

**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

Добавлено примечание ([A1]): У одних слева, у других справа?

Москва - 2025

## **Содержание**

<b>1 Техническое задание</b>	<b>3</b>
1.1 Описание задачи	3
1.2 Цели, задачи	5
1.3 Планируемый и фактический результат	6
1.4 Ход работ, роли участников команды	6
<b>2 Анализ предметной области и обзор литературы</b>	<b>9</b>
<b>3 Подробное описание подходов, методов и математических моделей, использованных при реализации проекта</b>	<b>10</b>
3.1 Марковская модель	10
3.1.1 Однородные марковские цепи	11
3.1.2 Поведение кредитных рейтингов как марковская цепь	12
3.1.3 Применение априорного распределения Дирихле	15
3.1.4 Обучение марковской модели	16
3.2 Модель классификации	17
3.2.1 Случайный лес (Random Forest)	19
3.2.2 Прогноз кредитного рейтинга как задача классификации	20
3.3 Спектральные характеристики	21
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 [11]. 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.	21
3.3.1 Параметр Херста	21
3.3.2 Матрица кросс-корреляции	22
3.4 Марковская цепь Монте-Карло (MCMC)	23
<b>4 Полученные результаты</b>	<b>23</b>
4.1 Данные	24
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.	24
4.2 Результаты baseline-алгоритмов	26
4.3 Результаты марковской модели	28
4.4 Результаты модели классификации	32
4.5 Сравнение качества прогнозов	32

4.6 Применение МСМС	35
4.7 Расчёт спектральных характеристик и матрицы кросс-корреляции	39
<b>5 Веб-интерфейс (Streamlit)</b>	<b>43</b>
<b>6 Дальнейшие шаги</b>	<b>44</b>
<b>7 Заключение</b>	<b>44</b>
<b>8 Приложение</b>	<b>45</b>
<b>Список литературы</b>	<b>45</b>
<b>1</b>	

# 1 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.

## 1.1 Описание задачи

Рассмотрим профиль клиента со стороны банка-консультанта, предоставляющего данные. Внутри банка существуют 2 уровня описания состояний, в которых может находиться клиент: агрегированные состояния

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-C	Финансовые обязательства находятся под угрозой невыполнения. Компании находятся в крайне тяжёлом положении.
D	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

CCC Высокая вероятность дефолта, серьёзные проблемы с выплатами

CC Дефолт практически неизбежен

C 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 обслуживания (такие наблюдения не несут стат. характера и пропускают в прогнозировании), или ушел в дефолт.

Для подсчета 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 некоторый период *вперёд*. Например, зная цепочку на конец 3-го месяца

## 1.2 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

Добавлено примечание ([A2]): Таблица ушла влево

Добавлено примечание ([A3]): После номера точки, а текст с маленькой буквы – это нормально?

### 1.3 Планируемый и фактический результат

**Планируемый результат.** Методология (алгоритм) для прогнозирования кредитного рейтинга на несколько периодов, а именно: произведен 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, а точнее: извлечение данных из открытых источников (открытый репозиторий Банка России), обработка данных, сбор описательных статистик, ресемплинг данных с шагом 1 месяц, моделирование марковских цепей 1-ого и 2-ого порядка, построение доверительных интервалов методом bootstrap, вычисление спектральных характеристик и матрицы кросс-корреляции, построение модели классификации методом машинного обучения, вычисление апостериорных распределений методом МСМС, реализация веб-интерфейса и его адаптации к новым функциям.

### 1.4 Ход работ, роли участников команды

Ход работы представлен в виде Таблицы 1.

Добавлено примечание ([A4]):

Добавлено примечание ([A5]): Методология, а потом в прошедшем времени. Наверное лучше в настоящем... выполнение, подсчет

Добавлено примечание ([A6]): Не нравится мне такая формула результата. Лучше написать что сделано, получено, чем что недоделано

Добавлено примечание ([A7]):

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 2. Спектральные характеристики 3. 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	Полноценная реализация модели классификации - случайного леса, <b>реализация</b> 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 <b>классификация</b>	Не завершено	Дополнительные факторы будут <b>получены</b> по мере окончание реализации парсера открытых <b>источников</b>
14	Сравнительный анализ для метода MCMC	Не начато	

## **2 Анализ предметной области и обзор литературы**

Все чаще люди из совершенно разных областей прибегают к математическим моделям для решения своих задач. Кредитные рейтинговые агентства (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 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 заёмщиков 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 учётом 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 объёмом 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.

### **3 Подробное описание подходов, методов и математических моделей, использованных при реализации проекта**

#### **3.1 Марковская модель**

Прежде чем перейти к непосредственному описанию использованной модели, приведём ряд основных определений и результатов, полученных для марковских процессов. Большинство излагаемой ниже теории можно найти в работе [5], [6], [15].

Добавлено примечание ([A8]): Надо указывать источники

### 3.1.1 Однородные марковские цепи

На вероятностном пространстве  $(\Omega, \mathcal{F}, P)$  случайный процесс  $\{X_t, t \in T\}$  с дискретным множеством состояний  $S$  называется *цепью Маркова*, если для любого  $n \in \mathbb{N}$  для произвольных моментов времени  $s_1 < s_2 < s_3 < \dots < s_n < s < t$  и произвольных состояний  $i_1, i_2, \dots, i_j, i, j \in S$

$$P(X_t = j | X_{s_n} = i_n, X_{s_{n-1}} = i_{n-1}, \dots, X_{s_1} = i_1) = P(X_t = j | X_s = i)$$

Переходными вероятностями марковской цепи называются функции

$$p_{i,j}(s, t) = P(X_t = j | X_s = i), \quad \text{где } s \leq t (s, t \in T), i, j \in S$$

Они удовлетворяют следующим условиям:

$$1) p_{i,j}(s, t) \geq 0, \forall i, j \in S$$

$$2) \sum p_{i,j}(s, t) = 1, \forall i \in S$$

$$3) \forall i, j \in S \text{ и } \forall s \leq u \leq t$$

$$p_{i,j}(s, t) = \sum_{k \in S_u} p_{i,k}(s, u) p_{k,j}(u, t) \quad (1)$$

Если  $p_{i,j}(s, t)$  зависят лишь от разности  $t - s$ , то цепь называется *однородной*. В этом случае  $p_{i,j}(s + s + u) = p_{i,j}(u)$ . А равенство (1) запишется как

$$p_{i,j}(s, t) = \sum_{k \in S_u} p_{i,k}(s) p_{k,j}(t)$$

или, в матричной форме,  $P(s + t) = P(s)P(t)$ , где  $P(t)$  – матрица, состоящая из элементов  $p_{i,j}(t)$ . Заметим, что если  $T$  – дискретно, то  $p_{i,j}(n)$  есть в точности элемент  $(P^n)_{i,j}$  где  $P^n$  –  $n$ -я степень матрицы  $P(I)$ . 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  $p_{i,j}(t)$ , вводится понятие *инфinitезимальной* матрицы  $Q$  (или как её называют в работе [3] – матрицы *генератор*) и используется результат А.Н.Колмогорова. А именно, если марковская цепь *стандартна*<sup>1</sup>, то существует правая производная

$$\frac{d^+ P(t)}{dt} = Q$$

$$\lim_{t \rightarrow 0^+} \frac{p_{i,j}(t) - p_{i,j}(0)}{t} = q_{i,j}$$

Добавлено примечание ([A9]): Падеж матрицы!

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

<sup>1</sup>  $P(t) \rightarrow I$  при  $t \rightarrow 0^+$ , где  $I$  – единичная матрица

вероятности за время  $t$

$$P'(t) = P(t)Q$$

формальное решение которых при  $P(0) = I$ , выглядит следующим образом:

$$P(t) = \exp(Qt) = \sum_{r=0}^{\infty} Q^r \frac{t^r}{r!}$$

Действительно, почленно продифференцировав решение по  $t$ , получим:

$$\frac{d}{dt} e^{Qt} = \sum_{r=1}^{\infty} \frac{Q^r t^{r-1}}{(r-1)!} = Q e^{Qt} = e^{Qt} Q$$

Полезно сразу записать прямую систему уравнений Колмогорова в следующем виде:

$$\pi'(t) = \pi(t)Q, \quad (2)$$

где  $\pi(t)$  — вектор — строка из  $\pi_i(t) = P(Xt = i)$ .

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

$$\widehat{\lambda}_{i,j} = \frac{N_{i,j}(T)}{\int_0^T Y_i(s) ds}$$

$\widehat{\lambda}_{i,j}$  — оценка интенсивности перехода из  $i$  в  $j$

$N_{i,j}(T)$  — количество наблюдаемых переходов из  $i$  в  $j$  на интервале времени  $[0, T]$

$Y_i(s)$  — 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.

### 3.1.2 Поведение кредитных рейтингов как марковская цепь

Согласно процессам, описанным в множестве литературы на данную тему, кредитный рейтинг может находиться в сгруппированных состояниях, а также в детализированных, как было описано в таблице. 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. (взят из презентации на представлении проекта)

Добавлено примечание ([A10]): Поведение как марковская цепь — не звучит

Добавлено примечание ([A11R10]):

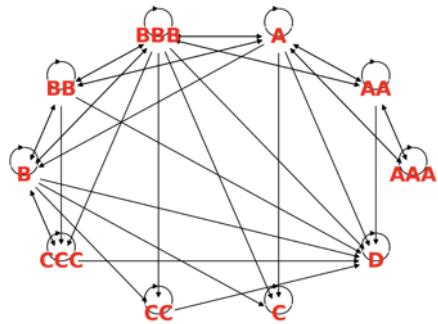


Рис. 1: Граф переходов состояний кредитного рейтинга

Рассмотрим простой пример из [3]. Пусть на момент времени  $t = 0$  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 за 1 год. Если рассматривать частоты переходов, то мы получим просто матрицу вида:

Добавлено примечание ([A12]): За какое время?

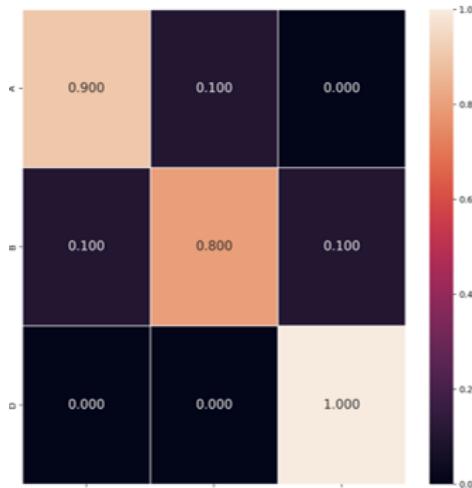


Рис. 2: Термальная карта матрицы переходных вероятностей дискретного подхода

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

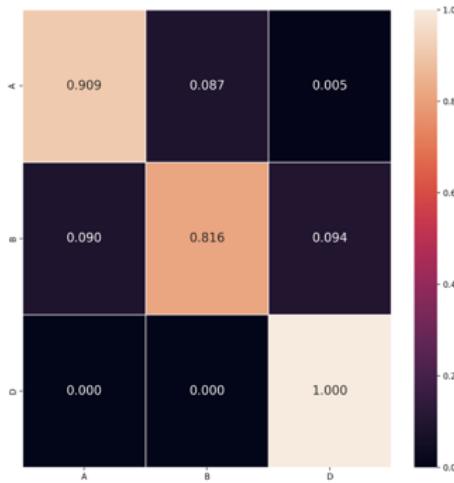


Рис. 3: Термальная карта матрицы переходных вероятностей непрерывного подхода

На данном этапе мы полагаем, что цепь 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 рассматривать процесс как на цепь Маркова 2 порядка — процесс, удовлетворяющий условию

$$P(X_t = j | X_s = i, X_{s_n} = i_n, \dots, X_{s_1} = i_1) = P(X_t = j | X_s = i, X_{s_n} = i_n)$$

для любого  $n \in \mathbb{N}$ , для произвольных моментов времени  $s_1 < s_2 < s_3 < \dots < s_i < s < t$  и произвольных состояний  $i_1, i_2, \dots, i_n, i, j \in S$ . Другими словами, цепь Маркова 2-го порядка будет обладать «памятью», и будущее состояние будет зависеть от настоящего и одного прошлого состояния. Чтобы реализовать это, в качестве пространства состояний  $S'$  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:

$$S' = \{AA, AB, AD, BA, BB, BD, DA, DB, DD\}$$

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.

### 3.1.3 Применение априорного распределения Дирихле

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

- При малом числе наблюдений  $N_i$ , оценки нестабильны;
- Если  $N_{i,j} = 0 \rightarrow p_{i,j} = 0$ , что может привести к вырожденности модели (например, при прогнозе  $P^n$ );
- 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 [7] 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 учётом наблюдаемых данных. В работе [8] изучается время расчёта оценки матрицы переходных вероятностей для байесовского подхода и подхода с непрерывным временем.

$$p_i = (p_{i,1}, p_{i,2}, \dots, p_{i,K})$$

$$p_i \sim Dir(a_{i,1}, a_{i,2}, \dots, a_{i,K})$$

где

$a_{i,j} > 0$  - априорные параметры

$a_{i,j} = 1$  - равномерное распределение (удобно использовать без экспертных оценок)

После наблюдения  $N_{i,j}$  переходов, апостериорное распределение так же будет Дирихле:

$$\widehat{p_{i,j}} = E[p_{i,j}|N_{i,j}] = \frac{N_{i,j} + a_{i,j}}{\sum_{k=1}^K (N_{i,k} + a_{i,k})}$$

Что позволяет:

- гибко задавать априорные веса
- переходит в частотную оценку при  $a_{i,j} \rightarrow 0$ .
- не равна нулю, если  $a_{i,j} \neq 0$

Возвращаясь к предыдущей задачи из п.3.1.2 матрица переходных вероятностей будет иметь вид:

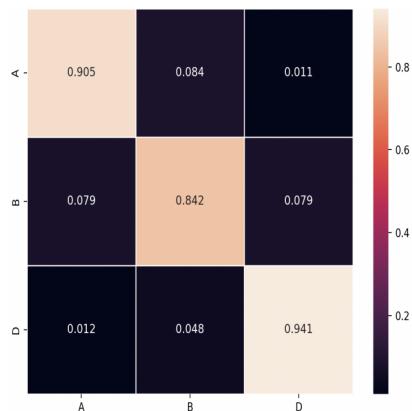


Рис. 4: Тепловая карта  
матрицы переходных вероятностей  
Байесовского подхода

### 3.1.4 Обучение марковской модели

Опишем принцип обучения и получения оценок в марковской модели для однородного случая. Как мы помним, для идентификации марковской цепи достаточно задать семейство матриц  $\{P(T)\}_{t \in R_+}$ . Согласно уравнениям Колмогорова,  $P(t)$  можно получить взятием матричной экспоненты от инфинитезимальной матрицы  $Q$  со множителем  $t$ , поэтому задача сводится к оценке элементов этой матрицы  $q_{i,j}$ . В выбранном нами пакете для этого используется метод максимального правдоподобия.

Представим, что на вход модели подаётся  $M$  цепочек. Для каждой  $i$ -ой цепочки нам известны  $(t_{i,1}, t_{i,2}, \dots, t_{i,N})$  и состояния  $(x(t_{i,1}), x(t_{i,2}), \dots, x(t_{i,N}))$  наблюдаемые в эти моменты. Мы можем записать правдоподобие цепочки в общем виде :

$$L(Q) = \prod_{i=1}^M L_i(Q)$$

Отсюда же можно напрямую получить оценку интенсивностей перехода  $(q_{i,j})$

Остаётся решить задачу безусловной оптимизации:

$$L(Q) \rightarrow \max_Q$$

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

$$-\ln L(Q) \rightarrow \min_Q$$

Собственно, данная задача при дифференцировании по  $q_{i,j}$  и приравниванию к нулю получаем выражение для оценки компонент матрицы  $Q$ , согласно работе [3], [9], [16].

### 3.2 Модель классификации

Прежде чем перейти к описанию использованного алгоритма, приведём общую постановку задачи классификации. Основную часть повествования ниже можно найти в источниках [10], [13], [14].

Пусть имеется множество обучающих объектов  $X$ , относящихся некоторым образом к известным классам  $Y$ . Один объект  $x \in X$  относится ровно к одному классу  $y \in Y$ . Пусть также имеется множество контрольных объектов  $\bar{X}$ , у которых отнесение к классам  $Y$  неизвестно. Задачу классификации можно сформулировать как поиск отображения  $a$  из множества объектов  $X \cup \bar{X}$  в множество классов  $Y$ . То есть, польза заключается в возможности отнести к классу именно контрольные объекты.

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

Добавлено примечание ([A13]): а почему такой порядок? Не по возрастанию?

1. **Accurасу.** Она измеряет долю правильно классифицированных образцов по отношению ко всем классифицированным образцам.

$$Accuracy = \frac{\text{Количество правильно распознанных объектов}}{\text{Общее кол - во объектов}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Непоказательна, если существует явный дисбаланс классов. Например, перед нами стоит задача — предсказать будет ли следующим состоянием 1(Default) или 0(не приведёт 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).

$$Accuracy = 0.95 = 95\%$$

Может создать ложное впечатление, что модель "хороша",

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

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

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.

3. **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.

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

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 тяжёлом финансовом состоянии. Это важно для сохранения доверия к методологии агентства и минимизации ошибок первого рода при автоматизированной классификации кредитного риска.

4. **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.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

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

- (a) Если все объекты помечены положительно, то *Recall* будет высоким, а *Precision* страдает.
- (b) 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.

**5. Матрица ошибок** — это показатель успешности классификации, где классов два или более. В случае бинарной классификации это таблица с 4 различными комбинациями сочетаний прогнозируемых и фактических значений.

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

- (a) True Positive (истинно-положительное решение) – прогноз совпал с реальностью, результат положительный
- (b) False Negative (ложноотрицательное решение) – ошибка 2-го рода, модель предсказала отрицательный результат, но на самом деле он положительный
- (c) False Positive (ложноположительное решение) – ошибка 1-го рода, модель предсказала положительный результат, а на самом деле он отрицательный
- (d) True Negative (истинно-отрицательное решение) – прогноз совпал с реальностью, результат отрицательный

### 3.2.1 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:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N, x_i \in \mathbb{R}^d, y_i \in Y = \{y_1, y_2, \dots, y_K\}$$

где  $x_i$  - вектор признаков объекта (например, финансовых или макроэкономических характеристик компаний), а  $y_i$  - метка класса, соответствующая одному из  $K$  кредитных рейтингов.

2. Алгоритм строит ансамбль из  $T$  решающих деревьев:

$$\mathcal{F} = \{h_t(x)\}_{t=1}^T, h_t: \mathbb{R}^d \rightarrow Y, t \in \{1, 2, \dots, T\}$$

Каждое дерево  $h_t$  обучается на случайной подвыборке  $\mathcal{D}_t \subseteq \mathcal{D}$  полученной методом бутстрапа (случайной выборкой с возвращением).

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

$$\mathcal{F}_m \subset \{1, 2, \dots, d\}, |\mathcal{F}_m| = m \ll d$$

Выбирается оптимальное разбиение пространства:

$$(j^*, r^*) = \arg \min_{j \in \mathcal{F}_m, r \in R} I(S_{left}, S_{right})$$

$$S_{left} = \{(x_i y_i) \in S | x_i^j \leq r\}$$

$$S_{right} = \{(x_i y_i) \in S | x_i^j > r\}$$

где

$I$  - критерий информативности. Во многих задачах классификации рассматривается как индекс Джини.

$j^*$  - индекс выбранного признака

$r^*$  - порог, по которому мы делим множество объектов в текущем узле.

4. Предсказание ансамбля осуществляется через голосование деревьев:

$$f(x) = \arg \max_{c \in C} \left[ \frac{1}{T} \sum_{t=1}^T \mathbb{I}(h_t(x) = c) \right]$$

где

$\mathbb{I}(\cdot)$  - индикатор 1, если предсказание совпало с  $c$ , иначе 0.

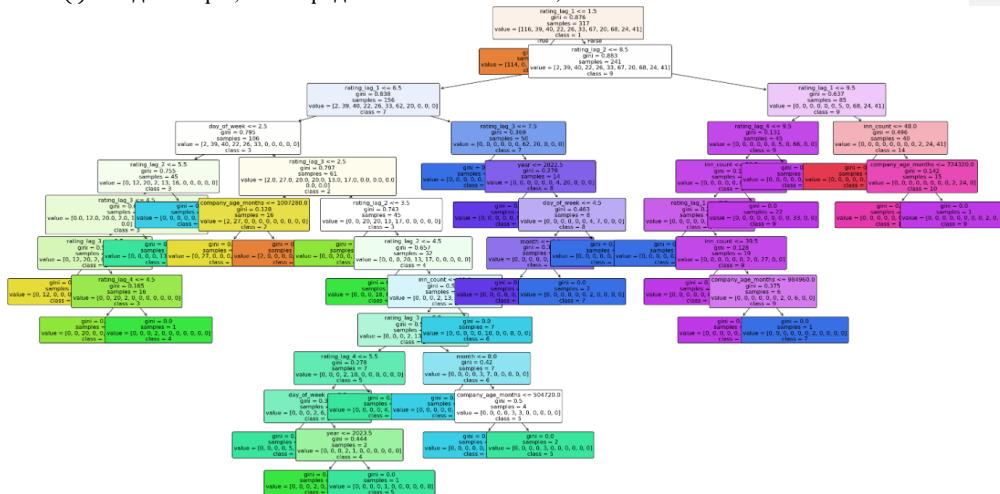


Рис. 5 :Пример дерева

### 3.2.2 Прогноз кредитного рейтинга как задача классификации

В качестве множества ответов  $Y$  выступает состояние из  $S$ . В обучающей выборке для

прогноза  $t$ -ого месяца каждой цепочке  $(s_1, \dots, s_{l-t})$  длины  $l$  ставится в соответствие пара  $(x_i, y_i)$ , где  $y_i$ — последнее известное состояния  $s_l$ ,  $x_i$ — вектор признаков, получающийся из информации о состояниях  $(s_1, \dots, s_{l-t})$ .

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 кредитных рейтингов за предыдущие 12 месяцев.
- Год - месяц - день

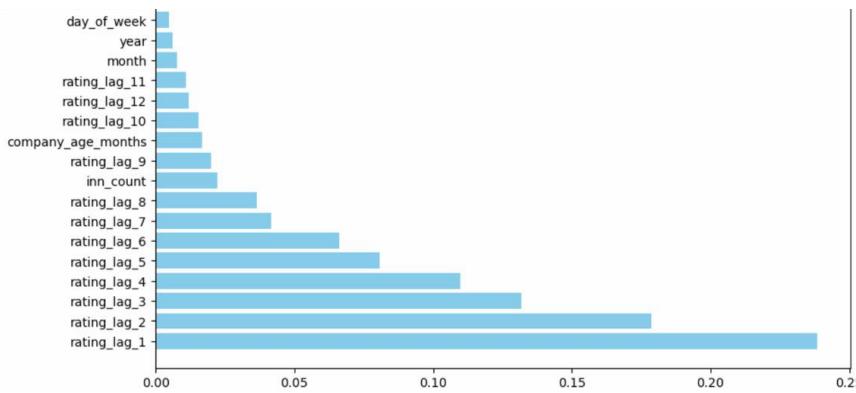


Рис. 6: Важность признаков после обучения

### 3.3 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 [11]. 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.

#### 3.3.1 Параметр Херста

Параметр Херста  $H \in (0,1)$ — 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 компаний.

Параметр Херста показывает:

- $H = 0.5$ : поведение похоже на случайное блуждание;

- $H > 0.5$ : персистентное поведение — рост (или падение) рейтинга, скорее всего, продолжится;
- $H < 0.5$ : антиперсистентное поведение — рост сменяется падением и наоборот.

### 1. Метод нормализованного размаха (R/S-анализ)

Пусть  $\{r_t\}_{t=1}^T$  — числовой временной ряд. Для каждого масштаба  $n, n < T$  временной ряд разбивается на  $\left[\frac{T}{n}\right]$  непересекающихся сегментов длины  $n$ . Для каждого сегмента вычисляется:

1. Среднее значение:  $r_{\bar{n}} = \frac{\sum_{t=1}^n r_t}{n}$
2. Кумулятивные отклонения:  $X_t = \sum_{k=1}^t r_k - r_{\bar{n}}$
3. Размах:  $R(n) = \max_{1 \leq t \leq n} X_t - \min_{1 \leq t \leq n} X_t$
4. Нормализованный размах:  $\frac{R(n)}{S(n)}$

Далее из уравнения находится параметр Херста:

$$\log\left(\frac{R(n)}{S(n)}\right) = H \cdot \log(n) + \log(c)$$

### 2. Оценка через спектральный наклон

Второй подход основан на анализе плотности спектральной мощности сигнала.

Предполагается, что спектр  $S(f)$  подчиняется степенному закону:

$$S(f) \sim \frac{1}{f^\beta}, f \rightarrow 0$$

где

$S(f)$  - мощность сигнала на частоте  $f$ ,

$\beta$  - наклон в логарифмической шкале.

Параметр Херста при этом выражается через наклон  $\beta$  следующим образом:

$$H = \frac{1 + \beta}{2}$$

#### 3.3.2 Матрица кросс-корреляции

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

$$C \in \mathbb{R}^{N \times N}$$

Где каждый элемент  $C_{i,j}$  определяется как нормированное ковариационное выражение:

$$C_{i,j} = \frac{1}{T} \sum_{t=1}^T \frac{(r_t^i - \bar{r}_i)(r_t^j - \bar{r}_j)}{\sigma_i^2 \sigma_j^2}$$

Данная матрица  $C$  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 С имеет вид:

$$C = U\Sigma V^T$$

где

$U = [u_1, u_2, \dots, u_N], V = [v_1, v_2, \dots, v_N]$  - ортонормированные матрицы собственных векторов.

$\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_N)$  - диагональная матрица сингулярных значений, упорядоченных по убыванию:  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_N \geq 0$

Каждое  $\sigma_k$  отражает вклад  $k$ -ой компоненты в общую структуру кросс-корреляций. Чем больше значение  $\sigma_1$ , тем ярче выражен единый (глобальный) фактор, влияющий на рейтинговое поведение разных компаний.

### 3.4 Марковская цепь Монте-Карло (МСМС)

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 (МСМС). В частности, теорию по данному методу можно найти в работе [12]

Пусть даны наблюдаемые данные  $\{r_t\}_{t=1}^T$  - кредитные объекты некоторого объекта длины  $T$ .

$\theta \in \Theta \subseteq \mathbb{R}^d$  - вектор параметров модели

$p(r_{1:T}|\theta)$  - функция правдоподобия

$p(\theta)$  - априорное распределение параметров

Задача: сгенерировать выборку из апостериорного распределения:

$$p(\theta|r_{1:T}) = \frac{p(r_{1:T}|\theta)p(\theta)}{p(r_{1:T})}$$

Вместо точного вычисления распределения  $p(\theta|r_{1:T})$  метод МСМС строит марковскую цепь  $\{\theta^n\}_{n=1}^N$ , стационарное распределение которой совпадает с нужным  $p(\theta|r_{1:T})$ .

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  $p_{i,j}$

## 4 Полученные результаты

## 4.1 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 7.

Все рейтинги, присвоенные объекту		
Значение кредитного рейтинга	Наименование КРА	Дата публикации кредитного рейтинга
ruAA-	АО "Эксперт РА"	17.12.2024
AA-(RU)	AKPA (AO)	19.04.2024
AA-(RU)	AKPA (AO)	18.05.2023
AA-(RU)	AKPA (AO)	14.03.2023
AA-(RU)	AKPA (AO)	27.05.2022
AA-(RU)	AKPA (AO)	07.06.2021
A+(RU)	AKPA (AO)	16.06.2020
A+(RU)	AKPA (AO)	24.06.2019
A+(RU)	AKPA (AO)	28.06.2018
A+(RU)	AKPA (AO)	28.06.2017

Рис. 7: Пример данных из открытого источника

Вид изначальных данных:

id	name_com	ro_type	identifier	identifier_type	is_in	rating	forecast	onal_idents	release	l_release	agency	status	onal_id	impany_id	comment	last_action	date	scale_
1	Закрытое	FLSG - ли	782549698ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
2	Закрытое	FINS - срп	77060532ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
3	Закрытое	FINS - срп	814084842ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
4	Акционер	BNFC - не	770985754ИИНН	A-ru		ООО "НКР					NW - перি ##### Кредит. р							
5	Акционер	BNFC - не	660800044ИИНН	ruAA		АО "Эксп					NW - пери ##### Кредит. р							
6	Авен-ир-2	Общество с ТБНД - офф	771851193ИИНН	A.III		АО "Эксп					АО "Эксп				AF - Подт ##### Кредит. р			
7	АВТОНОМ	CO - проп	213006522ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
8	АВТОНОМ	CO - проп	213006522ИИНН	ruA		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
9	АВТОНОМ	CO - проп	213006522ИИНН	ruA		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
10	АВТОНОМ	CO - проп	213006522ИИНН	A+III		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
11	АВТОНОМ	BNFC - не	773426665ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
12	АВТОНОМ	BNFC - не	773426665ИИНН	ruBBB		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
92	Акционер	FINS - срп	772853462ИИНН	ruBBB		АО "Эксп					АО "Эксп				WD - отзы ##### Кредит. р			
93	Акционер	FINS - срп	772853462ИИНН	ruBBB		АО "Эксп					АО "Эксп				WD - отзы ##### Кредит. р			
13	АВТОНОМ	FMFO - м	213006529ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
14	АВТОНОМ	FMFO - м	213006529ИИНН	ruBBB+		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
15	АВТОНОМ	FMFO - м	213006529ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
16	АВТОНОМ	FMFO - м	275066729ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
17	АВТОНОМ	FMFO - м	275066729ИИНН	Рейтинг о		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			
18	АВТОНОМ	FMFO - м	275066729ИИНН	A+.mfi		АО "Эксп					АО "Эксп				WD - Отзыв ##### Кредит. р			

Рис. 8: 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. 9 продемонстрировано распределение всех сфер с числовым и процентным содержанием наблюдений:



Рис. 9: Распределение всех наблюдений

А так же по каждому из КРА в отдельности:

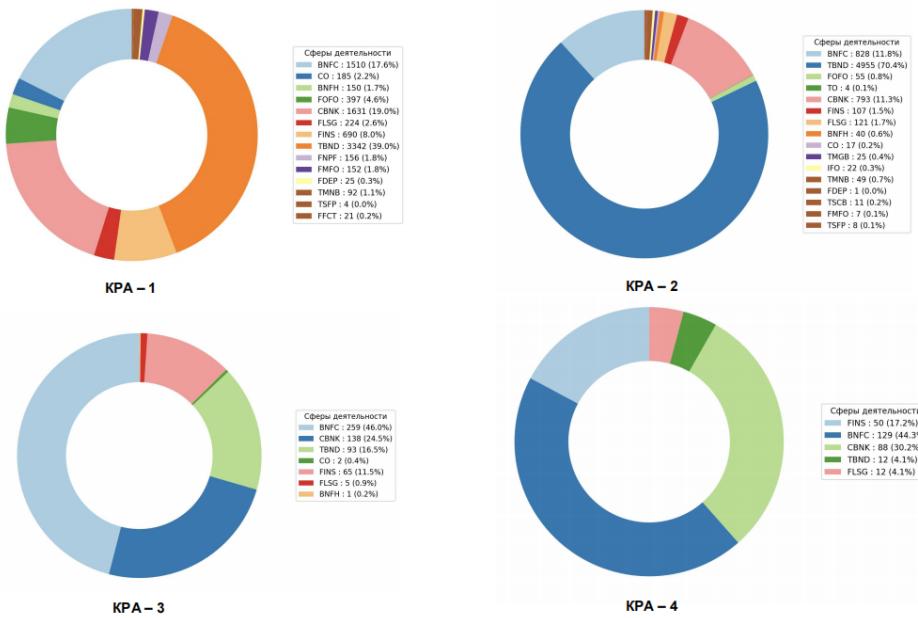


Рис. 10: Распределение всех наблюдений по каждому КРА

После ,так называемого, ресемплинга кол-во данных значительно увеличилось. В расчёт не шли какие либо облигации и сферы, чё обозначение начиналось с Т (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. 11:

agency	ro_type	resampled_dates	implied_waiting_lag	sting_lag	year	month	by_of_week	country	age_d										
Expert RA FINS - ctpi		31.12.2022	3	3	3	3	3	3	3	3	3	3	3	3	2022	12	5	77	2324160
Expert RA FINS - ctpi		31.01.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	1	1	77	2388800
Expert RA FINS - ctpi		28.02.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	2	1	77	2409120
Expert RA FINS - ctpi		31.03.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	3	4	77	2453760
Expert RA FINS - ctpi		30.04.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	4	6	77	2496960
Expert RA FINS - ctpi		31.05.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	5	2	77	2541600
Expert RA FINS - ctpi		30.06.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	6	4	77	2584800
Expert RA FINS - ctpi		31.07.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	7	0	77	2629440
Expert RA FINS - ctpi		31.08.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	8	3	77	2674080
Expert RA FINS - ctpi		30.09.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	9	5	77	2717280
Expert RA FINS - ctpi		31.10.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	10	1	77	2761920
Expert RA FINS - ctpi		30.11.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	11	3	77	2805120
Expert RA FINS - ctpi		31.12.2023	3	3	3	3	3	3	3	3	3	3	3	3	2023	12	6	77	2849760
Expert RA FINS - ctpi		31.01.2024	3	3	3	3	3	3	3	3	3	3	3	3	2024	1	2	77	2894400
Expert RA FINS - ctpi		29.02.2024	3	3	3	3	3	3	3	3	3	3	3	3	2024	2	3	77	2936160
Expert RA FINS - ctpi		31.03.2024	3	3	3	3	3	3	3	3	3	3	3	3	2024	3	6	77	2980800
Expert RA FINS - ctpi		30.04.2024	4	3	3	3	3	3	3	3	3	3	3	3	2024	4	1	77	3024000
Expert RA FINS - ctpi		31.05.2024	4	4	3	3	3	3	3	3	3	3	3	3	2024	5	4	77	3068640
Expert RA FINS - ctpi		30.06.2024	4	4	4	3	3	3	3	3	3	3	3	3	2024	6	6	77	3111840
Expert RA FINS - ctpi		31.07.2024	4	4	4	4	3	3	3	3	3	3	3	3	2024	7	2	77	3156480
Expert RA FINS - ctpi		31.08.2024	4	4	4	4	4	3	3	3	3	3	3	3	2024	8	5	77	3201120
Expert RA FINS - ctpi		30.09.2024	4	4	4	4	4	4	3	3	3	3	3	3	2024	9	0	77	3244320
Expert RA FINS - ctpi		31.10.2024	4	4	4	4	4	4	3	3	3	3	3	3	2024	10	3	77	3288960
Expert RA FINS - ctpi		30.11.2024	4	4	4	4	4	4	4	3	3	3	3	3	2024	11	5	77	3332160

Рис. 11: Пример данных для модели классификации

**О предобработке, делении выборки и метриках.** Для обработки цепочек перед обучением, деления на обучающие и тестовые данные использовался язык программирования Python (библиотеки numpy, pandas, sklearn). Для подсчета интересующих метрик и формирования прогнозов на  $N$  периодов вперед были написаны пользовательские функции. 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).

## 4.2 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  $A, B, A, B$  прогноз на 3 месяца вперед будет  $B, B, B$ , так как последнее известное состояние было  $B$ . Заметим, что при таком подходе мы никогда не будем прогнозировать состояния  $D$ , 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. 12:

Тип компл:	Прогноз	Accuragcy	F1-score
BNFC/BNF	12	50%	49,75%
BNFC/BNF	24	50,33%	50,01%
BNFC/BNF	36	53,71%	51,02%
BNFC/BNF	48	57,22%	53,83%
CBNK/CO	12	35,57%	31,94%
CBNK/CO	24	37,25%	33,32%
CBNK/CO	36	36,94%	33,98%
CBNK/CO	48	37,42%	34,31%
FOFO/FNF	12	43,28%	37,88%
FOFO/FNF	24	45,29%	39,32%
FOFO/FNF	36	46,62%	41,33%
FOFO/FNF	48	48,53%	42,50%

Рис. 12: Результаты наивного классификатора

### 4.3 Результаты марковской модели

Реализация модели, основанной на марковской цепи, происходила на языке программирования 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 13.

metric	average	agency	field	result	model	forecast	comment
f1_score	macro	CRA_1	BNFC - нефинансовая ко	80,04% TCMC	1_year		
f1_score	macro	CRA_1	BNFH - холдинговая ком	97,98% TCMC	1_year		
f1_score	macro	CRA_1	CBNK - кредитная органи	81,57% TCMC	1_year		
f1_score	macro	CRA_1	CO - прочий вид организ	89,88% TCMC	1_year		
f1_score	macro	CRA_1	FOFO - прочие финансов	90,53% TCMC	1_year		
f1_score	macro	CRA_1	FNPF - негосударственн	77,46% TCMC	1_year		
f1_score	macro	CRA_1	FINS - страховая организ	77,89% TCMC	1_year		
f1_score	macro	CRA_1	FMFO - микрофинансова	79,27% TCMC	1_year		
f1_score	macro	CRA_1	FLSG - лизинговая компа	84,35% TCMC	1_year		
f1_score	macro	CRA_1	FDEP - депозитарий	100,00% TCMC	1_year	length = 5 (co	
f1_score	macro	CRA_1	FFCT - юридическое лицо	100,00% TCMC	1_year	length = 4 (co	
f1_score	macro	CRA_1	BNFC - нефинансовая ко	72,16% Dirichlet	1_year		
f1_score	macro	CRA_1	BNFH - холдинговая ком	46,85% Dirichlet	1_year		
f1_score	macro	CRA_1	CBNK - кредитная организ	74,63% Dirichlet	1_year		
f1_score	macro	CRA_1	CO - прочий вид организ	60,04% Dirichlet	1_year		
f1_score	macro	CRA_1	FOFO - прочие финансов	62,18% Dirichlet	1_year		
f1_score	macro	CRA_1	FNPF - негосударственн	61,50% Dirichlet	1_year		
f1_score	macro	CRA_1	FINS - страховая организ	62,66% Dirichlet	1_year		
f1_score	macro	CRA_1	FMFO - микрофинансова	52,93% Dirichlet	1_year		
f1_score	macro	CRA_1	FLSG - лизинговая компа	41,37% Dirichlet	1_year		
f1_score	macro	CRA_1	FDEP - депозитарий	44,44% Dirichlet	1_year	length = 5 (co	
f1_score	macro	CRA_1	FFCT - юридическое лицо	0,00% Dirichlet	1_year	length = 4 (co	
f1_score	macro	CRA_1	BNFC - нефинансовая ко	79,24% Second_order	1_year		
f1_score	macro	CRA_1	BNFH - холдинговая ком	84,48% Second_order	1_year		
f1_score	macro	CRA_1	CBNK - кредитная организ	74,67% Second_order	1_year		
f1_score	macro	CRA_1	CO - прочий вид организ	96,81% Second_order	1_year		
f1_score	macro	CRA_1	FOFO - прочие финансов	85,28% Second_order	1_year		
f1_score	macro	CRA_1	FNPF - негосударственн	72,00% Second_order	1_year		
f1_score	macro	CRA_1	FINS - страховая организ	87,98% Second_order	1_year		
f1_score	macro	CRA_1	FMFO - микрофинансова	75,31% Second_order	1_year		
f1_score	macro	CRA_1	FLSG - лизинговая компа	81,51% Second_order	1_year		
f1_score	macro	CRA_1	FDEP - депозитарий	100,00% Second_order	1_year	length = 5 (co	
f1_score	macro	CRA_1	FFCT - юридическое лицо	100,00% Second_order	1_year	length = 4 (co	

Рис. 13: 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. 14, 15 и 16 для каждой марковской цепи (здесь состояния из укрупненной группы):

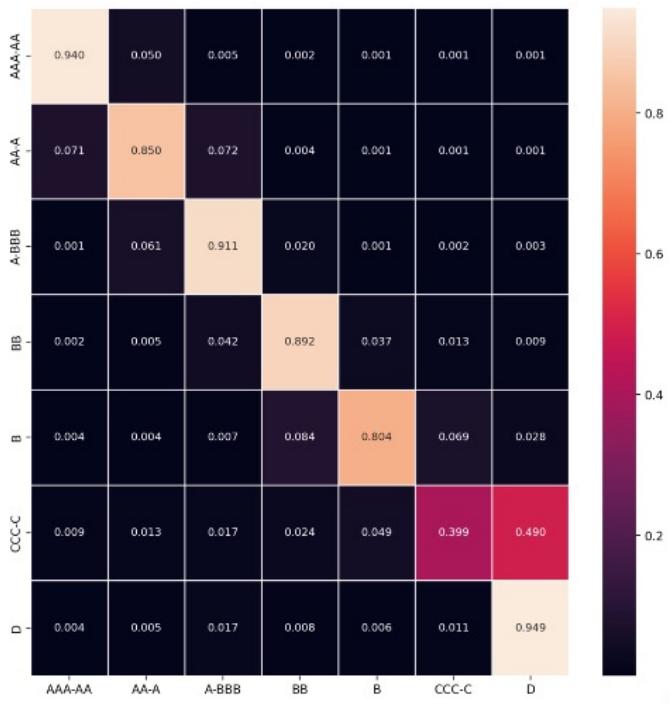


Рис. 14: Тепловая карта (heatmap) марковской цепи 1-ого порядка с априорным распределением Дирихле



Рис. 15: Тепловая карта (heatmap) марковской цепи 1-ого порядка  
с непрерывным временем

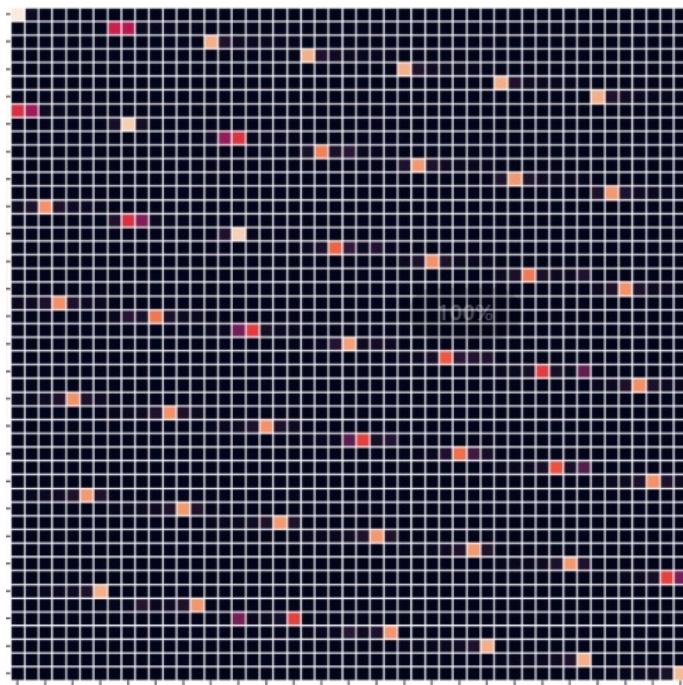


Рис. 16: Тепловая карта (heatmap) марковской цепи 2-ого порядка

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.

#### 4.4 Результаты модели классификации

Аналогичный расчет проведем для модели классификации - случайного леса. На Рис. 17 представлены результаты расчета матрики *F-measure*:

f1_score	macro	CRA_1	BNFC - нефт	91,91%	Random_fc1_year
f1_score	macro	CRA_1	BNFH - хол	85,26%	Random_fc1_year
f1_score	macro	CRA_1	CBNK - кре	95,89%	Random_fc1_year
f1_score	macro	CRA_1	CO - прочи	85,11%	Random_fc1_year
f1_score	macro	CRA_1	FLSG - лизи	96,07%	Random_fc1_year
f1_score	macro	CRA_1	FOFO - про	93,11%	Random_fc1_year
f1_score	macro	CRA_1	FINS - стра	92,48%	Random_fc1_year
f1_score	macro	CRA_1	FMFO - ми	71,43%	Random_fc1_year
f1_score	macro	CRA_1	FFCT - юри	48,39%	Random_fc1_year
f1_score	macro	CRA_1	FDEP - деп	100,00%	Random_fc1_year

Рис. 17: 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 крупненные группы, то средний результат по *F-measure* и *accuracy* изменится 88,91% и 95,45% соответственно. Мы не рассматривает *recall* и *precision*, т.к. ранее было упомянуто, 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 деятельности, КРА.

#### 4.5 Сравнение качества прогнозов

Подробные результаты сравнения качества прогнозов наивного классификатора, случайного леса и марковских моделей представлены на Рис. 18. Метрики выбраны те же, что и рассматривались ранее (*accuracy* / *F-measure*). Результаты приведём усреднённые по всем КРА и областям, а также в Приложении будут продемонстрированы результаты отдельно по каждому КРА и сфере деятельности.

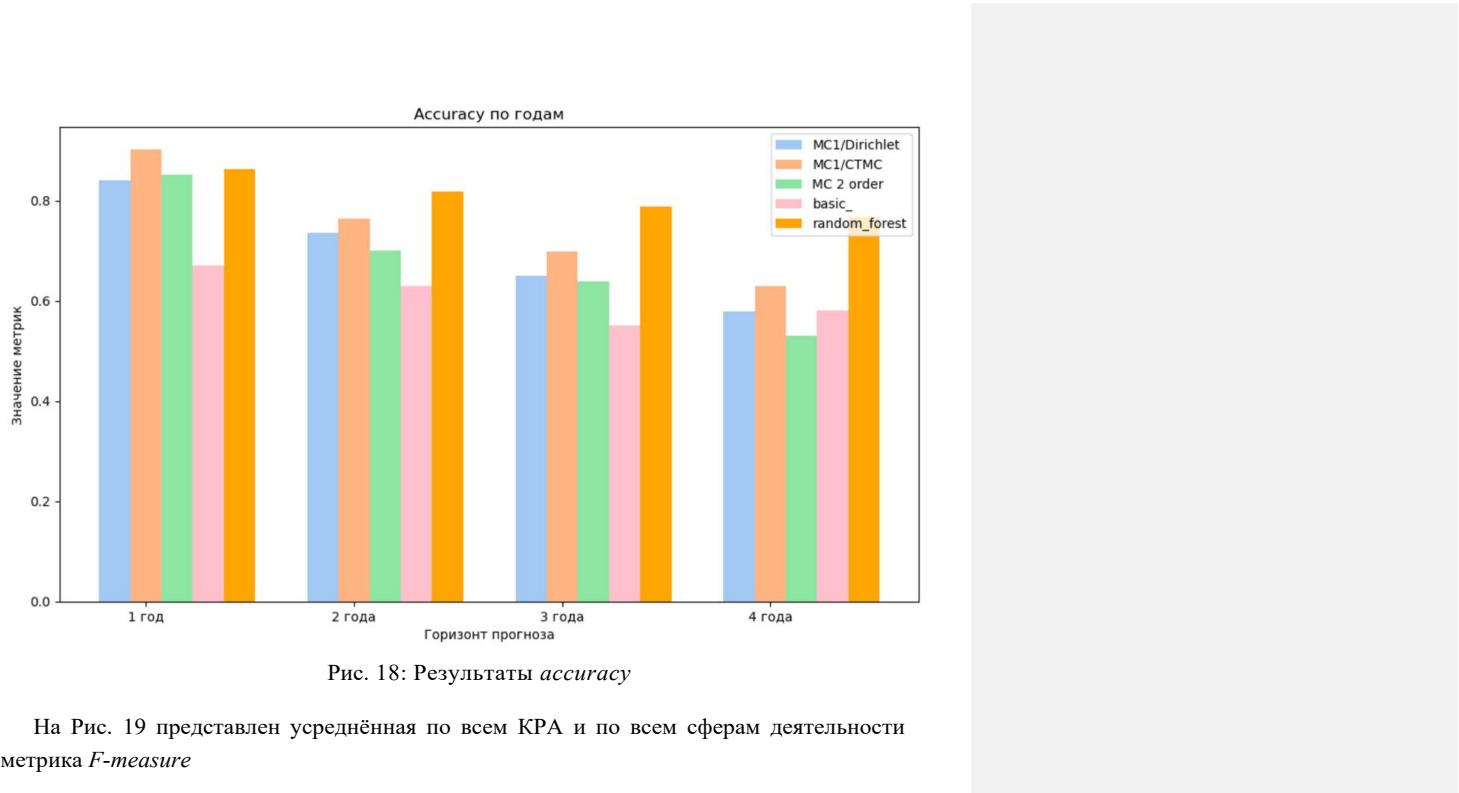


Рис. 18: Результаты *accuracy*

На Рис. 19 представлен усреднённая по всем КРА и по всем сферам деятельности метрика *F-measure*

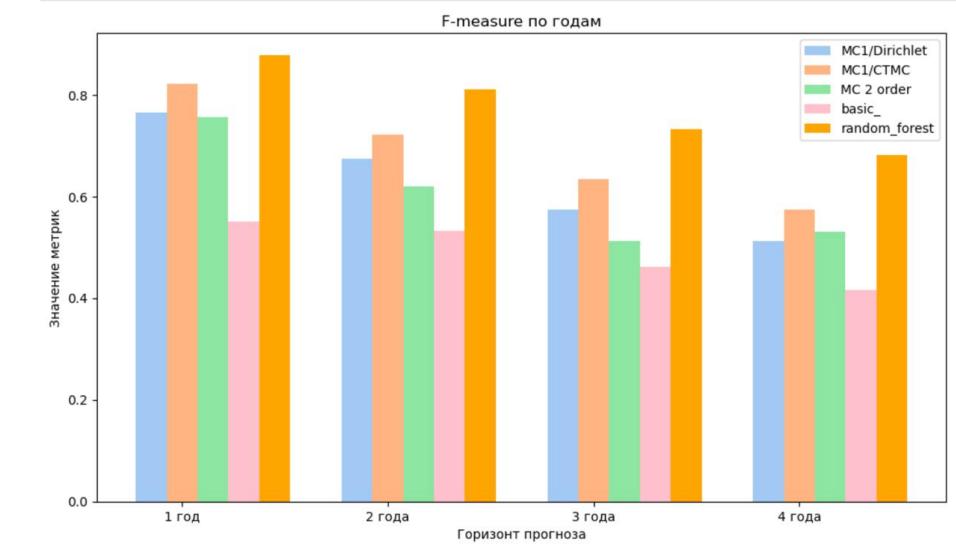


Рис. 19: Результаты *F-measure*

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 исключением 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. 20 продемонстрирована таблица относительной «деградации» (Relative Degradation) методов прогнозирования.

Метод	1 год	4 год	?_abs	?_rel (%)
Dirichlet	0.7655	0.5124	0.2531	33.07%
СТМС	0.8228	0.5753	0.2475	30.09%
2-й порядок	0.7566	0.5304	0.2262	29.89%
Basic	0.5506	0.4165	0.1341	24.35%
Random Forest	0.8779	0.6814	0.1965	22.38%

Рис. 20: Результаты относительной «деградации» способности прогнозирования существующих методов

- Dirichlet и СТМС 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 идеи, она должна показывать примерно схожую ситуацию (поведение), что и *F-measure*.

Метод	1 год	4 год	?_abs	?_rel (%)
Dirichlet	0.8406	0.5795	0.2611	31.06%
CTMC	0.9015	0.6296	0.2719	30.17%
2-й порядок	0.8520	0.5304	0.3216	37.73%
Basic	0.6706	0.5817	0.0889	13.26%
Random Forest	0.8628	0.7676	0.0952	11.04%

Рис. 21: Результаты относительной «деградации» способности прогнозирования 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.

#### 4.6 Применение MCMC

Немало важным следует отметить применение MCMC в данной задаче. Данный метод позволил посмотреть насколько сильно наши кредитные рейтинги «разбрасываются» и посмотреть на их распределение. На Рис. 22, 23, 24, 25 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)

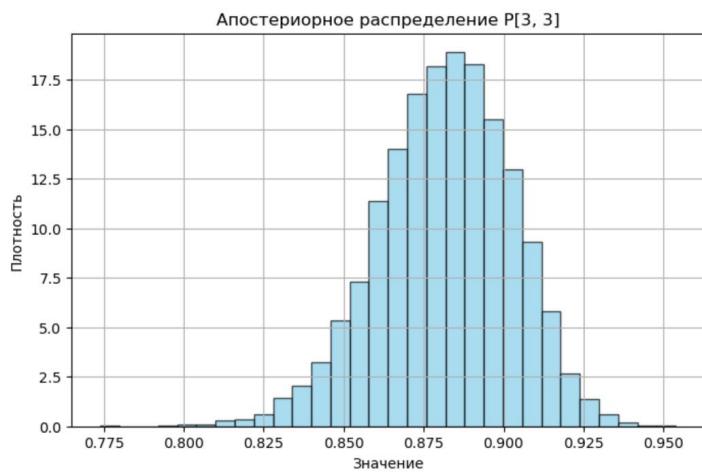


Рис. 22: Распределение вероятности перехода из 4 состояния в 4

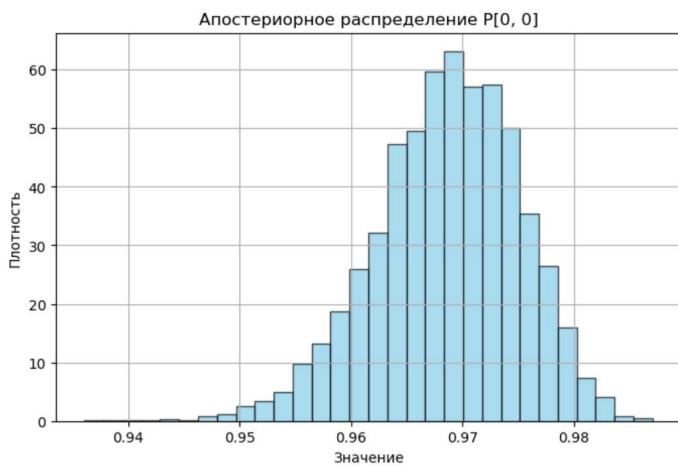


Рис. 23: Распределение вероятности перехода из 1 состояния в 1

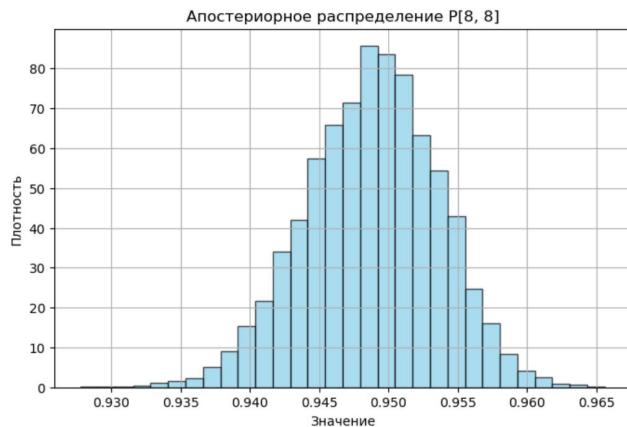


Рис. 24: Распределение вероятности перехода из 9 состояния в 9

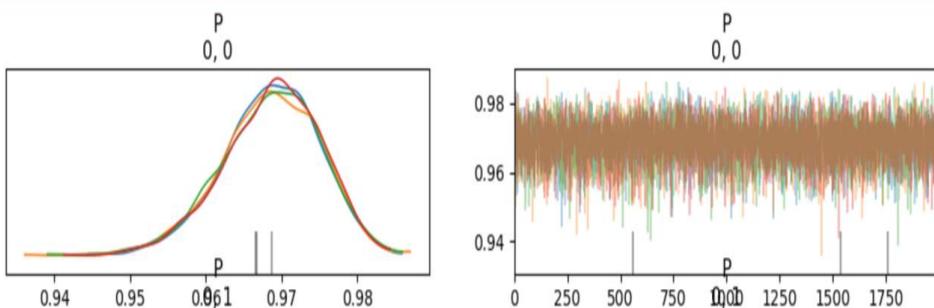


Рис. 25: Сэмплирование вероятностей перехода из 1 в 1 состояние  
при разных «эпохах»

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-го порядка. Однако проверить сходимость - задача на будущее.

На данном этапе реализовано сравнение доверительного интервала методом MCMC и bootstrap. В общем случае при большом кол-ве сэмплов матрица, полученная методом MCMC, 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. 26 и Рис. 27 продемонстрированы соответствующие вероятности:

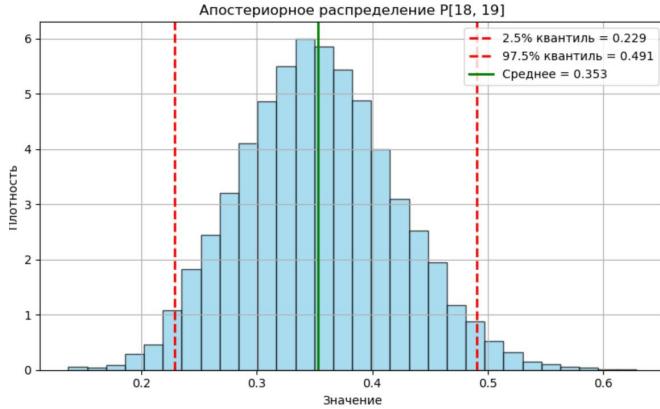


Рис. 26: Распределение вероятности перехода методом МСМС

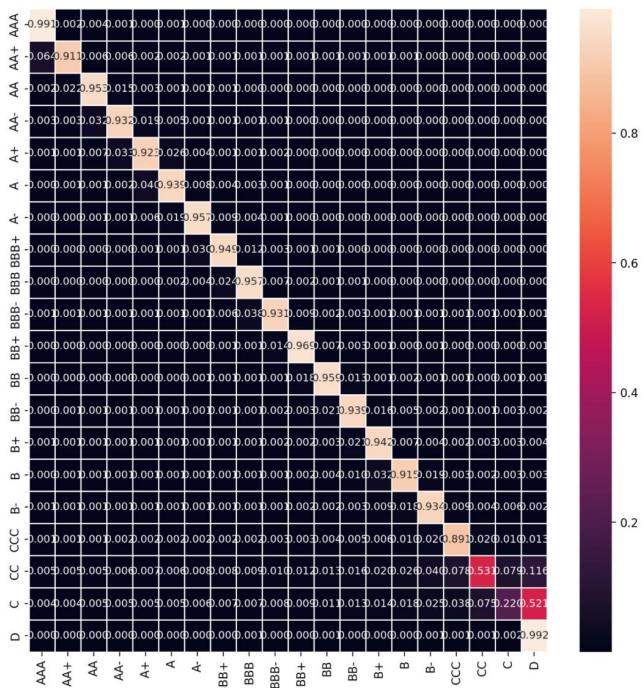


Рис. 27: Термальная карта марковской цепи 1-ого порядка

Соответствующий доверительный интервал, высчитанный методом bootstrap получился равным: (0.5038, 0.5564)

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.

#### 4.7 Расчёт спектральных характеристик и матрицы кросс-корреляции

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

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:

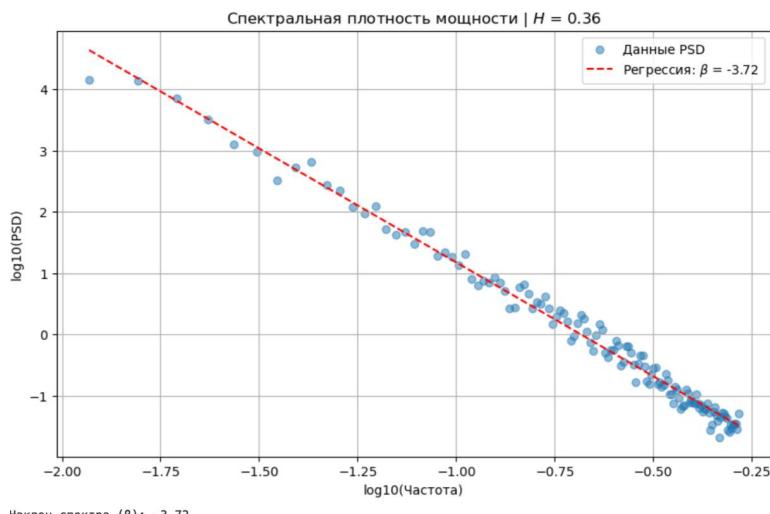


Рис. 28: Наклон спектра

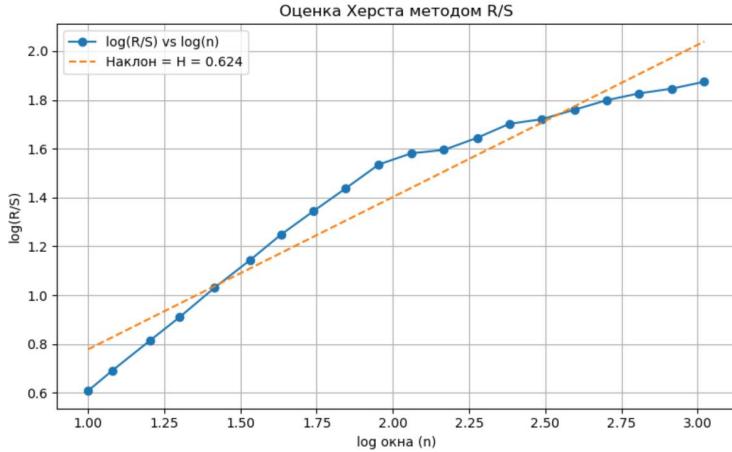


Рис. 29: R/S анализ

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 пунктах, 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 сохраняющейся 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. 30)

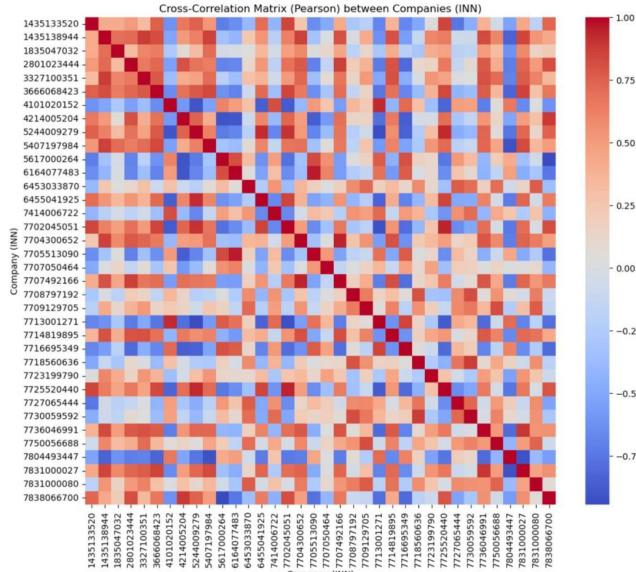


Рис. 30: Матрица кросс-корреляции

В пакете seaborn есть встроенная функция для кластеризации объектов на основе их коэффициента корреляции (Рис. 31)

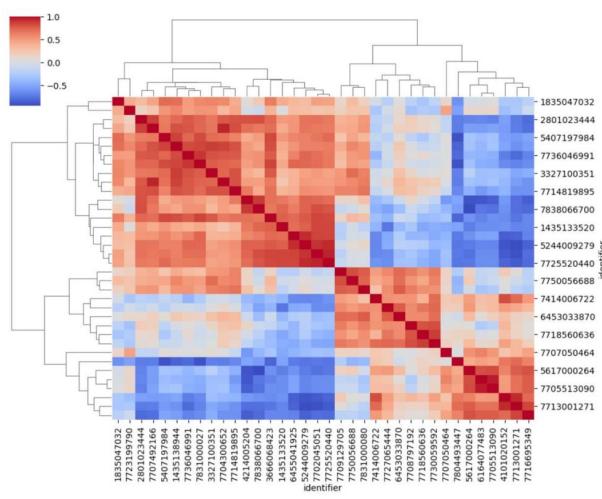


Рис. 31: Кластеризация матрицы кросс-корреляции

Однако нас больше интересует прикладная кластеризация посредством сингулярного разложения (SVD). Получим следующее (Рис. 32)

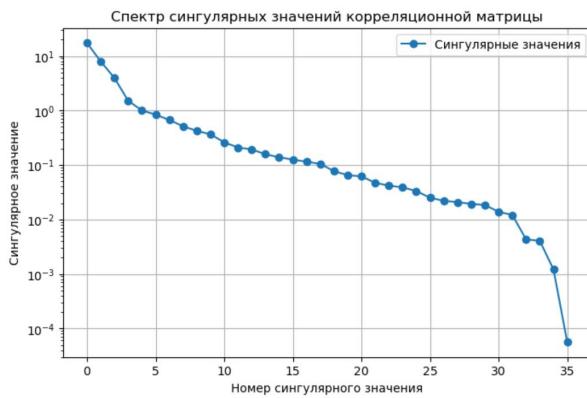


Рис. 32: Спектр сингулярных значений

А теперь посмотрим на главные компоненты и кластеризуем их (Рис. 33). 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)

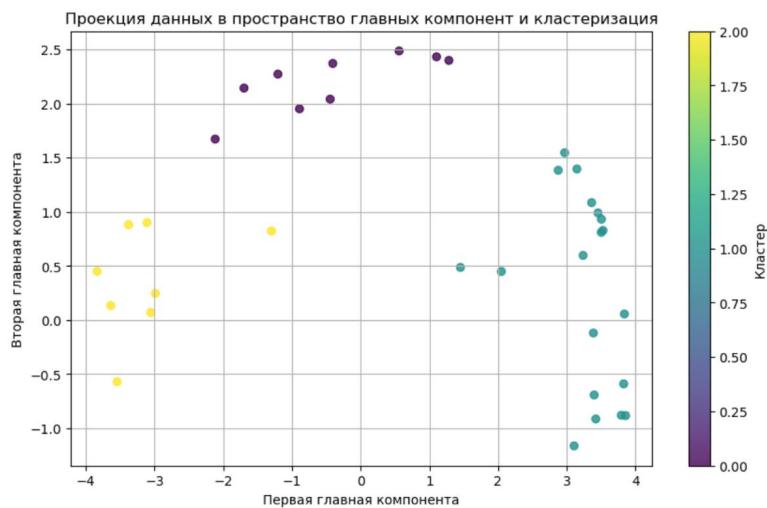


Рис. 33: Кластеризация компаний

В данном пункте проделана довольно интересная и немаловажная работа. Для более представительной выборки данные методы позволяют 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.

## 5 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. 34, 35, 36)

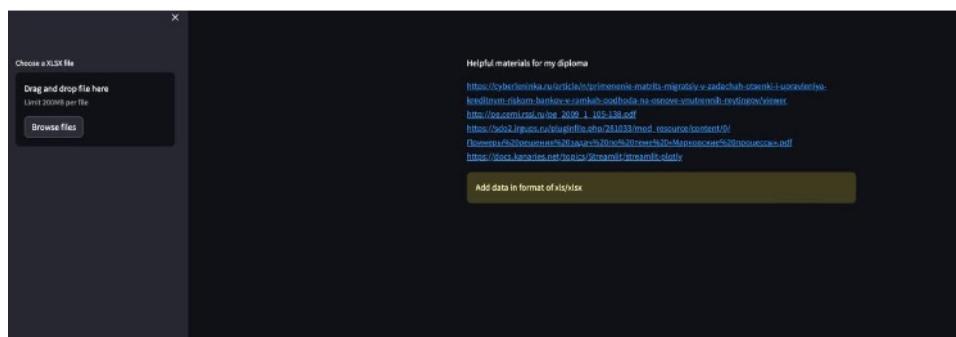


Рис. 34: Начальная страница

A screenshot of the Streamlit application showing a more complex interface. On the left, there is a sidebar with various dropdown menus for selecting columns: 'Choose agency column', 'Choose date column', 'Choose identifier column', 'Choose rating column', and 'Choose type column'. Below these are 'Choose scale column' and 'Choose one agency to check'. The main area shows a large table with columns 'Unnamed: 0', 'id', 'score', and 'name'. The table lists 13 rows of data, each with a checkbox next to it. The data includes names like 'Национальное Кредитное Организация Центральный Контроль ТОО', 'АКЦИОНЕРНОЕ ОБЩЕСТВО «ЛУГАНСК», 'Полпромбанк (Инженерное общество)', etc. At the bottom of the table, there is a note: 'Область с ограниченной ответственностью "Экобанк"'. The sidebar also includes dropdowns for 'Choose type of companies' (with options 'СБНК кредит' and 'СО - производст') and 'Choose one agency to check' (with options 'AKRA' and 'МФО').

Рис. 35: Настройка параметров

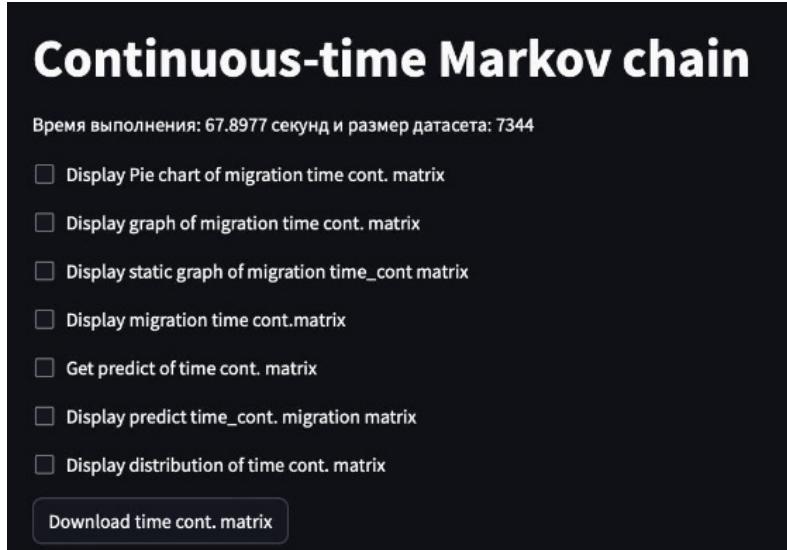


Рис.36: Дополнительный функционал расчёта

## 6 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 MCMC, 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).

## 7 Заключение

Основное достижения данного проекта состоит в разработке алгоритмов предсказания типа карточного клиента на несколько будущих периодов вперёд. Алгоритмы были основаны на случайном лесе и марковской модели. Ключевой особенностью проекта является изучение и предсказание всех состояний кредитных рейтингов, в то время как на основе приведённого 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. Также 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.

## 8 Приложение

[GitHub](#)

[Результаты](#)

[Постер проекта](#)

## Список литературы

- [1] Mahsa Tavakoli, Rohitash Chandra, Fengrui Tian, and Cristián Bravo Multi-Modal Deep Learning for Credit Rating Prediction Using Text and Numerical Data Streams // *Applied Soft Computing*. - 2025. - №171
- [2] Aiyegebeni G., Li Y., Annan J., Adebayo F. Credit Rating Prediction Using Different Machine Learning Techniques // *International Journal of Data Science and Advanced Analytics*. – 2023
- [3] David Lando Credit Risk Modeling: Theory and Applications. - 1 изд. - *United Kingdom: Princeton University Press*, 2004. – 329 с.
- [4] Kutenko, S. and Ozerov, K. (2024). Approaches to Default Probability Estimation of Credit Rating Agencies' Rating Scales. *Russian Journal of Money and Finance*, 83(4), pp. 98–118.
- [5] А.В.Булинский and А.Н.Ширяев. *Теория случайных процессов*. ФИЗМАТЛИТ, 2005.
- [6] David Lando Analyzing rating transitions and rating drift with continuous observations // *Journal of Banking & Finance*. - 2002. - №26. - С. 423-444
- [7] Peter Eichelsbacher, Ayalvadi Ganesh Bayesian inference for Markov chains // *Journal of Applied Probability*. - 2002
- [8] Л.В. Кирьянова, Д.В. Липатов. Статистический анализ изменчивости экономических рейтингов компаний // *Экономика и предпринимательство*. № 7; 2025 г., стр 1087 - 1091
- [9] Л.В. Кирьянова, Д.В. Липатов. Статистическое исследование изменения рейтинга компаний,

занимающихся экономической деятельностью. *Экономика и предпринимательство*. № 8; 2024 г., стр 1286 - 1290

- [10] Parisa Golbayani, Ionut, Florescu, and Rupak Chatterjee A comparative study of forecasting Corporate Credit Ratings using Neural Networks, Support Vector Machines, and Decision Trees // *The North American Journal of Economics and Finance*. - 2020. - №54
- [11] Христофоров А.В. *ТЕОРИЯ СЛУЧАЙНЫХ ПРОЦЕССОВ В ГИДРОЛОГИИ*. - М.: издательство МГУ, 1994 - 139 с.
- [12] Das S.R., Fan R., Geng G. Bayesian Migration in Credit Ratings Based on Probabilities of Default // *The Journal of Fixed Income*. – 2002
- [13] Han B. Evaluating Machine Learning Techniques for Credit Risk Management: An Algorithmic Comparison // *Proceedings of the 5th International Conference on Signal Processing and Machine Learning*. – 2025.
- [14] Wu H.-C., Wu Y.-T. Evaluating Credit Rating Prediction by Using the KMV Model and Random Forest // *Kybernetes*. – 2016. – Vol. 45, No. 10. – P. 1637–1651
- [15] Ивченко Г.И., Медведев Ю.И *Введение в математическую статистику*. - 1 изд. - М.: ЛКИ, 2010. - 600 с.

