

Тренировка по машинному обучению

Определение категорий покупок в чеке









Владимир Семеновых

Олег Акимов

26.05.2018

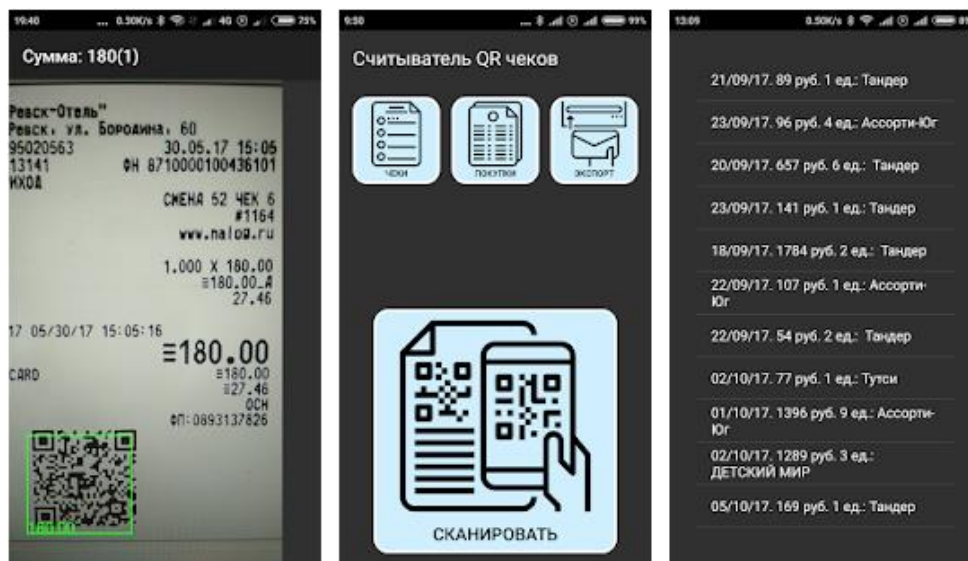
Соревнование OpenDataScience: Категоризация покупок

Private Leaderboard

#	Δ pub	Team Name	Kernel	Team Members	Score ?
1	—	Семеновых Владимир Николае...			0.29960
2	▲ 2	Акимов Олег Александрович			0.30161
3	—	Artgor [ods_cheater]			0.31077
4	▼ 2	Титаевская Наталья Анатольев...			0.31194
5	▲ 3	Коломиец Сергей Иванович			0.31665
6	▼ 1	Самсонов Максим			0.32768
7	—	off rating			0.33176
		Продвинутое линейное решение			0.34560

Описание задачи

Недавно на чеках появились QR коды. Пока еще не все с этим знакомы, но по информации из этого кода можно получить полное содержание чека. Это дает возможность вести расходы, учитывая каждый отдельный товар, включая расходы, сделанные наличными. Как следствие наличия полной информации, можно анализировать изменения характера расходов и инфляцию по собственной продуктовой корзине. Названия товаров не стандартизованы: у одного товара в разных магазинах существенно отличаются названия; отдельные слова могут сокращаться; названия могут содержать опечатки. В магазины постоянно добавляются новые товары. Это делает простое составление каталога всех товаров с категориями нереалистичным. Задача, которую предлагается решить — это разбиение всех покупок чека по небольшому набору понятных человеку категорий.



Примеры данных

check_id		name	category	price	count
0	0	*3479755 TRUF.Конф.кр.корп.гл.вк.шок180г	Чай и сладкое	49.00	2.000
1	0	3408392 ECONTA Мешки д/мусора 30л 30шт	Для дома	21.00	1.000
2	0	3260497 ЯШКИНО Рулет С ВАР.СГУЩ. 200г	Чай и сладкое	39.00	1.000
3	0	3300573 Пакет ПЯТЕРОЧКА 65x40см	Упаковка	4.00	1.000
4	0	3413607 ЗЕР/СЕЛ.Сухари с изюмом 250г	Чай и сладкое	35.00	1.000
5	0	3221388 ШАРЛ.Печенье вафел.рассыпч.225г	Чай и сладкое	38.00	1.000
6	0	*97452 ПРОСТ.Кефир 3,2% 930г	Молочка	55.00	1.000
7	0	57575 MILFORD Зам.сахара доз. 650таб	Бакалея	119.00	1.000
8	0	29880 ПИСК.Ацидоб.2.2%сл.пюр-пак0.5л	Молочка	34.00	1.000
9	1	ШОКОЛАДНОЕ ЯЙЦО 20Г КИНДЕР СЮРПРИЗ ФЕРРЕ	Дети	49.00	1.000

check_id		shop_name	datetime	sum
0	3947	Тандер	2018-02-03 12:40:00	222.00
1	3948	Нытва-Фарм	2018-02-02 10:59:00	137.00
2	3949	Агроторг	2018-02-03 12:33:00	160.00
3	3950	NaN	2018-01-31 15:38:00	265.50
4	3951	Бэст Прайс	2018-02-03 11:34:00	413.00
5	3953	Лента	2018-02-01 21:57:00	728.00
6	3954	Лента	2018-02-01 21:59:00	796.00
7	3955	Кронар	2018-02-03 14:45:00	168.00
8	3956	АТАК	2018-02-03 15:31:00	19.00
9	3957	Пятью пять	2018-02-03 15:21:00	811.00

Описание данных

Тренировочная выборка (13682 записи из чеков): ID чека, Название товара из чека, Категория товара, Цена товара, Количество товара

Файл с чеками, из которого взяты позиции для тренировочной выборки (2042 чека): ID чека, Название магазина, Время покупки, Суммарная стоимость товаров

Тестовая выборка (3000 записей из чеков)

Файл с чеками для тестовой выборки (933 чека)

3 дополнительных каталога с товарами (суммарно 250745 товарных позиций): Описание товара, Категория

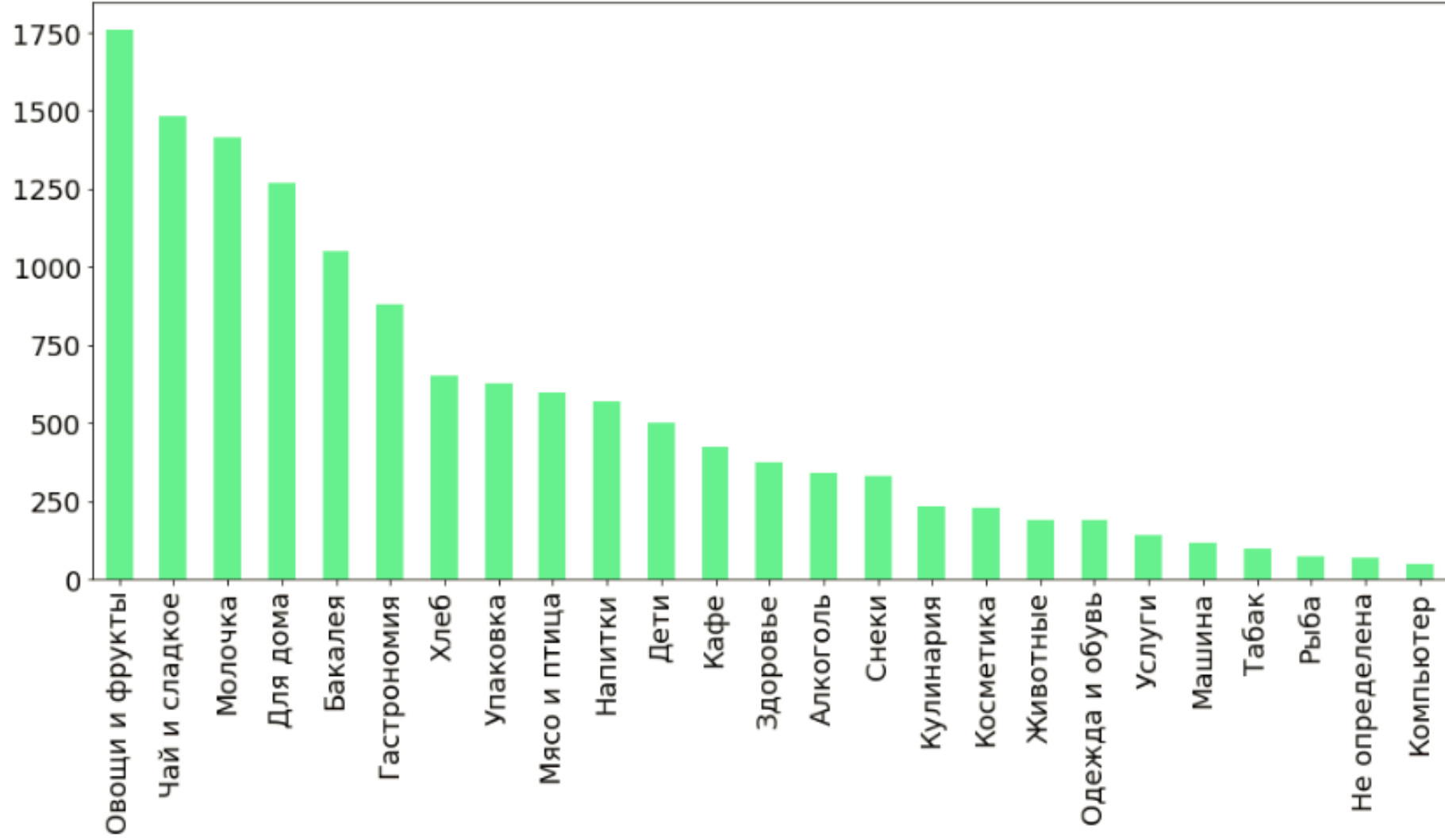
Прайс лист аптеки (57850 позиций)

Список блюд и цен из кафе (151144 позиции)

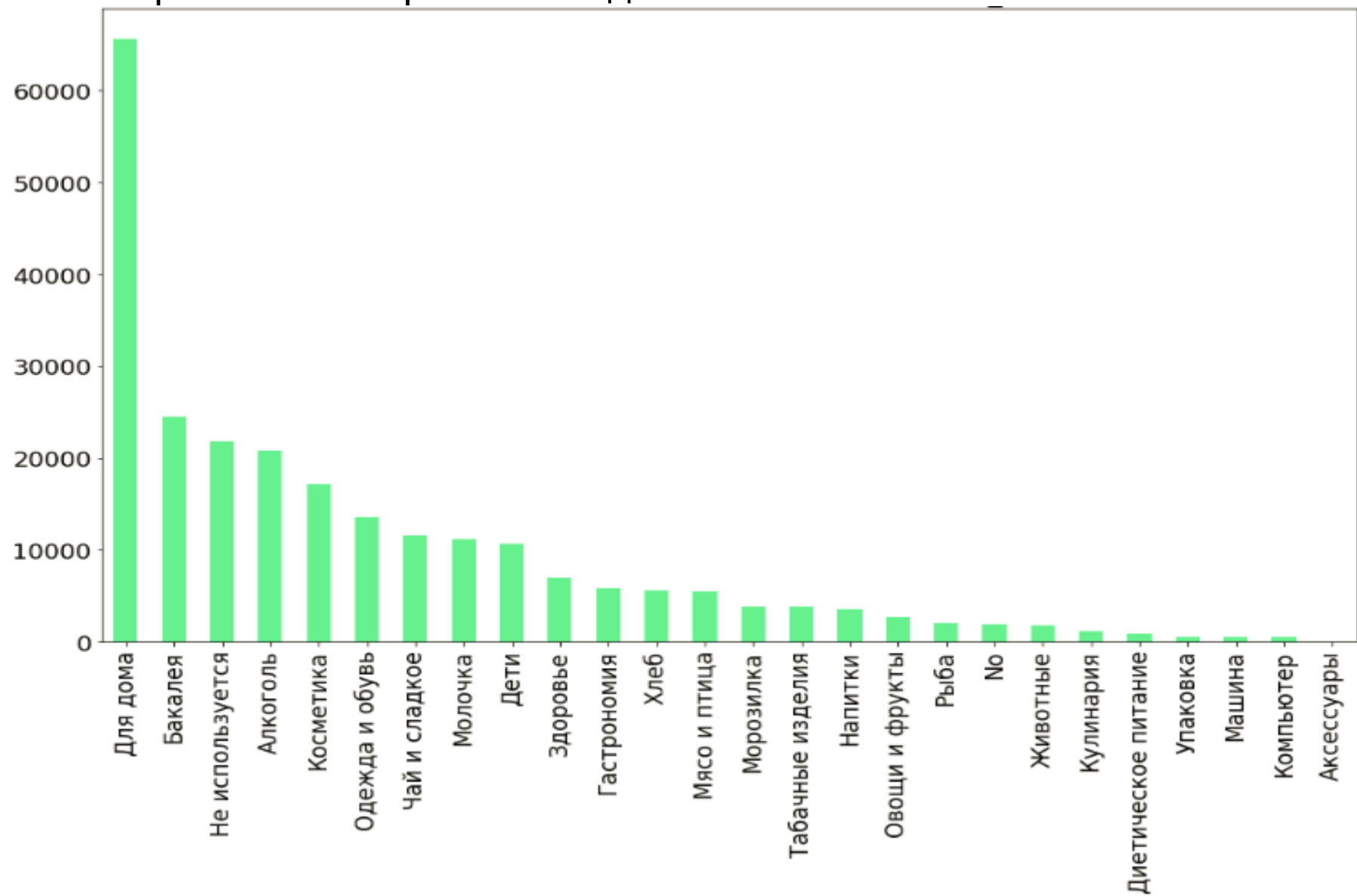
Прайс-лист строительного магазина (22203 позиции)

Прайс-лист детского магазина (24717 позиций)

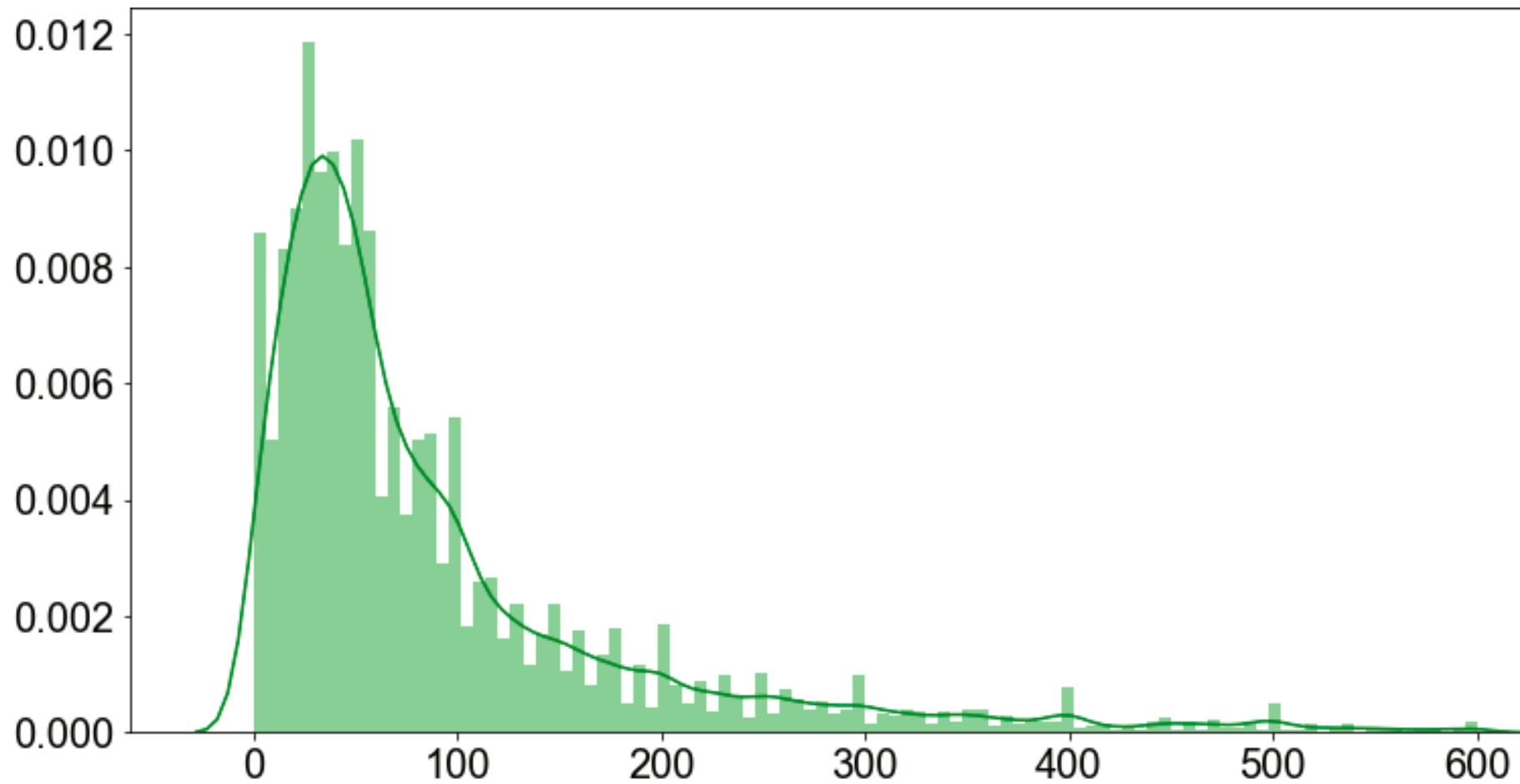
Распределение товаров по категориям в TRAIN



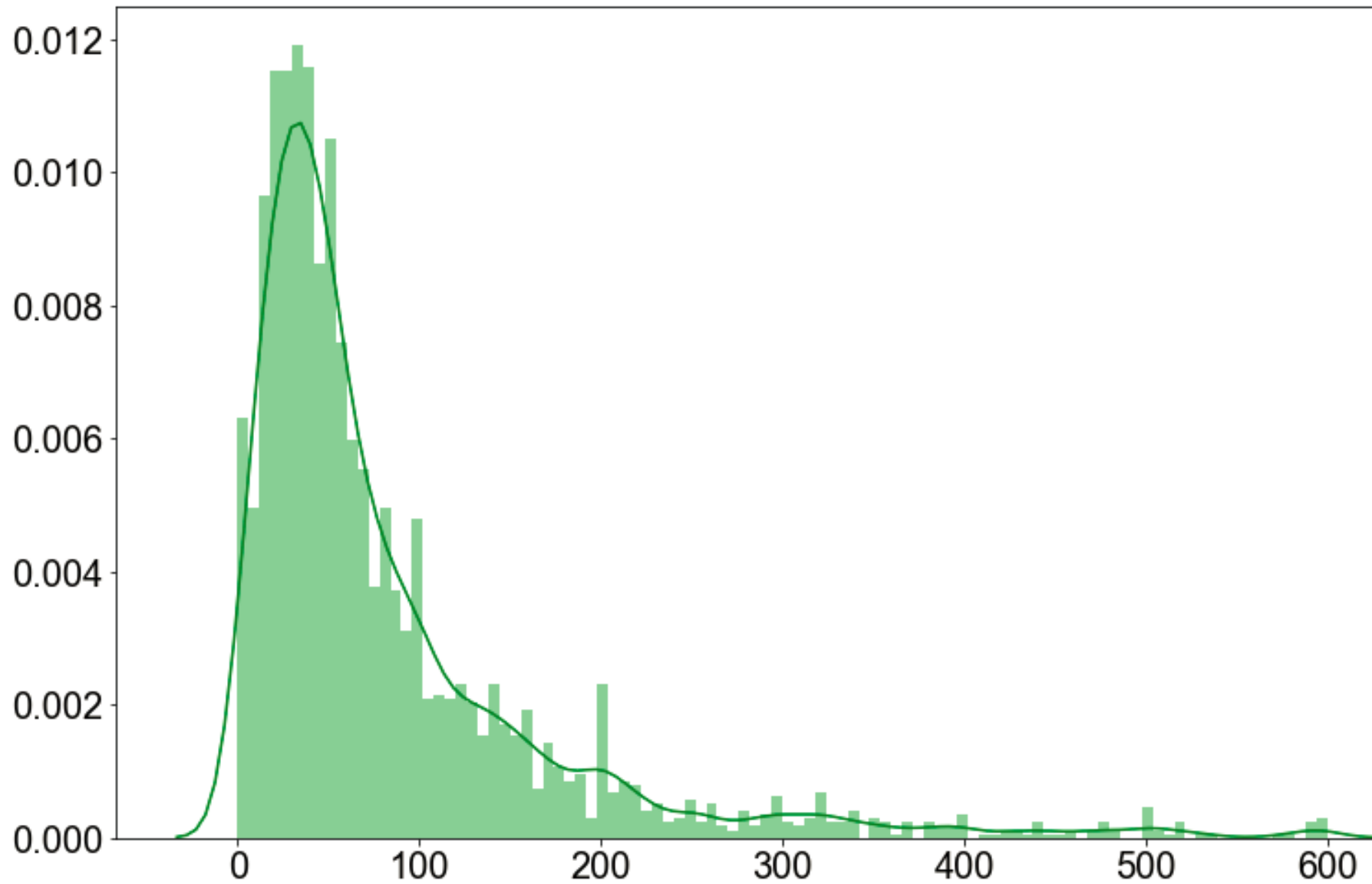
Распределение товаров по категориям в 3-х дополнительных каталогах



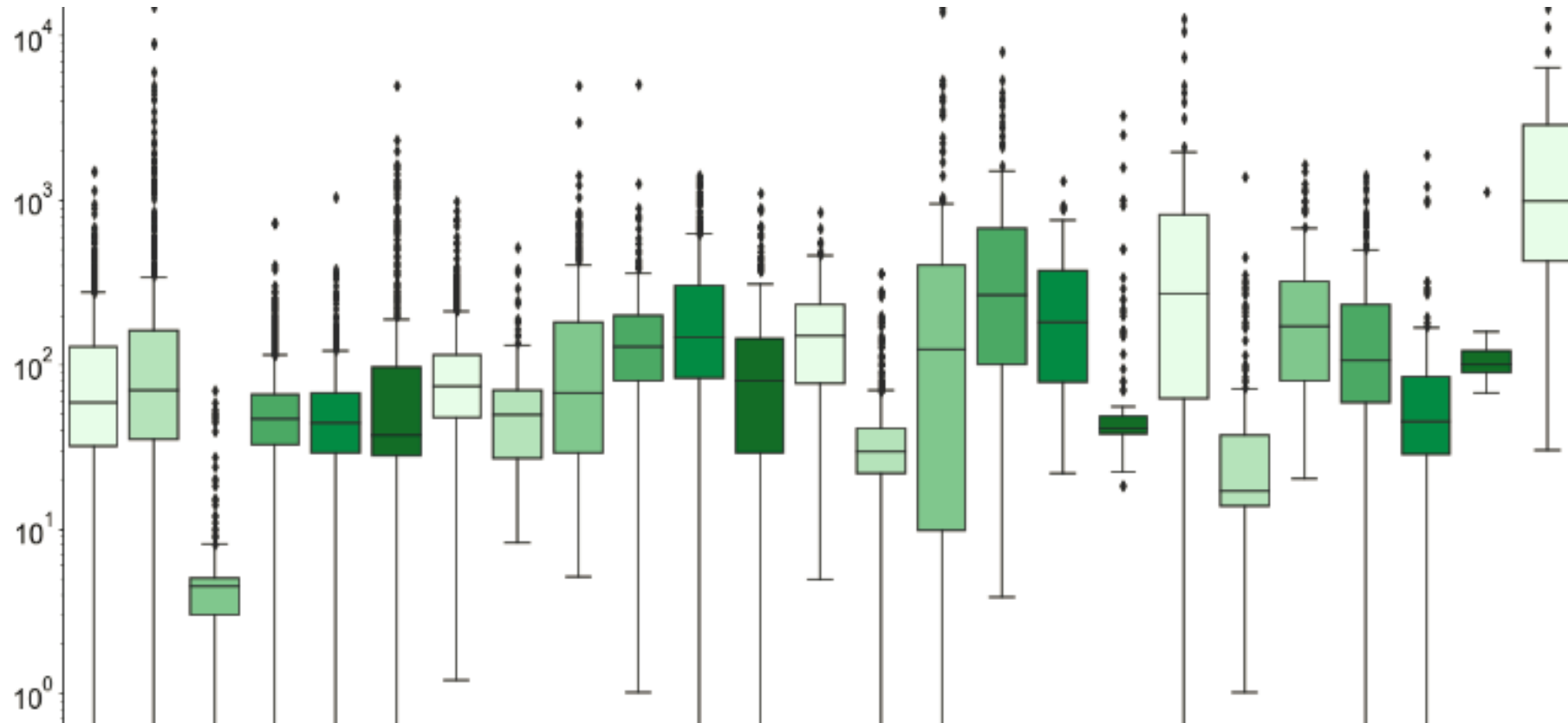
Распределение цен в TRAIN (до 600 рублей)



Распределение цен в TEST (до 600 рублей)



Распределение цен по категориям в TRAIN



Распределение по количеству купленного товара

Для некоторых категорий значения поля «Количество» всегда целое:

Косметика, Услуги, Упаковка, Табак, Дети, Одежда и обувь, Для дома, Животные, Напитки, Здоровье, Компьютер

Для некоторых категорий значение поля «Количество» в большинстве случаев (минимум 90%) целое:

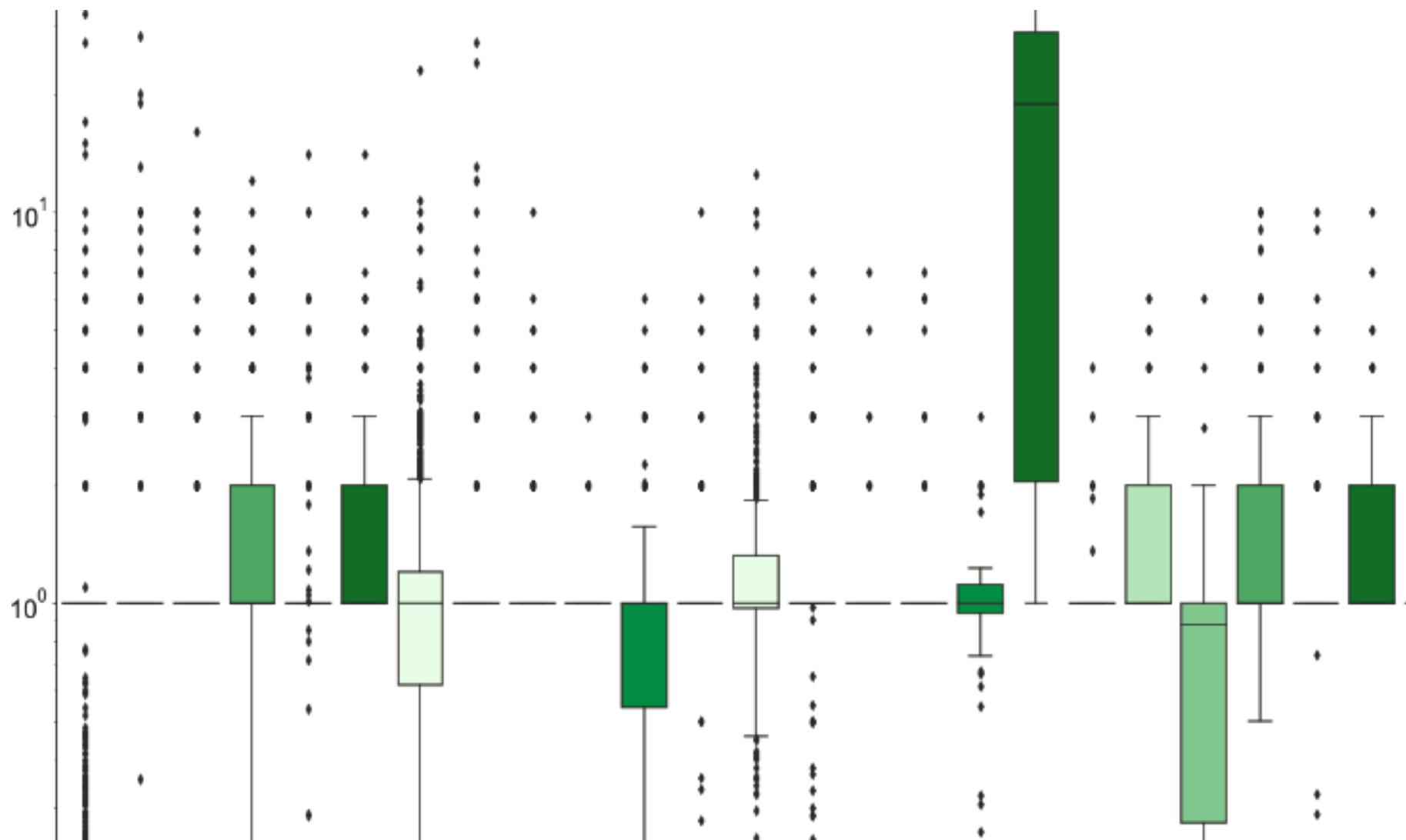
Молочка, Бакалея, Кафе, Алкоголь, Снеки, Не определена, Хлеб, Чай и сладкое

Для некоторых категорий значение поля «Количество» целое в 45-70 % случаев:

Гастрономия, Рыба, Мясо и птица, Машина, Кулинария

И только для категории «Овощи и фрукты» количество чаще дробное

Распределение по количеству купленного товара



Сопутствующие категории

Товары из некоторых категорий часто могут встречаться в одном чеке, например:

- Овощи – Мясо
- Молоко – Мясо
- Чай и сладкое – Снеки
- Молоко – Гастрономия и т.д.

Некоторые категории никогда не встречались в одном чеке в TRAIN, например:

- Для дома – Кафе
- Здоровье – Кафе
- Здоровье – Компьютер
- Рыба – Компьютер и т.д.

Подход к решению

XGBoost над 7-ю группами признаков:

- 1) Logit на двух «мешках слов» по основной выборке**
- 2) Logit на «мешке слов» по 3-м дополнительным каталогам**
- 3) Multinomial Naive Bayes на «мешке слов» на основном каталоге**
- 4) Multinomial Naive Bayes на «мешке слов» по 3-м дополнительным каталогам**
- 5) Усредненные вероятности для всех товаров каждого чека, посчитанные на основной выборке**
- 6) Усредненные вероятности для всех товаров каждого чека, посчитанные на 3-х дополнительных каталогах**
- 7) Суммарное количество товаров по каждому чеку**

Загрузка данных

1) Т.к. в тесте отсутствуют товары дороже 5000 рублей, из тренировочного сета удалены товары дороже 5000 рублей, чтобы алгоритм не настраивался на выбросы.

2) Соединяем 3 дополнительных датасета в один (некоторые категории нужно переименовать, чтобы они сошлись, например, Табак)

3) Из дополнительного датасета удаляем часть категорий, которые вносили шум:

- Пустая категория
- Аксессуары
- No
- Не используется

Первые 4 группы признаков

1) Logit на основной выборке:

- «Мешок слов» (n-grams 1-2, min 5, word) на исходной строке + результаты SnowballStemmer
- «Мешок слов» (n-grams 2-5, min 3, char)
- OneHotEncoder: магазин (по которым есть не менее 10 чеков), ценовая категория, час, день недели
- Цена и Количество продукта

2) Logit на 3-х дополнительных каталогах:

- «Мешок слов» (n-grams 2-5, min 3, char)

3) Multinomial Naive Bayes на основной выборке:

- «Мешок слов» («из коробки»)

4) Multinomial Naive Bayes на 3-х дополнительных каталогах :

- «Мешок слов» («из коробки»)

Последние 3 группы признаков

1) Усреднение вероятностей всех товаров по каждому чеку.

Используются вероятности посчитанные для 1-й группы признаков (основная выборка)

2) Усреднение вероятностей всех товаров по каждому чеку.

Используются вероятности посчитанные для 2-й группы признаков (3 дополнительных каталога)

3) Количество товаров в чеке

Подход к решению

XGBoost над 7-ю группами признаков:

- 1) Logit на двух «мешках слов» по основной выборке**
- 2) Logit на «мешке слов» по 3-м дополнительным каталогам**
- 3) Multinomial Naive Bayes на «мешке слов» на основном каталоге**
- 4) Multinomial Naive Bayes на «мешке слов» по 3-м дополнительным каталогам**
- 5) Усредненные вероятности для всех товаров каждого чека, посчитанные на основной выборке**
- 6) Усредненные вероятности для всех товаров каждого чека, посчитанные на 3-х дополнительных каталогах**
- 7) Суммарное количество товаров по каждому чеку**

Подход к решению 2

Перепроцессинг:

- Замена латинских символов русскими
- Выделение кода товара из названия позиции в чеке
- Выделение весовых товаров
- Генерация всевозможных агрегатов по магазинам и чекам
- Синусное и косинусное преобразование времени
- Построение TF-IDF матрицы посимвольно без ограничения на чисто признаков
- Генерация KNN фичей
- Нормализация данных

Подход к решению 2

1. Моделирование:

- 1) Logit на основной выборке с тонкой настройкой гиперпараметров:
 - TF-IDF
 - Табличные + сгенерированные фичи
 - KNN фичи
- 2) Logit на каталогах:
 - TF-IDF
- 3) XGBoost с оооооочень долгим подбором гиперпараметров посредством hyperopt на кросс-валидации:
 - Усредненные предсказания Logit
 - Сгенерированные фичи
- 4) ПРОФИТ!

Спасибо за внимание!



Open
Data
Science

