Урок 4

Введение в АБ-тесты

4.1. Что такое АБ-тестирование

4.1.1. Тестирование идей

Во всех сферах бизнеса постоянно требуется улучшать его ключевые показатели. Идеи возможных улучшений могут возникать, с одной стороны, исходя из анализа рынка, поведения пользователей, их потребностей, а с другой — из знания об устройстве бизнеса, желаемом направлении развития.

Хочется уметь как-то проверять такие идеи: их много, а применять нужно только самые успешные. Существует много различных методов проверки. Прежде всего, идея должна соответствовать здравому смыслу. Как правило, это устанавливается в разговоре с коллегами, и принимается решение: стоит ли дальше заниматься идеей или от неё можно отказаться уже сейчас.

Затем бывает полезно протестировать идею на реальных пользователях. Как правило, такие проверки стоят дорого из-за привлечения людей, которые тратят на проверку свое время. Однако такой способ позволяет получить более релевантную оценку, потому что тестирование происходит на тех же пользователях, что

ляет получить более релевантную оценку, потому что тестирование происходит на тех же пользователях, что и будут потом пользоваться услугой.

Существует ряд требований, предъявляемых к таким проверкам. С одной стороны, хочется, чтобы они были достоверными, то есть условия, в которых проверяется идея, были максимально приближены к реальным. С другой стороны, не хочется тратить на это очень много времени и денег.

В качестве проверки качества идеи может выступать опрос пользователей: заранее подготавливаются какие-то вопросы, а затем людей просят на них ответить.

4.1.2. Фокус-группы

«Фокус-группы» — это ещё один метод проверки идеи. Набирается небольшая группа пользователей, и от каждого их них поступает очень много информации: что он чувствует по отношению к новому баннеру; как именно на него кликает; производятся наблюдения, как долго его взгляд задерживается на каких-то деталях сайта. По сравнению с опросами, данные, полученные с помощью фокус-групп, глубже, однако такие исследования нельзя провести на большом количестве людей, поскольку это дорого.

Казалось бы, фокус-группы — почти идеальный механизм проверки бизнес-идей. Однако есть следующий нюанс: не всегда удается с точностью воссоздать те самые условия, в которых пользователи будут использовать продукт. Известный пример — тестирование количества сахара в напитке Кока-Кола. В рамках исследования была собрана фокус-группа, которой предложили на выбор два напитка: со стандартным содержанием сахара и увеличенным. Людям нужно было попробовать оба и выбрать тот, который им больше понравился. В результате этого исследования выяснилось, что большее количество людей предпочитает Кока-Колу с увеличенным количеством сахара. По результатам исследования количество сахара в напитке увеличили и предложили его более широкой аудитории. Неожиданно продажи упали. Эксперты стали разбираться, почему так произошло, ведь в рамках фокус-группы было показано, что больше сахара — это вкуснее. Выяснилось, что это исследование проходило не в тех самых условиях, в которых люди обычно пьют Кока-Колу. Если речь идёт всего лишь об одном стакане, то, действительно, людям нравится большее количество сахара. Однако если напиток употребляется постоянно в больших количествах, то больше сахара — хуже. Кажется, что это логично: сложно выпить большое количество очень сладкого напитка. Поэтому, когда проводятся исследования на фокус-группах, нужно следить за тем, чтобы условия были максимально приближены к настоящим.

Однако это не всегда возможно.

4.1.3. АБ-тестирование

Проблема приближения условий к настоящим решается полностью в рамках другого способа оценки идей: «A/B testing». «A/B testing» — это способ проверки идей непосредственно в боевых условиях. Пользователям предлагают новую функциональность или новый элемент дизайна именно в тех условиях, в которых они взаимодействуют с продуктом.

Очевидный плюс А/Б-теста в том, что, с одной стороны, условия больше похожи на настоящие. С другой стороны, финансовые риски максимально снижены: А/Б-тестирование проводится на небольшой группе пользователей, поэтому максимальные потери можно заранее спрогнозировать.

4.2. Где используется АБ-тестирование

В области IT A/B-тесты используются практически повсеместно. Любые сайты, независимо от их профиля, используют механизм A/B-тестирования для принятия решения о внесении каких-то изменений. Кроме того, такой метод используется в приложениях, играх и вообще во всем, что взаимодействует с конечным пользователем.

Во всех этих сферах нужно тестировать такие вещи, как изменения в дизайне, изменения функциональности, новые возможности для пользователя или изменения в алгоритмах.

С одной стороны, хочется как можно меньше времени тратить на эксперименты и проверить множество идей сразу. С другой стороны, изменения разных типов плохо тестируются одновременно. Может оказаться, что разные изменения просто технически несовместимы друг с другом. Или, например, одно из изменений действует на бизнес-показатель положительно, а другое — отрицательно, и, применяя их одновременно, не получится разделить эти два эффекта.

Такие техники тестирования идей применяются не только в IT, но и в довольно неожиданных местах, например, для оптимизации работы государственных органов. При правительствах США и Великобритании есть небольшие группы, которые совместно с психологами-бихевиористами выдвигают гипотезы о том, как

люди взаимодействуют с государством, и на основании таких гипотез проводятся эксперименты: например, небольшие изменения в дизайне налоговой формы, или изменения в способе записи на донорство или трансплантацию органов. Такие вещи, которые легко и очень дешево можно поменять, оказывается, приводят к тому, что государство может сэкономить миллионы.

В целом, процесс А/Б-тестирования распадается на две большие части. Первая часть — это планирование эксперимента: как именно будет выглядеть А/Б-тест, как пользователей будут делить на группы, сколько будет длиться тест и многие другие вопросы, связанные с этим. Вторая часть — это непосредственно проверка гипотез, принятие решения о том, положительно или отрицательно влияют изменения на бизнес в целом.

4.3. Метрики

4.3.1. Что такое метрики

Итак, для того, чтобы показать эффективность идеи, нужно провести эксперимент, в котором она применяется, и при этом получить улучшение некого показателя. Этот показатель необходимо выбрать перед тем как проводить эксперимент. Речь идёт о выборе метрик. Часто для того, чтобы показать, что состояние бизнеса улучшилось, а именно, что его ключевые показатели изменились в ожидаемом направлении, нужно выбрать некоторые метрики, связанные напрямую с состоянием бизнеса.

Чаще всего это метрики, связанные непосредственно с деньгами, или аудиторные метрики. Однако при их применении возникают проблемы. Во-первых, часто их сложно измерить, во-вторых, они бывают достаточно грубыми и практически не реагируют на небольшие изменения в функциональности или дизайне. Кроме того, во многих случаях требуется очень много времени, чтобы измерить интересующие метрики. Например, после внесения изменения в сайт с арендой квартир, требуется узнать, как увеличилось количество людей, рекомендующих этот сервис своим знакомым. Чтобы измерить этот показатель, может понадобиться целый год, потому что люди переезжают не так уж часто.

4.3.2. Промежуточные метрики

Это важное замечание приводит к идее использования так называемых прокси-, или промежуточных метрик. Это такие метрики, которые, с одной стороны, достаточно чувствительны, чтобы измерять их в рамках А/Б-тестирования, а с другой стороны, хорошо согласуются с теми бизнес-показателями, которые в реальности требуется измерить. Например, при внесении изменений в сайт с арендой квартир в качестве такой метрики может быть использована метрика «среднее количество визитов на сайт в день», «среднее количество уникальных пользователей» или «количество шеров сайта в социальных сетях». Эти метрики обладают требуемыми свойствами промежуточных метрик.

4.3.3. Оффлайн-тестирование

Важным этапом при принятии решения о том, следует ли проверять изменение в A/B-тестинге, является оффлайн-тестирование. В рамках оффлайн-тестирования можно по историческим данным проверить, как определённые изменения сказываются на поведении пользователей.

Допустим, изменён алгоритм ранжирования результатов по запросам пользователей при поиске квартир на сайте недвижимости. В этом случае можно поступить следующим образом: проанализировать запросы пользователей в прошлом (если известно, что пользователи искали на сайте ранее), и эмулировать выдачу новым алгоритмом.

Таким образом, с одной стороны, доступна информация о том, что пользователи искали и на какие позиции они кликали, какие ответы показались им релевантными. С другой стороны, теперь имеется новое ранжирование, новые результаты поиска. Совместив эти данные, можно проверить, на какие позиции теперь приходятся клики пользователей. В результате можно получить такие метрики, как «средняя позиция клика». Если значение этой метрики уменьшается, то, наверное, внесённое изменение хорошее. Значит, имеет смысл протестировать его в онлайне. А если, например, новый алгоритм ранжирования совсем не находит те результаты, которые пользователи посчитали релевантными, то, возможно, этот алгоритм не стоит тестировать на реальных пользователях.

Возникает следующая иерархия метрик: предварительные метрики, измеряемые до начала эксперимента, экспериментальные, на основании которых принимается решение о том, хорошее изменение или плохое, и

бизнес-метрики, которые измеряются в самом конце и на изменение которых направлены все действия.

4.4. Дизайн эксперимента

4.4.1. Стратификация и рандомизация

Для проведения эксперимента требуется небольшая группа пользователей, которой будут предъявлены изменения. Для того, чтобы результаты, полученные на этой небольшой группе можно было обобщать на всех пользователей, группа должна быть репрезентативной. Это значит, что её структура должна совпадать со структурой набора всех пользователей. Например, если известно, что 2/3 пользователей продукта — женщины, то в экспериментальной группе должно быть 2/3 женщин.

Таким образом, при построении экспериментальной группы можно выделять какие-то важные свойства пользователей: например, возраст, или другие интересующие характеристики, — а затем искусственно делать так, чтобы в экспериментальной группе были ровно такие же доли по разным подгруппам, как и среди пользователей в целом. Такой подход называется стратификацией.

Другой подход, в каком-то смысле противоположный ему, — это рандомизация. Если набирать пользователей в экспериментальную группу абсолютно случайно, то в среднем она получится такого же состава, как и вся генеральная совокупность пользователей. Дополнительный плюс рандомизации заключается в том, что при этом экспериментальная группа пользователей выравнивается со генеральной совокупностью по всем возможным показателям, а не только по тем, которые показались важными.

4.4.2. Связанные выборки

В некоторых случаях оказывается важным измерить, как на пользователя влияет несколько воздействий сразу, например, как он реагирует на сайт без изменений и на сайт с изменениями. Такой дизайн эксперимента называется парным, или связанным. Выборки результатов получаются не зависимые, а связанные, и это очень выгодно в ситуациях, когда измеряемый показатель имеет большую индивидуальную дисперсию (то есть пользователи очень сильно отличаются по этому показателю).

При связанном дизайне эксперимента зачастую оказывается важным, в каком порядке пользователю предъявляются разные варианты. Для того, чтобы снять влияние порядка, можно использовать дизайн крест-накрест: половине пользователей показать сначала новый вариант, потом — старый, а другой половине — наоборот.

4.4.3. Проведение нескольких экспериментов сразу

Одновременно можно проводить большое количество экспериментов. Но в этой ситуации не возникает никаких проблем только до тех пор, пока каждый пользователь участвует в одном эксперименте. Если существует вероятность, что каждый пользователь попадает сразу в несколько экспериментальных групп, нужно внимательно следить за тем, чтобы эти эксперименты друг другу не противоречили.

Например, известна история о том, как Google тестировал 41 оттенок синего в цвете ссылок в поисковой выдаче. Если допустить, что одновременно еще проводился бы эксперимент о том, как выбрать цвет страницы или цвет, на фоне которого показываются эти ссылки, очевидно, что они не должны быть тех же самых цветов, что и текст ссылок, иначе пользователь просто не сможет ничего прочитать. То есть пример подбора одновременно цвета текста и цвета фона, на котором он показывается, — это пример экспериментов, которые друг с другом не сочетаются.

4.5. Устойчивость

Одно из важнейших требований к А/Б-тестированию, которое обязательно должно быть заложено в дизайн эксперимента, — это требование устойчивости. В данном случае под устойчивостью понимается следующее: во-первых, хочется не видеть значимых изменений там, где их на самом деле нет, во вторых, если какие-то значимые изменения есть, то хочется, чтобы они отражались на метриках.

Это очень просто понять на примере. Пусть в эксперименте участвуют две одинаковые версии сервиса. В данном случае, конечно же, на всех метриках хочется видеть одинаковый результат. С другой стороны, если в одну из версий внесены некоторые значимые изменения, то, конечно, хочется увидеть это на метриках и убедиться, что по всем метрикам есть значимый прирост.

4.5.1. Обратный эксперимент

Казалось бы, требования устойчивости очень простые, логичные и должны всегда выполняться. Однако на практике это часто не так. Например, есть некоторый сайт, который позволяет производить поиск по некоторому специфичному контенту, например, по объявлениям о продаже/аренде недвижимости. Можно изменить дизайн и проверить, как это повлияло на поведение пользователей: правда ли, что новый дизайн им нравится больше, и они начинают более активно пользоваться сервисом. Можно сделать очень простое изменение, например, перекрасить кнопку поиска из синего цвета в зеленый. В данном случае легко понять, как будет выглядеть А/Б-тестирование. Пользователей разобьют на тестовую и контрольную группу. Одной группе будут показывать кнопку синего цвета (старый дизайн), а другой группе пользователей — кнопку зеленого цвета (новый дизайн). Далее можно подсчитать онлайн-метрику (количество нажатий на эту кнопку), и посмотреть, как она изменилась. Часто в таких экспериментах можно наблюдать следующий эффект: количество кликов будет больше в контрольной группе (с новым дизайном). Трактовка этого может быть двоякой. Те пользователи, которые часто пользуются сайтом и уже привыкли к тому, что кнопка имеет синий цвет, могут удивиться, что что-то изменилось, и захотеть проверить, изменился ли только дизайн или, может быть, и поведение. Соответственно, они могут начать чаще нажимать на кнопку просто из любопытства. В данном случае важно убедиться, что наблюдаемые метрики не учитывают это изменение как значимое.

Для того, чтобы обезопасить себя от ситуации, в которой незначимые изменения принимаются за значимые, можно поступить следующим образом. Пусть классический А/В-тест показал, что новый дизайн лучше, то есть кнопка зеленого цвета больше нравится пользователям. По измеряемым метрикам (например, доле кликов или длине сессий) наблюдаются значимые улучшения. Логично сделать следующее: выбрать новый дизайн и применить его для всех пользователей, то есть показывать всем пользователям кнопку зеленого цвета. Однако можно поступить несколько хитрее: выбрать небольшую группу пользователей (например, меньше 1 %) и продолжать показывать им старый дизайн после выкатки нового. То есть будет отдельно существовать некоторая группа пользователей, которая в течение какого-то существенного промежутка времени будет видеть старый дизайн. Это позволит в течение большего срока рассчитывать те же самые метрики, что и в процессе А/Б-тестирования. После этого можно снова сравнить поведение пользователей, которые

видят новый дизайн, с теми, кто видит старый дизайн. В данном случае, если значимые изменения не будут наблюдаться, то можно сделать вывод о том, что, во-первых, дизайн эксперимента не позволяет отличать незначимые изменения от значимых (иначе результат первоначального А/Б-тестирования был бы таким же), а во-вторых, можно оставить любой дизайн, потому что пользователи их не отличают.

Такая техника называется обратным экспериментом. Ее идея заключается в том, что после классического А/Б-тестирования эксперимент продолжается: выделяется некоторая маленькая группа пользователей, которые продолжают видеть старое решение. Такой подход предоставляет возможность убедиться в том, что изменения действительно значимы и приводят к ожидаемому эффекту.

4.5.2. А/А-тестирование

Казалось бы, технология обратного эксперимента способна решить все проблемы и помочь очевидным образом отличать значимые изменения от незначимых.

Однако пусть А/Б-тестирование проводится часто и каждый раз демонстрирует значимые изменения метрик на целевой и контрольной группе. Тестируемые нововведения запускаются в технологию «обратный эксперимент», и оказывается, что на самом деле изменений нет. Конечно же, эта ситуация является крайне нежелательной, потому что на проведение А/Б-теста и обратного эксперимента тратится много времени. Возникает вопрос: можно ли заранее убедиться в том, что дизайн эксперимента (в частности, размер контрольной и целевой групп, а также длительности эксперимента) позволяет отличать значимые изменения от незначимых.

Для того, чтобы эту задачу решить, применяется технология А/А-тестирования. Она работает следующим образом. Пусть принято решение о проведении А/Б-тестирования нового алгоритма (например поиска или рекомендации). В таком случае классический А/Б-тест выглядел бы следующим образом: пользователей разделили бы на контрольную и тестовую группу и показывали бы разные алгоритмы в разных группах, после чего сравнили бы метрики, рассчитанные на разных группах.

Перед тем, как запускать классический А/Б тестинг, можно поделить пользователей на группы точно так же, как это сделали бы в рамках А/Б тестинга, но обеим группам демонстрировать один и тот же алгоритм. Этот эксперимент должен длиться ровно столько же, сколько бы длился А/Б тестинг. В результате нужно определить, видны ли значимые изменения на интересуемых метриках. Если значимые изменения

нужно определить, видны ли значимые изменения на интересуемых метриках. Если значимые изменения не видны, то это хорошо, потому что эксперимент не показывает значимые изменения там, где их нет. В противоположной ситуации, если вдруг появятся значимые изменения, это должно наводить на мысль, что в дизайне эксперимента что-то не так: например, эксперимент длится недостаточно долго, или пользователи неправильно разбиты на группы. В любом случае, это повод задуматься об ошибках в дизайне эксперимента.

4.5.3. Размер выборки

Итак, метрика, дизайн эксперимента, метод его анализа выбраны, вся экспериментальная инфраструктура в достаточной степени устойчива и практически всё готово к запуску эксперимента. Единственный вопрос, на которой остается ответить, — это как долго эксперимент должен длиться, и сколько пользователей должно быть в тестовой выборке, чтобы можно было с уверенностью ответить на поставленные вопросы.

Задача определения необходимого объема выборки тесно связана с тем, какой именно статистический инструмент будет использоваться для ее анализа. Для каждого конкретного критерия подбор необходимого объема выборки делается своим способом.

Для того, чтобы понять, какой объем выборки необходим, нужно зафиксировать некоторые параметры. Во-первых, минимальный размер эффекта, который хочется измерить. То есть, насколько большие отклонения от значения по умолчанию (показатель, который сохраняется, если изменения никак не влияют на пользователей) хочется наблюдать в эксперименте.

Следующий показатель, который необходимо зафиксировать, — это допустимые вероятности ошибок первого и второго рода. В А/Б-тестах, как правило, выдвигается нулевая гипотеза, что никакие примененные изменения не повлияли на пользователей, и она проверяется против альтернативы, что изменения как-то повлияли. Ошибкой первого рода в этой ситуации будет отвержение неверной нулевой гипотезы, то есть принятие изменений, которые на самом деле не влияют на пользователей. Ошибка второго рода — это, наоборот, отклонение действительно хороших и влияющих на пользователей изменений. В статистике, как правило, вероятность ошибки первого рода — 0.05, а вероятность ошибка второго рода — 0.2. В конкретном эксперименте стоимости ошибок первого и второго рода могут быть существенно разными, поэтому часто может оказаться выгодно вручную выбрать эти пороги.

Наконец, когда размер эффекта и допустимые вероятности ошибок зафиксированы, можно выбрать статистический критерий и использовать калькулятор мощности этого критерия. Вообще, для всех статистических критериев между собой связаны несколько величин: тип альтернативы, размер эффекта, размер выборки и допустимые вероятности ошибок первого и второго рода. Если зафиксировать какие-то из этих величин, то можно рассчитать оставшиеся, используя калькулятор мощности.