Аффинитивный анализ чеков покупок

Аффинитивный анализ — это метод в области данных, который используется для выявления закономерностей или ассоциаций между элементами в больших наборах данных. Это один из наиболее популярных методов для анализа данных, связанных с покупками, например, в розничной торговле или онлайн-торговле. Аффинитивный анализ помогает найти товары, которые часто покупаются вместе, а также изучать другие взаимосвязи между элементами данных.

Основные понятия аффинитивного анализа: ассоциативные правила, support (поддержка), confidence (доверие), lift (подъём).

В рамках проведенного анализа был выполнен аффинитивный анализ чеков покупок с целью выявления скрытых закономерностей в поведении покупателей. Задача заключалась в исследовании взаимосвязей между товарами, которые часто покупаются вместе, и построении моделей для улучшения маркетинговых стратегий и оптимизации товарных предложений.

Для анализа использовались данные чеков покупок, которые были собраны со студентов нашей группы. В процессе работы с данными мы выполнили несколько ключевых шагов:

* Сбор данных: Собраны данные по транзакциям и товарам, купленным в одном чеке, включая информацию о времени покупки, количестве товаров и ценах.
* Предобработка данных: Данные были очищены от выбросов, затем преобразованы в нужный формат для дальнейшего анализа.
* Смёрдживание данных: Для комплексного анализа мы объединили несколько источников данных в один итоговый файл. Это позволило учитывать как отдельные товары, так и информацию о покупательских предпочтениях на разных уровнях.
* Использование Google Colab: Для обработки и анализа данных был использован Google Colab, где мы реализовали код на Python с использованием библиотек Pandas и ML-algorithms для построения моделей.

Алгоритм Apriori — это один из самых популярных алгоритмов для поиска ассоциативных правил в базе данных транзакций. Он был предложен Аграфи, Султаном и Шивашакти в 1994 году и является основным методом для анализа частых наборов и поиска ассоциативных правил в больших наборах данных, например, при анализе покупок.

У данного алгоритма есть несколько принципов:

* Принцип "если это частое, то это подмножество частого": Алгоритм основан на важном принципе, который гласит, что если набор товаров является частым (то есть его поддержка превышает заданный порог), то все его подмножества также должны быть частыми. Это позволяет значительно уменьшить количество проверок и сделать алгоритм более эффективным.
* Итеративный процесс: Алгоритм работает итеративно. Он начинает с поиска частых одиночных элементов (например, одиночных товаров) и постепенно расширяет их до более сложных комбинаций.

Association Rule Mining (или Анализ ассоциативных правил) — это метод анализа данных, используемый для обнаружения интересных связей или закономерностей между элементами в больших наборах данных. Этот метод применяется в различных областях, включая маркетинг, финансовый анализ, интернет-торговлю и многие другие.

Цель Association Rule Mining — найти правила, которые описывают зависимости между элементами данных. Например, в ритейле это может быть правило, которое говорит: "Если покупатель купил молоко, он с высокой вероятностью купит хлеб". Эти правила называются ассоциативными правилами.

Алгоритм FP-growth (Frequent Pattern Growth) — это один из самых эффективных алгоритмов для решения задачи поиска часто встречающихся наборов элементов (частых паттернов) в больших наборах данных, таких как транзакционные базы данных. Он является альтернативой алгоритму Apriori, но работает намного быстрее и эффективнее, особенно когда данные имеют большое количество транзакций.

Этот график показывает связь продукта "Бананы" с другими продуктами, которые покупают вместе с ним. Из результата видно, что продукты, такие как "Сыр", "Газировка", "Алкоголь", "Вода", "Молоко", "Хлеб", "Картофель", "Курица" и "Колбаса", чаще всего встречаются в одних чеках с "Бананами".

Толщина линий на графике указывает на интенсивность связи (чем толще линия, тем чаще эти продукты покупаются вместе). Такой анализ помогает выявить популярные товарные комбинации, которые могут быть использованы для улучшения выкладки товаров в магазинах или создания акций.

Аналогичный график построим для продукта "Вода". По нему также можно сделать выводы о том, какие продукты покупают вместе с ним.

График показывает частоту совместной покупки 20 самых популярных товаров. Наиболее сильная взаимосвязь наблюдается между базовыми продуктами, такими как "Молоко" и "Хлеб", "Курица" и "Картофель". Это указывает на типичные повседневные покупки. Более слабые связи видны между менее часто совмещаемыми продуктами, такими как "Конфеты" и "Сливки".

На графике видно, что "Молоко" и "Хлеб" являются лидерами по количеству покупок, за ними следуют "Бананы" и "Сыр". Это указывает на то, что данные продукты наиболее востребованы среди покупателей. Остальные продукты, такие как "Алкоголь", "Вода" и "Газировка", занимают средние позиции, что говорит о популярности базовых и повседневных товаров.

График показывает, что наиболее часто покупаются товары с низкой средней ценой (до 500 рублей), тогда как дорогие продукты (выше 2000 рублей) приобретаются значительно реже. Это подтверждает, что цена сильно влияет на спрос.

График показывает тенденцию роста общих расходов на покупки в течение времени, с пиками в середине ноября. Это может указывать на сезонные покупки или акционные периоды. Небольшие спады между пиками могут отражать завершение активных периодов покупок.

График показывает распределение итоговой стоимости чеков. Большинство транзакций сосредоточено в диапазоне до 500 рублей, что указывает на низкую стоимость средних покупок. Среднее значение (315.73 руб.) выше медианы (150 руб.), что свидетельствует о наличии редких, но более крупных затрат, которые смещают среднее. Квартильные значения (Q1: 83.25 руб., Q3: 305.50 руб.) показывают значительное скопление данных в нижней части распределения.

Смоделируем алгоритм Apriori. Сначала сгенерируем набор часто используемых элементов, а затем правила ассоциации, используя набор часто используемых элементов. Нам нужно убедиться, что мы сгенерировали матрицу со значениями 0/1, представляющими наличие транзакции этого элемента.

Данная таблица содержит результаты анализа ассоциативных правил. Каждый столбец представляет собой различные метрики и характеристики, связанные с выявленными ассоциативными правилами, такими как поддержка, доверие, подъём и другие.

antecedents

Это левая часть правила, или предпосылка. В данном случае, это набор товаров, который должен быть в корзине покупок для того, чтобы произошло событие, описанное правой частью правила.

consequents

Это правая часть правила, или следствие. Это товар (или товары), который покупатель с вероятностью приобретет, если он уже купил товары из левой части.

antecedent support

Это поддержка для левой части правила (предпосылки). Поддержка показывает, насколько часто товары из предсказуемой части (antecedents) встречаются в транзакциях.



consequent support

Это поддержка для правой части правила (consequents). Показывает, насколько часто товары из правой части (consequents) встречаются в транзакциях.

support

Это поддержка всего ассоциативного правила A⇒BA \Rightarrow BA⇒B, то есть процент транзакций, которые содержат как товары из левой, так и из правой части правила.



confidence

Это доверие к правилу, то есть вероятность того, что при наличии товаров из левой части правила покупатель также приобретёт товары из правой части.



lift

Это подъём (или lift) — метрика, показывающая, насколько сильнее связаны товары в правой части правила с товарами в левой части, по сравнению с их независимостью. Если lift > 1, то товары в правой части с большей вероятностью покупаются вместе с товарами из левой части. Если lift = 1, то товары независимы.



representativity

Это метрика, которая отражает, насколько правило соответствует выборке. В идеале, она должна быть близка к 1, что означает, что правило точно соответствует данным.

leverage

Leverage измеряет отклонение от независимости между элементами. Это метрика для проверки того, насколько больше (или меньше) события происходят вместе, чем ожидалось бы при случайном распределении.



conviction

Conviction (или убеждённость) — это мера, которая оценивает, насколько сильным является правило. Conviction показывает, насколько часто можно ожидать, что правило не будет выполняться.



zhangs\_metric

Zhang's Metric — это метрика, предложенная для оценки значимости правила с учетом разницы в вероятностях.

jaccard

Jaccard Index — это коэффициент сходства между наборами AAA и BBB. Он вычисляется как отношение пересечения наборов к их объединению.



certainty

Это метрика, аналогичная confidence, но учитывающая вероятности невыполнения правила.

kulczynski

Это метрика, которая оценивает силу ассоциации, комбинируя confidence и support.



Количество продуктов (antecedents)

Это количество товаров в левой части правила (предпосылке).

Левая часть (antecedents)

Это перечень товаров из левой части правила (предпосылки).

Правая часть (consequents)

Это перечень товаров из правой части правила (следствие).

Эта таблица позволяет глубоко анализировать ассоциативные правила между товарами, оценивая их частоту, вероятность совместной покупки, степень зависимости и другие метрики.

На графике представлены популярные пары продуктов и количество транзакций, в которых они одновременно встречались. Самая частая комбинация — [Алкоголь, Огурцы], она встречается 8 раз. Также востребованы пары [Сыр, Рис], [Хлеб, Масло] и [Молоко, Бананы] с 6 транзакциями.

Менее популярные комбинации, такие как [Хлеб, Молоко] и [Помидоры, Огурцы], встречаются 1-2 раза. Этот анализ полезен для планирования размещения товаров и акций, ориентированных на совместные покупки.

Теперь сгенерируем тепловую карту ассоциативных правил для минимального порога поддержки в 1%.

Этот график представляет собой тепловую карту ассоциаций продуктов (Lift), построенную на основе данных о совместной покупке продуктов. Она отображает связи между продуктами и их относительную силу (поддержка ассоциации), выраженную через метрику Lift.

Горизонтальная ось (правая часть - consequents): отображает продукты или их комбинации, которые являются "следствием" в ассоциации.

Вертикальная ось (левая часть - antecedents): отображает продукты или их комбинации, которые являются "предшествующими" в ассоциации.

Каждая ячейка на пересечении строки и столбца показывает значение Lift для соответствующей пары "antecedent → consequent".

Справа находится цветовая шкала, где синий цвет (более светлый) соответствует низким значениям Lift. Красный цвет (более насыщенный) обозначает высокие значения Lift. Чем выше значение Lift, тем сильнее ассоциация между конкретными продуктами.

В каждой ячейке указано числовое значение Lift. Например, для "Курица, Морковь" и "Лук, Морковь" Lift равен 14.17.

На графике видны области с высокими значениями Lift, что указывает на частое совместное приобретение соответствующих товаров. Например, связь "Сыр, Хлеб → Молоко" имеет высокое значение Lift (32.60).

Яркие красные ячейки, такие как в сочетаниях "Картофель, Масло → Соль" и "Колбаса, Бананы, Молоко → Хлеб", выделяются как самые сильные ассоциации.

На представленной тепловой карте выделены 15 наиболее значимых ассоциаций между продуктами, измеряемых с использованием метрики Lift.

Сильнейшие ассоциации (максимальные значения Lift):

Ассоциации с насыщенным красным цветом и самым высоким значением Lift: "Соль, Лук → Соль": Lift = 46.57. Это самая сильная связь на графике. Она указывает на то, что при наличии соли и лука в корзине соль почти гарантированно оказывается покупаемым товаром.

Выраженные ассоциации с высокими значениями Lift (35–40): "Соль → Хлеб": Lift = 36.22. Соль в корзине часто сопровождается хлебом. Это может отражать популярность этих товаров в повседневных покупках. Хлеб часто дополнительно появляется как сопутствующий товар.

Средние ассоциации (30–35): "Яйца, Сыр → Хлеб": Lift = 32.60. Эта связь показывает, что сочетание яиц и сыра нередко дополняется хлебом. Это логично для приготовления завтраков или бутербродов. Также встречается связка "Бананы, Колбаса → Хлеб" с таким же Lift. "Бананы, Сыр → Хлеб": Lift = 31.05. Присутствие этих продуктов в корзине указывает на высокую вероятность покупки хлеба. "Масло → Хлеб": Lift = 31.05. Ассоциация подтверждает, что масло часто покупается вместе с хлебом. "Молоко, Морковь → Хлеб": Lift = 31.05. Это интересная связь, указывающая на то, что хлеб часто становится дополнением к молоку и моркови, возможно для приготовления закусок.

Менее выраженные, но значимые ассоциации (28–30): "Яйца → Хлеб": Lift = 28.98. Продукты завтрака, такие как яйца, часто покупаются с хлебом. "Бананы, Яйца, Сыр → Хлеб": Lift = 28.98. Эта связь укрепляет тренд покупок на завтрак, где яйца, бананы, и сыр сопровождаются хлебом.

Замеченные комбинации категорий продуктов: Хлеб и "утренние" продукты: Хлеб часто связан с продуктами, которые могут быть частью завтрака или быстрых перекусов, такими как яйца, сыр, масло. Соль и универсальные товары:  
Соль чаще всего ассоциируется с хлебом и луком. Это может отражать кулинарные привычки — соль, лук, и хлеб используются в большом количестве блюд. Бананы и хлеб: Бананы появляются в нескольких связках (с сыром, колбасой) и везде дополняются хлебом. Это может говорить о популярности сочетания фруктов и хлеба в корзинах.

Далее стоит представить диаграммы, на которых изображены визуализации связей между наборами продуктов (ассоциаций), построенных на основе анализа правил ассоциативного обучения.

Это одноуровневая визуализация ассоциаций. Она отображает связи между продуктами, представленных как "Antecedent" (слева) и "Consequent" (справа).

Antecedent (Предшествующий продукт): Товары, которые уже находятся в корзине покупателя.

Consequent (Последующий продукт): Товары, которые с высокой вероятностью добавляются в корзину в результате.

Это двухуровневая визуализация ассоциаций, где используется два уровня Antecedent (Antecedent 1 и Antecedent 2), которые приводят к Consequent.

Antecedent 2 и 1: Представляют два разных товара, которые могут одновременно находиться в корзине.

Consequent: Товар, добавляемый после них.

Этот график демонстрирует связь между несколькими наборами товаров (Antecedent 1, Antecedent 2, Antecedent 3) и их последствиями (Consequent). Линии представляют правила ассоциации, соединяющие товары.

Каждая связь показывает, что если покупают товары из левой группы (Antecedents), то с определённой вероятностью покупают товар из правой группы (Consequent).

Здесь связи выглядят более сложными, с большим количеством пересечений.

Этот график углубляет анализ, добавляя ещё один уровень Antecedent (Antecedent 4).

Более детализированный разбор, показывающий, как группы товаров (до четырёх уровней Antecedents) связаны с Consequent.

Линии менее плотные, но выражают более глубокие уровни связей между товарами.

Если разбирать итоговый результат, то можно сказать, что на графике представлена параллельная координатная визуализация ассоциативных правил между товарами. Каждый столбец отображает продукты, которые могут быть включены в левую (Antecedent) или правую (Consequent) часть правил. Толщина линий между продуктами указывает на силу связи (Lift).

Основные наблюдения:

* Продукты, такие как "Молоко", "Хлеб" и "Яйца", часто выступают как основные связующие элементы.
* Связи между продуктами в правилах имеют высокую плотность, что говорит о частых совместных покупках.
* Некоторые товары, такие как "Печенье" и "Колбаса", демонстрируют значительные ассоциации с множеством других продуктов.

Это указывает на значимость анализа совместных покупок для оптимизации продаж и маркетинга.

Данный график показывает следующее:

* Продукты, такие как Колбаса, Сыр, Хлеб, и Яйца, имеют высокую взаимосвязь, так как они часто покупаются вместе.
* Молоко и Яблоки также выступают как популярные связанные товары.
* Связи между категориями помогают выявить основные группы продуктов, что может быть использовано для планирования акций или улучшения расположения товаров в магазине.

На представленном графе отображена сеть связей между элементами (продуктами или объектами) на основе выявленных ассоциаций. Узлы (черные точки) представляют продукты, а линии между ними показывают наличие связи или ассоциации, основанной на частоте совместного появления в транзакциях.

Ключевые моменты:

* Узлы с большим числом соединений указывают на продукты, часто ассоциируемые с другими.
* Слабые связи (менее плотные линии) могут указывать на менее значимые ассоциации.
* Общая структура графа демонстрирует, как продукты взаимосвязаны, что может быть полезным для оптимизации выкладки или анализа потребительских предпочтений.

В ходе работы был выполнен аффинитивный анализ покупательских чеков с использованием данных о транзакциях. Основной задачей являлось выявление закономерностей в покупательских предпочтениях и построение моделей ассоциативных правил для анализа связей между товарами. Для этого были проведены этапы предобработки данных, визуализации, анализа частотных наборов и построения ассоциативных правил.

На этапе предобработки данные были очищены и приведены к удобному формату. Количество и стоимость товаров в транзакциях были агрегированы, а для последующего анализа была создана бинарная матрица товаров и транзакций. Эта матрица стала основой для построения моделей ассоциативных правил.

Визуализация данных включала построение различных графиков, таких как распределение общей стоимости транзакций, тренды покупательской активности по времени и тепловые карты совместной покупки товаров. Например, тепловая карта показала соотношение частот совместной покупки товаров и позволила выделить основные популярные пары продуктов. Это дало возможность визуально оценить степень ассоциативности между товарами и обнаружить ключевые взаимосвязи.

С помощью алгоритма Apriori были выявлены частотные наборы товаров, которые часто встречаются в одной транзакции. На их основе были сгенерированы ассоциативные правила, отражающие, какие товары покупаются вместе, с указанием таких метрик, как lift, confidence и support. Эти метрики позволили оценить силу связей между товарами и их значимость в анализе.

Для улучшения наглядности правил была построена тепловая карта ассоциаций, отображающая lift для наиболее значимых пар товаров. Визуализация с использованием PyARMViz дала возможность проследить ассоциации в виде параллельных категорий, что облегчило понимание сложных взаимосвязей в данных.

В ходе анализа были выявлены ключевые пары товаров с наибольшим lift, что говорит о сильной ассоциации между ними. Например, такие товары, как "Хлеб" и "Молоко", "Яйца" и "Сыр", чаще всего встречались в одной транзакции. Это указывает на то, что данные продукты имеют высокую вероятность покупки вместе, что может быть полезно для оптимизации выкладки товаров в магазинах или создания маркетинговых акций.

Дополнительно, анализ распределений стоимости транзакций и их изменения во времени помог выявить особенности поведения покупателей. Большинство транзакций имели небольшую общую стоимость, но также были зафиксированы высокие траты, что свидетельствует о редких крупных покупках.

Итогом работы стало создание инструмента для выявления и визуализации покупательских паттернов. Этот инструмент может быть полезен для повышения продаж через рекомендации или планирование выкладки товаров в торговых точках. Проделанный анализ подтвердил эффективность использования методов анализа частотных наборов и ассоциативных правил для обработки данных о транзакциях.