Тестовое задание для кандидатов на предстажировку Case Lab ML

Есаревой Елены

В рамках тестового задания необходимо разработать веб-сервис для оценки комментариев (отзывов) к фильмам.

Описание данных:

В качестве исходных данных использовался открытый набор данных, который содержит в себе отзывы о фильмах, а также соответствующие им оценки рейтинга.

https://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

---

Этапы выполнения работы:

- Загрузка данных
- Исследовательский анализ данных
- Обучение моделей
- Тестирование лучшей модели
- Разработка веб-сервиса на базе фреймворка Django

Загрузка данных

Из представленных данных сформировали датасеты.

Данные в датасетах соответствуют описанию.

Количество сток - 25 000, Количество столбцов - 4.

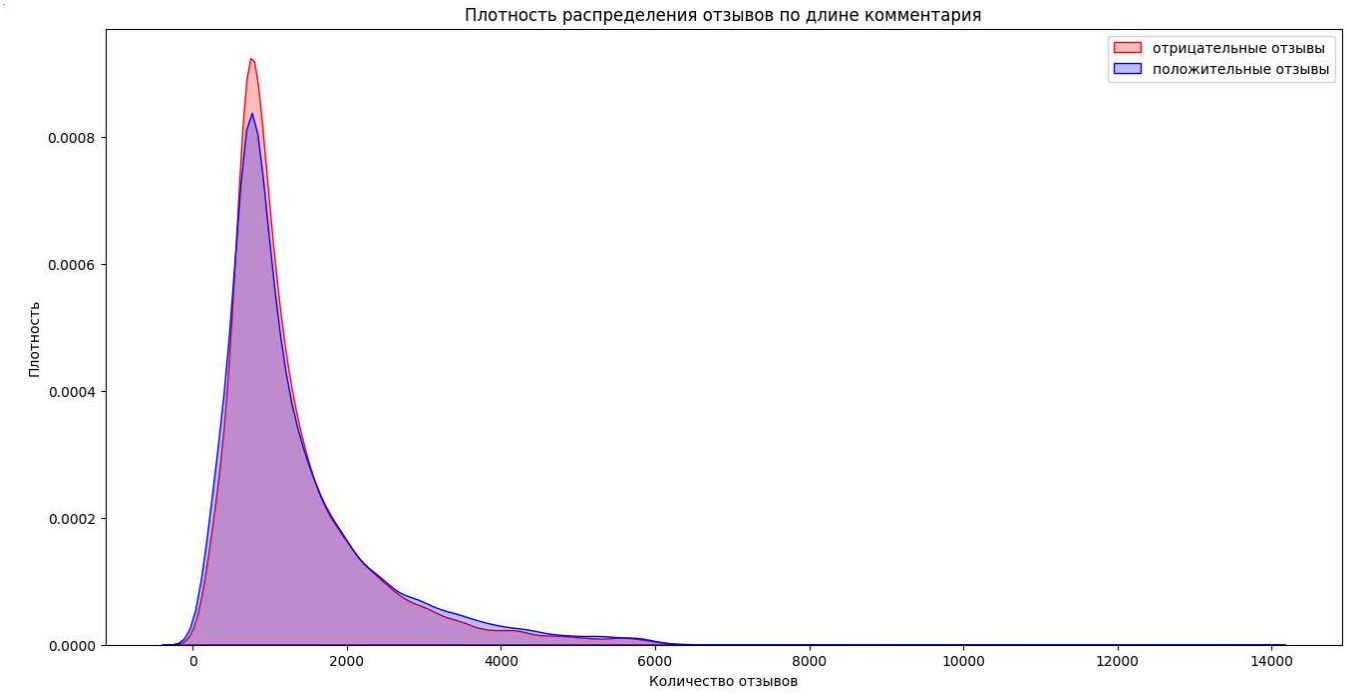
Тип данных указан верно, пропусков и дубликатов в датафреймах нет, название столбцов сответсвуют стилю написания snake_case

	id	rating	text	pos_neg
0	3937	8	The early career of Abe Lincoln is beautifully presented by Ford. Not that anyone alive has seen footage of the real Lincoln, but Fonda, wearing a fake nose, is uncanny as Lincoln, with the voice, delivery, walk, and other mannerisms - exactly as one would imagine Lincoln to have been. Ford, in the first of three consecutive films he made with Fonda, is at the top of his form, perfectly evoking early 19th century America. The story focuses on a pair accused of murder that Lincoln defends and the courtroom scenes are quite well done. The supporting cast includes many of Ford's regulars. This was Alice Brady's last film, as she died months after its release.	1
1	10290	8	Bruce Almighty is the best Jim Carrey work since The Truman Show, and was a pleasant surprise after some of his recent "Hey Hollywood - look how good I can act!" box office disappointments. It's great to see Jim recognizing and embracing his strengths. He won't get an Academy Award but the film itself will last longer than many of the "awarded films" of the Academy. He is at the top of his form in this most recent film - it's like the return of an old friend. Carrey, Freeman, and Aniston all do a great job together - comfortable in their comedy roles, superb comic timing, and obviously having fun together but without the "hey mom - look how funny I am" type of comedy. A real surprise was Steven Carrell as Carrey's nemesis (Carrell of The Daily Show fame), who walked away with some the best and funniest scenes of the film. I laughed harder at Carell than anyone else in the past three years. I can foresee the religious nuts in the US will be up-in-arms over the treatment of God, but the bottom line of the film is true to all major theological beliefs - we are masses of protoplasms trying to get through our short lives by exercising our free will. Without Married With Children t o complain about, this will likely become a target of people with misplaced priorities (who know the types - men adorned in gold watches on Sunday morning and late nigh television, selling prayers to God). And, again, about 0.5% of the country will care and 80% of the media will report it. The bottom line: this a purely entertaining film, each audience member laughingly wondering what they would do, and a feel-good feeling at the movie conclusion. A walk down any major street in America has to confirm that God has a tremendous sense of humor. What better comic genius to remind us of that than Jim Carry. Thanks again, Jim -- it's GREAT to have you back!!	1
2	11917	8	Yeah, the archetype of a simple but inspirational movie. The very end when the entire crowd in the stadium gets up and the people raise their hands gives me a chill whenever I see it. That's just brilliant. Joseph is wonderful as the lonely and sad kid who has so far been disappointed by anyone and anything in his life. The way he interacts with Danny Glover and tries to make him believe in the magic and the angels is funny and exhilarating. A very nice family movie with - I concede - a rather corny happy end. But hey, it doesn't really matter, the movie retains its basic quality by the good acting and the inspirational themes.	1

Исследовательский анализ данных

Для признака text:

- Видим, что отзывы написаны на английском языке. Для отзывов используется разный регистр и неинформативные знаки. Сделали копию датасета и привели текст к нижнему регистру, очистили от ненужных символов и лемматизировали (привели слова к начальным формам).
- В датасет с тренировочными данными добавили признак, длина комментария. Определили, что длина самых длинных комментариев превышает 10 000 знаков. Сделали вывод, положительность/отрицательность отзыва не зависит от его длины.

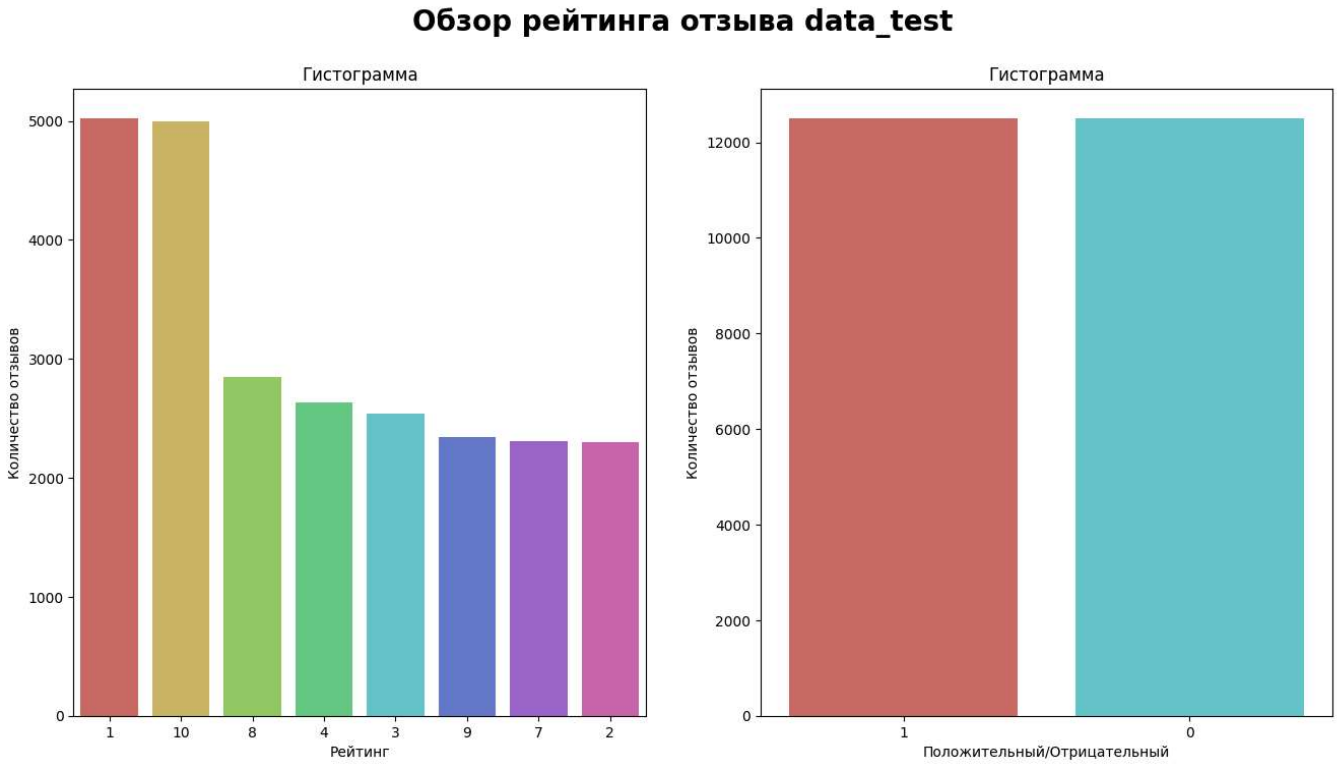


Для признака pos_neg - целевой признак

- Содержит два значения 0 и 1.
- Дисбаланса классов нет.

Для признака rating - целевой признак

- содержит 10 значений от 1 до 10.
- Значения 1 и 10 встречаются чаще остальных



Закодировали признаки с помощью TFIDF

Обучение моделей

Задача классификации

Предсказание бинарного признака pos_neg

Для поиска лучшей модели:

- была проведена подготовка данных
- в качестве моделей рассматривались: LogisticRegression(), CatBoostClassifier(), DecisionTreeClassifier()
- для моделей были подобраны гиперпараметры
- для перебора параметров использовали функцию grid
- в качестве метрики оценки модели была определена метрика f1_score

	Модель	f1_score
0	LogisticRegression	0.892375
2	CatBoostClassifier	0.836440
1	DecisionTreeClassifier	0.569362

Лучшей моделью оказалась:

- LogisticRegression(C = 5, penalty = 'l2', solver = 'liblinear')
- Метрика лучшей модели на тренировочной выборке: 0.89

- Метрика f1_score на тестовых данных: 0.87

Задача регресии

Предсказание признака raiting

Для поиска лучшей модели:

- была проведена подготовка данных
- в качестве моделей рассматривались: LinearRegression(), CatBoostRegressor(), DecisionTreeRegressor(), LGBMRegressor(), RandomForestRegressor()
- для моделей были подобраны гиперпараметры
- для перебора параметров использовали функцию grid
- в качестве метрики оценки модели была определена метрика MAE

	Модель	MAE
3	LGBMRegressor	2.104534
1	CatBoostRegressor	2.538928
0	LinearRegression	2.799353
4	RandomForestRegressor	2.877506
2	DecisionTreeRegressor	3.794577

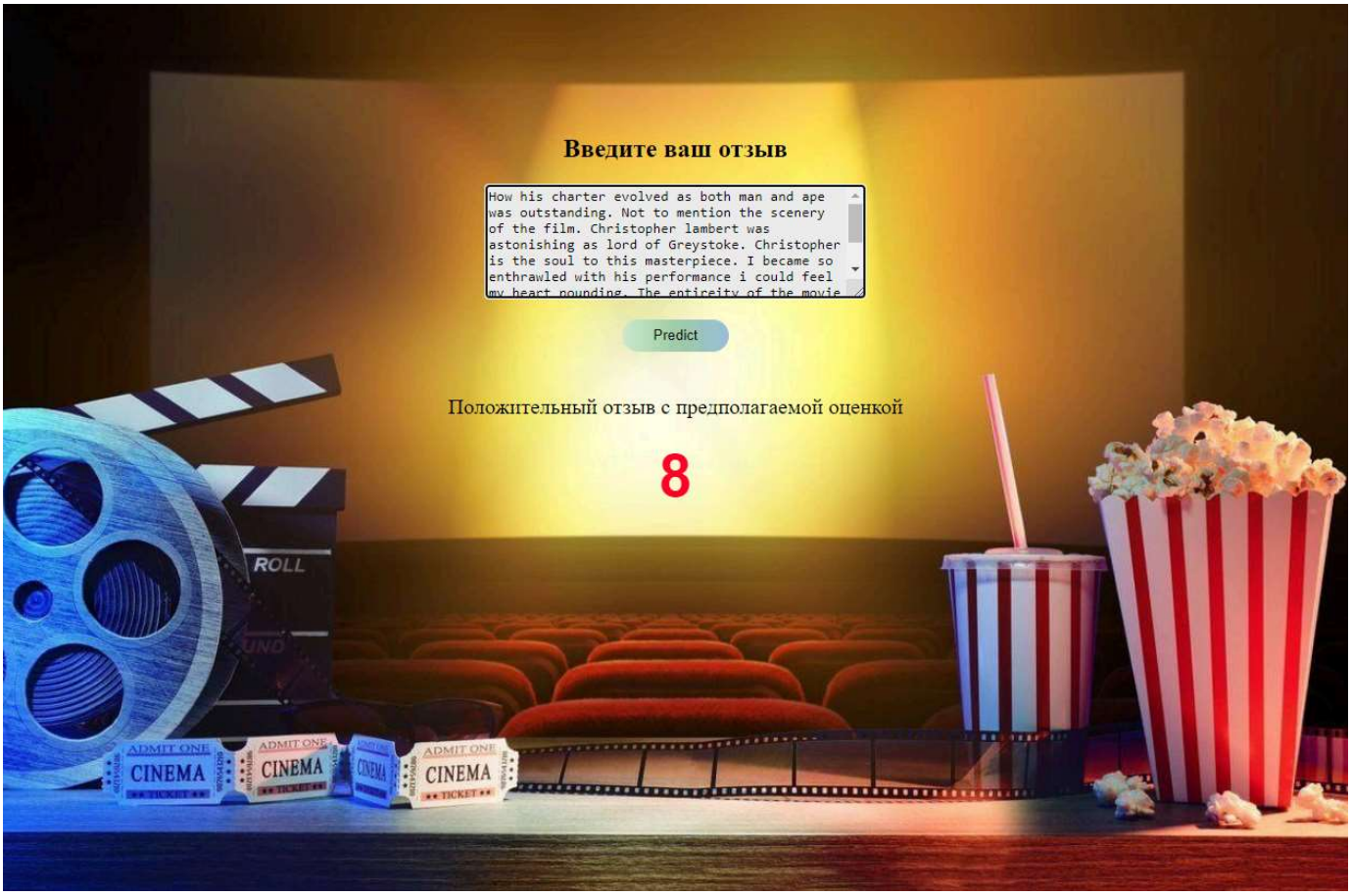
Лучшей моделью оказалась:

- LGBMRegressor(random_state=RANDOM_STATE, n_estimators = 200, max_depth = 9)
- Метрика лучшей модели на тренировочной выборке: 2.1
- Метрика MAE на тестовых данных: 1.79

Рекомендации по улучшению предсказаний моделей

Для повышения точности предсказаний модели можно рассмотреть использование языковой модели BERT или нейронных сетей. Однако для их применения потребуется больше вычислительных ресурсов.

Готовый сайт с предсказанием рейтинга отзыва.


Введите ваш отзыв
How his charter evolved as both man and ape was outstanding. Not to mention the scenery of the film. Christopher lambert was astonishing as lord of Greystoke. Christopher is the soul to this masterpiece. I became so enthrawled with his performance i could feel my heart pounding. The enticeity of the movie
Predict
Положительный отзыв с предполагаемой оценкой
8
ROLL
ADMIT ONE
CINEMA
TICKET