

Грязнов Сергей Александрович

канд. пед. наук, доцент, декан

ФКОУ ВО «Самарский юридический институт ФСИН России»

г. Самара, Самарская область

CHERRY PICKING: МАНИПУЛЯЦИЯ ДАННЫМИ

***Аннотация:** статья посвящена исследованию избирательного подхода в аналитике данных под названием «cherry picking». Данный подход понимается как выборочное и предвзятое извлечение данных или информации для последующего анализа. Сделан вывод, что избирательный подход, выборочное извлечение и анализ точек данных, которые поддерживают желаемый результат, игнорируя при этом противоречивую информацию, представляет собой значительный риск для целостности и объективности анализа.*

***Ключевые слова:** cherry picking, анализ данных, аналитический вывод, сбор данных, манипуляция, продвижение интересов.*

Термин «cherry picking» (в данном контексте выбор вишни или по аналогии – снятие сливок) становится все более распространенным в анализе данных, где целостность и объективность принятия решений имеют ключевое значение. Cherry picking – выборочное извлечение данных или информации и последующий анализ этого набора для выявления конкретных точек, которые соответствуют предвзятым представлениям или желаемым результатам. Этот прием может использоваться намеренно или непреднамеренно, но в любом случае он приводит к предвзятым или ошибочным выводам.

Чтобы лучше понять значение и последствия подхода cherry picking, важно углубиться в его определение и рассмотреть влияние, которое он оказывает на процессы принятия решений. Следует отметить, что cherry picking присутствует в различных областях: от научных исследований до политического дискурса, где данные используются для поддержки аргументов, манипулирования общественным мнением и продвижения личных интересов.

Однако именно в контексте анализа данных «выбор вишни» создает серьезные проблемы. Выборочно отбирая точки данных, которые поддерживают заранее определенный результат, аналитики вносят значительную погрешность в анализ (ищут или интерпретируют данные таким образом, чтобы подтвердить уже существующие убеждения или гипотезы). Поэтому подобные предубеждения подрывают объективность и препятствуют обнаружению альтернативных идей. «Выбор вишни» может происходить в различных формах:

- выборочное включение данных: выбор включения только подмножества данных, подтверждающих определенный вывод, и исключение данных, которые ему противоречат;

- искажение данных: представление данных в вводящей в заблуждение форме, например, с использованием неподходящих масштабов или методов визуализации, которые искажают истинную природу данных.

- переоснащение моделей: создание слишком сложных моделей, которые хорошо соответствуют обучающим данным, что приводит к плохому обобщению новых, невидимых данных.

- p-hacking: манипулирование данными или статистическими тестами для получения статистически значимого результата, даже если он не отражает истинные основные взаимосвязи [1].

Выборочный подход игнорирует ценность всестороннего анализа, а также потенциально подрывает доверие к инициативам и стратегиям, основанным на данных. Ниже рассмотрим несколько примеров «выбора вишни».

Так, можно заметить, что результаты анализа в бизнес-аналитике часто бывают положительными. Однако выделяя положительные, и игнорируя отрицательные отзывы, маркетологи создают искаженное представление об общем успехе продукта (услуги) и уровне удовлетворенности клиентов [2].

В научных исследованиях избирательность данных может привести к ошибочному научному дискурсу (когда исследователи выбирают данные или эксперименты, подтверждающие их гипотезы, игнорируя при этом противоречивую

информацию). Такой предвзятый подход препятствует прогрессу в области исследований.

Средства массовой информации также могут выбирать данные для поддержки определенных нарративов. Например, новостное издание может сообщать только о положительных аспектах деятельности политического кандидата, игнорируя отрицательные моменты. Таким образом, отобранные данные могут использоваться для манипулирования информацией в поддержку определенной повестки дня. Это ведет к ложным обещаниям и политике, не направленной на решение реальных проблем.

Аналитики данных должны стремиться обеспечить объективность аналитического процесса. Этого можно достичь путем систематического подхода к сбору данных, отдавая приоритет комплексным наборам данных над выборочными и обеспечивая прозрачность процесса отбора. Использование надежных аналитических и статистических методов может помочь выявить и смягчить последствия *cherry picking*. Используя соответствующие статистические тесты и методологии, аналитики могут гарантировать, что их выводы основаны на всеобъемлющей и репрезентативной выборке данных, что снижает риск предвзятости и искаженного анализа.

Таким образом, аналитики данных играют решающую роль в поддержании целостности и достоверности анализа. Содействуя прозрачности отчетности, включая раскрытие всех соответствующих данных, используемых в анализе, аналитики обеспечивают критическую оценку и облегчают принятие решений на основе данных. В свою очередь, прозрачная отчетность гарантирует, что лица, принимающие решения, осознают ограничения, предвзятости и потенциальные ловушки, связанные с *cherry picking*.

Важным инструментом анализа данных является визуализация, поскольку она помогает ясно и кратко представить результаты анализа. Визуализация может помочь выявить закономерности, тенденции и взаимосвязи в данных, которые могут быть неочевидны при использовании таблиц или текста. Однако при этом необходимо убедиться, что визуализация точна. Например, гистограмма, на

которой сравниваются продажи различных продуктов, может вводить в заблуждение, если шкала непоследовательна.

Этика – неотъемлемая часть ответственного анализа данных. Этичное исследование включает в себя получение информированного согласия от участников исследования, защиту их частной жизни и конфиденциальности, а также предотвращение причинения вреда отдельным лицам или сообществам.

Как определить, что данными манипулируют? Одним из наиболее распространенных признаков мошенничества являются исключительно положительные данные. Если кто-то представляет данные, которые лишь подтверждают озвученные аргументы, игнорируя при этом данные, которые им противоречат, то, скорее всего, это – cherry picking. Чтобы выявить необъективность, следует найти недостающие данные или данные, которые были исключены из анализа. Если данные неполны или предвзяты, то вполне вероятно, что аргумент также является предвзятым [3].

Как сказано выше, «выбор вишни» часто предполагает выбор данных, которые поддерживают определенную модель или тенденцию, игнорируя при этом данные, которые ей противоречат. Чтобы это выявить следует найти закономерности в данных и убедиться, что они соответствуют общей тенденции. Если данные противоречивы или не соответствуют общей тенденции, то возможно, аргумент является пристрастным.

Также важную роль играет контекст данных (данные могут быть вырваны из контекста или представлены в вводящем в заблуждение виде). Например, исследование, изучающее эффективность конкретного лекарства на пациентах, может дать разные результаты, если оно проводится в разных странах или с разными демографическими группами.

Следовательно, избирательный подход, выборочное извлечение и анализ точек данных, которые поддерживают желаемый результат, игнорируя при этом противоречивую информацию, представляет собой значительный риск для це-

лостности и объективности анализа. Чтобы снизить риски, организации и аналитики должны уделять приоритетное внимание прозрачности, надежным аналитическим методам и приверженности объективному выбору данных.

Ответственный анализ данных необходим для принятия обоснованных решений, обеспечения справедливости и укрепления доверия. Проявляя прозрачность, принимая во внимание контекст, соблюдая этику и сотрудничая, можно гарантировать, что анализ данных будет ответственным и окажет положительное влияние.

Список литературы

1. Andrade C. HARKing, Cherry-Picking, P-Hacking, Fishing Expeditions, and Data Dredging and Mining as Questionable Research Practices. *J Clin Psychiatry*. 2021 Feb 18; 82 (1):20f13804. doi: 10.4088/JCP.20f13804. PMID: 33999541.

2. Татаринов К.А. Психологические инструменты манипуляции поведением потребителя / К.А. Татаринов, А.Ф. Зверев [Электронный ресурс]. – Режим доступа: <https://www.elibrary.ru/item.asp?id=43913863> (дата обращения: 17.04.2024)

3. Бушуева Л.И. О манипуляциях статистическими данными / Л.И. Бушуева, М.В.Карманов, В.И. Кузнецов [Электронный ресурс]. – Режим доступа: <https://www.elibrary.ru/item.asp?id=48418349> (дата обращения: 17.04.2024).