База №2 Тинькофф:
Заказы билетов в кино

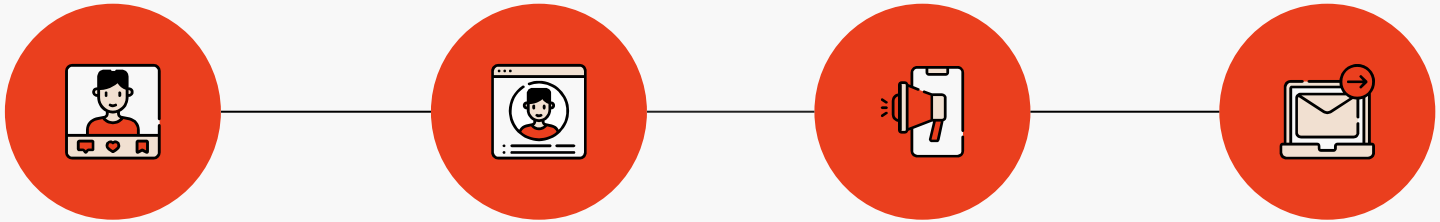
Предугадывание количества билетов в одном заказе

Исследование команды:
“DANONETKИ”

Актуальность

Тинькофф, оценивая количество билетов в потенциальном заказе, сможет использовать эти данные для увеличения прибыли. К примеру, проводить промоакции.

Dataset



Данные заказов билетов на сервисе «Тинькофф Афиша»

2020 год

525271 запись

46 переменных

ИВ: Зависит ли количество билетов в заказе от каких-либо характеристик?

- **Основная гипотеза:** Количество билетов в одном заказе можно предсказать, имея определенные характеристики клиента, кинотеатра и фильма.

- **Гипотеза 2:** Так как раз наша гипотеза заключается в том, что на фильмы с ограничением меньше 16 ходят больше детей, чем на взрослые фильмы, и поэтому там больше билетов в заказе

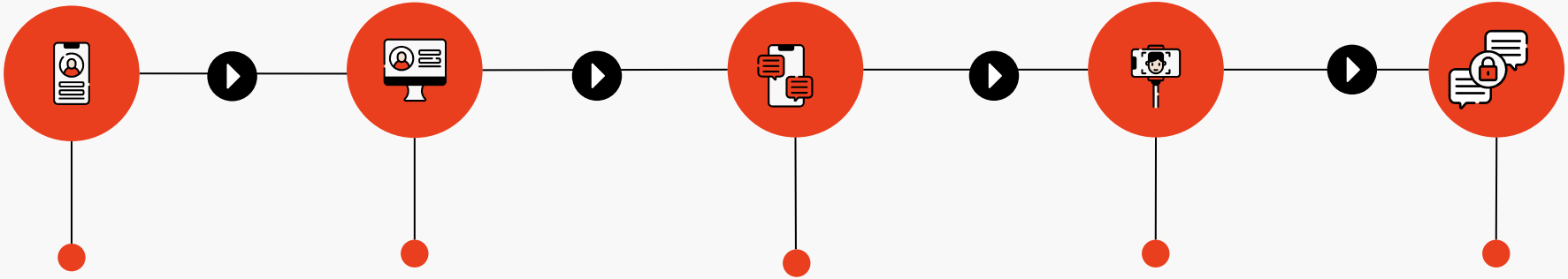
Механизм 1

- Данные характеристики предоставляют достаточную информацию для построения модели предсказания количества билетов => наша модель может предсказывать количество билетов, основываясь только на этих данных

Механизм 2

- Возрастное ограничение на фильм - менее 16 лет => ходят больше детей и подростков => подростки ходят в кино компаниями, а дети ходят с родителями => больше билетов в одном заказе

Этапы исследования



Анализ существующих исследований

Предварительный анализ данных

Анализ переменных

Мат.часть (Корреляции, регрессия, класстеризация, группировка, нейронная сеть)

Интерпретация результатов

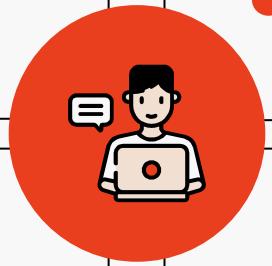
Исследуемые переменные

Возрастные ограничения

- 0+, 6+, 12+
- 16+, 18+

Жанр

- 19 жанров



Стоимость билетов в условных единицах

Дата сеанса

- Выходной или будний

Местоположение кинотеатра

- Продолжительность фильма

Введенные переменные:

- session_date

Выходной или будний день принимает значение 0,1

counts_of_repeats_by_people

Использовал ли клиент Тинькоff Афишу >1 раза за данный промежуток времени

Работа с DATASET-ом

1) убрали показ трансляций, оперу, театр (**1427 записей**)

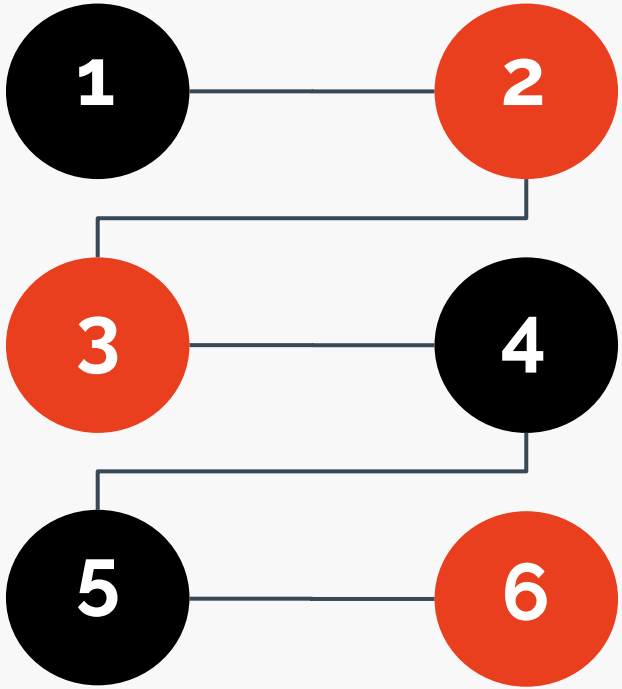
2) удалили строки без значений в исследуемых переменных→ (NaN's) (**5392**)

3) обработали фильмы без рейтинга (добавили им **актуальный рейтинг**)

4) избавились от фильмов без жанров и от жанров без фильмов (**21572 строк**)

5) Привели названия фильмов к **одному формату**

6) обработали **выбросы** по ценам билетов (**10131**)

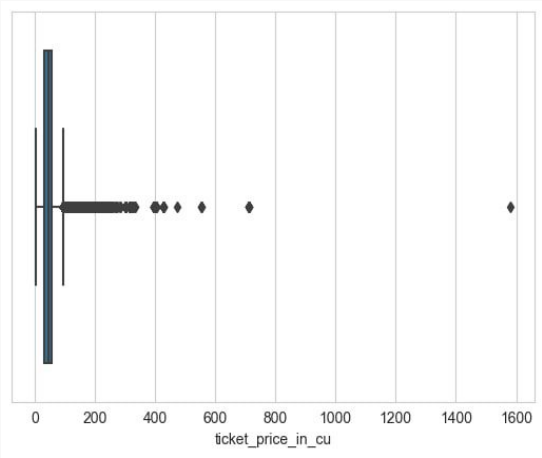


Работа с DATASET-ом

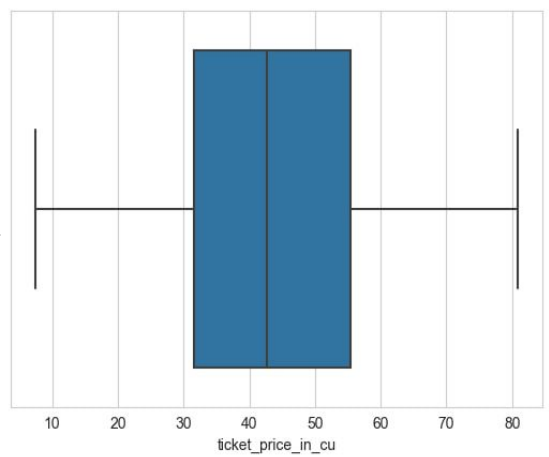
'genre_is_fairy_tale'
'genre_is_animation'
'genre_is_humor'

→ Фильмов таких жанров нет в нашем DATASET-е

было:



стало:



Удалено строк: 10131 (выбросы)

Корреляция между количеством билетов и приведенными переменными

	session_date	movie_duration	movie_age_restriction	number_of_tickets	ticket_price_in_cu	retrurn
session_date	1.000000	-0.031901	-0.041456	0.058846	0.190925	-0.012996
movie_duration	-0.031901	1.000000	0.173229	-0.043108	0.090409	-0.045889
movie_age_restriction	-0.041456	0.173229	1.000000	-0.133129	0.088947	0.003109
number_of_tickets	0.058846	-0.043108	-0.133129	1.000000	0.024746	-0.042591
ticket_price_in_cu	0.190925	0.090409	0.088947	0.024746	1.000000	-0.022793
retrurn	-0.012996	-0.045889	0.003109	-0.042591	-0.022793	1.000000

Анализ данных: Нейронная сеть

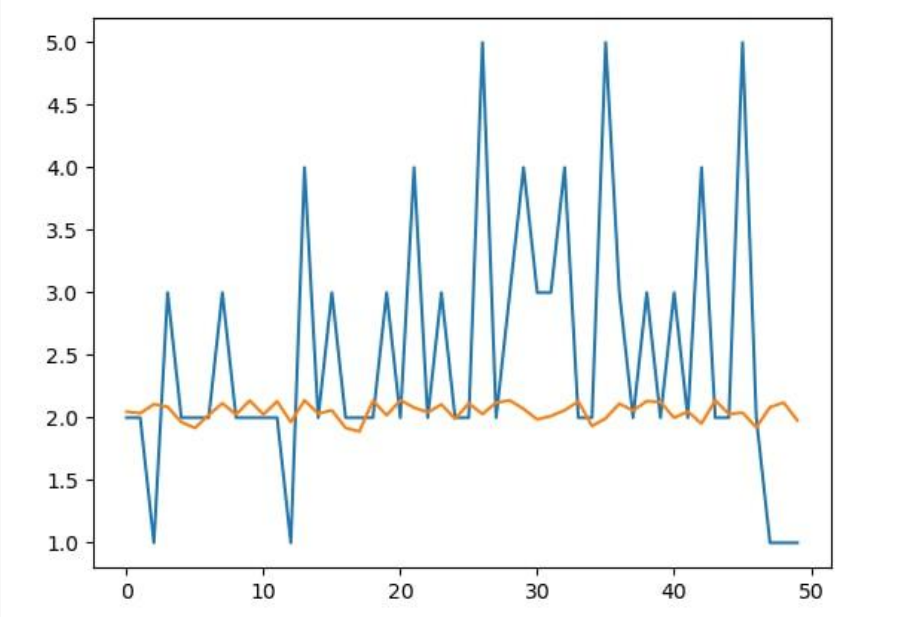
Общая

Пробуем обучить нейросеть на всём массиве данных

По группам

Пробуем разбить базу данных на группы для лучшего прогноза нашей сети

Нейронная сеть: вся база данных



Синяя полоса → реальное количество билетов в заказе
Оранжевая полоса → прогноз нейронной сети

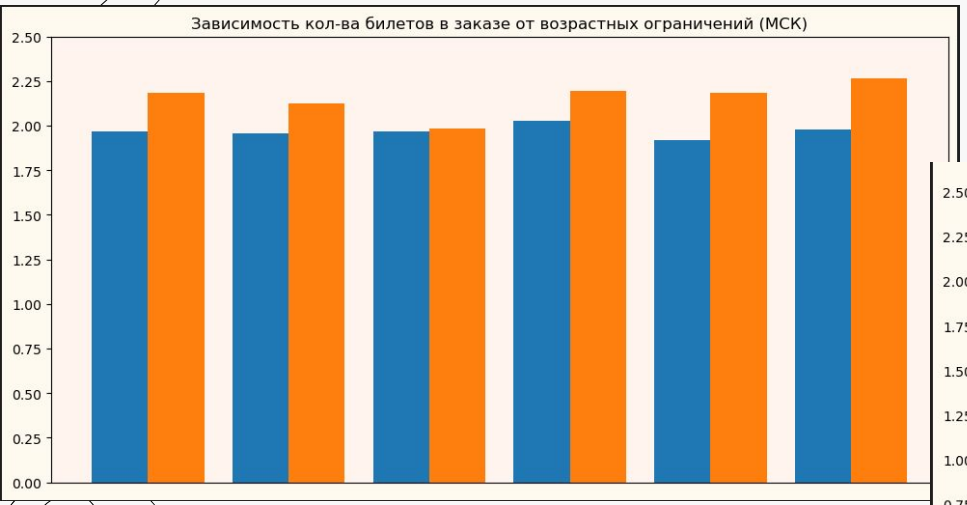
Средняя ошибка примерно 0,5 → это является посредственным результатом

Нейронная сеть: разбиение по группам

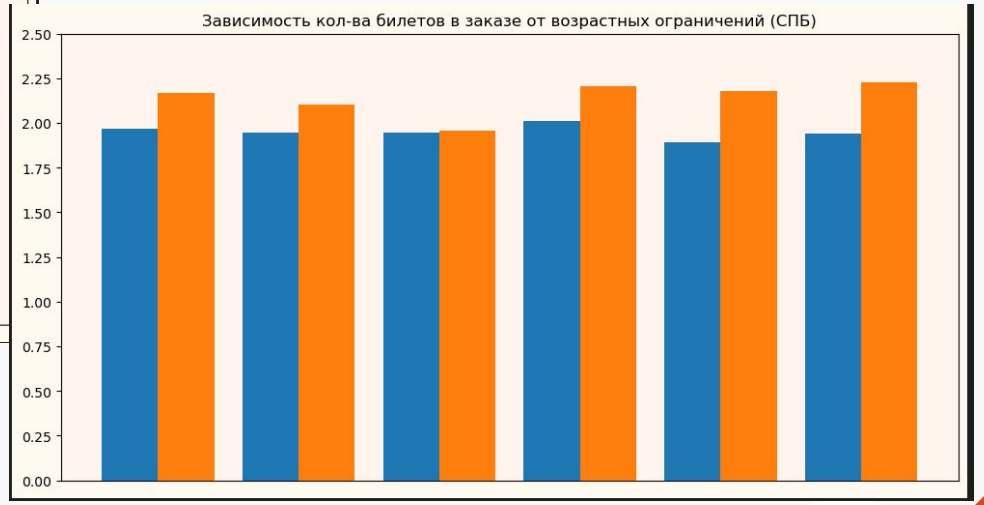
Первая попытка разбиения по группам с помощью кластеризации оказалась неуспешной

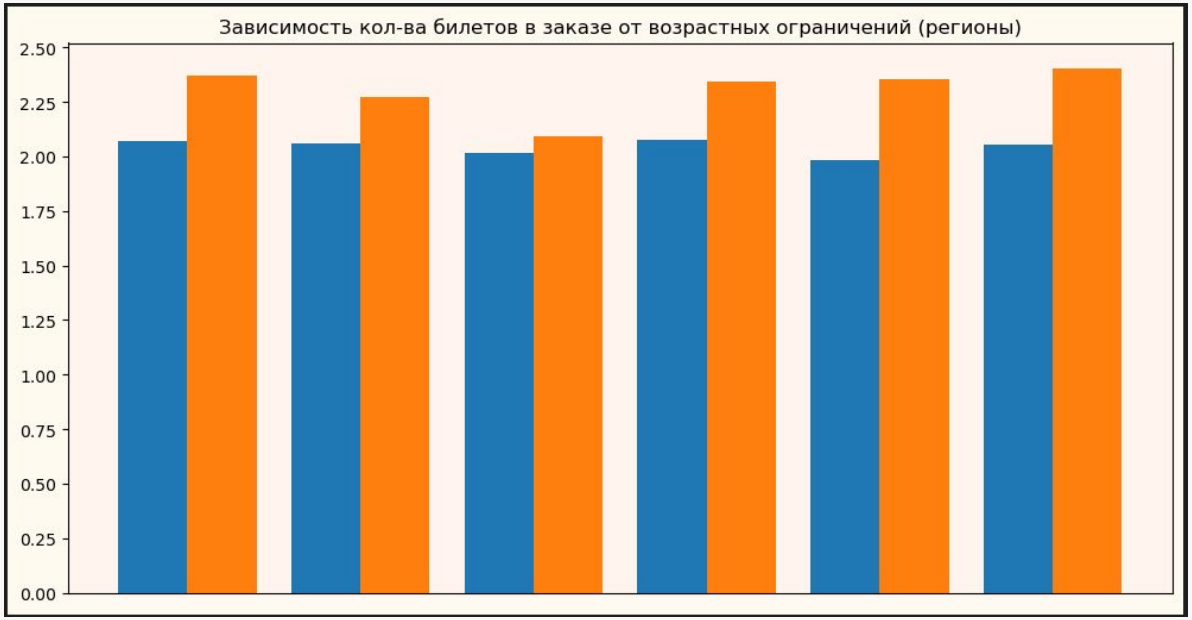
Успешная попытка: Разбиение по группам по возрастному ограничению на фильмы
→ 1 группа - заказы по фильмам, у которых возрастное ограничение меньше 16 (0+, 6+, 12+)
→ 2 группа - больше 16 (16+, 18+)

Почему мы выбрали такой метод группировки?



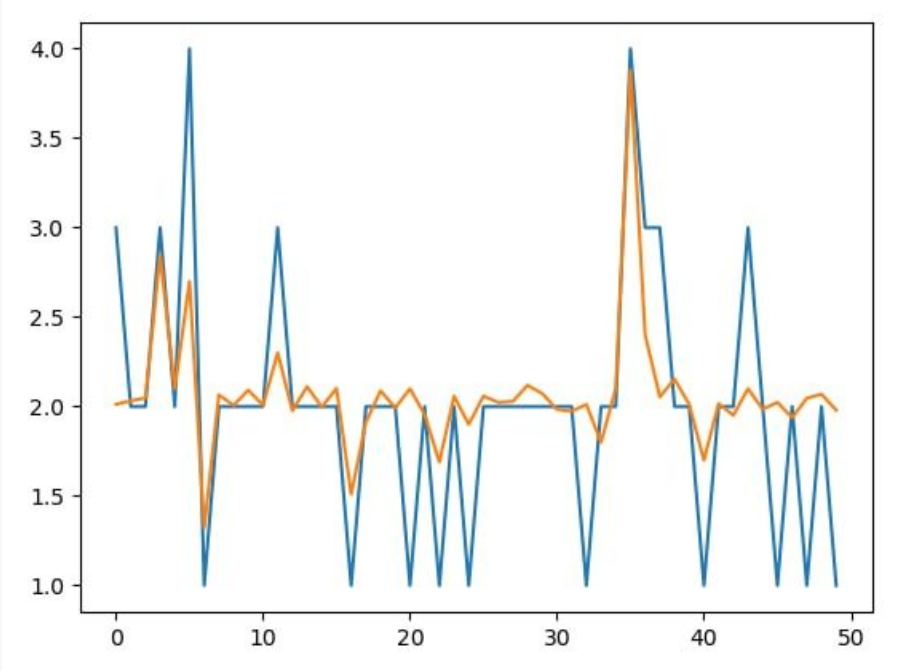
- Среднее кол-во билетов в заказе на фильм 16+, 18+
- Среднее кол-во билетов в заказе на фильмы 0+, 6+, 12+





Вывод: в каждой из сравниваемых выборок разница составляет около 15%

Результат работы нейросети при используемой нами группировке:



Синяя полоса → реальное количество билетов в заказе
Оранжевая полоса → прогноз нейронной сети

Средняя ошибка примерно 0,28 → что является хорошим результатом

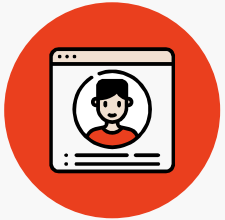
Каких переменных не хватает?

Бюджет фильма

Числовой коэффициент

Индекс рекламной кампании

Бюджет, способы продвижения



Информация о создателях фильма

Режиссер, актеры, съемочная группа

Выводы

Наша основная гипотеза подтвердилась. При разбиении нашего датасета на группы, мы смогли обучить нашу нейронную сеть предсказывать количество билетов в заказе с допустимой ошибкой.

Рекомендации по улучшению проекта:

1. Добавление переменных, таких как бюджет фильмов, индекс рекламных кампаний, информация о создателях фильма, возраст и пол клиента.

2. Поиск более оптимальной группировки.

3. Улучшение алгоритма нейронной сети.

В главных ролях:

Ден Косинов

-злодей британец

Илья Лукин

-самый сексуальный мужик в мире

Никуша Кузьмина

-горячая чикса

Костя Литвинов

-недопонятый гений

Глеб Карташов

-какой-то мужик

Дополнительные данные