

Отчет по проекту

Прогнозирование популярности музыкальных треков Spotify с помощью машинного обучения

Данные: [high_popularity_spotify_data.csv](#)

Инструменты: Python, Pandas, NumPy, Scikit-learn, Matplotlib, Jupyter Notebook

1. Краткое описание работы

В данной работе изучалось, какие характеристики музыкальных треков влияют на их популярность на платформе Spotify. **Основная цель** — научиться предсказывать популярность композиций, используя методы машинного обучения.

Работа включала две задачи:

1. регрессию — прогноз числового значения популярности трека;
2. классификацию — определение, станет ли трек хитом или нет.

Для этого была реализована собственная модель линейной регрессии с использованием стохастического градиентного спуска (SGD), а также применены готовые модели из библиотеки scikit-learn: логистическая регрессия и Random Forest. Качество моделей оценивалось с помощью метрик MSE и ROC-AUC.

2. Введение

Музыкальные стриминговые сервисы, такие как Spotify, собирают большое количество данных о треках. Эти данные позволяют анализировать вкусы слушателей и выявлять факторы, влияющие на популярность музыки. Одним из ключевых показателей является параметр `track_popularity`, отражающий интерес аудитории к конкретной композиции.

Основные задачи исследования:

- Проанализировать данные и подготовить их к моделированию.
- Реализовать линейную регрессию «с нуля», чтобы лучше понять принцип работы градиентного спуска.
- Исследовать, как скорость обучения влияет на процесс сходимости модели.
- Сравнить линейные модели и ансамблевые методы в задаче определения успешных треков.

3. Данные и их подготовка

3.1. Описание датасета

В работе использовался набор данных [high popularity spotify data.csv](#), содержащий 1686 треков и 29 признаков.

Для моделирования были выбраны наиболее информативные характеристики:

Акустические параметры: энергичность (energy), танцевальность (danceability), громкость (loudness), акустичность (acousticness), инструментальность (instrumentalness), живость (liveness), наличие речи (speechiness), эмоциональная окраска (valence);

Целевая переменная:

для регрессии — числовая популярность трека;

для классификации — бинарный признак («хит» / «не хит»), полученный на основе порога популярности.

3.2. Предобработка данных

Перед обучением моделей данные были подготовлены следующим образом:

Пропущенные значения были удалены. При этом размер датасета не изменился, что говорит о его хорошем качестве.

Данные разделены на обучающую и тестовую выборки в пропорции 80/20.

Все признаки были стандартизированы с помощью `StandardScaler`.

Масштабирование необходимо, так как признаки имеют разные единицы измерения, а линейные модели и градиентный спуск чувствительны к этому.

4. Модели и методы

4.1. Линейная регрессия с SGD

Для задачи регрессии была реализована собственная модель линейной регрессии, обучаемая с помощью стохастического градиентного спуска.

Основные особенности модели:

- обучение происходит по мини-батчам;
- используется функция потерь MSE;
- параметры модели обновляются на каждой итерации.
- Алгоритм включал инициализацию весов, перемешивание данных, вычисление предсказаний, расчет градиентов и обновление параметров.

4.2. Модели классификации

Для задачи классификации использовались:

- логистическая регрессия — простой и интерпретируемый линейный классификатор;

- Random Forest — ансамблевый метод, способный выявлять сложные нелинейные зависимости.

5. Эксперименты и результаты

5.1. Влияние скорости обучения

Было проверено несколько значений learning rate: 0.1, 0.01 и 0.001.

Результаты показали, что:

- слишком большое значение приводит к нестабильному обучению;
- слишком маленькое — замедляет сходимость;
- оптимальным оказалось значение 0.01, при котором ошибка стабильно уменьшалась.

Анализ коэффициентов модели показал:

- положительное влияние громкости и танцевальности;
- отрицательное влияние акустичности и инструментальности.

5.2. Сравнение моделей классификации

Качество моделей оценивалось с помощью ROC-кривых.

Логистическая регрессия показала приемлемый базовый результат.

Random Forest продемонстрировал более высокое значение ROC-AUC, что говорит о лучшем качестве классификации.

6. Итоговые выводы

В ходе работы был выполнен полный цикл анализа данных и машинного обучения.

Основные выводы:

- Реализация линейной регрессии с SGD позволила на практике понять влияние масштабирования и скорости обучения.
- На популярность треков сильнее всего влияют энергичность, громкость и танцевальность.
- Для практических задач прогнозирования популярности наиболее эффективной моделью оказался Random Forest, показавший лучшие результаты в классификации.