Игра №1

Принимаю решения на уровне значимости α = 0.05

Период проведения каждого эксперимента – 4 дня (на 4 день принимается решение о релизе или остановке)

Распределение уровней сложности активных экспериментов стараюсь делать равномерным.

Эксперимент, у которого в первый день было очень маленькое значение p-value приняла сразу, т.к. для принятия нулевой гипотезы и, соответственно, решении остановить эксперимент, p-value должно увеличиться почти в 100 раз и я посчитала это маловероятным. К тому же, эксперимент не крупный, поэтому в случае ошибки, она не будет дорогой.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Эксперименты, у которого на третий день p-value было нулевым, принимала, не дождавшись четвертого дня, из-за максимально низкого значения p-value почти к концу установленного срока.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

У крупных экспериментов с пограничными значениями p-value < α продлевала эксперимент еще на один день, чтобы не допустить ошибку первого рода и не запускать в релиз изменения, которые могут оказаться неудачными. Это важно, т.к. крупный релиз сильно влияет на прибыль.

Итог:

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Сравнивая полученную выручку 585 844 с пороговой в 450 000, можно сделать вывод, что стратегия принятия решений при уровне значимости α = 0.05 и сроком проведения каждого эксперимента в 4 дня оказалась успешной.

Попробую применить другую стратегию для выявления наиболее результативной.

Игра №2

Принимаю решения на уровне значимости α = 0.05 для крупных экспериментов, 0.06 – для средних и 0.1 – для мелких.

Приняла такое решение, т.к. чем мельче эксперимент, тем меньше его влияние и тем, я посчитала, важнее увеличить мощность, чтобы точнее находить удачные эксперименты. Увеличение ошибки первого рода не окажет сильного негативного влияния на доход, т.к. ее совершение означает, что новый функционал для пользователей по удобности такой же, как и старый, и не будет иметь негативного эффекта на пользователей.

Период проведения каждого эксперимента уменьшила до трех дней (на 3 день принимается решение о релизе или остановке).

Распределение уровней сложности активных экспериментов стараюсь делать равномерным.

Эксперименты, у которых было очень маленькое значение p-value (p-value < 0.001) принимаю, не дождавшись 3 дня, т.к. для принятия нулевой гипотезы и, соответственно, решении остановить эксперимент, p-value должно увеличиться минимум в 50-100 раз и я посчитала это маловероятным.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Эксперименты, у которых на второй день наблюдалось p-value < α и которые показывали положительный результат, я принимала, т.к. маловероятно, что при продолжении эксперимента статистически значимые изменения в лучшую сторону поменяется на статистически значимые изменения в худшую сторону.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Эксперименты, у которых на второй день наблюдалось p-value < α и которые показывали негативный результат, я останавливала, т.к. маловероятно, что при продолжении эксперимента статистически значимые изменения в худшую сторону поменяется на статистически значимые изменения в лучшую сторону.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Эксперименты, у которых на второй день наблюдалось p-value 1 (p-value > 0.8), я останавливала, т.к. маловероятно, что при продолжении эксперимента явно выраженное отсутствие статистически значимых изменений поменяется на статистически значимые изменения в лучшую сторону.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Итог:

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Сравнивая полученную выручку 727 403 с пороговой в 450 000 и с выручкой, полученной в игре №1 (585 844), можно сделать вывод, что изменения в стратегии о сокращении срока проведения каждого эксперимента до трех дней и сокращении срока отмены эксперимента или релиза в случаях, когда на соответствующий день результаты были статистически значимы, до двух дней оказалась успешной.

В данной стратегии игры немного измененные уровни значимости для средних и мелких экспериментов не повлияли на принятие решений, т.к. не было ситуаций, при которых решения, принятые на новом или старом уровне значимости, были бы различны. Поэтому я решила проверить еще одну стратегию с более значительными изменениями уровня значимости.

Игра №3

Принимаю решения на уровне значимости α = 0.05 для крупных экспериментов, 0.1 – для средних и 0.3 – для мелких.

Период проведения каждого эксперимента оставляю равным трем дням (на 3 день принимается решение о релизе или остановке).

Распределение уровней сложности активных экспериментов стараюсь делать равномерным.

Эксперимент принимается за счет нового уровня значимости (при уровне значимости, принятом в предыдущих экспериментах, он бы был отклонен):

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Средние и мелкие эксперименты, у которых было сильно маленькое значение p-value (p-value < 0.01) принимаю, не дождавшись 3 дня, т.к. для принятия нулевой гипотезы и, соответственно, решении остановить эксперимент, p-value должно увеличиться минимум в 5-10 раз и я посчитала это маловероятным. (Крупные эксперименты в случае ошибки первого рода будут оказывать более сильное влияние на пользователей, поэтому на них это решение не распространяется)

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Большие эксперименты, у которых было очень маленькое значение p-value (p-value < 0.001) принимаю, не дождавшись 3 дня, т.к. для принятия нулевой гипотезы и, соответственно, решении остановить эксперимент, p-value должно увеличиться минимум в 50-100 раз и я посчитала это маловероятным.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Эксперименты, у которых на второй день наблюдалось p-value < α и которые показывали негативный результат, я останавливала, т.к. маловероятно, что при продолжении эксперимента статистически значимые изменения в худшую сторону поменяется на статистически значимые изменения в лучшую сторону.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Итог:

Изображение выглядит как текст, снимок экрана, Шрифт, дизайн

Автоматически созданное описание

Сравнивая полученную выручку 816 954 с пороговой в 450 000 и с выручкой, полученной в игре №2 (727 403), можно сделать вывод, что изменения в стратегии об уменьшении уровня значимости для средних экспериментов до α = 0.1 и для маленьких до α = 0.3 оказалась успешной.

Вывод:

Самой успешной оказалась стратегия, принятая в третьей игре.

В этой стратегии были взяты разные уровни значимости α для экспериментов разного масштаба, а именно α = 0.05 для крупных релизов, α = 0.1 для небольшого изменения функционала и α = 0.3 для мелких доработок. Это позволило повысить мощность некоторых тестов и чаще находить изменения там, где они есть, без значимых негативных последствий для пользователей.

Длительность эксперимента была принята в 3 дня, а там, где раньше этого срока были выявлены статистически значимые результаты с маленькой вероятностью дальнейшего их изменения, решение принималось на второй день; в первый день принимались решения только относительно экспериментов, где p-value равнялось нулю. В результате более быстрое распространение новых фичей на пользователей положительно повлияло на итоговую прибыль компании.

Таким образом, аналитикам данного маркетплейса следует применять стратегию, описанную в игре №3.