

ОГЛАВЛЕНИЕ

Об авторе	8
Благодарности	10
От издательства.....	11
Введение	12
Для кого эта книга	14
Как читать эту книгу	15
Глава 1. Как мы принимаем решения	17
Четыреста сравнительно честных способов	20
Чему можно научиться у Amazon?.....	21
Аналитический паралич.....	23
Погрешности — правило штангенциркуля.....	25
Принцип Парето	26
Можно ли принимать решения только на основе данных?.....	27
Глава 2. Делаем анализ данных	31
Артефакты анализа данных	33
Бизнес-анализ данных	34
Гипотезы и инсайты	34
Отчеты, дашборды и метрики.....	37
Артефакты машинного обучения.....	41
Артефакты инженерии.....	43
Кто анализирует данные	45
Идеальная кнопка	46

Продать аналитику внутри компании.....	48
Конфликт исследователя и бизнеса	49
Недостатки статистического подхода в аналитике	51
Глава 3. Строим аналитику с нуля	55
Первый шаг	56
Выбираем технологии.....	57
Поговорим об аутсорсе	61
Наем и увольнения.....	64
Кому подчиняются аналитики	67
Должен ли руководитель аналитики писать код.....	68
Управление задачами.....	71
Как управлять романтиками	76
Глава 4. Делаем аналитические задачи	79
Как ставить задачи аналитикам	80
Как проверять задачи.....	83
Как тестировать и выкладывать изменения в рабочую систему	87
Как защищать задачу перед инициатором.....	87
Нужно ли уметь программировать?	88
Датасет.....	90
Описательная статистика	91
Графики	94
Общий подход к визуализации данных	98
Парный анализ данных.....	100
Технический долг.....	101
Глава 5. Данные	103
Как собираются данные.....	104
Big Data.....	105
Связность данных	106
Много данных не бывает	107
Доступ к данным	108

Качество данных	110
Как проверяется и контролируется качество данных	112
Типы данных	114
Форматы хранения данных	116
Способы получения данных.....	120
Глава 6. Хранилища данных.....	121
Зачем нужны хранилища данных.....	122
Слои хранилища данных.....	125
Какие бывают хранилища.....	126
Как данные попадают в хранилища	130
Hadoop и MapReduce.....	132
Spark	136
Оптимизация скорости работы	140
Архивация данных и устаревание.....	141
Мониторинг хранилищ данных	143
Личный опыт.....	144
Глава 7. Инструменты анализа данных.....	147
Электронные таблицы.....	148
Сервисы блокнотов	149
Инструменты визуального анализа	151
Пакеты статистического анализа данных.....	153
Работа с данными в облаках.....	154
Что такое хорошая отчетная система	156
Сводные таблицы	159
OLAP-кубы.....	164
Корпоративные и персональные BI-системы.....	168
Мой опыт	170
Глава 8. Алгоритмы машинного обучения	171
Типы ML-задач	174
Метрики ML-задач	178

ML изнутри	182
Линейная регрессия	183
Логистическая регрессия	184
Деревья решений.....	186
Ошибки обучения	188
Как бороться с переобучением	190
Ансамбли.....	194
Глава 9. Машинное обучение на практике	197
Как изучать машинное обучение	198
Соревнования по ML	200
Искусственный интеллект	202
Необходимые преобразования данных	204
Точность и стоимость ML-решения	207
Простота решения	208
Трудоемкость проверки результата	210
Mechanical Turk / Yandex Toloka	210
ML и большие данные	212
Recency, Frequency и Monetary.....	212
Последний совет.....	215
Глава 10. Внедрение ML в жизнь: гипотезы и эксперименты.....	217
Гипотезы	218
Планируем тест гипотезы	220
Что такое гипотеза в статистике	222
Статистическая значимость гипотез.....	226
Статистические критерии для р-значений	229
Бутстрэп	232
Байесовская статистика	234
А/Б-тесты в реальности	237
A/A-тесты	240
Еще несколько слов о А/Б-тестах	242

Что делать перед А/Б-тестом	244
Конвейер экспериментов	245
Глава 11. Этика данных	247
Как за нами следят	248
Хорошее и плохое использование данных	255
Проблема утечки данных	258
Этика использования данных	260
Как защищают пользовательские данные	263
Глава 12. Задачи и стартапы	267
Веб-аналитика в рекламе	268
Внутренняя веб-аналитика	271
Маркетинг на основе баз данных	276
Стартапы	280
Личный опыт	284
Глава 13. Строим карьеру	291
Старт карьеры	292
Как искать работу	294
Требования к кандидатам	296
Вы приняли оффер	298
Как развиваться и работать	298
Когда менять место работы	302
Нужно ли все знать?	304
Эпилог	308
Список литературы	309

12

ЗАДАЧИ И СТАРТАПЫ



В этой главе я расскажу о проблемах, которые стоят перед современными компаниями в e-commerce, а также о способах их решения. Думаю, вам также пригодится мой опыт создания Retail Rocket в качестве одного из учредителей.

ВЕБ-АНАЛИТИКА В РЕКЛАМЕ

Веб-аналитика — это предметная область, которая изучает поведение людей в интернете. Веб-аналитику я делю на две части: оценка эффективности рекламы и анализ взаимодействия пользователей с сайтом.

Начну с анализа эффективности интернет-рекламы. Есть крылатая фраза Джона Ванамейкера (1838–1922) — легендарного американского коммерсанта, революционера в торговле (он открыл первый универсам и первым применил ценники) и отца современной рекламы: «Я знаю, что половина моего рекламного бюджета расходуется впустую, вот только не знаю, какая именно».

Раньше я искренне считал, что именно интернет-реклама положит конец пустому расходованию денег и станет намного эффективнее рекламы на телевидении и в печати. Например, вы показали в телеэфире ролик — как теперь измерить его эффективность? Есть несколько способов: от изменения графика продаж в момент показа рекламы до опроса аудитории с целью узнать, насколько повысилась осведомленность. Для печатной рекламы, помимо этих методов, существует еще один, более точный — использование промокодов на скидку или подарок. По числу введенных промокодов можно определить условную эффективность рекламы.

С интернет-рекламой все стало проще. Все ссылки помечаются специальными тегами, например utm-метками. Обратите на них внимание, когда кликаете на рекламе. После перехода на сайт на компьютер пользователя записываются так называемые куки-файлы (cookies), по которым сайт узнает этого посетителя, когда он туда вернется. С помощью этого механизма можно отследить

покупки пользователя, сделанные через несколько дней или недель после перехода с рекламы. Не правда ли, что это выглядит намного точней, чем при традиционной оффлайн-рекламе? Именно так я наивно и считал в далеком 2005 году, когда только начал заниматься оценкой эффективности рекламы в онлайне. Тогда не было такого количества рекламы и перекрестных переходов, поэтому ее влияние отслеживалось хорошо.

В наши дни рекламы стало не просто много, а очень много, и пользователь перед покупкой порой делает несколько переходов с разных источников рекламы. Вначале он может искать что-то в поисковике, перейти на сайт интернет-магазина с поисковой рекламы, сделать в магазине пару кликов, уйти с сайта. Через несколько дней он может вернуться на сайт с так называемой ретаргетинговой рекламой (например, этим занимается уже известная нам Criteo), зарегистрироваться в магазине, бросить товар в корзину и уйти с сайта. Скорее всего, через несколько часов или даже минут он (или она) получит письмо — «вы забыли оформить заказ, ваш товар уже в корзине». Пользователь возвращается на сайт магазина из письма и совершает заказ. Внимание, вопрос: благодаря какой рекламе пользователь сделал покупку? Кажется очевидным, что если бы не его первый переход из поисковой системы, магазин точно не получил бы заказ. Но как быть с остальными двумя — ретаргетинговой рекламой и письмом с просьбой завершить заказ? Действительно ли они повлияли на результат в этой цепочке переходов?

В стандартных инструментах веб-аналитики обычно выигрывает последний клик (last click attribution). В нашем примере это письмо о забытом заказе, но его бы не было без первых двух переходов. Это называется проблемой реатрибуции — когда разные источники рекламы «бьются» между собой за заказ. Как посчитать эффективность рекламы, если было несколько разных переходов с источниками рекламы перед целевым действием, например заказом? Чтобы ответить на этот вопрос наверняка, нужно провести А/Б-тест — половине людей показывать ретаргетинг, другой — нет. Половине людей отправить письмо, другой — нет. А если эффективность ретаргетинга и email зависят друг от друга? В теории

можно было бы сделать сложный многофакторный тест — но на практике это невыполнимо. А/Б-тесты такого типа в интернет-рекламе — очень сложные и достаточно дорогие, так как приходится отключать часть интернет-рекламы, а это падение выручки. Многие великие умы бьются над созданием альтернативных способов расчета эффективности рекламы. Возможно, рано или поздно они выработают систему, в основе которой будет лежать некий вероятностный подход: например, давать больший вес начальным переходам. Чтобы построить такую модель, нужно сделать много А/Б-тестов, которые обойдутся очень дорого, но при этом все равно получить некую частную, а не общую модель, которую невозможно распространить на всю индустрию.

В рекламной веб-аналитике вы еще встретитесь с двумя терминами — сквозная аналитика и когортный анализ. Под сквозной аналитикой обычно понимают работу с клиентом на индивидуальном уровне: от показа рекламы до отгрузки заказа отслеживания последующих действий заказчика. Это делается с помощью уникальных идентификаторов клиента (ID), с помощью которых его «ведут» в разных системах, от рекламных до логистических. Благодаря этому можно считать затраты на рекламу и обработку заказов с точностью вплоть до индивидуального клиента, пусть и с некоторым приближением.

Когорта в маркетинге — это группа людей, которые совершили определенное действие в заданный промежуток времени. Под когортным анализом подразумевается отслеживание таких однородных групп клиентов. Самое главное его назначение — расчет LTV (Life Time Value), количества денег, которые приносит клиент за определенный промежуток времени. Предположим, вы определили, что этот период будет составлять три месяца, и решили считать LTV первого числа каждого месяца (рис. 12.1). Каждый расчетный месяц аналитик будет «смотреть» на клиентов, которые совершили свой первый заказ или регистрацию три месяца назад, и считать их покупки за эти три месяца, потом делить это число на число клиентов. Для такого расчета нельзя использовать клиентов, которые совершили первое действие четыре или два месяца назад.

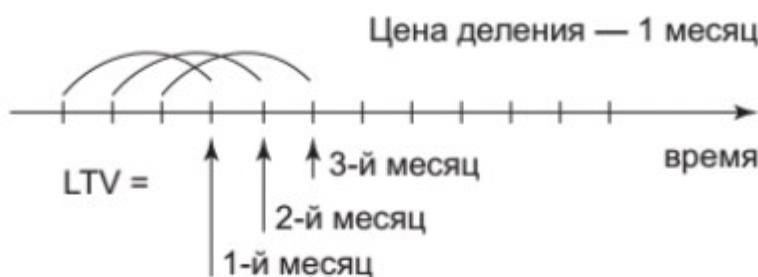


Рис. 12.1. Расчет LTV

ВНУТРЕННЯЯ ВЕБ-АНАЛИТИКА

Внутренней веб-аналитике сайта уделяется не так много внимания, как рекламной — на рекламу тратится куда больше денег, чем на сайт, поэтому руководство хочет знать, насколько эффективно они потрачены. А ведь действия посетителя на сайте, которые как раз являются объектом внутренней аналитики, очень важны. В этот анализ входят: воронка продаж, анализ заполнения форм и анкет, мерчандайзинг, функционализм сайта, карты кликов, запись действий пользователя (например, Яндекс.Вебвизор). Используя эти инструменты, можно гораздо лучше понимать свою аудиторию.

Воронка продаж выглядит почти как обычная воронка — посетитель сайта «проваливается» по ней, пока не достигнет целевого действия, например заказа. В среднестатистическом интернет-магазине конверсия посетителя в заказ составляет обычно один процент, то есть лишь каждый сотый посетитель доходит до дна воронки продаж и совершает покупку. Улучшению этого показателя уделяется очень много времени, ведь если растет конверсия сайта, то вы зарабатываете больше при тех же затратах на рекламу. Хотя рекламе рознь: можно гнать на сайт небольшой поток почти готовых покупателей или большую толпу посетителей, подавляющее большинство которых уйдет с сайта сразу. В первом случае конверсия может быть высокой, во втором низкой, но и стоить первый вариант будет дороже. Поэтому я не сторонник «меряться» конверсиями, более важный показатель — средняя стоимость привлеченного заказа (Cost per Order). Он позволяет объективно сравнить экономику двух интернет-магазинов в первом

приближении. Воронку продаж можно также рассматривать как последовательность микрошагов из целевых действий:

1. Сделал хотя бы один клик после перехода (non-bounced visitor).
2. Добавил товар в корзину.
3. Нажал кнопку «оформить заказ» (checkout).
4. Оформил заказ.

Оптимизируя каждый шаг, можно увеличить число посетителей, которые доходят до конца воронки.

Анализ мерчандайзинга — это самое лучшее, что я узнал о внутренней веб-аналитике, когда изучал систему Omniture (ныне Adobe) SiteCatalyst. Анализ мерчандайзинга — это способ оценки эффективности виртуальных полок интернет-магазина. Сайт интернет-магазина включает в себя несколько типов страниц: главная, поиск, страница категории товаров, страница информации о товаре, корзина, шаги заказа и личный кабинет пользователя. На каждом типе страниц размещаются блоки товаров (рис. 12.2) — например, горизонтальная линия из пяти ротируемых товаров или большой блок списка товаров на страницах категории. В любом товарном

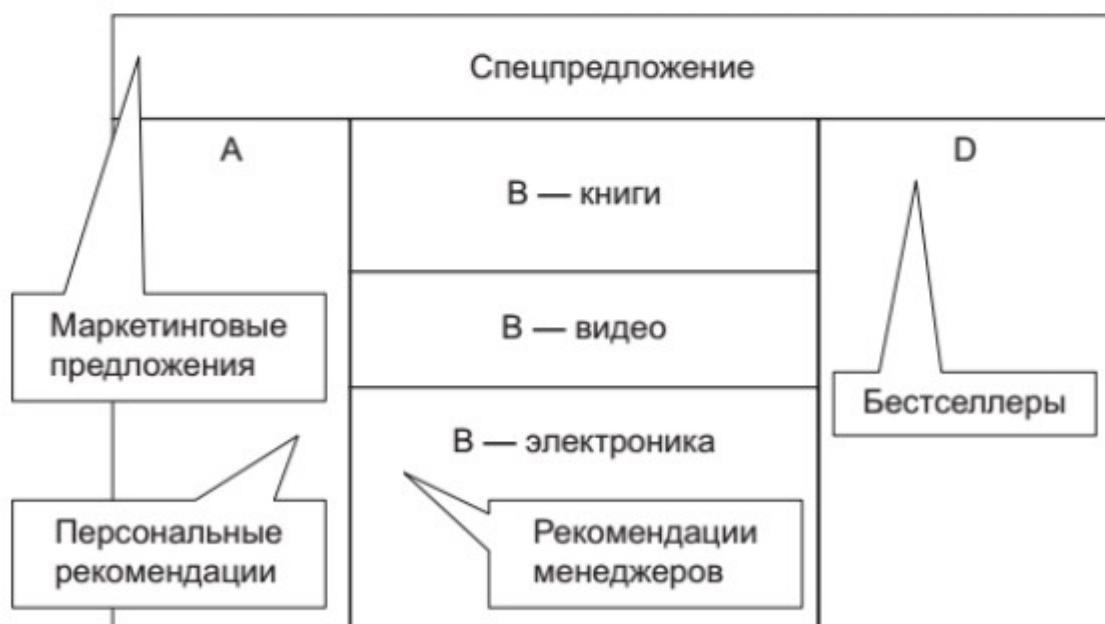


Рис. 12.2. Пример мерчандайзинга сайта интернет-магазина

блоке товар подается со следующими атрибутами: картинка, сниппет с небольшой информацией, цена, название товара, кнопка добавления в корзину или быстрого заказа. Что можно делать с дизайном подачи товара в блоке? Можно увеличить картинку, убрать какие-то элементы. А вот посчитать, что изменилось в метриках, можно с помощью анализа мерчандайзинга, где аналогом обычной полки в магазине будет блок товаров в интернет-магазине.

Сам анализ работает следующим образом: все ссылки на товары (картинки, названия, кнопка добавления в корзину) помечаются специальными невидимыми тегами, где могут быть указаны тип страницы (главная, поиск и другие), название блока (горизонтальный, листинг), тип ссылки (картинка, название, кнопка добавления в корзину). Для каждого клика на таком блоке система запоминает, на каком товаре в каком блоке какой пользователь кликнул. Затем система в течение заранее установленного времени (например, 24 часа) следит за пользователем, что он будет делать с этим товаром после клика. Если пользователь добавил его в корзину или заказал, то эта метрика будет приписана к тому невидимому тегу, который был при клике. На выходе вы можете получить следующую статистику (табл. 12.1).

Таблица 12.1. Расчет эффективности мерчандайзинга

Название тега	Клики	Добавления в корзину	Заказы
Главная / Рекомендации первой строки / Картинка товара	1000	350	20
Поиск / Листинг / Название товара	15 000	3 000	500
Страница товара / Рекомендации аналогов / Картинка товара	50 000	15 000	2000

Обычно я выгружаю такую статистику в Excel, разбиваю тег на три поля (тип страницы, тип блока и тип ссылки) и получаю возможность легко решать следующий круг задач:

- Каков вклад в продажи каждого типа страницы? Например, 15 лет назад я вычислил, что страница поиска Ozon.ru дает половину от всех добавлений в корзину на сайте.
- Каков вклад рекомендательных блоков в продажи? На момент моего ухода из Ozon.ru система рекомендаций обеспечивала около 38 % всех добавлений в корзину.
- Откуда чаще покупают — после клика на картинке товара или на его названии? Тогда я выяснил, что чаще кликают на изображении, но названия товаров дают больше продаж.

Когда аналитик может это считать, у компании появляется неограниченное поле для экспериментов «а что, если»: увеличить картинки товаров, убрать картинки из поиска, поменять местами блоки товаров, изменить алгоритм рекомендаций в блоке товаров. Если у вас есть метрики мерчандайзинга, появляется гораздо больше возможностей для модификации сайта.

Напишу про некоторые нюансы этого типа анализа. Во-первых, там есть такая же проблема реатрибуции тегов, как и в рекламе: пользователь через поиск на сайте кликнул на товаре, через некоторое время он кликнул на том же товаре в блоке рекомендаций и купил его. К чему атрибуцировать товар — к странице поиска или блоку рекомендаций на сайте? Есть две стратегии: выиграл первый и выиграл последний. В первом случае этот заказ получит страница поиска, во втором — блок рекомендаций. Однозначного ответа на вопрос, какая стратегия лучше, нет. Я лично предпочитаю вариант «выиграл первый». Во-вторых, вычисления для анализа мерчандайзинга намного более затратны по сравнению с анализом рекламы. Из-за этого Omniture SiteCatalyst отказался поднимать время слежения за действиями пользователя с 24 часов до 7 дней, и мне пришлось пользоваться метрикой добавления в корзину, а не заказа, потому что в течение двадцати четырех часов после первого визита на сайт человек, как правило, не делает заказ, но успевает положить товар в корзину. Обращайте внимание, как вендоры веб-аналитики работают с мерчандайзингом: у Яндекс.Метрики такого нет и не планируется, у Google Analytics есть Enhanced

Ecommerce, у Adobe Analytics есть анализ мерчандайзинга [114]. Я изучал документацию по внедрению двух последних систем и могу сказать, что в Adobe Analytics это сделано намного лучше, чем в Google Analytics. Я сам заимствовал эту идею и написал свой алгоритм расчета, который используется и по сей день компанией Retail Rocket для вычисления эффективности рекомендаций на сайтах клиентов.

Карта кликов на странице – интересный инструмент, но ее нужно очень серьезно настраивать, если работа идет с динамическими блоками, когда товары там ротируются. Я обычно старался заменять ее на анализ мерчандайзинга, а саму карту рисовать в редакторе. Это позволяло мне сделать усредненную карту кликов для страницы товара, когда самих товаров около 500 тысяч. Никакая карта кликов сама по себе с этим не справится, а анализ мерчандайзинга может.

Еще один полезный инструмент – «видеозапись» действий пользователя. Его умеет делать Яндекс.Метрика, сам инструмент называется вебвизор. Он сохраняет все действия небольшой части пользователей, включая движения мыши. Потом вы можете просмотреть такие записи в интерфейсе программы. Это напомнило мне книгу Пако Андерхилла «Как заставить их покупать». В этой книге автор рассказывает, как он расставляет огромное количество камер в магазинах клиентов, сутками смотрит видеозаписи, дает рекомендации, как изменить пространство магазина, чтобы больший процент посетителей совершили покупку. Точно так же можно использовать и вебвизор. К сожалению, инструмент недооценен либо по причине слабой информированности, либо из-за неудобства в использовании. Этот способ – хорошая альтернатива дорогим системам юзабилити, например трекерам глаз.