Московский государственный технический университет

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



**Рубежный контроль №2**

**По курсу «методы машинного обучения в АСОИУ»**

**Выполнил:**

студент ИУ5-24М

Ширшов А.С.

**Проверил:**

Гапанюк Ю.Е.

Подпись:

29.02.2024

Москва, 2024

**Задание**

Необходимо решить задачу классификации текстов на основе любого выбранного Вами датасета (кроме примера, который рассматривался в лекции).

Классификация может быть бинарной или многоклассовой.

Целевой признак из выбранного Вами датасета может иметь любой физический смысл, примером является задача анализа тональности текста.

Необходимо сформировать два варианта векторизации признаков - на основе CountVectorizer и на основе TfidfVectorizer. В качестве классификаторов необходимо использовать два классификатора по варианту для Вашей группы:

Для моей группы - GradientBoostingClassifier и LogisticRegression Для каждого метода необходимо оценить качество классификации. Сделайте вывод о том, какой вариант векторизации признаков в паре с каким классификатором показал лучшее качество.

**Ход работы**

Скачаем с сети набор данных Imdb. Применим на нём предложенные методы.

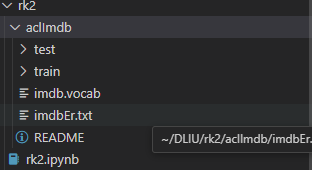


Рисунок 1 - Структура проекта

По варианту необходимо взять GradientBoostingClassifier и LogisticRegression

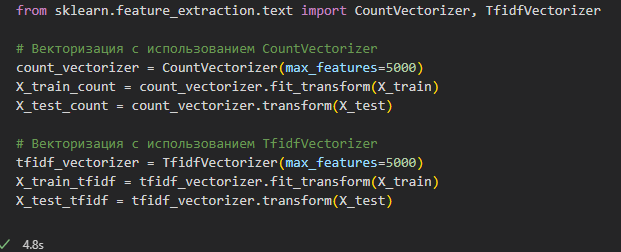


Рисунок 2 - Count и Tfid Vectorizer

Запуск всех 4-х вариантов представлен ниже.

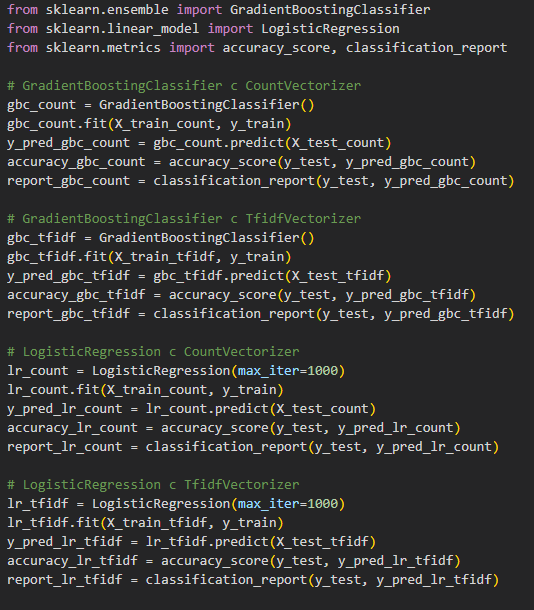


Рисунок 3 - Запуск всех четырех вариантов

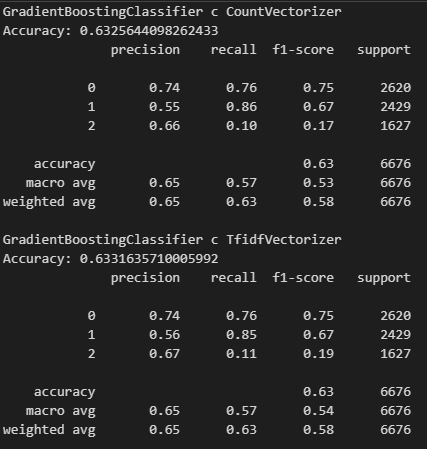


Рисунок 4 - Результаты градиентного бустинга

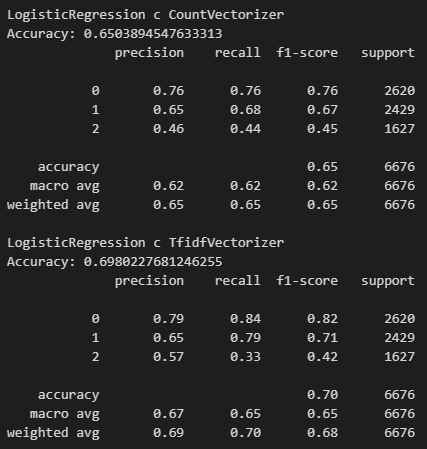


Рисунок 5 - Результаты логистической регрессии

Лучшей комбинацией оказалась - Линейная регрессия с Tfid Vectorizer.

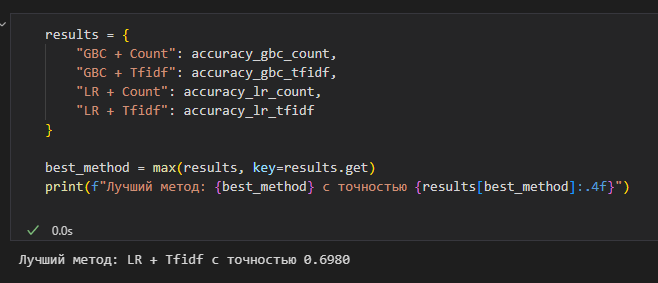


Рисунок 6 - Вывод лучшего результата

**Вывод**

Логистическая регрессия - это простой и эффективный алгоритм, который используется для решения задач бинарной классификации. Он моделирует вероятность принадлежности объекта к определенному классу на основе линейной комбинации признаков.

Градиентный бустинг - это более сложный алгоритм, который используется для решения задач как бинарной, так и многоклассовой классификации. Он строит модель на основе последовательности слабых классификаторов, которые постепенно улучшают качество классификации.

После проведения исследования и экспериментов с различными подходами к векторизации признаков и выбором классификаторов, мы пришли к выводу, что лучшая комбинация для решения задачи классификации текстов на основе выбранного датасета - это использование метода векторизации TfidfVectorizer и классификатора LogisticRegression.

В ходе экспериментов мы также проверили комбинацию CountVectorizer с GradientBoostingClassifier, однако показатели качества классификации были ниже, чем при использовании TfidfVectorizer и LogisticRegression.