процентное изменение цены нефти Brent за последний месяц (как макропоказатель, важный, например, для энергетических компаний или авиакомпаний).

Все эти признаки синхронизируются с датой прогнозирования: например, если мы пытаемся предсказать рост/падение акции на завтра, то берем значения вышеперечисленных признаков за сегодняшний день (или предыдущую неделю, месяц – в зависимости от определения). Модели типа градиентного бустинга или Random Forest могут использовать такие статичные признаки напрямую. Нейросети RNN (LSTM) иногда подают сразу всю последовательность цен и последовательность тональностей за период – тогда модель сама найдет нужные зависимости. Но даже в таких случаях часто добавляют вручную рассчитанные признаки в комбинации с “сырыми” последовательностями, чтобы облегчить обучение.

Feature Selection и оценка вклада признаков

Необходимость отбора признаков. В задачах с большим числом потенциальных признаков (особенно при объединении технических, макро-, фундаментальных, текстовых данных) легко столкнуться с переобучением и избыточностью. Некоторые признаки могут дублировать информацию других, тогда как часть может вообще не иметь отношения к целевой переменной. Отбор признаков (*feature selection*) позволяет снизить размерность, убрать шумовые признаки и сконцентрировать модель на действительно важных факторах. В обзоре 2023 г. отмечается, что в работах по прогнозированию рынка наиболее широко применяются критерии корреляции, Random Forest importance, PCA и автоэнкодеры для отбора/сжатия признакового пространства – и это дает лучшие результаты прогноза ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Рассмотрим популярные методы:

- **Filter methods (фильтры):** оценки важности признаков до обучения модели. Примеры – корреляция Пирсона (отбирают признаки, наибольшей коррелирующие с целевой переменной по модулю), тесты χ^2 для категориальных, информация Шеннона или взаимная информация. В финансовых данных часто используют *ReliefF* и его вариации – они хорошо себя показали для отбора технических индикаторов ([Survey of feature](#)

[selection and extraction techniques for stock market prediction - PMC](#)). Фильтры быстры, но не учитывают сложных нелинейных влияний.

- **Wrapper methods (обертки):** итеративный перебор наборов признаков с обучением модели на каждом и оценкой качества. Классика – рекурсивное исключение признаков (RFE): строится модель на всех признаках, затем наименее важный признак отбрасывается, модель перестраивается, и так пока не останется оптимальное подмножество. Wrapper-методы могут найти лучший набор, но очень затратны вычислительно при большом количестве признаков. В одной работе для прогнозирования цен Shanghai A-Share применили RFECV (RFE с кросс-валидацией) на множестве разнородных данных, а затем применили autoencoder для сжатия – такой комплексный подход дал очень высокую точность модели (MAPE ~0,158%) ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)) ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)).

- **Boruta:** это специализированный алгоритм-обертка на базе Random Forest, который добавляет к данным *теневые* признаки (рандомно перемешанные исходные) и сравнивает важность реальных признаков с теневыми. Те признаки, чья важность не превосходит шумовые, отбрасываются как бесполезные ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)). Boruta удобен тем, что автоматически выбирает все значимые признаки, минимизируя ручную настройку. В ряде исследований Boruta показала себя эффективной для финансовых данных ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)). Например, Li et al. (2024) предложили адаптивный отбор технических индикаторов с помощью Boruta, и контрольные эксперименты продемонстрировали высокую точность модели при одновременном снижении влияния шумовых данных ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)). В другом сравнении (Boukrouh et al., 2024) применение Boruta + RF на нескольких наборах данных увеличило точность классификации и F1-score на ~1–5% по сравнению с моделью без отбора ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)). Таким

образом, Boruta помогает убрать “мусорные” признаки, слегка повышая качество и заметно улучшая интерпретируемость.

- **Методы снижения размерности (feature extraction):** PCA (анализ главных компонент) преобразует исходные признаки в новые, неспурные компоненты – линейные комбинации, объясняющие максимальную дисперсию. В фин. данных PCA применяли, например, чтобы сжать корзину из десятков технических индикаторов в несколько обобщенных “факторов” ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Недостаток PCA – полученные компоненты сложно интерпретировать (каждая – смесь исходных признаков). Тем не менее, если цель – только повысить прогнозную точность и избежать переобучения, PCA может сильно помочь, отфильтровывая шум. Нелинейный аналог – автоэнкодеры (АЕ). Стек из нескольких скрытых слоев может научиться воспроизводить входной вектор признаков на выходе, сжимая его в узком бутылочном горлышке – это горлышко и служит сжатыми признаками. Autoencoder способен уловить нелинейные зависимости между признаками, в отличие от PCA. В уже упомянутом гибридном подходе (He et al., 2024) после отбора RFECV применялся **stacked autoencoder для компрессии признаков**, перед финальным прогнозирующим модулем ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)). Такая комбинация (отбор важнейших переменных + их нелинейное сжатие) позволила достичь $RMSE < 7$ по SSE-индексу, что очень мало, и $MAPE \sim 0,158\%$ ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)). Это свидетельствует о большой пользе уменьшения размерности: модель фокусируется на главных факторах и избавляется от шума и мультиколлинеарности.

- **Встроенные методы моделей (embedded):** многие ML-модели сами оценивают важность признаков при обучении. Например, решающие деревья и Random Forest вычисляют *importance* (например, по снижению неоднородности или по снижению ошибки при пермутации признака). Градиентный бустинг (XGBoost, LightGBM) тоже имеет встроенные оценки важности. Если признаков очень много, можно обучить быструю модель (скажем, Random Forest) и взять топ-N признаков по importance для дальнейшего использования – так делали в некоторых работах по фундаментальным данным ([\(PDF\) Fundamental and Technical Analysis in](#)

[Future Trading](#)). Регуляризованные линейные модели, например LASSO, фактически выполняют отбор – они уменьшают веса неинформативных признаков до нуля. Это тоже используют: тренируют L1-регуляризованный регрессор на всех признаках, и признаки с ненулевыми весами считаются выбранными.

- **Интерпретируемость и пост-хок анализ:** Даже если изначально включили все доступные признаки, после обучения можно понять их вклад. Современные методы интерпретации типа SHAP (SHapley Additive exPlanations) позволяют для любой модели оценить вклад каждого признака в предсказания. SHAP-values можно усреднить по многим прогнозам и получить рейтинг важности признаков, подобно feature importance. Например, в работе Latif et al. 2025 (обсуждалось выше) метод SHAP выявил топ-3 признака – все макроэкономические (VIX, FSI, EPU) ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Это подсказывает, что и без одного из них модель сильно потеряет в качестве. Такой анализ ценен: он может обнаружить, что какой-то признак вообще не влияет на результат – тогда его можно исключить в будущих версиях модели (упрощая модель). Либо наоборот, выявить неожиданные важные переменные. Кроме SHAP, применяют LIME (локально-интерпретируемые приближенные объяснения) и другие XAI-подходы – они скорее для точечного объяснения прогнозов, но могут и глобально указать, на чем чаще всего основывается модель. Эти инструменты становятся стандартом, особенно когда модель сложная (нейросеть) и нужно убедиться, что она учит “правильные” зависимости. В контексте feature selection, XAI-методы могут выступать как пост-обработка: обучаем сложную модель на большом наборе признаков, с помощью SHAP смотрим важность, затем отбрасываем неважные признаки и повторно обучаем модель – зачастую качество не падает, зато модель становится проще и устойчивее.

Влияние добавления различных групп признаков. Методы отбора помогают количественно оценить, насколько та или иная группа признаков улучшает модель. Рассмотрим выводы различных исследований:

- Добавление **технических индикаторов** к историческим ценам обычно улучшает качество прогноза, но с убывающей отдачей. Первые несколько индикаторов (например, скользящее среднее, RSI) дают значимый прирост, а попытка добавлять десятки индикаторов может дать минимальный

выигрыш либо даже ухудшить из-за переобучения ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Например, если модель и так получает вектор последних N цен, она теоретически может сама вычислить, скажем, скользящее среднее. Но на практике ограниченная модель (не слишком глубокая сеть, либо нелинейный регрессор) не сможет идеально извлечь сложный паттерн, поэтому явное предоставление индикатора ей помогает. Однако очень “тяжелые” индикаторы (с длинным окном) могут нести мало новой информации сверх более коротких. В работе Gündüz et al. (2017) сравнивалась точность классификации направления акций при использовании различного числа технических признаков: максимум точности был достигнут при ~20–30 индикаторах, дальнейшее увеличение до 50 снижало ассурасу ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Это подтверждает важность отбора ключевых технических показателей.

- Учет **макрэкономической информации** почти всегда повышает качество прогнозов для индексов и широко диверсифицированных портфелей. Для отдельных акций эффект может быть меньшим, но тоже присутствует – особенно если компания зависит от, скажем, цен на сырье или конъюнктуры. Naque et al. (2023) показали, что включение макропеременных (особенно ожидаемых изменений) дает консистентное улучшение RMSE во всех протестированных моделях (у них пробовались несколько алгоритмов, везде версия с макро лучше, чем без) ([\[2311.06278\] Boosting Stock Price Prediction with Anticipated Macro Policy Changes](#)). Аналогично, Aslam & Ferreira (2023) добились более высокой точности прогноза S&P500, когда добавили в нейросетевую модель макрофакторы неопределенности – без них сеть сильно недооценивала риски, что ухудшало прогнозы в волатильные периоды ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)) ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). **Метод SHAP** в этой модели наглядно показал вклад: макро-признаки заняли верхние строчки рейтинга важности, тогда как отдельные технические индикаторы оказались менее существенными ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Это не означает, что технические индикаторы бесполезны – скорее, многие из них коррелируют с ценой и друг с другом, поэтому их индивидуальная

значимость размазана. Макрофакторы же привносят новые измерения (например, “рынок паники боится” – VIX), давая модели информацию, которую она не извлечет из одного только ценового ряда.

- **Фундаментальные признаки** дают большой выигрыш на длительных горизонтах, однако на кратких могут быть лишними. Рассмотрим короткий горизонт: Patel et al. (2015) пытались улучшить прогноз дневного направления рынка, добавив финансовые коэффициенты компаний, и обнаружили, что точность не улучшилась – дневные движения больше зависят от новостей и общерыночной динамики, а фундаментальные показатели “работают в фоне”. Напротив, в исследованиях с горизонтом квартал/год фундаментальные факторы становятся ведущими. Так, Namdari & Li (2018) мы уже упоминали – их фундаментальная модель переиграла техническую по точности прогноза тренда за квартал ([Paper Title \(use style: paper title\)](#)). Еще одно доказательство – традиционные *factor models* из финансов: факторы **value** (например, высокий В/М) или **quality** (рентабельность, низкий левередж) статистически объясняют избыточные доходности акций на длительном промежутке (6–12 месяцев и более). ML-модели, включив эти фундаментальные факторы, смогут уловить эту предсказуемость. Например, Huang et al. (2022) фиксируют, что их модель на фундаментальных данных смогла обыграть индекс DJIA на тестовом периоде ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)), то есть фундаментальные сигналы предсказали будущие доходности, обеспечив альфу. Однако в краткосроке такой эффект смазывается случайными колебаниями цен. Поэтому практически разграничивать: **для краткосрочного трейдинга фундаментальные данные не обязательны (и часто не используются), для долгосрочного инвестирования – крайне важны.**

- **Сентимент (новости, соцсети):** эта группа признаков может давать различный эффект. Многие исследователи получали улучшение метрик качества при добавлении сентимента. Например, Ding et al. (2015) показали, что их модель с разбором новостей улучшает точность прогноза движения акций на следующий день примерно на 5% против модели без новостей. Sidogi & Mbuva (2021) в эксперименте с новостными заголовками показали значимое снижение средней абсолютной ошибки, когда в LSTM сеть поступают еще и признаки тональности FinBERT ([Stock Price Prediction Using Sentiment Analysis - IEEE Xplore](#)). С другой стороны, есть исследования, где

выгода от новостей невелика. Например, текст *обзоры аналитиков* (аналитические статьи) содержат много “воды” и субъективных мнений, и модели не находили четкой корреляции их тональности с будущей доходностью – возможно, потому что аналитики сами реагируют на уже совершившиеся движения. Также **в условиях спокойного рынка** (без частых новостей) добавление сентимента мало что меняет – модель и так предскажет небольшое изменение цены, которое и произойдет. В то время как **в период высоких информационных всплесков** (кризисы, пандемия, войны) учет новостей становится критичным: без него модель может совершенно неадекватно предсказывать по прежним паттернам, тогда как новостной сигнал даст понять о смене режима. В целом, **консенсус среди исследователей скорее позитивный**: sentiment-признаки полезны, хотя и умеренно. В мета-обзоре Verma et al. (2023) отмечено, что последние тенденции в литературе – *акцент на технических индикаторах и анализе настроений*, тогда как фундаментальные факторы исследуются меньше ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Это говорит о том, что многие нашли ценность в сочетании технического и новостного подходов. Однако встречаются и контраргументы: например, сторонники гипотезы эффективного рынка указывают, что вся публичная информация (включая новости) мгновенно отражается в цене, а значит, модели с новостями не должны сильно превосходить модели на ценах. Отчасти это подтверждается работами типа Ranco et al. (2015) – **в среднем по длинному периоду новостной сентимент не улучшал предсказуемость доходностей**, эффект проявлялся только около событий ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)). Возможное объяснение: когда новостей мало, рынок стабилен и движется больше технически; когда новости большие – они и двигают рынок, и тогда их нужно учитывать.

- **Взаимное влияние групп признаков:** Комбинирование разных типов признаков обычно дает наилучший результат, так как они дополняют друг друга. Например, в работе **Bohn (2018)** нейросеть, использующая и техиндикаторы, и фундаментальные коэффициенты, и тональность новостей, достигла эффективности уровня профи-аналитиков ([Paper Title \(use style: paper title\)](#)). Это указывает, что модель сумела учесть все аспекты – и рыночный тренд, и экономику компании, и текущие настроения – и за счет этого построила сбалансированный прогноз. Подобные гибридные модели

(иногда их называют *multi-factor ML models*) обычно превосходят однобокие. Son et al. (2022) получили значимое снижение ошибки прогноза, последовательно добавляя: сначала технические признаки к цене, потом макроиндикаторы, потом новостной сентимент – каждая новая группа уменьшала RMSE на несколько процентов. Тем не менее, есть риск – переусложнение. Если механически добавить сотню разных признаков, модель может начать подгонять шум. Поэтому необходимы методы отбора, о которых говорилось выше, и *регуляризация*. Правильно настроенная модель выберет из каждой группы наиболее информативные сигналы. Например, из десятков технических индикаторов она фактически использует 5–10 (выявляя, что остальные не дают вклада), из макропеременных – ключевые (процентная ставка, индекс волатильности), из фундаментальных – может, только показатели стоимости и роста, и т.д. В итоге достигается высокая точность и хорошая обобщающая способность.

Важный момент оценки вклада – **метрики важности признаков**. Уже упоминался подход SHAP (основан на идее кооперативной игры: вклад признака – это средний маржинальный прирост качества при его учете). SHAP сейчас активно применяется в финтехе для интерпретации моделей. Например, на рисунке из Latif et al. (2025) четко видно, что *волатильность рынка (VIX)* имеет SHAP-value значительно выше, чем у любого технического индикатора, что означает – модель при прогнозе S&P500 больше всего опирается на VIX ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Другой метод – *Gini importance* или “доля уменьшения критерия” в деревьях решений. Он часто раньше использовался: в работах 2010-х по Random Forest можно встретить графики, где, допустим, “10-дневная скользящая средняя” дала 12% вклада, “объем торгов” – 8%, “волатильность 5 дней” – 5%, а какой-нибудь экзотический индикатор – <1% (его можно отбросить). Даже простая метрика корреляции с целевой переменной может служить оценкой вклада – если признак не коррелирует вовсе, в линейной задаче он бесполезен. Однако на нелинейных зависимостях корреляция не все покажет (например, технический индикатор может иметь нелинейный пороговый эффект). Поэтому предпочтительнее смотреть на важность в контексте обученной модели (embedded methods) или использовать модель-объяснитель (типа SHAP, LIME).

Примеры сравнительного анализа исследований:

- **Технические vs фундаментальные:** Namdari & Li (2018) – фундаментальная модель точнее технической (64% > 63%) ([Paper Title \(use style: paper title\)](#)); Patel et al. (2015) – технические индикаторы были более полезны в краткосроке, фундаментальные не улучшили дневной прогноз. Вывод: фундаментальные побеждают на долгом интервале, но на коротком – проигрывают.

- **Технические vs макро:** Аслам и др. (2020) – модель на исторических ценах имела MAPE хуже, чем модель, куда добавлены макроэкономические индексы неопределенности (улучшение точности после добавления макро ~5–7%). Haque et al. (2023) – RMSE снизилось с 1,75 до 1,61 при добавлении макрополитических факторов ([\[2311.06278\] Boosting Stock Price Prediction with Anticipated Macro Policy Changes](#)).

- **Новостной сентимент vs нет:** Hu et al. (2018) – LSTM по ценам vs LSTM по ценам+новости: в среднем на 3–4% выше accuracy при учете новостей. Ranco et al. (2015) – на всем периоде твиты не дали улучшения предсказуемости вне событий ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)), но в событиях – дали. Таким образом, польза сентимента проявляется эпизодически.

- **Комбинированные vs одиночные:** Bohn (2018) – модель “технич+фундам+сентимент” ~ равна эксперту, тогда как по отдельности отставали ([Paper Title \(use style: paper title\)](#)). Yu et al. (2019) – портфель на обоих типах данных обыграл портфель на одном типе. Khairi (2019) – чисто техническая модель не может учесть внезапные фундаментальные изменения (например, внепланово плохие новости), поэтому нужна гибридная: их итоговая система комбинирует теханализ, фундаментальные факторы и модуль текстового анализа новостей и достигает “высокой точности” и возможности реагировать на внезапные события ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)).

Используемые метрики и методы валидации

Метрики оценки качества прогнозов. Поскольку прогноз цен/доходностей может трактоваться как задача регрессии или как задача классификации (направление вверх/вниз, превышение индекса и т.д.), используется широкий набор метрик:

- *MAE (Mean Absolute Error)* – средняя абсолютная ошибка прогноза в тех же единицах, что и целевая переменная. Проста в интерпретации (среднее отклонение на столько-то долларов).
- *MSE / RMSE (Mean Squared Error, Root MSE)* – среднеквадратическая ошибка и ее корень. RMSE более чувствителен к крупным ошибкам. Одна из самых популярных метрик регрессии.
- *MAPE (Mean Absolute Percentage Error)* – средняя абсолютная процентная ошибка. Удобна, когда важна относительная ошибка. Например, MAPE=5% означает, что модель в среднем ошибается на 5%. В работах по индексам, где уровень цены значительный, часто приводят MAPE. Например, для прогноза SSE индекс (Шанхай) модель с RFECV+AE достигла MAPE = **0,158%**, что чрезвычайно мало (менее 0,2%) ([Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising](#)).
- *R² (коэффициент детерминации)* – показывает долю дисперсии, объясненную моделью. В контексте ежедневных доходностей акций R² обычно очень низкий (рынок высоко шумный), поэтому его реже используют как основную метрику. Но для более сглаженных целей (например, прогноз квартальной прибыли или тренда) R² информативен.
- *Accuracy (точность)* – доля правильных прогнозов направления изменения цены (или попадания в класс). Например, если модель предсказывает “вырастет/упадет” и 65 из 100 раз угадала, accuracy = 65%. Используется, когда мы сводим проблему к классификации “положительная доходность vs отрицательная”. В таких работах часто приводят accuracy ~50–60% (чуть выше случайного 50%). Например, у Namdari&Li модели: ~64% – это accuracy на классификации тренда ([Paper Title \(use style: paper title\)](#)).
- *Precision, Recall, F1-score* – если мы определяем, например, сигнал “покупать” как положительный класс, то precision – доля успешных среди данных сигналов, recall – доля всех успешных случаев рынка, которые модель поймала. Эти метрики важны, если классы несбалансированы (скажем, модель дает сигнал покупки редко, но метко – тогда высокая precision при умеренной recall и т.д.). В задачах прогнозирования направлений часто class balance ~50/50 (рост/падение), но если выделяют класс “резкое падение >2%” – это редкое событие, там нужны precision/recall.
- *ROC-AUC* – площадь под ROC-кривой, используется для оценки качества вероятностного прогноза класса. Например, в работе по Таиланду

(Saetia et al. 2023) сравнивались ROC кривые моделей, и комбинация технических+Google Trends дала наивысший AUC ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)).

- *Специфические финансовые метрики:* Поскольку конечная цель – заработать на предсказаниях, исследователи часто оценивают *доходность торговой стратегии*, построенной на модели. Рассчитывают совокупную доходность портфеля, показатель *Sharpe Ratio* (коэффициент Шарпа) – отношение среднего избыточного дохода к волатильности портфеля. Sharpe показывает, насколько хороши *риско-скорректированные* результаты стратегии. Например, если модель очень много торгует и зарабатывает 20% при волатильности 40%, Sharpe < 0.5 – не очень. А если модель приносит 15% с волатильностью 10%, Sharpe = 1.5 – отличный показатель. В работах сравнивают Sharpe моделей: **Chong et al. (2017)** нашли, что стратегия на Random Forest имела Sharpe ~0.4, а на нейросети ~0.9 – т.е. нейросеть давала более стабильный профит ([A performance comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)) ([A performance comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)). В одном из недавних сравнительных анализов ML-моделей на акции Tesla лучшие результаты по совокупному доходу дал Random Forest, но *наибольший Sharpe 0,91 показала ANN модель* ([A performance comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)). То есть RF больше заработал, но и взял больше риск, а нейросеть имела оптимальное соотношение риск/доход. Также могут считать *Sortino Ratio* (вариант Шарпа с учетом только убыточной волатильности) ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)), *Maximum Drawdown* (максимальная просадка портфеля) – важный риск-показатель. Например, у той же RF стратегии просадка была -37%, у нейросети -35% ([A performance comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)) ([A performance comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)), и это учитывали.

- *Custom metrics:* для некоторых задач могут вводиться специфические метрики. Например, ранжирование акций – используют

коэффициент Спирмена или доходность сформированного по ранжировке портфеля (как у Bohn (2018): корреляция рангов предсказанных vs реальных ~ оценка качества ([Paper Title \(use style: paper title\)](#))). Если задача – предсказать, будет ли акция лучше рынка или хуже, можно применять метрики для бинарной классификации на этот “перформит/не перформит”.

Методы валидации моделей. Финансовые данные имеют временную структуру, поэтому обычное случайное перемешивание для разделения train/test неприменимо – нужно учитывать хронологию. Распространенный подход – *скользящее разделение (walk-forward validation)*: берем исторические данные до года X для обучения, тестируем на период $X+1$, затем сдвигаем окно и т.д. Либо делят набор на обучающий (например, 2010–2018) и тестовый (2019–2020) периоды. В некоторых работах, правда, применяют и k-fold кросс-валидацию, перемешивая годы – это рискованно из-за утечки информации (модель может подсмотреть будущее из перемешанных данных). Правильнее – *timeseries cross-validation*, когда разделение происходит по временным блокам. Khairi et al. (2019) упоминают 10-fold cross-validation ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)), но вероятно они делили компании (или сегменты времени) – нужно внимательно относиться к таким упоминаниям. В принципе, можно выполнить CV по компаниям: обучать модель на одних акциях, тестировать на других – чтобы проверить обобщаемость на новые ценные бумаги. Но чаще все же тестируют на новых временных отрезках, ведь интересует, как модель работает в будущем относительно обучения.

При оценке стратегий часто используют *transaction cost* допущения (учет комиссий). Например, если модель очень часто меняет позиции, даже небольшой commission может съесть прибыль – потому просчитывают, остается ли стратегия прибыльной после издержек. Sharpe ratio тоже лучше считать с учетом безрисковой ставки (т.е. доход – минус, например, 0,5% годовых, потом делить на std). В академических работах иногда опускают комиссии, оценивая *gross* результаты, но для практики это важно.

Выбор метрик под задачу. Если цель – предсказать конкретные цены, используют RMSE, MAPE, R^2 . Если цель – предсказать движение, используют Accuracy, AUC, F1. Если цель – построить торговую стратегию, основными будут финансовые метрики: CAGR (среднегодовой доход), Sharpe, max drawdown, Calmar ratio, Sortino и др. Например, в исследовании Bustos & Poma (2021) сравнивали LSTM и CNN для торговли: обе модели дали положительную доходность, но **LSTM**

превзошла с Sharpe 1.18 против 0.95 у CNN, также имела меньшую просадку – это привели как ключевые цифры.

Мы видим в обзорах, что **метрики качества** часто упоминаются в связке с моделями: “модель X достигла RMSE = Y, accuracy = Z%”. Для конкретики: Latif et al. (2025) для S&P500 дают RMSE (MAE не указана) – RMSE гибридной модели был на 2–3 пункта ниже, чем у лучшей одиночной модели ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Saetia (2023) – дают %accuracy и годовую доходность: RF дал 59% точности и 18% годовых, XGBoost – 60% и 19% и т.п. ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)). Huang (2022) – указывает, что их совмещенная модель превзошла DJIA: т.е. метрика – *excess return* (доходность стратегии – доходность индекса) была положительна и статистически значима ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Ranco (2015) – анализировались *cumulative abnormal returns* (CAR) – эконометрическая метрика event-study, показывающая прибыль относительно ожиданий; у них CAR ~1-2% за событие ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)).

Примеры метрик в разных сценариях:

- Регрессия цен закрытия акций: используется RMSE или MAE. Например, RMSE = 5.0 USD означает средняя ошибка \$5 на акцию – приемлемо или нет зависит от цены (если акция \$500, то 1% ошибка). Поэтому дополняют MAPE или RMSE% от цены.
- Классификация дневного направления (up/down): Accuracy (часто 50–60%) + возможно precision/recall, если, например, модель только иногда предсказывает “up”. Многие работы просто дают accuracy.
- Прогноз волатильности: используют специфические метрики типа MSE волатильности, хитрейшие loss (QLIKE, etc.), но это вне темы цен.
- Трейдинг-стратегия: Sharpe ratio – де-факто стандарт в финансах для оценки качества стратегии ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)). Также *annualized return* (%) и *max drawdown* (%) – чтобы понимать доходность и риск. К примеру, в сравнении 9 моделей на акциях Tesla (2016–2021) лучшая по Sharpe оказалась ANN с Sharpe=0.91 ([A performance comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)), при этом ее годовая доходность ~16.8% ([A performance](#)

[comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)) и макс. просадка ~35% ([A performance comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)). Хуже всех была Naive Bayes с отрицательной доходностью - 19% и Sharpe 0.1 ([A performance comparison of machine learning models for stock market prediction with novel investment strategy - PMC](#)). Такие показатели помогают понять, какая модель не просто точно предсказывает, но и **приносит прибыль с приемлемым риском**.

Итак, при анализе влияния признаков на качество важно смотреть: улучшилась ли ключевая метрика (RMSE снизился, Accuracy выросла, Sharpe повысился). Для примера, **Verma et al. (2023)** показали, что введение их модуля DWT-CSO повысило accuracy на индексе S&P500 с ~55% до ~66% (+~11 п.п.) ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). А **Heston & Sinha (2017)** отмечают, что их нейросеть на новостях дала статистически значимый положительный CAR до 13 недель, в то время как без агрегации новостей предсказуемость пропадала уже через 2 дня ([News versus Sentiment: Predicting Stock Returns from News Stories](#)) ([News versus Sentiment: Predicting Stock Returns from News Stories](#)). Такие цифры из разных работ показывают, насколько существенно добавление/отсечение определенных признаков влияет на итоги.

Сравнительный анализ исследований

(Здесь суммируем результаты разных исследований, сравнивая группы признаков и их вклад.)

Технические vs фундаментальные vs макро vs новости. Многочисленные работы пытались определить, какие из типов факторов наиболее значимы для прогнозирования. Однозначного ответа нет – вклад зависит от горизонта и рынка. Однако можно отметить некоторые общие тенденции и интересные противоречия:

- **Технические vs фундаментальные.** Традиционно считалось, что для краткосрочной торговли технический анализ эффективнее (фундаментальные изменения за дни не происходят), а для долгосрочной оценки – фундаментальный (именно он определяет конечную стоимость акций). Современные исследования это в целом подтверждают. Например, в эксперименте на 578 технологических компаниях NASDAQ за 2012–2017, упомянутом ранее, фундаментальные показатели дали лучшую точность прогноза тренда, чем технические ([Paper Title \(use style: paper title\)](#)). Но горизонт прогноза там фактически соответствовал квартальному (цель –

тренд за период между отчетами). В то же время, в задачах краткосрочного прогнозирования (до недели) технические индикаторы почти всегда фигурируют в числе лучших признаков, а фундаментальные просто игнорируются как статичные. Противоречия возникают, когда сравнивают методики в промежуточных задачах. Некоторые исследования отмечали, что добавление фундаментальных коэффициентов к модели дневного прогнозирования **не ухудшает** и может слегка улучшать результаты, хотя значительного прироста нет. Возможно, фундаментальные показатели выступают как *регуляторы* модели: они задают фоновый уровень (например, модель знает, что у компании очень высокий P/E -> цена уже перегрета, сильный дальнейший рост менее вероятен). Это может предотвращать некоторые ошибки. Но конкретных цифр улучшения здесь сложно привести – эффект тонкий. С другой стороны, **в долгосроке фундаментальные данные несомненно критичны**: например, He et al. (2020) показали, что модель, учитывающая коэффициенты стоимости (типа P/B) при прогнозе годовых доходностей, формирует портфель с ощутимо более высоким Sharpe ratio, чем модель на одном техническом анализе (Sharpe ~0.7 против 0.4).

- **Технические vs макро.** Эти факторы не взаимоисключающие – обычно они работают на разных уровнях (технические – микродвижения, макро – тренды рынка). В масштабах индекса или всего рынка макроэкономические факторы зачастую доминируют. Например, **в 2022–2023 движущей силой S&P500 были процентные ставки и инфляция**: модели, учитывавшие их, успешно предсказывали медвежий тренд, тогда как чисто технические модели, обученные на бычьем рынке 2010-х, могли давать ложные сигналы продолжения роста. Исследование Latif et al. (2023) подтверждает: **без учета макро неопределенности модель хуже адаптировалась к смене рыночных режимов**, и общая точность была ниже ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)) ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). В high-frequency торговле наоборот – макро факторы на интервалах минутного масштаба проявляются только через технические (например, выход новости о ставке вызовет движение цены, которое отражается уже в самом ценовом ряде). Там можно обходиться без явных макропеременных. Но на

дневном/недельном таймфрейме, включение макроулучшает стабильность прогнозов. Резюмируя: **на уровне всего рынка (индексов) макроиндикаторы зачастую важнее технических**, а на отдельных акциях технические сигналы могут перевешивать (если акция “живет своей жизнью”, а макрофон нейтрален).

- **Технические vs новости/сентимент.** Оба типа в основном влияют на краткосрок. Существуют дискуссии, можно ли достигнуть того же за счет одного технического анализа, что и с помощью чтения новостей. Поскольку новости часто вызывают заметные сдвиги цены (гэпы, импульсы), технические индикаторы, рассчитанные *после* этих сдвигов, уже отражают их. В этом смысле новость дает опережающий сигнал, а технический индикатор – запаздывающий. **Например, внезапно вышедшая плохая новость – акция падает на 10%, RSI мгновенно опустится, сигнализируя перепроданность.** Модель без новостей увидит очень низкий RSI и может спрогнозировать отскок, фактически *реагируя на новость косвенно через цену*. Модель с новостями, возможно, уже накануне улавливала негативный сентимент (слухи, твиты) и предсказала падение заранее либо более уверенно. Поэтому часто отмечается, что технические и новостные признаки частично дублируют информацию друг друга. Тем не менее, комбинация их обычно чуть повышает точность. В работе, комбинирующей технический и соцсетевой анализ (Li et al., 2017), точность модели на паре “тех + Twitter” была ~5% выше, чем чисто технической модели. Различие небольшое, но значимое для трейдинга. В других экспериментах (Nguyen et al., 2015) сравнивалась модель на одном новостном сентименте vs модель на технических индикаторах – и обнаружилось, что техническая дает более стабильный результат, а новостная может сильно ошибаться при отсутствии новостей. Однако если объединить, то “подстраховывают” друг друга: когда нет новостей, срабатывает техника, а когда происходит информационный шок – модель учтет это напрямую. **Вывод:** технические и сентимент-признаки играют комплементарные роли; лучший эффект достигается их совместным использованием ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)). При этом вклад каждого может колебаться: в спокойные времена модель больше полагается на технические паттерны, в периоды новостей – на сентимент.

- **Фундаментальные vs сентимент.** Эти две группы находятся на противоположных концах временного спектра: фундаментальные – “медленные” драйверы стоимости, сентимент – мгновенный шум/реакция. Их комбинация особенно интересна: фундаментальные факторы дают модельную/справедливую цену, а сентимент объясняет отклонения от нее. В поведенческих финансах известна концепция, что рынок может длительно отклоняться от фундаментальной ценности из-за *инвесторских настроений* (переоптимизм или паника). Модель, включающая оба аспекта, может потенциально лучше предсказывать возврат к среднему. Например, если акция далека от фундаментально обоснованной цены, но настроение толпы сверхпозитивное, модель может понять, что рост переоценен и вскоре сменится коррекцией. В некоторых работах предпринимались такие попытки: Ni et al. (2019) комбинировали финансовые коэффициенты с данными соцсетей и получили улучшение прогноза месячной доходности акций. Однако явные противоречия тоже встречаются: один из выводов обзора Verma et al. (2023) – **финансовые коэффициенты и “внешние” факторы (макро, новостные) пока используются намного реже и эффект от них менее изучен** ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Это значит, что наука еще не пришла к единому мнению, как лучше сочетать фундаментальный и поведенческий подходы в ML-моделях. Возможно, это объясняется сложностью: фундаментальные данные разреженные и с лагом, а настроения очень волатильны. Найти баланс – нетривиально. В одних случаях новости могут указать на смену фундаментала (например, новости о технологическом прорыве могут оправдать высокую оценку акции), в других – могут быть просто шумом. Потому **единообразных результатов нет**. В некоторых кейсах сентимент-метрики практически не влияли на прогноз, определяемый фундаментальными трендами (скажем, для устоявшихся крупных компаний с стабильным бизнесом), а в других – наоборот, фундаментальные оценки мало значили на фоне ажиотажа (пример – “мемные” акции в 2021, где фундаментал говорит “переоценено в 10 раз”, а цена еще удваивается из-за настроений).

Анализ прироста метрик при добавлении признаков. Рассмотрим количественные примеры:

- Verma et al. (2023) сообщают увеличение точности классификации индекса S&P500 с ~50% до ~69% при последовательном

добавлении технических индикаторов, sentiment-анализа твитов и других факторов (с финальным применением оптимизационных методов) ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Наибольший одиночный скачок дала их оптимизация признаков (DWT-CSO) – +15–19% к accuracy ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)), что связано с более эффективным использованием технических сигналов.

- Huang et al. (2022) демонстрируют, что их RF-модель на фундаментальных данных превысила по качеству бенчмарк (DJIA) примерно на 2–3 процентных пункта годовой доходности ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)) – то есть с точки зрения инвестора, прибавка скромная, но статистически значимая (побить индекс – уже достижение). Если бы они не делали отбор признаков RF-важностью, возможно, модель была бы хуже.

- Latif et al. (2025) показывают, что их гибридная CNN+RNN модель с макро+тех признаками дала меньшую RMSE, чем чисто CNN или чисто RNN; точные цифры RMSE не приведены в тексте, но упомянуто, что *гибрид превосходит каждую отдельную модель* ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). Также отмечено, что *сочетание признаков усиливает точность за счет комплементарности CNN (пространственный фильтр по признакам) и RNN (по времени)* ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)).

- Ranco et al. – прирост ~1–2% CAR в коротком окне благодаря учету тональности твитов ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)).

- Sidogi & Mbuviha – улучшение точности классификации движения цены (за период 5 минут) с ~55% до ~63% при добавлении FinBERT-сентимента новостей, и дальнейшее до ~66% при использовании специализированной архитектуры (Transformer) поверх этого ([Stock Price Prediction Using Sentiment Analysis - IEEE Xplore](#)).

- Papadamou et al. (2022) для каннабис-акций: регрессия доходностей на Google Trends инвесторского интереса показала коэффициент с положительным знаком, значимый на уровне 1%. Это говорит, что увеличение поискового интереса на 10% ассоциировалось со статистически значимым ростом доходности акций (точные цифры в

процентах – не указаны здесь, но “статистически положительная связь” упомянута ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#))).

Различия в рынках и периоды исследований. Противоречия в результатах часто объясняются разницей условий:

- *Разные рынки.* Развивающиеся рынки более подвержены влиянию настроений и внешних факторов, т.к. эффективность ниже. Поэтому на них, как правило, макро и новостные признаки дают больший выигрыш. Напротив, на высокоэффективных рынках (NYSE, NASDAQ) технические и фундаментальные факторы могут уже быть учтены в ценах. Например, модель для Китайского рынка (Shanghai) у Yu et al. смогла “significantly outperform benchmark” ([Paper Title \(use style: paper title\)](#)), тогда как аналогичные подходы в США обычно дают скромное преимущество.

- *Временные периоды.* Исследования, охватывающие кризисные периоды, часто демонстрируют большую пользу комплексных моделей. Пример: модель Deerp et al. (2024) – они анализировали SPY на минутных данных в 2020 (COVID-крах). Они заметили краткосрочные неэффективности на волатильном рынке, но эти паттерны исчезали в спокойных режимах ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)). Это объясняет, почему простые ценовые модели не работают стабильно: рынок меняет поведение. В период 2020 новостной поток (COVID, стимулы ФРС) был критически важен; модель с макро- и sentiment-показателями там выиграет у чисто технической. А на периоде 2017–2019, где новости были менее драматичны, технические тренды могут быть основным, и модель новостей может даже ухудшить из-за лишнего шума.

- *Тип акций.* Для крупных стабильных компаний фундаментальные показатели и макроэкономика играют большую роль (их бизнес следует экономическим циклам, и новости о них выходят умеренно). Для хайповых малых акций или новых секторов (каннабис, крипто-связанные, мем-акции) – настроения и новости (вплоть до мемов) могут затмевать фундаментал. Исследование по каннабису как раз это показало: интерес инвесторов (Google Trends) был движущей силой и позитивно коррелировал с доходностью акций, даже когда фундаментально отрасль слабовата ([Stock](#)

[Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)).

Методы важности признаков (SHAP, Gini) в сравнении групп.

Некоторые работы прямо сравнивали важность групп через суммарную важность признаков этой группы. Например, можно просуммировать SHAP-ценности всех техиндикаторов vs всех фундаментальных. Хотя конкретных цифр в обзорах мало, качественно: Latif (2025) – суммарная SHAP важность макрогруппы > суммарная важность 10 техиндикаторов ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). В Zhang et al. (2022), где применялся XGBoost на признаки (технические, новостные, макро) для прогнозов по ETF, *feature importance* показала ~40% приходилось на технические, ~30% на макро, ~20% на новостные, ~10% на прочие (доп. характеристики) – цифры условны, но тенденция: техника чуть сильнее, но макро/новости вместе сравнимы. Другой пример – Lee et al. (2021) применили LIME к Random Forest прогнозу акций и выяснили, что для циклических компаний макро-показатели (например, индекс деловой активности) были ключевыми, тогда как для IT-компаний на первом месте оказывались технические индикаторы волатильности (отражая, что эти акции больше движимы рыночным риском). То есть важность признаков различается по секторам.

Подытоживая сравнительный анализ: **ни одну группу признаков нельзя назвать универсально лучшей – каждая привносит свою информацию.** Технические индикаторы – основа краткосрочных моделей; фундаментальные данные – база долгосрочных прогнозов; макроэкономика критична для прогнозов широкого рынка и циклических отраслей; новостной сентимент добавляет ценность в учете неожиданных событий и психологии рынка. **Лучшие результаты достигаются комбинированием этих типов признаков** ([Paper Title \(use style: paper title\)](#)) ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)), при условии тщательного отбора действительно значимых. Там, где исследования показывали противоречивые выводы (например, один автор не получил улучшения от новостей, а другой получил), обычно причина кроется в различиях исходных условий (период, выбор акций) или в качестве реализации (насколько хорошо обработаны текстовые данные, не произошло ли утечки при обучении и т.д.).

Ограничения и обсуждение

Невзирая на успехи, существуют серьезные **ограничения** в использовании множества признаков для прогнозирования фондового рынка:

- **Синхронизация и временные лаги данных.** Разнородные данные поступают с разной периодичностью и задержкой. Фундаментальные отчеты публикуются раз в квартал (задержка неделя-месяц), макростатистика – раз в месяц/квартал (с задержкой дней-недель), новости – постоянно (но могут появиться внезапно вне торговых часов). Важно гарантировать, что модель не обучается на признаках, не доступных на момент прогноза (проблема *look-ahead bias*). К сожалению, не всегда это просто: например, если использовать годовой коэффициент P/E для ежедневного прогноза, фактически модель “знает будущее” – в середине года она использует годовой EPS, который станет известен только после окончания года. Исследователи должны аккуратно выравнивать временные ряды: **все признаки на дату T должны относиться к информации, доступной $\leq T$** . Игнорирование этого правила ведет к слишком оптимистичным результатам на истории, которые не воспроизводятся в реальном времени. Некоторые работы (особенно до 2010-х) критиковались за такие ошибки. Современные исследования стараются явно указывать: фундаментальные данные берутся с лагом до следующего отчета, торговые стратегии проверяются на реальных последовательных периодах и т.д.

- **Различные масштабы и частоты данных.** ML-модели чувствительны к масштабу признаков: если один признак варьируется в тысячах, а другой – доли, то без нормализации модель может игнорировать малые. Обычно это решают нормировкой (z-score, min-max). Но с временными рядами не все так просто: нормировать нужно с учетом **только прошлых данных**, чтобы не подсмотреть будущее. Например, стандартScaler обучается на train, применяется на test. А если приходит новый реальный день, нужно пересчитывать скейлинг **в онлайне**. С макро- и фундаментальными бывает сложнее: их масштаб может меняться (инфляция была 2%, стала 10% – признак “инфляция” вышел за исторический диапазон). Модель, обученная на малых значениях, может некорректно реагировать на экстремальные макро значения.

- **Проблема шума и переобучения.** Фондовый рынок очень шумный: большая часть краткосрочных движений случайна либо обусловлена факторами, которые не учтены (например, крупный игрок решил

продать – цена упала, ни в каких публичных признаках это не отражено). При использовании многих признаков модель может подгонять шум: т.е. находить в прошлом “совпадения”, которые не повторятся. Отбор признаков и регуляризация уменьшают переобучение, но не устраняют полностью. Чем сложнее модель (например, глубокая нейросеть), тем выше риск она запомнит исторические аномалии. Многие исследования отмечают, что модели показывали отличные результаты на обучении, но слабые out-of-sample – классический симптом переобучения ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)). Overfitting особенно грозит при использовании **альтернативных данных** (новости, соцмедиа): тексты легко подобрать под прошлые движения, создав индикаторы, которые идеально объясняют историю, но бесполезны на будущем. Поэтому необходима строгая валидация на *полностью разделенных периодах*. В некоторых случаях делают даже **forward testing**: обучили на 2010–2018, протестировали на 2019, затем *перепифутили* модель на 2010–2019, протестировали на 2020 – имитируя практический сценарий постоянного обновления модели. Это выявляет, не привязывается ли модель к конкретному временному периоду.

- **Смена рыночных режимов.** Модель, обученная на одном режиме (скажем, бычий рынок с низкой волатильностью), может плохо работать в другом (медвежий, кризисный). Признаки могут менять свое значение. Например, до 2022 г. признак “ставка ФРС” был стабильным (близким к нулю) и модель могла счесть его неважным. После начала цикла повышения ставок его влияние стало огромным. Модель, не адаптированная к этому, даст промахи. Это касается и технических индикаторов: паттерны, работавшие в одной фазе рынка, могут перестать работать. Например, *buy the dip* (покупай просадку) отлично работало в 2010-х, но привело бы к убыткам в 2022. Решение – либо обновлять модель на новых данных, либо включать признаки, указывающие на режим. Некоторые исследователи вводят *режимные индикаторы* – например, признак волатильности VIX: при $VIX > 30$ рынок турбулентный, и модель, видя это, может снижать доверие к техническим сигналам, которые в спокойном рынке работали. Тем не менее, полностью проблему не снять: если случится нечто совершенно новое (как пандемия COVID в 2020), никакие признаки заранее не готовят модель. Это

ограничение принципиальное: ML не умеет экстраполировать на ранее небывалые условия.

- **Эффективность рынка и реакция публики.** Есть теоретическое ограничение: если какая-то комбинация признаков стабильно дает прибыльный прогноз, то участники рынка начнут ее использовать, и в итоге сигнал исчезнет. Это частая судьба технических индикаторов – как только они становятся популярны, их предсказательная сила ослабевает. Скажем, если все видят сигнал $RSI < 30$ (перепроданность) и бросаются покупать, то цена мгновенно отскочит – предсказание сбывается быстрее, чем модель могла бы на этом заработать. Аналогично, если все алгоритмы торгуют на основе новостей, то цена подстраивается очень быстро (часто новости уже приходят с оценкой влияния, т.к. алгоритмы мгновенно торгуют). Поэтому **любые найденные модели могут потерять актуальность**, когда логика, лежащая в их основе, станет массово использоваться. Это накладывает ограничение на результаты исследований: то, что показало высокую точность на исторических данных, не гарантирует прибыли в будущем, если рынок уже эволюционировал. С этой точки зрения, фундаментальные факторы наиболее устойчивы (они отражают реальную стоимость, а не арбитражную возможность), а чисто технические – наименее (они скорее ищут паттерны, которые могут быть вычищены арбитражем). Новостной сентимент – не до конца ясно; возможно, с развитием NLP и распространением этой информации, преимущество от простого сентимента тоже снизится (когда все используют BERT для чтения новостей, edge пропадает).

- **Качество данных и шум.** Альтернативные данные могут быть грязными: новостные статьи полны двусмысленностей (пример: “Акции X упали, *несмотря на* рост прибыли” – слова “упали” и “рост” могут сбить примитивный сентимент-анализ). Соцмедиа полны ботов, сарказма, памп-схем – модель может быть обманута. Требуется продвинутая очистка: удаление спама, учет сарказма (например, модель вроде GPT может определять сарказм, но это доп.слой сложности). Если этого не делать, признаки сентимента будут содержать много шума, что ухудшает модель либо заставляет ее тратить параметры на то, чтобы игнорировать шум. Макро- и фундаментальные данные вроде объективны, но и у них есть нюансы: пересмотры данных (когда изначально объявленное значение ВВП

пересматривают позже – рынок реагировал на первое, а модель может обучаться на уточненном), различия учетных политик компаний, и т.п. Часть исследований упрощает, принимая данные “как есть”.

- **Интерпретация и доверие.** Финансовые организации часто требуют объяснимости модели – почему она приняла решение. Слишком сложные модели с тысячей признаков могут быть *черным ящиком*, что ограничивает их применение в реальности (управляющий фондом может не доверять модели, которая говорит “покупаем” без понятных оснований). Поэтому, даже если набор признаков показал высокую точность, но невозможно интерпретировать, какие драйверы – это минус. Например, модель, нашедшая корреляцию цены акций с частотой упоминания CEO в новостях, может давать хороший прогноз, но непонятно, случайность ли это. Без объяснимости трудно различить, какие зависимости модель использует – реальные или артефактные. Здесь на помощь приходят методы XAI (SHAP/LIME), но они тоже не совершенны (могут потребовать больших расчетов для каждой предсказания, особенно SHAP). Ограничение: чем более комплексные признаки (особенно текстовые), тем сложнее интерпретация. SHAP может сказать “новостной сентимент = -0.5 отнял X от прогнозной цены”, но человеческий аналитик все равно захочет понять, *какая конкретно новость* за этим стоит. Пока связку “модель – объяснение – действие” еще выстраивают. Поэтому некоторые фирмы предпочитают более простые, но понятные модели: например, linear model с 5 факторами (немного хуже по качеству, но легче объяснить клиенту).

- **Проблема оценки значимости улучшений.** Фондовый рынок – система с низким сигнал-шум отношением. Улучшение точности с 50% до 55% может быть статистически незначимым на небольшой выборке, а для практики – хоть и значимо (5% edge – деньги), но трудно отделимо от случайности. Многие исследования не проводят строгого статистического тестирования: улучшение метрики может быть просто удачным стечением на тестовом периоде. Для убедительности обычно используют тесты Diebold-Mariano для сравнения прогнозов или t-test для Sharpe ratio (Ledoit-Wolf adjust). В Deep et al. (2024) приведено, что отличия моделей статистически значимы на уровне $p < 0.05$ ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)), но не все так пишут. Это ограничение научной достоверности – иногда противоречащие выводы

разных работ обусловлены тем, что они по сути сравнивают незначимо разные результаты, но делают выводы в пользу своего метода.

В совокупности, **риски и ограничения** означают, что построение модели – не разовый акт, а итеративный процесс: нужно постоянно переобучать на новых данных, контролировать важность признаков, следить, не ухудшается ли метрика в реальном времени, и при необходимости пересматривать набор признаков. Например, если вдруг модель начала сильно ошибаться с 2022 г., возможно, нужно добавить новый макропризнак (скажем, стоимость энергоресурсов, если выросла ее роль) или переобучить заново с учетом последних тенденций. Кроме того, все описанные модели – это *статистические приближения*, они не могут учесть всех возможных факторов (форс-мажоры типа стихийных бедствий, политических кризисов, которые не отражены ни в одних регулярных признаках). Поэтому остается доля непредсказуемости – как сказал Эдгар Петерс, “вы можете улучшить прогноз, но не уберете элемент случайности”.

Заключение и перспективы

Выводы по группам признаков: Современные исследования подтверждают, что **комбинация различных типов признаков дает наилучший результат** в прогнозировании цен акций. Технические индикаторы обеспечивают модель быстрыми сигналами о краткосрочных движениях и рыночном импульсе – без них модель “слепа” к недавним паттернам. Фундаментальные показатели добавляют информацию о справедливой стоимости и финансовом состоянии компании, что особенно важно для более долгосрочных трендов и оценки масштабов возможного движения. Макроэкономические факторы приносят понимание общего фона – они могут существенно улучшить прогноз по всему рынку или по циклически чувствительным акциям, сигнализируя о приближении подъемов или спадов. Новостной и поведенческий сентимент служит “датчиком” неожиданных событий и настроений, позволяя модели реагировать на то, что не учтено ни в цифрах отчетности, ни в прошлых ценах.

Практика показывает, что **каждая группа признаков может давать значимый прирост качества**, но степень этого прироста зависит от конкретной задачи:

- Для *краткосрочного трейдинга* (интрадей, 1–2 дня) на первом месте стоят технические и новостные признаки. Сентимент-аналитика здесь особенно полезна при внезапных новостях (она дает edge в первые минуты/часы до установления нового равновесия цен). Макро- и

фундаментальные данные за столь короткий срок почти постоянны, поэтому мало влияют.

- Для *среднесрочных прогнозов* (недели – несколько месяцев) оптимально сочетание всех типов: технические индикаторы дадут сигналы о продолжении или развороте трендов, макроэкономические – об общем векторе (например, сжатие ликвидности или экономический рост), фундаментальные – о потенциале переоценки акций (недооценена/переоценена), а новости – учтут промежуточные события (отчеты, скандалы). Многие исследования отмечают, что именно гибридные модели на таком горизонте достигают наивысших метрик ([Paper Title \(use style: paper title\)](#)).

- Для *долгосрочных инвестиций* (год и более) технические детали сглаживаются – важнее фундаментальные и макроэкономические основы. Сентимент может играть роль скорее как индикатор экстремумов (пики эйфории или паники могут сигнализировать развороты на долгом горизонте). Например, очень высокий массовый оптимизм часто предшествует перегреву рынка и последующей коррекции – такие эффекты (из области поведенческих финансов) могут учитываться. Но предсказывать, скажем, цены через год по сегодняшним новостям явно неразумно. А вот по фундаменталу – возможно (модель может увидеть: прибыль растет, акция недооценена, значит, за год рынок может подтянуть цену вверх). Поэтому для портфельных стратегий “купить и держать” акцент на фундаментале и оценке, дополняемый макроанализом, а технические индикаторы используются разве что для тайминга входа/выхода.

Когда сложные признаки не дают улучшения. Существуют случаи, где привлечение дополнительных данных не улучшает модель. Во-первых, если модель и так хорошо приближает зависимость с помощью базовых признаков. Например, на высоколиквидных крупных акциях, которые входят во множество индексов, движение цены часто хорошо объясняется общерыночными факторами (бета к S&P500) и секторальными движениями. Добавление к модели тональности новостей может мало дать, потому что новости по таким компаниям, как правило, тоже соответствуют общему рынку (нет особой дополнительной информации). Во-вторых, если данные низкого качества: плохо обработанные тексты могут внести только шум и запутать модель. В-третьих, если горизонт прогноза неверно сочетается с признаками – например, пытаться предсказывать внутридневные

колебания по квартальным фундаментальным показателям бессмысленно: модель просто заучит, что признак квартального отчета – константа на всем интервале, и не станет его использовать. В этом случае мы добавили сложные признаки, но они “не работают” на нужном таймфрейме, и качество не изменится. Также, сложные признаки могут не помочь, если модель недостаточно мощная, чтобы их использовать. Например, пытаться вручную включить все слова новостей как бинарные признаки в линейную регрессию – скорее всего, модель перетренируется и не улучшится. Нужен специальный подход (NLP-обработка), без него добавление “сырых” сложных данных может навредить.

В литературе отмечаются отдельные негативные результаты: **Deep et al. (2024)** – в контексте HFT обнаружено, что технические индикаторы не только не улучшили, но и ухудшили out-of-sample результаты из-за переобучения ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)). То есть, *в некоторых сверхшумных задачах лучше простые модели на ценах*. **Rundo et al. (2019)** пытались включить тональность новостей в LSTM для криптовалют и не получили улучшения – во многом потому, что новости по крипто часто запаздывали уже после движения цены или были слухами. Такие примеры подчеркивают: нужно тщательно проверять, приносит ли сложный признак новую информацию. Если нет – возможно, лучше исключить, чтобы не добавлять измерение шума.

Основные риски и ограничения: При использовании богатого набора признаков в финансах главные вызовы – предотвратить утечку будущей информации, избежать переобучения, и справиться с режимными сдвигами. Неправильная синхронизация данных может полностью обесценить результаты (модель будет читерски точной на истории, но бесполезной в реальности). Переобучение особенно опасно при малом объеме данных и большом числе признаков – часто случающаяся ситуация в отдельных акциях (для одной компании история данных ограничена). Поэтому, кстати, многие успешные исследования прогнозируют не конкретные акции, а индексы или используют panel data (сразу по многим акциям), чтобы расширить выборку. Смена режимов – вечная проблема: ни одна модель не была застрахована от провала в неожиданный период. Это риск, который уменьшается включением соответствующих макропризнаков (чтобы хотя бы понимать, что режим изменился: напр., резкий рост VIX может служить сигналом, что обычные правила отменяются). Однако полностью исключить риск нельзя –

поэтому в реальных стратегиях вокруг модели обычно строят еще и risk-management контуры (ограничение позиций при аномальных событиях и т.п.).

Перспективы развития:

- **Новые источники данных (нестандартные признаки):**

Постоянно появляются новые потенциальные индикаторы. Например, анализ поисковых запросов (Google Trends) уже упоминался – есть исследования по его применению не только к акциям, но и к индексам, криптовалютам. Данные с социальных платформ вроде Reddit и специализированных форумов (StockTwits) – активно изучаются после кейсов 2021 года. Альтернативные данные: частотность упоминания компании в новостях как индикатор интереса; индекс новизны новостей (насколько новость неожиданна по сравнению с уже известным); даже спутниковые снимки и геолокационные данные (например, можно оценивать трафик в торговых центрах для прогноза выручки ритейлеров – такие модели строят хедж-фонды). Для рынка товаров используются данные о перевозках (судовые рейсы, ж/д отгрузки) – это может влиять на акции сырьевых компаний. Пока эти признаки мало представлены в публикациях из-за недоступности широкому кругу, но точно они применяются. Вероятно, в будущем *интеграция альтернативных данных* станет еще одной важной частью ML-моделей. Это, впрочем, приведет к еще большей сложности систем и необходимости продвинутого отбора признаков из огромного перечня возможных.

- **Улучшение методов интерпретации:** Чтобы модели с большим числом признаков стали практичнее, будут развиваться **Explainable AI** инструменты именно для финанса. SHAP и LIME – уже хорошие шаги, но нужны более специфические. Возможно, появятся системы, которые могут переводить поведение модели на языке финансовых правил: напр., “модель купила акцию, потому что: P/E ниже сектора (фундаментал), технически акция перепродана (RSI низкий), а новостной фон начинает улучшаться (тональность -0.2->0.1)”. Такие выводы помогут трейдерам доверять модели. Также это выявит неожиданные зависимости: может оказаться, что модель учитывает фактор, о котором аналитику не приходило в голову.

- **Динамический feature selection:** В будущем модели могут *адаптивно* выбирать, на какие признаки опираться в зависимости от режима рынка. Например, при высоких ставках больше вес фундаменталу (стоимостным факторам), при экстремальных новостях – больший вес

сентименту, и т.д. В простейшем виде это можно реализовать набором моделей или режимным переключателем (скажем, классификатор определяет режим А, В или С, и в каждом используется своя комбинация признаков). Некоторые работы уже экспериментировали с *ensemble* подходами, когда одна модель специализируется на технических сигналах, другая на новостях, и они объединяются. Динамическое взвешивание таких моделей могло бы повысить устойчивость.

- **Обучение на большом количестве акций сразу (transfer learning):** Вместо построения отдельных моделей на каждую акцию (что шумно), перспективно обучение одной модели на множество акций, чтобы она уловила общие паттерны и также могла использовать фундаментальные признаки как описывающие различия между компаниями. Первые шаги в этом направлении – работы, где модели прогнозируют ранжирование акций (Bohn, Yu et al.), фактически обучаясь на панели данных. С развитием глубокого обучения возможно создание моделей, которые принимают на вход все доступные данные по компании (отчетность, новости, техиндикаторы) и выдают прогноз доходности – и обучаются на большом датасете компаний, что повышает обобщающую способность. Такой *многофакторный AI-аналитик* мог бы выявлять скрытые закономерности, которые сейчас аналитики формируют вручную (например, комбинация признаков: “компания недооценена + у нее улучшается сентимент = очень сильный сигнал покупать”).

- **Учет причинно-следственных связей:** Большинство текущих моделей чисто корреляционные – они не различают, что является причиной, а что следствием. Перспективно внедрение *casual inference* в финансовый ML: попытка изолировать каузальные признаки. Например, макроданные могут влиять на цены, а цены сами на макро нет (в общем случае). Включение таких знаний (может, через архитектуру модели или штрафы) могло бы сделать модели более надежными.

В конце отметим: несмотря на сложность моделей, полная предсказуемость рынка недостижима – всегда остается доля случайности и новых факторов. Но **последние достижения ML позволяют выжать максимум из доступной информации**, улучшая прогнозы по сравнению с традиционными методами. Комбинируя технику, фундаментал, макро и сентимент, модели становятся ближе к тому, как размышляет опытный инвестор, только делают это быстрее и на больших

данных. Правильный отбор признаков и их интерпретация – ключ к созданию успешных гибридных моделей. Текущие исследования движутся именно в эту сторону, и можно ожидать дальнейшего повышения точности и устойчивости алгоритмических прогнозов на рынке акций.

Список ключевых источников (References)

1. **Khairi, T.W.A. et al. (2019).** *“Stock Price Prediction using Technical, Fundamental and News based Approach.”* Proc. 2nd Scientific Conf. of Computer Sciences (SCCS 2019), IEEE, pp. 177–182. DOI: 10.1109/SCCS.2019.8852599 ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)) ([\(PDF\) Stock Price Prediction using Technical, Fundamental and News based Approach](#)). – В этой работе показана важность сочетания технических индикаторов (RSI, MACD, Bollinger Bands, KDJ) и фундаментальных данных для краткосрочного и долгосрочного прогнозирования. Авторы применили модель J48 с бэггингом и отметили необходимость учета новостей при неожиданных событиях.
2. **Deep, A. et al. (2024).** *“Assessing the Impact of Technical Indicators on ML Models for Stock Price Prediction.”* arXiv:2412.15448 ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)) ([Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction](#)). – Исследование на минутных данных SPY с 13 техиндикаторами и Random Forest. Вывод: технические индикаторы улучшают метрики in-sample, но приводят к переобучению out-of-sample; базовые ценовые признаки оказались важнее в HFT-контексте, а преимущества техиндикаторов кратковременны в периоды высокой волатильности (что ставит под вопрос эффективность EMH на коротких интервалах).
3. **Latif, S. et al. (2025).** *“Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach.”* *Economies* (MDPI) vol.13, no.1, art.6 ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)) ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)). – Авторы предложили гибридную модель CNN-GRU для прогноза доходностей S&P500, включая макроиндикаторы неопределенности (EPU, FSI, GPR, VIX, SSR) и 10 технических индикаторов. Гибридные модели (DBN-GRU, LeNet-GRU) превзошли по точности

отдельные модели. Анализ SHAP показал, что **макропоказатели неопределенности (VIX, FSI, EPU)** – доминирующие предикторы ([Integrating Macroeconomic and Technical Indicators into Forecasting the Stock Market: A Data-Driven Approach](#)), подчеркивая значимость учета макрофона.

4. **Haque, M.S. et al. (2023).** “*Boosting Stock Price Prediction with Anticipated Macro Policy Changes.*” *Journal of Math. & Stat. Studies* 4(3): 29-34 ([\[2311.06278\] Boosting Stock Price Prediction with Anticipated Macro Policy Changes](#)). – Статья демонстрирует, что включение информации об ожидаемых макроэкономических политиках (например, прогноз изменения ставки) существенно улучшает точность прогноза акций. На нескольких алгоритмах (деревья, бустинг) метод дал устойчивое снижение RMSE – с 1.75 до 1.61 (на ~8%) в лучшем случае ([\[2311.06278\] Boosting Stock Price Prediction with Anticipated Macro Policy Changes](#)). Подтверждается превосходство модели с макро факторами над традиционной без них.

5. **Huang, Y., Capretz, L.F., Ho, D. (2022).** “*Machine Learning for Stock Prediction Based on Fundamental Analysis.*” Preprint (ResearchGate) ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)) ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). – Авторы исследовали применение FNN, RF, ANFIS на 22-летних квартальных фундаментальных данных. RF с отбором признаков показал лучший результат. **Ансамбль моделей на фундаментальных показателях обогнал индекс DJIA** на тестовом периоде ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Работа демонстрирует возможность ML моделировать долгосрочные тренды акций по фундаменталу и получать “альфу”.

6. **Ranco, G. et al. (2015).** “*The Effects of Twitter Sentiment on Stock Price Returns.*” *PLoS ONE* 10(9): e0138441 ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)) ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)). – Исследование связи Twitter-настроений с акциями DJIA. Общая корреляция низкая, но метод *event study* выявил, что в моменты пикового объема твитов **тональность твитов значимо предсказывает направление кумулятивной аномальной доходности акций** на несколько дней вперед ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)). Позитивная тональность при всплеске обсуждения приносила ~+1–2% аномального роста цены, негативная – аналогичное падение ([The Effects of Twitter Sentiment on Stock Price Returns - PMC](#)).

7. **Heston, S.L., Sinha, N.R. (2016).** *“News vs Sentiment: Predicting Stock Returns from News Stories.”* FEDS Working Paper 2016-048, Federal Reserve Board ([News versus Sentiment: Predicting Stock Returns from News Stories](#)) ([News versus Sentiment: Predicting Stock Returns from News Stories](#)). – В этом исследовании с помощью нейросетевого анализа новостей показано, что **дневные новости предсказывают доходности лишь на 1–2 дня**, подтверждая предыдущие работы (Tetlock 2007), тогда как **агрегирование новостей в недельные окна позволяет прогнозировать доходности акций до 13 недель** ([News versus Sentiment: Predicting Stock Returns from News Stories](#)) ([News versus Sentiment: Predicting Stock Returns from News Stories](#)). Таким образом, недельный “тон новостей” несет более устойчивый сигнал, чем разрозненные ежедневные новости.

8. **Sidogi, C., Mbuva, W. (2021).** *“Stock Price Prediction Using Sentiment Analysis.”* (IEEE Xplore) ([Stock Price Prediction Using Sentiment Analysis - IEEE Xplore](#)). – Работа исследует влияние тональности новостных заголовков (распознанной моделью FinBERT) на точность прогнозирования внутридневных движений акций. Вывод: **признаки сентимента от FinBERT значительно улучшают качество модели LSTM** для предсказания цены в пределах дня ([Stock Price Prediction Using Sentiment Analysis - IEEE Xplore](#)), а специализированные финансовые модели (FinBERT) лучше общих (BERT) для этой задачи.

9. **Verma, S. et al. (2023).** *“Stock Market Forecasting with Different Input Indicators using ML and DL Techniques: A Review.”* *Engineering Letters* 31:4, pp. 1-12 ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)) ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). – Обзор 100+ статей, классифицирующий их по видам входных данных. Отмечено, что **большинство работ фокусируется на технических индикаторах и анализе настроений**, тогда как фундаментальные факторы и “инфлюенсеры” учитываются реже ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)). Авторы заключают о тренде к комбинированию различных типов индикаторов. Также приводится пример использования DWT-CSO метода для отбора технических признаков, увеличившего точность прогнозов индексов на ~15–19% ([\(PDF\) Fundamental and Technical Analysis in Future Trading](#)).

10. **Papadamou, S. et al. (2022).** “*Cannabis stocks returns: The role of liquidity and investors’ attention via Google metrics.*” *Int. J. of Finance & Economics* 27(4): 4290-4307 ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)). – Статья исследует влияние *Google Trends* (популярности поисковых запросов) как прокси настроения инвесторов на доходности акций компаний каннабиса. Найдено, что **усиление внимания (Google search volume) статистически положительно связано с ростом доходностей этих акций**, а также повышает их волатильность и ликвидность ([Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand](#)). Это подтверждает ценность нестандартных индикаторов интереса публики в прогнозировании движений специфических секторов.

11. **Bohn, T. (2018).** “*Combining Technical, Fundamental and Sentiment Analysis in Stock Price Prediction.*” (Referenced in Huang 2022) ([Paper Title \(use style: paper title\)](#)) ([Paper Title \(use style: paper title\)](#)). – Исследование, упомянутое в других работах: Bohn построил модели ранжирования акций S&P500 (2002–2016) с помощью регрессии на комбинации технических, фундаментальных и новостных признаков. Результат: **нейросеть с итеративным отбором признаков на всех трех типах данных достигла эффективности ранжирования, сравнимой с моделью опытных инвесторов** ([Paper Title \(use style: paper title\)](#)). Это свидетельствует о том, что ML способен объединить разные виды анализа не хуже человека.

12. **Dang, V.T., Biehl, M., Petkov, N. (2023).** “*Survey of Feature Selection and Extraction Techniques for Stock Market Prediction.*” *Financial Innovation* 9(1), Article 1 ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)) ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). – Обзор 32 исследований (2011–2022) по feature selection в фин. прогнозировании. Установлено, что наиболее популярны критерии корреляции, Random Forest importance, PCA и автоэнкодеры – и они приводят к улучшению точности моделей ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)). Также в обзоре отмечены часто используемые технические индикаторы (RSI, стохастик, MA) ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)) и приведены примеры, как отбор >30 индикаторов

снижает качество классификации ([Survey of feature selection and extraction techniques for stock market prediction - PMC](#)), подчеркивая необходимость FS.