Итоговое задание по программе профессиональной переподготовки

Data Science: обработка естественного языка

Рауф Мамедов
DSN-CP-4

Постановка задачи для машинного обучения

1

Описание проекта

Исходная задача:

Построение классификатора для предсказания ассортиментной категории магазина по названию и описанию товара. Набор данных для анализа хранится по адресу: https://www.kaggle.com/datasets/saurabhshahane/ecommerce-text-classification

Актуальность задачи, ее место в предметной области:

Задача относится к области тематического моделирования и её решение позволяет проводить автоматическую классификацию товара по описанию, снижая затраты времени на занесения товара в соответствующий раздел ассортиментного каталога. По своей сути задача относится к задачам мультикласоовой классификации

Целевая метрика:

Ввиду сбалансированности классов, в качестве целевой метрики была выбрана точность классификации.

Анализ данных

2

Анализ данных

Откуда данные? Какие аналогичные решения существуют?

1. Для анализа использованы данные соревнования на Kaggle
2. Представлены данные в качестве документов относящиеся к четырём категориям товаров: «Домоводство», «Книги», «Одежда и аксессуары» и «Электроника». Распределение документов по категориям приведено на диаграмме справа
3. Полученый датасет, содержит текстовое описание товара и разметку класса.
4. Корпус представлен 50424 документами, среди который 44.9% составляют дубликаты.
5. Корпус документов содержит единственный пропуск, Обработка данных заключалась в удалении стоп-слов п также не-ASCII символов и подготовке набора признаков содержащих как TFI-DF лемматизированных слов, так и мешок символов документов, которые также могут иметь различную встречаемость в документах (например числовых значений может быть больше или меньше в документах различных тематик). Исследование описательной статистики показало, что после удаления дубликата сбалансированность классов не нарушается, а следовательно мы можем использовать более компактную очищенную версию датасета. После очистки от дубликатов, датасет содержит 27802 документа.

Методика реализации

3

Каким образом решали задачу?

Какие способы реализации были использованы?

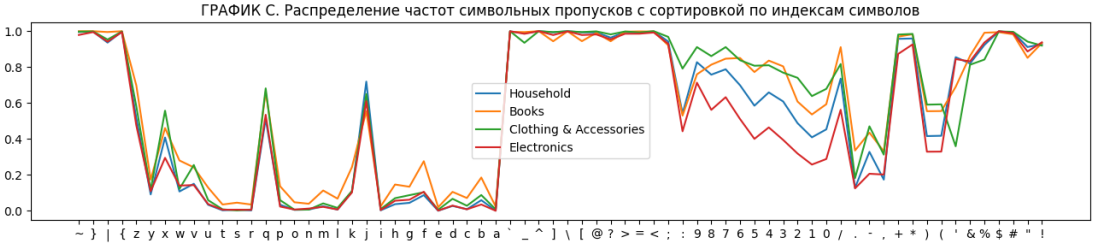
1. После подготовки и преобразования данных, были проведены следующие эксперименты с различными алгоритмами машинного обучения (линейные модели, деревья решений, ансамблевые методы):

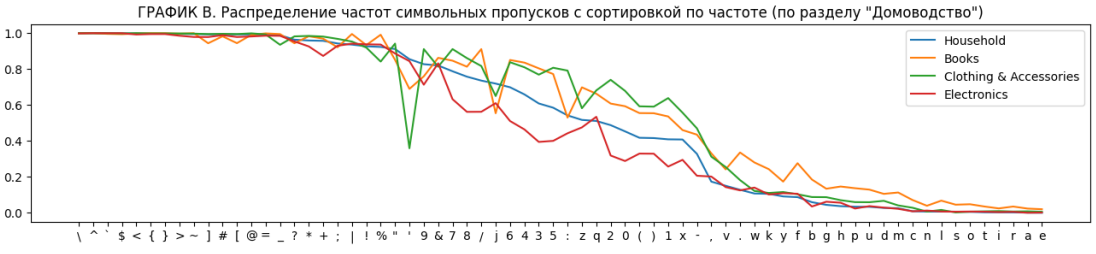
A. Эксперименты на признаках встречаемости символов

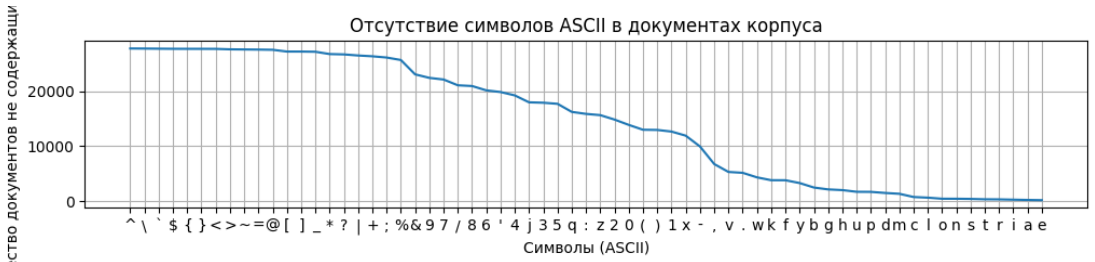
B. Эксперименты на признаках символов объединённых с матрицей TFI-DF

2. Были использованы описания алгоритмов в руководствах пользователя библиотек SCI-Kit Learning и XGBoost.
3. В результате экспериментов была выбрана модель превосходящая по точности предсказания иные модели при значениях гиперпараметров заданных по умолчанию. Далее было осуществлена оптимизация молели настройкой её гиперпараметров осуществлявшаяся с помощью библиотеки Optuna.

Модели машинного обучения на признаках символов







Проведено исследование данных с определением частоты встречаемости символов в документах различных тематик. На графике различимы 3 основные области, в результате анализа которых мы приходим к интересным выводам:

А) Низковариативная область редкой встречаемости символов (спецсимволы и арифметические действия, а также скобки кроме круглых). Интересно что в эту группу попадает знак сложения, который было бы ожидаемо видеть среди часто встречающихся символов.

Б) Область высокой вариативности распределения отсутствий символов. Эту область образуют буквы (x, q и j, для них фиксируются аномально высокие, нехарактерные для букв частоты пропусков), знаки препинания, а также все цифры. Интересно что очень контрастна вариативность распределения цифр, при сохранении трендов изменения частот встречаемости. Чаще всего цифры ожидаемо встречаются в документах с описаниями электроники и реже всего в описаниях одежды и аксессуаров. Также очень интересно аномально низком количества пропусков симовла "одинарные кавычки" для раздела Одежда и аксессуары, возможно она отражает имена собственные в притяжательном падеже.

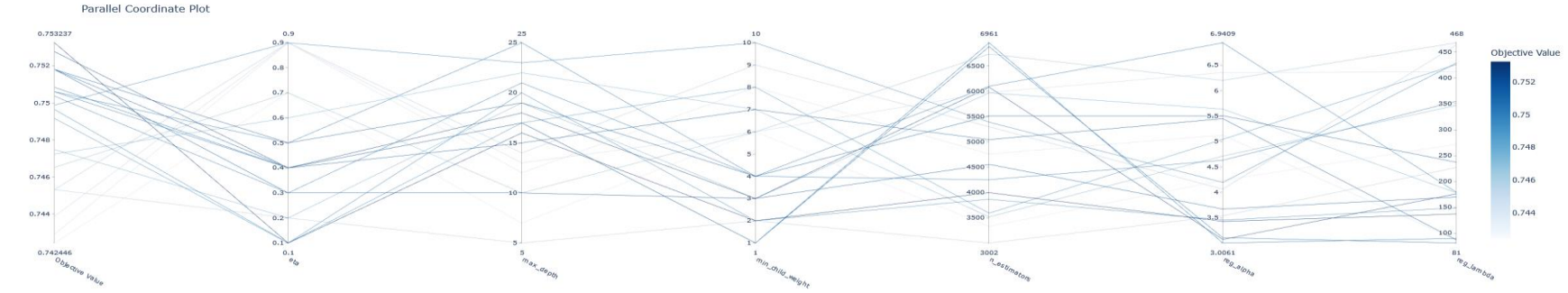
В) Область стационарности (высокая встречаемость символов) представленная только буквами, здесь наблюдается малая вариативность частот по буквам. Вариативность по классам тоже незкая, за исключением более высокого количества пропущеннных букв для раздела Книги.

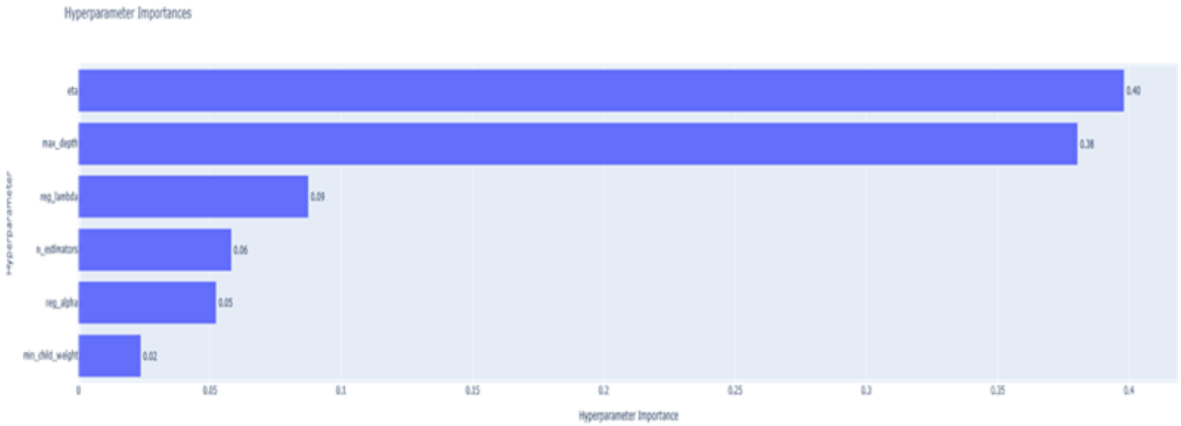
Модели машинного обучения на признаках символов

Используемая модель	train	val
XGBoost	0.919	0.740
Extra Trees Classifier	1.000	0.718
Random Forest	1.000	0.715
Gradient Boosting Classifier	0.713	0.677
Bagging Classifier	0.988	0.677
Logistic Regression	0.612	0.606
Decision Tree	1.000	0.574
Linear SVC	0.556	0.558
SGD Classifier	0.552	0.548
Ridge Classifier	0.528	0.524
Perceptron	0.522	0.522
Multinomial NB	0.499	0.500
Passive Aggressive Classifier	0.447	0.438
Gaussian NB	0.391	0.385

На признаках подсчёта признаков было построено несколько классификаторов линейного, древесного и ансамблевого типа, а также классификатор экстремального градиентного спуска. Последняя из моделей хорошо себя показала при использовании с параметрами по умолчанию. Оптимизация проводилась с помощью библиотеки Optuna.

Модели машинного обучения на признаках символов





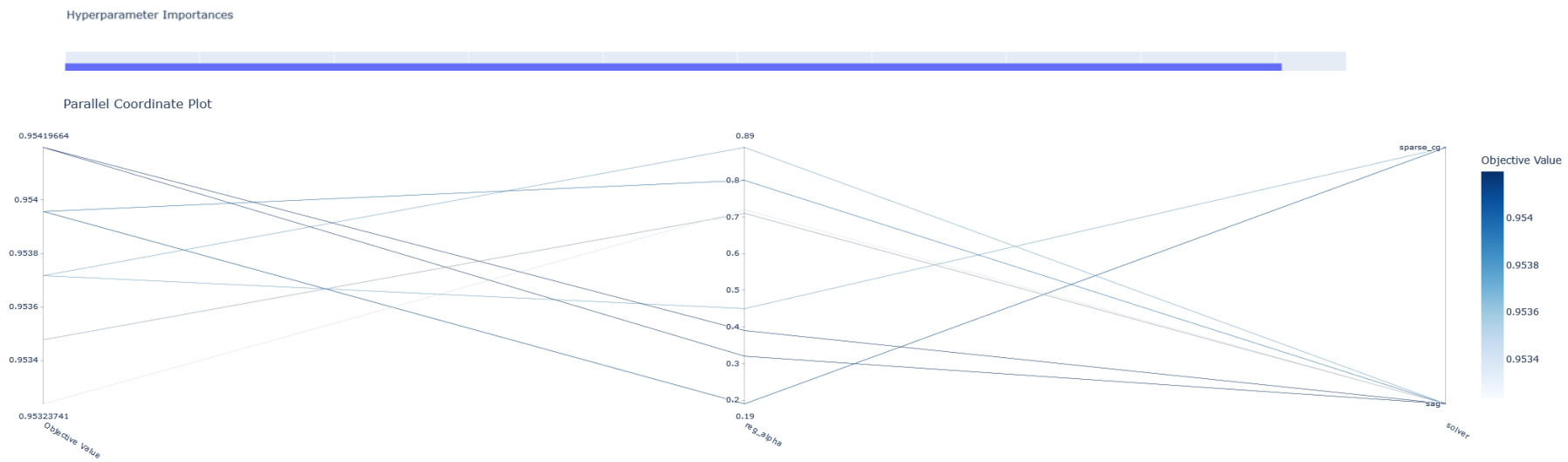
Модели машинного обучения на всех признаках

Какие способы реализации были использованы?

Используемая модель	Обучение	Валидация
Ridge Classifier	0.987	0.954
Linear SVC	0.971	0.944
XGBoost	0.978	0.931
Extra Trees Classifier	1.000	0.931
Passive Aggressive Classifier	0.925	0.914
Random Forest	1.000	0.913
Gradient Boosting Classifier	0.919	0.902
Bagging Classifier	0.991	0.896
Decision Tree	1.000	0.865
Multinomial NB	0.762	0.741
Logistic Regression	0.620	0.614
SGD Classifier	0.562	0.560
Perceptron	0.419	0.417

- Лучшими моделями оказались линейные модели (гребневой классификации и линейный классификатор на методе опорных векторов), а также метод экстремальных градиентов, и Классификатор дополнительных деревьев, который при этом сильно переобучился, и возможно имеет перспективы улучшения за счёт подбора гиперпараметров.
- Также неплохо отработали Пассивно-агрессивный линейный классификатор и алгоритм случайного леса
- Остальные ансамблевые и линейные классификаторы отработали плохо. Логистическая регрессия на полном наборе данных сработала значительно хуже чем XGBoost на малом наборе признаков (символы).
- Была предпринята попытка улучшить точности предсказания алгоритма Ridge Classifier путём настройки гиперпараметров. Небольшой рост метрики точности свидетельствует об успешной работе алгоритма на имеющемся датасете без существенного изменеия значения гиперпараметров заданных по умолчанию.

Модели машинного обучения на всех признаках



Итоги обучения

4

Модели машинного обучения на признаках символов

Оптимальные настройки гиперпараметров модели

index	Params
alpha	0.24
class weight	null
Copy X	true
Fit intercept	true
Max iter	null
positive	false
Random state	null
solver	sag
tol	0.0001

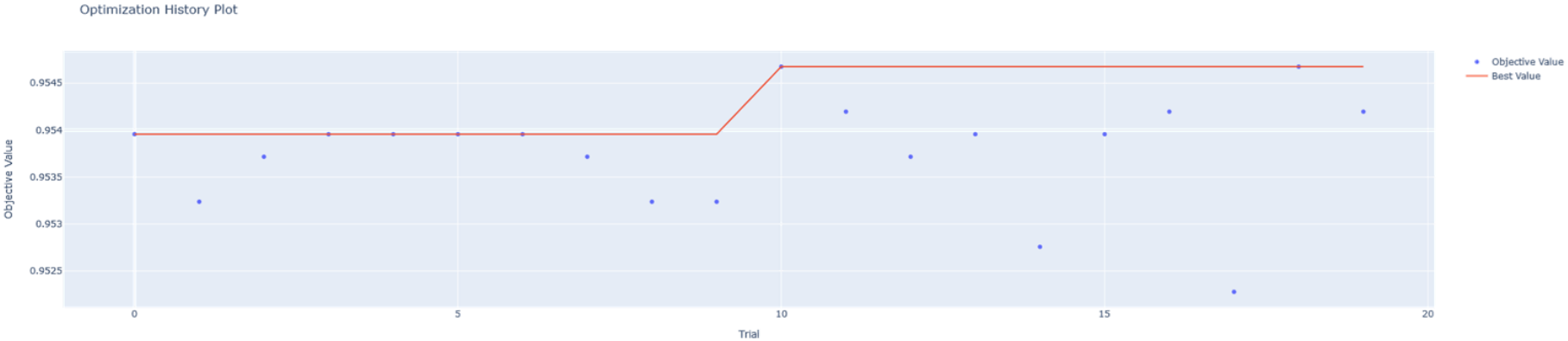
В результате оптимизации гиперпараметров модели получены следующие значения метрик:

Index	precision	recall	f1-score	support
0	0.79	0.68	0.73	1841
1	0.75	0.81	0.78	866
2	0.76	0.82	0.79	784
3	0.61	0.72	0.66	680
accuracy	0.74	0.74	0.74	0.74
macro avg	0.73	0.76	0.74	4171
weighted avg	0.75	0.74	0.74	4171

Описание итоговой модели

Оптимальные значения гиперпараметров модели гребневой классификации: **alpha: 0.24, solver: sag**

Описание результатов обучения с полученными метриками



Как мы видим на графиках. Модель достаточно быстро достигает максимального значения. Как мы и предполагали, точность является удачной метрикой для данного набора данных, для классов «Книги» и «Электроника» незначительно уступают значению точности, а значение f1 практически равно равно значению метрики точности. При этом, для раздела «Одежда и Аксессуары» мы наблюдаем очень высокие значения этих трёх метрик.

index	precision	recall	f1-score	support
0 Household	0.97	0.94	0.95	1 639
1 Books	0.93	0.97	0.95	902
2 Clothing & Accessories	0.98	0.98	0.98	849
3 Electronics	0.93	0.95	0.94	781
accuracy	0.95	0.95	0.95	0.95
macro avg	0.95	0.96	0.96	4 171
weighted avg	0.96	0.95	0.95	4 171

Выводы

5

1. Лучшая модель: Гребневая регрессия

2. Удачные эксперименты:
A) На всем множестве признаков: Ridge Classifier, Linear SVC, XGBoost, Extra Trees Classifier
B) На множестве признаков-символов, неплохую производительность показали алгоритмы XGBoost Extra Trees Classifier Random Forest

3. Неудачные эксперименты: Multinomial NB, Logistic Regression, SGD Classifier, Perceptron

4. Дальнейшие пути развития и улучшения решения: проверка возможности снижения размерности признаков в матрице TFI-DF. Также планируется оценить возможность повысить точность прогноза применения стеккинга и голосующего классификатора.

5. На чём следует сосредоточить усилия: полученная модель высокочувствительна к классу «Одежда и Аксессуары» и редко ошибается, однако при попытке выявления товаров из классов Книги и Электроника, модель чаще совершает ошибки I рода, также ошибки II рода чаще встречаются в случае класса «Домоводство». Следовательно необходимо повысить качество

Спасибо за внимание!