УДК 330.46 EDN: TIULIP

Лепило Н. Н., Катан К. С.
Донбасский государственный технический университет
**E-mail: nnlepilo@mail.ru*

МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ ДЛЯ АНАЛИЗА ЭКОНОМИЧЕСКИХ ДАННЫХ

В статье рассмотрены и проанализированы современные методы снижения размерности, особенности их использования для анализа данных. В качестве примера реализации методов на языке Python использован известный набор данных, содержащий информацию о клиентах банка и параметрах проводимой с ними маркетинговой кампании. Разработаны рекомендации по использованию методов в сфере анализа экономических данных.

Ключевые слова: *снижение размерности, анализ данных, набор данных, сингулярное разложение, независимые компоненты, многомерное масштабирование.*

Проблема и ее связь с научными и практическими задачами. В современных условиях рыночной экономики и ее цифровой трансформации резко возрастают объемы накопленной информации, что связано с развитием цифровых технологий и автоматизацией бизнес-процессов [1, 2]. С ростом количества данных усложняется извлечение и визуализация необходимой информации. Методы снижения размерности данных позволяют уменьшить количество измерений (столбцов), обеспечивая возможность сохранения наиболее важной информации. При этом возможна потеря некоторых деталей, но итоговая структура данных становится более простой и удобной для анализа и сопоставления.

Актуальность применения методов сокращения размерности при анализе экономической информации можно объяснить следующими причинами:

– рост объемов экономических данных создает значительные трудности их обработки, анализа и визуализации;

– уменьшение числа признаков приводит к упрощению вычислений за счёт удаления мультиколлинеарности, сокращая время и ресурсы, затрачиваемые на анализ и хранение необходимой информации;

– снижение размерности способствует выявлению скрытых закономерностей в данных.

Постановка задачи. ***Целью*** статьи является анализ современных методов снижения размерности и разработка рекомендаций по их применению в сфере анализа экономических данных.

Для достижения поставленной цели необходимо решение следующих ***задач***:

– рассмотреть и проанализировать существующие методы снижения размерности;

– рассмотреть реализацию этих методов с помощью современных программных продуктов;

– разработать рекомендации по использованию рассмотренных методов в сфере анализа экономических данных.

Методика исследования. В работе использованы методы системного подхода, моделирования, анализа данных, машинного обучения, статистические методы.

Изложение материала. Будем считать сложным набор данных, содержащий более девяти атрибутов, поскольку американским психологом Д. Миллером обнаружена закономерность, что человек способен держать в памяти не более 9 элементов [3]. При работе с большим количеством информации ее необходимо структурировать.

Цель использования методов снижения размерности — упростить сложные наборы данных, уменьшив количество признаков (атрибутов, переменных) при сохранении

важной информации. При рассмотрении методов снижения размерности использован набор данных bank.csv — известный набор данных, содержащий информацию о клиентах и параметрах проводимой с ними маркетинговой кампании [4]. Все атрибуты этого набора данных можно разделить на группы.

Первая группа включает данные о клиенте (демография и финансовое поведение):

– age — возраст клиента (целое число);

– job — сфера занятости (категориальный);

– marital — семейное положение;

– education — уровень образования — primary (начальное), secondary (среднее), tertiary (высшее), unknown (неизвестный);

– default — имеется ли просроченная задолженность (yes, no);

– balance — средний годовой баланс на счете (в евро). Может быть отрицательным;

– housing — есть ли ипотечный кредит (yes, no);

– loan — есть ли персональный кредит (yes, no).

Вторая группа включает данные о текущей маркетинговой кампании:

– contact — тип средства связи — cellular (мобильный), telephone (стационарный);

– day — последний день контакта (число месяца);

– month — последний месяц контакта (jan, feb, mar, ..., dec);

– duration — длительность последнего контакта в секундах. Этот атрибут сильно влияет на целевую переменную (чем дольше разговор, тем выше шанс успеха), но его использование для прогнозирования в реальной жизни некорректно, так как длительность звонка неизвестна до его начала. Его часто исключают или используют с осторожностью.

Третья группа включает следующие данные о предыдущих маркетинговых кампаниях и активностях текущей кампании:

– campaign — количество контактов с клиентом (с учетом последнего контакта) в ходе текущей маркетинговой кампании;

– pdays — число дней, прошедших с момента последнего контакта с клиентом в предыдущей кампании. Если клиент раньше не контактировался, то указывается –1;

– previous — количество контактов с клиентом до текущей маркетинговой кампании;

– poutcome — результат предыдущей кампании — success (успех), failure (неудача), other (другое), unknown (неизвестное).

Предварительный анализ рассматриваемого набора данных показал, что он содержит 7 числовых признаков и 9 категориальных, т. е. большая часть информации представлена в категориальном виде. Целевой переменной является бинарная переменная deposit — подписался ли клиент на срочный вклад (yes, no).

Такая ситуация характерна для большинства экономических данных — как правило, большая часть информации представлена в категориальном виде. Для набора данных, в котором есть категориальные переменные, даже обычный корреляционный анализ выполнить нельзя, так как линейную связь между категориальными переменными (факторами, которые нельзя представить в виде числа) измерить невозможно.

Для реализации алгоритмов снижения размерности будем использовать язык Python, специализированные библиотеки которого содержат инструменты, облегчающие их практическое применение [4].

Методы снижения размерности подразделяются на линейные и нелинейные. Линейные методы направлены на поиск исходных закономерностей и линейных зависимостей в данных, позволяя представить их в пространстве меньшей размерности. Среди наиболее популярных методов этой группы можно выделить:

– метод главных компонент (Principal Component Analysis, PCA);

– метод усеченной декомпозиции сингулярных значений (Truncated Singular Value Decomposition, TruncatedSVD);

– анализ независимых компонент (Independent Computing Architecture, ICA).

Метод PCA трансформирует сложный набор данных в простое представление, выделяя наиболее значимые направления вариации [5]. В результате использования метода выделяются главные компоненты — линейные комбинации исходных признаков, упорядоченные по убыванию объясняемой дисперсии. Дисперсия используется в качестве показателя для оценки разброса значений в столбце. Объекты с более высокой дисперсией содержат больше информации, тогда как нулевая дисперсия указывает на полное отсутствие информативности — именно поэтому данный показатель имеет ключевое значение. Первая главная компонента захватывает максимальную дисперсию данных, вторая — максимум оставшейся дисперсии и т. д.

Принцип работы метода PCA основан на поиске компактного представления исходных данных в пространстве меньшей размерности. При этом ключевой критерий — максимизация общей дисперсии данных, то есть сохранение максимально возможного объёма информации об изменчивости данных. В результате структура данных по-прежнему сохраняется, но их описание существенно упрощается за счёт сокращения размерности. Особенностями метода являются:

– первые несколько компонент сохраняют большую часть информации (часто 80–95 %);

– PCA работает только с линейными взаимосвязями между переменными;

– требует предварительной стандартизации данных для корректных результатов.

Метод TruncatedSVD основан на сингулярном разложении и сохраняет только ряд верхних сингулярных значений и соответствующие им сингулярные векторы, обрезая остальные [6]. Будучи тесно связанным с PCA, данный метод может трактоваться как его разновидность. При этом он эффективнее справляется с задачей преобразования в плотные представления разреженных матриц, содержащих много нулевых элементов. Это аналог PCA, но для нечисловых данных, особенно после их преобразования в разреженную матрицу (например, через One-Hot Encoding). В отличие от PCA, TruncatedSVD не требует центрирования данных и работает напрямую с разреженными матрицами. Он является особенно эффективным методом уменьшения размерности при обработке больших наборов данных и отлично подходит для категориальных данных.

Его основные преимущества:

– может использоваться с разреженными и плотными матрицами;

– помогает снизить влияние шума или избыточности в данных, что улучшает точность моделей машинного обучения.

Из недостатков метода можно отметить:

– не работает хорошо с данными, имеющими сложные нелинейные взаимосвязи между признаками;

– может быть вычислительно дорогостоящим, особенно для больших наборов данных.

Основные результаты работы программы, реализующей методы PCA и TruncatedSVD на примере набора данных bank.csv, показаны на рисунке 1, а их визуализация — на рисунке 2. Установлено, что оба метода выделяют в качестве самого важного фактора историю взаимодействия с клиентом. PCA дополнительно выявляет социально-экономический профиль: возраст, баланс, семейное положение — влияющий на поведение, что важно для таргетирования. TruncatedSVD больше реагирует на категориальные признаки, чем на числовые. Объясненная дисперсия у этого метода составила 100 %. Низкая суммарная объяснённая дисперсия у PCA (28 %) говорит о том, что данные имеют нелинейную структуру.

Преимуществами метода являются выявление кластеров, уменьшение размерности, гибкость. Из недостатков можно отметить субъективность результатов, поскольку они зависят от выбора мер близости, чувствительность к выбросам и вычислительную сложность.

```
Числовые признаки: ['age', 'balance', 'day', 'duration', 'campaign', 'pdays',
'previous']
Категориальные признаки: ['job', 'marital', 'education', 'default', 'housing', 'loan',
'contact', 'month', 'poutcome']
✅ Размер закодированной матрицы: (11162, 42)

🚀 Выполняется PCA...
Объяснённая дисперсия (PCA): [0.16580939 0.11623407]
Суммарно: 28.2%
🚀 Выполняется TruncatedSVD...
Объяснённая дисперсия (TruncatedSVD): [0.58392152 0.41607848]
Суммарно: 100.0%

🔍 Анализ влияния признаков:

Топ-5 признаков по модулю нагрузки в PCA:
  PC1: ['pdays', 'previous', 'poutcome_unknown', 'campaign', 'day']
  PC2: ['age', 'balance', 'marital_single', 'marital_married', 'job_retired']

Топ-5 признаков по модулю нагрузки в TruncatedSVD:
  SV1: ['poutcome_unknown', 'marital_married', 'housing_yes', 'pdays',
'education_secondary']
  SV2: ['pdays', 'previous', 'housing_yes', 'marital_married', 'education_secondary']

Корреляция между компонентами:
  PC1 vs SV1: -0.929
  PC2 vs SV2: -0.119
⚠️ Методы выделяют разные аспекты данных
```

Рисунок 1 — Основные результаты работы программы, реализующей методы PCA и TruncatedSVD

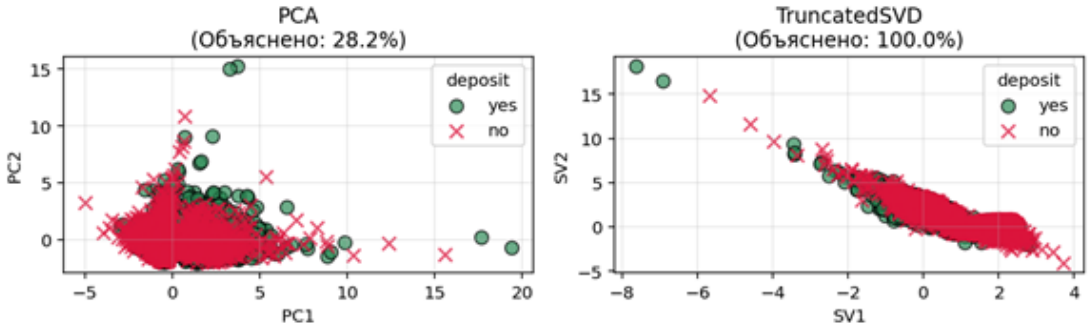

Рисунок 2 — Визуализация результатов

Анализ независимых компонент направлен на то, чтобы разделить смешанные сигналы на их первоначальные источники [7]. В рамках ICA принимается, что эти источники не зависят друг от друга, то есть воздействие одного на другой отсутствует. Выбрав нужное число независимых компонентов, их можно использовать как компактное представление данных. При этом каждый компонент фиксирует свой уникальный аспект данных, что и позволяет сократить размерность представления. Основные результаты работы программы, реализующей метод ICA, показаны на рисунке 3.

Преимуществами метода являются выявление кластеров, уменьшение размерности, гибкость. Из недостатков можно отметить субъективность результатов, поскольку они зависят от выбора мер близости, чувствительность к выбросам и вычислительную сложность.

```
ICA завершён. Найдено 4 независимых компонент.
🔍 Значимые нагрузки в компонентах (по модулю > 0.3):
IC1:
  loan: 0.434
  balance: -0.402
  campaign: 0.378
  default: 0.328
IC2:
  month: -0.427
  housing: -0.404
  contact: -0.378
IC3:
  poutcome: 0.384
  pdays: -0.380
  previous: -0.315
IC4:
  marital: 0.508
  age: -0.504

🧠 Интерпретация компонент (на основе нагрузок):
  IC1: Финансовая активность: долговая нагрузка, активность кампании
  IC2: Сезонность и коммуникация: время контакта, тип связи
  IC3: История взаимодействий: успех предыдущей кампании
  IC4: Демографический профиль: возраст и семейное положение
```

Рисунок 3 — Основные результаты работы программы, реализующей метод ICA

ICA выделил четыре независимых компоненты, каждую из которых можно интерпретировать как важный аспект поведения клиента. Эти компоненты позволяют:

– IC1 — сегментировать клиентов по финансовому состоянию (рискованные и стабильные);

– IC2 — оптимизировать время и канал связи (звонить летом, использовать мобильную связь);

– IC3 — прогнозировать открытие депозита, поскольку наиболее важный признак — прошлый успех;

– IC4 — персонализировать предложения (молодым одиноким — одни, семейным — другие).

Преимуществами метода ICA являются выделение независимых источников в смешанных сигналах, очистка данных от шума и артефактов, выявление аномалий, улучшение интерпретации данных. Из недостатков можно отметить большой объем вычислений и предположения о том, что источники негауссовы и смешиваются линейно.

Методы уменьшения нелинейной размерности нацелены на выявление и сохранение сложных нелинейных взаимосвязей в данных при их проецировании в пространство меньшей размерности. Далее рассмотрены три наиболее распространённых подхода к нелинейному уменьшению размерности в трёхмерном пространстве.

Метод стохастического соседского вложения с t-распределением (t-Distributed Stochastic Neighbor Embedding, t-SNE) преобразует сходства между данными в значения вероятностей и в дальнейшем стремится минимизировать расхождение между распределениями вероятностей в высокоразмерном и низкоразмерном пространствах. Этот алгоритм применяется для упрощения и наглядного представления сложных многомерных данных. Он достигает поставленной цели с помощью сравнения степени сходства между точками данных в исходном высокоразмерном пространстве и целевом низкоразмерном. На основе этих сравнений формируется вероятностное распределение, которое количественно отражает выявленные сходства, стараясь их сделать максимально похожими. Алгоритм действует итеративно,

корректируя расположение точек данных в пространстве низкой размерности до тех пор, пока оно не станет максимально приближённым к распределению в исходном пространстве высокой размерности [8].

Основными особенностями метода являются:

– t-SNE сохраняет локальную структуру данных (похожие точки в высокомерном пространстве остаются близко в низкоразмерном пространстве);

– эффективен при визуализации данных высокой размерности, но может искажать глобальную структуру данных, поскольку не учитывает линейные зависимости, а лишь их близость в исходном пространстве.

Метод равномерного приближения и проекции (Uniform Manifold Approximation and Projection, UMAP) строит граф ближних соседей на основе высокоразмерных данных и проектирует его в низкоразмерное пространство, минимизируя разницу между локальными структурными отношениями [8]. Метод можно рассматривать как более мощного родственника t-SNE. В отличие от t-SNE, он не только выявляет нелинейные отображения, позволяющие сохранить целостность кластеров, но и делает это с более высокой скоростью. Кроме того, UMAP демонстрирует лучшее сохранение глобальной структуры данных по сравнению с t-SNE. Глобальная структура — это степень близости между похожими сегментами данных (например, группами клиентов). Она отражает, насколько «похожи» друг на друга разные кластеры в уменьшенном пространстве. При этом локальная структура характеризует степень кластеризации одного и того же сегмента клиентов в уменьшенном пространстве.

Особенностями метода являются:

– UMAP сохраняет глобальную структуру данных, в отличие от t-SNE;

– может применяться на больших наборах данных, например, с более чем 1 млн. измерений.

Преимуществами метода являются выявление кластеров, уменьшение размерности, гибкость. Из недостатков можно отметить субъективность результатов, поскольку они зависят от выбора мер близости, чувствительность к выбросам и вычислительную сложность. Основные результаты работы программы, реализующей методы UMAP и t-SNE, показаны на рисунке 4.

Реализация этих методов показала, что значение среднего межклассового расстояния у t-SNE составило 74,29, а у UMAP — 6,77, что свидетельствует о том, что t-SNE лучше разделяет классы. Это объясняется тем, что t-SNE оптимизирует локальную структуру, пытаясь максимально отделить кластеры. UMAP, хотя и сохраняет более глобальную топологию данных, в данном случае даёт менее выраженную сепарацию между классами. В обоих методах первая компонента отделяет клиентов без истории взаимодействия (poutcome_unknown (неизвестный посетитель), pdays = –1) от тех, кто уже участвовал в кампаниях.

Признаки, влияющие на первую компоненту, в UMAP и t-SNE имеют противоположные знаки корреляции, поскольку методы UMAP, t-SNE и другие алгоритмы снижения размерности не фиксируют направление осей. UMAP использует графы ближайших соседей, t-SNE — вероятности парных расстояний — они могут выбрать противоположное направление. Знаки в снижении размерности не имеют абсолютного значения — они зависят от направления оси, которое выбирается случайно. Направление оси различается, но логическая структура совпадает. Если сравнить основные признаки по модулю корреляции, то одни и те же признаки доминируют в обеих моделях.

Вторая компонента («Личностный профиль») выделяет в качестве ключевого фактора семейное положение, а также влияние возраста. В итоге более склонны к открытию депозита пожилые, состоятельные, женатые клиенты, часто из категории пенсионеров, а менее склонны — молодые, холостые, возможно менее финансово стабильные.

```
Среднее межклассовое расстояние:
  UMAP:  6.770
  t-SNE: 74.290
✅ t-SNE лучше разделяет классы

🔍 Интерпретация компонент через корреляцию:

Топ-5 признаков по корреляции с UMAP:
  Компонента 1: poutcome_unknown(0.89), pdays(-0.78), previous(-0.57),
contact_unknown(0.51), poutcome_success(-0.48)
  Компонента 2: month_may(-0.53), age(0.53), marital_single(-0.52),
marital_married(0.43), job_retired(0.35)

Топ-5 признаков по корреляции с t-SNE:
  Компонента 1: poutcome_unknown(-0.79), pdays(0.72), contact_unknown(-0.54),
previous(0.52), poutcome_success(0.41)
  Компонента 2: marital_single(-0.66), age(0.59), marital_married(0.54),
job_retired(0.35), campaign(0.28)
```

Рисунок 4 — Основные результаты работы программы, реализующей методы UMAP и t-SNE

Метод многомерного масштабирования (Multidimensional Scaling, MDS) применяется для визуализации различия или сходства между наблюдениями в наборе данных. MDS пытается найти проекцию данных, которая минимизирует различия между расстояниями в исходном пространстве и расстояниями в низкомерном пространстве [9, 10]. Похожие наблюдения в данном представлении располагаются ближе друг к другу, а отличающиеся находятся на большом удалении. Многомерное масштабирование сочетает преимущества линейных и нелинейных методов снижения размерности, в зависимости от выбранных параметров и алгоритмов. При любом подходе оно нацелено на сохранение расстояния между точками данных, обеспечивая сохранение этих расстояний с уменьшением размерности.

Существуют следующие основные типы MDS:

– классический — предполагает метрические расстояния, неприменим для оценок прямого различия;

– метрический — расстояния между двумя точками на выходе устанавливаются как можно ближе к данным сходства или несходства;

– неметрический — фокусируется на упорядочивании данных. Алгоритмы пытаются сохранить порядок расстояний и ищут монотонную связь между расстояниями в пространстве и сходством или несходством.

В ходе реализации получена несходимость неметрического MDS за 300 итераций. Это довольно распространённая ситуация, особенно на реальных данных, таких как используемый набор данных. Этот метод часто требует более 1000 итераций, что приводит к долгому выполнению, и всё равно может не сойтись полностью, поэтому при корректной работе метрического MDS его лучше не использовать.

Для оценки качества MDS использованы показатели коэффициентов детерминации и ранговой корреляции. В результате применения метода получены умеренные корреляции (по модулю меньше 0,6), ни один признак не доминирует. MDS — нелинейный метод, поэтому интерпретация через корреляцию — упрощение. На основе двух компонент можно выделить четыре типа клиентов:

– молодые, одинокие, с низким балансом, звонят в начале месяца, долго разговаривают;

– пожилые, состоятельные, с историей неудачных контактов, редко звонят;

– молодые, активные, с короткой историей взаимодействия;

– пожилые, с долгой историей, возможно, постоянные клиенты.

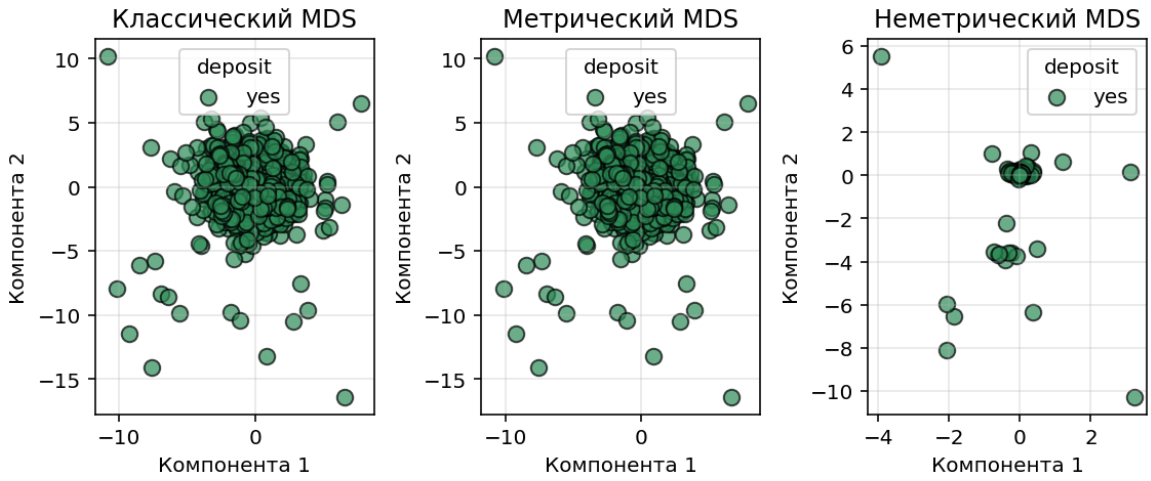

Рисунок 5 — Сравнение типов MDS

Молодые клиенты более склонны к общению, но могут иметь низкий доход. Они подходят под микрокредиты, бонусы, привлечение через длительное общение. Пожилые с poutcome_unknown (неизвестный посетитель) — потенциально недооцененный сегмент, поэтому стоит провести таргетированную кампанию. Клиенты с высоким duration (длительность последнего контакта) более вовлечённые, и их нужно направлять на депозиты.

Сравнение типов MDS показано на рисунке 5, из которого четко видно, что неметрический MDS не сходится.

Преимуществами метода являются выявление кластеров, уменьшение размерности, гибкость. Из недостатков можно отметить субъективность результатов, поскольку они зависят от выбора мер близости, чувствительность к выбросам и вычислительную сложность.

Выводы и направление дальнейших исследований. Выполненные исследования позволяют сделать следующие выводы:

1. Метод главных компонент (PCA) рекомендуется использовать при анализе данных, если требуется уменьшить их размерность и можно создать линейные комбинации с максимальным сохранением информации. При анализе экономической информации это особенно важно при исследовании финансовых показателей предприятий. Если данные разреженные или высокой размерности, то лучше использовать метод TruncatedSVD.

2. Метод ICA целесообразно использовать, когда нужно разделить многомерный сигнал на аддитивные независимые компоненты (факторы), например, при анализе временных рядов или анализе вклада отраслей в экономику. В отличие от традиционных методов (например, PCA), ICA фокусируется на статистической независимости, а не на дисперсии.

3. Что касается нелинейных методов, то метод t-SNE лучше использовать для анализа высокоразмерных данных, где важна локальная структура, например, для прогнозирования, в банковском телемаркетинге. Если надо сохранить как локальную, так и глобальную структуру данных, то лучше использовать UMAP (например, для описания структуры внешнеэкономических связей). Метод MDS позволяет анализировать сложные данные, где нужно исследовать закономерности или взаимосвязи, например, для выявления кластеров или исследования отношений, в макроэкономике для анализа эволюции экономических показателей.

Результаты проведенных исследований можно использовать в практической деятельности при выполнении анализа экономических данных.

Список источников

1. Городнова Н. В. Развитие цифровой экономики: теория и практика // Вопросы инновационной экономики. 2021. Т. 11. № 3. С. 911–928. DOI: 10.18334/vinec.11.3.112227 EDN RBVHJM

2. Дмитриев А. П., Лейба С. Ш. Стремительный рост цифровых данных: анализ мировых трендов и прогноз развития в России // Региональная и отраслевая экономика. 2024. № 1. С. 141–152. DOI: 10.47576/2949-1916.2024.1.1.019 EDN PCNILO

3. Miller G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information // Psychological Review. 1956. Vol. 63 (2). P. 81–97. URL: https://colab.ws/articles/10.1037%2Fh0043158. DOI: 10.1037/h0043158

4. Bachmann J. M. Bank Marketing Dataset [Electronic resource] : Kaggle : [website]. [2026]. URL: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset?resource=download&select=bank.csv.

5. Shlens J. A. Tutorial on Principal Component Analysis // arXiv preprint. 2014. arXiv:1404.1100v1. 12 p.

6. Van der Maaten L., Hinton G. Visualizing Data using t-SNE // Journal of Machine Learning Research. 2008. № 9. P. 2579–2605.

7. Hyvärinen A., Oja E. Independent component analysis: algorithms and applications // Neural Networks. 2000. Vol. 13. № 4–5. P. 411–430. EDN AHTWPL

8. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction // Journal of Open Source Software. 2018. Vol. 3. № 29. P. 861. DOI: 10.21105/joss.00861

9. Adams H., Blumstein M., Kassab L. Multidimensional scaling on infinite metric measure spaces // arXiv preprint. 2019. arXiv:1907.01379v1. 13 p.

10. Kruskal J. B. Nonmetric multidimensional scaling: a numerical method // Psychometrika. 1964. Vol. 29. № 2. P. 115–129. DOI: 10.1007/bf02289694 EDN GVZKRU

Рекомендована к печати д.э.н., проф. каф. ИСИБ ДонГТУ Бизяновым Е. Е., к.т.н., доц. каф. ЭКиПС ЛГУ им. В. Даля Велигурой А. В.

Статья поступила в редакцию 30.01.2026.

СВЕДЕНИЯ ОБ АВТОРАХ

Лепило Наталья Николаевна, *канд. техн. наук, доцент, доцент каф. информационных технологий*
Донбасский государственный технический университет,
г. Алчевск, Россия, e-mail: nnlepilo@mail.ru

Катан Карина Станиславовна, *ассистент каф. информационных технологий*
Донбасский государственный технический университет,
г. Алчевск, Россия

***Lepilo N. N., Katan K. S.** *(Donbass State Technical University, Alchevsk, Russia, *e-mail: nnlepilo@mail.ru)*

DIMENSIONALITY REDUCTION METHODS FOR ECONOMIC DATA ANALYSIS

The article examines and analyzes modern dimensionality reduction methods and the specifics of their use for data analysis. As an example of implementing methods in Python, there has been used a well-known dataset which contains information about bank customers and parameters of the marketing campaign carried out with them. Recommendations have been developed for the use of methods in the field of economic data analysis.

Key words: *dimension reduction, data analysis, dataset, singular decomposition, independent components, multi-dimensional scaling.*

References

1. Gorodnova N. V. Developing the digital economy: theory and practice [Razvitie cifrovoj ekonomiki: teoriya i praktika]. Innovation Economics. 2021. Vol. 11. No. 3. Pp. 911–928. DOI: 10.18334/vinec.11.3.112227 EDN RBVHJM

2. Dmitriev A. P., Lejba S. Sh. Rapid growth of digital data: analysis of world trends and predictiont of development in Russia [Stremitel'nyj rost cifrovyh dannyh: analiz mirovyh trendov i prognoz razvitiya v Rossii]. Regional and Sectoral Economics. 2024. No. 1. Pp. 141–152. DOI: 10.47576/2949-1916.2024.1.1.019 EDN PCNILO

3. Miller G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. 1956. Vol. 63(2). Pp. 81–97. URL: https://colab.ws/articles/10.1037%2Fh0043158. DOI: 10.1037/h0043158

4. Bachmann J. M. Bank Marketing Dataset. Kaggle. 2026. URL: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset?resource=download&select=bank.csv.

5. Shlens J. A. Tutorial on Principal Component Analysis. arXiv preprint. 2014. arXiv:1404.1100v1. 12 p. DOI: 10.48550/arXiv.1404.1100

6. Van der Maaten L., Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008. No. 9. Pp. 2579–2605.

7. Hyvärinen A., Oja E. Independent component analysis: algorithms and applications. Neural Networks. 2000. Vol. 13. No. 4–5. Pp. 411–430. EDN AHTWPL

8. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Journal of Open Source Software. 2018. Vol. 3. No. 29. Pp. 861. DOI: 10.21105/joss.00861

9. Adams H., Blumstein M., Kassab L. Multidimensional scaling on infinite metric measure spaces. arXiv preprint. 2019. arXiv:1907.01379v1. 13 p. DOI: 10.48550/arXiv.1907.01379

10. Kruskal J. B. Nonmetric multidimensional scaling: a numerical method. Psychometrika. 1964. Vol. 29. No 2. Pp. 115–129. DOI: 10.1007/bf02289694 EDN GVZKRU

INFORMATION ABOUT THE AUTHORS

Lepilo Natalya Nikolaevna**,** *PhD in Engineering, Assistant Professor of the Department of Information Technologies*
Donbass State Technical University,
Alchevsk, Russia, e-mail: nnlepilo@mail.ru

Katan Karina Stanislavovna, *Assistant lecturer of the Department of Information Technologies*
Donbass State Technical University,
Alchevsk, Russia