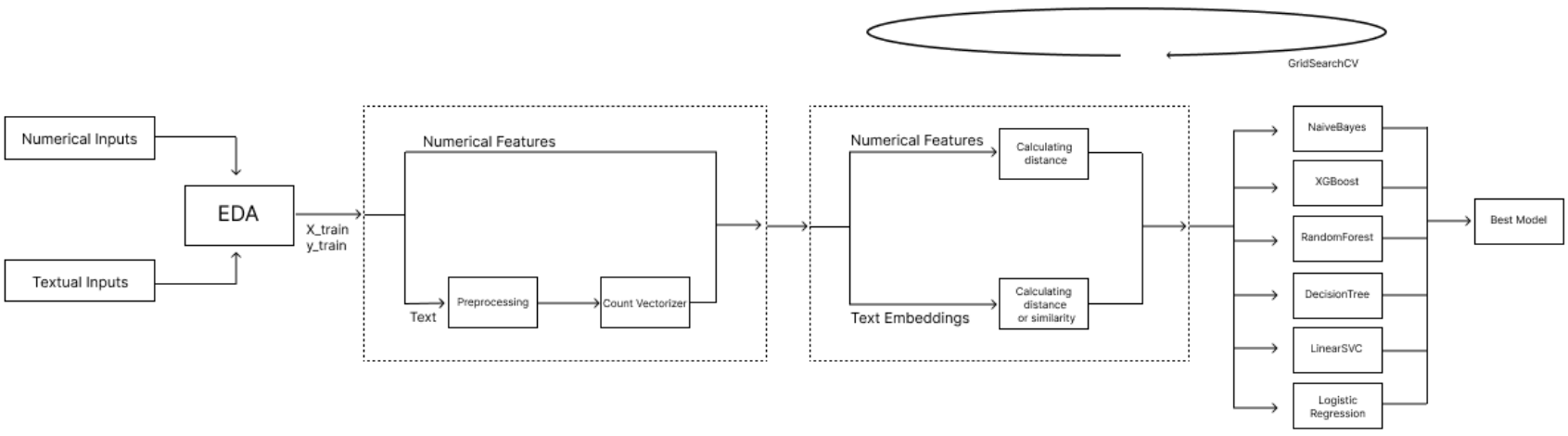
Описание метода классификации пар

Общее представление описанного ниже метода можно рассмотреть на следующей картинке



Сбор данных после разметки

После разметки пар с помощью веб интерфейса удалось собрать 2054 пары.

Полученная таблица имеет следующий вид:

pair_id	id_1	id_2	match_type
1966	6212	6213	1
488	665	4582	3
2005	2569	3833	1
1547	1276	7185	2
345	3835	6034	3

а также следующие статистики:

	products	pos. pairs	neg. pairs
smartphones	7500	497	1557
books	6878	347	725
TV	2903	1050	741
toys (LEGO)	9865	3512	2533

На следующем шаге для полученных пар выбиралась вся информации из таблицы products в базе данных по id товаров.

В итоге данные принимают следующий вид

pair_id	id_1	title_1	brand_1	price_1	description_1	specifications_1	marketplace_1	id_2	title_2	brand_2	price_2	description_2	specifications_2	marketplace_2	match_type
608	371	Смартфон Samsung Galaxy S22 Ultra SM-S908 128G...	Samsung	103790.0	Заводская упаковка вскрыта для проверки качест...	{'беспроводные интерфейсы': 'bluetooth; nfc; w...	wb	6904	6.6" Смартфон Samsung Galaxy M33 128 ГБ зеленый	Samsung	24999.0	Смартфон Samsung Galaxy M33 128 ГБ воплощает в...	{'гарантия от продавца': '12 мес.', 'модель': ...	dns	3
2041	3107	Смартфон Samsung SM-F936B Galaxy Z Fold 4	Samsung	138990.0	Смартфон Samsung SM-F936B Galaxy Z Fold 4,цвет...	{'беспроводные интерфейсы': 'поддержка bluetoo...	wb	6999	7.6" Смартфон Samsung Galaxy Z Fold4 512 ГБ бе...	Samsung	124999.0	Если вы хотите получить в свое распоряжение дв...	{'fm радио': 'нет', 'led-индикатор уведомлений...	dns	1
605	4375	Мобильный телефон PPA-LX1 CRUSH GREEN	Huawei	22789.0	Huawei P Smart 2021 – стильный смартфон в проч...	{'вес товара с упаковкой (г)': '330 г', 'высот...	wb	7186	6.1" Смартфон Samsung Galaxy S23 128 ГБ зеленый	Samsung	74999.0	Смартфон Samsung Galaxy S23 основан на однокри...	{'fm радио': 'нет', 'nfc': 'есть', 'автофокуси...	dns	3
1844	2569	Смартфон Note 12 2023 8/128 ГБ	Infinix	13812.0	Компания Infinix представляет новейшую модель ...	{'беспроводные интерфейсы': 'bluetooth; wi-fi'...	wb	7072	6.7" Смартфон Infinix NOTE 12 2023 128 ГБ серый	Infinix	12999.0	Смартфон Infinix NOTE 12 2023 продемонстрирует...	{'fm радио': 'есть', 'nfc': 'есть', 'автофокус...	dns	1
960	1673	Смартфон iPhone 14 128 ГБ eSIM фиолетовый	Apple	78019.0	Операционная система. Операционная система: iO...	{'версия операционной системы': '16', 'вес тов...	wb	7659	6.1" Смартфон Apple iPhone 14 Pro 1000 ГБ фиол...	Apple	156399.0	У iPhone 14 Pro 6,1--дюймовая панель. Пиковая ...	{'гарантия от продавца': '12 мес.', 'модель': ...	dns	3

	pairs with different marketplace	pos. pairs with different marketplace

smartphones	210	43

Препроцессинг

После того, как все данные собраны, реализуется следующий подход для построения модели классификации.

На вход алгоритму поступают как текстовые (название, бренд, описание, характеристики), так и числовые (цена) данные.

Работа с текстовыми данными:

Текстовые данные необходимо предобработать, для этого осуществляются следующие шаги:

- приведение текста к нижнему регистру
- замена спец. комбинаций единым видом (ГБ→ gb)
- удаление множественных пробельных символов между словами
- токенезация текстов
- удаление токенов, являющимися стоп словами или пунктуационными символами, т.к. эти данные не принесут необходимой информации, а только ухудшат качество
- Приведение слова к нормальной форме
- из столбцов с характеристиками товаров были получены только не пустые значения без ключей (названий характеристик)
- для товаров с пустой строкой в названии, определялось название как первые 5 слов из описания

Для токенезации текстов используется ToktokTokenizer из библиотеки nltk, т.к. он позволяет не убирать пунктуационные символы, если они являются частью слова (напр. диагональ 5.5).

Для приведения слова к нормальной форме используется лемматизация, т.к. этот подход лучше работает с данными на русском языке. Для применения лемматизации используется MorphAnalyzer из библиотеки pymorphy2.

Численные данные (цена), не обрабатывались, т.к. они уже предобрабатывались на этапе парсинга маркетплейсов.

Построение векторного пространства

После того, как текстовые данные были предобработаны, их необходимо перевести в векторное представление.

Для построения векторного представления был выбран базовый алгоритм bag-of-words. (также проведены сравнения с tf-id, word2vec, doc2vec, fasttext)

Особенностью построения векторного представления на существующих данные является множественное количество столбцов, в которых содержится похожая, но все же отличающаяся информация. Поэтому, каждая модель для построения векторного пространства обучалась только на определенном столбце. Кроме того, данные об одном признаке содержатся не в одном столбце (напр. title), а в двух (title_1, title_2), поэтому предварительно необходимо объединить столбцы по признакам и удалить повторяющиеся элементы.

В результате чего удается получить 4 модели построения векторного пространства (для названий, для бренда, для описаний и для характеристик), а также обучить и трансформировать данные с помощью обученных моделей.

```
{'title': CountVectorizer(binary=True, min_df=2),
 'brand': CountVectorizer(binary=True, min_df=2),
 'description': CountVectorizer(binary=True, min_df=2),
 'specification_values': CountVectorizer(binary=True, min_df=2)}
```

title_wordocc_1	title_wordocc_2	brand_wordocc_1	brand_wordocc_2	description_wordocc_1	description_wordocc_2	specification_values_wordocc_1	specification_values_wordocc_2
(0, 32)\t1\n (0, 264)\t1\n (0, 418)\t1\n ...	(0, 16)\t1\n (0, 264)\t1	(0, 22)\t1	(0, 22)\t1	(0, 92)\t1\n (0, 155)\t1\n (0, 366)\t1\n ...	(0, 25)\t1\n (0, 330)\t1\n (0, 499)\t1\n ...	(0, 81)\t1\n (0, 95)\t1\n (0, 124)\t1\n (...	(0, 11)\t1\n (0, 24)\t1\n (0, 25)\t1\n (0...
(0, 179)\t1\n (0, 305)\t1	(0, 0)\t1\n (0, 42)\t1\n (0, 80)\t1\n (0,...	(0, 64)\t1	(0, 73)\t1	(0, 0)\t1\n (0, 48)\t1\n (0, 78)\t1\n (0,...	(0, 16)\t1\n (0, 219)\t1\n (0, 425)\t1\n ...	(0, 60)\t1\n (0, 65)\t1\n (0, 85)\t1\n (0...	(0, 11)\t1\n (0, 41)\t1\n (0, 112)\t1\n (...
(0, 0)\t1\n (0, 16)\t1\n (0, 139)\t1\n (0...	(0, 9)\t1\n (0, 120)\t1\n (0, 125)\t1\n (...	(0, 73)\t1	(0, 73)\t1	(0, 16)\t1\n (0, 65)\t1\n (0, 130)\t1\n (...	(0, 25)\t1\n (0, 59)\t1\n (0, 338)\t1\n (...	(0, 11)\t1\n (0, 24)\t1\n (0, 41)\t1\n (0...	(0, 6)\t1\n (0, 41)\t1\n (0, 63)\t1\n (0,...

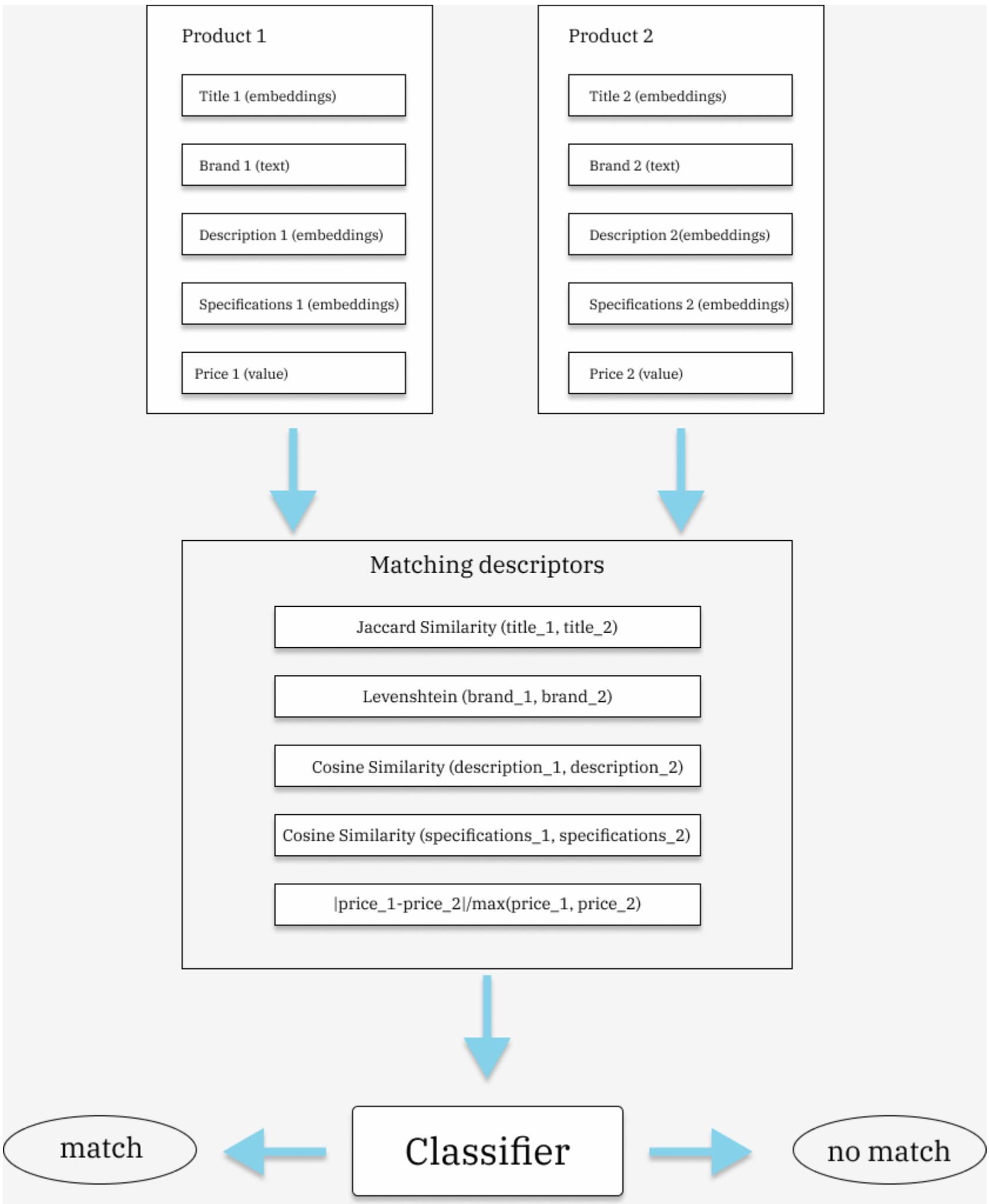
Создание новых признаков для модели

Так как полученные векторные представления для стобцов не могут быть отправлены модель, то для построение датасета, который будет подаваться в модель, считается близость между векторами или значениями по следующим правилам:

- Для длинных строк (описание, характеристики) считается cosine similarity
- Для коротких строк (название) считается Jaccard similarity
- Для строк, в основном состоящих из одного слова (бренд) считается расстояние Левенштейна
- Для цены считается относительная разность по формуле

$$\frac{|priceA - priceB|}{\max(priceA, priceB)}$$

Если визуализировать выщеперечисленные правила, то получается следующая картина:



Построение модели

После того, как датасет для обучения построен, данные разбиваются на train и test в соотношении 70% на 30%, причем в каждом из них есть данные как о positive парах, так и о negative парах в равном отношении.

Для проведения экспериментов и нахождения лучшей модели используются такие модели как: NaiveBayes, XGBoost, RandomForest, DecisionTree, LinearSVC, LogisticRegression, все из которых способны решать задачу классификации. Для каждой модели был подобран список параметров, а также список их значений, для нахождения лучших параметров модели.

Пример:

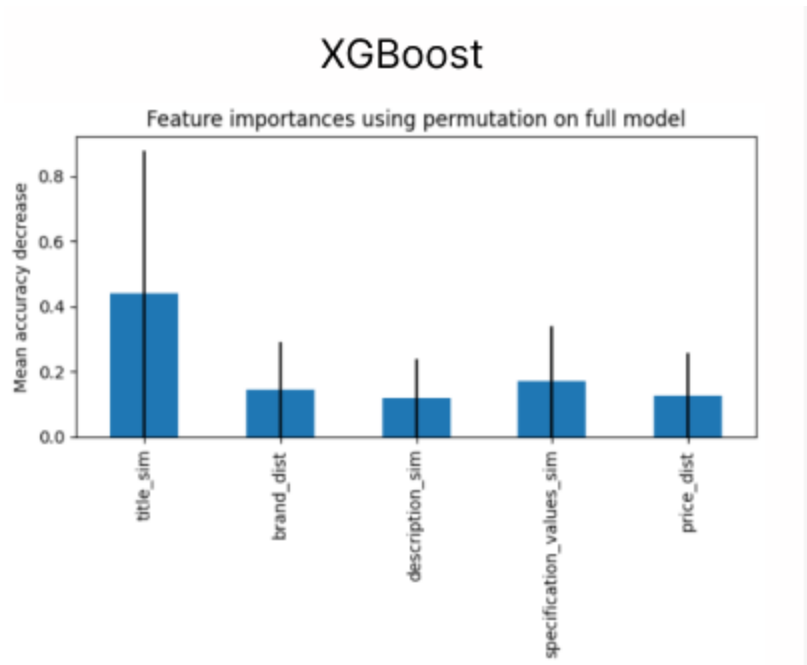
```
'DecisionTree':  {'clf':DecisionTreeClassifier(random_state=42),
                  'params':{'max_features': ['sqrt', 'log2', None],
                   'max_depth': [2,4,7,10],
                   'min_samples_split': [2, 5, 10, 20],
                   'min_samples_leaf': [1, 2, 4, 8],
                   'class_weight':[None, 'balanced']
                   }},
```

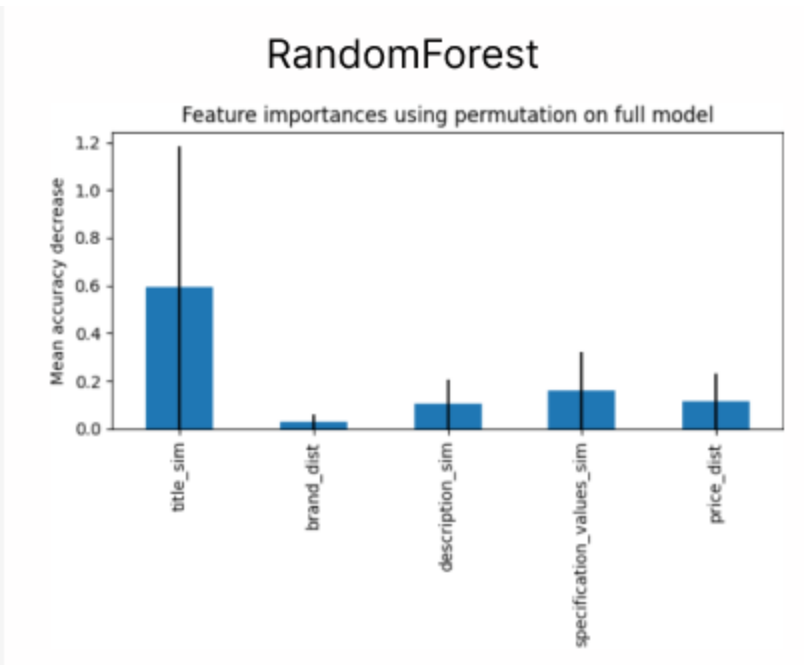
Нахождение лучших параметров было реализовано с помощью GridSearchCV для перебора всех возможных пар параметров, с использованием метрики f1-score, для выбора лучшей модели по этой метрике.

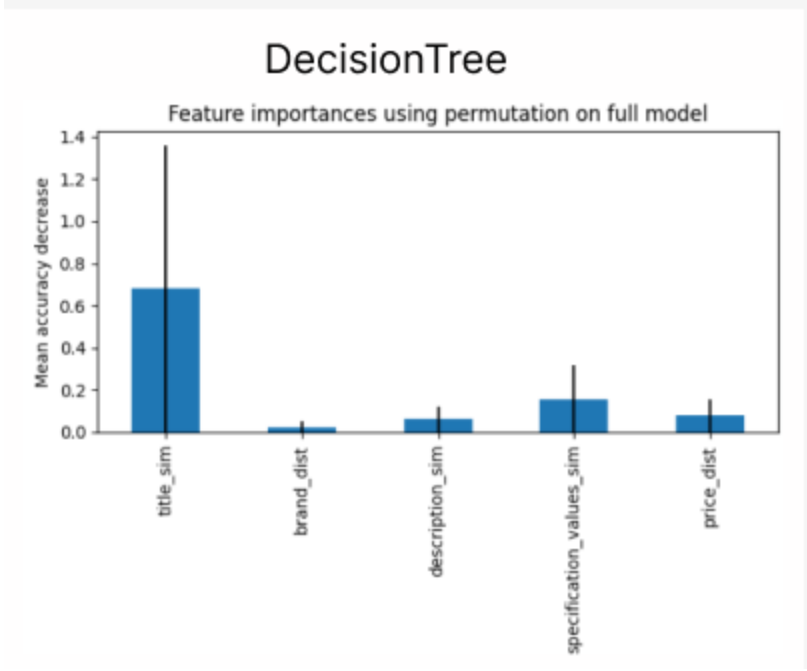
В результате перебора, получились следующие лучшие параметры для моделей:

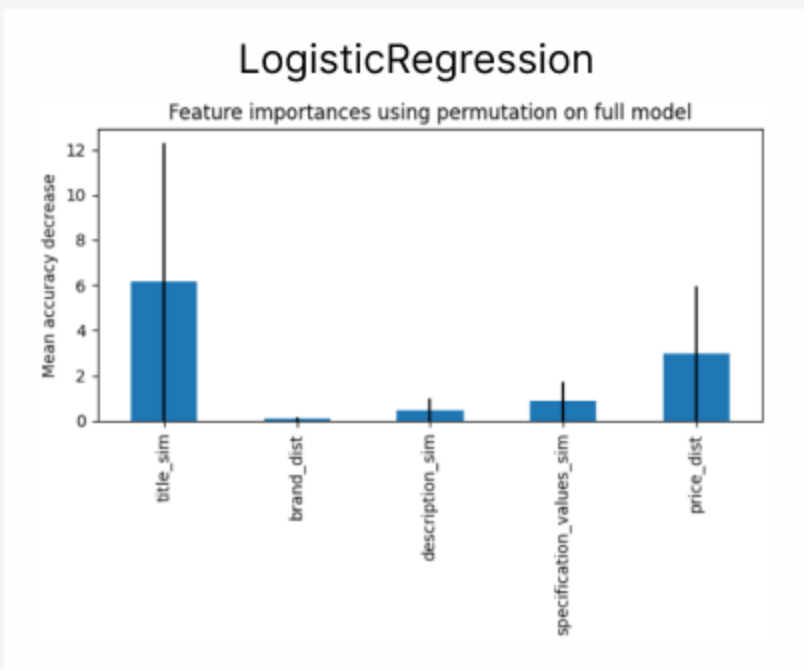
XGBoost	RandomForest	DecisionTree	LinearSVC	LogisticRegression
'subsample': 0.5, 'scale_pos_weight': 3, 'reg_lambda': 4.5, 'reg_alpha': 0, 'n_estimators': 100, 'min_child_weight': 1, 'max_depth': 10, 'learning_rate': 0.1, 'gamma': 0.5, 'colsample_bytree': 0.6	'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': None, 'max_depth': 10, 'class_weight': None	'min_samples_split': 2, 'min_samples_leaf': 8, 'max_features': None, 'max_depth': 10, 'class_weight': None	'class_weight': None 'C': 10	'class_weight': None, 'C': 100

Также была рассмотрена важность признаков для каждой модели

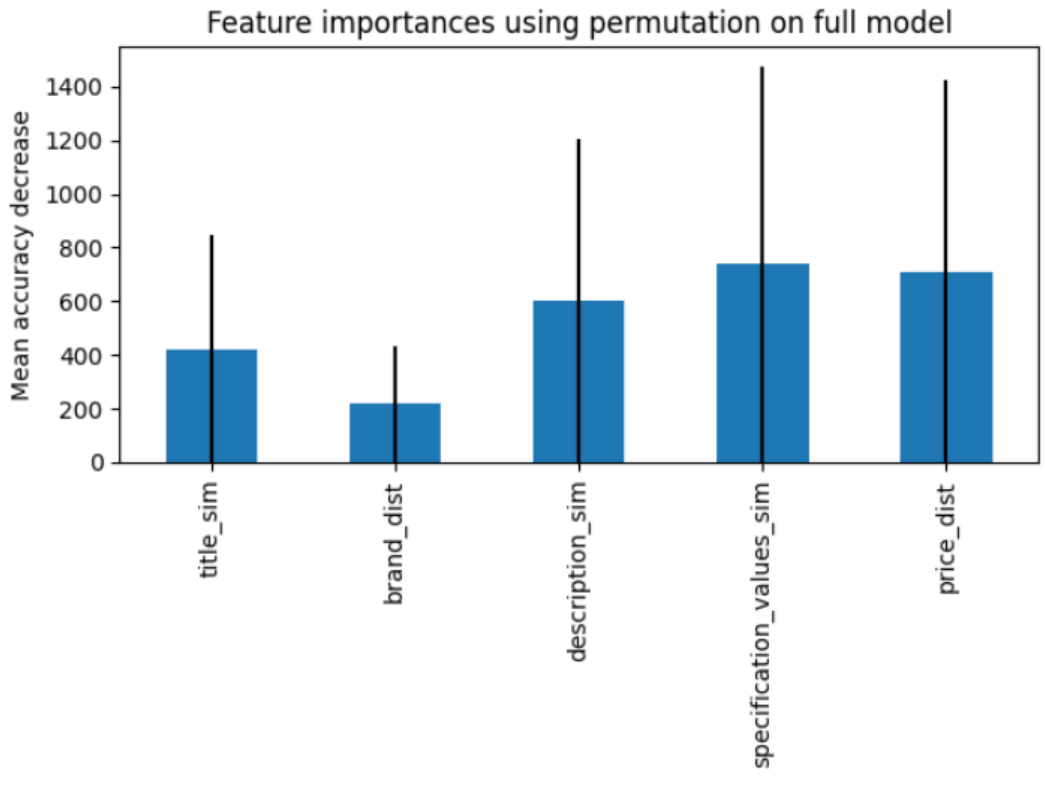








(добавить LightGBM!, там совсем другой график)



Проанализировав полученные графики, можно заметить, что близость названий товаров является наиболее значимым атрибутом, а расстояние между брендами можно отнести к менее значимым, что также можно отметить и про близость описаний (скорее всего в данном случае, в описаниях хранится слишком много шума).

Нахождение лучшей модели

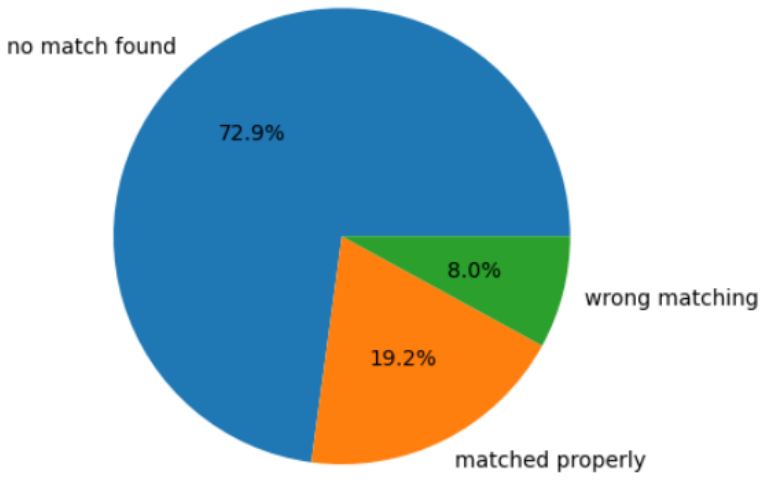
После получения лучших парметров для каждой модели, ищется лучшая модель по метрикам precision, recall, f1-score, accuracy, полученных после тестирования на тестовых данных.

В результате тестирования моделей получилась следующая таблица, показывающая как хорошо отработала модель на тестовых данных. Из таблицы видно

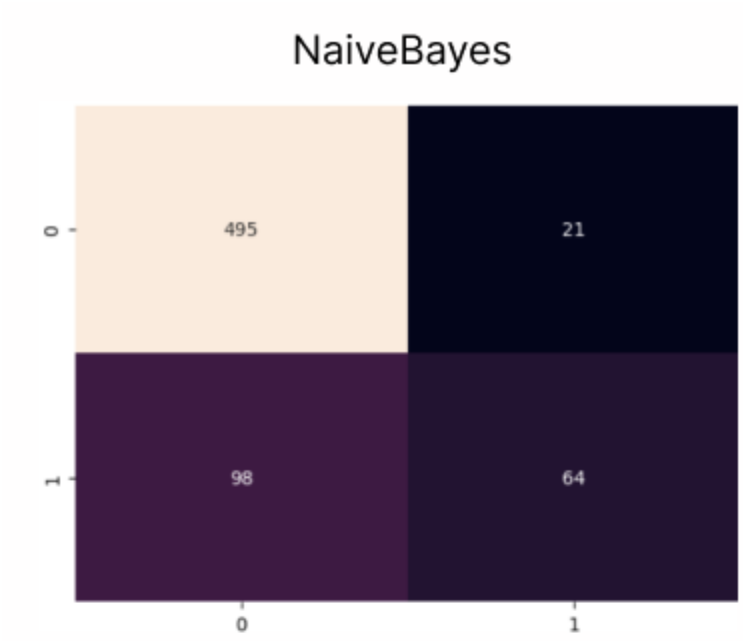
	precision	recall	f1-score	accuracy
NaiveBayes	0.79	0.68	0.71	0.82
XGBoost	**0.90**	**0.88**	**0.89**	**0.92**
RandomForest	0.88	0.85	0.86	0.90
DecisionTree	0.87	0.84	0.85	0.90
LinearSVC	0.84	0.83	0.83	0.88
LogisticRegression	0.86	0.83	0.84	0.89
LightGBM	0.90	0.85	0.87	0.91

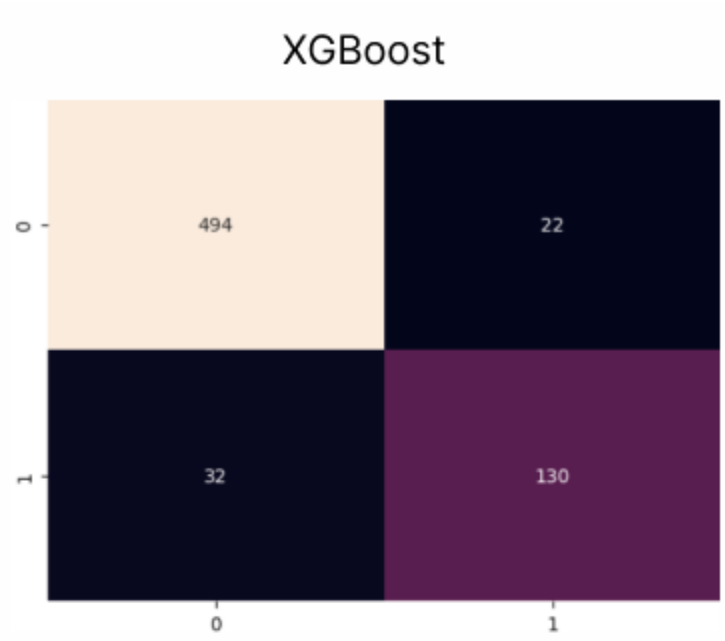
Из таблицы видно, что модель XGBoost показывает лучшие результаты на тестовых данных, но и нельзя сказать, что остальные модели справились сильно хуже.

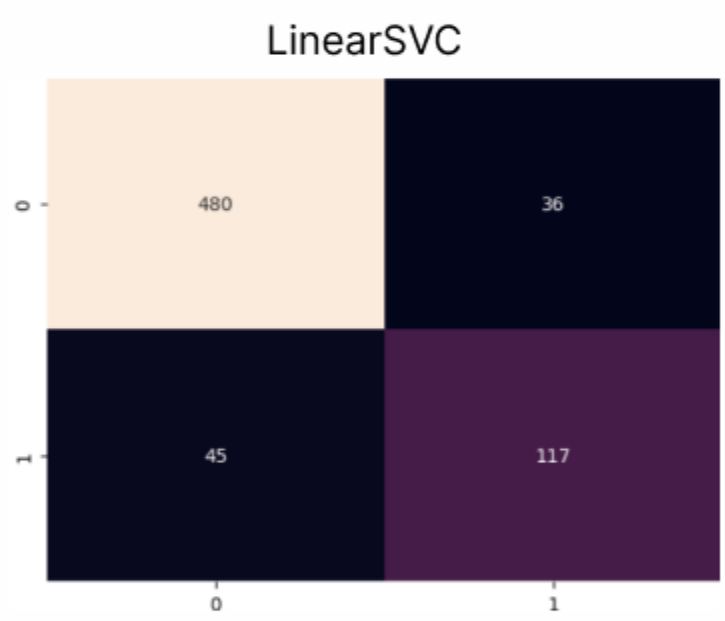
Распределение предсказаний для модели XGBoost

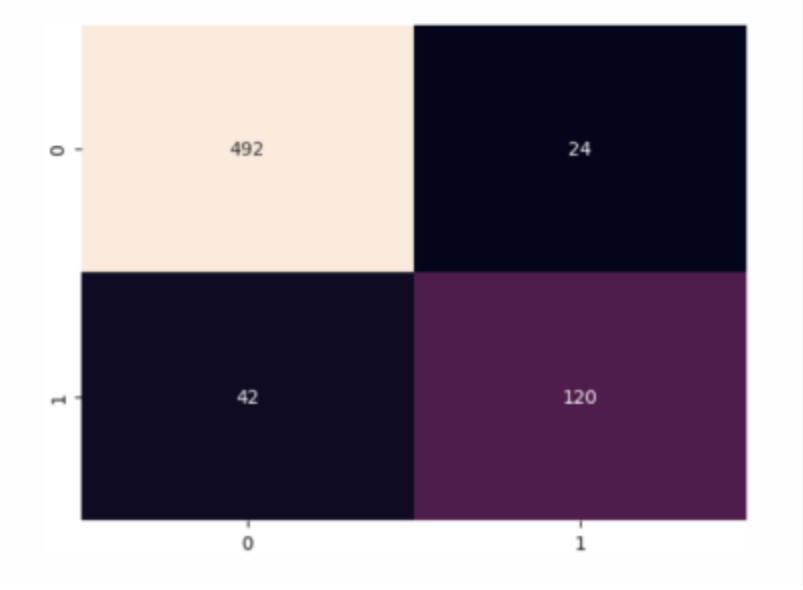


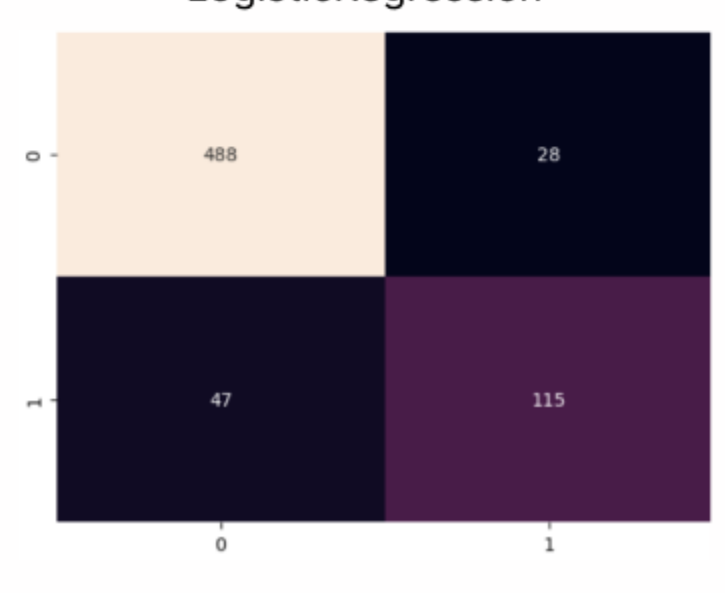
Для каждой модели были построены матрицы ошибок, которые видны на рисунке ниже.











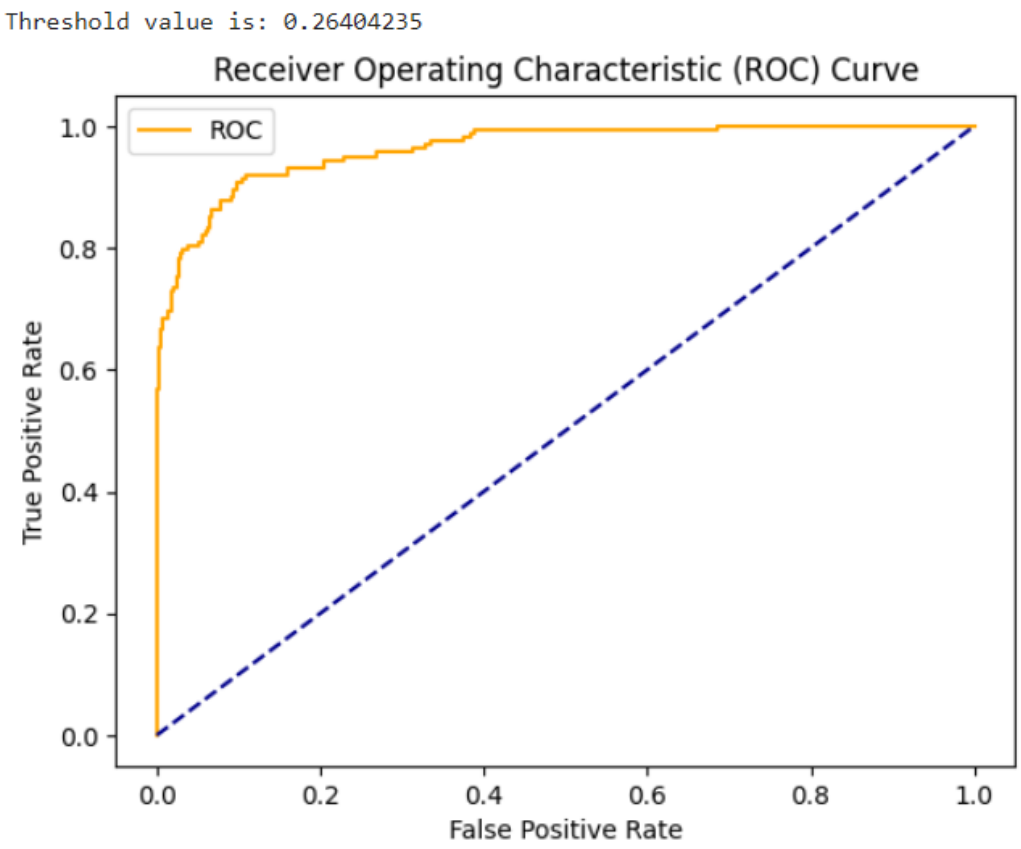
Из представленных матриц, можно заметить, что даже при хорошем проценте угалывания положительного или отрицательного объекта, в основном модели больше ошибаются предсказывая негативный класс на самом деле позитивным элементам, нежели ошибаясь в другую сторону. На данном этапе важно решить, что важнее: чтобы модель предсказывала как можно больше позитивных элементов, но при этом чаще ошибалась или чтобы модель находила меньше элементов, но ошибалась реже.

С названиями характеристик

	precision	recall	f1-score	accuracy
NaiveBayes	0.79	0.68	0.71	0.82
XGBoost	0.87	0.85	0.86	0.90
RandomForest	0.87	**0.87**	0.87	0.91

	precision	recall	f1-score	accuracy
DecisionTree	0.86	0.86	0.86	0.90
LinearSVC	0.85	0.84	0.84	0.89
LogisticRegression	0.85	0.83	0.84	0.89
LightGBM	**0.92**	0.86	**0.89**	**0.92**

больше картинок: так мы выбирали threshold, но лучше не стало (у некоторых моделей вырос показатель recall, это говорит о …, но в общем случае показатели ухудшились)



	precision	recall	f1-score	accuracy
NaiveBayes	0.79	0.68	0.71	0.82
XGBoost	**0.85**	**0.90**	**0.87**	**0.90**
RandomForest	0.81	**0.85**	0.82	0.86
DecisionTree	0.79	0.84	0.81	0.85
LinearSVC	0.84	0.83	0.83	0.88
LogisticRegression	0.79	0.83	0.80	0.85
LightGBM	**0.85**	0.87	0.86	0.89

После того как убрали description (стало хуже, не убираем)

	precision	recall	f1-score	accuracy
NaiveBayes	0.79	0.68	0.71	0.82
XGBoost	0.81	**0.88**	0.84	0.87
RandomForest	0.80	0.86	0.82	0.86
DecisionTree	0.83	0.84	0.84	**0.88**
LinearSVC	0.84	0.83	0.83	0.88
LogisticRegression	0.78	0.83	0.80	0.84
LightGBM	**0.84**	0.87	**0.85**	**0.88**

После того как убрали brand

	precision	recall	f1-score	accuracy
NaiveBayes	0.79	0.68	0.71	0.82
XGBoost	0.84	**0.87**	0.86	0.89
RandomForest	0.83	0.87	0.84	0.88
DecisionTree	0.80	0.84	0.82	**0.86**
LinearSVC	0.84	0.82	0.83	0.88
LogisticRegression	0.80	0.85	0.82	0.86

	precision	recall	f1-score	accuracy
LightGBM	**0.86**	0.86	**0.86**	**0.90**

Описание - первые 5 слов (улучшения не дало)

	precision	recall	f1-score	accuracy
NaiveBayes	0.62	0.67	0.57	0.59
XGBoost	0.81	**0.88**	0.84	0.87
RandomForest	0.80	0.87	0.82	0.85
DecisionTree	0.81	0.85	0.83	**0.87**
LinearSVC	0.82	0.81	0.82	0.87
LogisticRegression	0.80	0.85	0.82	0.85
LightGBM	**0.86**	0.86	**0.86**	**0.90**

конкатенация признаков

	precision	recall	f1-score	accuracy
title				
NaiveBayes	**0.86**	0.68	0.71	0.84
XGBoost	0.77	0.82	0.79	0.83
RandomForest	0.71	0.76	0.72	0.76
DecisionTree	0.70	0.76	0.71	0.76
LinearSVC	0.81	0.82	0.82	0.87
LogisticRegression	0.82	0.83	0.82	0.87
title+description				
NaiveBayes	0.83	0.73	0.76	0.85
XGBoost	0.82	0.84	0.83	0.87
RandomForest	0.80	0.79	0.79	0.85
DecisionTree	0.71	0.77	0.73	0.77
LinearSVC	0.85	0.81	0.83	0.88
LogisticRegression	0.84	**0.86**	**0.85**	**0.89**
title+brand				
NaiveBayes	0.84	0.66	0.7	0.83
XGBoost	0.78	0.83	0.80	0.84
RandomForest	0.76	0.78	0.77	0.83
DecisionTree	0.70	0.76	0.71	0.75
LinearSVC	0.84	0.84	0.84	0.88
LogisticRegression	0.84	0.84	0.84	0.88
title+brand+description				
NaiveBayes	0.83	0.74	0.77	0.85
XGBoost	0.80	0.83	0.81	0.86
RandomForest	0.71	0.76	0.73	0.78
DecisionTree	0.71	0.76	0.72	0.76
LinearSVC	0.84	0.85	**0.85**	0.88
LogisticRegression	0.84	**0.86**	**0.85**	**0.89**
title+description+specifications				
NaiveBayes	0.81	0.78	0.79	0.86
XGBoost	0.82	0.83	0.83	0.87
RandomForest	0.80	0.81	0.80	0.85
DecisionTree	0.76	0.79	0.77	0.82
LinearSVC	0.83	0.83	0.83	0.88
LogisticRegression	0.84	0.84	0.84	**0.89**

	precision	recall	f1-score	accuracy
title+brand+description+specifications				
NaiveBayes	0.81	0.78	0.79	0.86
XGBoost	0.83	0.82	0.82	0.87
RandomForest	0.78	0.81	0.79	0.84
DecisionTree	0.75	0.79	0.77	0.82
LinearSVC	0.83	0.83	0.83	0.87
LogisticRegression	0.85	0.85	**0.85**	**0.89**

tf-idf vectorizer (очень близкое качество с изначальной моелью у LightGBM)

	precision	recall	f1-score	accuracy
NaiveBayes	0.79	0.68	0.71	0.82
XGBoost	0.86	0.86	0.86	0.90
RandomForest	0.85	**0.85**	0.85	0.89
DecisionTree	0.85	0.84	0.85	0.89
LinearSVC	0.85	0.82	0.83	0.88
LogisticRegression	0.86	0.82	0.84	0.89
LightGBM	**0.91**	**0.85**	**0.88**	**0.92**

word2vec (хорошие результаты)

	precision	recall	f1-score	accuracy
NaiveBayes	0.80	0.68	0.71	0.83
XGBoost	**0.93**	**0.85**	**0.88**	**0.92**
RandomForest	**0.93**	**0.85**	**0.88**	**0.92**
DecisionTree	0.90	**0.85**	0.87	0.91
LinearSVC	0.78	0.76	0.77	0.84
LogisticRegression	0.78	0.73	0.75	0.83
LightGBM	0.91	**0.85**	**0.88**	**0.92**

doc2vec (это лучший результат)

	precision	recall	f1-score	accuracy
NaiveBayes	0.80	0.68	0.71	0.83
XGBoost	0.88	0.86	0.87	0.91
RandomForest	**0.93**	**0.88**	**0.90**	**0.93**
DecisionTree	0.86	0.83	0.84	0.89
LinearSVC	0.79	0.75	0.77	0.84
LogisticRegression	0.79	0.73	0.76	0.84
LightGBM	0.90	0.85	0.87	0.91

fasttext

	precision	recall	f1-score	accuracy
NaiveBayes	0.80	0.68	0.71	0.83
XGBoost	0.89	0.86	0.87	0.91
RandomForest	**0.89**	**0.86**	**0.88**	**0.91**
DecisionTree	0.84	0.81	0.83	0.88
LinearSVC	0.81	0.73	0.76	0.85
LogisticRegression	0.80	0.74	0.76	0.84
LightGBM	0.89	0.85	0.87	0.91

SentenceTransformers

	precision	recall	f1-score	accuracy
NaiveBayes	0.80	0.68	0.71	0.83
XGBoost	0.89	0.85	0.87	0.91
RandomForest	**0.93**	**0.85**	**0.88**	**0.92**
DecisionTree	0.90	0.81	0.85	0.90
LinearSVC	0.87	0.81	0.84	0.89
LogisticRegression	0.85	0.82	0.83	0.88
LightGBM	0.89	0.84	0.86	0.91

senttr with cos sim on all features (странно, по ощущениям должно было быть лучше, может остальные признаки хорошо влияют)

	precision	recall	f1-score	accuracy
NaiveBayes	0.80	0.68	0.71	0.83
XGBoost	**0.92**	**0.85**	**0.88**	**0.92**
RandomForest	0.90	0.84	0.86	0.91
DecisionTree	0.87	0.83	0.85	0.89
LinearSVC	0.82	0.77	0.79	0.86
LogisticRegression	0.82	0.77	0.79	0.86
LightGBM	0.88	0.84	0.86	0.90

Пример 1 предсказания положительного класса, для случайно выбранного объекта:

2264	TECNO POVA Neo 2/6.82/1640*720/4+64ГБ	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
2267	TECNO POVA Neo 2/6.82/1640*720/4+64ГБ	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
2512	TECNO POVA Neo 2/6.82/1640*720/4+64ГБ	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
2529	Смартфон Tecno Pova Neo 2 4/64 Гб	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
3192	Смартфон TECNO Pova Neo 2 64GB	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
3207	Смартфон Tecno Pova Neo 2 4/64GB	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
3484	Смартфон Tecno Pova Neo 2, 4/64 ГБ	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
3485	Смартфон Tecno Pova Neo 2, 4/64 ГБ	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
3820	Смартфон Tecno Pova Neo 2 / (4ГБ/64ГБ)	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
3835	Смартфон Tecno POVA NEO 2 LG6n 4/64 ГБ	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
4159	Смартфон Tecno POVA NEO 2 LG6n 4/64 ГБ	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
4175	Смартфон Tecno Pova Neo 2 / (4ГБ/64ГБ)	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
4862	Смартфон Tecno Pova Neo 2 / (4ГБ/64ГБ)	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
5236	Смартфон TECNO Pova Neo 2 64GB	Смартфон TECNO POVA Neo 2 4/64 ГБ	1
5427	TECNO Pova 2 Neo 4+64 ГБ	Смартфон TECNO POVA Neo 2 4/64 ГБ	1

Пример 2 предсказания положительного класса, для случайно выбранного объекта:

	title_1	specification_values_1	title_2	specification_values_2	match
63	poco m5 4gb 128gb green	bluetooth wi fi nfc, adnroid 12, 201 г, usb ty...	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	1
1135	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	1
2747	poco m5 4gb 128gb black	bluetooth wi fi nfc, adnroid 12, 201 г, usb ty...	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	1
3793	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	1
5078	xiaomi poco m5 4 128gb black	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	1
5589	xiaomi poco m5 4 128gb black	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	1
5638	poco m5 4gb 128gb yellow	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	1
5639	poco m5 4gb 128gb green	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	poco m5	bluetooth 5.3 wi-fi 2.4 5.0 ггц nfc ик-порт, a...	1