Содержание

[Глава 1. Теоретические основы рейтингового ранжирования компаний 2](#_Toc39749010)

[1.1 Обзор основных концепций составления рейтинговых систем 2](#_Toc39749011)

[1.2 Линейные методы решения задачи двоичной классификации 6](#_Toc39749012)

[1.3 Нелинейные методы решения задачи двоичной классификации, основанные на параллельном обучении 11](#_Toc39749013)

[1.4 Нелинейные методы решения задачи двоичной классификации, основанные на последовательном обучении 18](#_Toc39749014)

[1.5 Метрики качества решений, кросс-валидация и рейтинговая шкала 24](#_Toc39749015)

[Глава 2. Конструирование, обработка и отбор факторов для построения моделей 31](#_Toc39749016)

[2.1 Описание выборки компаний, используемых для анализа 31](#_Toc39749017)

[2.2 Конструирование целевой переменной модели 36](#_Toc39749018)

[2.3 Конструирование, обработка и отбор факторов внутреннего контура модели 43](#_Toc39749019)

[2.4 Конструирование, обработка и отбор факторов макроконтура модели 52](#_Toc39749020)

[2.5 Конструирование, обработка и отбор факторов рыночного контура модели 57](#_Toc39749021)

[Глава 3. Эмпирические результаты исследования 63](#_Toc39749022)

[3.1 Настройка гиперпараметров линейных моделей 63](#_Toc39749023)

[3.2 Настройки гиперпараметров моделей случайного леса. 66](#_Toc39749024)

[3.3 Настройка гиперпараметров моделей градиентного бустинга. 73](#_Toc39749025)

[3.4 Результаты рейтинговой оценки компаний 83](#_Toc39749026)

[Приложения 92](#_Toc39749027)

# Глава 1. Теоретические основы рейтингового ранжирования компаний

## 1.1 Обзор основных концепций составления рейтинговых систем

Зачастую люди понимают концепцию рейтинга недостаточно точно и путают ее с ранжированием. Разница заключается в том, что построение рейтинга базируется на системе факторов, в то время как ранжирование осуществляется на основе одного показателя. Другими словами, рейтинг, в сравнении с ранжированием, характеризуется более комплексным подходом.[[1]](#footnote-1) Например, если мы говорим о месте, которое занимает компания в списке «Топ 100 компаний по балансовой стоимости», то мы говорим о ранжировании, если же для оценки места компании на шкале мы используем множество факторов, таких как активы, капитальные затраты, рентабельности активов и любые иные, то мы говорим о рейтинге.

Составление рейтинга финансовых инструментов компаний или же компаний в целом – процесс дорогостоящий. Это объясняется необходимостью для рейтинговых компаний систематически собирать необходимые для оценки данные и заботиться о своей репутации, которая может быть поставлена под сомнение в любой момент.[[2]](#footnote-2)Среди всех рейтинговых агентств можно выделить 3 ключевых игрока: Moody’s, Fitch, S&P’s. Эти компании доминируют на мировой арене.

Обычно в финансах используются кредитные рейтинги. Они предоставляют инвесторам информацию о кредитоспособности компании. Методология составления таких рейтингов не до конца прозрачна, в основном, из-за вопросов, связанных с коммерческой тайной компаний, их публикующих.

Разберем существующие подходы конструирования рейтинговых систем. Интуитивно их можно разделить на две группы:

1. Основанные на математическом моделировании;
2. Основанные на мнении экспертов.

Под рейтингом, основанным на экспертном мнении, подразумевается рейтинговая система, построенная на основе экспертных оценок некоторых параметров, например, такого как инвестиционный климат. Эти оценки могут быть подкреплены количественными методами, однако веса перед факторами в итоговой модели выставляются, исходя из субъективных суждений. Подобные рейтинги широко применяются в тех случаях, когда доступ к информации сильно лимитирован и большинство имеющихся параметров имеют качественную, а не количественную природу. Рейтинговые модели, основанные на экспертном суждении, широко применяются в венчурной индустрии, например, при оценке привлекательности стартапов.[[3]](#footnote-3)

Описанная методология не подходит для целей текущего исследования. Компании, входящие в индекс S&P500, публикуют большой объем количественной информации, который имеет смысл обрабатывать, используя математические методы, а не субъективные суждения.

Далее рассмотрим рейтинговые модели, предполагающие наличие вычислений. Их можно грубо поделить на две категории[[4]](#footnote-4):

1. Модели составного рейтинга
2. Модели рейтинга на основе регрессионного анализа

Идея составного рейтинга заключается в следующем: оценивается ранг объектов по каждому фактору отдельно, а затем на основе полученных значений вычисляется интегральный рейтинг.[[5]](#footnote-5) Этот подход позволяет создавать динамические рейтинги. Для этого нужно использовать в рейтинговой модели не только текущие значения факторов, но и их предыдущие значения.

В качестве примера можно привести реализацию типичного составного рейтинга. В первую очередь осуществляется процесс вычисления следующей величины:

– модифицированное значение фактора j на объекте k

– значение фактора j на объекте k в момент времени t

Рассчитанная величина учитывает влияние значений фактора j в периоды, предшествующие исследованию, а также определяет место объекта i среди иных объектов ранговой системы по фактору j.

Далее для каждого объекта выборки складываются модифицированные значения факторов D и получается вектор интегральных показателей:

– значение интегрального показателя для объекта j

Полученный вектор величин преобразуется в вариационный ряд и разделяется на R интервалов по квантилям. Каждому объекту присваивается рейтинг в зависимости от того, к какому интервалу относится значение его интегрального показателя.

Учитывая, что природа влияния факторов на полезность не всегда известна заранее, предположение о положительной корреляции значения фактора и функции полезности ставить под вопрос эффективность использования подобных моделей в рамках данной работы. Также нужно отметить, что метод составных рейтингов основан на относительных, а не абсолютных значениях. Следовательно, использовать качественные факторы, превращенные в бинарные переменные, невозможно.

Что касается регрессионных моделей, то в данном случае существует два основных подхода для их использования. Первый подход применяется в том случае, если существует выборка компаний, которая уже имеет рейтинги, и исследователю необходимо расширить область применения этих рейтингов на иные компании, которые этих рейтингов не имеют. Второй подход применяется в том случае, если в распоряжении исследователя отсутствует обучающая совокупность компаний с рейтингами. В такой ситуации для генерирования рейтингов используются регрессионные модели бинарного выбора.

Фактически, задачи, решаемые в рамках рассмотренных выше подходов, являются задачами классификации. Разница между ними заключается в том, что в первом случае речь идет о задаче множественной классификации, а во втором о бинарной. В рамках первого подхода метками классов является дискретное множество объектов, представляющее собой значения оценок, которые присваиваются компаниям авторитетными рейтинговыми агентствами, например, если речь идет о кредитных рейтингах, такими как Moody’s, Fitch, S&P’s.

Первый подход реализовать в рамках данного исследования невозможно ввиду того, что доступ к существующим рейтингам инвестиционной привлекательности компаний крайне ограничен. Существуют небольшие компании, которые публикуют подобные рейтинги достаточно открыто, однако методология построения таких рейтингов закрыта, а качество сомнительно.

В рамках второго подхода объясняемая переменная может принимать лишь два значения 0 или 1, неудача или успех. Этот подход наиболее релевантен, так как он обеспечивает возможность свободно выбирать, что будет считаться успехом, а что неудачей. В рамках данной работы будем считать, что компания является успешной, если ее ожидаемая доходность превышает среднеотраслевую доходность.

Говоря о моделях бинарной классификации, мы можем использовать опыт банковского сектора, где эти модели используются чаще, поскольку дефолты банков существенно влияют как на экономику, так и на клиентов (Карминский, Костров, Мурзенков, 2012). На примере банков мы видим, что эти модели легко превращаются в рейтинговые модели, где рейтинг присваивается в соответствии с оценочной вероятностью дефолта с использованием мастер-шкалы. Пример такого подхода приведен в (Моргунов, 2017) и (Карминский, Полозов, 2016). Этот подход будем использовать в дальнейшем, поскольку он позволяет использовать категориальные переменные и не требует наличия существующего рейтинга, присвоенного рейтинговым агентством. Далее более подробно разберем методы решения поставленной задачи, а также варианты ее верификации.

## 1.2 Линейные методы решения задачи двоичной классификации

Методы решения задач двоичной классификации можно разделить на два основных вида:

1. методы линейной классификации;
2. методы нелинейной классификации.

Суть методов линейной классификации заключается в том, что данное семейство алгоритмов пытается разделить признаковое пространство, описывающее объекты выборки, на два полупространства с помощью гиперплоскости так, чтобы в одном полупространстве находились объекты класса 1, а в другом класса -1. Если гиперплоскость можно провести таким образом, что объекты разделятся на классы без ошибок, то выборка объектов называется линейно разделимой.

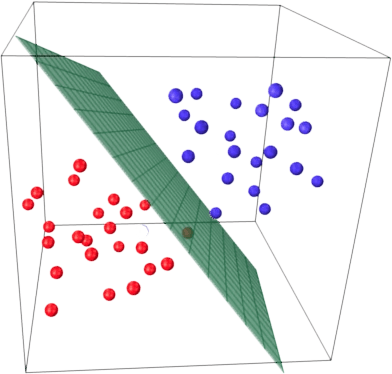


Рис. Визуализация работы линейного классификатора.

Формализуем задачу линейной классификации. Пусть - пространство ответов, – пространство объектов,— признаковое описание объекта, – обучающая выборка.

Классическим методом решения задачи линейной классификации является метод линейного дискриминанта Фишера, который заключается в минимизации среднеквадратичной ошибки:

В результате получается вектор весов:

Откуда выходит, что если , то объект относится к классу 1, иначе к классу -1. Иными словами получается следующее выражение:

Где:

* – вектор ответов классификатора на объектах выборки;
* Sign – функция, возвращающая знак своего аргумента.

Основная проблема данного алгоритма заключается в том, что он предоставляет метки класса, но не вероятность принадлежности объекта классу, а в задаче составления рейтинга способность алгоритма определять вероятность принадлежности объекта классу является критичной. Другими словами, необходимо ввести функцию, которая будет переводить интервал (0,1) на множество всех действительных чисел:

В таком случае, можно решать задачу линейной регрессии, в которой строится оценка не для условного матожидания , а для .

В статистике такое семейство моделей называется обобщенными линейными моделями. В задаче бинарной классификации в качестве часто используется сигмоида.

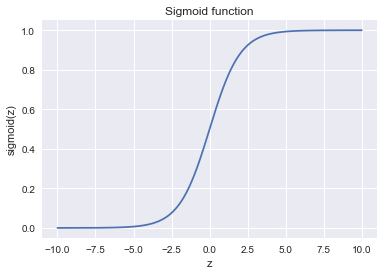


Рис. Функция сигмоида

Выбор функции сигмоида в качестве функции g обусловлен нижеследующим. Пусть – условная вероятность события , а – отношение вероятностей этого события, которое определяется следующим образом:

Ясно, что если вероятность рассматриваемого события – это величина, которая лежит в интервале от 0 до 1, то отношение вероятностей – это величина, которая лежит в интервале от 0 до ∞. Если вычислить логарифм отношения вероятностей, то легко заметить, что рассчитанная величина будет лежать на множестве всех вещественных чисел, т.е. . Следовательно, если

То:

Из вышесказанного следует, что рассмотренный алгоритм прогнозирует вероятность отнесения объекта к классу +1 с помощью сигмоид-преобразования линейной комбинации вектора весов модели и вектора признаков объекта.

Настраивание модели осуществляется с помощью метода максимального правдоподобия. Принцип максимизации правдоподобия приводит к минимизации логистической функции потерь, которая принимает следующий вид:

Основным преимуществом рассмотренного алгоритма является его простота и скорость работы, однако стоит понимать, что линейные методы классификации строят достаточно тривиальную разделяющую поверхность – гиперплоскость, которая не справляется с более сложными закономерностями, встречающимися в данных. Самый известный пример, в котором данные нельзя поделить на два полупространства гиперплоскостью, получил название «the XOR problem". В рассматриваемой задаче бинарной классификации классы представлены вытянутыми по диагоналям и пересекающимися облаками точек.

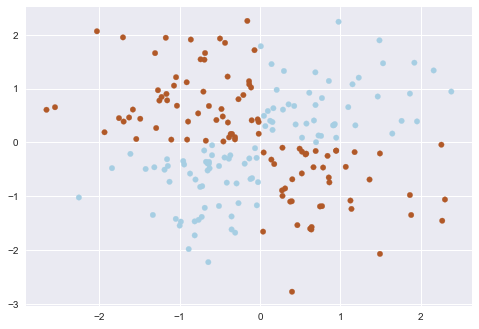


Рис. Иллюстрация «the XOR problem»

Очевидно, что нельзя провести прямую так, чтобы без ошибок отделить один класс от другого. Поэтому логистическая регрессия плохо справляется с такой задачей. На рисунке, представленном ниже, наглядно отображается предсказательная мощность различных алгоритмов машинного обучения и их интерпретируемость.

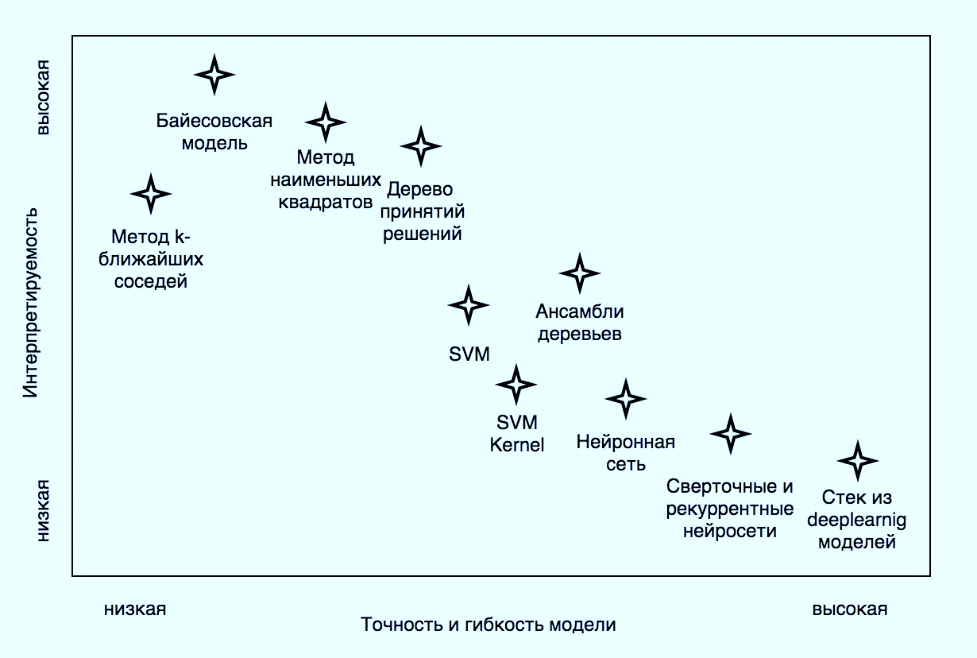


Рис. Зависимость гибкости алгоритма машинного обучения и интерпретируемости полученной модели.

Из рисунка видно, что линейные модели обладают высоким уровнем интерпретируемости, однако они недостаточно точные и гибкие в сравнении с иными моделями машинного обучения. Поэтому далее рассмотрим нелинейные методы классификации, а именно алгоритмы случайного леса и градиентного бустинга, которые также будут использоваться в создании рейтинга инвестиционной привлекательности компаний.

## 1.3 Нелинейные методы решения задачи двоичной классификации, основанные на параллельном обучении

Случайный лес – нелинейный алгоритм машинного обучения, составными частями которого являются деревья принятия решений, объединённые в ансамбль[[6]](#footnote-6).

Дерево принятия решений, как алгоритм машинного обучения, представляет собой совокупность логических правил, на основе которых производится классификация объектов по классам, исходя из их признакового описания так, как это представлено на рисунке ниже.

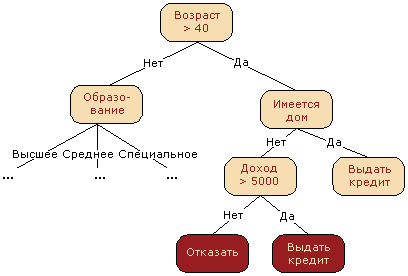


Рис. 1 Структура дерева принятия решений

Критерием разбиения структуры данных в узле на подмножества выступает такой показатель, как энтропия Шеннона. Этот показатель определяется для системы с N возможными состояниями по следующей формуле:

*Где:*

*– вероятность нахождения системы в состоянии i*

Чем более однородно подмножество – тем меньше энтропия, и наоборот, чем выше энтропия, тем больше хаоса в системе. Следовательно, уменьшение энтропии приводит к повышению информации в системе. Прирост информации (IG) по признаку Q при разбиении выборки рассчитывается следующим образом:

*Где:*

*– значение показателя энтропии до разбиения выборки по признаку Q*

*– число объектов в подвыборке i после разбиения выборки*

*– значение показателя энтропии в подвыборке i после разбиения*

*q – число подвыборок после разбиения*

Многие популярные алгоритмы построения деревьев принятия решений, например, такие как ID3 и C4.5, основаны на принципе жадной максимизации показателя IG, другими словами, на каждой итерации выбирается такой признак, разделение по которому приводит к наибольшему приросту информации в системе. Дальше процедура повторяется рекурсивно, пока энтропия не окажется равной нулю или какой-то малой величине (если дерево не подгоняется идеально под обучающую выборку во избежание переобучения)[[7]](#footnote-7).

Одной из главных проблем деревьев принятия решений является их нестабильность, которая приводят к тому, что незначительные изменения в структуре входных данных могут существенно повлиять на алгоритм построения дерева. Решать эту проблемы призваны ансамбли. Ансамбль алгоритмов – метод, который использует несколько обучающих алгоритмов с целью получения лучшей эффективности прогнозирования, чем можно было бы получить от каждого обучающего алгоритма по отдельности.[[8]](#footnote-8) Классическим примером пользы ансамблей в задачах классификации является теорема Кондорсе «о жюри присяжных».

Если каждый член жюри присяжных имеет независимое мнение, и если вероятность правильного решения члена жюри больше 0.5, то тогда вероятность правильного решения присяжных (R) возрастает с увеличением количества членов жюри, и стремится к единице.

*Где:*

*М – количество членов жюри присяжных*

*p – вероятность верного решения конкретного эксперта*

*m – минимальное большинство членов жюри*

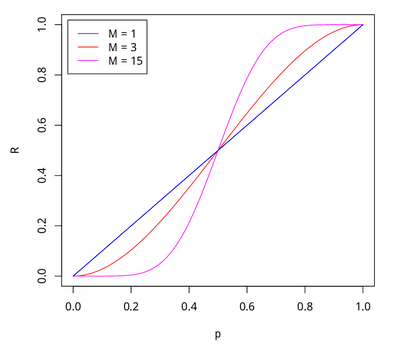


Рис. Распределение вероятности правильного решения жюри

На рисунке представлена вероятность правильного решения жюри присяжных в зависимости от вероятности, с которого конкретный эксперт дает верный прогноз, при разном числе членов жюри.

Простейшим видом ансамблей деревьев принятий решений является бэггинг[[9]](#footnote-9). Работа данного алгоритма основана на методе бутстрэпа, суть которого заключается в нижеследующем. Из исходной совокупности последовательно равновероятно с возвращением извлекаются элементы, формируя новую выборку. Повторяя процедуру N раз, генерируется новых выборок, на основе которых можно оценивать различные статистики исходного распределения. Графически описанный процесс представлен на рисунке ниже.

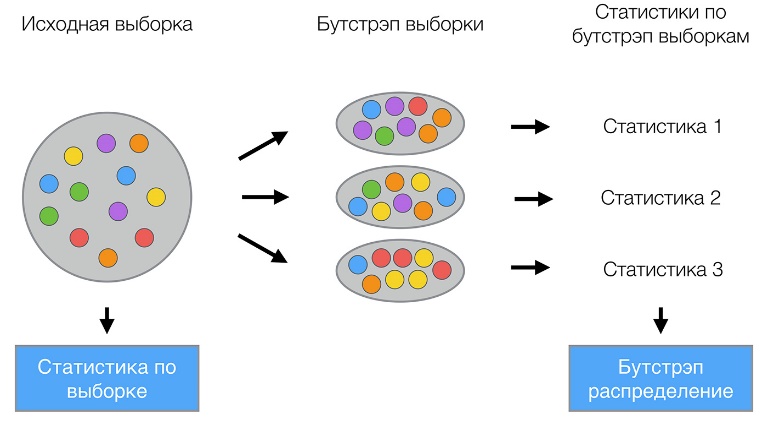


Рис. Иллюстрация инструмента бутсрэп

Под бэггингом понимается алгоритм, который принимает ответы деревьев принятия решений, обученных на бутстрапированных выборках исходной совокупности, и выдает окончательное решение, например, на основе простого большинства. Графическая иллюстрация алгоритма представлена на рисунке ниже.

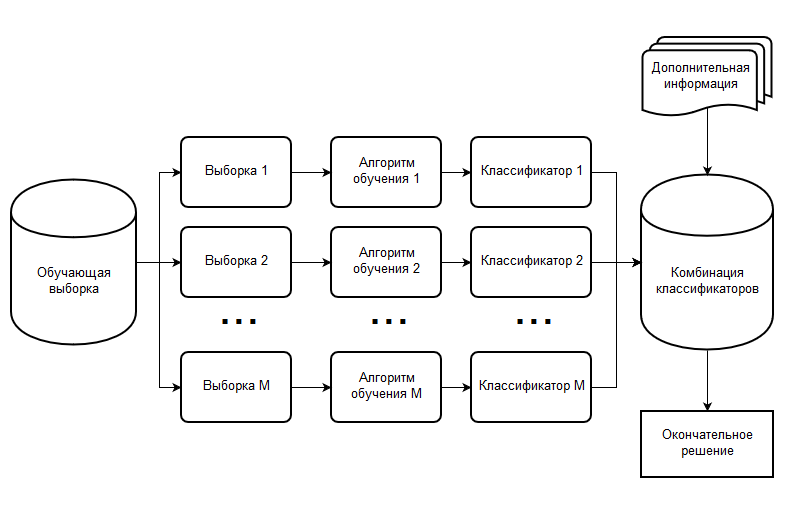


Рис. Графическая реализация алгоритма бэггинга

Преимущество бэггинга перед решающими деревьями можно показать, используя следующие выкладки. Допустим, решается задача регрессии и при этом имеется базовых алгоритмов, настроенных на решение поставленной задачи. Также представим, что существует истинная зависимость, которая выражается с помощью функции , тогда ошибку каждого алгоритма на объектах можно записать в следующей форме:

А также можно записать математическое ожидание среднеквадратичной ошибки в следующей форме:

Средняя ошибка полученных алгоритма имеет вид:

Если же предположить, что ошибки не коррелированы и не смещены, и построить новую функцию регрессии, аналогично тому, как это осуществляется в алгоритме бэггинга, усредняя ответы уже построенных алгоритмов следующим образом:

То среднеквадратичная ошибка в этом случае примет следующий вид:

Таким образом, при использовании бэггинга ошибка предсказания сокращается в n раз. Эффективность бэггинга объясняется тем, что алгоритмы обучаются на разных выборках, однако на практике далеко не всегда получается выполнить предпосылки некоррелированности ошибок, что приводит к тому, что ошибка снижается не так существенно. Для решения этой проблемы алгоритм бэггинга был усовершенствован Адель Катлером и Лео Брейманом, которые добавили в алгоритм метод случайных подпространств[[10]](#footnote-10). Этот метод предполагает обучение деревьев на различных подмножествах признакового описания объектов, которые выделяются случайным образом. Итоговый вариант алгоритма получил название случайного леса.

Как уже было сказано выше, при решении данной задачи классификации требуется не просто предсказать метку класса, но получить вероятность соответствующей метки, на основе которой будет строиться рейтинговая шкала. Линейные модели, например логистическая регрессия, описанная выше, дают неплохие оценки вероятности принадлежности объекта к классу. Однако нелинейные модели могут давать плохие оценки вероятностей класса. Следовательно, их необходимо калибровать.

Хорошо откалиброванные классификаторы – это вероятностные классификаторы, для которых выходные данные метода прогнозирования могут напрямую интерпретироваться как уровень достоверности. Например, хорошо откалиброванный классификатор должен классифицировать объекты выборки так, чтобы среди тех объектов, для которых он дал значение предиката близкое к 0,8, примерно 80% объектов фактически принадлежали к положительному классу. Следующий график сравнивает, насколько хорошо откалиброваны вероятностные прогнозы различных классификаторов:

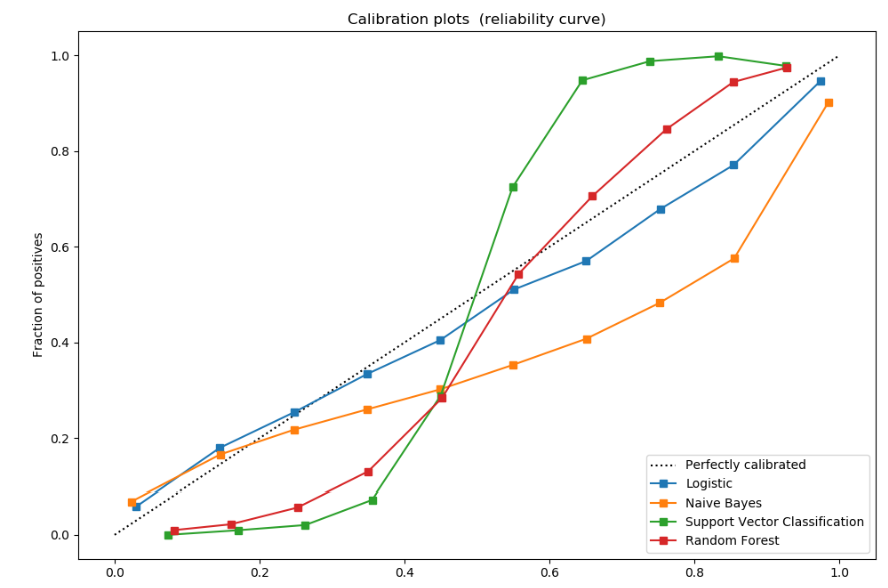


Рис. Диаграмма надежности классификаторов

Из рисунка видно, что алгоритм случайного леса показывает пики вероятностей в районе 20% и 90%, в то время как вероятности, близкие к 0 или 1, очень редки. Объяснение этому дают Niculescu-Mizil и Caruana[[11]](#footnote-11) в своих работах. У таких алгоритмов, как бэггинг и случайные леса, которые усредняют прогнозы базового набора моделей, могут возникать трудности при прогнозировании вероятностей классов около 0 и 1. Эта проблема возникает в следствие того, что дисперсия предиктов базовых моделей смещает прогнозы от 0 и 1. Например, если модель должна прогнозировать вероятность класса равную 0, то единственный способ, которым она может быть достигнута – это если все деревья в мешках предсказывают ноль. Однако если вспомнить суть алгоритма случайного леса, в котором каждое дерево обучается на бутстрапированной выборке и на случайном подпространстве признакового описания объекта, то вполне очевидно, что вероятность того, что хотя бы один из базовых алгоритмов проголосует за 1, тем самым сдвинув среднее предсказание ансамбля от 0, достаточно высока. В результате калибровочная кривая, также называемая диаграммой надежности[[12]](#footnote-12), имеет характерную сигмовидную форму, указывая на то, что классификатор может больше доверять своей «интуиции» и возвращать вероятности ближе к 0 или 1.

Выделяют два подхода для калибровки вероятностных прогнозов, а именно параметрический подход, основанный на сигмоидальной модели Платта[[13]](#footnote-13) и непараметрический подход, основанный на изотонической регрессии. Оба этих подход реализованы в библиотеке scikit learn[[14]](#footnote-14) и используются в работе.

## 1.4 Нелинейные методы решения задачи двоичной классификации, основанные на последовательном обучении

Нелинейный метод решения задачи двоичной классификации, основанный на параллельном обучении и описанный выше, имеет ряд проблем. Первая проблема заключается в том, что алгоритм случайного леса является ненаправленным жадным алгоритмом, т.е. каждое новое дерево строится отдельно от всех остальных, поэтому для успешной работы алгоритма требуется построение большого числа деревьев.

Вторая проблема вытекает из первой. Так как суть алгоритма заключается в построении большого количества глубоких переобученных деревьев, это, естественно, требует немалого количества вычислительных ресурсов, особенно если речь идет о гигантских выборках с большим количество признаков. При этом в данном случае нельзя просто ограничить глубину решающих деревьев, так как это приведет к тому, что алгоритм станет неспособен улавливать сложные связи и закономерности в данных.



Рис. Разделяющая поверхность, которую строит случайный лес, ограниченный по глубине

Одним из вариантов решения проблем, описанных выше, является использование бустинга. Под бустингом понимается подход, в рамках которого базовые алгоритмы строятся друг за другом, а не параллельно, при этом каждый алгоритм настраивается таким образом, чтобы исправлять ошибки построенной композиции в целом.

Благодаря тому, что построение композиций в бустинге является направленным, достаточно использовать простые базовые алгоритмы, например неглубокие деревья.

На примере задачи классификации алгоритм бустинга можно описать следующим образом. Пусть задана функция потерь , где y – истинный ответ, z – прогноз алгоритма на некотором объекте. В задаче классификации примером функции потерь может быть рассмотренная выше логистическая функция потерь:

В начале построения композиции по методу градиентного бустинга нужно ее инициализировать, то есть построить первый базовый алгоритм . Этот алгоритм не должен быть сколько-нибудь сложным и не стоит тратить на него много усилий. Можно использовать алгоритм, который всегда возвращает метку самого распространенного класса в обучающей выборке.

Обучение базовых алгоритмов происходит последовательно. Пусть к некоторому моменту обучены алгоритмов , то есть композиция имеет вид:

Теперь к текущей композиции добавляется еще один алгоритм . Этот алгоритм обучается так, чтобы как можно сильнее уменьшить ошибку композиции на обучающей выборке:

Сначала имеет смысл решить более простую задачу: определить, какие значения должен принимать алгоритм = на объектах обучающей выборки, чтобы ошибка на обучающей выборке была минимальной:

Другими словами, необходимо найти такой вектор сдвигов s, который будет минимизировать функцию . Поскольку направление наискорейшего убывания функции задается направлением антиградиента, его можно принять в качестве вектора s:

Компоненты вектора сдвигов s, фактически, являются теми значениями, которые на объектах обучающей выборки должен принимать новый алгоритм , чтобы минимизировать ошибку строящейся композиции.

Обучение , таким образом, представляет собой задачу обучения на размеченных данных, в которой – обучающая выборка, и используется, например, квадратичная функция потерь:

Следует обратить особое внимание на то, что информация об исходной функции потерь , которая не обязательно является квадратичной, содержится в выражении для вектора оптимального сдвига s. Поэтому для большинства задач при обучении можно использовать квадратичную функцию потерь.

В рамках данной работы будут использоваться следующие реализации, описанного выше алгоритма:

XGBoost – одна из самых популярных и эффективных реализаций алгоритма градиентного бустинга на деревьях на 2019-й год.[[15]](#footnote-15)

CatBoost – библиотека для метода машинного обучения, основанная на градиентном бустинге.[[16]](#footnote-16)

LightGBM – открытая программная библиотека, разработанная компанией Яндекс.[[17]](#footnote-17)

Калибровка прогнозных вероятностей, предсказанных рассмотренными реализациями алгоритма, осуществляется аналогично алгоритму случайного леса. После обучения всех рассмотренных алгоритмов (линейных и нелинейных) получается матрица ответов алгоритмов на объектах выборки. Далее необходимо объединить эти предсказания в итоговый результат.

Стекинг – наиболее популярный метод ансамблирования алгоритмов. Идея стекинга заключается в обучении нескольких разных алгоритмов и передаче их результатов на вход последнему, который принимает итоговое решение.

Для формализации решения задачи стекинга введем следующие обозначения. Пусть – обучающая выборка, – базовый классификатор, использующийся для построения метапризнака. – функция обучения классификатора на . – функция, которая предсказывает метку класса для классификатором . – некоторый метаклассификатор. – метапризнак, полученный классификатором для выборки . – финальное предсказание стекинга для валидационной выборки.

В простейшем виде получение предсказания для тестовой выборки P с помощью стекинга выглядит следующим образом:[[18]](#footnote-18)

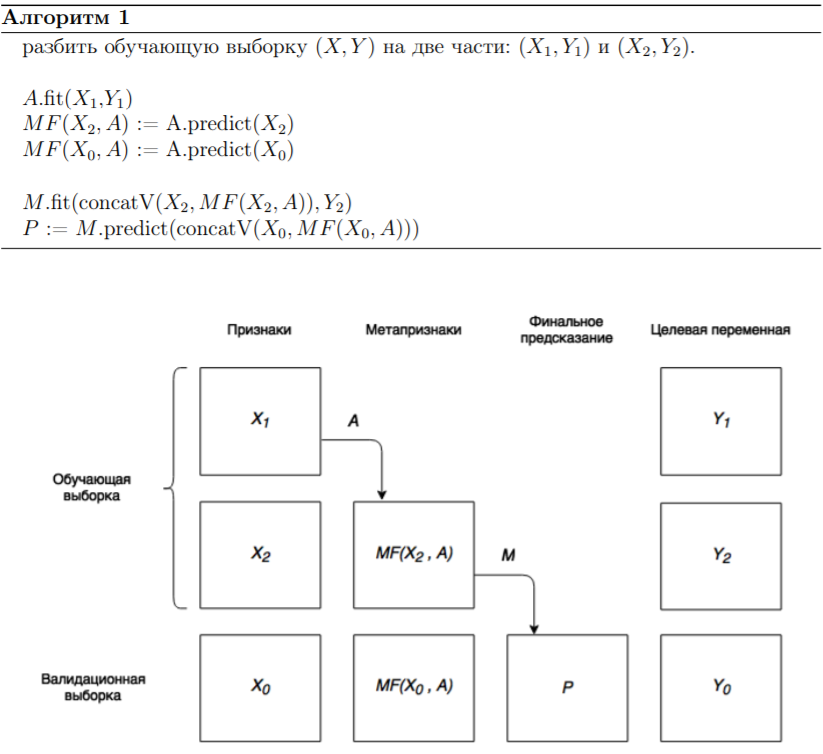


Рис. Алгоритм 1, стекинг по схеме hold-out[[19]](#footnote-19)

Недостатком этого варианта стекинга является то, что обучается только на части обучающей выборки, а другая часть им не используется. Чтобы избежать этого, мы можем повторить алгоритм, поменяв местами (, ) и (, ). В таком случае получим два предсказания для валидационной выборки, которые можно усреднить. В качестве метаклассификатора будет использоваться линейный алгоритм. Именно на предсказания метаклассификатора будет строится рейтинговая шкала.

Далее рассмотрим метрики качества алгоритмов, которые используются в задачах классификации, а также способы верификации полученных результатов. Уточним методологию составления рейтинговой шкалы.

## 1.5 Метрики качества решений, кросс-валидация и рейтинговая шкала

В задаче бинарной классификации, в которой метки принадлежат множеству {-1, 1}, объекты с меткой 1 будем называть положительными, а с меткой -1 – отрицательными. Базовый алгоритм возвращает произвольное вещественное число, далее для удобства называемое вероятностью принадлежности метки классу, которое с помощью порога вероятности переводится в бинарный ответ:

Наиболее очевидной мерой качества в задаче классификации является доля правильных ответов (accuracy):

Данная метрика, однако, имеет существенный недостаток. Если взять порог меньше минимального значения прогноза на выборке или больше максимального значения, то доля правильных ответов будет равна доле положительных и отрицательных ответов соответственно. Таким образом, если в выборке 950 отрицательных и 50 положительных объектов, то при тривиальном пороге мы получим долю правильных ответов 0.95. Это означает, что доля положительных ответов сама по себе не несет никакой информации о качестве работы алгоритма , и вместе с ней следует анализировать соотношение классов в выборке.

Следовательно, в случае с несбалансированными классами одной доли правильных ответов недостаточно – необходима еще одна метрика качества. Для начала введем понятие матрицы ошибок. Матрица ошибок в задаче двоичной классификации – это способ разбить объекты на четыре категории в зависимости от комбинации истинного ответа и ответа алгоритма (см. таблица)

Таблица. Матрица ошибок

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | True Positive (TP) | False Positive (FP) |
|  | False Negative (FN) | True Negative (TN) |

Гораздо более информативными критериями являются точность (precision) и полнота (recall):

Точность показывает, какая доля объектов, выделенных классификатором как положительные, действительно является положительными. Полнота показывает, какая часть положительных объектов была выделена классификатором.

Отметим, что точность и полнота не зависят от соотношения размеров классов. Даже если объектов положительного класса на порядки меньше, чем объектов отрицательного класса, данные показатели будут корректно отражать качество работы алгоритма.

Существует несколько способов получить один критерий качества на основе точности и полноты. Один из них – F-мера, гармоническое среднее точности и полноты:

Среднее гармоническое обладает важным свойством – оно близко к нулю, если хотя бы один из аргументов близок к нулю. Именно поэтому оно является более предпочтительным, чем среднее арифметическое.

Выше были рассмотрены такие показатели, как точность, полнота и F-мера, которые характеризуют качество работы алгоритма при конкретном значении порога . Однако зачастую интерес представляет лишь вещественный алгоритм , а порог выбирается позже в зависимости от требований к точности и полноте. В таком случае возникает потребность в измерении качества семейства моделей .

Широко используется такая интегральная метрика качества семейства, как площадь под ROC-кривой (Area Under ROC Curve, AUC-ROC). Рассмотрим двумерное пространство, одна из координат которого соответствует доле неверно принятых объектов (False Positive Rate, FPR), а другая – доле верно принятых объектов (True Positive Rate, TPR):

Каждый возможный выбор порога t соответствует точке в этом пространстве. Всего различных порогов имеется ℓ + 1. Максимальный порог даст классификатор с TPR = 0, FPR = 0. Минимальный порог даст TPR = 1 и FPR = 1. ROC-кривая – это кривая с концами в точках (0, 0) и (1, 1), которая последовательно соединяет точки, соответствующие порогам . Площадь под данной кривой называется AUC-ROC, и принимает значения от 0 до 1. Если порог может быть подобран так, что алгоритм не будет допускать ошибок, то AUC-ROC будет равен единице; если же ранжирует объекты случайным образом, то AUC-ROC будет близок к 0.5.

Критерий AUC-ROC имеет большое число интерпретаций – например, он равен вероятности того, что случайно выбранный положительный объект окажется позже случайно выбранного отрицательного объекта в ранжированном списке, порожденном .

Далее разберемся с проблемой переобучения. Допустим при решении задачи классификации был построен некоторый алгоритм, например линейный классификатор, причем доля ошибок на объектах из обучающей выборки была равна 0,2, и такая доля ошибок является допустимой. Но поскольку алгоритм не обладает обобщающей способностью, нет никаких гарантий, что такая же доля ошибок будет для новой выборки. Вполне может возникнуть ситуация, что для новой выборки ошибка станет равной 0,9. Это значит, что алгоритм не смог обобщить обучающую выборку, не смог извлечь из нее закономерности и применить их для классификации новых объектов. При этом алгоритм как-то смог подогнаться под обучающую выборку и показал хорошие результаты при обучении без извлечения истинной закономерности. В этом и состоит проблема переобучения.

Глубже понять проблему переобучения можно на данном примере. На следующем графике изображена истинная зависимость, объекты обучающей выборки и переобученный алгоритм:



Рис. Переобученный классификатор

Восстановленная зависимость дает идеальные ответы на всех объектах обучающей выборки, но при этом в любой другой точке сильно отличается от истинной зависимости. Такая ситуация называется переобучением. Алгоритм слишком сильно подогнался под обучающую выборку ценой того, что он будет давать плохие ответы на новых точках.

Выявить переобучение, используя только обучающую выборку, невозможно, поскольку и хорошо обученный, и переобученный алгоритмы будут хорошо ее описывать. Необходимо использовать дополнительные данные.

Существуют несколько подходов к выявлению переобучения:

* + Отложенная выборка. Часть данных из обучающей выборки не участвуют в обучении, чтобы позже проверять на ней обученный алгоритм.
  + Кросс-валидация, несколько усложненный метод отложенной выборки.

Самый простой способ оценить качество алгоритма – использование отложенной выборки. В этом случае следует разбить выборку на две части: первая из двух частей будет использоваться для обучения алгоритма, а вторая, тестовая выборка, – для оценки его качества, в том числе для нахождения доли ошибок в задаче классификации

Более системный подход – кросс валидация. В этом случае выборка делится на k блоков примерно одинакового размера. Далее по очереди каждый из этих блоков используется в качестве тестового, а все остальные – в качестве обучающей выборки. После того, как каждый блок побывает в качестве тестового, будут получены k показателей качества. В результате усреднения получается оценка качества по кросс-валидации. В качестве иллюстрации традиционный алгоритм кросс-валидации можно представить следующим образом:

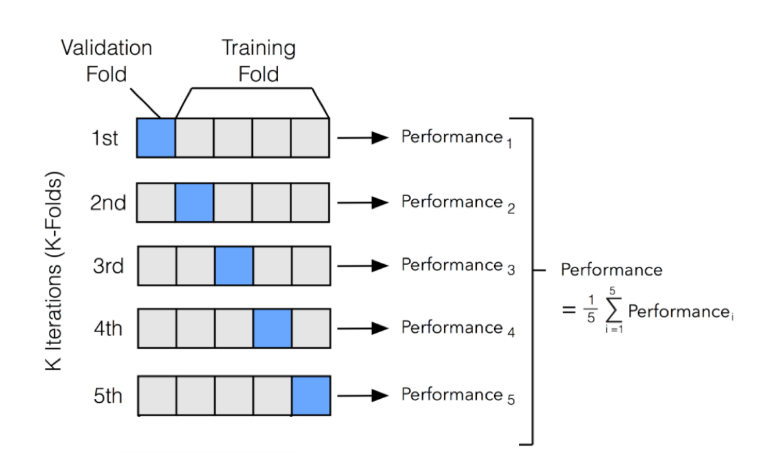
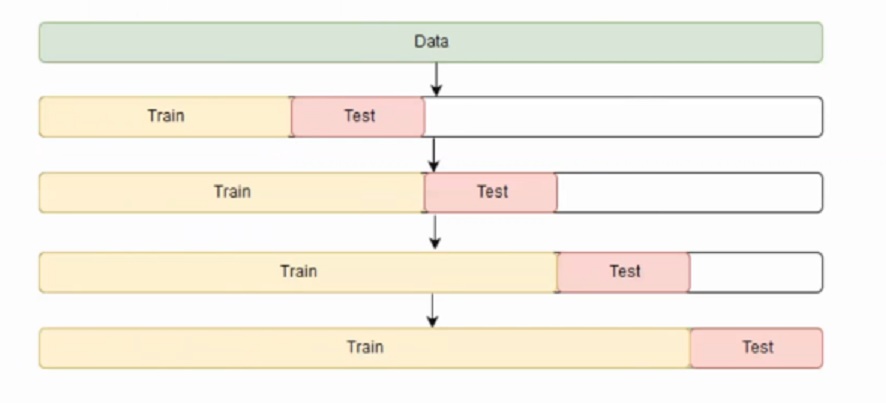


Рис. Иллюстрация принципа кросс-валидации

Однако следует понимать, что финансовые данные имеют временную структуру, с которой традиционная кросс-валидация на k блоков справляется недостаточно хорошо.

Чтобы точно смоделировать «среду прогнозирования реального мира, в котором мы находимся в настоящем, и прогнозировать будущее»[[20]](#footnote-20) (Tashman 2000), прогнозист должен скрывать все данные о событиях, которые происходят в хронологическом порядке после событий, используемых для подгонки модели. Таким образом, вместо использования перекрестной проверки в k-кратном порядке, для данных временного ряда используется перекрестная проверка с удержанием, когда подмножество данных (разделенных по времени) зарезервировано для проверки производительности модели. Например, см. Рисунок 1, где данные набора тестов поступают в хронологическом порядке после обучающего набора.



После обучения модели и верификации решения получаем вектор вероятностей, предсказанных моделью. Этот вектор необходимо превратить в вариационный ряд, где элементы отсортированы по неубыванию. Для полученного распределения случайных величин строим квантили уровня , где – количество равномерных градаций в интервале [0,1], а – номер интервала. Эти квантили делят интервал на R частей, которые определяются следующим образом:

Для каждого объекта рейтинговая оценка формируется в зависимости от попадания выборочного значения предсказанной вероятности для рассматриваемого объекта в один из приведенных выше интервалов.

В рамках данной работы разделим вариационный ряд на 10 интервалов, каждый из которых будет соответствовать значению рейтинга. Так объекту, попавшему в 10 интервал, будет присваиваться значение рейтинга 10, которое указывает на максимальную привлекательность рассматриваемого объекта для инвестирования.

Нужно понимать, что сложность используемого алгоритма не сможет повысить качество предсказаний, если выбраны неверные факторы, описывающие объекты выборки. Поэтому вторая глава данной работы полностью посвящена конструированию объясняющих переменных.

# Глава 2. Конструирование, обработка и отбор факторов для построения моделей

## 2.1 Описание выборки компаний, используемых для анализа

Вполне очевидно, что для конструирования признаков, обладающих высокой информативностью, необходимо полнее понять суть объектов, которые они должны описывать. Объектами данного исследования является компании, входившие в индекс S&P500 за период с 2000 года по 2018 год. S&P500 – фондовый индекс, формирующийся из 500 публичных компаний, торгующихся на фондовых биржах США, таких как NASDAQ и Фондовая биржа Нью Йорка. Капитализация и ликвидность акций являются основными критериями отбора компаний в индекс. А также авторы индекса стремятся сохранить репрезентативность выборки для каждой отдельной отрасли экономики США.

Ключевое отличие индекса S&P500 от индекса Доу Джонса заключается в том, что его значение рассчитывается, исходя из капитализации компаний с поправкой на free-float, которые в него входят, в то время как значение индекса Доу Джонса зависит от стоимости акций компаний, формирующих его. Другими словами, индекс S&P500 отражает не динамику движения цен акций, а изменения в структуре фондового рынка США.

Компании, формирующие индекс S&P500, разделены на сектора согласно глобальному стандарту отраслевой классификации (далее по тексту GICS[[21]](#footnote-21)). Стандарт CIGS был разработан в 1999 году компаниями MSCI Inc. и Standard & Poor's Financial Services LLC. Стандарт кластеризует экономику США на 11 ключевых секторов:

1. Энергетический сектор (Energy). Энергетический сектор представлен отраслями добывающей промышленности, которые специализируются на добыче нефти, газа и иных видов топлива, а также отраслями, обслуживающими их.
2. Сектор сырья и материалов (Materials). Сектор представляют компании, занимающиеся разведкой, разработкой и переработкой сырья. Например, компании металлургической, химической и лесной промышленностей.
3. Промышленный сектор (Industrials). В промышленный сектор входя компании занятые в производстве готового продукта. Например, компании строительной и обрабатывающей промышленностей.
4. Потребительский сектор (Consumer Discretionary). Компании данного сектора занимаются реализацией потребительских товар, которые потребители могут избегать без каких-либо серьезных последствий для их благополучия. Например, компании, представляющие ресторанный бизнес, являются частью потребительского сектора.
5. Сектор потребительских товаров и услуг (Consumer Staples). В отличие от товаров, реализуемых компаниями потребительского сектора, многие товары компаний сектора потребительских товаров и услуг являются жизненно необходимыми и приобретаются потребителями вне зависимости от их социального положения и достатка. Например, компании пищевой промышленности традиционно относят к сектору потребительских товаров и услуг.
6. Сектор здравоохранения (Health Care). Данный сектор представлен компаниями, предоставляющими медицинские услуги и все, что с ними связано.
7. Финансовый сектор (Financials). Компании данного сектора предоставляют широкий спектр финансовых услуг компаниям иных отраслей.
8. Сектор информационных технологий (Information Technology). Сектор информационных технологий состоит из компаний, которые предоставляют программное обеспечение, аппаратное или полупроводниковое оборудование.
9. Сектор услуг связи (Communication Services). Компании данного сектора предоставляют весь спектр услуг, связанных с интернетом и навигацией.
10. Utilities (Сектор коммунальных услуг). Компании данного сектора предоставляют доступ к основным удобствам, таким как вода, канализация, электричество и другим.
11. Сектор недвижимости (Real Estate). До 2011 года данный сектор был частью финансового сектора. Основными сегментами сектора являются жилая недвижимость, коммерческая недвижимость и промышленная недвижимость. Эти три сегмента представлены публично торгуемыми инвестиционными фондами недвижимости (REITs). В рамках данной работы сектор недвижимости не будет отделяться от финансового сектора для обеспечения однородности данных на разных временных горизонтах.

Анализируя долю значения индекса, которая приходится на каждый сектор экономики США, можно определить ключевые для определенного периода сектора. Ниже представлена гистограмма, отражающая динамику изменения доли секторов в индексе S&P500 за 18 лет. По каждому году рассчитано усредненное значение индекса.

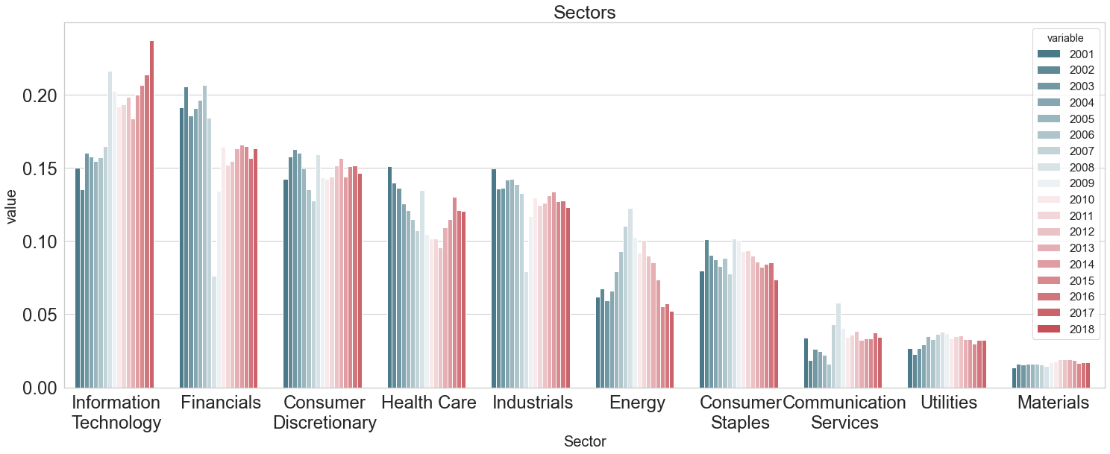
****

Рис. 1 Гистограмма распределения долей секторов в индексе S&P500

Гистограмма показывает, что наименьший вклад в индекс вносят компании таких секторов, как Communication services, Materials, Utilities. Их совокупное присутствие в индексе оставалось фактически неизменным и не превышало более 15% от значения индекса за последние 18 лет. На основе этих данных можно сделать вывод, что данные отрасли обладают наименьшим влиянием на экономику США. Сектора Consumer Staples и Industrials за последние 18 лет не продемонстрировали существенного изменения своей доли в индексе. Их доли составили 10 и 13 процентов соответственно. Значение доли финансового сектора в индексе S&P500, начиная с 2008 года, планомерно снижается. Вероятно, подобные изменения связаны не только с финансовым кризисом 2008 года, но и с развитием финтех подразделений IT компаний, которые начинают выполнять функции ранее считавшиеся классически финансовыми. Тем не менее, доля финансового сектора по-прежнему высока и на 2018 год составляет 17% от индекса. В отличие от доли финансового сектора, доля сектора информационных технологий в индексе после краха доткомов в начале 2000-х годов планомерно наращивается. На конец 2018 года она составляет более 20%. Вполне очевидно, что на данный момент сектор информационных технологий является центральным для экономики США. За рассматриваемый период изменения долей секторов Energy и Health Care отрицательно скоррелированы. За период с 2001 по 2009 гг. доля сектора Energy в индексе выросла на 6 процентных пунктов, в то время как доля сектора Health Care сократилась на 5 процентных пунктов. За период с 2010 по 2018 гг. произошло обратное перераспределение долей между данными секторами.

Также для целей работы полезно понимать, какова средняя капитализация компаний в разрезе секторов. Гистограмма, представленная ниже отражает динамику изменения средней стоимости компаний в рамках основных секторов за период с 2001 по 2018 года.

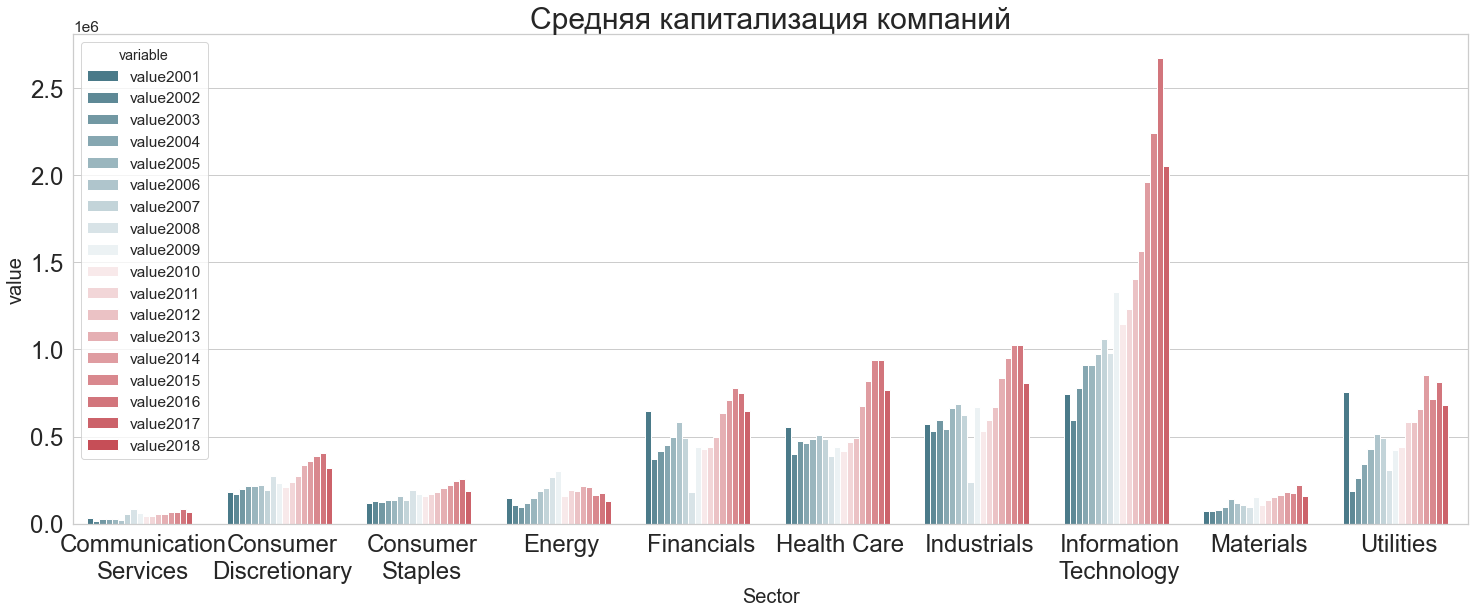


Рис 2. Гистограмма распределения средней капитализации компаний по секторам за рассматриваемый период.

Согласно данным, представленным рисунке, средняя стоимость компании, принадлежащей сектору информационных технологий, экспоненциально растет год от года и превышает среднюю стоимость иных секторов более чем в 2 раза. На данном этапе можно предположить, что компании этого сектора должны находиться достаточно высоко в рейтинге инвестиционной привлекательности. Относительно высокой капитализацией обладают компании таких секторов, как Health Care, Financials, Industrials, Utilities. Интересно, что несмотря на то, что сектор Utilities занимает наименьшую долю по капитализации в индексе, средняя компания этого сектора имеет достаточно высокую капитализацию. Вероятно, это указывает на серьезный уровень монополизации отрасли.

Получив представление об объектах выборки, участвующих в анализе, перейдем непосредственно к конструированию признаков, их описывающих, а также инициализации целевой переменной модели. В целях структуризации конструирование признаков проводится в рамках трех контуров:

1. Контура внутренних факторов компаний;
2. Контура макрофакторов;
3. Контура технических показателей.

Подобное разделение позволяет охватить весь спектр компонент, влияющих на формирование стоимости компании.

В качестве инструментов анализа в работе используются такие инструменты, как GitHub[[22]](#footnote-22), JupyterLab, Jupyter Notebook и Python. Jupyter Notebook - инструмент для создания аналитических отчетов, позволяющий хранить код, комментарии, изображения, формулы и графики, а релиз JupyterLab поддерживает отображение и редактирование таких форматов данных, как CSV, JSON, PDF, Vega и тд. Используемый в работе язык программирования Python позволяет писать алгоритмы обработки данных с использованием различных фреймворков, системных утилит и приложений для автоматизации действий.

## 2.2 Конструирование целевой переменной модели

Цель данного исследования заключается в создании модели, которая способна, используя некий набор входных данных, на выходе определять вероятности того, что доходность конкретной компании окажется выше среднеотраслевого уровня. Другими словами, модель должна на определенном уровнем значимости гарантировать инвестору получение аномальной доходности, т.е. доходности выше простого инвестирования в индекс. Очевидно, что подобная задача идет в разрез с гипотезой эффективного рынка[[23]](#footnote-23), согласно которой цены акций подчиняются закону случайного блуждания и, как следствие, не могут быть предсказаны. Однако в ряде исследований (Bartov, Givoly, & Hayn, 2002; Kasznik & McNichols, 2002 Сюрпрайз) были получены статистически значимые результаты, согласно которым доходность компаний напрямую зависит от публикуемой ими отчетности. Из опросов топ менеджеров компаний так же ясно, что руководители считают, что прибыль их компании должна соответствовать или превышать ожидаемую рыночную прибыль для того, чтобы цены акций их компаний увеличивались или находились на том же уровне.[[24]](#footnote-24) Интересно, что экономические последствия публикации неожиданной для инвесторов информации в отчетности компаний не обязательно сразу заметны, поскольку рынку может потребоваться некоторое время, чтобы отразить их предполагаемое экономическое влияние[[25]](#footnote-25). На основе данного тезиса строится методология большого количества исследований, которые пытаются выявить, какие из факторов наиболее сильно влияют на будущую доходность компаний, закладывая в свое исследование идею о том, что цена акции подстраивается под произошедшие изменения не моментально, а с некоторым лагом.[[26]](#footnote-26) В общем и целом, методология данного исследования также эксплуатирует описанный феномен.

В качестве доходности акций используется показатель общей доходности акции (TSR), который учитывает прирост капитала и дивиденды при измерении общего дохода, приносимого акцией инвестору, и является одним из наиболее популярных показателей для оценки привлекательности вложений в компанию с точки зрения миноритарного акционера.[[27]](#footnote-27) Показатель рассчитывается на основе рыночной информации по следующей формуле:

*Где*

*– дивиденды текущего периода*

*– рыночная капитализация предыдущего периода*

*– рыночная капитализация текущего периода*

API «Alpha Vantage»[[28]](#footnote-28) позволило получить доступ к свободно распространяемой рыночной информации по анализируемым компаниям. Были скачаны ежемесячные значения капитализации компаний, входящих в индекс S&P500 за последние 20 лет.

Для определения эффективного периода реакции рынка на публикуемую информацию о компаниях и макроэкономической ситуации рассчитываются доходности за три периода: за квартал, за 6 месяцев и за год. На рисунке ниже представлены гистограммы распределения рассчитанных доходностей, а также изображены распределения, аппроксимирующие построенные гистограммы (распределения Стьюдента), базирующееся на математическом ожидании выборки доходностей, их стандартном отклонении и числе степеней свободы. Также построены нормальные распределения на базе математических ожиданий и среднеквадратичных отклонений рассматриваемых данных. Так же для каждого периода построены графики QQ-plot.

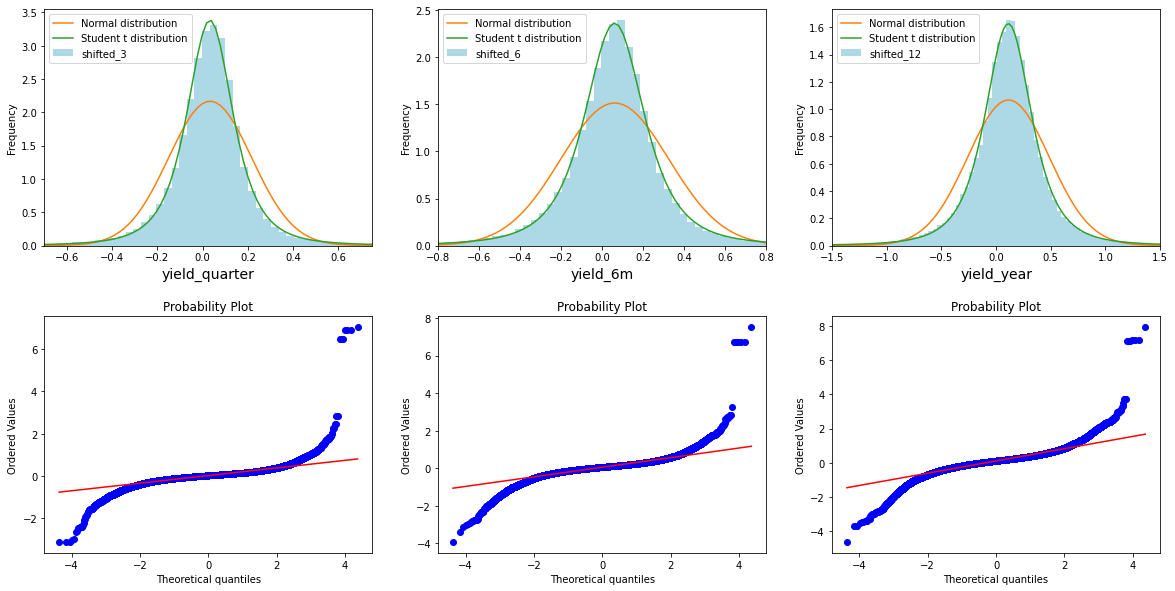


Рис. 3 Распределение квартальных, полугодовых и годовых доходностей

Из рисунка видно, что распределение доходностей компаний индекса S&P500 далеки от нормального, что подтверждают графики Q-Q plot. При детальном рассмотрении становится очевидно, что распределения доходностей, в сравнении с нормальным распределением, имеют более тяжелые хвосты, и их можно аппроксимировать распределением Стьюдента. При построении регрессии подобная ситуация может привести к снижению качества оценок модели, особенно чувствительны к выбросам линейные алгоритмы классификации. Следовательно, необходимо удалить наблюдения с выбросами.

При работе с выбросами важно понимать, что оценка стандартного отклонения для распределения, содержащего выбросы, рассчитанная с помощью среднеквадратичного отклонения будет смещена. Поэтому будем использовать робастные к выбросам оценки, полученные с помощью расчета медианного абсолютного отклонения (MAD).

В среднеквадратичном отклонении расстояния от среднего значения возводятся в квадрат, поэтому большие отклонения имеют более высокий вес, и, следовательно, выбросы могут сильно влиять на него. В MAD небольшое число выбросов не оказывает существенного влияния на итоговый результат, поэтому оно может выступать в качестве оценки стандартного отклонения выборки.

Где k – постоянный коэффициент, который зависит от распределения, стандартное отклонение которого определяется. Для нормального распределения k = 1.4826.

где:

– обратная функция к функции квантиля стандартного нормального распределения (функция распределения (CDF)).

Коэффициент функции распределения равен 3/4, так как интервал ±MAD покрывает 50% (между квантилями 1/4 и 3/4) функции плотности нормального стандартного распределения, т.е.

Следовательно:

А так как:

Получается, что

Или

Откуда получается, что

Построим график доходностей случайной компании и визуализируем границы интервала 3 сигма, определенного с помощью медианного абсолютного отклонения. Значения доходностей, которые лежат за границами этого интервала будем считать выбросами.

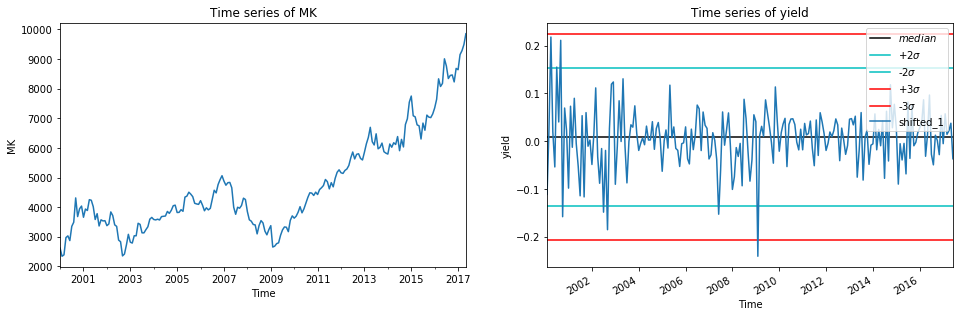


Рис 3. Визуализация выбросов доходностей

Из графика видно, что присутствуют наблюдения, которые лежат за пределами отведенного интервала. Именно эти наблюдения и формируют тяжелые хвосты на исходных распределениях доходностей. Их удаление должно способствовать нормализации распределений.

Удалим выбросы из данных по правилу, определенному выше, и еще раз построим гистограммы распределения исследуемых величин.

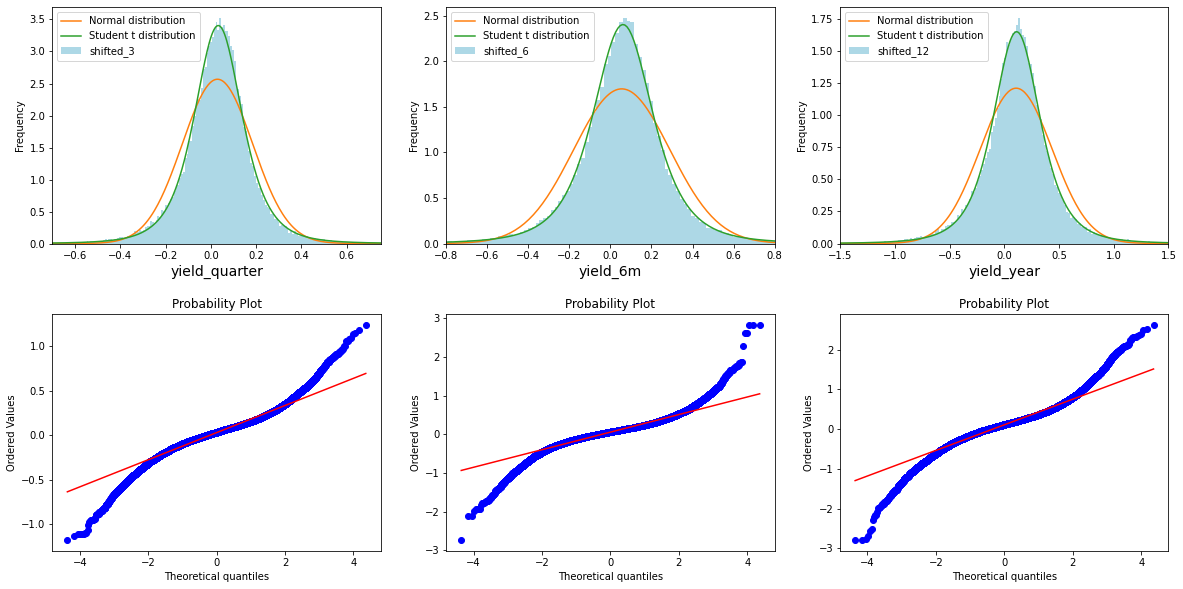


Рис. Гистограммы распределения квартальных, полугодовых и годовых обработанных доходностей

Из рисунка видно, что распределения доходностей все еще не нормальны, однако теперь предположение о нормальности выполняется значительно лучше.

Используя очищенные доходности компаний, рассчитаем среднеотраслевые доходности. На рисунке ниже представлена гистограмма, визуализирующая изменения доходности секторов по годам.

Рис. Гистограмма

На основе представленных показателей можно рассчитать вектор целевых переменных. Целевая переменная принимает значение равное 1 в том случае, если доходность компании в момент времени t превышает доходность сектора, к которому она относится, и 0 в обратной ситуации на выбранном временном интервале. На рисунке ниже представлена гистограмма, отражающая сбалансированность рассматриваемого показателя.

Рис. Гистограмма

## 2.3 Конструирование, обработка и отбор факторов внутреннего контура модели

База финансовых данных, формирующих контур внутренних факторов, была получена от фирмы, специализирующейся на управлении активами клиентов, Signet Financial Management. Аналитики Signet FM структурируют портфель клиентов на основе исторических исследований рынка и анализа текущей коньюктуры. Основным полем работы компании является рынок ценных бумаг США. Компания предлагает разные типы портфелей, которые представлены в таблице 1:

**Таблица. 1** «Виды предлагаемых портфелей Signet FM»

|  |  |
| --- | --- |
| **Вид портфеля** | **Описание** |
| Консервативный | Предназначен для инвесторов, которые ищут максимальный доход при условии минимальных рисков |
| Уравновешенный | Инвесторы стремятся к росту и доходам, принимающую нормальную волатильность |
| Умеренный рост | Для инвесторов, которые осознают риски и готовы терпеть разумную волатильность с целью получения максимальной выгоды |

Компаний не занимается спекулятивными краткосрочными сделками и сосредоточена на долгосрочном инвестировании. Следовательно, главная задача компании – грамотно анализировать изменчивый рынок и предугадывать среднесрочные и долгосрочные тренды движения рынков капитала. В ходе работы компания систематизировала существенный объем публично доступных финансовых данных по компаниям, входящим в индекс S&P500, а именно, данные из ежегодно публикуемых отчетов о финансовом положении компаний, данные из отчетов о совокупном доходе и отчетов о движении денежных средств. Полный перечень показателей перечислен в приложении 1.

При анализе данных становится очевидно, что ряд показателей имеет существенное количество пропусков в наблюдениях. Главная причина пропусков заключается в том, что некоторые компаний, которые находились в индексе S&P500 в левой границе исследуемого временного интервала, выбывают из него ближе к правой границе исследуемого временного интервала. Отдельно стоит выделить пропуски в тех показателях, которые рассчитывались для одной совокупности компаний, входящих в выборку, и не рассчитывались для другой в определенный момент времени. На рисунке 1 представлена гистограмма, отражающая долю пропусков в наблюдениях для таких показателей в общем количестве временных интервалов.

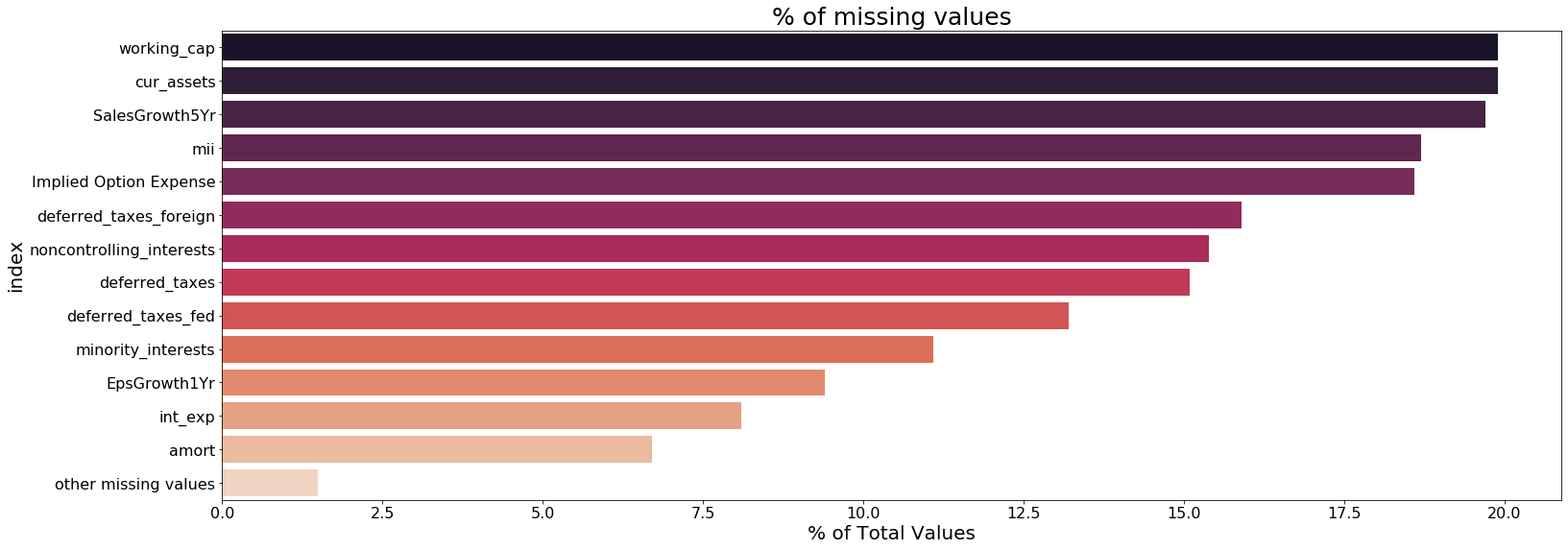


Рис 1. Доля пропущенных наблюдений

Для устранения пропусков в наблюдениях в рамках данной работы используется метод сплайн-интерполяции.[[29]](#footnote-29) В математической области численного анализа сплайн-интерполяция является формой интерполяции, где интерполант представляет собой особый тип кусочно-полиномиального типа, называемый сплайном. Сплайн-интерполяция часто предпочтительнее полиномиальной, потому что погрешность интерполяции может быть небольшой даже при использовании полиномов низкой степени для сплайна. Сплайн-интерполяция позволяет избежать проблемы феномена Рунге[[30]](#footnote-30). Феномен Рунге представляет собой проблему колебаний на краях интервала, возникающую при использовании полиномиальной интерполяции с многочленами высокой степени по множеству равноотстоящих точек интерполяции.

Решив проблему пропусков, сгенерируем признаки, на основе которых инвесторы потенциально могут делать выбор о покупке или продаже актива.

В ходе анализа литературы, посвященной предсказанию будущей доходности компаний фондового рынка США, были выявлены следующие наиболее популярные фундаментальные факторы, используемые исследователями в своих работах:

* Log dividend-price ratio (Lin, Q., 2018; Campbell et all, 2006 и другие)
* Log earnings-price ratio (Rapach et all, 2016; Lin, Q., 2018 и другие)
* Log dividend-payout ratio (Lamont,1998; Rapach et all, 2016 и другие)
* Book-to-market ratio (Pontiff and Schall, 1998; Lin, Q., 2018и другие)

А также, согласно современным представлениям о анализе финансового состояния компании, можно выделить три проекции финансовой успешности компании, а именно проекцию ликвидности, проекцию текущей экономической эффективности и проекцию сбалансированности[[31]](#footnote-31).

Проекция ликвидности отражает достаточность поступлений денежных средств и формирования денежных потоков для удовлетворения интересов заинтересованных групп. Главный вопрос, на который должен получить ответ аналитик: способна ли компания генерировать денежные средства, способна ли компания расплачиваться вовремя по операционным и финансовым обязательствам. В рамках проекции ликвидности рассчитываются такие показатели, как коэффициент текущей ликвидности, коэффициент покрытия процентов, доля краткосрочного долга в общем долге.

Коэффициент текущей ликвидности (Current Ratio) показывает достаточно ли у компании ресурсов для выполнения своих краткосрочных обязательств и рассчитывается по формуле, представленной ниже:

Коэффициент покрытия процентов (ICR) отражает способность компании выполнять свои процентные платежи и рассчитывается по формуле, представленной ниже:

Доля краткосрочного долга в общем долге рассчитывается по формуле, представленной ниже:

Проекция текущей эффективности позволяет сопоставить текущие затраты ресурсов, выраженные в денежной форме с получаемыми текущими выгодами. Главный вопрос: насколько эффективно используются ресурсы. В рамках проекции текущей эффективности рассчитываются такие показатели, как коэффициенты рентабельности, темпы роста основных доходных статей отчета о прибылях и убытках, экономическая добавленная стоимость (EVA) и коэффициент недооценки.

Норма операционной прибыли отражает долю операционной прибыли в общих доходах компании и рассчитывается по формуле:

Рентабельность задействованного капитала (ROCE) — это финансовый коэффициент, который измеряет прибыльность компании и эффективность использования ее капитала. Другими словами, коэффициент измеряет, насколько хорошо компания получает прибыль от своего капитала. Коэффициент ROCE считается важным показателем рентабельности и часто используется инвесторами при отборе подходящих кандидатов для инвестиций. Рентабельность задействованного капитала рассчитывается по формуле:

Экономическая добавленная стоимость (EVA) — один из финансовых показателей компании, основанный на остаточной стоимости, рассчитанной путем вычета стоимости привлечения капитала компании из ее операционной прибыли, скорректированной на налоги на кассовой основе. Эта мера была разработана консалтинговой фирмой Stern Value Management, изначально входившей в состав Stern Stewart & Co.[[32]](#footnote-32) Показатель рассчитывается по формуле:

*где:*

*Инвестированный капитал (Invested Capital) – это общая сумма денег, привлеченных компанией путем выпуска ценных бумаг и облигаций;*

*Рентабельность инвестированного капитала (ROIC) – коэффициент эффективности, который предназначен для измерения процентного дохода, получаемого инвесторами в компании от инвестированного ими капитала.*

*Средневзвешенная стоимость капитала (WACC) – средний уровень затрат компании на привлечение и дальнейшее обслуживание капитала из различных источников.*

Показатель WACC рассчитывается на основе рыночной информацию по следующей формуле:

*где:*

*– стоимость i-ого источника, формирующего капитал компании;*

*– доля i-ого источника, формирующего капитал компании.*

Коэффициент недооценки рассчитывается на основании двух величин: рыночной стоимости компании в момент публикации финансовой отчетности компании и внутренней стоимости этой компании, которая рассчитывается на основе модели дисконтирования денежных потоков. В виде формулы коэффициент можно представить следующим образом:

*где:*

*– коэффициент переоценки/недооценки;*

*– внутренняя стоимость компании;*

*– рыночная стоимость компании.*

Коэффициент недооценки в каком-то роде является мерой для определения времени, которое необходимо рынку для реакции на поступающую информацию.

Проекция сбалансированного роста диагностирует целесообразность роста бизнеса (выручки) и сбалансированность основных финансовых пропорций (роста активов, прибыли, денежных средств). Главный вопрос – целесообразен ли рост компании? В рамках проекции сбалансированного роста рассчитывается индекс устойчивого роста и строится матрица качества роста.[[33]](#footnote-33)

Индекс устойчивого роста – показатель, разработанный И. В. Ивашковской и Е. Л. Животовой, который отражает степень сбалансированности операционной и стратегической эффективности компании, которая является ключевым драйвером стабильного развития. Показатель рассчитывается по следующей формуле:

*где:*

*k – количество лет наблюдений*

*– средний темп роста продаж*

В целях управления параметрами качества роста компании Ивашковская также разработала аналитическую матрицу («матрица качества роста»), представленную ниже (рис. 1):



Рис. 1 Матрица качества роста

1. По оси абсцисс откладывается значение спреда доходности инвестиционного капитала, рассчитываемое как разница между ROCE и WACC. Значение спрэда учитывает альтернативные издержки, которые несет компания при привлечении капитала, и отвечает за операционную эффективность.
2. По оси ординат откладывается значение устойчивого темпа роста (SGR), отвечающее за стратегическую эффективность компании.
3. Находится среднее значение темпов роста выручки и среднее значение спрэда в рамках отрасли. Таким образом, появляется возможность разделить компании, входящие в выборку на 4 группы.

Матрица разделена на четыре квадранта: Q1 – сбалансированный рост, Q2 – сфокусированный рост, Q3 – агрессивный рост и Q4 – догоняющий рост.

В квадрант Q1 попадают компании с наиболее качественным ростом, так как оба эффекта – стратегический и финансовый – в данном случае сбалансированы. Следовательно, есть все основания предполагать, что темпы роста выручки в данном случае будут стабильно высокими.

Противоположностью компаний, которые входят в квадрант Q1, являются компании, входящие в квадрант Q4. Исходя из текущего положения этих фирм, нельзя предположить, что их темпы роста в ближайшее время станут выше среднеотраслевых значений.

Можно предположить, что фирмы, которым присущ агрессивный рост (Q3), более перспективны ввиду того, что в исследованиях, проведенных И. Ивашковской, компании с агрессивной политикой роста с большей вероятностью перемещались в квадрант Q1, нежели компании с сфокусированным ростом (Q2). Подобную закономерность можно объяснить ловушкой прибыли, в которую попадают компании, использующие модель сфокусированного роста.

Систематизированный перечень рассчитанных факторов внутреннего контура представлен в таблице 1.

Таблица 1. Описательная характеристика факторов внутреннего контура

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Название** | **Описание** | **Count** | **median** | **mean** | **std** |
| **Current\_ratio** | Коэффициент текущей ликвидности | 6828.00 | 4.27 | 39.65 | 61.55 |
| **ICR** | Коэффициент покрытия процентов | 6828.00 | 9.72 | 154.45 | 421.72 |
| **Std\_coef** | Доля краткосрочного долга в общем долге | 6432.00 | 0.10 | 0.17 | 0.21 |
| **EBIT/Sales** | Норма операционной прибыли | 6828.00 | 0.16 | 0.12 | 2.09 |
| **ROCE** | Рентабельность задействованного капитала | 6231.00 | 0.12 | 0.13 | 0.10 |
| **EVA** | Экономическая добавленная стоимость | 6710.00 | 242.69 | 734.08 | 2672.72 |
|  | Коэффициент недооценки |  |  |  |  |
| **Sales\_growth** | Темпы роста продаж | 6775.00 | 7.02 | 9.22 | 20.01 |
| **SGI** | Индекс устойчивого роста |  |  |  |  |
| **Cvadrant** | Квадрант матрицы Ивашковской | 6828.00 | 2.00 | 2.15 | 1.47 |
| **Net\_Assets** | Чистые активы |  |  |  |  |
| **BV/MV** | Мультипликатор отношения балансовой стоимости к рыночной капитализации | 5957.00 | 0.87 | 1.69 | 2.69 |
| **EBITDA/EV** | Мультипликатор отношения EBITDA к EV |  |  |  |  |
| **Debt/EBITDA** | Мультипликатор отношения долга к EBITDA |  |  |  |  |
| **Div/MV** | Мультипликатор отношения дивидендных выплат к рыночной стоимости компании |  |  |  |  |
| **Div Payout Ratio (D/E)** | Коэффициент выплаты дивидендов |  |  |  |  |

Выбранные фактора могут иметь высокий уровень корреляции, однако прежде удалим выбросы, присутствующие в данных, аналогично объясняемой переменной, так как они могут повлиять на оценки корреляционного анализа. Модели линейной регрессии чувствительны к линейной зависимости факторов, так как она приводит к численной неустойчивости при обращении матрицы объект-признак, что, в свою очередь, приводит к росту дисперсии оценок регрессии.

Для проверки наличия линейной зависимости между факторами построим корреляционную матрицу, используя формулу корреляции Пирсона.



Рис. 1 Корреляционная матрица факторов внутреннего контура

Из рисунка видно, что выбранные факторы не содержат сильных линейных связей между собой. Поэтому оставим все без изменения. К имеющимся абсолютным значениям факторов добавим относительные величины, отражающие изменения данных величин в течение времени.

После рассмотрения факторов внутреннего контура перейдем к анализу факторов макроконтура.

## 2.4 Конструирование, обработка и отбор факторов макроконтура модели

С помощью базы данных «Federal Reserve Economic Data» был получен доступ к макроэкономическим показателям экономики США. Эти данные будут составлять контур макрофакторов, влияющих на доходность компаний. Ниже представлена таблица с описательными статистиками полученных данных.

Таблица.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Лейбл** | **Описание** | **Медиана** | **Мат. ожидание** | **STD** |
| **GDP** | Реальный ВВП в ценах базового года. | 15,710.62 | 15,837.48 | 1,506.12 |
| **CPI** | Индекс потребительских цен | 217.38 | 215.52 | 22.93 |
| **M2** | Агрегатор денежной массы М2 | 8,481.25 | 9,035.53 | 2,799.10 |
| **DebtF** | Государственный долг США | 12,693,493.60 | 12,923,075.25 | 5,227,591.64 |
| **PromIndex** | Индекс промышленного производства (IPI) | 100.33 | 99.37 | 5.37 |
| **Saves** | Личные сбережения населения в процентах от располагаемого личного дохода | 6.40 % | 6.04 % | 1.59 |
| **HouseInc** | Медианный заработок домохозяйств США | 59,331.41 $ | 59,358.50 $ | 2,125.47 |
| **dif\_DSG330** | Разница между доходностью 30-летних и 3-х месячных казначейский облигаций США | 2.87 % | 2.70 % | 1.32 |
| **UNRATE** | Количество безработных людей в процентах от общей рабочей силы | 5.60 % | 6.11 % |  |

Дадим краткую характеристику некоторым используемым показателям. Реальный ВВП представляет собой меру, которая отражает стоимость всех товаров и услуг, произведенных экономикой в данном году с поправкой на инфляцию, выраженную в ценах базового года.

Индекс потребительских цен измеряет изменения в уровне цен средневзвешенной рыночной корзины потребительских товаров и услуг, купленных домашними хозяйствами. В исследованиях Rapach et all, 2016, Campbell and Vuolteenaho, 2004 были получены статистически значимые результаты, подтверждающие предиктивную силу данного показателя

М1 агрегатор состоит из (1) валюты за пределами Казначейства США, Федеральных резервных банков и хранилищ депозитарных учреждений; (2) дорожных чеков небанковских эмитентов; (3) депозитов до востребования; и (4) других чековые депозитов (OCDs), которые состоят в основном из проектных счетов кредитных союзов. M2 агрегатор состоит из M1 плюс: (1) сберегательных депозитов (которые включают депозитные счета денежного рынка или MMDA); (2) срочных депозитов малого номинала (срочные депозиты на сумму менее 100 000 долл. США); и (3) остатков в розничных паевых инвестиционных фондах (MMMF).

Разница между доходностью 30-летних и 3-х месячных казначейский облигаций США отражает воспринятие рисков субъектами экономических отношений. В исследованиях Fama and French, 1988; Lin, Q., 2018 были получены статистически значимые результаты, подтверждающие предиктивную силу данного показателя.

Индекс промышленного производства — это экономический показатель, который измеряет реальный объем производства в обрабатывающей промышленности, добыче полезных ископаемых и коммунальных услугах.

Для проверки наличия линейной зависимости между факторами построим корреляционную матрицу, используя формулу корреляции Пирсона.

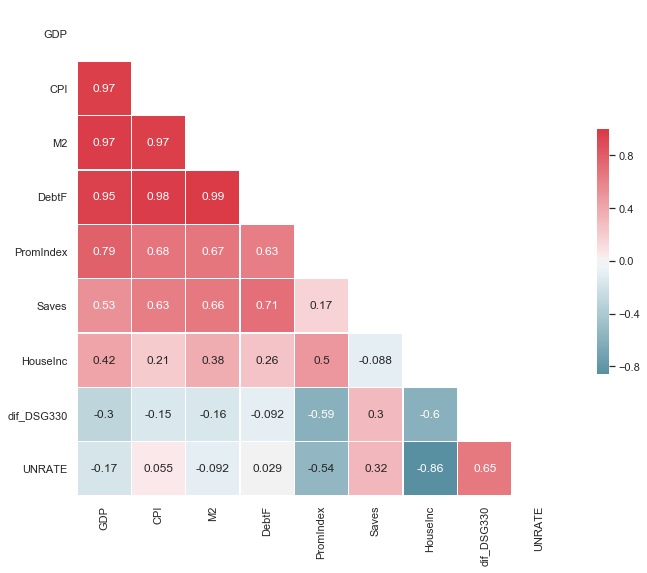


Рис. 1 Корреляционная матрица макроэкономических факторов

Из рисунка один видно, что наблюдается сильная положительная линейная связь между такими факторами, как индекс потребительских цен (CPI), агрегатор денежной массы М2 (M2), государственный долг США (DebtF), промышленный индекс (PromIndex). А также наблюдается сильная отрицательная линейная связь между такими факторами, как уровень безработицы (UNRATE) и Медианный заработок домохозяйств США (HouseInc).

Если для древовидных моделей наличие коллинеарных признаков не является помехой, то для линейных моделей подобная ситуация критична и приводит к росту дисперсии оценок коэффициентов модели. Решать проблему мультиколлинеарности факторов можно несколькими способами. Самый очевидный заключается в удалении высоко коррелированных факторов. Этот способ отличается простотой, а также он позволяет пользоваться всеми свойствами, которые вытекают из теоремы Гаусса-Маркова, при условии выполнения иных предпосылок теоремы. Другой способ заключается в добавлении в модель L1 / L2 регуляризации. Использование регуляризации лишает нас возможности пользоваться свойствами теоремы Гаусса-Маркова и проверять гипотезы о значимости коэффициентов. Третий способ заключается в использование метода главных компонент. Применение данного способа не приводит к нарушению предпосылок теоремы Гаусса-Маркова, однако приводит к утрате возможности интерпретации коэффициентов перед факторами. Ввиду того, что корреляция между факторами более 90%, в данном случае разумнее удалить столь сильно скоррелированные признаки из анализа.

После удаления линейно зависимых факторов получаем следующую корреляционную матрицу.

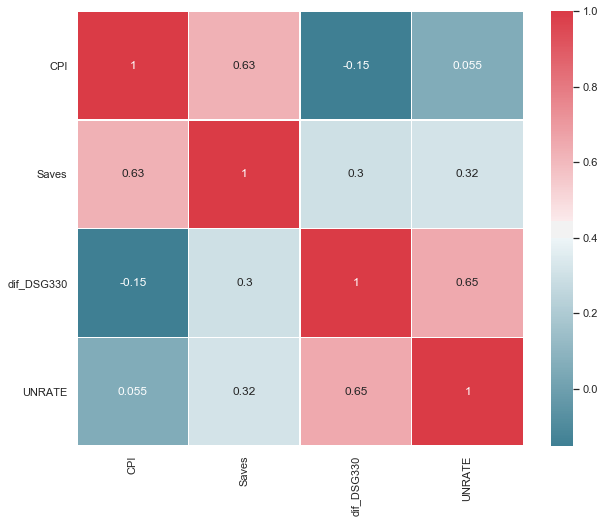


Рис. 2 Модифицированная корреляционная матрица макро факторов

Из изначального набора макроэкономических факторов остались следующие переменные: индекс потребительских цен (CPI), уровень сбережений населения (Saves), Разница между доходностью 30-летних и 3-х месячных казначейский облигаций США (dif\_DSG330) и уровень безработицы (UNRATE). Корреляционная матрица демонстрирует линейную зависимость умеренной силы между такими факторами, как уровень сбережений населения и индекс потребительских цен, а также dif\_DSG330 и уровень безработицы, но так как коэффициенты корреляции менее 0.7, то было принято решение оставить данные факторы для дальнейшего включения в регрессионную модель с использованием регуляризации.

Аналогично объясняемой переменной, проверим выбранные макроэкономические факторы на выбросы, так как они могут привести к снижению качества модели линейной регрессии (рисунок 1).



Рис.

Из рисунка видно, что анализируемые показатели содержат выбросы. После удаления выбросов показатели макроконтура готовы к использованию в модели.

## 2.5 Конструирование, обработка и отбор факторов рыночного контура модели

В техническом анализе технический индикатор - это величина, рассчитанная на основе исторических значений цен активов и объемов торгов по ним, которая предназначена для прогнозирования направления движения финансового рынка.[[34]](#footnote-34) Основная часть академический исследований в значительной степени опирается на фундаментальные переменные для прогнозирования доходности акций компаний фондового рынка США, при этом относительно мало внимания уделяется техническим индикаторам, широко используемым практиками.

В работах Lin, 2018, Nelly et all., 2014 в ходе исследования предиктивной силы технических индикаторов были получены статистически значимые результаты, согласно которым индикаторы, основанные на правиле импульса и скользящих средних, превышают по своей предсказательной способности большинство традиционно используемых для анализа фундаментальных показателей.

В данной работе используется 14 технических индикаторов, которые основаны на трех популярных торговых стратегиях, преследующих тренд. Первая стратегия основана на правиле импульса (MOM), которое генерирует сигнал покупки или продажи в конце месяца t, сравнивая текущую цену акций с ее уровнем m месяцев назад:

*Где:*

*– текущий уровень цен акции i*

*– уровень цен акции i m месяцев назад*

*– сигнал на покупку (1) или продажу (0) актива i*

Технически этот индикатор показывает наличие или отсутствие тренда, а также направление этого тренда. Положительные значения индикатора говорят о восходящем тренде, отрицательные значения – о нисходящем. Смена знака индикатора говорит о смене тренда. Индикатор MOM с импульсом m месяцев определяется как MOM (m). В рамках данной работы m принимает значение горизонтов прогнозирования, а именно 6, 9 и12 месяцев.

Вторая стратегия основана на схождении и расхождении двух скользящих средних с различными периодами (MACD). Данный технический индикатор был разработан основателем инвестиционной компании Signalert Corporation Джеральдом Аппелем и предназначен для анализа силы и направления тренда, а кроме этого, для выявления разворотных точек на ценовом графике. Сигнал формируется исходя из следующего неравенства:

*Где:*

*– сигнал на покупку (1) или продажу (0) актива i*

*– значение скользящей средней длины s <l в момент времени t*

*– значение скользящей средней длины l> s в момент времени t*

При этом значениеопределяется следующим образом:

*Где:*

*m – параметр определяющий длину скользящей средней*

*– значение стоимости актива i в момент времени t*

Соответствующий индикатор с длинами MA s и l определяется как . В данной работе вычисляются месячные сигналы для s = 1, 2, 3 и l = 9, 12.

Последняя стратегия основана на правиле торгового объема (VOL)[[35]](#footnote-35), поскольку изменение объема является еще одной полезной мерой, которая часто используется для определения рыночных тенденций. Торговый сигнал с использованием торгового объема определяется следующим образом:

*Где:*

*– сигнал на покупку (1) или продажу (0) актива i*

*– значение скользящей средней длины s <l в момент времени t, рассчитанной по показателю «on-balance» volume (OBV)[[36]](#footnote-36)*

*– значение скользящей средней длины l> s в момент времени t, рассчитанной по показателю «on-balance» volume (OBV)*

При этом значениеопределяется следующим образом:

*где:*

*m – параметр определяющий длину скользящей средней*

*– показатель «on-balance» volume (OBV) актива i в момент времени t*

Показатель «on-balance» volume (OBV) — трендовый индикатор, рассчитываемый по данным объема торгов по следующей формуле:

*где:*

*– объем торгов по i активу в момент времени j*

*– бинарная переменная, принимающая значение 1, если цена актива в период j-1 оказалось меньше текущей, и -1 в противном случае*

Интуитивно понятно, что рост недавней цены в сочетании с относительно высоким недавним объемом торгов обычно указывает на сильную положительную рыночную тенденцию и, таким образом, генерирует сигнал на покупку. Напротив, снижение недавней цены вместе с относительно высоким недавним объемом торгов обычно сигнализирует о сильной негативной тенденции рынка.

Соответствующий индикатор с длинами MA s и l определяется как . В данной работе вычисляются месячные сигналы для s = 1, 2, 3 и l = 9, 12.

Кроме технических индикаторов в рамках рассматриваемого контура рассчитываются такие рыночные индикаторы, как short interests’ rate (SIR) (Y. Zhang et all, 2019; Lin, Q., 2018 и другие) и stock return volatility (SRV) (Mele, 2007 и Goyal and Welch (2008) и другие).

Short interests’ rate – доля акций компании, которые были проданы без покрытия за определенный промежуток времени, от общего объема сделок за этот промежуток времени. Если SIR действительно содержит информацию о будущих рыночных доходностях, мы ожидаем, что более высокие значения SIR предскажут более низкие будущие доходности.

Stock return volatility (SRV) – показатель волатильности, используемый Nelly et all., 2014.

*где:*

*– годовая волатильность актива в момент времени t,*

*– месячная доходность актива в момент времени t,*

*– робастная волатильность актива в момент времени t.*

Goyal and Welch, 2008 использовали схожую меру волатильности. Они измеряли месячную волатильность как сумму квадратов ежедневных доходностей акций в течение месяца, однако эта мера приводила к серьезным выбросам в ряде временных отрезков. Показатель SRV, используемый в работе Nelly et all., 2014 позволяет сгладить выбросы и дает более правдоподобные результаты оценки.

В таблице ниже представлена описательная статистика рассчитанных технических индикаторов до бинаризации.

Таблица

|  |  |  |  |
| --- | --- | --- | --- |
| **Индикатор** | **Mean** | **Median** | **Std** |
| MA (1,9) |  |  |  |
| MA (1,12) |  |  |  |
| MA (2,9) |  |  |  |
| MA (2,12) |  |  |  |
| MA (3,9) |  |  |  |
| MA (3,12) |  |  |  |
| MOM (6) |  |  |  |
| MOM (9) |  |  |  |
| MOM (12) |  |  |  |
| VOL (1,9) |  |  |  |
| VOL (1,12) |  |  |  |
| VOL (2,9) |  |  |  |
| VOL (2,12) |  |  |  |
| VOL (3,9) |  |  |  |
| VOL (3,12) |  |  |  |
| SIR |  |  |  |
| SRV |  |  |  |

# Глава 3. Эмпирические результаты исследования

## 3.1 Настройка гиперпараметров линейных моделей

Под гиперпараметрами алгоритма понимаются такие параметры, поиск которых не может быть осуществлен в рамках работы лишь с обучающей выборкой. Количество гиперпараметров может варьироваться в зависимости от специфики используемого алгоритма.

В данной работе подбор гиперпараметров моделей осуществляется на основе метода кросс-валидации, суть которого описана в первой главе. Исходная выборка разделяется на две основные части: часть обучения и часть теста. Часть обучения исходной выборки представлена данными за период с 2000 года по 2014 год, на которых с помощью метода кросс-валидации происходит настройка гиперпараметров модели. Часть теста выборки представлена данными за период с 2015 года по 2017 год. На данной части выборки производится тестирование адекватности работы алгоритма на основе подобранных параметров.

Первая модель, подлежащая настройке, представляет собой логистическую регрессию. При работе с линейными моделями производится настройка таких гиперпараметров, как метод регуляризации и коэффициент регуляризации.

Регуляризация — это наиболее очевидный способ уменьшить сложность модели, с целью предотвращения переобучения. Как правило, указанная задача решается путем добавления априорной информации в условие задачи. Например, следующим образом:

*Где:*

*– коэффициент регуляризации*

*– метод регуляризации*

Обычно в качестве метода регуляризации используется либо L1 – регуляризатор, либо L2 – регуляризатор. L1-регуляризатор штрафует модель, используя функцию модулей весов модели, а L2 – регуляризатор штрафует модель, используя функцию квадратов весов модели.

На рисунке ниже представлен частичный перебор гиперпараметров моделей с указанием средней точности предсказания по каждому параметру на кроссвалидации.

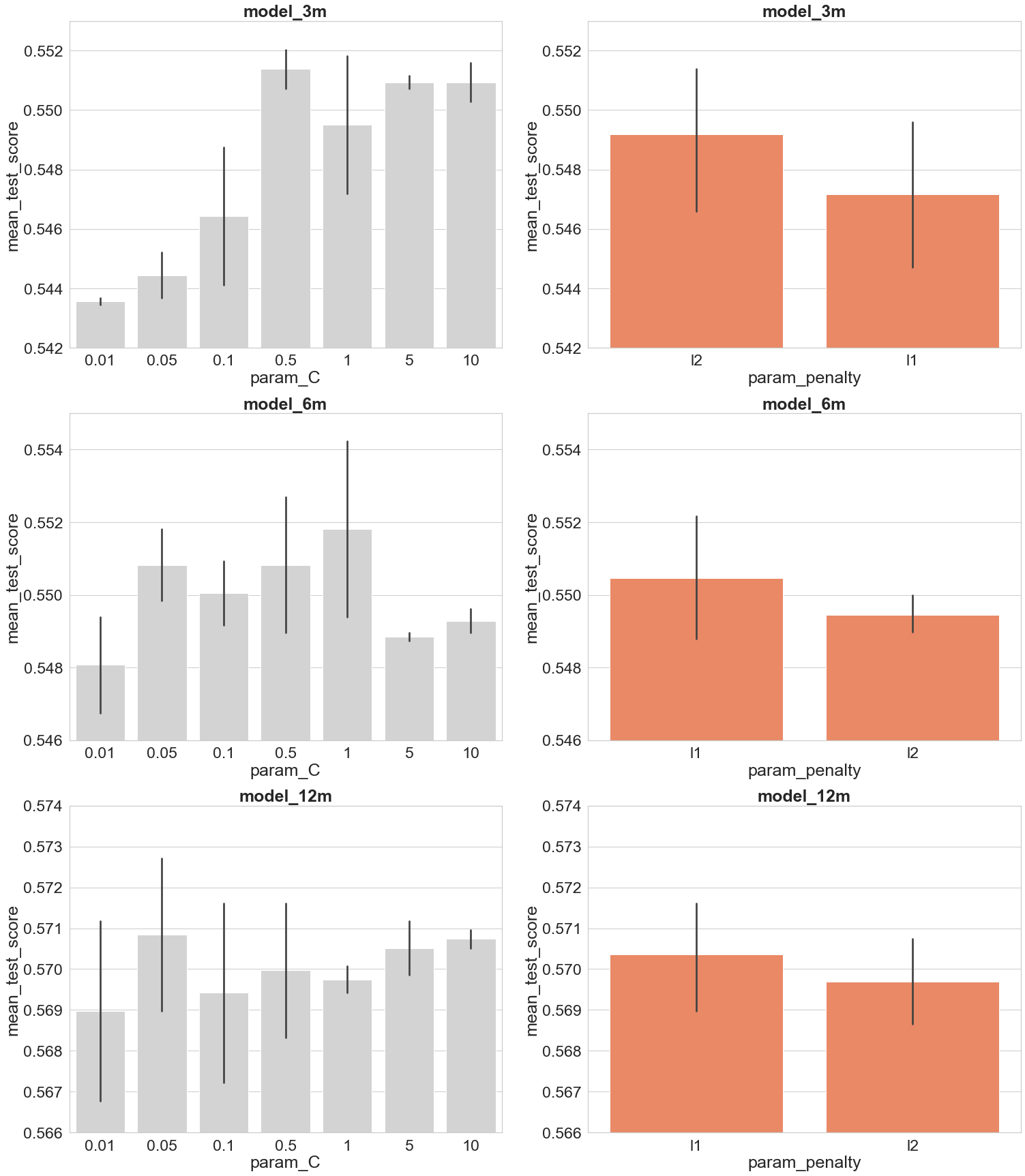


Рис. Полный перебор гиперпараметров линейных моделей

Используя данные, представленные на рисунке, можно сделать следующие выводы:

Во-первых, модель с лагом в 3 месяца показывает наилучшие результаты при использовании L2-регуляризации, при этом коэффициент регуляризации равен 0,5.

Во-вторых, модель с лагом в 6 месяцев показывает наилучшие результаты при использовании L1-регуляризации, при этом коэффициент регуляризации равен 1.

В-третьих, модель с лагом в 12 месяцев показывает наилучшие результаты при использовании L1-регуляризации, при этом коэффициент регуляризации равен 10.

Для выбора наилучшего временного лага проверим качество моделей, построенных с использованием лучших гиперпараметров, на данных за 2015 год. В качестве метрики качества используется ROC-кривая.

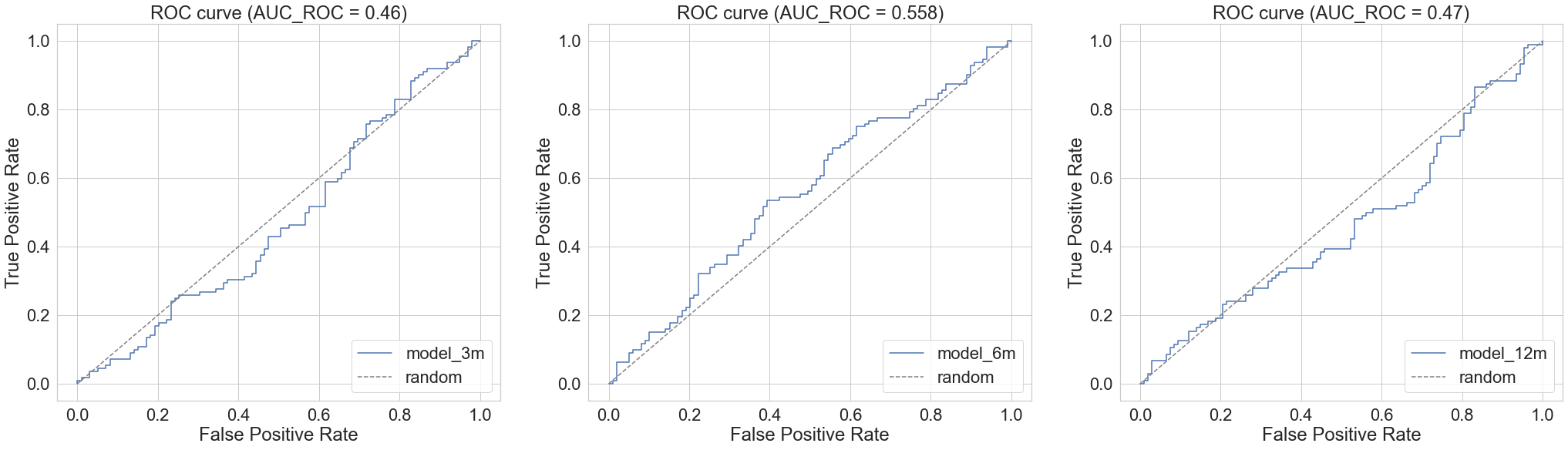


Рис. Проверка качества построенных моделей на данных за 2015 год

Исходя из данных, которые представлены на рисунке, очевидно, что для семейства линейных алгоритмов лаг в 6 месяцев является наиболее оптимальным, ROC\_AUC при этом равен 56%. При этом так же очевидно, что модель не способна давать адекватные прогнозы на годовых и квартальных временных интервалах, так как качество предсказания в обоих случаях ниже качества алгоритма случайного выбора.

## 3.2 Настройки гиперпараметров моделей случайного леса.

Далее проведем настройку гиперпараметров алгоритма случайного леса. В рамках данной работы в качестве настраиваемых гиперпараметров алгоритма случайного леса выступают такие параметры, как количество деревьев, глубина деревьев, число элементов в листе дерева, число подпространств, на котором обучаются деревья.

В сравнении с линейной регрессией, случайный лес куда более сложный и ресурсоемкий алгоритм. Провести полный перебор его гиперпараметров не представляется возможным ввиду ограниченности вычислительных мощностей. Чтобы решить данную проблему, в качестве промежуточного шага используется алгоритм жадной оптимизации, который призван сократить количество значений гиперпараметров, подлежащих дальнейшему перебору.

На каждой итерации алгоритма жадной оптимизации параметры модели разбиваются на две части – изменяемую и дефолтную. В качестве дефолтных параметров выступают изначальные параметры модели, предоставляемые библиотекой sklearn[[37]](#footnote-37). Изменяемая часть представлена гиперпараметром, настройка которого производится на данной итерации алгоритма. Осуществляется последовательное измерение качества модели на кроссвалидации для разных значений оптимизируемого параметра при фиксированной дефолтной части. Далее три лучших значения исследуемого параметра отправляются на следующий этап перебора параметров для выявления синергетических эффектов.

В первую очередь проводим тестирование качества модели при изменении количества деревьев, составляющих модельный ансамбль.

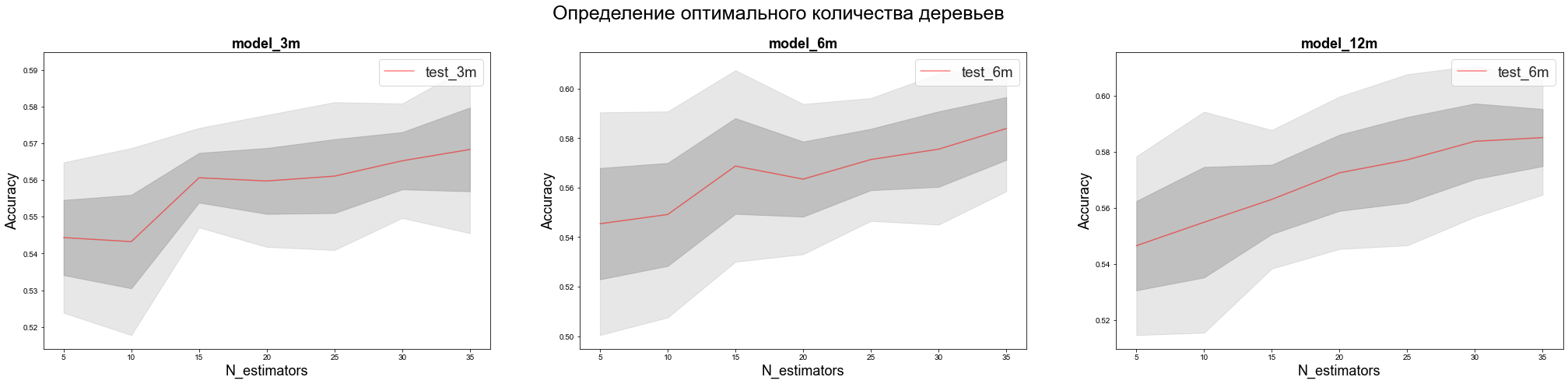
****

Рис. Определение оптимального числа деревьев в ансамбле для моделей случайного леса

Из рисунка видно, что качество моделей последовательно растет при росте количества деревьев, входящих в ансамбль. Однако в данной ситуации необходимо действовать очень осторожно ввиду того, что чрезмерное количество базовых эстиматоров приводит к росту коррелированности их ответов, что противоречит идее случайного леса и ведет к переобучению алгоритма. Поэтому на данной итерации нецелесообразно использовать более 35 деревьев в рамках одного ансамбля.

Дале проводим тестирование качества модели при изменении глубины деревьев, входящих в ансамбль.

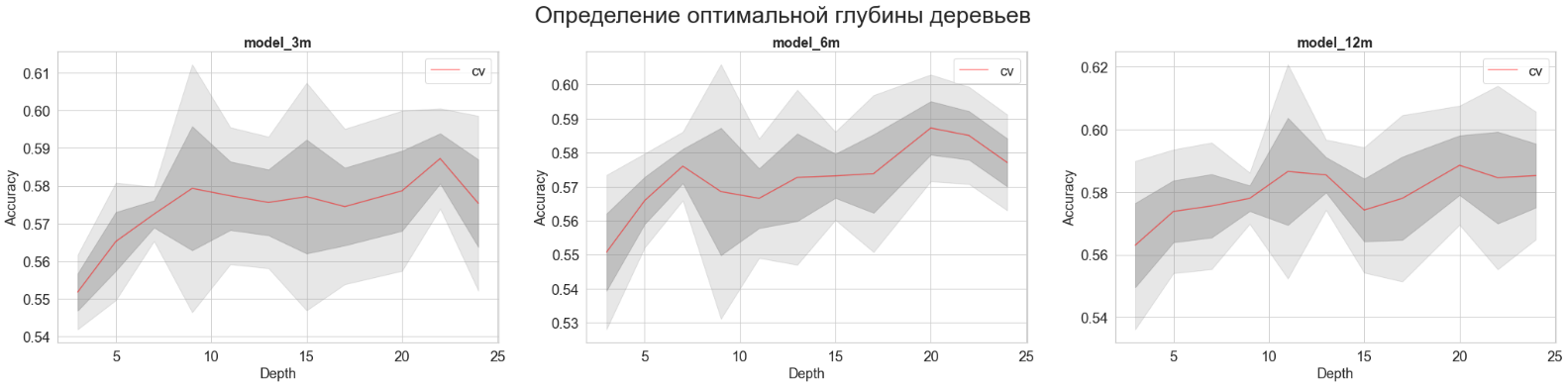
****

Рис. Определение оптимальной глубины деревьев для моделей случайного леса

Из рисунка видно, что рост глубины деревьев после 10 итерации замедляется и выходит на асимптоту. Следовательно, оптимальное количество деревьев не может быть меньше 10. Три параметра с лучшими результатами для каждой модели отправляются на следующий этап для полного перебора.

Далее определяем оптимальное количество элементов в листе каждого дерева. Данный параметр является регуляризатором для деревьев. Рост показателя снижает качество модели на обучающей выборке, но в то же время повышает ее устойчивость и отлично борется с переобучением.

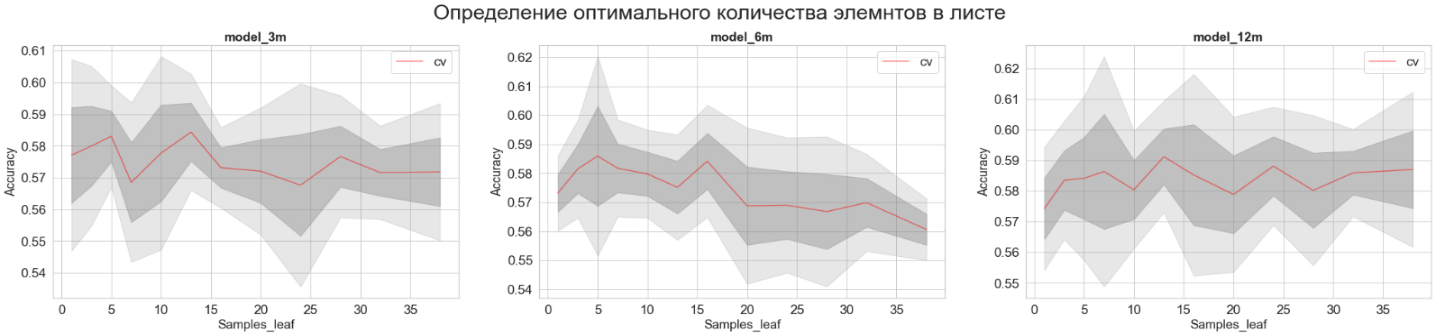
****

Рис. Определение оптимального числа элементов в листе дерева для моделей случайного леса

Из рисунка видно, что увеличение числа элементов в листе дерева не приводит к существенному снижению качества алгоритма, поэтому в данном случае на следующий этап оценки отправляются значения параметра выше 20.

Завершающим этапом жадной оптимизации является настройка числа подпространств. Число подпространств характеризует количество факторов, на котором обучаются различные деревья. Данный показатель так же входит в группу регуляризационных гиперпараметров и способствует снижению вероятности переобучения модели.

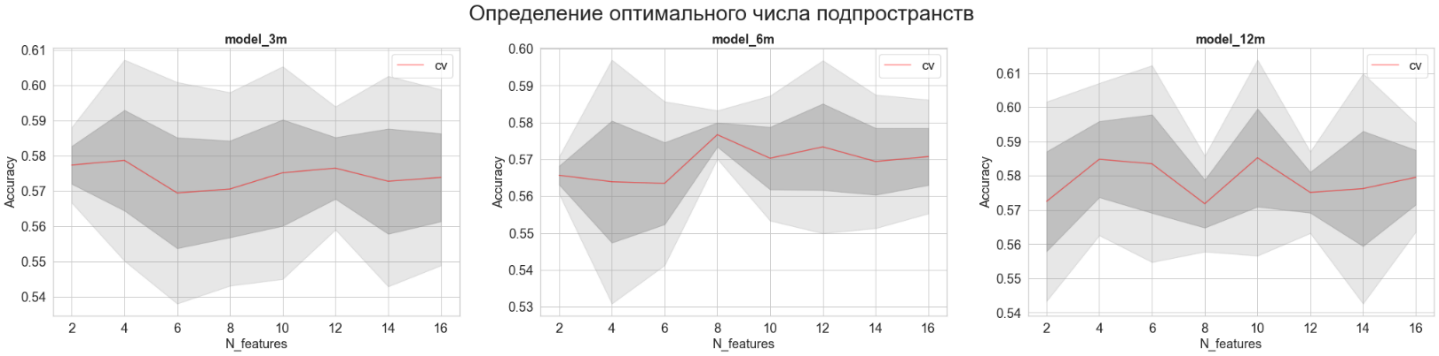
****

Рис. Определение оптимального числа подпространств для моделей случайного леса

Выбрав наиболее эффективные значения гиперпараметров с помощью алгоритма жадной оптимизации, проведем их полный перебор, чтобы определить наиболее эффективное сочетание для каждой модели.

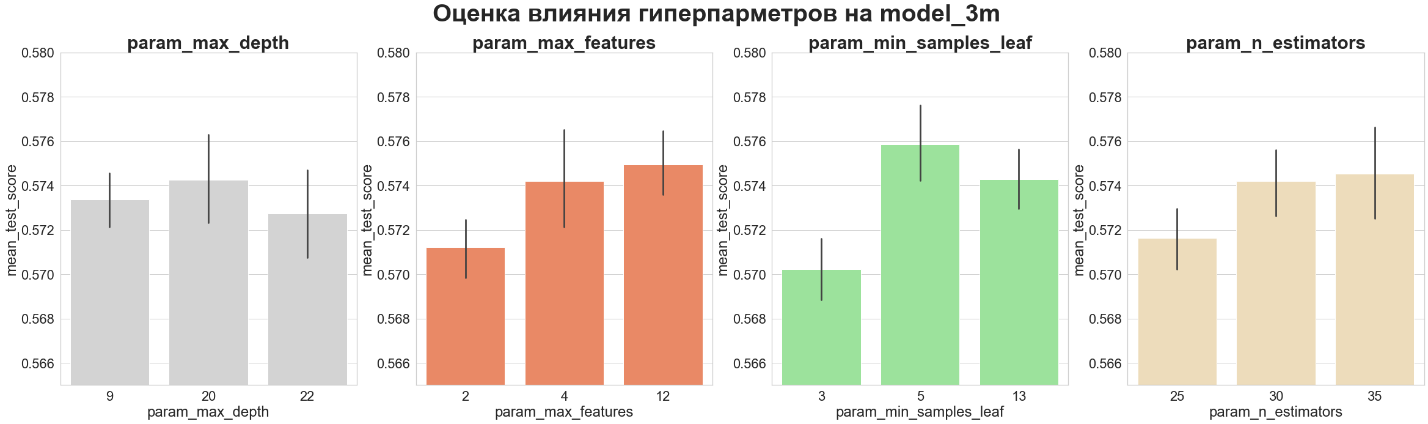
****

Рис. Полный перебор гиперпараметров модели случайного леса с лагом в три месяца

Из диаграмм, представленных на рисунке видно, что наиболее оптимальные параметры для модели с лагов в три месяца представлены следующей комбинацией: оптимальная глубина деревьев равна 20 итерациям, оптимальное количество подпространств равно 12 факторам, оптимальное количество элементов в листе равно 12 элементам, оптимальное количество деревьев равно 30 деревьям.

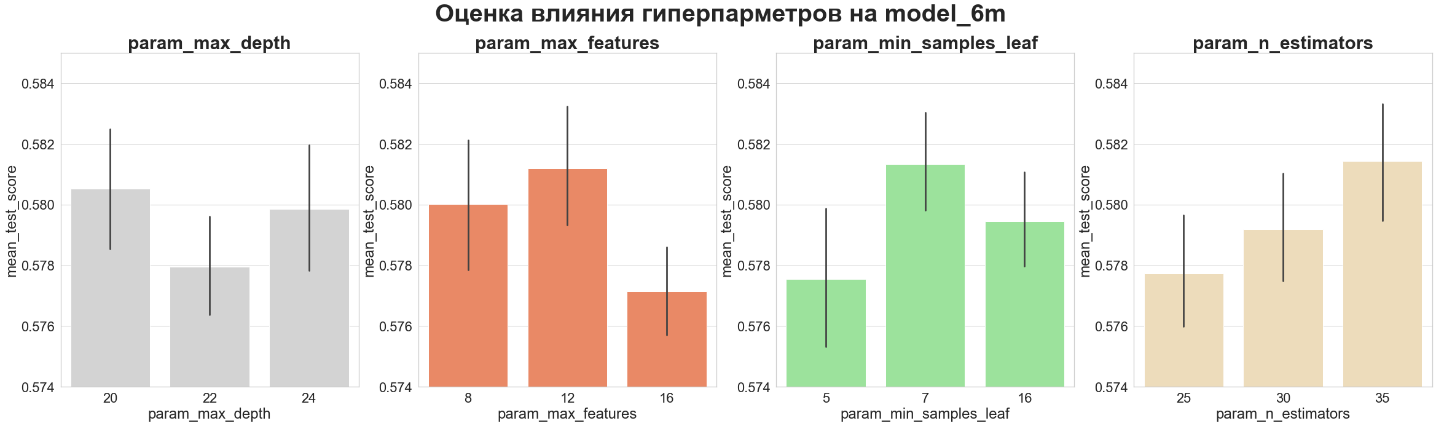
****

Рис. Полный перебор гиперпараметров модели случайного леса с лагом в шесть месяцев

Из диаграмм, представленных на рисунке видно, что наиболее оптимальные параметры для модели с лагом в шесть месяцев представлены следующей комбинацией: оптимальная глубина деревьев равна 20 итерациям, оптимальное количество подпространств равно 12 факторам, оптимальное количество элементов в листе равно 7 элементам, оптимальное количество деревьев равно 35 деревьям.

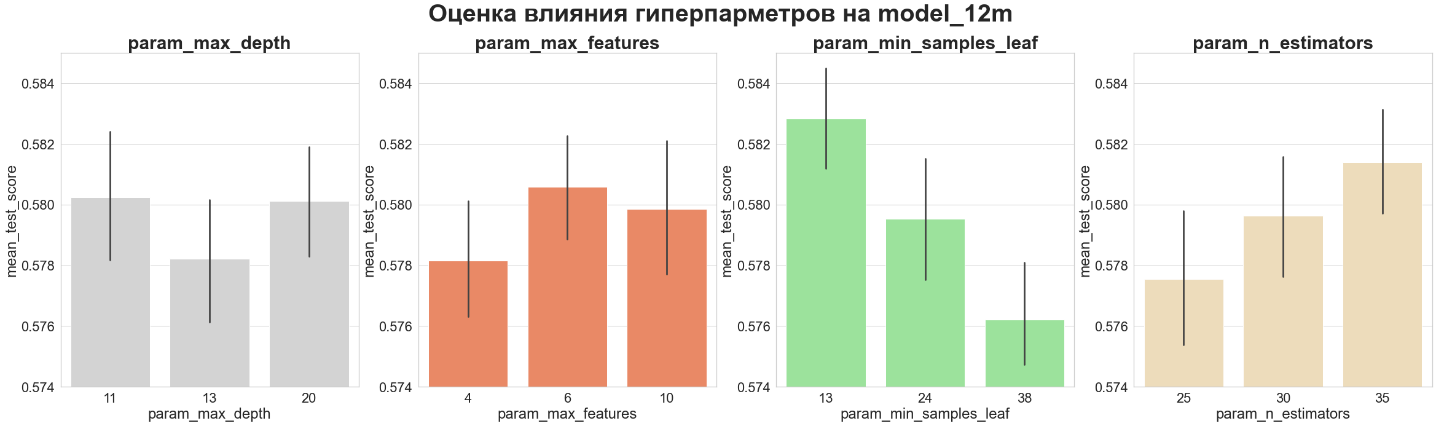
****

Рис. Полный перебор гиперпараметров модели случайного леса с лагом в двенадцать месяцев

Из диаграмм, представленных на рисунке видно, что наиболее оптимальные параметры для модели с лагом в двенадцать месяцев представлены следующей комбинацией: оптимальная глубина деревьев равна 11 итерациям, оптимальное количество подпространств равно 6 факторам, оптимальное количество элементов в листе равно 13 элементам, оптимальное количество деревьев равно 35 деревьям.

После настройки гиперпараметров моделей случайного леса проводится калибровка прогнозируемых алгоритмом вероятностей. На рисунке ниже представлены кривые прогнозируемых вероятностей до и после калибровки.

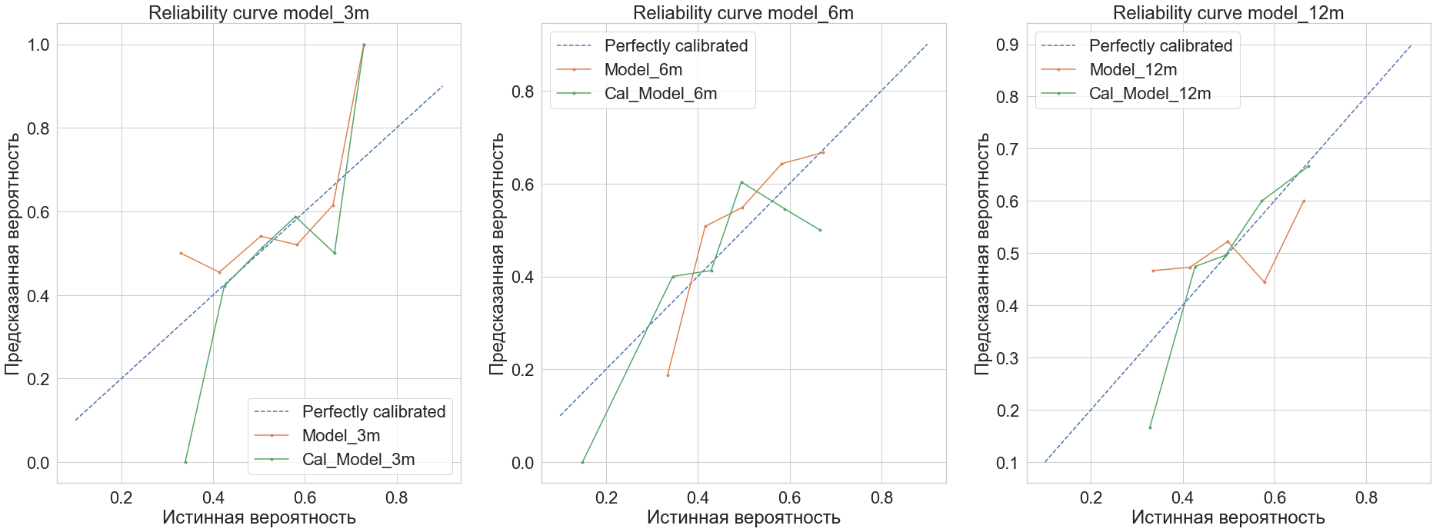
****

Рис. Калибровка вероятностей моделей случайного леса

Основываясь на данных, представленных на рисунке, можно сделать вывод, что калибровка вероятностей для модели с лагом в три месяца не приводит к существенному улучшению качества предсказанных вероятностей. Однако для моделей с лагом в шесть и двенадцать месяцев имеет смысл проводить калибровку вероятностей.

Для выбора наилучшего временного лага проверим качество откалиброванных моделей, построенных с использованием лучших гиперпараметров, на данных за 2015 год. В качестве метрики качества используется AUC-ROC.

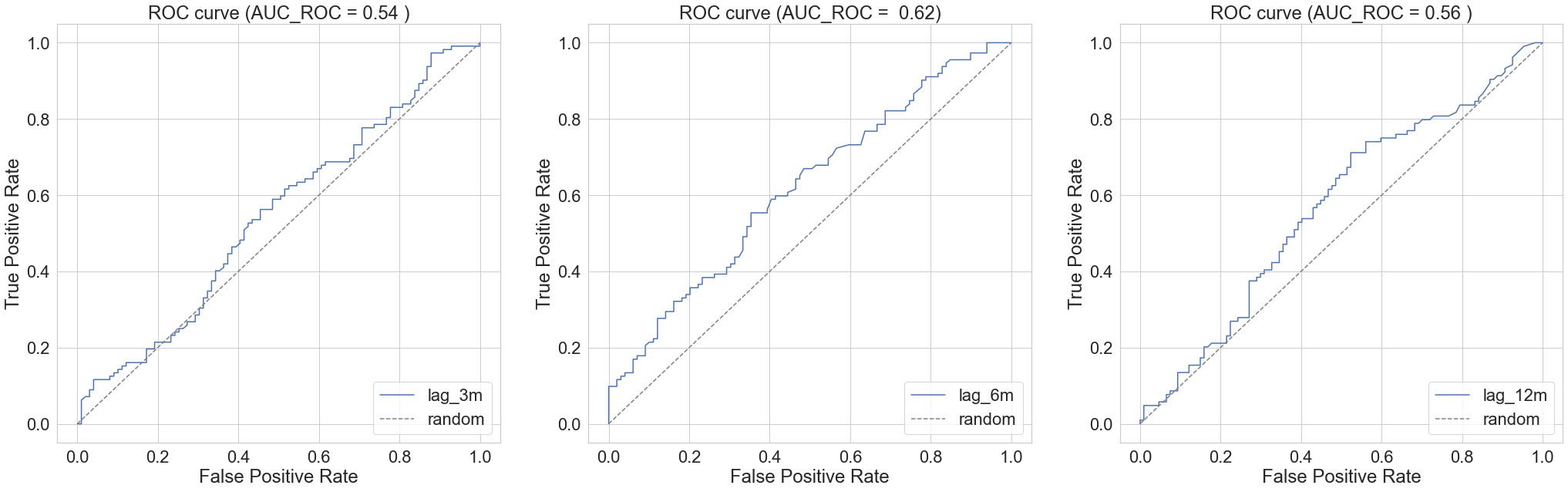
****

Рис. Проверка качества моделей случайного леса на данных за 2015 год

Исходя из данных, которые представлены на рисунке, очевидно, что для семейства деревянных алгоритмов лаг в 6 месяцев является наиболее оптимальным, AUC-ROC при этом равен 62%. При этом видно, что качество моделей с лагом в 3 и 12 месяцев существенно уступает модели с лагом в 6 месяцев и составляет 54 и 56 процентов соответственно.

Опираясь на данные, полученные при обучении алгоритма случайного леса на шестимесячном интервале предсказания, построим диаграмму значимость факторов, используя модель SHAP[[38]](#footnote-38). На рисунке представлены топ-15 наиболее значимых факторов. Они отсортированы по значимости от наиболее важного к наименее важному фактору сверху вниз. По горизонтальной оси для каждого фактора отложены значения SHAP-value, увеличение которых приводит к росту вероятности принадлежности объекта классу 1. Каждое наблюдение, участвующее в процессе обучения модели, представлено в виде точки на горизонтальной оси, при этом большим значениям фактора соответствует оттенок красного цвета, меньшим значениям соответствуют синие тона.

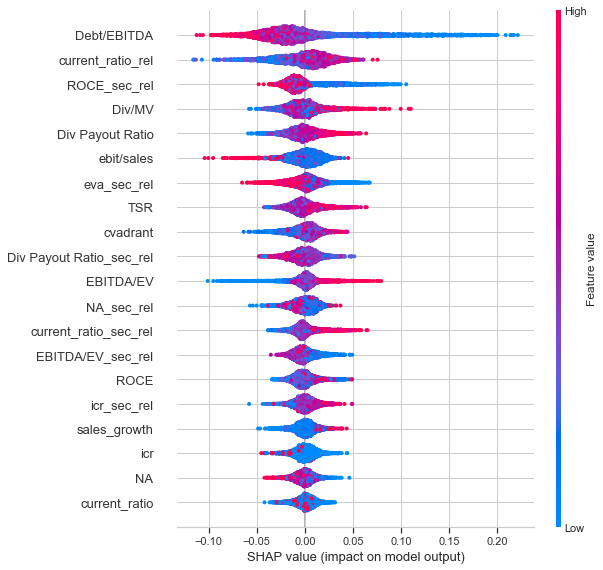
****

Рис. Диаграмма модели SHAP для алгоритма случайного леса

Коэффициент отношение долга к показателю EBITDA является наиболее важным фактором по версии модели SHAP. При этом объекты выборки, характеризующиеся высокими значениями данного показателя, с меньшей вероятностью будут отнесены алгоритмом к классу 1, нежели объекты с низкими значениями показателя. Вторым по значимости фактором выступает коэффициент отношения текущей ликвидности компании к среднеотраслевому коэффициенту ликвидности. Исходя из данных, предоставленных моделью SHAP, компании, коэффициент ликвидности которых превышает среднеотраслевой уровень, с большей вероятностью продемонстрируют доходность выше среднеотраслевой на шестимесячном интервале. Так же согласно данным, представленным на диаграмме, более высокие дивидендные выплаты компаний приводят к повышению вероятности отнесения компании к классу 1. Отлично сепарирует выборку мультипликатор отношения EBITDA к EV, значение мультипликатора прямо пропорционально вероятности отнесения компании к классу 1.

## 3.3 Настройка гиперпараметров моделей градиентного бустинга.

Рассмотрим третий класс моделей, подлежащих настройке. В рамках данной работы в качестве настраиваемых гиперпараметров алгоритма градиентного бустинга выступают такие параметры, как глубина деревьев, вес каждого дерева в общем ансамбле, параметр регуляризации гамма, доля выборки, на которой обучаются алгоритмы на каждой итерации, доля подпространств, на котором обучаются деревья на каждой итерации, и шаг градиентного спуска «eta». Настройка параметров алгоритма градиентного бустинга осуществляется по схеме аналогичной настройке параметров алгоритма случайного леса.

В первую очередь проводим тестирование качества модели при изменении глубины деревьев, составляющих модельный ансамбль.

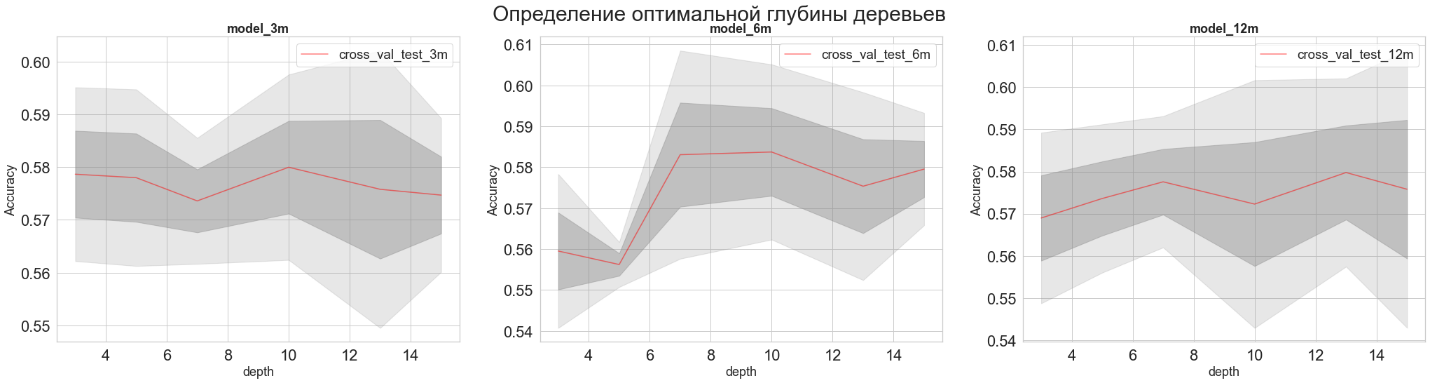
****

Рис. Определение оптимальной глубины деревьев для моделей градиентного бустинга

Из рисунка видно, что наилучшие результаты модели показывают при значениях параметра в диапазоне от 7 до 12 итераций.

Далее определяем оптимальное количество элементов в листе каждого дерева. Данный параметр является регуляризатором для деревьев. Рост показателя снижает качество модели на обучающей выборке, но в то же время повышает ее устойчивость и отлично борется с переобучением.

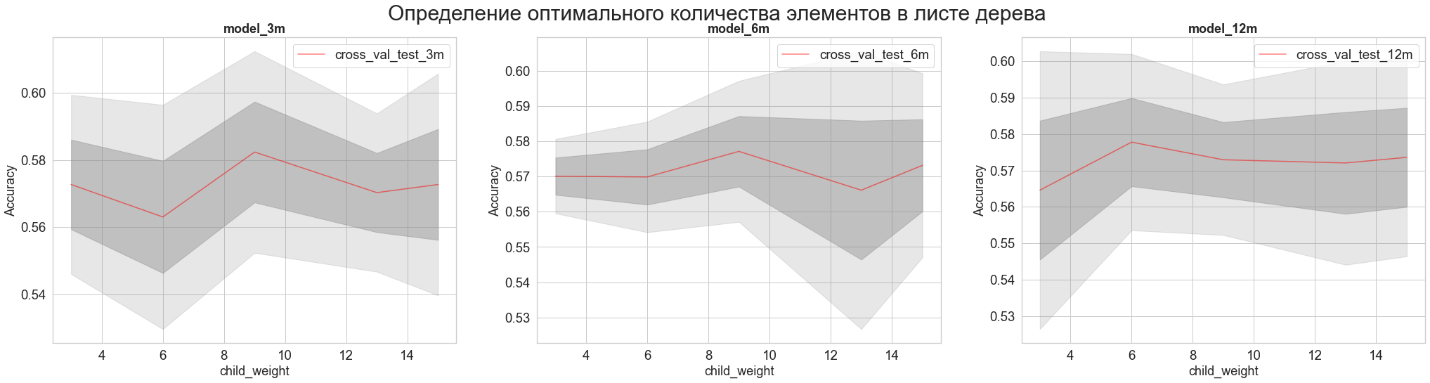
****

Рис. Определение оптимального количества элементов в листе дерева для моделей градиентного бустинга

Из рисунка видно, что наилучшие результаты модели показывают при значениях параметра в диапазоне от 10 до 14 элементов в каждом листе дерева.

Следующим параметром, подлежащим отбору, выступает параметр регуляризации гамма. Гамма – минимальное снижение значения функции потерь, необходимое для дальнейшего разбиения вершины дерева. Большие значения параметра приводят к более консервативным моделям.

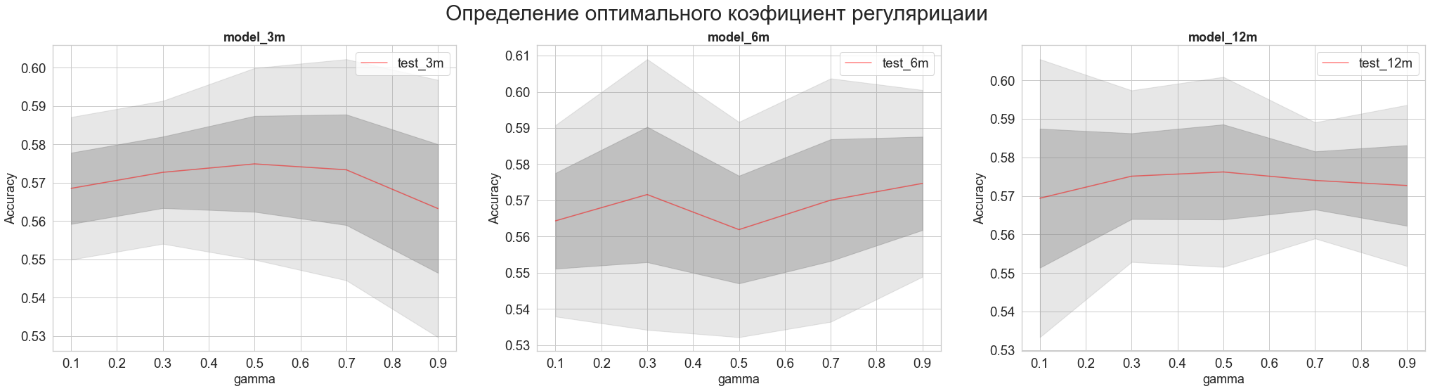
****

Рис. Определение оптимального коэффициента регуляризации gamma для моделей градиентного бустинга

Из рисунка видно, что наилучшие результаты модели показывают при значениях параметра гамма в диапазоне от 0.4 до 0.8.

Далее осуществляется оптимизация доли подпространств, на котором обучается каждое дерево. Если доля подпространств, использующихся при обучении каждого дерева меньше 1, то выбирается случайная подвыборка из общей совокупности, что помогает в борьбе с переобучением.

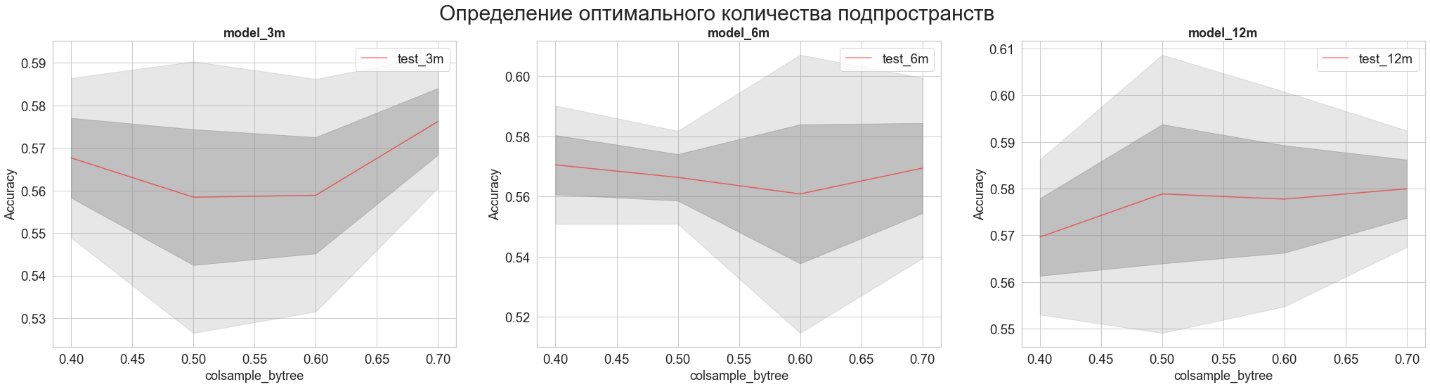
****

Рис. Определение оптимального количества подпространств для моделей градиентного бустинга

Из рисунка видно, что наилучшие результаты модели показывают, если каждое дерево обучается на доле подпространств в диапазоне от 50 до 60 процентов.

Следующий параметр описывает долю объектов выборки, на которой обучается каждое дерево и несет в себе смысл аналогичный параметру доли случайных подпространств.

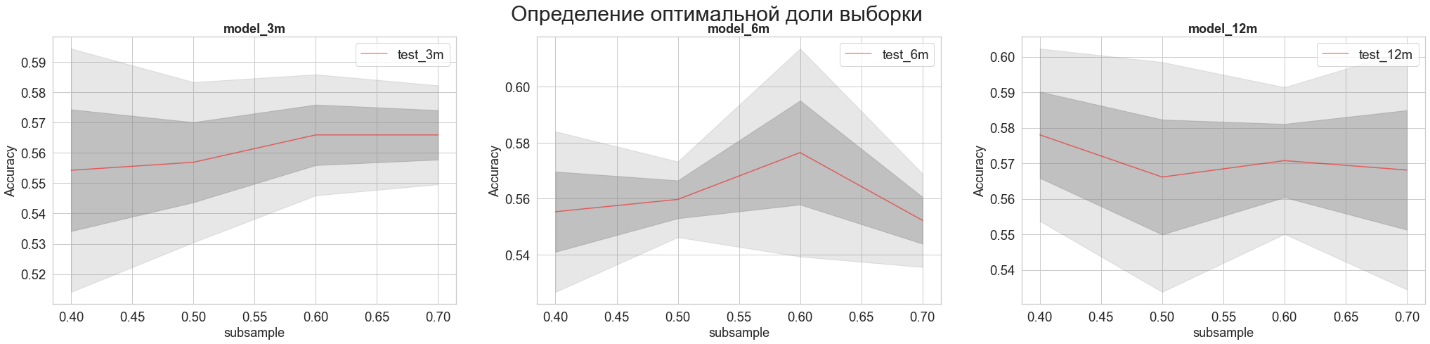
****

Рис. Определение оптимальной доли выборки для обучения для моделей градиентного бустинга

Из данных, представленных на рисунке видно, что наилучшие результаты модели показывают, если каждое дерево обучается на доле объектов выборки в диапазоне от 50 до 60 процентов.

Последним параметром настраивается величина шага градиентного спуска «eta», которая описывает темп обучения. Данный параметр используется для предотвращения переобучения за счёт "сокращения" весов базовых алгоритмов, делая модель более консервативной. Чем меньше параметр eta, тем больше нужно итераций для обучения модели с хорошим качеством.

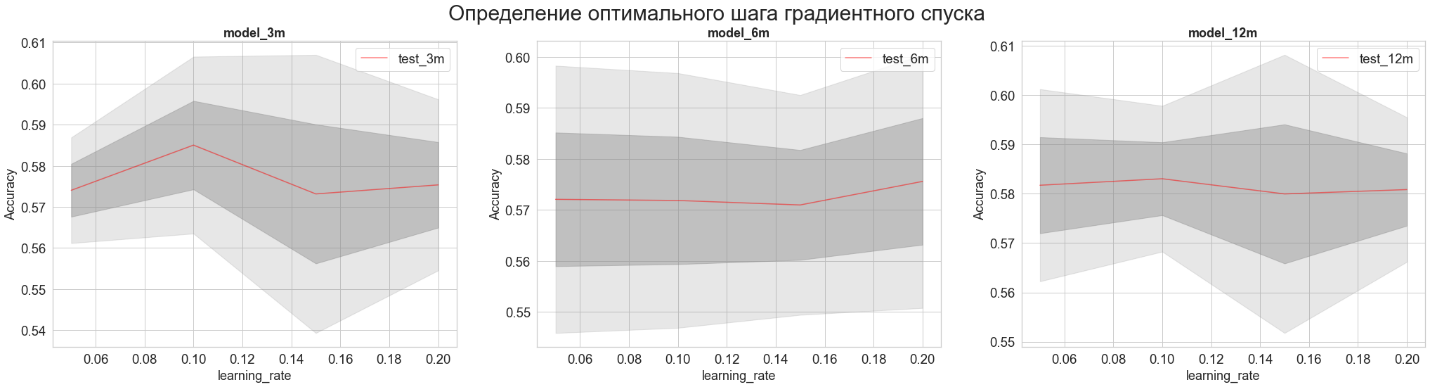
****

Рис. Определение оптимального шага градиентного спуска для моделей градиентного бустинга

Из данных, представленных на рисунке видно, что модели, построенные на трех и двенадцатимесячных лагированных доходностях, показывают наилучшие результаты при значении параметра «eta» в диапазоне от 0,06 до 0,1. Модель, построенная на шестимесячных лагированных доходностях, показывает наилучшие результаты при значении параметра «eta» в диапазоне от 0,16 до 0,2.

Выбрав наиболее эффективные значения гиперпараметров с помощью алгоритма жадной оптимизации, проведем их полный перебор, чтобы определить наиболее эффективное сочетание для каждой модели.

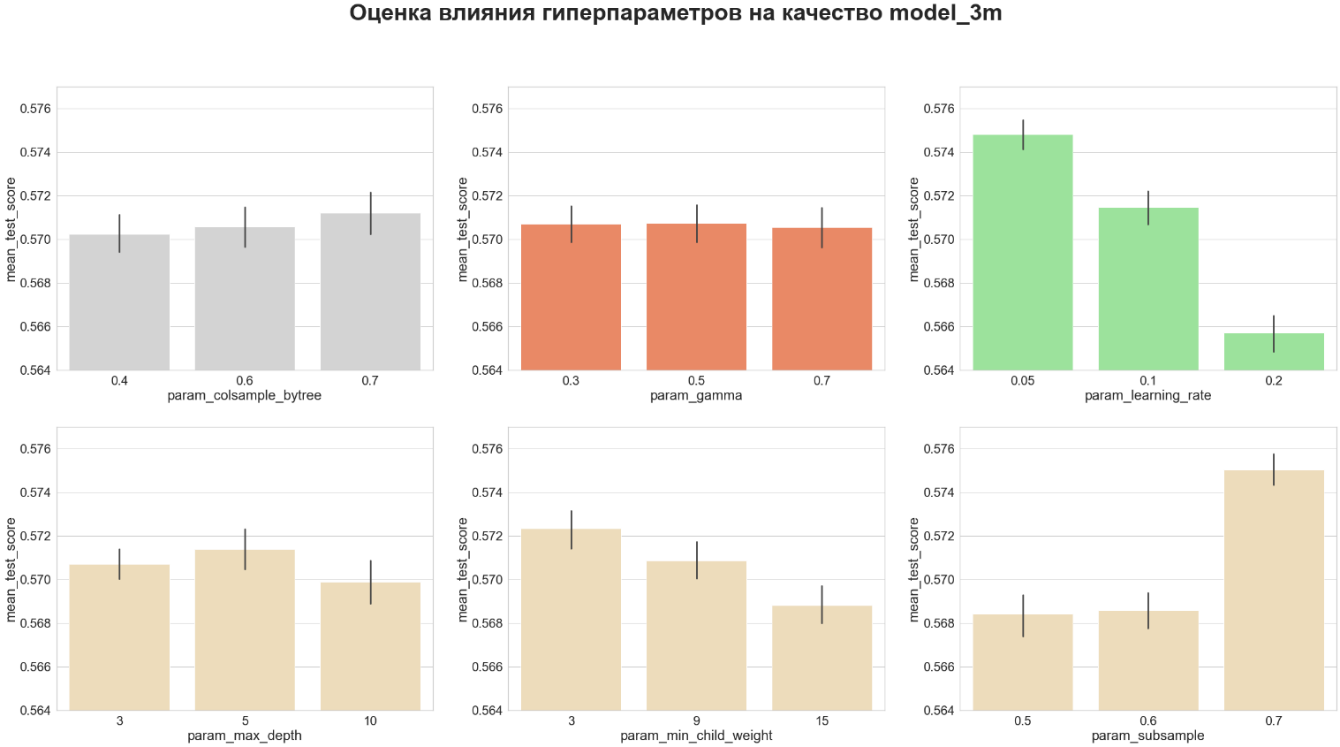


Рис. Полный перебор гиперпараметров модели градиентного бустинга с лагом в 3 месяца

Из диаграмм, представленных на рисунке видно, что наиболее оптимальные параметры для модели с лагом в три месяца представлены следующей комбинацией: оптимальная глубина деревьев равна 5 итерациям, оптимальное количество подпространств равно 70 процентам, оптимальное количество объектов выборки, на которой обучается алгоритм равно 70 процентам, оптимальное количество элементов в выборке равно 9 элементам, наиболее оптимальный параметр гамма равен 0,5, а шаг градиентного бустинга равен 0,05.

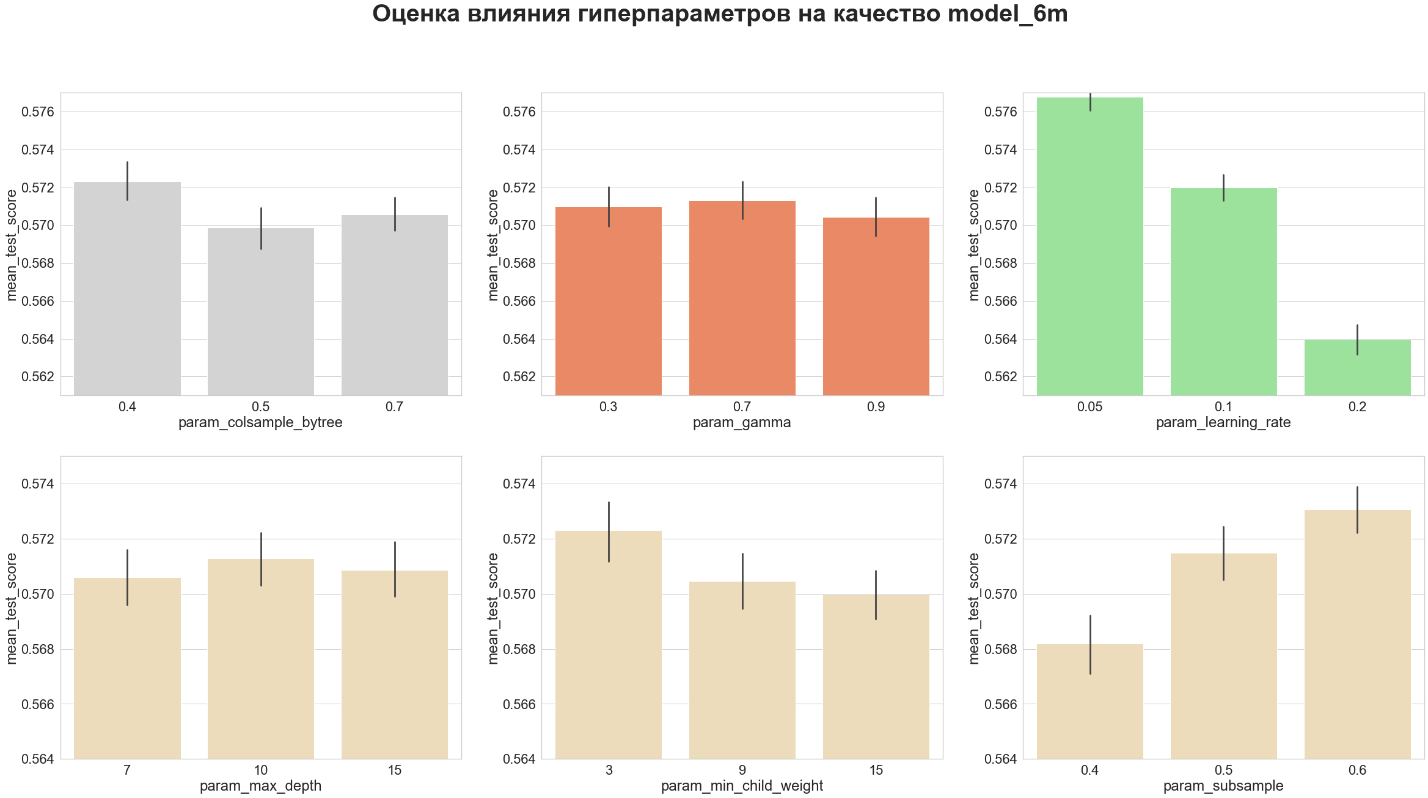


Рис. Полный перебор гиперпараметров модели градиентного бустинга с лагом в 6 месяцев

Из диаграмм, представленных на рисунке видно, что наиболее оптимальные параметры для модели с лагом в три месяца представлены следующей комбинацией: оптимальная глубина деревьев равна 10 итерациям, оптимальное количество подпространств равно 40 процентам, оптимальное количество объектов выборки, на которой обучается алгоритм равно 60 процентам, оптимальное количество элементов в листе равно 9 элементам, наиболее оптимальный параметр гамма равен 0,7, а шаг градиентного бустинга равен 0,05.

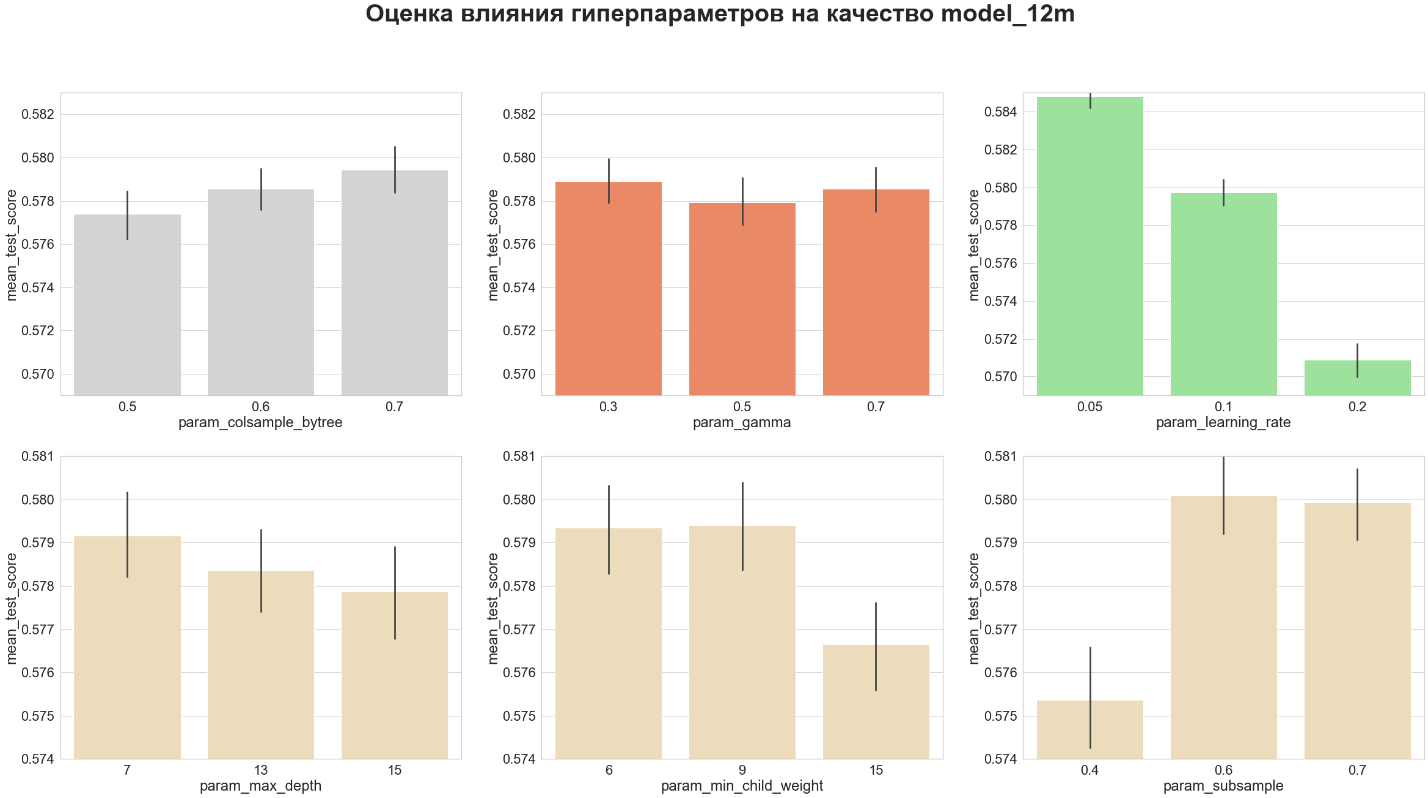


Рис. Полный перебор гиперпараметров модели градиентного бустинга с лагом в 12 месяцев

Из диаграмм, представленных на рисунке видно, что наиболее оптимальные параметры для модели с лагом в три месяца представлены следующей комбинацией: оптимальная глубина деревьев равна 7 итерациям, оптимальное количество подпространств равно 70 процентам, оптимальное количество объектов выборки, на которой обучается алгоритм равно 60 процентам, оптимальное количество элементов в листе равно 9 элементам, наиболее оптимальный параметр гамма равен 0,3, а шаг градиентного бустинга равен 0,05.

Для выбора наилучшего временного лага проверим качество моделей, построенных с использованием лучших гиперпараметров, на данных за 2015 год. В качестве метрики качества используется ROC-AUC.

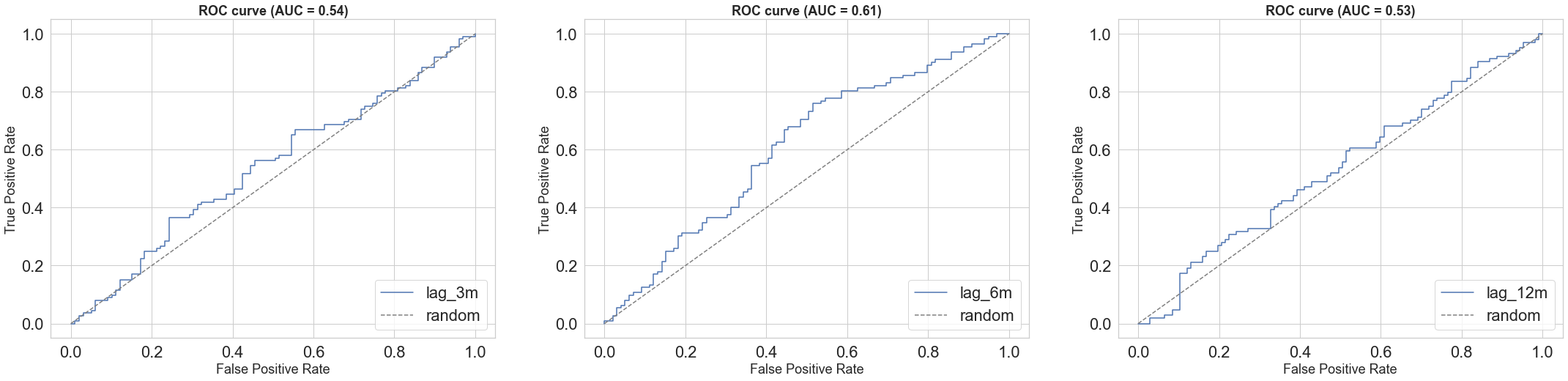
****

Рис. Проверка качества моделей градиентного спуска на данных за 2015 год

Исходя из данных, которые представлены на рисунке, видно, что для семейства алгоритмов градиентного бустинга лаг в 6 месяцев является наиболее оптимальным, AUC-ROC при этом равен 61%. Модели, построенные с лагом в 3 и 12 месяцев, демонстрируют качество немногим выше качества алгоритма случайного выбора.

Опираясь на данные, полученные при обучении алгоритма градиентного бустинга на шестимесячном интервале предсказания, построим диаграмму значимость факторов, используя модель SHAP.

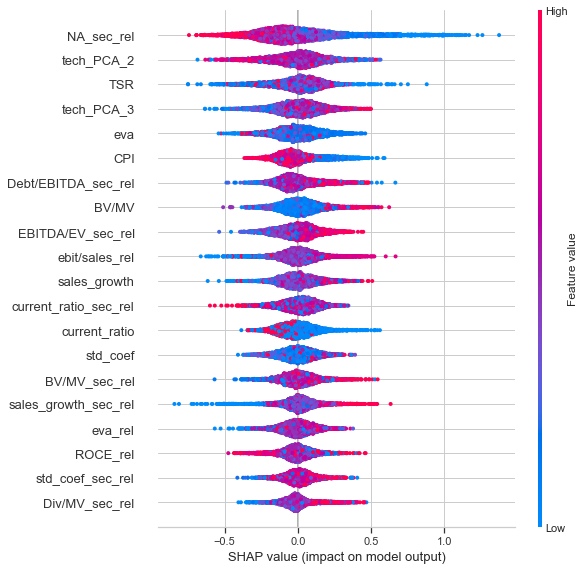
****

Рис. Диаграмма модели SHAP для алгоритма градиентного бустинга

Коэффициент отношение долга к показателю EBITDA является наиболее важным фактором по версии модели SHAP. При этом объекты выборки, характеризующиеся высокими значениями данного показателя, с меньшей вероятностью будут отнесены алгоритмом к классу 1, нежели объекты с невысокими показателями. Вторым по значимости фактором выступает коэффициент отношения текущей ликвидности компании к среднеотраслевому коэффициенту ликвидности. Исходя из данных, предоставленных моделью SHAP, компании, коэффициент ликвидности которых превышает среднеотраслевой уровень, с большей вероятностью продемонстрируют доходность выше среднеотраслевой на шестимесячном интервале. Так же согласно данным, представленным на диаграмме, для таких показатели, как dividends/MV, EBITDA\EV высокие значения приводят к повышению вероятности принадлежности объекта к классу 1.

Исходя из тестирования алгоритмов на данных за 2015 год, становится ясно, что оптимальный горизонт предсказания моделей составляет 6 месяцев с даты публикации отчетностей компаниями. Протестируем качество работы алгоритмов на данных за 2016 и 2017 года.

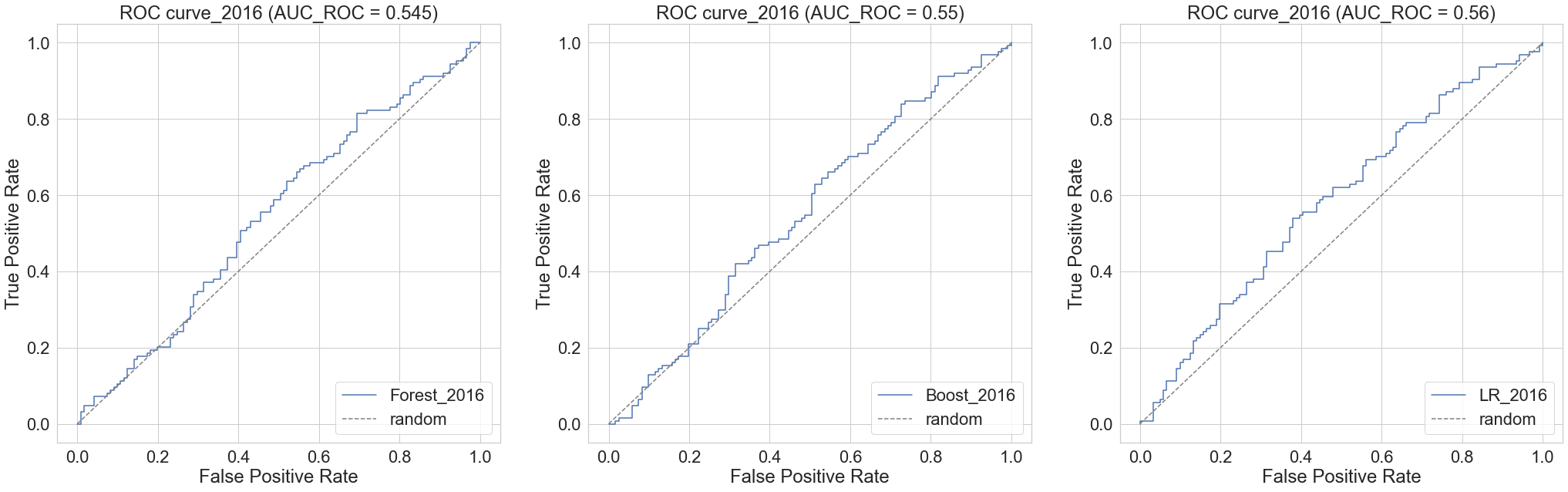
****

Рис. Качество работы алгоритмов градиентного спуска на данных за 2016 год

Анализ качества работы алгоритмов за 2016 год указывает на незначительно снижение предсказательной силы алгоритмов на данном временном интервале, наилучший результат демонстрирует модель линейной регрессии, AUC\_ROC которой составляет 56%. Далее проанализируем качество работы алгоритмов на данных за 2017 год.

****

Рис. Качество работы алгоритмов градиентного спуска на данных за 2017 год

Из рисунков видно, что качество предсказаний алгоритмов подвержено достаточно серьезным колебаниям год от года. Качество работы алгоритма случайного леса на данных за 2017 год возросло на 7% в сравнении с его качеством на данных за 2016 год. Качество работы градиентного бустинга так же увеличилось, однако чуть скромнее, всего на 1%. В то же время AUC-ROC логистической регрессии снизился на 1,5%. Для решения данной проблемы воспользуемся алгоритмом стекинга, суть которого была описана в первой главе. Он позволяет стабилизировать результаты работы алгоритмов и предоставляет интегральную оценку предсказанных вероятностей.

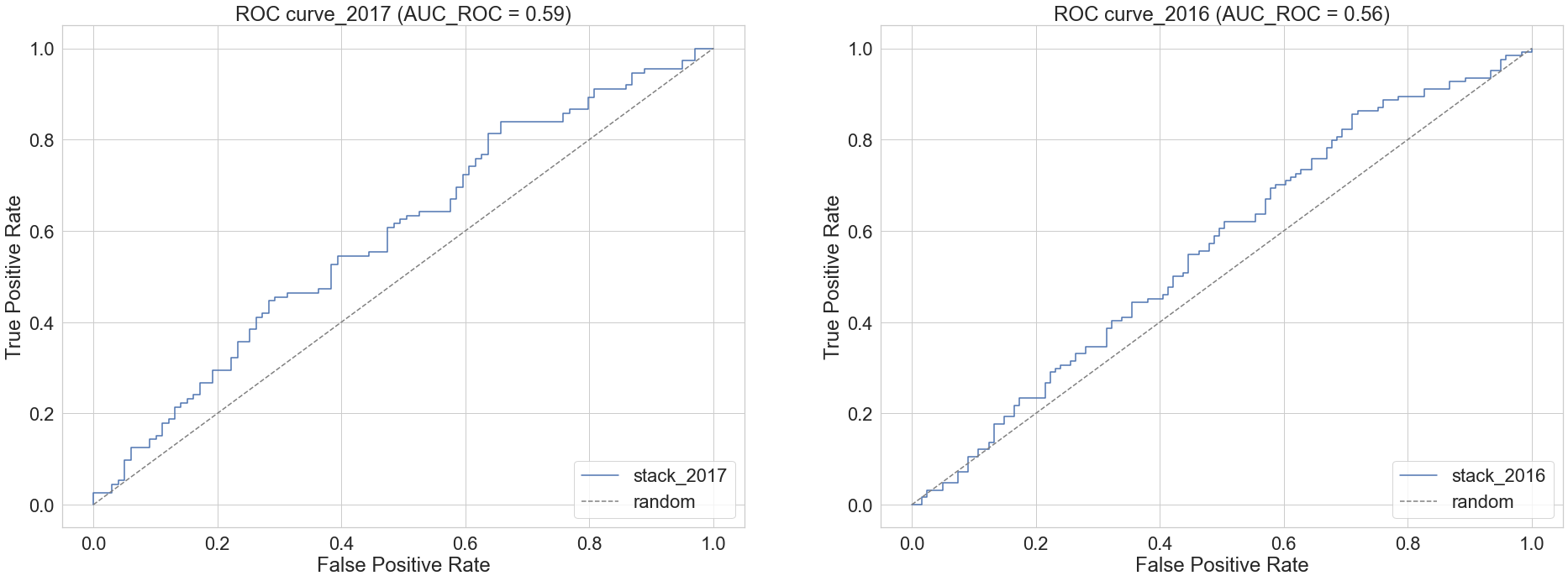
****

Рис. Качество предсказания алгоритма стекинга на данных за 2016 и 2017

Из рисунка видно, что качество алгоритма стекинга год от года изменяется на 3 процента, что указывает на стабильность его работы. ROC-AUC алгоритма на данных за 2016 год составляет 56%, за 2017 год – 59%.

Используя вероятности, предсказанные алгоритмом, составим рейтинг инвестиционной привлекательности компаний за 2016 и 2017 года, а также протестируем инвестиционную стратегию, основанную на комбинации принципов современной портфельной теорию и рейтинговых оценках.

## 3.4 Результаты рейтинговой оценки компаний

Согласно методологии составления рейтинга, описанной в первой главе, проранжированный по предсказанным вероятностям ряд компаний разделяется на 10 интервалов по квантилям распределения. В приложениях 2 и 3 представлены таблицы, в которых отражены 20 наиболее перспективных компаний по данным за 2016 и 2017 года соответственно по версии алгоритма стекинга. В таблицах так же указаны доходности компаний и доходности секторов, к которым рассматриваемые компании относятся. Ниже представлены гистограммы, в которых отражено разбиение компаний, вошедших в топ 20 по секторам за 2017 и 2016 года.

Большинство компаний, вошедших в топ 20, представляют 2-3 сектора экономики США. Следовательно, учитывая основополагающие правила построения инвестиционных стратегий, в основе которых лежит принцип диверсификации рисков, необходимо увеличить число компаний, представляющих иные сектора экономики. Для реализации поставленной задачи за сектором закрепляется доля в итоговой выборке топ 20 инвестиционно-привлекательных компаний в том случае, если хотя бы одна компания из этого сектора получила оценку алгоритма выше 50%. При этом доли между секторами в выборке распределяются пропорционально количеству компаний сектора, перешагнувших порог 50% вероятности успеха по версии алгоритма стекинга. Обновленные перечни компаний, которые вошли в топ-20 инвестиционно-привлекательных компаний за 2016 и 2017 года, представлены в приложениях номер 4 и 5.

На основе разработанного перечня инвестиционно-привлекательных компаний сформируем инвестиционной портфель и протестируем его доходность. В качестве теоретической базы воспользуемся основными положениями современной портфельной теории. Согласно портфельной теории Марковица инвестор принимает решение о выборе акций в портфель, исходя из правила "ожидаемая доходность-дисперсия доходности", другими словами, инвестор хочет минимизировать риск, который связан с вложением, и максимизировать ожидаемую доходность. При этом ожидаемая доходность портфеля представляет собой линейную комбинацию ожидаемых доходностей ценных бумаг, входящих в этот портфель и рассчитывается по формуле

,

*Где:*

*– доля i бумаги в портфеле;*

*– ожидаемая доходность i-ой бумаги в портфеле.*

А дисперсия портфеля равна следующей величине ввиду того, что доходности бумаг, входящих в портфель скоррелированы:

V (R),

Где:

– ковариация бумаги i-ой с бумагой j-ой

*– доля i-ой бумаги в портфеле;*

*– доля j-ой бумаги в портфеле;*

Под эффективным портфелем понимается портфель, который обеспечивает минимальное значение риска при заданном уровне ожидаемой доходности. Все множество портфелей, удовлетворяющих указанному условию, формируют границу эффективных портфелей.

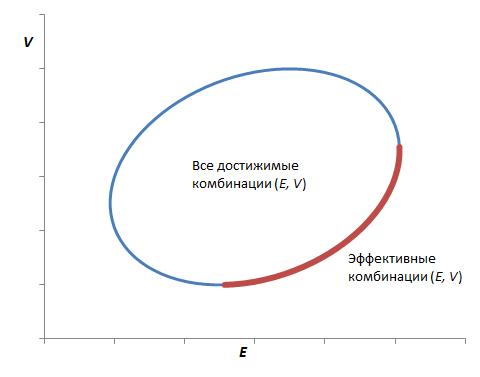


Рис. Граница эффективных портфелей

Портфельная теория Марковица завоевала доверие многих инвесторов и является одним из наиболее простых и эффективных инструментов конструирования инвестиционных портфелей. Однако очевидный недостаток данного метода заключается в том, что решения о внесении или не внесении акции в портфель принимаются исключительно исходя из ретроспективного анализа, результаты которого подвержены серьезным колебаниям при изменении временных рамок анализа.

Для повышения эффективности описанного метода в рамках данной работы для поиска эффективной границы портфелей предлагается вместо ожидаемой доходности компании использовать интегральный показатель, представляющий собой комбинацию вероятности успеха компании в будущем, которая была предсказана алгоритмом стекинга, и ожидаемой доходности, рассчитанной на 3-летнем интервале. Рассматриваемая величина объединит в себе компоненты как предиктивного, так и ретроспективного анализов.

В качестве инструмента для объединения показателей ожидаемой доходности и предсказанной вероятности используется F-мера, которая представляет собой гармоническое среднее рассматриваемых показателей. Среднее гармоническое обладает важным свойством: оно близко к нулю, если хотя бы один из аргументов близок к нулю. Именно поэтому оно является более предпочтительным вариантом, чем среднее арифметическое. Формула расчета F-меры представлена ниже.

,

*Где:*

*– вероятность, предсказанная алгоритмом для бумаги i.*

Ввиду того, что вероятности, предсказанные алгоритмом, располагаются в интервале от 0,5 до 1, а ожидаемые доходности компаний, приблизительно, в интервале от 0 до 0,5, для повышения интерпретируемости получаемых результатов необходимо перевести рассматриваемые величины в однородные интервалы. Для этого удалим из анализа компании, ожидаемая доходность которых, рассчитанная на 3-х летнем интервале, ниже нуля, а для оставшихся компаний применим масштабирование текущих значений их показателей в интервал от 0.5 до 1. Для этого воспользуемся следующей формулой:

*Где:*

*– минимальное значение масштабируемой выборки*

*– максимальное значение масштабируемой выборки*

*– верхняя граница целевого интервала*

*– нижняя граница целевого интервала*

Получив целевые интервалы и рассчитав на них показатель F-меры, построим эффективную границу портфелей путем решения следующей системы уравнений:

*Где:*

*– целевое значение риска портфеля*

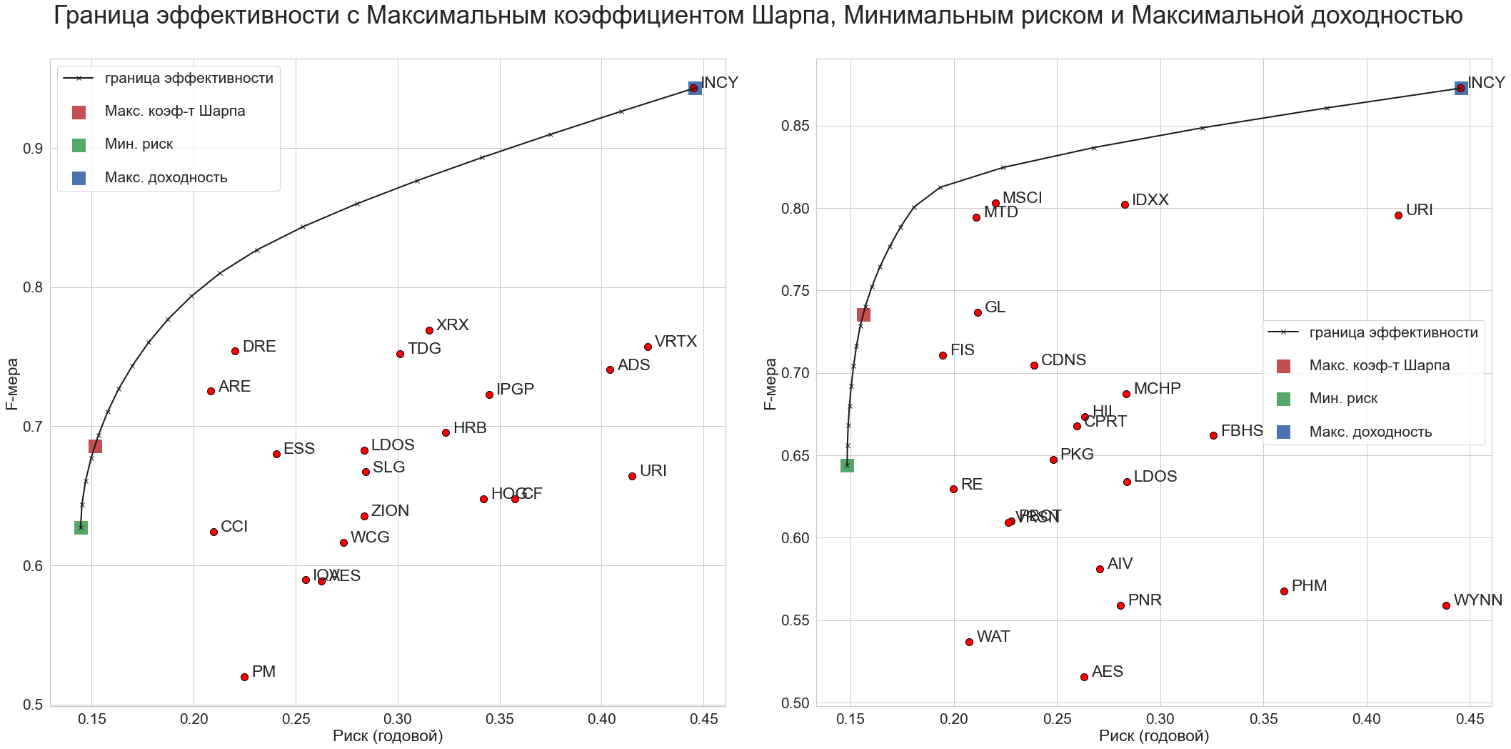
Так как система имеет бесконечно много решений, найдем пограничные точки границы эффективности и проложим между ними кривую, корректируя ее форму, рассчитывая промежуточные комбинации бумаг на границе эффективности. В итоге, получим нижеследующий результат. 

Рис. Эффективные границы портфелей за 2017 год (слева) и за 2016 год (справа)

Все множество портфелей, расположенное на границе эффективности, потенциально может быть использовано инвестором в качестве варианта для инвестирования. Выбор зависит только от отношения инвестора к риску. В данной работе выбор конкретного портфеля осуществляется на основе коэффициента Шарпа, исходя из принципа его максимизации. Коэффициент Шарпа рассчитывается по следующей формуле:

На рисунках, представленных ниже, отражено распределение весов акций в портфеле с максимальным коэффициентом Шарпа за 2017 и 2016 года.

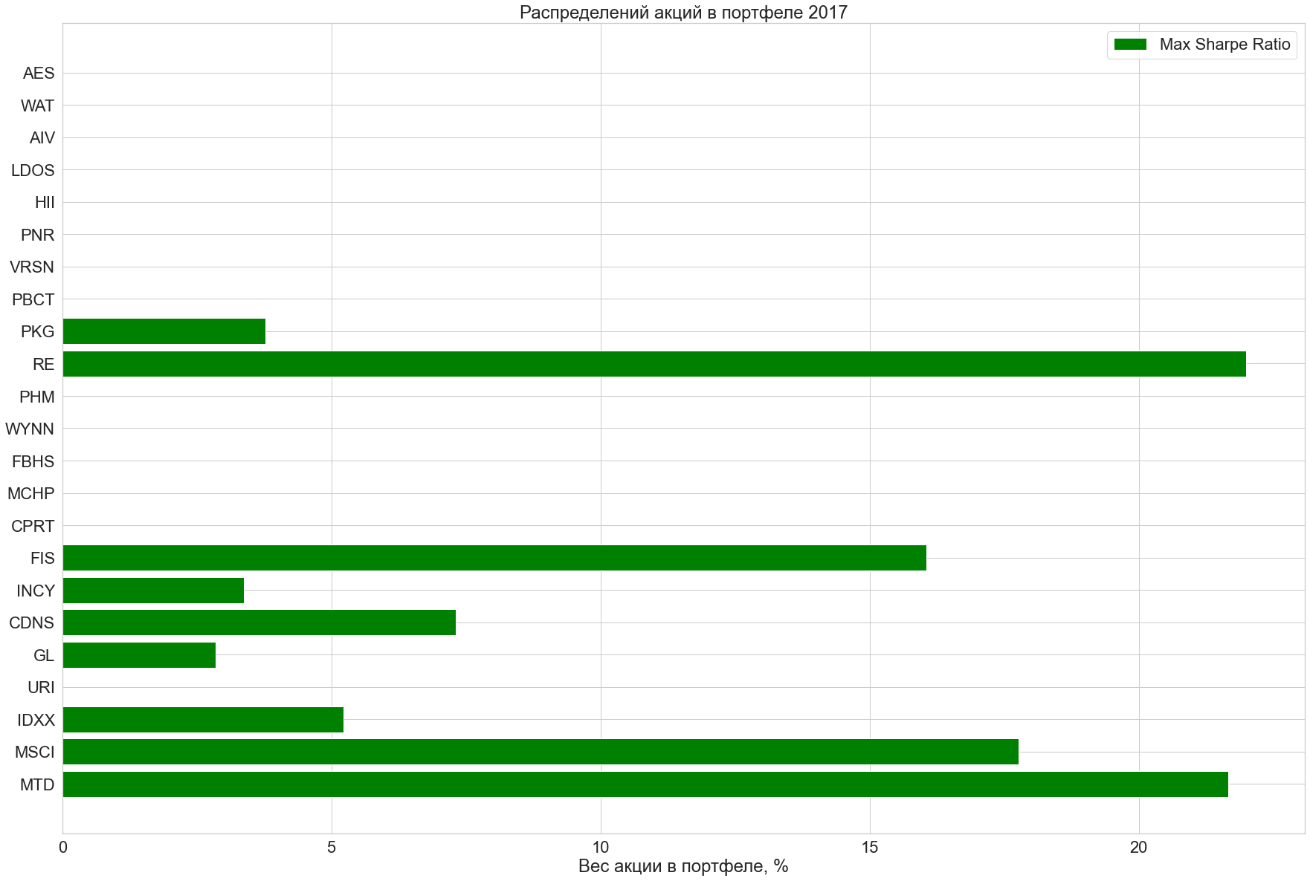


Рис. Распределение акций в портфеле Шарпа за 2017 год

Портфель Шарпа за 2017 год включается в себя 9 компаний. Доля каждой компании в портфеле не превышает 25%.

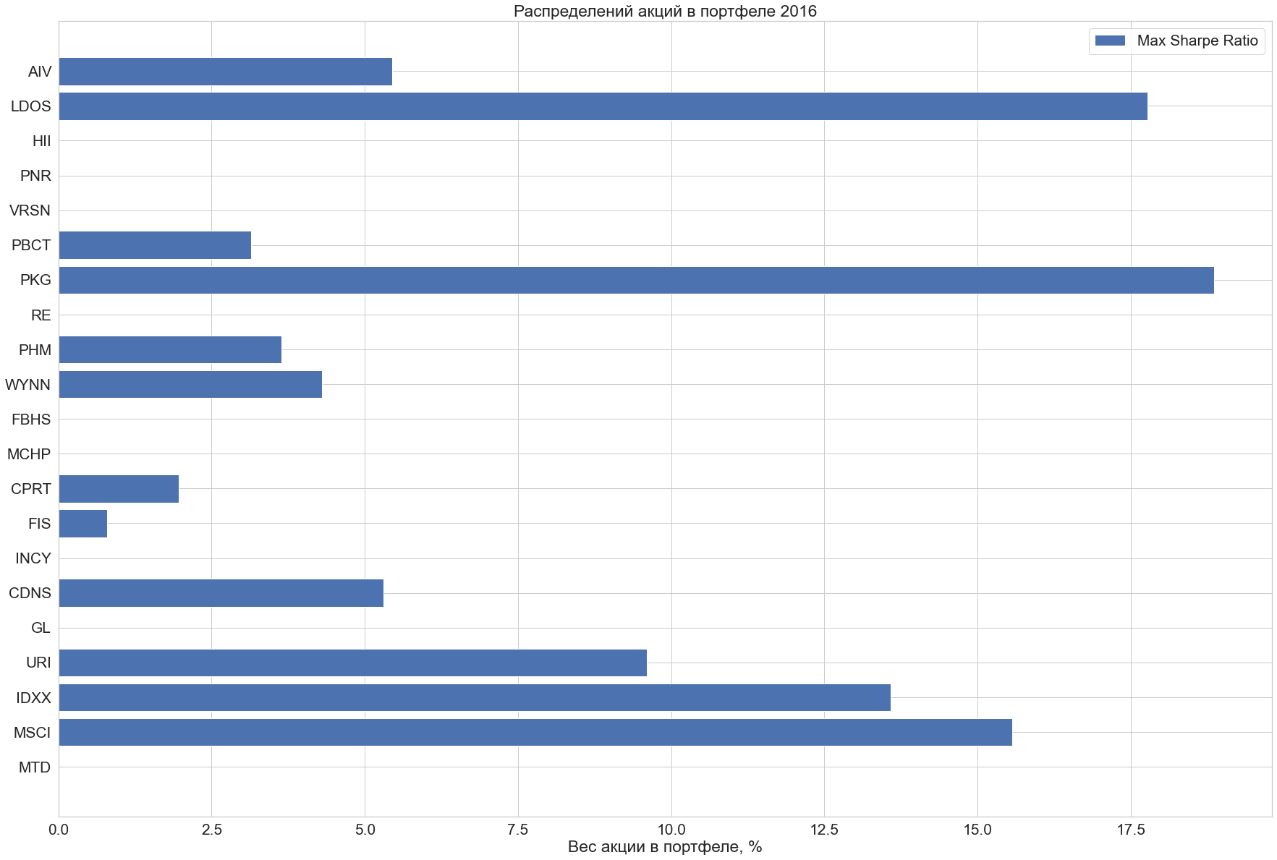


Рис. Распределение акций в портфеле Шарпа за 2016 год

Портфель Шарпа за 2016 год более диверсифицирован сравнении с 2017 годом. В портфель вошло 12 компаний. Доля каждой компании в портфеле не превышает 20%.

Далее протестируем доходность портфелей, составленных на основе модернизированной портфельной теории. Основная проблема предложенного алгоритма предсказания вероятностей, на базе которых выстраивается рейтинг компаний, заключается в том, что эффективный горизонт предсказания алгоритма составляет всего 6 месяцев. Следовательно, перед инвестором встает вопрос о том, что делать оставшиеся 6 месяцев до начала следующего цикла работоспособности алгоритма. В рамках данной работы производится реинвестирование средств в индекс S&P500, однако возможны и любые иные манипуляции. На рисунках ниже отражена динамика доходности портфелей, составленных на основе метода, описанного выше, за 2016 и 2017 года.

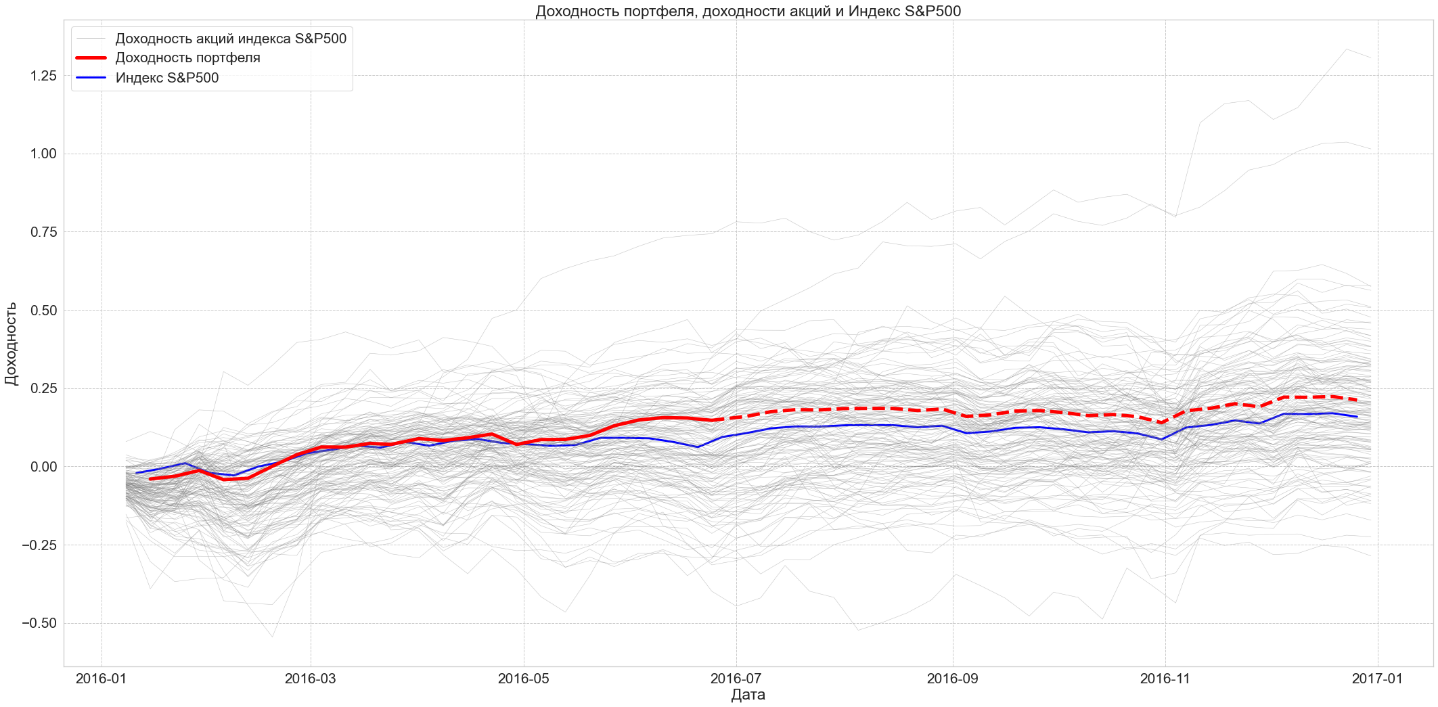


Рис. Динамика доходности портфеля за 2016 год

Из рисунка видно, что первые пять месяцев доходность индекса была приблизительно равна доходности портфеля. Однако за первую половину 2016 года доходность портфеля превысила доходность индекса на 4%. Годовая доходность после реинвестиций в индекс составила 24%.

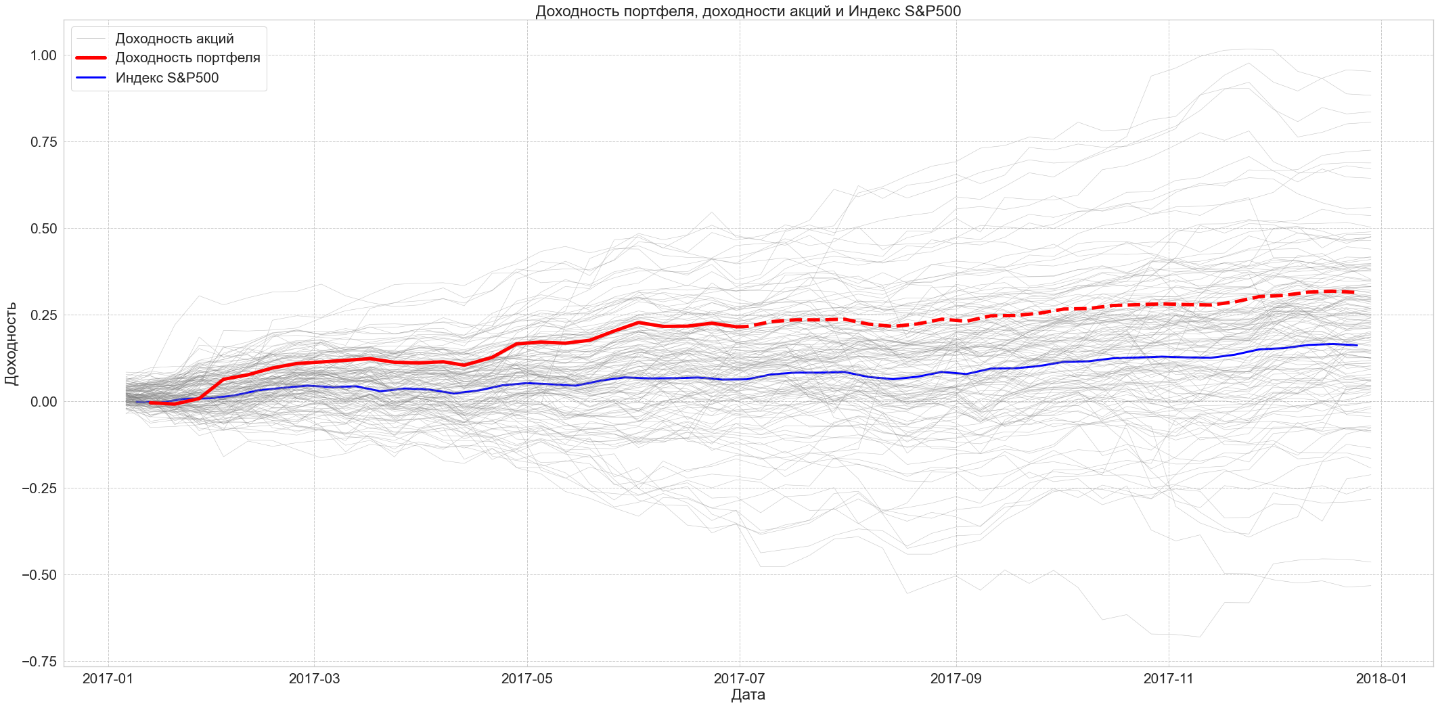


Рис. Динамика доходности портфеля за 2017 год

Доходность портфеля за первую половину 2017 года составила 25%, тем самым превысив доходность индекса более чем на 17%. Годовая доходность предложенной стратегии после реинвестиции в индекс составила 33%. Продемонстрированные результаты указывают на состоятельность и устойчивость разработанной стратегии.

## Приложения

Приложение 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Показатель** | **Число наблюдений** | **Медиана** | **Среднее** | **STD** |
| total\_assets | 7,459 | 11,334.00 | 52,181.06 | 184,136.58 |
| equity | 7,459 | 3,927.95 | 10,307.38 | 21,629.12 |
| ebit | 7,459 | 950.00 | 2,657.19 | 5,885.45 |
| Invested\_capital | 7,459 | 6,815.00 | 18,757.83 | 44,764.57 |
| other\_liabilities | 7,459 | 616.00 | 10,457.79 | 51,276.12 |
| net\_income | 7,459 | 509.00 | 1,465.08 | 3,695.05 |
| op\_inc\_after\_dep | 7,459 | 950.00 | 2,657.19 | 5,885.45 |
| sales | 7,459 | 6,470.60 | 17,677.30 | 35,807.99 |
| stockholders\_equity | 7,459 | 3,964.98 | 10,518.98 | 22,525.88 |
| STD | 7,456 | 202.57 | 5,352.10 | 34,689.95 |
| income\_taxes | 7,456 | 208.68 | 626.00 | 1,811.83 |
| Retained Earnings | 7,454 | 2,147.75 | 7,551.77 | 20,651.60 |
| EarningsYld | 7,452 | 4.62 | 2.27 | 102.92 |
| OperatingMargin | 7,452 | 15.95 | 3.45 | 579.61 |
| ROA | 7,452 | 5.23 | 5.59 | 11.09 |
| SalesYld | 7,452 | 51.61 | 80.64 | 110.62 |
| tangible\_equity | 7,451 | 1,468.37 | 4,903.44 | 16,970.20 |
| book\_value\_per\_share | 7,449 | 14.94 | 178.19 | 3,152.92 |
| preference\_stock | 7,449 | - | 211.88 | 1,895.75 |
| liabilities\_total | 7,435 | 6,657.00 | 41,489.55 | 166,250.71 |
| LTD | 7,433 | 2,223.04 | 7,964.63 | 24,986.44 |
| employees | 7,417 | 15.85 | 47.97 | 118.91 |
| dividends | 7,403 | 157.75 | 579.71 | 1,418.43 |
| acc\_receivables | 7,395 | 953.00 | 13,884.67 | 73,081.77 |
| SalesGrowth1Yr | 7,389 | 7.10 | 62.49 | 4,306.77 |
| op\_act\_net\_CF | 7,384 | 949.16 | 2,630.28 | 6,970.98 |
| acc\_payable | 7,383 | 568.76 | 14,410.58 | 88,211.86 |
| GrossMargin | 7,358 | 38.99 | 35.96 | 196.93 |
| COGS | 7,358 | 3,608.12 | 11,570.72 | 26,753.14 |
| Other\_exp | 7,358 | 1,175.50 | 3,523.24 | 7,783.20 |
| capex | 7,351 | 232.00 | 995.43 | 2,443.35 |
| inventories | 7,343 | 363.69 | 3,582.99 | 24,884.96 |
| PriceBook | 7,295 | 2.95 | 8.62 | 92.66 |
| ROE | 7,295 | 14.01 | 18.11 | 212.11 |
| cash | 7,254 | 569.65 | 2,264.00 | 7,161.80 |
| intangible\_assets | 7,187 | 1,273.00 | 5,595.74 | 14,213.34 |
| interests\_paid\_net | 7,042 | 123.13 | 560.81 | 2,707.20 |
| tax | 7,042 | 244.70 | 625.90 | 2,177.10 |
| marketable\_sec | 7,022 | - | 145.41 | 1,268.48 |
| op\_in\_before\_dep | 7,006 | 1,393.38 | 3,653.15 | 7,352.76 |
| amort | 6,958 | 294.73 | 873.08 | 1,969.51 |
| int\_exp | 6,858 | 128.68 | 452.56 | 1,993.41 |
| EpsGrowth1Yr | 6,756 | 11.11 | 24.28 | 1,002.87 |
| minority\_interests | 6,634 | - | 158.18 | 1,214.84 |
| deferred\_taxes\_fed | 6,471 | 93.70 | 327.60 | 830.00 |
| deferred\_taxes | 6,329 | 138.00 | 1,413.62 | 3,995.41 |
| noncontrolling\_interests | 6,307 | - | 357.12 | 2,082.76 |
| deferred\_taxes\_foreign | 6,274 | 28.69 | 251.08 | 1,179.31 |
| Implied Option Expense | 6,073 | - | 22.66 | 102.71 |
| mii | 6,061 | - | 41.54 | 314.47 |
| SalesGrowth5Yr | 5,991 | 38.54 | 480.32 | 21,285.29 |
| cur\_assets | 5,978 | 2,758.05 | 6,628.20 | 11,626.96 |
| working\_cap | 5,978 | 662.65 | 1,819.34 | 5,845.84 |

Таблица. Рейтинг компаний за 2017 год.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Компания** | **Сектор** | **Предсказанная вероятность** | **Доходность компании через 6 месяцев** | **Доходность сектора через 6 месяцев** | **rating** | **Верно?** |
| Mettler-Toledo Int. | Health Care | 66.4% | 22.9% | 8.3% | 10 | 1 |
| MSCI Inc. Class A | Financials | 65.1% | 18.7% | -0.6% | 10 | 1 |
| IDEXX Laboratories Inc. | Health Care | 64.9% | 6.2% | 8.3% | 10 | 0 |
| Globe Life Inc. | Financials | 63.0% | -2.0% | -0.6% | 10 | 0 |
| Cadence Design Systems | Information Technology | 62.3% | 24.5% | 13.7% | 10 | 1 |
| Incyte Corporation | Health Care | 61.6% | 11.4% | 8.3% | 10 | 1 |
| Fidelity National Information Services | Information Technology | 61.2% | 13.3% | 13.7% | 10 | 0 |
| Copart Inc. | Industrials | 60.3% | 39.8% | 3.8% | 10 | 1 |
| Fortune Brands Home & Security Inc. | Industrials | 59.8% | 7.6% | 3.8% | 10 | 1 |
| Wynn Resorts Limited | Consumer Discretionary | 58.7% | 37.5% | 3.4% | 10 | 1 |
| PulteGroup Inc. | Consumer Discretionary | 58.4% | 10.6% | 3.4% | 10 | 1 |
| Everest Re Group Ltd. | Financials | 58.3% | 7.5% | -0.6% | 10 | 1 |
| Packaging Corporation of America | Materials | 58.0% | 19.7% | 3.1% | 10 | 1 |
| People's United Financial | Financials | 57.9% | -5.3% | -0.6% | 10 | 0 |
| Cincinnati Financial Corporation | Financials | 57.7% | 4.7% | -0.6% | 10 | 1 |
| VeriSign Inc. | Information Technology | 57.5% | 20.6% | 13.7% | 10 | 1 |
| Pentair plc | Industrials | 57.5% | 6.5% | 3.8% | 10 | 1 |
| Huntington Ingalls Industries | Industrials | 57.1% | -3.5% | 3.8% | 10 | 0 |
| Leidos Holdings Inc. | Information Technology | 57.0% | 9.6% | 13.7% | 10 | 0 |

Таблица

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Компания** | **Сектор** | **Вероятность** | **Доходность компании через 6 месяцев** | **Доходность сектора через 6 месяцев** | **rating** | **Верно?** |
| Xerox Holdings Corporation | Information Technology | 72.59% | 2.51% | 14.19% | 10 | 0 |
| Alexandria Real Estate | Financials | 71.74% | 39.06% | 14.78% | 10 | 1 |
| Duke Realty Corporation | Financials | 71.71% | 32.13% | 14.78% | 10 | 1 |
| IPG Photonics Corporation | Information Technology | 71.25% | 5.84% | 14.19% | 10 | 0 |
| Harley-Davidson Inc. | Consumer Discretionary | 70.74% | 17.06% | 8.17% | 10 | 1 |
| Incyte Corporation | Health Care | 70.68% | 10.27% | 9.52% | 10 | 1 |
| Alliance Data Systems | Information Technology | 70.14% | -3.92% | 14.19% | 10 | 0 |
| Allegion PLC | Industrials | 68.86% | 12.06% | 11.80% | 10 | 1 |
| SL Green Realty Corp. | Financials | 68.17% | 29.16% | 14.78% | 10 | 1 |
| Twitter Inc. | Information Technology | 68.09% | 7.26% | 14.19% | 10 | 0 |
| Essex Property Trust | Financials | 68.01% | 8.30% | 14.78% | 10 | 0 |
| H&R Block Inc. | Consumer Discretionary | 67.84% | -6.31% | 1.67% | 10 | 0 |
| Zions Bancorporation N.A. | Financials | 67.59% | 36.44% | 14.78% | 10 | 1 |
| AES Corporation | Utilities | 67.56% | 20.74% | 9.68% | 10 | 1 |
| Amazon.com Inc. | Consumer Discretionary | 67.44% | 33.77% | 8.17% | 10 | 1 |
| Crown Castle | Financials | 67.43% | 10.26% | 14.78% | 10 | 0 |
| UDR Inc. | Financials | 67.31% | 7.11% | 14.78% | 10 | 0 |
| Vertex Pharmaceuticals | Health Care | 67.25% | 10.59% | 9.52% | 10 | 1 |
| United Rentals Inc. | Industrials | 66.78% | 41.32% | 11.80% | 10 | 1 |
| Leidos Holdings Inc. | Information Technology | 66.75% | 66.50% | 14.19% | 10 | 1 |
| Arista Networks Inc. | Information Technology | 66.62% | 16.47% | 14.19% | 10 | 1 |

1. (Karminsky, Polozov, 2016) [↑](#footnote-ref-1)
2. (Han, Pagano, Shin, 2012) [↑](#footnote-ref-2)
3. (Vlasov, Abrekov, 2018) [↑](#footnote-ref-3)
4. Тут чел из рейтингов [↑](#footnote-ref-4)
5. (Karminsky, Polozov, 2016). [↑](#footnote-ref-5)
6. Breiman, Leo. Random Forests // Machine Learning, 45(1), 5-32, 2001 [↑](#footnote-ref-6)
7. ОДС [↑](#footnote-ref-7)
8. Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978 [↑](#footnote-ref-8)
9. Breiman, Leo. Bagging predictors // Machine learning 24.2 (1996): 123-140 [↑](#footnote-ref-9)
10. Skurichina M., Duin R. P. W. Limited bagging, boosting and the random subspace method for linear classifiers // Pattern Analysis & Applications. 2002. Pp. 121–135. [↑](#footnote-ref-10)
11. Predicting Good Probabilities with Supervised Learning, A. Niculescu-Mizil & R. Caruana, ICML 2005 [↑](#footnote-ref-11)
12. On the combination of forecast probabilities for consecutive precipitation periods. Wea. Forecasting, 5, 640–650., Wilks, D. S., 1990a [↑](#footnote-ref-12)
13. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods

    John C. Platt, ADVANCES IN LARGE MARGIN CLASSIFIERS, 1999 [↑](#footnote-ref-13)
14. <https://scikit-learn.org/stable/index.html> [↑](#footnote-ref-14)
15. Tianqi Chen, Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System [↑](#footnote-ref-15)
16. <https://catboost.ai/> [↑](#footnote-ref-16)
17. <https://lightgbm.readthedocs.io/en/latest/> [↑](#footnote-ref-17)
18. Jahrer, Michael. Netflix Prize report 2009 // http://elf-project.sourceforge. net/CombiningPredictionsForAccurateRecommenderSystems.pdf [↑](#footnote-ref-18)
19. <http://www.machinelearning.ru/wiki/images/5/56/Guschin2015Stacking.pdf> [↑](#footnote-ref-19)
20. A nested cross-validation procedure provides an almost unbiased estimate of the true error. (Varma and Simon 2006) [↑](#footnote-ref-20)
21. <https://www.msci.com/gics> [↑](#footnote-ref-21)
22. https://github.com/gracikk-finance/gracikk [↑](#footnote-ref-22)
23. Fama E.F. Efficient capital markets: a review of theory and empirical work // J. Finance. 1970. Т. 25. № 2. С. 383–417 [↑](#footnote-ref-23)
24. (Graham, Harvey, & Rajgopal, 2005) Сюрпрайз [↑](#footnote-ref-24)
25. (Bernard & Thomas, 1990). [↑](#footnote-ref-25)
26. D. E. Rapach “Short interest and aggregate stock returns,” [↑](#footnote-ref-26)
27. Теплова новое в фин аналитике [↑](#footnote-ref-27)
28. Ссылка на API [↑](#footnote-ref-28)
29. *de Boor, Carl.* A Practical Guide to Splines. — New York: Springer-Verlag, 1978. [↑](#footnote-ref-29)
30. [*Runge, Carl*](https://en.wikipedia.org/wiki/Carl_David_Tolm%C3%A9_Runge) (1901), "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten", Zeitschrift für Mathematik und Physik, **46**: 224–243. available at [www.archive.org](https://archive.org/details/zeitschriftfrma12runggoog) [↑](#footnote-ref-30)
31. (Теплова 2012) [↑](#footnote-ref-31)
32. http://www.sternstewart.com/?content=proprietary&p=cov) [↑](#footnote-ref-32)
33. (Ивашковская, Животова) [↑](#footnote-ref-33)
34. Колби, Роберт. Энциклопедия технических индикаторов рынка = The Encyclopedia of Technical Market Indicators. — М.: Альпина Паблишер, 2011. — 840 с. — ISBN 978-5-9614-1443-1. [↑](#footnote-ref-34)
35. Barron et al. 2005 [↑](#footnote-ref-35)
36. J.Financ.58,1393–1413. Granville,J.E.,1963.Granville'sNewKeytoStockMarketProfits.PrenticeHall,NewYork. [↑](#footnote-ref-36)
37. sklearn [↑](#footnote-ref-37)
38. Модель шап ссылка [↑](#footnote-ref-38)