Выявление кластеров компаний с определенными стратегиями роста

Еремина Дарья и Прокудина Ника

10 ноября 2023 г.

Содержание

1 Введение

Эта работа представляет собой детальное объяснение и описание методов машинного обучения, использованных в рамках выполнения контрольного задания. Важно отметить, что часть аспектов, отраженных здесь, описывается и в файле .ipynb, однако некоторые из них будут рассматриваться более подробно здесь.

Приведенные методы не только описаны словесно, но и иллюстрированы с целью обеспечения более глубокого понимания работы алгоритмов. Кроме того, в данном тексте представлено описание признаков в используемом датасете, без которого результаты могут быть труднопонимаемыми. Также включен список источников, использованных нами для более глубокого вникания в тему.

2 Постановка задачи

Название нашей темы звучит как «Identifying Clusters of Companies with Specific Growth Strategies». Из самого заголовка становится ясным, что наше исследование связано с задачей кластеризации, то есть разделения объектов выборки на группы-кластеры таким образом, чтобы объекты, находящиеся внутри одного кластера были схожи между собой. Например, объекты-векторы признаков располагаются близко друг к другу в n-мерном пространстве, где n - количество признаков у наших объектов. Объекты из разных кластеров, напротив, имеют существенные различия (рис. 1). Отличие от задачи классификации заключается в том, что при решении задачи кластеризации нет данных о том, к какому классу должен относиться объект. Чаще всего, как было и в нашем случае, неизвестно даже число классов, поэтому его необходимо определять самостоятельно.

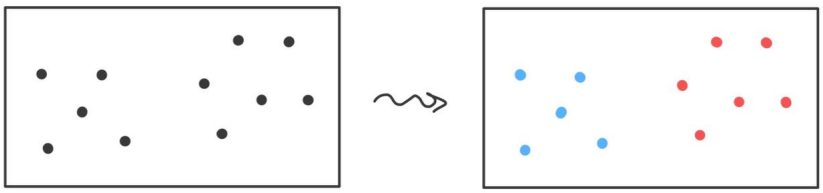

Рис. 1: Принцип кластеризации

3 Датасет

Наша работа базировалась на анализе датасета «Inc 5000s Fastest Growing Companies from 2007-2019», включающего в себя списки и рейтинги наиболее быстрорастущих компаний за последние 12 лет по мнению журнала Inc. Каждая строка в датасете соответствовала рейтинговой позиции компании за определенный год. У объектов были следующие признаки:

1. **year** – год выпуска рейтинга, из которого взята соответствующая строка
2. **rank** – место в рейтинге, которое заняла компания
3. **city** – город, в котором находится компания
4. **growth** – рост выручки в процентах за последние 3 года
5. **workers** – количество работников в компании
6. **company** – название компании
7. **state_s** – аббревиатура штата, в котором находится компания

8. **state_l** – полное название штата, в котором находится компания
9. **revenue** – выручка компании за последние 3 года
10. **yrs_on_list** – сколько раз компания попадала в рейтинг
11. **industry** – сфера, в которой работает компания
12. **metro** – metropolitan area, тоже обозначает местонахождение компании (объединение небольших населенных пунктов вокруг мегаполиса вместе с мегаполисом)

4 Предобработка и EDA

1. Удаление повторяющихся компаний:

Поскольку в датасете присутствуют данные за последние 12 лет, некоторые компании встречаются неоднократно. Однако, так как решается задача кластеризации и отсутствует целевая переменная, повторения могут исказить результаты. Поэтому было принято решение удалить повторяющиеся компании из списка.

2. Удаление лишних признаков:

Некоторые признаки не предоставляют полезной информации для кластеризации, включая повторяющуюся информацию о местонахождении, количество лет в списке, год и название компании, которое является идентификатором.

3. Удаление строк с выбросами:

Были удалены строки с данными, содержащими слишком большие значения переменных «growth» и «revenue», чтобы избежать искажений в результатах кластеризации.

4. Преобразование категориальных переменных:

Категориальные переменные были преобразованы в числовой формат при помощи Label Encoding для того, чтобы использовать их в алгоритмах машинного обучения.

5. Нормировка данных:

Была выполнена нормировка данных, так как используемые алгоритмы машинного обучения считают расстояния между объектами или

кластерами. Нормировка помогает избежать искажений в результате больших значений, например, выручки, которая может достигать нескольких миллионов.

5 Гипотеза и выделение важных признаков

После предобработки данных, мы задумались о том, какие признаки могут влиять на итоговый результат. Мы выдвинули гипотезу, что компании можно разделить на группы в соответствии с их стратегиями роста, учитывая особенности переменных «growth» и «revenue» следующим образом: некоторые компании больше фокусируются на увеличении выручки («revenue»), в то время как другие ориентированы на увеличение роста компании за год («growth»). Разница заключается в том, что для компании с высокой выручкой, чтобы увеличить ее в 2 раза, может потребоваться значительное увеличение в абсолютных значениях, в то время как для компании с низкой выручкой, чтобы достичь того же роста в процентном соотношении, требуется гораздо меньшее увеличение. Кроме того, возможно, компании стремятся попасть в список Inc для увеличения видимости. По этой причине этот признак может оказаться важным. Тем не менее такой признак как число работников, хоть и важный, не учитывается, так как сильно скоррелировано с выручкой.

Для того чтобы результаты можно было проще интерпретировать, попробуем использовать только признаки «rank», «growth», «revenue», а после получения результатов посмотрим, есть ли какая-то связь между разделением компаний на кластеры и их распределением по штатам или сферам.

6 Поиск оптимального числа кластеров при помощи the elbow method

Для определения оптимального числа кластеров мы использовали метод «локтя» (the elbow method). Этот метод основан на анализе зависимости между числом кластеров и суммой квадратов расстояния между объектами каждого кластера и центром кластера, который определяется как среднее всех объектов кластера. Если мы увеличиваем число кластеров, сумма квадратов расстояния уменьшается, однако, с определенного момента, уменьшение становится менее заметным.

Если построить график зависимости суммы квадратов расстояния от числа кластеров (см. Рисунок 2), можно заметить точку, в которой скорость уменьшения суммы квадратов расстояния замедляется. Приблизительно в этой точке и будет оптимальное число кластеров (если представить, что график – это рука, то оптимальная точка будет похожа на локоть – elbow, поэтому метод так и называется: мы как будто ищем «локоть» нашего графика)

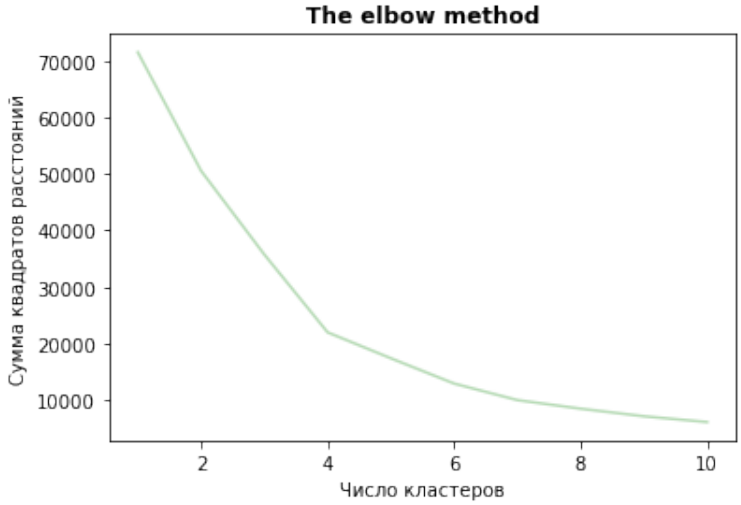

Рис. 2: Метод «Локтя»

7 Используемые методы

В нашей работе мы применили два основных метода кластеризации для анализа данных:

7.1 Агломеративная иерархическая кластеризация:

Агломеративная иерархическая кластеризация – метод, основанный на пошаговом объединении ближайших кластеров. На начальном этапе каждый объект считается отдельным кластером, затем на каждом шаге ближайшие кластеры объединяются до тех пор, пока не достигнется желаемое количество кластеров.

Процесс объединения в агломеративной иерархической кластеризации зависит от расстояния между кластерами. В модели scikit-learn, за это отвечают параметры:

- `metric` – метрика расстояния (в данном случае, используется евклидово расстояние: $d(A, B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$).
- `linkage` – указывает, как выбираются оптимальные для объединения кластеры. В зависимости от выбранной метрики (среднее, минимальное или максимальное расстояние), стремимся минимизировать соответствующее значение, определяя оптимальные пары кластеров для объединения.

В нашем исследовании мы применили значение по умолчанию для параметра `linkage`, выбрав «ward», так как это обеспечивает результат, который

можно интерпретировать как стратегии роста. В методе ward происходит объединение кластеров следующим образом: для каждого кластера рассчитывается его центр (среднее всех объектов в кластере). Затем вычисляется сумма квадратов расстояний от каждого объекта выбранного кластера до его центра. После этого посчитаем, что произойдёт, если мы объединим эти два кластера: то есть вычислим новый центр и сумму квадратов расстояния от каждого объекта нового кластера, который состоит из двух старых, до его центра. Две кластерные группы объединяются, если разница между этими суммами минимальна. Процесс продолжается до достижения необходимого количества кластеров.

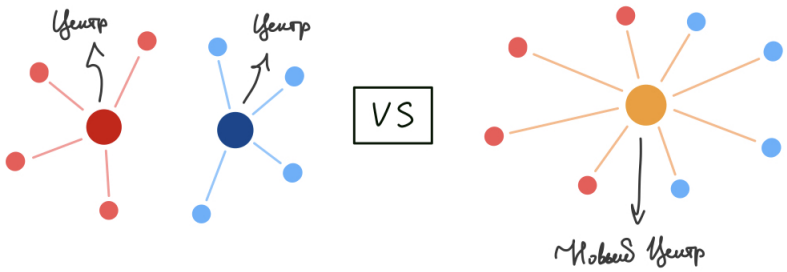

Рис. 3: Иллюстрация процесса агломеративной кластеризации.

7.2 K-means:

Метод k-means существенно отличается от предыдущего метода на уровне идеи. Его основная идея заключается в следующем: предположим, что нам нужно получить n кластеров. Алгоритм начинается с выбора случайных центров кластеров, затем объекты присваиваются ближайшим кластерам, и центры пересчитываются на основе средних значений объектов в каждом кластере. Процесс повторяется до сходимости.

В библиотеке scikit-learn, которую мы применяем, используется немного модифицированный алгоритм k-means++, который способствует более эффективному выбору начальных центров кластеров.

Данная процедура в методе k-means++ выглядит следующим образом: начальный центр выбирается случайным образом из наших объектов. Затем для каждого последующего центра мы рассчитываем квадрат расстояния от каждого объекта до уже выбранных центров, и следующий центр выбирается случайным образом из оставшихся объектов с вероятностью,

пропорциональной квадратам расстояний от объекта до ближайшего центра.

Простыми словами, мы стремимся выбирать центры, которые максимально удалены от уже выбранных центров, избегая случаев, когда центры слишком близки друг к другу или слишком далеко от объектов выборки.

При использовании алгоритма k-means в библиотеке scikit-learn мы применяем модификацию, известную как greedy k-means++. В ней все начальные центры, кроме первого, выбираются несколько раз для улучшения результатов кластеризации. В конце концов, мы получаем «оптимальный» центр, при котором сумма квадратов расстояний от объектов до ближайших центров минимальна.

После выбора начальных центров мы переходим к формированию первых кластеров. Каждый объект помещается в кластер с центром, находящимся ближе всего. После создания первых n кластеров обновляется положение центров. Важно отметить, что теперь центры не обязательно являются объектами из нашей выборки.

Каждый центр кластера i пересчитывается как среднее всех объектов, принадлежащих кластеру i. После этого обновленные центры используются для пересоздания кластеров, где каждый объект помещается в кластер с ближайшим обновленным центром. Эти шаги повторяются до тех пор, пока центры перестают изменяться, сигнализируя о достижении сходимости алгоритма.

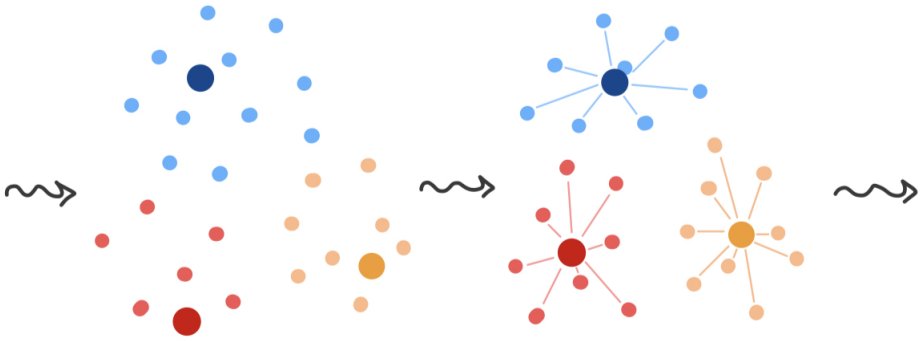

Рис. 4: Иллюстрация процесса алгоритма k-means.

8 Результаты

После проведения кластерного анализа с применением агломеративной иерархической кластеризации и алгоритма k-means с улучшением выбора изначальных центров, мы получили схожие результаты для обеих моделей.

Рассмотрим выделенные кластеры:

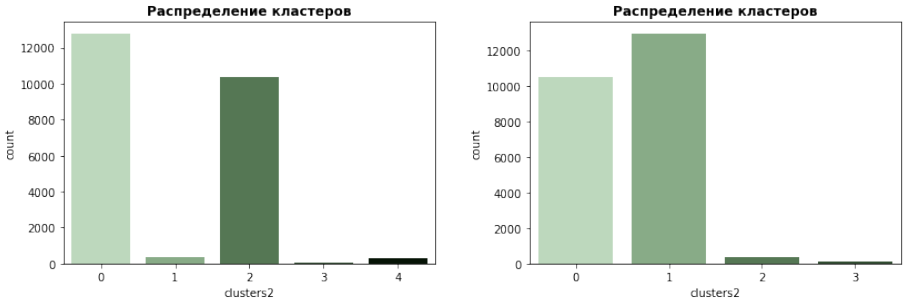

Рис. 5: Разделение компаний на 4 и 5 кластеров

- **Максимальный фокус на growth:** Компании в данном кластере сильно ориентированы на увеличение выручки за последние 3 года в процентах. Это, вероятно, свидетельствует о том, что они являются относительно новыми, возможно, стартапами, которым проще добиться высокого роста.

- **Основной фокус на growth (нижний рейтинг):** Компании в этом кластере также сосредоточены на росте, однако они расположены ниже в рейтинге по сравнению с предыдущим кластером. Вероятно, эти компании существуют дольше, но всё еще ставят целью увеличение среднегодового роста.

- **Максимальный фокус на выручке:** Компании этого кластера ориентированы на максимизацию выручки. Возможно, это компании класса люкс или те, которые уже имеют высокий уровень выручки, и им важнее поддерживать стабильность и рост за счет расширения бизнеса.

- **Основной фокус на выручке (низкие показатели):** Подобные компании также ориентированы на увеличение выручки, но не смогли достичь таких высоких результатов. Возможно, их цели также касаются чего-то неотраженного в датасете.

- **Неопределенность в росте:** При добавлении 5 кластера выделяются компании, которые не достигли относительных высоких результатов ни в росте, ни в выручке по сравнению с остальными. Вероятно, эти компании фокусируются на чем-то, что не отражено в датасете.

Таким образом, компании могут фокусироваться на увеличении среднегодового роста или выручки, а также комбинировать эти цели с факторами, не учтенными в датасете, такими как расширение линейки продуктов или услуг, расширение географии деятельности и другие стратегии роста.

9 Источники

1. https://www.inc.com/inc5000/2023 (дата обращения: 9.10.23)
2. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (дата обращения: 9.10.23)
3. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html (дата обращения: 9.10.23)
4. https://scikit-learn.org/stable/modules/clustering.html (дата обращения: 9.10.23)
5. https://www.scikit-yb.org/en/latest/api/cluster/elbow.html (дата обращения: 9.10.23)
6. https://academy.yandex.ru/handbook/ml (дата обращения: 9.10.23)
7. https://stats.stackexchange.com/questions/572409/intuitive-explanation-of-wards-method#~:text=Ward%27s%20procedure%20is%20a%20variance,summed%20for%20all%20the%20objects (дата обращения: 9.10.23)