CORPORACIÓN
FAVORITA

Grocery Sales Forecasting

kaggle

Команда

- Ahmet Erdem
- Andrey Filimonov

slonoschildpad : 3rd place / 1675

Задача

- Предсказать объем продаж каждого товара в каждом магазине в течение будущих 16 дней
- Public Leaderboard : первые 5 дней
- Private Leaderboard : последние 11 дней

Метрика

- Normalized Weighted Root Mean Squared Logarithmic Error (NWRMSLE)

$$NWRMSLE = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(\ln(\hat{y}_i + 1) - \ln(y_i + 1)\right)^2}{\sum_{i=1}^{n} w_i}}$$

- w – веса продуктов (1.25 для скоропортящихся, 1 для остальных)

Данные

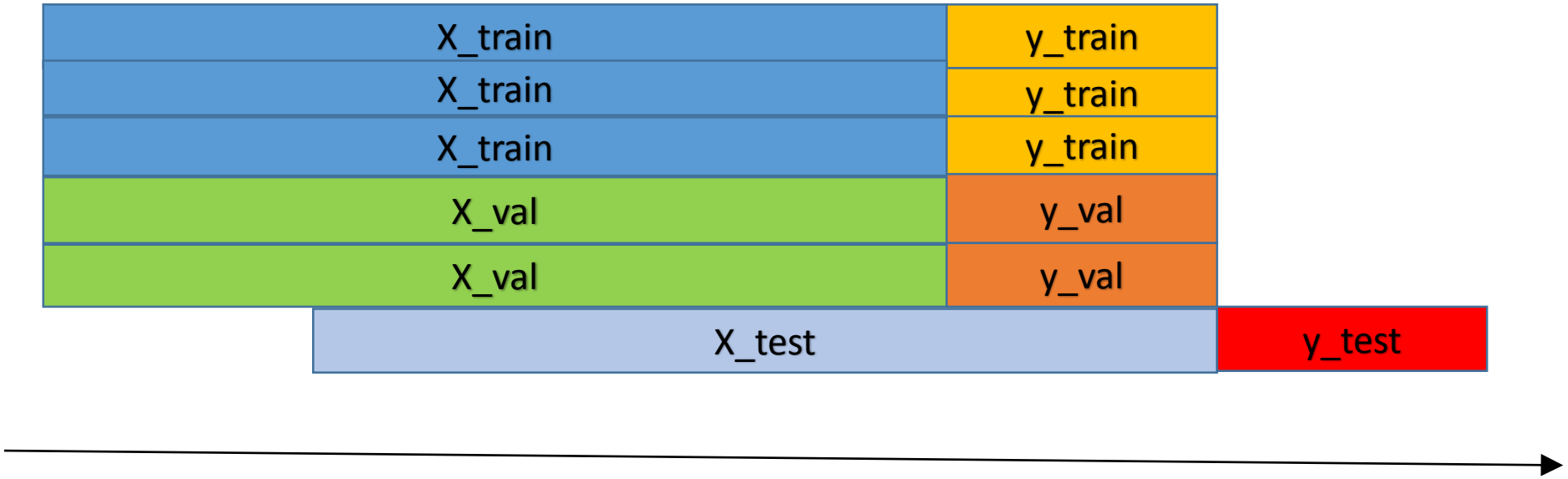
- История продаж товаров и промо-акций с 2013 года, ~170К пар (магазин, товар)
- Признаки товаров (семейство, класс, является ли скоропортящимся)
- Признаки магазинов (город, штат, тип, кластер)
- Количество транзакций в магазинах по дням
- Выходные дни (государственные и местные)
- Цены на нефть

Различия train и test

- Записи с onpromotion = True отсутствуют в train
- ~40K пар (магазин, товар) отсутствуют в train

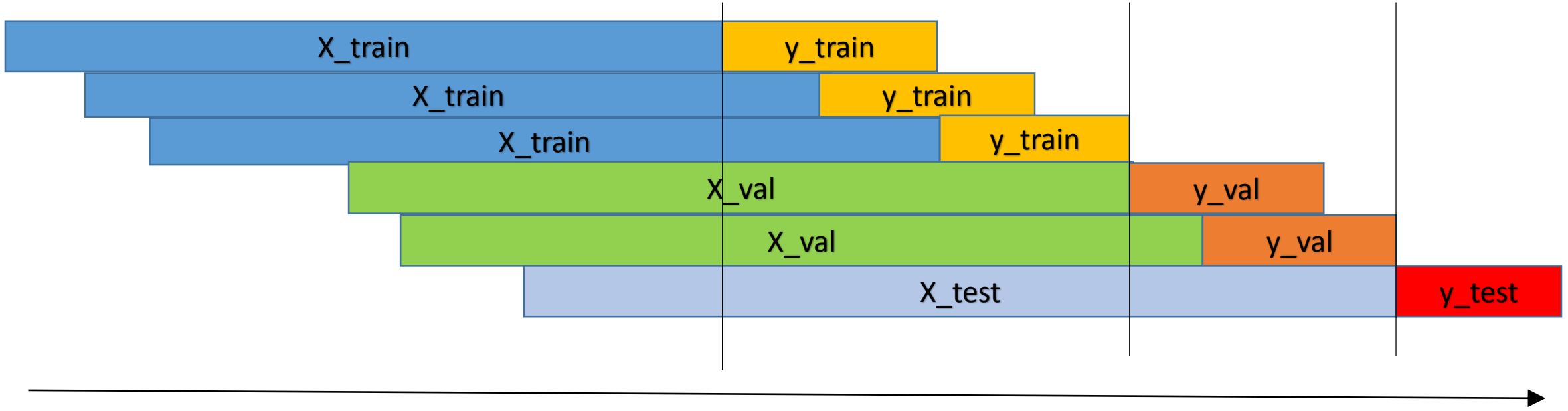
Валидация временных рядов

- Рискованно, есть заглядывание в будущее



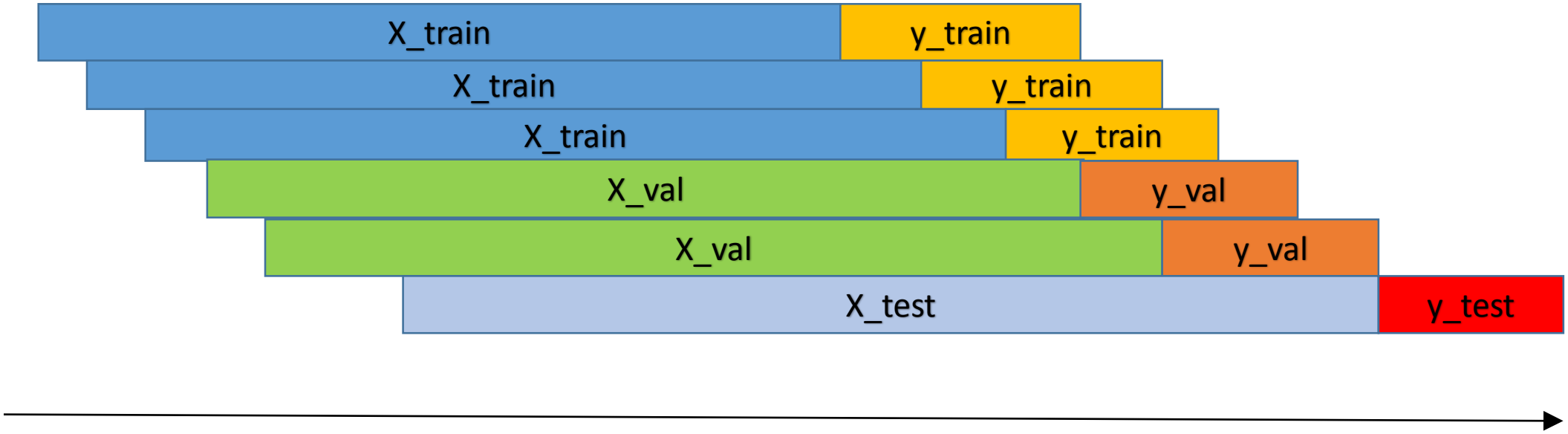
Теоретически правильный вариант

- Нет утечки будущей информации



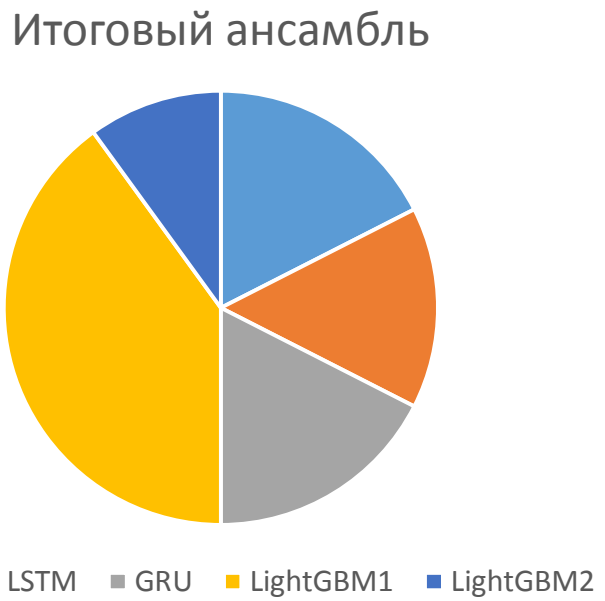
Как валидировались мы

- Шаг нарезки = 1 неделя



Решение

- 50% : 3 NN Ensemble (CNN, LSTM, GRU), seq2seq approach, weekly slices (Andrey Filimonov)
- 40% : LightGBM, monthly slices (Ahmet Erdem)
- 10% : Another LightGBM, based on public kernel, weekly slices (Andrey Filimonov)

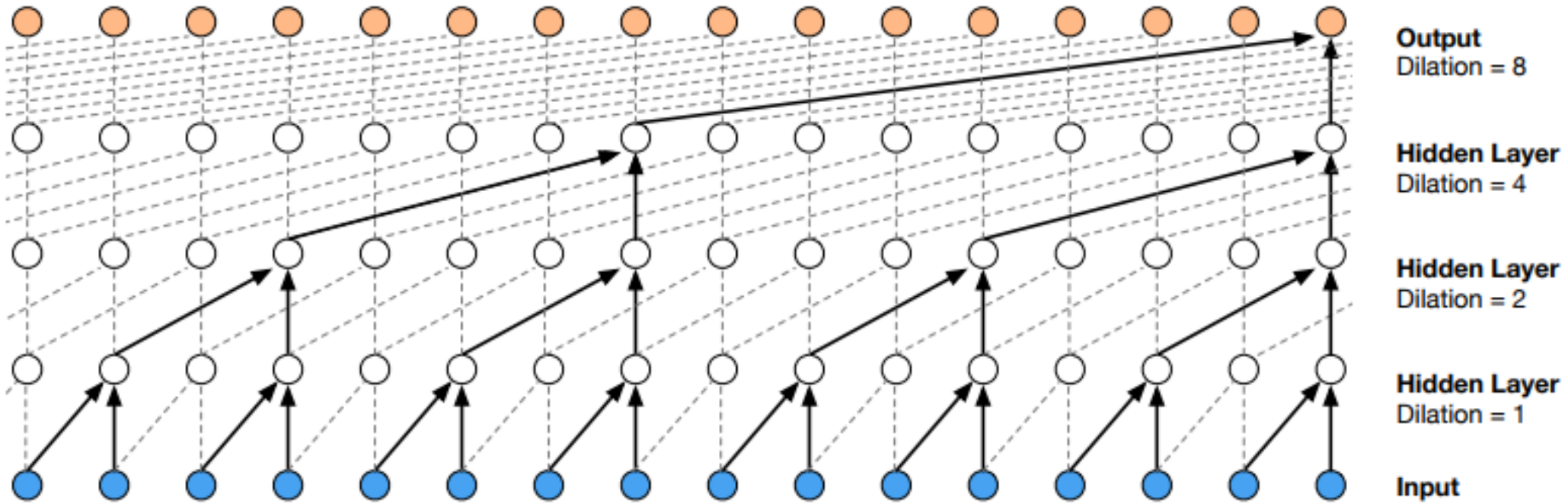


Нейросети

- Input sequences : 80..90 days, output sequences : 16 days

- 3 top layers (recurrent or convolution)
- 2-3 bottom fully-connected layers
- CNN with dilated convolutions : 2-8-32

Stack of dilated convolutions



Предобработка данных:

- Sequences : shape = (item_store_pairs_num, history_len, 4)

1: логарифмированная история продаж по дням
2: история промо-акций
3: среднее логарифмов продаж товара по всем магазинам
4: среднее логарифмов продаж всех товаров в данном магазине

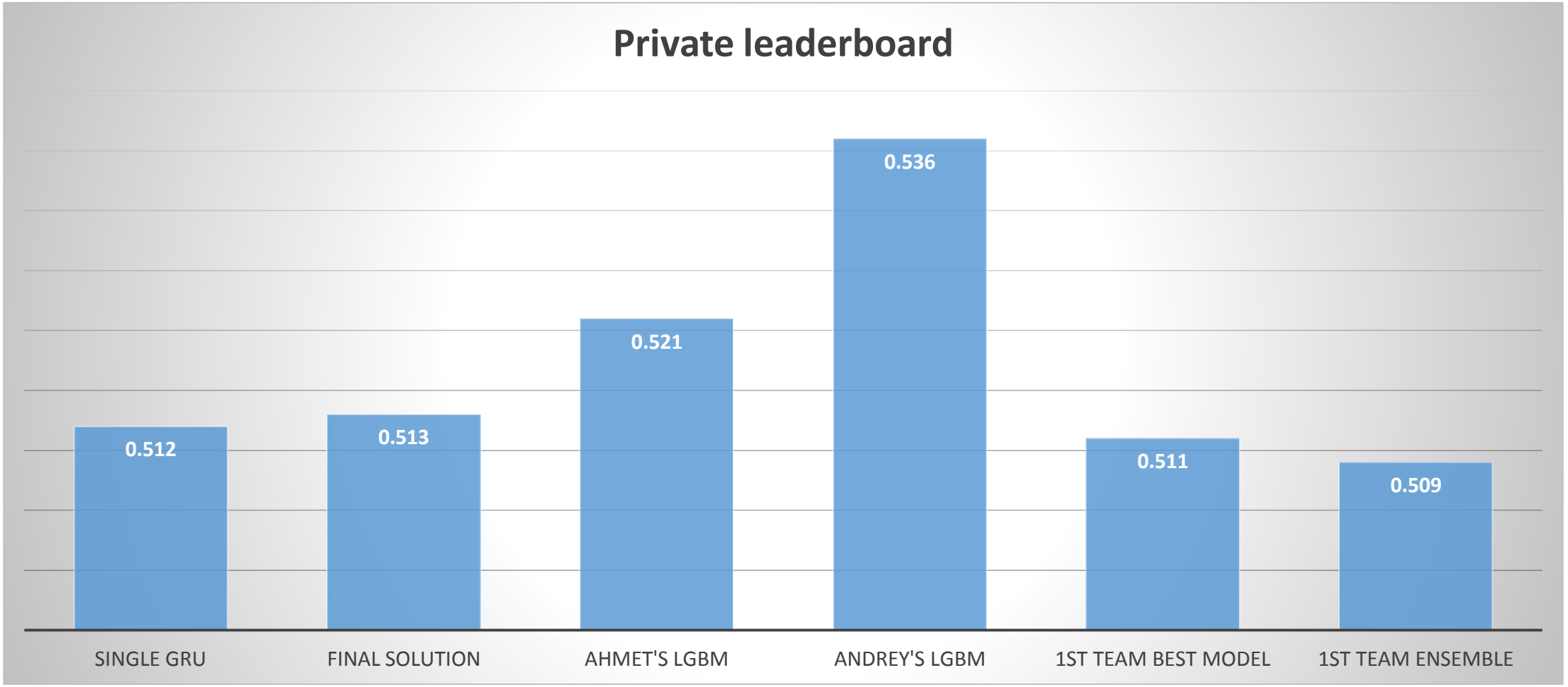
Дополнительные данные

- На полносвязные слои нейросети также подавались:
    1. Будущие промо-акции за 39 предсказываемых дней
    2. Embeddings категориальных признаков товаров и магазинов

Технические подробности

- Нейросети: 1070ti + 1080ti (последние 3 дня)
- Градиентный бустинг: Kaggle kernels

Результаты отдельных моделей



Что сработало

- Усреднение всех эпох (неожиданно)
- Переобучение на максимально свежих данных
- Усреднение нескольких обучений (ожидаемо)
- Подход seq2seq
- Нарезка данных по неделям
- Dilated convolutions

Что не сработало

- Заполнение пропусков в onpromotion (команды, не делавшие этого, оказались выше)
- Попытки предсказывать пары (магазин, товар), отсутствующие в train
- Назначение повышенных весов последним 11 предсказываемым дням
- Наша неудачная реализация градиентного бустинга
- Дополнительные данные: праздники, цены на нефть, Google trends…

Неожиданное

- Усреднение всех эпох оказалось самым эффективным
- Не работала почти никакая явная регуляризация
- Onpromotion adjustment не работает на private
- Ансамбль из 50% удачных моделей и 50% неудачных показывает практически те же результаты, что и состоящий из одних удачных
- Относительно простая доработка публичного скрипта победившей командой

Решение победителей :

- Авторы: Eureka, weiwei, infzero

- model_1 : 0.506 / 0.511 , 16 lgb models trained for each day
- model_2 : 0.507 / 0.513 , 16 nn models trained for each day
- model_3 : 0.512 / 0.515, 1 lgb model for 16 days
- model_4 : 0.517 / 0.519, 1 nn model based on @sjv's code

- Взвешенные средние от рядов продаж и промо-акций
- Разности средних со сдвигом
- Количество дней с момента первой/последней продажи за период
- Количество дней с момента первой/последней промо-акции за период
- Обучение на короткой истории (последние 1.5 месяца)