

# A/B-контекст от samokat.tech

## Оглавление

1 Описание задачи .....	1
2 Ответы на продуктовые вопросы .....	2
2.1 Как определить, какой продавец мошенник, а какой — нет? Какие ещё могут быть схемы мошенничества?.....	2
2.1.1 Как определить, какой продавец мошенник, а какой — нет? .....	2
2.1.2 Какие ещё могут быть схемы мошенничества? .....	3
2.2 Какие продуктовые фишки могут помочь нашим клиентам избежать неприятных ситуаций с мошенничеством? .....	3
2.3 Через какую механику мошенник узнает контакты покупателя? Что можем сделать, чтобы усложнить жизнь фродерам? .....	4
3 Дизайн A/B-теста .....	4
3.1 EDA - exploratory data analysis .....	5
3.1.1 registration_date .....	5
3.1.2 activation_date .....	6
3.1.3 merchant_id, type, ind_frod.....	6
3.2 Определение гипотезы и метрик .....	7
3.3 Разделение на группы и проведение A/A-теста .....	7
3.4 Определение MDE и сроков проведения эксперимента .....	9
3.5 Результаты дизайна A/B-теста.....	10

## 1 Описание задачи

Распространённая и самая известная на рынке схема мошенничества: мошенник берёт данные ИНН из открытых источников, регистрирует на него компанию на площадке маркетплейса. Далее выставляет ходовой товар на продажу, например iPhone 14, на 50% дешевле рынка. Когда покупатель оформляет заказ на маркетплейсе, продавец пишет покупателю, что у него есть другой сайт, где можно купить товар ещё дешевле. Покупатель соглашается и оплачивает заказ по ссылке продавца в обход площадки. После этого продавец пропадает. Покупатель остается без товара. Такие продавцы очень сильно портят репутацию маркетплейса, и необходимо их выявлять как можно быстрее, лучше всего в процессе регистрации.

Для этого ML-команда предложила к использованию модель по автоопределению продавцов-мошенников, которую нам предлагается внедрить в процесс регистрации продавца.

Мы хотим ввести новый функционал, но как data-driven компания не можем это сделать без тестирования и проверки качества новой функциональности через A/B-тест.

## **2 Ответы на продуктовые вопросы**

### **2.1 Как определить, какой продавец мошенник, а какой — нет? Какие ещё могут быть схемы мошенничества?**

#### **2.1.1 Как определить, какой продавец мошенник, а какой — нет?**

Выделим основные признаки мошенника на маркетплейсе:

1. Использование чужого ИНН из открытых источников
2. Продажа товаров ниже рыночной стоимости
3. Продажа товаров популярных брендов средней и высокой стоимости
4. Наиболее вероятная схема работы: «Доставка от продавца (DBS)»

Первый этап выявления мошенника на основе данных, получаемых при регистрации продавца:

- проверка валидности ИНН
- проверка совпадения введенных ФИО с ФИО, принадлежащих введенному ИНН
- проверка номера телефона (различные сервисы с отзывами, наименованием контакта)
- если продавец ИП сравнить выбранную категорию товаров с открытыми ОКВЭД

Однако, последние два способа не могут гарантировать является ли продавец мошенником.

Второй этап после предоставления копий документов, проверка:

- подлинности документов, насколько это может быть возможно
- совпадения персональных данных в предоставленных документах (паспорт, ИНН)

Третий этап после создания карточки товара, дополнительно проверять продавцов с популярными товарами.

Исходя из полученного опыта работы с мошенниками, выявить основной пласт товаров, которые используются для завлечения покупателей.

Для этих товаров рассчитать среднерыночную стоимость (внутри маркетплейса, по возможности, использовать цены с других источников), как вариант, если товар новый его

стоимость определена производителем и достаточно стабильна в начале продаж, можно отталкиваться от нее.

Далее, если стоимость товара отличается от рыночной более чем на 30-50% производить дополнительную проверку продавца.

Если у нас есть распределение цен товара, проводить дополнительную проверку при цене отличающейся от средней на  $2\sigma$ .

Дополнительная проверка может заключаться в видеозвонке с продавцом для дополнительной проверки его документов (так делают некоторые букмекерские конторы для вывода денежных средств с баланса), проверке по видеосвязи наличия товара. Однако, такой способ требует человеческого ресурса.

Наиболее вероятная схема работы мошенников «Доставка от продавца (DBS)», т.к. такой способ не предполагает передачу товара на склад маркетплейса и при заказе товара продавец получает персональные данные покупателя. Возможно, стоит ввести усиленную проверку для данной схемы работы продавцов.

### **2.1.2 Какие ещё могут быть схемы мошенничества?**

- Отправка другого товара, часто схожего по весу и упакованного в фирменную упаковку
- Отправка брака или реплики
- Возможно схожие варианты с описанной в задаче, отправка письма:
  - об оплате дополнительной доставки (более быстрой)
  - о возврате части средств (необходимо ввести данные карты)
  - получение подарка или скидки (необходимо ввести данные карты)
  - о входе в аккаунт и необходимости сменить пароль по ссылке

### **2.2 Какие продуктовые фишки могут помочь нашим клиентам избежать неприятных ситуаций с мошенничеством?**

Предупреждение пользователя (всплывающее окно) при стоимости товара ниже рыночной цены. Можно не проводить дополнительную проверку демпингующих продавцов, а попробовать предупреждать покупателей о том, что товар продается по заниженной цене.

Предупреждение пользователя (всплывающее окно), если продавец новый и у него не было успешных продаж. Однако, могут возникнуть сложности с стартом продаж на площадке у нормальных продавцов, так как у них будут бояться заказывать.

Дополнительно предупредить их о возможных рисках. В случае, если продавец им пишет на почту необходимо сообщить в службу поддержки и ничего не оплачивать на сторонних сервисах.


Таким образом, можно снизить количество жертв мошенников, однако, полностью это проблему не решит.

### 2.3 Через какую механику мошенник узнает контакты покупателя? Что можем сделать, чтобы усложнить жизнь фродерам?

При работе по схеме «Доставка от продавца (DBS)» продавец получает следующую информацию о покупателе (рисунок 2.1).

В данном случае продавец получает e-mail покупателя. Как вариант, организовать подменный e-mail по аналогии с телефоном и проверять отправляемые письма от продавца на наличие ссылок, ключевых слов и т.п.

#### Подробная информация о заказе

В правой части информационного поля заказа находится кнопка . Нажмите на неё, чтобы увидеть подробные данные о заказе:

- **Адрес доставки** — данные, которые должны быть указаны в логистических и бухгалтерских документах.
- **Интервал выдачи заказов.** В этом поле указан интервал выдачи заказа покупателю согласно плану-графику.
- **Покупатель** — ФИО покупателя.
- **Номер доставки** — номер, который использует система Мегамаркета для всего заказа покупателя. Он может потребоваться, если вам будут поступать обращения от покупателей.
- **Заказ оформлен** — дата и время оформления заказа на Мегамаркете.
- **Сумма заказа** — цена товара с учётом всех скидок, которую должен заплатить или уже заплатил за товар покупатель.
- **Email** — email покупателя.
- **Статус оплаты** — информация об оплате заказа.
- **Телефон** — контактный телефон покупателя.



**В личном кабинете мы показываем временный номер покупателя**

Это нужно, чтобы реальные номера покупателей не использовались для спам-рассылок и массовых звонков. Этот номер предназначен исключительно для звонков и SMS-сообщений, связанных с доставкой заказа.

Рисунок 2.1 – Информация о заказе при работе по схеме «Доставка от продавца (DBS)»

### 3 Дизайн А/В-теста

Задача: проверить эффективность модели по автоопределению продавцов-мошенников.

### 3.1 EDA - exploratory data analysis

Перед тем как переходить непосредственно к дизайну A/B-теста рассмотрим предоставленные данные:

- registration\_date: дата регистрации (дата получения мерчант айди, продавец зарегистрировался на площадке)
- activation\_date: дата активации (продавец прошел все этапы регистрации и может продавать)
- merchant\_id: айди продавца
- type: форма организации бизнеса
- ind\_frod: индекс мошенника

#### 3.1.1 registration\_date

Пропущенных дат в данной колонке нет, однако, присутствует 175 значений равных 01-01-1970. Данное значение соответствует нулю, можно предположить, что при регистрации произошел сбой и дата не записалась. Рассмотрим количество регистраций в месяц (за исключением 1970 года), рисунок 3.1.

По данным видно, что выборка продавцов за 2023 год. Выделяются два месяца с малым количеством регистраций – февраль в большей степени и сентябрь в меньшей. Можно предположить, что именно в эти месяцы были регистрации с датой 01-01-1970. Если у данных пользователей есть дата активации между февралем и сентябрем проставим им февральскую дату регистрации, после августа – сентябрем. Выглядит немного лучше, рисунок 3.1.

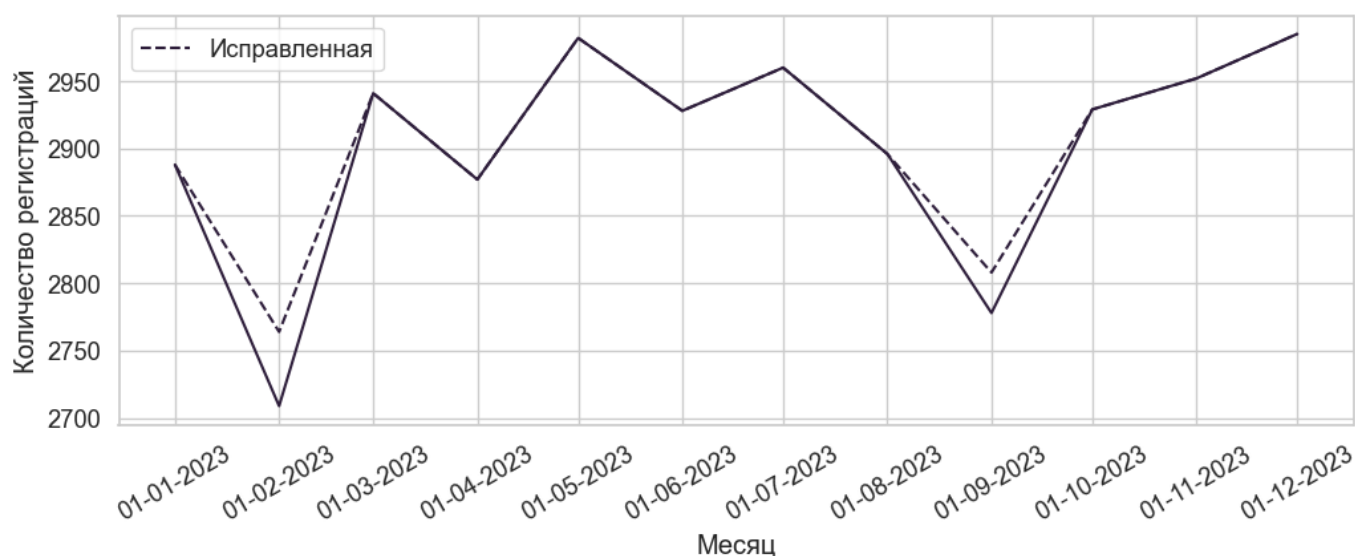


Рисунок 3.1 - Количество регистраций продавцов по месяцам

Пользователей с `registration_date = 01-01-1970` и без даты активации уберем из выборки (90 шт). Если бы записей было много, стоило бы задуматься над тем, чтобы их оставить.

### 3.1.2 activation\_date

Дат активаций 1970 годов нет. 14701 значение без даты (после удаления в прошлом шаге 90 строк, осталось 14611). Количество активаций продавцов по месяцам рисунок 3.2. Имеются 5 пользователей с активацией в 2022 году, исправим год на 2023, тогда хронология сохраняется (регистрация затем активация). После этого остается 14 пользователей с датой активации раньше даты регистрации. Уберем их из выборки. Пользователей с активацией в 2024 году оставим. Количество активаций продавцов по месяцам рисунок 3.2.

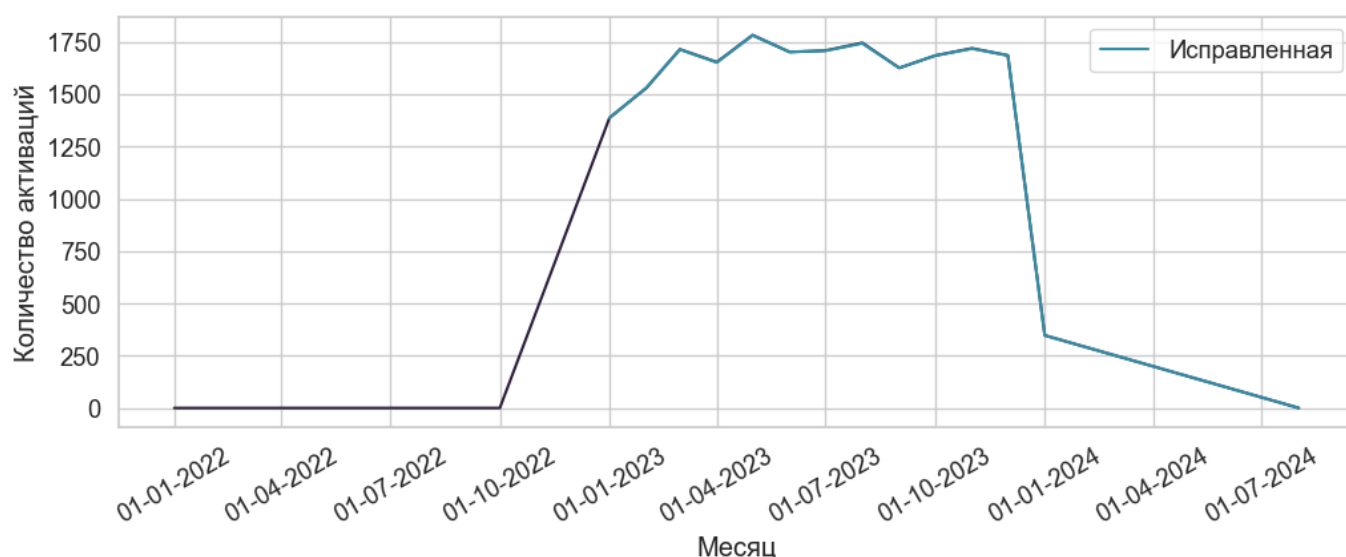


Рисунок 3.2 – Количество активаций продавцов по месяцам

### 3.1.3 merchant\_id, type, ind\_frod

После предыдущих шагов обработки осталось 34896 строк.

Все значения `merchant_id` уникальны, пропусков нет.

Колонка `type` включает только два показателя: IE и LLC, пропусков нет.

`ind_frod` включает 699 не определенных значений, 2819 мошенников, 31378 нормальных продавцов.

Так как 699 пользователей не прошли онбординг и мы не можем оценить их принадлежность к мошенникам (`ind_frod = NaN`), данных продавцов исключим из выборки.

Итоговый датафрейм содержит 34197 строк. Из них 8,24% продавцов определены, как мошенники.

### 3.2 Определение гипотезы и метрик

Гипотеза:

- модель по автоопределению продавцов-мошенников с большей вероятностью определяет мошенника.

Метрики:

- core-метрика: доля мошенников. Модель должна, как минимум, не хуже ручной проверки определять продавца, как мошенника. Однако, нас будет интересовать тот факт, что модель работает лучше.

- warning-метрика: количество регистраций. Возможно, модель будет определять добросовестных продавцов, как мошенников, соответственно, будет проблематично зарегистрироваться. Ожидаем, что количество регистраций не изменится.

- гроху-метрика: доля мошенников до активации продавца. Если общая доля мошенников не будет отличаться от контрольной группы, возможно, что модель лучше определяет мошенников именно на этапе регистрации, что является ее преимуществом, так как продавец не успевает выйти на площадку и испортить репутацию маркетплейса. Возможно, стоит отдельно рассмотреть ООО и ИП, так как они отличаются по доле обнаруженных мошенников.

Зафиксируем стандартные значения  $\alpha = 0,05$ ,  $\beta = 0,2$  и мощности  $= (1 - \beta) = 0,8$ .

### 3.3 Разделение на группы и проведение A/A-теста

Разделим пользователей на группы с помощью хеширования с солью и проведем A/A тест. Пусть 0 – контрольная группа (17186 продавцов: IE = 12524, LLC = 4662), 1 – тестовая (17011 продавцов: IE = 12120, LLC = 4891), выборки получились схожие по размеру. Доля мошенников в контрольной группе = 8,37%, в тестовой = 8,11%.

При запуске 10000 симуляций A/A-тестов, в которых на каждой итерации формируется подвыборка без повторения в 1500 продавцов из двух групп. Проводятся сравнения этих подвыборок t-testом. В итоге получаем процент p-values меньше либо равных 0.05: 4.83% с доверительным интервалом: (4.3%, 5.2%), где 5% входит в интервал, соответственно, система сплитования работает корректно. Распределение полученных p-values приведено на рисунке 3.3

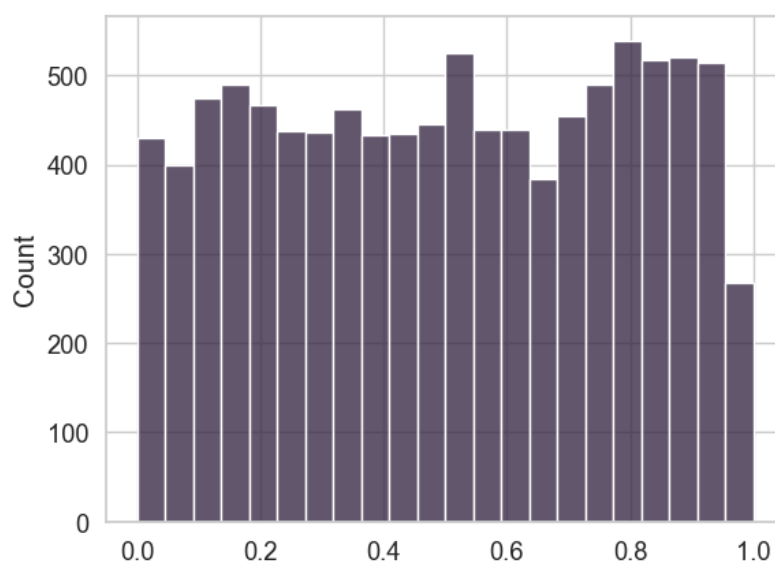


Рисунок 3.3 – Распределение p-values при симуляции A/A-теста после сплитования

Смоделируем биномиальное распределение для каждой группы (рисунок 3.4):

- размеры выборок равные количеству регистраций в июне для каждой группы

- вероятность посчитаем за год для каждой группы отдельно

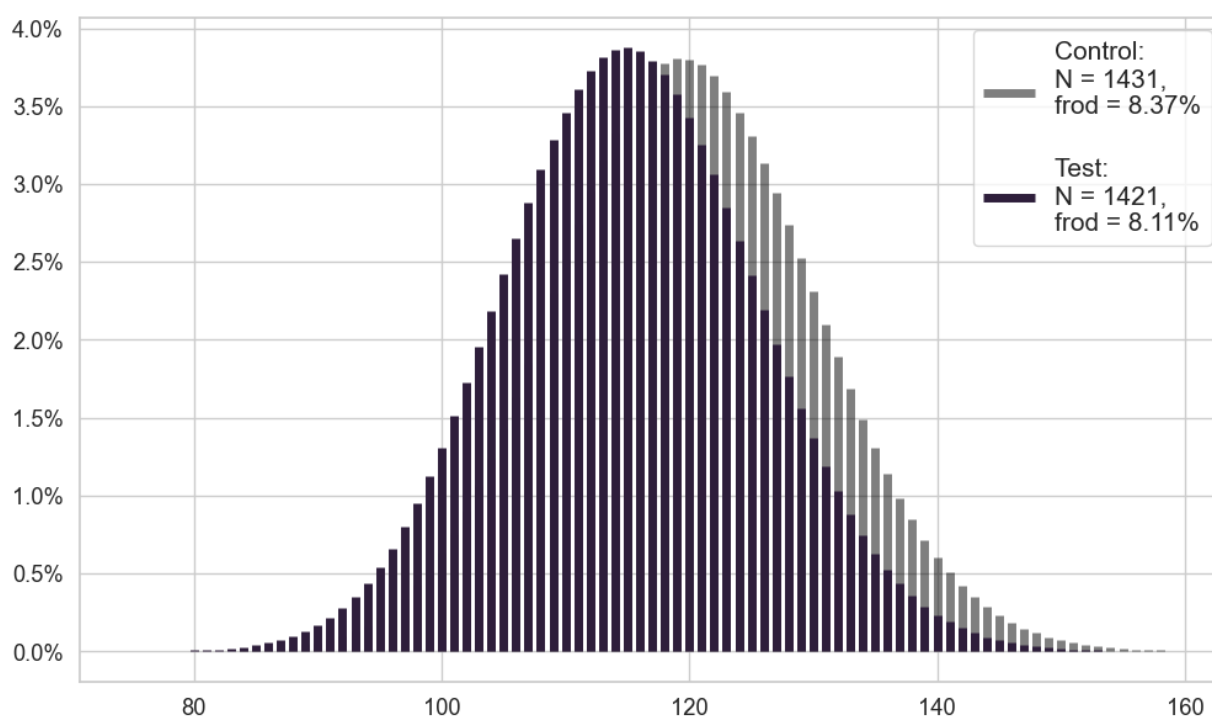


Рисунок 3.4 – Биномиальное распределение с характеристиками полученных групп

Различие между группами видно, однако, симуляция A/A-теста показывает, что  $\alpha = 0,05$  входит в доверительный интервал p-values.



### 3.4 Определение MDE и сроков проведения эксперимента

В качестве симуляции используем биномиальное распределение. Чтобы приблизить симуляцию к реальным данным с использованием системы сплитования зададим для контрольной и тестовой групп вероятность определения мошенника равную среднегодовой в группе: 8,37% для контрольной и 8,11% для тестовой.

Размер выборок будем симулировать от 100 до 2000 продавцов с шагов в 10. Так как после сплитования группы немного различаются, учтем это для тестовой группы. Для этого размер выборки для тестовой группы умножим на коэффициент, соответствующий разнице, и округлим вверх.

Размер эффекта симулируем как прибавление к 8,11% (годовая для тестовой) по 0,5%, до 14,11%.

Исходя из заданных условий получим распределение мощности от количества продавцов при различном MDE (рисунок 3.5)

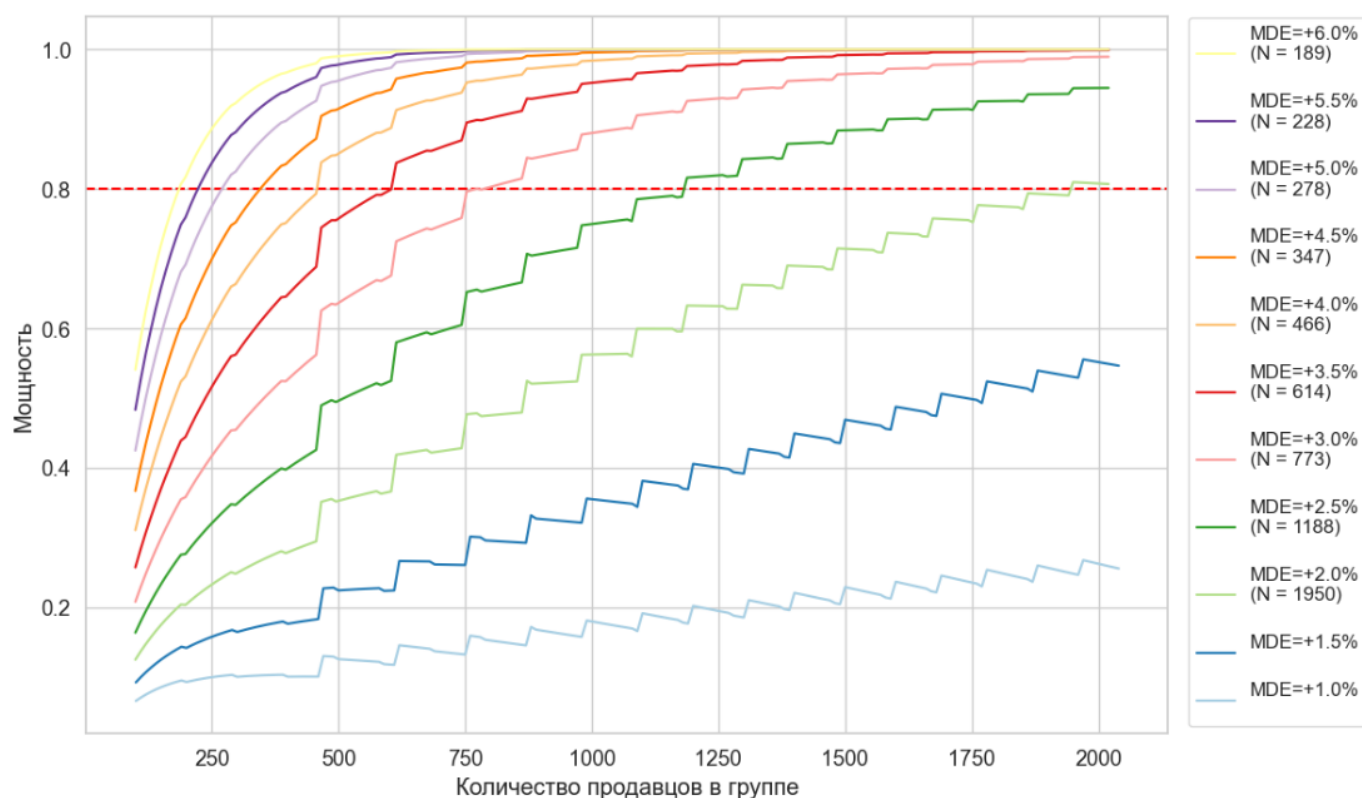


Рисунок 3.5 – Распределение мощности от количества продавцов при различном MDE

Для определения оптимального срока проведения эксперимента посчитаем сколько продавцов регистрируется на маркетплейсе в неделю. Распределение приведено на рисунке 3.6.

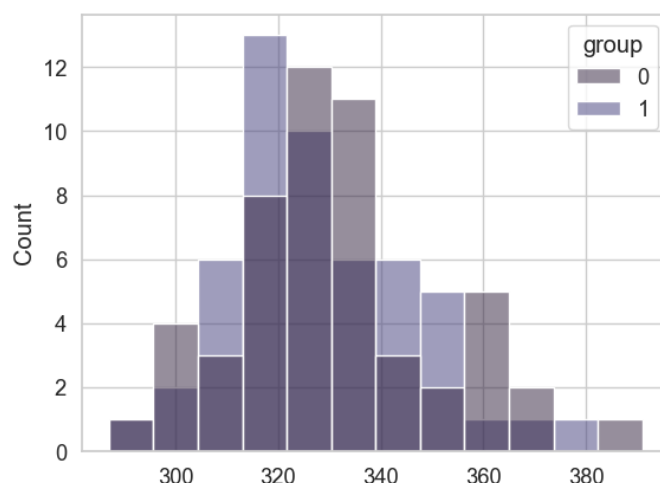


Рисунок 3.6 – Распределение регистраций в неделю

Среднее количество регистраций в неделю в контрольной группе =  $330,5 \pm 39,9$  ( $\pm 2\sigma$ ).

Среднее количество регистраций в неделю в тестовой группе =  $327,1 \pm 36,5$  ( $\pm 2\sigma$ ).

Рассмотрим худший случай, когда у нас будет минимальное количество регистраций каждую неделю: тестовая -  $2\sigma = 290$  регистраций.

Получим следующие длительности эксперимента при 290 регистраций в неделю:

- MDE = 10.1% (+2.0), длительность A/B-теста: 6.7 недель
- MDE = 10.6% (+2.5), длительность A/B-теста: 4.1 недель
- MDE = 11.1% (+3.0), длительность A/B-теста: 2.7 недель
- MDE = 11.6% (+3.5), длительность A/B-теста: 2.1 недель
- MDE = 12.1% (+4.0), длительность A/B-теста: 1.6 недель
- MDE = 12.6% (+4.5), длительность A/B-теста: 1.2 недель
- MDE = 13.1% (+5.0), длительность A/B-теста: 1.0 недель
- MDE = 13.6% (+5.5), длительность A/B-теста: 0.8 недель
- MDE = 14.1% (+6.0), длительность A/B-теста: 0.7 недель

### 3.5 Результаты дизайна A/B-теста

Оптимальные сроки проведения эксперимента: 2-4 недели. Такая длительность позволит зафиксировать долю мошенников в тестовой группе равную 10,6-11,6% (+2,5-3,5%), изначально доля равна 8,1%.

Для фиксации доли мошенников в тестовой группе в 10,1% (+2%) потребуется значительное увеличение сроков эксперимента – около 7 недель.

Jupyter Notebook с расчетами на github:

[https://github.com/4n1k1n/AB\\_contest\\_samokat](https://github.com/4n1k1n/AB_contest_samokat)