КЛАСТЕРИЗАЦИЯ РАЙОНОВ МОСКВЫ ПО ФАКТОРАМ СОЦИАЛЬНОЙ СРЕДЫ И ЖИЛИЩНО-КОММУНАЛЬНОГО ХОЗЯЙСТВА

CLUSTERIZATION OF MOSCOW DISTRICTS BY SOCIAL ENVIRONMENT AND HOUSING AND COMMUNAL SERVICES FACTORS

УДК 004.9

Якушева Е.В., студент магистратуры 1 курс, Факультет бизнес-информатики и управления комплексными системами, НИЯУ МИФИ, Россия, г. Москва
Научный руководитель: к.т.н., доц. **Киреев В.С.**

Yakusheva E.V., 1st year Master's student, Faculty of Business Informatics and Management of Complex Systems, National Research Nuclear University MEPhI, Russia, Moscow
Scientific adviser: Ph.D., Assoc. **Kireev V.S.**

Аннотация: В работе представлены результаты кластеризации районов Москвы на основе открытых данных. Используются данные по количеству долговых обязательств по единому платежному документу (ЕПД), средней сумме долга по ЕПД и количеству семей, получающих субсидию в городе Москве. Для кластеризации используется инструмент RapidMiner Studio. К исходным данным было применено три метода кластерного анализа: k-средних, DBSCAN и EM-кластеризация. В качестве критерия оценки качества полученных результатов использовались индекс Дэвиса-Болдина, среднее внутри- и межкластерное расстояния. В исследовании использовались данные по 126 районам Москвы. В результате были выделены 7 кластеров по долговому

уровню жителей района по ЕПД и количеству семей, получающих субсидии. Также по каждому кластеру была дана интерпретация результата.

Annotation: The study presents the results of clustering Moscow districts based on open data. The data on the number of debt obligations under a single payment document (SPD), the average amount of debt under the SPD, and the number of families receiving subsidies in the city of Moscow are used. The RapidMiner tool is used for clustering. Three methods of cluster analysis were applied to the original data: k-means, DBSCAN, and EM-clustering. As a criterion for assessing the quality of the results obtained, the Davis-Boldin index, average intra- and inter-cluster distances were used. The study used data from 126 districts of Moscow. As a result, 7 clusters were identified according to the debt level of the district residents according to the SPD and the number of families receiving subsidies. Also, for each cluster, an interpretation of the result was given.

Ключевые слова: кластеризация RapidMiner, k-средних, DBSCAN, EM Clustering, районы Москвы.

Key words: RapidMiner clustering, k-means, DBSCAN, EM Clustering, Moscow districts.

Введение

Использование методов кластерного анализа для выявления общих характеристик изучаемых объектов позволяет упростить анализ и принятие решений по каждому кластеру в отдельности.

Результатом решения задачи кластеризации является отнесение каждого объекта к какому-либо классу. Однако решение, полученное разными методами, может отличаться, и заключение по распределению объектов должно основываться и на содержательном анализе.

В работе рассматривается сегментация районов Москвы с помощью программного продукта RapidMiner Studio с применением кластерного анализа с цель распределения исходного массива районов по кластерам в зависимости от

материального положения семей и долгов по единым платёжным документам (ЕПД).

Описание данных

В работе использовались данные Портала открытых данных Правительства Москвы. Для исследования были использованы следующие данные по районам Москвы с используемыми обозначениями:

1. Район (district).
2. Округ (okrug).
3. Среднее число долгов в месяц по единому платёжному документу (ЕПД) за 2020 г. (avg_n_debts_per_month).
4. Средняя сумма долга в месяц по единому платёжному документу (ЕПД) за 2020 г. (avg_sum_debt_per_month).
5. Число жителей (n_residents).
6. Площадь района в кв. м. (area).
7. Среднее число семей в месяц, получающих субсидии (avg_n_families_subsidies_per_month).

Проведение кластерного анализа с помощью RapidMiner

Проведём кластеризацию районов по количеству и сумме долгов по ЕПД и количеству семей с субсидиями. Реализован кластерный анализ с применением метода k-средних, метода DBSCAN и EM-алгоритма.

Рассмотрим алгоритм k-means. В таблице 1 представлены значения среднего внутрикластерного расстояния и индекса Дэвиса-Болдина. Как видно из таблицы 1 при числе кластеров равного 6 индекс Дэвиса-Болдина имеет минимальное значение, при дальнейшем увеличении числа кластеров показатель увеличивается.

Таблица 1.

Сравнение критериев качества кластерного решения k-means

Число кластеров	Среднее внутрикластерное расстояние	Индекс Дэвиса-Болдина
2	-1,804	-1,104
3	-1,353	-1,037
4	-1,087	-1,084

5	-0,892	-0,972
6	-0,802	-0,970
7	-0,751	-1,023
8	-0,659	-1,080

Для определения оптимального числа кластеров было выдвинуто предположение о том, что каждый параметр может иметь три значения – высокое, среднее и низкое. Однако при детальном рассмотрении стало понятно, что отсутствуют данные о большом количестве семей с субсидиями, низким числом долгов и высокой средней суммой долга, а также большим количеством долгов при низкой сумме долга и малым количеством семей с субсидиями. В связи с этим для рассмотрения результатов выполнения кластеризации выбрано шесть и семь кластеров.

На рисунке 1 представлены диаграммы рассеяния методом k-средних для шести кластеров, а на рисунке 2 – для семи кластеров.

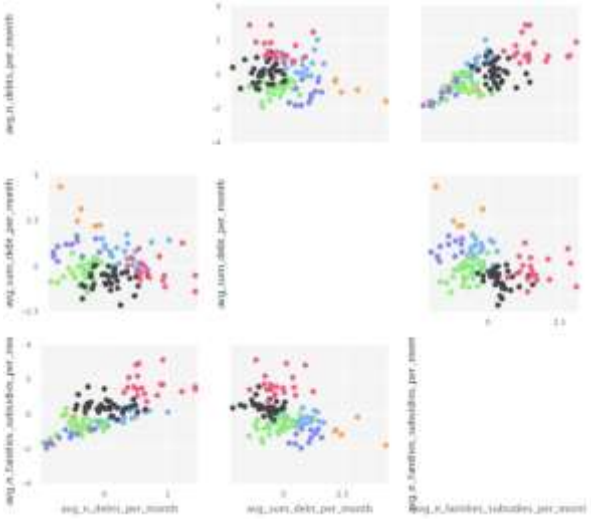

Рисунок 1. Диаграммы рассеяния для шести кластеров, полученных методом k-средних

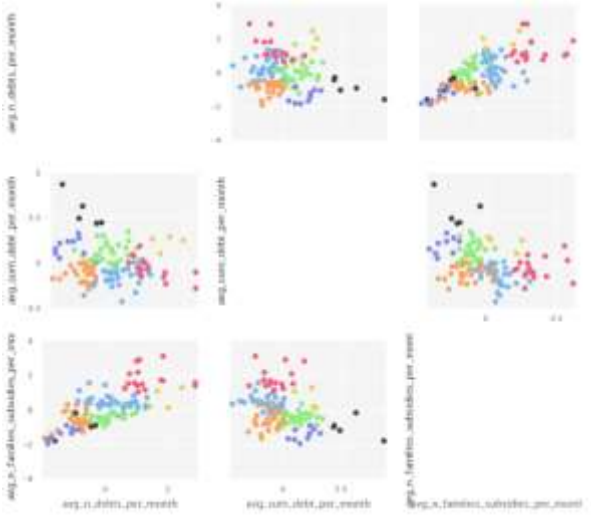

Рисунок 2. Диаграммы рассеяния для семи кластеров, полученных методом k-средних

В таблице 2 представлены центроиды и расшифровка цветов шести кластеров.

Таблица 2.

Таблица центроидов для шести кластеров, полученных методом k-средних

Атрибут	cluster_0 зелёный	cluster_1 голубой	cluster_2 оранжевый	cluster_3 красный	cluster_4 синий	cluster_5 чёрный
avg_n_debts_per_month	-0,75	0,58	-0,86	1,46	-1,21	0,09

avg_sum_debt_per_month	-0,28	0,67	2,89	-0,20	0,99	-0,74
avg_n_families_subsidies_per_month	-0,67	-0,21	-1,01	1,68	-1,33	0,43

В таблице 3 представлены центроиды и расшифровка цветов семи кластеров.

Таблица 3.

Таблица центроидов для семи кластеров, полученных методом k-средних

Атрибут	cluster_0 голубой	cluster_1 синий	cluster_2 жёлтый	cluster_3 зелёный	cluster_4 оранжевый	cluster_5 красный	cluster_6 чёрный
avg_n_debts_per_month	0,257	-1,401	1,905	-0,004	-0,939	1,385	-0,856
avg_sum_debt_per_month	-0,673	0,954	1,136	0,568	-0,561	-0,311	2,887
avg_n_families_subsidies_per_month	0,412	-1,434	0,666	-0,462	-0,658	1,801	-1,008

Дополнительно к результатам шести кластеров добавился седьмой, описывающий ситуацию, когда все три параметра имеют значения выше среднего. Поэтому k-mean с семью кластерами больше подойдёт для кластеризации.

Кластеризация с помощью метода DBSCAN проводилась со следующими параметрами: epsion=0,5 и минимальное число точек=2. Диаграммы рассеяния после применения метода представлены на рисунке 3.

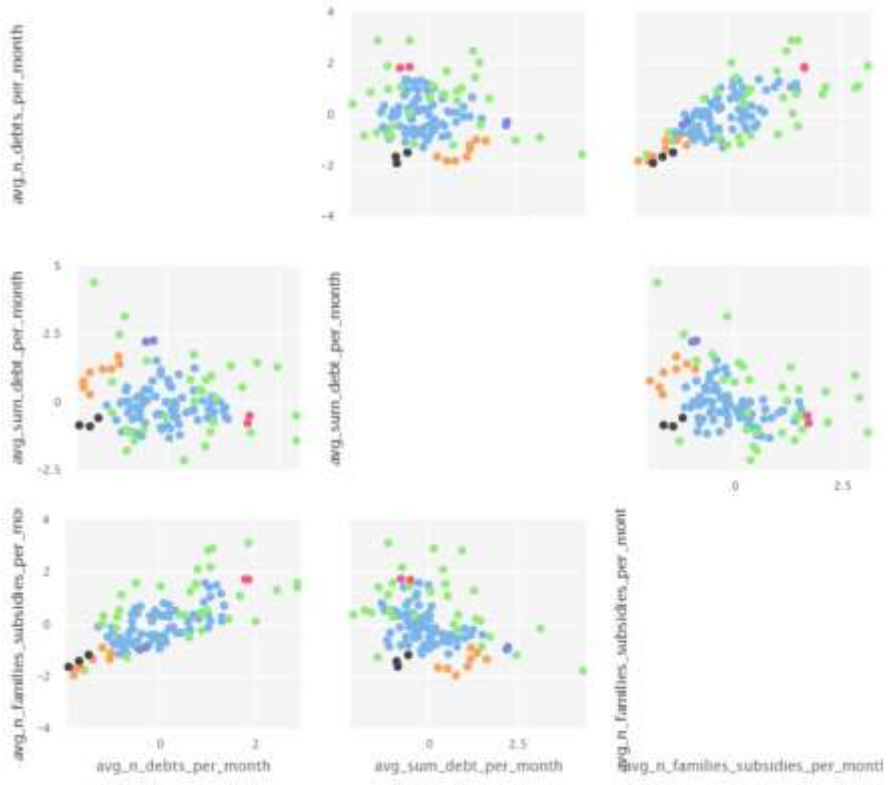

Рисунок 3. Диаграммы рассеяния для кластеров, полученных методом DBSCAN

В таблице 4 представлены критерии качества применения метода DBSCAN. По результатам применения алгоритма получены пять кластеров, при этом большое среднее значение межкластерного расстояния достигается за счёт первых двух кластеров, поэтому данный алгоритм для решения задачи не подойдёт.

Таблица 4.

Критерии качества кластерного решения DBSCAN

Показатель	Значение
Среднее межкластерное расстояние	-92.932
Среднее межкластерное расстояние cluster_0	-89.738
Среднее межкластерное расстояние cluster_1	-111.172
Среднее межкластерное расстояние cluster_2	-6.038
Среднее межкластерное расстояние cluster_3	-0.962
Среднее межкластерное расстояние cluster_4	-0.173
Среднее межкластерное расстояние cluster_5	-0.269

Если рассмотреть полученные результаты, то в самый большой кластер (голубой) попали районы с наиболее средними показателями по всем трём параметрам, имеющие наибольшую плотность. В зелёный кластер попали

районы, сильно отличающиеся по одному или нескольким параметрам от общей массы. Например, среднее число долгов в месяц высокое, при этом средняя сумма долга низкая или наоборот.

С помощью данного метода были отобраны районы, выбивающиеся из общей массы районов, при этом получить более детальное разбиение районов для последующего анализа не удалось.

Кластеризация с помощью метода EM Clustering была проведена для числа кластеров, указанных в таблице 5. Для указанного числа кластеров было рассчитано среднее межкластерное расстояние.

Таблица 5.

Сравнение критериев качества кластерного решения EM Clustering

Число кластеров	Среднее межкластерное расстояние
2	-123,374
3	-79,586
4	-72,102
5	-55,881
6	-44,022

При дальнейшем увеличении числа кластеров алгоритм объединяет все элементы в один кластер. Поэтому кластеризация с помощью метода EM Clustering проводилась со следующими параметрами: k=6. Диаграммы рассеяния после применения метода представлены на рисунке 4.

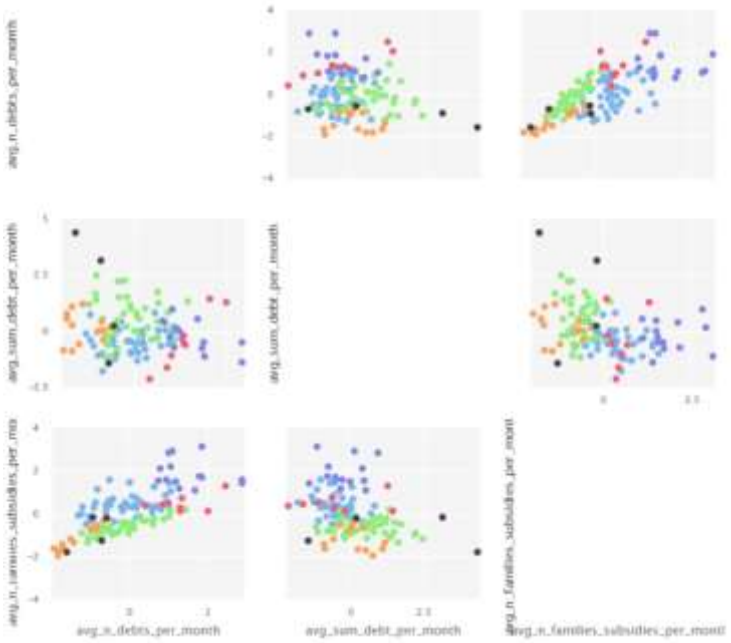

Рисунок 4. Диаграммы рассеяния для шести кластеров, полученных методом EM Clustering

Алгоритм EM Clustering показал схожий результат с алгоритмом k-средних. Различия можно увидеть для граничных районов. Проинтерпретируем полученный результат.

Интерпретация результатов

Для метода k-means с числом кластеров равным семи получилось следующее распределение районов. В первый кластер cluster_0 (голубой) входят Академический, Бабушкинский, Бирюлево Восточное, Богородское, Вешняки, Восточное Дегунино, Головинский, Западное Дегунино, Ивановское, Измайлово, Коптево, Крюково, Лосиноостровский, Нагатинский Затон, Новогиреево, Орехово-Борисово Северное, Орехово-Борисово Южное, Очаково-Матвеевское, Перово, Печатники, Рязанский, Северное Измайлово, Северное Медведково, Солнцево, Таганский, Текстильщики, Теплый Стан, Тимирязевский, Филевский парк, Фили-Давыдково, Ховрино, Царицыно, Чертаново Северное, Чертаново Центральное, Щукино, Южное Медведково, Южное Тушино, Ярославский районы. Районы данного кластера имеют низкие значения по средней сумме задолженности по ЕПД, средние значения по количеству среднемесячных долгов по ЕПД и количеству семей, получающих субсидии. Как выяснили, количество долгов по ЕПД коррелирует с количеством семей, получающих субсидии. При этом сумма долга низкая, что может говорить о дисциплинированности людей, которые предпочитают не иметь долгов по ЕПД.

Во второй кластер cluster_1 (синий) вошли районы: Беговой, Внуково, Красносельский, Куркино, Мещанский, Молжаниновский, Ростокино, Северный, Якиманка. Данные кластер характеризуется низким числом среднемесячных долгов по ЕПД, сумма долга находится на среднем уровне, число семей, получающих субсидии мало. В эту группу вошли как престижные районы, так и районы, находящиеся за пределами МКАД. Люди, проживающие в престижных районах, имеют средства для оплаты, ответственно относятся к своим средствам, вовремя оплачивая счета. Проживающие в районах,

находящихся за пределами МКАД, могут быть с достатком, так же как в предыдущем случае следят за своевременной оплатой, либо не имеют высокого дохода, но также ценят денежные средства.

В третий кластер cluster_2 (жёлтый) вошли районы: Кунцево, Пресненский, Раменки, Южное Бутово. Основные черты данного кластера: большое число среднемесячных долгов по ЕПД при средней сумме долга, и большое число семей с субсидиями. Сумма долга выше, чем у следующего, четвёртого кластера, по одной из версий, может быть более высокие доходы жителей районов, по другой – беспечность по отношению к финансам.

В четвёртый кластер cluster_3 (зелёный) вошли районы: Алексеевский, Аэропорт, Басманный, Бескудниковский, Войковский, Гагаринский, Дмитровский, Косино-Ухтомский, Крылатское, Лефортово, Лианозово, Ломоносовский, Марьина роща, Матушкино, Некрасовка, Новокосино, Ново-Переделкино, Обручевский, Останкинский, Савелки, Савеловский, Силино, Старое Крюково, Тверской, Тропарево-Никулино, Хамовники, Хорошевский, Черемушки, Южнопортовый. Для четвёртого кластера характерны средние показатели по всем трём параметрам. Из-за среднего числа семей с субсидиями, количество долгов велико, однако средняя сумма долга возможна из-за престижности некоторых районов и возможной расточительности жителей.

В пятый кластер cluster_4 (оранжевый) вошли районы: Алтуфьевский, Бирюлево Западное, Братеево, Бутырский, Восточное Измайлово, Восточный, Донской, Зябликово, Капотня, Котловка, Левобережный, Марфино, Метрогородок, Москворечье-Сабурово, Нагатино-Садовники, Нагорный, Нижегородский, Покровское-Стрешнево, Преображенское, Проспект Вернадского, Свиблово, Северное Бутово, Соколиная гора, Сокольники. Низкие показатели по количеству долгов и сумме долга по ЕПД, при этом в районах среднее число семей, получающих субсидии. Жители данных районов наиболее дисциплинированные.

В шестой кластер cluster_5 (красный) вошли районы: Бибирево, Выхино-Жулебино, Гольяново, Зюзино, Коньково, Кузьминки, Люблино, Марьино,

Митино, Можайский, Отрадное, Северное Тушино, Строгино, Хорошево-Мневники, Чертаново Южное. Для данного кластера характерно большое число долгов в месяц с низкой суммой долга, но с большим числом семей, получающих субсидии. Семей с субсидиями в районах много, поэтому и долгов в районе много.

В седьмой кластер cluster_6 (чёрный) вошли районы: Арбат, Даниловский, Дорогомилово, Замоскворечье, Сокол. Кластер характеризуется средним числом долгов, но высокой суммой долга, число семей с субсидиями мало. Район интересен, тем, что при престижности районов, сумма долга велика, это может означать безответственность людей, пренебрежение к правилам или расточительность.

Такие районы, как Замоскворечье, Дорогомилово, Сокол, Северный, Беговой, Мещанский, Кунцево и некоторые другие при использовании алгоритма EM Clustering были отнесены к другим кластерам. Для более точного отнесения районов к определённому кластеру возможно рассмотреть и другие параметры, характеризующие район с точки зрения долговых обязательств.

Заключение

По итогам исследования были получены схожие результаты для алгоритмов k-means и EM Clustering, с семью и шестью кластерами соответственно. Однако из-за выделения дополнительного кластера по методу k-means определение кластеров оказалось более приближенным к реальности. Поэтому на основе кластеров, полученных с помощью алгоритма k-средних, было проанализировано отнесение районов к определённому кластеру. Была выявлена положительная корреляция между числом семей района, получающих субсидии, и количеством долгов по ЕПД в районе. Зависимости между числом семей и суммой долга не найдено, но были выявлены районы у которых число семей с субсидиями велико, однако сумма долга находится на среднем уровне или ниже. Также присутствует обратная ситуация, когда число семей с субсидиями в районе мало, но сумма долга по ЕПД значительно превышает

средний показатель. Это можно объяснить разным социальным статусом районов.

Использованные источники:

1. Тюрин, Ю.Н. Анализ данных на компьютере [Текст] : учебное пособие для вузов / Ю. И. Тюрин , А. А. Макаров. - изд. 4-е, перераб. - Москва: Форум, 2011, 2010. - 367 с.
2. Портал открытых данных Правительства Москвы. [Электронный ресурс]. URL: https://data.mos.ru/ (дата обращения: 12.05.2021).
3. RapidMiner. [Электронный ресурс]. URL: https://rapidminer.com/ (дата обращения: 12.05.2021).
4. Hofmann M. RapidMiner. Data Mining Use Cases and Business Analytics Application / Markus Hofmann, Ralf Klinkenberg – FL, 2014 - International Standard Book Number-13: 978-1-4822-0550-3 (eBook - PDF).

Used sources:

1. Tyurin, Yu.N. Analysis of data on a computer [Text]: textbook for universities / Yu. I. Tyurin, AA Makarov. - ed. 4th, rev. - Moscow: Forum, 2011, 2010.- 367 p.
2. Portal of open data of the Government of Moscow. [Electronic resource]. URL: https://data.mos.ru/ (date of access: 12.05.2021).
3. RapidMiner. [Electronic resource]. URL: https://rapidminer.com/ (date of access: 12.05.2021).
4. Hofmann M. RapidMiner. Data Mining Use Cases and Business Analytics Application / Markus Hofmann, Ralf Klinkenberg - FL, 2014 - International Standard Book Number-13: 978-1-4822-0550-3 (eBook - PDF).