**Федеральное государственное автономное образовательное учреждение высшего**

**образования**

**"Национальный исследовательский университет "Высшая школа экономики"**

**"Московский институт электроники и математики им А. Н. Тихонова"**

**Отчет**

**по проектной работе**

Изучение временных рядов рейтингов компаний, занимающихся экономической деятельностью

**Выполнено: студентом гр. МСМТ-243** Липатов Данила Вячеславович

**Руководитель проекта:**

Зотов Леонид Валентинович

**Консультант проекта:**

Игнатовская Валерия Анатольевна

Москва - 2025

**Содержание**

[1 Техническое задание 3](#_Toc1803)

[1.1 Описание задачи 3](#_Toc4366)

[1.2 Цели, задачи 5](#_Toc14545)

[1.3 Планируемый и фактический результат 6](#_Toc23633)

[1.4 Ход работ, роли участников команды 6](#_Toc7399)

[2 Анализ предметной области и обзор литературы 9](#_Toc10127)

[3 Подробное описание подходов, методов и математиче- ских моделей, использованных при реализации про- екта 10](#_Toc31052)

[3.1 Марковская модель 10](#_Toc27174)

[3.1.1 Однородные марковские цепи 11](#_Toc29655)

[3.1.2 Поведение кредитный рейтингов как марковская цепь 12](#_Toc22202)

[3.1.3 Применение априорного распределения Дирихле 15](#_Toc12315)

[3.1.4 Обучение марковской модели 16](#_Toc22540)

[3.2 Модель классификации 16](#_Toc21864)

[3.2.1 Случайный лес (Random Forest) 19](#_Toc2424)

[3.2.2 Прогноз кредитного рейтинга как задача классификации 20](#_Toc1487)

[3.3 Спектральные характеристики 20](#_Toc12385)

[3.3.1 Параметр Херста 21](#_Toc9755)

[3.3.2 Матрица кросс-корреляции 22](#_Toc6599)

[3.4 Марковская цепь Монте-Карло (MCMC) 22](#_Toc24086)

[4 Полученные результаты 23](#_Toc6817)

[4.1 Данные 23](#_Toc21069)

[4.2 Результаты baseline-алгоритмов 26](#_Toc12363)

[4.3 Результаты марковской модели 27](#_Toc6834)

[4.4 Результаты модели классификации 31](#_Toc13126)

[4.5 Сравнение качества прогнозов 31](#_Toc28698)

[4.6 Применение МСМС 34](#_Toc15391)

[4.7 Расчёт спектральных характеристик и матрицы кросс-корреляции 38](#_Toc19968)

[5 Веб-интерфейс (Streamlit) 42](#_Toc32323)

[6 Дальнейшие шаги 43](#_Toc9402)

[7 Заключение 44](#_Toc4294)

[8 Приложение 44](#_Toc24253)

[Список литературы 44](#_Toc27955)

# Техническое задание

Каждому рейтинговому агентству и финансовому институту необходимо отслеживать динамику кредитных рейтингов компаний, чтобы своевременно оценивать кредитные риски, прогнозировать вероятность дефолта и принимать обоснованные решения по управлению портфелем. Если рассматривать корпоративные и государственные облигации, то инвестор рискует понести убытки в случае ухудшения рейтинга или дефолта эмитента. Поэтому важно понимать, как и почему меняются рейтинги во времени. Каждая компания характеризуется своим кредитным профилем, который зависит от множества факторов: финансовых показателей, макроэкономической ситуации, отраслевых трендов, регуляторных изменений и прочего. С течением времени состояние компании может улучшаться или ухудшаться, и, соответственно, меняться её рейтинг. Для адекватной оценки инвестиционного риска и построения устойчивых стратегий важно уметь анализировать вероятности переходов между уровнями рейтингов, выявлять закономерности и учитывать влияние внешних и внутренних факторов.

## Описание задачи

В общей сложности КРА публикуют конкретные кредитные рейтинги с подуровнями, которые соответствуют прогнозу кредитного рейтинга «+» - повышение, «-» понижение, без доп. знака - стабильность. Но в открытых источниках Банка России есть условная группировка таких рейтингов на 6 состояний (без учёта дефолта). Рассмотрим детализированные состояния и укрупненные подробнее:

|  |  |
| --- | --- |
| AAA-AA | Наивысший рейтинг. Минимальный кредитный риск. Финансовые обязательства практически безрисковы. |
| AA-A | Высокая способность выполнять долговые обязательства, но чувствительность к ухудшению макроэкономической ситуации выше. |
| A-BBB | Приемлемая кредитоспособность. Обязательства выполняются, но чувствительность к изменениям в экономике и финансовых условиях значительно выше. |
| BB | Финансовые обязательства могут быть выполнены, но существуют значительные риски. |
| B | Финансовые обязательства исполняются, но кредитоспособность крайне чувствительна к ухудшению условий. |
| CCC-С  D | Финансовые обязательства находятся под угрозой невыполнения. Компании находятся в крайне тяжёлом положении.  Не выполнение долговых обязательств (дефолт). Это может быть технический или окончательный дефолт. |
|  |  |

и детализированные состояния.

|  |  |
| --- | --- |
| AAA | Кредитоспособность на исключительном уровне. Фактически отсутствие кредитного риска |
| AA+ | Очень высокая надёжность, чуть ниже AAA |
| AA | Высокая устойчивость к рискам |
| AA- | Устойчивый, но чувствительный к изменениям в экономике |
| A+ | Надёжный, но выше чувствительность к рыночным условиям |
| A | Высокая способность к обслуживанию долга |
| A- | Всё ещё инвестиционный уровень, но ближе к спекулятивному |
| BBB+ | Приемлемая кредитоспособность, умеренный риск |
| BBB | Пограничный инвестиционный уровень |
| BBB- | Последняя ступень перед спекулятивным уровнем |
| BB+ | Спекулятивный уровень, умеренный риск дефолта |
| BB | Повышенный риск, но возможна устойчивость |
| BB- | Высокий риск при неблагоприятных условиях |
| B+ | Значительный кредитный риск, но обязательства исполняются |
| B | Высокий риск, возможны просрочки |
| B- | Крайне чувствительная ситуация, дефолт вероятен |
| ССС | Высокая вероятность дефолта, серьёзные проблемы с выплатами |
| СС | Дефолт практически неизбежен |
| С | Компания на грани неплатёжеспособности |
| D | Не выполнение долговых обязательств (дефолт). Это может быть технический или окончательный дефолт |

Кредитные рейтинги присваиваются неравномерно, то есть за полгода может быть дважды присвоен, а может и не присвоен вовсе в течение одного года. Важно понимать, что кредитные рейтинги действуют в течение 1 - 2 календарного года ( с условием того, что компания может отказаться от кредитного рейтинга, а потом снова иметь его). Тем не менее, жизнь клиента можно записать в виде так называемой «цепочки»:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Номер месяца жизни клиента | 01.01.2017 | 25.04.2017 | 25.04.2018 | 20.05.2019 | 01.01.2020 |
| Состояние | AA- | AA- | A+ | A- | BBB+ |

Цепочка заканчивается в случае, если клиент попал в состояние D (Default), а также NR (Not Rated) если отказывается от обслуживание (такие наблюдения не несут стат. характера и пропускают в прогнозировании) , или ушёл в дефолт.

Для подсчёта ожидаемого кредитного риска, ожидаемых прибыли и убытка по кредитным продуктам, кредитно рейтинговым агентствам необходимо предсказывать, в каком состоянии (кредитном рейтинге) будет находиться каждый клиент в будущем. Наша задача заключается в прогнозировании состояний клиентов на некотный период вперёд. Например, зная цепочку на конец 3-го месяца

## Цели, задачи

Целью проекта является разработка алгоритма прогнозирования типа кредитного рейтинга компании на несколько будущих периодов

Для достижения цели ставятся следующие задачи:

1. обзор и анализ источников по теме исследования, определение путей решения задач исследования
2. изучение возможных состояний кредитных рейтингов и переходов между ними
3. изучение и применение марковской модели с дискретным и непрерывным временем, а также второго порядка
4. изучение и применение в рамках методов классификации с помощью моделирования в Python на доступных данных
5. проведение тестирования, оценка качества моделей
6. вычисление спектральных характеристик
7. сравнение результатов

## Планируемый и фактический результат

**Планируемый результат.** Методология (алгоритм) для прогнозирования кредитного рейтинга на несколько периодов, а именно: произведён ресемплинг данных, произведена группировка компаний по агентствам и типам компаний, для каждой группы компаний выбрана лучшая модель, посчитаны спектральные характеристики и матрица кросс-корреляции, улучшена модель классификации на основе дополнительных переменных

**Фактический результат.** Проработаны почти все пункты проекта, за исключением обновлённой модели классификации. Так же была произведена кластеризация компаний и спрогнозированы кредитные рейтинги для этих кластеров. Дополнительно был улучшен веб интерфейс для прямого взаимодействия с расчётами, без необходимости изменения кода программы. Ведётся работа с извлечением дополнительных данных, необходимых для модели классификации.

## Ход работ, роли участников команды

Ход работы представлен в виде Таблицы [1](#_bookmark6).

Таблица 1: План проекта

|  |  |  |  |
| --- | --- | --- | --- |
| № | Наименование задач | Сроки  выполнения | Примечания |
| 1 | Выбор темы проекта, формулировка цели и  задач | 01.11.2024-  30.11.2024 |  |
| 2 | Анализ литературных источников, реализация парсера | 01.12.2024-  10.12.2024 |  |
| 3 | Сбор описательных статистик, построение начальных моделей - марковских цепей 1-ого порядка, построение доверительных интервалов методом bootstrap | 31.12.2024-  01.02.2025 |  |
| 4 | Представление проекта | 07.02.2025 |  |
| 5 | Реализация baseline-алгоритмов:   1. Прогноз на основе последнего известного состояния   Выбор метрик для оценки моделей | 01.02.2025-  15.02.2025 |  |
| 6 | Реализация алгоритмов прогнозирования:   1. Алгоритм на основе однородной марковской цепи 1-го порядка   в непрерывном времени и с априорным распределением   1. Спектральные характеристики 2. Марковская цепь 2-ого порядка | 05.02.2025-  31.03.2025 |  |
| 7 | Подготовка постера, уточнение данных до 2016 года, улучшение реализованных подходов, начало подготовки к реализации случайного леса | 31.03.2025-  22.04.2025 |  |
| 8 | Постерная сессия | 22.04.2025 |  |
| 9 | Полноценнная реализация модели классификации - случайного леса, релазиация МСМС, итерация по спектральным характеристикам, улучшение веб-интерфейса | 26.04.2025-  10.05.2025 |  |
| 10 | Сравнительный анализ подходов  по ключевым метрикам | 10.05.2025-  15.05.2025 |  |
| 11 | Написание отчета | 15.05.2025-  29.05.2025 |  |
| 12 | Защита проекта | 06.06.2025 |  |
| 13 | Добавление внешних факторов для модели классификацияя | Не завершено | Дополнительные факторы будут получшены по мере окончание реализации парсера открытых источяников |
| 14 | Сравнительный анализ для метода МСМС | Не начато |  |

# Анализ предметной области и обзор литературы

Все чаще люди из совершенно разных областей прибегают к математическим моде- лям для решения своих задач. Кредитные рейтинговые агенства (КРА) – не исключение. Существует множество исследовательских работ, посвященных применению математических методов в кредитно рейтинговой отрасли. В последние годы наблюдается стремительный рост интереса к применению алгоритмов машинного обучения для автоматизации и повышения точности прогнозов кредитных рейтингов. Одной из таких работ является «*Multi-modal deep learning for credit rating prediction using text and numerical data streams»* [1]. В этой работе исследуется применение многомодальных моделей глубокого обучения для прогнозирования кредитных рейтингов компаний, объединяя числовые финансовые показатели и текстовые данные. Авторы анализируют различные архитектуры и стратегии слияния данных, включая раннее и промежуточное объединение. Результаты показывают, что модели, основанные на сверточных нейронных сетях (CNN) с использованием стратегий слияния, превосходят другие подходы в точности прогнозирования.

В этом исследовании [2] авторы разработали подход к прогнозированию кредитных рейтингов клиентов с использованием различных алгоритмов машинного обучения. Были применены и сравнены по точности, полноте и площади под кривой ошибок (AUC) следующие модели: решающие деревья, случайный лес, метод опорных векторов (SVM) и логистическая регрессия. Также проведён анализ важности признаков, где ключевыми оказались: статус клиента, продолжительность кредита, кредитная история, сумма кредита, сбережения, наличие других заёмщиков, имущество и продолжительность занятости. Результаты показали, что алгоритм SVM наиболее эффективно предсказывает плохие кредиты, достигая точности 79,7%, AUC 0,76 и точности 0,88. После оптимизации модели AUC увеличился до 78%. Это исследование демонстрирует потенциал алгоритмов машинного обучения в прогнозировании кредитных рейтингов клиентов и может иметь значительные последствия для банковской и финансовой отрасли, позволяя более точно и эффективно предсказывать кредитные рейтинги и снижать риск дефолтов и финансовых потерь.

В книге David Lando [3] рассматриваются фундаментальные методы анализа кредитного риска, включая когортный метод и метод дюрации для построения матриц переходных вероятностей между кредитными рейтингами. Эти методы позволяют аналитикам количественно оценивать, как рейтинги заемщиков изменяются с течением времени, что важно для управления портфелем и оценки вероятности дефолта. Когортный метод предполагает фиксированную когорту компаний и отслеживание изменений их рейтингов на протяжении заданного горизонта, тогда как метод дюрации учитывает разные моменты рейтинговых изменений с поправкой на время пребывания в исходном состоянии. В книге проводится сравнение этих подходов, анализируются их преимущества и ограничения при работе с эмпирическими данными, а также описывается, как они могут быть использованы в рамках регуляторных требований (например, Basel II/III). Данная работа служит ценным источником как для исследователей, так и для практиков, обеспечивая математически строгое и в то же время прикладное понимание моделей кредитного риска.

Также следует отметить работу российских экспертов по данной теме, которые поднимают проблематику выбора метода для получения лучшей модели, В данной работе [4] авторы анализируют два подхода к построению матриц миграций кредитных рейтингов: классический когортный метод и непрерывный (дюрации) метод. На симулированных и реальных данных (рейтинги российских компаний за 2016–2024 гг.) показано, что непрерывный метод превосходит когортный по точности и стабильности оценки вероятностей дефолта. В исследовании смоделирована ежедневная миграция 5000 объектов между 8 рейтингами с учетом отзыва и дефолта. Показано, что когортный метод систематически завышает вероятность дефолта, особенно в высоких рейтинговых категориях, тогда как непрерывный метод точнее отражает динамику кредитного качества, снижая среднюю абсолютную ошибку (MAE) на 0,37 п. п. Также подтверждено, что для оценки на малых выборках (до 500 наблюдений) непрерывный метод сохраняет преимущество по сравнению с когортным. Это исследование актуализирует необходимость применения непрерывных моделей на развивающихся рынках с ограниченным объемом данных, таких как российский, и может способствовать более адекватной оценке рисков и резервов в банковском и инвестиционном анализе.

Вообще говоря, российская кредитно рейтинговая отрасль активное развитие получила с 2016 года, когда в данной отрасли начался активный рост кредитного рейтингования. На данный момент до сих пор развиваются методы для анализа и прогнозирования кредитных рейтингов, поэтому данная работа является актуальной для дальнейшего развития и применения тех или иных методов, как на российском рынке, так и на международном.

# Подробное описание подходов, методов и математиче- ских моделей, использованных при реализации про- екта

## Марковская модель

Прежде чем перейти к непосредственному описанию использованной модели, приведем ряд основных определений и результатов, полученных для марковских процессов. Боль- шинство излагаемой ниже теории можно найти в источниках.

### Однородные марковские цепи

На вероятностном пространстве случайный процесс с дискретным множеством состояний называется *цепью Маркова*, если для любого для произ- вольных моментов времени и произвольных состояний

*Переходными вероятностями* марковской цепи называются функции

Они удовлетворяют следующим условиям:

2. и

(1)

Если зависят лишь от разности , то цепь называется *однородной*. В этом случае . А равенство ([1](#_bookmark15)) запишется как

или, в матричной форме, , где P(t) – матрица, состоящая из элементов

. Заметим, что если T – дискретно, тоесть в точности элемент где

— *n*-я степень матрицы *P(1)*. То есть, зная текущее состояние *i* и матрицу перехода за 1 шаг, можно получить вероятность перехода из *i* в *j* за *n* шагов.

В случае непрерывного времени для отыскания , вводится понятие *инфинитезималь- ной* матрицы *Q* (или как её называют в работе [3] - матрица генератор)и используется результат А.Н.Колмогорова. А именно, если марковская цепь *стандартна*[1](#_bookmark0), то существует правая производная

Если пространство состояний S конечно и марковская цепь стандартна, то справедливы обе системы уравнений Колмогорова (прямая и обратная), позволяющие найти переходные

1

вероятности за время *t*

формальное решение которых при выглядит следующим образом [[4](#_bookmark63), Гл.4]:

Действительно, почленно продифференцировав решение по t, получим [[21](#_bookmark80), Гл. 1]:

Полезно сразу записать прямую систему уравнений Колмогорова в следующем виде:

(2)

где

В работе [3] для непрерывного времени используется очень удобная формула для оценки элементов матрицы *Q* методом максимального правдоподобия (MLE)

*-* оценка интенсивности перехода из *i* в *j*

*-* количество наблюдаемых переходов и *i* в *j* на интервале времени *[0,T]*

*-* число объектов, находящихся в состоянии *i* в момент времени *s*

Данный подход очень удобен в использовании, а полученная матрица *Q* так же раскладывается в ряд Тейлора, как из определения выше.

### Поведение кредитный рейтингов как марковская цепь

Согласно процессам, описанным в множестве литературе, кредитный рейтинг может находиться в следующих состояниях, что были определены выше. Будем считать, что множество состояний цепи S = {«AAA-AA», «AA-A», «A-BBB», «BB», «B», «CCC-C», «D»}, или можно рассмотреть состояния, как на изображении далее.

Состояние «D» является поглощающим, то есть объект не может перейти в другое

состояние после попадания в Default. Граф переходов состояний представлен на рис.[1](#_bookmark19). (взят из презентации на представлении проекта)

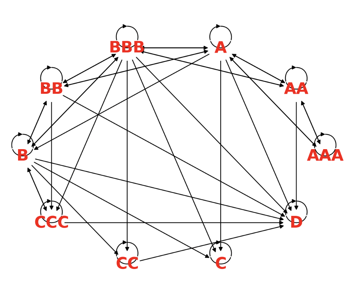


Рис. 1: Граф переходов состояний кредитного рейтинга

Рассмотрим простой пример из [3]. Пусть на момент времени десять объектов имели рейтинг А, десять объектов рейтинг В и 0 D (т.е. всего 3 состояния). Через 1 месяц компания из А сменила рейтинг на В, через 2 месяца компания из В сменила рейтинг на А и ещё через 6 месяцев компания из В ушла в дефолт (D). Для такой задачи надо построить матрицу переходных вероятностей. Если рассматривать частоты переходов, то мы получим просто матрицу вида:

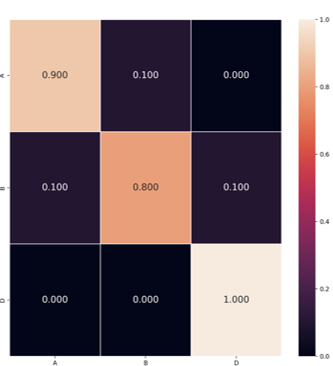


Рис. 2: Тепловая карта

матрицы переходных вероятностей

дискретного подхода

Однако, если мы воспользуемся марковской цепью с непрерывным временем, то получим уже иную матрицу переходных вероятностей:

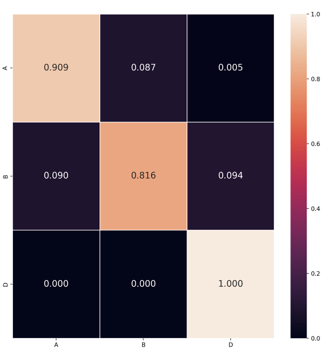


Рис. 3: Тепловая карта

матрицы переходных вероятностей

непрерывного подхода

На данном этапе мы полагаем, что матрица у нас однородна, так как никаких дополнительных факторов, зависящих от времени напрямую мы не наблюдаем. У нас есть лишь фактические рейтинги, с которыми мы и можем работать.

В некоторых случаях также полезно смотреть на процесс как на цепь Маркова 2 по- рядка — процесс, удовлетворяющий условию

для любого , для произвольных моментов времени

и произвольных состояний . Другими словами, цепь Маркова 2- го порядка будет обладать «памятью», и будущее состояние станет зависеть от настоящего и

одного прошлого состояния. Чтобы реализовать это технически, в качестве пространства состояний нужно взять упорядоченные 2-кортежи значений S и рассмотреть процесс на этом пространстве как цепь Маркова 1-ого порядка. При этом множество состояний цепи

Маркова сильно возрастает, для случая с 3 состояний:

Отметим здесь же, что всякая цепь Маркова первого порядка может считаться цепью Маркова второго порядка (это следует из определения выше), поэтому такое рассмотрение шире с точки зрения получения интересных результатов.

### Применение априорного распределения Дирихле

В отличии от обычного подхода, описанного раннее (с дискретным временем), байесовский подход с априорным распределением Дирихле позволяет использовать регуляризацию оценок, так как:

· При малом числе наблюдений , оценки нестабильны;

· Если , что может привести к вырожденности модели (например, при прогнозе );

· Невозможно задать **экспертные предположения** о характере миграции рейтингов (например, близкие переходы более вероятны).

Именно поэтому рассматривается новая идея - использование байесовского подхода с априорным распределением Дирихле. В работе [6] авторы предлагают байесовскую непараметрическую модель для оценки матриц миграции рейтингов и вероятностей дефолта, которая связана с распределением Дирихле. Этот подход позволяет гибко моделировать переходы между рейтингами с учетом наблюдаемых данных. В работе [7] изучается время расчета оценки матрицы переходных вероятностей для байсовского подхода и подхода с непрерывным временем.

где

- априорные параметры

- равномерное распределение (удобно использовать без экспертных оценок)

После наблюдения переходов, апостериорное распределение так же будет Дирихле:

Что позволяет:

· гибко задавать априорные веса

· переходит в частотную оценку при .

· не равна нулю, если

### Обучение марковской модели

Опишем принцип обучения и получения оценок в марковской модели для однородного случая. Как мы помним, для идентификации марковской цепи достаточно задать семей- ство матриц . Согласно уравнениям Колмогорова, можно получить взятием матричной экспоненты от инфинитезимальной матрицы со множителем поэтому задача сводится к оценке элементов этой матрицы. В выбранном нами пакете для этого используется метод максимального правдоподобия.

Представим, что на вход модели подается цепочек. Для каждой цепочки нам известны и состояниянаблюдаемые в эти моменты. Мы можем записать правдоподобие цепочки в общем виде :

Отсюда же можно напрямую получить оценку интенсивностей перехода ()

Остается решить задачу безусловной оптимизации:

В в общем виде ищется минимум минус логарифма функции правдоподобия, то есть:

Собственно, данная задача при дифференцировании по и приравниванию к нулю получаем выражение для оценки компонент матрицы *Q,* согласно работе [3], [5]*.*

## Модель классификации

Прежде чем перейти к описанию использованного алгоритма, приведем общую по- становку задачи классификации. Основную часть повествования ниже можно найти в источниках [[5](#_bookmark64)], [[22](#_bookmark81)], [[13](#_bookmark72)], [[7](#_bookmark66)].

Пусть имеется множество *обучающих* объектов *X*, относящихся некоторым образом

к известным классам *Y*. Один объект относится ровно к одному классу *.* Пусть также имеется множество *контрольных* объектов , у которых отнесение к классам

*Y* неизвестно. Задачу классификации можно сформулировать как поиск отображения

из множества объектов в множество классов *Y*. То есть, польза заключается в

возможности отнести к классу именно контрольные объекты.

Предположим, что мы получили алгоритм , решающий нашу задачу. Следующая не менее важная задача — оценить качество этого алгоритма. Верно ли мы угадываем классы? В зависимости от исходной выборки , числа классов в и цели, на которую направлена задача классификации, могут использоваться совершенно разные метрики. Ниже приводятся наиболее распространенных метрики для бинарного классификатора (отрицательный/положительный объект) с указанием случаев, когда их целесообразно использовать.

1. **Accuracy.** Она измеряет долю правильно классифицированных образцов по отно- шению ко всем классифицированным образцам.

Непоказательна, если существует явный дисбаланс классов. Например, перед нами стоит задача - предсказать будет ли следующим состоянием 1(Default) или 0(не приведет к дефолту), и пусть мы имеем:

· Всего 1000 компаний.

· Только 50 из них — дефолтные (положительный класс).

· Оставшиеся 950 — не дефолтные (отрицательный класс).

Классификатор, который всегда предсказывает "0" (т.е. без дефолта), даст:

· TP = 0 (никто из дефолтных не найден),

· TN = 950 (все недефолтные предсказаны правильно),

· FP = 0,

· FN = 50 (все дефолтные пропущены).

Может создать ложное впечатление, что модель "хороша",

1. **Recall(полнота).** Измеряет долю правильно классифицированных положительных образцов по отношению к общему количеству реальных положительных образцов. Она помогает оценить способность модели обнаруживать положительные образцы

В задачах классификации с выраженным дисбалансом классов, метрика *Recall* приобретает особую значимость. Например, если кредитно-рейтинговое агентство (КРА) строит модель, задача которой — выявлять переходы компаний в состояние дефолта (Default), следует учитывать, что дефолты представляют собой редкие события в общем распределении рейтингов. Модель, фокусирующаяся на правильной классификации стабильных рейтингов, может демонстрировать высокую общую точность, при этом практически не обнаруживая реальные дефолты. Метрика *Recall* в этом случае показывает, насколько полно классификатор выявляет целевой (дефолтный) класс. Таким образом, *Recall* помогает оценить, насколько эффективно модель КРА справляется с задачей распознавания критических случаев ухудшения кредитного качества, даже если такие события встречаются редко.

1. **Precision(точность).** Измеряет долю правильно классифицированных положительных образцов по отношению к общему количеству классифицированных положительных образцов. Эта метрика показывает, насколько точно модель идентифицирует положительные образцы.

С другой стороны, метрика *Precision* становится ключевой, когда важно, чтобы предсказанные дефолты действительно были обоснованными. Для КРА это означает, что ложные предсказания дефолта (ошибочное присвоение дефолтного рейтинга компании с приемлемым финансовым положением) могут повлечь за собой репутационные и регуляторные последствия, а также искажение рейтинговой статистики. Высокий уровень *Precision* говорит о том, что среди всех компаний, которым модель присвоила статус Default, действительно большинство находятся в тяжелом финансовом состоянии. Это важно для сохранения доверия к методологии агентства и минимизации ошибок первого рода при автоматизированной классификации кредитного риска.

1. **F-measure(F-мера).** Является гармоническим средним между полнотой и точностью. Она позволяет объединить эти две метрики в одно число, чтобы получить общую оценку производительности модели.

Существуют два экстремальных случая, которые можно рассмотреть, чтобы лучше понять влияние *F-measure* в зависимости от значений *Precision* и *Recall*:

* 1. Если все объекты помечены положительно , то *Recall* будет высоким, а *Precision* страдает.
  2. Если помечено очень мало объектов, как положительные, и при этом все они правильно помечены (т.е. метим единичкой только тех, в ком уверены однозначно), то *Precision* будет высоким, но *Recall* низким.

1. **Матрица ошибок** -– это показатель успешности классификации, где классов два или более. В случае бинарной классификации это таблица с 4 различными комбина- циями сочетаний прогнозируемых и фактических значений.

|  |  |  |
| --- | --- | --- |
| Прогноз | Реальность | |
| + | - |
| + | True Positive | False Positive |
| - | False Negative | True Negative |

* 1. True Positive (истинно-положительное решение) – прогноз совпал с реальностью, результат положительный
  2. False Negative (ложноотрицательное решение) – ошибка 2-го рода, модель пред- сказала отрицательный результат, но на самом деле он положительный
  3. False Positive (ложноположительное решение) – ошибка 1-го рода, модель пред- сказала положительный результат, а на самом деле он отрицательный
  4. True Negative (истинно-отрицательное решение) – прогноз совпал с реально- стью, результат отрицательный

### Случайный лес (Random Forest)

Представляет собой ансамбль решающих деревьев, каждое из которых обучается независимо на случайной подвыборке из исходных данных, а итоговое решение принимается агрегированием предсказаний всех деревьев:

1. Пусть дана обучающая выборка:

где - вектор признаков объекта (например, финансовых или макроэкономических характеристик компании), а - метка класса, соответствующая одному из *K* кредитных рейтингов.

1. Алгоритм строит ансамбль из *T* решающих деревьев:

Каждое дерево обучается на случайной подвыборке полученной методом бутстрапа (случайной выборкой с возвращением).

1. При построении каждого дерева в каждом узле случайным образом выбирается подмножество признаков

Выбирается оптимальное разбиение пространства:

где

- критерий информативности. Во многих задач классификации рассматривается как индекс Джини.

- индекс выбранного признака

- порог, по которому мы делим множество объектов в текущем узле.

1. Предсказание ансамбля осуществляется через голосование деревеьев:

где

- индикатор 1, если предсказание совпало с *с,* иначе 0.

### Прогноз кредитного рейтинга как задача классификации

В качестве множества ответов Y выступают состояния из *S.* В обучающей выборке для прогноза месяца каждой цепочке длины *l* ставится в соответствие пара где– последнее известное состояния – вектор признаков, получающийся из информации о состояниях.

В нашем случае, в качестве обучающей выборки X используются следующие признаки:

* + - * длина цепочки (срок «жизни» объекта)
      * Кол-во компаний на момент времени *t*
      * rating lag - Последовательность кредтных рейтингов за предыдущие 12 месяцев.
      * Год - месяц - день

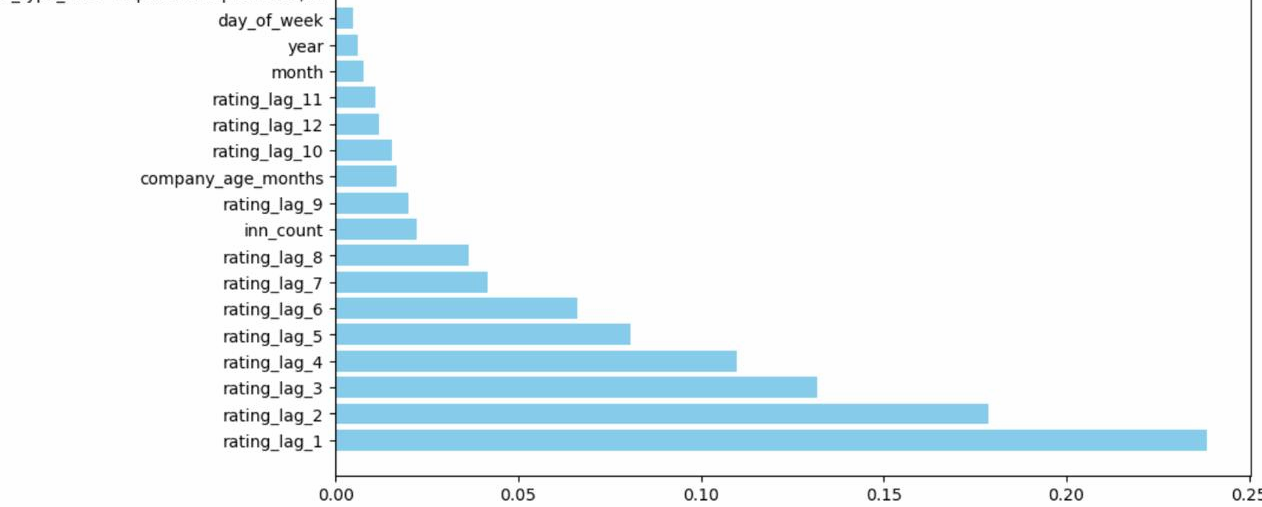


Рис. 4: Важность признаков после обучения

## Спектральные характеристики

## Дополнительно приведем формализацию подходов, связанных с частотным анализом поведения рейтингов во времени и межфирменными зависимостями на основе корреляционной структуры. Соответствующие идеи можно найти, в частности, в работе [10]. Из основных идей выделим нахождение параметра Херста, который поможет определить тенденцию наших данных к зависанию в состоянии *i* или наоборот его скорому изменению, а также зависимость тех или иных компаний и их кластеризацию на основе теории *SVD.*

### Параметр Херста

Параметр Херста — это числовая характеристика временного ряда, отражающая степень его **долгосрочной памяти (персистентности)**. Он используется для количественной оценки, **насколько текущие изменения сигнала зависят от его поведения в прошлом**. В нашей задаче параметр Херста рассчитывается для временного ряда, представляющего собой **долгосрочную ретроспективную последовательность кредитных рейтингов компании**.

Параметр Херста показывает:

* : поведение похоже на случайное блуждание;
* : персистентное поведение — рост (или падение) рейтинга, скорее всего, продолжится;
* : антиперсистентное поведение — рост сменяется падением и наоборот.
  1. **Метод нормализованного размаха (R/S-анализ)**

Пусть – числовой временной ряд. Для каждого масштаба временной ряд разбивается на непересекающихся сегментов длины nnn. Для каждого сегмента вычисляется:

1. Среднее значение:
2. Кумулятивные отклонения:
3. Размах:
4. Нормализованный размах:

Далее из уравнения находится параметр Херста:

* 1. **Оценка через спектральный наклон**

Второй подход основан на анализе плотности спектральной мощности сигнала. Предполагается, что спектр *S(f)* подчиняется степенному закону:

где

- мощность сигнала на частоте *f*,

- наклон в логарифмической шкале.

Параметр Херста при этом выражается через наклон следующим образом:

### Матрица кросс-корреляции

Для анализа линейных связей между этими рядами строится матрица кросс-корреляции

Где каждый элемент определяется как нормированное ковариационное выражение:

Данная матрица С будет либо ту же размерность, сколько N цепочек мы имеем, либо меньше, в зависимости от разнообразности данных и кол-ва переходов внутри цепочек.

**Сингулярное разложение (SVD)** для матрицы C имеет вид:

где

*-* ортонормированные матрицы собственных векторов.

- диагональная матрица сингулярных значений, упорядоченных по убыванию:

Каждое отражает вклад *k*-ой компоненты в общую структуру кросс-корреляций. Чем больше значение , тем ярче выражен единый (глобальный) фактор, влияющий на рейтинговое поведение разных компаний.

## Марковская цепь Монте-Карло (MCMC)

В рамках байесовского подхода к построению вероятностных моделей классификации и прогнозирования рейтингов, возникает необходимость в численном приближении апостериорного распределения параметров. В задачах, где аналитическое выражение этой апостериорной функции невозможно получить в замкнутом виде, используется метод Монте-Карло на базе марковских цепей (MCMC).

Пусть даны наблюдаемые данные - кредитные объекты некоторого объекта длины *T.*

*-* вектор параметров модели

- функция правдоподобия

- априорное распределение параметров

Задача: сгенерировать выборку из апостериорного распределения:

Вместо точного вычисления распределения метод МСМС строит марковскую цепь, стационарное распределение которой совпадает с нужным .

После «разогрева» (burn-in) цепь начинает производить зависимые, но корректно распределенные выборки из апостериорного распределения.

Также данный метод хорошо подходит для получения оценок в виде доверительного интервала каждой вероятности перехода

# Полученные результаты

## Данные

## Все данные были получены из открытых источников Банка России. В частности, данный репозиторий имеет всевозможные кредитные рейтинги с 2016 года и по настоящий день. Для того, чтобы их можно было использовать, был написан специальный скрипт на языке Python с использованием библиотек Pandas / Selenium / urllib , а также для успешного прохождения проверки на «робота» (капчи) была реализована сверточная нейронная сеть (CNN) в основе которой лежало около 10000 уникальных капч, использовавшихся для обучения и решения тех самых капч при парсинге, что существенно уменьшило время парсинга, а также достижение полного автоматизма.

Пример выгрузки на рисунке [5](#Данные).

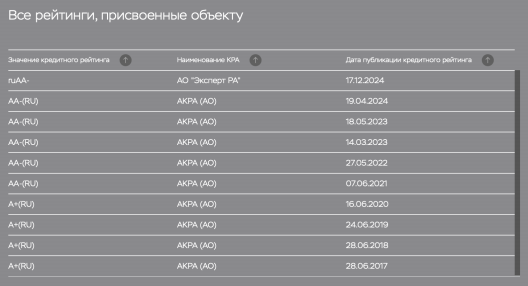
[](#Данные)

Рис. 5: Пример данных из открытого источника

Вид изначальных данных:



Рис. 6: Пример данных из открытого источника после парсинга

Прежде чем демонстрировать результаты моделирования, опишем некоторые особенности в данных. Всего мы имеем 7740 наблюдений (не облигаций) и 8591 наблюдений по облигациям на 31.12.2024 (последний парсинг в 2024 году). На рис. 7 продемонстрировано распределение всех сфер с числовым и процентным содержанием наблюдений:

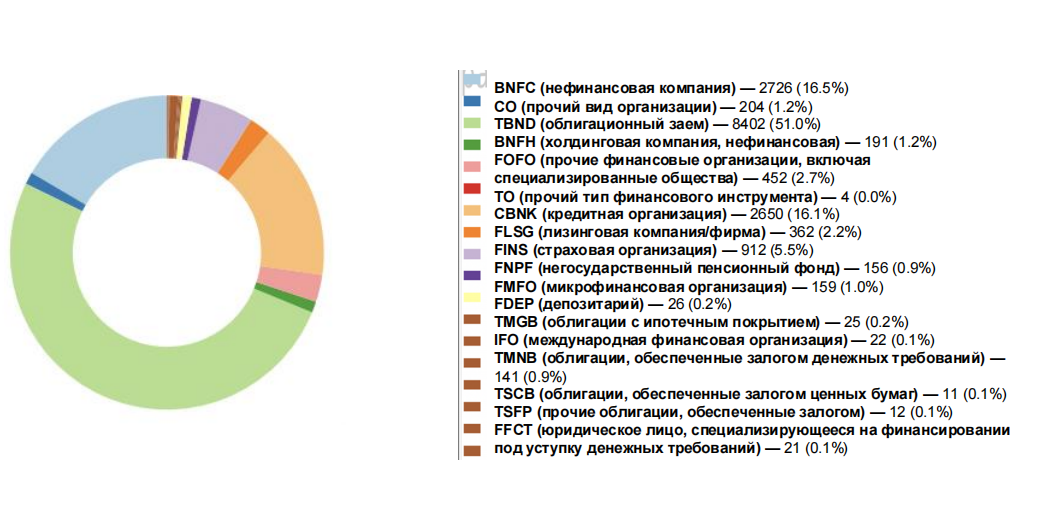


Рис. 7: Распределение всех наблюдений

А так же по каждому из КРА в отдельности:

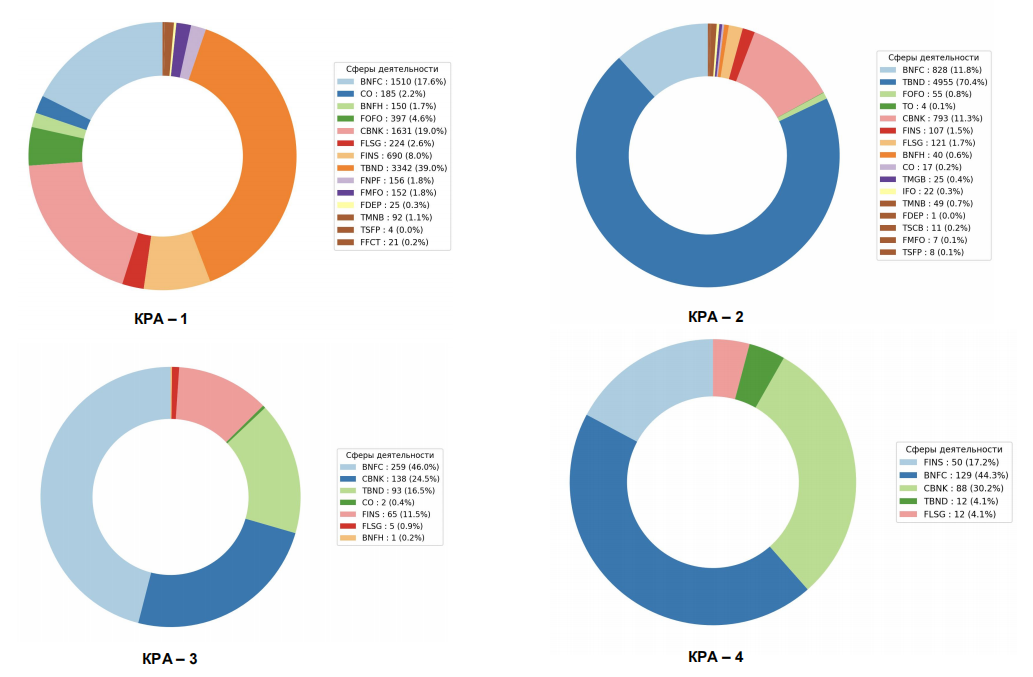


Рис. 8: Распределение всех наблюдений по каждому КРА

После ,так называемого, ресемпилнга кол-во данных значительно увеличилось. В расчёт не шли какие либо облигации и сферы, чьё обозначение начиналось с T (облигации). Всего мы получили 99180 наблюдений помесячных. К ним применялись два типа группировки, описанных в самом начале.

Так же есть данные уже для модели классификации. Приме таких данных продемонстрирован на Рис. 9:



Рис. 9: Пример данных для модели классификации

**О предобработке, делении выборки и метриках.** Для обработки цепочек перед обучением, деления на обучающие и тестовые данные использовался язык программиро- вания Python (библиотеки numpy, pandas, sklearn). Для подсчета интересующих метрик и формирования прогнозов на периодов вперед были написаны пользовательские функ- ции. Код содержится в файлах на GitHub.

Деление данных на обучающую и тестовую выборки происходило в соотношении 80:20 (это делалось для честного сравнения с классификатором).

## Результаты baseline-алгоритмов

Прежде чем строить нетривиальные модели и оценивать непосредственно их качество прогнозов, разумно построить простые базовые модели, которые позволят сформировать baseline для более сложных. Часто на практике оказывается, что простые триггерные модели могут работать не хуже продвинутых алгоритмов, но затраты на их создание, поддержку и развитие намного меньше. Реализация baseline алгоритмов проведена на языке программирования Python в среде PyCharm и Jupyter Notebook.

**Наивный классификатор.** Суть этого подхода заключается в том, чтобы копировать последнее известное состояние как прогноз. Например, для цепочки прогноз на 3 месяца вперед будет так как последнее известное состояние было Заметим, что при таком подходе мы никогда не будем прогнозировать состояния поскольку работаем только с открытыми на момент прогноза цепочками. В нашей задаче прогноз строился в горизонте 1 / 2 / 3 / 4 года. Как правило, в каждый из следующих годов отличных от первого вероятность попадания в D повышается для низких рейтингов, которые в общем случае описываются как «Спекулятивный».

**Результаты base-line алгоритма (наивный классификатор).** Результаты приведены ниже. В целом, как и ожидалось для разных типов компаний (некоторые собраны в группу, так как местами не хватало наблюдений для вычисления метрик качества. Результаты приведены на Рис. 10:



Рис. 10: Результаты наивного классификатора

## Результаты марковской модели

Реализация модели, основанной на марковской цепи, происходила на языке програм- мирования Python в среде Pycharm. В результате в рамках каждого кластера на тестовых данных были получены оценки для цепочек на горизонт 1/2/3/4 года. На основе спрогнозированных состояний вычислялась точность (доля верных ответов) и *F-measure*

В первую очередь обсудим модель, основанную на однородной марковской цепи 1-го порядка и 2-ого порядка для 1-ого КРА с прогнозом на 1 год (12 месяцев). Подробные результаты представлены на рисунке [11](#_bookmark39).



Рис. 11: Результаты марковской цепи 1-ого порядка и 2-ого порядка

Сразу стоит отметить, что у нас есть прогнозы всего по 5-4 компаниям, что очень и очень плохо. Как и ранее, в конечный результаты эти наблюдения были сгруппированы в одно целое. Также выводились матрица переходных вероятностей на Рис. 12, 13 и 14 для каждой марковской цепи (здесь состояния из укрупненной группы):

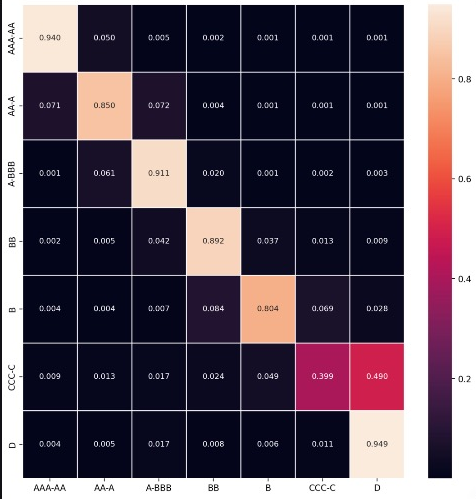


Рис. 12: Тепловая карта (heatmap) марковской цепи 1-ого порядка

с априорным распределением Дирихле



Рис. 13: Тепловая карта (heatmap) марковской цепи 1-ого порядка

с непрерывным временем

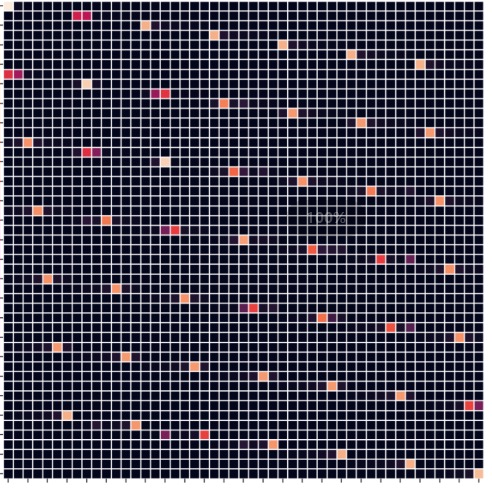


Рис. 14: Тепловая карта (heatmap) марковской цепи 2-ого порядка

Не смотря на то, что многие клетки темного цвета - почти все из них больше нуля. Значения для марковской цепи 1-ого порядка достаточно близки друг к другу, что является хорошим результатом, так как по своим особенностям они могут расходиться и достаточно сильно, если исходный набор данных ресемплирован с меньшим шагом.

Также стоит отметить, что полный список результатов доступен в Приложении и отдельным файлом в GitHub репозитории данного проекта.

## Результаты модели классификации

Аналогичный расчет проведем для модели классификации - случйного леса. На Рис. 15 предстален результаты расчета матрики *F-measure:*



Рис. 15: Результаты модели классификации

Стоит отметить, что в отличии от марковских цепей, случайный лес обработал даже совсем плохие наблюдения, в одном случае, как и марковские цепи он правильно классифицировал 100 % компаний (всего их 4) , но в случае с остальными - он дал результаты сильно ниже - 48% и 71% соответственно. В среднем мы получили показатель по всем отраслям примерно 86%, что является довольно высоким показателем. Однако, если убрать выбросы, объединив их в укрупненные группы, то средний результат по *F-measure* и *accuracy* изменится 88,91% и 95,45% соответственно. Мы не рассматривает *recall* и *precision*, т.к. раннее было уопянуто, что *F-measure* является совокупностью данных метрик.

Также в Приложении и GitHub репозитории размещен файл со всеми метриками для каждой группировки рейтингов и для разных сфер дейтельности, КРА.

## Сравнение качества прогнозов

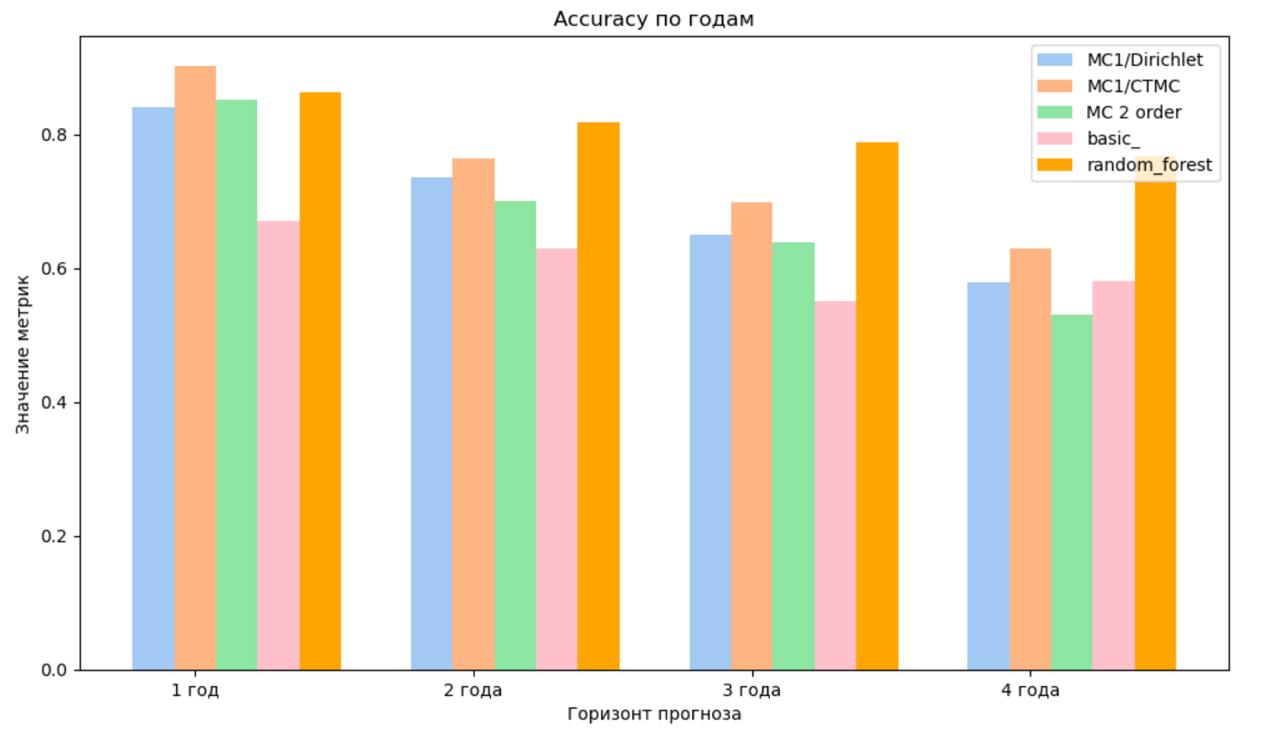
Подробные результаты сравнения качества прогнозов наивного классификатора, случайного леса и марковских моделей представлены на Рис. [16](#_bookmark47). Метрики выбраны те же, что и рассматривались раннее (*accuaracy / F-measure*). Результаты приведём усреднённые по всем КРА и областям, а так же в Приложении будут продемонстрированы результаты отдельно по каждому КРА и сфере деятельности.

Рис. 16: Результаты *accuracy*

На Рис. 17 представлен усреднённая по всем КРА и по всем сферам деятельности метрика *F-measure*

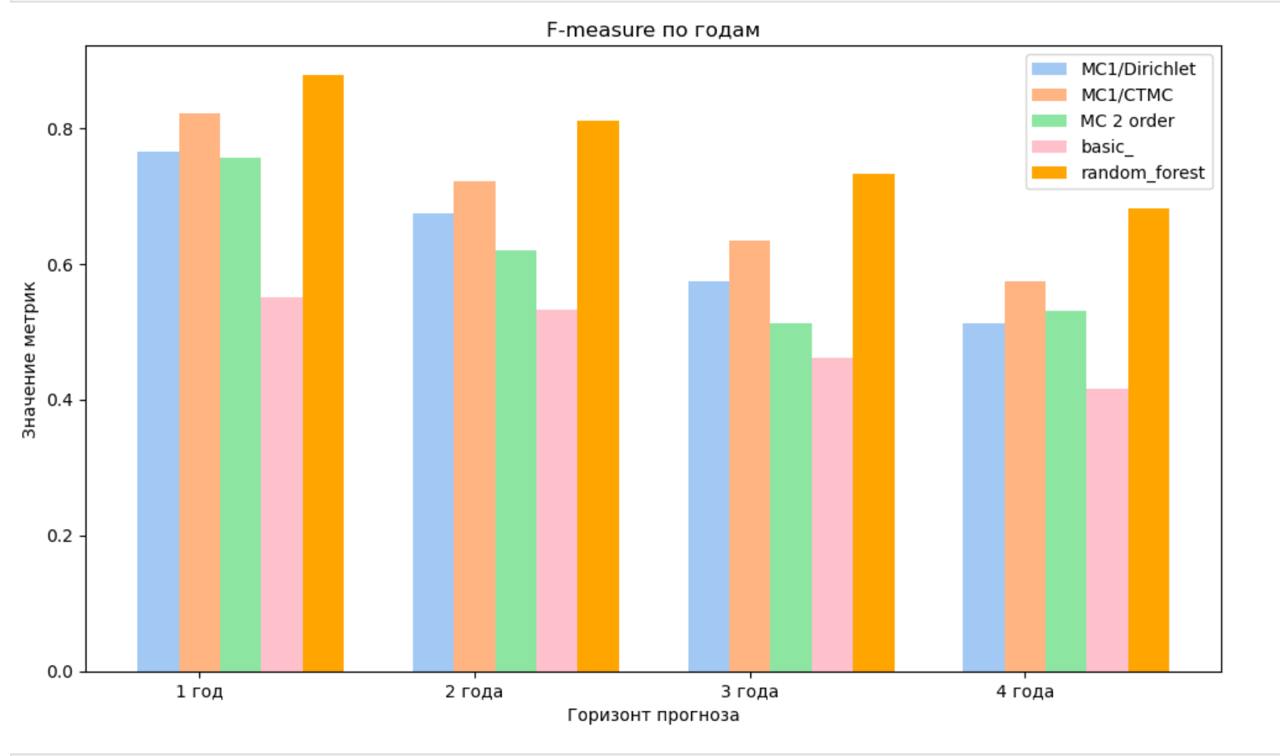


Рис. 17: Результаты *F-measure*

Что видно из этих двух графиков? По метрике accuracy и горизонту 1 год явно выделяется марковская цепь 1-ого порядка с непрерывным временем и случайный лес. В работе [7], как и ранее было сказано, рассматривается время работы и относительная сходимость двух методов. Это важно, когда мы предполагаем однородность данных без необходимости ресемплинга с меньшим шагом, так как длины «цепочек» увеличиваются. Поэтому марковская цепь 1-ого порядка с априорным распределением Дирихле также является довольно мощным инструментом для анализа и прогнозирования кредитных рейтингов. Марковская цепь 2-ого порядка также показала свою эффективность в горизонте 1 год. В горизонте 2 и 3 года тенденция схожая по марковским цепям, хотя случайный лес показывает большую эффективность при прогнозировании.

В случае с *F-measure*, которая, в общем случае, является более показательной при дисбалансе классов, ситуация примерно та же, за исключениям явного превосходства случайного леса. Во всех горизонтах прогнозирования данный метод имеет наибольшую эффективность и его способность к прогнозированию «падает» меньше, чем у альтернативных методов. На Рис. 18 продемонстрирована таблица относительной «деградации» (Relative Degradation) методов прогнозирования.



Рис. 18: Результаты относительной «деградации»

способности прогноизрования существующих методов

* Dirichlet и CTMC показывают наибольшую относительную деградацию (~30–33%) — возможно, они хуже захватывают долгосрочную динамику;
* Random Forest показывает наименьшую относительную деградацию (22%), что говорит о его устойчивости к увеличению горизонта прогноза;
* Basic модель(наивный классификатор) держится плохо по абсолютным значениям, но по относительному падению даже немного лучше, чем вероятностные модели;
* Модель второго порядка страдает от переобучения или недостатка данных при длинных цепочках.

Аналогичную таблицу приведём для метрики *accuracy*, так как по идее, она должна показывать примерно схожую ситуацию (поведение), что и *F-measure.*

·

Рис. 19: Результаты относительной «деградации»

способности прогноизрования существующих методов

* Наибольшую деградацию по *accuracy* демонстрирует модель второго порядка (почти 38%), что может указывать на переобучение или высокую чувствительность к изменению временного горизонта.
* Dirichlet и CTMC — также деградируют примерно на 30%, что схоже с F1-score.
* Basic модель по абсолютным значениям слабее, но деградирует сравнительно медленно (13.3%).
* Random Forest снова демонстрирует лучшую устойчивость к увеличению горизонта прогнозирования: падение *accuracy* менее 11.1% за 4 года — это наиболее стабильный результат.

Примерно схожую ситуацию мы наблюдаем в укрупненной группировке кредитных рейтингов. Результаты данных расчётов будут продемонстрированы в Приложении и в GitHub репозитории проекта.

В общих чертах однозначно можно заявить, что методы машинного обучения демонстрируют наибольшую эффективность при предсказании моделей.

Так же стоит отметить, что марковская цепь 1-ого порядка с непрерывным временем так же демонстрирует неплохие способности к прогнозированию на небольших горизонтах 1-3 года. В целом это согласуется со всеми экспертными оценками при расчёте матриц переходных вероятностей на эталонных рейтингах.

## Применение МСМС

Немало важным следует отметить применение МСМС в данной задаче. Данный метод позволил посмотреть насколько сильно наши кредитные рейтинги «разбрасываются» и посмотреть на их распределение. На Рис. 20, 21, 22, 23 представлены распределения выборочных вероятностей переходов, семплированных методом NUTS (\*По умолчанию в PyMC (начиная с версии 4.x) используется метод NUTS (No-U-Turn Sampler) — это адаптивный вариант гибридного метода Гамильтона (HMC, Hamiltonian Monte Carlo)

·

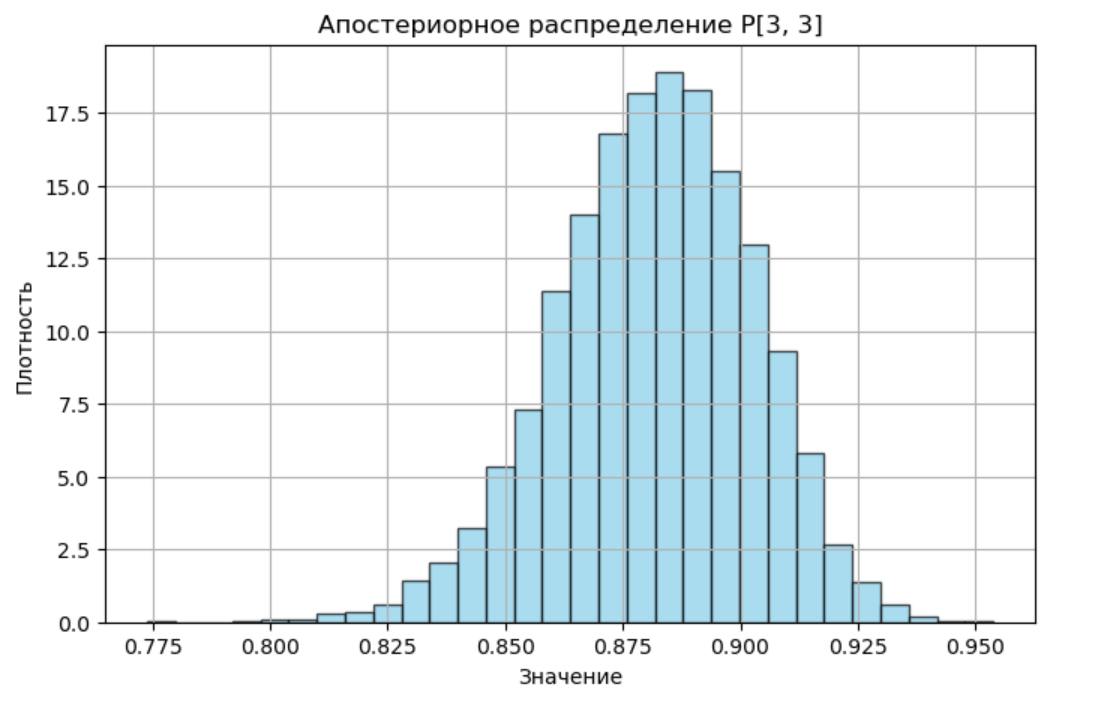


Рис. 20: Распределение вероятности перехода из 4 состояния в 4

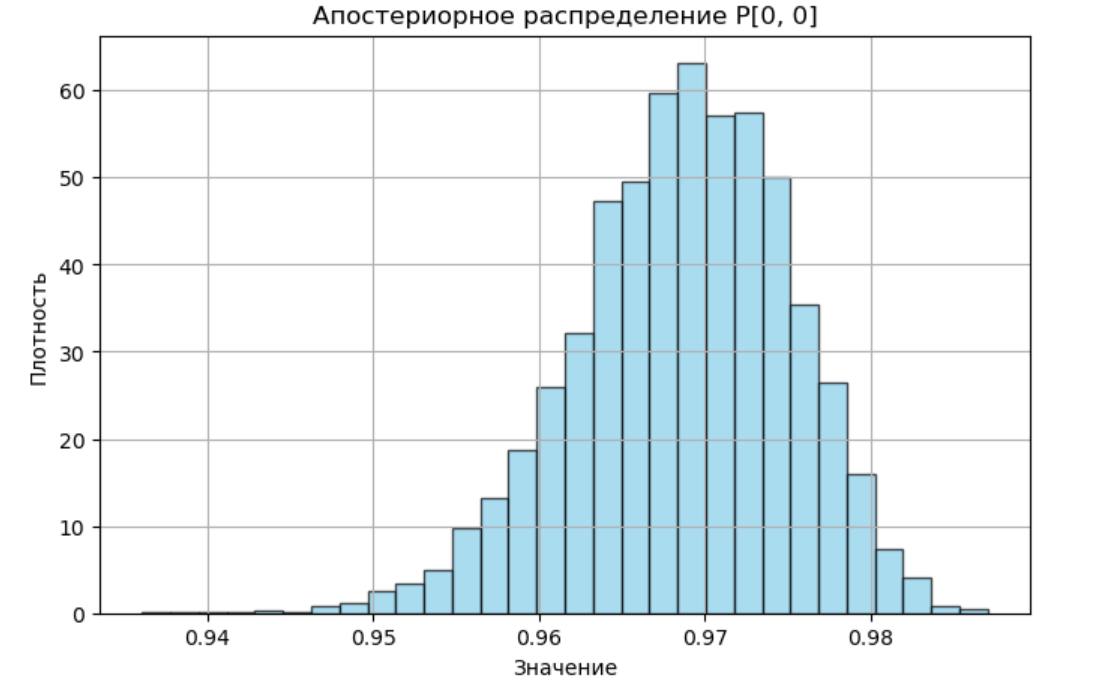


Рис. 21: Распределение вероятности перехода из 1 состояния в 1

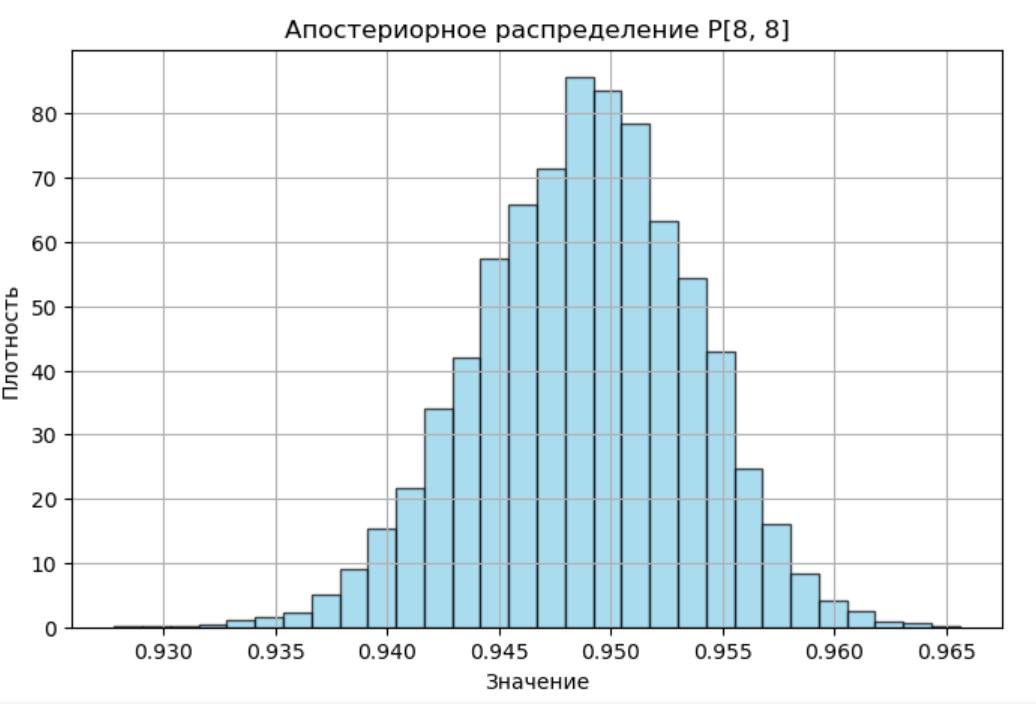


Рис. 21: Распределение вероятности перехода из 9 состояния в 9

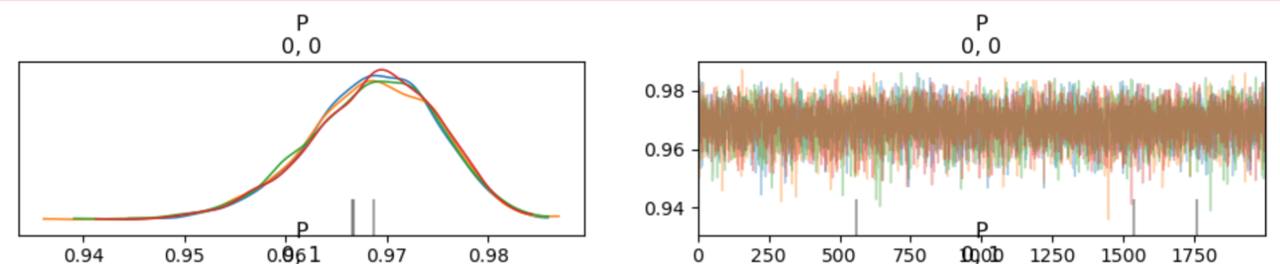


Рис. 22: Сэмплирование вероятностей перехода из 1 в 1 состояние

при разных «эпохах»

Данные апостериорные распределения можно строить абсолютно для всех вероятностей переходов и для всех сфер деятельности. В данном случае выбраны наиболее интересные переходы по соответствующим кредитным рейтингам, где 0,1,2,..,19 - индекс кредитного рейтинга согласно одной из таблиц п.1.1

Матрица, полученная данным методом очень близка к тому, что получается при использовании марковской цепи 1-ого порядка. Однако проверить сходимость - задача на будущее.

На данном этапе реализовано сравнение доверительного интервала методом МСМС и bootstrap.

В общем случае при большом кол-ве семплов матрица, полученная методом МСМС, должна быть близка к тому, что получается иными методами, в нашем случае наблюдаются отклонения, например, при вычислении распределения вероятности перехода из состояния С в состояние D (18 и 19 индекс соответственно). На Рис. 23 и Рис. 24 продемонстрированы соотвествующие вероятности:

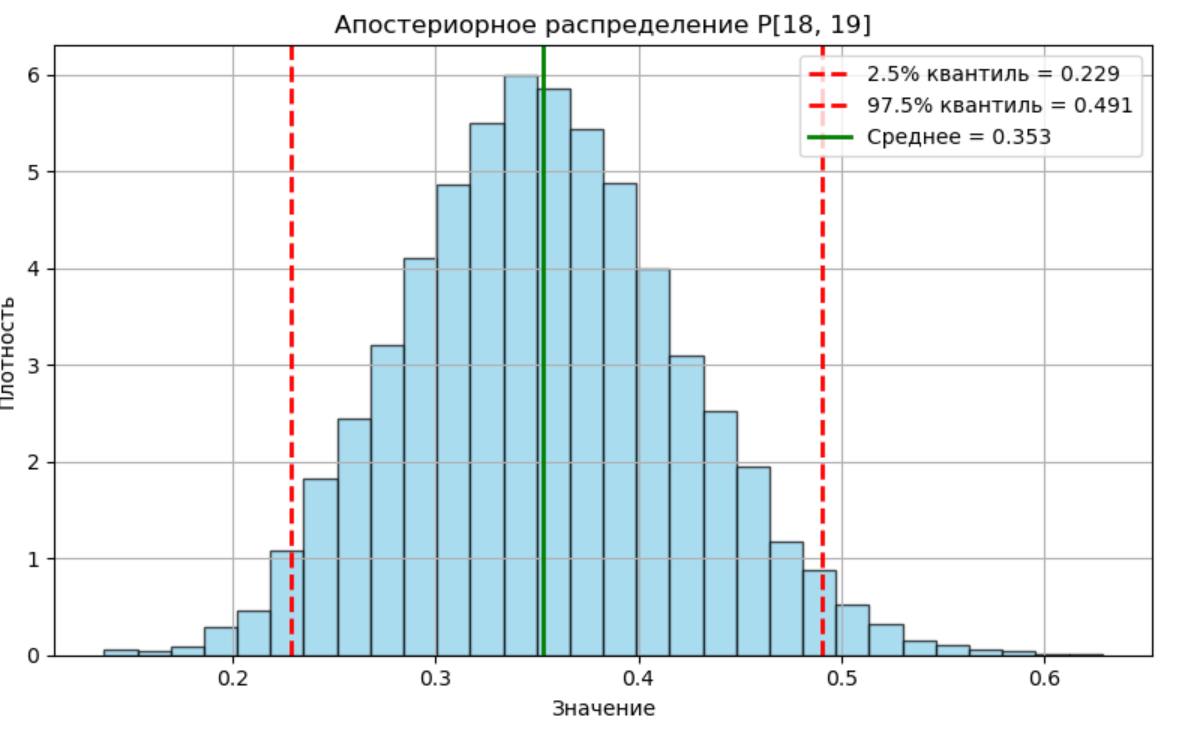


Рис. 23: Распределение вероятности перехода методом МСМС

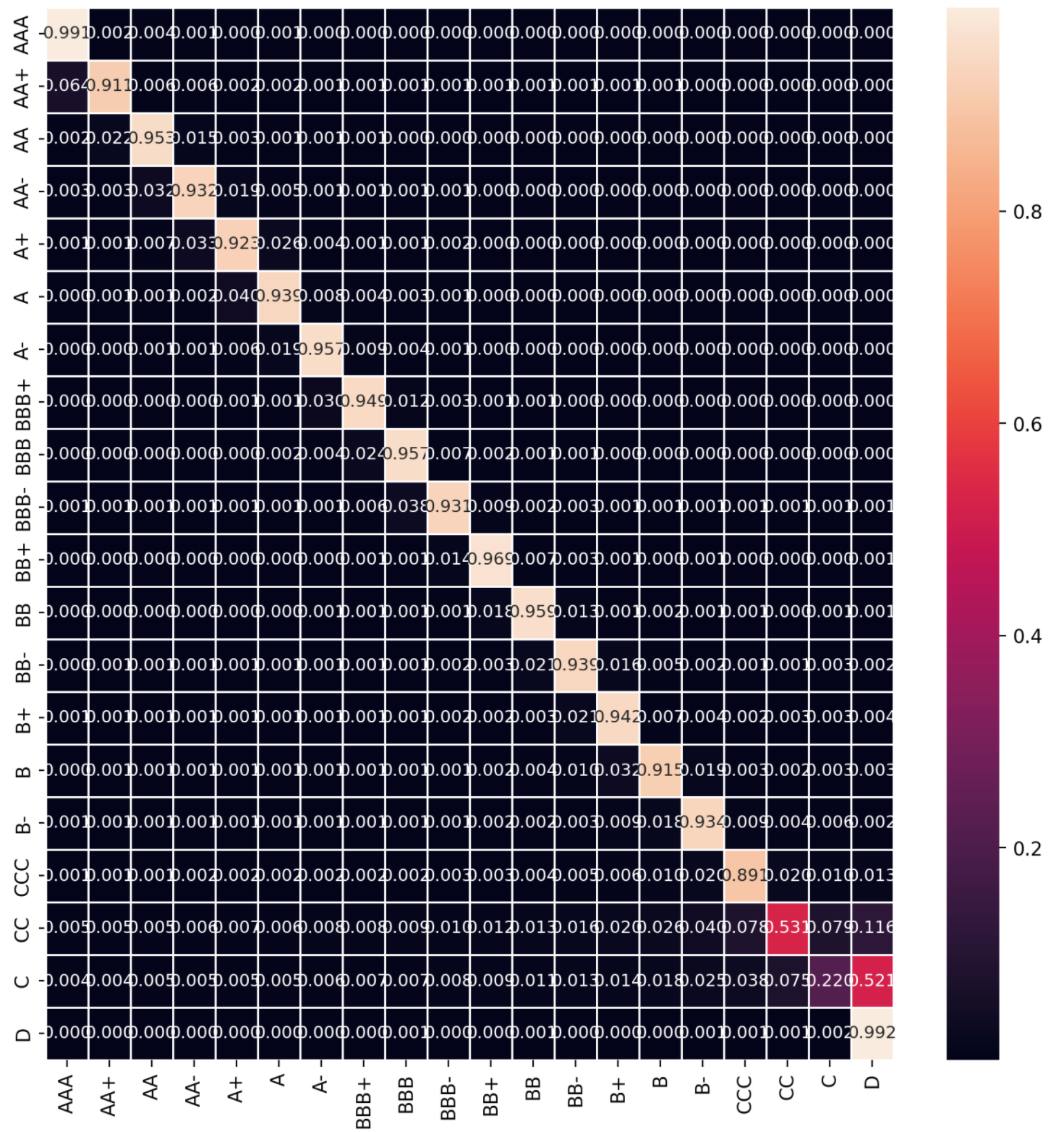


Рис. 24: Тепловая карта марковской цепи 1-ого порядка

Соответсвующий доверительный интервал, высчитанный методом bootstrap получился равным:

В целом, у данного похода есть огромный потенциал для использования в контексте кредитных рейтингов, поскольку он предоставляет широкий спектр методов построения распределений , а также является довольно гибким при использовании, так как не требует огромного кол-ва данных для семплирования, прекрасно работает с шумами, а также почти всегда точнее строит доверительные интервалы, в отличии от того же метода bootstrap.

## Расчёт спектральных характеристик и матрицы кросс-корреляции

На основе п.3.3 был рассчитан параметр Херста для наиболее представительных наборов «цепочек».

На основе всех расчётов была получена очень схожая картина, которая, в целом, демонстрирует общее поведение нашего набора данных.

Параметр Херста рассчитывался двумя способами и, как следствие, представлены два результата:

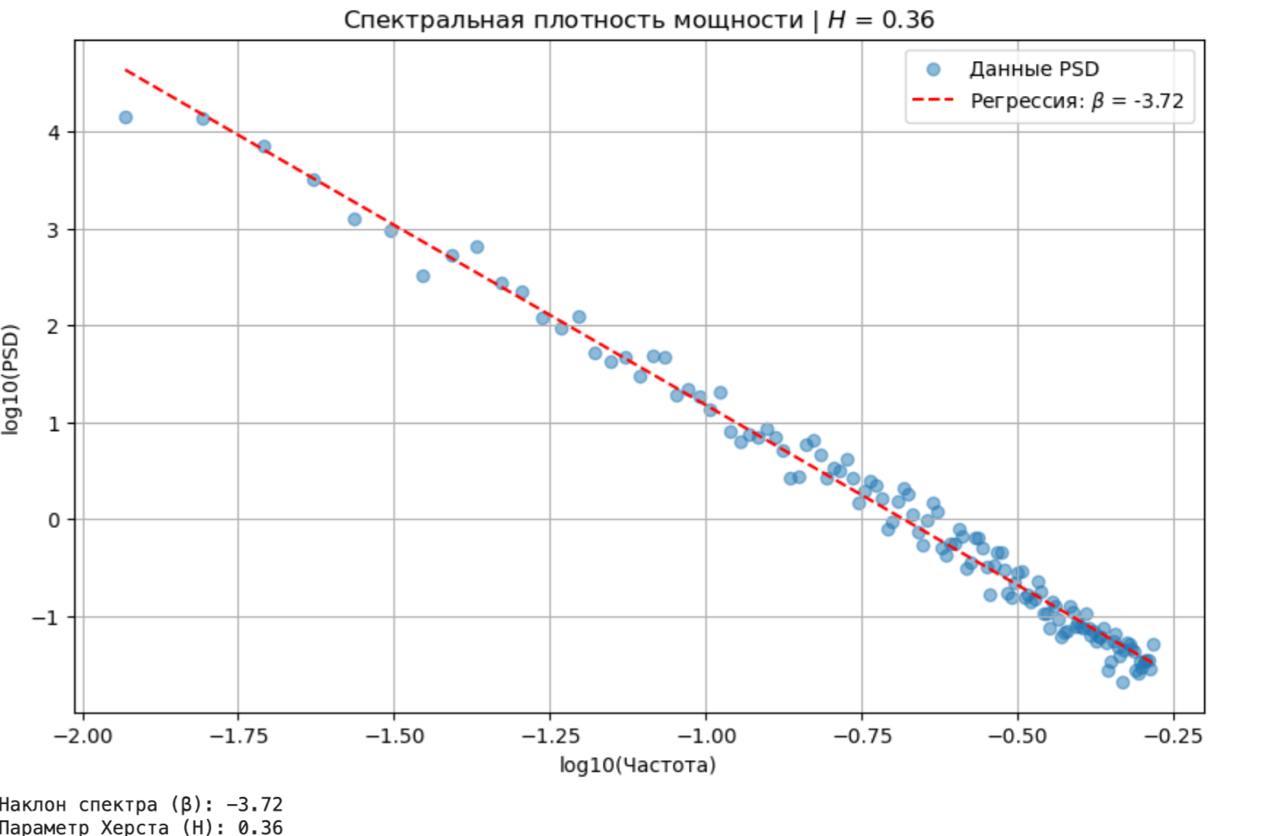


Рис. 25: Наклон спектра

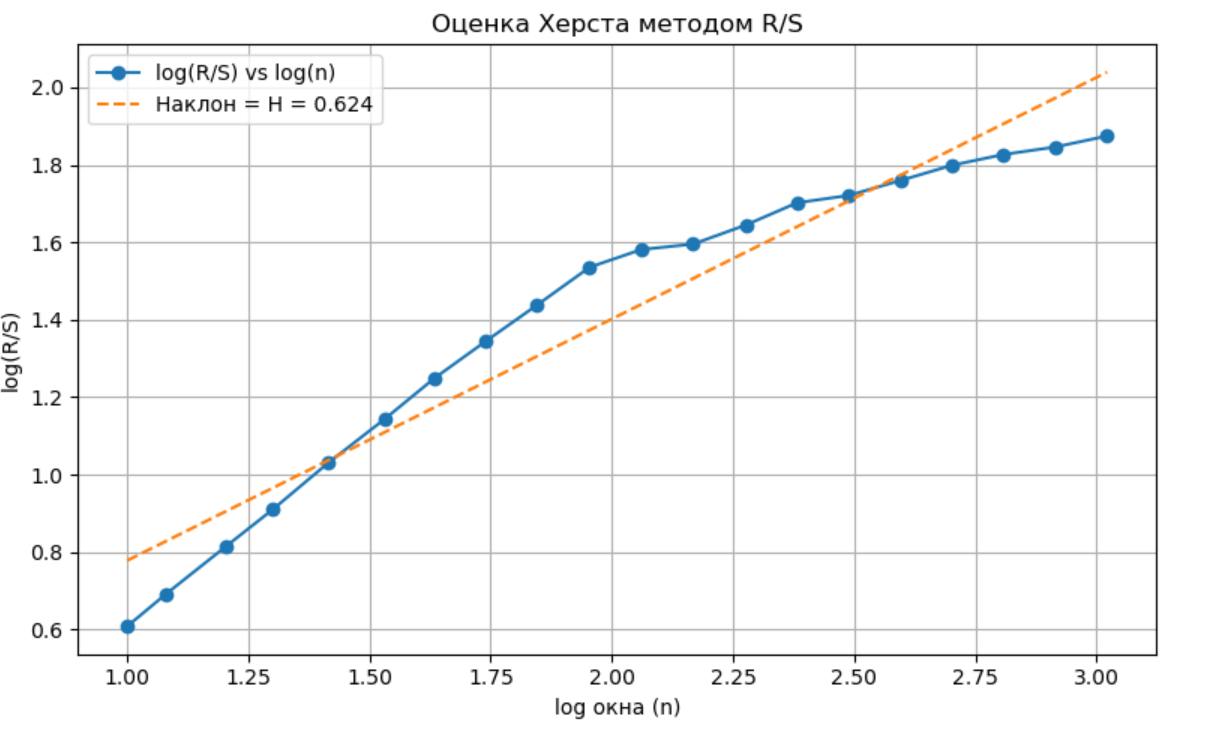


Рис. 26: R/S анализ

Два подхода дали совершенно разные результаты.Стоит отметить, что метод R/S как раз описан в работе [7] как наиболее релевантный, хотя из практики, наклон спектра также является мощным инструментом для анализа. Однако немало важным фактором играет то, как мы высчитываем данный наклон - через накопленную сумму (кумулятивную) или напрямую через ряд. От этого меняется исходная формула. Тем не менее, опираясь на визуальное представление рейтингов, а так же расчёты, проведённые в других пунктам, можно смело сказать, что параметр Херста для наших данных должен быть > 0.5, что совпадает с R/S анализом и следующим из него выводом о сохраняющийся тенденции исходных данных задерживаться в состояниях (либо, если рейтинг вырос, он продолжит расти, и наоборот).

Матрица кросс-корреляции вычислялась по формуле из п.3.3.2. На тех же данных, по которым вычислялся параметр Херста, так как они наиболее представительные, получим следующую матрицу (Рис. 27)

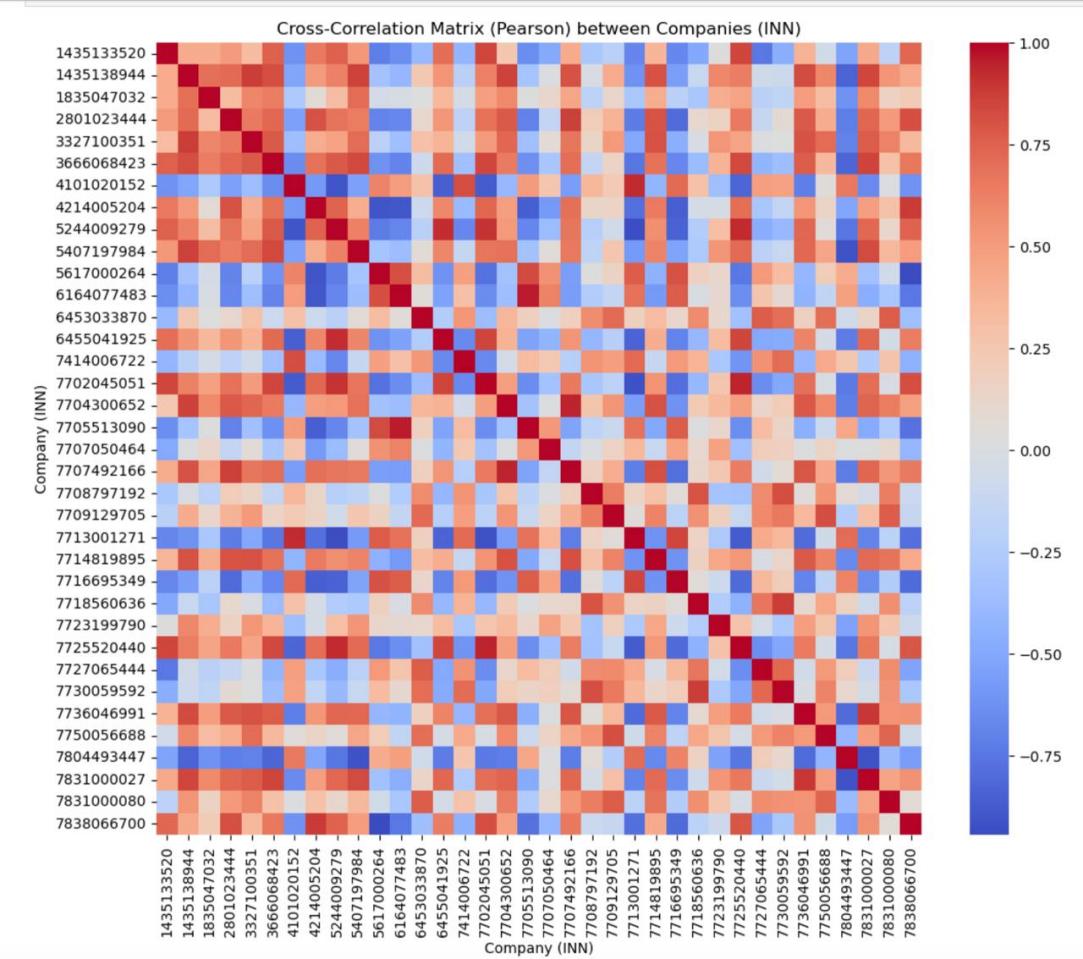


Рис. 27: Матрица кросс-корреляции

В пакете seaborn есть встроенная функция для кластеризации объектов на основе их коэффициента корреляции (Рис. 28)

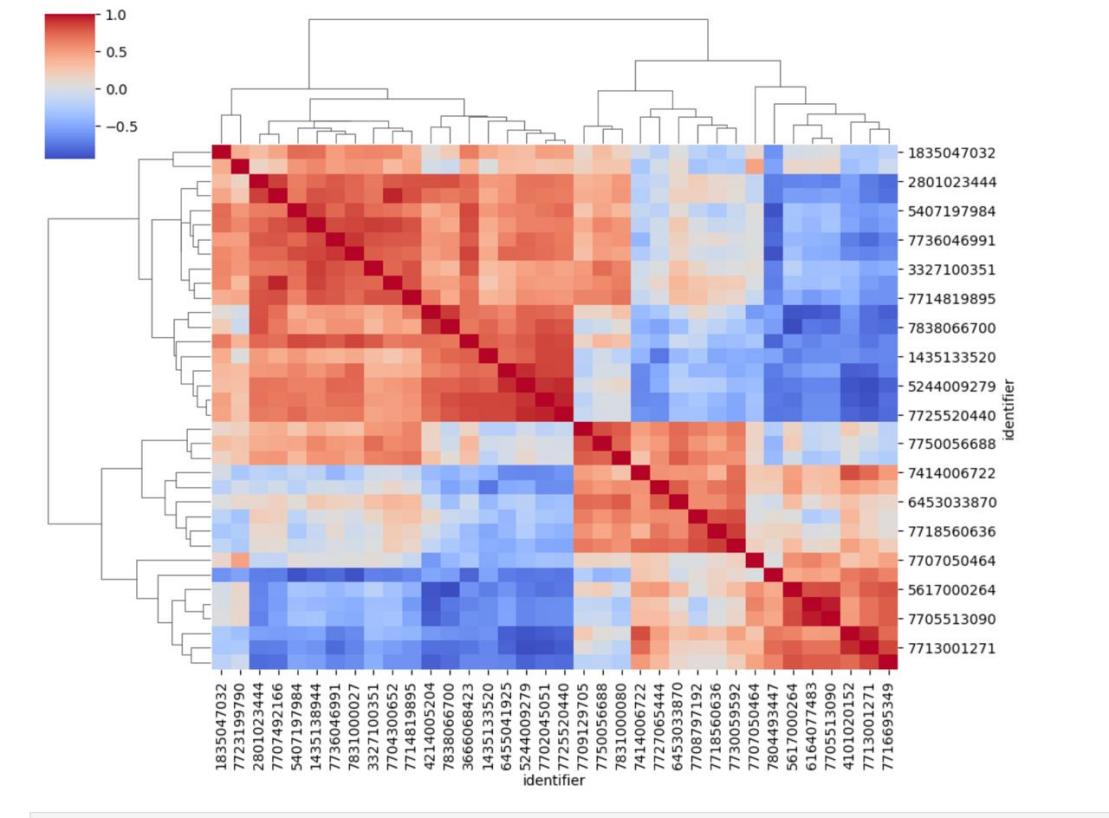


Рис. 28: Кластеризация матрицы кросс-корреляции

Однако нас больше интересует прикладная кластеризация посредством сингулярного разложения (SVD). Получим следующее (Рис. 29)

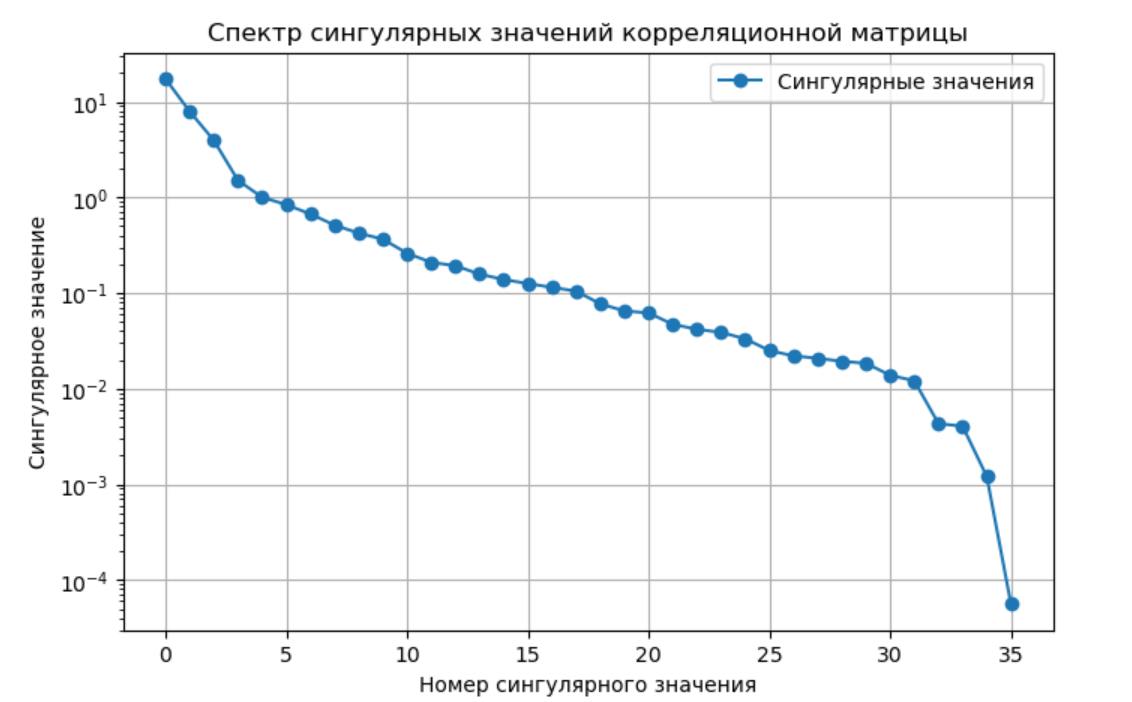


Рис. 29: Спектр сингулярных значений

А теперь посмотрим на главные компоненты и кластеризируем их (Рис. 30). Кластеры можно находить через метод локтя, например. Однако мы использовали фиксированное значение (k = 3). Это аргументировано тем, что больше кластеров выделить для наших данных просто бессмысленно (эмпирическое наблюдение)

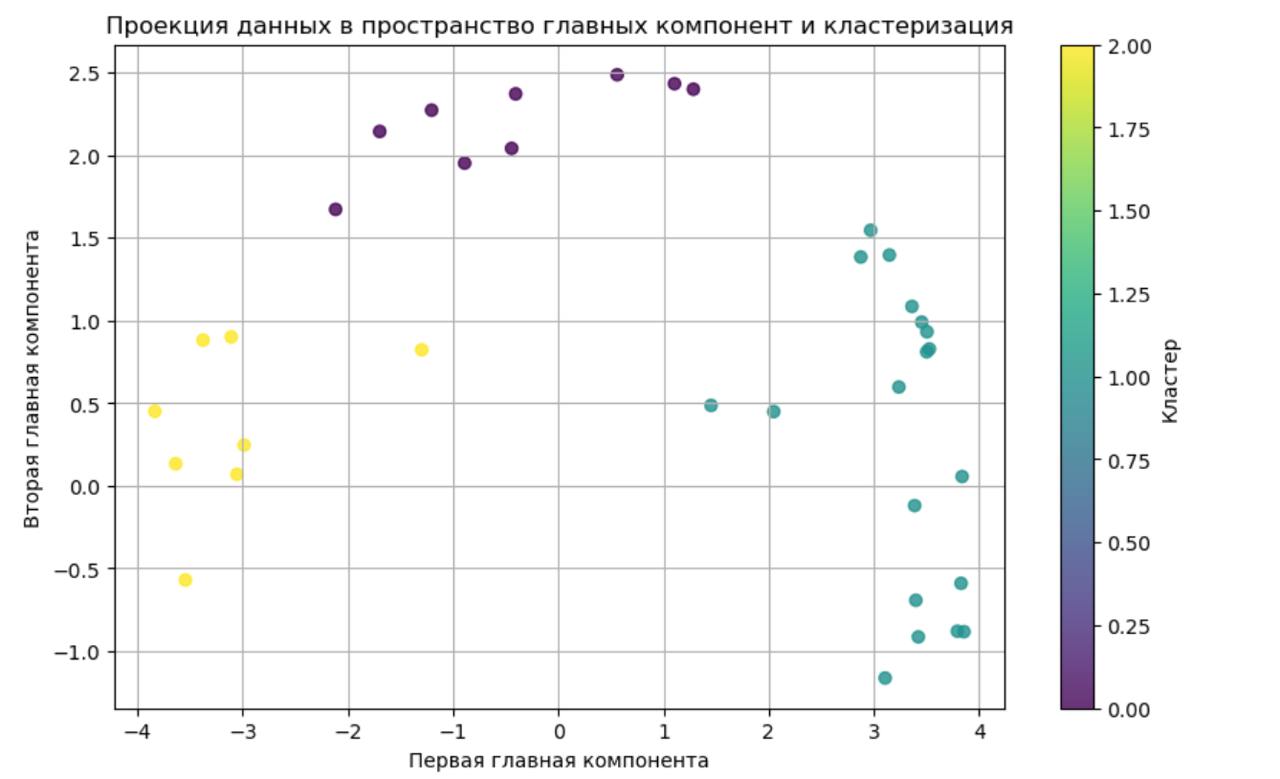


Рис. 30: Кластеризация компаний

В данном пункте проделана довольно интересная и немаловажная работа. Для более представительной выборки данные методы позволят пронаблюдать и проанализировать наглядное поведение компаний и кредитных рейтингов (и не только). В нашем случае все эти компании и вправду поделились в таком соответствии, как мы их разбили на кластеры. Одни из них и вправду меняли свой рейтинг при понижении / повышении другой компании, а третья группа, чьё поведение является неоднозначным и кто переходил за короткий промежуток времени в соседний рейтинг и обратно.

# Веб-интерфейс (Streamlit)

Для быстрого взаимодействия без необходимости постоянно менять код была придумана идея реализовать веб-интерфейс, чтобы каждый желающий мог использовать его для расчётов не только кредитных рейтингов, а так же любых других данных. Для этого был выбран веб-фреймворк streamlit, который имеет минималистичный веб дизайн и не требует знаний веб разработки для его реализации. Достаточно ознакомиться с базовой документацией [20] и начать реализацию своего собственного сервиса.

Каждая новая функция сразу же внедряется непосредственно в этот веб сервис. Так же данный сервис позволяет гибко использовать кэширование ресурсоемких вычислений при помощи декоратора @st.cache\_data.

Пример некоторых функций данного веб сервиса (Рис. 31, 32, 33)

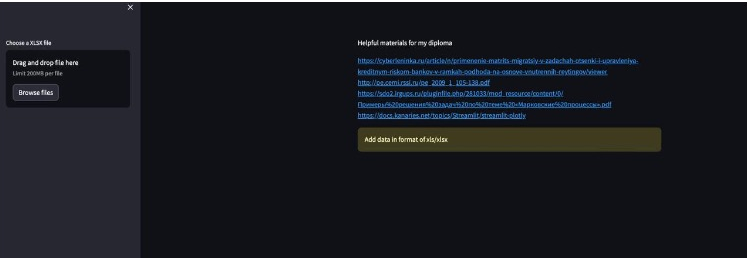


Рис. 31: Начальная страница

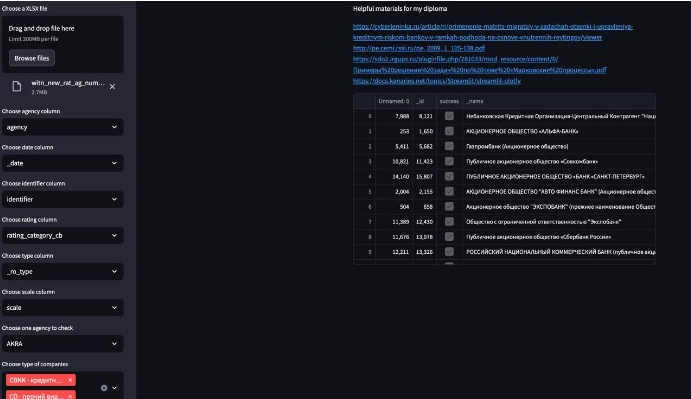


Рис. 32: Настройка параметров

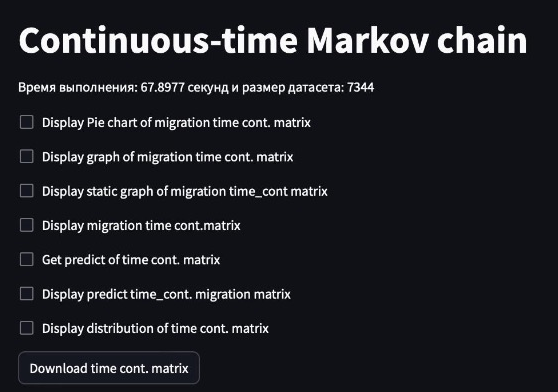


Рис. 33: Дополнительный функционал расчёта

# Дальнейшие шаги

В дальнейшем планируется использовать дополнительные факторы для дообучения модели классификации, а также рассмотреть иные от случайного леса методы машинного обучения. Также планируется реализовать сравнительный анализ метода МСМС, так как он является довольно мощным и интересным инструментом для прогнозирования и анализа не только кредитных рейтингов, но и временных рядов в целом. Немаловажным является идея улучшения марковской цепи 2-ого порядка, так как в ней заметен потенциал, однако раскрыть его полностью не удалось.

Хочется отметить, что существует достаточно много скрытых факторах в таких данных, поэтому интересно посмотреть на то, как ведут себя скрытые марковские модели (HMM).

# Заключение

Основное достижения данного проекта состоит в разработке алгоритмов предсказания типа карточного клиента на несколько будущих периодов вперед. Алгоритмы были основа- ны на случайном и марковской модели. Ключевой особенностью проекта является изучение и предсказание всех состояний кредитных рейтингов, в то время как на основе приведенного анализа литературы видно, что работы в мире посвящены предсказанию лишь отдельных состояний.

В рамках работы были изучены и реализованы новые для российского рынка подходы, которые несут в себе большой потенциал. В рамках работы с марковской моделью была построена цепь Маркова 2-го порядка, которая местами демонстрирует улучшение результатов по сравнению с цепью 1 порядка. Стоит отметить, что при обогащении исходных данных, данная работа имеет потенциал в развитии методологии не только по прогнозированию, но и по классификации кредитных рейтингов на базе дополнительных факторов.

Исполнители данного проекта благодарят С.А. Аксенова и Л. А. Маниту за ценные идеи и советы по реализации.

# Приложение

[GitLab проекта](https://git.miem.hse.ru/dynamics_of_card_client_profile_migrations/dynamic) [Подробные результаты](https://docs.google.com/spreadsheets/d/1znplgD5iQXW32lsg25YMPOdwudMYUZvkKDO6f70v9k0/edit?usp=sharing) [Постер проекта](https://drive.google.com/file/d/1-zsF22ciJoE-VIEM_jsmJT-O36p4NY9H/view)

# Список литературы

1. The Comprehensive R Archive Network, msm: Multi-state markov and hidden markov models in continuous time. <https://cran.r-project.org/web/packages/msm/>. Accessed: 2023-01-01.
2. D. AL-Najjar, N. Al-Rousan, and H. AL-Najjar. Machine learning to develop credit card customer churn prediction. *Journal of Theoretical and Appllied Electronic Commerce Research*, 2022.
3. G. I. W. Claude Sammut. *Encyclopedia of Machine Learning*. 2010.
4. D. R. Cox and H. D. Miller. *The Theory of Stochastic Process*. Chapman & Hall/CRC, 1965.
5. J. H. Friedman. Stochastic gradient boosting. 1999.
6. N. G., Y. Chen, L. Zhanga, and Y. Guo. Credit card customer analysis based on panel data clustering. *Elsevier Ltd*, 2012.
7. C. T. Ganesh Ramachandra Kini. Phase transitions for one-vs-one and one-vs-all linear separability in multiclass gaussian mixtures. *University of California, Santa Barbara*.
8. E. Koˇcenda and M. Vojtek. Default predictors and credit scoring models for retail banking.

*CESIFO WORKING PAPER*, 2009.

1. G. Marshall and R. H. Jones. Multi-state models and diabetic retinopathy. *Stat Med*, 1995.
2. G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 2011.
3. A. Rindos, S. Woolet, I. Viniotis, and K. Trivedi. Exact methods for the transient analysis of nonhomogeneous continuous time markov chains. pages 121–133. Springer, Boston, MA, 1995.
4. D. R´egis and R. Artes. Using multi-state markov models to identify credit card risk.

*Production*, 2016.

1. S. Sharma. Classification and regression trees: The use and significance of trees in analytics.

*Journal on Recent Innovation in Cloud Computing, Virtualization Web Applications*, 2022.

1. K. S. Trivedi. *Probability and statistics with reliability, queuing, and computer science applications*. Wiley, 2016.
2. K. S. Trivedi, A. Bobbio, and J. Muppala. *Reliability and Availability Engineering*. Cambridge University Press, 2017.
3. A. M. V.V. Zhuravleva. Simplified silhouette parameter for assessing the quality of cluster structure. *ltai State University (Barnaul, Russia)*.
4. I. Yeh and C. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 2009.
5. H. W. Youguo Li. A clustering method based on k-means algorithm. *Physics Procedia*.
6. D. Zakrzewska and J. Murlewski. Clustering algorithms for bank customer segmentation.

*5th International Conference on Intelligent Systems Design and Applicationss*, 2005.

1. А.В.Булинский and А.Н.Ширяев. *Теория случайных процессов*. ФИЗМАТЛИТ, 2005.
2. Б.П.Демидович. *Лекции по математической теории устойчивости*. Наука, 1967.
3. Е.А.Соколов. *Основы машинного обучения*. 2022.
4. СЕЛУКОВ Д.А. нахождение оптимального числа кластеров методом локтя. *Перм- ский национальный исследовательский политехнический университет*.