ИИ в бизнесе

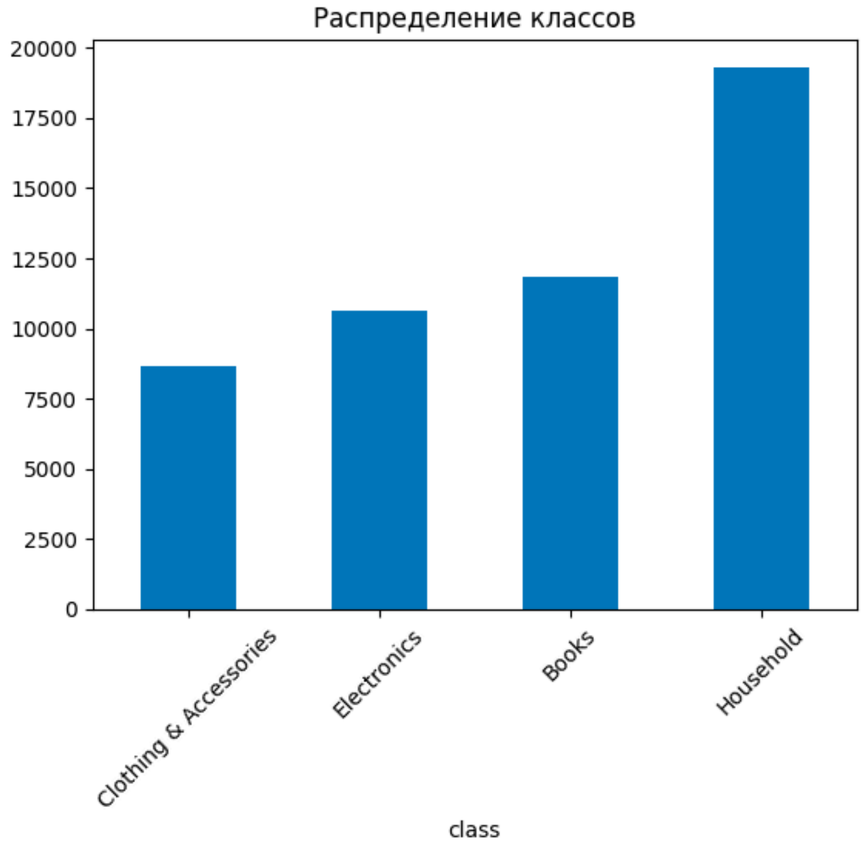
Классификация описания товаров для маректплейса

Курушин Федор, Горбунов Максим

EDA

Классы

- Данные - описание товаров
- Задача: классификация товара по одной из категорий



EDA

Распределения

- Типичное для настоящих данных логнормальное распределение

EDA

Распределения

- Типичное для настоящих данных логнормальное распределение

Data processing

Text processing

- Обрезка по кол-ву слов 2 % и 97 % перцентилями
- Приведение к нижнему регистру
- Удаление пунктуации
- Удаление стоп-слов
- Лемматизация
- Label encoding

Data processing

Sampling

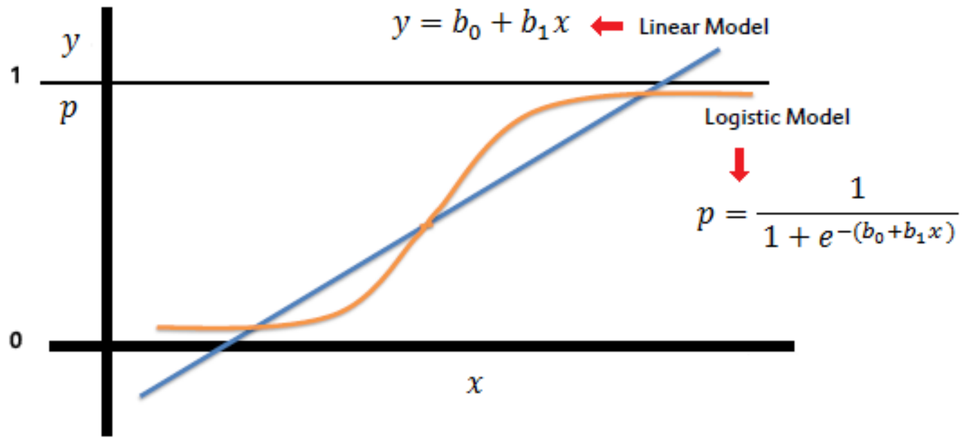
- Random under sampling
- Train test split shuffled, stratified

EDA

- Данные - описание товаров
- Задача: классификация товара по одной из категорий

Logistic Regression

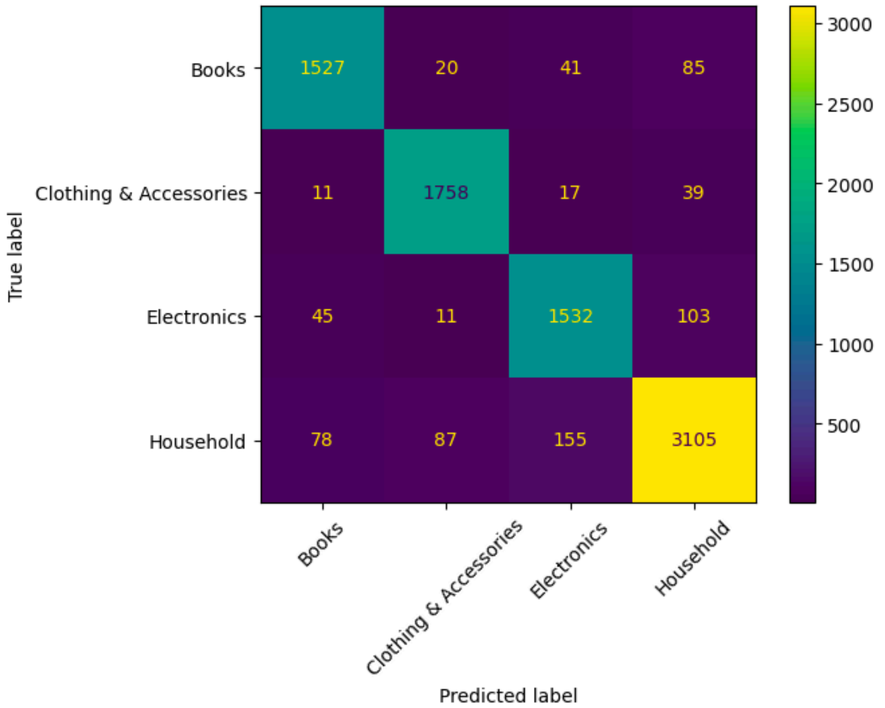
- Time frequency - inverse document frequency: **tf-idf(t, d) = tf(t, d) * idf(t)**
- **LogReg**



Logistic Regression

Результат с Under Sampling

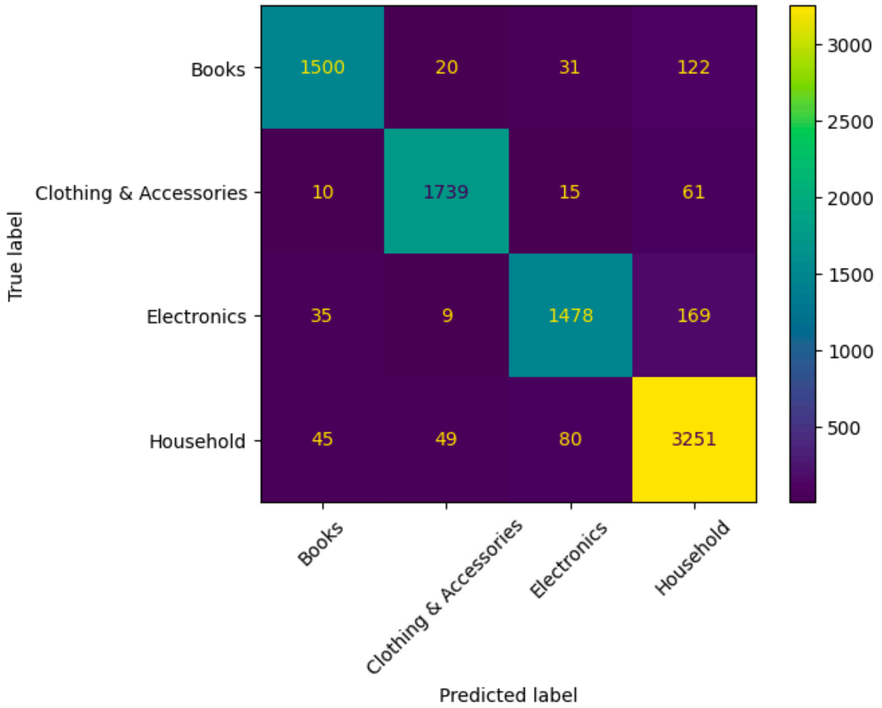
	precision	recall	f1-score	support
Books	0.91	0.92	0.92	1669
Clothing & Accessories	0.97	0.94	0.95	1879
Electronics	0.91	0.88	0.89	1759
Household	0.90	0.94	0.92	3307
accuracy			**0.92**	8614
macro avg	0.92	0.92	**0.92**	8614
weighted avg	0.92	0.92	**0.92**	8614



Logistic Regression

Результат без семплирования

	precision	recall	f1-score	support
Books	0.90	0.94	0.92	1590
Clothing & Accessories	0.95	0.96	0.95	1817
Electronics	0.87	0.92	0.90	1604
Household	0.95	0.90	0.93	3603
accuracy			**0.93**	8614
macro avg	0.92	0.93	**0.92**	8614
weighted avg	0.93	0.93	**0.93**	8614



Logistic Regression

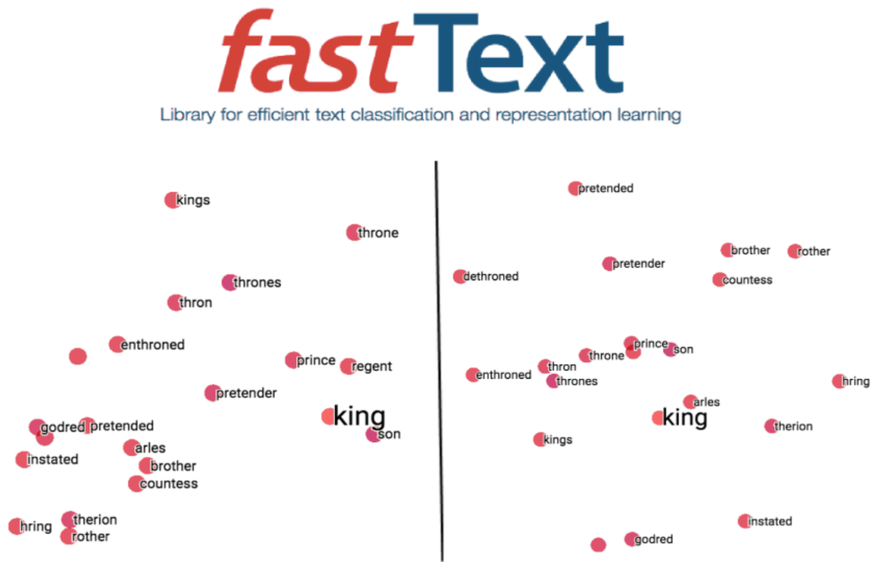
Вывод:

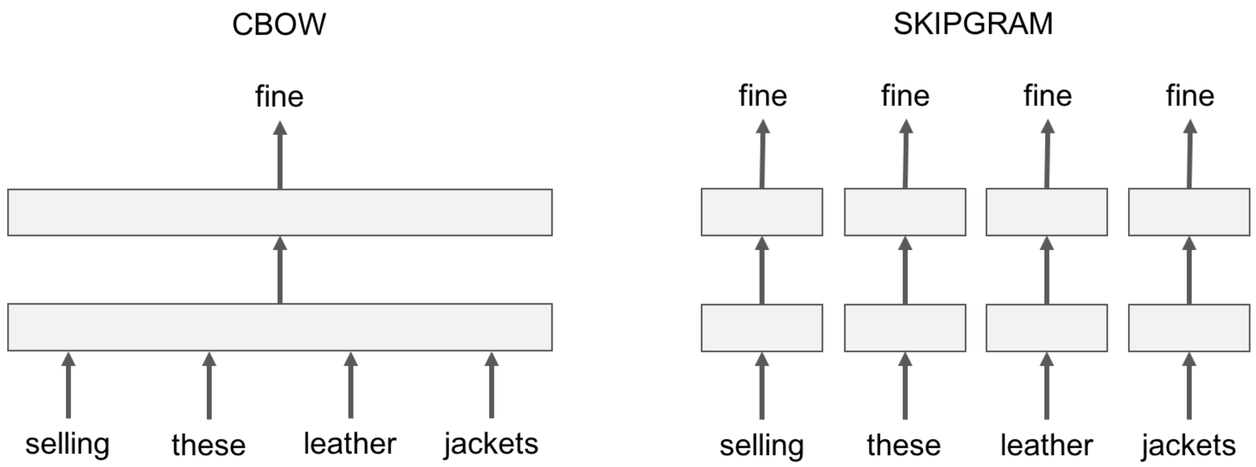
- Результат с/без сэмплирования достаточно схож, но как можно видеть отсутствие сэмплирование дает больший f1.
- Прореживание доминирующего класса - это потеря данных.
- По сути это делается для взвешивания, и взвешивание во многих моделях LogReg не исключение можно делать напрямую, не меняя данные. Расплата за взвешивание - это модель, скор в которой теперь стал ещё дальше от вероятности. Если нам нужно предсказывать вероятности, стратегия со взвешиванием на 100% ухудшит наши результаты.

FastText

Фреймворк

- Для построение эмбеддингов слов при помощи модели word2vec, на базе алгоритмов cbow и skipgram



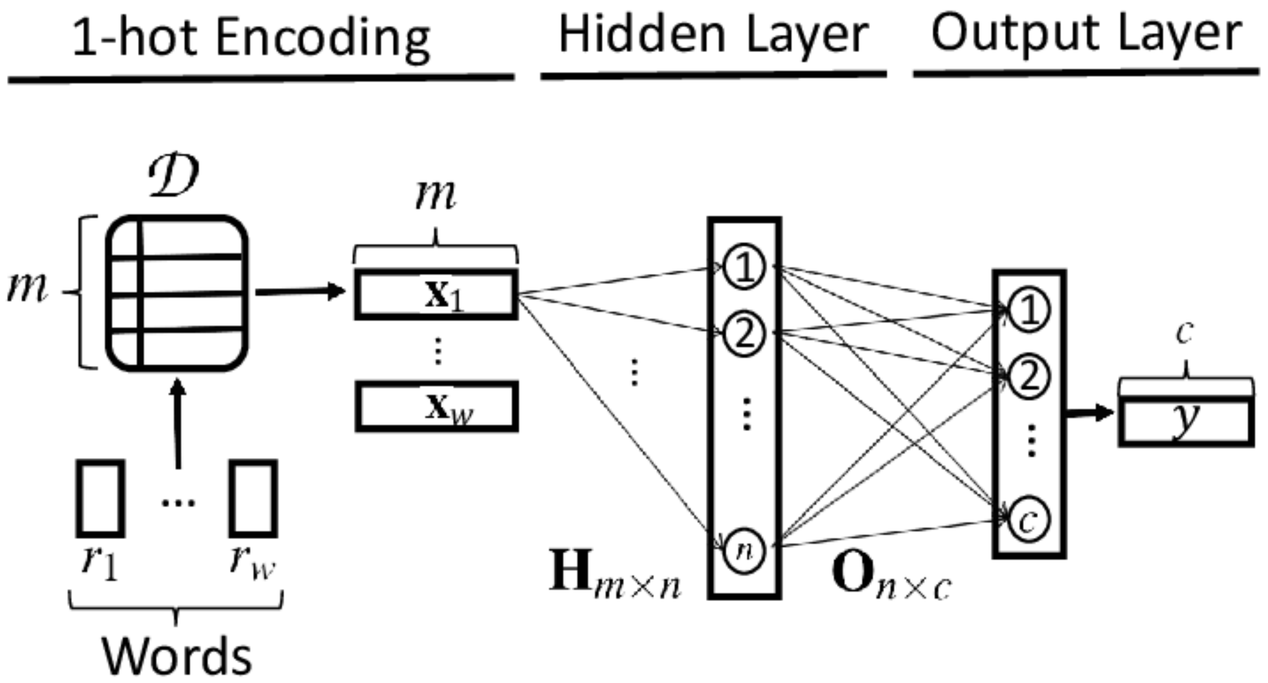


I am selling these fine leather jackets

FastText

Фреймворк

- Классификации текстов



FastText

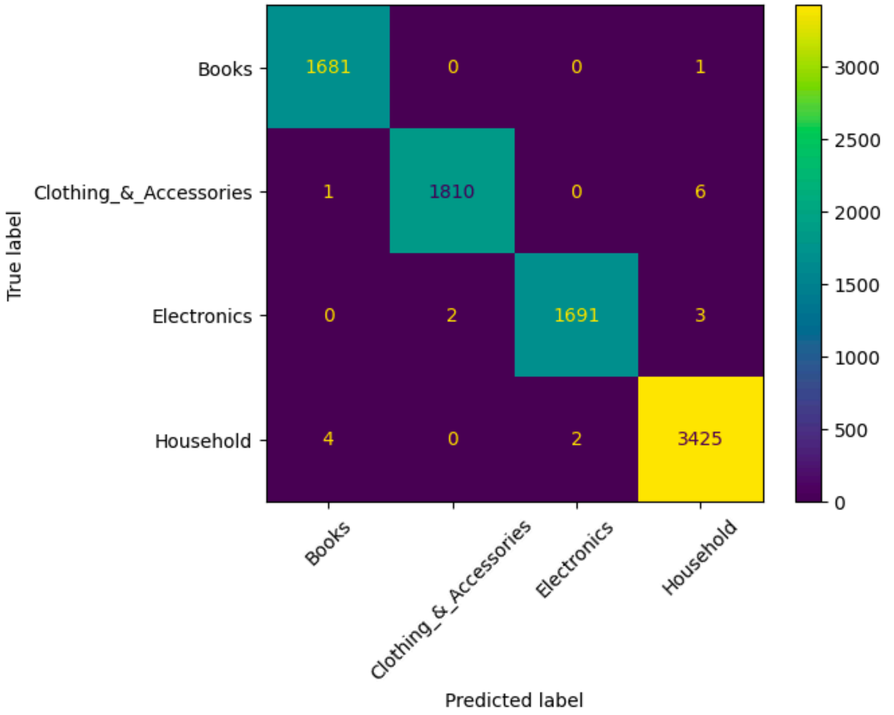
Обучение

- Обертка пайплайна обработки в python скрипт
- Перебор различных вариаций гиперпараметров: lr, epoch
- Автотюнинг модели

FastText

Обучение

	precision	recall	f1-score	support
Books	1.00	1.00	1.00	1682
Clothing_&_Accessories	1.00	1.00	1.00	1817
Electronics	1.00	1.00	1.00	1696
Household	1.00	1.00	1.00	3431
accuracy			1.00	8626
macro avg	1.00	1.00	1.00	8626
weighted avg	1.00	1.00	1.00	8626



Выводы:

- ReSampling может не приносить желаемого результата, лучше использовать взвешенные метрики. Если же хочется предсказывать редкий класс, то можно использовать более низкий threshold.
- **FastText** и **LogReg** являются двумя популярными инструментами методами машинного обучения для задачи классификации текста. Тем не менее, fasttext имеет более сложную архитектуру, учитывает контекст, и использует word-embeddings перед обучением модели классификации.
- Благодаря чему, fasttext в этой задаче дает гораздо более высокий результат