0.1 Introduction

В начале рассмотрим задачу детекции антифрода на маркетплейсе, возможные причины и проблемы, которые могли бы возникнуть в проуессе разработки, а также фичи, которые можно было бы подавать модели машинного обучения в качестве создания алгоритма, поогающего решать данную задачу.
Если рассматривать задачу в целом, то можно сказать, что на маркетплейсах могут быть распространены различные формы мошенничества, которые могут варьироваться в зависимотси от агента, то есть можно рассматривать мошеничесво как и со стороны партнеров , то есть селлеров , так , на самом деле , и на стороне самих покупателей, что в целом может негативно отразиться на обеих сторонах, то есть испортить как и, допустим, метрику CSAT как и на стороне покупателей, что может увеличить churn rate и снизить retention , такой же эффект может быть и на стороне самих продовцов.

0.2 Fraud Detection System

Целью фродовых транзакций является извлечение незаконной выгоды и нанесение ущерба пользователям и платформе. Если рассматривать общие методы в целом, то можно определить несколько категорий, которые включают в себя действия реселлеров, конкурирующих магазинов, пунктов самовывоза и кардеров. Рассмотрим в общем каждый из методов совершения фрода:
Первый и один из наиболее популярных способов является создание поддельных аккаунтов продавцов, которые предлагают популярные товары по заниженным ценам и требуют оплаты вне платформы, после чего исчезают с деньгами, не отправляя товар. Реселлеры в основном действуют так, создают множество аккаунтов для получения максимальных скидок и бонусов, нарушая правила сервиса, разбивают заказы на мелкие, что приводит к увеличению нагрузки на логистику, так как, например, один реселлер может создавать сотни заказов в день, что в итоге приводит к тому, что клиенты получать заказы не вовремя, так как слоты доставки ограничены, что также плохо обображается на уровне счастья клиентов. Кроме того, реселлеры иногда используют пункты выдачи заказов (ПВЗ) маркетплейса или его партнеров как склады, то есть они могут заказывать товары с постоплатой и забирать только те, которые удалось продать на стороне, что приводит к возврату большого количества товара на склад и снижению доступности для других покупателей, также снижая прибыль маркетплейса в целом, то есть GMV на комиссию.
Следующая проблема касается недобросовестных конкурентов, которые создают фиктивные заказы на товары других поставщиков, отправляя их в отдаленные регионы, например, в Сибирь или на Дальний Восток. Товары долго находятся в пути или на складах, выводя средства конкурентов из оборота и создавая логистическую нагрузку, что в общем приводит к убыткам и снижению способности платформы оперативно выполнять заказы для реальных клиентов.
Пункты самовывоза часто участвуют в мошенничестве, зарабатывая на комиссиях за обработку заказов, схема, например, может быть следующей, находясь в малопосещаемых местах, создают фиктивные заказы для получения минимальных комиссий. Эти заказы редко выкупаются, что создает дополнительную логистическую нагрузку и резервирует стоки, предназначенные для реальных клиентов.

А также самый опасный вид фрода включает кардеров, которые используют украденные данные и банковские реквизиты для совершения мошеннических транзакций, что несет репутационные и финансовые риски для платформы. Отмены таких платежей банками приводят к финансовым потерям и штрафам от платежных систем, если уровень мошенничества превышает допустимые нормы.
Фродовые транзакции также могут быть и со стороны покупателей, которые могут совершить манипуляции с возвратами, когда покупатели возвращают использованные или поддельные товары, требуя возврат средств, и использование украденных кредитных карт для покупок, что приводит к убыткам для продавцов.
Почти все виды мошенников действуют схожим образом: они создают множество новых аккаунтов, используют VPN для сокрытия своего местоположения, применяют режим инкогнито в браузерах, а также вымышленные имена и адреса доставки, что помогает обходить системы безопасности и оставаться незамеченными, а также иногда используют эмуляторы, чтобы получить бонусы за скачивание приложений. Эти специфические методы иллюстрируют уникальные внутренние проблемы, с которыми сталкивается маркетплейс в борьбе с мошенничеством. Мошенники также могут использовать автоматизированные скрипты и боты для массовой регистрации аккаунтов и выполнения транзакций, что усложняет их отслеживание и блокировку, а также использорвание поддельынх телефонных номеров дл повторной регистрации и получение определенных бонусов. Эти разнообразные стратегии требуют от маркетплейса постоянного совершенствования механизмов выявления и предотвращения мошеннической активности.
Для каждой из таких транзакций, для борьбы с такими действиями маркетплейсы применяют машинное обучение для выявления аномалий, системы верификации пользователей, анализ поведения для отслеживания подозрительных действий, также строит различные скоринговые модели и кластеризацию, чтобы на основе всех данных, которые были связаны с активностью пользователя на маркетплейсе, включая авторизацитю пользователей, размещенные товары, оценка настоящей суммы покупок, а также сотрудничество с правоохранительными органами для преследования мошенников.

0.3 Machine Learning Algorithms

Рассмотрим чуть подробнее какие алгоритмы машинного обучения могли бы быть применены в этой задачи. Сначала только скажим про то, как можно в гначале сделать разметку фродовых транзакций для обучения модели, в первую очередь для выявления нужно обращать внимания на жалобы и обращения клиентов. Исторические случаи выявленного мошенничества, внутренние проверки и расследования.Эти алгоритмы помогают анализировать поведение пользователей и выявлять аномалии, характерные для мошеннической активности. Одним из пнаиболее простых примеров, может быть Логистическая регрессия используется для задачи бинарной классификации6 следовательно мы можем и использовать этот алгоритм для определение фрода, предварительно собрав данные по истории транзакций пользователя, Средней сумме транзакций, Количество транзакций в день/месяц, время проведения транзакций, локации проведения транзакций (IP-адреса, географические координаты), использование различных устройств (мобильный, настольный), Возраст аккаунта, также можно попробовать применить более сложные алгоритмы ,такие как Решающее дерево, которое спользует иерархию условий для принятия реше-

ний, что делает его понятным и легким для визуализаци, включив Категориальные данные о транзакциях (тип товара, способ оплаты), Историю изменений аккаунта (изменение адреса, контактных данных), Частоту и время входа в аккаунт, Метаданные об устройстве (операционная система, браузер), Поведение пользователя на сайте (время на сайте, страницы посещения), а также , например, использовать градиентный бустинг, который улучшает производительность за счет последовательного обучения слабых моделей и их комбинации, также прибавив Исторические данные о выявленных мошеннических аккаунтах, а также использрование кластеризации для поиска аномалии в данных либо постави задачу, как поиск новизны. Можно также использровать нейронные сети, которые могут выявлять сложные нелинейные зависимости в данных, что делает их мощным инструментом для детекции фрода, также , например, использовав в них полный набор всех доступных данных, Временные ряды данных о транзакциях, данные о поведении пользователей в реальном времени (для моделей с временными рядами, таких как LSTM или GRU), а также можно рассматривать задачу как задачу детекции аномалии, используя такие алгоритмы как Isolation Forest и One-Class SVM. После того, когда злоумышленные действия выявлены, необходимо немедленно применять меры противодействия, что может включать блокировку аккаунтов мошенников и отмену подозрительных заказов, чтобы предотвратить дальнейшие убытки и защитить интересы партнеров и клиентов. Помимо реактивных мер, важно предпринимать проактивные шаги для уменьшения рисков и повышения безопасности на платформе. Внедрение предоплаты, дополнительных проверок при подозрительных заказах и блокировка подозрительных платежей позволяет минимизировать потенциальные убытки и поддерживать доверие пользователей.

0.4 What product features we can implement

Для того, чтобы помочь клиентам избежать неприятных ситуаций с мошенничеством, можно внедрить в платформу следущие фичи 6 которые могут также включасть подтверждение личности,например, двухфакторная аутентификация или проверка по SMS, для того, чтобы убедиться, что каждый аккаунт принадлежит реальному человеку , впоследствие снижая вероятность мошенничества. Еще одна фича, которую можно было бы внедрить - это Мониторинг активности, то есть возможность отслеживать активность на своем аккаунте, например, уведомления о входах из новых устройств или необычных действиях,что позволяет быстро реагировать на подозрительные события. Еще одна фича - это внедрение безопасных методо оплаты или совершении транзакций только через маркетплейс как через посрелника, предоставление клиентам безопасных методов оплаты, таких как платежные системы с защитой покупателя или возможность оплаты через платежные системы. Еще одна довольна интересная фича , это внедрение на платформу в качесвае 6 например, онбординга пользователя обучающие материалы и советы о том,например, что может помочь распознавать мошеннические схемы и как действовать в случае подозрительных ситуаций, для того, чтобы избежать риски и принимать осознанные решения, а также внедрение системы обратной связи и жалоб.

0.5 Что можно сделать для усложнения жизни фродерам

Мошенники могут получать контактные данные покупателей через различные методы, включая взлом баз данных, фишинг и социальную инженерию, а также атаки на уязвимые точки в системе. Взлом баз данных может происходить через эксплуатацию слабых мест в защите информации или через использование украденных учетных данных сотрудников. Фишинговые атаки обычно включают в себя отправку фальшивых сообщений или звонков, представляющихся от имени маркетплейса или продавцов, чтобы выманить личные данные у пользователей. Атаки на уязвимые точки могут быть направлены на слабые места в программном обеспечении или на ошибки веб-приложений, которые могут позволить злоумышленникам получить доступ к базам данных.
Для усложнения жизни мошенников можно принять ряд мер, например, одним из важных шагов может является улучшение безопасности данных путем внедрения современных методов шифрования и двухфакторной аутентификации. Кроме того, системы мониторинга и обнаружения аномальной активности помогут выявлять подозрительное поведение и автоматически блокировать. Обучение сотрудников и клиентов о методах защиты личной информации и распознавании мошеннических схем также играет важную роль в борьбе с мошенничеством. Регулярное обновление систем и процедур, а также усиление правил обработки данных, помогут минимизировать риски утечки информации и повысить уровень безопасности платформы.

0.6 Общий пайплан

Общий пайплан А/Б теста выглядит следующим образом:
1. Формулируем гипотезу
2. Выбираем целевую, прокси и контр-метрики
3. Определяем сегмент АВ теста
4. Запускаем эксперимент
В данном контексте при сравнение двух алгоритмов машинного обучения, цель A/B теста — сравнить производительность двух различных алгоритмов или настроек одного алгоритма на одном и том же наборе данных.

0.7 Цель

Оценить эффективность нового алгоритма автоопределения фродеров на этапе регистрации продавцов на маркетплейсе

0.8 Гипотеза

Использование ML-модели для определения фродеров на этапе регистрации уменьшит количество мошеннических продавцов и повысит безопасность транзакций на платформе без увеличения времени проверки новых продавцов.

0.9 Метрики

Какие метрики можно использовать при автоопределение фрадеров на платформе? Сначала можно посмотреть на какой-то супер общий пайплайн дерева метрик и дальше на более низком уровне, специфически для нашего кейса: Во-первых, существуют два ти-

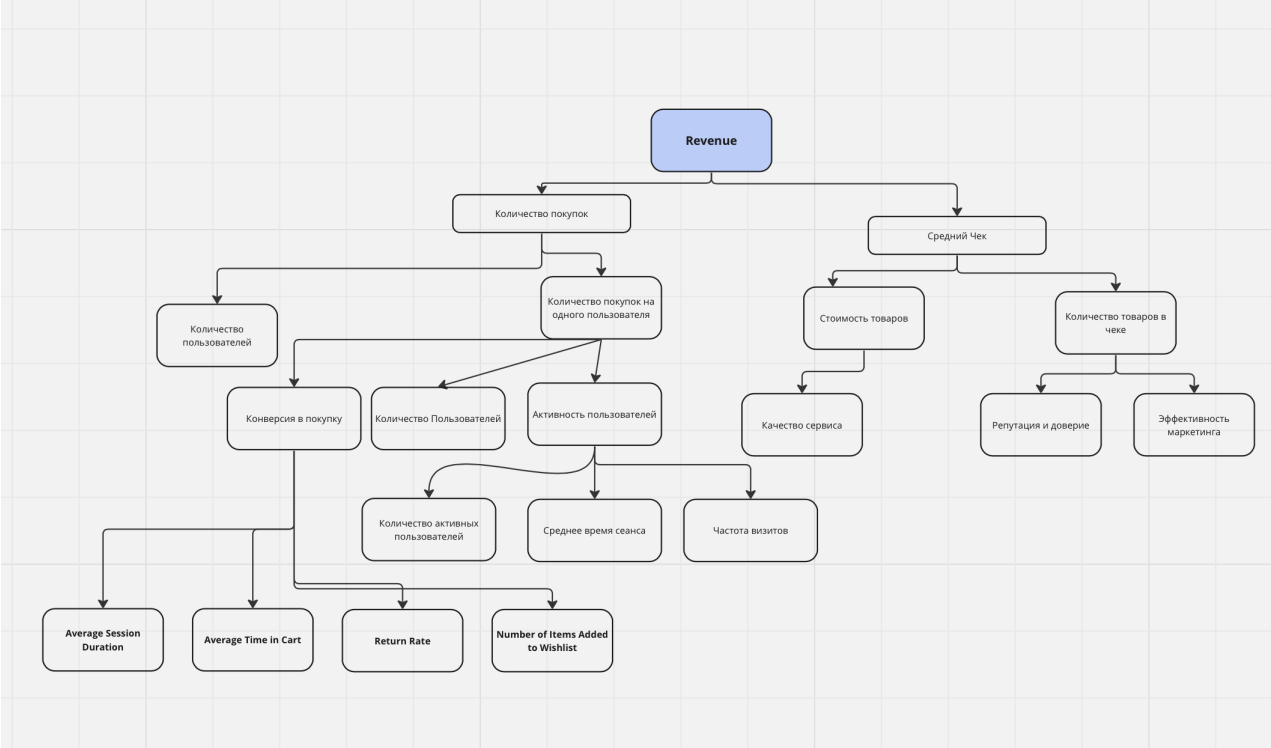

Рис. 1: After introduction of mixed pension system

па метрик - это оффлайн и онлайн метрики для антифрод системы. Оффлайн метрики используются для оценки качества модели на исторических данных или в оффлайн режиме, когда все данные уже доступны для анализа. В качесве них как нижнеуровневые метрики для аналитики мл продукта можно оценивать следующие параметры: Точность (Accuracy) - это одна из основных метрик, оценивающая долю правильно классифицированных транзакций (фродовых и нормальных) от общего числа транзакций. Однако, в случае с несбалансированными данными, где фродовых транзакций значительно меньше, точность может дать неправильное представление о качестве модели. Другие важные метрики включают точность отклика (Precision), полноту (Recall), F1-меру, ROC-кривую и AUC, а также матрицу ошибок (Confusion Matrix), которая помогает понять, в каких областях модель ошибается. Оценивая нижнеуровневые онлайн метрики, которые используются для оценки производительности модели в реальном времени, когда она применяется к новым данным в режиме онлайн. Доля ложных отклонений (False Positive Rate) и доля ложных пропусков (False Negative Rate) оценивают процент ошибок модели в классификации нормальных и фродовых транзакций соответственно. Время обработки и скорость обнаружения важны для оперативной реакции на потенциальные мошеннические операции, а процент отклоненных транзакций позволяет оценить эффективность принятых мер по борьбе с мошенничеством в реальном времени. Если смотреть на верзнеуровневые метрики, можно выделить следующие:

Процент мошеннических аккаунтов, выявленных на этапе регистрации (доля выявленных фродеров к общему числу зарегистрированных продавцов).
Процент успешных транзакций (доля успешных транзакций к общему числу транзакций)
Время проверки нового продавца (среднее время от момента подачи заявки до завершения проверки)
А также рассмотреть несколько вспомогательных метрик для анализа:
Общий уровень удовлетворенности пользователей (оценка удовлетворенности покупателей и продавцов)

Среднее количество невыявленных Количество мошенических транзаций
Conversion Rate продавцов: Доля зарегистрированных продавцов, которые активно начали торговать на платформе.
Доля жалоб покупателей от всех, сделанных покупок на маркетплейсе : Снижение количества жалоб на мошенничество.
В данном эксперименте в начале я предлагаю рассматривать две ключевые метрики - это Доля жалоб покупателей (Customer Complaint Rate). Доля жалоб покупателей (Customer Complaint Rate) вычисляется по следующей формуле:

$$Complaint_Rate = \left(\frac{Number_of_complaints}{Number_of Purchases_made}\right) \times 100\% \tag{1}$$

Например, если за месяц на маркетплейсе было сделано 100,000 покупок, и из них было получено 500 жалоб:

$$Complaint_Rate = \left(\frac{500}{100000}\right) \times 100\% = 0.5\% \tag{2}$$

Для анализа влияния новой модели автоопределения мошенников на долю жалоб в A/B-тесте, необходимо рассчитать метрику отдельно для контрольной и экспериментальной групп. Для контрольной группы (A):

$$Complaint_Rate_A = \left(\frac{Complaints_A}{Total_Purchases_A}\right) \times 100\% \tag{3}$$

Для экспериментальной группы (B):

$$Complaint_Rate_B = \left(\frac{Complaints_B}{Total_Purchases_B}\right) \times 100\% \tag{4}$$

Предположим, что для контрольной группы было сделано 50,000 покупок и получено 300 жалоб, а для экспериментальной группы было сделано 50,000 покупок и получено 200 жалоб: Для контрольной группы (A):

$$Complaint_Rate_A = \left(\frac{300}{50000}\right) \times 100\% = 0.6\% \tag{5}$$

Для экспериментальной группы (B):

$$Complaint_Rate_B = \left(\frac{200}{50000}\right) \times 100\% = 0.4\% \tag{6}$$

Таким образом, новая модель помогает снизить долю жалоб с 0.6% до 0.4%.

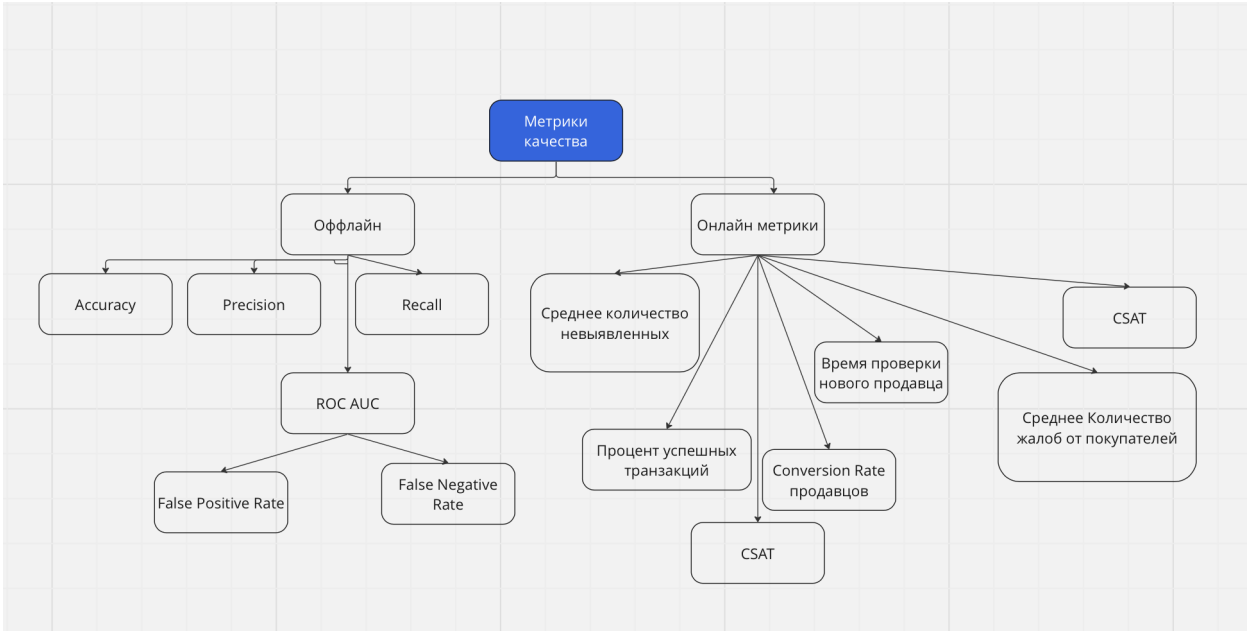

Рис. 2: After introduction of mixed pension system

0.10 Разделение на группы

1. Группы:
Контрольная группа (Group A):
Продавцы, зарегистрированные без использования нового алгоритма, используя текущий метод проверки.
Экспериментальная группа (Group B):
Продавцы, зарегистрированные с использованием нового алгоритма автоопределения фродеров.
При разделении на группы можно применить для выявления потенциальных мошеннических операций и проверки алгоритма мл несколько видов сплитов (смотри на схему). Один из таких методов - **сплит по историческим данным**, позволяющий анализировать изменения в паттернах мошеннической активности с течением времени. Другой метод - **сплит по географическим параметрам**, разделяет данные на группы в зависимости от географических характеристик для выявления мошенничества в конкретных регионах, а также, например, разделение по типам транзакций, чтобы выявлять аномалии в сравнении с типичными операциями, а также сплит по поведенческим признакам, который помогает обнаруживать аномальное поведение пользователей, также можно выделить сплит по результатам моделирования, который позволяет сосредоточить внимание на наиболее подозрительных группах пользователей или транзакций. Можно также в начале определить propensity scores, прежде всего, определяются характеристики или признаки, которые могут влиять на вероятность мошенничества, такие как поведенческие паттерны, географические данные или история транзакций. Затем строится модель, которая на основе этих признаков прогнозирует вероятность мошенничества для каждой транзакции или пользователя и дальше делаем метч, то есть полученные оценки используются для разделения данных на группы с различными уровнями риска мошенничества. Например, транзакции с высокими оценками могут быть отнесены к группе с высоким риском, что позволяет

проводить дополнительные проверки или анализировать их более внимательно.

0.11 Размер выборки

2. Размер выборки:
Определить минимальный размер выборки для каждой группы, основываясь на требуемой статистической мощности и минимально значимом различии в основных метриках. Можно использовать A/B тест калькуляторы для определения точного размера выборки, но мы посмотрим в Питоне.
3. Рандомизация:
Продавцы, подающие заявки на регистрацию, случайным образом распределяются между контрольной и экспериментальной группами. Рандомизация должна быть независимой и случайной, чтобы минимизировать смещения и увеличить мощность теста плюс все предпоссылки А/Б могут ломаться
4. Период тестирования:
Установить период тестирования, достаточный для сбора значимого объема данных (например, 4-6 недель), рассчеты в ноутбкуке.

0.12 Процедура тестирования

1. Этап регистрации:
При подаче заявки на регистрацию продавца, заявки рандомно распределяются между двумя группами.
Для группы А применяется текущий метод проверки.
Для группы В применяется новый алгоритм автоопределения фродеров.
2. Мониторинг и сбор данных:
Постоянно мониторить и собирать данные по основным и вспомогательным метрикам для обеих групп, также строить дешборыды дл отслеживония осноных показателей.
Регулярно проверять корректность и полноту собранных данных.
3. Анализ данных:
Сравнить основные метрики между двумя группами с использованием статистических методов (например, t-тест для средних значений, ^{2} тест для долей).
Оценить, есть ли статистически значимые различия в основных метриках между группами.
Анализировать вспомогательные метрики для более глубокого понимания эффектов алгоритма.
4. Интерпретация и выводы:
Если новая ML-модель показывает значительное снижение процента мошеннических аккаунтов при минимально увеличенном времени проверки и низких уровнях ложных срабатываний, алгоритм можно считать эффективным.
Оценить влияние на общую удовлетворенность пользователей, чтобы удостовериться, что алгоритм не создает негативных последствий для честных продавцов и покупателей.

Методы Проверки и Критерии

Ratio Test

Формула для ratio test:

$$\hat{R} = \frac{\hat{p}_B}{\hat{p}_A} \tag{7}$$

Проверка гипотезы $H_0 : R = 1$:

$$Z = \frac{\log(\hat{R})}{\sqrt{\frac{1}{n_A \hat{p}_A (1-\hat{p}_A)} + \frac{1}{n_B \hat{p}_B (1-\hat{p}_B)}}} \tag{8}$$

Delta Method

Вариация логарифма отношения:

$$Var(\log(\hat{R})) \approx \frac{1}{n_A \hat{p}_A (1-\hat{p}_A)} + \frac{1}{n_B \hat{p}_B (1-\hat{p}_B)} \tag{9}$$

Стандартное отклонение:

$$\sigma_{\log(\hat{R})} = \sqrt{Var(\log(\hat{R}))} \tag{10}$$

Bootstrap

Метод бутстреппинга:

- Сгенерировать множество бутстреппинг-выборок из данных контрольной и экспериментальной групп.
- Рассчитать долю жалоб для каждой бутстреппинг-выборки.
- Построить эмпирическое распределение разности долей жалоб и доверительные интервалы.

Коррелируемость Верха и Низа - Распределение Коши

Анализ коррелируемости двух зависимых величин:

$$(X_i, Y_i) \tag{11}$$

где X_i и Y_i — доли жалоб для контрольной и экспериментальной групп. Распределение Коши используется для анализа таких зависимостей.

0.13 Внедрение

1. Пилотное внедрение:
На основе результатов A/B-теста провести пилотное внедрение алгоритма на ограниченном сегменте рынка для дополнительной валидации.

2. Полное внедрение:
При успешных результатах пилотного внедрения, алгоритм автоопределения фродеров может быть полностью интегрирован в процесс регистрации на маркетплейсе.Можно также параллельно при раскатке запустить ухудшающий эксперимент, и смотреть как целевая метрика меняется 3. Непрерывный мониторинг:
Постоянно отслеживать эффективность алгоритма после полного внедрения и при необходимости вносить улучшения и корректировки.