УДК 004.021

КЛАСТЕРНЫЙ АНАЛИЗ ПОВЕДЕНИЯ ПОКУПАТЕЛЕЙ

Филипьев Андрей Владимирович

Ассистент;
ГБОУ ВО МО «Университет «Дубна»;
Институт системного анализа и управления;
e-mail: avfilipev@gmail.com.

В статье рассматривается проблема сложности анализа потребительского опыта. Применены алгоритмы кластеризации и классификации для определения вкусовых предпочтений покупателей. Определены кластеры, построенные на истории продаж продуктов, созданных по собственным рецептам пользователей. На основе кластерных меток обучена модель классификации. В результате получена модель, показывающая вкусовые предпочтения покупателей и изменения с введением нового продукта.

Ключевые слова: кластерный анализ, классификация, обучение с учителем, обучение без учителя.

CLUSTER ANALYSIS OF CUSTOMERS BEHAVIOR

Filipyev Andrey

Assistant;
Dubna State University;
Institute of the system analysis and management;
e-mail: avfilipev@gmail.com.

This article describes the problem of complexity of the analysis of customer orders. Clustering and classification algorithms were applied to examine the taste preferences of clients. Clusters are based on the sales history of products created by the customers by their own recipes. The classification model is taught by cluster marks. The result is a model showing the taste preferences of buyers and changes with the introduction of a new product.

Keywords: Cluster analysis, classification, supervised learning, unsupervised learning.

Введение

Данная статья посвящена анализу потребительского поведения с помощью методов машинного обучения. Рост компании и количества клиентов затрудняет аналитикам изучение продаж с помощью привычных им инструментов. С введением возможности для потребителя создать продукт по собственному рецепту становится невозможным анализировать предпочтения и изменения в поведении клиентов, так как продукт может состоять из большого числа потенциальных ингредиентов. Кластерный анализ позволяет разделить эти рецепты на понятные для аналитика группы. Использование модели классификации, обученной на метках, полученных из кластерного анализа, позволяет проследить поведение клиента в его вкусовых предпочтениях.

Аналитика и актуальность проблемы

Покупатели в потребительском поведении привыкли к повседневной кастомизации продукции: удаление ингредиентов, изменение первоначального рецепта или же создание собственного продукта. Пример такого поведения - заказ пицц, приготовленных по собственному набору ингредиентов. Пользователь, открывая сайт компании-пиццерии, может выбрать уже существующий продукт или, используя конструктор, создать свой собственный рецепт.

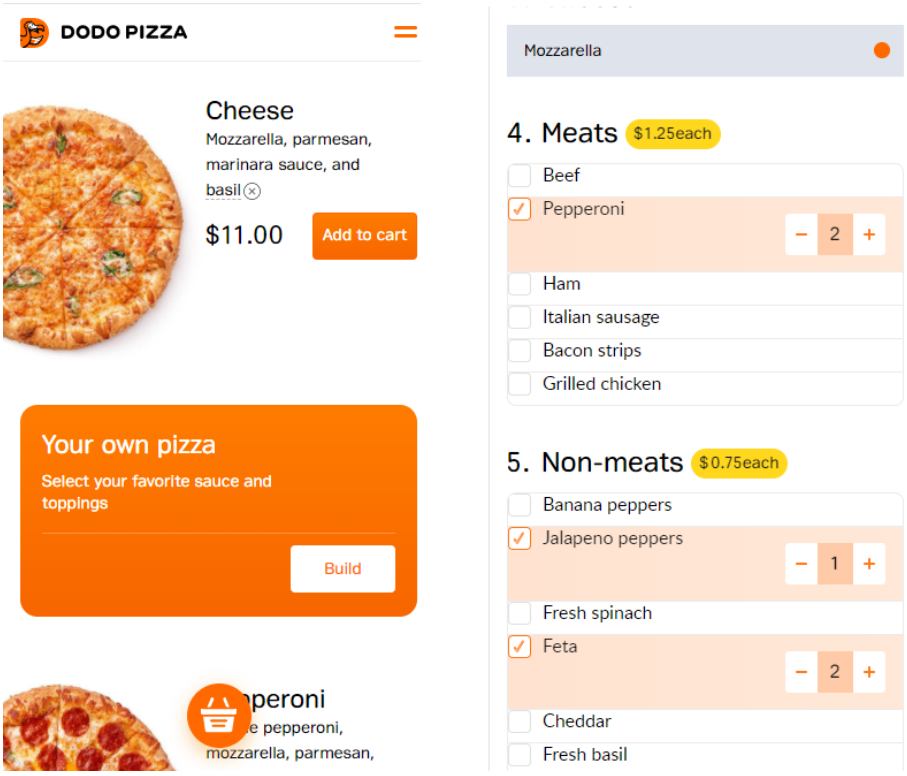

Рис. 1. Конструктор продукта с собственным рецептом

Анализ рынка

На момент написания статьи под брендом компании Додо Пицца открыто более 300 пиццерий в 10 странах, большинство которых открыто на российском рынке. В США открыто 2 пиццерии и этот рынок рассматривается как один из ключевых рынков компании Додо в перспективе нескольких лет.

Доля кастомизированных продуктов за январь 2018 года составила 24,78%, из которых 14,3% это пиццы, созданные по собственному рецепту через конструктор сайта. Это существенный процент, приближающийся к топовым по количеству продаж продуктам, таким как пицца «Пепперони».

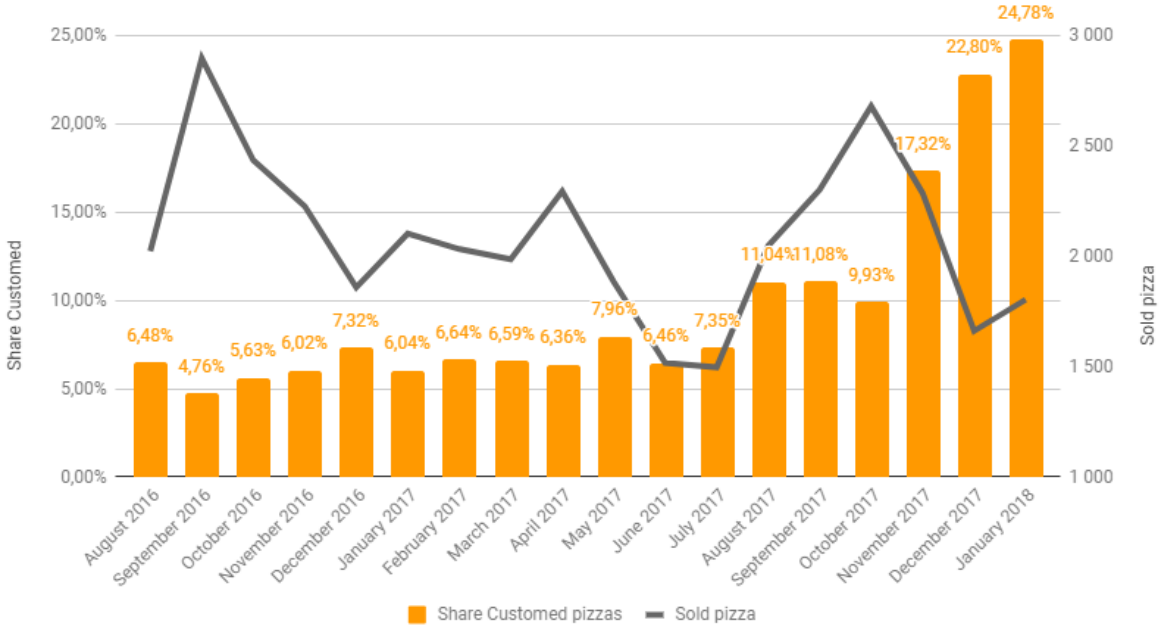

Рис. 2. График сравнения доли продаж продуктов с собственным рецептом, относительно других продуктов

Для оценки объемов потенциальных продаж продукта по собственным рецептам, можно воспользоваться статистикой по количеству открытых пиццерий, предоставляющих схожую продукцию. По данным ресурса *entrepreneur.com* под брендом Papa John's по франшизе в США открыто 2731 пиццерия [1].

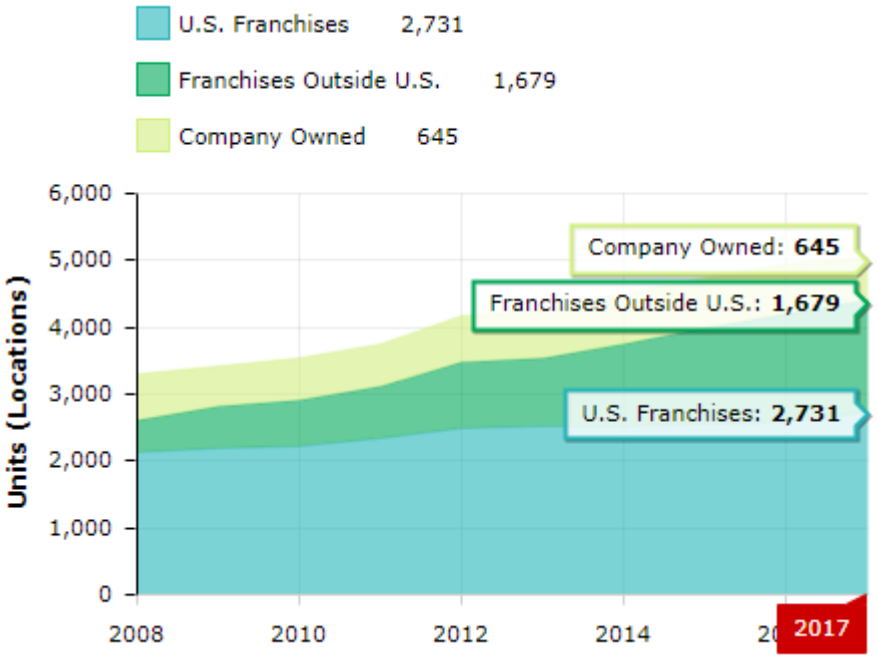

Рис. 3. Количество открытых точек продаж компании Papa John's в США

Развитие сети пиццерий Додо можно сравнить с крупными компаниями. Такой вывод можно сделать, изучив годовой отчет компании DP Eurasia перед акционерами [2], на графике видны темпы развития на Российском рынке конкурирующих сетей.

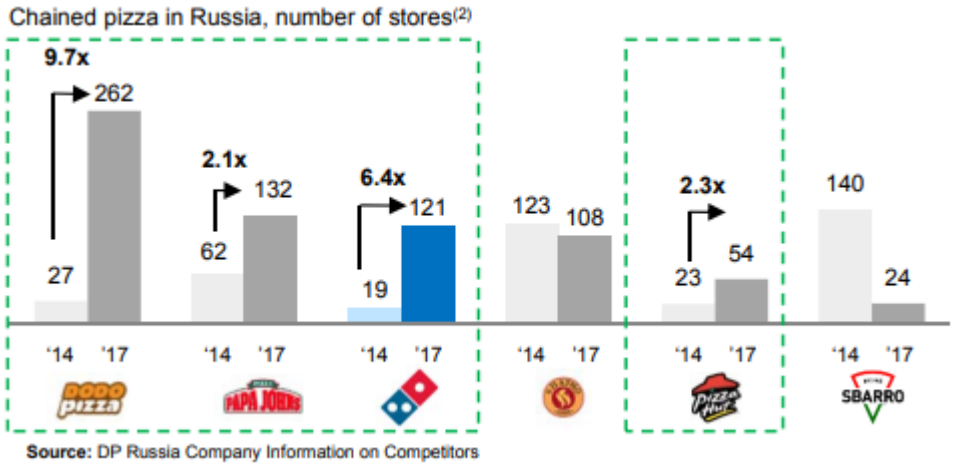

Рис. 4. Сравнение открытых точек продаж в России

Если предположить, что в ближайшие годы сеть пиццерий Додо вырастет на рынке США до 1000 открытых точек, и взять количество проданных пицц за февраль 2018 года, созданных по собственным рецептам пользователей, это около 300 штук, в условиях отсутствия традиционного маркетинга [3], то даже при самых минимальных оценках число продаж этого продукта будет не менее 300 000 в месяц. В год потенциальное количество продаж этого продукта может быть более 3600000 штук, при средней стоимости в $14, сумма выручки только с этого продукта будет более $50 000 000.

В отличии от традиционных рецептов пицц, которые предлагаются покупателю, продукт, называемый «Своя пицца» не имеет конкретного состава ингредиентов. То есть, если знать среднее количество продаж пиццы «Пепперони», то можно рассчитать количество запасов, необходимых на ее производство. В случае «Своей пиццы» задача сильно усложняется, нельзя заранее предугадать какие составы пицц будут куплены, следовательно придется увеличивать количество всех 28 возможных ингредиентов на складе. Неумение работать с данными по продажам ведет к следующим рискам:

- порча запасов ингредиентов и повышение денежных потерь;
- необходимость остановить продажу продукта, что приводит к потере выручки;
- невозможность анализа предпочтений покупателей;
- потеря лояльных покупателей и т.д.

Сложность анализа кастомизации

Кроме продукта «Своя пицца» пользователь может удалить из заказываемого продукта с готовым рецептом какие-то ингредиенты. Например, удалить из пиццы «Додо» лук, оливки или другие составляющие. Такие пиццы называются кастомизированные, и так же имеют сложность при анализе, приходится вручную обрабатывать таблицы информации по поводу изменений составов, группировать их на основе мнения одного человека. Такие продукты сильно отличаются от концепции создания “Своей пиццы”, но имеют похожие проблемы с большим разнообразием потенциально убираемых ингредиентов. Возможность удаления ингредиентов запущена и на рынке РФ, но несет меньше потенциальных рисков, чем продукты с собственным рецептом.

Опираясь на опыт аналитиков, строящих отчеты по кастомизированным продуктам, на схожесть идеи изменения состава пиццы, или добавления новых ингредиентов, была разработана концепция отчета для продукта «Своя пицца». После понимания какую информацию необходимо получать из отчета, были выстроены следующие шаги в работе.

Группа	Основа группы	Кол-во	Сумма
Группа 1	Бекон+пеп+лук (+N)		
Группа 1	+Базилик		
Группа 1	+Оливки		
Группа 1	....		
Группа 1	+N		
Группа 2	Бекон+пеп (+N)		
Группа 3	Пеп+Базилик+сыр (+N)		

(+N) - плюс несколько ингредиентов, не входящих в основу

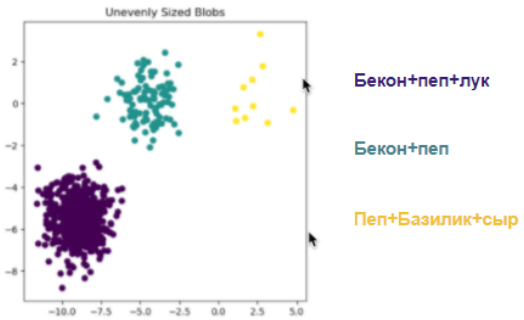

Рис. 5. Шаблон отчета для анализа продуктов с собственным рецептом

Перед поиском структуры в данных о проданных продуктах «Своей пиццы», предполагалось, что отчет должен содержать следующую информацию:

1. Разбить рецепты на группы, то есть кластеры, которые будут состоять основываться на двух-трех основных ингредиентах.
2. Содержать информацию о второстепенно добавленных ингредиентах внутри кластера.
3. Визуально отображать группы.

Такая информация позволила бы управляющим пиццерий лучше понимать какие продукты покупают пользователи, чтобы выстраивать процесс закупки ингредиентов и производства оптимальным способом.

Анализ данных о продажах продукта с собственным рецептом

Исходные данные для анализа можно представить в виде таблицы, каждая отдельная строка которой содержит информацию по отдельно купленному продукту «Своя пицца». В каждой отдельной колонке находится информация о каждом добавленном ингредиенте в конкретную пиццу.

Если представить, что X_i – это отдельная строка, то она будет соответствовать отдельно проданной пицце по собственному рецепту. Тогда X_{ij}– это количество конкретного ингредиента, добавленного в этот рецепт.

Задачу анализа продуктов, проданных по своим рецептам можно сформулировать следующим образом:

Необходимо найти множество объектов $Y_{1,\dots,l}$, и проставить множеству объектов X_i метки из множества Y таким образом, чтобы объекты из множества X, то есть рецепты, с одинаковыми метками были похожи друг на друга.

AddedIngredientName	BBQ sauce	Bacon	Banana peppers	Basil	Black olives	Buffalo	Buffalo sauce	Cheddar	Cherry tomatoes	Chicken	...	Marinara sauce	Mozzarella	Mushrooms	Parmesan
OrderCompositionId															
74621	0	0.0	1.0	0.0	0.0	0	0	0.0	0.0	0.0	...	0	0	0.0	0.0
74650	0	0.0	0.0	0.0	0.0	0	0	0.0	0.0	0.0	...	0	0	1.0	1.0
74687	0	3.0	0.0	0.0	0.0	0	0	0.0	0.0	0.0	...	0	0	0.0	0.0
74750	0	0.0	0.0	0.0	0.0	0	0	0.0	0.0	0.0	...	0	0	0.0	0.0
74776	0	0.0	0.0	1.0	1.0	0	0	0.0	0.0	0.0	...	0	0	0.0	0.0
74829	0	0.0	0.0	1.0	0.0	0	0	0.0	1.0	0.0	...	0	0	0.0	0.0
74853	0	1.5	1.0	0.0	0.0	0	0	0.0	0.0	0.0	...	0	0	0.0	0.0
74864	0	1.5	0.0	0.0	0.0	0	0	0.0	0.0	0.0	...	0	0	0.0	0.0
74880	0	0.0	0.0	0.0	0.0	0	0	0.0	0.0	0.0	...	0	0	1.0	0.0
74972	0	0.0	0.0	0.0	0.0	0	0	0.0	1.0	0.0	...	0	0	0.0	0.0

Рис. 6. Вид исходных данных для анализа

Алгоритм K-Means

Первым методом кластеризации был выбран популярный алгоритм KMeans. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров [4]:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 , \quad (1)$$

где k – число кластеров, S_i – полученные кластеры, $i=1,2,\dots,k$ и μ_i – центры масс векторов $x_j \in S_i$. Этот алгоритм достаточно простой, но при этом и не точный, помимо этого имеет ряд проблем:

- чувствительность к выбору начального приближения;
- результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен;
- число кластеров надо знать заранее.

Для того, чтобы определить начальное количество кластеров можно воспользоваться алгоритмом *The Elbow Method*. Это частично визуальный метод выбора начального количества кластеров. Число

кластеров выбирается таким образом, что после увеличения количества кластеров не дает существенного улучшения модели анализа. То есть при добавлении количества кластеров до выбранного значения, с каждым шагом дает гораздо больше информации, чем последующее добавление кластеров. Визуально на графике это будет видно в виде более резкого изгиба.

Как видно на графике ниже, явного выделения количества кластеров не происходит. Поэтому было выбрано более оптимальное количество кластеров – 6.

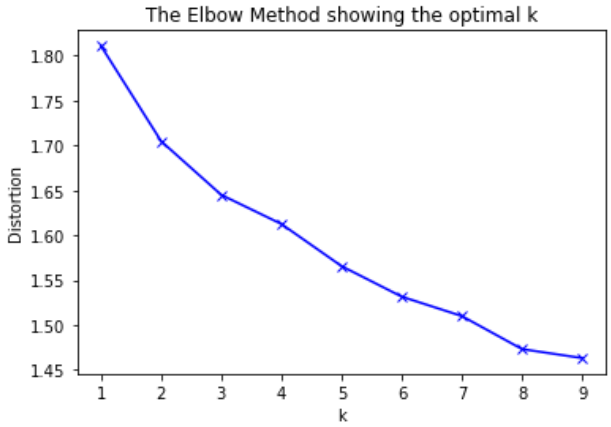

Рис. 7. Результат использования алгоритма The Elbow Method

Задав начальное значение количества кластеров 6 и воспользовавшись методом *KMeans* были получены метки рецептов своих пицц. После подробного изучения полученных кластеров, были сделаны выводы, что распределение объектов между кластерами не имеет какой-то явной логической закономерности.

Иерархическая кластеризация

После полученных результатов работы алгоритма *KMeans* было решено воспользоваться другим подходом для решения задачи. Иерархическая кластеризация – это комплекс алгоритмов, использующих разделение крупных кластеров на более мелкие или объединение мелких в более крупные. Соответственно, выделяют разделительную (дивизивную) и агломеративную (объединительную) кластеризации [6]. Зная предметную область анализа, было предположено, что агломеративный подход будет более логичен в решении задачи поиска кластеров среди продуктов с разными рецептами, так как в основе агломеративного похода лежит идея, что изначально каждый объект относится к отдельному кластеру. Каждый отдельный рецепт – это продукт, созданный персонально человеком. И объединяя похожие друг на друга рецепты можно найти кластеры, которые будут описывать похожие друг на друга предпочтения в рецептах разных людей. Визуально этот метод можно представить следующим образом [7]:

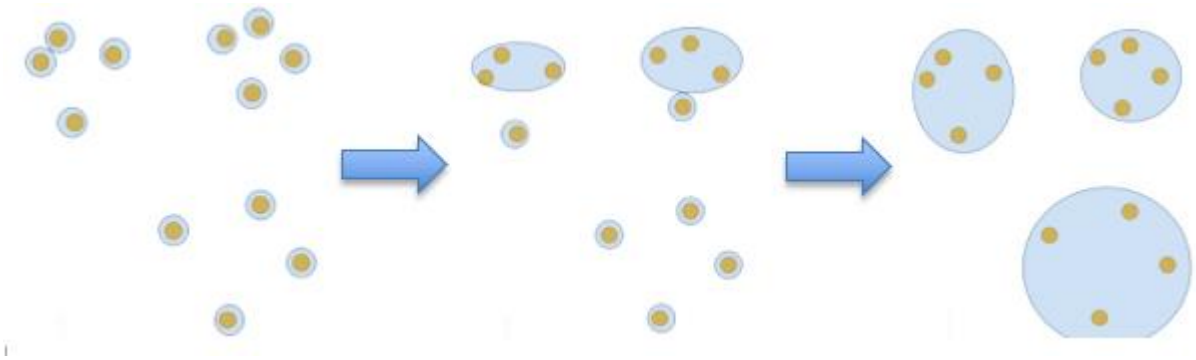

Рис. 8. Пример объединения объектов в кластеры с использованием агломеративных методов

Чтобы оценить, какое количество кластеров выбрать в иерархической кластеризации можно построить дендрограмму. Дендрограмма показывает степень близости отдельных объектов и кластеров,

а также наглядно демонстрирует в графическом виде последовательность их объединения или разделения. Количество уровней дендрограммы соответствует числу шагов слияния или разделения кластеров[8].

На графике отображена дендрограмма, построенная на исходных данных. Разными цветами помечены разные кластеры, с выбранным количеством равным 6. Это количество было выбрано визуально, так как описывает разделение объектов на кластеры таким образом, чтобы внутри себя они включали максимально похожие на друг друга объекты.

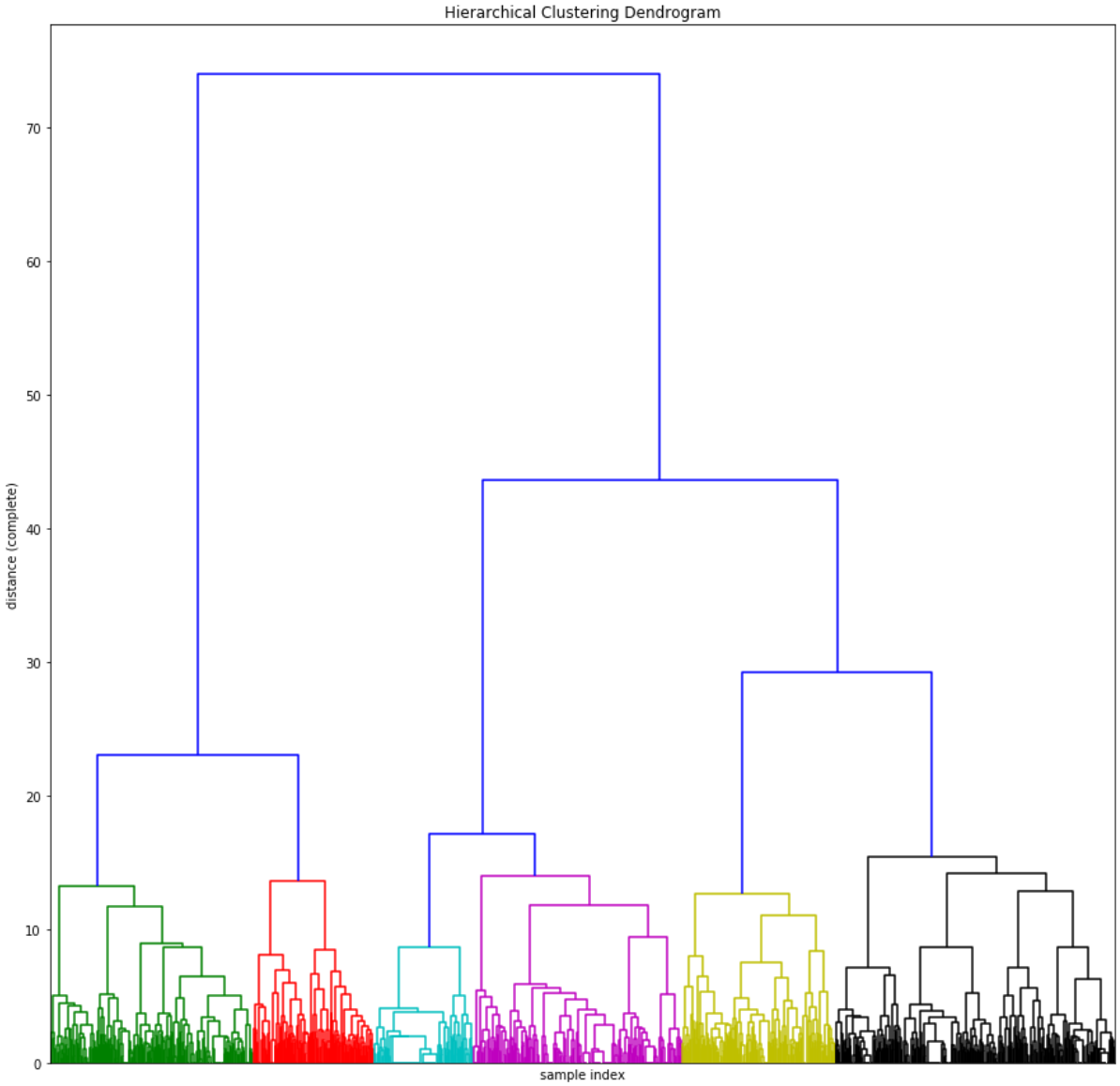

Рис. 8. Дендрограмма разделения на кластеры данных о продажах

Дальше в исследовательской работе были проведены эксперементы с помощью библиотеки *sklearn.cluster.AgglomerativeClustering* с разным набором параметров. Были рассмотрены следующие параметры связи:

- *Ward* – имеет тенденцию искать сферические кластеры, имеющие прочную связь внутри и очень отдаленные от других групп. Одной из характеристик является поиск кластеров схожих размеров. Работает только с Евклидовым расстоянием.

- *Complete* – связывает кластеры, используя их последующие наблюдения, то есть их самые отдаленные объекты. То есть кластеры, созданные с помощью этого метода, состоят в большинстве из очень похожих наблюдений, что делает результирующие группы довольно компактными.

- *Average* – образует кластеры, используя их центроиды и игнорируя их границы. Метод создает более большие группы, чем Complete. Кроме того, кластеры могут быть разных размеров и форм, в отличии от метода Ward.

Помимо параметров связи были использованы разные метрики дистанций:

- Евклидово пространство.
- *Manhattan*, похож на Евклидово, но дистанция считается по сумме абсолютных значений разницы между измерениями.
- *Cosine*, подходит для ситуаций, когда в выборке используется очень большое количество параметров.

Для визуального анализа результатов экспериментов был выбран метод *T-SNE*, который является одним самых современных подходов к визуализации многомерных объектов. *Stochastic Neighbor Embedding* основывается на сохранении пропорций расстояний между объектами [5]:

Предположим у нас есть три объекта x_i, x_j и x_k и известно, что расстояние между *i*-м и *j*-м объектом в α раз больше, чем расстояние между *i*-м и *k*-м объектом. Тогда в методе *SNE* требуется, чтобы в новом маломерном пространстве это свойство сохранялось.

Проведя эксперименты и проанализировав результаты, было выявлено, что параметр метрики *Ward* лучше всего подходит для кластеризации продуктов, проданных по собственным рецептам пользователей. Эксперименты с этими параметрами были проведены с разделением на 4, 5 и 6 кластеров. Визуальный подход дает сделать вывод, что количество кластеров 6 оптимально для исходных данных.

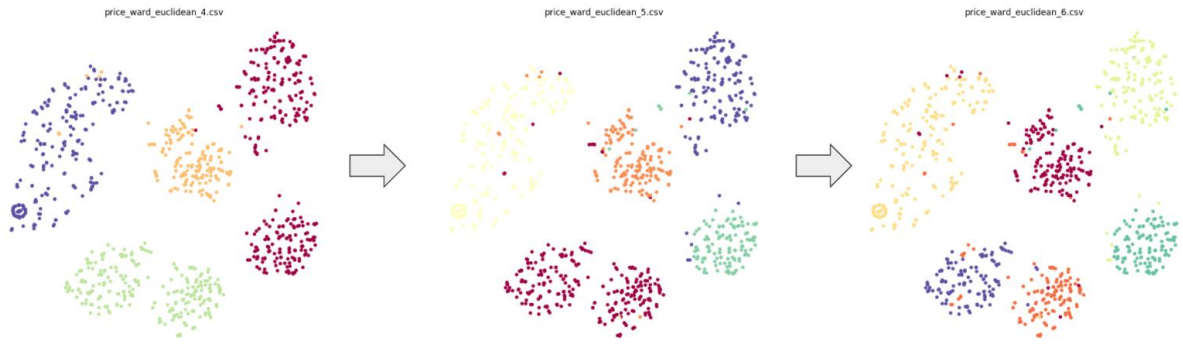

Рис. 9. Результат кластеризации объектов

При подробном изучении кластеров, были выделены основные составляющие рецептов в разных кластерах:

- *Cluster* 0: *Chicken, Bacon* + второстепенные ингредиенты, которые пользователи добавляют реже (Дальше +*N*).
- *Cluster* 1*: Basil, Italian Sausage +N*
- *Cluster* 2: *Mushrooms, Pepperoni +N*
- *Cluster* 3: *Pepperoni +N*
- *Cluster* 4: *Pepperoni, Italian Sausage +N*
- *Cluster* 5: *Italian Sausage, Bacon +N*

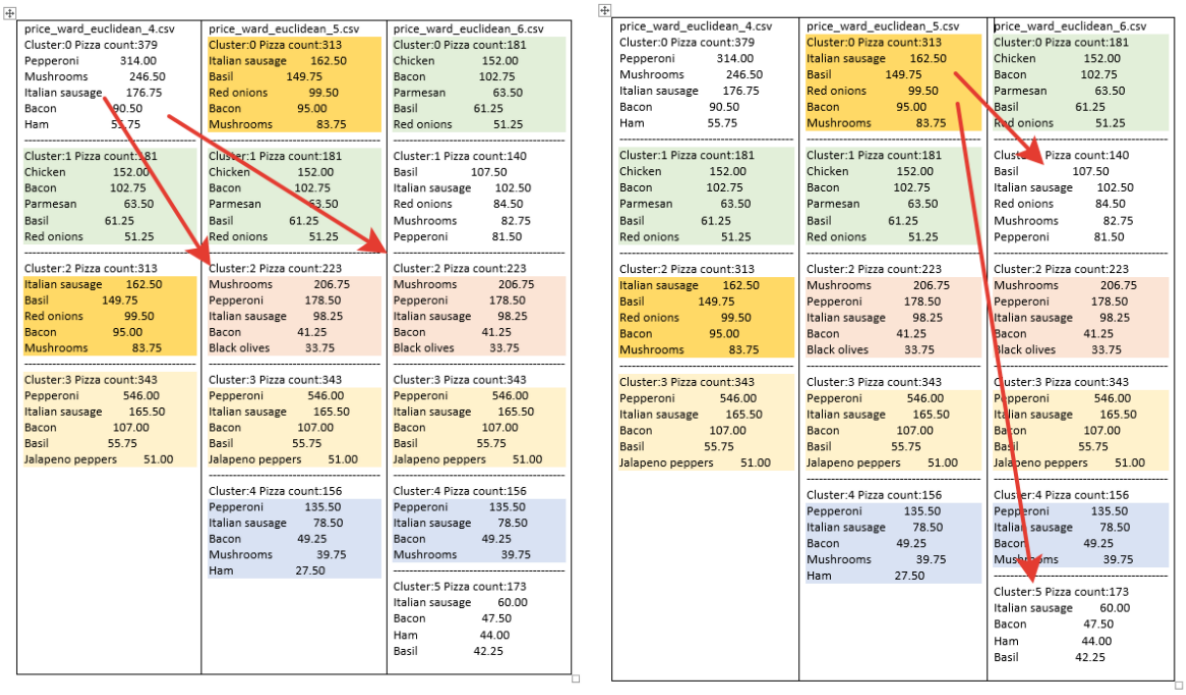

Рис. 10. Составы продуктов и их деление по кластерам

При увеличении количества кластеров, основа рецепта остается устойчивой, из чего можно сделать вывод о правильности выбора алгоритма кластеризации и подбора параметров.

Визуализировал цены в кластерах по месяцам, были получены следующие распределения цен:

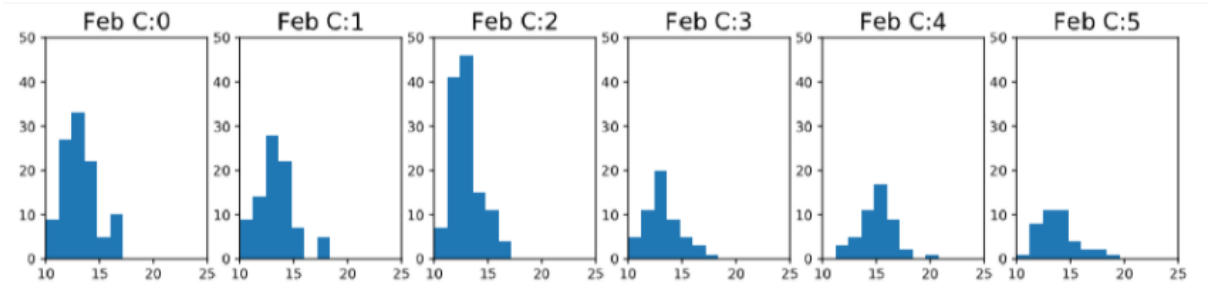

Рис. 11. Распределение цен по кластерам проданных продуктов в феврале 2018

На графиках видно, что в феврале в *Cluster 1 (Mushrooms & Pepperoni)* продано наибольшее количество, но цены на эти пиццы были не самыми высокими. В *Cluster 4 (Pepperoni +Italian Sausage)* продано меньше пицц, но цены на конечный продукт были выше, чем в *Cluster 1*.

Используя метки, полученные в ходе кластерного анализа, была обучена модель классификации и с ее помощью проставлены метки продуктам с готовыми рецептами. В результате удалось построить цепочки покупок клиентами. Пример построенной последовательности потребительского потребления: 21-(R3)-100-(R5)-12-(R5)-113-(R5)-138-(O5)-15-(O5)-38-(O5)-16-(O5)-29-(O5)-7-(O5) (2), где отдельное число – количество дней, прошедших между заказами. Символ O – заказ с продуктом на основе собственного рецепта, *R* – заказ продукта с готовым рецептом. Номер рядом с символом – порядок метки, к которому относится продукт.

Заключение

Кластерный анализ и последующее обучение модели классификации позволили построить цепочки поведения. Подход анализа данных с применением методов машинного обучения позволил вы-

явить зависимость во вкусовых предпочтениях пользователей от продуктов, созданных по их собственным рецептам. Эти данные используются для создания персонализированных маркетинговых кампаний с большей ориентацией на вкусы клиентов.

Отчет, построенный на таком анализе позволяет лучше планировать заготовки ингредиентов для продаж пиццы данного типа и избегать непредвиденных издержек.

Дальнейшее исследование последовательностей потребительского опыта позволит разработать рекомендательную систему и встроить ее в основное программное обеспечение. Когда пользователь будет переходить в конструктор для заказа продукта по собственному рецепту, можно предлагать ему начальные ингредиенты, которые входят в основу определенных кластеров, выбранных на основе его истории заказов. Это облегчит выбор для покупателя, и может повысить конечную стоимость продукции, что напрямую будет влиять на выручку сети пиццерий.

Список литературы

1. Papa John's Int'l. Inc. — [Электронный ресурс]. URL: https://www.entrepreneur.com/franchises/papajohnsintlinc/282662.
2. Preliminary Results for the Year ended 31 December 2017. — [Электронный ресурс]. URL: http://www.dpeurasia.com/Dosyalar/report/dpeu-2017-prelims-presentation-vf.pdf.
3. How to market a restaurant…without marketing it. — [Электронный ресурс]. URL: https://dodopizzastory.com/2017/09/12/how-to-market-restaurant/.
4. Метод k-средних. — [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_k%D1%81%D1%80%D0%B5%D0%B4%D0%BD%D0%B8%D1%85.
5. Метод t-SNE. — [Электронный ресурс]. URL: https://www.coursera.org/learn/unsupervised-learning/lecture/Bn22S/mietod-t-sne.
6. Иерархическая кластеризация — [Электронный ресурс]. URL: https://basegroup.ru/community/glossary/hierarc-clust.
7. Агломеративная иерархическая кластеризация — [Электронный ресурс]. URL: https://www.coursera.org/learn/unsupervised-learning/lecture/STrfh/aghlomierativnaia-iierarkhichieskaia-klastierizatsiia.
8. Дендрограмма. — [Электронный ресурс]. URL: https://basegroup.ru/community/glossary/dendrogram.