Содержание

Содержание

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 между этими концепциями 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 к оценке компании или иной сущности.[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] 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. Вполне очевидно, что оценки вероятности банкротства выставляются, исходя из анализа количественных показателей деятельности

[1] (Karminsky, Polozov, 2016)
[2] (Han, Pagano, Shin, 2012)

рейтингуемой компании. Существует достаточное количество исследований, цель которых заключается в разработке методологии определения вероятности банкротства компании на основе количественных показателей, например, **блаббла.** Результаты исследований, представленные в перечисленных работах, полезны для построения методологии рейтинговой оценки инвестиционной привлекательности компаний несмотря на то, что решаемые задачи, в целом, диаметрально противоположны. Вместо оценки вероятности дефолта компании предпринимается попытка составить рейтинг, который бы отражал ожидаемые перспективы компаний.

Разберем существующие подходы к конструированию 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]

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.

[3] (Vlasov, Abrekov, 2018)

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

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

Формализуем основные понятия, используемые при работе с ними.

Участник рейтинговой системы – объект, подлежащий оценке в данной рейтинговой системе. Участников системы принято нумеровать с помощью целочисленных индексов.

Фактор – неотрицательное число, характеризующее один из аспектов (свойств, показателей полезности) участника рейтинговой системы, в заданном числовом диапазоне. При работе с моделями составного рейтинга предполагается, что увеличение значения фактора соответствует увеличению полезности (ценности) участника рейтинговой системы и наоборот, увеличение полезности ведет к увеличению фактора участника системы. В дальнейшем, будем полагать, что каждый участник рейтинговой системы, характеризуется одинаковым числом факторов n и будем в дальнейшем обозначать j-ый индикатор участника с номером i через X_{ij}.

Учитывая, что природа влияния фактора 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.

Целевая функция - правило F, по которому для участника рейтинговой системы i на основании учета значений всех n его факторов приписывается некоторое неотрицательное число. Очевидно, функция F является функцией n аргументов. Задача функции состоит в том, чтобы привести совокупность из n

[4] Тут чел из рейтингов

индикаторов к одному числу b с целью его последующего сравнения с такими же числами, относящимися к другому объекту исследования.

К функции полезности необходимо предъявить следующие требования:

1. Целевая функция должна быть неотрицательной, т.е.

$$F(Y_{i1}, \dots, Y_{in}) \geq 0, \forall n$$

2. В нулевой точке, т.е. когда все аргументы одновременно равны нулю, целевая функция должна быть равна нулю:

$$F(0_{i1}, \dots, 0_{in}) = 0$$

3. Для моделей составного рейтинга также должно выполняться требование возрастания функции полезности по все аргументам функции.

$$\frac{dF}{dY_{ij}} > 0,$$

Рейтинг – неотрицательное число b_i являющееся результатом применения целевой функции к совокупности всех индикаторов участника номер i рейтинговой системы. Таким образом, рейтинг определяется на основании соотношения:

$$b_i = F(Y_i) = F(Y_{i1}, \dots, Y_{in})$$

Определив понятийную базу, рассмотрим более подробно основы составных рейтингов.[5] Идея составного рейтинга заключается в следующем: оценивается ранг объектов по каждому фактору отдельно, а затем на основе полученных значений вычисляется интегральный рейтинг. Этот подход является комплексным, поскольку он дает нам возможность создавать динамический рейтинг. 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:

[5] (Karminsky, Polozov, 2016).

$$D(x_{k,j}) = \frac{2}{13}\sum_{t=1}^{12}\frac{t}{12}\frac{x_{k,j}^{t}}{\sum_{i=1}^{N} x_{i,j}^{t}}$$

$D(x_{k,j})$ – модифицированное значение фактора j на объекте k

$x_{k,j}^{t}$ – значение фактора j на объекте k в момент времени t

Полученная 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 складываем модифицированные значения факторов D и получается вектор интегральных показателей:

$$Int_k = \sum_{j=1}^{L} D(x_{k,j})$$

Int_k – 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.

Кроме описанных выше недостатков этого метода, подход составных рейтингов не подходит для целей данной работы еще потому, что он основан на относительных, а не абсолютных значениях. Это означает, что данный подход не позволяет использовать качественные факторы, превращенные в бинарные переменные.

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 нашей выборке иметь много двоичных переменных и не иметь существующего рейтинга, присвоенного рейтинговым агентством.

Далее более подробно разберем методы решения поставленной задачи, а также варианты ее верификации.

Красивый цвет!!

```
colormap = plt.cm.RdBu
plt.figure(figsize=(14,12))
plt.title('Pearson Correlation of Features', y=1.05, size=15)
sns.heatmap(train.astype(float).corr(),linewidths=0.1,vmax=1.0,
            square=True, cmap=colormap, linecolor='white', annot=True)
```

https://www.kaggle.com/arthurtok/introduction-to-ensembling-stacking-in-python

https://medium.com/@Gafarov/new-opportunities-in-machine-learning-c62b93ccd064

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.

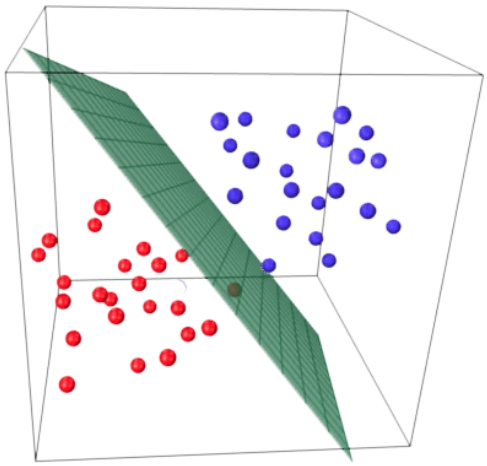


Рис. Визуализация работы линейного классификатора.

Формализуем задачу линейной классификации. Пусть $\mathbb{Y} = \{-1,1\}$ - пространство ответов, $\mathbb{X}$ – пространство объектов, $x^1, \dots, x^d$ – признаковое описание объекта, $\mathbb{X} = (x_i y_i)_{i=1}^{l}$ – обучающая выборка.

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

$$Q(w, X) = \frac{1}{l}\sum_{i=1}^{l}(\langle w, x_i\rangle - y_i)^2$$

В результате получается вектор весов:

$$w = argmin_w Q(w, X)$$

Откуда выходит, что если $\langle w, x_i\rangle > 0$, то объект относится к классу 1, иначе к классу -1. Иными словами получается следующее выражение:

$$a(\vec{x}) = sign(\vec{w}^T x),$$

Где:

- $a(\vec{x})$ – 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:

$$g : (0,1) \mapsto \mathbb{R}$$

В таком случае, можно решать задачу линейной регрессии, в которой строится оценка не для условного матожидания $E(y|x)$, а для $g(E(y|x))$.

$$g\big(E(y|x)\big) \approx \langle w, x \rangle$$

$$E(y|x) \approx g^{-1}(\langle w, x \rangle)$$

В статистике такое семейство моделей называется обобщенными линейными моделями. В задаче бинарной классификации в качестве g^{-1} часто используется сигмоида.

$$g^{-1}(\langle w, x \rangle) = \frac{e^{\langle w,x \rangle}}{1 + e^{\langle w,x \rangle}}$$

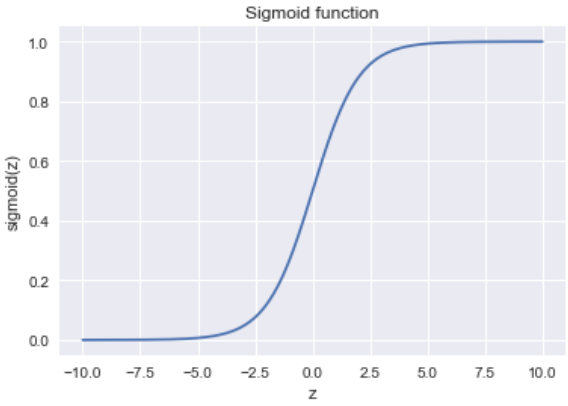


Рис. Функция сигмоида

Выбор функции сигмоида в качестве функции g обусловлен нижеследующим. Пусть $P_+(y_i = 1|\vec{x}_i, \vec{w})$ – условная вероятность события $y_i = 1$, а $OR_+(\mathrm{X})$ – отношение вероятностей этого события, которое определяется следующим образом:

$$OR_+(y_i = 1|\vec{x}_i, \vec{w}) = \frac{P_+(y_i = 1|\vec{x}_i, \vec{w})}{1 - P_+(y_i = 1|\vec{x}_i, \vec{w})}$$

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. $\log OR_+(y_i = 1|\vec{x}_i, \vec{w}) \in \mathbb{R}$. Следовательно, если

$$\log OR_+(y_i = 1|\vec{x}_i, \vec{w}) = \langle w^T x \rangle$$

$$OR_+(y_i = 1|\vec{x}_i, \vec{w}) = e^{\langle w^T x \rangle}$$

То:

$$P_+(y_i = 1|\vec{x}_i, \vec{w}) = \frac{OR_+}{1 + OR_+} = \frac{e^{\langle w^T x \rangle}}{1 + e^{\langle w^T x \rangle}} = g^{-1}(\langle w, x \rangle)$$

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:

$$\mathcal{L}_{log}(X, \vec{y}, \vec{w}) = \sum_{i=1}^{l} \log\left(1 + e^{-y_i w^T x_i}\right)$$

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.

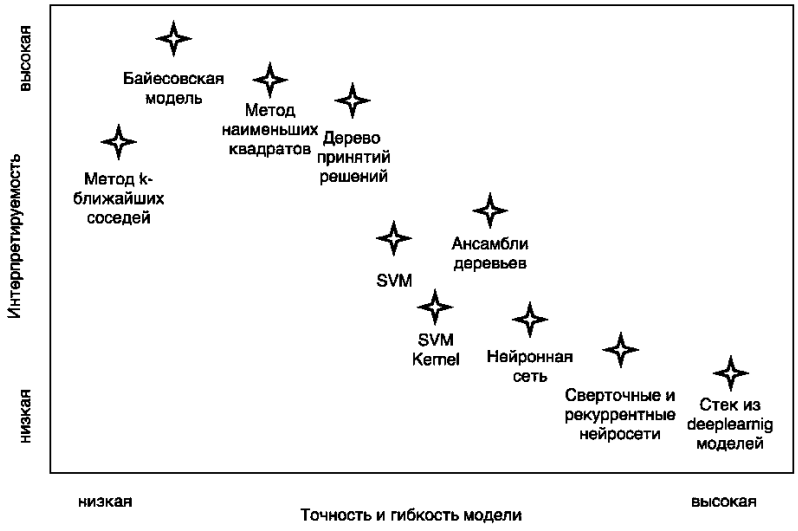


Рис. Зависимость гибкости алгоритма машинного обучения и интерпретируемости полученной модели.

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].

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:

[6] Breiman, Leo. Random Forests // Machine Learning, 45(1), 5-32, 2001

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Где:

p_i *– 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:

$$IG(Q) = S_0 - \sum_{i=1}^{q} \frac{N_i}{N} S_i$$

Где:

S_0 *– значение показателя энтропии до разбиения выборки по признаку* Q

N_i *– число объектов в подвыборке* i *после разбиения выборки*

S_i *– 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].

Одной из главных проблем деревьев принятия решений является их нестабильность, которая приводят к тому, что незначительные изменения в

[7] ОДС

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] 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.

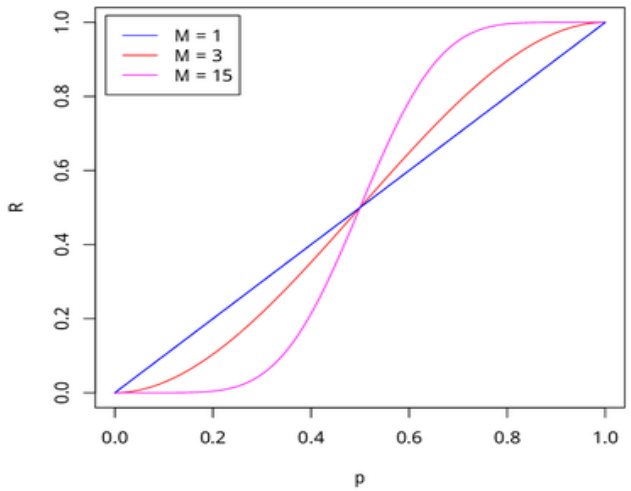
$$R = \sum_{i=m}^{M} C_M^i p^i (1-p)^{M-i}$$

Где:

M – количество членов жюри присяжных

p – вероятность верного решения конкретного эксперта

m – минимальное большинство членов жюри



[8] Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978

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]. 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 $X_1, \dots, X_N$ новых выборок, на основе которых можно оценивать различные статистики исходного распределения. Графически описанный процесс представлен на рисунке ниже.

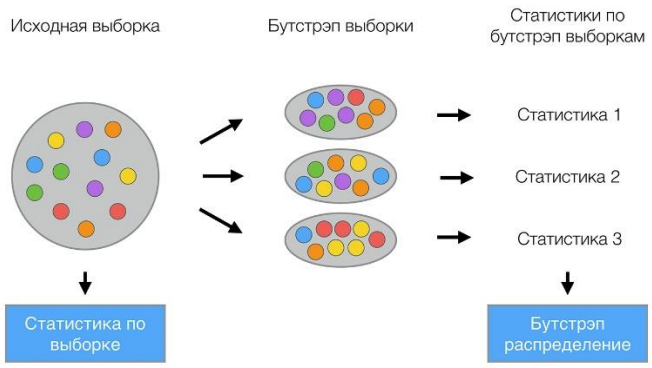


Рис. Иллюстрация инструмента бутсрэп

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.

[9] Breiman, Leo. Bagging predictors // Machine learning 24.2 (1996): 123-140

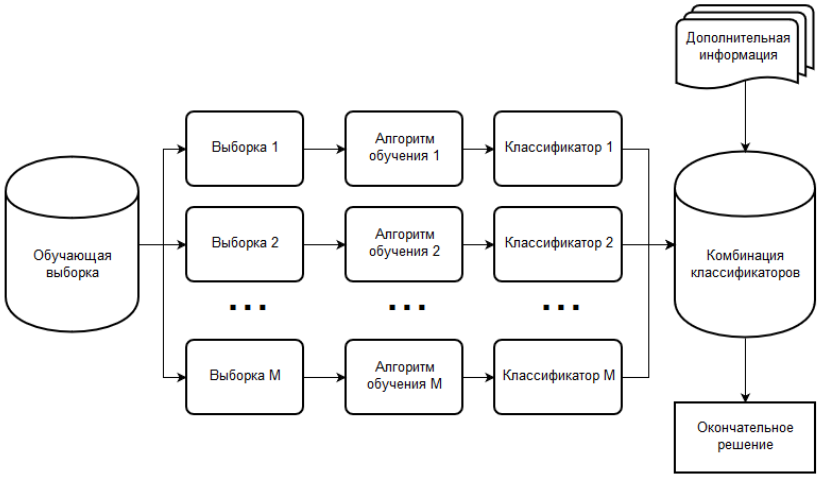


Рис. Графическая реализация алгоритма бэггинга

Преимущество бэггинга перед решающими деревьями можно показать, используя следующие выкладки. Допустим, решается задача регрессии и при этом имеется $b_1, \dots, b_N$ базовых алгоритмов, настроенных на решение поставленной задачи. Также представим, что существует истинная зависимость, которая выражается с помощью функции $y(x)$, тогда ошибку каждого алгоритма на объектах можно записать в следующей форме:

$$\varepsilon_i(x) = b_i(x) - y(x), i = 1, \dots, n$$

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

$$E_x(b_i(\mathrm{x}) - \mathrm{y}(\mathrm{x}))^2 = E_x\varepsilon_i^2(x)$$

Средняя ошибка полученных алгоритма имеет вид:

$$E_{alg} = \frac{1}{n}E_x\sum_{i=1}^{n}\varepsilon_i^2(x)$$

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:

$$a(x) = \frac{1}{n}\sum_{i=1}^{n} b_i(x)$$

То среднеквадратичная ошибка в этом случае примет следующий вид:

$$E_{bag} = E_x\,(1/n \sum_{i=1}^{n} b_i(x) - y(x))^2 = E_x\left(1/n \sum_{i=1}^{n} \varepsilon_i\right)^2 =$$

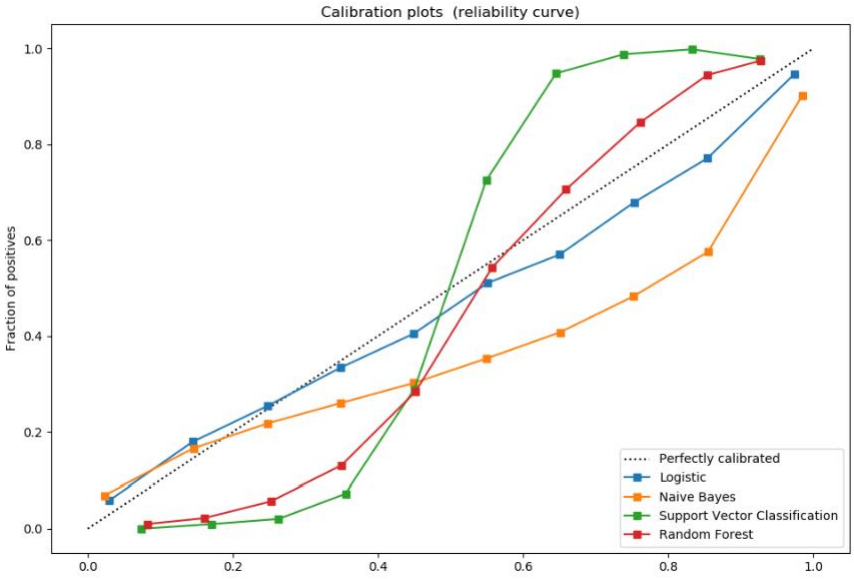
$$= 1/n^2\, E_x\left(\sum_{i=1}^{n} \varepsilon_i^2(x) + \sum\nolimits_{i\neq j} \varepsilon_i(x)\varepsilon_j(x)\right) = \frac{1}{n^2} E_x \sum_{i=1}^{n} \varepsilon_i^2(x) = \frac{1}{n} E_{alg}$$

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]. 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.

[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.

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:



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] 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

[11] Predicting Good Probabilities with Supervised Learning, A. Niculescu-Mizil & R. Caruana, ICML 2005

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], 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 Платта13 и непараметрический подход, основанный на изотонической регрессии. Оба этих подход реализованы в библиотеке scikit learn[14]

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

---

[12] On the combination of forecast probabilities for consecutive precipitation periods. Wea. Forecasting, 5, 640–650., Wilks, D. S., 1990a

[13] Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods John C. Platt, ADVANCES IN LARGE MARGIN CLASSIFIERS, 1999

[14] https://scikit-learn.org/stable/index.html

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.

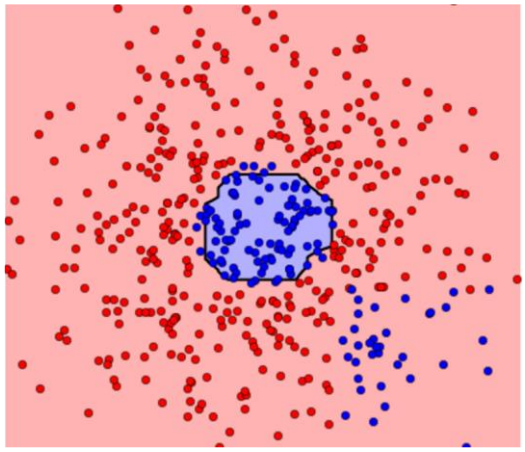


Рис. Разделяющая поверхность, которую строит случайный лес, ограниченный по глубине

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 $L(y, z)$, где y – истинный ответ, z – прогноз алгоритма на некотором объекте. В задаче классификации примером функции потерь может быть рассмотренная выше логистическая функция потерь:

$$L(y, z) = \log(1 + \exp(-yz))$$

В начале построения композиции по методу градиентного бустинга нужно ее инициализировать, то есть построить первый базовый алгоритм $b_1(x)$. 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.

$$b_1(x) = argmax_{y \in \mathbb{Y}} \sum_{i=1}^{l} [y_i = y]$$

Обучение базовых алгоритмов происходит последовательно. Пусть к некоторому моменту обучены $N-1$ алгоритмов $b_1(x), \dots, b_{N-1}(x)$, то есть композиция имеет вид:

$$a_{N-1}(x) = \sum_{i=1}^{N-1} b_i(x)$$

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

$$\sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + b(x_i)) \to \min_b$$

Сначала имеет смысл решить более простую задачу: определить, какие значения $s_1, \dots, s_l$ должен принимать алгоритм $b_N(x_i) = s_i$ на объектах обучающей выборки, чтобы ошибка на обучающей выборке была минимальной:

$$F(s_i) = \sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + s_i) \to \min_s$$

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

$$s = -\nabla \mathrm{F} = \begin{pmatrix} -L'_z(y_1, a_{N-1}(x_1)) \\ \dots \\ -L'_z(y_l, a_{N-1}(x_l)) \end{pmatrix}$$

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

Обучение $b_N(x)$, таким образом, представляет собой задачу обучения на размеченных данных, в которой $\{(x_i, s_i)\}_{i=1}^{l}$ – обучающая выборка, и используется, например, квадратичная функция потерь:

$$b_N(x) = argmin_b \frac{1}{l} \sum_{i=1}^{l} (b(x_i) - s_i)^2$$

Следует обратить особое внимание на то, что информация об исходной функции потерь $L(y, z)$, которая не обязательно является квадратичной, содержится в выражении для вектора оптимального сдвига s. Поэтому для большинства задач при обучении $b_N(x)$ 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]

CatBoost – библиотека для метода машинного обучения, основанная на градиентном бустинге.[16]

LightGBM – открытая программная библиотека, разработанная компанией Яндекс.[17]

Калибровка прогнозных вероятностей, предсказанных рассмотренными реализациями алгоритма, осуществляется аналогично алгоритму случайного

[15] Tianqi Chen, Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System
[16] https://catboost.ai/
[17] https://lightgbm.readthedocs.io/en/latest/

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 (X_0, Y_0) – обучающая выборка, $\mathbb{A}$ – базовый классификатор, использующийся для построения метапризнака. $\mathbb{A}.fit(X, Y)$ – функция обучения классификатора $\mathbb{A}$ на (X, Y). $\mathbb{A}.predict(X)$ – функция, которая предсказывает метку класса для X классификатором $\mathbb{A}$. $\mathbb{M}$ – некоторый метаклассификатор. $\mathcal{MF}(X, A)$ – метапризнак, полученный классификатором $\mathbb{A}$ для выборки X. P – финальное предсказание стекинга для валидационной выборки.

В простейшем виде получение предсказания для тестовой выборки P с помощью стекинга выглядит следующим образом:[18]

[18] Jahrer, Michael. Netflix Prize report 2009 // http://elf-project.sourceforge.net/CombiningPredictionsForAccurateRecommenderSystems.pdf

Алгоритм 1

разбить обучающую выборку (X, Y) на две части: (X_1, Y_1) и (X_2, Y_2).

A.fit(X_1, Y_1)
$MF(X_2, A) :=$ A.predict(X_2)
$MF(X_0, A) :=$ A.predict(X_0)

M.fit(concatV$(X_2, MF(X_2, A)), Y_2)$
$P := M$.predict(concatV$(X_0, MF(X_0, A)))$

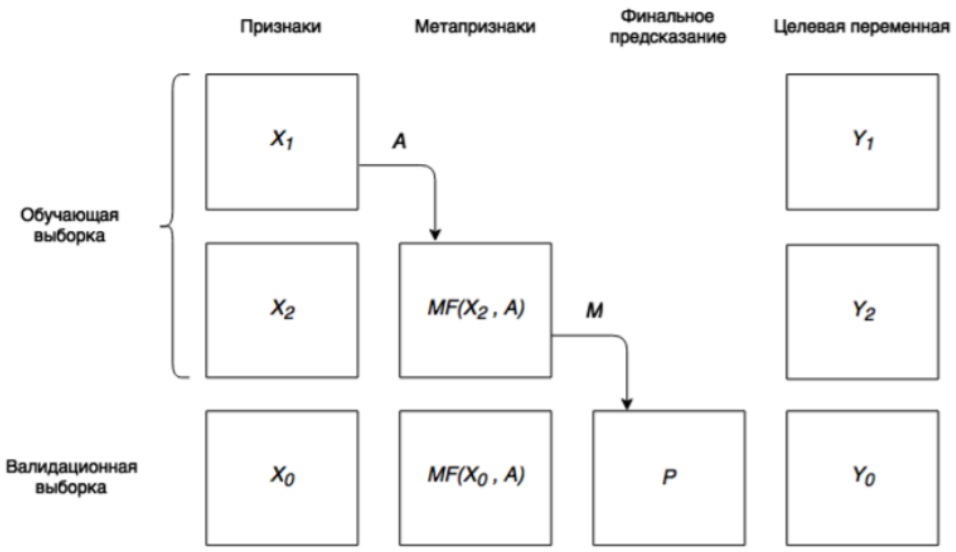


Рис. Алгоритм 1, стекинг по схеме hold-out[19]

Недостатком этого варианта стекинга является то, что $\mathbb{M}$ обучается только на части X_2 обучающей выборки, а другая часть X_1 им не используется. Чтобы избежать этого, мы можем повторить алгоритм, поменяв местами (X_1, Y_1) и (X_2, Y_2). В таком случае получим два предсказания для валидационной выборки, которые можно усреднить. В качестве метаклассификатора будет использоваться алгоритм градиентного бустинга, а именно XGBoost. 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.

[19] http://www.machinelearning.ru/wiki/images/5/56/Guschin2015Stacking.pdf

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 t переводится в бинарный ответ:

$$a(x) = [b(x) > t]$$

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

$$\text{accuracy} = \frac{\sum_{i=1}^{l}[a(x_i) = y_i]}{l}$$

Данная метрика, однако, имеет существенный недостаток. Если взять порог t меньше минимального значения прогноза $b(x)$ 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 $t = \max_i b(x_i)$ мы получим долю правильных ответов 0.95. Это означает, что доля положительных ответов сама по себе не несет никакой информации о качестве работы алгоритма $a(x)$, 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

	$y = 1$	$y = 0$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = 0$	False Negative (FN)	True Negative (TN)

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

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

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:

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

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 $a(x) = [b(x) > t]$ при конкретном значении порога t. Однако зачастую интерес представляет

лишь вещественный алгоритм $b(x)$, а порог выбирается позже в зависимости от требований к точности и полноте. В таком случае возникает потребность в измерении качества семейства моделей $\{ a(x) = [b(x) > t], t \in R\}$.

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):

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{TP + FN}$$

Каждый возможный выбор порога t соответствует точке в этом пространстве. Всего различных порогов имеется $\ell + 1$. Максимальный порог $t_{max} = \max_i b(x_i)$ даст классификатор с TPR = 0, FPR = 0. Минимальный порог $t_{min} = \max_i b(x_i) - \varepsilon$ 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 $b(x_1) - \varepsilon, b(x_1), \dots, b(x_l)$. Площадь под данной кривой называется AUC-ROC, и принимает значения от 0 до 1. Если порог t может быть подобран так, что алгоритм $a(x)$ не будет допускать ошибок, то AUC-ROC будет равен единице; если же $b(x)$ 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 $b(x)$.

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:

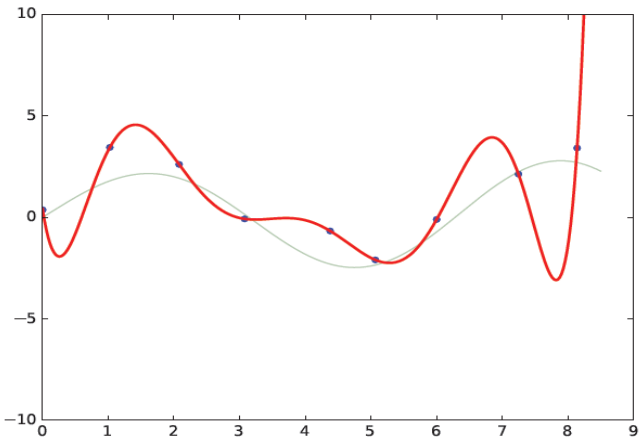


Рис. Переобученный классификатор

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.

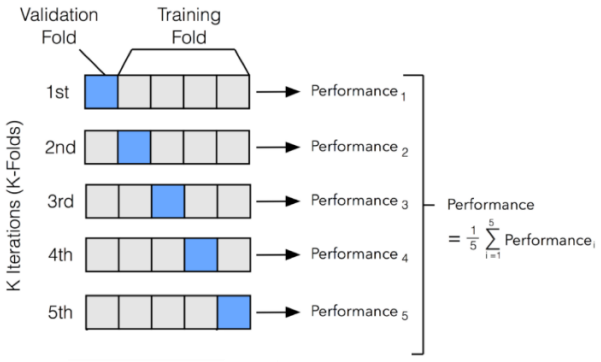
Выявить переобучение, используя только обучающую выборку, невозможно, поскольку и хорошо обученный, и переобученный алгоритмы будут хорошо ее описывать. Необходимо использовать дополнительные данные.

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.

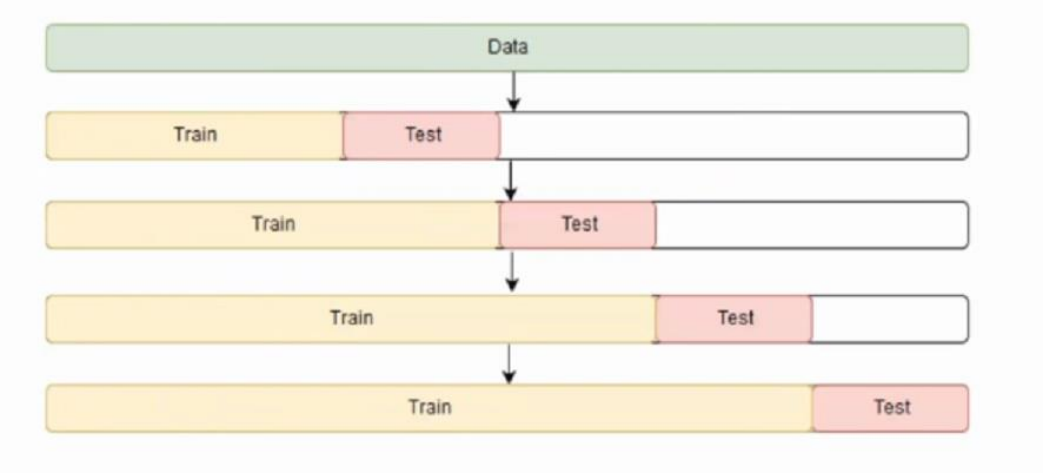
Самый простой способ оценить качество алгоритма – использование отложенной выборки. В этом случае следует разбить выборку на две части: первая из двух частей будет использоваться для обучения алгоритма, а вторая, тестовая выборка, – для оценки его качества, в том числе для нахождения доли ошибок в задаче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



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] (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 $q_{l/R}$, где R – количество равномерных градаций в интервале [0,1], а l – номер

[20] A nested cross-validation procedure provides an almost unbiased estimate of the true error. (Varma and Simon 2006)

интервала. Эти квантили делят интервал на R частей, которые определяются следующим образом:

$$D_1 = [0, q_{1/R}), D_2 = [q_{1/R}, q_{2/R}), \dots, D_R = [q_{R-1/R}, 1)$$

Для каждого объекта балльная 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.

Глава 2. Эмпирические результаты рейтингового ранжирования компаний

2.1 Описание выборки

Основой исследования является датасет, который состоит из ключевых финансовых показателей компаний, входящих в индекс S&P500. S&P500 – фондовый индекс, формирующийся из 500 публичных компаний, торгующихся на фондовых биржах США, таких как NASDAQ и Фондовая биржа Нью Йорка. Критериями отбора компаний в индекс служат их капитализация и ликвидность. 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]). 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, специализирующимися 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.

[21] https://www.msci.com/gics

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 автомобильной индустрии, ресторанный бизнес и другие 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 данного 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.

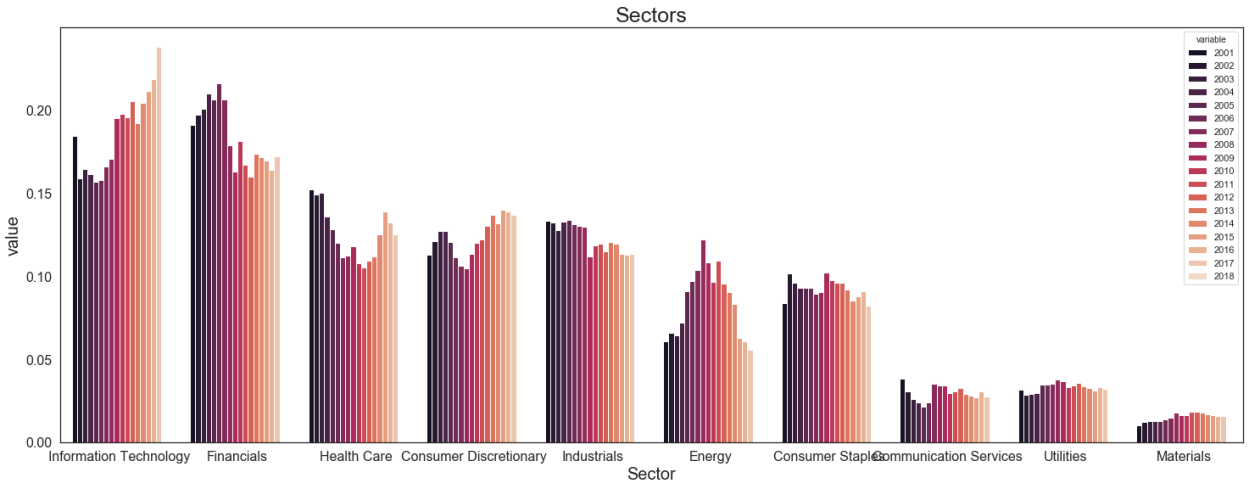


Рис. 1 Гистограмма распределения долей секторов в индексе S&P500

Гистограмма показывает, что наименьший вклад в индекс вносят такие сектора, как Communication services, Materials, Utilities. Их совокупное присутствие в индексе оставалось почти 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 2000х 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.

Для целей работы полезно понимать не только распределение капитализаций компаний, но и их количества по секторам. Гистограмма ниже отражает среднее количество компаний, которое приходится на сектор, за рассматриваемый период.

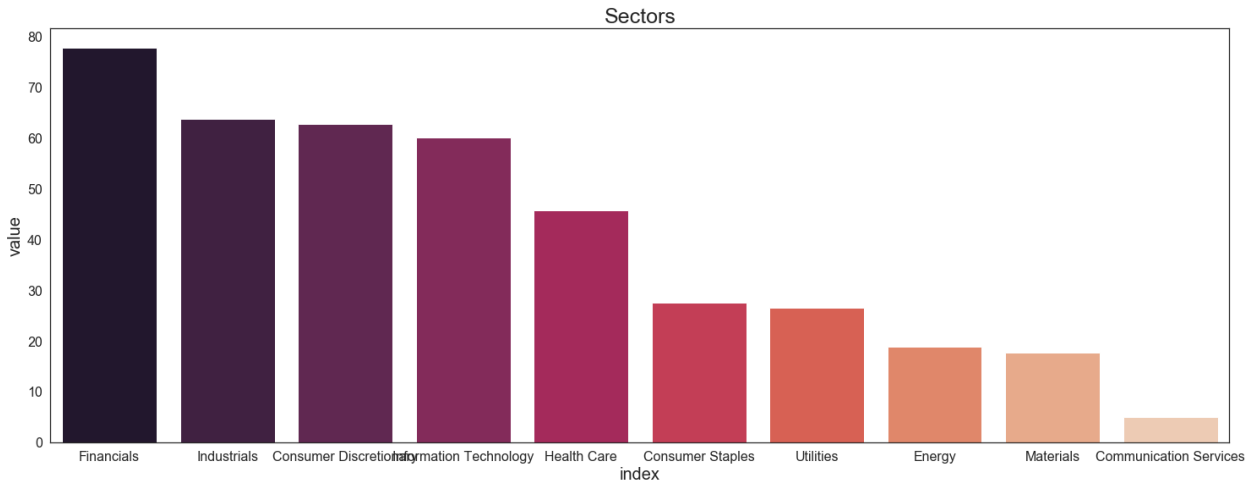


Рис 2. Гистограмма распределения компаний по секторам в среднем за рассматриваемый период.

Сравнивая, данные, представленные на рисунке 1 и на рисунке 2, можно прийти к следующим выводам:

1. Финансовые сектор один из крупнейших секторов экономики США, как с точки зрения доли капитализации компаний этого сектора в индексе S&P500, так и с точки зрения количества компаний, которые представляют этот сектор в индексе.

2. Компании, представляющие сектор информационных технологий, имеют наибольшую долю капитализации в индексе, однако их количество в индексе сравнительно невелико. Это говорит о том, что, в среднем, компания, которая относится к сектору информационных технологий имеет капитализацию выше, нежели любая другая компания.

3. Иная ситуация складывается для компаний индустриального сектора. Они обладают низкой капитализацией, однако в выборке их существенное количество.

Далее проведем предварительный анализ факторов, которые будут использоваться для построения регрессионных моделей. В целях исследования было принято решение разделить изучаемые факторы на три контура:

1. Контур внутренних факторов;

2. Контур макро факторов;
3. Контур новостного фона и технических показателей.

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

В качестве инструментов анализа в работе используются GitHub[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 Предобработка рыночных доходностей компаний S&P500

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, на сколь-либо длительном временном интервале. Очевидно, что подобная задача идет в разрез с гипотезой эффективного рынка[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

[22] https://github.com/gracikk-finance/gracikk
[23] Fama E.F. Efficient capital markets: a review of theory and empirical work // J. Finance. 1970. Т. 25. № 2. С. 383–417

зависит от публикуемой ими отчетности и, в частности, от их доходов. К тому же менеджеры 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] 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]. 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] 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] Показатель рассчитывается на основе рыночной информации по следующей формуле:

$$TSR = \frac{Div_t}{MK_{t-1}} + \frac{MK_t - MK_{t-1}}{MK_{t-1}}$$

Где

Div_t *– дивиденды текущего периода*

MK_{t-1} *– рыночная капитализация предыдущего периода*

[24] (Graham, Harvey, & Rajgopal, 2005) Сюрпрайз
[25] (Bernard & Thomas, 1990).
[26] D. E. Rapach “Short interest and aggregate stock returns,”
[27] Теплова новое в фин аналитике

MK_t – рыночная капитализация текущего периода

API «Alpha Vantage»[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: месяц, квартал и год. На рисунке ниже представлены гистограммы распределения рассчитанных доходностей.

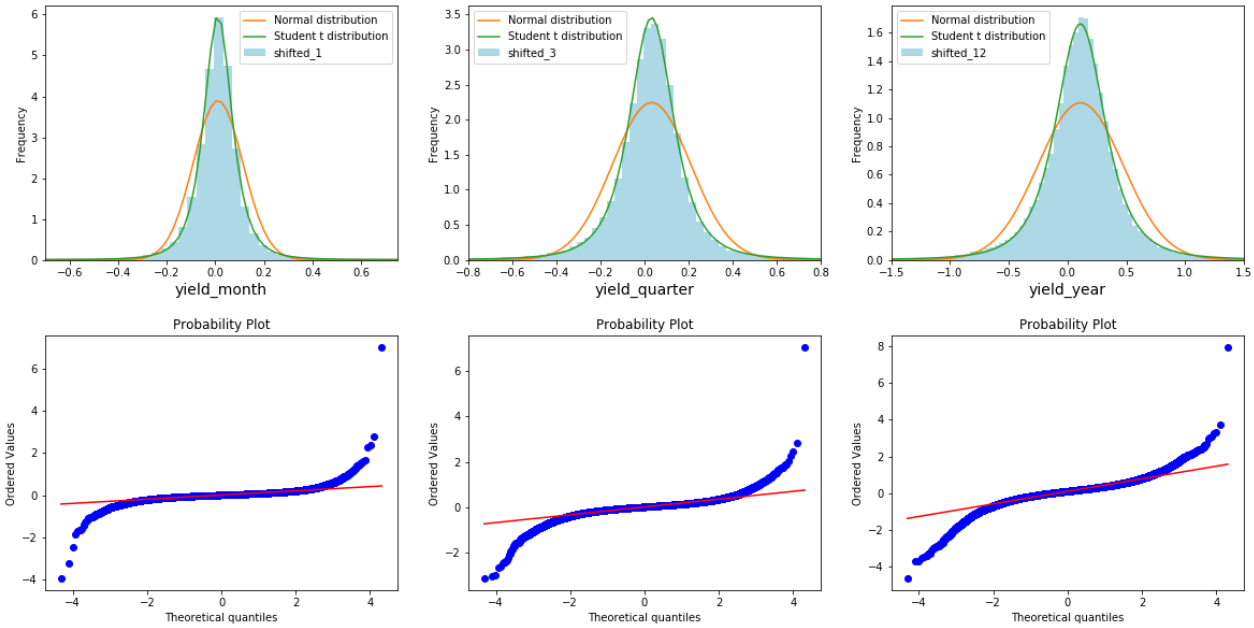


Рис. 3 Распределение месячных, квартальных и годовых доходностей

Из рисунка видно, что распределение доходностей компаний индекса S&P500 далеко 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

[28] Ссылка на API

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).

$$MAD = median(|X_i - \tilde{X}|)$$

$$\tilde{X} = median(X)$$

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.

$$\hat{\sigma} = k \cdot MAD$$

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

$$k = 1 / \left(\Phi^{-1} \left(\frac{3}{4} \right) \right) \approx 1.4826$$

где:

Φ^{-1} – 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.

$$\frac{1}{2} = P(|X - \mu| \leq MAD) = P\left(\left|\frac{X - \mu}{\sigma}\right| \leq \frac{MAD}{\sigma}\right) = P(|Z| < \frac{MAD}{\sigma})$$

Следовательно:

$$\Phi\left(\frac{MAD}{\sigma}\right) - \Phi\left(\frac{-MAD}{\sigma}\right) = \frac{1}{2}$$

А так как:

$$1 - \Phi\left(\frac{MAD}{\sigma}\right) = \Phi\left(\frac{-MAD}{\sigma}\right)$$

Получается, что

$$\frac{MAD}{\sigma} = \Phi^{-1}\left(\frac{3}{4}\right) = 0.67449$$

Или

$$\frac{\sigma}{MAD} = \frac{1}{\Phi^{-1}\left(\frac{3}{4}\right)} = 1.4826$$

Откуда получается, что

$$\hat{\sigma} = 1.4826 \cdot MAD$$

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.

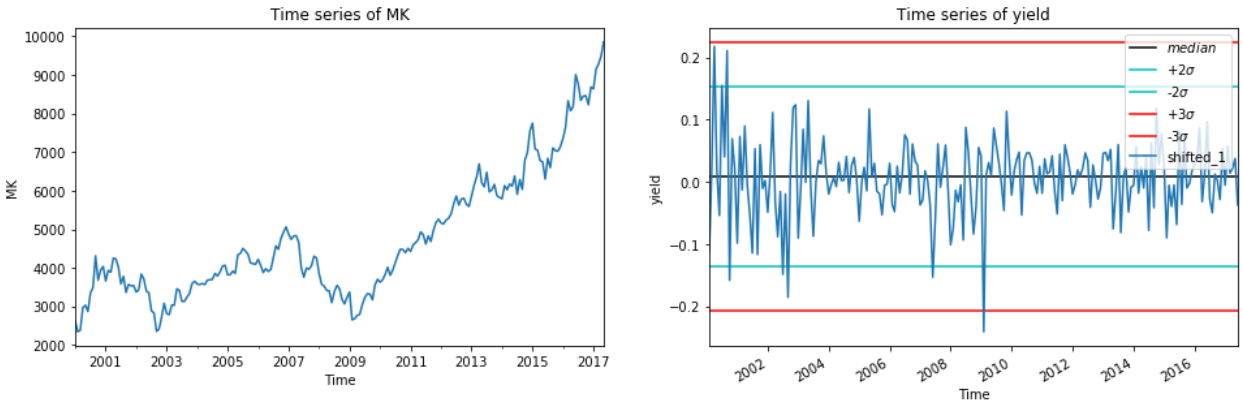


Рисунок 3. Визуализация выбросов доходностей

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.

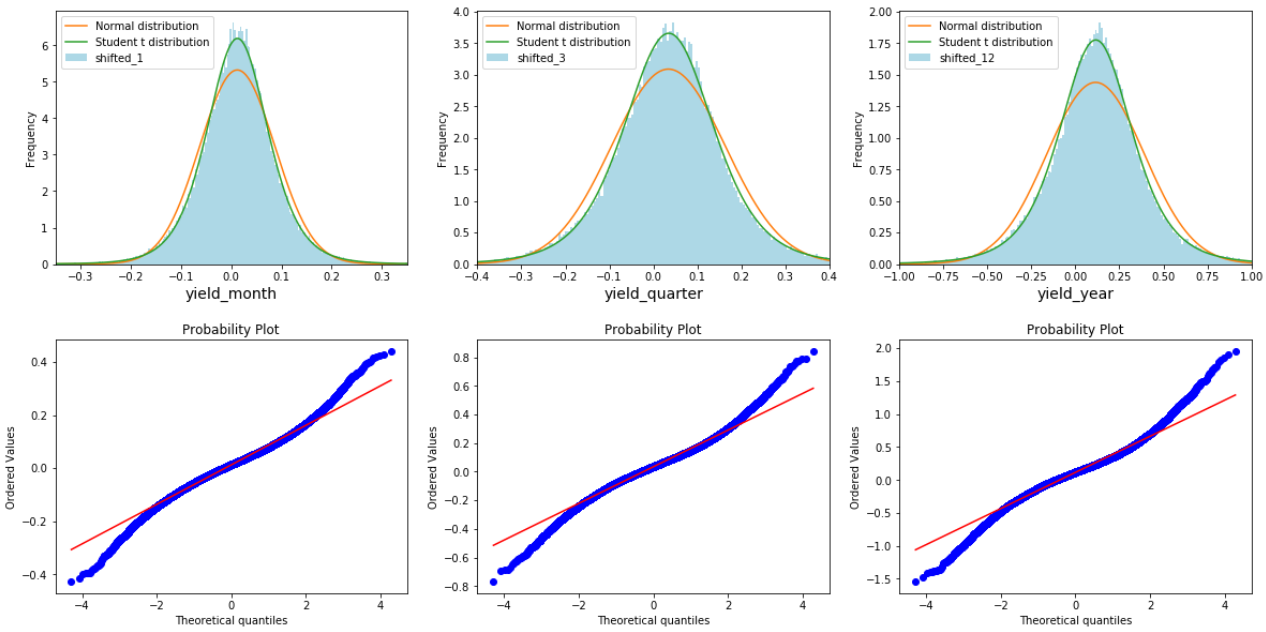


Рис. Гистограммы распределения месячных, квартальных и годовых обработанных доходностей

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 рисунке ниже представлена гистограмма, отражающая сбалансированность рассматриваемого показателя.

Рис. Гистограмма

2.3 Предобработка факторов внутреннего контура

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.

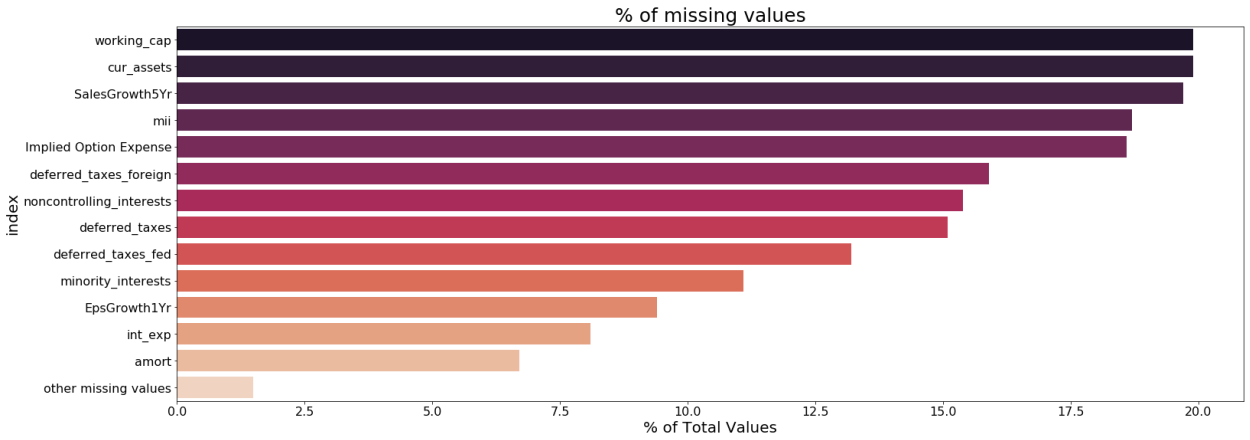


Рисунок 1. Доля пропущенных наблюдений

Для устранения пропусков в наблюдениях в рамках данной работы используется метод сплайн-интерполяции.[29] В математической области

[29] *de Boor, Carl.* A Practical Guide to Splines. — New York: Springer-Verlag, 1978.

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]. 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. 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].

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
ROCE	Рентабельность задействованного капитала	6231.00	0.12	0.13	0.10
sales_growth	Темпы роста продаж	6775.00	7.02	9.22	20.01
ebit/sales	Норма операционной прибыли	6828.00	0.16	0.12	2.09
gross_margin	Норма валовой прибыли	6732.00	39.12	37.03	198.72

[30] *Runge, Carl (1901), "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten", Zeitschrift für Mathematik und Physik, **46**: 224–243.* available at www.archive.org

[31] (Теплова 2012)

eva	Экономическая добавленная стоимость	6710.00	242.69	734.08	2672.72
BV/MV	Отношение балансовой стоимости к рыночной капитализации	5957.00	0.87	1.69	2.69
TSR	Показатель общей доходности акции	5748.00	0.15	0.20	0.37
cvadrant	Квадрант матрицы Ивашковской	6828.00	2.00	2.15	1.47

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:

$$Current\ Ratio = \frac{Оборотные\ активы}{Краткосрочные\ обязательства}$$

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

$$ICR = \frac{\text{Операционная прибыль}}{\text{Процентные расходы}}$$

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

$$\%\ STD = \frac{\text{Краткосрочные обязательства}}{\text{Долгосрочные обязательства}}$$

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, темпов роста основных доходных статей отчета о прибылях и убытках, показатель общей доходности акции (TSR) и экономическая добавленная стоимость (EVA), коэффициент недооценки.

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

$$\text{НОП} = \frac{\text{Операционная прибыль}}{\text{Выручка}}$$

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

$$\text{НОП} = \frac{\text{Валовая прибыль}}{\text{Выручка}}$$

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:

$$ROCE = \frac{\text{Операционная прибыль}}{\text{Активы} - \text{Краткосрочные обязательства}}$$

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] Показатель рассчитывается по формуле:

$$EVA = (ROIC - WACC) \times IC$$

где:

Инвестированный капитал (Invested Capital) – это общая сумма денег, привлеченных компанией путем выпуска ценных бумаг и облигаций. Показатель рассчитывается по следующей формуле:

$$IC = (\text{Обзятельства} + \text{Капитал}) - \text{Денежные средства и экв} - \text{ты}$$

ROIC – Рентабельность инвестированного капитала представляет собой коэффициент эффективности, который предназначен для измерения процентного дохода, получаемого инвесторами в компании от инвестированного ими капитала. Показатель рассчитывается по следующей формуле:

$$ROIC = \frac{\text{Операционная прибыль } \times (1 - \text{ставка налога})}{IC}$$

WACC – Средневзвешенная стоимость капитала отражает средний уровень затрат компании на привлечение и дальнейшее обслуживание капитала из различных источников. Показатель рассчитывается на основе рыночной информацию по следующей формуле:

$$WACC = r_i \times w_i$$

где:

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

w_i *– 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

[32] http://www.sternstewart.com/?content=proprietary&p=cov)

основе модели дисконтирования денежных потоков. В виде формулы коэффициент можно представить следующим образом:

$$K_{\text{нд/пр}} = \frac{P_{market} + P_{iternal}}{P_{iternal}}$$

где:

$K_{\text{нд/пр}}$ – *коэффициент переоценки/недооценки;*

$P_{iternal}$ – *внутренняя стоимость компании;*

P_{market} – *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 рассчитываются такие показатели, как коэффициент BV/BM, индекс устойчивого роста.[33]

Коэффициент BV/BM представляет собой отношение балансовой стоимости компаний к их рыночной капитализации.

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:

$$SGI = (1 + g_s) \times \frac{1}{k} \times \sum_{i=1}^{k} \max[0, (ROCE_i - WACC_i)]$$

[33] (Ивашковская, Животова)

где:

k *– количество лет наблюдений*

g_s *– 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.

Выбранные фактора могут иметь высокий уровень корреляции. 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.

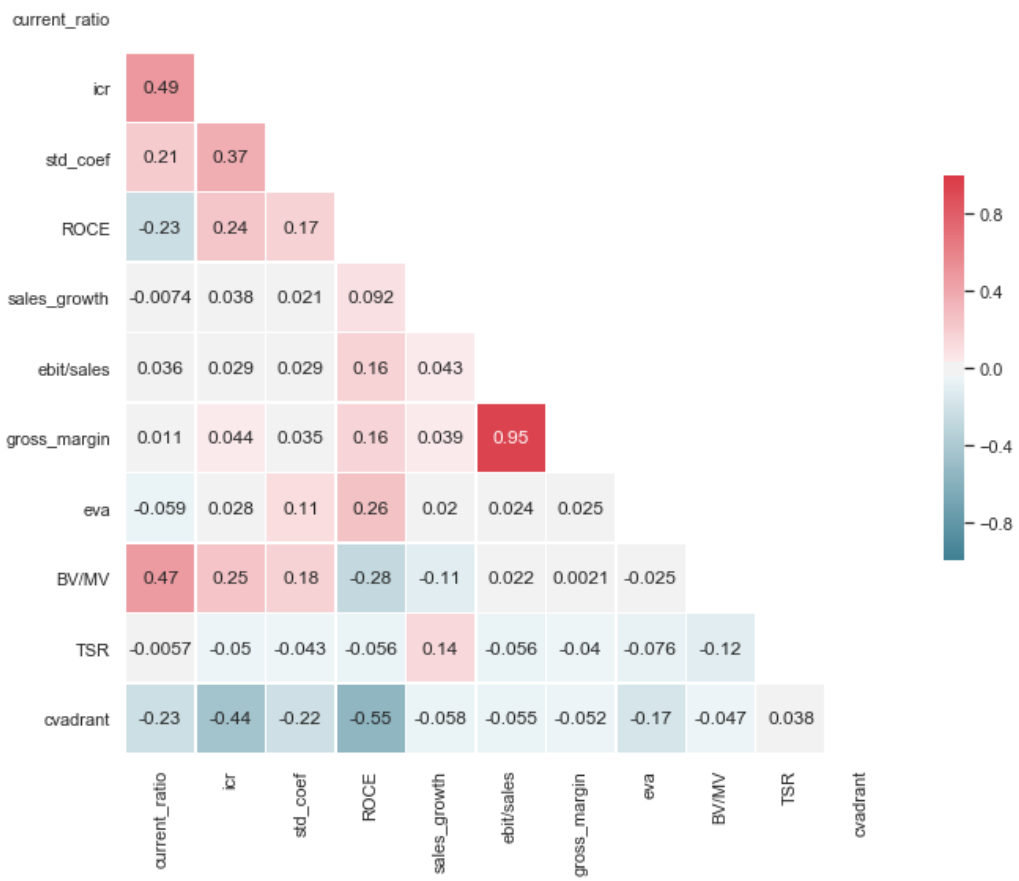


Рис. 1 Корреляционная матрица факторов внутреннего контура

Из рисунка видно, что наблюдается сильная линейная связь между такими факторами, как норма операционной прибыли (ebit/sales) и норма валовой прибыли (gross margin). Удалим показатель нормы валовой прибыли из анализа, так как у показателя нормы операционной прибыли больше наблюдений за рассматриваемый период.

После рассмотрения факторов внутреннего контура перейдем к анализу факторов макро контура.

2.4 Предобработка факторов макро контура

С помощью базы данных «Federal Reserve Economic Data» был получен доступ к макроэкономическим показателям экономики США. Эти данные будут составлять контур внешних факторов, 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 M2	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 %	1.76

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.

M1 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 отражается восприятие рисков субъектами экономических отношений.

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.

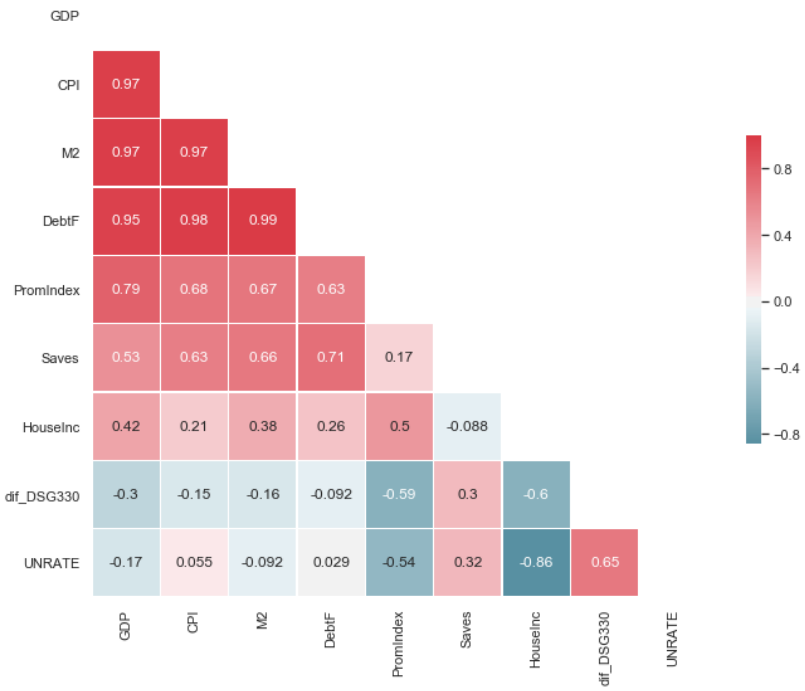


Рис. 1 Корреляционная матрица макроэкономических факторов

Из рисунка один видно, что наблюдается сильная положительная линейная связь между такими факторами, как индекс потребительских цен (CPI), агрегатор денежной массы M2 (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).

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. Так как на данной стадии основная задача исследования заключается в отслеживании влияния факторов на будущую доходность компании, линейно зависимые факторы будут удалены. А для повышения предсказательной мощности модели в дальнейшем будет использоваться метод главных компонент.

После удаления линейно зависимых факторов получаем следующую корреляционную матрицу.

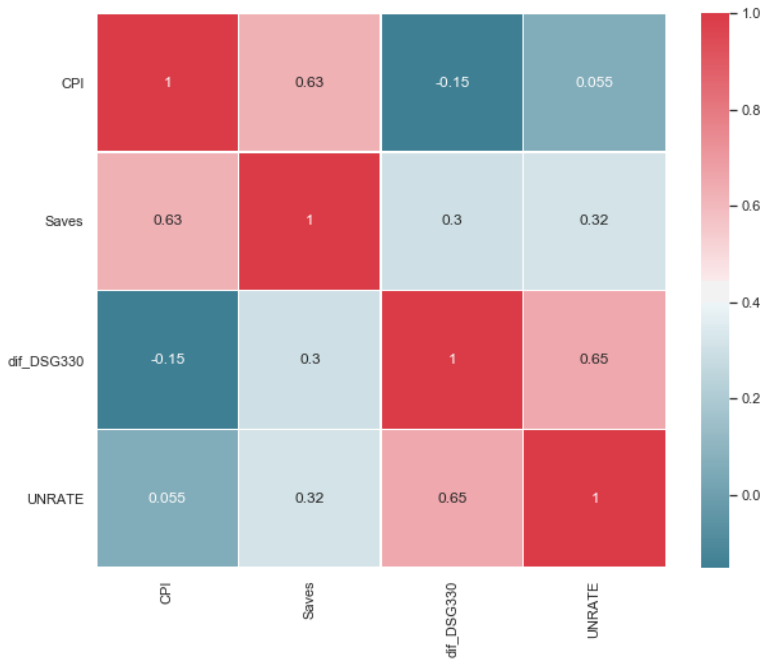
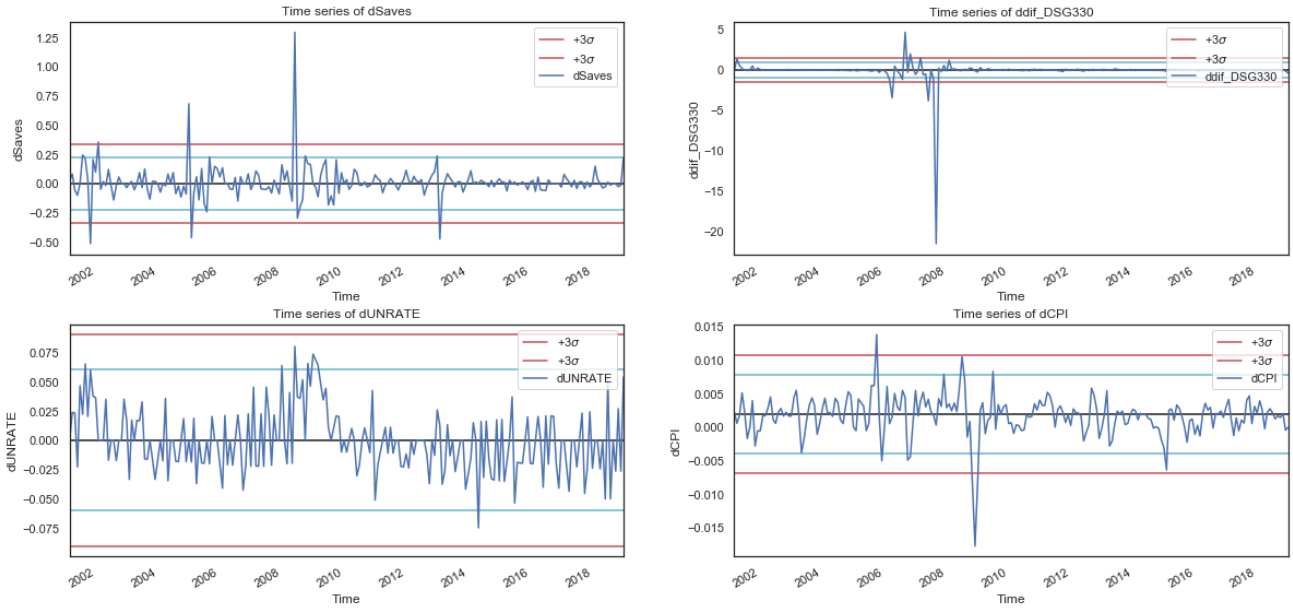


Рис. 2 Модифицированная корреляционная матрица макро факторов

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 равны менее 0.7, то было принято решение оставить данные факторы для дальнейшего включения в регрессионную модель.

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



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

2.5 Предобработка факторов контура новостного шума

С помощью инструмента Гугл трендс **был получен доступ к информационным каналам ПО каждой компании и общие макроэкономиечские запросы.**

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