Отчет по Проекту: Классификация Видеоигр

Проект: Классификация видеоигр на основе данных IMDb

Авторы:

Тұрғын Салтанат

Уршукова Томирис

Обзор Проекта и Подготовка Данных

Цель Проекта

Основная цель проекта — классификация видеоигр по уровню их **популярности** на основе **пользовательского рейтинга** (`User Rating`).

Датасет

Источник: Kaggle, файл `imdb_video_games.csv`.

Объем данных (исходный): 14682 записи.

Выбранные признаки для анализа:

- `User Rating` (Признак X) - числовая оценка пользователей игры.
- `Popularity` (Целевая переменная y) - числовой показатель популярности игры.

Предобработка: Были удалены строки с пропусками (NaN) в столбцах `Popularity` и `User Rating` для создания очищенного датасета `df_clean`.

Линейная Регрессия (Реализация с Нуля)

На первом этапе была реализована **одномерная линейная регрессия** для предсказания `Popularity` на основе `User Rating`.

Реализация и Обучение

Модель: Линейная регрессия, обученная с использованием градиентного спуска.

Метод: Batch Gradient Descent (пакетный градиентный спуск).

Функция потерь: Среднеквадратичная ошибка (MSE).

Результаты Обучения

Оптимальный Learning Rate: 0.01.

Найденные коэффициенты (после 500 эпох):

— Intercept (θ_0): 3397.360.
— Коэффициент (Slope, θ_1): 223.971.

Влияние Learning Rate на сходимость

На графике "Влияние learning rate на сходимость градиентного спуска" видно, что:

$\text{lr}=0.01$ показал лучшую сходимость к минимальному значению MSE.

Более высокие значения lr (например, $0.05, 0.1, 0.3$) привели к **расходимости** (значительному росту MSE) или нестабильному поведению, что подтверждается ошибками `RuntimeWarning: overflow encountered in square` и `RuntimeWarning: invalid value encountered in subtract`.

Логистическая Регрессия (Бинарная Классификация)

Задача была преобразована в **бинарную классификацию**, где целевой переменной `y` является бинарный класс, основанный на медиане популярности:

0: Популярность ниже медианы.

1: Популярность выше медианы.

Функционал (Реализован с нуля)

Сигмоидальная функция: $\sigma(z) = 1 / (1 + \text{np.exp}(-z))$.

Функция потерь: Бинарная кросс-энтропия (`Log Loss`).

Метод: Градиентный спуск с возможностью использования L2 регуляризации.

Результаты Обучения

Обучающая выборка (Train): 70% данных.

Тестовая выборка (Test): 30% данных.

Параметры: $\text{lr}=0.1$, $\text{epochs}=500$, $\text{L2\_reg}=0.01$.

Найденные коэффициенты:

- Intercept (θ_0): 0.851.
- Slope (θ_1): -0.136.

Метрики на Test (Логистическая регрессия)

Метрика	**Значениие**
Accuracy	0.613
Precision	0.726
Recall	0.373
F1-score	0.492
ROC AUC	0.698
Confusion Matrix	[[1106, 184], [820, 487]]

Классификация и Сравнение Моделей

Второй моделью для сравнения стало **Решающее Дерево (Decision Tree),** реализованное с помощью библиотеки `sklearn`.

Метрики на Test (Decision Tree)

Параметр: $\text{max\_depth}=5$.

Метрика	**Значениие**
Accuracy	0.641
Precision	0.653
Recall	0.613
F1-score	0.632
ROC AUC	0.693
Confusion Matrix	[[846, 426], [506, 801]]

Сравнение Моделей

Модель	Accuracy	F1-score	ROC AUC
Логистическая Регрессия	0.613	0.492	0.698
Decision Tree	0.641	0.632	0.693

Вывод: **Decision Tree** (при $\text{max\_depth}=5$) показало **лучшие** общие метрики (Accuracy, F1-score) на тестовой выборке.

Объяснение: Решающее дерево лучше работает на данном наборе данных (с единственным числовым признаком `User Rating`), вероятно, потому, что оно может строить **нелинейные пороговые** правила, которые более точно отделяют класс высокой популярности, чем простая линейная граница, используемая логистической регрессией.

Анализ Экспериментов (Влияние Гиперпараметров LogReg)

Был проведен анализ влияния **Learning Rate** и **количества Эпох** на сходимость функции потерь (`Log Loss`) для логистической регрессии.

Гиперпараметр	**Наблюдение**
Learning Rate (lr)	Более высокие значения lr (например, 0.1) приводят к **быстрой сходимости** Log Loss. Низкие значения ($\text{lr}=0.001$) сходят очень медленно, а слишком высокие ($\text{lr}=0.3$) могут вызывать нестабильность.
Эпохи (Epochs)	При $\text{lr}=0.1$ модель демонстрирует **полную сходимость** к минимальному Log Loss примерно к 300-й эпохе. Увеличение до 600 эпох не дает существенного прироста качества, подтверждая эффективность выбранного lr.