МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М.В. ЛОМОНОСОВА

Экономический факультет

Кафедра финансов и кредита

Курсовая работа по теме
«Модель ценообразования активов на основе методов машинного обучения»
(Asset pricing model based on ML methods)

Автор курсовой работы:
Студент группы Э-301
Барамия Никита Тенгизович

Научный руководитель:
Байбаков Владислав Игоревич

Москва 2020 г.

Оглавление

Введение

Машинное обучение является одним из самых бурно развивающихся направлений в прикладной статистике и компьютерных науках, нашедшее применение во многих сферах нашей жизни: чат-боты, распознавание изображений, подбор рекомендаций, медицина, кибербезопасность и так далее. Многие алгоритмы машинного обучения направлены на прогнозирование какой-либо переменной на основе признаков, предварительно проведя тренировку на исторических данных. Неудивительно, что с ростом количества данных машинное обучение обрело высокую практическую значимость в банковской и финансовой сферах.

В данной работе будет попытка рассмотреть применение методов машинного обучения в финансах: мы проанализируем использование композиционных (ансамблевых) методов в прогнозировании ожидаемой доходности акций, опираясь на классическую расширенную теорию CAPM и теорию арбитражного ценообразования, и попробуем оценить эффективность (точность) данных алгоритмов по сравнению с классическими подходами в эконометрике.

Глоссарий по машинному обучению

В начале работы предоставлю краткое объяснение некоторым понятиям, которые важны для понимания дальнейших действий в работе. На рисунке 1 можно увидеть всё многообразие методов машинного обучения.

Обучение с учителем (supervised learning) – раздел машинного обучения, целью которого является восстановление связи между объясняющими переменными (inputs) и объясняемой переменной (output).

Второй основной раздел – обучение без учителя (unsupervised learning), целью которого является поиск закономерностей и взаимосвязей между переменными, в частности, кластеризация объектов.

Композиционные или ансамблевые методы (ensemble methods) – мета-алгоритмы,

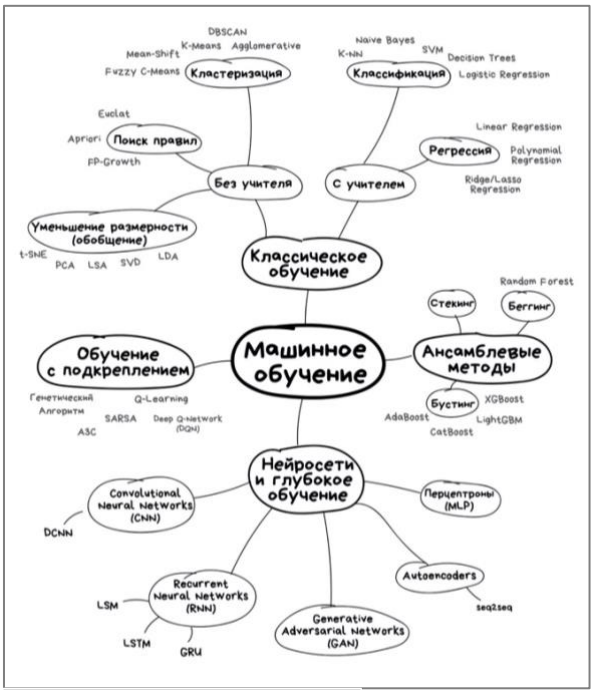


Рис. 1. Алгоритмы МО.

объединяющие несколько техник машинного обучения в одну прогнозирующую модель. Выделяют три семейства таких алгоритмов:

1. Бэггинг (bagging, bootstrap aggregating) – метод, при котором несколько алгоритмов обучается независимо друг от друга на одних и тех же данных (либо схожие алгоритмы обучаются на разных подвыборках), далее путём голосования определяется результат. Призван снижать дисперсию предсказания.
2. Бустинг (boosting) – метод, в котором каждый последующий алгоритм пытается исправить ошибки предыдущих алгоритмов. Призван снижать смещение предсказания.
3. Стекинг (stacking, stacked generalization) – метод, первым этапом которого является обучение на исторических данных, вторым этапом происходит обучение на предсказанных (моделями первого этапа) данных и производится финальное предсказание. Призван улучшать само предсказание.

Глубокое обучение – семейство алгоритмов машинного обучения, основанное на имитации работы человеческого мозга (широко известны, как нейронные сети). Обычно, для тренировки требуют очень большого массива данных.

Обзор литературы

Первые теории ценообразования активов

Одной из первых моделей, оценивающей ожидаемую доходность активов, считается модель CAPM (Capital Assets Price Model), представленная Джеком Трейнором (1961, 1962), Уильямом Шарпом (1964), Джоном Линтнером (1965) и Яном Моссиным (1966) с опорой на более раннюю работу Гарри Марковица 50-х годов о методики формирования инвестиционного портфеля (известная как портфельная теория Марковица). Зависимость выражена в виде простого линейного уравнения ожидаемой ставки доходности актива на премию рынка: $E(R_i) = R_F + \beta_i * \big(E(R_m) - R_f\big)$, где R_F – безрисковая ставка доходности, $E(R_m)$ – ожидаемая доходность рыночного портфеля. Премией за риск в данном случае называют $E(R_i) - R_F$. Данная модель является хорошей основой, но в связи с жёсткими предпосылками, обладает очень слабой прогностической способностью и, как показывает эмпирика, объясняет не всю вариацию в доходности активов.

Альтернативная модель была представлена Стивеном Россом в 1976 году и получила название теории арбитражного ценообразования (APT – Arbitrage Pricing Theory). В данном формате ожидаемая доходность моделируется как линейная зависимость от различных финансовых и макроэкономических факторов/индексов:

$E(R_i) = R_F + b_{i1}E(RP_1) + b_{i2}E(RP_2) + \ldots + b_{in}E(RP_n)$, где R_F всё та же безрисковая ставка доходности, RP_k – премия за риск k-го фактора. Таким образом, премия за риск линейно зависит от чувствительности актива к n факторам. В отличие от CAPM, в которой предполагается совершенная эффективность рынков, APT предполагает, что рынки иногда неверно оценивают ценные бумаги, то есть существует возможность арбитражных опций, прежде чем рынок снова придёт в равновесие. В какой-то степени, теорию арбитражного ценообразования можно рассматривать расширением CAPM, обладающим меньшим количеством предпосылок и допустимостью многофакторной зависимости, что даёт большие надежды с точки зрения эмпирического прогнозирования.

Развитие теории и эмпирические свидетельства

Roll, Ross 1980 провели первые эмпирические тесты в поддержку арбитражной модели. Connor, Korajczyk 1988, оценивая месячные доходности акций с применением метода главных компонент (выбор обусловлен превосходством количества ценных бумаг над периодами наблюдения), предоставили эмпирические доказательства, что модель APT является хорошей альтернативой CAPM. Арбитражная теория превзошла классическую в объяснении сезонности доходностей активов.

Priestley 1996 использовал три подхода/техники для отбора используемых факторов арбитражной модели. Лучший результат показала APT модель, полученная рекурсивным фильтром Калмана: годовая недооценка составила 1.63%. Модель опиралась на риск невыполнения обязательств, обменный курс, предложение денег, неожиданную инфляцию и рыночную премию.

Azeez, Yonezawa 2006, рассматривая японский рынок акций с использованием арбитражной теории, показали значимость таких макроэкономических факторов, как предложение денег, инфляция, индекс промышленного производств и обменный курс, в анализе доходности акций во всех рассматриваемых ими периодах. Kisman, Restiyanita 2015 провели сравнительный анализ двух классических моделей и пришли к тому, что простая модель APT с валовым внутренним продуктом и процентной ставкой объясняет большую долю дисперсии доходности акций индонезийской биржи по сравнению с CAPM. Однако не всё так однозначно. Джордан Френч (French, 2017) провёл эмпирическую проверку 5-ти макроэкономических факторов, используя данные для шести стран (США, Сингапур, Таиланд, Филиппины, Малайзия, Индонезия). Переменные, связанные с инфляцией, показали незначимость. Более того, на пяти из шести странах модель арбитражного ценообразования оказалась менее надёжным инструментом, чем модель CAPM.

Много продвижений сделано в сторону классической модели. Одним из известных расширений CAPM, призванной объяснить большую долю разброса доходности активов, является Fama–French 3-factor model (1992). Помимо премии рынка добавляются переменные SMB (Small Minus Big: предполагается, что, если в портфеле больше компаний с малой капитализацией, данный портфель должен превзойти рынок в долгосрочной перспективе) и HML (High Minus Low: позволяет учитывать ценные бумаги с высоким отношением балансовой стоимости к рыночной, которые генерируют более высокую прибыль по сравнению с рынком). Спустя несколько лет (Fama, French 1996) авторы опубликовали дополнительные эмпирические подтверждения справедливости их модели: она лучше справилась с объяснением аномалий доходности акций. Более того, она превзошла модель (однако тоже неплохую), основанную на переменных «прибыль / цена», «денежный поток / цена», «рост продаж», предложенных в другом исследовании (Lakonishok, Shleifer, Vishny 1994).

Fama, French 2008 провели эмпирическую проверку некоторых из факторов, пытающихся предсказать аномалии на рынке. В результате они получили, что рыночная капитализация имеет отрицательное влияние и положительное влияние HML. Lewellen провёл схожее свидетельство в 2014 году: он также использовал процедуру Фамы и МакБета (Fama–MacBeth regression, 1973). Исследователь моделировал инвестиционную деятельность, оценивая ожидаемую доходность акций по 15-ти характеристикам фирмы. Он получил достаточно сильную прогностическую силу модели.

Позднее в 2015 году те же профессора расширили модель ещё двумя факторами: RMW (прибыльность: компании, сообщающие о более высоких будущих доходах, имеют более высокую доходность на фондовом рынке) и CMA (показатель инвестиций: компании, направляющие прибыль на крупные проекты роста, вероятно понесут убытки на фондовом рынке). Согласно тестам, 5-факторная модель превзошла 3-факторную, разработанную раннее. Таким образом, их итоговую модель можно записать в виде:

$$R_{it} - R_{Ft} = \alpha_i + \beta_{1i}(R_{Mt} - R_{Ft}) + \beta_{2i}SMB_t + \beta_{3i}HML_t + \beta_{4i}RMW_t + \beta_{5i}CMA_t + \varepsilon_{it}$$

Методы отбора факторов и оценки качества прогнозирования

Lewellen, Nagel, Shanken 2010 в своём исследовании сделали вывод, что высокая объяснительная сила с точки зрения высокого коэффициента детерминации или небольших ошибок в оценке не является достаточным для оценки модели, и предложили несколько вариантов решения данного вопроса.

Несколько работ (Giglio, Xiu 2017 и Kelly, Pruitt, Su 2017) использовали метод главных компонент для определения важных прогностических факторов. Одним из результатов была высокая статистическая значимость (на уровне 1%) только 7 факторов: рыночная премия за риск, отношение балансовой стоимости к рыночной, отношение прибыли к цене, моментум и так далее.

Green, Hand, Zhang 2013 делают один из первых обзоров существующих прогностических сигналов (RPS – return predictive signals), опубликованных в период с 1970 по 2010. Позднее Harvey, Liu, Zhu 2016 выпускают статью с критикой: они оспаривают ценность и значимость предыдущих исследований и, помимо собранной ими таблицы из 316 уже найденных предполагаемо значимых факторов, заявляют более верным использовать уровень значимости, соответствующий $t = 3$, для подтверждения влияния переменной на ожидаемую доходность актива в дальнейших исследованиях. В работе Harvey, Liu 2019 исследователи попытались предоставить алгоритм для выбора определяющих прогностических факторов среди сотен, появившихся за последние десятилетия. Одним из результатов их методики является подтверждение доминирующей роли премии за рыночный риск, а также значимость прибыльности (хотя она и не так весома на фоне рыночного фактора).

Ценообразование активов и машинное обучение

Техники машинного обучения ещё не обрели широкого распространения в литературе, связанной с ценообразованием финансовых активов, но уже существует несколько работ, стоящие внимания.

Moritz Zimmermann 2016 обучали деревья решений для определения доходности акций. Данный метод позволяет иметь дело с большим количеством переменных и их потенциальными нелинейностями. Торговая стратегия, основанная на их способе, имеет информационный коэффициент в два раза выше, чем линейная регрессия Фамы-Макбета с учётом двухсторонних взаимодействий.

Messmer Gu, Kelly, Xiu 2018 в своей работе проводят масштабный сравнительный анализ методов машинного обучения по отношению к оцениванию премии за риск: линейные модели, разные техники понижения размерности, бустинг над деревьями, рандомные леса и так далее. В результате они определили два лучших метода: регрессионные деревья и нейронные сети.

Feng, Polson, Xu 2019 и Chen, Pelger, Zhu 2019 использовали методы глубоко обучения: определёнными способами они обучали нейронные сети, позволившие им

поймать нелинейные связи в предсказании доходности активов и достичь высокой точности в прогнозировании.

Таким образом, алгоритмы машинного обучения могут дать новый толчок в развитии анализа ценообразования активов, поиске паттернов и дальнейшем более точном предсказывании, что является востребованной задачей в индустрии.

Данные

С сайта yahoo были собраны ежедневные данные о стоимости акций компаний, включённых в составление фондового индекса S&P 500. Распределение данных компаний по секторам можно наблюдать ниже (рисунок 2). Будем предполагать возможную сонаправленность доходностей компаний из одного сектора, поэтому в модель будет включён бинарный контроль: фиктивные переменные принадлежности предприятия к определённому сектору.

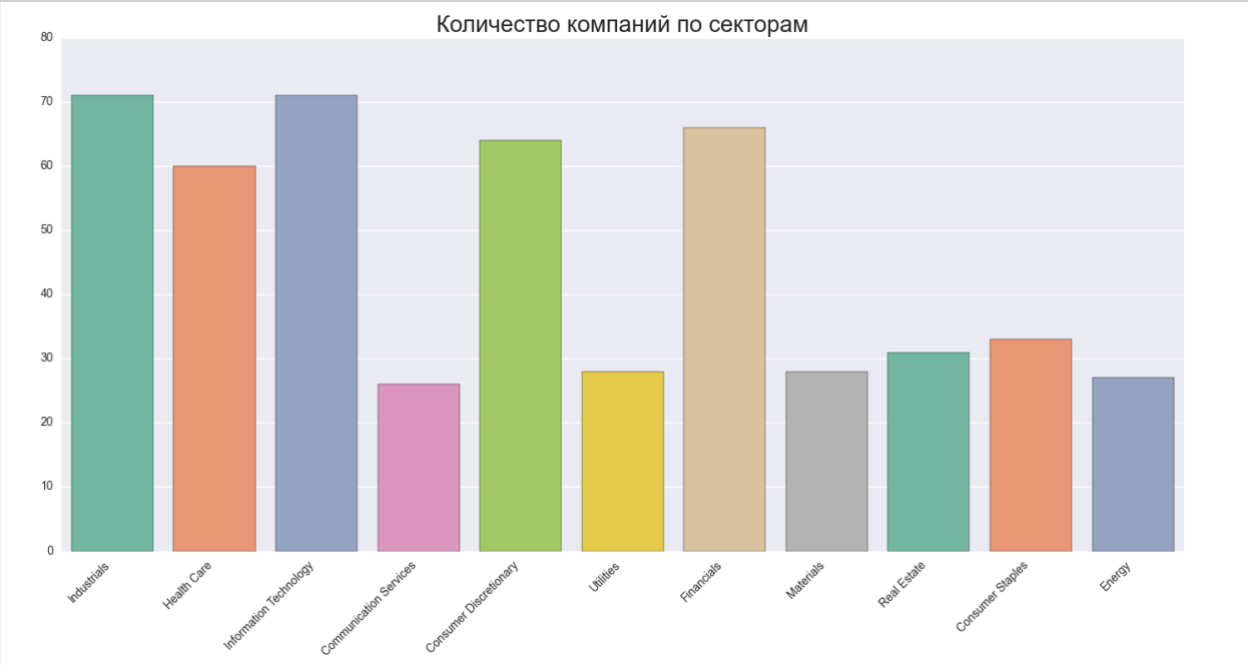


Рис. 2. Распределение компаний из S&P 500 по секторам.

Также были собраны характеристики из финансовых отчётностей компаний S&P500, включая квартальные статьи бухгалтерского баланса, отчёта о прибылях и убытках и движения денежных средств, стоимости предприятий, показатели роста показателей в данных статьях и прочие метрики. Полный список с описанием переменных представлен в приложении. Для сопоставимости дат между разными источниками, были отфильтрованы отчётности, чьи публикации не соответствуют концу кварталов. Также мы будем рассматривать только наблюдения после мирового экономического кризиса: в США спад

экономики закончился примерно в конце 2008 года, конечная выборка – с 2009 по 2020 год. Таким образом, мы имеем таблицу из 18091 наблюдений по 149 переменным[1].

Так как финансовые отчётности имеют квартальную периодичность, для сопоставимости мы будем рассматривать такую же периодичность и стоимости акций. В данной таблице остались 21162 наблюдения по стоимости акций. При подробном рассмотрении переменных в таблице с финансовыми отчётностями и другими метриками также была найдена переменная с квартальными стоимостями акций. Построение графиков (рисунок 3) показывает, что большинство значений разбросаны около нуля. Разница между двумя переменными невелика, поэтому будем рассматривать качество алгоритма на одном из датасетов (с б**о**льшим количеством наблюдений).

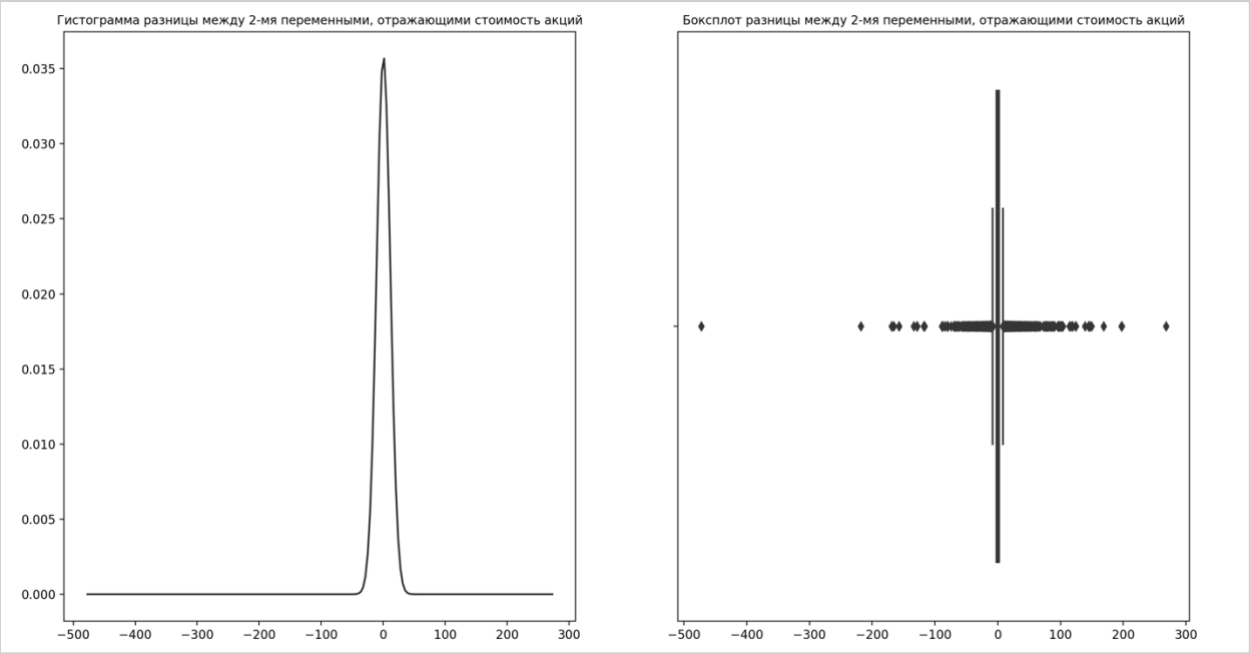


Рис. 3. Свидетельство о наличии разницы между двумя переменными.

Совокупность макроэкономических факторов влияют на доходность акций. Для того, чтобы учесть их колебания, мы используем собранные исследователями McCracken и Ng квартальные данные с FRED[2]. Аналогично используем года с 2009 по 2020, и таким образом мы обладаем информацией о 248 макро-переменных за каждый интересующий нас квартал.

Итого, при объединении всех таблиц мы получили два варианта дата-сетов, для самостоятельно собранных квартальных цен акций и взятых из источника финансовых отчётностей: 17545 и 17671 наблюдений соответственно, с аналогичным набором предполагаемо объясняющих переменных.

[1] Не все переменные имеют такое количество наблюдений в связи с наличием пропусков.
[2] В марте 2020 года появился подробный рабочий документ с описанием этих данных.

Обработка и отбор переменных

Нас интересует предсказание риск премий по акциям, поэтому мы сделаем данное преобразование: $ERP_t = \frac{P_t + D_t - P_{t-1}}{P_t} - R_{Ft}$, где P_t – цена акции в t-ом периоде, D_t – дивиденд за акцию в t-ом периоде, R_{Ft} – квартальная безрисковая ставка в t-ом периоде, за которую мы примем ставку по десятилетнему казначейскому векселю. Изменение динамики ряда можно увидеть ниже:

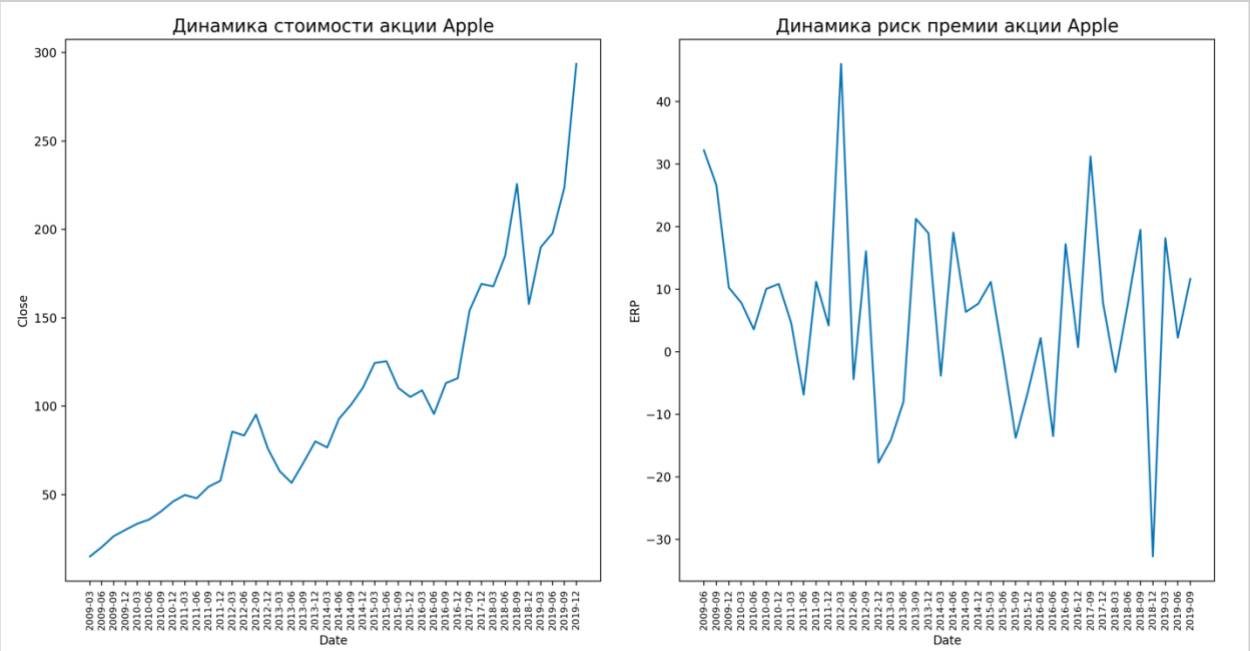


Рис. 4. Преобразование временного ряда (в качестве примера акции компании Apple)

Данное преобразование привело к созданию стационарного ряда (что также подтверждает тест Дики-Фуллера на 1%-ом уровне значимости). К стационарности мы приведём и макроэкономические показатели. Для них мы будем пользоваться таким алгоритмом:

1. Проводим тест исходного ряда на стационарность: если стационарен, переходим к следующему ряду, если нет:
2. Делаем преобразование вида $\frac{x_t - x_{t-1}}{x_{t-1}}$ и проверяем данный ряд на стационарность, если да, то сохраняем преобразование и переходим к следующем ряду, если нет:
3. Делаем преобразование вида $\frac{\log(x_t) - \log(x_{t-1})}{\log(x_{t-1})}$ и проверяем его на стационарность, если да, то сохраняем преобразование и переходим к следующем ряду, если нет:
4. Записываем, что ряд не стационарен и переходим к следующему ряду.

Можно было бы произвести более сложные преобразования (например, брать более высокий порядок интегрированности), но пришлось бы пожертвовать большим количеством наблюдений. По итогу алгоритма мы получили 29 рядов, которым не

потребовалась трансформации, 171 ряд, преобразованный первым способом, и 2 ряда, преобразованных вторым способом. 46 рядов не удалось привести к стационарному виду, их мы удалим: не будем рассматривать как объясняющие переменные.

Следуя работе Chen, Pelger 2019 года и используя наши квартальные отчётности компаний, создадим переменные, которые предполагаемо могут объяснить будущие доходности акций компаний. Сводная таблица этих переменных (31) представлена ниже:

Аббревиатура	Расшифровка
A2M	Отношение активов к рыночной капитализации
AC	Изменение в операционном капитале на балансовую стоимость
NOA	(Операционные активы – операционные обязательства) / лаг. все активы
ATO	Выручка / лаг. чистые операционные активы
B2M	Отношение балансовой стоимости к рыночной капитализации
C	Отношение наличности и SR инвестиций ко всем активам
CF2B	Свободный денежный поток к балансовой стоимости
CF2P	Свободный денежный поток к рыночной капитализации
CTO	Выручка к лаг. всем активам
D2A	Амортизация ко всем активам
D2P	Все дивиденды к рыночной капитализации
DPI2A	Изменение в основных средствах к лаг. всем активам
E2P	Прибыль к рыночной капитализации
FC2R	Отношение SG&A, R&D и рекламных расходов к выручке
OC2R	Отношение операционных расходов к выручке
I	Процентное изменение во всех активах
Lev	Отношение обязательств к обязательствам + собственный капитал
MCC	Процентное изменение рыночной капитализации
NSI	Лог изменение в количестве акций в обороте
OA	Изменение в неналичном оборотном капитале и амортизации к лаг всем активам
OL	Себестоимость и SG&A расходы ко всем активам
OP	Операционная прибыльность
PCM	Разница между выручкой и себестоимостью к выручке
PM	Операционная прибыль после амортизации к выручке
PROF	Валовая прибыль к балансовой стоимости
Q	Все активы + рыночная капитализация – наличность – SR инвестиции – предоплаченные налоги ко всем активам
RNA	Отношение операционной прибыли к лаг. операционным активам
ROA	Прибыль к лаг. всем активам
ROE	Прибыль к лаг. балансовой стоимости
S2P	Отношение выручки к рыночной капитализации
SGA2S	Отношение SG&A расходов к выручке

Таблица 1. Реплицированные специфичные для компаний признаки.

Помимо реплицированных переменных мы возьмём из исходных данных некоторые метрики и коэффициенты ростаз: в общей сложности мы обладаем 77 специфическими переменными компаний. К этому списку мы прибавим бетта коэффициент, рассчитанный как ковариация доходностей рынка S&P500 и акций компании, делённая на дисперсию доходности рынка. Эмпирические свидетельства показывают, что большему коэффициенту соответствует большие доходности.

Следующим шагом были удалены использованные столбцы, удалены пропуски и значения равные бесконечности (одной из причин данных явлений может быть деление на ноль при создании новых переменных).

При тщательном рассмотрении созданной переменной доходностей можно увидеть явные выбросы (рисунок 5). В связи с этим наложим ограничение на переменную: будем считать наблюдение нормальным, когда доходность по модулю меньше 40%. Это действие позволяет сделать более адекватное распределение наблюдаемых доходностей с потерей всего около 200-250 наблюдений.

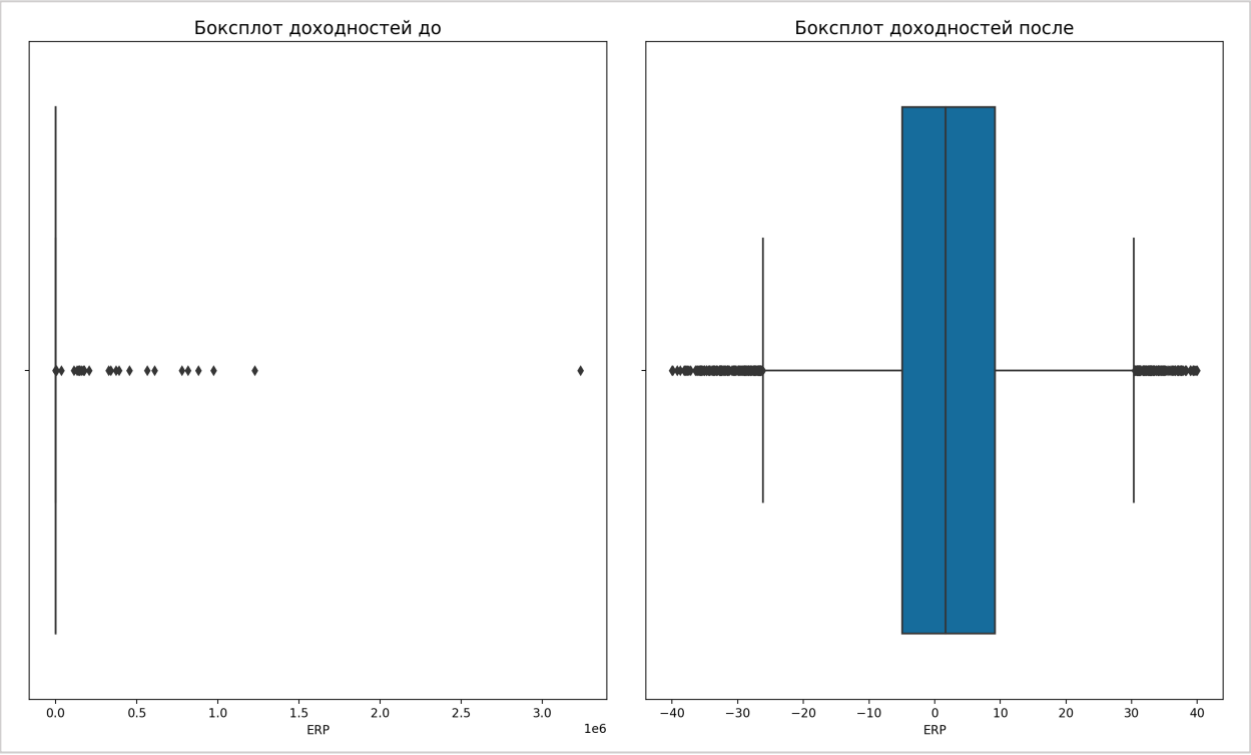


Рис. 5. Изменение распределения доходностей акций компаний S&P500.

Выбросы возможны и по объясняющим переменным, в связи с этим оценим нормированные значения переменных: выбросами будут считаться наблюдения, чьё значение по одной из объясняющих выходит за рамки 3-х по нормированной шкале.

3 Отмечены в приложении

Таким образом, у нас есть два дата сета с 9906 и 10069 наблюдениями: две переменные, соответствующие дате и компании (для идентификации доходностей), одна объясняемая переменная (доходность) и 288 переменных, которые, предположительно, могут предсказать будущие доходности акций компании.

Модели

Обзор выбранных алгоритмов

Эластичная сеть

Эластичной сетью (Elastic Net) является линейная модель регрессии с двумя регуляризаторами L1 и L2 (при отдельном использовании соответствуют лассо регрессии (LASSO) и гребневой регрессии (RIDGE) соответственно). Формально, задача оптимизации выглядит так:

$$\|y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|^2 \longrightarrow \max_{\beta}$$

Дерево решений

Дерево решений – алгоритм обучения с учителем, предсказание которого основано на обученных правилах для принятия решений. Решающее дерево создаётся путем рекурсивного разделения выборки – начиная с корневого узла (известного как первый родительский элемент), каждый узел может быть разбит на левый и правый дочерние узлы. Эти узлы затем могут быть дополнительно разделены, и они сами становятся родительскими узлами своих результирующих дочерних узлов. Нахождение оптимальных точек для ветвления происходит с помощью информационного критерия.

Адаптивный бустинг

Также известный как AdaBoost или Adaptive Boosting. Алгоритм представлен ниже:

1. Все тренировочные точки данных инициализируются с определёнными весами (для первого раза все точки имеют одинаковый вес).
2. Выбранный алгоритм (базовый) обучается на этих точках.
3. Рассчитывается взвешенный коэффициент ошибок – сумму отклонений ошибочных прогнозов из общего числа, у каждой ошибки есть свой вес, определённый на первом шаге. Также считаем вес для самого алгоритма на основе этого коэффициента.
4. Обновляем веса каждой точки, в зависимости от того, насколько неточно она была предсказана, и возвращаемся к первому шагу, пока не будет достигнуто определённое при инициализации требуемое количество базовых оценщиков.

Финальное предсказание будет взвешенным предсказанием всех ранее обученных алгоритмов. Дерево с б**о**льшим весом будет иметь б**о**льшее влияние на результат.

Случайный лес

Алгоритм метода Random Forest для выбранного заранее N таков:

1. Случайным образом выбираются подвыборки (N штук) из тренировочных данных.
2. На этих N подвыборках обучаются N решающих деревьев (одна подвыборка на одно дерево, каждое дерево обучается на случайном подмножестве признаков).
3. Делается финальное предсказание, основанное на среднем из N предсказаний.

Градиентный бустинг

Алгоритм метода Gradient Boosting для выбранного заранее N таков:

1. Решающее дерево обучается на тренировочной выборке.
2. Обученное дерево используется для предсказания на этой же самой выборке.
3. Вычисляются остатки модели как разница между фактическими и предсказанными значениями.
4. Данные остатки встают на место объясняемой переменной (обучение на остатках), алгоритм возвращается к первому шагу, пока не будет достигнуто определённое при инициализации требуемое количество деревьев.

Финальное предсказание является простой суммой предсказанных значений N решающих деревьев.

Итоговые постановки

Таким образом, мы проведём эмпирическую проверку, используя данные о 298 специфичных для компании и макроэкономических переменных, следующих вариантов:

I. Эластичная сеть со встроенным подбором гиперпараметров.
II. Решающее дерево с циклом, перебирающим глубину дерева и выбирающим самую оптимальную оценку.
III. AdaBoost с деревом решений в качестве базового оценщика (выбор глубины будет основываться на результате из предыдущего пункта, число оценщиков – 500)
IV. Случайный лес (число оценщиков – 500, глубина – по анализу прошлых пунктов)
V. Два варианта градиентного бустинга (обычный и histogram-based, который считается более быстрым), число оценщиков и глубина дерева аналогично пункту IV.

Результаты

Алгоритм I показал полное отсутствие предсказательной силы: коэффициент детерминации имеет почти нулевое значение на тренировочной выборке и отрицательное значение на тестовой выборке. Возможной причиной неудачи может быть множество нелинейных связей, которые модель не смогла уловить.

Алгоритм II оказался немного лучше, но далек от идеала: для лучшего варианта, глубины (4), R_2 на тренировочной выборке равен 0,131, на тестовой – 0,080. Модель оказалась чуть лучше, чем предсказание средним значением.

Алгоритм III мы также рассмотрели для разной глубины дерева (рисунок 6). Как мы можем видеть, алгоритм показал примерно параболическую зависимость оценки R_2 от глубины базового классификатора.

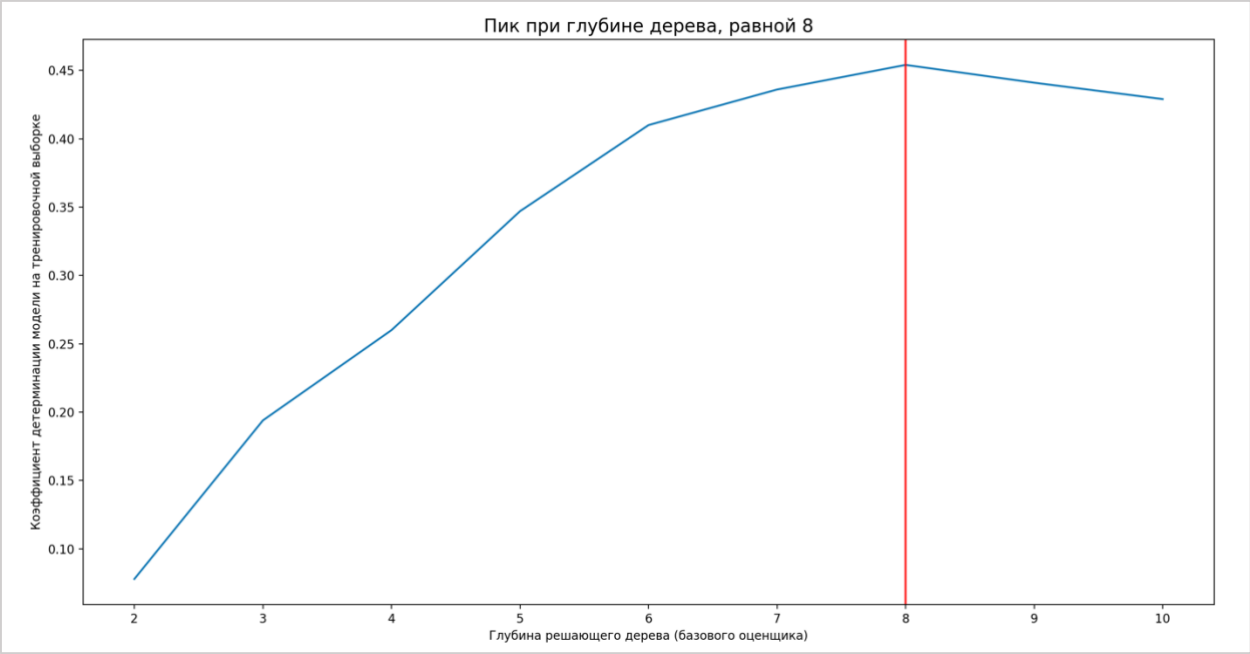


Рис. 6. Изменение качества адаптивного бустинга от глубины дерева решений.

Оценки адаптивного бустинга при лучшей глубине таковы: 0.807 на тренировочной выборке и 0.454 на тестовой. Хотя алгоритм и показал лучшую оценку на тестовой выборке именно при этой глубине, мы видим немалое переобучение алгоритма.

По алгоритму IV не получилось прийти к однозначным результатам (рисунок 7, см. ниже). Как можно увидеть, с увеличением глубины дерева решений случайный лес начинает справляться всё лучше и лучше. Дальнейшая проверка оказалась излишне времязатратной (и трудазатратной для ноутбука), однако уже и по данному графику можно наблюдать убывающий прирост в качестве.

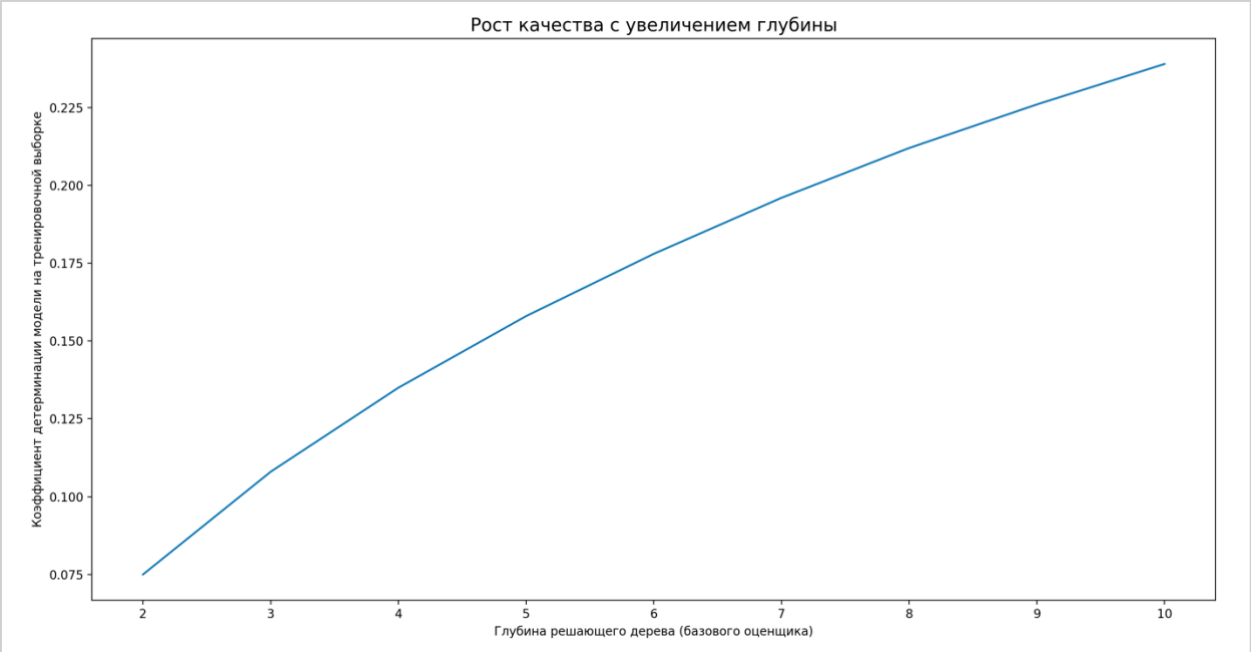


Рис. 7. Изменение качества случайного леса от глубины дерева решений.

Также алгоритму были предоставлена возможность самому выбрать глубину: пока все листья не станут чистыми или пока все листья не будут содержать менее чем 2 наблюдения. Коэффициент детерминации на тренировочной выборке равен 0.899, на тестовой – 0.281. Здесь мы тоже видим свидетельство сильного переобучения алгоритма, хотя и качество на тренировочной – наивысшее из всех рассмотренных глубин.

Алгоритм V показал впечатляющие результаты: обычный градиентный бустинг с глубиной 3 (по умолчанию) показал 0.843 и 0.606 на тренировочной и тестовой выборках соответственно. Ускоренный бустинг – 0.993 и 0.674, ограничение на глубины – нет.

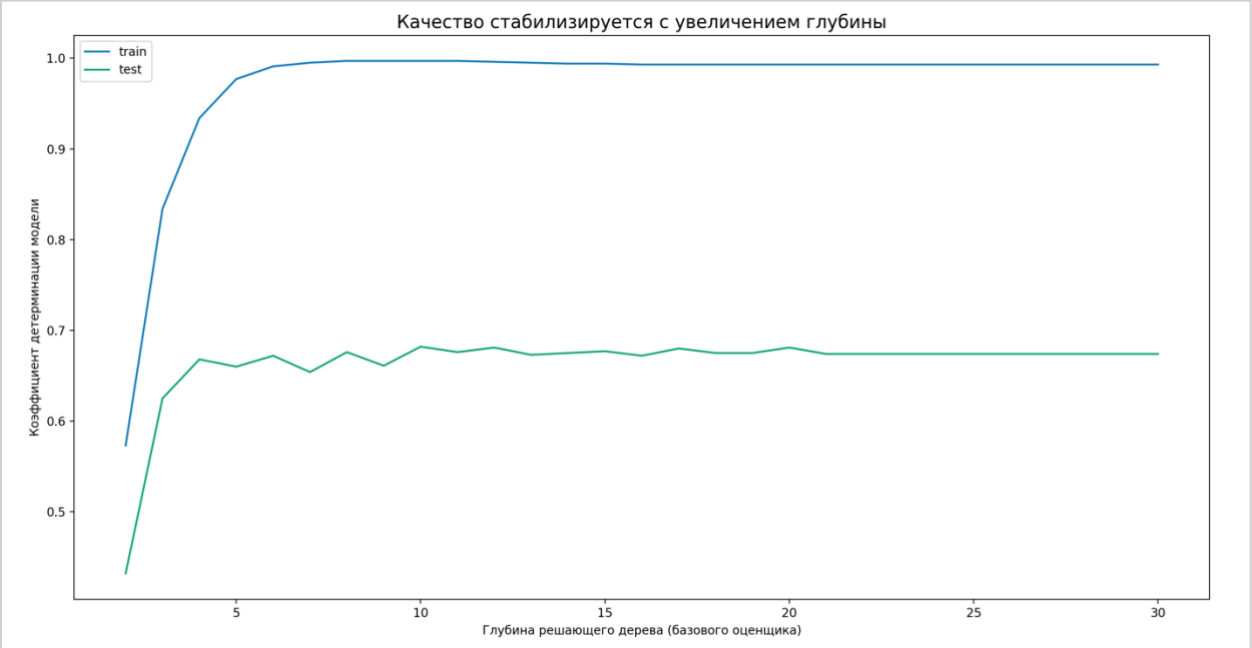


Рис. 8. Изменение качества градиентного бустинга от глубины дерева решений.

В связи с тем, что скорость данного оценщика достаточно велика, был проведён анализ качества модели в зависимости от глубины дерева с целью улучшить предсказание. Результаты представлены на графике (рисунок 8, см. выше). Мы можем наблюдать стабилизацию оценок, начиная с глубины 21 и больше. Значительно улучшить предсказание не вышло: для дальнейшего анализа будем использовать глубину 10 в связи с тем, что меньшая глубина больше способствует генерализации. Результат алгоритма на этой глубине равен 0.997 и 0.682 для тренировочной и тестовой соответственно. Это лучший результат среди всех рассмотренных нами алгоритмов и их вариаций. Изучим результаты данного оценщика подробнее.

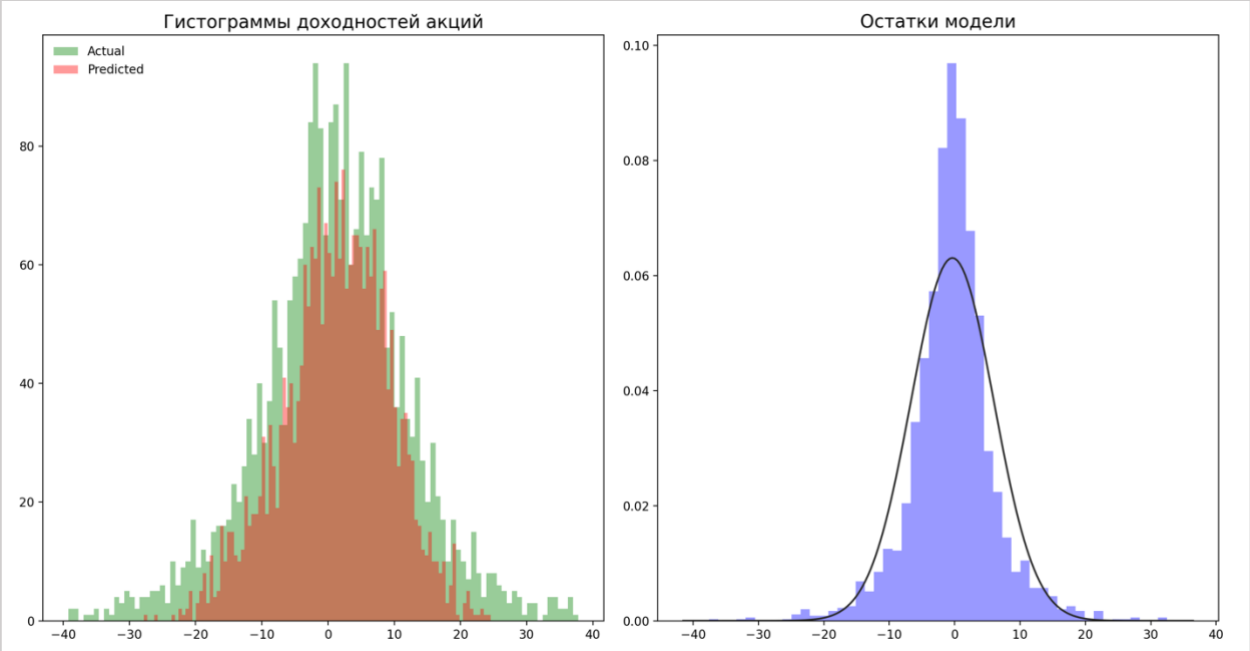


Рис. 9. Распределение предсказанных и фактических доходностей.

На рисунке 9 мы можем наблюдать, как распределены фактические и предсказанные доходности, а также разницу между ними в виде остатков. Достаточно много наблюдений были оценены с малой ошибкой: 485 из 2518 наблюдений предсказаны с точностью до 1%, 1742 наблюдения с точностью до 5%, 2244 – с точностью до 10%. Квантили распределения абсолютных значений ошибки таковы:

Квантили	0.05	0.25	0.50	0.75	0.95
Отклонение	0.23	1.37	3.10	5.93	13.67

Таблица 2. Квантили абсолютных отклонений.

Из данной таблицы мы видим, что 5% предсказаны с очень хорошей точностью: ошибка не больше 0.23%. Половина наблюдений предсказаны с точностью до 3%. Подводя итог, алгоритм показал достаточно мощную прогностическую силу.

Заключение

Мы применили несколько методов машинного обучения для предсказывания риск премий акций. Все методы обошли модифицированную линейную регрессию, которая не смогла справиться с анализом собранных данных. Нам удалось достичь неплохих результатов с помощью ансамблевых методов. В частности, лучшим из рассмотренных алгоритмов оказался градиентный бустинг (Histogram based Gradient Boosting). Многие наблюдения были предсказаны с высокой точностью.

Есть несколько вариантов проверки устойчивости результатов. В работе было использовано случайное разделение на тренировочную и тестовую выборки. Можно делать деление по датам доходностей, можно сделать сортировку по акциям: например, обучать алгоритм на всех компаниях кроме какой-то одной, и итерационно изучить, доходности чьих компаний менее или более предсказуемы. Можно взять более широкий временной горизонт, принимая в расчёт крупные кризисы, оказывающие влияние на бизнес-активность в США. Также стоит больше внимания обратить на предыдущие исследования и их способы проверки оценки качества и устойчивости прогноза.

Есть несколько путей по улучшению прогноза. Во-первых, нахождение новых переменных или преобразования уже имеющихся могут добавить информации для прогноза. Во-вторых, не были рассмотрены все комбинации гиперпараметров: анализ поведения прогноза в зависимости от определённых гиперпараметров могло бы увеличить точность. В-третьих, использование более качественных или б**о**льших по количеству данных может улучшить прогноз уже выбранных алгоритмов. В-четвёртых, можно обратить внимание на более сложные алгоритмы, как стекинг, или же обратиться к методам глубинного обучения.

Суммируя всё вышесказанное, машинное обучение открывает много возможностей в финансовой области и умелое применение алгоритмов может привести к прорывам в понимании ценообразования активов и их прогнозировании.

Приложение

Источники данных

1. https://research.stlouisfed.org/econ/mccracken/fred-databases/ – макро-статистика
2. https://finance.yahoo.com – данные по стоимостям акций
3. https://financialmodelingprep.com/developer/docs/ – финансовые отчётности и другие метрики, специфичные для компаний
4. https://www.investopedia.com – хороший сайт с разъяснением многих бухгалтерских и финансовых показателей

Описание характеристик компании[4]

Income statements (отчёт о доходах и расходах и дополнительные к нему статьи)		
№	Название	Расшифровка
1.	Revenue	Выручка
2.	Revenue Growth	Рост выручки
3.	Cost of Revenue	Расходы на создание выручки
4.	Gross Profit	Валовая прибыль (доход)
5.	R&D Expenses	Затраты на исследования и развитие
6.	SG&A Expense	Затраты продажные, общие и админ.
7.	Operating Expenses	Операционные расходы
8.	Operating Income	Операционные доходы
9.	Interest Expense	Расходы на выплату процентов
10.	Earnings before Tax (EBT)	Прибыль до выплаты налогов
11.	Income Tax Expense	Налог на доход
12.	Net Income - Non-Controlling int	ЧД с неконтролируемых долей
13.	Net Income - Discontinued ops	ЧД с прерванных операций
14.	Net Income	Чистый доход (ЧД)
15.	Preferred Dividends	Дивиденды по привилег. акциям
16.	Net Income Com	Чистый доход (ЧД)
17.	EPS	Прибыль на акцию
18.	EPS Diluted	Прибыль на акцию с учётом всех конвертируемых ценных бумаг
19.	Weighted Average Shs Out	Средневз. знач. размещённых акций
20.	Weighted Average Shs Out (Dil)	Средневз. знач. размещённых акций, с учётом всех ковертируемых бумаг
21.	Dividend per Share	Дивиденды на одну акцию
22.	Gross Margin	Маржа $= \frac{\text{Выручка} - \text{себестоимость } (COGS)}{\text{Выручка}}$
23.	EBITDA Margin	EBITDA, делённая на выручку
24.	EBIT Margin	EBIT, делённая на выручку

[4] Здесь и далее зелёным будут отмечены переменные, относящиеся к составлению объясняемых (y), оранжевым – относящиеся к составлению или использованию в качестве объясняющих (X).

25.	Profit Margin	Отношение прибыли к выручке
26.	Free Cash Flow margin	Free Cash Flow, делённая на выручку
27.	EBITDA	EBIT (operat. profit) + вся амортизация
28.	EBIT	Выручка − операционные расходы or прибыль + налоги + проценты
29.	Consolidated Income	Консолидированный доход
30.	Earnings Before Tax Margin	EBT, делённая на выручку
31.	Net Profit Margin	Чистый доход, делённый на выручку $\frac{Net\ income}{Revenue} = \frac{R-COGS-OperExp-Int-Taxes}{Revenue\ (R)}$

Balance sheet statements (статьи бухгалтерского баланса)		
№	Название	Расшифровка
1.	Cash and cash equivalents	Наличность и её эквиваленты
2.	Short-term investments	Краткосрочные инвестиции
3.	Cash and short-term investments	Наличность и SR инвестиции
4.	Receivables	Счета к получению
5.	Inventories	Запасы
6.	Total current assets	Все оборотные активы
7.	Property, Plant & Equipment Net	Основные средства
8.	Goodwill and Intangible Assets	Нематериальные активы
9.	Long-term investments	Долгосрочные инвестиции
10.	Tax assets	Налоги, оплаченные заранее
11.	Total non-current assets	Все внеоборотные активы
12.	Total assets	Все активы
13.	Payables	Счета к оплате
14.	Short-term debt	Краткосрочная задолженность
15.	Total current liabilities	Все оборотные обязательства
16.	Long-term debt	Долгосрочная задолженность
17.	Total debt	Все задолженности
18.	Deferred revenue	Доходы будущих периодов
19.	Tax Liabilities	Налоговые обязательства
20.	Deposit Liabilities	Обязательства по депозитам
21.	Total non-current liabilities	Все внеоборотные обязательства
22.	Total liabilities	Все обязательства
23.	Other comprehensive income	Доход, не вошедший в net income
24.	Retained earnings (deficit)	Нераспределённая прибыль
25.	Total shareholders equity	Общий собственный капитал
26.	Investments	Инвестиции
27.	Net Debt	Чистый долг
28.	Other Assets	Прочие активы
29.	Other Liabilities	Прочие обязательства

Cash flow statements (статьи отчётов о движении денежных средств)		
№	Название	Расшифровка
1.	Depreciation & Amortization	Амортизация (матер. и нематер.)
2.	Stock-based compensation	Оплата акциями сотрудникам
3.	Operating Cash Flow	Операционный денежный поток
4.	Capital Expenditure	Капитальные расходы
5.	Acquisitions and disposals	Приобретения и выбытия
6.	Investment purchases and sales	Инвест покупки и продажи
7.	Investing Cash flow	Инвестиционный денеж поток
8.	Issuance (repayment) of debt	Выдача (погашение) долга
9.	Issuance (buybacks) of shares	Выдача (покупка обратно) акций
10.	Dividend payments	Выплаты по дивидендам
11.	Financing Cash Flow	Финансовый денежный поток
12.	Effect of forex changes on cash	Влияние изменение форекс на наличность
13.	Net cash flow / Change in cash	Чистый денежный поток / изменение в наличности
14.	Free Cash Flow	Свободный денежный поток
15.	Net Cash/Marketcap	Чистый денежный поток / рыночная капитализация

Enterprise values (статьи, отражающие стоимость предприятия)		
№	Название	Расшифровка
1.	Stock Price	Цена акции
2.	Number of Shares	Количество акций
3.	Market Capitalization	Рыночная капитализация
4.	- Cash & Cash Equivalents	Наличность и его эквиваленты
5.	+ Total Debt	Общий долг
6.	Enterprise Value	Стоимость предприятия

Key metrics (ключевые метрики по мнению источника данных)		
№	Название	Расшифровка
1.	Revenue per Share	Выручка на акцию
2.	Net Income per Share	Чистый доход на акцию
3.	Operating Cash Flow per Share	Операционный денежный поток на акцию
4.	Free Cash Flow per Share	Свободный денежный поток на акцию
5.	Cash per Share	Наличность на акцию
6.	Book Value per Share	Балансовая стоимость на акцию
7.	Tangible Book Value per Share	Стоимость материал. активов на одну акцию
8.	Shareholders Equity per Share	Собственный капитал на акцию
9.	Interest Debt per Share	Долг по процентам на акцию
10.	Market Cap	Рыночная капитализация
11.	Enterprise Value	Стоимость предприятия
12.	PE ratio	Соотношение «цена-прибыль»

13.	Price to Sales Ratio	Соотношение «цена-выручка»
14.	POCF ratio	Соотношение «цена-операц. денежный поток»
15.	PFCF ratio	Соотношение «цена-свобод. денежный поток»
16.	PB ratio	Цена / балансовая стоимость
17.	PTB ratio	Цена / материал. баланс. стоим.
18.	EV to Sales	Стоим. предприятия / выручка
19.	Enterprise Value over EBITDA	Стоим. предприятия / EBITDA
20.	EV to Operating cash flow	Стоим. предприятия / операц. д.п.
21.	EV to Free cash flow	Стоим. предприятия / своб. д.п.
22.	Earnings Yield	Соотношение «прибыль-цена»
23.	Free Cash Flow Yield	Свобод денеж поток / цена
24.	Debt to Equity	Долг к капиталу
25.	Debt to Assets	Долг к активам
26.	Net Debt to EBITDA	Чистый долг к EBITDA
27.	Current ratio	Оборот. активы к обязательствам
28.	Interest Coverage	EBIT / расходы по процентам
29.	Income Quality	Операц. д.п. / чистый доход
30.	Dividend Yield	Соотношение «дивиденды-цена»
31.	Payout Ratio	Дивиденды / чистый доход
32.	SG&A to Revenue	SG&A / выручка
33.	R&D to Revenue	R&D / выручка
34.	Intangibles to Total Assets	Нематериальные ко всем активам
35.	Capex to Operating Cash Flow	Капитальные затраты к операц. д.п.
36.	Capex to Revenue	Капитальные затраты к выручке
37.	Capex to Depreciation	Капитальные затраты к аморт.
38.	Stock - based compensation to Revenue	Плата акциями к выручке
39.	Graham Number	$\sqrt{22.5 * \text{приб. на ак} * \text{балан. ст. на ак.}}$
40.	Graham Net-Net	Скорр. активы – все обязательства
41.	Working Capital	Оборот. активы – оборот. обязател.
42.	Tangible Asset Value	Оборотные активы – все обязательс.
43.	Net Current Asset Value	Оборот. активы – оборот. обязател.
44.	Invested Capital	Инвестиционный капитал
45.	Average Receivables	Ср. дебиторская задолженность
46.	Average Payables	Ср. кредиторская задолженность
47.	Average Inventory	Ср. инвентаризация
48.	Capex per Share	Капитальные затраты на акцию

Financial statement growths (показатели роста финансовой отчётности)		
№	Название	Расшифровка
1.	Gross Profit Growth	Рост валовой прибыли
2.	EBIT Growth	Рост EBIT
3.	Operating Income Growth	Рост операц. прибыли

4.	Net Income Growth	Рост чистой прибыли
5.	EPS Growth	Рост EPS
6.	EPS Diluted Growth	Рост DG
7.	Weighted Average Shares Growth	Рост WAS
8.	Weighted Average Shares Diluted Growth	Рост WASD
9.	Dividends per Share Growth	Рост DPS
10.	Operating Cash Flow growth	Рост OCF
11.	Free Cash Flow growth	Рост FCF
12.	Receivables growth	Рост R
13.	Inventory Growth	Рост I
14.	Asset Growth	Рост A
15.	Book Value per Share Growth	Рост BPS
16.	Debt Growth	Рост D
17.	R&D Expense Growth	Рост R&D E
18.	SG&A Expenses Growth	Рост SG&A E

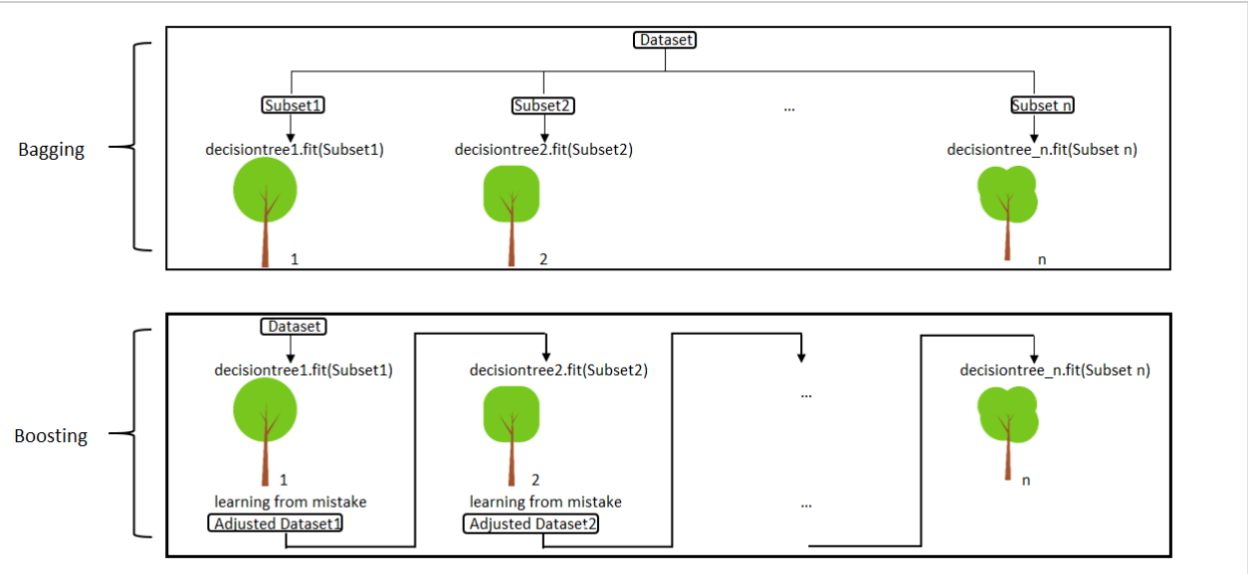
Описание макроэкономических переменных

Данная таблица хороша описана и полностью представлена в работе McCracken, Ng 2020, посвящённая собранным ими квартальным макро-данным. Помимо описания также предложены эталонные преобразования для всех рядов.

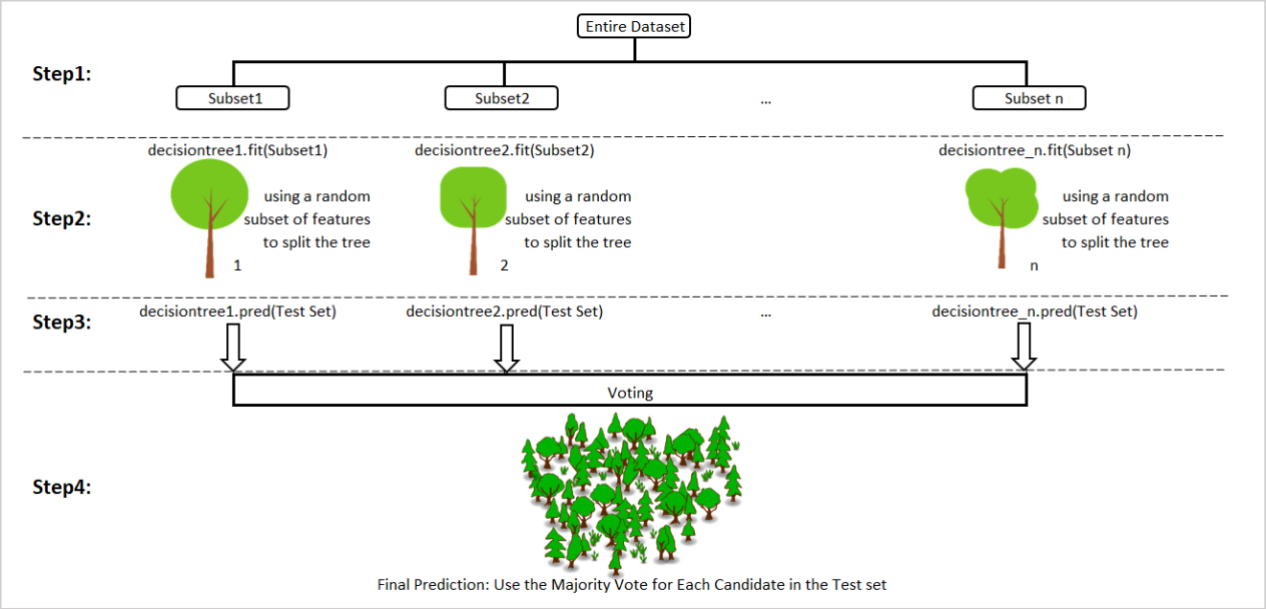
Ансамбли в картинках

Более подробно алгоритмы описаны в данной статье, здесь же в приложении размещу картинки из статьи, наглядно объясняющие работу некоторых композиционных методов:

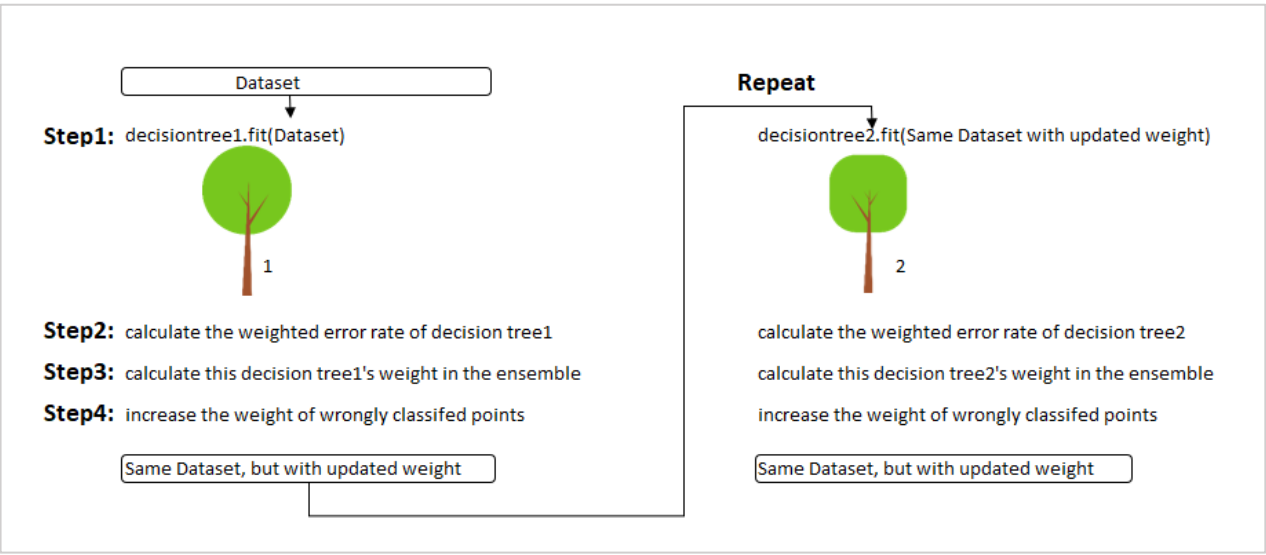
Bagging и Boosting

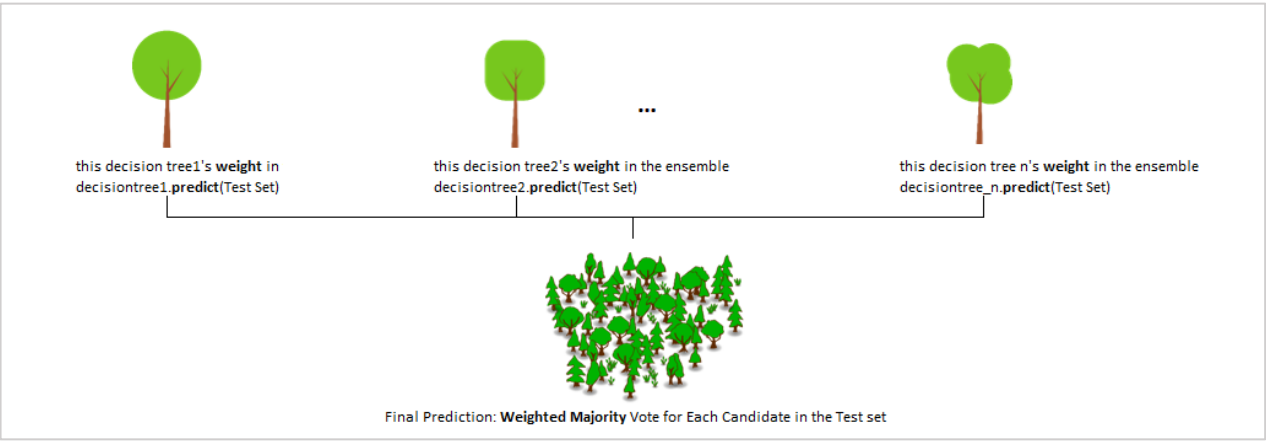


Random Forest

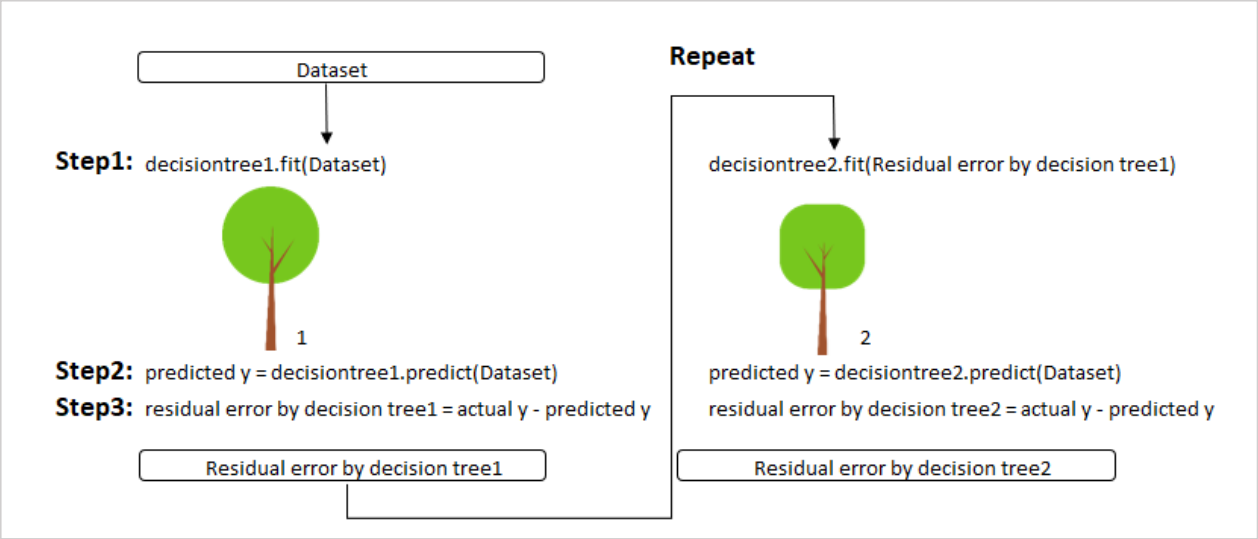


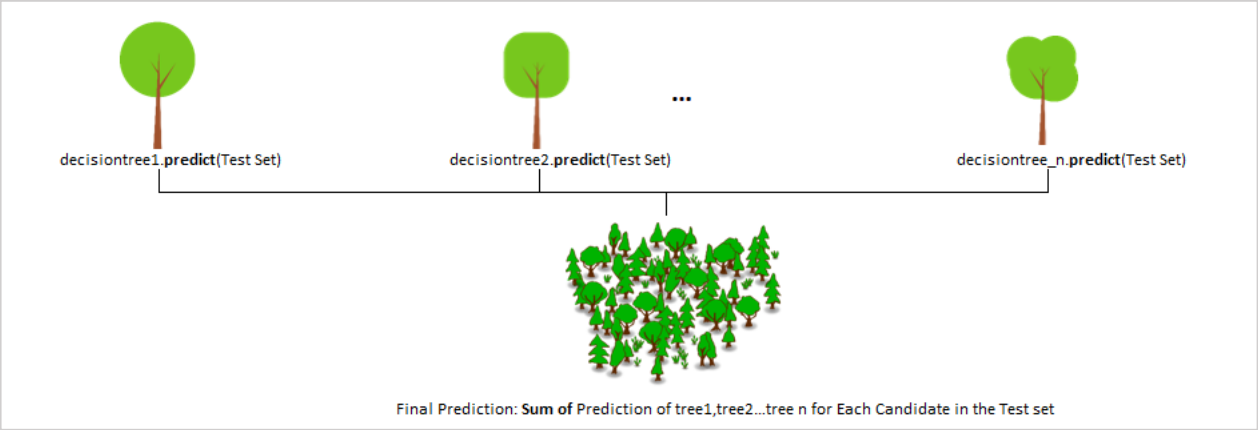
AdaBoost (Adaptive Boosting)





Gradient Boosting





Код работы

Для реплицирования действий и результатов, описанных в работе, оставляю здесь ссылку на репозиторий GitHub со всеми использованными данными и самим кодом, написанным на Python:

- https://github.com/NBar05/Coursework_in_Finance

Список литературы

Книги

1. Bishop C. M. Pattern recognition and machine learning. – springer, 2006.
2. Duffie D. Dynamic asset pricing theory. – Princeton University Press, 2010.
3. Back K. Asset pricing and portfolio choice theory. – Oxford University Press, 2010.
4. Lee C. F., Lee J. (ed.). Handbook of quantitative finance and risk management. – Springer Science & Business Media, 2010.
5. De Prado M. L. Advances in financial machine learning. – John Wiley & Sons, 2018.

Статьи

1. Roll R., Ross S. A. An empirical investigation of the arbitrage pricing theory //The Journal of Finance. – 1980. – Т. 35. – №. 5. – С. 1073-1103.
2. Connor G., Korajczyk R. A. Risk and return in an equilibrium APT: Application of a new test methodology //Journal of Financial Economics (JFE). – 1988. – Т. 21. – №. 2.
3. Fama E. F., French K. R. The cross-section of expected stock returns //the Journal of Finance. – 1992. – Т. 47. – №. 2. – С. 427-465.
4. Lakonishok J., Shleifer A., Vishny R. W. Contrarian investment, extrapolation, and risk //The journal of finance. – 1994. – Т. 49. – №. 5. – С. 1541-1578.
5. Fama E. F., French K. R. Multifactor explanations of asset pricing anomalies //The journal of finance. – 1996. – Т. 51. – №. 1. – С. 55-84.
6. Priestley R. The arbitrage pricing theory, macroeconomic and financial factors, and expectations generating processes //Journal of Banking & Finance. – 1996. – Т. 20. – №. 5. – С. 869-890.
7. Fama E. F., French K. R. The capital asset pricing model: Theory and evidence //Journal of economic perspectives. – 2004. – Т. 18. – №. 3. – С. 25-46.
8. Azeez A. A., Yonezawa Y. Macroeconomic factors and the empirical content of the Arbitrage Pricing Theory in the Japanese stock market //Japan and the world economy. – 2006. – Т. 18. – №. 4. – С. 568-591.
9. Fama E. F., French K. R. Dissecting anomalies //The Journal of Finance. – 2008. – Т. 63. – №. 4. – С. 1653-1678.
10. Welch I., Goyal A. A comprehensive look at the empirical performance of equity premium prediction //The Review of Financial Studies. – 2008. – Т. 21. – №. 4. – С. 1455-1508.
11. Lewellen J., Nagel S., Shanken J. A skeptical appraisal of asset pricing tests //Journal of Financial economics. – 2010. – Т. 96. – №. 2. – С. 175-194.

12. Green J., Hand J. R. M., Zhang X. F. The supraview of return predictive signals //Review of Accounting Studies. – 2013. – Т. 18. – №. 3. – С. 692-730.
13. Lewellen J. The cross section of expected stock returns //Forthcoming in Critical Finance Review. – 2014.
14. Oliveira M. R., Torgo L. Ensembles for time series forecasting. – 2014.
15. Fama E. F., French K. R. A five-factor asset pricing model //Journal of financial economics. – 2015. – Т. 116. – №. 1. – С. 1-22.
16. Kisman Z., Restiyanita S. M. The Validity of Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) in Predicting the Return of Stocks in Indonesia Stock Exchange //American Journal of Economics, Finance and Management. – 2015. – Т. 1. – №. 3. – С. 184-189.
17. Harvey C. R., Liu Y., Zhu H. … and the cross-section of expected returns //The Review of Financial Studies. – 2016. – Т. 29. – №. 1. – С. 5-68.
18. Moritz B., Zimmermann T. Tree-based conditional portfolio sorts: The relation between past and future stock returns //Available at SSRN 2740751. – 2016.
19. French J. Macroeconomic forces and arbitrage pricing theory //Journal of Comparative Asian Development. – 2017. – Т. 16. – №. 1. – С. 1-20.
20. Giglio S., Xiu D. Inference on risk premia in the presence of omitted factors. – National Bureau of Economic Research, 2017. – №. w23527.
21. Kelly B., Pruitt S., Su Y. Some characteristics are risk exposures, and the rest are irrelevant //Unpublished Manuscript, University of Chicago. – 2017.
22. Gu S., Kelly B., Xiu D. Empirical asset pricing via machine learning. – National Bureau of Economic Research, 2018. – №. w25398.
23. Chen L., Pelger M., Zhu J. Deep learning in asset pricing //Available at SSRN 3350138. – 2019.
24. Feng G., Polson N., Xu J. Deep learning in characteristics-sorted factor models //Available at SSRN 3243683. – 2019.
25. Harvey C. R., Liu Y. Lucky factors //Available at SSRN 2528780. – 2019.
26. Freyberger J., Neuhierl A., Weber M. Dissecting characteristics nonparametrically //The Review of Financial Studies. – 2020. – Т. 33. – №. 5. – С. 2326-2377.
27. McCracken M., Ng S. FRED-QD: A Quarterly Database for Macroeconomic Research. – National Bureau of Economic Research, 2020. – №. w26872.