

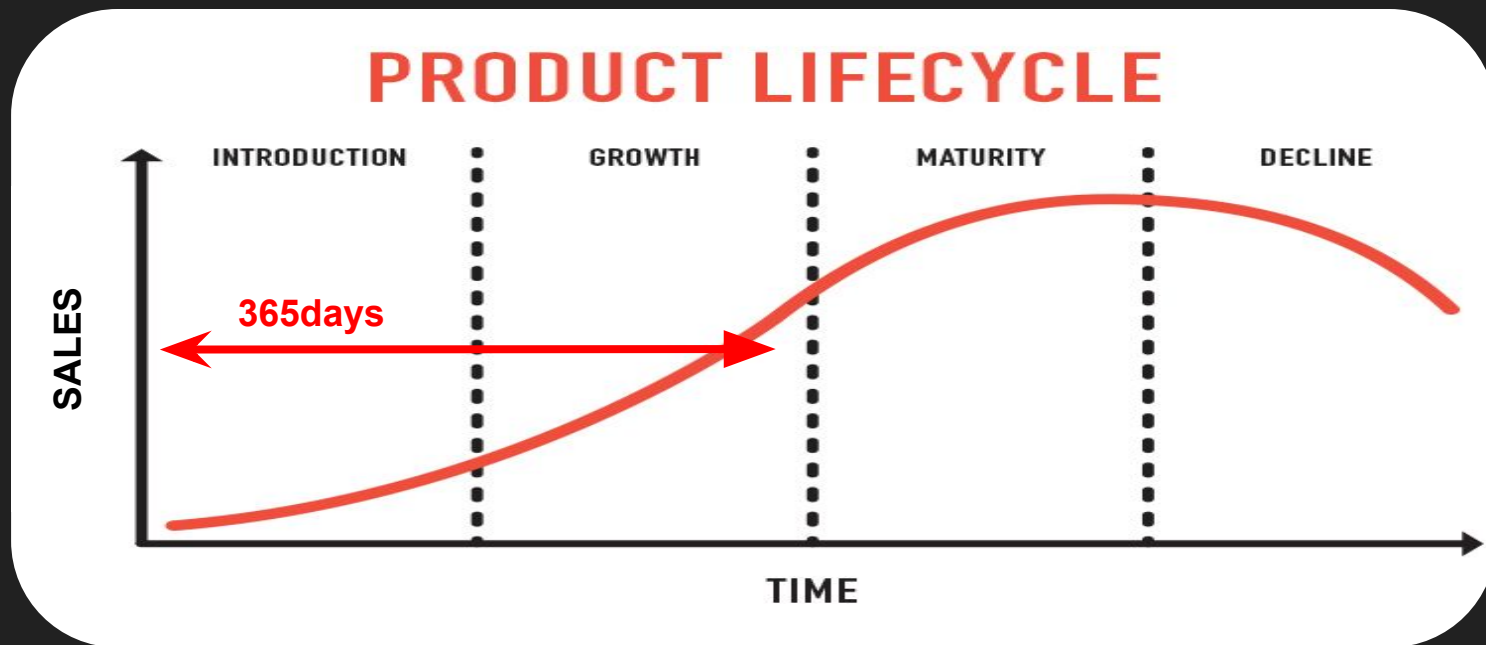
Долгосрочное прогнозирование спроса для нового ассортимента в ритейл

Команда № 11

Бочкарев Сергей Максимович
Копейкин Александр Дмитриевич
Федотов Кирилл Романович

“Новые товары”

Бизнес-правило: пара Товар|Магазин считается новой, пока с момента продаж Товара в Магазине прошло не более 28 дней.



Данные: построение обучающей выборки

Исходные данные: история продаж с 19 декабря 2016 года и атрибуты:

period_dt	Дата
location_id	Данные о местоположении
demand	Данные о спросе
price	Цена обычная и со скидкой
promo	Флаги на промоакции
authorization	Авторизованность магазина
num_consultant	Число консультантов

Предобработка данных

- 1) Соединили таблицы в один датасет
- 2) Удалили лишние символы
- 3) Удалили большую часть записей в трейне с пропущенными значениями
- 4) В тесте пропущенные заменили на среднее или медиану
- 5) Изменили формат времени

Построение фичей

Какие фичи были использованы:

- Изначальные данные
- ID типов продуктов и локаций магазина
- Флаги промоакций: сильная и маленькая скидка
- Изменение цены при скидке
- Лаговые переменные

Алгоритмы ML

Какие алгоритмы использовали:

- LGBM Regressor с лаговыми фичами
- Иерархическая модель: Уровни магазинов, товаров, относительный объем скидок.
- Ансамбль: voting ensemble с наивной моделью, предсказанием среднего и катбустом

История Submit и Результаты

Критерий ошибки MAE

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Основные улучшения:

Шаг	Private Score
Добавляли лаг, равный 28 или 35	0,70
Добавляли флаги на промоакции	0,65
Подобрали $lr=0,05$	0,48
Использовали иерархический временной ряд ETNA	0,37
Использовали ансамбль	0,29
<u>Бенчмарк</u>	<u>0.275</u>

Выводы

- Библиотека ETNA показала лучший результат по сравнению с LGBM
- Для лучшего качества следовало обучать по агрегированным группам товаров
- Нужно было пробовать строить Zero-inflated модель
- Следовало использовать перцентили, а также агрегированные данные о продажах по неделям в качестве фичей