



Факультет Экономических Наук

Экономика

Москва
2024

Кластерный анализ банков методами машинного обучения на данных открытых источников

Бочкарев Сергей Максимович БЭК201
Лашманов Валентин Денисович БЭК204
Мамедов Ильгар Габилович БЭК207



Контекст

Российская банковская система сталкивается с конъюнктурными проблемами с 2022 года:

- Обесценивание рубля
- Рост инфляции
- Масштабные санкции
- Интенсивный фискальный импульс
- Жесткая денежно-кредитная политика

Макроэкономические индикаторы в РФ



Аккумулирование рисков в банковском секторе

Коммерческие банки в РФ вынуждены:

- Работать в условиях высокой ставки ЦБ РФ
- Переориентировать услуги на российский бизнес
- Отказаться от предоставления услуг нерезидентам
- Повысить аппетит к риску в собственном портфеле
- Развивать сложные финансовые продукты

Идентификация риск-профиля

Банки сталкиваются с неопределенностью и специфичными рисками. Регулятору необходимо оценивать риск-профиль банков в дополнение к традиционным метрикам риска.

Определение его методами кластерного анализа



Цель:

Разработать инструмент для определения риск-профиля банков с использованием кластерного анализа на основе данных из открытых источников.

Общие задачи:

- Определить выборку исследуемых банков
- Собрать и агрегировать данные
- Объединить результаты кластеризации
- Оценить и проинтерпретировать результат

Бочкарев С. Кластеризация банковских портфелей:

- Отбор источников и показателей для кластерного анализа
- Выявить связи между элементами банковских портфелей
- Применить техники кластеризации банковских портфелей

Лашманов В. Кластеризация по фин. показателям:

- Найти показатели банков, определяющие риск-профиль
- Собрать данные о бизнес-моделях банков и их результатах
- Построить модель кластеризации на фин. показателях

Мамедов И. Кластеризация текстовых данных:

- Предобработка текста, вычленение важных фактов
- Преобразование текста в числовое представление
- Подбор алгоритма для разбиения данных и кластеризации



Гипотезы

Общие:

- Банки можно кластеризовать по риск-профилю.
- Разные методы дают схожие результаты, кластеры.
- Объединение данных повышает качество моделей.

Банковские портфели:

- Рискованность банковской деятельности снижается с диверсификацией продуктов.
- Если состав портфеля банка отличается от среднерыночного, то ниже системные риски.

Финансовые показатели:

- Банки с высокой доходностью имеют более высокую финансовую стабильность.
- Банки с волатильными показателями доходности менее финансово устойчивы.

Текстовые данные:

- Банки из кластера положительных настроений имеют ниже показатель финансовой стабильности.
- Банки, в новостях которых чаще затрагивается тема финансовых рынков, менее финансово устойчивы.

Практическая значимость

Использование службой анализов рисков ЦБ РФ для надзорной деятельности над коммерческими банками:

- Дополнительная оценка банковских рисков
- Выделение неявных паттернов деятельности банков
- Проведение стресс-тестирований

Работа расширяет набор применяемых методов, впервые объединяет разные форматы данных и применена для большого числа российских банков.

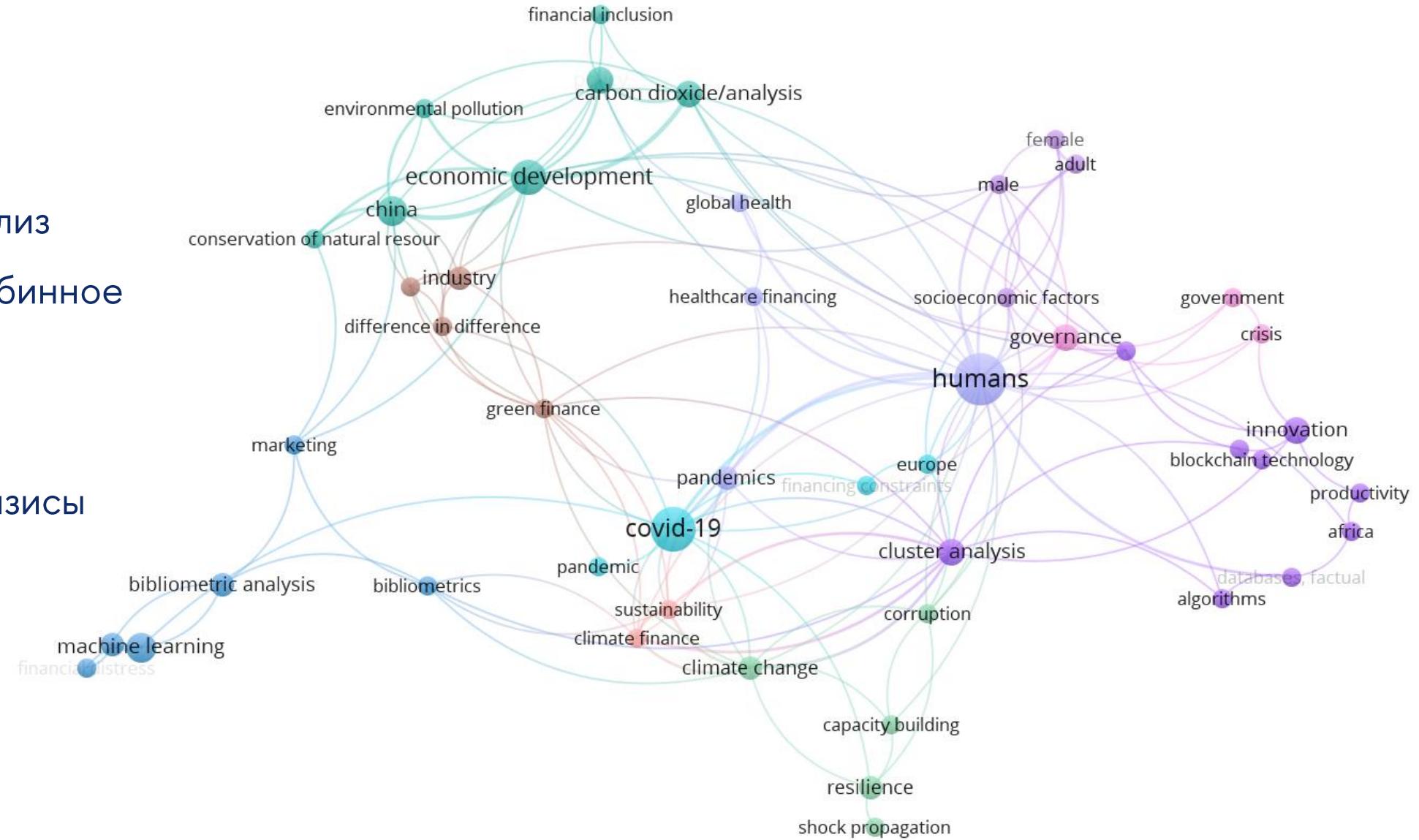
Рассматриваемый круг вопросов:

- Уровень концентрации банковского сектора в РФ
- Наполнение портфелей российских банков
- Диверсифицированность бизнес-моделей банков
- Приоритеты деятельности российских банков
- Продуктовый риск-менеджмент российских банков



Основные темы:

- Кластерный анализ
- Машинное / Глубинное обучение
- Covid-19
- Финансовые кризисы
- Клиенты
- Стабильность





Связь между портфелем, продуктами и риском и Политика банков по ведению портфелей

Автор(ы)	Название	Год	Важные выводы
Köhler M.	Which banks are more risky? The impact of business models on bank stability	2014	Влияние непроцентных доходов на бизнес-модель и риск-профиль банка. Исследование показало, что чем выше доля непроцентных доходов, тем выше устойчивость банка.
Altunbas Y., Manganelli S., Marques-Ibanez D.	Bank Risk During The Financial Crisis: Do Business Models Matter?	2011	Исследовался риск в зависимости от бизнес-модели банков во время кризиса 2007-2009 гг. На риск влияла структура активов, уровень секьюритизации и выдач кредитов.
Deloitte	Product risk management top of mind for banking regulator	2022	В документе подчеркивается рост требований к управлению продуктами и бизнес-моделью для противостояния рискам.
Huang, J., Chen, Z.	Optimal risk asset allocation of a loss-averse bank with partial information under inflation risk.	2021	Есть связь между неполнотой информации, инфляционными рисками и инвестиционным поведением банков. Банки повышают склонность к риску во время кризисов.
Lee, C. Chen, Pei-Fen; Zeng, Jhih-Hong	Bank Income Diversification, Asset Correlation, and Systemic Risk	2020	Исследовалась корреляция между видами деятельности и продуктов банков. Показано, что диверсификация доходов банков значительно повышает системный риск в банковском секторе
Farnè M.; Vouldis A.	Business models of the banks in the euro area	2017	Выделены четыре основные бизнес-модели банков еврозоны. Для этих моделей характерны разные уровни риска и показателей эффективности



Кластеризация банков по составу портфеля и бизнес-модели и роль целевой аудитории

Автор(ы)	Название	Год	Важные выводы
Lagasio, V., Quaranta, A.	Cluster analysis of bank business models: The connection with performance, efficiency and risk.	2021	Группировка банковских профилей рисков возможна по долям инструментов в портфеле банка. Было выделено 5 конкретных групп банков, таких как розничное или оптовое финансирование.
Zarutska, E., Novikova, L., Pavlov, R., Pavlova, T., Levkovich, O.	Evaluation of Ukrainian Banks' Business Models by the Structural and Functional Groups Analysis Method.	2022	Выявлены корреляции между элементами портфеля, в том числе в динамике. Это помогает охарактеризовать бизнес банка с учетом его склонности к риску.
Zhu, Z., Liu, N.	Early Warning of Financial Risk Based on K-Means Clustering Algorithm.	2021	Доказана большая эффективность ML-алгоритмов при установлении пороговых значений для принадлежности к банковским кластерам.
Hinchcliff M., Kyriazis E., McCarthy G., Mehmet M.	The moderating role of high- and low-involvement product types on customer loyalty and satisfaction in banking: an Australian perspective	2023	Изучено влияние банковских продуктов на лояльность и удовлетворенность клиентов. Продукты поделены на вовлекающие клиентов и нет.
Hanaki, N.	Risk misperceptions of structured financial products with worst-of payout characteristics revisited.	2021	На рискованность банка влияет тип банковских продуктов и профиль целевой аудитории. Сложные финансовые продукты следует предлагать профессиональным клиентам.



Автор(-ы)	Название	Год	Выводы
Cerchiello P., Nicola G., Rönnqvist S., Sarlin P.	Deep Learning for Assessing Banks' Distress from News and Numerical Financial Data	2018	Объединение текстов новостей с финансовыми показателями улучшает модель прогнозирования дефолта/кризиса.
Chen M., DeHaven M., Kitschelt I., Lee S. J., Sicilian M.	Identifying Financial Crises Using Machine Learning on Textual Data	2023	
Kriebel J., Stitz L.	Credit Default Prediction from User-Generated Text in Peer-to-Peer Lending Using Deep Learning	2021	Использование информации из текстов клиентов повышает качество модели.
Chiagoziem O., Muliaro J., Kyalo J.	Support Vector Machine for Sentiment Analysis of Nigerian Banks Financial Tweets.	2019	
Garcia S., Fernandez-Gavilanes M., Juncal-Martinez J., Francisco J., Barba O.	Identifying Banking Transaction Descriptions via Support Vector Machine Short-Text Classification Based on a Specialized Labelled Corpus	2020	ML модели обученные на текстовых данных способны давать высокое качество предсказаний.
Priola P., Lorenzini P., Tizzanini G., Zicchino L.	Measuring Central Banks' Sentiment and its Spillover Effects with a Network Approach	2021	



Автор(-ы)	Название	Год	Выводы
Omotosho, B.	Analysing User Experience of Mobile Banking Applications in Nigeria: A Text Mining Approach	2021	Используя алгоритмы тематического моделирования, можно понять основные проблемы банка.
Garcia-Mendez S., Arriba-Perez F., Barros-Vila A., Francisco J., Costa-Montenegro E.	Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with Latent Dirichlet Allocation	2023	
Loukas L., Stogiannidis I., Malakasiotis P. Vassos S.	Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance.	2023	Предобученные модели дают лучшее качество на маленьких корпусах текстах.
Yang L., Yingpeng M., Zhang Y.	Measuring Consistency in Text-based Financial Forecasting Models	2023	
Fatma S., Kaya E.	Academic Text Clustering Using Natural Language Processing	2022	
Subakti A., Murf H., Hariadi N.	The Performance of BERT as Data Representation of Text Clustering	2022	Числовые представления BERT дают устойчивые результаты в задачах классификации и кластеризации.
Mehta V., Bawa S., Singh J.	WEClustering: word embeddings based text clustering technique for large datasets	2021	



Авторы (-ы)	Название	Год	Выводы
Vladyslav Rashkovyan, Dmytro Pokidin	Ukrainian Banks' Business Models Clustering: Application of Kohonen Neural Networks	2016	<ul style="list-style-type: none">• SOM для кластеризации и визуализации• Разделение признаков на риск и бизнес категории• Визуализация миграции банков
Станик Наталья, Крайнюков Николай	Методы и показатели для кластеризации коммерческих банков	2020	<ul style="list-style-type: none">• Единственная статья по кластеризации российских банков• Использование абсолютных переменных
Bernardo P. Marques, Carlos F. Alves	Using clustering ensemble to identify banking business models	2020	<ul style="list-style-type: none">• Признаки отнормированы на Активы Нетто• Использование ансамбля из методов PAM, FCM, SOM• Классификация банков на устойчивые и нет
Mathieu Mercadier, Amine Tarazi, Paul Armand, Jean-Pierre Lardy	Banks' risk clustering using k-means: a method based on size and individual & systemic risks	2021	<ul style="list-style-type: none">• Использование риск-признаков (Z-score, RatioBS, CreditBS)• Использование метрики Силуэта для определения кол-ва кластеров• Использование признаков рыночного риска
Sepideh Aghajania; Reza Samizadehb	A Clustering Approach for Business Models of Iranian Banks; Analysis of Risks and Migrations	2023	<ul style="list-style-type: none">• Метод ансамбля дает более устойчивые разбиения• Использование риск-признаков вместе с показателями банков• Предсказание миграций банков из кластера в кластер



Репозиторий исследования на GitHub:

- Датасеты
- Дескриптивный анализ
- Визуализация
- Модели



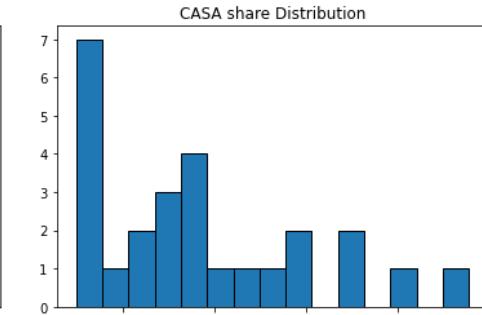
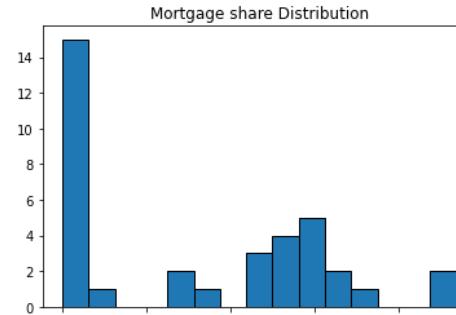
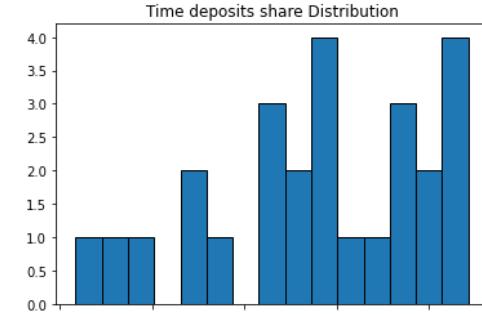
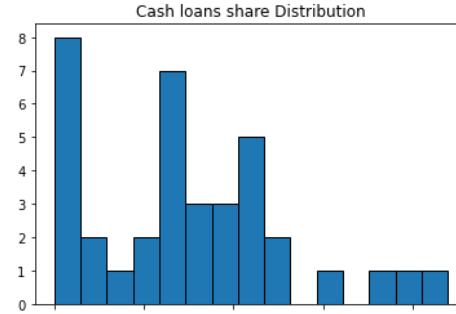


Банковские портфели, данные Frank RG, 2023 г.

- 40+ крупнейших банков
- Распределение активов и пассивов банков

Доли	Кредиты	Ипотеки	Авто кредиты	Кредитки, овердрафт	Товарные кредиты	депозиты	Тек. и сберег сч.	Эскроу-счета
count	37.000000	36.000000	37.000000	36.000000	15.000000	26.000000	26.000000	26.000000
mean	0.297009	0.319745	0.268852	0.091938	0.083042	0.563085	0.367313	0.069602
std	0.230447	0.310040	0.385231	0.169540	0.113842	0.242391	0.237230	0.151211
min	0.000000	0.000000	0.000000	0.00	0.000000	0.034421	0.099542	0.000000
25%	0.067708	0.000000	0.000014	0.003578	0.000244	0.442497	0.152453	0.000000
50%	0.286237	0.327013	0.104079	0.027472	0.029899	0.576417	0.322929	0.013644
75%	0.422366	0.578119	0.228803	0.095012	0.135990	0.756376	0.508184	0.092174
max	0.877907	0.938807	1.000000	0.843740	0.377919	0.884653	0.953335	0.768733
mlrd. rub.	Кредиты	Ипотеки	Авто кредиты	Кредитки, овердрафт	Товарные кредиты	депозиты	Тек. и сберег сч.	Эскроу-счета
mean	227.31682	329.47038	29.618823	62.217695	15.934247	734.91160	494.13016	38.866318
std	606.57625	1194.7647	44.335601	192.49853	16.074398	1893.5844	1257.1546	216.44317

Распределение ключевых статей



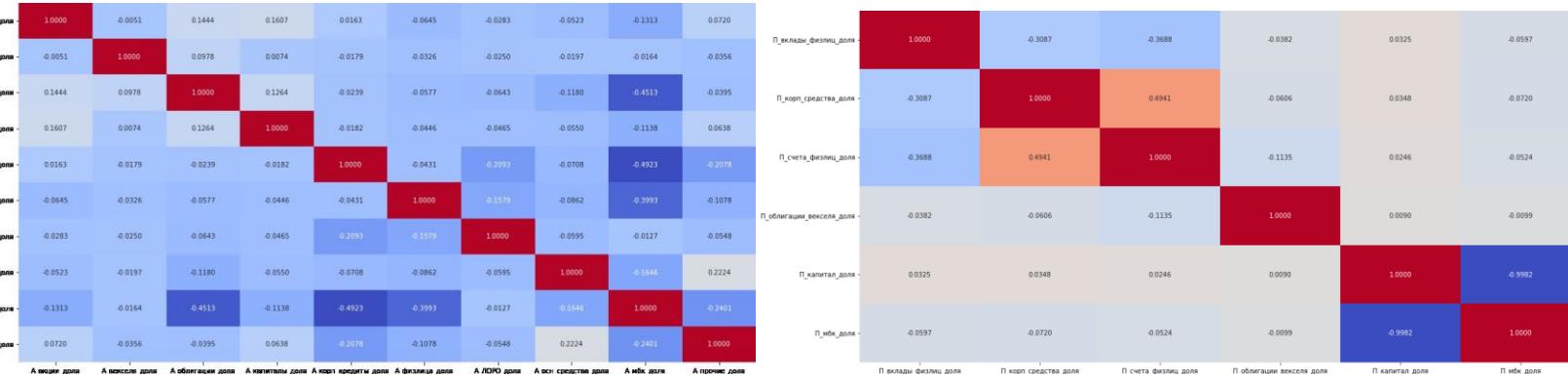
Источник: Frank RG



Банковские портфели, данные Banki.ru 2023 г.

Корреляционная матрица долей активов и пассивов, banki.ru

- 339 российских банков
- Распределение активов и пассивов банков
- Переведено в доли
- Мультиколлинеарности нет



Сводные статистики о статьях активов и пассивов в долях от совокупных активов и пассивов, banki.ru

Доли портфеля	Акции	Векселя	Облигации	Другие капиталы	Корп кредиты	Физлица	ЛОРО	Основные средства	МБК	Прочие	Вклады физлиц	Корп средства	Счета физлиц	Облигации и векселя	Капитал	МБК
count	339	339	339	339	339	339	339	339	339	339	339	339	339	339	339	339
mean	0,004	0,001	0,108	0,005	0,236	0,117	0,028	0,037	0,392	0,071	0,207	0,272	0,164	0,008	0,079	0,272
std	0,017	0,006	0,157	0,022	0,202	0,168	0,095	0,075	0,285	0,121	0,200	0,158	0,111	0,034	3,966	3,989
min	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	-72,609	0,000
25%	0,000	0,000	0,000	0,000	0,039	0,003	0,000	0,006	0,156	0,018	0,002	0,159	0,078	0,000	0,135	0,000
50%	0,000	0,000	0,027	0,000	0,206	0,042	0,000	0,016	0,339	0,037	0,180	0,258	0,147	0,000	0,218	0,000
75%	0,000	0,000	0,172	0,000	0,393	0,164	0,007	0,035	0,612	0,071	0,371	0,378	0,227	0,000	0,388	0,030
max	0,145	0,110	0,888	0,238	0,829	0,932	0,753	0,748	0,986	0,890	0,776	0,877	0,457	0,380	1,000	73,451

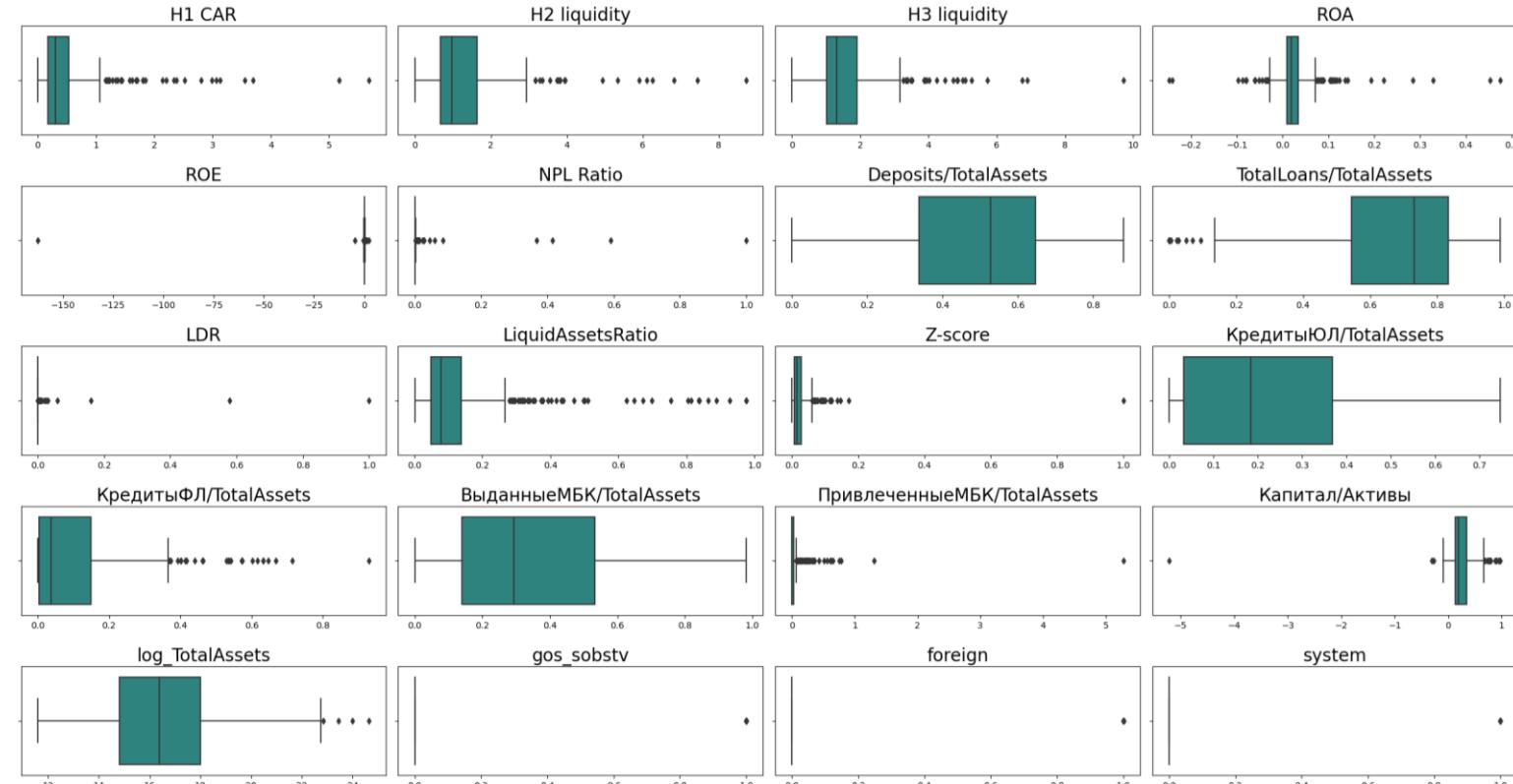
Источник: расчеты авторов



Признаки, использующиеся для кластеризации на финансовых показателях

- 2023 год
- 339 банков
- Источник:
[banki.ru/финансовые
показатели банков](http://banki.ru)
- К признакам Z-score,
LDR, NPL Ratio
применен
MinMaxScaler [0, 1]

Распределения финансовых показателей, ящиковыe диаграммы



* **gos_sobstv** – банк государственный (1) или частный (0)
* **foreign** – банк иностранный (1) или нет (0)
* **system** – банк системозначимый (1) или нет (0)



Признаки, использующиеся для кластеризации на финансовых показателях

$$\text{Deposits/TotalAssets} = \frac{\text{Вклады физических лиц} + \text{Средства предприятий и организаций}}{\text{Активы нетто}}$$

$$\text{TotalLoans/TotalAssets} = \frac{\text{Кредиты предприятиям и организациям} + \text{Кредиты физическим лицам} + \text{Выданные МБК}}{\text{Активы нетто}}$$

$$\text{LDR} = \frac{\text{Кредиты предприятиям и организациям} + \text{Кредиты физическим лицам} + \text{Выданные МБК}}{\text{Вклады физических лиц}}$$

$$\text{LiquidAssetsRatio} = \frac{\text{Высоколиквидные активы}}{\text{Активы нетто}}$$

$$\text{Кредиты ЮЛ/TotalAssets} = \frac{\text{Кредиты предприятиям и организациям}}{\text{Активы нетто}}$$



Признаки, использующиеся для кластеризации на финансовых показателях

$$\text{КредитыФЛ/TotalAssets} = \frac{\text{Кредиты физическим лицам}}{\text{Активы нетто}}$$

$$\text{ВыданныеМБК/TotalAssets} = \frac{\text{Выданные МБК}}{\text{Активы нетто}}$$

$$\text{ПривлеченныеМБК/TotalAssets} = \frac{\text{Привлеченные МБК}}{\text{Активы нетто}}$$

$$\text{Капитал/Активы} = \frac{\text{Капитал}}{\text{Активы нетто}}$$

$$\log_{\text{TotalAssets}} = \ln(\text{Активы нетто} + 1)$$

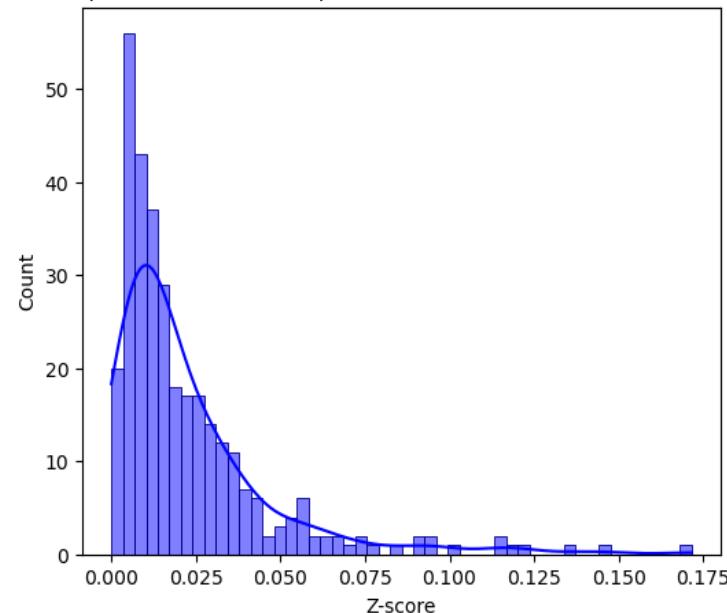
$$ROA = \frac{\text{Чистая прибыль}}{\text{Активы нетто}}$$

$$ROE = \frac{\text{Чистая прибыль}}{\text{Капитал (по форме 123)}}$$

Основной риск-признак:

$$Z - score = \frac{ROA_t + H1 \ CAR_t}{std(ROA_{t-3:t})}$$

Распределение Z-score без выброса
(банк ИНЭКО)



Источник: расчеты авторов

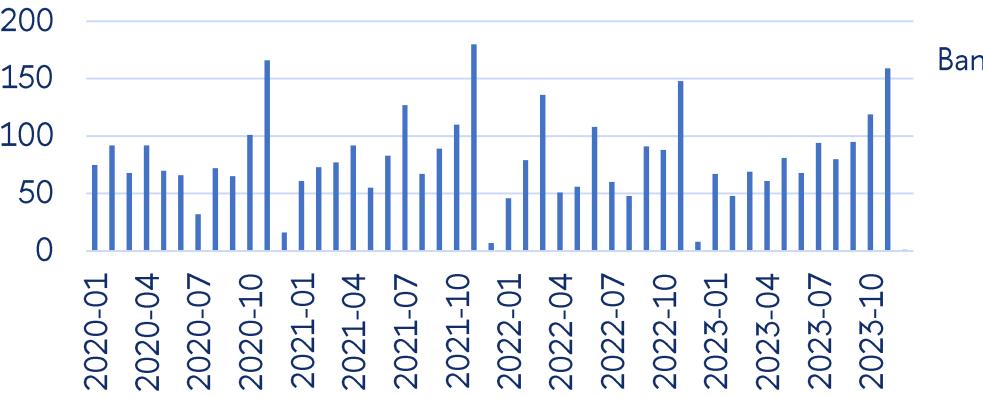


Новости на “Коммерсантъ”

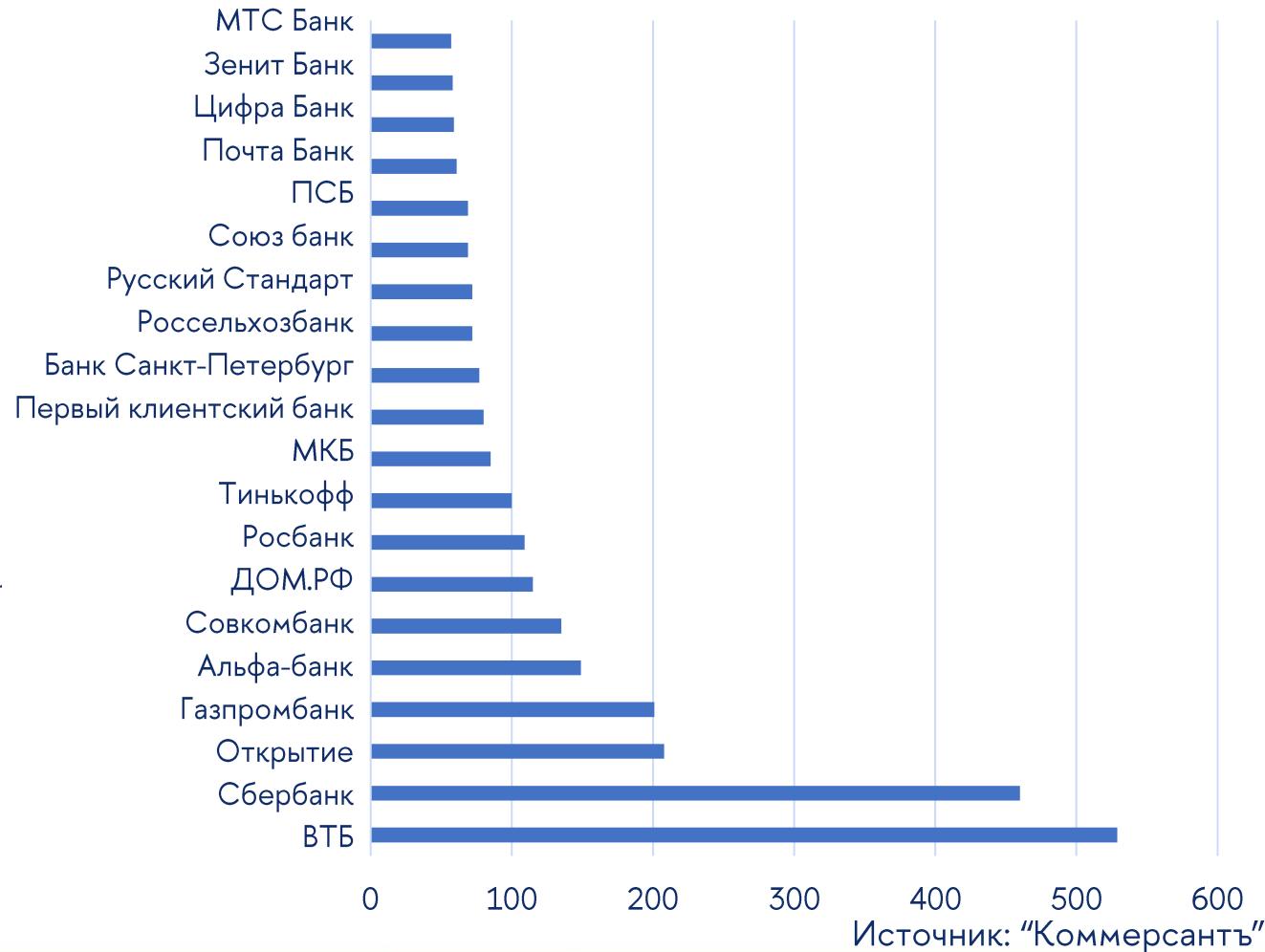
- Банков: 114
- Статей: 4050
- Начало периода: 2020-01-09
- Конец периода : 2023-12-01



Количество опубликованный
статьей



Топ-20 популярных банков по количеству статей на “Коммерсантъ”, 2020-2023



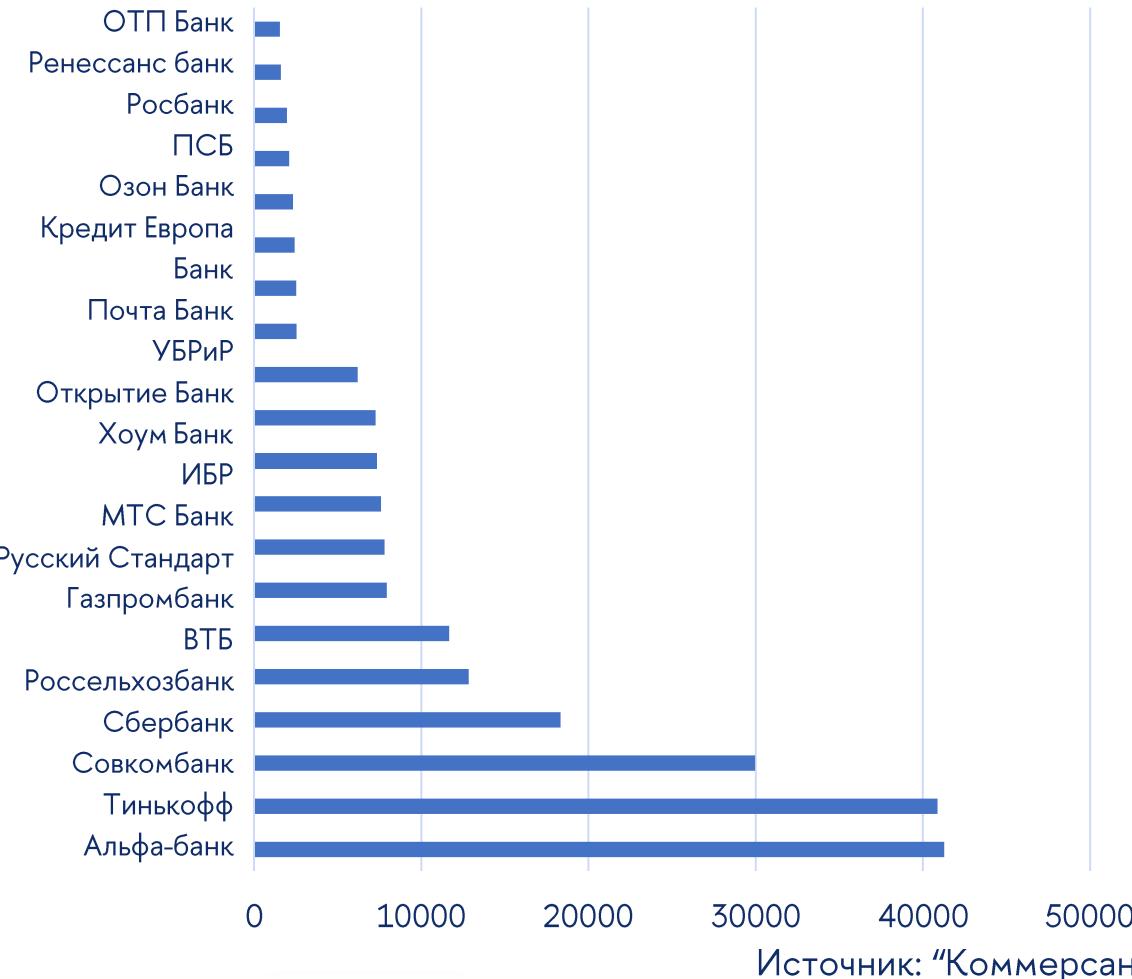


Отзывы на “Банки.ру”

- Банков: 165
- Отзывов: 234659
- Начало периода: 2019-01-02
- Конец периода: 2024-01-01

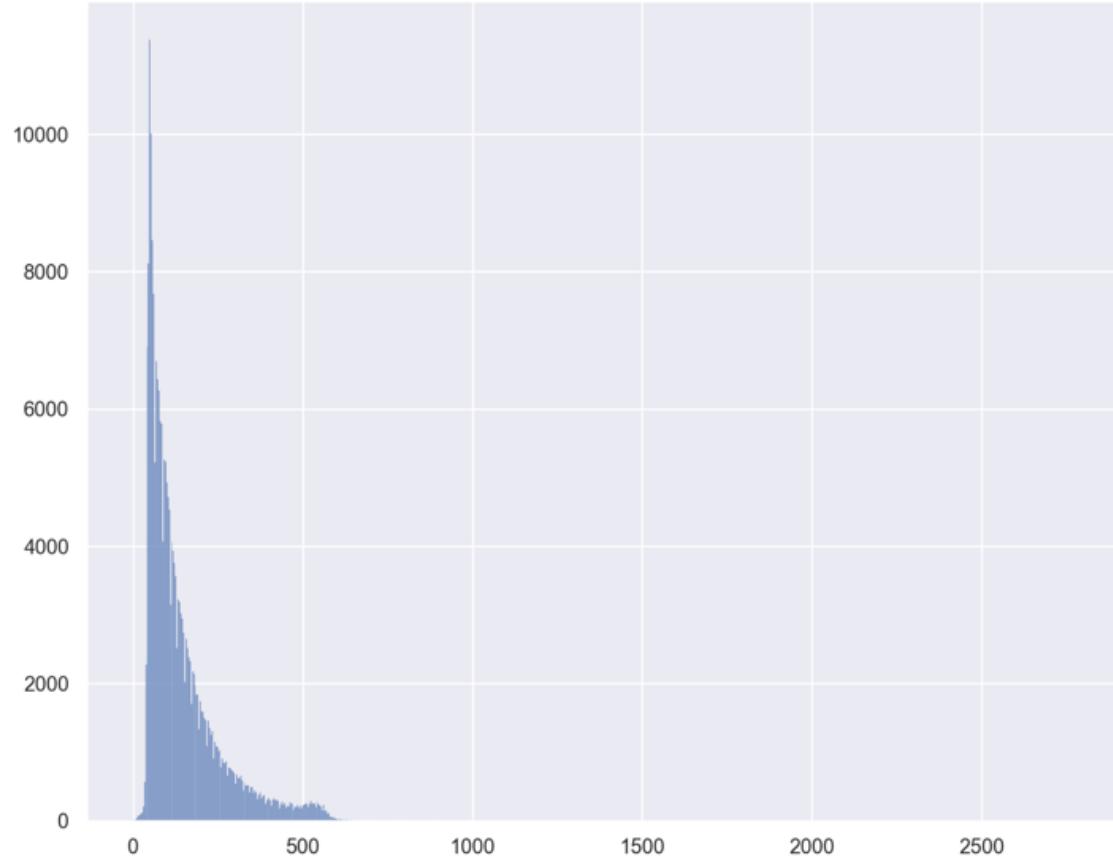


Топ-20 популярных банков по количеству отзывов на сайте banki.ru, 2020-2023



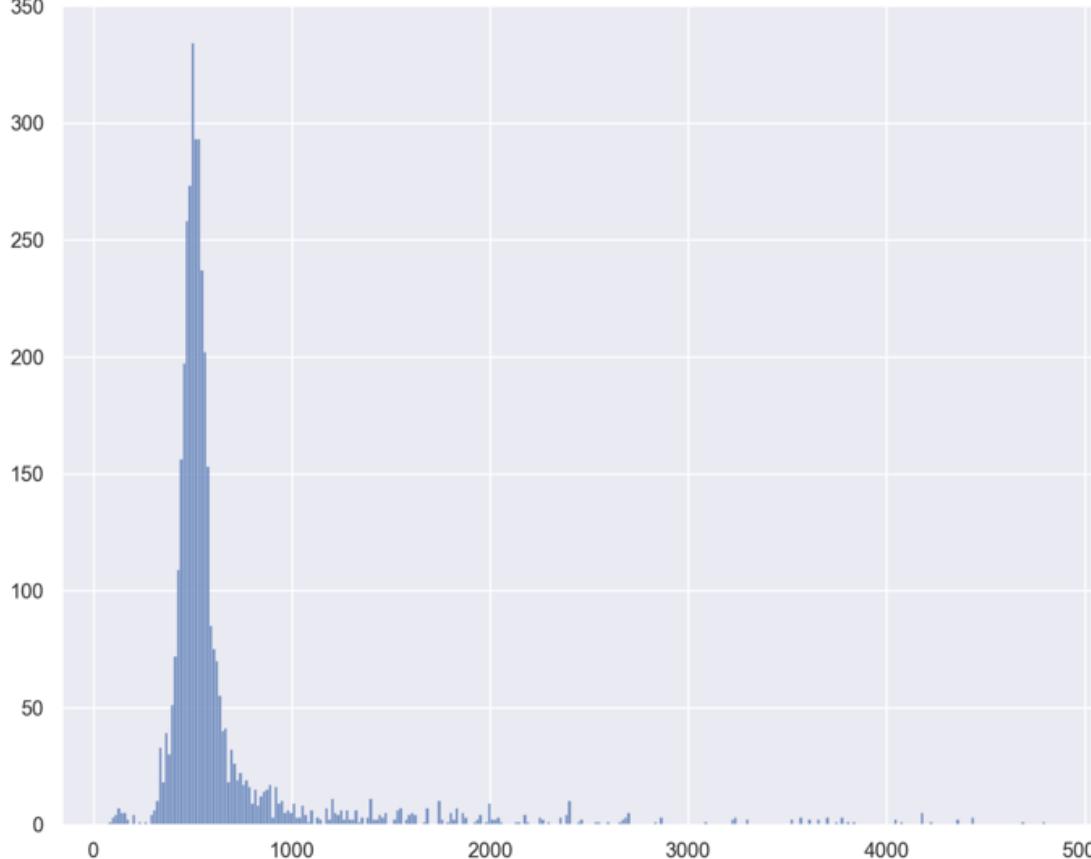


Распределение длины наблюдений в отзывах

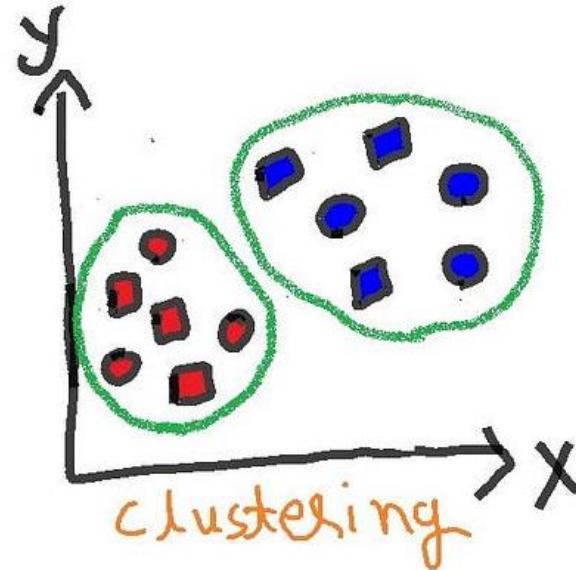
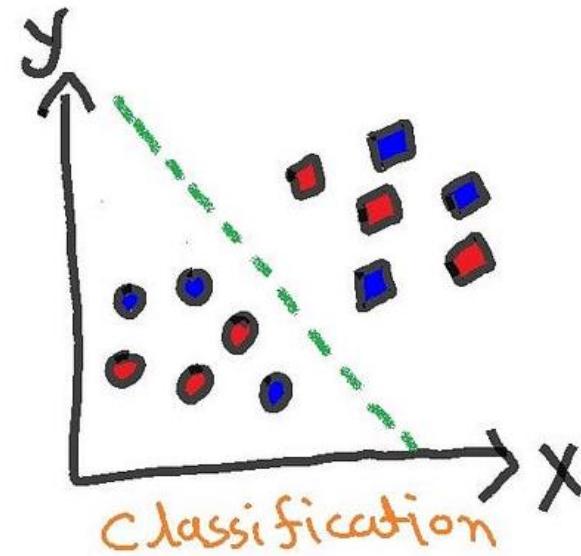


Источник: Banki.ru

Распределение длины наблюдений в новостных статьях



Источник: Коммерсантъ



Задача кластеризации

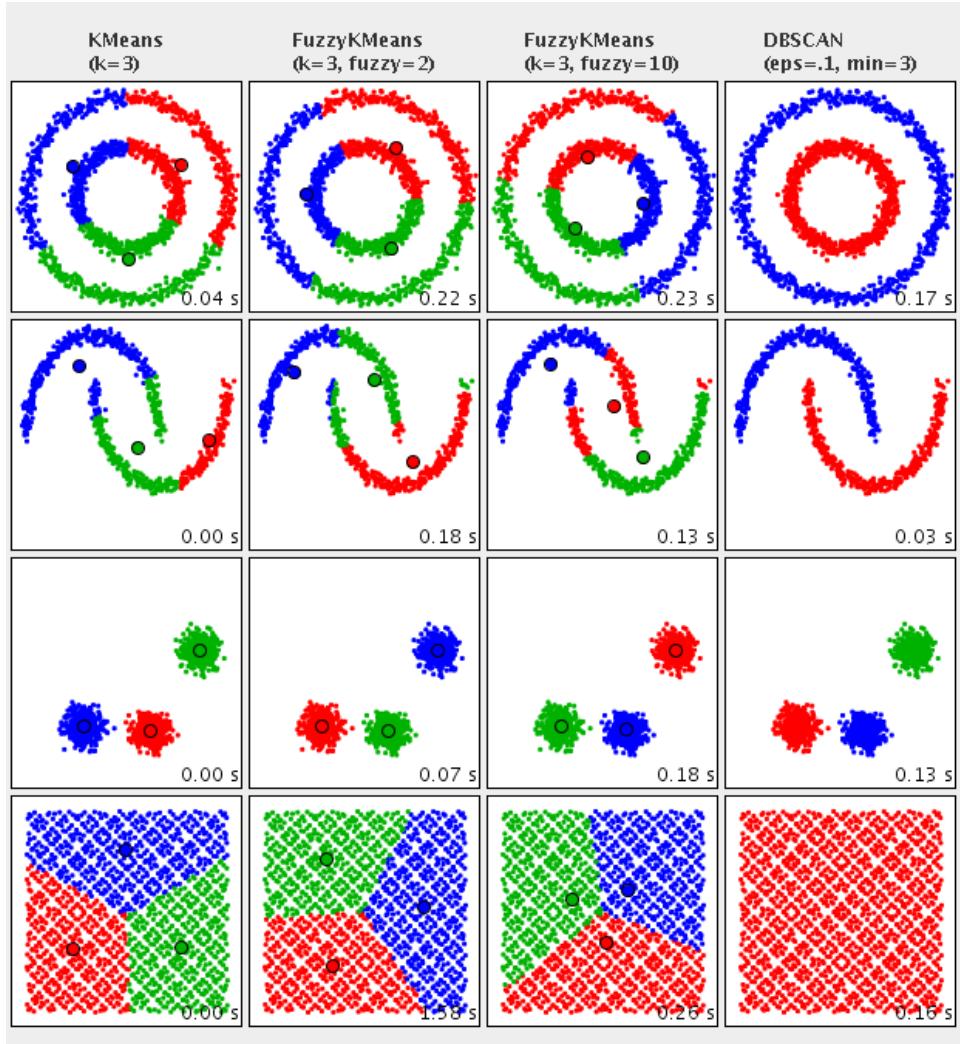
$X = (x_i)_{i=1}^n$ – выборка из n объектов

Модель a по любому объекту выдает
номер кластера, в который данный
объект попал:

$$a: X \rightarrow \{1, 2, 3, \dots, K\}$$

K – количество кластеров.

Если объекты x_i и x_j «похожи» $\Rightarrow a(x_i) = a(x_j)$



K-means:

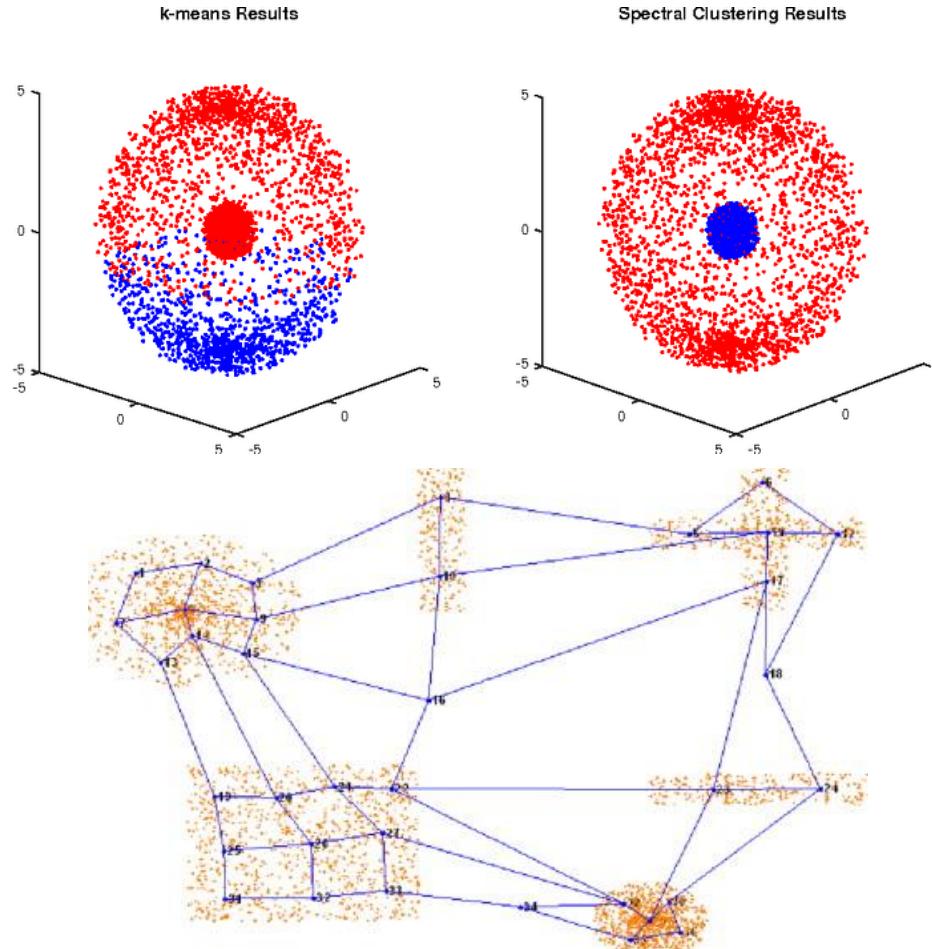
- + Быстрый
- + Легко распараллелить
- Зависит от инициализации
- Чувствителен к признакам разного масштаба
- Находит выпуклые кластеры

FCM:

- + Находит пересекающиеся кластеры
- + Быстрый
- Чувствителен к выбросам
- Чувствителен к инициализации числа кластеров и степени нечеткости

DBSCAN:

- + Находит кластеры сложной формы
- + Находит выбросы
- + Более простые гиперпараметры
- Медленный
- Плохо работает при разной плотности в данных



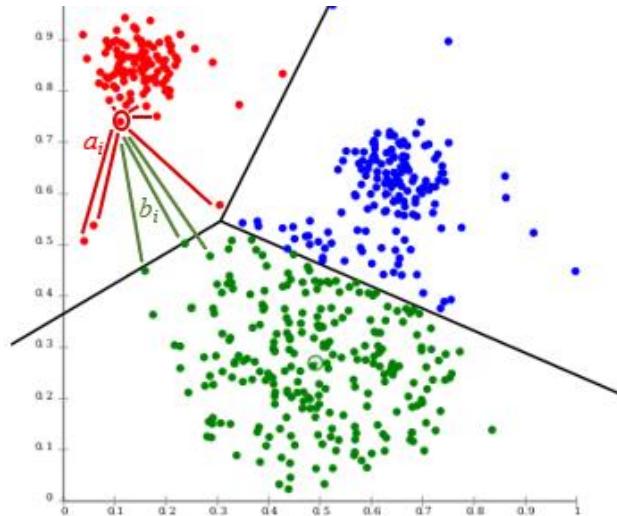
Спектральная кластеризация:

- + Может разделять нелинейно зависимые признаки
- + Способно выделять кластеры разной плотности и размера
- + Может работать с большой размерностью в данных
- Медленная (вычислительно дорогая)

SOM:

- + Много вариантов визуализации
- Медленный (вычислительно дорогая)
- Зависит от инициализации
- Нужно определять размер «карты»

Метрика Силуэта

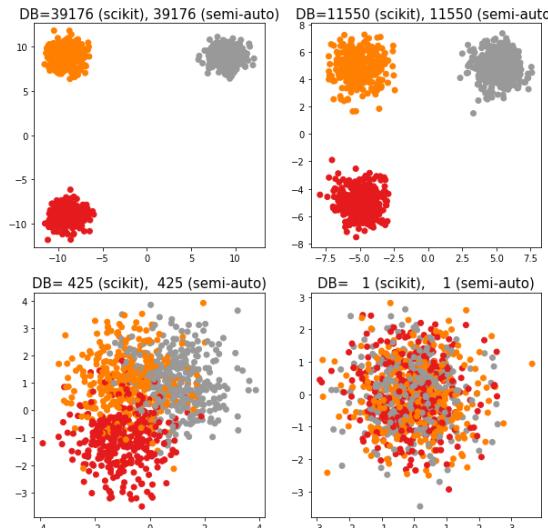


$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$$

$$a(x_i) = \frac{1}{|C_k|} \sum_{z \in C_k} \rho(x_i, z)$$

Чем больше, тем лучше [-1, 1]

Индекс Calinski-Harabasz



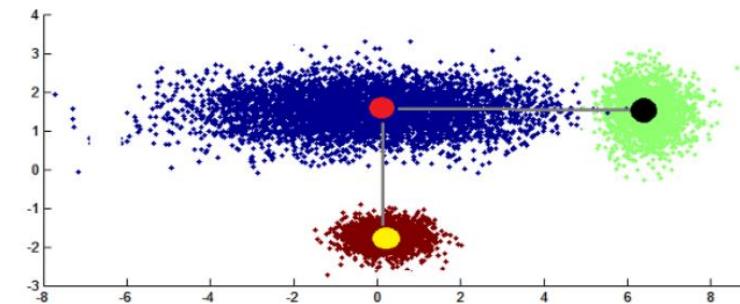
$$CH_{index} = \frac{tr(B_K)}{tr(W_K)} * \frac{n - K}{K - 1}$$

$$W_K = \sum_{q=1}^K \sum_{x_i \in C_q} (x_i - c_q)(x_i - c_q)^T$$

$$B_K = \sum_{q=1}^K |C_q| (c_q - c)(c_q - c)^T$$

Чем больше, тем лучше [0, + inf]

Индекс Дэвиса-Болдуина



$$DB_{index} = \frac{1}{K} \sum_{C_k \in C} \max_{C_l \in C \setminus C_k} \left(\frac{S(c_k) + S(c_l)}{\|c_k - c_l\|} \right)$$

$$S(C_k) = \frac{1}{|C_k|} \sum_{x_i \in C_k} (\|x_i - c_k\|)$$

Чем меньше, тем лучше [0, + inf]



Предварительная обработка



Необучаемые методы

Bag of words: наличие слова в качестве индикатора во всем словаре
Tf –idf: как BoW, но с корректировкой на уникальности слова

Численное представление

Обучаемые методы

Предобученная модель на большом
корпусе (ru-en вики)

- **Токенизация:** получение наилучшего разделения слов
- **Эмбединг:** вычисление наилучшего числового представления токена

Предсказание модели на нашем
корпусе

[ба,нк,про,лема,де,бет,ка,рта,gold,pla,tin,um]
[12, 1, 725, 16 10 1002, 10, 2, 9832, 727, 5, 103]

$$\begin{aligned} \text{ba} &\rightarrow X \in R^n \\ \text{nk} &\rightarrow X \in R^n \end{aligned}$$

n – размерность
численного
представления
токена



Основная задача

Кластеризация

- Получение числового представления текста
- Кластеризация методами МО

Тематическое моделирование

- Обучение LDA
- Определение главной темы текста
- Кластер текста = главная тема

Агрегация наблюдений

Макро усреднение

$obs_{bank,i} \rightarrow cluster_i$
 $obs_{bank,i+1} \rightarrow cluster_{i+1}$
...
 $obs_{bank,i+l} \rightarrow cluster_{i+l}$

Кластер банка
= самый популярный
кластер среди его
текстов

Микро усреднение

$obs_{bank,i} \rightarrow$
 $obs_{bank,i+1} \rightarrow AverageObs_{bank}$
 $obs_{bank,i+l} \rightarrow$

Кластер банка
= кластер его
среднего
представления



Композитный индекс надежности

Индекс является инструментом проверки результатов кластеризации. Он включает в себя:

- Кредитный рейтинг НРА, Эксперт РА, АКРА
- Объем активов
- Доля просроченной задолженности
- Санкционный статус банка
- Статус системно-значимого банка

Результаты рейтингования были проверены на устойчивость.

Средняя корреляция результатов с разными весами составила **0,87**.

С учетом экспертного мнения представителей ЦБ РФ и существующих практик, были подобраны следующие веса для рейтинга:

Объем активов	0,1
Статус системно-значимого банка	0,1
Санкционный статус	0,2
Доля просроченной задолженности	0,2
Кредитный рейтинг	0,4



BCubed

Формула:

$$\text{Correctness}(x, x') = \begin{cases} 1, & C(x) = C(x') \text{ и } L(x) = L(x') \\ 0, & \text{иначе} \end{cases}$$

$$\text{PrecisionBCubed} = \text{Avg}_x [\text{Avg}_{x':C(x)=C(x')} \text{Correctness}(x, x')]$$

$$\text{RecallBCubed} = \text{Avg}_x [\text{Avg}_{x':L(x)=L(x')} \text{Correctness}(x, x')]$$



$$BCubed = \frac{\text{PrecisionBCubed} * \text{RecallBCubed}}{\text{PrecisionBCubed} + \text{RecallBCubed}}$$

$C(x)$ – предсказанный
кластер
 $L(x)$ – истинная метка



BCubed

Свойства:

Однородность

$$Q \left(\begin{array}{c|cc} & \diamond & \diamond \\ \hline \times & & \diamond \\ \times & & \end{array} \right) < Q \left(\begin{array}{c|cc} & \diamond & \diamond \\ \hline \times & & \diamond \\ \times & & \end{array} \right)$$

Основное свойство разделения различных объектов на разные кластеры

Полнота

$$Q \left(\begin{array}{c|cc} & \times & \times \\ \hline \times & & \times \\ \times & & \end{array} \right) < Q \left(\begin{array}{c|cc} & \times & \times \\ \hline \times & & \times \\ \times & & \end{array} \right)$$

Один кластер не следует разбивать на несколько небольших

Rag-bag

$$Q \left(\begin{array}{c|cc|cc} \times & \times & \bullet & \circ \\ \times & \times & \triangleright & \star \\ \times & * & \odot & \square \end{array} \right) < Q \left(\begin{array}{c|cc|cc} \times & \times & \bullet & \circ \\ \times & \times & \triangleright & \star \\ \times & * & \odot & \square \end{array} \right)$$

Нетипичные наблюдения должны быть в одном кластере, чтобы остальные кластеры были однородными

Размер кластера vs количество

$$Q \left(\begin{array}{c|cc|cc} \times & \circ & \circ \\ \times & \star & \star \\ \times & \triangleright & \triangleright \\ \times & \odot & \odot \end{array} \right) < Q \left(\begin{array}{c|cc|cc} \times & \circ & \circ \\ \times & \star & \star \\ \times & \triangleright & \triangleright \\ \times & \odot & \odot \end{array} \right)$$

Лучше испортить один кластер, чтобы улучшить качество многих других

Результаты моделирования

Этапы моделирования:

- Ограничения на число и размер кластеров
- Подбор параметров и числа кластеров
- Выбор решения внутри модели
- Понижение размерности с помощью t-SNE

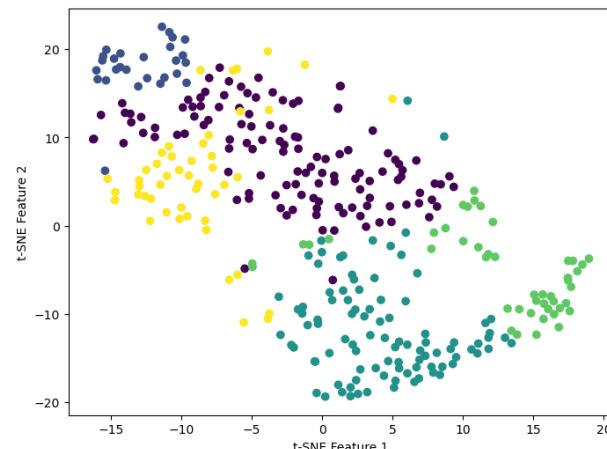
Кластеризация методом K-means показала лучшее качество при заданных условиях.

Получено 5 сбалансированных кластеров.

Результаты лучших моделей кластеризации по банковским

	K-means	DBSCAN*	SOM	Fuzzy C-means	Эталон
Число кластеров	5	8	4	5	—
Метрика Силуэта	0,26	-0,18	0,26	—	1
Индекс Davies-Bouldin	1,51	1,95	1,65	2,65	0
Индекс Calinski-Harabasz	83,03	12,88	81,27	61,79	+ ∞

* eps = 0.2, min_samples = 5



- «Кластер 0» – 121 банк
- «Кластер 1» – 27 банков
- «Кластер 2» – 95 банков
- «Кластер 3» – 46 банков
- «Кластер 4» – 48 банков



Интерпретация кластеров

Анализ центров кластеров	Интерпретация	Число банков	Примеры банков
Высокая доля корпоративного кредитования в активах (44%) Равномерное распределение вкладов физлиц (30%) корпоративных средств (27%)	Большие диверсифицированные банки, с большой долей корпоративного кредитования и финансирования крупных инфраструктурных проектов	121	Сбербанк, Банк Россия, ВТБ, Газпромбанк, ДОМ.РФ, Открытие, Банк Санкт-Петербург, Совкомбанк
Высокая доля кредитования физлиц в активах (56%) Высокая доля вкладов физлиц в пассивах (42%)	Розничные банки со значимой долей средств и кредитования физически лиц, а также с частичным привлечением средств корпоративных клиентов	27	Тинькофф Банк, МТС Банк, Почта Банк, Абсолют Банк, Хоум Банк
Высокая доля выданных межбанковских кредитов в активах (71%) Высокая доля корпоративных средств в пассивах (38%) Высокая доля счетов физлиц в пассивах (26%)	банки, которые в основном размещают средства в межбанковское кредитование. Средства привлекаются от физлиц и юрлиц. В кластере много региональных и несколько иностранных банков, обслуживающих клиентов.	95	Банк Йошкар-Ола, Банк Пермь, Донкомбанк, Великие Луки банк, БЭНК ОФ ЧАЙНА, Дойче Банк, Райффайзен Банк
Значимая доля ЛОРО-счетов в активах (11%) Высокая доля выданных межбанковских кредитов в активах (50%) Значимая доля прочих активов (15%) Преобладание капитала в пассивах (77%) Низкая доля корпоративных средств (9%) и средств физлиц (в сумме 16%)	Банки со специфическими целями, которые могут служить кэптивным банком для другой организации или миноритарным банком в крупной банковской группе – среди них много иностранных банков	46	Голдман Сакс Банк, Ситибанк, HSBC, Юнистрийм, НРД, МБА-Москва, СПБ Банк, ОЗОН Банк
Высокая доля вложений в облигации в активах (39%) Наименьшая доля выданных межбанковских кредитов в активах (14%) Наименьшая доля капитала в пассивах (18%)	Банки, размещающие средства в облигациях и корпоративных кредитов. Корпоративные банки, которые также привлекают много корпоративных средств в пассивы. У этих банков низкая доля капитала	48	Банк Синара, Уралсиб, Инвестторгбанк, ФИНСТАР Банк, Локо-Банк, Металлинвестбанк, Москомбанк



Выбор лучшей модели

Варианты спецификации алгоритмов кластеризации (540 вариантов)

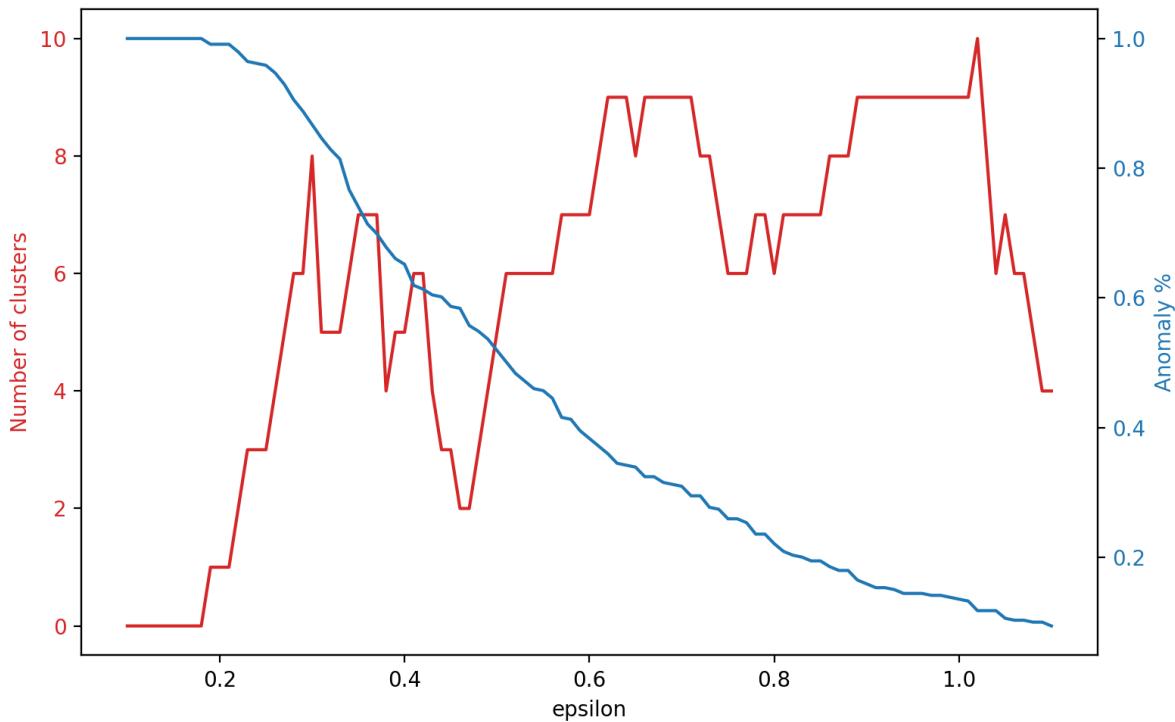
Алгоритм	K	scaling	cluster before dimreduce	pca_or_tsne	FCM fuzziness	SOM sigma_elast	SOM learnig_rate	davies_bouldin	calinski_harabasz_score	silhouette_score	f1_bcubed	prc_distributed
K-means	5	1	1					1,33	40,27	0,18	0,36	[0.5, 0.3, 0.2, 0.0, 0.0]
K-means	6	1	1					1,31	38,63	0,21	0,34	[0.5, 0.3, 0.1, 0.1, 0.0, 0.0]
DBSCAN	1	0	1					1,67	0,07	-0,38	0,98	[1.0, 0.0]
DBSCAN	1	0	1					1,67	0,07	-0,38	0,98	[1.0, 0.0]
fuzzy C-means	2	0	1		2			1,77	11,77	0,60	0,28	[0.9, 0.1]
fuzzy C-means	2	0	1		12			1,77	7,72	0,39	0,27	[0.7, 0.3]
Spectral	3	0	1					0,50	1,94	0,53	0,50	[1.0, 0.0, 0.0]
Spectral	4	0	1					0,73	1,43	0,32	0,41	[1.0, 0.0, 0.0, 0.0]
SOM	4	0	1			1	0,5	1,04	4086,31	0,27	0,23	[0.7, 0.3, 0.1, 0.0]
SOM	4	0	1			1	1	1,10	3600,53	0,49	0,26	[0.8, 0.1, 0.0, 0.0]

- K – количество кластеров в алгоритме
- scaling – применять или нет масштабирование данных до кластеризации (1 – да, 0 – нет)
- cluster before dimreduce – проведение кластеризации до понижения размерности или после (1 – до, 0 – после)
- pca_or_tsne – метод понижения размерности (PCA – Метод главных компонент или t-SNE)

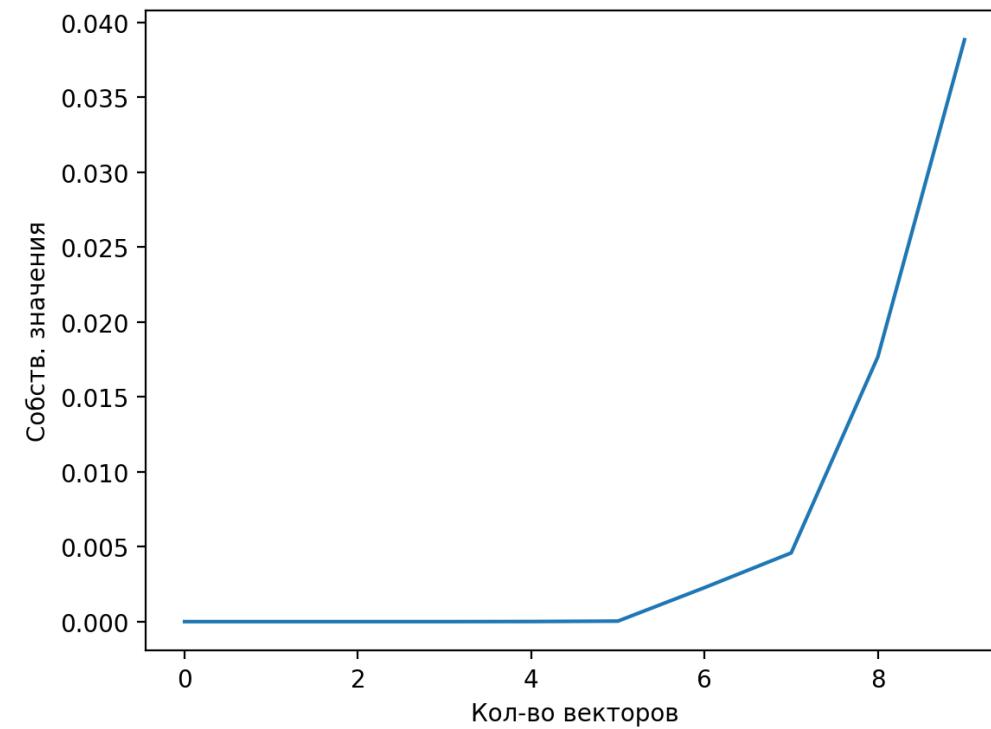


Выбор лучшей модели

Зависимость числа кластеров и процента аномалий
от параметра epsilon в алгоритме DBSCAN



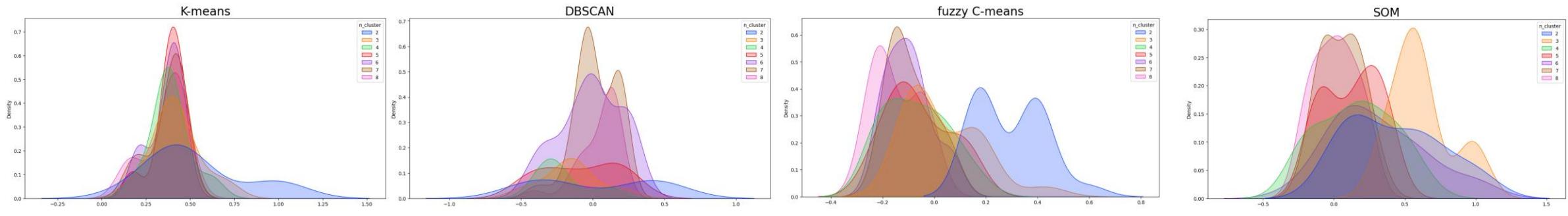
Собственные значения векторов
для спектральной кластеризации



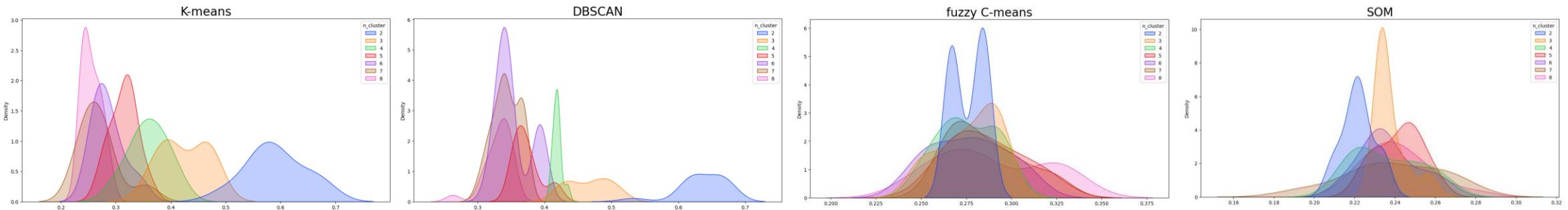


Выбор лучшей модели

Распределение метрики Силуэта в зависимости от алгоритма и числа кластеров



Распределение меры f1-bcubed в зависимости от алгоритма и числа кластеров



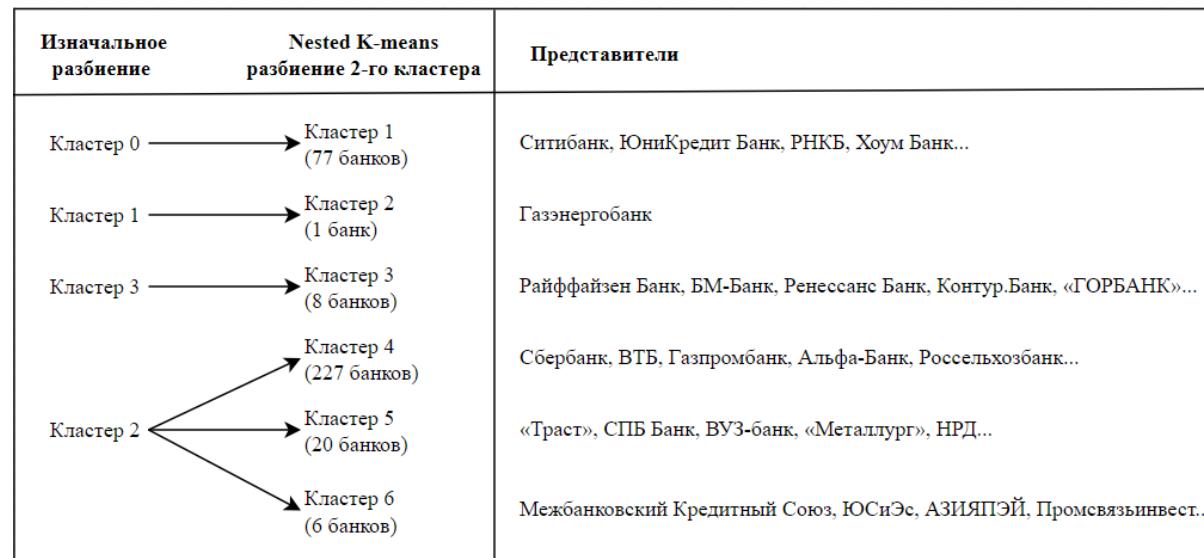


Лучшая модель

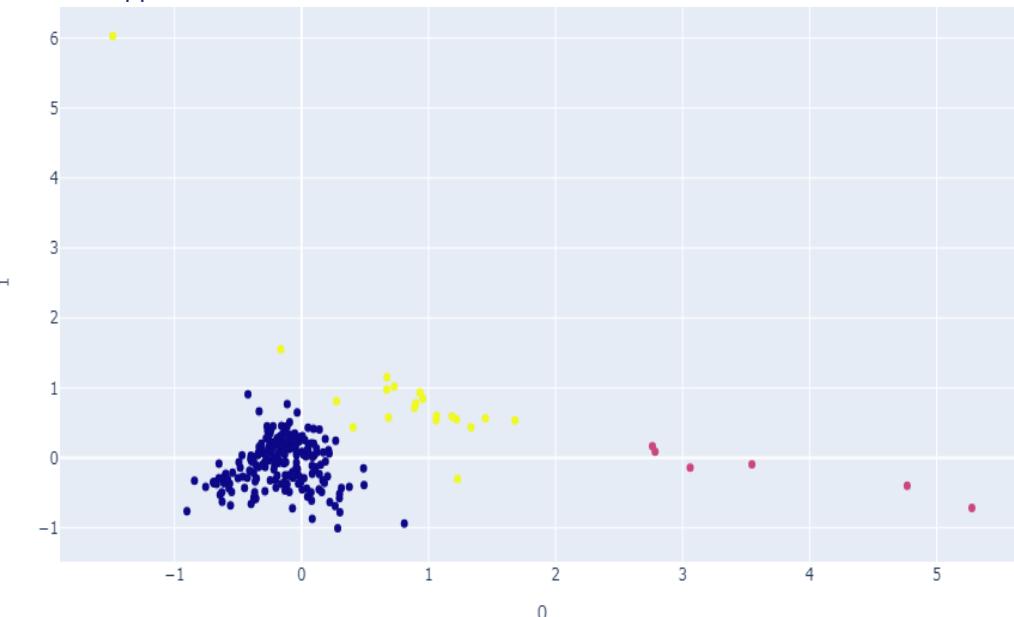
Конфигурация лучшей модели (K=4)

Алгоритм	K	scaling	cluster before dimreduce	pca_or_tsn e	FCM fuzziness	SOM sigma_elast	SOM learnig_rate	davies_bouldi n	calinski_harabasz_sc ore	silhouette_score	f1_bcubed	prc_distributed
K-means	4	0	0	pca				0,37	14334,94	0,62	0,38	[0.8, 0.2, 0.0, 0.0]

Итоговое разбиение выборки (K=6)



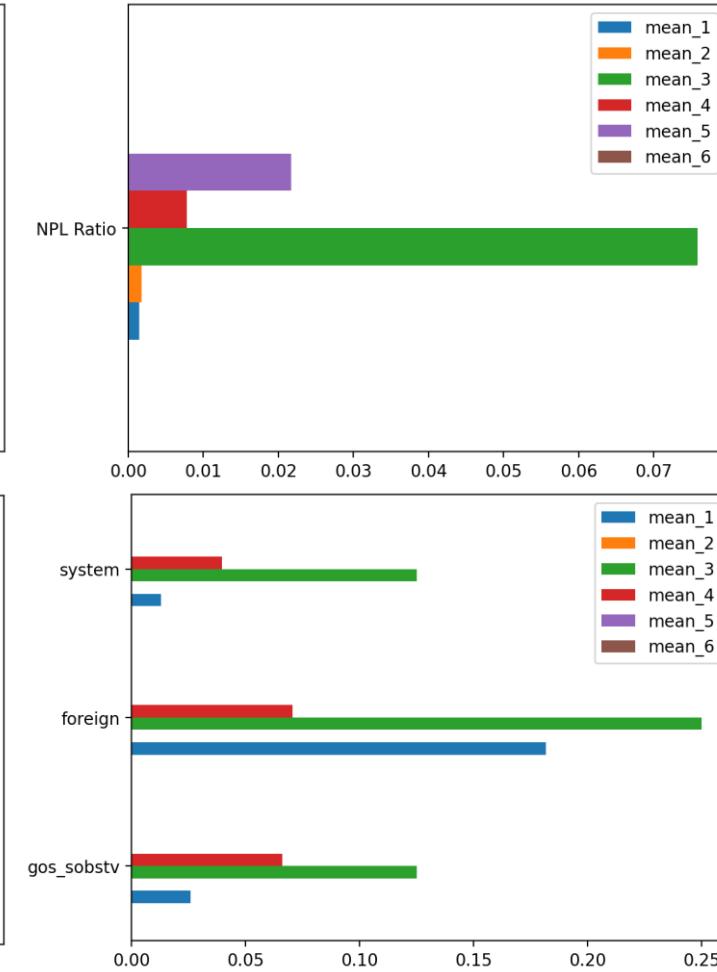
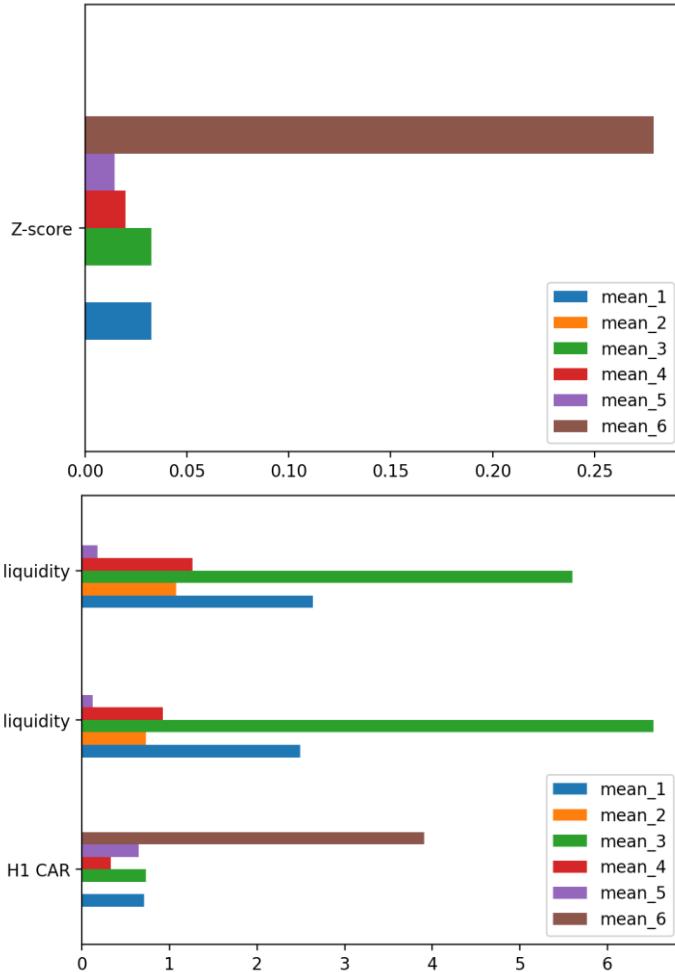
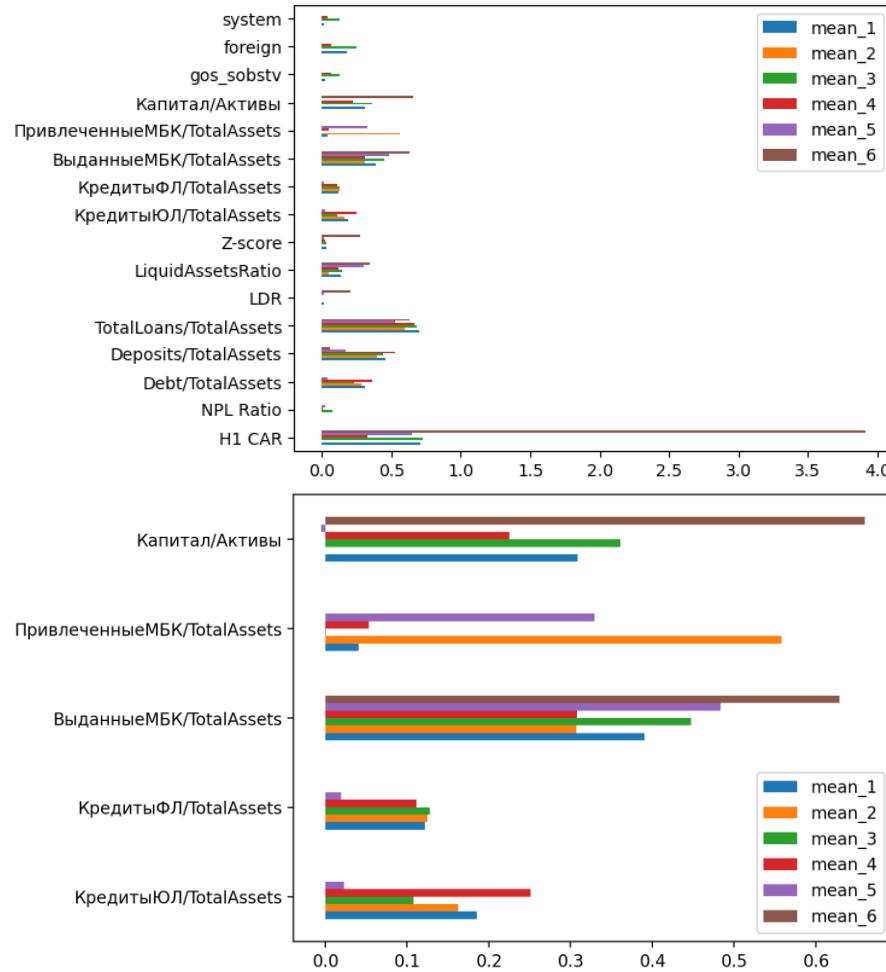
Визуализация распределения банков 2-го кластера на подкластеры методом главных компонент



Качество разбиения: DB score = 1.009, CH score = 3066, Silhouette score = 0.342, f1-bcubed = 0.289



Сравнение внутрикластерных средних значений





Новостной анализ: Поиск лучшей архитектуры

Результаты экспериментов кластеризации банков по новостям.

Nº	embed	clustering	transform	n cluster	averaging	davies bouldin	calinski harabasz	silhouette	f1 bcubed
1	tf-idf	spectr	stem	6	micro	2,47	5,16	0,09	0,35
2	tf-idf	spectr	lem	6	micro	2,45	5,12	0,08	0,35
3	tf_idf	k-means	lem	6	macro	2,41	1,88	-0,18	0,34
4	tf-idf	spectr	lem	6	macro	2,55	1,70	-0,21	0,33
5	tf-idf	fuzzy	lem	6	macro	3,57	2,10	-0,09	0,33
6	tf-idf	fuzzy	lem	6	micro	3,85	3,15	-0,17	0,33
7	tf-idf	spectr	stem	6	macro	2,37	1,63	-0,18	0,32
8	tf-idf	fuzzy	stem	6	micro	3,46	2,35	-0,12	0,32
9	bert	spectr		6	micro	1,70	14,53	0,18	0,32
10	bert	k-means		6	micro	2,07	13,63	0,08	0,31
11	tf-idf	k-means	stem	6	macro	3,09	3,30	-0,14	0,29
12	bert	k-means		6	macro	2,28	10,93	0,13	0,29
13	tf-idf	fuzzy	stem	6	macro	3,52	3,67	0,01	0,28

Источник: расчеты авторов

davies bouldin

2,47

silhouette

0,09

calinski harabasz

5,16

f1 bcubed

0,35



Новостной анализ: Spectral clustering + tf-idf

Кластер 0 (рискованные банки):

Ключевые слова: "финансовое оздоровление", "санкционный банк", "ЦБ".

6 банков: "Балтингвестбанк", "Банк Пересвет", "Банк Таврический", "Венец банк", "Газэнергобанк", "Инвестторгбанк", "Мособлбанк", "Севергазбанк"

Кластер 1 (рынок Азии):

Ключевые слова: "Китай", "иностранный", "санкция".

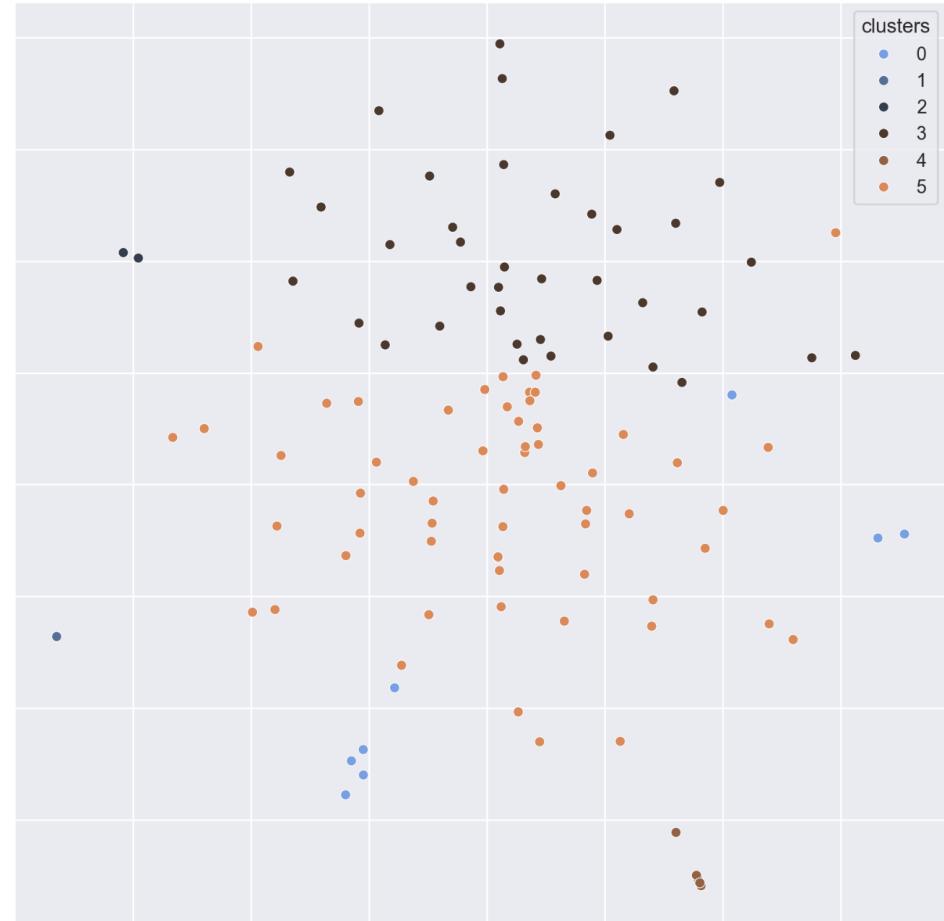
2 банка: "Акибанк" и "ББР Банк"

Кластер 2 (партнёрство авиакомпаний):

Ключевые слова: "utair", "авиаперевозки", "партнер"

2 банка: "Нико-Банк" и "Сургутнефтегазбанк"

Расположение кластеров TSNE по новостным статьям за 4 года



Источник: расчеты авторов



Новостной анализ: Spectral clustering + tf-idf

Кластер 3 (региональные банки):

Ключевые слова: "страна", "регион", "свой"

39 банков: "Банк Восточный", "Банк Держава", "Банк Екатеринбург", "Банк Казани", "Банк Левобережный", "Банк Оренбург"

Кластер 4 (автомобильные банки):

Ключевые слова: "автомобиль", "автокредитование", "машина"

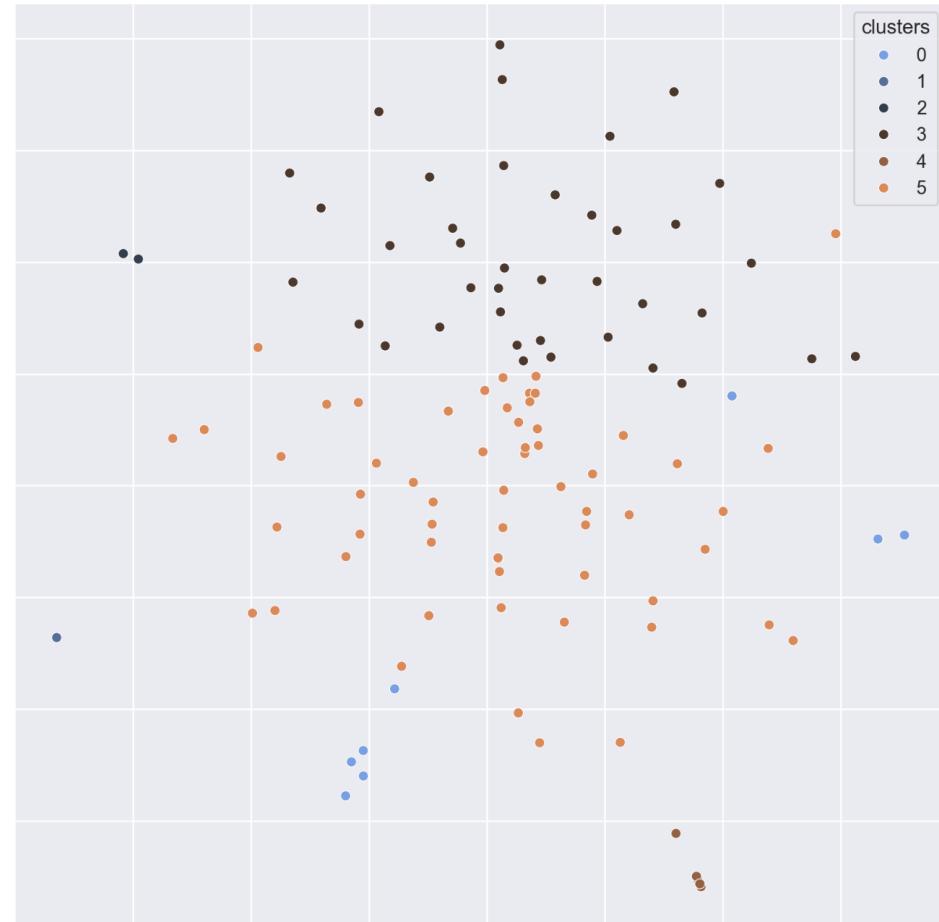
5 банков: "БМВ Банк", "МС Банк Рус", "Мерседес-Бенц Банк РУС", "Тойота Банк", "Фольксваген Банк РУС"

Кластер 5 (крупные банки):

Ключевые слова: "крупный", "биржа", "вырастать"

59 банков: "Альфа-Банк", "Банк ДОМ.РФ", "ВТБ", "Россельхозбанк", "Сбербанк"

Расположение кластеров TSNE по новостным статьям за 4 года



Источник: расчеты авторов

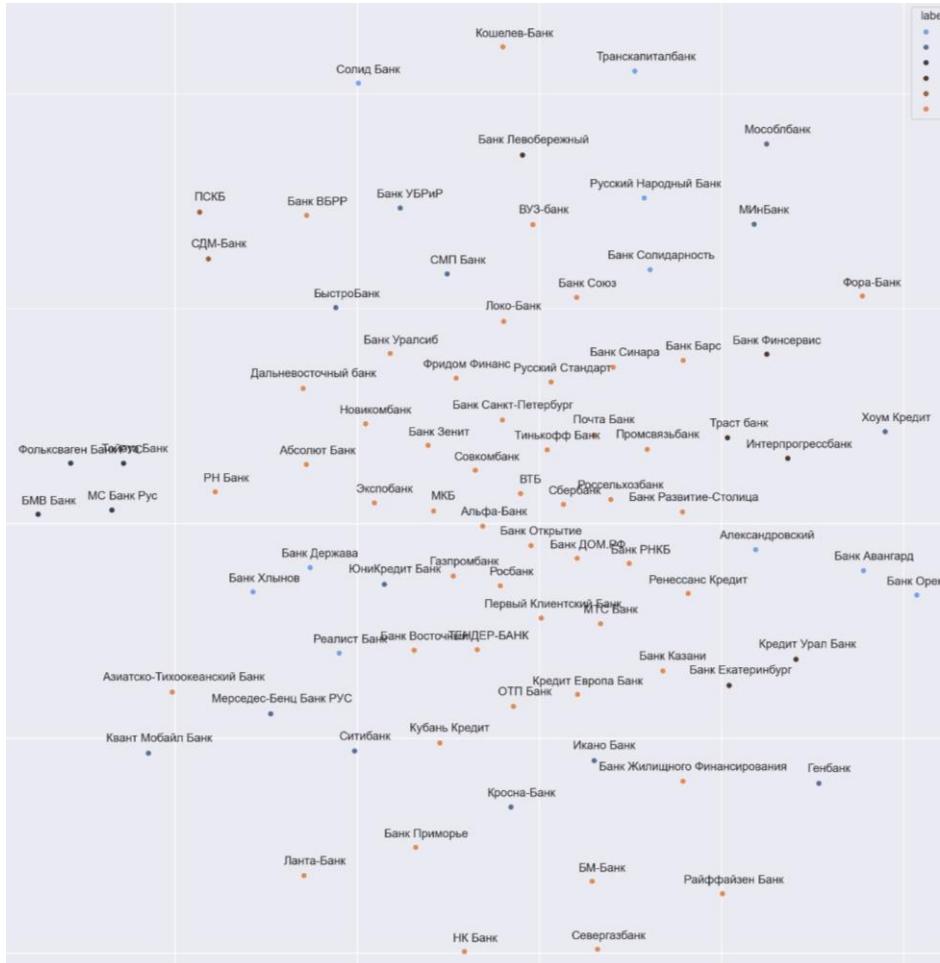


Кластерный анализ банков методами машинного обучения на данных открытых источников

Результаты

Новостной анализ: Spectral clustering + tf-idf

Расположение кластеров TSNE по новостным статьям за 2023 год

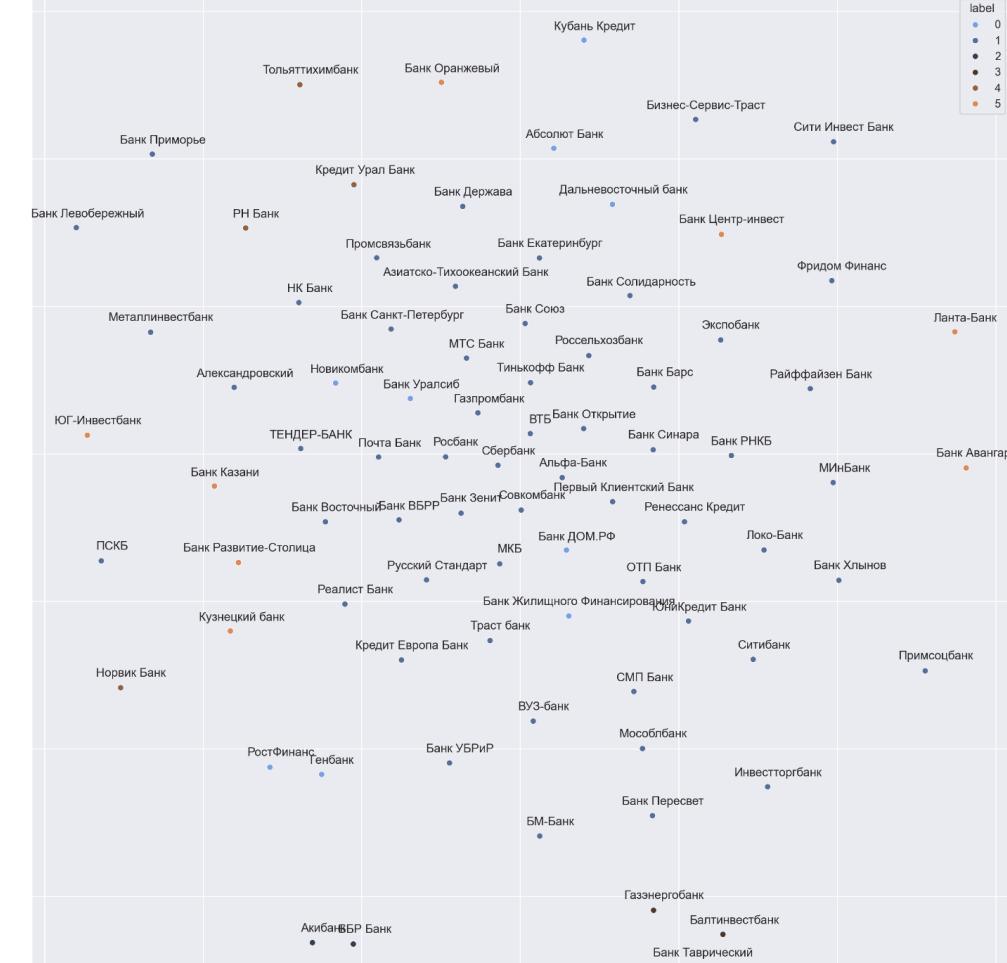


Источник: расчеты авторов

Кластерный анализ текстовых данных

39

Расположение кластеров TSNE по новостным статьям за 2022 год



Источник: расчеты авторов



Новостной анализ: LDA

Тема 1 – “Законность деятельности”:

“уголовное дело”, “УК РФ”, “адвокат”

Тема 2 – “Расширение стратегий”:

“строительство”, “актив”, “развитие”

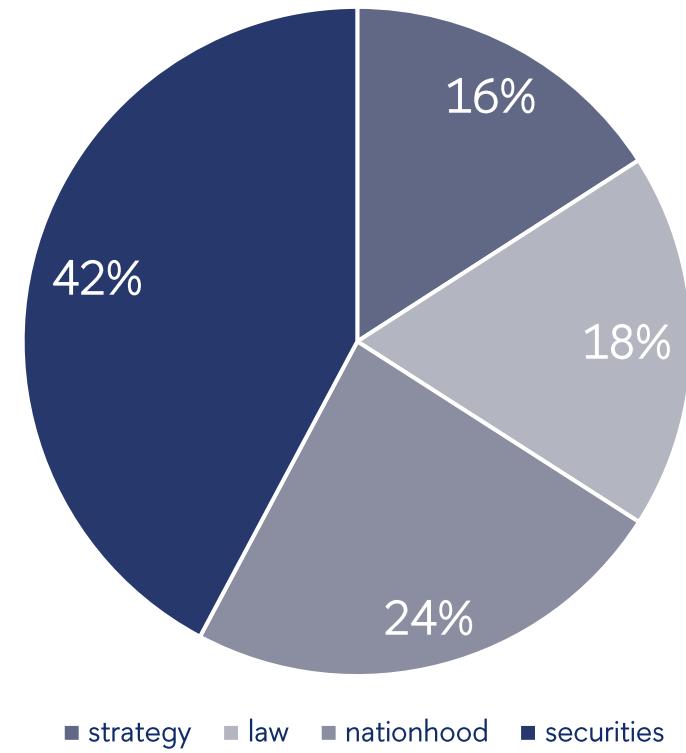
Тема 3 – “Участие государства”: “ЦБ”,

“Правительство”, “комиссия”

Тема 4 – “Финансовый рынок”: “биржа”,

“брокер”, “инвестор”

Распределение новостных тем за 2023 год



Источник: расчеты авторов



Анализ отзывов: Spectral clustering + RoBERTa

Кластер 0 (клиенты удовлетворены банком):

Средний рейтинг: 2,328
38 банков

Кластер 1 (клиенты неудовлетворены
банком):

Средний рейтинг: 1,963
80 банков

Кластер 2 (клиентам не нравится банк):

Средний рейтинг: 1,728
23 банков

Кластер 3 (клиенты довольны банком):

Средний рейтинг: 3,740
25 банков

davies bouldin silhouette calinski harabasz f1 bcubed

2,23

0,12

16,37

0,33

$$\begin{aligned}H_0: \mu_1 &= \mu_2 \\H_1: \mu_1 &\neq \mu_2\end{aligned}$$



$$\begin{aligned}\alpha &= 5\%: \\(-0.032, 0.295)\end{aligned}$$

Кластер 0 (клиенты удовлетворены банком):

Средний рейтинг: 2,328
38 банков: Нико-Банк, Ланта-Банк, Гарант-
Инвест

Кластер 1 (клиенты неудовлетворены
банком):

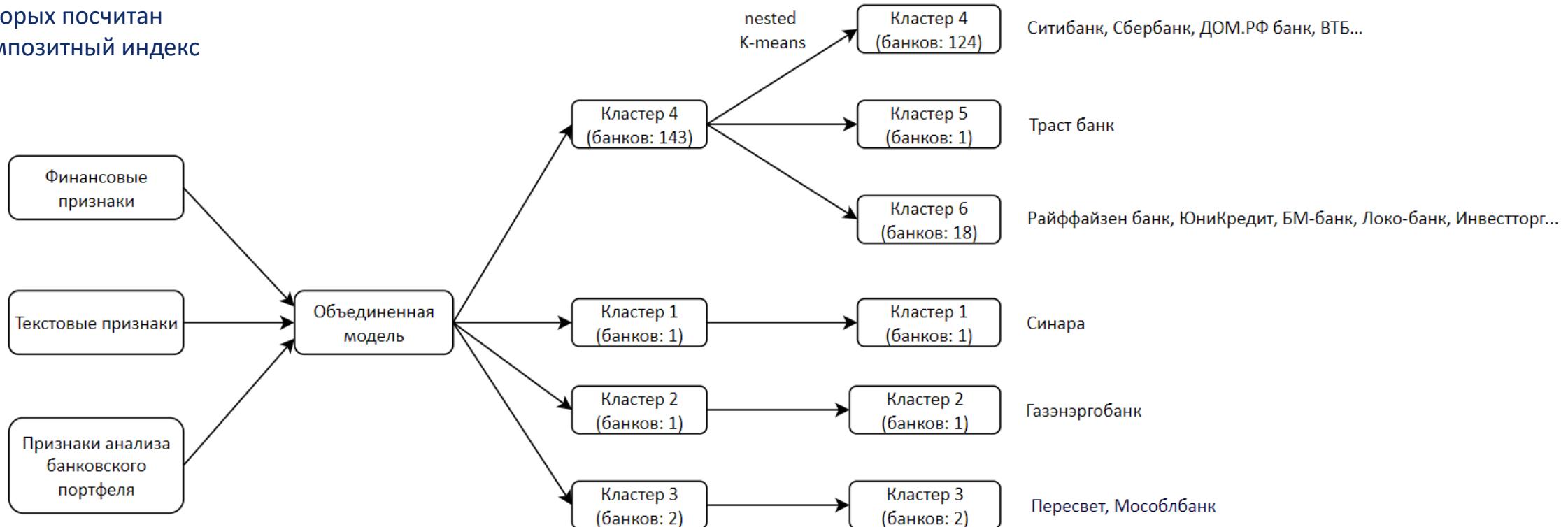
Средний рейтинг: 1,86
103 банков: Норвик Банк, Реалист Банк,
РостФинанс

Кластер 2 (клиенты довольны банком):

Средний рейтинг: 3,740
25 банков: Сбер, ВТБ, Газпромбанк, Ситибанк

147 банков, для
которых посчитан
композитный индекс

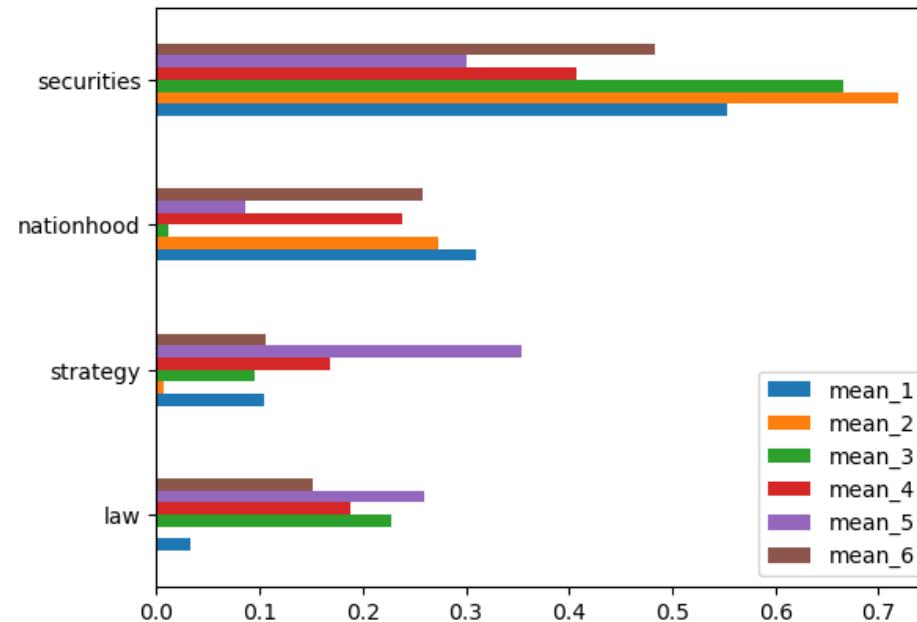
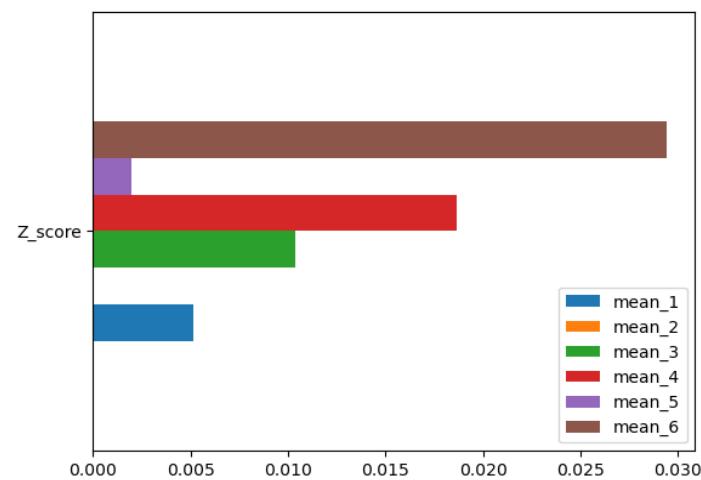
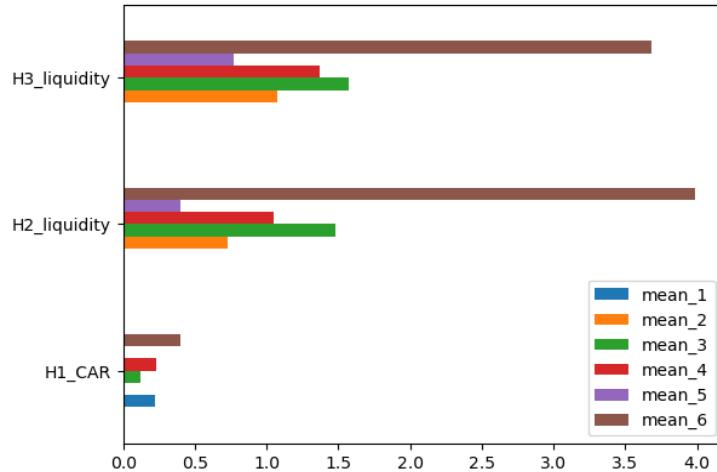
Объединенная модель



Качество разбиения: DB score = 0.379, CH score = 28120, Silhouette score = 0.475, f1-bcubed = 0.304

Улучшилось

Сравнение внутрикластерных средних значений для объединенной модели



Источник: расчеты авторов



Порядковая пробит модель

$$Y_i^* = X_i\beta + W_i\gamma + \varepsilon_i, \varepsilon_i \sim N(0, 1)$$

$$Y_i = \begin{cases} 1, & Y_i^* < c_1 \\ 2, & Y_i^* \in [c_1, c_2) \\ \dots \\ n, & Y_i^* \geq c_{n-1} \end{cases}$$

- Y_i – истинный класс i -го банка
 Y_i^* – латентная переменная i -го банка
 X_i – контрольные переменные i -го банка
 W_i – дамми переменная на предсказанный кластер i -го банка
 c_i – оцененные пороги i -го банка



Тест Бранта

H0: Parallel Regression Assumption

соблюдается

H1: Parallel Regression Assumption не

соблюдается

p-value 0.35



Parallel Regression
Assumption

	x2	df	probability
Omnibus	97,73	93,00	0,35
nps_cluster_1	8,73	3,00	0,03
nps_cluster_2	526,10	3,00	0,00
porfolio_cluster_1	1,81	3,00	0,61
porfolio_cluster_2	2,62	3,00	0,45
porfolio_cluster_3	10,21	3,00	0,02
porfolio_cluster_4	0,62	3,00	0,89
fin_cluster_1	13,26	3,00	0,00
fin_cluster_2	2,95	3,00	0,40
fin_cluster_3	2,67	3,00	0,45
fin_cluster_4	0,00	3,00	1,00
fin_cluster_5	0,00	3,00	1,00
law	7,06	3,00	0,07
securities	12,37	3,00	0,01
strategy	6,52	3,00	0,09
H1_CAR	-0,85	3,00	1,00
H2_liquidity	15,74	3,00	0,00
H3_liquidity	3,40	3,00	0,33
ROA	1,04	3,00	0,79
TotalLoans_TotalAssets	-0,98	3,00	1,00
Z_score	3,44	3,00	0,33
gos_sobstv	29,07	3,00	0,00
foreign	4,32	3,00	0,23
system	0,00	3,00	1,00
A_Shares	2,90	3,00	0,41
A_bonds	-15,43	3,00	1,00
A_capitals	80,39	3,00	0,00
A_loro_loans	-6,78	3,00	1,00
A_fixed_assets	25,18	3,00	0,00
P_deposits_individuals	10,11	3,00	0,02
P_corporate_funds	3,95	3,00	0,27
P_capitals	10,30	3,00	0,02



	orlogit		orprobit		binprobit		Im	
	Estimate	Std Error	Estimate	Std Error	Estimate	Std Error	Estimate	Std Error
nps_cluster_1	0,02	0,31	0,04	0,26	-0,30	0,35	1,42	2,26
nps_cluster_2	0,13	0,43	0,00	0,40	-0,80	0,59	-0,08	3,38
porfolio_cluster_1	-0,90 **	0,31	-0,77**	0,29	-0,91*	0,43	-8,41**	2,54
porfolio_cluster_2	-0,81 .	0,41	-0,95*	0,38	-1,05*	0,51	-7,01 .	2,96
porfolio_cluster_3	-0,74	1,34	-0,59	1,01	0,02	1,34	-1,09	8,81
porfolio_cluster_4	-0,72	0,48	-0,63	0,47	-0,79	0,60	-4,05	4,2
fin_cluster_1	-0,18	1,06	-0,32	0,96	1,03	1,10	0,83	7,58
fin_cluster_2	0,58	0,57	0,58	0,48	0,58	0,59	5,04	4,63
fin_cluster_3	-0,15	0,39	-0,07	0,35	-0,18	0,49	-1,45	3,35
fin_cluster_4	-9,61***	0,90	-9,61***	1,07	-5,27	414,65	-13,86 *	6,97
fin_cluster_5	12,38***	1,23	12,38	476,71	8,05	973,50	20,47 *	9,52
law	-0,50	0,77	-0,53	0,73	0,05	0,86	6,51	6,95
securities	-1,58*	0,67	-1,65*	0,65	-1,08	0,90	-16,89**	5,67
strategy	0,23	0,66	0,01	0,66	1,02	0,98	-0,95	4,94
AIC	436,55		441,20		188,90		1162,04	
BIC	541,22		545,87		284,59		1260,73	

Коды для уровня значимости: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 '' 1

Источник: расчеты авторов



Гетероскедастичность

В каждом кластере своя дисперсия ошибок.

Наличие той или иной темы в новостях влияет на дисперсию ошибок.

Het ordered probit:

$$Y_i^* = X_i\beta + W_i\gamma + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_i)$$

$$\sigma_i = e^{W_i\tau}$$

LR-тест:

$$2(LnL - LnL_{H0}) \sim \chi^2_{13}$$

p-value = 0.425.

Наличие гетероскедастичности от переменных кластеров и наличие тем в новостных статьях не было выявлено.



Средние предельные эффекты значимых изучаемых переменных на вероятность
попадание в конкретную группу надежности банка для порядковой логит модели.

	Group_0	Group_1	Group_2	Group_3	Group_4
porfolio_cluster_1	0.18**	0.09	-0.02	-0.1**	-0.15**
porfolio_cluster_2	0.15	0.08	-0.01	-0.09*	-0.13*
fin_cluster_4	0.82	-0.18***	-0.22***	-0.22***	-0.23***
fin_cluster_5	-0.18***	-0.18***	-0.20***	-0.22***	0.78***
securities	0.31*	0.13*	-0.02	-0.15 .	-0.26 .

Коды для уровня значимости: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Источник: расчеты авторов



Общие гипотезы

H1

Банки можно поделить на кластеры по риск-профилю.

Гипотеза **не отвергается**, так как есть кластеры, которые статистически значимо влияют на переменную надежности банка.

H2

Деление банков на кластеры с помощью разных методов дает схожие результаты.

Гипотеза **не отвергается**, о чем свидетельствуют получившиеся кластеры.

H3

При объединении типов данных кластеры формируются качественнее.

Гипотеза **не отвергается**. Метрики качества кластеризации имеют более высокие значения на объединенном датасете со всех трех направлений кластеризации.



Банковские портфели

H1 Рискованность банковской деятельности снижается с диверсификацией продуктов и групп клиентов.

Гипотеза **не отвергается**. Кластеры с банками из более диверсифицированных портфелей являются более надежными.

H2 Продуктовая полка банка должна отличаться от средневзвешенного портфеля по рынку для минимизации системных рисков.

Гипотеза **отвергается**. Среди значимых кластеров более приближенные к среднему портфелю банки являются более надежными.



Банковские портфели

portfolio_cluster_1 и portfolio_cluster_2 оказывают схожее влияние на группы, а также имеют схожий коэффициент.

$$H_0: \gamma_{\text{portfolio_cluster_1}} = \gamma_{\text{portfolio_cluster_2}}$$

$$H_1: \gamma_{\text{portfolio_cluster_1}} \neq \gamma_{\text{portfolio_cluster_2}}$$

LR-тест:

$$2(LnL - LnL_{H_0}) \sim \chi^2_1$$

- p-value = 0.6804 на порядковой пробит модели
- p-value = 0.8512 на порядковой логит модели

Основная гипотеза **не отвергается**, а значит данные два кластера выделяют одинаковый сегмент банков с точки зрения надежности.



Финансовые показатели

H1

Банки с повышенными показателями доходности имеют выше показатель финансовой стабильности.

Гипотеза **отвергается**. Среди значимых кластеров по финансовым показателям более надежными оказались банки с меньшими уровнями ROE, ROA.

H2

Банки с волатильными показателями доходности менее финансово устойчивы.

Гипотеза **не отвергается**. Действительно, среди значимых кластеров, группы банков с большим разбросом ROA и ROE имеют более низкий уровень надежности.



Текстовые данные

H1

Банки из кластера положительных настроений имеют ниже показатель финансовой стабильности.

Гипотеза **отвергается**, так как получившиеся кластеры не значимо влияют на переменную надежности.

H2

Банки, в новостях которых чаще затрагивается тема финансовых рынков, менее финансово устойчивы.

Гипотеза **не отвергается**, так как присутствие темы финансовых рынков значимо отрицательно влияет на переменную надежности.



Итог работы

Что сделано:

1. Изучена академическая литература
2. Собраны, обработаны и проанализированы данные для трех направлений работы
3. Применены и выбраны модели кластеризации
4. Получены и проинтерпретированы кластеры банков
5. Кластеры проверены на статистическую значимость
6. Собрана объединенная модель с лучшими результатами
7. Все поставленные гипотезы были проверены
8. Описание исследования и исходный код переданы представителям ЦБ

Ограничения работы:

- Ограниченный круг российских банков
- Пропуски и методы их заполнения
- Не использованы иные профессиональные источники
- Нет детализации показателей
- Не учтены показатели фондовых рынков
- Не применен голосующий ансамбль и другие методы кластеризации

Улучшения возможны в процессе работы службы анализа рисков ЦБ и будущих исследованиях



Бочкарев Сергей. Кластеризация банковских портфелей



Мамедов Ильгар. Кластеризация текстовых данных



Лашманов Валентин. Кластеризация по фин. показателям



