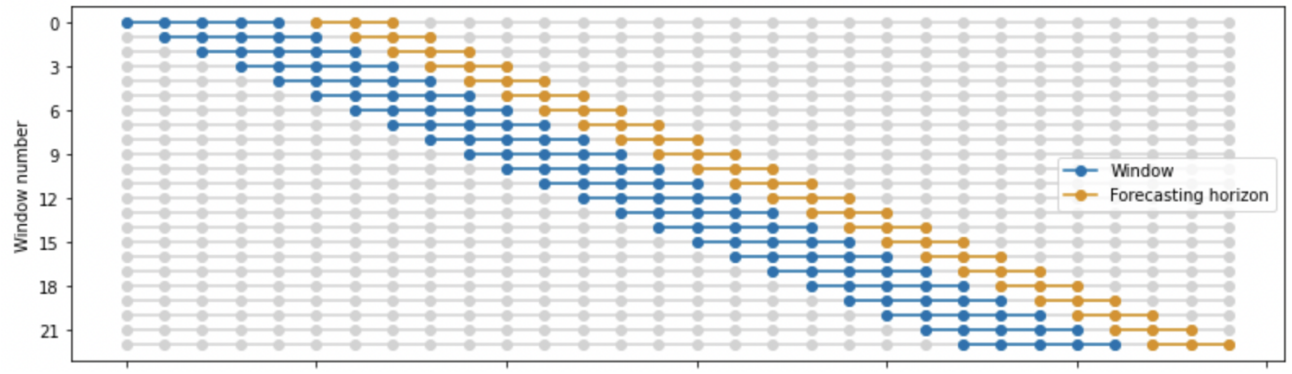
Огляд рішення команди AfterParty

Валідація:

Валідацію робимо time based k-fold стратегією, в якій в трейн потрапляє 1 місяць, а в валідацію 2 тижні (окремо рахуємо метрику на 1-й і 2-й тиждень, щоб змоделювати public/private розбиття на lb):



Тривіальне рішення (базовий підхід):

Для кожної дати і (geoCluster, SKU) пари робимо прогноз на всі 14 днів медіанним значенням продаж SKU в geoCluster за останній місяць.

Local validation		Leaderboard	
public (1st week)	private (2nd week)	public (1st week)	private (2nd week)
0.93600	0.93877	0.93858	~0.939 (estimate)

Model-based базове рішення:

Ознаки:

- geo_cluster_id
- sku_id
- median_sales - (медіанне значення кількості продаж sku_id в geo_cluster_id за попередній місяць)

- median_non_zero_sales - (медіанне значення кількості продаж sku_id в geo_cluster_id за попередній місяць без урахування днів, коли продажів не відбувалось)

Модель: градієнтний бустинг (lightgbm) з MAE loss і early stopping за цільовою метрикою змагання

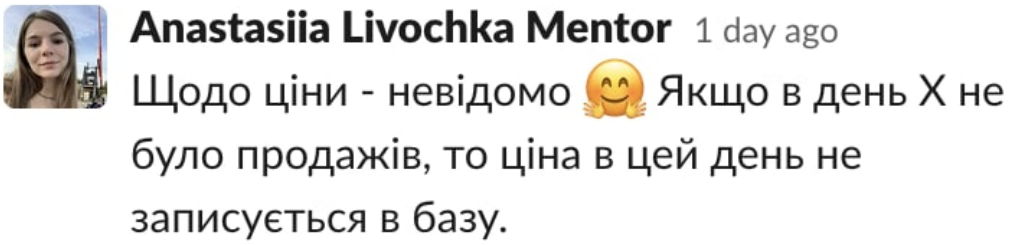
Local validation		Leaderboard	
public (1st week)	private (2nd week)	public (1st week)	private (2nd week)
0.911177	0.9186835	0.91128	~0.918 (estimate)

Додаткові ознаки:

Також додали інші фічі, такі як: city_id, category_id, product_type_id, brand_id, trademark_id, origin_country_id, commodity_group_id, згруповані по city_id/ category_id/ product_type_id/ brand_id/ trademark_id/ origin_country_id/ commodity_group_id медіани продажів, weekday, week_no (week of year), month та эмбеддінг вектори текстових ознак товарів, проте все це істотно на результат не вплинуло.

p.s. маппінг назв використаних колонок на первинні дивіться у файлі *col_mapping.json*
(генерується у ноутбуці FN-eda-data-cleaning-and-no-ml-baseline-v2.ipynb)

Meta Feature:



Оскільки в тестовому датасеті ціни заповнені всюди (а продажі, очевидно, не всюди були), то природним чином можна вирішити, що ціни в тесті дозаповнили

штучним чином й це нам може допомогти. Спробуємо досягти того ж ефекту на трейні і валідації.

Очевидні варіанти як можна “дозаповнити ціну” штучно:

1. для днів, в яких продажів товару sku_id в geo_cluster_id не було, дозаповнимо їх такою ж, як в день попереднього продажу в тесті, якщо ж попереднього продажу не було, то заповнимо ціною наступного продажу (ffill, bfill)
2. для днів, в яких продажів товару sku_id в geo_cluster_id не було, дозаповнимо їх такою ж як в день попереднього продажу в тесті, якщо ж попереднього продажу не було, то заповнимо ціною останнього продажу з трейну
3. для днів, в яких продажів товару sku_id в geo_cluster_id не було, дозаповнимо їх такою ж як в день попереднього продажу в тесті, якщо ж попереднього продажу не було, то заповнимо її деякою середньою/медіаною для sku_id серед інших geo_cluster_id в той же день

Оскільки в трейні ціни на більшість sku_id в рамках кожного geo_cluster_id досить часто змінюються, то за рахунок штучного характеру дозаповнення пропущених цін можна згенерувати корисний індикатор (price_change) - “чи змінилася ціна в цей день для sku_id в geo_cluster_id порівняно з минулим днем?”, який ідейно інтерпретується так:

- ціна відносно попереднього дня змінилась -> в цей день точно (100%) були продажі > 0
- ціна відносно попереднього дня не змінилась -> ймовірно продажів не було (але це не точно), оскільки скоріш за все це просто дублювання ціни з моменту останньої реальної продажі

Порахувавши статистики для price_change по різних періодах трейну і тесту, а також метрики на валідації і public leaderboard ми дійшли до висновку, що скоріш за все було обрано схему подібну до варіанта 3 для заповнення пропущених цін в тесті.

Зауважимо, що фічі пов`язані з price_change є даталіком і їх неможливо використати в продакшн режимі, оскільки в продакшені ми не знатимемо майбутніх цін (або знатимемо, але вони будуть формуватися за зовсім іншою логікою, і не можуть бути так використані).
Корисність (за метрикою на лідерборді) price_change фічі пояснюється не величиною зміни ціни, а саме природою цієї зміни (ми знаємо ціну, тільки якщо були продажі) і по суті для прогнозу кількості продажів в певний день ми використовуємо інформацію про те чи були вони взагалі.

Як тільки ми усвідомили причину покращення результатів через використання фічі price_change, ми повідомили про це представнику від організатора й отримали відповідь про те, що її використання є легальним.

Фінальне рішення:

Фінальним рішенням є ансамбль з lightgbm та feed-forward нейронної мережі з MAE loss і early stopping за цільовою метрикою змагання, що використовує всі вищенаведені фічі.

LightGBM:

Local validation		Leaderboard	
public (1st week)	private (2nd week)	public (1st week)	private (2nd week)
0.7384	0.7938	0.75180	~0.81 (estimate)

Feed-Forward Network:

Local validation		Leaderboard	
public (1st week)	private (2nd week)	public (1st week)	private (2nd week)
0.7669	0.8216	0.7598	~0.81 (estimate)

Ensemble:

Local validation		Leaderboard	
public (1st week)	private (2nd week)	public (1st week)	private (2nd week)
-	-	0.7482	~0.8 (estimate)