Teamestro

Хакатон 24/25

КОМАНДА 4

16 МАРТА

Структура данных

Количественные данные

- Время обработки запросов
- Время работы в группе
- Количество обработанных звонков и чатов
- Время, потраченное на чаты и звонки
- Рабочее время

Строчек: **170774**

Столбцов: **35**

Качественные данные

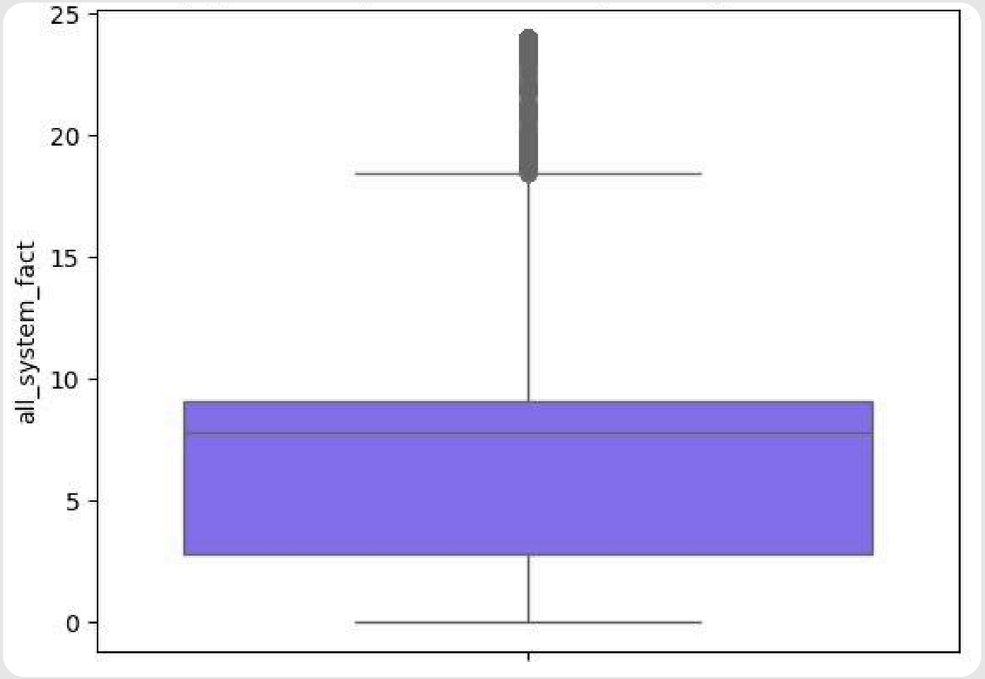
- Тип трудового договора
- Бизнес-направление
- Скилл-группа

Терминология

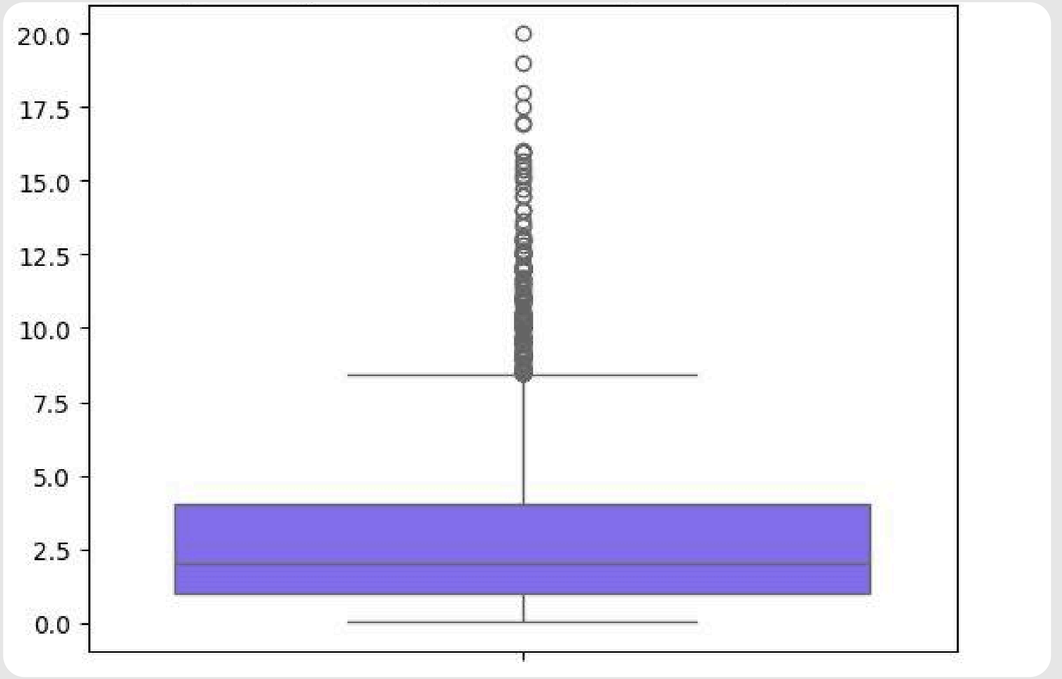
$$\text{Продуктивность} = \frac{\sum \text{количество запросов}}{\sum \text{время работы}}$$

$$\text{Нормированная продуктивность} = \frac{\text{Продуктивность человека}}{\text{Продуктивность скилл — группы}}$$

Предварительный анализ



Распределение фактического времени работы в часах



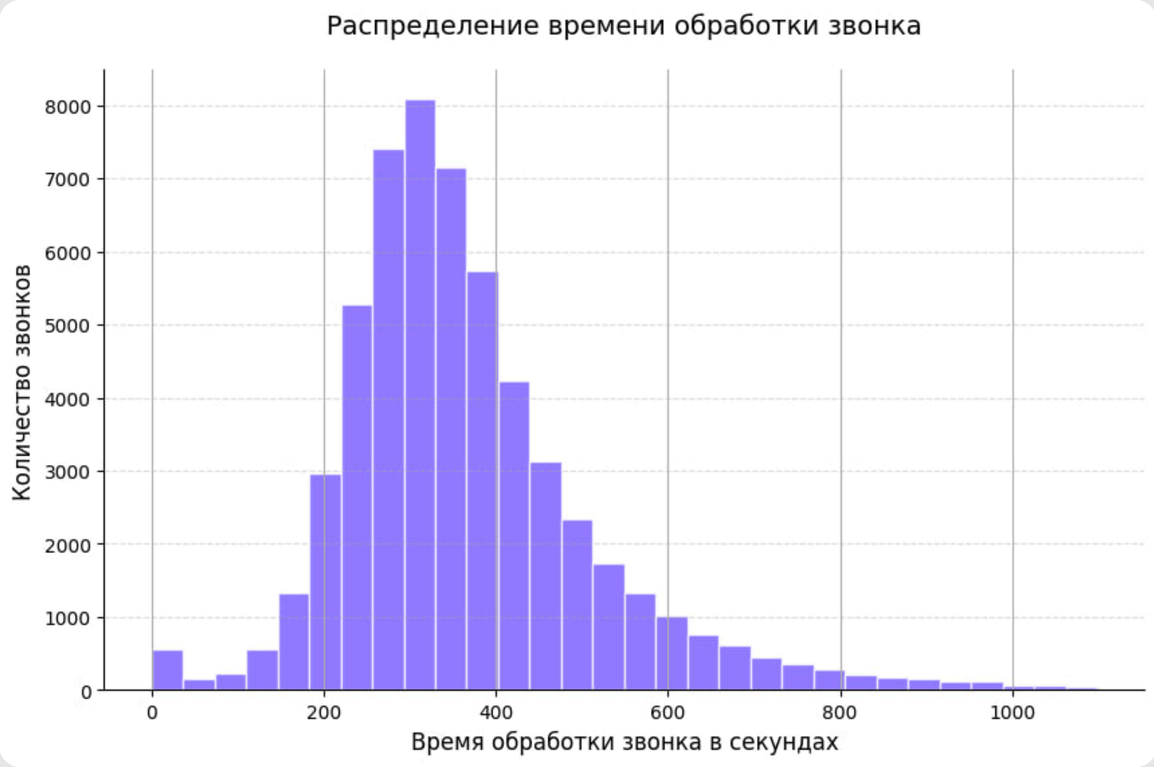
Разница all_smena_plan и суммы категории планирования в часах

0 сек
min

6.43ч
avg

23.95ч
max

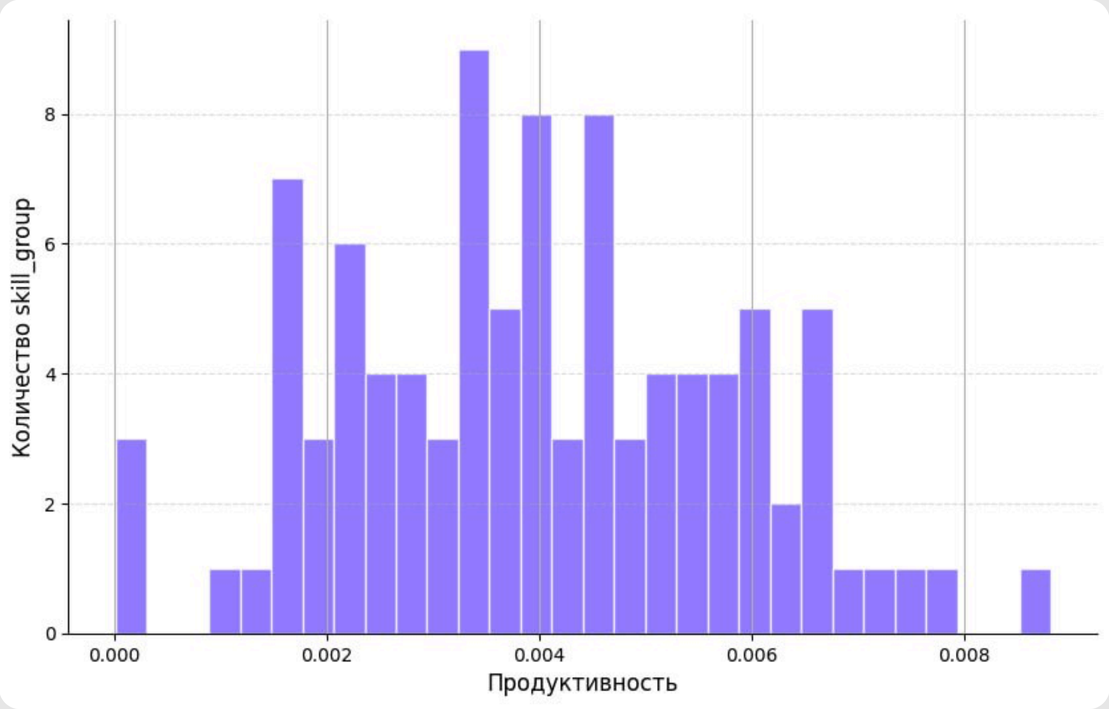
Предварительный анализ



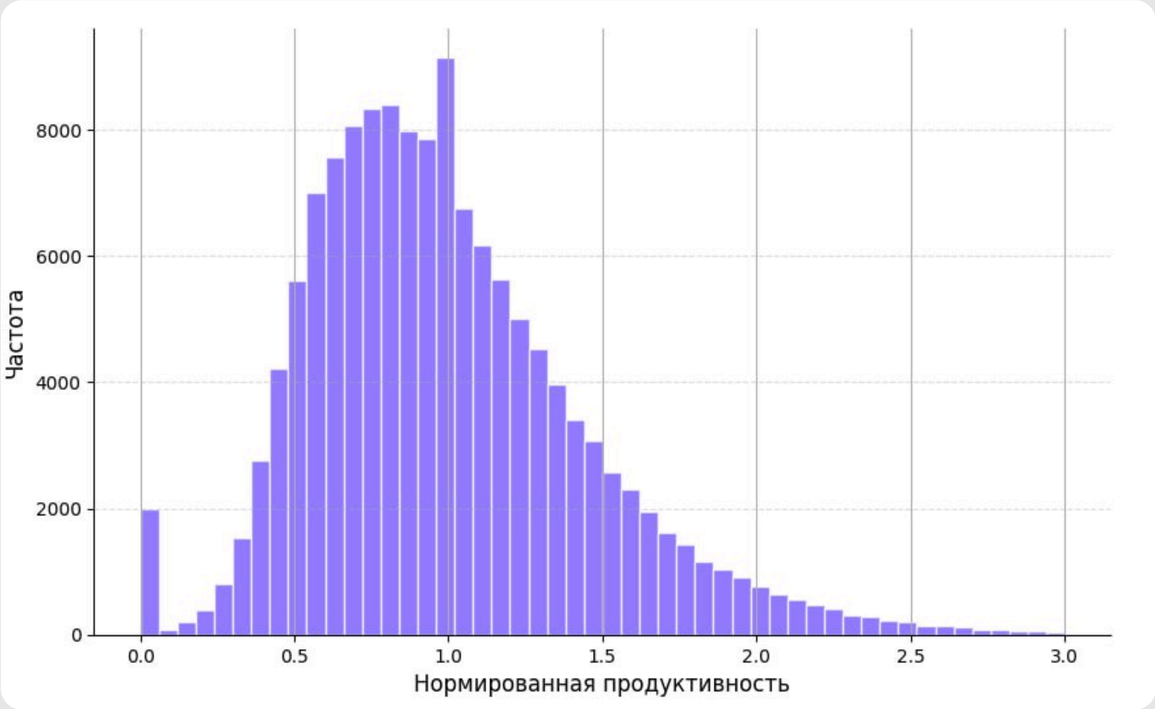
Медиана времени ответа на звонок 339 сек
Среднее время ответа на звонок 374 сек

Медиана времени ответа на запрос в чате 186 сек
Среднее время ответа на запрос в чате 229

Предварительный анализ



Распределение продуктивности по skill_group



Распределение нормированной продуктивности по работникам в определенной группе в определенный день

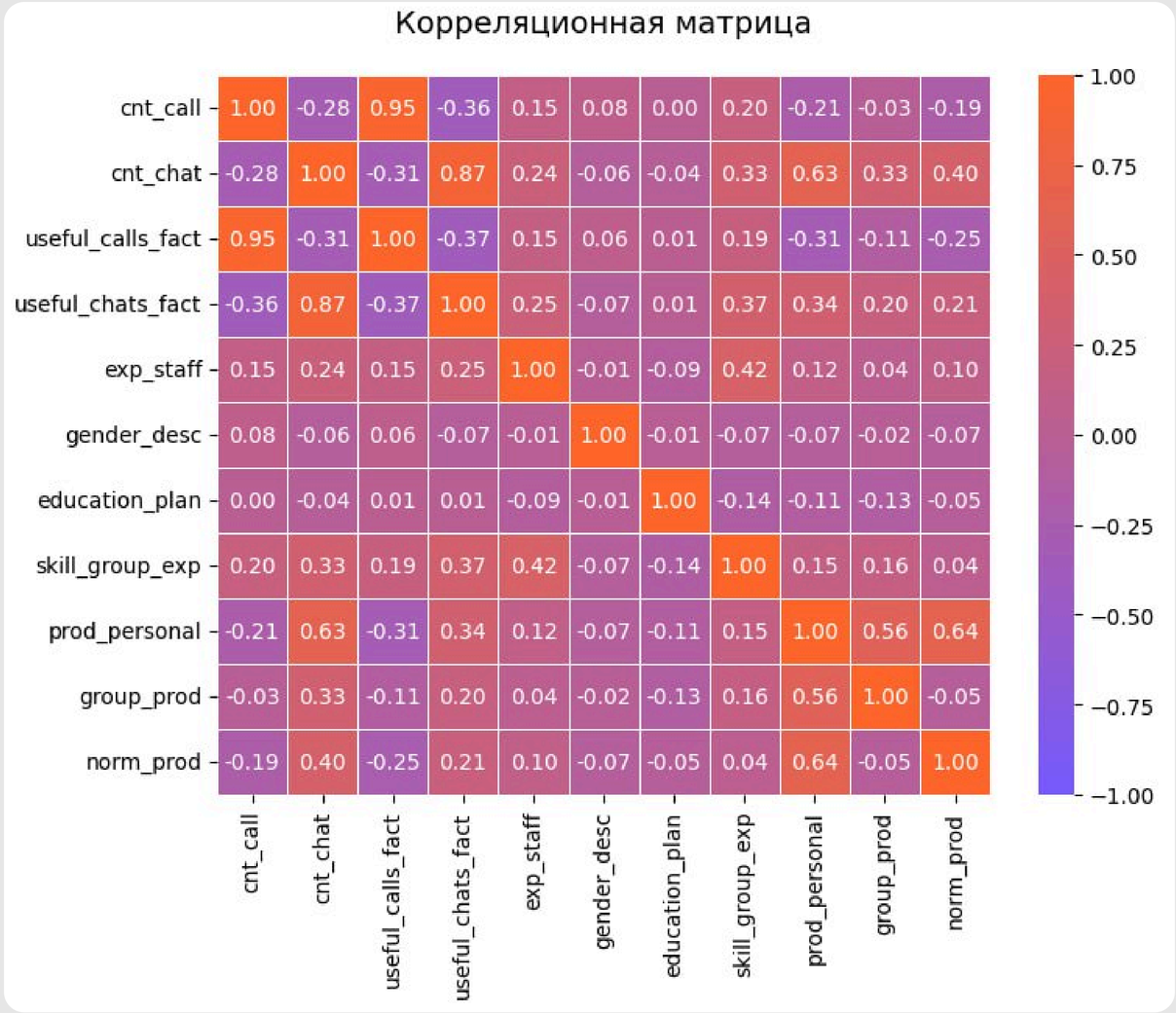
Предварительный анализ

Главные зависимости

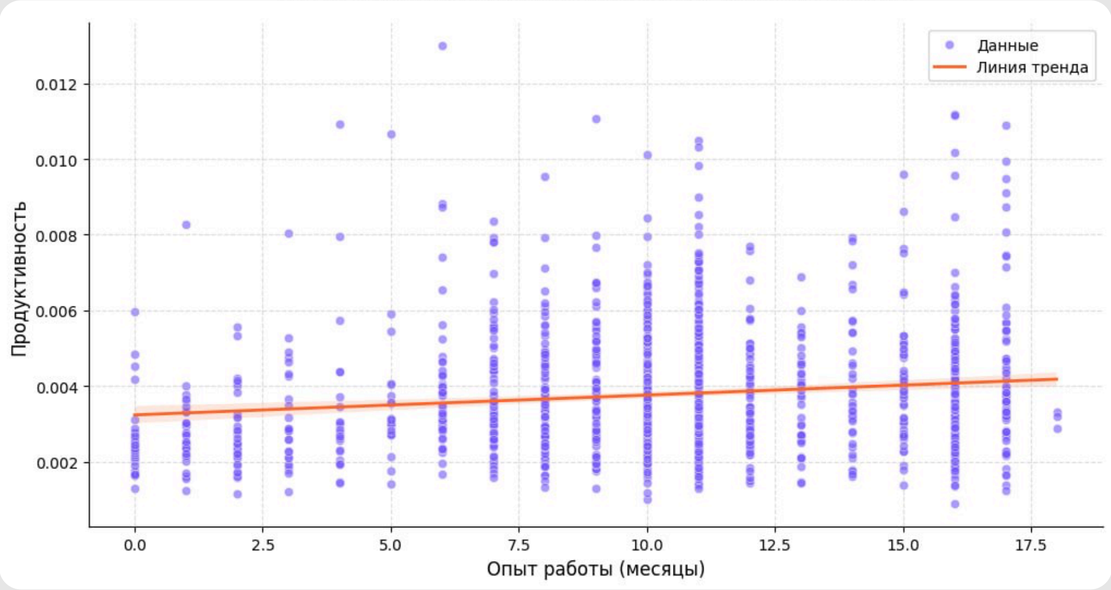
exp_staff, gender_desc, education_plan, skill_group_exp

- Это колонки, которые имеют влияние на продуктивность
- Из них самую большую коррелияцию с продуктивностью имеет колонка exp_staff. Но также положительное влияние имеет skill_group_exp

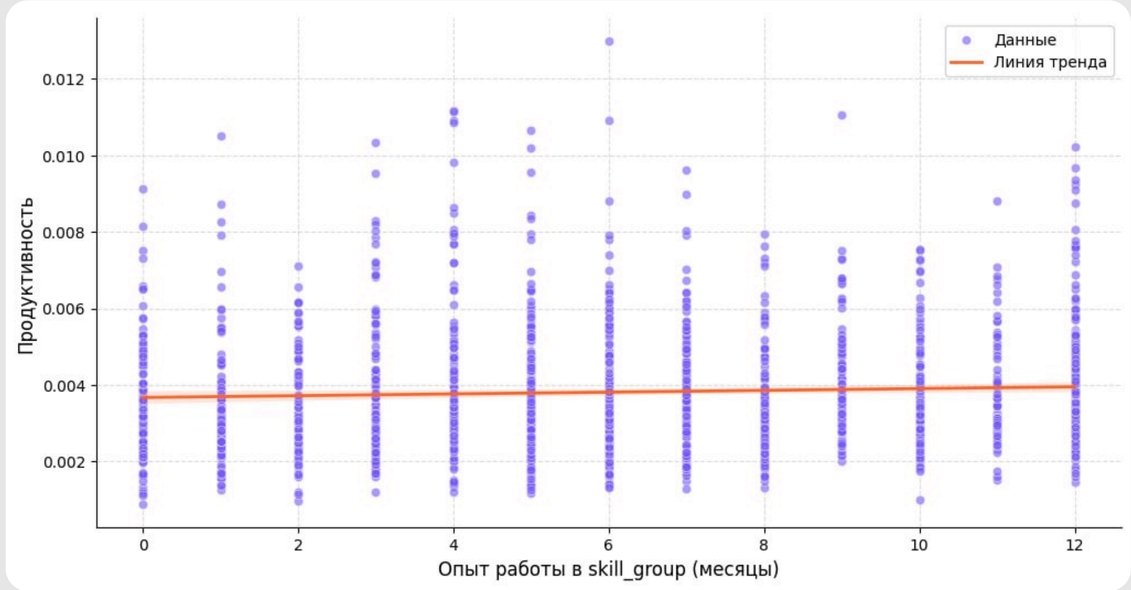
prod_personal - личная продуктивность
norm_prod - нормированная продуктивность



Предварительный анализ



Взаимосвязь между опытом работы и продуктивностью



Взаимосвязь между опытом работы в skill_group и продукивностью

Чистка данных от выбросов (1/2)

Действие 1

Из колонки 'emp_type' удаляем пропуски, т.к. в датасете есть люди, которые меняли свой тип трудоустройства, поэтому пропуски мы не можем однозначно заполнить.

Действие 2

Удаляем все кроме нижнего квартиля по различию (суммы всех категорий действий) и (общей работой) - все значения не больше 8 минут.

Действие 3

Делаем “Действие 2”, но для плана работы- все значения не больше 30 минут.

Действие 4

Удаляем строки в которых значение 'all_system_fact' или 'all_smena_plan' больше 13 часов, т.к. по закону нагружать сверхурочной работой больше 4 часов нельзя.

Чистка данных от выбросов (2/2)

Строчек: **148747**

Столбцов: **28**

Действие 5

В 'skill_group' заменяем пропуски на значения из 'planning_group_nm', если значение 'No', то удаляем, остальные пропуски удаляем

Действие 6

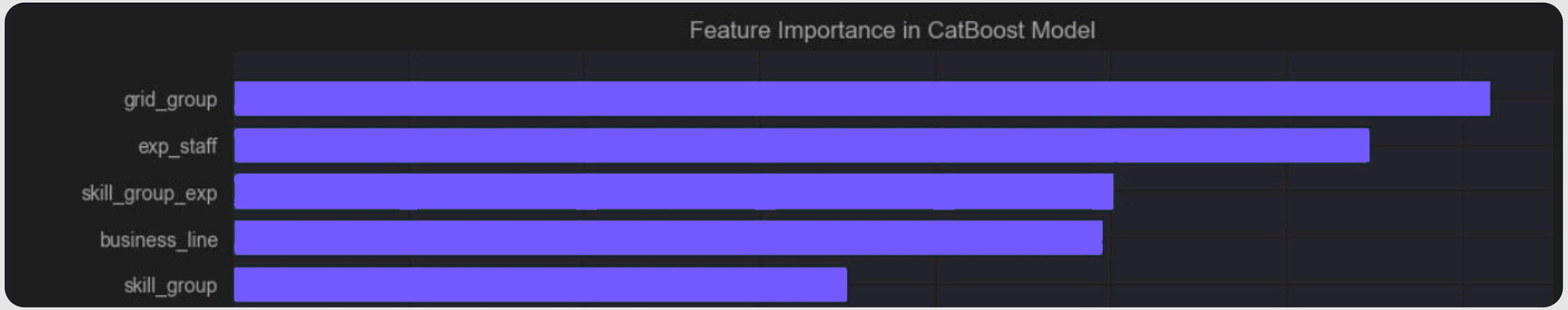
В колонках 'cnt_call' и 'cnt_chat' заменяем на нули, если 'useful_calls_fact' = 0 и 'useful_chats_fact' = 0 соотвественно. Оставшиеся пропуски удаляем

Действие 7

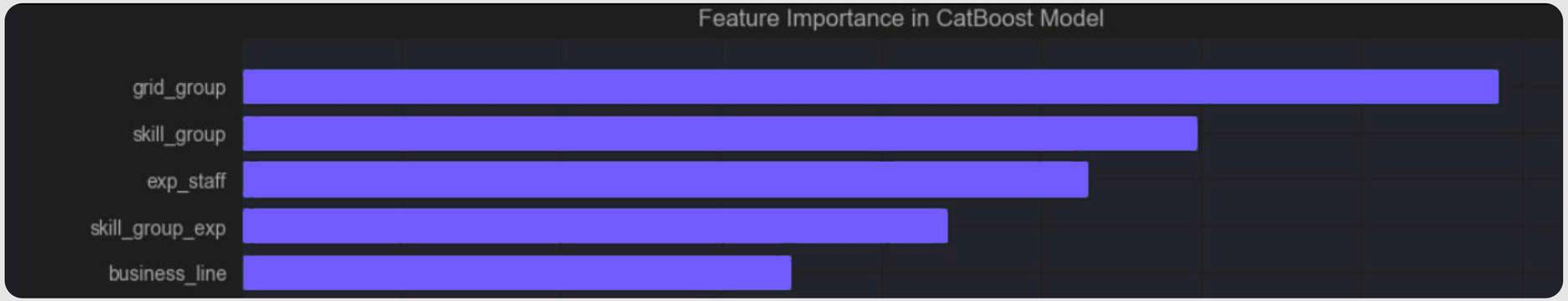
Оставшиеся колонки с пропусками ('planning_group_nm', 'residential_district_nm', 'residential_city_nm', 'residential_state_nm', 'residential_settlement_nm', 'dlg_time_call', 'dlg_time_chat') удаляем

Предварительный анализ

- На очищенных данных (chats и calls) обучили **CatBoost** на предсказание **продуктивности**. У модели с лучшей точностью получили **feature importance**
- **Exp_staff и skill_group_exp** находятся в топе **самых важных признаков** для модели

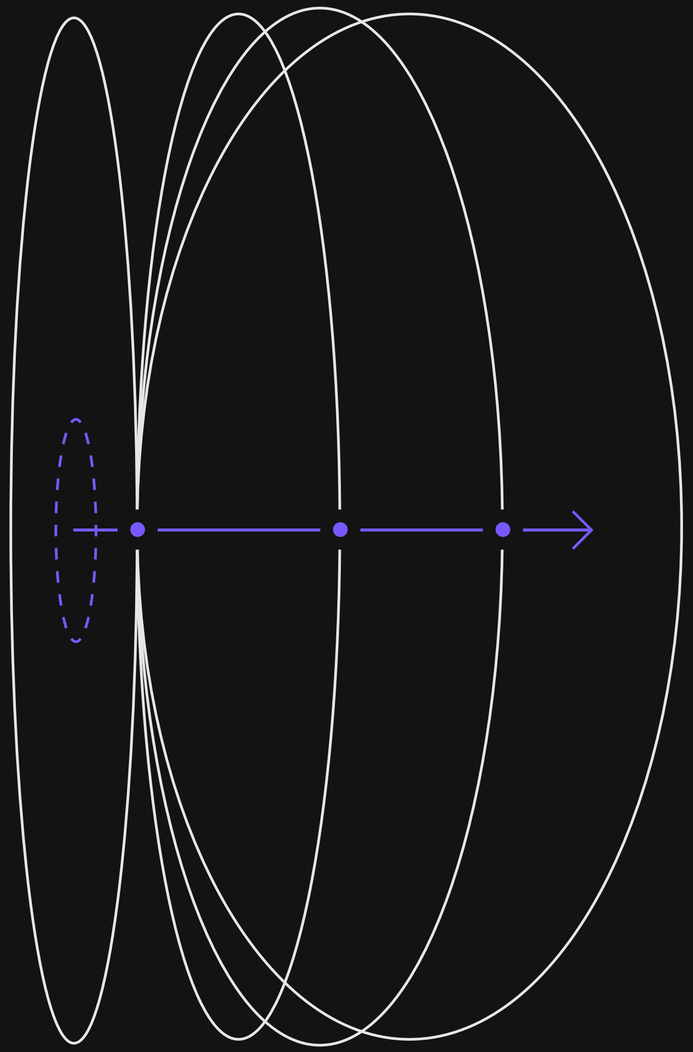


Чаты



Звонки

Выводы



- Нормируем продуктивность по skill-группам
- Разделяем call и chat
- В качестве исследуемого параметра выбираем skill_group_exp

Исследовательский вопрос и гипотеза

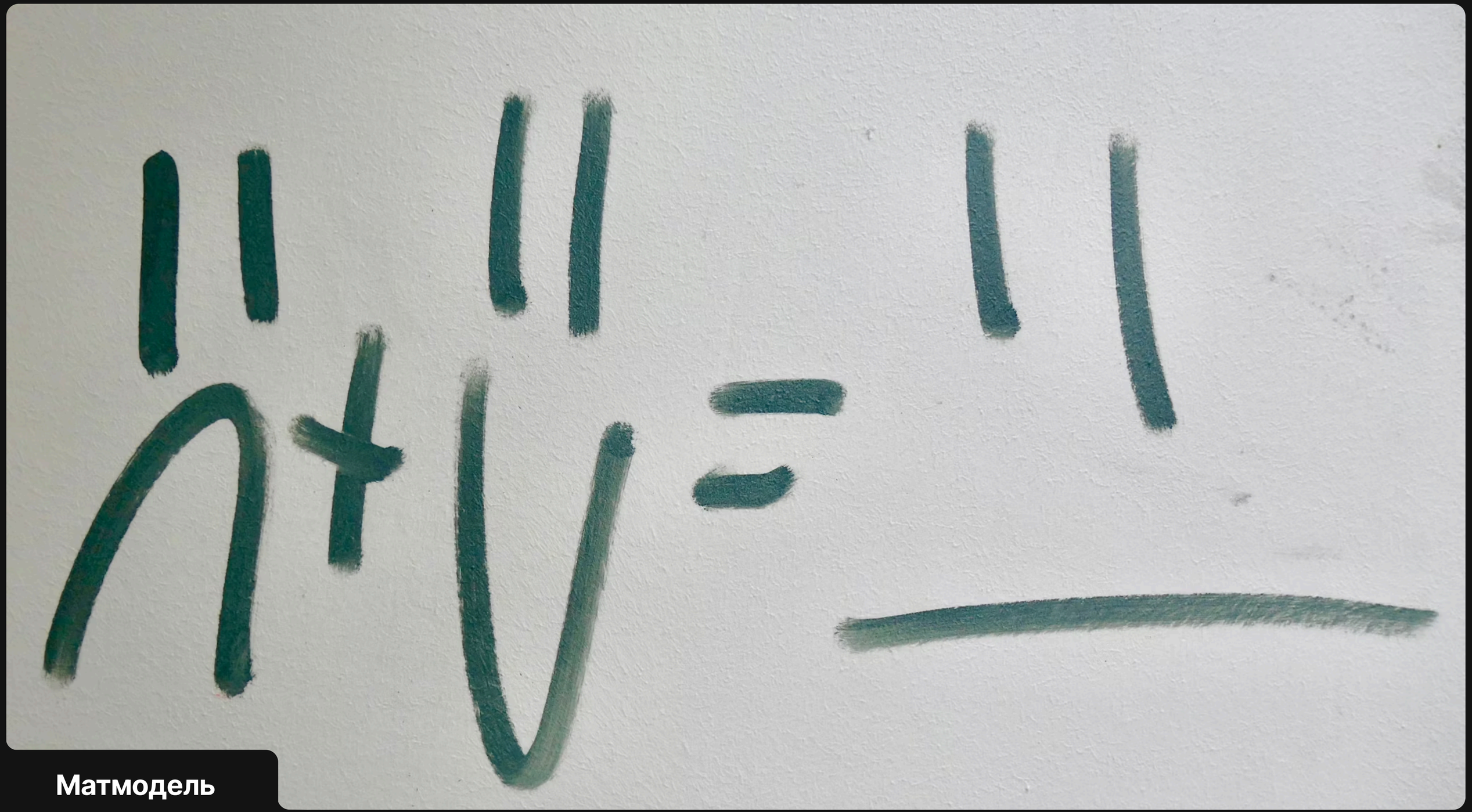
Исследовательский вопрос

- Какие показатели влияют на продуктивность сотрудников?

Гипотеза

- Продуктивность зависит от времени работы в группе, в которой сотрудник находиться

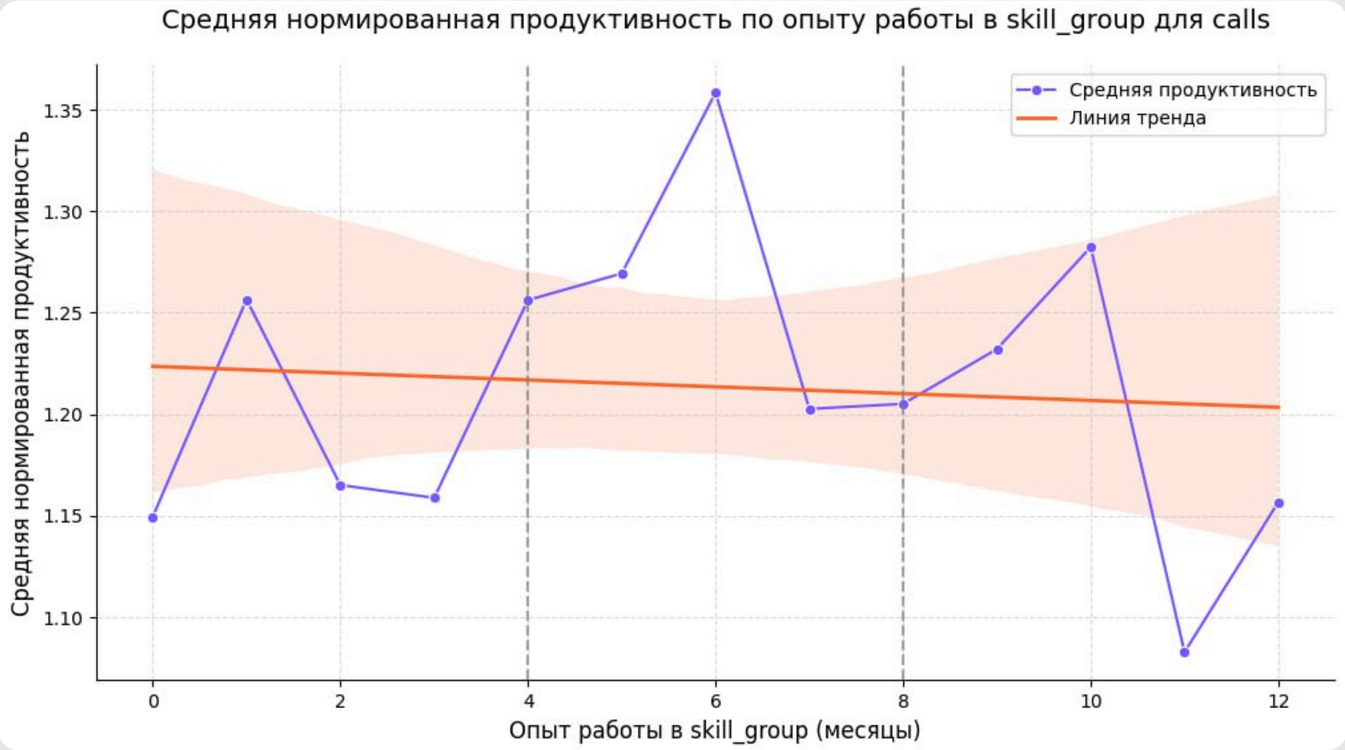
Механизм работы гипотезы



Матмодель

Математическая модель

для звонков

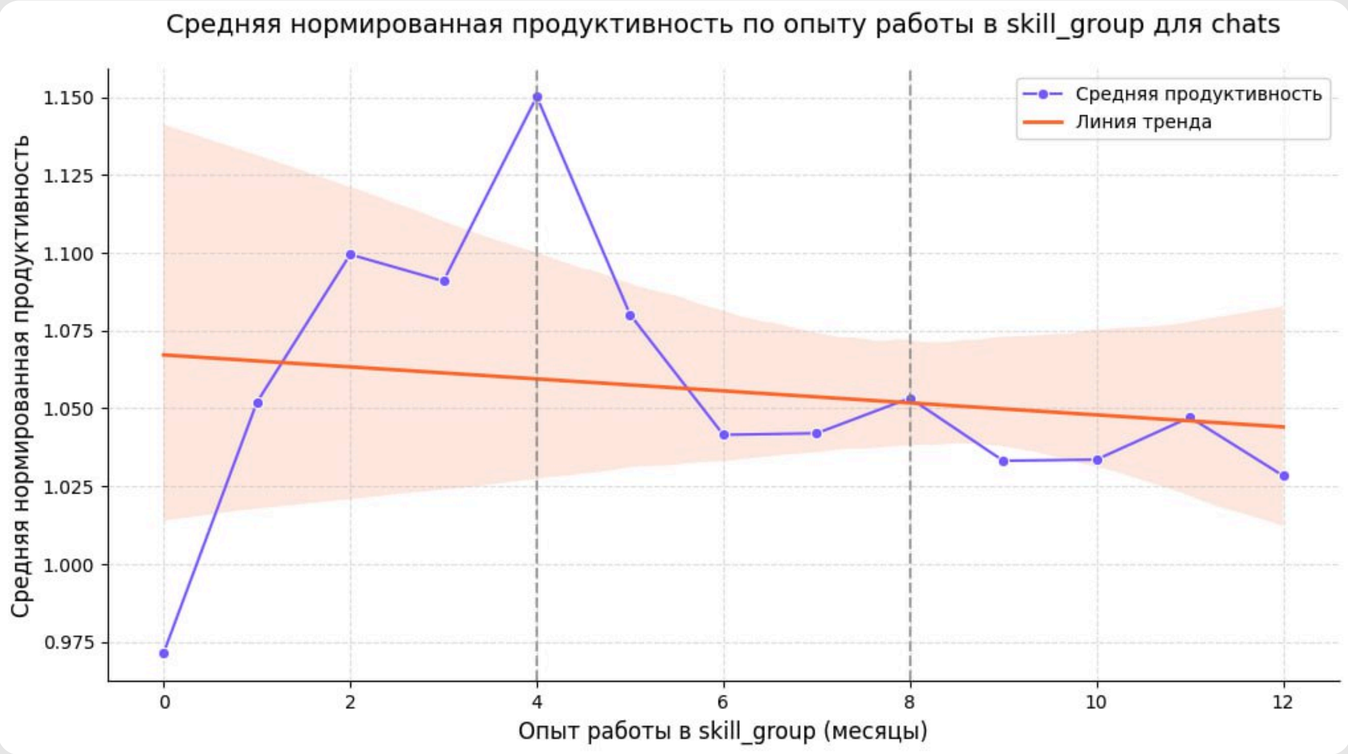


Разобьём выборку на 3 группы по стажу.

- Очень опытные - high
- Среднеопытные - meduim
- Малоопытные - low

Математическая модель

ДЛЯ ЧАТОВ



Сравним группы через тест Kruskal – Wallis

Сравним среднюю продуктивность по группам через тест Манни-Уитни

Результаты тестов

Звонки

Группа	W-статистика	P-value
Low	0.0577	0.0000
Medium	0.0284	0.0000
High	0.0268	0.0000

Тест Шапиро-Уилка

Статистика	P-value
155.7871	0.0000

Kruskal-Wallis H-тест

Результаты тестов

Звонки

Группа	U-статистика	P-value
Low vs Medium	171693718.5000	0.0000
Medium vs High	137220985.0000	0.0000
High vs Low	193897205.5000	0.0000

Тест Манна-Уитни

Low	Medium	High
1.1923010223858188	1.269292684465458	1.2062973668716939

Среднее для всех

Результаты тестов

Чаты

Группа	W-статистика	P-value
Low	0.2713	0.0000
Medium	0.1952	0.0000
High	0.2110	0.0000

Тест Шапиро-Уилка

Статистика	P-value
25.7977	0.0000

Kruskal-Wallis H-тест

Результаты тестов

Чаты

Группа	U-статистика	P-value
Low vs Medium	375063330.5000	0.0000
Medium vs High	321984589.0000	0.0606
High vs Low	408595879.0000	0.0024

Тест Манна-Уитни

Low	Medium	High
1.0543669035107854	1.0964586577230067	1.0382571914141079

Среднее для всех

Проверка на устойчивость

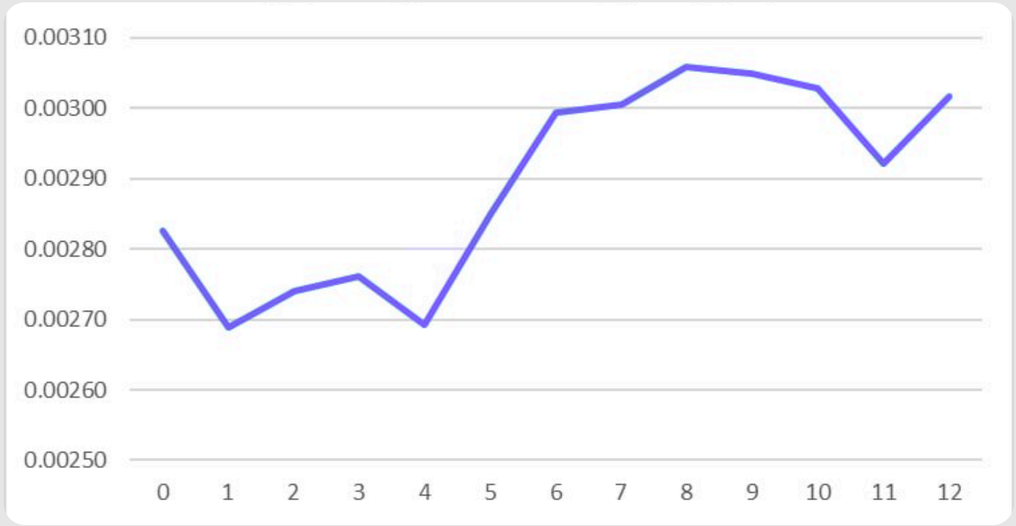
- Мы взяли только мужчин для проверки устойчивости
- Мы провели такие же 3 теста
- Все тесты подтвердили стат-значимость полученных результатов

Проверка на устойчивость

Для мужчин

Группа	Звонки	Чаты
Low	1.1684510391403513	1.0543669035107854
Medium	1.299843674258132	1.0964586577230067
High	1.2223585748957175	1.0382571914141079

Тесты Манна-Уитни и Kruskal-Wallis H-тест показали статзначимые результаты



Средняя нормированная продуктивность от времени работы в скилл-группе

Проверка на устойчивость

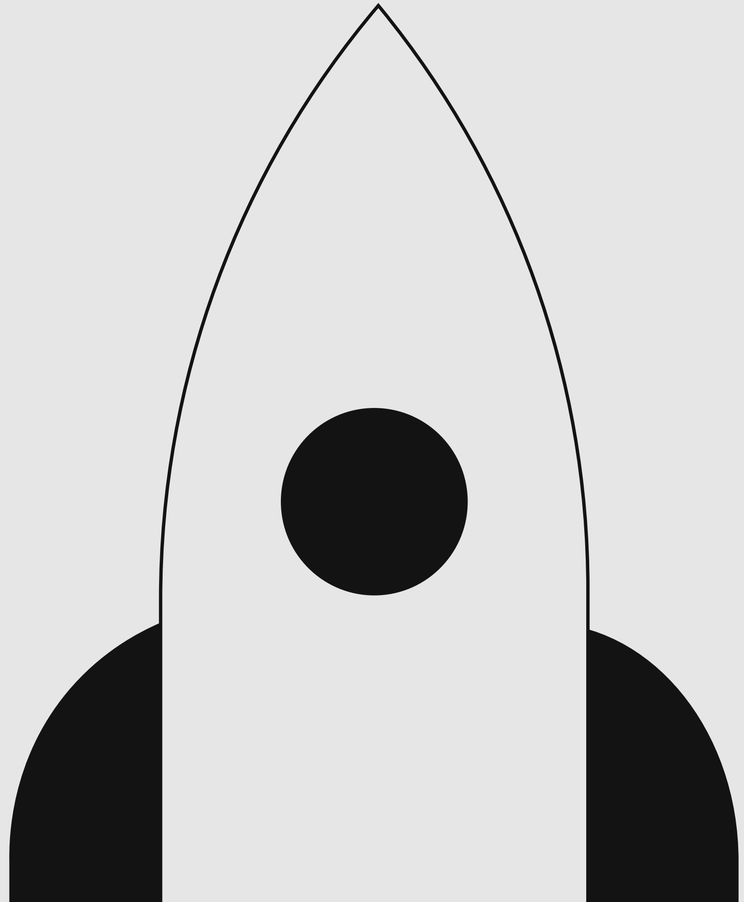
Для женщин

Группа	P_value для чатов	P-value для звонков
Low vs Medium	0.0015	0.0000
Medium vs High	0.8689	0.0722
High vs Low	0.0006	0.0000

Группа	Звонки	Чаты
Low	1.2191423140590865	1.0738964407577714
Medium	1.2319006359492872	1.0448576436229504
High	1.1674224966102984	1.030588359040341

Общее время возрастает по мере роста времени работы

Общий вывод



Лучший отдых — это смена деятельности

ЭТО БАЗА

Выводы из матмодели

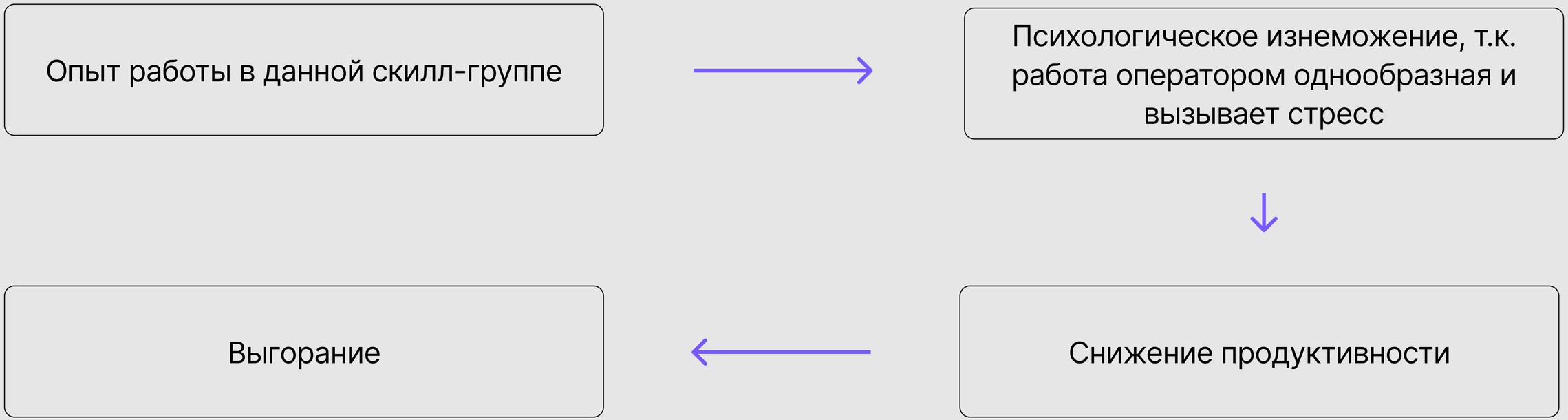
Для мужчин

Время работы в скилл-группы положительно влияет на мужчин. Со временем они выходят на плато продуктивности

Для женщин

У женщин продуктивность падает со временем работы в скилл группе

Альтернативный механизм



Источники: https://www.theflowspace.com/mental-health/brain/stress-women-vs-men-2938783

https://adme.media/articles/my-zadali-10-voprosov-operatoram-kol-centrov-i-oni-bez-prikras-rasskazali-vse-o-svoej-rabote-2247465/

Решение

Policy implication

Мобильность мужчин

Сохранять сотрудников в одном отделе (skill group)

Мобильность женщин

Ротация женщин между скилл-группами

Ограничения методики

- Не хватает понимания, какие звонки/чаты являются сложными
- Делается допущение, что распределение продуктивных и не продуктивных сотрудников по группам - равномерное
- Не учитываются данные о прошлом опыте работы, графике и др.
- Так как у нас есть данные опыта в skill-группе максимум за 12 месяцев, мы не можем точно экстраполировать логику на более длительный срок

НАША КОМАНДА

Владимир Габестро

Степан Буян

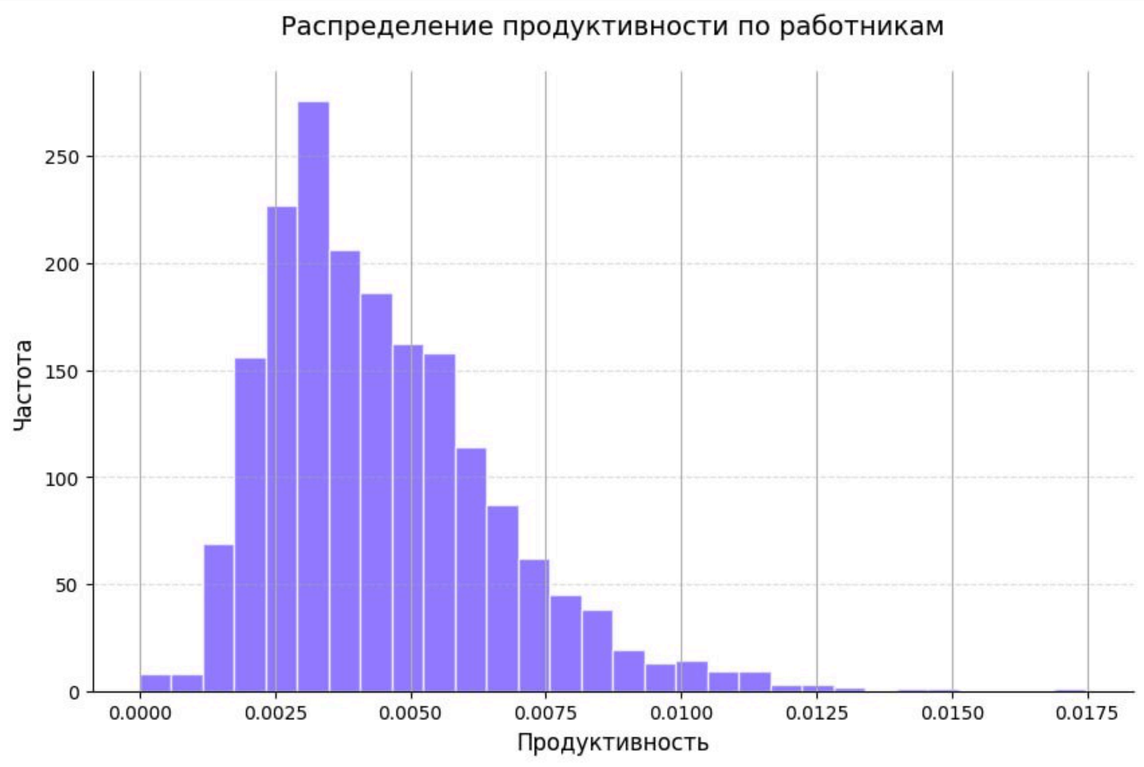
Пётр Палов

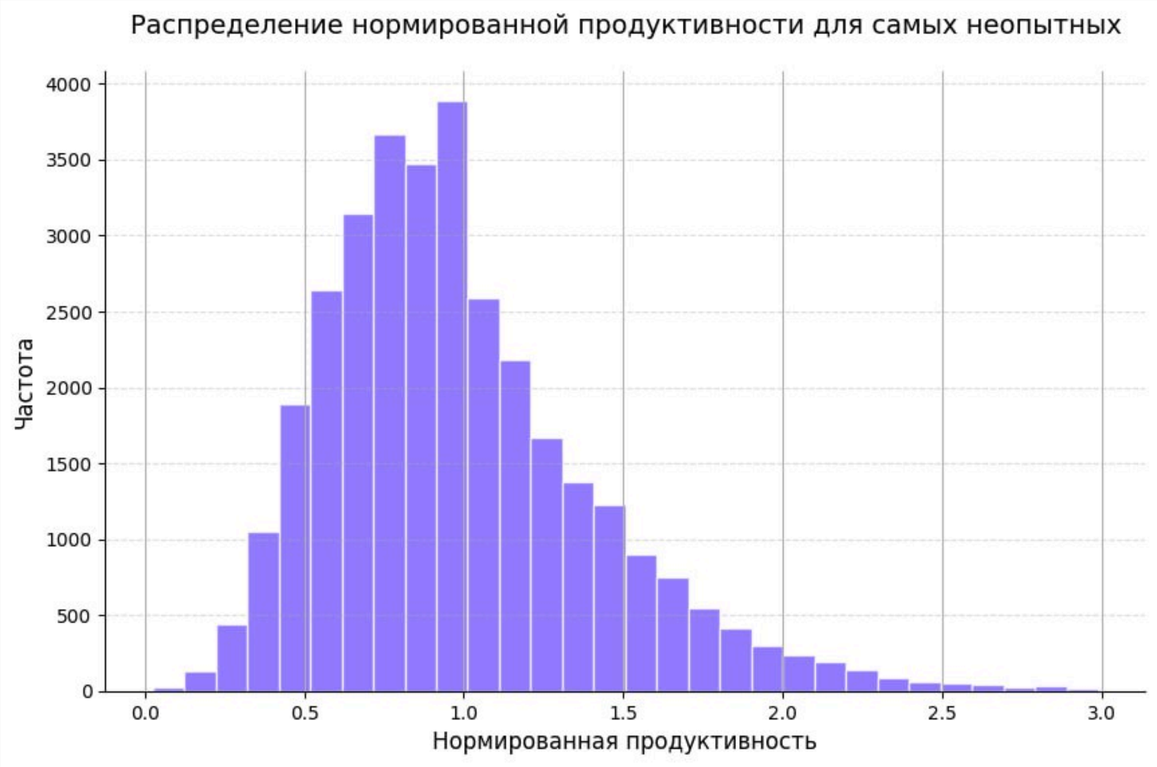
Вадим Прокофьев

Егор Бобров

Полина Никулина

Приложение


Распределение продуктивности по работникам
Частота
Продуктивность
0
50
100
150
200
250
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175


Распределение нормированной продуктивности для самых неопытных
Частота
Нормированная продуктивность
0
500
1000
1500
2000
2500
3000
3500
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Математическая модель

Тест Шапиро-Уилка

Этот тест проверяет нормальность гипотезы о том, что распределение величины нормально.

Процесс проходит в несколько этапов:

1. Определение сходства между наблюдаемым и нормальным распределениями. Для этого на наблюдаемое распределение накладывают нормальную кривую.
2. Вычисление процента сходства. Рассчитывается, какой процент выборки совпадает с нормальной кривой.
3. Определение вероятности получения такого процента сходства. Расчёт ведётся на основе предположения, что распределение в популяции точно нормальное (нулевая гипотеза).

Если значение p меньше выбранного альфа-уровня, то нулевая гипотеза отклоняется, и есть доказательства того, что проверяемые данные распределены ненормально. В противном случае нулевая гипотеза не отклоняется, и делается вывод, что нет статистически значимых доказательств того, что данные не принадлежат к нормально распределённой популяции.

Тест Левена

Принцип работы теста Левена заключается в сравнении средних различий между показателями в каждой группе с общим средним значением по всем группам.

Если различия между группами стабильны и похожи, то тест Левена предполагает, что группы имеют примерно одинаковую вариабельность.

Если различия между группами значительно отличаются, то это указывает на то, что вариабельность (разброс значений) больше различается между группами. Это может означать, что на различия между группами влияет что-то, кроме случайности.

Нулевая гипотеза теста Левена заключается в том, что сравниваемые группы имеют равные дисперсии в популяции. Если это верно, то, вероятно, будут обнаружены слегка разные дисперсии в выборках из этих популяций.

Важно отметить, что средние значения отдельных групп не влияют на результат теста, они могут отличаться.