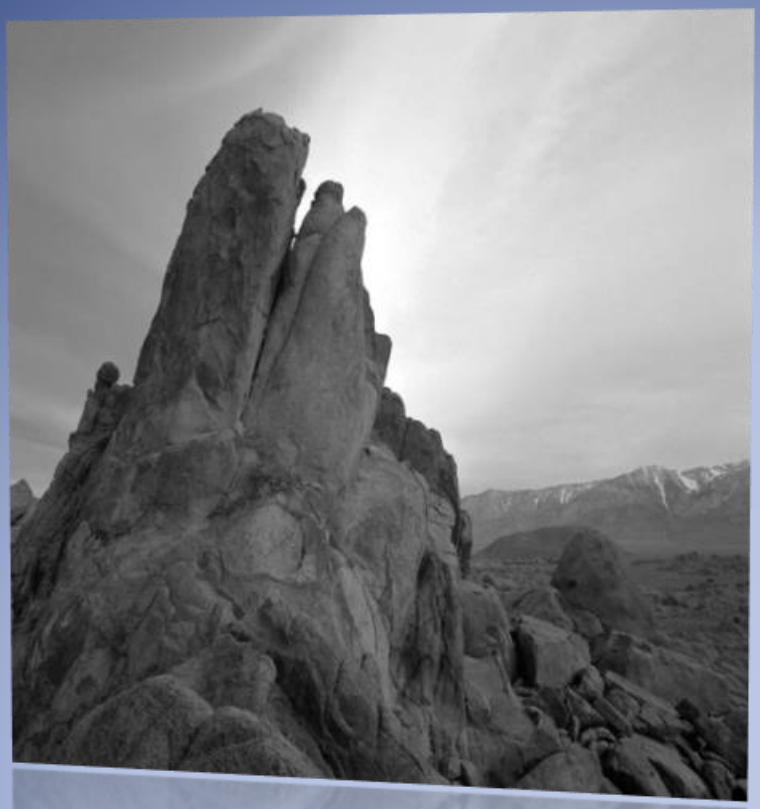
СТАНИМИР КАБАИВАНОВ



Интелигентни системи и самообучаващи се алгоритми за анализ на финансови инструменти

Интелигентни системи и самообучаващи се алгоритми за анализ на финансови инструменти

ПЪРВО ИЗДАНИЕ

СОФИЯ 2020

Д-р Станимир Иванов Кабаиванов – автор

Рецензенти:

проф. д.м.н. Иван Ганчев Иванов

доц. д-р Александър Красимиров Ефремов

Коректор:

гл. ас. Гергана Тодорова Иванова

Издател:

„Евдемония Продъкшън“ – гр. София

Съдържанието е свободно достъпно на http://www.kabaivanov.net

ISBN 978-619-7209-37-2

Съдържание

I. Въведение 7
1. Защо е необходимо това изследване? 9
2. Защо финансови инструменти? 10
3. Софтуер, софтуер, софтуер… 11
4. Терминологична джунгла 13
5. Обективност и пълнота на изследването 13
II. Обучение и самообучение 14
1. Контролирано обучение 16
2. Неконтролирано обучение 21
3. Смесено обучение 23
4. Обучение с утвърждаване 25
5. Източници на данни 26
6. Изходен код и възпроизводимост на резултатите 27
III. Контролирано обучение при анализа на финансови инструменти 28
1. Общи условия и допускания в изследването 29
2. Инструменти с фиксиран доход. Криви на доходността 30
2.1. Модел на Нелсон–Сиегел–Свенсон 32
2.2. Контролирано обучение и анализ на корпоративни облигации 37
2.3. Използвани в обучението алгоритми 44
2.4. Сравнение на точността на използваните алгоритми 49
3. Риск от неплатежоспособност. Дискриминантен анализ и самообучаващи се алгоритми 53
3.1. Модел на Алтман 53
3.2. Алтернативи за оценка на риска от неплатежоспособност 54
3.3. Сравнение на точността на използваните алгоритми 55
4. Изводи 59
5. Изходен код и възпроизводимост на резултатите 61

IV. Неконтролирано обучение при анализа на финансови инструменти 63
1. Финансова стабилност и стрес тестове 65
2. Методология на стрес тестовете 67
2.1. Алгоритми за класификация с неконтролирано обучение 70
2.2. Алгоритми за намаляване на размерността на входните данни.....79
2.3. Класификационни алгоритми и стрес тестове на EBA 84
2.4. Неконтролирано обучение и анализ на риска 94
3. Лихвени индекси и PCA анализ 98
3.1. PCA анализ и доходност на държавни ценни книжа.....99
3.2. PCA анализ и LIBOR 102
4. Изводи.....105
5. Изходен код и възпроизводимост на резултатите.....106
V. Смесено обучение при анализа на финансови инструменти 108
1. Алгоритми, използващи смесено обучение.....110
2. Приложимост на смесеното обучение при анализа на финансови инструменти 113
3. Анализ на борсов индекс с EM и смесено обучение 113
3.1. EM и оценка на разпределенията в отделните състояния.....115
3.2. EM и оценка на пазарните състояния 119
4. Изводи.....124
5. Изходен код и възпроизводимост на резултатите.....126
VI. Специфични проблеми на самообучаващите се алгоритми и начини за тяхното преодоляване.....128
1. Общи проблеми – общи решения 128
1.1. Добри практики при предварителната обработка на данни 129
1.2. Добри практики при избор на аналитични инструменти.....132
1.3. Добри практики при оценка на хиперпараметрите на избраните алгоритми 134

1.4. Добри практики при оценка на точността и валидирането на алгоритмите 136
2. Когато един инструмент не е достатъчен 138
VII. Вместо заключение 140
Списък на фигурите 142
Списък на таблиците 145
БИБЛИОГРАФИЯ 147

I. Въведение

Живеем в свят, който предлага все по-голяма степен на автоматизация и зависимост от различни системи за вземане на решения. Резултатите и трансформацията на начина, по който работим, живеем и в крайна сметка мислим, са свързани със съхранението, натрупването и подходящата обработка на голям обем от разнообразна по своя характер информация. Освен „цифрови изображения“ на заобикалящата ни реалност това огледало на нашата действителност предлага сериозни предизвикателства, интригуващи решения за част от тях и достатъчно неизследвани територии, които да предизвикат въображението на изкушения читател. За да се ориентираме добре в тази информационна среда имаме нужда от интелигентни помощни средства – отворени към променящите се външни условия и с възможност да се коригират въз основа на натрупаните опит и знания. Описаните характеристики изключително много приличат на начина по който хората постепенно и самостоятелно опознават света, което ни позволява да обозначим този тип системи като самообучаващи се. Както и при живите същества, тези процеси не протичат хаотично, а с помощта на определени правила – алгоритми. Плод именно на тяхното усвършенстване са някои от най-вълнуващите и полезни придобивки в нашето „цифрово ежедневие“.

Мащабите и скоростта, с която се развива анализът на данни, предопределят централната роля, която алгоритмите със способност да се адаптират и самообучават заемат при отделните етапи на изследване на събраната информация. Не може обаче да не отчитаме и факта, че основните ни интереси все още са свързани с физическото ни съществуване. Следователно светът на данните и тяхната обработка са ценни най-вече във връзка с ползите и облекченията, които носят в реалния свят.

Целта на това изследване е да проследи и анализира приложението на самообучаващите се алгоритми и интелигентните системи за обработка на информация за решаването на практически значими икономически проблеми като оценката на финансови инструменти. Изложението залага на достъпен и практически ориентиран подход, като водещ критерий за включването на отделни аналитични инструменти в проучването е тяхната

приложимост и възможност да отговорят на изискванията за бързодействие и стабилност при работа в реални условия.

Представените алгоритми предполагат добро познаване на основните концепции и принципи на статистическите изследвания. За читателите, които нямат необходимата теоретична основа, са предвидени специфични допълнения и препратки в отделните части на текста.

1. Защо е необходимо това изследване?

Анализът на данни и по дефиниция, и по необходимост залага на интердисциплинарен подход в търсенето на решения на практически значими проблеми. Причините за това са много, но е достатъчно да спомена, че отделните задачи носят в себе си спецификата на съответната област (транспорт, финанси, продажби и др.), като едновременно с това са подложени на въздействието на случайни фактори, неточности и грешки в регистрирането на данните, необходимост от съвместяването на числени и категорийни параметри и др.

Както повечето пазарно ориентирани дейности, така и анализът на данни е, освен всичко друго, и своеобразно състезание с цел да се постигнат максимално бързо точни, адекватни и полезни резултати. Устремени към победата, много често участниците в това състезание са склонни да жертват определени характеристики на крайното решение, но да бъдат първи или просто да покажат по-добри показатели за точност на базата на наличната информация. Редица изследвания ([1], [2], [3]), както и личните наблюдения на автора показват, че този тип поведение може да доведе до някои от следните неоптимални резултати:

- Акцент върху краткосрочното представяне на използваните аналитични инструменти, за да се открие този подход или негов вариант, които най-добре описват проблемната ситуация, без да се изследват неговата пригодност в бъдеще и стабилността на получените резултати.

Формалното доказателство за правилната работа на един алгоритъм и дългосрочната проверка на неговите качества са сложни и изискващи много ресурси задачи. Но това не означава, че приемането на даден метод трябва да бъде формализирано до сравнение на един показател за грешка, без да се коментират в дълбочина неговите предимства и ограничения.

- Стремеж да се търси „философският камък“ сред наличните инструменти – метод, който при наличието на достатъчен изчислителен ресурс и време може да намери решение на почти всяка задача, пред която сме изправени.

Много често имаме възможност да използваме повече от един инструмент за постигане на нашите цели. Някои от аналитичните средства, с които разполагаме, са с по-широко приложение или просто са по-мощни. Но това не означава, че винаги трябва да прибягваме до тяхната употреба, особено когато има по-проста, но достатъчно добра алтернатива.

- Ограничаване на свободата на изследване до функционалността, предоставяна от наличните (и най-често стандартни) софтуерни пакети.

Този тип поведение отразява нещо повече от обикновена инертност, като много често желанието да се придържаме към изпробвани и проверени решения ни ограничава и не позволява да реализираме пълния потенциал на знанията, с които разполагаме. А с времето и скоростта на попълване на нашите собствени познания става зависима от темпа, с който разработчиците добавят нови функции към използваните от нас софтуерни пакети. Казано по друг начин – от господари ние се превръщаме в затворници на програмното си обезпечение.

- Желание да се комбинират известни вече решения, така че с малки промени те да се напаснат към спецификите на конкретния анализиран проблем.

Няма нищо нередно в това да се използват вече известни резултати и начини за справяне със сходни проблеми. Когато обаче търсенето на решение се превърне в издирване на достатъчно на брой елементи, с които да сглобим „аналитичен Франкенщайн“, това е ясен сигнал, че сме стигнали момента, в който безкритично отдаваме предпочитание на бързината пред качеството.

2. Защо финансови инструменти?

Едва ли е необходимо да се аргументира изрично практическата значимост на анализа и оценката на финансови инструменти. Прогнозите на Делойт ([4]) за пазара на финансови и консултантски услуги през 2020 г. са свързани както с по-доброто използване на технологиите, така и с преоткриването на съществуващите пазари с помощта на нови продукти. Превръщането на желанието за успех и растеж в реалност е свързано с

доброто познаване на утвърдените аналитични инструменти, тяхното комбиниране и адаптиране към променящите се пазарни условия.

От чисто практическа гледна точка фокусирането върху финансовите инструменти и трансакции носи няколко важни предимства:

- решаваните задачи могат лесно да бъдат разбрани от икономисти с различна специализация, доколкото основните инструменти са широко известни и добре познати;
- информацията за процесите, протичащи на пазарите, е достъпна, което позволява разглежданите алгоритми да бъдат проверени и подложени на независима оценка от интересуващите се читатели;
- промените в цените на финансовите инструменти са резултат от едновременното въздействие на разнообразни фактори, което позволява да се изследват и наблюдават в реално време интересни феномени и относително по-редки събития.

Извън посочените предимства има още един важен аргумент, свързан с начина, по който самообучаващите се алгоритми и системи за интелигентен анализ на информация въздействат върху хората. Реакциите варират от пълно отрицание – „това не може да работи“, до безрезервно възприемане – „това е магия и работи“, но при всички случаи отсъства безразличието. За да може обаче предимствата и недостатъците на тези аналитични инструменти да бъдат трезво и прагматично да бъдат оценени, е необходимо да разбираме истинската им същност, ограниченията и допусканията, без които тяхното прилагане не е оправдано. В този ред на мисли е формулирана и последната задача на този текст – да опита да покаже, че едно изследване не е задължително да бъде скучно, сухо и отблъскващо за всички, освен за малката група хора, работещи върху точно същата проблематика.

3. Софтуер, софтуер, софтуер…

Приложението на самообучаващите се алгоритми предполага използването на всички предимства, които съвременната изчислителна техника ни предлага. Ние сме облагодетелствани да разполагаме с възможности за обработка на информация и извършване на изчисления, които нямат аналог в записаната до момента история. Това разнообразие

обаче има своя цена и тя е свързана с нуждата от предварително запознаване с различните възможности и избор на тази от тях, която е най-подходяща за решаване на конкретния проблем.

Последното твърдение не е случайно и означава, че е необходимо да познаваме повече от един софтуерен инструмент, за да можем да отговорим на предизвикателствата, пред които сме изправени. Не е възможно обаче в рамките на едно изследване да бъдат представени и приложени всички програмни продукти, които реализират самообучаващи се алгоритми или са насочени към оценката на финансови инструменти. Поради тази причина настоящото изложение се придържа към следните основни принципи при избора на софтуерни продукти:

- предпочитани са възможностите с отворен код и неограничаващи лицензи, защото те дават възможност на читателите да експериментират, както и да използват заложените в тях идеи според своите нужди;
- реализацията на самообучаващите се алгоритми в това изследване е преди всичко на Python[1], R[2] и C++, като нарочно търсен ефект е комбинацията от няколко платформи;
- всички изходни кодове са свободно достъпни на адрес https://github.com/drnmy/mlf;
- предпочитани са безплатни и доказани източници на данни като гаранция, че експериментите в изследването могат да бъдат възпроизведени.

Отделни части от изследването изискват повече време за обучение на използваните алгоритми или за обработка на информацията от провежданите експерименти. При всички случаи обаче необходимите ресурси за самостоятелна проверка на получените резултати са съобразени с техниката, която средностатистическият читател има на разположение.

[1] https://www.python.org/, като използваната версия е 3.8.3.
[2] https://www.r-project.org/ , като използваната версия е 4.0.2.

4. Терминологична джунгла

Една от особеностите, с които неизбежно се е сблъсквал всеки читател, изкушен от анализа на данни, е наличието на разнообразни български преводи на широко използваните английски термини и съкращения. Примерите за директно привнасяне на обозначения са много, включително и масово разпространеното име „машинно обучение“ (machine learning). В повсеместната си употреба, този термин се прикрепя към различни неща – от изкуствен интелект до статистически анализ. В известен степен това го компрометира до степен, при която е по-добре да бъдат обозначени конкретните аналитични инструменти, вместо общото и силно размито по съдържание понятие.

Целта на изследването не е да добави нови елементи в пъстрата палитра от названия. Затова към всеки един термин, който няма единодушно прието и утвърдено съответствие на български език, ще бъде добавяно и международно признатото понятие.

5. Обективност и пълнота на изследването

Представените в изследването резултати са получени при обработката на публично достъпна пазарна информация. Основната цел е да бъдат безпристрастно и пълно анализирани ползите от прилагането на самообучаващи се алгоритми при решаването на финансови проблеми. Изводите и заключенията, представени в текста, изразяват само и единствено личното виждане на автора, не са обвързващи и не представляват препоръка за покупка или продажба на ценни книжа.

II. Обучение и самообучение

Марси Дрискол определя обучението като „*трайна промяна в човешкото поведение или потенциал…[което] е следствие от взаимодействието на обучаващия се с околния свят*“ ([5]). Основните характеристики на тази дефиниция могат да бъдат пренесени и в контекста на това изследване, стига да заменим човешкото поведение с това на използвания алгоритъм, а околния свят - с данните, които обработваме. Теоретичната основа е важна, защото предимствата и недостатъците на обучението, с които се сблъскваме в нашето ежедневие, ще дадат своето отражение и когато прилагаме различни възможности за анализ на финансова информация. Не може обаче да спрем до философските аспекти на понятието, защото за решаването на конкретни задачи е необходимо да ползваме и ясно определени инструменти.

За нуждите на това изследване ще разделим обучението на четири вида според това как се реализира то и какво бихме могли да очакваме като резултат:

- контролирано обучение;
- неконтролирано обучение;
- смесено обучение;
- обучение с утвърждаване.

Посоченото разделение е резултат от комбинирането на следните критерии:

- наличие на предварително известни решения на анализирания проблем;
- изисквания към получените решения.

За да демонстрираме различните подходи към решаването на изследваните проблеми, както и какви са особеностите на получените резултати при различите типове обучения, е необходимо да им дадем по-строга дефиниция.

Нека имаме съвкупност $X = \{(x_i)\}_{i=1}^{N} = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(D)} \\ \vdots & \ddots & \vdots \\ x_N^{(1)} & \dots & x_N^{(D)} \end{pmatrix}$, описваща изследвания проблем, като всеки елемент от $x_i = (x_i^{(1)}, x_i^{(2)}, .., x_i^{(D)})$ представя отделна характеристика. Векторите $x^{(k)} = \begin{pmatrix} x_1^k \\ \vdots \\ x_N^k \end{pmatrix}$ могат да имат различен смисъл и да представят, както числени, така и категорийни, и номинални параметри на изследвания обект. Приемаме, че са в сила следните допускания:

- елементите на x_i могат да имат различен смисъл и размерност;
- за всяко $k = 1..D$, стойностите в $x^{(k)}$ се отнасят до една и съща характеристика на изследвания проблем.

Да разгледаме един пример, при който x_i съдържа характеристиките на една облигация. Тогава отделните параметри биха могли да бъдат *{емитент, номинал, дата на падежа, купонов процент, цена на обратно изкупуване, дюрация}*. Всеки един от описаните елементи има различен смисъл и размерност – например номиналът и цената на обратното изкупуване могат да бъдат в щатски долари, докато дюрацията е в години, а емитентът може да бъде просто борсов код (т.е. символен низ или код). Това, което второто допускане ни казва, е, че при наличието на повече от една облигация винаги първият елемент ще бъде свързан с емитента. Например:

{04HA, 1000, 28.06.2025, 3.50%, 0, 4.56},

{3D4A, 100, 26.11.2026, 5.75%, 0, 5.23}.

Допускането е важно, защото ни позволява да анализираме всеки от параметрите, описващи проблема, и да проследим дали има някаква връзка между него и други интересуващи ни характеристики. Ако посочените две допускания не са изпълнени, то или трябва да обработим предварително данните, с които разполагаме, или да се запитаме възможно ли е да извлечем максимална полза от информация, която не е структурирана по никакъв начин и нямаме идея какво точно ни казват отделните елементи.

1. Контролирано обучение

Международно признатият термин за този тип обучение е supervised learning. На български език са възможни няколко различни превода (напр. „подпомагано“, „обучение с учител“ и др.). Скромното мнение на автора е, че „контролирано“ е подходящият термин, доколкото във всеки един момент може да проверим/проконтролираме как се справя конкретно решение. Това е възможно благодарение на характерната особеност на контролираното обучение, че освен с $\{(x_i)\}_{i=1}^{N}$ ние разполагаме с още една съвкупност, означавана с $Y = \{(\mathrm{y}_i)\}_{i=1}^{N}$, която съдържа информация за желаните или предварително известни изходи. За да се елиминират потенциалните неточности в представянето на отделните алгоритми и резултати, във всяка една ситуация ще използваме наименованията, посочени в Таблица 1.

Таблица 1. Важни понятия, използвани в изследването

Наименование	***Бележки***
Входни данни	$\{(x_i)\}_{i=1}^{N}$, $\{(\mathrm{y}_i)\}_{i=1}^{N}$ – всички релевантни на решаваната задача данни, с които разполагаме и които използваме към настоящия момент.
Обучаващо множество	Данни, използвани за обучение на алгоритмите: $\{(x_i)\}_{i=1}^{P}, \{(\mathrm{y}_i)\}_{i=1}^{P}, P < N$.
Тестово множество	Данни, използвани за проверка на точността на алгоритмите: $\{(x_i)\}_{i=1}^{Q}$, $\{(\mathrm{y}_i)\}_{i=1}^{Q}$, където обикновено $\mathrm{Q} < P$, а тестовото и обучаващото множество нямат общи елементи.
Валидиращо множество	Данни, използвани за допълнително валидиране на получените резултати: $\{(x_i)\}_{i=1}^{R}$, $\{(\mathrm{y}_i)\}_{i=1}^{R}$, $R \sim Q < N$, а тестовото, валидиращото и обучаващото множество нямат общи елементи.

Работата на алгоритмите използващи контролирано обучение, е да намерят една функция f – такава, че:

$$f: \mathrm{X} \xrightarrow{yields} \mathrm{У}. \qquad (1)$$

Тъй като възможностите за f допускат повече от едно възможно решение, обикновено зависимост (1) е съпроводена със следните ограничения:

- пространството за избор на f е предварително определено и отговаря на ясни критерии (например f е линейна функция). Прилагането на f води до $\widehat{y_i} = f(x_i)$;
- оценката за точността на създадения модел може да се представи с функция на „загубите“ - $l(\hat{y},y)$, която зависи от емпирично наблюдаваните и теоретично получените стойности за y.

В зависимост от това как се проявяват тези ограничения, съществуват различни алгоритми, които използват контролирано обучение. В Таблица 2 са представени най-широко използваните от тях, без, разбира се, да се претендира за изчерпателност в изброяването. Подробна информация за всеки един от посочените алгоритми може да се намери в ([6]) и ([7]).

Таблица 2. Алгоритми, използващи контролирано обучение

Алгоритъм	***Бележки***
Регресионен анализ	В тази група попадат линейните и нелинейните регресионни модели ([8], [9]), като специално внимание заслужава логистичната регресия в случаите, свързани с решаване на класификационни задачи.
Дървета на решенията	Алгоритмите от тази група могат да се използват както за решаване на класификационни проблеми, така и в случаи, когато резултатите са числени променливи. От гледна точка на анализа на финансови инструменти особен интерес представлява възможността получените дървовидни структури да бъдат съпоставени с процеса на вземане на решения.
Машина с поддържащи вектори	Освен за проблеми, свързани с класификация, решенията от тази категория могат да допринесат съществено при изграждането на хибридни модели (например с невронни мрежи) и като алтернатива на обикновената регресия. Както и при дърветата на решенията, важни за практиката са и получените резултати, и тяхното по-широко тълкуване като „поле на решенията“. То може да ни даде допълнителна представа за риска, с който са

	съпроводени дадено действие или инвестиционна стратегия.
Линеен дискриминантен анализ	Въпреки че този вид анализ най-често се свързва с модела на Алтман ([10]), възможностите на ЛДА са много по-широки. Освен за оценка на финансовото състояние на отделна компания, алгоритмите от тази категория са подходящи инструменти за оценка и на рейтинга на облигации, решения, свързани с промени в капиталовата структура, и др.
KNN класификация	Освен за класификация KNN може да бъде използван и като компонент за създаване на хибридни модели. Определен интерес за финансовите приложения представлява фактът, че ако използваме алгоритъма с цел прогнозиране, особено силно влияние върху търсената стойност биха оказвали сходните, а не всички наблюдения.
Невронни мрежи (някои видове)	Невронните мрежи са широко (дори в определени ситуации прекалено широко) използвано средство за анализ. В контекста на оценката на финансови инструменти те намират приложение при моделирането на динамични редове, класификация на ценните книги и др. В това изследване вниманието е насочено преди всичко върху коректното приложение на невронните мрежи, както и върху възможностите за използването им съвместно с други от изброените алгоритми.

Контролираното обучение има една характеристика, много близка до фундаменталните понятия, с които работят участниците на финансовите пазари. А именно – риска и естествения стремеж той да бъде минимизиран. Зависимост (1) позволява да изберем функцията f така, че $f(x) = P(y|x)$, като например логистичната регресия е типичен пример[3]. Възможно е обаче f да бъде избрана така, че $f(x, y) = P(x, y)$, като линейният дискриминантен анализ е класически пример за това.

[3] Резултатите от линейната регресия могат да се тълкуват като P(y = 1|x).

Независимо от това кой подход е следван за избора на f, трябва да отговорим и на второто ограничение, а именно – да ги подредим съобразно с определен критерий. Тъй като за всяко $\{(x_i, y_i)\}_{i=1}^{N}$ ще имаме и $l_i = l(\hat{y}_i, y_i)$, може да дефинираме риска, свързан с избор на f, като сума от натрупаните до момента „загуби“ в резултат на нейното използване. Ако отчитаме и броя на наблюденията, с които работим, тогава:

$$\text{Риск} = \frac{\sum_{i=1}^{N} l_i}{N} = \frac{\sum_{i=1}^{N} l(\hat{y}_i, y_i)}{N} \qquad (2)$$

Необходимо е да направим уточнение във връзка с понятията *загуба* и *риск*, защото те носят малко по-различен смисъл спрямо установеното им значение във финансите. Преди всичко тук става дума за отклонение между наблюдаваните и изчислените стойности. Тази разлика може да има и парично изражение, което би позволило да я съпоставим с финансовите загуби, но не е задължително стойностите да бъдат едни и същи. От друга страна, представянето на риска като осреднена стойност насочва вниманието ни по-скоро към показатели като VaR и CVaR (въпреки че не се препокрива изцяло с тяхната дефиниция), отколкото към отклонения спрямо някаква очаквана стойност.

Вапник посочва в ([11]) два принципа, които могат да подпомогнат избора ни на *f*:

- Принцип за минимизиране на емпиричния риск.

Съгласно с този принцип най-добрият избор ще бъде този, който минимизира риска от (2), откъдето и търсенето на f е на практика процес на оптимизация. Тази идея е привлекателна поради поне две причини:

  - методите за оптимизация са добре известни и съществуват много техни софтуерни реализации, което прави решението достъпно и бързо;
  - постигането на най-доброто възможно съответствие с наличните данни е източник на допълнителна увереност и показва, че сме направили всичко възможно предвид наличната информация.

При липса на достатъчно данни обаче, както и когато пространството за избор на f предполага наличието на много възможности, прилагането на този принцип може да доведе до прекомерно нагаждане (overfitting) спрямо

наличната информация. Като резултат ще се получи модел, който описва множеството от данни, върху които е обучаван, много добре, но има слабо представяне при постъпването на нова информация.

- Принцип за минимизиране на структурния риск.

Аристотел ([12]) твърди, че „при равни други условия може да смятаме за превъзхождаща тази идея, която стъпва върху по-малък брой постулати или допускания“. Тази идея, по-известна в съвремието с интерпретацията си като „бръснача на Оукъм“, намира отражение в принципа за минимизиране на структурния риск. За да се решат проблемите, свързани с прекомерното нагаждане, в процеса на оптимизация се правят промени, като в крайна сметка функцията на риска придобива вида:

$$\text{Риск}' = \frac{\sum_{i=1}^{N} l(\hat{y}_i, y_i)}{N} + \gamma\, g(x, y), \qquad (3)$$

където $g(x, y)$ е коригиращата функция, а γ е коефициент, който определя значимостта на корекцията. Според стойността на γ този подход може да обхване различни случаи и да отговори на съответните цели:

- при $\gamma = 0$ изразът (3) е идентичен с (2), което означава, че резултатите ще бъдат същите както при оптимизация на емпиричния риск;
- стойности на $\gamma < 0$ не носят полза във връзка с нашето желание да избегнем прекомерното нагаждане;
- при стойност на $\gamma > 0$ имаме възможност за избор между отместване и разсейване на получените резултати. Високите стойности на γ придават по-голямо значение на коригиращия фактор и следователно могат да доведат до по-голямо отместване на нашите резултати.

Последните разсъждения повдигат въпроса как може да бъде определена стойността на γ, като обичайно това се случва с помощта на допълнителни проверки, които включват данни извън обучаващото множество.

Интерес представляват възможните финансови отражения на този коефициент. Ако допуснем, че има исторически данни, така че множеството, върху което се обучават алгоритмите, е достатъчно голямо, тогава γ може да се използва като индикатор доколко е допустимо да смятаме, че бъдещето ще следва същите правила за развитие и промяна,

както миналото. Високи стойности на коефициента, получени в резултат на допълнителните тестове, биха означавали и необходимостта да се отчитат съществени промени с развитие на времето. Без да измества останалите показатели за риск, γ може да предостави ценна информация за вземащите решение на финансовите пазари.

2. Неконтролирано обучение

За разлика от контролираното обучение в този случай разполагаме единствено с характеристиките $\{(x_i)\}_{i=1}^{N}$, без да имаме информация за $\{(y_i)\}_{i=1}^{N}$. При това посочените в предходната част допускания за вектора x_i остават в сила.

Липсата на информация за $\{(y_i)\}_{i=1}^{N}$ означава, че не е възможно да се търси по аналог функция *f*, която както в (1) да изразява $f(x) = P(y|x)$, или $f(x, y) = P(x, y)$. Възможно е обаче да се търси априори вероятностно разпределение $P_X(x)$, така че $f(x) = P_X(x)$. В Таблица 3 са представени някои от най-разпространените алгоритми, използващи неконтролирано обучение, както и тяхната възможна употреба в областта на анализа на финансови инструменти.

Таблица 3. Алгоритми, използващи неконтролирано обучение

Алгоритъм	***Бележки***
Намаляване на размерността • PCA • CCA	Намаляването на размерността е от съществено значение за моделите, които включват множество зависими икономически променливи. Оптимизацията и управлението на инвестиционни портфейли е една от областите, в които тези алгоритми намират практическо приложение ([13], [14]).
Алгоритми за клъстеризация • K-Means • DBSCAN • Optics • Смесени модели	Алгоритмите за клъстеризация могат да се прилагат за решаване на различни финансови задачи, като например класификация на компании и ценни книжа ([15]), оценка на деривативни инструменти ([16]). Освен изброените възможности, особен интерес представлява анализът и групирането при

	представянето на колективни инвестиционни схеми (напр. взаимни фондове), оценка на стила на управление на портфейлните мениджъри и др.
Невронни мрежи	Адаптивността на невронните мрежи ги прави ефикасен инструмент и за решаване на задачи, свързани с неконтролирано обучение. Например самоорганизиращите се мрежи на Кохонен могат да бъдат използвани за изследване и групиране на финансовите показатели на финансови институции ([17]). Успехът на GAN невронните мрежи при генерирането на изображения е предпоставка за използването им при генериране на данни за оценка на инвестиционни портфейли и стратегии. Това би могло да ускори и подобри в значителна степен Монте Карло анализа на финансови инструменти ([18]).
Анализ за аномалии и отклонения • LOF • Изолирана гора	Отклоненията са важни за оценката на финансови инструменти, от една страна - защото представят редки събития, а от друга - защото обикновено остават недобре проучени от традиционните средства за анализ. Класическите примери за оценка на аномалии най-често са свързани с анализ на отделни трансакции (например в търсене на злоупотреба с кредитни карти). Възможно е обаче предварителната обработка на данните с търсене на необичайни отклонения да се използва като средство за повишаване на точността на последващите анализи ([19], [20]).

Неконтролираното обучение може да подпомогне значително анализа на финансови инструменти, особено в комбинация с други средства. Освен че имат специфични за отделните алгоритми недостатъци, които трябва да познаваме, този тип изследвания могат да ни изправят и пред някои общовалидни затруднения:

- Съпоставяне на резултатите от приложението на алгоритмите с изискванията и очакванията на финансовата теория.

Този проблем е повече от търсене на тривиално съответствие. Причината е, че независимо от това колко добри са получените резултати, те трябва да могат да бъдат обяснени точно и достоверно с помощта на терминологията на съответната проблемна област. Когато става дума за анализ на финансови инструменти, това означава, че те трябва да бъдат „преведени“ на език, който финансистите разбират добре.

- Използването им съвместно с други аналитични инструменти натоварва допълнително целия процес заради необходимите изчисления.

Допълнителните изчисления зависят от избрания алгоритъм, като специфични затруднения могат да възникнат при обработката на изключително големи по обем данни и/или при необходимостта да се предостави много бързо решение.

3. Смесено обучение

Смесеното обучение, както подсказва името му, има характеристики, които са междинни по отношение на двата вече представени подхода. Най-често те се изразяват в това, че разполагаме освен с набор от характеристики $\{(x_i)\}_{i=1}^{N}$, още така и с относително малък брой крайни стойности $\{(y_i)\}_{i=1}^{k<<N}$. Причините, поради които липсват крайните стойности за всички наблюдения, с които разполагаме, могат да бъдат различни – от икономическа целесъобразност (скъп процес по предварителна обработка и определяне на y_i) до исторически причини (не е събирана такава информация в миналото). Така описаната ситуация ни предоставя следните възможности:

- да подобрим точността на получените резултати спрямо неконтролираното обучение, като използваме достъпните крайни/желани стойности;
- да продължим да използваме данните, за които липсват стойностите $\{(y_j)\}_{j=k+1}^{N}$.

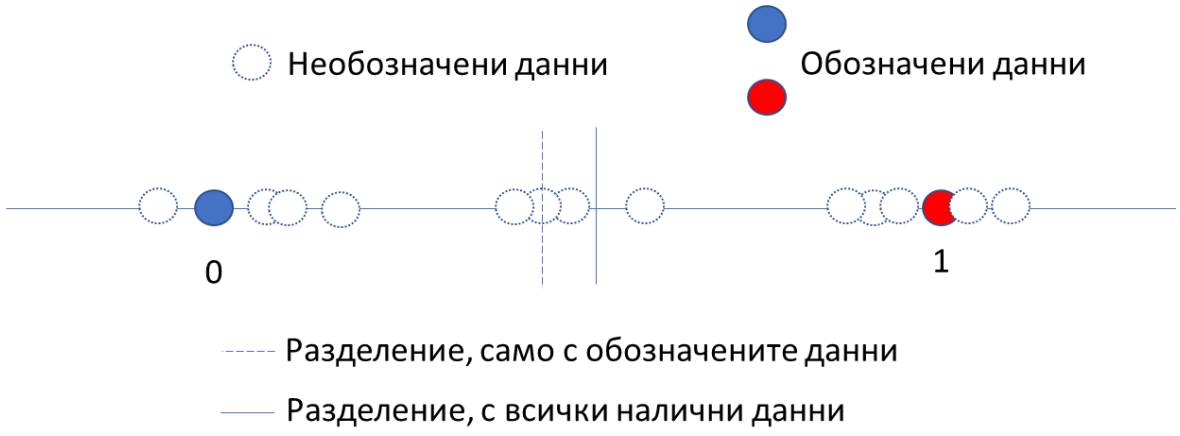
За да използваме в максимална степен предимствата, предоставени ни под формата на известните стойности $\{(y_i)\}_{i=1}^{k << N}$, можем да изберем между следните възможности:

- да се опитаме да изчислим липсващите стойности $\{(y_j)\}_{j=k+1}^{N}$;
- да се опитаме да намерим съответствието $X \to Y$, като разчитаме на случаите, за които вече са известни $\{(y_i)\}_{i=1}^{k << N}$.

Когато се налага да използваме смесено обучение, обикновено целта е свързана с прилагането на получените модели върху данни, които не са известни към момента (т.е. по същество са тестови). Поради тази причина, съзнателно или не, ние прилагаме индуктивен подход и в крайна сметка използваме и двете описани възможности.

Фигура 1 представя ползите от прилагането на смесено обучение в контекста на един изключително прост пример при наличието на един признак и две групи – обозначени с 0 и 1. Плътно оцветените стойности са предварително известни, но както се вижда, те са много по-малко на брой (2) спрямо общо наличните стойности (15). Ако използваме контролирано обучение, то тогава границата, при която бихме класифицирали едно наблюдение като 0 или 1, е представена с пунктираната отсечка (напр. 0.5). С помощта на смесено обучение тази граница би се преместила до позицията, обозначена с плътна вертикална отсечка.

Разликата в двете стойности е всъщност ефектът от прилагането на смесено обучение. Използваната дума неслучайно не е „полза“, защото няма гаранция, че прилагането на обсъжданите алгоритми в крайна сметка ще доведе до по-добро решение.



Фигура 1. Смесено обучение – примерен ефект от приложението му

Таблица 4 представя накратко някои възможни приложения на смесеното обучение в контекста на анализа на финансови инструменти. Специфично предимство, което изследването на финансовите пазари има, е, че относително рядко се налага задължителна работа с данни, които налагат използването на смесено обучение.

Таблица 4. Алгоритми, използващи смесено обучение

Алгоритъм	***Бележки***
Генеративни модели	Специално внимание в контекста на това изследване заслужават възможните приложения на този тип модели заедно с EM алгоритъма ([21]) за оценка на разпределението на възвръщаемостта на отделни инструменти като комбинация от няколко статистически разпределения.
Смесени SVM (S3VM)	Характеристиките на смесените SVM модели позволяват да се подкрепи анализът на стартиращи компании или рискови инвестиции, при които има сериозна динамика и относително малко регистрирани случаи в определена област.
Граф модели	В настоящото изследване няма да използваме тези алгоритми, въпреки че предварителното дефиниране на връзките би могло да помогне при моделирането им в определени индустрии, а след това например и при оптимизацията на инвестиционни портфейли.

4. Обучение с утвърждаване

При този тип обучение целта е посредством последователност от „награди“, или най-общо стимули, да бъдат усвоени определени политики, които да позволят на създадения модел да реагира по най-добрия възможен начин на промени в околната среда. Така векторите с характеристики x_i са всъщност параметрите на средата, в която работи алгоритъмът, и съответно те се променят с времето.

Според начина, по който самообучаващият се алгоритъм конструира вътрешното представяне на околната среда, различаваме два вида обучение с утвърждаване:

- с предварително установени правила (модел);

- без предварително установени правила (модел).

В условията на развити финансови пазари, установени правила и теоретични финансови модели задаването на предварително установен модел е по-ефективен и оправдан подход. Даяна и Нив в ([22]) акцентират върху това, че алгоритмите без установени правила са статистически по-слабо ефективни от тези, при които предварително е зададен модел. Трябва да се отчита и фактът, че ако последният е погрешен, то това се явява сериозно ограничение за постигането на добри резултати. Препоръчително е да се направи проверка и с двата варианта, като при наличие на големи отклонения в получената точност да се разгледа подробно обосновката на използваните правила и ограничения.

Обучението с утвърждаване може да бъде от полза при симулирането на различни инвестиционни стратегии и реакции на пазара, при постъпване на нова информация. Благодарение на способността да проследява еволюцията на отделните решения, този тип алгоритми дават възможност за оценка на действия, чиито последствия се проявяват по-късно във времето.

Подробното представяне на свързаните с обучението с утвърждаване аналитични инструменти и алгоритми надхвърля целите на изследването. Поради тази причина няма да се спираме повече върху неговите възможности и недостатъци.

5. Източници на данни

Свободно достъпните данни са предпочитани при всяка от ситуациите, разглеждани в този текст. Основната причина е, че така се осигурява възможност за повторна проверка на получените резултати и постепенно разширяване на анализа. Най-често използваните източници на финансова информация са:

- Yahoo Finance, като за целта се разчита на функционалност от следните софтуерни пакети:
    - Pandas DataReader (pandas_datareader) в Python;
    - Quantmod в R.
- Quandl, като за целта се разчита на едноименния пакет в Python и R.

За макроикономически показатели се използват динамичните редове предоставяни от ЕЦБ (ECB Statistical Data Warehouse[4]) и "достъпвани" с помощта на пакетите pandaSDMX[5] в Python и ecb[6] в R. Подробна класификация и информация за отделните групи показатели може да се намери в [23], [24].

6. Изходен код и възпроизводимост на резултатите

В края на всяка част от изследването са посочени изходните кодове на R и Python, използваните източници на данни и библиотеки, които са необходими за възпроизвеждане на представените резултати. Когато са използвани данни, които не могат да бъдат директно "достъпени" онлайн, те са достъпни в същото хранилище, в което се намира изходният код.

[4] https://sdw.ecb.europa.eu/
[5] https://pandasdmx.readthedocs.io/en/latest/index.html
[6] https://cran.r-project.org/web/packages/ecb/index.html

III. Контролирано обучение при анализа на финансови инструменти

Приложенията на контролираното обучение при анализа на финансови инструменти са твърде разнообразни, за да бъдат събрани всички в рамките на това изследване. Вместо стремеж към изчерпателно представяне на всички възможности, което би повторило части от [25] и [26], в изложението е следван подход, при който се търси цялостен отговор на проблем, свързан с инвестиции на финансовите пазари. Основните предимства на този избор са свързани с това, че:

- Освен особеностите на отделните алгоритми и модели се представят и възможностите за тяхната съвместна и логически обоснована употреба.

Съвместното приложение на различни алгоритми е свързано не само с нуждата от балансиране на техните предимства и недостатъци, но и с възможностите да се отчетат по-добре индивидуалните характеристики на всеки изследван финансов инструмент.

- Изследването на решения, насочени към отделни видове инструменти и едновременно с това към различни етапи от инвестиционния процес, е предпоставка за установяване на точното съответствие между обучаващите се алгоритми и финансовата теория.

Най-често анализът на ценни книжа не е самоцел, а част от общата стратегия на участниците на финансовите пазари. Следователно използваните алгоритми трябва да улесняват реализацията на плановете на инвеститора. Това предполага, че изборът на аналитични инструменти следва да бъде съобразен освен с информацията, която получаваме от изследваните данни, също и с финансовите правила, на които е подчинена инвестиционната стратегия на участниците на пазара.

- Инвестиционните решения използват, но не се ограничават само до резултатите от анализа на възможните алтернативи.

Обвързването на получените в изследването резултати с цялостния процес по вземане на решения позволява да се разшири тяхната приложимост. Не бива да се пренебрегва и фактът, че по този начин използващите отделните алгоритми получават по-пълна представа в какъв

контекст ще бъдат прилагани техните продукти. Като отчитаме, че съществена част от практиците в областта на обучаващите се и самообучаващите се системи са по образование и призвание математици и информатици, връзката на изложението с принципите на финансовата теория може да бъде само от полза.

1. Общи условия и допускания в изследването

В последващото изложение разглеждаме желанията на един рационален инвеститор, разполагащ с ограничени средства K_0, които желае да инвестира в рамките на предварително ясен времеви хоризонт t. За да сведе броя на вариантите, с които разполага, до една разумна стойност, този homo economicus ([27], [28]) въвежда допълнителни ограничения описани подробно в Таблица 5.

Таблица 5. Допълнителни ограничения, наложени от инвеститора

Ограничение	***Обосновка***
Срокът на отделните инвестиции трябва да се намира в границите от 24 часа до t години. Максималната продължителност на разглеждания период (t) е 30 години.	Въпреки че съществуват инвестиционни възможности с продължителност извън посочения интервал, това допускане позволява да обхванем най-често търгуваните инструменти.
Инструментите, разглеждани като реална инвестиционна алтернатива, трябва да се търгуват на развити и утвърдени пазари.	Това ограничение отчасти опростява приложението на различните алгоритми, а едновременно с това позволява да се провери съответствието на получените резултати със съвременната финансова теория.
Като инвестиционна възможност се разглеждат: (а) ценни книжа с фиксиран доход, (б) обикновени и привилегировани акции и (в) дериватни инструменти.	Извън полезрението на примерния инвеститор остават алтернативите, свързани с недвижими имоти, произведения на изкуството и др. Но те по дефиниция са изключени от обхвата на изследването.
Диверсификация на пазари в различни страни и региони е	Въвеждаме това ограничение, за да намалим броя на възможните комбинации от активи и

допустима, ако техният брой не надвишава три.	необходимото време за анализ на данните ([29]). Това обаче не ограничава възможностите на разглежданите алгоритми, нито пречи при изследването на техните предимства и недостатъци.
Размерът на инвестициите е ограничен до първоначално определената стойност K_0 без изисквания за използване на ливъридж извън нормално допустимия за съответните финансови инструменти.	Целта на това ограничение е да се постигне по-лесно сравнимост между ефективността на разглежданите алгоритми при анализ на инструменти, които допускат различно използване на ливъридж.
Трансакционните разходи не се отчитат освен в случаите, където това е изрично посочено. Неявно се въвежда допускането, че ефектът от този вид разходи е пренебрежимо малък, което невинаги отговаря на реалните пазарни условия ([30]).	

2. Инструменти с фиксиран доход. Криви на доходността

Инструментите с фиксиран доход са една от основните инвестиционни алтернативи. Те са интересни дори и за участниците на финансовите пазари, които не възнамеряват да вложат средства в този тип ценни книжа, защото дават ориентир за потенциално постижимата доходност при различен времеви хоризонт. Нещо повече – като изследваме начина, по който доходността се променя с увеличаване на матуритета на инструментите с фиксиран доход, става възможно да се отговори на следните въпроси:

- Какви са преобладаващите очаквания сред участниците на пазара за бъдещето?

Зад това твърдение стои идеята, че ако пазарите са ефективни и отразяват очакванията на рационалните икономически субекти, доходността от тях за съответните периоди е най-добрият индикатор за това какво реално се случва в икономиката.

- Какви промени може да очакваме в кредитирането за съответните бъдещи периоди?

Връзката с бъдещото кредитиране е индиректна, но много важна. Нека разгледаме една търговска банка, която си осигурява ресурс, предлагайки относително краткосрочни депозити и предоставяйки заеми с обичайно по-дълъг срок. При равни други условия тази институция би имала икономически стимул да разшири кредитирането, ако положителната разлика между дългосрочната и краткосрочната доходност нарасне.

Конкретни отговори на поставените въпроси могат да се получат, като изследваме зависимостта между доходност и матуритет. Графичното представяне на тази зависимост е известно като крива на доходността и представлява един от основните аналитични инструменти, използвани от участниците на пазарите на ценни книжа с фиксиран доход. Построяването на тази зависимост изглежда като тривиална задача, защото информацията за очакваната доходност (при определено равнище на риск) за съответните периоди е лесна за наблюдение и е достъпна. Динамиката на пазарите води до честа промяна във формата на кривите на доходността и извежда на преден план нуждата от модел, който да обясни характера на наблюдаваните изменения. Ролята на самообучаващите се алгоритми в този контекст е свързана преди всичко с решаването на две задачи:

- да подпомогнат избора на най-подходящия за съответния пазар и инструмент модел;
- да оценят параметрите на избрания модел и анализират промените, които са настъпили с тях във времето.

И в двата случая има контролирано обучение, тъй като разполагаме както с входните данни (матуритет на съответните ценни книжа), така и с желаните резултати (доходност).

Практиката показва, че зависимостта между доходността и матуритета не е линейна. Моделите, които се опитват да обяснят каква точно е тя и защо кривите на доходността имат точно формата, която наблюдаваме, могат да се разделят на три категории ([31]):

- Регресионни модели, които съпоставят доходността на купонови облигации с техния матуритет.

Получените резултати след това могат да бъдат отнесени и към облигациите с нулев купон, но страдат от един съществен недостатък –

купонови облигации с един и същи матуритет може да имат различна доходност поради разлики в размера на купоновите плащания.

- Емпирични модели, които се опитват да преодолеят недостатъците на предходната група посредством дефиниране на функционална зависимост, отразяваща дисконтираните купонови плащания.

Специфично предимство на тази категория модели е, че могат да се използват различни функционални зависимости. Това позволява да се търсят решения, които имат различна степен на сложност.

- Равновесни модели, които се стремят да обяснят промените в изследваните криви на доходността, като ги обвържат с икономически фактори и промените в тях.

Въпреки че този тип решения са привлекателни от гледна точка на резултатите, тип „формула от затворен тип“, те не са толкова гъвкави и по-трудно могат да се адаптират към специфични промени, характерни само за определен период от време.

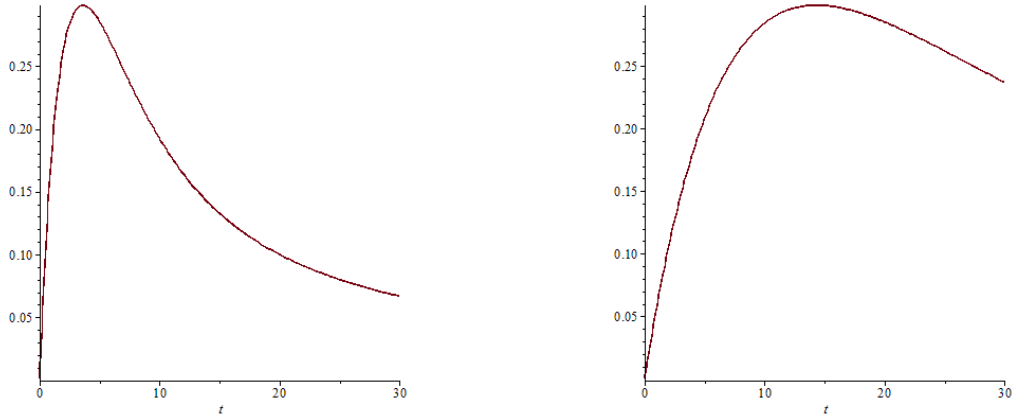
2.1. Модел на Нелсон–Сиегел–Свенсон

Моделът съществува в няколко разновидности, като интерес за изследването представляват преди всичко оригиналната му формулировка в [32], както и разширението на Свенсон в [33]. Параметрите на зависимост (4) са общо четири – β_0, β_1, β_2 и γ, като всеки един от тях обяснява влиянието на специфични фактори върху доходността.

$$r(t) = \beta_0 + \beta_1 \left[\frac{1 - e^{-t/\gamma}}{t/\gamma}\right] + \beta_2 \left[\frac{1 - e^{-t/\gamma}}{t/\gamma} - e^{-t/\gamma}\right]. \qquad (4)$$

Така например първият параметър – β_0, не зависи от матуритета (времето t) и следователно може да се приеме за израз на дългосрочната очаквана доходност от инвеститорите. β_1 е свързан с краткосрочните ефекти, оказващи влияние върху доходността, защото с нарастване на t стойността на $\left[\frac{1 - e^{-t/\gamma}}{t/\gamma}\right]$ намалява много бързо (експоненциално). Третият параметър –

β_2, също е функция на времето, но благодарение на факта, че $\left[\frac{1-e^{-t/\gamma}}{t/\gamma} - e^{-t/\gamma}\right]$ се променя по начина, показан на Фигура 2, може да се приеме за израз на средносрочните икономически ефекти върху доходността.



Фигура 2. Графично представяне на $\left[\frac{1-e^{-t/\gamma}}{t/\gamma} - e^{-t/\gamma}\right]$ *и* γ*=2 (вляво), и* γ*=4 (дясно),*

Последният параметър – γ, определя относителното значение на всеки един от посочените ефекти – краткосрочни, средно- и дългосрочни. Визуално това се изразява във формата на кривата и „изместването“ на максимума във времето, както е показано на Фигура 2. Точната оценка на всички параметри на (4) позволява да се решат следните проблеми:

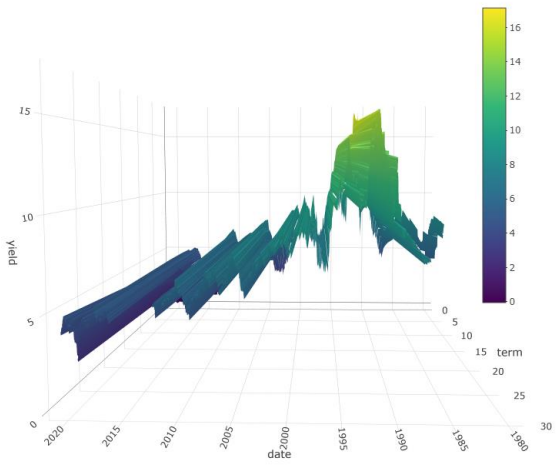
- да се определи кой от тях има най-висока стойност към определен момент, като по този начин на практика оценим преобладаващото мнение сред участниците на пазара;
- да проследим как стойностите на всеки параметър се променят с времето, като по този начин проследим еволюцията в очакванията на инвеститорите;
- да прогнозираме формата на кривата, както и очакваната доходност в бъдещи периоди.

В разширението на модела, предложено от Свенсон са добавени още два параметъра – γ_2 и β_3, а вместо γ обозначаваме съответния параметър от модела на Нелсон-Сиегел с γ_1:

$$r(t) = \beta_0 + \beta_1\left[\frac{1 - e^{-t/\gamma_1}}{t/\gamma_1}\right] + \beta_2\left[\frac{1 - e^{-t/\gamma_1}}{t/\gamma_1} - e^{-t/\gamma_1}\right] + \beta_3\left[\frac{1 - e^{-t/\gamma_2}}{t/\gamma_2} - e^{-t/\gamma_2}\right]. \quad (5)$$

Смисълът на новите параметри може да бъде проследен по същия начин, както и този за β_2 и γ. Оценка за стойностите на тези параметри може да получим по различни наични, като например MLE или минимизиране на $\sum(r(t) - \widehat{r(t)})^2$, при предварително зададени ограничения за β_0, β_1, β_2, β_3, γ_1 и γ_2.

Фигура 3 представя историческите промени в доходността на ДЦК на САЩ за последните малко повече от 40 години. Ако трябва да отчитаме тези изменения за книжа с различен матуритет, тогава визуалното представяне не е добър инструмент: първо – защото е необходимо да работим с триизмерно изображение (период, доходност, матуритет), и второ – защото е много трудно да отделим краткосрочната динамика. Посочените проблеми могат да бъдат преодолени, ако разгледаме освен историческите данни за доходността и промените в параметрите на моделите на Нелсон–Сиегел (NS) и Нелсон–Сиегел–Свенсон (NSS).



Фигура 3. Исторически промени в доходността на ДЦК на САЩ

На Фигура 4 са представени промените в коефициентите на NSS модела, оценени с данни за доходността на ДЦК на САЩ в периода януари 1981 – юни 2020 г. Освен факта, че получените стойности показват съществени

изменения през изследвания период, допълнително внимание заслужават следните резултати:

- Ако сравним коефициента на вариация (CV, RSD[7]) на отделните параметри, най-голяма е стойността му при β_2, следван от β_1.

Този ефект не е необичаен, защото коефициентите са свързани с краткосрочните очаквания, които инвеститорите имат. В подкрепа на горното твърдение може да се посочи, че през кризисната 2008 г. абсолютната стойност на коефициента на вариация на β_1 нараства значително спрямо стойността за целия наблюдаван период ([34]).

- Както може и да се очаква, относителните колебания са най-малки при β_0, измерващи промените във влиянието на дългосрочните фактори.

Оценките на отделните параметри за представени в Таблица 6.

Таблица 6. Промени в параметрите на NSS модела

Показател	***Стойност***
януари 1981 г. – юни 2020 г.	
CV(β_0)	0.8148377
CV(β_1)	–5.752443
CV(β_2)	–23.73349
CV(β_3)	2.559231
януари 2008 г. – януари 2009 г.	
CV(β_0)	4.403384
CV(β_1)	20.58448
CV(β_2)	16.5954
CV(β_3)	0.8779798

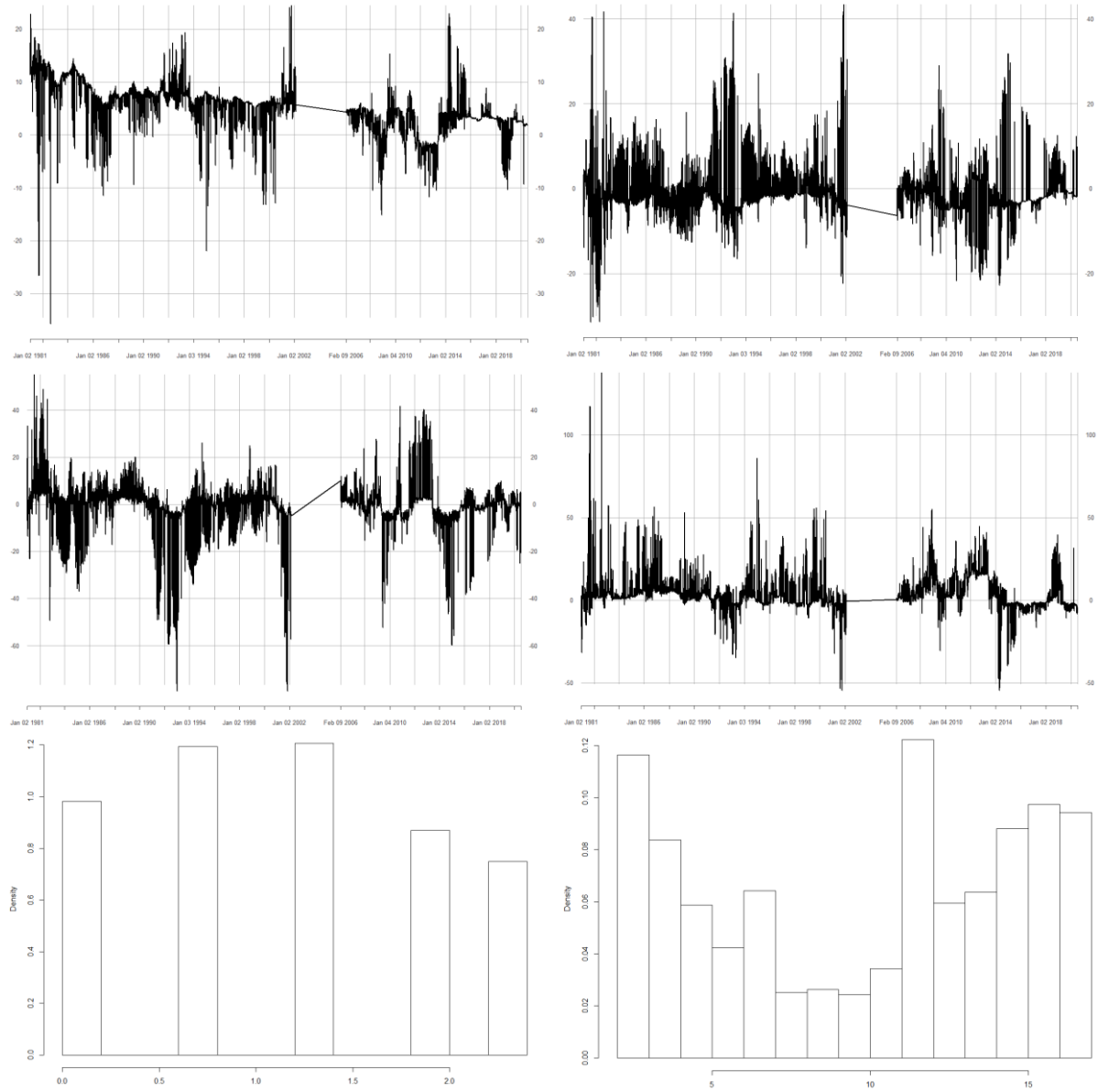
- Хистограмите в последния ред на Фигура 4 показват значително по-голямо разнообразие в стойностите на параметъра γ_2 спрямо γ_1.

Причината е, че γ_2 участва в последното събираемо в NSS и съответно показва ефектите на промяна във формата и „извиването“ на кривата на доходността.

[7] Relative Standard Deviation.

Стъпките за контролирано обучение в конкретния случай могат да бъдат представени като стандартна оптимизационна процедура, а интервалите с възможни стойности на γ_1 и γ_2 са определени предварително. Критерият за избор на решение е свързан с минимизиране на квадрата на грешките (остатъчните стойности).

Тъй като функцията в NS и NSS моделите е предварително известна, то може директно да се търси нейният екстремум. Когато диференцирането е проблем, могат да се използват възможните решения, описани в ([35]).



Фигура 4. Промени в параметрите на NSS за периода 1981 - 2020 г. β_0 (първи ред, ляво), β_1 (първи ред, дясно), β_2 (втори ред, ляво), β_3 (втори ред, дясно), хистограма на γ_1 (трети ред, ляво) и на γ_2 (трети ред, дясно).

2.2. Контролирано обучение и анализ на корпоративни облигации

Пазарът на корпоративни облигации много често играе ролята на „резервен“ или „втори“ избор, когато определяме изследователските си приоритети. Както показват стойностите от Фигура 5, подобен подход е необоснован и не би позволил да се анализират редица интересни финансови проблеми. Ако приемем предизвикателството да изследваме корпоративни облигации с помощта на самообучаващи се алгоритми, трябва да предложим отговори на следните въпроси:

- Как да се интегрират специфичните характеристики за отделните емисии облигации в отделните етапи на анализа?

Специфичните характеристики, които трябва да бъдат отчетени, попадат в следните категории, като всяка от тях на практика налага отделни ограничения върху средствата, които можем да използваме:

  - ликвидност на пазара и потенциални проблеми, свързани с времето, необходимо за приключване на трансакция при определена цена;

  - валидност на обявените последни котировки, които при липса на активна търговия може да се отнасят за периоди като например няколко дни назад.

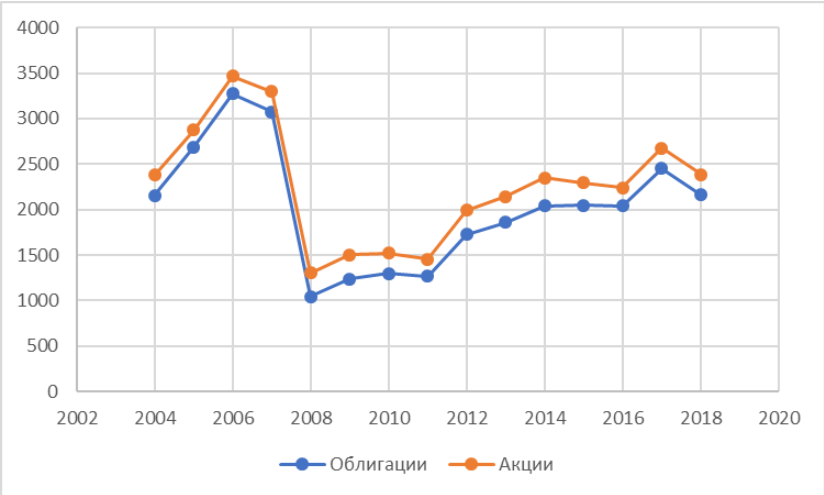
- Как да отчетем особеностите в начина на търгуване на облигации?

Особеностите на търговията могат да бъдат както чисто технически, така и с отношение към ликвидността и ефективността на пазара. Например, TRACE системата на FINRA има ограничение, че трансакциите с всички облигации, отговарящи на нейните изисквания, трябва да бъдат обявени в рамките на 15 минути от сключването им[8]. В рамките на изследването тези ограничения не са критични, тъй като данните и моделите не изискват незабавна реакция. Освен това, както посочва обзорният доклад на FINRA, над 80% от трансакциите се обявяват много по-бързо ([36]).

[8] https://www.finra.org/sites/default/files/TRACE_Overview.pdf

- Как могат да се отчетат при анализа на доходността специфичните рискове и показателите за тях?

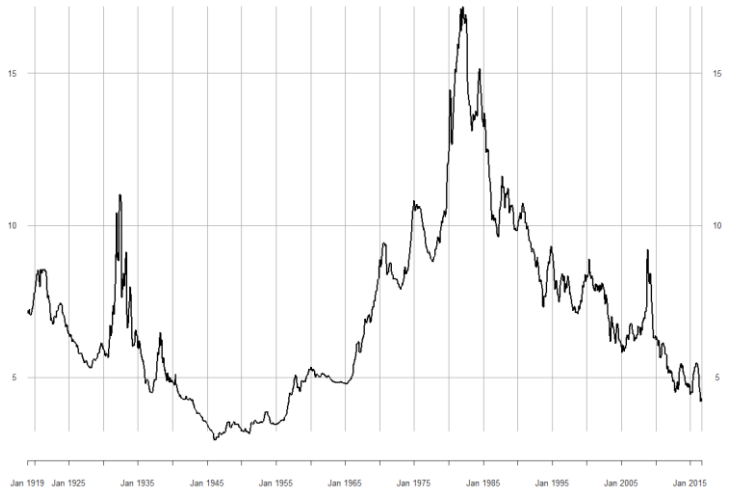
Специфичните рискове могат да бъда отчетени различно, но най-лесният за приложение начин е свързан с комбинираното изследване на облигации с един и същи кредитен рейтинг. Този подход не елиминира възможността за грешки, свързани с некоректно определен рейтинг, но позволява да се опрости много селекцията на ценни книги и предварителната обработка на данните.



Фигура 5. Обем на новите емисии корпоративни облигации и акции в САЩ в млрд. USD ([37])

За целите на това изследване ще бъдат използвани данните за възвръщаемостта на корпоративни облигации с кредитен рейтинг BAA, определен от Moody's. Данните са достъпни след регистрация в Quandl[9] с идентификатор *FED/RIMLPBAAR_N_M* и са представени на Фигура 6. Динамичният ред има достатъчно дълга история, която включва и последната голяма финансова криза – от 2008 г. Това е предпоставка да се изследват качествата на използваните аналитични инструменти в различни етапи от развитието на пазара.

[9] https://www.quandl.com



Фигура 6. Исторически данни за Moody's yield on seasoned corporate bonds (Рейтинг: BAA)

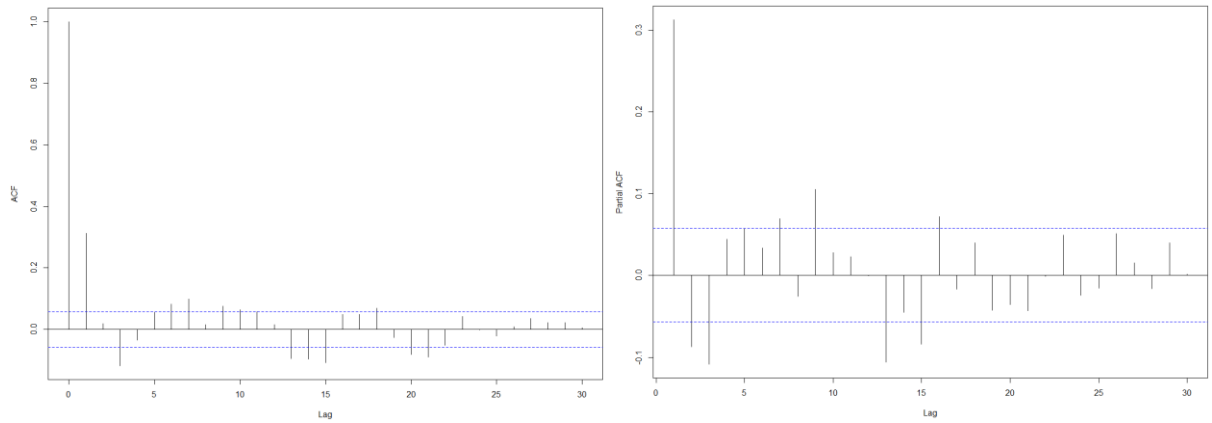
Използването на готов индекс не позволява да направим анализ дали отделните дългови ценни книжа са подценени или надценени. Това ни лишава от възможността да изградим инвестиционни стратегии, основани на по-точната ни оценка на тези облигации, които да донесат печалба на инвеститора от точка 1 на стр. 28. За сметка на това е възможно да намерим изгодни възможности, основани на коректното прогнозиране на бъдещите изменения в доходността на дълговите ценни книжа с рейтинг BAA. Това свежда решаваната задача до изследване и прогнозиране на динамичния ред с доходността на индекса и дава възможност да отговорим на два важни въпроса:

- дават ли определени предимства самообучаващите се алгоритми спрямо други, традиционни аналитични инструменти за моделиране на динамични редове;
- в какво се изразява цената, която трябва да платим, за да използваме самообучаващите се алгоритми – време за подготовка и изчисления.

Не е възможно в рамките на това изследване да сравним всички традиционни средства за анализ на динамични редове с избраните самообучаващи се алгоритми, за да отговорим на първия въпрос. Поради тази причина ще използвам един от най-често използваните модели – ARIMA, като следваните стъпки могат да бъдат повторени и за други модели.

Проверката за стационарност с разширения тест на Дики – Фулър показва, че е необходимо да работим с първите разлики на разглеждания

динамичен ред (Dickey – Fuller = -1.9691, Lag order = 10, p-value = 0.5914 за оригиналните стойности; Dickey – Fuller = -8.1771, Lag order = 10, p-value < 0.01 за първите разлики на динамичния ред).



Фигура 7. Автокорелация (ляво) и частична автокорелация (дясно) на първите разлики в доходността на облигациите с рейтинг BAA

Стойностите на автокорелацията и частичната корелация са допълнителен ориентир за параметрите на ARIMA модела. След проверка по стойността на информационния критерий на Акайке оптималната размерност е ARIMA(2,1,2). В Таблица 7 са показани различни критерии за точност на авторегресивния модел с движещи се средни, като целта е не само да има повече от един показател за сравнение със самообучаващия се алгоритъм, а и защото видовете грешки имат различно отношение към отделните инвестиционни стратегии.

Таблица 7. Характеристика на оптималния ARIMA модел

Параметър	***Стойност***
AIC	–341.23
AICc	–341.18
BIC	–315.9
ME	–0.001883951
RMSE	0.2081958
MAE	0.119313
MPE	–0.05745414
MAPE	1.602857
MASE	0.9308054

Общият вид на модела след оценка на параметрите му е представен в (6), като в скобите са зададени стандартните грешки за отделните параметри.

$$r_{t_n} = \underset{(0.1658)}{0.4750} r_{t_{n-1}} - \underset{(0.0815)}{0.4193} r_{t_{n-2}} - \underset{(0.1701)}{0.1443} \gamma_{t_{n-1}} + \underset{(0.0769)}{0.3236} \gamma_{t_{n-2}} \qquad (6)$$

За да сравним поведението на различни самообучаващи се алгоритми, ще проверим точността на два модела с невронни мрежи:

- MPL (Multilayer perceptron) модел ([38]);
- ELM (Extreme learning machine) модел ([39]).

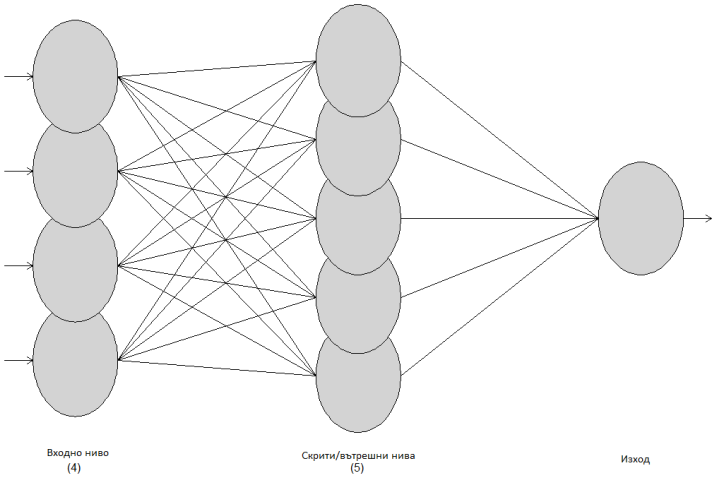
Причините да бъдат избрани точно тези възможности, са свързани с това, че MPL и обучението с обратно разпространение на грешката (backpropagation) са изключително широко използвани. Това е един от най-добре познатите варианти за контролирано обучение на невронни мрежи. ELM моделът е включен с основната идея да покаже как биха изглеждали точността и времето, необходимо за калибриране на една по-проста схема, защото често се налага да разрешаваме проблеми, които имат не само функционални, но и времеви ограничения. В тези случаи всяка хилядна от секундата е от значение и вниманието ни по необходимост се насочва към по-малко ресурсоемките алгоритми.

За да бъде сравнението обективно, то е разделено на две части – сравнение на постигнатата точност в процеса на обучение, който включва всички налични данни, и оценка на точността при разделяне на данните на две части – за обучение и тестово множество. Представените в Таблица 7, Таблица 8 и Таблица 9 стойности се отнасят към първата част на проверката, докато дискусията в точка 2.4. разглежда по-реалистичната втора част.

Като входни данни в MLP и ELM модела са използвани само стойностите от разглеждания динамичен ред, без да се добавят икономически показатели от какъвто и да е вид. По този начин може да се направи съпоставка с първоначалния ARIMA модел. Активиращите функции и в двете невронни мрежи са сигмоидни ($f(x) = \frac{1}{1+e^{-x}}$), като не са правени опити с друг вид активиращи функции. Това ограничава възможностите ни при анализа на динамичния ред, но позволява да сравняваме времето необходимо за обучение – в случая – с MLP и ELM.

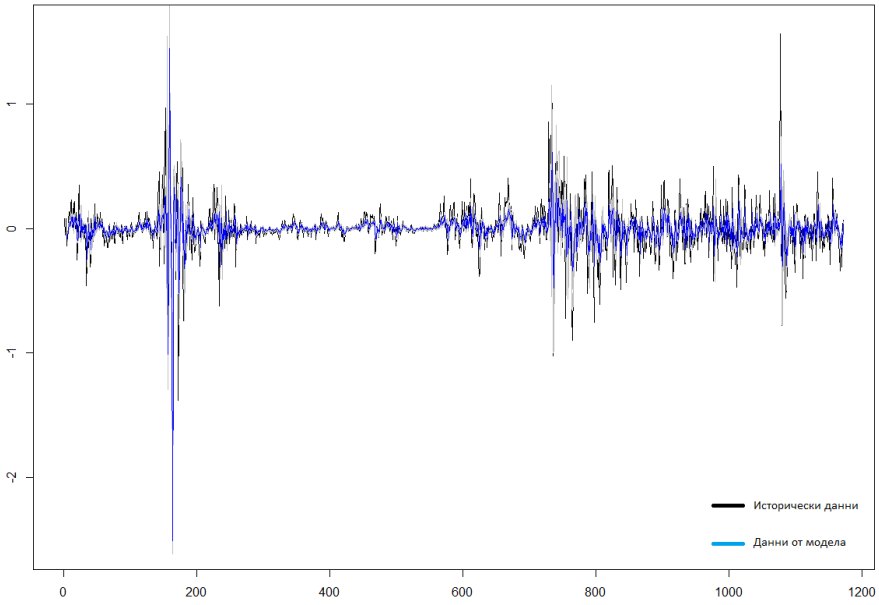
Мрежата, използвана за MLP модела, е представена на Фигура 8, като броят на невроните е получен с помощта на оптимизационна процедура и сравняване на различни варианти, но с ограничение за брой входни

елементи, които са близки до допустимите размерности на ARIMA. Последното условие е ограничаващо, но е необходимо, за да може да поставим отделните варианти при относително равностойни условия. В крайна сметка е избрана тази топология на мрежата, която ни води до най-малка грешка. Трябва да се отбележи, че този подход не би бил подходящ при по-голям брой възможности, тъй като на практика се калибрират и сравняват голям брой мрежи. За целите на конкретното изследване обаче той е достатъчно бърз и може да бъде прилаган без нужда от специална техника.



Фигура 8. Невронна мрежа, използвана за MLP модела

Резултатите от процеса на обучение са показани в Таблица 8, като не са изрично изчислявани показателите за информационна стойност. Сравнението на отделните изследвани модели може да се направи спрямо стойността на показателите за грешка. В този случай основното сравнение е между ME, RMSE и MAE.



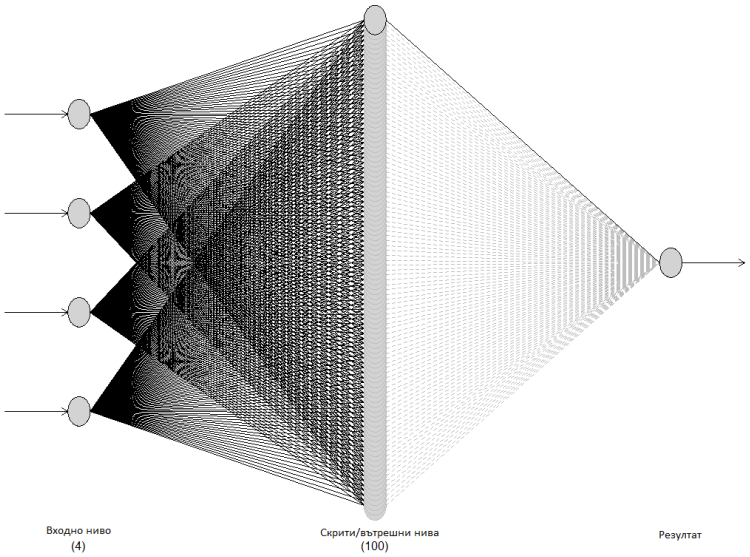
Фигура 9. Исторически и прогнозни данни за MLP модела

И по трите показателя MLP моделът демонстрира по-добро представяне спрямо ARIMA. Това е валидно спрямо поставените условия и при положение че правим обучението и прогнозирането върху всички данни, с които разполагаме. По-точна картина може да се получи при сравняване на поведението на различните подходи върху „нови“ данни, които не са използвани в процеса на контролирано обучение. Точно такъв тип сравнение е предмет на дискусия в част 2.4.

Таблица 8. Характеристики на оптималния MLP модел

Параметър	***Стойност***
ME	–0.07281846
RMSE	0.1757413
MAE	0.0899
MPE	–0.186024
MAPE	1.228387

Топологията използвана за ELM модела, е представена на Фигура 10 и изцяло съответства на класическото определение като система с единично скрито ниво.



Фигура 10. Невронна мрежа, използвана за ELM модела

Резултатите представени в Таблица 9, очаквано показват, че стойността на грешката е по-висока, отколкото при MLP модела.

Таблица 9. Характеристики на разглеждания ELM модел

Параметър	***Стойност***
ME	–0.01182363
RMSE	0.2220233
MAE	0.3391743
MPE	–0.001951198
MAPE	0.04557141

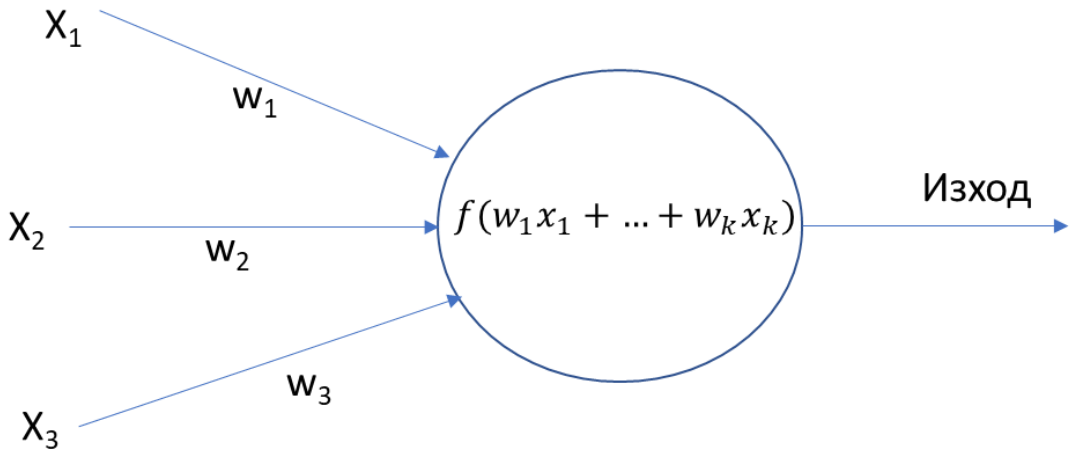
2.3. Използвани в обучението алгоритми

Съществуват различни подходи, с чиято помощ можем да определим какви да бъдат относителните тегла на връзките между отделните елементи на невронната мрежа. В разгледаните до момента модели се използват два алгоритъма за контролирано обучение.

1) С обратно разпространение на грешката (backpropagation).

Този алгоритъм има дълга история ([40], [41]), която заедно с демонстрираните качества и универсалност са предпоставка за превръщането му в най-широко използвания аналитичен инструмент за обучение на невронни мрежи. Основната идея е с помощта на градиентно

спускане (gradient descent) да се адаптират относителните тегла на връзките между отделните елементи на мрежата, така че да се достигне минимална стойност на предварително дефинираната функция на загубите (респ. да се минимизират грешките). От гледната точка на единичен елемент този процес изглежда по начина, показан на Фигура 11.



Фигура 11. Принцип на работа на отделен елемент в невронната мрежа

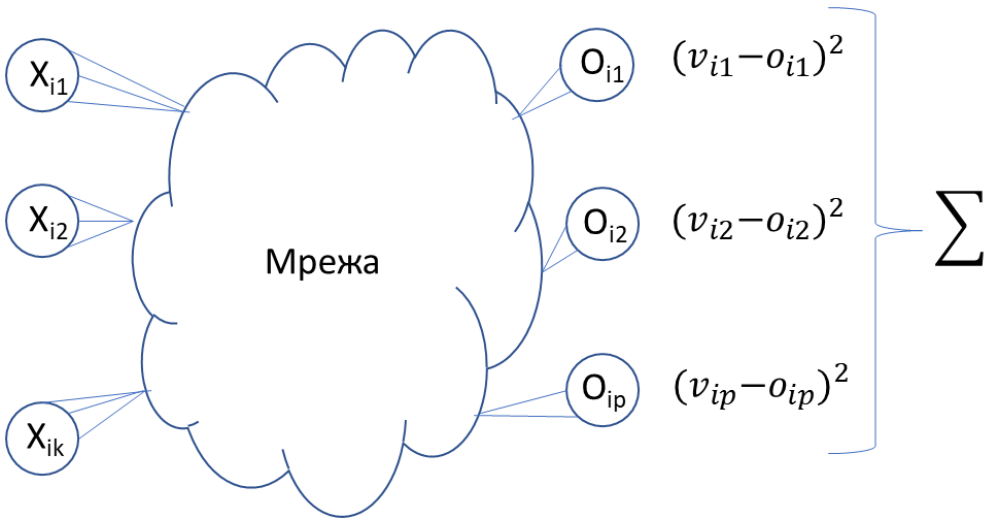
Тъй като алгоритъмът разчита на градиента на функцията на загубите (респ. грешките), това означава, че тази функционална зависимост трябва да отговаря на две условия – да бъде непрекъсната и да е диференцируема. Това е причината една от предпочитаните зависимости за $f(w_1x_1 + ... + w_kx_k)$ да е сигмоидната функция.

$$f(x) = \frac{1}{1+e^{-(x)}}, \quad (7)$$

$\frac{df(x)}{dx} = f(x)(1 - f(x)) = \frac{e^{-x}}{(1+e^{-x})^2}$, откъдето

$$f(w_1x_1+\ldots+w_kx_k) = \frac{1}{1 - e^{-(w_1x_1+\ldots+w_kx_k)}} \quad (8)$$

Ако в мрежата има само един елемент, то тогава целта на алгоритъма ще бъде да минимизираме грешката, която се получава, като изходът се сравни с очакваната стойност – т.е. $E = e(\|v - f(w_1x_1+\ldots+w_kx_k)\|)$, където v е предварително известната стойност и е част от осъществяваното контролирано обучение. При наличие на повече от един елемент, отразяването на грешката е представено на Фигура 12, където v_{ij} е реална стойност на резултат j за наблюдение i, а o_{ij} е резултатът от работата на модела. В анализа на облигациите j има стойност 1, тъй като от значение е само каква би била стойността на съответния индекс.



Фигура 12. Отразяване на грешката при повече от един неврон

Алгоритъмът търси минимум на функцията на загуба (9). Тъй като o_{ij} зависят както от входните стойности x_{ij}, така и от отделните теглови коефициенти на връзките между невроните, отразяването на грешките е свързано с промяна на тези тегла:

$$E = \frac{1}{2}\sum_{j=1}^{p}\left\|v_{ij} - o_{ij}\right\|^2 \qquad (9)$$

Стъпките по промяна на теглата, като отчитаме необходимостта да минимизираме (9), са:

1) определяме градиента на функция E: $\nabla E = (\frac{dE}{dw_1}, \frac{dE}{dw_2}, \ldots, \frac{dE}{dw_l})$;

2) коригираме всеки един от тегловите коефициенти $\Delta w_i = -\gamma \frac{dE}{dw_i}$, където коефициентът γ е допълнителен параметър на процеса на контролирано обучение;

3) повтаряме отново, докато не достигнем определеното гранично условие.

При достигането на минимума на E, $\nabla E = 0$ и последващи корекции на тегловите коефициенти не биха довели до промени в точността на получените резултати.

След като три стъпки се прибавят към първоначалната обработка на данните (необходима, за да се получи o_{ij}), ще се получи и общият вид на обучаващия алгоритъм:

а) изчисляване на крайния резултат $O_i = (o_{i1}, \ldots, o_{ip})$ за данните $X_i = (x_{i1}, \ldots, x_{ik})$;

б) изчисляване на грешката съобразно с израз (9);

в) изчисляване на градиента за изходното ниво;

г) изчисляване на градиента с връщане назад за скритите нива;

д) корекция на всички теглови коефициенти;

е) повторение на стъпките, докато достигнем определените гранични условия.

Последното твърдение обикновено е свързано с две неща, определени като гранични условия – или докато се премине през предварително дефиниран брой итерации с наличните данни, или докато достигнем ситуация, при която грешката не намалява. Разбира се, може да има и други предпоставки за прекратяване на обучението, но тези две са най-често срещани.

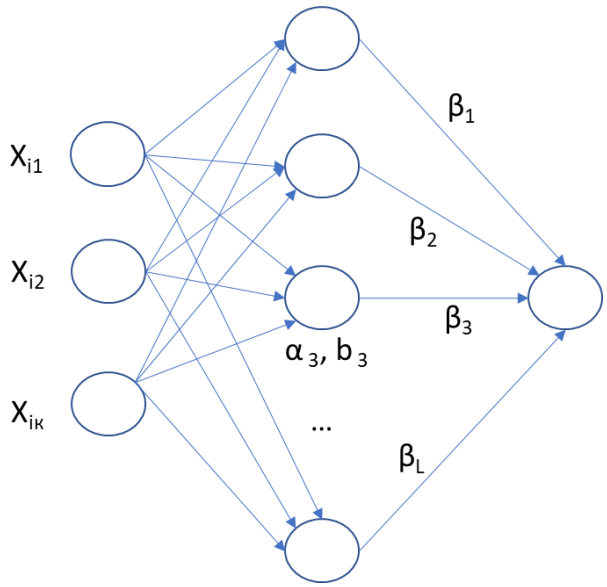
2) ELM алгоритъм с право (директно) разпространение на информацията (feedforward).

ELM мрежите се стремят да разрешат редица проблеми, които съпътстват използването на обучението с обратно разпространение на грешката, като например попадането в локален екстремум (Фигура 13), необходимите сериозни изчислителни ресурси и способността за „прекомерно нагаждане“ (overfitting) към характеристиките на наличните данни.



Фигура 13. Локален и глобален минимум за $\mathrm{E} = \mathrm{e}(\|v - f(w_1 x_1 + \ldots + w_k x_k)\|)$

За да постигнат целите си, ELM мрежите използват алгоритъм от три стъпки, като всеки един от свързаните неврони се представя с активиращата си функция, тегла на връзките с елементите от входния слой (α), с елементите от изходното ниво (β) и значимост за елементите от скритото ниво (b). Това е възможно, при условие че в рамките на модела има едно скрито ниво и топологията съответства на показаната на Фигура 14.



Фигура 14. Топология на ELM мрежа (едно скрито ниво)

Като използваме обозначенията, свързани с предходния пример, но с различна активираща функция за отделните елементи – $g(x, \alpha, b)$, ще получим:

$$o_i = \sum_{j=1}^{L} \beta_j g(x_i, \alpha_j, b_j). \qquad (10)$$

Ако използваме матричен запис, тогава (10) може да се представи като:

$$\begin{bmatrix} o_1 \\ o_2 \\ \vdots \\ o_N \end{bmatrix} = \begin{bmatrix} g(x_1, \alpha_1, b_1) & \cdots & g(x_1, \alpha_L, b_L) \\ \vdots & \ddots & \vdots \\ g(x_N, \alpha_1, b_1) & \cdots & g(x_N, \alpha_L, b_L) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_L \end{bmatrix}, \text{ или} \qquad (11)$$

$$O = G\beta. \qquad (12)$$

Зависимост *(12)* дава възможност да се представи в съкратена форма целият алгоритъм на обучение в три стъпки:

1) Инициализират се по случаен начин тегловите коефициенти α_i и стойностите на значимост b_i за всеки един от елементите на скритото ниво ($i = 1..L$).
2) Изчисляваме матрицата на G, като използваме активиращата функция g и стойностите от стъпка 1.
3) Обновяваме тегловите коефициенти β съгласно с $\beta = G^g O$. G^g е псевдо-обратната матрица на G (обобщено обратна матрица, която отговаря на четирите условия на Пенроуз ([42])).

Трите стъпки в алгоритъма водят до това, че обучението протича много по-бързо, отколкото в случая с обратно разпространение на грешката. Освен това, тъй като не се използва градиентното спускане, отпада изискването активиращата функция да бъде диференцируема.

2.4. Сравнение на точността на използваните алгоритми

Сравнението на използваните алгоритми може да се направи по няколко критерия. Кой точно от тях ще има решаващо значение за избора ни на аналитичен инструмент, зависи и от допълнителните ограничения на решавания проблем. В рамките на това изследване няма изисквания за бързодействие, което означава, че водещ показател е постигнатата точност, измерена със съответните показатели за грешка.

Последното твърдение, въпреки че изглежда логично и недвусмислено, на практика поражда редица проблеми. Те попадат в една от следните групи.

- Проблеми свързани с избор на конкретен показател за измерване на грешката.

Изборът на показател за оценка на грешката заслужава отделна част от изследването. За момента е достатъчно да се отчете, че не може да става дума за „перфектна“ метрика, без да знаем предварително за какви цели ще използваме нашия модел.

- Проблеми, свързани с разнопосочни сигнали при различните показатели за грешка.

В контекста на предходното твърдение, ако един показател не може да претендира за абсолютно предимство пред останалите, е логично да се опитаме да използваме няколко различни метрики. От една страна това ще ни позволи да комбинираме предимствата на различните метрики, но от друга – поставя въпроса за избор, при положение че някои от тях отдават предпочитания на един алгоритъм и модел, а други – на алтернативни възможности.

За разрешаването на този тип противоречия е уместно да се приложат следните подходи:

  - предварително да бъде избрана водещата метрика и въз основа на нейните стойности да се направи избор на оптимален алгоритъм и модел;
  - да се дефинира функция на загубите, която може да измерва както информационните щети, така и финансовите вреди. В зависимост от вида на тази функционална зависимост може да бъде направен обоснован избор на модел.
- Проблеми, свързани с несиметрично отражение на грешките спрямо постигнатите крайни резултати.

Несиметричното отражение на грешките спрямо постигнатите финансови резултати е функция на това, че инвестиционните стратегии и съответно загубите/печалбите не са еднакво чувствителни на грешки, при които „надхвърляме“ истинската стойност, и такива, при които получаваме по-малки от нея резултати. До каква степен този ефект ще даде отражение върху крайното решение, зависи преди всичко от това по какъв начин са формирани позициите. Изразено с финансова терминология, това ни води към един значително по-важен въпрос – дали е уместно да използваме измерители на риска, които се отнасят по един и същи начин към неточностите (напр. стандартното отклонение), или да прибегнем към показатели, отчитащи асиметрията (напр. CVaR и VaR).

Асиметрията в постигнатите финансови резултати означава, че ще сме принудени да използваме и съответните специфични метрики за грешка. Ако приемем, че стандартното отклонение е достатъчно добър измерител на риска, тогава бихме могли да се задоволим и със стандартните MAE и RMSE.

Таблица 10. Бързодействие на сравняваните алгоритми

Алгоритъм	***Време за обучение и калибиране***
ARIMA	0.247931 секунди
MLP	477.66648 секунди
ELM	45.61625 секунди

Таблица 10 съдържа измерените времена за обучение върху едни и същи входни данни, но с използването на различни алгоритми. Напълно очаквано най-бързият алгоритъм е ARIMA, като времетраенето на обучителният процес е около ¼ от секундата. С много голяма разлика след него е ELM мрежата, която изисква почти 180 пъти по-дълъг период. Последен, а следователно най-неефективен алгоритъм от гледна точка на бързодействие, е MLP мрежата. В този случай времето за обучение е около 8 минути, което не е изненадващо, особено като се има предвид, че в рамките на MLP модела освен всичко друго беше извършена и неколкократна проверка за най-добра топология на изградената система.

Преди да представим функционалната точност на изследваните варианти, трябва да направим едно важно уточнение, свързано със стойностите в Таблица 10. Преди всичко представеното бързодействие се базира на стандартна реализация на всички алгоритми. Не са използвани възможностите за паралелизация на работата на невронните мрежи, което би ускорило много тяхното обучение. Вторият фактор, който трябва да бъде отчетен, е свързан с това, че подготовката не е процес, който се повтаря непрекъснато. Според това колко често трябва да бъдат отново обучени мрежите, може да се окаже, че необходимото време е без значение – защото тренирането на модела се случва, преди да започне същинската му работа. В заключение може да обобщим, че изчислителната ефективност е много важен компонент, формиращ решението кой алгоритъм да бъде използван, но не трябва да се абсолютизира.

За да е пълно сравнението между отделните алгоритми, е формирана тестова извадка, която съдържа 17.1% от общия брой случаи. Всички елементи от проверовъчното множество са взети от края на динамичния ред. Алгоритмите са обучени (със същата размерност и структура), като е направена прогноза, покриваща периода на тестовата извадка. Резултатите са представени в Таблица 11.

Таблица 11. Показатели за грешка на сравняваните алгоритми

Показател	***ARIMA***		***MLP***		***ELM***	
	Обучаващо	*Тестово*	*Обучаващо*	*Тестово*	*Обучаващо*	*Тестово*
ME	-0.001884	0.0011584	-0.0003975	0.00075414	-0.0002911	0.002052
RMSE	0.2081958	0.2169027	0.1771316	0.1706527	0.2220233	0.230142
MAE	0.119313	0.196517	0.1324795	0.1041684	0.1281466	0.020852

Представянето на отделните аналитични инструменти не се различава драматично, което се дължи на факта, че имаме относително кратък прогнозен хоризонт и не сравняваме абсолютните постигнати финансови резултати, а само показателите за грешка. Въпреки това представените резултати са средство, с което да подчертаем следните ползи, които приложението на самообучаващи се алгоритми може да донесе при анализа на дългови ценни книжа:

- Като допълнителен аналитичен инструмент, който може да обхване и нелинейни зависимости.

Допускането за нелинейни зависимости е важно, защото позволява да се отговори по-точно на редица икономически процеси, които наблюдаваме в действителност. Ако просто го пренебрегнем, рискуваме да получим сериозни отклонения, когато инвестиционният ни хоризонт се разширява.

- Като модел, който може да бъде сравнително лесно допълнен, за да включи нови променливи.

При необходимост да разширим анализа на изследваните процеси с включване на нови променливи, може да използваме и VAR/VARMA модели ([43]). Промените обаче биха били много по-лесни за първоначална проверка, ако използваме MLP или ELM, отколкото ако се налага да заменим ARIMA модела.

- Като инструмент, който може да се използва за проверка за наличие на определени зависимости (шаблони), които се разпростират в по-дълъг период от време.

Моделирането на дългосрочни зависимости във времето обикновено се реализира с помощта на рекурентни невронни мрежи и по-различни алгоритми за обучение. Но то е реална възможност, която може да подпомогне икономическите ни анализи.

3. Риск от неплатежоспособност. Дискриминантен анализ и самообучаващи се алгоритми

Освен за анализ на динамични редове контролираното обучение може да бъде успешно използвано и за решаване на класификационни задачи. В контекста на направения анализ на корпоративни облигации, тази функционалност може да бъде използвана с цел:

- да се оцени рискът от изпадане на емитента в неплатежоспособност;
- да се провери дали определеният кредитен рейтинг съответства на този, определен на сходни компании, и ако има разлики, да се акцентира върху факторите, допринасящи за тях.

Приведеният списък с приложения в никакъв случай не може да претендира за изчерпателност. Но е достатъчен, за да изследваме дали самообучаващите се алгоритми могат да подобрят оценките на безспорно важния риск от неплатежоспособност.

3.1. Модел на Алтман

Съществуват разнообразни модели за анализ на възможността от фалит ([44], [45], [46]), като в рамките на това проучване ще използваме класическия анализ на Алтман ([10]).

Като използва множествен дискриминантен анализ (MDA) и съвкупност от шестдесет и шест фирми, половината от които са в състояние на неплатежоспособност, Алтман конструира известния Z-score, съгласно зависимост (13).

$$Z = 0.12X_1 + 0.14X_2 + 0.33X_3 + 0.006X_4 + 0.999X_5, \text{ където} \qquad (13)$$

X_1 е съотношението на оборотния капитал спрямо общите активи; X_2 е неразпределената печалба спрямо общите активи; X_3 е печалбата преди данъчно облагане и лихвени плащания спрямо общите активи; X_4 е пазарната стойност на СК на компанията спрямо балансовата стойност на общите задължения; X_5 е стойността на продажбите спрямо общите активи. Спрямо средните стойности на отделните групи и характеристиките на медицентровете, формирани при клъстеризацията, Z показателят на дадена

фирма може да покаже дали нейната финансова стабилност е по-скоро добра, или лоша.

При друга съвкупност стойностите на коефициентите пред всяка от променливите в (13) ще се променят, като ще се изместят и медицентровете на отделните групи. Но принципите на работа на множествения дискриминантен анализ ще останат валидни. За да сравним ефективността на модела на Алтман с други възможности, използващи самообучаващи се алгоритми, ще проверим точността на прогнозите за неплатежоспособност върху публично достъпни данни. За разлика от дискусията върху различните изследователски инструменти, изложена в Част III т. 2.4., тук проблемите с избора на показател за сравнение са по-лесно разрешими. Подобно на грешките от първи и втори род, с които се сблъскваме при проверката на хипотези, така и анализът на риска от неплатежоспособност може да завърши с два типа неточности – компания, която е стабилна, да бъде определена като заплашена от фалит и финансово нестабилна компания да бъде определена като безпроблемен случай.

3.2. Алтернативи за оценка на риска от неплатежоспособност

Като алтернатива на модела на Алтман ще приложим няколко възможности, при които има контролирано обучение:

- регресионен модел, при който независимите променливи съответстват на променливите в оригиналната формулировка на Алтман;
- дискриминантен анализ – като препратка към използваните аналитични инструменти в оригиналния модел;
- машина с поддържащи вектори (Support Vector Machine – SVM) във вариант с предварително определени теглови коефициенти – като демонстрация как може да се реши проблемът с липсата на баланс във входните данни;
- невронна мрежа, която използва обучение с обратно разпространение на грешката, но с краен резултат, който е класификация по групи, а не прогноза за стойността на дадена променлива.

Причините да се ограничи изборът на алтернативни подходи само до посочените възможности, са свързани преди всичко с това, че те са

достатъчни, за да демонстрират силните и слабите страни на самообучаващите се алгоритми. Приложението на SVM ([7], [47]) с предварително определени тегла хвърля допълнителна светлина върху начините да се определи кой тип грешки водят до по-съществени негативни последици.

3.3. Сравнение на точността на използваните алгоритми

Точността на използваните алгоритми ще бъде анализирана с помощта на публично достъпна база от данни за полски компании[10]. Тя проследява статуса на фирмите в продължение на 5 години, като за всяка са зададени ключови финансови съотношения и статус през съответния период (фалит). Разпределението на случаите е посочено в Таблица 12.

Таблица 12. Характеристики на данните за фалирали компании

Година	***Фалирали компании***	***Нормално състояние***	***Общо***
1	271	6756	7027
2	400	9773	10 173
3	459	10 008	10 503
4	515	9277	9792
5	410	5500	5910

В Таблица 13 са представени обобщените резултати от приложението на различните самообучаващи се алгоритми върху данните за първата година. Както в предходните части общата съвкупност е разделена на обучаващо (85%) и тестово множество (15%).

Преди стартиране на контролираното обучение отделните финансови показатели са предварително обработени с цел премахване на единични екстремни стойности, които се наблюдават в отделни компании. Тъй като става въпрос за малко на брой случаи (по един за показател), тези стойности са заменени със средната за крайния квартил. Подобно калибриране не променя по никакъв значим начин крайните резултати.

[10] https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data#

Таблица 13. Характеристики на сравняваните подходи

Вид	***Параметри след обучението***	***Точност (тестови данни)***
Logit	*const* X_1 X_2 X_3 X_4 X_5 -2.8895 *-0.3837* 0.08851 -2.73636 -0.04904 0.02534 (0.1166) *(0.2102)* (0.1566) 0.41101 (0.0304) (0.046) *** ***	Обща точност: 0.9605 Pr/Ref 0 1 0 96% 3.89% 1 0.1% > 0.1% AUC: 0.7097815
MDA	Процент на обяснената между групова вариация: X_1 X_2 X_3 X_4 X_5 85.94 99.89 99.99 100 100	Pr/Ref 0 1 0 96% 3.89% 1 0.1% > 0.1%
LDA	Коефициенти пред отделните променливи: X_1 X_2 X_3 X_4 X_5 -1.7219 0.02113 -0.7972 -0.03522 0.13078	Pr/Ref 0 1 0 96% 3.89% 1 0.1% > 0.1% AUC: 0.6109158
ANN	Топологията и тегловите коефициенти след обучението на мрежата са представени на Фигура 16. При проведените експерименти вариантът при който няма скрито ниво неврони, показва най-висока точност (от проверените случаи с 1 и 2 скрити нива, съдържащи между 4 и 10 неврона). Това е обяснимо, защото скритите нива биха помогнали в случаите, когато е необходимо да имаме нелинейно разделяне между отделните групи фирми.	Pr/Ref 0 1 0 95% 3.4% 1 0.86% 0.74%
SVM	SMV-Kernel: sigmoid. Брой на поддържащите вектори 438 -> (221 217)	Pr/Ref 0 1 0 94.7% 3.5% 1 1.4% 0.4%

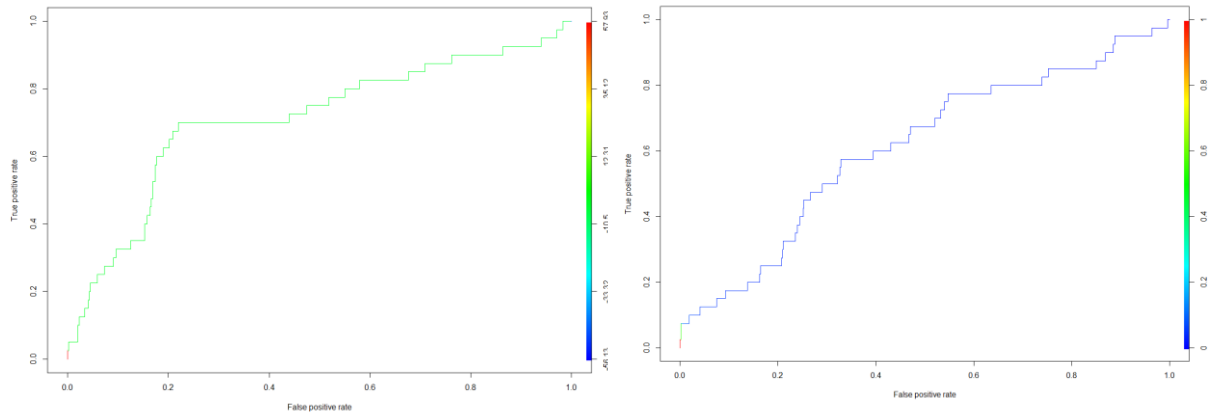
Привидно високият процент на обща точност не ни гарантира добре свършена работа, като основната причина за това е дисбалансът в отделните случаи (само 3.85% са фалиралите компании в извадката). Възможно е да се справим с този проблем, като подберем по подходящ начин значително по-малко здрави фирми, но целта на изследването ни е да проверим как се справят различните аналитични инструменти, без да се налага да се инвестират значителни усилия в подготовка на входните данни. Таблица 13 и Фигура 15 показват, че и в случая с дискриминантния анализ и при logit моделите има изразен недостатък с коректната класификация на фалиралите фирми като такива. Съществуват няколко възможности да подобрим получените резултати.

- Да включим допълнителни финансови показатели, като по този начин разширим модела на Алтман.

Този път е логичен и позволява да проверим кои от наличните финансови съотношения са определящи за прогнозирането на неблагоприятните последици от фалит. Но едновременно с това включването на нови променливи ще разруши възможността да сравняваме получените резултати с други изследвания върху валидността на модела на Алтман.

- Да разширим обема на данните, с които разполагаме и/или да ги подложим на предварителна обработка.

Това обаче невинаги е възможно и изисква допълнителни усилия, като освен аргументацията за прилагането на конкретния аналитичен инструмент става необходимо да се защити изборът на метод за предварителна обработка на данните.



Фигура 15. ROC крива на logit модела (вляво) и на LDA (вдясно)

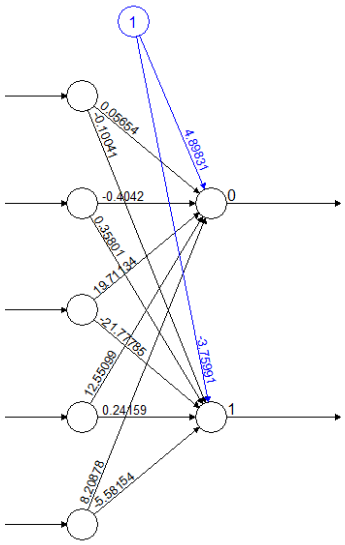
Получените резултати позволяват да сравним отделните алгоритми, като освен за приложимостта им към конкретната проблемна ситуация те свидетелстват и за наличието на следните ефекти:

- Също както и при резултатите от точка 2.4., не се наблюдават отчетливо големи разлики в точността на отделните разглеждани варианти.

Това в никакъв случай не означава, че няма значение кой аналитичен инструмент би предпочел да използва инвеститорът. Има забележим ефект по отношение на дела на групата компании, които са фалирали и са класифицирани предварително като такива. За практическата употреба на подобен тип резултати това означава, че инвеститорите могат предварително да планират и отбягват да влагат средства в такива компании или че могат да се възползват от възможността за къси продажби. Ако се

тръгне от позиция, при която не се разглежда конкретна инвестиционна стратегия, а сравнението се извършва само на базата на стойността на грешките, това означава, че е възможно отново да предпочетем по-прости за използване и обяснение аналитични инструменти.

- Невронната мрежа позволява да направим по-точна прогноза именно за специфичната група „1–1“, но далеч не може да обхване мнозинството от реалните фалити с помощта само на избраните пет финансови съотношения.



Фигура 16. Топология на използваната в анализа невронна мрежа

При анализа на различни топологии надеждни и точни резултати показа представената на Фигура 16 структура, която не съдържа скрити нива. Обяснението е свързано и с това, че в проведените тестове стремежът е да се придържаме към възможността за линейно разделяне на използваните данни, тъй като това позволява сравнимост с работата на другите прилагани алгоритми.

- Предварителната обработка на данните (мащабиране, изчистване на екстремните стойности) е изключително важна стъпка към постигането на по-добри резултати.

В направените експерименти са използвани само стъпки по изчистване на екстремните стойности и предварително мащабиране за невронната мрежа. Причините да не бъдат предварително обработени входните данни са свързани с това да се направи сравнение между отделните алгоритми при условия, които не създават изкуствени предимства за който и да е от тях.

4. Изводи

В тази част бяха представени само малка част от възможните приложения на самообучаващите се алгоритми в подкрепа на анализа на ценни книжа с фиксиран доход. Времевите ограничения пред изследването не позволяват да бъдат подробно разгледани другите потенциални проблеми, за които може да се потърсят автоматизирани и по-точни решения. Дискутираните модели позволяват да подчертаем определени характеристики на самообучаващите се алгоритми, които са важни и съдържат сериозни възможности за допълване и разширяване на традиционните средства за анализ:

- Самообучаващите се алгоритми предоставят възможност да комбинираме модели, които са задвижвани от предварително известно теоретично познание, с такива, които се основават на характеристиките на разполагаемите данни.

Съвместяването на теоретичното познание с това, което данните ни казват, е изключително полезно, защото позволява да проверим дали имаме право да използваме съответния абстрактен модел. В случаите, когато това не е така, получените стойности на показателите за грешка могат да ни насочат каква алтернатива е подходяща. Особено важно е да се подчертае, че самообучаващите се алгоритми рядко могат да ни дадат обосновката *„защо“* даден теоретичен модел не е подходящ – това все пак остава задача за този, който провежда изследването.

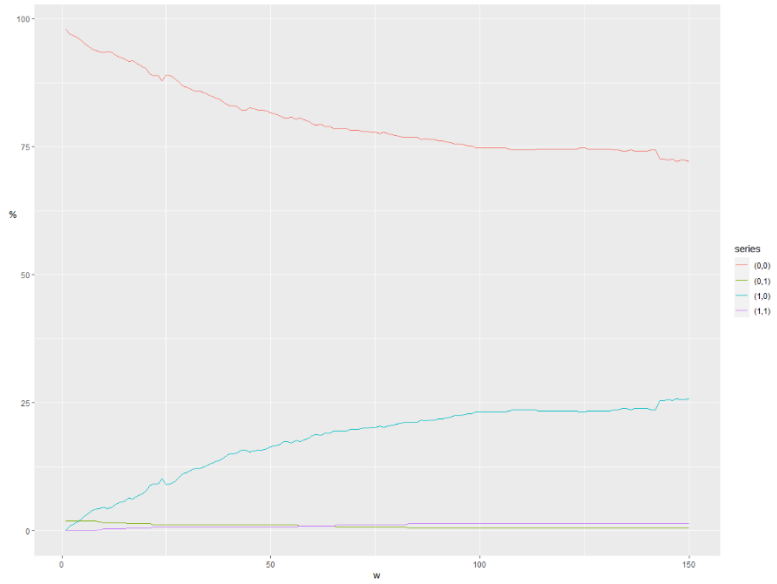
- Характеристиките на невронните мрежи ги правят ефективен инструмент при изследването на динамични редове. В това си качество те могат да допълват стандартните иконометрични модели, както и да служат като специфична проверка на получените с тях резултати.

Резултатите, свързани с приложението на MLP и ELM при анализа на индекс проследяващ доходността на корпоративни *Baa* облигации, показаха, че въпреки по-ниските стойности на грешка за първия от тези модели подобрението спрямо ARIMA не е чак толкова голямо. Този резултат подчертава наблюдението, че различните динамични редове имат и специфични особености, което изисква да се сравнят няколко подхода, за

да бъде избран този, който е най-подходящ за конкретната проблемна ситуация.

- Класификационните проблеми, свързани с представянето и финансовото здраве на различни компании, могат да бъдат изследвани с редица самообучаващи се алгоритми.

SVM алгоритмите, представени като отделна алтернатива, дават възможност за гъвкава форма на анализ, при която е възможно да се обработват данни със съществен дисбаланс между отделните категории (Фигура 17 представя тези случаи при повишаващо се тегло за по-малката съвкупност).



Фигура 17. Случаи в четирите възможни категории и различни теглови коефициенти

- Самообучаващите се алгоритми имат своите ограничения и на тях не бива да се гледа като на магическо решение, което е способно да се справи с произволен проблем.

Точността на използваните модели е близка, което означава, че ако измерителите на грешката са адекватни на инвестиционната стратегия, е разумно да се прилага най-простият и лесен за обяснение модел.

5. Изходен код и възпроизводимост на резултатите

Изходният код на R, използван за анализа на кривите на доходността и корпоративните облигации, заедно с необходимите библиотеки са представени в Таблица 14. При избора на готови компоненти в случая водещ критерий е удобството за възпроизвеждане на крайните резултати, а не ефективността на работата. Това се отнася най-вече за експериментите, използващи невронни мрежи, защото при тях ползата от специализирани библиотеки като TensorFlow[11], Theano[12] и FANN[13] е непосредствена и най-лесно забележима. Подготовката на изброените компоненти изисква допълнителни усилия, като често се достига до ситуации, при които времето за инсталиране и конфигуриране е сравнимо с това на реалното използване на продукта (особено когато целта е да се възползваме от предимствата на графичните ускорители за паралелна обработка на данните). Именно това е причината в изследването да не се включат софтуерни библиотеки, зависещи от конкретни хардуерни решения (напр. видеокарти с поддръжка за CUDA платформа).

Таблица 14. Изходен код и библиотеки използвани в Част III

Елемент	***Бележки***
Изследване на кривите на доходността	
Изходни кодове	https://github.com/drnmy/mlf (Папка III)
Източник на данни	Quandl (изисква безплатна регистрация)
Използвани библиотеки в R	
Quandl	https://www.quandl.com/tools/r
forecast	https://cran.r-project.org/web/packages/forecast/forecast.pdf
gglot2	https://ggplot2.tidyverse.org/
tseries	https://cran.r-project.org/web/packages/tseries/tseries.pdf
YieldCurve	https://cran.r-project.org/web/packages/YieldCurve/YieldCurve.pdf
DescTools	https://cran.r-project.org/web/packages/DescTools/index.html
Metrics	https://cran.r-project.org/web/packages/Metrics/Metrics.pdf

[11] https://www.tensorflow.org/
[12] http://deeplearning.net/software/theano/
[13] http://leenissen.dk/fann/wp/

xts	https://cran.r-project.org/web/packages/xts/xts.pdf
Библиотеки, използвани при анализа с невронни мрежи	
nnfor	https://cran.r-project.org/web/packages/nnfor/nnfor.pdf
nnet	https://cran.r-project.org/web/packages/nnet/nnet.pdf

В тази, както и в следващите части на изследването простите експерименти са реализирани с R. За Python са запазени примерите, които позволяват лесно мащабиране най-вече от гледна точка на увеличаване на обема на входните данни или които отнемат твърде много време за изпълнение. Причината за това изкуствено наложено разделение е, че възпроизвеждането на получените резултати е малко по-лесно в R, тъй като почти не изисква предварителна подготовка и инсталирането на допълнителен софтуер.

IV. Неконтролирано обучение при анализа на финансови инструменти

Неконтролираното обучение напомня много силно на начина, по който хората се учат да изпълняват определени задачи. Нека за момент си представим, че трябва да подредим склад с автомобилни части, но нямаме никаква представа за устройството и начина на работа на автомобилите. Как може да се разреши проблемът, пред който сме изправени, ако нямаме нито време, нито шанс да започнем да четем подходящите книги? Ако сме рационални, една възможна последователност от стъпки е:

- да подредим кутиите по форма и тегло на отделни групи;
- да ги отворим и да подредим съдържанието им по материал на изработка – пластмаса, метал и т.н.;
- да сравним отворените детайли и да ги подредим на групи по тяхната собствена форма и визуална прилика.

Този процес може да продължи доста дълго време – докато всеки един детайл не попадне в подходящата група (което обаче би означавало и да има такива с по един – единствен елемент) или докато не преценим, че наличните до момента съвкупности са достатъчно хомогенни. Който и от двата критерия да изберем обаче, ще бъде налице една обща характеристика на целия процес – че подредбата е направена без външна помощ и без да се знаят предварително характерните признаци на отделните авточасти.

Същата логика и принципи на работа са заложени, макар и с различна реализация, при алгоритмите, използващи неконтролирано обучение. Таблица 15 представя основните разлики между самообучаващите се алгоритми от двете категории. В контекста на анализа на финансови инструменти тези особености разкриват една много важна възможност – да се изследват проблеми, при които е ясно, че има различни категории и групи, но не знаем предварително какви са техните характеристики. Страничен „ефект“ на това приложение е, че могат да се анализират и случаи, при които липсват достатъчно исторически данни.

Таблица 15. Основни разлики между контролирано и неконтролирано обучение

Контролирано обучение	***Неконтролирано обучение***
Алгоритмите се обучават „кое какво е“ въз основа на голям брой предварително известни и ясно определени случаи.	Структурата и шаблоните, които се формират, са получени изцяло въз основа на наличните данни.
Отделните категории са известни отпреди.	Отделните категории не са известни.
Входните данни са придружени с очакван/желан резултат.	Входните данни не са придружени с очакван/желан резултат.
В ролята на „учител“ влизат предварително определените входни данни.	Алгоритъмът се обучава самостоятелно, като използва разликите в признаците на отделните наблюдения от входните данни.

И все пак какви биха могли да бъдат причините да прибегнем до неконтролирано обучение, особено като отчитаме изобилието от финансови и икономически данни? Основните двигатели на широкото приложение на този тип самообучаващи се алгоритми са:

- Цената, свързана с получаването на достатъчни по обем извадки, позволяващи контролирано обучение.

В този случай под общия термин „цена“ следва да се разбират не само паричните средства, но и времето, което трябва да се инвестира в правилното оразмеряване и оценка на отделните елементи от обучаващото множество. Като отчитаме динамиката на финансовите процеси и необходимостта информацията да се допълва и осъвременява (съответно алгоритмите да се преобучават), общите разходи не са малки.

- Конкуренцията и търсенето на нови възможности за анализ са свързани с използването на нови източници на данни.

Високите потенциални печалби, които един добре работещ алгоритъм може да донесе, когато се прилага на финансовите пазари, привличат много изследователи, инвестиционни фондове и посредници. В резултат на нарастващата конкуренция аналитичните инструменти стават все по-разнообразни и отдавна надхвърлят представите ни, свързани с анализ на

динамични редове. Използването на данни от социални медии като Twitter и Facebook ([48]) предполага: 1) работа с неструктурирана информация и 2) бърза реакция. И двете изисквания не допускат възможността за евтино допълване на получената информация с контролни стойности.

- Необходимостта да работим с промени с голяма размерност на възможните резултати;

Този проблем, известен с популярното си название „проклятието на размерността“ ([49]), е сериозно препятствие, когато се сблъскаме с широк набор от възможни резултати. Нека да си представим, че имаме N наблюдения, като всяко едно от тях може да заеме K различни стойности. Тогава общо ще се сблъскаме с пространство на възможните състояния на нашата система с размерност N.K, като това произведение може да има наистина големи стойности за множество реални задачи. Ако имаме предварително известни правила, по които се движи и развива изследваната система, ние много лесно можем да ограничим възможностите, с които да работим. В случаите, когато те липсват или не е сигурно, че действат, е възможно да прибегнем до самообучаващите се алгоритми, за да получим събирателен образ на изследваните данни с по-ниска размерност.

- Нуждата да получим безпристрастно мнение и да избегнем изместванията, които са следствие от индивидуалните предпочитания или уникалния опит на анализаторите.

Макар и да не може да бъде точно оценена, тази възможност е много полезна, особено когато разполагаме с разнопосочни и/или субективни мнения на отделни изследователи. Тогава самообучаващят се алгоритъм може да играе ролята на безпристрастна проверка (без да се забравя възможността той също да е неточен) и допълнително „мнение“.

1. Финансова стабилност и стрес тестове

Финансовата стабилност на компаниите е от изключителна важност за тяхната нормална работа и възможността им да генерират положителни нетни парични потоци и финансови резултати. Развитието на паричните отношения и капиталовите пазари освен това води до по-висока обвързаност на отделните фирми. Тази зависимост може да има положителен аспект, при

който специализацията и допълващите се продукти подпомагат растежа на компаниите и тяхното оцеляване. Но тя може да има и отрицателни последици в случаите, когато проблемите в една фирма се разпростират върху нейните партньори, клиенти и доставчици.

Когато предоставяните услуги са финансови и компанията е публична и достатъчно голяма, трудностите, с които се сблъсква тя, може да се превърнат в системен проблем. Ролята на регулаторните органи е да не допуснат това да се случи, въпреки че, както историята ни показва, те невинаги успяват да се справят с предизвикателствата на разнообразните финансови продукти и отношения. Самообучаващите се алгоритми могат да бъдат много полезни при решаването на класификационни проблеми в този контекст, като допълват традиционните финансови модели и очертават по-ясно границите между отделните категории и групи участници на пазара. Една от възможностите, демонстриращи ползите от неконтролираното обучение, е свързана с провеждането на стрес тестове за оценка на банковата стабилност.

Въпреки че самите стрес тестове не са свързани с конкретен финансов инструмент, резултатите от тяхното провеждане са изключително важни за инвеститорите, включително и за събирателния образ на участник на финансовите пазари, описан на стр. 29. Причините за това са, че от една страна, банките, са свързани с общата стабилност на цялата финансова система, а от друга, че част от тях са активни участници на пазарите. Различните форми и подходи за провеждане на стрес тестове са подробно анализирани в [50], [51] и [52]. За целите на това изследване ще бъдат използвани само данните от проведеното през 2018 г. изследване от EBA (European Banking Authority) на 48 банкови институции в 15 страни от ЕС и ЕИЗ[14].

Причината да се ограничи изследването до данните отпреди две години, е, че планираните за 2020 г. стрес тестове бяха отложени за идващата 2021 г., като основната причина за това е неовладяната към момента пандемия от COVID-19[15].

[14] https://eba.europa.eu/risk-analysis-and-data/eu-wide-stress-testing/2018
[15] https://eba.europa.eu/eba-launches-2020-eu-wide-stress-test-exercise

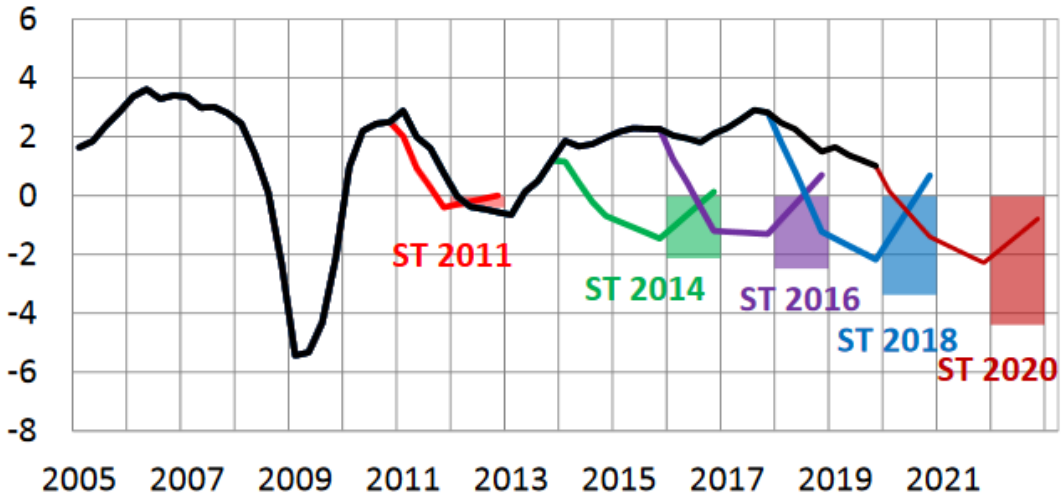
2. Методология на стрес тестовете

Целта на това изследване не е да прави критичен анализ на методологията на стрес тестовете. Дискутираният в [53] подход съдържа редица привлекателни черти, свързани с отчитането на нелинейните ефекти от излагането на риск. Въпреки това ще се придържаме към основните положения, използвани от ЕВА, тъй като целта е да се покаже по какъв начин самообучаващите се алгоритми могат да подобрят вече използваните инструменти, а не да се предложи алтернатива.

Провеждането на стрес тестовете е свързано с анализ на финансовите показатели на анализираните кредитни институции при два основни сценария за развитие на икономиката през следващите три години (за оценката през 2018 г. за начало се приема краят на 2017 г.). Първата – „базова“ възможност разглежда най-вероятното развитие на икономиката въз основа на наличната към момента информация. Обикновено при този възможен сценарий не се предполага възникването на специфични негативни последици. Както показва и самото наименование, основната цел на тази възможност е да служи като отправна точка, за да се сравнят ефектите върху стабилността на кредитните институции с алтернативните варианти. В стрес теста от 2018 г. параметрите на този сценарий са разработени от Европейската централна банка (ЕЦБ).

Втората – „вредна“ – възможност разглежда негативен вариант на бъдещо развитие, като целта е да се поставят в хипотетична „стресова“ ситуация анализираните банки и да се оценят ефектите върху тяхната ликвидност и стабилност. При разглеждането на този сценарий е особено важно до каква степен ще приемем, че е реалистично икономическите параметри да се влошат. В стрес теста от 2018 г. параметрите на този сценарий са разработени от Европейския съвет за системен риск (European Systemic Risk Board – ESBR). На Фигура 18 и Фигура 19 са представени основните показатели, заложени в негативния сценарий, както и тяхната промяна спрямо реалното развитие на икономиката през последното десетилетие. Очаквано се вижда, че през текущата година са заложени по-екстремни варианти предвид негативните последици от COVID-19 пандемията и несигурността, свързана с времето, необходимо за намирането на дълготрайно решение. Като отчитаме състоянието на икономиките към средата на 2020 г., както и развитието на заболеваемостта, може да се

очаква, че ревизираните стойности през следващата година няма да бъдат по-оптимистични.



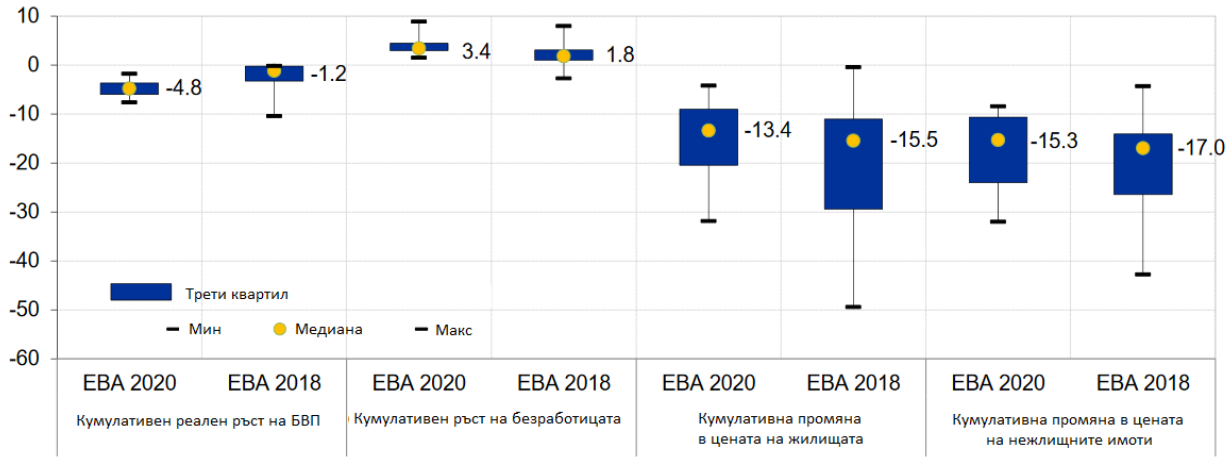
Фигура 18. Прогнозираните спадове в БВП (в проценти) за вредните сценарии при стрес тестовете в периода 2011–2020 г. Източник: ЕВА[16]

Икономическите параметри в рамките на негативния сценарий са от изключително значение за коректното провеждане на стрес тестовете. Но дори и в ситуация, при която е възможно да се направи перфектна прогноза за бъдещето, това няма да е достатъчно, за да се гарантират точни резултати от теста. Причината е свързана с начина, по който отделните изследвани банки са зависими от промени в макроикономическите параметри ([54], [55], [56], [57]).

Важна особеност на стрес тестовете на ЕВА е, че техният изход не е дихотомен – издържал/не издържал. Вместо това основната им цел е да предоставят информация за финансовата стабилност и до каква степен може да приемем, че дадена банка е устойчива. Причини резултатите да се структурират по този начин са:

- неблагоприятният сценарий е преди всичко само предположение и като такова не можем да твърдим дали ще се реализира, или ще бъдем свидетели на още по-сериозни от предвидените спадове в БВП и цените на активите, с които са обезпечени кредитите;

[16] https://eba.europa.eu/sites/default/documents/files/document_library//2020%20EU-wide%20stress%20test%20-%20Macroeconomic%20scenario.pdf



Фигура 19. Оценка на ключови икономически показатели при сценариите за стрес тестове през 2018 и 2020 г. Източници: ЕВА, ЕСВ[17]

- разглежданите два варианта на развитие на икономическата среда не изчерпват всички възможности, което оставя много неизвестни;

- финансовите показатели на проверяваните финансови институции се базират на техните отчети, което означава, че в най-добрия случай се анализират текущите стойности, които обаче се съотнасят към тригодишен бъдещ период.

Когато посочените особености се съпоставят с предимствата на неконтролираното обучение, може да се оформят следните основни области, в които да се търси рационализиране на анализа и подобряване на точността на резултатите:

- Групиране на участващите кредитни институции по начин, който да разкрие в най-голяма степен приликите и различията между тях.

Не съществува гаранция, че финансовите съотношения и показателите за кредитиране са достатъчни, за да обхванат всички особености на проучваните банки. В контекста на макропруденциалните тестове индивидуалните разлики не са в центъра на нашето внимание и съответно добавянето на нови показатели не носи толкова ползи. При оценката на устойчивостта на отделните институции обаче тази допълнителна

[17] https://eba.europa.eu/sites/default/documents/files/document_library//2020%20EU-wide%20stress%20test%20-%20Macroeconomic%20scenario.pdf

информация може да помогне да се очертаят по-добре границите между отделните участници, както и да се оцени по-добре рискът.

- Комбиниране на показатели с различен тип (числени, ординални), без предварително да се налагат ограничения за общата форма на зависимост (линейна или нелинейна).

Използването на показатели от различен тип не е привилегия на самообучаващите се алгоритми. Последните обаче позволяват да подходим към изследването на една нова съвкупност от наблюдения с по-малък брой предварителни допускания.

- Намаляване на размерността на използваните данни.

Тази възможност не противоречи на предходните, тъй като става дума за намаляване на размерността на съвкупности, при които наблюденията имат множество характеристики със сходен характер. Например при провеждането на стрес тестовете един от потенциалните въпроси, на които може да търсим отговор, е – изложени ли са банките на съществени рискове от промяна във формата на кривата на доходността. Има няколко възможности да получим задоволителен отговор, като една от тях, свързана с модела на Нелсъл–Сиегел–Свенсон, вече беше разгледана в предходната част на изследването. Прецизната оценка на ефектите от промени в лихвените проценти изисква да бъдат прогнозирани промените в доходността на инструменти с различен матуритет. Голяма част от разглежданите динамични редове са силно корелирани помежду си, което означава, че има възможност да комбинираме техните ефекти и по този начин да намалим размерността на входните данни.

2.1. Алгоритми за класификация с неконтролирано обучение

Процесът на класификация е свързан с причисляването на изследваните обекти, според техните признаци към определена група. Съществуват различни начини, по които може да се определи към кое (под)множество принадлежи едно наблюдение, когато се налага да използваме неконтролирано обучение. Най-често използваните от тях са представени в Таблица 16, като извън обхвата на настоящото изследване остават редица възможности, описани в [58], [59], [60]. За всички изследвани финансови

показатели неявно е прието допускането за ексклузивност – т.е. едно наблюдение може да принадлежи най-много на един клъстер. Съществено предимство е възможността да работим с числени показатели, което елиминира нуждата да разглеждаме възможности с препокриване или размиване на отделните групи.

Таблица 16. Алгоритми за класификация с неконтролирано обучение

Алгоритъм	***Особености***
Алгоритми за клъстеризация	
K-Means	Един от най-често използваните алгоритми. Приложим и в случаите на голямо множество от данни, като тогава се прибягва до употребата на евристични алгоритми за ускоряване на клъстеризацията.
Йерархична клъстеризация	Приложим с различни метрики за „дистанция“ между отделните наблюдения, като алгоритъмът може да работи дори и без да разполага със стойностите за отделните характеристики – стига да са ясни разликите между наблюденията (т.е. отделните „дистанции“).
DBSCAN	Важно предимство на този алгоритъм е, че не изисква предварително дефиниране на броя на клъстерите, като едновременно с това може да формира групи с различна форма. Методът обаче е силно чувствителен към правилния избор на функция, с която да се определи дистанцията между отделните елементи.
Смесени модели (Mixed models)	Приложимостта на този подход се определя до голяма степен от това дали е възможно предварително да обосновем какво разпределение следват елементите от даден клъстер и колко са отделните групи. Тогава може да направим опит да оценим параметрите на разпределението за всяка отделна група.
Алгоритми с невронни мрежи	
Генеративни състезателни мрежи (GAN)	Въпреки че приложенията, с които GAN придобиват масова популярност, са свързани с генерирането на изображения и медицинската диагностика, дискриминантната част на GAN модела може да се използва с цел класификация ([61]).

Самоорганизиращи се мрежи (SOM)	Самоорганизиращите се мрежи предоставят гъвкава възможност за групиране на наблюденията по различен брой показатели, включително и в областта на финансите и анализа на банкови институции ([17]).

За изследване на възможностите, които неконтролираното обучение предоставя, от всички изброени алгоритми ще използваме само K-Means клъстеризацията[18] и самоорганизиращите се мрежи. Сравнението в точността на изброените възможности заслужава отделен анализ, но за момента е по-важно как групирането може да подпомогне традиционния анализ, а не толкова коя от новите възможности дава по-добри резултати.

Преди да коментираме работата на избраните възможности, да определим какви са параметрите на решаваната задача. Входните данни ще обозначим като съвкупността от *n* наблюдения, всяко едно от които съдържа *p* на брой характеристики:

$$X = \{x_1, .., x_n\},$$
$$x_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right). \quad (14)$$

Крайната цел е тези наблюдения да бъдат асоциирани с k на брой клъстери, като според особеностите на използвания алгоритъм стойността на k може да не е предварително известна. Резултатът от работата на клъстеризацията е съвкупност от характеристики на получените групи:

$$C = \{c_1, .., c_k\},$$
$$c_i = (c_{i1}, c_{i2}, \dots, c_{im}), m \leq p. \quad (15)$$

2.1.1. K-Means клъстеризация

K-Means клъстеризацията разчита на представянето на всяка една група с помощта на една точка – нейния център. Това означава, че всяко едно от

[18] Строго погледнато наименованието K-Means е описание на проблема, а не е име на алгоритъм. Стъпките, следвани в процеса на клъстеризация, са известни като алгоритъм на Лойд (Lloyd [95]).

наблюденията, класифицирани в съответната група, ще се асоциира и представлява от центъра на тази група, въпреки че между него и въпросния център ще има разлика. Ако тази разлика се представи като функция на разстоянието – $d(x, c)$, тогава всеки един от елементите x_i ще принадлежи на този клъстер, за който стойността на $d(x, c)$ е най-малка. Нека $f_C(x) = \underset{c \in C}{arg\,\min} d(x, c)$ определя съответствието на отделните наблюдения към дадена група. Процесът по клъстеризация може да се сведе до намирането на $C = \{c_1, .., c_k\}$ такива, че да се минимизира сумата от всички разстояния. Таблица 17 показва различни начини, по които може да се представи тази сума и които водят до различни решения на дадения проблем.

Таблица 17. Критерии за минимизиране на разстоянието и варианти на K-Means

Критерий за минимизиране	***Бележки***
$\sum_{x \in X} d(f_C(x), x)^2$	При този широко използван вариант на групиране се минимизира вариацията вътре в рамките на отделния клъстер.
$\sum_{x \in X} d(f_C(x), x)$	Тази форма на групиране позволява да се намерят централните точки на клъстерите, така че те да бъдат компактни.
$\underset{x \in X}{max}\, d(x, c)$	Оптимизацията по този критерий позволява фокусът да бъде изместен към най-лошите (отдалечени) случаи. Графично разликата е представена на Фигура 20.

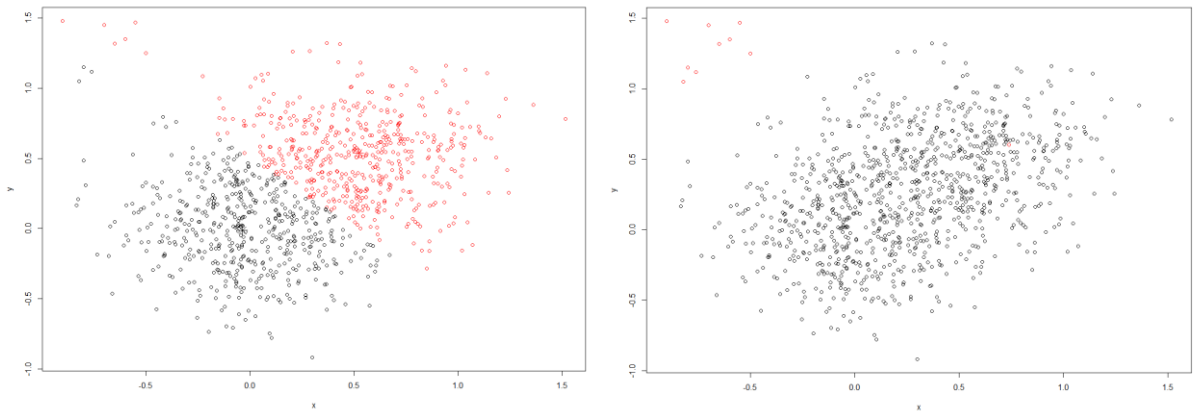
При изследване на стабилността на банковите институции е относително лесно да се установи групата на тези участници в анализа, които имат силно отклоняващи се показатели (напр. изключително висок риск или големи експозиции към определен тип активи). По-трудно е да се разграничат отделните сегменти сред останалите участници. Поради тази причина при проверката с реални данни е предпочетен стандартният вариант с минимизиране на дистанцията, повдигната на квадрат.

Алгоритъмът на Лойд, с чиято помощ се осъществява групирането, може да се представи като последователност от четири стъпки:

1) Избираме по псевдослучаен начин k стойности, представящи отделните клъстери, така че $C \subset X$.
2) За всички $x_i \in X$ определяме $f_C(x_i)$, а следователно и към кой клъстер принадлежи x_i.
3) За всеки един от получените клъстери определяме представящата го точка, като осредняваме дистанцията на елементите в него.
   Т.е. $c_i = avg\{x_j \in X \mid f_C(x_j) = c_i\}$.
4) Ако отделните центроиди не се променят, тогава алгоритъмът приключва работа. В противен случай продължава отново с точка 2).

Броят на повторенията, необходими за достигане на крайната точка в описаната последователност, може да бъде намален (съответно бързодействието на алгоритъма да бъде подобрено), ако първоначалните стойности се избират по предварително обмислена стратегия. Обикновено се прилагат три подхода за това:

- генерирани случайни стойности за отделните c_i, в който случай не можем да разчитаме на съществени подобрения в бързодействието;
- разделяме по случаен начин обучаващото множество X на k части, като съответният медицентър c_i се реализира с осреднената стойност на характеристиките от съответната част.

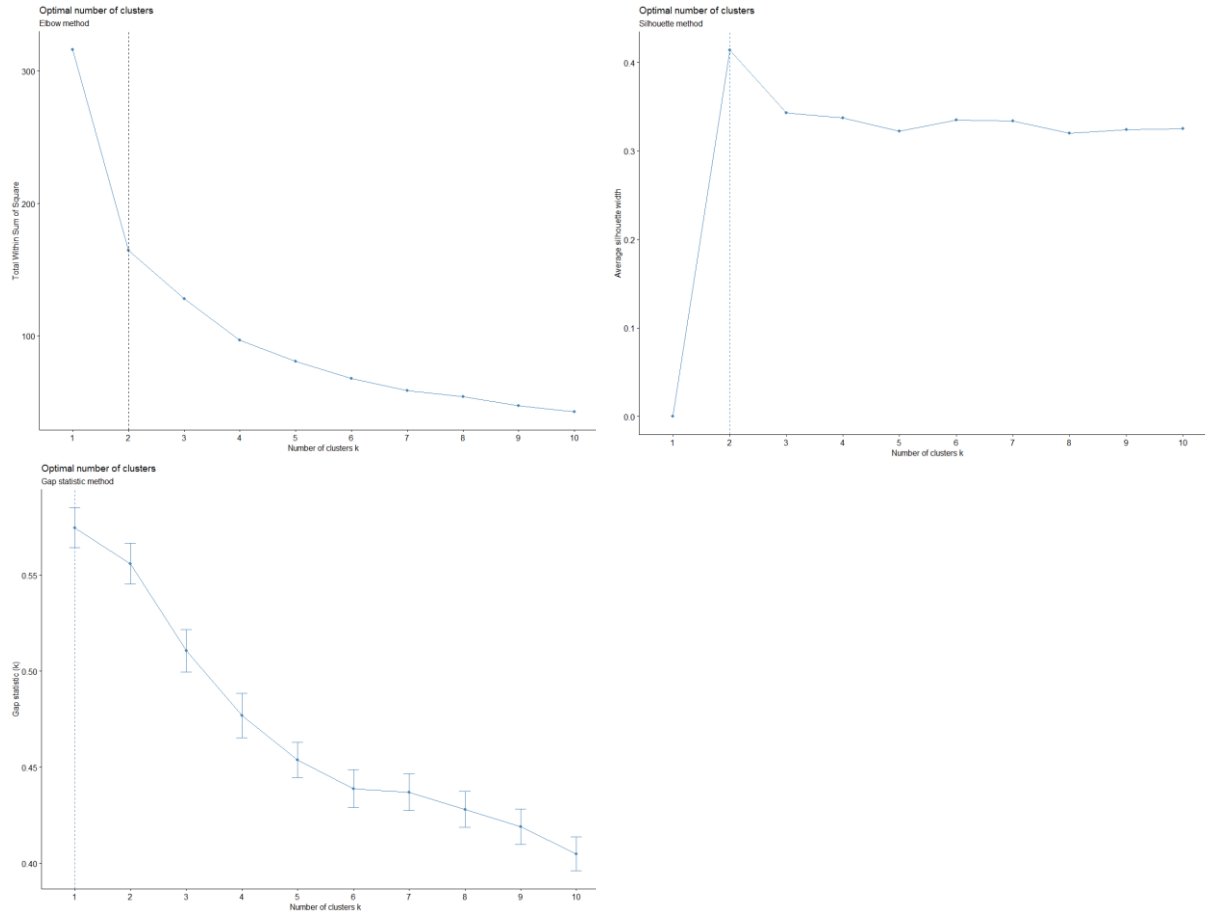


Фигура 20. Разлики в клъстеризацията между K-Means (вляво) и K-Center (вдясно)

Специфично ограничение при K-Means е, че броят на отделните клъстери трябва да е предварително известен. Тъй като няма достатъчно основание да фиксираме стойност предварително, в рамките на изследването ще бъдат използвани три метода за намиране на оптималния брой групи при ограничение, че техният брой е в интервала от 2 до 10:

- Elbow метод ([62]);
- Gap statistic метод ([63]);
- Silhouhette метод ([64]).

Фигура 21 представя резултатите от работата на трите метода върху данните, използвани за демонстрация на разликите между K-Center и K-Means. Като отчитаме, че при изкуствено създадените данни имаме две големи групи плюс добавена малка съвкупност от екстремни стойности, може да добием представа за това как се търси оптималният брой клъстери.



Фигура 21. Проверка за оптимален брой клъстери на симулираните данни от Фигура 20

Приложението на посочените три метода е една от възможностите да се определи k. Освен оптимизационните процедури друг важен аргумент за броя клъстери са финансовите и икономическите аргументи. По този начин крайното решение за разделяне на анализираните институции може да отразява не само фактическите разлики между наблюденията, но и желанията на тези, които провеждат изследването и в крайна сметка ще превъплътят неговите резултати в конкретни решения и политики.

Третият важен параметър, свързан с приложението на K-Means, освен решението колко да бъдат клъстерите и каква оптимизация ще се приложи, е за вида на функцията на разстоянието – $d(x, c)$. Най-често използваната зависимост е свързана с Евклидовото разстояние (метрика) - (16).

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2} \qquad (16)$$

При наличието на разнородни по тип данни е възможно да бъдат използвани и други функционални зависимости, като например разстояние на Gower:

$$d(x_i, x_j) = \frac{\sum_p \theta_{ijk} d_{ijp}}{\sum_p \theta_{ijp}}, \qquad (17)$$

където разликата между отделните характеристики d_{ijp} зависи от техния вид и θ_{ijp} са тегловите коефициенти на значението на всяка характеристика. Конкретните форми на метриките d_{ijp} са посочени в (18).

$d_{ijp} = \frac{|x_{ip} - x_{jp}|}{R_p}$, където R_p е размахът на p-та характеристика от числен тип;

$d_{ijp} = \frac{|t_{ip} - t_{jp}|}{R_p}$, където $t_{ijp} = \frac{(r_{ip} - 1)}{\max(r_{ip}) - 1}$ е трансформацията на ранга r_{ip} ординалната характеристика x_{ip}; (18)

$d_{ijp} = \begin{cases} 0, & \text{ако } x_{ip} = x_{jp} \\ 1, & \text{ако } x_{ip} \neq x_{jp} \end{cases}$ за логически или факторни характеристики.

Използването на алтернативни измерители за разстоянието е особено полезно, когато се налага да се анализират социални фактори и променливи с категориен характер ([65]). Информацията от стрес тестовете на ЕВА се състои от утвърдени финансови показатели и съответно е във вид на числени променливи. Това дава основание да се използва Евклидовото разстояние в процеса на клъстеризация.

2.1.2. SOM клъстеризация

Принципите на работа на самоорганизиращите се мрежи (SOM) са подробно описани в [66], но за целите на това изследване особен интерес представлява способността им да групират наличните наблюдения, без да са налични контролни/желани стойности. Неконтролираното обучение в SOM е възможно благодарение на това, че тегловите коефициенти на връзките между отделните неврони в мрежата изразяват значимостта и силата на тяхната свързаност.

На всяка стъпка от етапа на обучение се изчислява Евклидовото разстояние между характеристиките на съответното наблюдение и тегловите коефициенти на отделните неврони:

$$I(i) = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}, \qquad (19)$$

$$w_q(i) = \{w_{q1}, w_{q2}, \ldots, w_{qk}\}, \qquad (20)$$

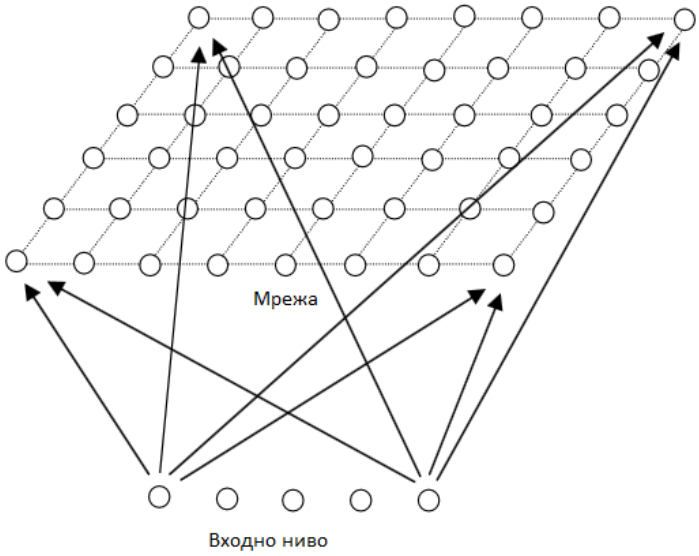
$$\delta(I(i), w_q(i)) = \sqrt{\sum_{m=1}^{p} (d_m - w_{im})^2}, \qquad (21)$$

$$w_q(i+1) = w_q(i) + f(B, q, i)\, L(i)\, [I(i) - w_q(i)], \qquad (22)$$

където:

- $w_q(i) = \{w_{q1}, w_{q2}, \ldots, w_{qk}\}$ представя тегловите коефициенти на неврона q на стъпка i; $I(i) = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ са стойностите на отделните признаци, които системата използва в обучението;
- $\delta(I(i), w_q(i))$ е Евклидовото разстояние между тегловите коефициенти на q-тия неврон и входните параметри;
- $f(B, q, i)$ е т. нар. „функция на съседство“, която определя дистанцията между неврона q и текущия неврон с минимално Евклидово разстояние.

Последният елемент в (22) е $L(i)$, който определя скоростта, с която системата се обучава.



Фигура 22. Примерна самоорганизираща се мрежа

Алгоритъмът, с чиято помощ се осъществява неконтролираното обучение на самоорганизиращите се мрежи, може да бъде представен като последователност от пет стъпки:

1) Инициализация на тегловите коефициенти на отделните неврони с псевдослучайни стойности.
2) Избор на случаен елемент от обучаващото множество[19].
3) Определяне на неврона, който има най-малка стойност на Евклидовото разстояние (или съответната метрика) спрямо елемента от стъпка 2).
4) Обновяваме теглата на съответния неврон съобразно със зависимост *(22)*.
5) Ако установим, че промените в тегловите коефициенти са много малки, или иначе казано – топологията на мрежата вече не се променя, тогава може да прекратим работата на алгоритъма. В противен случай се повтаря отново стъпка 2).

Специфично приложение SOM може да намери и при решаване на проблеми с размерността на входните данни. Особено полезни самоорганизиращите се мрежи могат да бъдат в случаите, когато се налага нелинейно разделяне на входните данни – ситуация, с която обичайните инструменти като PCA не могат да се справят.

[19] Съществува и модифициран вариант на обучаващия алгоритъм, при който последователно се избират един по един всички елементи от обучаващото множество.

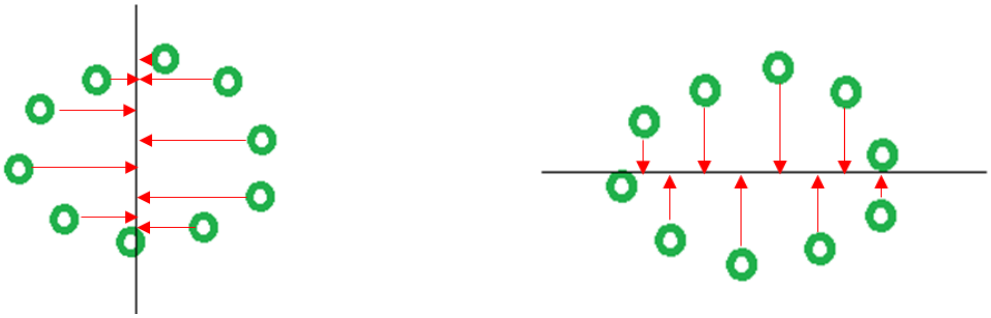
2.2. Алгоритми за намаляване на размерността на входните данни

Намаляването на размерността на входните данни е от съществено значение, когато работим с реални проблеми и в конкуренция с други изследователи или автоматични системи за анализ. Първият аргумент в подкрепа на това твърдение – че по-бързият и точен алгоритъм винаги има предимство, трудно може да бъде оспорен. Но освен конкурентни предимства, свързани с времето за пресмятане, намаляването на размерността може да позволи прилагането на алгоритми, които иначе изискват прекалено много ресурси.

За нуждите на настоящото изследване ще бъдат използвани само два от възможните подходи за намаляване на размерността – анализът на главните компоненти (PCA) и самоорганизиращите се мрежи. Причината да ограничим представянето само до тези два алгоритъма, е свързана с това, че те са изключително популярни и достатъчно лесни за използване.

2.2.1. Анализ на главните компоненти (PCA)

Основната идея на PCA анализа е показана в изключително опростения пример на Фигура 23. Да предположим, че трябва да изберем една от двете възможности, за да опишем разположението на точките, така че да получат координати само по вертикалната или само по хоризонталната ос. Разбира се, най-добре би било, ако опишем точките с координати по две перпендикулярни оси, но нашата специално зададена цел е да намалим размерността и да използваме само една от тях.



Фигура 23. Проста демонстрация на PCA анализа

Ако подходим рационално, би трябвало да изберем тази възможност, която показва най-добре разликата между отделните точки (в този случай това е дясната страна на графиката). Следователно би трябвало да предпочетем този вариант, който максимализира вариацията (отклоненията/разликите) между отделните наблюдения. Когато размерността на данните е по-голяма, графичното представяне не е сполучливо, за да преценим как би могло да се намали тя. Вместо това PCA анализът използва ковариацията и коефициентите на корелация между отделните стойности, за да определи кои от тях са силно свързани и съответно промените в тях могат да бъдат съвместно представени.

Работата на алгоритъма може да бъде описана в следните три стъпки:

1) Стандартизиране на отделните променливи, за да сме сигурни, че всяка една от тях ще може да „допринесе“ по аналогичен начин при последващия анализ.

Причината да се налага стандартизиране, е, че PCA е много чувствителен към вариацията на входните данни. Следователно, ако не извършим предварителна обработка и има две променливи, които се променят съответно в интервала от 1 до 1000 и от 0 до 1, влиянието на първата от тях ще бъде непропорционално голямо. След преобразуването на входните стойности с (23) сравнимият мащаб ще елиминира този негативен ефект:

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}. \qquad (23)$$

В изследването предпочетената форма за стандартизиране на входните данни е (23), а не минимум-максимум нормализация (24), защото преди всичко целта ни е да уловим тези елементи, които имат най-голяма вариация:

$$x' = \frac{x_i - min(x)}{max(x) - min(x)} \qquad (24)$$

2) Изчисляване на вариационно-ковариационната матрица като израз на потенциалните връзки между отделните променливи.

Ако имаме p отделни променливи, то тази матрица ще има размерност pxp, като елементите по главния диагонал ще са вариацията на всяка променлива:

$$CV = \begin{bmatrix} cov(x_1, x_1) & \cdots & cov(x_1, x_p) \\ \vdots & \ddots & \vdots \\ cov(x_p, x_1) & \cdots & cov(x_p, x_p) \end{bmatrix}. \quad (25)$$

3) Изчисляване на собствените вектори и собствените стойности на получената на стъпка 2) матрица.

За собствените вектори и собствените стойности ще припомним, че ако имаме една матрица A с размерност pxp, вектор стълб v с размерност 1xp, тогава от произведението:

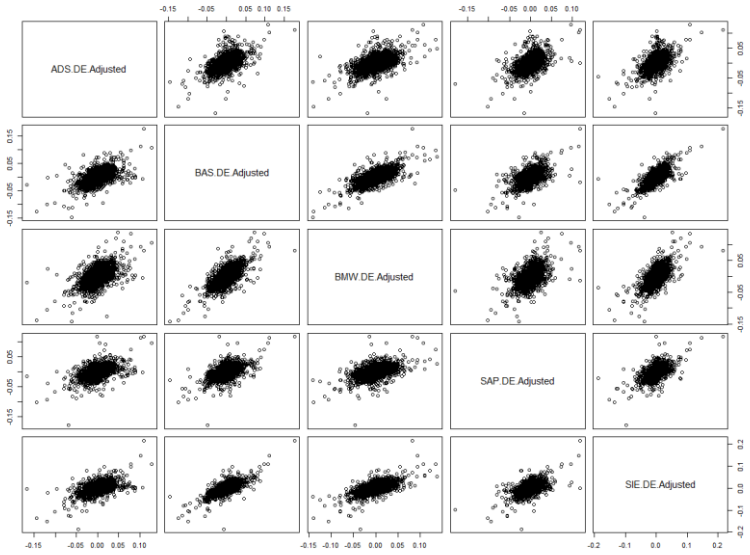
$$Av = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}, \quad (26)$$

$$\theta_i = \sum_{j=1}^{p} a_{ij} v_i. \quad (27)$$

Казваме, че ако $Av = \theta = \gamma v$, то v е собствен вектор[20] на линейната трансформация A, а числото γ е собствената ѝ стойност. Какво е общото между собствените вектори и стойности, и главните компоненти? По определение главните компоненти са линейна комбинация от оригиналните променливи, така че получените резултати не са корелирани.

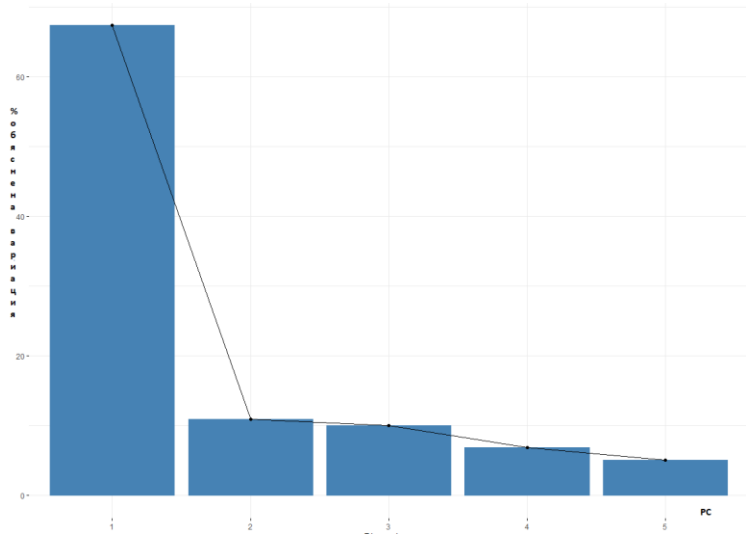
Освен да посочи основните компоненти, PCA анализът се стреми да ги дефинира по такъв начин, че максимално голяма част от промените на оригиналните стойности да са свързани с първия основен компонент. Съответно всеки следващ носи по-малко информация относно поведението на оригиналните стойности. Целта на цялата последователност от стъпки е да изберем определен брой от получените главни компоненти и като използваме тях вместо първоначалните входни данни, да ограничим размерността на решавания проблем. Като пример нека да разгледаме съвкупност от дневните възвръщаемости на пет борсово търгувани компании в периода 01.02.2007 г. – 27.07.2020 г., показани на Фигура 24.

[20] Тъй като имаме произведението Av, по-точно би било да се говори за десен собствен вектор.



Фигура 24. Разпределение на дневните възвръщаемости от примерната съвкупност

В зависимост от желаната степен на точност може да изберем първите *n* собствени вектора. Колкото повече смаляваме размерността на проблема, толкова по-голяма става загубата на информация, изразена като дял от вариацията в първоначалната съвкупност, която не можем да обясним с новите данни.



Фигура 25. PCA анализ върху възвръщаемостта на портфейл от пет акции

Така например, ако намалим размерността от пет (колкото са динамичните редове за възвръщаемостта на всяка акция) на три, това ще

означава използването на нови данни, които обясняват 88.2% от вариацията в първоначалната съвкупност – както е показано в Таблица 18.

Таблица 18. Характеристика на основните вектори от примера с петте акции

Параметър	PC1	PC2	PC3	PC4	PC5
Стандартно отклонение	1.836	0.735	0.70645	0.58214	0.50193
Дял от вариацията в началната съвкупност	0.674	0.108	0.09982	0.06778	0.05039
Обща вариация в началната съвкупност	0.674	0.782	0.882	0.9496	1

Както е при всеки един от използваните до момента аналитични инструменти, така и анализът на главните компоненти има редица предимства и недостатъци. Те трябва да бъдат отчитани, преди да се вземе решение за прилагането му, както и на етапа на тълкуване на получените резултати. Основните слабости, които заслужават специално внимание са:

- Загуба на информация.

Всяко намаляване на броя на собствените вектори, които използваме, е съпроводено със загуба на информация, защото, както показва и примерът в Таблица 18, дори и последните елементи носят определена информация за промените на стойностите в първоначалната съвкупност. Като положителна страна на PCA може да се спомене, че все пак изследователят има възможност да избере колко да бъде тази загуба.

- Нуждата от стандартизация на променливите, преди да бъдат намерени собствените вектори и стойности.

Стандартизацията означава, че категорийните променливи трябва да бъдат преобразувани в числени такива. Освен допълнителното изчислително време, което е необходимо, този факт трябва да бъде отчитан допълнително.

- Трудности при интерпретирането на получените нови променливи.

Въпреки че главните компоненти са линейна комбинация от първоначалните променливи, тяхното тълкуване е по-трудно, отколкото това на индивидуалните характеристики. Което означава, че след прилагане

на PCA анализ трябва да се обърне специално внимание на коректното тълкуване на получените резултати спрямо първоначалните променливи.

- Проблеми при невъзможност за линейно разделяне и нелинейни промени в зависимостите.

Тези недостатъци могат да бъдат преодолени с използването на kernel PCA (kPCA) с подходящата функционална зависимост. Представеният до момента анализ на главните компоненти може да се разглежда и като частен случай на kPCA, при условие че функционалната зависимост на ядрото (kernel) е линейна. Допълнителна информация за kPCA може да се намери в [67].

2.3. Класификационни алгоритми и стрес тестове на EBA

Преди клъстеризацията данните на EBA ще бъдат подложени на предварителна обработка с цел да се гарантира консистентност и сравнимост на получените резултати. Конкретно, преди да се изпълни алгоритъмът за клъстеризация, входните параметри са:

1) Филтрирани, така че от всички 809136 елемента, предоставени в базата данни на EBA, да останат само тези, които представят общото състояние на отделната банка.

По този начин се лишаваме от възможността да се направи сравнение на управлението и финансовото състояние на отделните кредитни институции по държави, но този проблем е извън обхвата на изследването. При необходимост би могло към филтрите с приложения код да се добави и проверка за определена страна, което не променя следващите стъпки от изследването.

2) Добавени са филтри към входните елементи, така че да се използват общите експозиции на кредитната институция, а не тези, предоставени във връзка с отделни регламенти.
3) Получената съвкупност е разделена спрямо периодите, за които се отнасят отделните наблюдения.

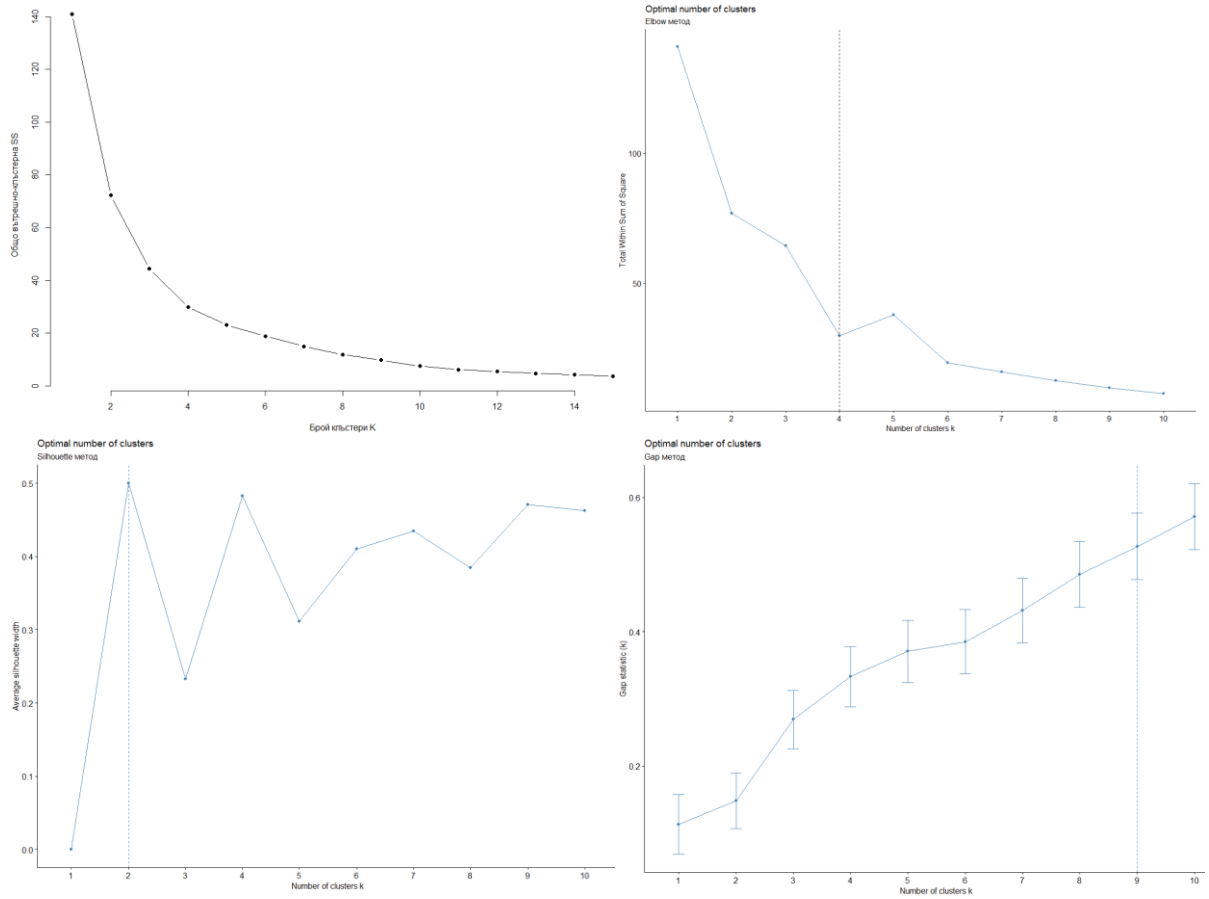
Тази стъпка ни дава възможност да сравним получените клъстери и тяхната промяна през отделните периоди, включени в стрес теста. Поради

относително краткия времеви хоризонт са анализирани данните само към началния и крайния период.

4) Получената съвкупност е разделена спрямо сценариите, за които се отнасят отделните наблюдения (реални и симулирани).

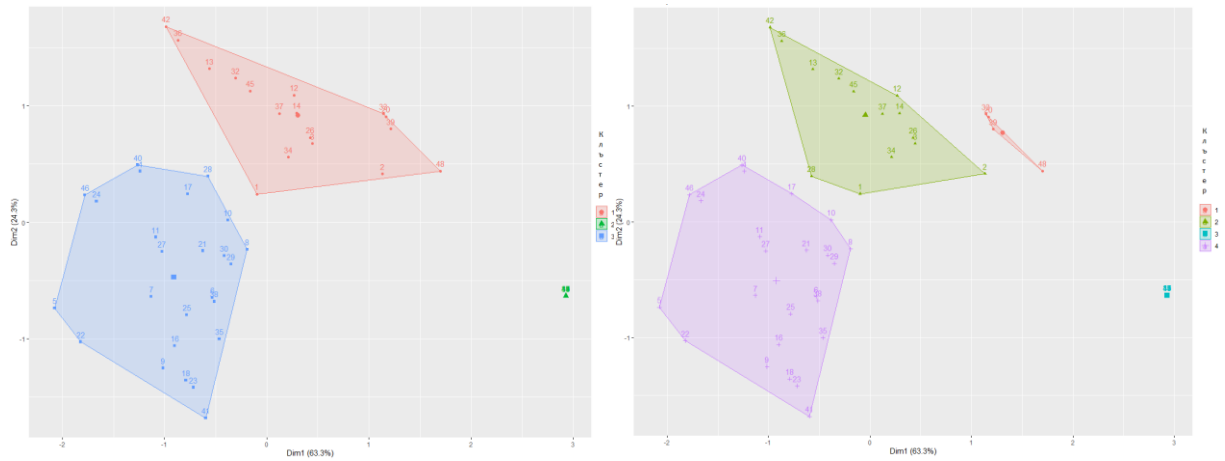
Допълнителното сравнение по тези показатели показва как биха се променили позициите на кредитните институции както вътре в съответния клъстер, така и при евентуално преминаване от една група към друга. За целите на изследването са разгледани и трите възможности – текущо състояние, базов и негативен сценарий.

Фигура 26 представя промяната на квадратите на вътрешните отстояния в клъстерите при увеличаване на техния брой. Въз основа на трите приложени метода – elbow, silhouette и gap statistics, можем да заключим, че оптималният брой клъстери е между два и пет. В последващото изложение ще използваме вариант с три и четири групи.



Фигура 26. Оптимален брой клъстери спрямо края на 2017 г. на базата на текущото състояние

Резултатите съответно с три и четири клъстера са показани на Фигура 27, като са използвани три показателя: *SEC – Exposure values (actual/restated figures), SEC – Risk Exposures amounts (actual/restated and projected figures), SEC – Impairments (actual/restated and projected figures)*. Информацията е анонимизирана, като отделните институции са представени само с пореден номер.



Фигура 27. Клъстеризация по данни към края на 2017 г. при три и четири области

Резултатите, представени на фигурата, могат да подпомогнат традиционните стрес тестове в няколко различни направления:

- Определяне на съвкупността от показатели, които най-добре позволяват да се установят разликите между участващите в стрес теста банки.

Основната цел на стрес тестовете е да направи оценка на финансовата стабилност на кредитните институции и да предотврати проблеми, които е възможно да прераснат в сериозни кризи. Освен като инструмент за подпомагане на контрола обаче те могат да се използват и като източник на информация за задълбочен анализ на процесите, протичащи във финансовата система и даващи отражение върху цялостното поведение на изследваните компании. Сравнението между съвкупността от показатели, които са заложени в методиката на провежданите тестове, и тези, които отразяват най-добре разликите между отделните участници, е добра възможност за проверка на адекватността на заложените в проучването стъпки.

- Разграничаване на отделните групи от кредитни институции, което може да е основа за диференциран подход не само при провежданите

тестове, но и след това – при необходимост от предписания за по-добро управление на риска.

Групирането на участниците в стрес тестовете носи много ползи, но още по-важно е как то се променя при условията на различните сценарии. Това позволява да се направи прогноза за начина, по който различни промени в икономическата обстановка ще се отразят на целия пазар, а не само на отделни банки.

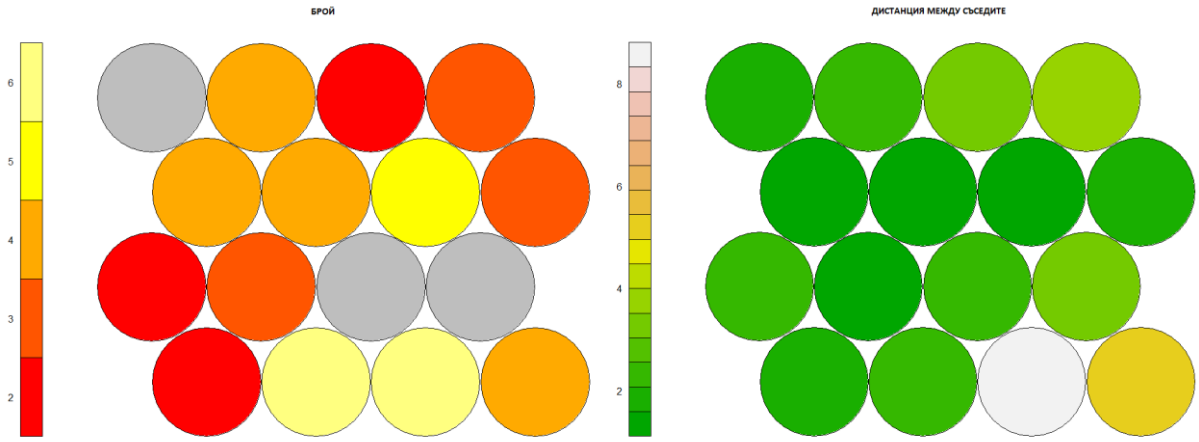
- Определяне на отклоненията вътре в отделните групи, което може да е основание за подробен бъдещ анализ на кредитните институции, участващи в рамките на отделен клъстер.

Компактното разположение вътре в отделните клъстери е не просто желан ефект от използването на правилен брой групи. То съдържа индиректни сигнали по какъв начин целенасочените регулативни промени и политики ще се отразят на отделната съвкупност, както и дали е обосновано да се приеме, че ефектът ще бъде идентичен.

За да оценим ефективността на K-Means, ще сравним с резултатите, получени с помощта на самоорганизиращи се мрежи на Кохонен. Използвани са същите входни стойности, което позволява да се направи директна съпоставка между отделните групи.

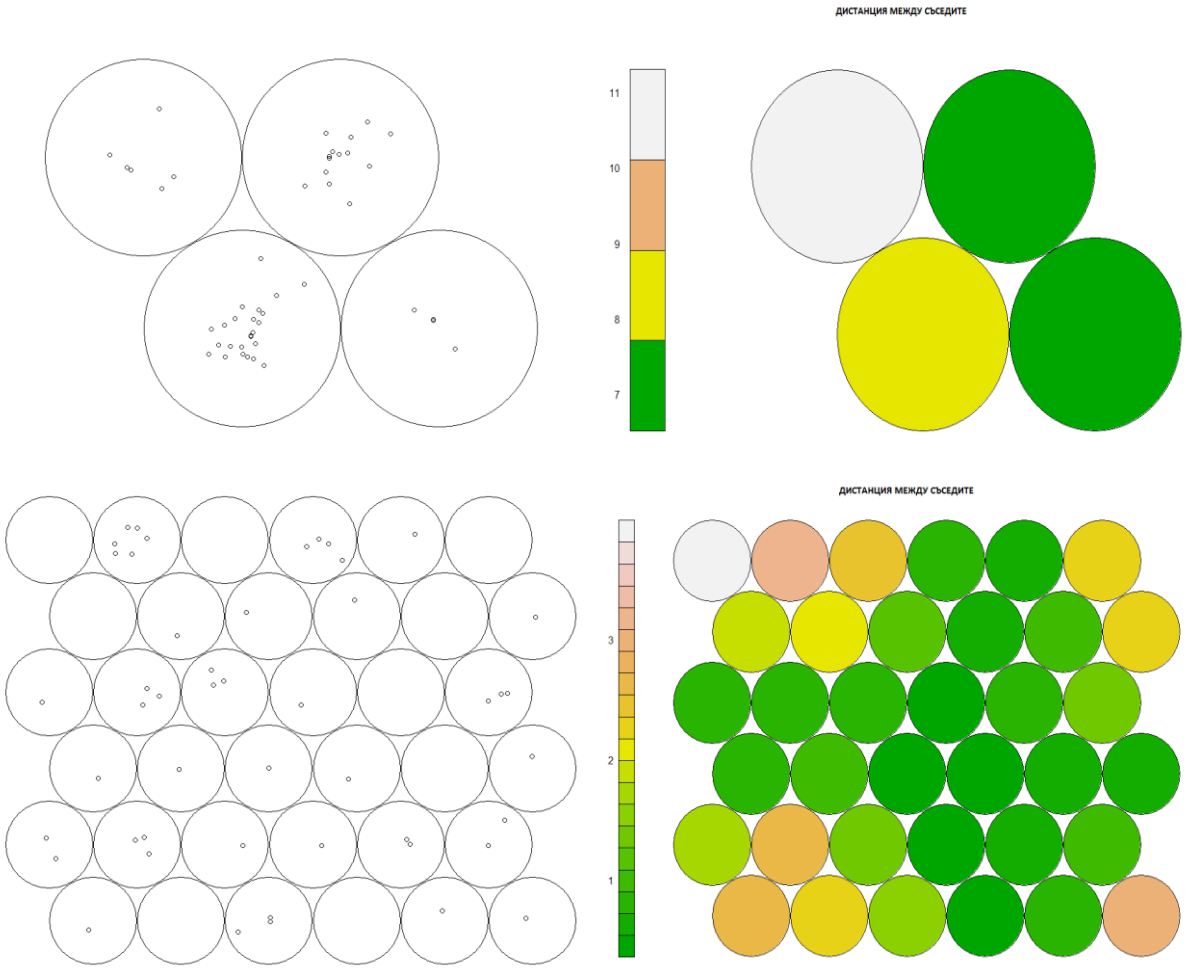
На Фигура 28 са представени отстоянията между отделните елементи при мрежа с топология 4x4 (в дясната част на графиката), както и броят на отделните институции, класифицирани по всяка от шестнадесетте възможни категории (в лявата част на графиката). Кое обаче налага да използваме точно невронна мрежа с топология 4x4 (съответно 16 неврона в цялата мрежа) и защо да не приложим модел с друга размерност?

За да определим оптималната топология на самоорганизиращата се мрежа, може да подходим по различни начини, като в рамките на това изследване са използвани два от тях. Първият е да се търси изравняване на броя на желаните категории, което, ако сравняваме с K-Means с четири клъстера, означава да се избере топология със същия брой неврони. При квадратна мрежа това води до структуриране на SOM с размерност 2x2, както е показано в горния ред на Фигура 29.



Фигура 28. Резултати от клъстеризацията с помощта на самоорганизиращи се мрежи

Вторият вариант е да се използва предложената в [68] зависимост, според която броят на невроните би трябвало да бъде $D = 5\sqrt{N}$, където N е броят на наблюденията, използвани в анализа.



Фигура 29. Резултати от самоорганизиращата се мрежа при размерност 2x2 и 6x6

При квадратна самоорганизираща се мрежа това би определило, че размерността ще бъде $\sqrt{D}$. В рамките на изследваните данни от стрес теста през 2018 г. правилото подсказва, че броят на невроните следва да е 35, което след закръгляне води до квадратна мрежа 6x6 неврона.

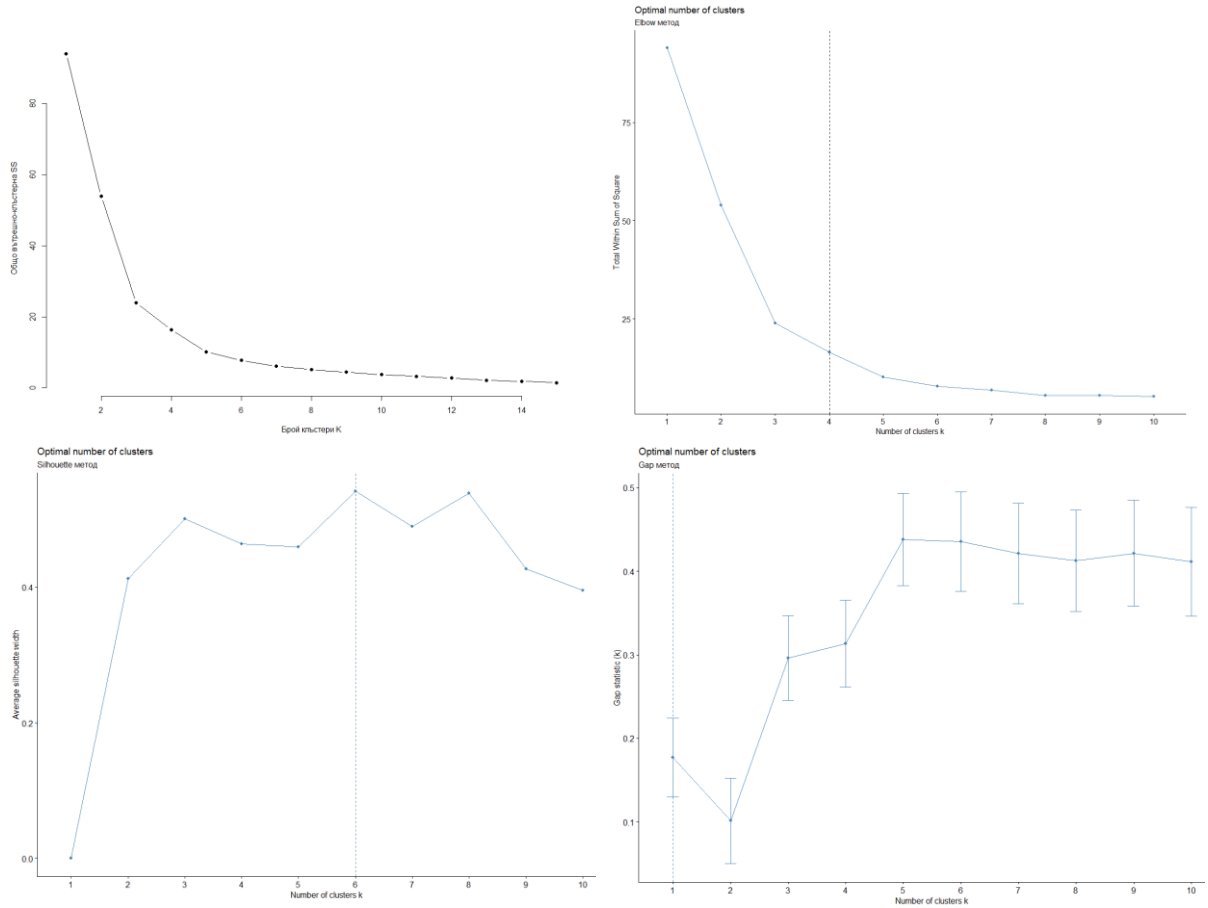
Резултатите от работата на самоорганизиращите се мрежи позволяват да направим следните заключения:

- Представянето на получените групи в случая със самоорганизиращата се мрежа дава алтернативен поглед върху различията в изследваните кредитни институции.

При ниска размерност на мрежата може да очакваме (както потвърждават и резултатите), че получените групи ще бъдат близки до тези, които се оформят от K-Means. Ако обаче се увеличи значително броят на невроните, то тогава резултатите от SOM ще съдържат допълнителна информация за преходите, които се осъществяват от една група към друга (тъй като по дефиниция на работа, когато едно наблюдение води до класифицирането му в определена група, това се отразява и на близките до съответния неврон възли).

- При малката размерност 2x2 не се наблюдават съществени разлики в резултатите на K-Means алгоритъма и SOM.

Близките резултати предоставят възможност да използваме по-добре познатия и по-лесен за тълкуване метод – K-Means. Това не означава, че самоорганизиращите се мрежи нямат място в други подобни изследвания, още повече като се отчита фактът, че размерността на данните, с които работим, е много ниска. В последващите експерименти, свързани с приложението на самообучаващи се алгоритми за подпомагане на стрес тестовете, основно внимание ще бъде обърнато на резултатите на K-Means клъстеризацията.



Фигура 30. Оптимален брой клъстери при прогноза за края на 2020 г., базов сценарий

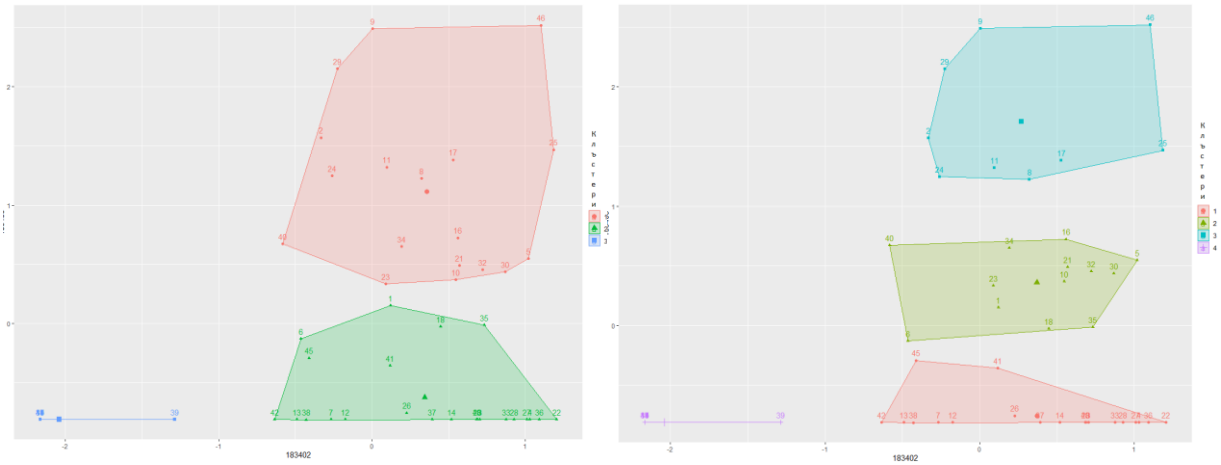
Резултатите от изследването за оптимален брой клъстери при базисния сценарий, показани на Фигура 31, позволяват да направим следните изводи:

- Оптималният брой групи е по-висок спрямо първоначалните данни.

Основните причини, които водят до по-висок брой групи, са свързани с различния начин, по който прогнозираният сценарий се отразява на финансовите показатели на изследваните компании. Въпреки различния брой групи обаче няма значителни промени в стойностите на отделните медицентрове.

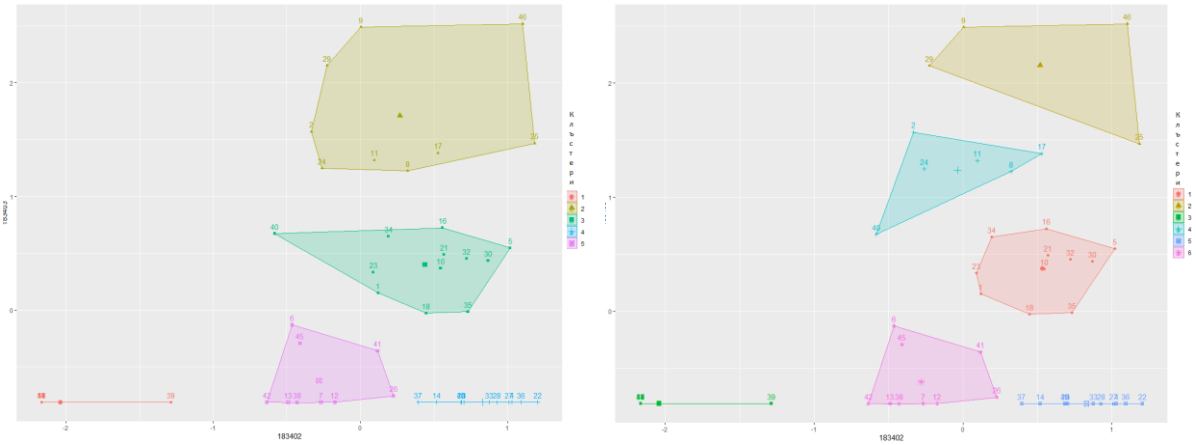
- От използваните в анализа променливи само последните две имат значим принос при оформянето на групите.

Този резултат показва, че размерността на данните може да бъде намалена, като променливите без значим ефект върху позиционирането на кредитните институции в клъстерите се изключат от анализа. Трябва да се има предвид, че този тип оптимизация на входните данни не е същият, както при PCA, въпреки че и двата водят до по-прост модел.



Фигура 31. Клъстеризация по данни към края на 2020 г., базов сценарий при три и четири области

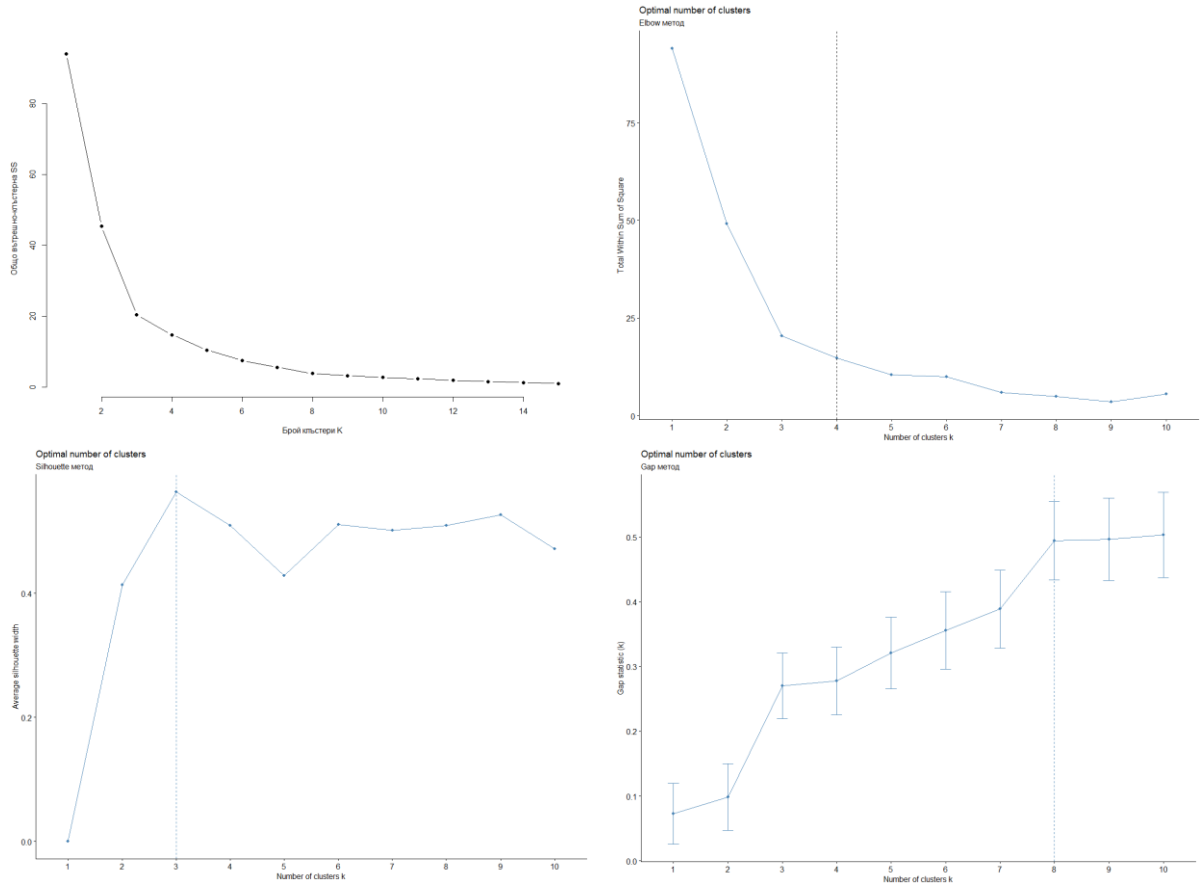
Представените на Фигура 32 резултати показват и визуално, че разделянето на изследваните институциите на по-голям брой групи позволява да се постигне по-висока хомогенност вътре в самите клъстери, както и да се опише по-добре ефектът от разглеждания сценарий върху банките. Разбира се, при достатъчно висок брой подмножества винаги може да се постигне по-компактно представяне на отделните части. Ценният ефект от приложението на алгоритъма е, че ни посочва, че увеличението не е необходимо да бъде голямо.



Фигура 32. Клъстеризация по данни към края на 2020 г., базов сценарий при пет и шест области

При негативния сценарий съвсем очаквано оптималният брой на клъстерите е по-малък. Една от основните причини е, че когато е налице сериозен спад в икономиката, отрицателните ефекти се отразяват на всички участници във финансовата система, което води и до сближаване на техните показатели. Трябва да се отбележи, че оптималният брой на клъстерите

остава по-висок от две, което означава, че необходимостта от прецизна и консистентна система от икономически и регулаторни мерки се запазва.



Фигура 33. Оптимален брой клъстери при прогноза за края на 2020 г., негативен сценарий

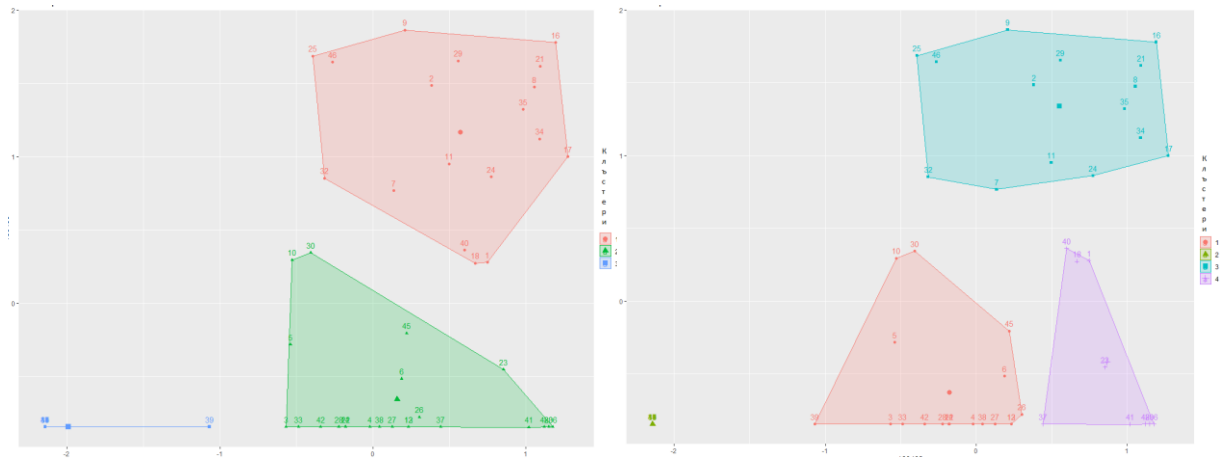
Важно уточнение към представените резултати и очаквания за края на 2020 г. е, че те не отчитат (тъй като публикуването на данните и прогнозите е в края на 2018 г.) влиянието на кризата с COVID-19. Проведените в края на месец юни 2020 г. симулации и тестове на Европейската централна банка (ЕЦБ – [69]) предвиждат корекция на първоначалните възможности, като вече предвидените спадове в БВП са съобразени с последиците от пандемията.

Таблица 19. Промени в БВП, отчитащи пандемията от COVID-19 ([69])

Сценарий	2020	2021	2022
Базисен	–8.7%	+5.3%	+3.3%
Негативен	–12.6%	+3.3%	+3.8%

Промените в начина на работа на всички икономически агенти, предизвикани от пандемията, посочват два недостатъка на механизмите, с

които се осъществяват стрес тестовете – фокусиране върху малък брой добре разработени сценарии, както и разглеждането на състоянието на банковите институции отделно. В този случай, както се посочва в [53], настъпването на малко вероятни, но много значими събития може да доведе до компрометиране на получените резултати.



Фигура 34. Клъстеризация по данни към края на 2020 г., негативен сценарий при три и четири области

Самообучаващите се алгоритми могат да помогнат за преодоляване на посочените недостатъци по следните начини:

- Като предоставят възможност да се изследват от различна гледна точка банките, които имат близки параметри и едновременно с това са свързани.

По този начин клъстеризацията може да подпомогне анализа на взаимната зависимост на банките. Липсата на оценка за тези зависимости би довела до увеличаване във вариацията/грешката на получените резултати и в крайна сметка до неправилни решения.

- Като разглеждат движението на отделните участници от една група в друга или вътре в отделния клъстер.

Тази гледна точка ни предоставя възможност да сравняваме позицията на отделната институция и промените в нея спрямо сходните участници на пазара. Освен че избягваме от необходимостта да имаме „твърда“ абсолютна база за сравнение (която може да не е адекватна за всички разглеждани случаи), по този начин се поставя акцент върху относителното представяне на банките.

- Представянето на отделните групи с характеристиките на техните медицентрове е по-добра възможност, отколкото простото осредняване на индивидуалните показатели, когато се налага да представим промените в група банки.

2.4. Неконтролирано обучение и анализ на риска

Приложението на самообучаващите се алгоритми за подобряване на точността на стрес тестовете е само една от възможностите, които неконтролираното обучение предлага. Акцентирането върху основните предимства на този тип обучение позволява да се решават проблеми, свързани с анализа на риска като цяло.

Макар и непълна, класификацията от Фигура 35 на различните предизвикателства, с които трябва да се справи една кредитна институция, дава представа за широкия спектър от приложения на неконтролираното обучение. За пример могат да се посочат разнообразните системи за оценка на риска при отпускане на заеми, анализ на трансакциите с кредитни карти, както и за моделиране на потребителското поведение и сегментиране на пазара.

Подробният анализ на всяко едно от приложенията ще доведе до неоправдано увеличаване на обема на изследването. Поради тази причина, вместо да се разглеждат предимствата за индивидуалните възможности, ще акцентираме върху предизвикателствата, свързани със самообучаващите се алгоритми:

- Представяне и възможност за обяснение на получените резултати.

Липсата на логично и обосновано обяснение защо изградените модели дават един или друг резултат, е сериозна пречка при тяхната практическа употреба. Независимо дали става дума за правно основание (например да може да се обоснове защо се отказва издаването на кредитна карта или отпускането на кредит при оспорване на решението), или за бизнес логика, третирането на самообучаващите се алгоритми като „черни кутии“ прави много трудно тяхното широко приложение. За да бъдат избегнати описаните проблеми, може да се използват следните принципи:

- При наличие на повече от една възможност предпочитанията да се насочат към по-простите модели с детерминистично поведение и лесна за обяснение вътрешна структура.



Фигура 35. Частичен списък на рисковете, стоящи пред една кредитна институция

- Използване на тези характеристики на входните данни, които имат установен глобален принос към крайните резултати (напр. водят до консистентно увеличаване или намаляване в стойностите на зависимите променливи). Този принцип е особено важен при нелинейните и моделите с градиентно спускане.

- При сложните модели (например невронни мрежи с много възли и сложна топология) може да се използват съпоставки с аналитични инструменти, като например стойности на Ауман – Шапли ([70], [71]). По този начин е възможно да се потвърди, че резултатите имат „смисъл“ и от гледна точка на кооперативните игри и рационалните интереси на всеки един от участниците.

- Хиперпараметри и интегриране на нови елементи в анализа.

Като „хиперпараметри“ означаваме всички елемент от аналитичния модел, които не са свързани с разполагаеми данни, а произтичат от избрания самообучаващ се алгоритъм. Такива например са броят на невроните и нивата в невронните мрежи, броят на клъстерите в K-Means. Тъй като трябва да бъдат дефинирани предварително, това означава, че освен да обясним зависимостите, залегнали в основата на изграждания модел, трябва да се обосноват всички тези допълнителни характеристики. Съществуват три принципни подхода, с които може да се осигури достатъчно добра аргументация на избраните стойности за отделните хиперпараметри:

- експертни оценки, които обаче трябва да бъдат проверени допълнително за обективност;
- проверка с помощта на търсене за най-добрата стойност на съответния параметър в предварително дефинирана област – неутрален от гледна точка на възможен субективизъм подход, но чувствителен към начина на търсене;
- въз основа на аналитични модели и съществуващи изследвания, които са насочени към същите или близки проблеми.

Възможно е и комбинирането на посочените подходи, така че да се осигури по-стабилна обосновка на избраните параметри.

- Изместване (bias) на получените резултати.

Причините за изместване на получените резултати могат да бъдат разнообразни, но в контекста на неконтролираното самообучение особено значение имат две от тях – свързани с избора на конкретен алгоритъм и свързани със специалните характеристики на входните параметри (например анализ на групи от хора, при които има специални условия, свързани с отпускането на заеми). Можем да приемем, че вторият източник

на изместване е свързан с предварително определени правни и регулаторни механизми, върху които нямаме влияние. За да се елиминират проблеми, свързани с избор на алгоритъм, могат да се приложат следните принципи:

- предварително проучване и елиминиране на тези аналитични инструменти, които не отговарят на особеностите на проблема;
- сравнение и състезателен принцип при подбора на самообучаващи се алгоритми, като при идентични резултати се отдава предпочитание на по-простите възможности (както е и в настоящото изследване).

- Избор на входни параметри и фалшиви зависимости.

Както резултатите трябва да имат логично обяснение, така и входните данни следва да бъдат логични и свързани с решаваната задача. Ако това правило се нарушава, рискуваме да попаднем в ситуация, при която има прекомерно обучение и добри резултати с тренировъчното множество, но ниска ефективност на изградения модел извън него. В допълнение, ако използваме нелогични (или чисто технически параметри – например съотношения без икономически и финансов смисъл) характеристики за вход на изграждания модел, рискуваме да бъдем подведени да търсим зависимости там, където ги няма. Опасността да бъдем подведени от фалшиви връзки, нараства заедно с увеличаване на размерността на модела, което налага използването на допълнителни проверки, за да се убедим в коректното тълкуване на резултатите ([72]).

- Приложимост в реални условия.

Приложимостта в реални условия изисква да се отговори на редица въпроси, които често надхвърлят чисто научната дискусия относно качествата на изградените модели. Допълнителните изисквания, наложени от практиката, като например време за реакция, необходимост от поддръжка и обновяване на софтуера, стабилност на работа и автоматична употреба (без необходимост от висококвалифицирани кадри, които да наблюдават и подпомагат самообучаващата се система), могат сериозно да ограничат приложението на неконтролираното обучение. Решенията на проблемите, свързани с реалната употреба на изградените системи, могат да бъдат както технологични, така и принципни – например изграждането на по-прости и

стабилни модели дори и с цената на малко по-висока (но все пак приемлива) неточност.

- Възможности за (ре)калибриране на моделите.

Възможностите за автоматично повторно калибриране на параметрите на моделите са често желан елемент от процес на самообучение. От гледна точка на регулациите този факт може да се превърне в кошмар, който да предизвика цялостен отказ от подобни системи. Причината е, че ако е възможно самообучаващият се алгоритъм да се променя под въздействие на нопопостъпващата информация, това поражда поне два сериозни проблема:

  - Какво да се направи с периодичното одобряване, сертифициране и диагностициране на съответната система.

Непрекъснато адаптиращ се алгоритъм би означавал и постоянна необходимост от преглед и сертифициране на цялата система. Това не е възможно практически, нито е финансово изгодно. Поради тази причина може да се фиксират определени периоди, след които да се извършва преглед и обучение отново, но по този начин се губи ефектът от динамичното пренастройване на самообучаващия се метод. Втората възможност е да се определят прагови стойности, като промени в структурата и параметрите на използвания алгоритъм настъпват само след преминаването им. Цената на това решение отново е свързана с ограничаване на адаптивността на изградената система.

  - Как да се гарантира, че в процеса на първоначално калибриране не са използвани данни, свързани с временни (локални) промени в наблюдаваните финансови параметри.

Първоначалното калибриране на модела е изключително важна стъпка, като проверката за качеството на използваната информация има индиректен, но важен принос за ползата от крайните резултати.

3. Лихвени индекси и PCA анализ

Приложението на метода на главните компоненти ще бъде демонстрирано върху данни за доходността на американските държавни ценни книжа с различен матуритет и върху исторически стойности за

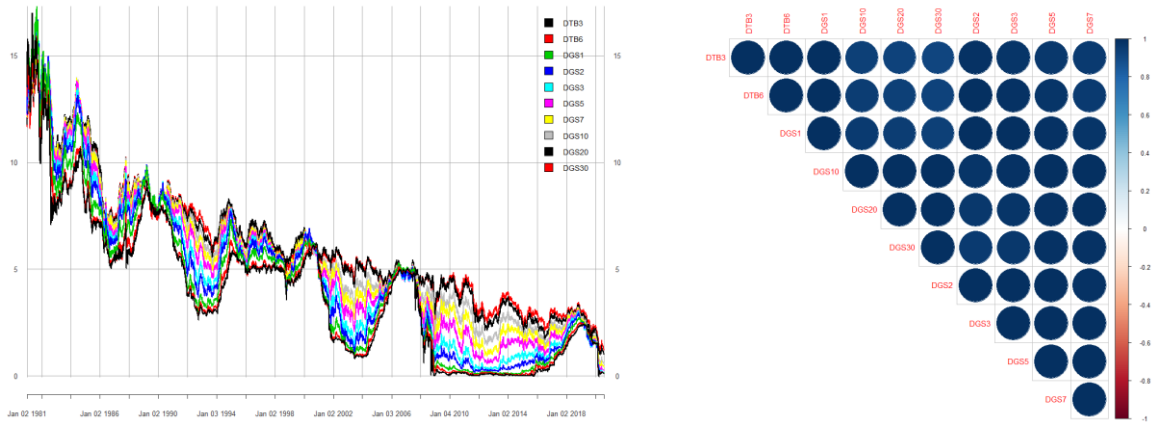
LIBOR. Причината да се използват тези икономически показатели, е да се провери ефектът от PCA в различни ситуации и с различна продължителност на динамичните редове. Допълнителен аргумент за този избор е, че ДЦК предоставят възможност да се анализират матуритети с продължителност до 30 години, докато LIBOR е ориентиран към по-кратки периоди (до една година).

Таблица 20. Характеристики на входните данни, използвани за PCA анализа

Показател	ДЦК	LIBOR, базиран на USD
Начален период	01.01.1981 г.	02.01.2001 г.
Краен период	30.07.2020 г.	30.07.2020 г.
Брой стойности	9686	4816
Матуритети включени в изследването	3M, 6M, 1Y, 2Y, 3Y, 5Y, 7Y, 10Y, 20Y, 30Y	OTN, 1WK, 1M, 2M, 3M, 6M, 12M

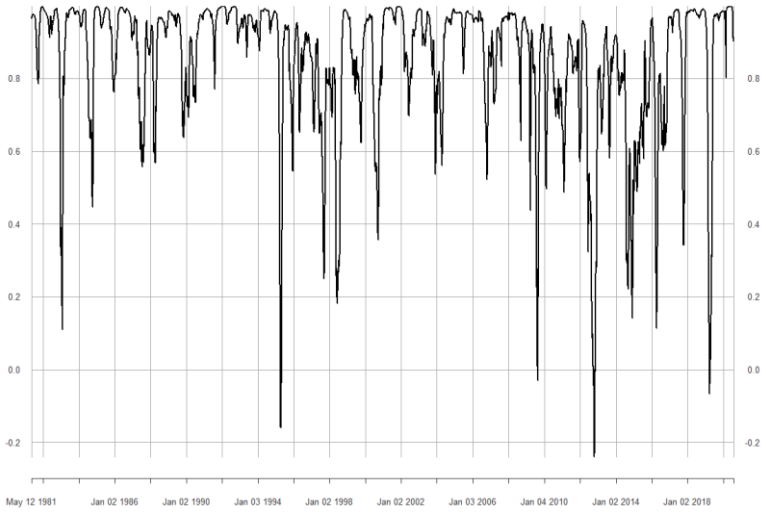
3.1. PCA анализ и доходност на държавни ценни книжа

Изключително високата корелация между доходността на ценните книжа, участващи в експеримента, е предпоставка за успешно използване на анализа на главните компоненти. Преди да бъдат представени и неговите резултати, трябва да направим едно уточнение, което кореспондира с възможно ограничение в PCA. Стойностите на корелацията са изчислени на базата на всички налични 9686 наблюдения, което означава, че резултатите може да бъдат други, ако използваме по-малък времеви интервал.



Фигура 36. Възвръщаемост на различните ДЦК и корелация в тяхната доходност за изследвания период

Описаният ефект е видим на Фигура 37, която представя промените в коефициента на корелация, но при условие че той се изчислява само върху последните 90 стойности (т.е. интервал от 3 месеца). Хипотетичният инвеститор, чиито изисквания бяха дефинирани в началото на изследването, има достатъчно свободно определен инвестиционен хоризонт и това оправдава използването на дългосрочните зависимости. Когато се налага да ограничим нашите анализи само до времеви отрязък с относително малка продължителност, тогава и резултатите от PCA анализа ще бъдат различни.



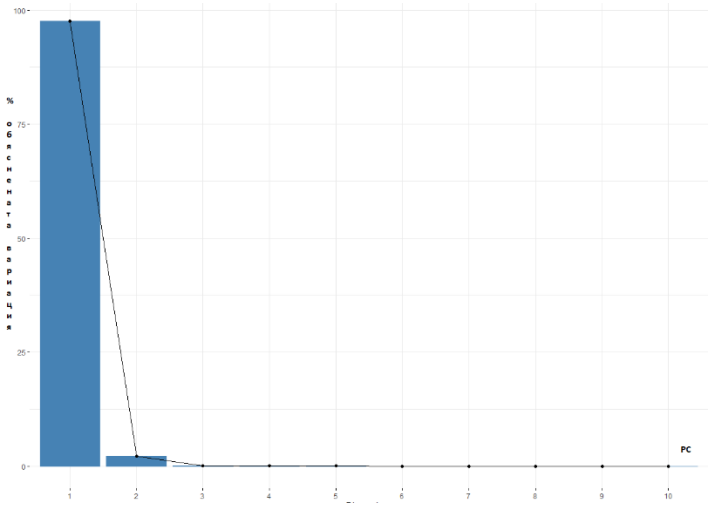
Фигура 37. Промени в коефициента на корелация между доходността на ЦК с най-малък и най-голям матуритет спрямо период от 3 месеца

Високата корелация се вижда и при получените с помощта на PCA резултати. Първият главен компонент съответства на 97.68% от наблюдаваните колебания. Заедно с втория главен компонент общо могат да обяснят 99.82%. Вече получените резултати от PCA анализа могат да се използват, за да допълнят нашето разбиране за промените в наблюдаваните рентабилности по няколко различни начина:

- Като изследваме отделните главни компоненти и ги обвържем с това какъв тип изменения те представляват.

Тъй като главните компоненти са линейна комбинация от първоначалните променливи, то те могат да бъдат тълкувани във връзка с кривата на доходността, коментирана в Част II. При това отделните компоненти могат да послужат при разлагането на промените в тази крива – изместване или извиване.

- Като помогнат за редуциране на размерността на данните, с които работим, когато доходността от облигациите е само част от по-мащабен модел.



Фигура 38. PCA анализ, приложен към държавните ценни книжа

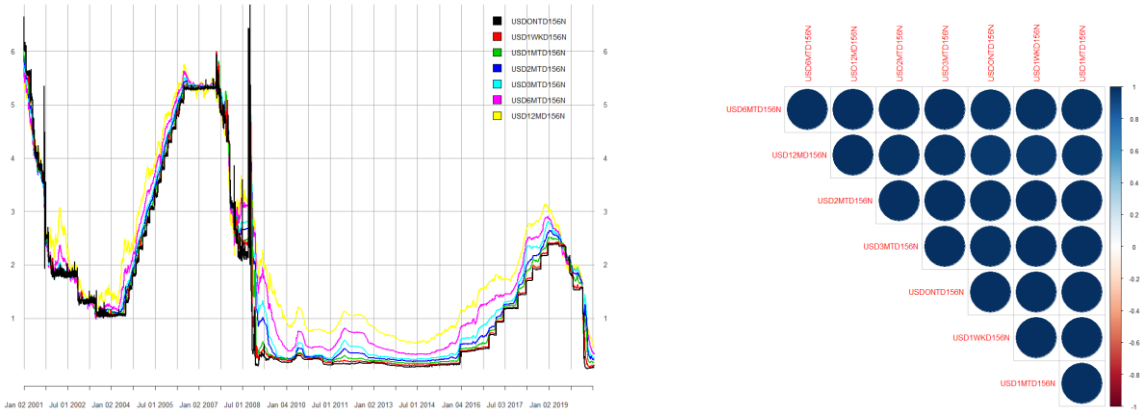
За икономически изследвания, които зависят не само от абсолютната стойност на доходността на държавните ценни книжа, но и от нейната промяна с матуритета, тези резултати са много добра възможност да се снижи размерността на моделите и да се ускори тяхната работа.

Таблица 21. PCA анализ на доходността на американски ДЦК

Параметър	**Стандартно отклонение**	**Дял от вариацията в началната съвкупност**	**Обща вариация в началната съвкупност**
PC1	3.1254	0.9768	0.9768
PC2	0.46197	0.02134	0.99817
PC3	0.11632	0.00135	0.99952
PC4	0.04713	0.00022	0.99974
PC5	0.0369	0.00014	0.99988
PC6	0.02388	0.00006	0.99994
PC7	0.01595	0.00003	0.99996
PC8	0.01378	0.00002	0.99998
PC9	0.01089	0.00001	0.99999
PC10	0.00757	0.00001	1

3.2. PCA анализ и LIBOR

Провеждането на същия експеримент със стойностите на LIBOR води до аналогични резултати. Всъщност, като отчитаме краткосрочния характер на данните, с които разполагаме, може да се очаква, че ефектите от PCA анализа ще бъдат също толкова силно изразени. Представените на Фигура 40 промени в корелацията на овърнайт и 12–месечният LIBOR са пресметнати въз основа на последните 90 дни, за да се постигне сравнимост с анализа на американските ДЦК, макар че максимален матуритет от една година предполага и по-кратка продължителност на този „прозорец“.



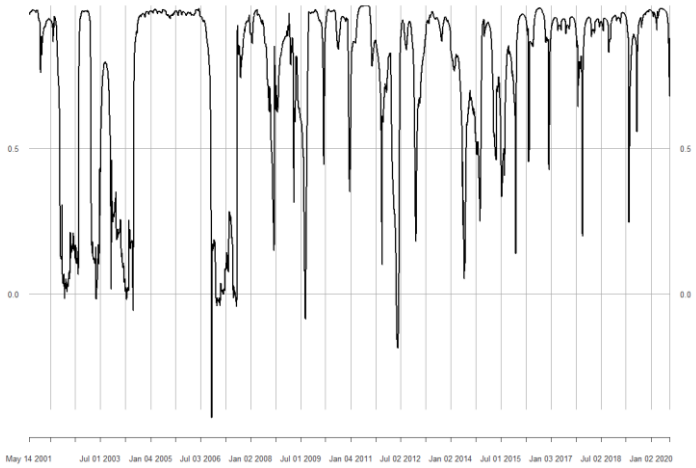
Фигура 39. LIBOR и корелация между стойностите за различен матуритет

Резултатите от PCA анализа, представени в Таблица 22 и на Фигура 41, потвърждават нашите очаквания, че предвид по-късия общ период се наблюдава също толкова силна зависимост, и с помощта на първия главен компонент се обясняват почти 99% от промените в LIBOR.

Таблица 22. Резултати от PCA анализа на LIBOR

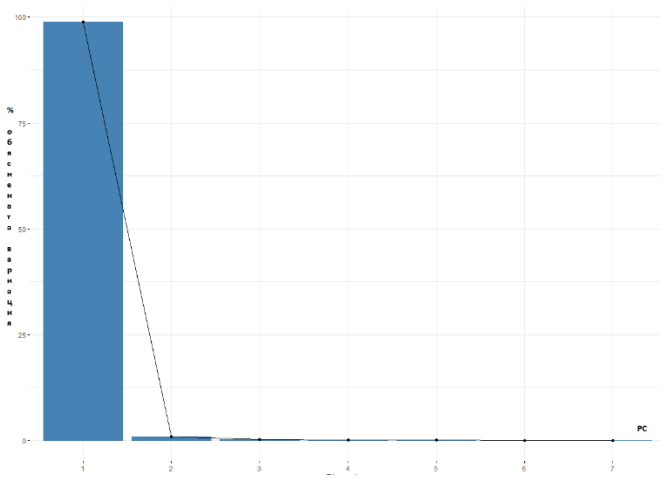
Компонент	**Стандартно отклонение**	**Дял от вариацията в началната съвкупност**	**Обща вариация в началната съвкупност**
PC1	2.6317	0.9894	0.9894
PC2	0.2411	0.0083	0.9977
PC3	0.11008	0.00173	0.99942
PC4	0.05114	0.00037	0.99979
PC5	0.03189	0.00015	0.99994
PC6	0.0165	0.00004	0.99998
PC7	0.0123	0.00002	1

Както и в случая с държавните ценни книжа, отделните главни компоненти може да се използват за оценка на ефектите, които водят до промяна в кривата на доходността. Предимствата на кратките периоди в случая са свързани с това, че PCA резултатите за LIBOR са подходящи за оценка на краткосрочните промени.



Фигура 40. Промени в коефициента на корелация между овърнайт и 12M LIBOR спрямо период от 3 месеца

Преди да приключим експеримента, свързан с приложение на PCA за анализ на доходността, е редно да посочим няколко важни ограничения на метода. Ако те не бъдат правилно отчетени, това може да доведе до „предоверяване“ и в крайна сметка до финансови загуби.



Фигура 41. Резултати от анализа на главните компоненти, приложен към LIBOR

Както много често се случва, ограниченията са свързани с неявните предположения, които съпътстват използваните аналитични инструменти.

В контекста на това изложение и проведените експерименти най-важните от тях са:

1) Допускането, че първоначалните данни, които използваме за анализа на главните компоненти, достатъчно добре представят промените в доходността (или лихвените проценти, както е в случая с LIBOR). Въпреки че това допускане няма отношение към стъпките в PCA, то е изключително важно за качеството на крайните резултати. Защото, ако входните данни, които най-често са основани на исторически стойности, не описват добре възможните промени в лихвите и доходността, тогава допълнителната загуба на информация, свързана с избора на ограничен брой главни компоненти, ще доведе до още по-голяма грешка.

Една от възможностите да се преодолеят ограниченията, свързани с качеството на изходната точка, от която започва анализът е освен исторически стойности да се използват и симулирани такива. Генерирането на тези допълнителни случаи може да стане с помощта на предварително избран модел, така че да се обхванат случаи, липсващи в историческата извадка. Страничен положителен ефект от този подход е, че с генерираните данни може да се избегнат грешки в резултат на субективен и неправилен избор на определен времеви интервал, от който са взети историческите данни.

2) Решението за избор на определен брой главни компоненти. Ако запазим всички, тогава ще можем да представим с точност всяка една от кривите на доходност, които са налице във входните за PCA анализа данни (но това оставя отворен въпроса с точността, описан в предходната точка), обаче няма да решим въпроса за размерността. Ако изберем твърде малко главни компоненти, това ще доведе до ниска размерност на модела, но и до висока неточност.

Няма точна „рецепта“ дали е по-добре да се използват по-малко, или повече главни компоненти, както няма и определен праг какъв процент на обяснена вариация е „достатъчен“. Като се отчита итеративният характер на повечето изследвания, свързани с анализ на данни, препоръчителната стратегия е да се направи опит с по-голям брой главни компоненти, като постепенно се свива размерността, докато се намери желаният баланс.

4. Изводи

В тази част бяха представени част от алгоритмите за клъстеризация и обработка на данни, които разчитат на неконтролирано обучение. Проведените експерименти и анализ на икономически и финансови проблеми са малка, но важна част от възможните приложения, които носят реални ползи и позволяват да се открояят по-добре характеристиките на проучваните процеси. Основните резултати може да бъдат обобщени по следния начин:

- Алгоритмите, използващи неконтролирано обучение, могат да помогнат в групирането на икономическите агенти въз основа на финансови и нефинансови показатели.

Тази стъпка предоставя ценна допълнителна информация при провеждането на сравнителен анализ, независимо дали той е насочен към оценка на финансовата стабилност на отделните участници, или към проверка до каква степен те имат сходни икономически показатели.

- При малък брой неврони самоорганизиращите се мрежи на Кохонен имат резултати, които са близки до K-Means при сравняване на кредитни институции по данни от стрес теста на EBA през 2018 г.

Проведеният експеримент с част от показателите, събирани от EBA в рамките на стрес тестовете от 2018 г., демонстрира, че може да се използва и по-простият и лесен за обяснение вариант, когато резултатите са сходни.

- Използването на различни показатели за дистанция между отделните наблюдения дава възможност за комбиниране на качествени и количествени показатели.

Резултатите в това изследване са получени с използване на Евклидово разстояние, което е оправдано предвид наличието на числени показатели за финансовото състояние на банките. При необходимост от включване на категорийни променливи K-Means е приложим, както е демонстрирано в [65].

- PCA анализът може да окаже ценно съдействие при намаляване на размерността на използваните модели.

Изследването на сложни икономически процеси предполага включването на голям брой променливи, което може да се превърне в сериозен проблем както от гледна точка на необходимото време за пресмятане, така и спрямо нуждата от обяснение на получените резултати. В този случай PCA може да подобри ефективността на аналитичните инструменти както в ролята си на средство за намаляване на размерността, така и като самостоятелен подход, който насочва вниманието към ефектите, обяснявани от съответните главни компоненти.

5. Изходен код и възпроизводимост на резултатите

Изходният код на R, използван за експериментите и изследванията, свързани с неконтролирано обучение, заедно с необходимите библиотеки са представени в Таблица 23. Обучението на самоорганизиращата се мрежа е извършено, без да се използват библиотеки за паралелни изчисления. Предвид малкия обем на входните параметри и стойности това няма съществено отражение върху бързодействието на цялата система.

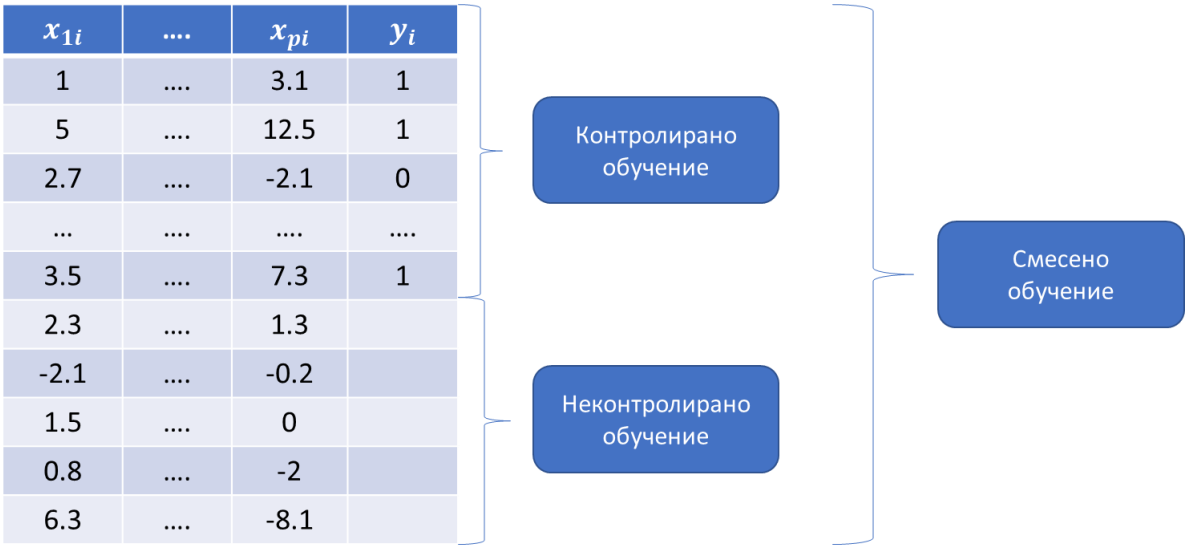
Таблица 23. Изходен код и библиотеки, използвани в Част IV

Елемент	***Бележки***
Изследване на кривите на доходността	
Изходни кодове	https://github.com/drnmy/mlf (Папка IV)
Източник на данни	Quandl (изисква безплатна регистрация)
Използвани библиотеки в R	
Quandl	https://www.quandl.com/tools/r
forecast	https://cran.r-project.org/web/packages/forecast/forecast.pdf
gglot2	https://ggplot2.tidyverse.org/
tseries	https://cran.r-project.org/web/packages/tseries/tseries.pdf
tidyverse	https://cran.r-project.org/web/packages/tidyverse/index.html
cluster	https://cran.r-project.org/web/packages/cluster/cluster.pdf
factoextra	https://cran.r-project.org/web/packages/factoextra/index.html

Kernlab	https://cran.r-project.org/web/packages/kernlab/index.html
Библиотеки, използвани при анализа с невронни мрежи	
kohonen	https://cran.r-project.org/web/packages/kohonen/kohonen.pdf

V. Смесено обучение при анализа на финансови инструменти

Тази форма помага да се комбинират предимствата на контролираното с възможностите, предоставяни от неконтролираното обучение. Когато разполагаме със съвкупност с частично налични крайни/желани резултати, както е показано на Фигура 42, възможностите за тяхното изследване са – или да използваме за контролирано обучение известните y_i, или да не се възползваме от тях въобще, като преминем към неконтролирано обучение.



Фигура 42. Връзка между разгледаните до момента типове обучение

И при двете описани възможности е налице загуба на информация, която смесеното обучение се опитва да елиминира, като оползотворява y_i, за да постигне по-добри резултати, отколкото е възможно с помощта на изцяло неконтролирано обучение. Историята на тази идея може да се проследи назад във времето до втората половина на 60-те години на XX век, когато в редица изследвания ([73], [74]) се използват т.нар. „обвиващи алгоритми“ за решаването на проблема с липсващите y_i. Тяхната цел е да обработят всички входни данни стъпка по стъпка:

1) За данните с известни y_i се извършва контролирано обучение с помощта на предварително избран алгоритъм.
2) Полученият модел се използва за прогнозиране на желаната стойност на едно или повече от наблюденията с неизвестни y_i.
3) Изпълняваме отново стъпка 1), но вече с допълнените на втората стъпка данни.

Привлекателна черта на описания алгоритъм е, че той е относително прост и лесен за реализация. Същевременно е налице силна чувствителност спрямо избрания метод за контролирано обучение от стъпка 1), както и към начина, по който се оценяват и минимизират грешките. По тази причина описаните три стъпки може да имат различен ефект и съответно ефективността от включването на наблюденията без известни крайни стойности може да носи различна полза.

Погледнато от чисто философска гледна точка, смесеното обучение има редица предимства – колкото и малко да са наблюденията с предварително известни стойности за y_i, те все пак носят информация и би трябвало да ни помогнат да получим по-добри резултати, отколкото ако не разполагахме с тях. За практически цели не е толкова лесно да се превърне това твърдение в реалност. Причините за това са няколко, като основните от тях са:

- Начинът, по който се дефинира смесеното обучение спрямо допусканията за данните, с които разполагаме;

Ситуацията, изобразена на Фигура 42, е една от най-често срещаните при разглеждания вид обучение, но далеч не е единствената възможна. Съществуват няколко варианта с различни допускания какво точно се съдържа в характеристиките на наблюденията, с които разполагаме:

  - Наблюдения с известни крайни стойности и такива без (Фигура 42).
  - Наблюдения без точно определени крайни стойности, но с предварително ясни ограничения;

Пример за такъв случай би била ситуация, в която разполагаме с информация за това кои наблюдения водят до различни крайни резултати, но без да знаем какви точно са те. Или ограничения, които определят границите на y_i, но отново без да фиксират техните стойности.

  - Наблюдения без точно определени крайни стойности, но с обосновано допускане за разпределението на желаните резултати.

При наличието на толкова разнообразни възможности е логично да възникне въпросът за приложимостта и ефективността на смесеното обучение. Както останалите видове, така и този вид обучение разчита на определени допускания ([75]) и тяхното познаване е важно, за да се избегне

прилагането на неподходящи за конкретните проблеми аналитични инструменти:

- непрекъснатост, свързана с това, че ако имаме входни характеристики с близки стойности $x_1 = (x_{11}, \ldots, x_{1p})$ и $x_2 = (x_{21}, \ldots, x_{2p})$, тогава и резултатите y_1 и y_2 също ще бъдат близки;
- склонност към клъстеризация, свързана с това, че ако имаме две наблюдения, принадлежащи към един клъстер, то тогава вероятността те да бъдат от един и същи тип/клас е висока;
- съкратимост, която е свързана с възможността да избегнем вече коментираното „проклятие на размерността“.

Идеята за съкратимост може да бъде илюстрирана, ако си представим, че искаме да изследваме финансови данни, които са породени от процес с малък брой характеристики. Вместо да анализираме съвкупността от крайни стойности, за които, ако нямаме идея от процеса, ще се получи много висока размерност на решавания проблем, може да фокусираме нашите усилия върху определяне на параметрите на генериращия процес. Това ще доведе до решаване на задача с много по-малка сложност и съответно до по-бързо получаване на адекватни резултати.

1. Алгоритми, използващи смесено обучение

Не е възможно, а и не е необходимо за нуждите на изследването да се представят подробно всички алгоритми, използващи смесено обучение. За да бъдат правилно разбрани ограниченията на отделните възможности, трябва да се посочат особеностите поне на отделните класове аналитични инструменти.

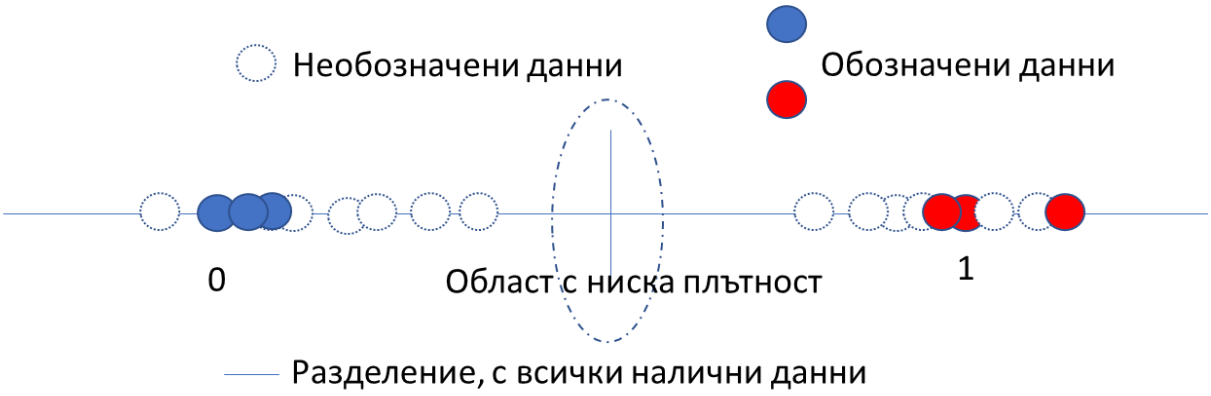
- Генеративни модели (generative models).

Идеята на генеративните модели е, че данните могат да бъдат описани с помощта на комбинация от параметрични модели (например като комбинация от нормални разпределения). Характеристиките на тези модели могат да бъдат намерени с помощта на оптимизационни алгоритми, като най-често използваният е Expectation-Maximization (EM) ([21]). Важно предимство на този подход е, че знанието, с което разполагаме за проблема, може директно да се включи при неговото разрешаване (в примера под

формата на допускането коя фамилия разпределения да използваме). Едновременно с това, ако направеният избор е грешен, това автоматично ограничава ползите, които можем да получим от генеративния модел.

- Модели с разделяне на базата на плътност (low-density separation).

Моделите, разчитащи на разделението на базата на плътност изхождат от предположението, че данните, с които разполагаме, са „ясно разделени“ (т.е. има забележима и значима разлика между отделните случаи). Така, ако трябва да поставим „демаркационна линия“ между отделните наблюдения, можем да предполагаме, че тя ще бъде в регион с малко на брой случаи (т.е. в област с ниска плътност). Ако противното е вярно, то това би нарушило допускането за ясно разделение, защото в граничната зона ще се наблюдават много „размесени“ случаи, които имат различни характеристики.



Фигура 43. Групиране с модел, използващ разделяне на базата на плътност

Един от най-често прилаганите самообучаващи се алгоритми, основан на този тип разделяне, е S3VM (Semi-Supervised Support Vector Machine)[21], разгледан подробно в [76].

- Граф базирани модели (graph based models).

Тази група алгоритми разчитат на представянето на отделните наблюдения като възли в граф, ребрата на който се обозначават („претеглят“) със стойности в зависимост от разликата между свързаните елементи[22]. При определяне на теглата за възли, при които липсват данните

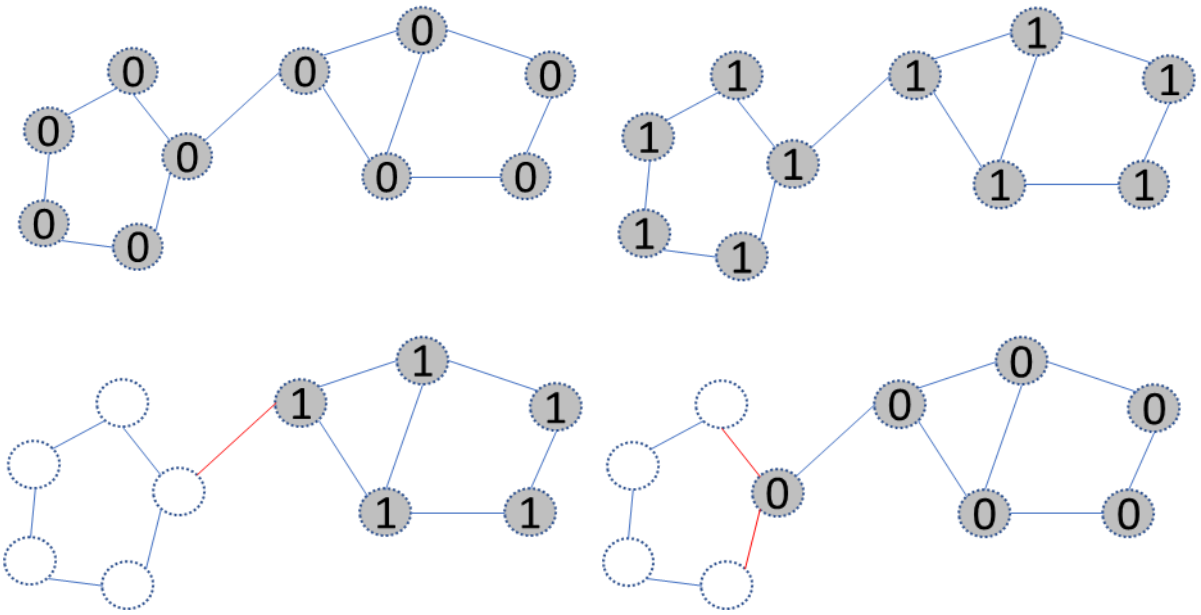
[21] Вж. също TSVM (Transductive Support Vector Machine) в [11].
[22] По този начин липсата на връзка би означавала безкрайно голяма разлика.

за y_i, се използва принципът за минимизиране на съвкупното разстояние от всички пътища, свързващи съответните елементи.

На Фигура 44 е показана работата на един граф базиран модел, при който трябва да класифицираме данните в две категории – „1“ и „0“. Ако дефинираме функцията, която искаме да минимизираме, като:

$$E(y) = \frac{1}{2}\sum_{i,j} w_{ij}(y_i - y_j)^2, \qquad (28)$$

тогава има две благоприятни възможности за стойностите на съседните възли (1-1 и 0-0), както и две неблагоприятни (1-0 и 0-1). При липса на каквато и да е информация за резултати y_i вариантите, при които всички елементи са маркирани като единица или като нула (представени с горните два графа на Фигура 44), водят до минимално $E(y) = 0$ – защото $\sum_{i,j} w_{ij}(y_i - y_j)^2 = 0$.



Фигура 44. Пример за граф базиран алгоритъм и класификация 1-0

Когато имаме съвкупност от наблюдения с известни y_i, както и такива с липсващи стойности, тогава *E(y)* може да бъде минимизирано с последователни промени в присвояването на „1“ или „0“ за възлите, където няма предварителна информация. По какъв начин ще се променят поставените *y*, зависи от началното състояние на графа (например в долната лява част на Фигура 44 ще имаме $\sum_{i,j}(y_i - y_j)^2 = 1$, а в дясната – $\sum_{i,j}(y_i - y_j)^2 = 2$).

Съществува още една принципна възможност, при която просто се игнорират наличните стойности за y_i, като по този начин обучението се превръща в неконтролирано. На неконтролираното обучение беше посветена отделна част от изследването и то няма да бъде разглеждано отново тук.

2. Приложимост на смесеното обучение при анализа на финансови инструменти

Практическата приложимост на алгоритмите, разчитащи на смесено обучение, зависи от две предпоставки:

- Дали са изпълнени допусканията, описани в точка 1.

Допусканията определят и съвкупността от алгоритми, между които може да се направи избор при решаването на конкретния проблем.

- Дали данните с липсващи резултати/крайни стойности превишават значително като брой тези с известни y_i.

Съществуват редица случаи, в които е икономически неизгодно да се прегледат и класифицират/оценят всички данни, а това може да се направи само за една малка част от тях. Причините могат да бъдат както финансови (скъпоструващи експерти), така и технологични – например времето за реакция е много малко, за да позволи преглед на информацията от човек.

3. Анализ на борсов индекс с EM и смесено обучение

Разпределението на възвръщаемостта на отделни ценни книжа и на борсовите индекси е важна тема, предмет на анализ в редица изследвания ([77], [78], [79]). Изследванията в тази посока дават възможност да се отговори на два критични за успешните инвестиционни стратегии въпроса:

- Как разпределението на реално наблюдаваните възвръщаемости се отклонява от нормалното.

Отклоненията спрямо нормалното разпределение при възвръщаемостта на ценните книжа е проблем както при оценката на риска, така и при формирането на цялостна инвестиционна стратегия. Емпиричните проверки, които показват наличието на положителен ексцес ([80], [81]), а в някои случаи и на асиметрия, поставят на изпитание допусканията, с които са свързани редица широко използвани финансови модели. Една от възможностите, които позволяват да се обяснят наблюдаваните отклонения, е свързана с прилагането на Скритите Марковски модели (Hidden Markov Model – HMM - [83], [83]). За да представим промените на пазара в този контекст, допускаме, че той може да бъде моделиран с помощта на Марковски процес, но при неизвестни („скрити“) състояния. Икономическата обосновка на отделните фази, както и възможните преходи от едно положение в друго са връзката, която е необходима, за да свържем аналитичния инструмент с финансовата теория.

- Какви са причините за тези отклонения и възможно ли е те да бъдат моделирани и изследвани.

Причините за наблюдаваните отклонения могат да бъдат различни, като се започне от времевия хоризонт ([84]) и се стигне до конкретни промени в икономическата среда. Подробните постфактум обяснения са важни, защото подобряват нашето разбиране за начините, по които се променят пазарите. От гледна точка на дефинирания в началото на изследването инвеститор обаче много по-съществени са точните прогнози. За практическата приложимост на моделите е разумно подробните причини за пазарните колебания да бъдат обединени и анализирани като група. По този начин не само се спестяват усилия, но и се намалява възможността да се изразходват усилия за моделирането на промени с чисто случаен характер.

Ако комбинираме HMM с идеята за групиране на причините, които водят до по-висок ексцес и асиметрия в доходността на ценните книжа, можем да представим характеристиките на наблюдаваната възвръщаемост, като резултат от „наслагването“ на няколко отделни разпределения. Всяко от тях е свързано с определена фаза от развитието на пазара, а с течение на времето се наблюдават и преходи между отделните състояния. В този контекст асиметрията и ексцеса на наблюдаваните стойности са функция както на разпределенията в отделните фази, така и на вероятността за преход между тях.

3.1. EM и оценка на разпределенията в отделните състояния

За да проверим хипотезата, че промените на финансовите пазари могат да бъдат ефективно моделирани с помощта HMM и комбинация от разпределения, ще разгледаме възвръщаемостта на два от водещите борсови индекса – S&P500 и DAX. Има няколко аргумента в подкрепа точно на този избор:

- и двата индекса отразяват състоянието на развити капиталови пазари и водещи икономики – САЩ и Германия;
- свободно достъпна е историческа информация с достатъчна продължителност и висока честота на наблюденията;
- и двата борсови индекса са обект на нестихващ интерес, което позволява да се направи сравнение на получените резултати с други изследвания.

Както показват стойностите в Таблица 24 и диаграмите на Фигура 45, разпределенията на възвръщаемостите се отклоняват от нормалното, като най-съществен е този ефект по отношение на ексцеса. И двата теста – на Андерсон – Дарлинг и на Шапиро – Уилк, категорично отхвърлят хипотезата за нормалност. Следователно няма да бъде достатъчно, ако в последващия анализ използваме единствено първите два централни момента, за да опишем разпределението на възвръщаемостта.

Ако приемем, че в разглеждания исторически период пазарите са преминавали през различни фази, тогава може да се провери дали емпирично наблюдаваното разпределение се описва добре като комбинация от няколко разпределения – всяко едно, представящо съответното състояние на пазара. Не е необходимо отделните етапи да се характеризират с разпределения от една и съща фамилия, но подобно допускане би опростило значително изследването. Използваният анализ е валиден дори и ако предположението не е основателно, като в този случай е необходима допълнителна стъпка, за да оценим характеристиките на всяко едно икономическо състояние. Така например в периоди на подем можем да приемем, че определящо е нормалното разпределение на възвръщаемостта, а в случай на криза – Student’s t.

Независимо кой вариант използваме, първото условие, което трябва да дефинираме е какъв е броят на различните фази. Има няколко възможности

да направим това – въз основа на анализ на историческите събития и периоди; под формата на оптимизационна процедура, като се сравнява грешката при различен брой състояния. В изследването е използвана първата възможност, като са определени три различни фази – подем, криза, устойчиво развитие.

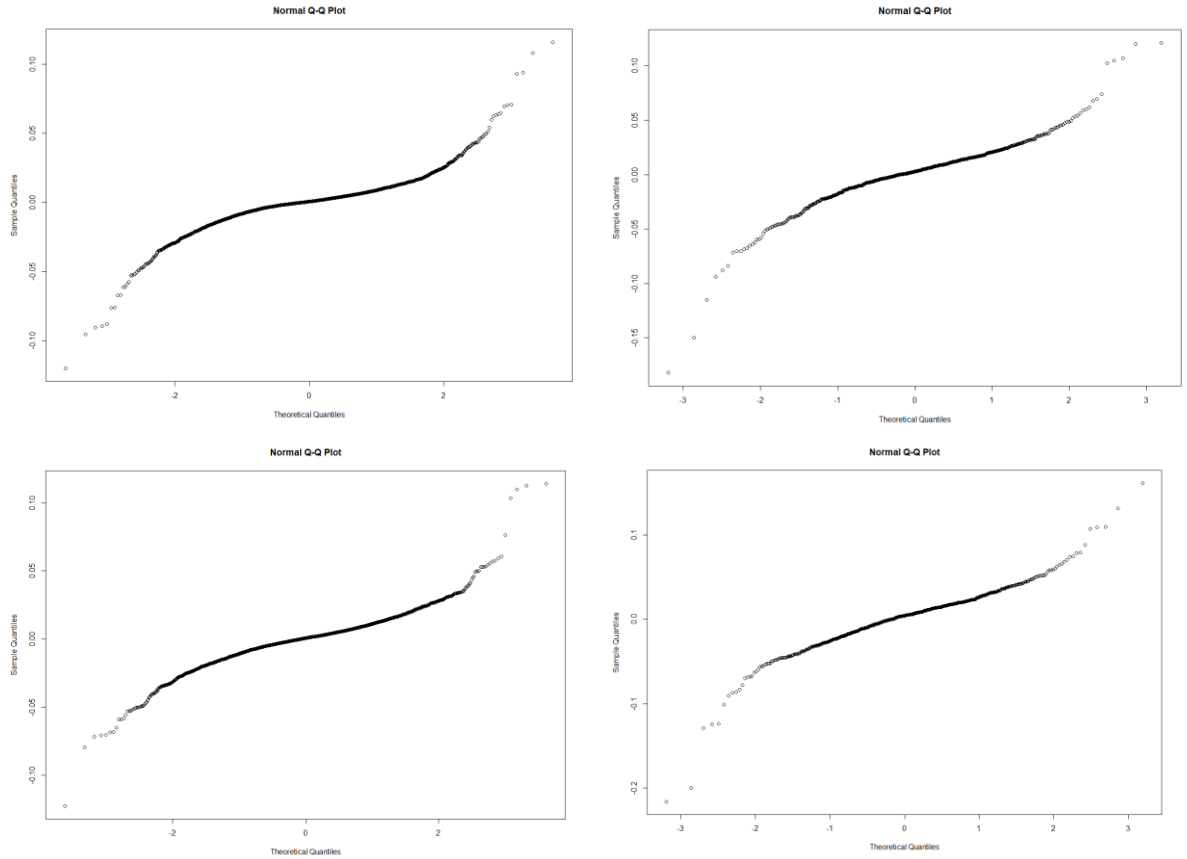
Таблица 24. Характеристики на изследваните динамични редове

Параметър	***S&P 500***	***DAX***
Начална дата	03 януари 2007 г.	01 януари 2007 г.
Крайна дата	04 август 2020 г.	04 август 2020 г.
Брой наблюдения	3421	3446
Възвръщаемост на дневна база		
Средна	0.0003346794	0.0002860084
Стандартно отклонение	0.01316212	0.01425071
Коефициент на асиметрия	–0.2546156	–0.01130767
Коефициент на ексцес	15.78465	11.24446
Anderson – Darling тест	A = 3286, p-value < 2.2e-16	A = 3399.8, p-value < 2.2e-16
Shapiro – Wilk тест	W = 0.85482, p-value < 2.2e-16	W = 0.92056, p-value < 2.2e-16
Възвръщаемост на седмична база		
Средна	0.001537656	0.001413639
Стандартно отклонение	0.02602078	0.03192917
Коефициент на асиметрия	–0.7389025	–0.7949187
Коефициент на ексцес	10.56055	10.02441
Anderson – Darling тест	A = 668.32, p-value < 2.2e-16	A = 692.49, p-value < 2.2e-16
Shapiro – Wilk тест	W = 0.90566, p-value < 2.2e-16	W = 0.9252, p-value < 2.2e-16

Възможно е да разширим броя на състоянията, като това би дало по-добра възможност моделът да отчита и по-краткосрочни промени в изследваните индекси. Едновременно с това големият брой възможни преходи от една фаза в друга води до необходимостта от по-голям брой данни, за да може да оценим параметрите на всеки един етап, както и до по-сложни за обяснение и практическа употреба резултати.

Допълнителен аргумент в подкрепа на тезата за запазване на относително малък брой състояния е дългият инвестиционен хоризонт. Ако хипотетичният участник на финансовите пазари и потребител на

резултатите от анализа има достатъчно дългосрочни намерения, за него биха били важни преди всичко по-сериозните промени в развитието на пазара. Съответно по-лесните за обосновка и обяснение модели имат предимство в тази ситуация.



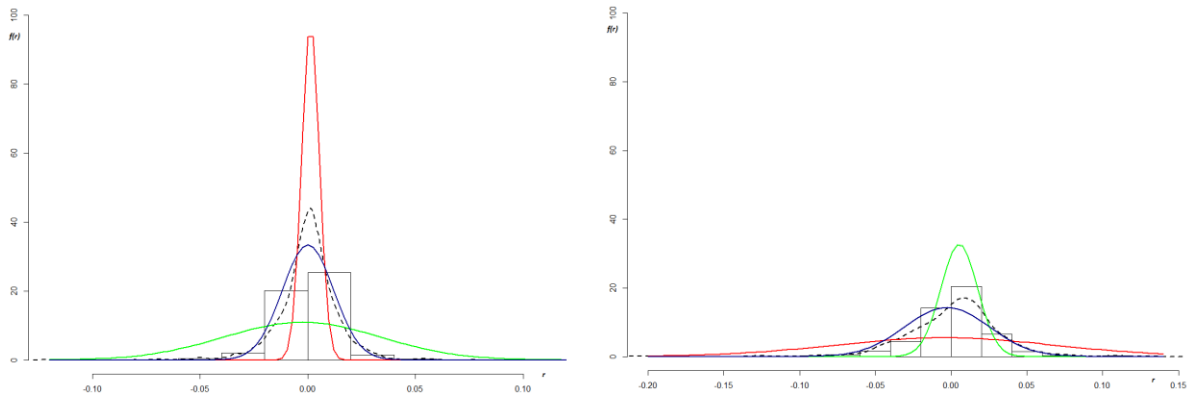
Фигура 45. Q-Q диаграма на възвръщаемостите на дневна (ляво) и седмична (дясно) база за S&P500 (горе) и DAX (долу)

Резултатите от работата на EM алгоритъма, представени на Фигура 46 и Фигура 47, показват разликите в характеристиките на отделните нормални разпределения, които биха съответствали на трите състояния. Основното предимство, което получаваме при използването на този аналитичен инструмент, не е само възможността да моделираме по-точно наблюдаваните възвръщаемости, а да вземем решение съобразно с параметрите, които най-точно отговарят на текущото състояние на пазара.

За да бъдат наистина практически приложими получените резултати обаче, има три важни въпроса, които трябва да намерят своя отговор:

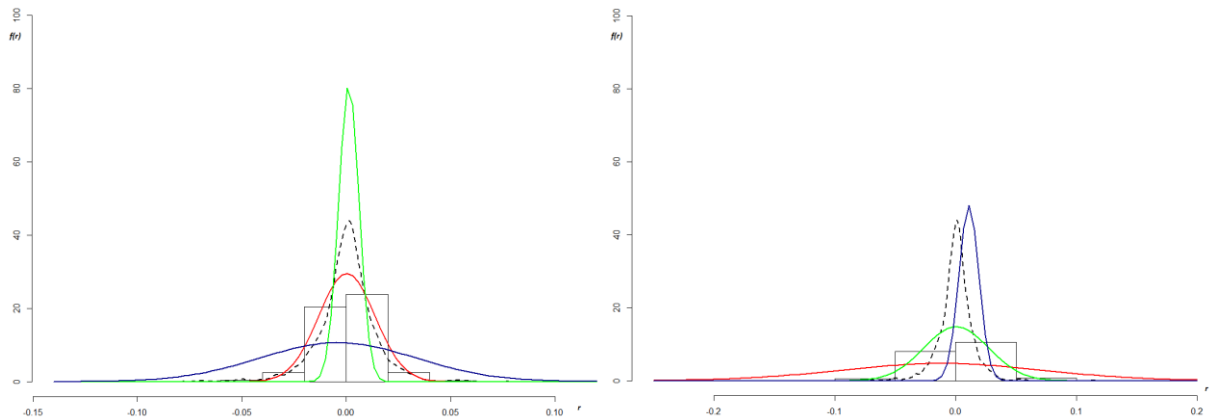
- Дали броят на състоянията, които EM алгоритъмът използва, е правилен.

- Дали фамилията разпределения, избрани, за да характеризират всяко едно състояние, е адекватна на неговите особености;
- можем ли да определим по достатъчно надежден начин в кое състояние се намираме в момента.



Фигура 46. Дневна (ляво) и седмична (дясно) възвръщаемост за DAX, представени като микс от три нормални разпределения

Частичен отговор на първия въпрос вече беше даден в контекста на инвестиционния хоризонт и търсенето на баланс между точност и възможности за обобщаване на пазарните промени. Формална проверка дали избраният брой фази е достатъчен, може да се направи с помощта на разделяне на входните данни на части и оценка на възможностите за моделиране поотделно. Втора възможност да дадем отговор на този въпрос, е да проследим как се променя подобието при прилагане на EM с различен брой състояния.



Фигура 47. Дневна (ляво) и седмична (дясно) възвръщаемост за DAX, представени като микс от три нормални разпределения

Отговорите на следващите два въпроса имат много общо, защото, за да направим проверка дали избраната фамилия разпределения наистина има отношение към поведението на ценните книжа и индекси в дадена ситуация, трябва да можем предварително да определим кога тя настъпва, както и кога вече е налице преход към друго състояние. По своята същност това е класификационен проблем и съответно неговото решение може да бъде подпомогнато с прилагането на самообучаващи се алгоритми. Трябва да се отбележи, че предимствата на автоматизираната класификация не са свързани само с точността на групиране, а също така и със скоростта на пресмятане. Ако сравним самообучаващия се алгоритъм с друг подход – например експертна оценка, то за последната са нужни значително повече време, усилия и финансови ресурси.

3.2. EM и оценка на пазарните състояния

Смесеното обучение може да подобри класификацията на състоянията, като използва следните прости принципи:

1) Историческите данни обикновено съдържат достатъчно информация за минали събития, които могат да бъдат определени с висока точност като състояние на подем, криза или устойчиво развитие.

Не е необходимо целият набор от исторически наблюдения да бъде обработен и класифициран – и в това е едно от основните предимства, които ни предоставя смесеното обучение. Ако изберем вариант, при който състоянията се класифицират изцяло с неконтролирано обучение, тогава рискуваме да не можем да определим достатъчно точно границите и преходите между отделните етапи. Ако се използват само предварително класифицирани данни, тогава, освен разходите за ресурсите необходими за тази обработка, има опасност да получим изместени резултати. Основната причина за това е, че относително рядко е налице пълен консенсус между експертите кога започва и приключва една криза. За сметка на това има периоди, за които няма спор в каква фаза се намират пазарите. Този факт може да помогне, защото позволява да бъдат класифицирани само тези периоди, за които съществува предварителен консенсус какво е състоянието. Самообучаващите се алгоритми могат да определят момента

на преход, като обработят отделните наблюдения като липсващи (немаркирани) данни.

2) Определянето на принадлежността на всяко едно наблюдение към определено състояние е бавно, като не е целесъобразно да се случва ръчно.

За да се избегне необходимостта от цялостна обработка на входните данни, може да се посочи състоянието само на част от наблюденията – например тези, свързани с кризисни периоди или бурен ръст. Останалите наблюдения могат да бъдат автоматично класифицирани от самообучаващия се алгоритъм. При работа с данните за двата пазарни индекса са използвани следните маркери – „0“ – за неизвестно/липсващо състояние, „1“ – за бурен ръст, „2“ – за устойчив растеж, „3“ – за криза.

3) Ефектите от работата на смесеното обучение могат да бъдат оценени, като се използват следните два показателя – разлика в броя на случаите, поставени във всяка група, спрямо неконтролирано обучение, както и промени в параметрите на разпределенията, характеризиращи всяка една фаза.

Резултатите от работата на EM клъстеризацията се съдържат в Таблица 25 и Таблица 26, като за смесеното обучение са предварително анализирани следните периоди – финансовата криза от пролетта на 2008 г. (за начало условно се приема краят на месец април), началото на 2014 г. до средата на 2015 г., както и средата на 2011 г. до края на първата половина на 2013 г. Сравнението на резултатите, получени с помощта на двата подхода, позволява да направим изводи не само относно общата точност на моделите, но и относно начините и честотата, с която пазарите преминават от едно състояние в друго.

Възможно е и друго групиране на периодите в зависимост от това колко прецизни предварителни анализи са направени на историческото развитие на изследваните индекси. При провеждането на експериментите в това изследване е използван консервативен подход, като целта е да се обозначат само безспорните периоди на криза и растеж. Причините за този избор са, че целта е да се покаже ефектът от работата на самообучаващите се алгоритми, а не да се провери една или друга теория за началото на последните финансови кризи. Подробен анализ на последното може да бъде намерен в [85] и [86].

Използваният за клъстеризация модел допуска наличието на предварително определен брой състояния, всяко от които се описва с нормално разпределение. Следователно функцията, която се оптимизира, е:

$$f(x|\gamma) = \sum_{k=1}^{K} \pi_k \varphi(x|\mu_k, \sigma_k), \qquad (29)$$

където π_k определят „пропорцията“, в която участва съответното нормално разпредение, K е броят на състоянията, а стойностите на π_k отговарят на ограниченията, дефинирани в (30).

Таблица 25. Резултати от EM клъстеризацията за S&P 500

Показатели	***Неконтролирано обучение***	***Смесено обучение***
Дневна база		
Разпределение A		
μ	–0.0029523856	–0.007257876
σ	0.034771148	0.0246556672
π	0.08244	0.17785
Разпределение B		
μ	0.0011478242	–0.004043481
σ	0.003796932	0.0007277789
π	0.39486	0.02426
Разпределение C		
μ	0.0002456637	0.002164707
σ	0.011286845	0.0080411242
π	0.52270	0.79788
Седмична база		
Разпределение A		
μ	–0.004975891	–0.004246673
σ	0.07215403	0.04062243
π	0.05258	0.2663
Разпределение B		
μ	–0.002355711	0.003683646
σ	0.02729122	0.01959564
π	0.41743	0.4575
Разпределение C		
μ	0.005285098	0.003627864
σ	0.01197810	0.01335885
π	0.52998	0.2761

$$\sum_{k=1}^{K} \pi_k = 1, \qquad (30)$$

$$0 < \pi_k < 1, \text{ за k=1..K}$$

Сравнението на резултатите, получени с помощта на двата подхода, позволява да направим изводи не само относно общата точност на моделите, но и относно начините и честотата, с която пазарите преминават от едно състояние в друго.

Таблица 26. Резултати от EM клъстеризацията за DAX

Показатели	***Неконтролирано обучение***	***Смесено обучение***
Дневна база		
Разпределение A		
μ	–0.0040380249	–0.0019364007
σ	1.336288e-03	0.0004133678
π	0.06204	0.2568
Разпределение B		
μ	0.0003119126	0.0014364546
σ	1.741621e-04	0.0001475934
π	0.64484	0.5104
Разпределение C		
μ	0.0011485813	0.0002212286
σ	2.177782e-05	0.0000838704
π	0.29312	0.2328
Седмична база		
Разпределение A		
μ	0.0004208206	0.004425888
σ	0.026870205	0.02153100
π	0.78159	0.2549
Разпределение B		
μ	–0.0111900610	0.002762041
σ	0.083611903	0.02726170
π	0.05984	0.4670
Разпределение C		
μ	0.0110578100	–0.003616735
σ	0.008050366	0.04417197
π	0.15856	0.2781

Основните заключения, свързани с приложението на смесеното обучение за анализ на прехода между различни състояния на пазарните индекси, могат да бъдат обобщени по следния начин:

- Изборът на подходящ алгоритъм за смесено обучение може да подобри оценката както за относителната значимост на отделните състояния, така и за техните параметри.

Този ефект е видим ясно, когато сравним отделните π_k за неконтролирано и смесено обучение от Таблица 25 и Таблица 26, като в случая става дума за съпоставяне на стойностите за един и същи индекс и една и съща продължителност (дневна и седмична). Неконтролираното обучение показва тенденция да поставя в един клъстер екстремните отклонения, което в случая дава много по-ниска стойност за π на съответния сценарий. От една страна – това може да доведе до подценяване на риска, а от друга – до грешни инвестиционни решения, тъй като е възможно да класифицираме неправилно състоянието на пазара.

Разбира се, остава отворен въпросът дали тези екстремни отклонения си заслужава да бъдат определени като самостоятелна и отделна фаза. Решението дали това да се случи, изисква допълнителен анализ и съобразяване с конкретната инвестиционна стратегия. В рамките на проведения експеримент се използват само три състояния, което за нуждите на изследването е оправдано, но за други цели може да се окаже, че води до твърде грубо разграничение.

- Коректната предварителна класификация на част от използваните данни е изключително важна за правилната работа на смесеното обучение.

Изкушаваща е идеята, след като работим с исторически данни, да маркираме колкото се може повече от тях. Логиката на подобно поведение е, че след като информацията се отнася за минали периоди, то тя би трябвало вече да е коректно и изцяло обработена. На практика е възможно да има проблеми с пълното обозначаване на историческите данни, като това не означава, че пазарите не са ефективни. Причините могат да бъдат както чисто технически – например свързани с честотата на данните и броя на

трансакциите[23], така и принципни – промяната на пазара е резултат и от очаквания на инвеститорите. А да се улови точният момент, в който очакванията се променят, е отделна и още по-трудна задача.

Използваният в експеримента метод за оценка на историческите наблюдения е основан на идеята да се маркират само тези моменти, за които има консенсус, че представят определена фаза от пазарното развитие. От една страна – това позволява да се избегне изместването заради липсата на обективност или пристрастия на външния фактор (експерт) оценяващ данните. От друга – позволява неявно да се възложи допълнителна задача на самообучаващия се алгоритъм – да очертае границите и момента на преход от едно състояние към друго.

- Фамилията от статистически разпределения, които описват отделните състояния, е изключително важна за точността на изградените модели;

Резултатите, представени в Таблица 25 и Таблица 26, са получени с помощта на допускането, че нормалното разпределение, макар и с различни параметри, може да опише добре поведението на пазара в различните състояния. Това допускане е много ограничаващо и не особено адекватно, особено за крайните случаи на рязък спад или бурно развитие, защото не може да обхване добре екстремните промени, които се наблюдават в тези ситуации. Въпреки тези недостатъци нормалното разпределение позволява да се покаже работата на алгоритмите със смесено обучение, както и да се подчертаят техните достойнства и недостатъци. Това оправдава употребата му в тази част от изследването.

4. Изводи

Смесеното обучение позволява да извлечем максимална полза от ситуации, при които има голям обем от данни, но само за част от тях разполагаме с желания/крайния резултат. Когато сме поставени пред избор дали да отхвърлим необозначените наблюдения, или да използваме изцяло

[23] Когато става дума за общия случай, нищо не гарантира, че пазарите са силно ликвидни и сделки точно с тези изследвани активи се случват непракснато. Пример за това е например идиосинкратичната липса на ликвидност на някои сегменти от пазара на облигации.

неконтролирана класификация, разгледаните в тази част алгоритми предоставят компромисен вариант. Ако разгледаме множеството от предварително известни резултати в по-широк смисъл, това увеличава обхвата на смесеното обучение, като позволява да се разглеждат проблеми с дефинирани ограничения или такива само с изискване за наличие на определена фамилия разпределение на резултатите.

Съществуват редица практически проблеми, за решаването на които е необходимо да се използва хибриден подход. Примери за такива задачи са анализът на новини за оценка на пазарните очаквания; класификацията на уеб ресурси; визуално разпознаване; класификация на нови събития и обекти. Във всеки един от изброените случаи разполагаме с голям брой наблюдения, чието подробно изследване би изисквало прекалено много ресурси и време. Едновременно с това, ако разчитаме само на вече обозначените данни, това би означавало доброволно да се лишим от ценна и полезна информация. Изследването на новопостъпилата информация е специфичен случай, който също заслужава специално внимание. При анализа на ценни книжа това е необходимост, тъй като пазарните промени е възможно да включват ефекти, които не са наблюдавани преди това.

От трите основни типа модели в изследването беше използван само един вариант, основан на EM алгоритъм и HMM. Целите на проведения експеримент бяха свързани с това да се провери възможността за асоцииране на промените във водещи борсови индекси с различни състояния на пазара. Резултатите от това изследване имат практическа значимост, защото позволяват 1) да се определи текущото състояние, а според използвания конкретен модел – и вероятността за преход към друга фаза; 2) да се изградят по-сложни инвестиционни стратегии, които отчитат етапа от развитие на финансовите пазари. Независимо от това, че анализираните исторически стойности са на борсовите индекси S&P 500 и DAX, разглежданият метод и предложените подобрения към отделните стъпки може да бъдат прилагани и към друг тип ценни книжа и пазарни сегменти.

Получените с помощта на EM резултати демонстрираха пригодността на смесеното обучение за анализ на състоянието на пазара. Сравнението с клъстеризация, без да се използват обозначените желани стойности, показа следните предимства на разглеждания подход:

- възможност да се адаптира лесно към различни проблемни ситуации, като позволи в изследването да се включат различни експертни мнения и исторически анализи;
- според вида на решавания проблем и изпълнението на основните допускания за приложимост на смесеното обучение, може да бъдат подбрани различни алгоритми, които да отговорят на установените ограничения и да се съобразят с обема на разполагаемата информация;
- гъвкавостта, свързана с избор на различни разпределения, които да представят различните състояния, позволява да се представи с по-малко ограничаващи допускания реалното състояние на пазарите.

Смесеното обучение ни предоставя още една възможност да погледнем на решаваните задачи под друг ъгъл, като се възползваме в максимална степен от наличната информация. Също както и другите разгледани алгоритми то не е универсален инструмент, който може автоматично да разреши всички проблеми и да постигне чудеса. Но познаването и прилагането му има важно практическо значение най-малкото защото начинът на работа е много близък до маниера, по който самите ние се учим.

5. Изходен код и възпроизводимост на резултатите

Изходният код на R, използван за експериментите и изследванията, свързани със смесено обучение, заедно с необходимите библиотеки са представени в Таблица 27. Един от специфичните проблеми при използването на EM алгоритмите е визуализацията на получените резултати. Най-често използваната възможност е графично представяне на функцията на плътността на съответното разпределение. Това решение има няколко важни недостатъка, тъй като е много трудно директно да се сравняват резултатите, получени с помощта на различни аналитични инструменти.

Таблица 27. Изходен код и библиотеки, използвани в Част V

Елемент	***Бележки***
Изследване на кривите на доходността	
Изходни кодове	https://github.com/drnmy/mlf (Папка V)
Източник на данни	Quandl (изисква безплатна регистрация), Yahoo Finance

Използвани библиотеки в R	
Quandl	https://www.quandl.com/tools/r
forecast	https://cran.r-project.org/web/packages/forecast/forecast.pdf
gglot2	https://ggplot2.tidyverse.org/
tseries	https://cran.r-project.org/web/packages/tseries/tseries.pdf
mixtools	https://cran.r-project.org/web/packages/mixtools/index.html
nortest	https://cran.r-project.org/web/packages/nortest/index.html
moments	https://cran.r-project.org/web/packages/moments/index.html
plotGMM	https://cran.r-project.org/web/packages/plotGMM/index.html
Библиотеки използвани при клъстеризация със смесено обучение	
EMCluster	https://cran.r-project.org/web/packages/EMCluster/index.html

VI. Специфични проблеми на самообучаващите се алгоритми и начини за тяхното преодоляване

Последната част на изследването е посветена на специфичните проблеми, които съпътстват употребата на самообучаващи се алгоритми. Тези трудности се повтарят, много често независимо от избрания алгоритъм и без да са част от него. Въпреки това, ако не бъдат разрешени по подходящ начин, те могат да доведат до некоректни резултати и да лишат от смисъл останалите стъпки в изследването. Според това към кой етап от проучването се отнасят проблемите, те могат да бъдат класифицирани като свързани с:

- предварителната обработка на данните;
- избор на конкретен аналитичен инструмент и алгоритъм;
- оценка на хиперпараметрите на избрания алгоритъм;
- оценка на точността и валидация на използваните аналитични инструменти.

Всяка една от тези категории има установени и доказали се с времето „добри практики“, които могат да помогнат и да ни предпазят от допускането на грешки. Това не означава, че с използването на натрупания опит можем да се справим лесно с всяка една проблемна ситуация – необходимостта да се обмислят стъпките в изследването и да сме креативни, е винаги актуална. Но когато вече има готово решение, много по-разумно е усилията да се насочат към комбинирането на най-подходящите за случая възможности, отколкото да се открива отново колелото.

1. Общи проблеми – общи решения

Общите решения са важни по две причини: първо те правят решението лесно разбираемо, тъй като са отдавна известни; второ – обикновено те са отдавна популярни и апробирани, което не изисква допълнителна аргументация защо трябва да бъдат използвани. За да запазим изложението кратко, общите проблеми и начините за справяне с тях са представени в табличен вид.

Има специална причина тази част от изложението да е последна и тя е, че дискутираните тук правила и препоръки са свързани с първите стъпки на

едно изследване. Бидейки част от предварителното планиране и обмисляне на анализа, много често добрите практики остават без изрично описание. Отчасти това се дължи на допускането, че те са толкова добре известни, че няма нужда да бъдат подчертавани отново. Надеждата на автора е, че като ги постави в края на този текст, това ще подчертае тяхното важно значение.

1.1. Добри практики при предварителната обработка на данни

Предварителната обработка на данни може да бъде условно разделена на две части:

- Трансформация на наличните характеристики, така че да отговорят на изискванията на прилаганите алгоритми (напр. стандартизация).

Не е задължително алгоритмите да изискват нормализиране на данните, за да бъде включена тази стъпка в анализа. Има ситуации, в които обаче е задължително изпълнението на някоя от описаните в Таблица 28 стъпки.

Таблица 28. Добри практики при трансформация на наличните характеристики

Добра практика	***Бележки***
Нормализация	Нормализацията е достатъчно добре позната стъпка, за да бъде коментирана подробно отново. Освен когато е задължително изискуема от използвания алгоритъм, смисълът от нейното прилагане е свързан най-вече с това, че може да осигури по-бърза работа при търсенето на оптимум (напр. при градиентното спускане).
Стандартизация (z-нормализация)	Освен при използването на неконтролирано обучение стандартизацията е препоръчително да се използва и в случаите, когато има малко на брой екстремни стойности, както и ако разпределението на съответната характеристика е близо до нормалното. Специално внимание заслужава нормализацията, когато целта е да се справим с екстремните стойности. Технологията, с която се елиминира тяхното влияние, не е

	проблем. Но по-важно е дали имаме основание да направим това, или екстремните стойности са индикатор за специфично състояние на изследваните процеси. Отговор на този въпрос може да се даде в контекста на изследвания проблем и наличната информация.
Работа с липсващи стойности	Някои от възможните подходи при липсващи стойности бяха разгледани в Част V. Има случаи, при които дефицитите се изразяват в непълни данни – следствие от технически проблем или просто липсващо наблюдение за даден период. Липсите могат да бъдат преодолени, като: - „празнините“ бъдат заместени със средната стойност на съответния показател; - за стойност се използва средата на интервала, в който очакваме да се изменя съответният показател; - липсващите стойности се заместят нарочно със стойности, които са извън нормалния интервал за съответната характеристика. Информация за по-сложните начини за допълване на данните може да бъде намерена в [87]. Също както и при самообучаващите се алгоритми, за предпочитане е при възможност за избор да предпочетем по-простата за прилагане и обяснение възможност.

- Промяна на наличните данни, така че да се получат нови характеристики, които да бъдат използвани в следващите стъпки на анализа.

Този тип трансформация е по-интересен, тъй като включва в себе си елементи на предварителното планиране и наличното знание за проблема. В този смисъл трудно може да се говори за абсолютно валидна последователност от действия, които да следваме – най-малкото защото решенията са контекстово-зависими.

Съществуват обаче редица добри практики, които могат да бъдат следвани, когато става дума за преобразуване на отделни характеристики от входните данни. Най-често срещаните от тях са описани в Таблица 29.

Таблица 29. Добри практики при формиране на нови характеристики в данните

Добра практика	***Бележки***
Кодиране на категорийни променливи	Тази стъпка е необходима, когато избраните алгоритми изискват работа само с числени променливи, а във входните данни има и категорийни такива. В този случай препоръчително е кодировката да се осъществи с увеличаване на размерността, а не с директно номериране. *Пример:* Нека в данните да има променлива, свързана с образование и три категории – средно, висше, докторантура. Обикновено категориите за този тип показатели са повече, но за простота да предположим, че са само три. *Препоръчително кодиране:* средно – (1, 0, 0) висше – (0, 1, 0) докторантура – (0, 0, 1) Както се вижда от примера, при кодирането ние увеличаваме размерността на задачата (вместо една променлива вече имаме три съдържащи 0 или 1). Изкушаващо е вместо това да изберем кодиране от типа средно (1), висше (2), докторантура (3), но това реално води до грешни резултати. Причината е, че подобно обозначение довежда до неявно дефиниране на относителни зависимости (напр. висше е два пъти повече от средно и т.н.). А подобни отношения не отговарят на действителността.
Окрупняване на числени променливи в групи и категории	По същество идеята на тази стъпка е противоположна на предходната. Целта на групирането най-често е да се позволи

	обучението да се извърши с по-малък брой наблюдения. Аргументите, на които се базира обединяването на отделните стойности в обща категория, са, че те поотделно нямат по-висока стойност за решаване на проблема, отколкото категорията. *Пример:* Нека предположим, че изследваме ефекта „усмивка на волатилността“ при финансови кол опции и искаме да изолираме дали има други фактори освен характеристиките на самата опция, влияещи върху този ефект. *Възможно групиране:* Вместо да отчитаме разликата между двете стойности като числена променлива, можем да я поставим в три категории – по-голяма, по-малка и равна на нула. По този начин може да се установи връзка със състоянията на опцията „извън пари“, „в пари“ и „на пари“.

1.2. Добри практики при избор на аналитични инструменти

Правилното решение за избор на аналитичен инструмент има водеща роля в осигуряването на точни и адекватни резултати. Можем да очакваме, че то е плод на задълбочен анализ и сравнение между наличните възможности, но много често практиката показва, че силен ефект имат предпочитанията и предишният опит на изследователите. Естественото желание да прилагаме тези средства, които познаваме най-добре, е предпоставка за въвеждането на изкуствени ограничения и за влошаване на точността на получените резултати.

Има различни пътища, по които да стигнем до избор на подходящ аналитичен инструмент:

- Да приложим всички подходящи възможности и да изберем тази от тях, която има най-добри показатели за точност при съответния проблем.

Проблемите, свързани с подобен подход, са два – необходими са много повече време и усилия за да приложим всички подходящи алгоритми (а в определени случаи е трудно да поставим и границата къде свършват те); предварително да определим най-подходящия показател за точност съобразно с решаваните задачи.

- Да изберем аналитичен инструмент съобразно с предварително определен набор от критерии.

Този подход не ни лишава от възможността да бъдем съзидателни, но придава структурираност на процесите по вземане на решения. В Таблица 30 са представени някои от критериите, които могат да подпомогнат класирането на аналитичните инструменти, като подредбата по значимост е съобразно с вижданията на автора.

Таблица 30. Примерни критерии за избор на самообучаващи се алгоритми

Критерий	***Бележки***
Възможност за обяснение на резултатите	Този фактор е изключително важен, защото, ако няма разумно и адекватно обяснение на получените резултати, тяхната приложимост е ограничена.
Характеристики на входните данни	Основните характеристики на входните данни, които определят насоките при избор на алгоритми, са наличието на нелинейни зависимости, което би елиминирало редица възможности, както и комбинацията между числени и категорийни данни.
Размерност на входните данни	Размерността на входните данни е важен фактор, който определя бързодействието и приложимостта на отделните алгоритми.
Скорост на обучение	Скоростта на обучение може да се окаже критичен показател за отхвърляне на дадена възможност. При твърде дълъг процес на подготовка и нужда от бърза реакция е възможно да предпочетем по-неточни, но по-бързи решения.
Скорост на работа и анализ на новопостъпила информация	Дори когато обучението е еднократно и се извършва преди реалното прилагане на модела, скоростта на работа е важна характеристика на алгоритмите. За финансови приложения, като алгоритмична търговия в реално време, бързата реакция е ключов параметър.

1.3. Добри практики при оценка на хиперпараметрите на избраните алгоритми

Хиперпараметрите определят поведението и точността на използваните самообучаващи се алгоритми. Тъй като по дефиниция те са зададени предварително, самият алгоритъм не променя техните стойности, а се опитва да намери най-доброто решение при зададените външни ограничения. Това оставя отворена възможността да се експериментира с различни стойности на хиперпараметрите, докато се достигне желаното ниво на точност за изграждания модел.

В Таблица 31 са представени основните възможности за оптимизация на параметрите на използваните аналитични инструменти. Всеки от изброените подходи е свързан с нуждата от инвестиция на допълнителни ресурси и време. Размерът на допълнителните усилия може да бъде намален, ако предварително сме наясно с всички ограничения и недостатъци на избрания алгоритъм и локализираме възможните стойности на хиперпараметрите предварително. Представените в таблицата подходи са подредени според необходимите усилия за тяхното прилагане, като се започва от най-простия вариант.

Таблица 31. Стратегии за оптимизация на хиперпараметрите при самообучаващи се алгоритми

Подход за оптимизация	***Бележки***
Ограничено изчерпателно търсене (grid search)	Простотата на този подход е, че просто трябва да проверим изчерпателно възможните стойности на хиперпараметрите в предварително определените граници. Ключовите моменти в този случай са два – как да се определи критерият за сравнение (т.е. оценката на точността) и как да се ограничи броят на възможностите за проверка.
Случайно търсене	Предимството на случайното търсене е, че вместо да се сравняват всички възможни варианти, може да оценяваме само случайна извадка от възможните стойности. Ако предварително имаме възможност да определим разпределението, което следват

	изследваните хиперпараметри, тогава търсенето може да бъде значително по-ефективно.
Байес оптимизация	Идеята на работа на Байесовата оптимизация е да изгради вероятностен модел, свързващ стойностите на хиперпараметрите с качествата на получените крайни резултати. След като с последователност от опити и стъпки може да се установи как би изглеждала една такава функционална зависимост, тогава вече може да се търси с традиционните подходи и съответният екстремум.
Оптимизация с градиентно спускане	За разлика от спускането, което беше дискутирано в Част III.2.3., в този случай градиентът се изчислява спрямо оптимизираните хиперпараметри. Предимствата на подхода са свързани с по-бързото достигане до резултат, като отново специално внимание трябва да се обърне на проверката дали това не е локален екстремум.
Еволюционна оптимизация	Еволюционната оптимизация черпи идеи от генетичните алгоритми и възможността да намерим по-добър резултат, като последователно „кръстосваме“ и променяме („мутираме“) съществуващите решения.

Оптимизацията на хиперпараметрите е полезен елемент от цялостния анализ, тъй като позволява да се използват в максимална степен предимствата на самообучаващите се алгоритми. Нейното приложение обаче не отменя изискването получените резултати да имат логично и адекватно обяснение. В противен случай може да бъдем подведени от спорадични зависимости, а тогава оптимизацията на хиперпараметрите ще бъде единствено още една стъпка в „маскирането“ на същинския проблем.

За да избегнем негативните последици от последния описан сценарий, е особено важно да се извърши адекватна и безпристрастна оценка на точността на използваните алгоритми, както и резултатите да бъдат валидирани по подходящ начин.

1.4. Добри практики при оценка на точността и валидирането на алгоритмите

Оценката на точността и валидирането на получените резултати е една от последните стъпки в анализа, но има изключително важно значение за целия процес. Много често в желанието да се покаже колко добре работят самообучаващите се алгоритми, тази част се свежда до привеждането единствено до изчисляване на отделни показатели за грешка, което не е излишно, но е крайно недостатъчно. Това е причината да се обърне специално внимание на проблемите, свързани с валидирането и проверката на точността.

Един от фундаменталните въпроси, които нерядко се тълкуват превратно, е свързан с разделянето на входните данни на отделни части. Обичайна практика е те да се разпределят в две множества – обучаващо, в което попадат преобладаващата част от наблюденията, и тестово, с чиято помощ да се провери точността на получените резултати. Това би било изцяло оправдано, ако предварително са известни оптималните стойности за всички хиперпараметри на алгоритъма. Когато това условие не е изпълнено, по-добре би било, ако информацията, с която разполагаме, се оформи в три множества:

- обучаващо множество – което обхваща най-голяма част от данните (обикновено над 70 – 75%) и се използва за трениране на избрания алгоритъм;
- тестово множество – което обхваща част от останалите наблюдения и се използва за оценка на точността на получения модел;
- валидиращо множество – което обикновено включва сходен брой наблюдения както тестовото, но се използва за избор на самообучаващ се алгоритъм и за оценка на неговите хиперпараметри.

Оценката за точността и качествата на вече избраните алгоритми зависи от това какви показатели за грешка се използват. Този въпрос беше вече коментиран в Част II от изследването, като тук само ще допълним, че въпросните измерители трябва да могат да отразят успешно показаните на Фигура 48 типични проблеми.



Фигура 48. Типични проблеми при използването на самообучаващи се алгоритми. Източник: [88]

По своята същност борбата с първите два проблема, показани на Фигура 48, може да се представи като търсене на:

- подходящ баланс между прекалено прости и прекалено сложни алгоритми;
- входни данни и характеристики, които позволяват да се улови същността на изследвания проблем и да се извлече достатъчно информация за неговите особености.

В това търсене активно могат да се прилагат различни техники, като например L1 (по-познато като Lasso метод) и L2 (също известна като Ridge регресия) регуляризация ([89], [90]). Посредством избора на различна форма на „наказание“, свързано с увеличаване на размерността на модела, на практика може да се подпомогне етапът по избор на характеристики (вкл. създаване на нови такива), които да участват в окончателния модел.

Показателите за оценка на грешките вече бяха коментирали в Част III, като тук ще се ограничим само да систематизираме възможните варианти и връзката им с типа на решаваните задачи. Специално внимание трябва да се обърне на контекстово зависимите оценки на загубите.

В анализа на финансови инструменти тези показатели имат най-близък до практиката смисъл, защото показват как отделните самообучаващи се алгоритми се справят с подпомагането на инвестиционните стратегии. За да се постигне съпоставимост при отделните изследвания, е необходимо освен контекстово зависимите да бъдат прилагани и някои от стандартните измерители. Именно такъв подход беше използван и в проведените експерименти с контролирано обучение в Част III.

Таблица 32. Показатели за оценка на грешки и точност на самообучаващите се алгоритми

Класификационни проблеми	***Регресионен анализ, динамични редове***
Матрица на грешките (Confusion matrix)	Абсолютни показатели за грешка: ME, MSE, MAE, RMSE
Дял на правилно класифицираните случаи	Относителни показатели за грешка: MPE, MAPE, MASE
AUC и площ под ROC кривата	Контекстово зависима оценка на загубите
Контекстово зависима оценка на загубите	

2. Когато един инструмент не е достатъчен

Съвременните пазарни отношения са сложни и се променят бързо. Навлизането на нови технологии и финансови инструменти създава възможности, които не са присъствали в предходните периоди. Това променя условията на работа на участниците на пазара, а оттам и изискванията към използваните аналитични инструменти. Комбинацията от множество фактори и взаимната обвързаност на финансовите пазари водят до необходимост от по-сложни аналитични инструменти. Възможна е ситуация, при която никой от познатите алгоритми да не е в състояние да предостави задоволително решение на поставените задачи.

Една възможност да излезем от затруднението, е да използваме комбинация от алгоритми, разчитайки, че съвместната им работа ще преодолее индивидуалните недостатъци и ще ни позволи да постигнем по-висока точност, отколкото ако използваме участващите инструменти поотделно. Възможностите за реализация на подобни комбинации са разгледани подробно в [91] и [92]. При анализа на финансови инструменти от дискутираните възможности особено отношение към практиката имат следните два подхода:

- съвместяването на различни алгоритми и избор на един от тях в зависимост от характеристиките на решавания проблем (т. нар. “bucket of models”);

- прилагане на множество от алгоритми, които се обучават с наличните данни, а впоследствие се използва друг инструмент, за да комбинира техните прогнозни стойности в окончателен резултат – т.нар. “stacking” подход.

Практическата приложимост на тези подходи е важна, защото може да се съпостави с различните етапи от вземане на решения по планиране, изпълнение и приключване на различните инвестиционни стратегии. Използването на комбинация от алгоритми не е „безплатно“ и това най-често се изразява в:

- увеличени изисквания откъм ресурсите, необходими за провеждане на изследването;
- нарастване на сложността на получения краен модел, което прави по-трудно неговото аргументиране и обяснение.

Повишената сложност на комбинираните алгоритми действа като специфичен ограничител. Това не е задължително свързано с негативни последици, защото не позволява създаването на комплексни системи, които да се възприемат просто като „черни кутии“.

VII. Вместо заключение

В това изследване бяха разгледани възможностите, предимствата и недостатъците на редица самообучаващи се алгоритми за анализ на финансови инструменти. Огромното разнообразие както на аналитични инструменти, така и на инвестиционни възможности не позволява да бъде подробно анализирана всяка една ситуация, в която интелигентната обработка на данни би донесла ползи. Направените експерименти и изследвания на различните типове обучение могат да очертаят благоприятните за използване на самообучаващи се алгоритми условия:

- наличие на голям обем от данни, които описват сложни икономически и финансови процеси и отразяват влиянието на разнообразни по характер фактори;
- нуждата от бърза и адекватна реакция на нова информация, която има директно отношение към успеха на инвестиционните решения;
- необходимост от проследяване на хетерогенни финансови инструменти и процеси, което предполага гъвкавост и адаптивност на изследователските средства.

Списъкът с условия може да бъде разширен и допълнен, но дори и в този си вид е свидетелство да широкото приложение, което самообучаващите се алгоритми имат. В рамките на изследването бяха проведени експерименти, които демонстрират успешното използване на всеки един от основните три типа обучение за решаване на практически значими финансови проблеми.

Успехите и предимствата на самообучаващите се алгоритми не бива да бъдат приемани безрезервно. Има поне две важни правила, които не трябва да бъдат забравяни:

- Успешното приложение на алгоритмите стъпва върху адекватен и задълбочен предварителен анализ и подготовка. Можем да твърдим, че качеството на крайните резултати зависи колкото от избрания аналитичен инструмент, толкова и от предварително свършената работа.
- При идентични резултати и близка точност за предпочитане е да не се усложнява излишно анализът. Спазването на това правило се отплаща, като спестява усилия и ресурси за аргументиране и обясняване на получените резултати.

В заключение можем да кажем, че самообучаващите се алгоритми ни предоставят набор от мощни и полезни аналитични инструменти. Но какво точно ще построим с тяхна помощ, както и колко устойчиво ще бъде то на теста на времето, зависи преди всичко от нас самите.

Благодарност

Монографията е част от работата по проект ФП19-ФИСН-014, финансиран от Фонд „Научни изследвания“ при Пловдивския университет „Паисий Хилендарски“.

Списък на фигурите

Фигура 1. Смесено обучение – примерен ефект от приложението му 24
Фигура 2. Графично представяне на $1 - e - t\gamma t\gamma - e - t\gamma$ и γ=2 (вляво), и γ=4 (дясно), ... 33
Фигура 3. Исторически промени в доходността на ДЦК на САЩ 34
Фигура 4. Промени в параметрите на NSS за периода 1981 - 2020 г. $\beta 0$ (първи ред, ляво), $\beta 1$ (първи ред, дясно), $\beta 2$ (втори ред, ляво), $\beta 3$ (втори ред, дясно), хистограма на $\gamma 1$ (трети ред, ляво) и на $\gamma 2$ (трети ред, дясно). 36
Фигура 5. Обем на новите емисии корпоративни облигации и акции в САЩ в млрд. USD ([37]) 38
Фигура 6. Исторически данни за Moody's yield on seasoned corporate bonds (Рейтинг: BAA) 39
Фигура 7. Автокорелация (ляво) и частична автокорелация (дясно) на първите разлики в доходността на облигациите с рейтинг BAA 40
Фигура 8. Невронна мрежа, използвана за MLP модела 42
Фигура 9. Исторически и прогнозни данни за MLP модела 43
Фигура 10. Невронна мрежа, използвана за ELM модела 44
Фигура 11. Принцип на работа на отделен елемент в невронната мрежа 45
Фигура 12. Отразяване на грешката при повече от един неврон 46
Фигура 13. Локален и глобален минимум за $\mathrm{E} = \mathrm{e}(v - f(w1x1+\ldots+wkxk))$ 47
Фигура 14. Топология на ELM мрежа (едно скрито ниво) 48
Фигура 15. ROC крива на logit модела (вляво) и на LDA (вдясно) 57
Фигура 16. Топология на използваната в анализа невронна мрежа 58
Фигура 17. Случаи в четирите възможни категории и различни теглови коефициенти 60
Фигура 18. Прогнозираните спадове в БВП (в проценти) за вредните сценарии при стрес тестовете в периода 2011–2020 г. Източник: EBA 68
Фигура 19. Оценка на ключови икономически показатели при сценариите за стрес тестове през 2018 и 2020 г. Източници: EBA, ECB 69
Фигура 20. Разлики в клъстеризацията между K-Means (вляво) и K-Center (вдясно) 74
Фигура 21. Проверка за оптимален брой клъстери на симулираните данни от Фигура 20 75
Фигура 22. Примерна самоорганизираща се мрежа 78

Фигура 23. Проста демонстрация на PCA анализа 79
Фигура 24. Разпределение на дневните възвръщаемости от примерната съвкупност 82
Фигура 25. PCA анализ върху възвръщаемостта на портфейл от пет акции 82
Фигура 26. Оптимален брой клъстери спрямо края на 2017 г. на базата на текущото състояние 85
Фигура 27. Клъстеризация по данни към края на 2017 г. при три и четири области 86
Фигура 28. Резултати от клъстеризацията с помощта на самоорганизиращи се мрежи 88
Фигура 29. Резултати от самоорганизиращата се мрежа при размерност 2x2 и 6x6 88
Фигура 30. Оптимален брой клъстери при прогноза за края на 2020 г., базов сценарий 90
Фигура 31. Клъстеризация по данни към края на 2020 г., базов сценарий при три и четири области 91
Фигура 32. Клъстеризация по данни към края на 2020 г., базов сценарий при пет и шест области 91
Фигура 33. Оптимален брой клъстери при прогноза за края на 2020 г., негативен сценарий 92
Фигура 34. Клъстеризация по данни към края на 2020 г., негативен сценарий при три и четири области 93
Фигура 35. Частичен списък на рисковете, стоящи пред една кредитна институция 95
Фигура 36. Възвръщаемост на различните ДЦК и корелация в тяхната доходност за изследвания период 99
Фигура 37. Промени в коефициента на корелация между доходността на ЦК с най-малък и най-голям матуритет спрямо период от 3 месеца 100
Фигура 38. PCA анализ, приложен към държавните ценни книжа 101
Фигура 39. LIBOR и корелация между стойностите за различен матуритет 102
Фигура 40. Промени в коефициента на корелация между овърнайт и 12M LIBOR спрямо период от 3 месеца 103
Фигура 41. Резултати от анализа на главните компоненти, приложен към LIBOR 103
Фигура 42. Връзка между разгледаните до момента типове обучение 108

Фигура 43. Групиране с модел, използващ разделяне на базата на плътност 111
Фигура 44. Пример за граф базиран алгоритъм и класификация 1-0 112
Фигура 45. Q-Q диаграма на възвръщаемостите на дневна (ляво) и седмична (дясно) база за S&P500 (горе) и DAX (долу) 117
Фигура 46. Дневна (ляво) и седмична (дясно) възвръщаемост за DAX, представени като микс от три нормални разпределения 118
Фигура 47. Дневна (ляво) и седмична (дясно) възвръщаемост за DAX, представени като микс от три нормални разпределения 118
Фигура 48. Типични проблеми при използването на самообучаващи се алгоритми. Източник: [88] 137

Списък на таблиците

Таблица 1. Важни понятия, използвани в изследването 16
Таблица 2. Алгоритми, използващи контролирано обучение 17
Таблица 3. Алгоритми, използващи неконтролирано обучение 21
Таблица 4. Алгоритми, използващи смесено обучение 25
Таблица 5. Допълнителни ограничения, наложени от инвеститора 29
Таблица 6. Промени в параметрите на NSS модела 35
Таблица 7. Характеристика на оптималния ARIMA модел 40
Таблица 8. Характеристики на оптималния MLP модел 43
Таблица 9. Характеристики на разглеждания ELM модел 44
Таблица 10. Бързодействие на сравняваните алгоритми 51
Таблица 11. Показатели за грешка на сравняваните алгоритми 52
Таблица 12. Характеристики на данните за фалирали компании 55
Таблица 13. Характеристики на сравняваните подходи 56
Таблица 14. Изходен код и библиотеки използвани в Част III 61
Таблица 15. Основни разлики между контролирано и неконтролирано обучение 64
Таблица 16. Алгоритми за класификация с неконтролирано обучение 71
Таблица 17. Критерии за минимизиране на разстоянието и варианти на K-Means 73
Таблица 18. Характеристика на основните вектори от примера с петте акции 83
Таблица 19. Промени в БВП, отчитащи пандемията от COVID-19 ([69]) . 92
Таблица 20. Характеристики на входните данни, използвани за PCA анализа 99
Таблица 21. PCA анализ на доходността на американски ДЦК 101
Таблица 22. Резултати от PCA анализа на LIBOR 102
Таблица 23. Изходен код и библиотеки, използвани в Част IV 106
Таблица 24. Характеристики на изследваните динамични редове 116
Таблица 25. Резултати от EM клъстеризацията за S&P 500 121
Таблица 26. Резултати от EM клъстеризацията за DAX 122
Таблица 27. Изходен код и библиотеки, използвани в Част V 126
Таблица 28. Добри практики при трансформация на наличните характеристики 129
Таблица 29. Добри практики при формиране на нови характеристики в данните 131

Таблица 30. Примерни критерии за избор на самообучаващи се алгоритми 133
Таблица 31. Стратегии за оптимизация на хиперпараметрите при самообучаващи се алгоритми..... 134
Таблица 32. Показатели за оценка на грешки и точност на самообучаващите се алгоритми 138

БИБЛИОГРАФИЯ

[1] D. Dunson, "Statistics in the big data era: Failures of the machine," *Statistics & Probability Letters,* vol. 136, pp. 4-9, 2018.

[2] C. Tong, "Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science," *The American Statistician,* vol. 73, no. 1, pp. 246-261, 2019.

[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, "Concrete Problems in AI Safety," *arXiv:1606.06565,* vol. July 2019, pp. 1-29, 2016.

[4] D. Dannemiller and S. Collins, "2020 investment management outlook," in *Deloitte Insights,* New York, Deloitte Development LLC, 2020, p. 32.

[5] M. Driscoll, From Psychology of Learning for Instruction, London: Pearson, 2005, p. 504.

[6] R. Mello and M. Ponti, Machine Learning - A Practical Approach on the Statistical Learning Theory, Cham: Springer, 2018, p. 377.

[7] И. Иванов и В. Танов, Алгоритми за анализ на големи данни и приложения (Machine Learning), София: Авангард Прима, 2019.

[8] С. Кабаиванов, Иконометрия за финансисти, София: Евдемония продъкшън, 2014, p. 138.

[9] А. Ефремов, Идентификация на многомерни системи • Линеен подход за оценяване, София: ДАР-РХ, 2014, p. 342.

[10] E. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance,* vol. 23, no. 4, pp. 589-609, 1968.

[11] V. Vapnik, The Nature of Statistical Learning Theory, vol. 2en, New York: Springer, 2000, p. 185.

[12] G. Mure, Posterior Analytics, e-Book: The University of Adelaide, 2007.

[13] Y. Peng, P. H. Melo Albuquerque, I. F. do Nascimento and J. V. Freitas Machado, "Between Nonlinearities, Complexity, and Noises: An Application on Portfolio Selection Using KernelPrincipal Component Analysis," *Entropy,* vol. 21, no. 376, pp. 1-27, 2019.

[14] В. Марковска, Управление на инвестиционни портфейли от реални активи, София: Евдемония продъкшън, 2019, p. 172.

[15] S. Husmann, A. Shivarova and R. Steinert, "Company classification using machine learning," *arXiv:2004.01496,* p. 16, 2020.

[16] B. Salvador, C. Oosterlee and R. van der Meer, "Financial option valuation by unsupervised learning with artificial neural networks," *arXiv preprint arXiv:2005.12059,* p. 23, 2020.

[17] V. Konsulov and S. Kabaivanov, "Phenomenon of fraud? Lessons learned from turmoil in Bulgarian banks in 2014," *VSIM Journal,* vol. 1, no. 9, pp. 8-16, 2015.

[18] P. Atanasov and S. Kabaivanov, "Modelling the innovative potential of companies in the pharmaceutical industry in Bulgaria," *AIP Conference Proceedings,* vol. 2048, no. 1, pp. 1-10, 2018.

[19] L. Carson, R. Thulasiram and D. Bondarenko, "An Efficient System for Detecting Outliers from Financial Time Series," in *BNCOD 2006: Flexible and Efficient Information Handling,* Belfast, 2006.

[20] В. Марковска, (Не)възможните проекти, София: Евдемония Продъкшън, 2018, p. 150.

[21] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B.,* vol. 39, no. 1, pp. 1-38, 1977.

[22] P. Dayan and Y. Niv, "Reinforcement learning: The Good, The Bad and The Ugly," *Current Opinion in Neurobiology,* vol. 18, no. 2, pp. 185-196, 2008.

[23] М. Русева, Системата на национални сметки и макроикономическите статистически изследвания, Пловдив: Университетско издателство „Паисий Хилендарски", 2015, pp. 1-174.

[24] A. Zahariev, M. Zveryakov, S. Prodanov, G. Zaharieva, P. Angelov, S. Zarkova and M. Petrova, "Debt management evaluation through support vector machines: On the example of Italy and Greece," *Entrepreneurship and Sustainability Issues,* vol. 7, no. 3, pp. 2382-2393, 2020.

[25] F. Fabozzi (Editor), Encyclopedia of Financial Models, 3 Volume Set, New York: Wiley, 2012, p. 2100.

[26] C. Sammut and G. Webb, Encyclopedia of Machine Learning and Data Mining, New York: Springer US, 2017, p. 1335.

[27] L. Rittenberg and T. Tregarthen, Principles of Microeconomics, Nyack, NY: Flat World Knowledge, L.L.C., 2009, p. 540.

[28] I. Popchev and I. Radeva, "Risk Analysis – an Instrument for Technology Selection.," *Engineering Sciences,* vol. 4, no. 1, pp. 5-20, 2019.

[29] I. Popchev, D. Orozova and S. Stoyanov, "IoT and Big Data Analytics In E-Learning," in *Big Data, Knowledge and Control*, Sofia, 2019.

[30] F. Li, T.-M. Chow, A. Pickard and Y. Gar, "Transaction Costs of Factor-Investing Strategies," *Financial Analysts Journal,* vol. 75, no. 2, pp. 62-78, 2019.

[31] Y. Stander, "Yield Curve Models," in *Yield Curve Modeling. Finance and Capital Markets Series*, London, Palgrave Macmillan, 2005, pp. 33-52.

[32] C. Nelson and A. Siegel, "Parsimonious Modeling of Yield Curves," *Journal of Business*, vol. 60, no. 4, p. 473–489, 1987.

[33] L. Svensson, "Estimating and Interpreting Forward Interest Rates: Sweden1992–1994," *IMF Working paper,* vol. 1994, no. 114, p. 76, 1994.

[34] I. Ivanov, S. Kabaivanov and B. Bogdanova, "Stock market recovery from the 2008 financial crisis: The differences across Europe," *Research in International Business and Finance*, vol. 37, pp. 360-374, 2016.

[35] R. Brent, Algorithms for Minimization without Derivatives, Englewood Cliffs, New Jersey: Prentice-Hall, 1973, p. 206.

[36] FINRA, "TRACE: The Source for Real-Time Bond Market Transaction Data," FINRA, Washington DC, 2017.

[37] SIFMA, "Capital Markets Fact Book," SIFMA, New York, 2019.

[38] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems,* vol. 2, no. 4, p. 303–314, 1989.

[39] G.-B. Huang, "What are Extreme Learning Machines? Filling the Gap Between Frank Rosenblatt's Dream and John von Neumann's Puzzle," *Cognitive Computation*, vol. 7, no. 13, p. 263–278, 2015.

[40] H. Kelley, "Gradient theory of optimal flight paths," *ARS Journal*, vol. 30, no. 10, p. 947–954, 1960.

[41] D. Rumelhart, G. Hinton and R. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, p. 533–536, 1968.

[42] R. Penrose, "A Generalized Inverse for Matrices," *Proc. Cambridge Phil. Soc.,* vol. 51, no. 1, pp. 406-413, 1955.

[43] H. Lütkepohl, "Vector Autoregressive Moving Average Processes," in *New Introduction to Multiple Time Series Analysis*, Berlin, Heidelberg, Springer, 2005, pp. 419-446.

[44] D. Ackerer and D. Filipovic, "Linear credit risk models," *Finance and statistics*, vol. 24, p. 169–214, 2020.

[45] J. Fons, "Using Default Rates to Model the Term Structure of Credit Risk," *Financial Analysts Journal*, vol. 50, no. 5, pp. 25-32, 1994.

[46] B. Barnard, "Bond-Valuation and Default-Risk Re-examined: A Post Reduced-Form Model Based on Integer Migration and Portfolio Migration," *Available at SSRN: https://ssrn.com/abstract=3427127*, pp. 1-57, 2019.

[47] B. Vrigazova and I. Ivanov, "Tenfold bootstrap procedure for support vector machines," *Computer Science,* vol. 21, no. 2, pp. 241-257, 2020.

[48] Y.-T. Huang and P.-F. Pai, "Using the Least Squares Support Vector Regression to Forecast Movie Sales with Data from Twitter and Movie Databases," *Symmetry,* vol. 12, no. 4, p. 625, 2020.

[49] R. Bellman, Dynamic programming, New Jersey: Princeton University Press, 1957, p. 365.

[50] K. Dowd, "Central bank stress tests: Mad, bad, and dangerous," *Cato J.,* vol. 35, p. 507, 2015.

[51] T. Schneider, P. Strahan and J. Yang, "Bank Stress Testing: Public Interest or Regulatory Capture?," National Bureau of Economic Research, Cambridge, MA, 2020.

[52] M. Gerhardt and R. Vander Vennet, "European bank stress test and sovereign exposures," *Applied Economics Letters,* vol. 24, no. 17, pp. 972-976, 2017.

[53] N. Taleb, E. Canetti, T. Kinda, E. Loukoianova and C. Schmieder, "A New Heuristic Measure of Fragility and Tail Risks: Application to Stress Testing," IMF Working paper, 2012.

[54] I. Ignatov, "Automatic Budget Stabilization in the EU during 2007-2017," in *International Conference "Economic challenges: development, welfare, integration*, UNWE, Sofia, 2019.

[55] I. Ignatov, "The Choice of an Approach to Fiscal Consolidation," *Annual of Sofia University "St. Kliment Ohridski",* vol. 18, no. 1, pp. 131-144, 2020.

[56] С. Вачков, „Дигитални рефлексии върху банковата организация,“ в *Годишник на ИУ - Варна*, Варна: 88, Наука и икономика, 2018, pp. 129-230.

[57] Б. Манов, Основи на финансите - Пари и банково дело, София: Нов живот, 2012, p. 172.

[58] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks,* vol. 16, no. 3, pp. 645-678, 2005.

[59] K.-L. Wu and M.-S. Yang, "Alternative c-means clustering algorithms," *Pattern recognition,* vol. 35, no. 10, pp. 2267-2278, 2002.

[60] X. Li and Z. Fang, "Parallel clustering algorithms," *Parallel Computing,* vol. 11, no. 3, pp. 275-290, 1989.

[61] E. Cortes, "Plant Disease Classification Using Convolutional Networks and GenerativeAdverserial Networks," Stanford University, 2017.

[62] R. Thorndike, "Who Belongs in the Family?," *Psychometrika,* vol. 18, no. 4, pp. 267-276, 1953.

[63] R. Tibshirani, G. Walther and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Statist. Soc. B,* vol. 63, no. 2, pp. 411-423, 2000.

[64] P. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Computational and Applied Mathematics,* vol. 20, p. 53–65, 1987.

[65] S. Kabaivanov, K. Roberts and S. Kovacheva, "Machine learning assisted social system analysis: Youth transitions in five South and East Mediterranean countries," in *AMEE'20,* Sozopol, 2020.

[66] T. Kohonen, "Self-Organized Formation of Topologically Correct Feature Maps," *Biological Cybernetics,* vol. 43, pp. 59-69, 1982.

[67] R. Rosipal, M. Girolami, L. Trejo and A. Cichocki, "Kernel PCA for feature extraction and de-noising in nonlinear regression," *Neural Computing & Applications,* vol. 10, no. 3, pp. 231-243, 2001.

[68] J. Tian, M. Azarian and M. Pecht, "Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm," in *European conference of the prognostics and health management society*, Nantes, 2014.

[69] European Central Bank, "Euro area banking sector resilient to stress caused by coronavirus, ECB analysis shows," 28 07 2020. [Online]. Available: https://www.bankingsupervision.europa.eu/press/pr/date/2020/html/ssm.pr200728~7df9502348.en.html. [Accessed 29 07 2020].

[70] R. Aumann and L. Shapley, Values of Non-Atomic Games, New Jersey: Princeton: Princeton Univ. Press, 1974, p. 323.

[71] J.-F. Mertens, "Values and Derivatives," *Mathematics of Operations Research*, vol. 5, no. 4, p. 523–552, 1980.

[72] J. Fan and W.-X. Zhou, "Guarding against Spurious Discoveries in High Dimensions," *Journal of Machine Learning Research,* vol. 17, no. 1, pp. 1-34, 2016.

[73] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11, no. 3, p. 363–371, 1965.

[74] S. Fralick, "Learning to recognize patterns without a teacher," *IEEE Transactions on Information Theory*, vol. 13, no. 1, p. 57–64, 1967.

[75] O. Chapelle, B. Schölkopf and A. Zien, Semi-Supervised Learning, Cambridge, Massachusetts: The MIT Press, 2006, p. 524.

[76] K. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information processing systems*, The MIT Press, 1999, pp. 368-374.

[77] P. Chunhachinda, K. Dandapani, S. Hamid and A. Prakash, "Portfolio selection and skewness: Evidence from international stock markets," *Journal of Banking & Finance,* vol. 21, no. 2, pp. 143-167, 1997.

[78] F. Aparicio and J. Estrada, "Empirical distributions of stock returns: European securities markets, 1990-95," *The European Journal of Finance*, vol. 7, no. 1, pp. 1-21, 2001.

[79] I. Antoniou, V. Ivanov, V. Ivanov and V. Zrelov, "On the log-normal distribution of stock market data," *Physica A: Statistical Mechanics and its Applications,* vol. 331, no. 3-4, pp. 617-638, 2004.

[80] R. Harris and C. Küçüközmen, "The Empirical Distribution of UK and US Stock Returns," *Journal of Business Finance and Accounting,* vol. 28, no. 5-6, pp. 715-740, 2001.

[81] A. Kumar and P. A. Raja, "Modeling commodity market returns: the challenge of leptokurtic distributions," in *Advances in Analytics and Applications*, Singapore, Springer, 2019, pp. 203-224.

[82] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *ieee assp magazine,* vol. 3, no. 1, pp. 4-16, 1986.

[83] S. Eddy, "What is a hidden Markov model?," *Nature biotechnology,* vol. 22, no. 10, pp. 1315-1316, 2004.

[84] E. Fama and K. French, "Long-Horizon Returns," Fama-Miller Center for Research in Finance, Chicago Booth Paper No. 17-17, 2017.

[85] W. Quinn and J. Turner, Boom and Bust: A Global History of Financial Bubbles, Cambridge: Cambridge University Press, 2020, p. 296.

[86] L. Baiardi, M. Costabile, D. De Giovanni, F. Lamantia, A. Leccadito, I. Massabó and A. Staino, "The Dynamics of the S&P 500 under a Crisis Context: Insights from a Three-Regime Switching Model," *Risks,* vol. 8, no. 3, p. 71, 2020.

[87] B. Efron, "Missing data, imputation, and the bootstrap," *Journal of the American Statistical Association,* vol. 89, no. 426, pp. 463-475, 1994.

[88] A. Burkov, The Hundred-Page Machine Learning Book, Amazon, 2019, p. 160.

[89] C. Cortes, M. Mohri and R. Afshin, "L2 regularization for learning kernels," *arXiv preprint 1205.2653,* pp. 1-5, 2012.

[90] J. Kolluri, V. Kotte, M. Phridviraj and S. Shaik Razia, "Reducing Overfitting Problem in Machine Learning Using Novel L1/4

Regularization Method," in *4th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2020.

[91] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, Springer, Berlin, Heidelberg,, 2000.

[92] C. Zhang and Y. Ma, Ensemble machine learning: methods and applications, Berlin, Heidelberg: Springer Science & Business Media, 2012, p. 332.

[93] E. Fama and K. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 3, pp. 3-56, 1993.

[94] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, p. 129–137, 1982.

[95] А. Велева и М. Русева, *Съвкупностният подход в статистиката,* Пловдив: УИ - ПУ "Паисий Хилендарски", 2016, pp. 1-380.