ПРОГНОЗИРОВАНИЕ ЧИСЛА СТРИМОВ ДЛЯ МУЗЫКАЛЬНЫХ ТРЕКОВ

КОМАНДА 5: АНДРЕЙ ЕРИЧЕВ, АНАСТАСИЯ ИВАНОВА, ЕЛИЗАВЕТА КУЗЬМИНЫХ, ПАВЕЛ СМИРНОВ

РУКОВОДИТЕЛЬ: АЛИСА ШИКАНЯН

ВШЭ Г. МОСКВА 2025

Цель и Задачи проекта

Цель

Исследовать зависимости числа стримов музыкального трека на платформе Spotify от набора признаков трека и на основании найденных зависимостей решить задачу линейной регрессии для прогнозирования числа стримов

Задачи

Выбрать датасет

Провести EDA

Проверить гипотезы

Построить модель

Дать рекомендации

Разведочный анализ данных. EDA

Источник данных

Kaggle: Spotify and YouTube Music dataset

Статистика по 10 лучшим песням различных исполнителей на Spotify и их видеоклипам YouTube

Структура данных

Датасет состоит из 16 полей и 20718 строк

Количественные признаки:

- danceability
- energy
- key
- loudness
- speechiness
- acousticness
- instrumentalness
- liveness
- valence
- tempo
- duration_min
- stream

Категориальные признаки:

- artist
- track
- album
- album_type

Очистка данных

Удалены следующие данные:

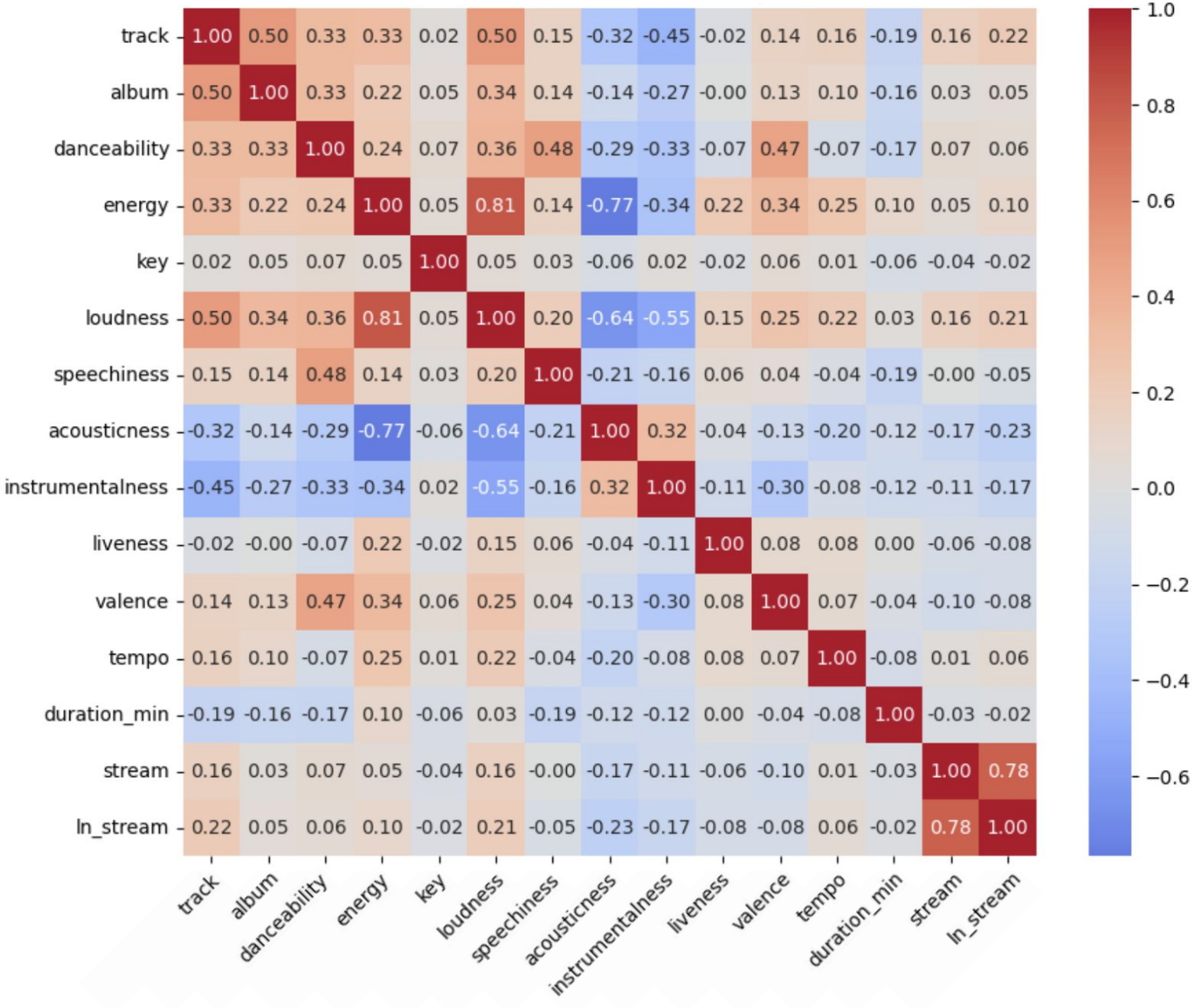
- Строки с пропусками
- Треки с белым шумом
- Аудиокниги
- Альбомы типа Compilation
- Треки длительностью более 7 минут
- Значения loudness вне допустимого диапазона (больше 0)

Удалены выбросы (по боксплоту):

- duration_min
- energy
- danceability
- tempo

Датасет после EDA и очистки состоит из 16 полей и 17968 строк

Корреляционная матрица



Разведочный анализ данных. EDA

Выводы по признакам

- Сильная положительная корреляция между **energy** и **loudness** (0.81)
- Сильная отрицательная связь между **energy** и **acousticness** (-0.77)
- Слабая положительная связь между логарифмом среднего количества прослушиваний (**in_stream**) и количеством композиций исполнителя (**track**)

Целевая переменная **stream** не имеет сильной связи с переменными.

Проверка гипотез

Треки из альбомов набирают больше стримов

Подтверждено критерием Манна–Уитни

Треки с приглашённым артистом (feat) набирают статистически значимо больше прослушиваний *

По результатам t-теста Уэлча

*доп.задание

Комбинации параметров связаны с числом стримов

Подтверждено тестом Пирсона

Количество треков у артиста не влияет на стримы

По результатам линейной регрессии

Отбор признаков для моделей

Музыкальные признаки

Внутренние характеристики композиции

Danceability	Energy	Valence
Speechiness	Instrumentalness	Acousticness
Tempo	Loudness	Duration_min

Аналитические признаки

Факторы продвижения трека

is_feat	**1** – *если трек совместный* **0** – *иначе*
is_album	**1** – *если трек входит в альбом* **0** – *иначе*

Как выбирали признаки?

- По результатам EDA и статистических тестов
- По результатам проверки гипотез

Проблемы, с которыми столкнулись перед обучением моделей

- Нет сильной связи между целевой переменной и параметрами
- Не хватает дополнительных признаков (is_feat, is_album)

Зависимая переменная

Логарифм количества прослушиваний

Построение моделей. Результаты и метрики

1. Линейная регрессия

Линейная регрессия на всех признаках
R^2 на train: 7.99%
R^2 на test: 7.93%

↓

Мультиколлинеарность
Убрали признаки с VIF >> 10
R^2 на train: 7.14%
R^2 на test: 6.99%

↓

R^2 снизился
Пробуем регуляризацию, лучше не стало
R^2 на train: 7.14%
R^2 на test: 6.99%

↓

Добавляем полиномиальные признаки
Лучше справляется Ridge

↓

Регрессия с полиномиальными признаками и Ridge-регуляризацией
R^2 на train: 10.55%
R^2 на test: 9.37%

2. Бустинговые модели

Light Gradient Boosting (LGBM)
R^2 на train: 17.3%
R^2 на test: 10.49%

↓

Подбор параметров через Optuna
n_estimators = 87, max_depth = 13
R^2 на train: 37.7%
R^2 на test: 12.91%

↓

Стекинговая модель
XGBoost + CatBoost + LGBM
R^2 на train: 49.03%
R^2 на test: 13.9%

R^2 значительно вырос как на train, так и на test…

НО! Модель явно переобучилась

Выводы

Энергичные, танцевальные треки, выпущенные как часть альбома или с участием приглашенного артиста, набирают больше прослушиваний

Целевая переменная stream не имеет сильной связи с другими переменными

Комбинации параметров имеют более сильную корреляцию с числом стримов,
чем отдельные параметры

Качество текущих моделей необходимо повышать за счет дополнительных характеристик
(жанры, год выхода композиции, дата релиза и т.д.)

Рекомендации

Текущие модели показывают **неудовлетворительные результаты**

Нет сильной связи между набором признаков трека и количеством прослушиваний

Что можно улучшить?

Датасет Yambda

Идеи новых фич

Частотность прослушиваний треков в чартах

Поведение пользователей в течение первых дней после релиза

Средняя динамика роста прослушиваний

Прогнозирование числа стримов для музыкальных треков

СЛЕВА ВЫ ВИДИТЕ НАШУ КОМАНДУ №5 ЗА РАБОТОЙ НАД ПРОЕКТОМ

	Андрей Еричев	Анастасия Иванова	Елизавета Кузьминых	Павел Смирнов
Сбор данных	Взяли готовый датасет			
EDA	+	+	+	+
Гипотезы	+	+	+	+
Построение модели	+			+
Доп. задание	+			
Презентация		+	+	+

Распределение обязанностей

	R^2 train	R^2 test	MSE train	MSE test
Линейная регрессия	7.99%	7.93%	2.4174	2.4307
Линейная регрессия без danceability, energy	7.14%	6.99%	2.4398	2.4554
Регуляризация + полиномиальные фичи	10.55%	9.37%	2.3501	2.3926
LightGBM	15.35%	9.87%	1.4913	1.5425
LightGBM + полиномиальные признаки	17.3%	10.49%	1.4741	1.5372
LGBM+Optuna	37.7%	12.91%	1.2794	1.5163
Стекинговая модель LGBM+XGBoost+CatBoost	49.03%	13.9%	1.1572	1.5077

ПРИЛОЖЕНИЕ //
Таблица с результатами моделей