Исследование данных:

**Демин Георгий**: внешняя информация доступна только для 800 пользователей

**Павел Белов**: Отбирая 5000 популярных товаров мы “теряем” информацию о “новинках”. Они не могут быть популярными из-за малого времени продаж, но с точки зрения бизнеса интересны для пользователей.

Проверил на тестовых периодах для моделей 1 и 2 уровня:

data['item\_id'].nunique() # (89051)

data\_train\_lvl\_1['item\_id'].nunique() # (83685)

data\_val\_lvl\_1['item\_id'].nunique() # (27649)

data\_val\_lvl\_2['item\_id'].nunique() # (24329)

np.intersect1d(data\_train\_lvl\_1['item\_id'].unique(), data\_val\_lvl\_1['item\_id'].unique()).shape[0] # (24469)

np.intersect1d(data\_val\_lvl\_1['item\_id'].unique(), data\_val\_lvl\_2['item\_id'].unique()).shape[0] # (17646)

Первые две строки показывают, что 5366 товаров из изначальной базы данных ни разу не были куплены за период до последних 9 недель. Наша модель первого уровня ничего не будет о них знать во время тренировки, хотя предположительно это - новинки.

Пересечение валидационных выборок 1 и 2 уровня еще сильнее отличается.

Практически ровно 10000 товаров, на которых будет обучаться бустинговая модель не войдут в валидационную выборку, но там появится порядка 6000 “свежих” товаров, о которых известны только данные из item\_features.

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*

\*