Отчёт

- Задача: спрогнозировать количество проданных товаров, имея данные о количестве продаж сагрегированные по разным срезам.
- Описание итогового решения:
  - В качестве таргета используем 'y'
  - Отбрасываем фичи с 31 по 60, так как они являются полной копией фич с 1 по 30.
  - От id возьмем остаток от деления на 100 и создадим новые признаки, равенство остатка конкретному значению от 0 до 99.
  - Уберем лишние признаки. Во-первых id (вместо него уже добавили новые признаки), во-вторых временные признаки (предсказывать-то нужно все равно на будущее) и шифт(он не влияет на ответ, просто сообщает какой давности данные в фичах).
  - В качестве регрессора будем использовать XGBRegressor, параметры которого будем подбирать последовательно, оставляя те, при которых SMAPE получается наилучшим.
  - фича "f30"с некоторым почти постоянным коэффициентом равна значению 'y' для того же ID, но на $shift$ недель назад. Что можем оценим таким образом, а для остальной части test применим наш обученный на train XGBRegressor.
- Рассказ о подходах:
  - В начале делал все тоже самое только совсем не работал с признаками.
  - Затем убрал временные признаки и шифт - качество стало получше.
  - Затем преобразовал категориальный $item\_id$ в 100 булевых признаков, после чего качество еще немного выросло.
  - В итоге заметил эту читерскую закономерность со связью "f30"и 'y', и качество на тестовой выборке выросло в три раза (потому что оценить таким образом можно примерно $\frac{2}{3}$ данных).
- код в прикрепленном файле contest.ipynb
- Оценка качества:
  - Для оценки качества делал кросс-валидацию, причем так чтобы время всех элементов обучающей выборки было меньше времени элементов тестирующей, потому что некорректно исопльзовать данные "из будущего"для прогнозированния данных "из прошлого". Число сплитов выбрал равным пяти.

– Когда делал попытку без использования связи "f30"и "y средний SMAPE при кросс-валидации получился 29, а в leaderboard SMAPE было равно 27, то есть было близко к моему значению. Когда делал попытки с использованием этой связи SMAPE в leaderboard было примерно в три раза меньше, чем SMAPE при кросс-валидации на обучающей выборке. Это опять же из-за того, что где-то $\frac{2}{3}$ данных мы предсказываем почти точно. То есть SMAPE полученное мной на обучающей выборке неплохо коррелировала со SMAPE в leaderboard.