Отчёт по построению моделей для предсказания популярности треков Spotify

1 Введение

В данной работе рассматривается задача предсказания популярности музыкальных композиций на основе различных признаков.

2 Кросс-валидация

Для оценки использовалась стратифицированная кросс-валидация по бинам целевой переменной (потому что распределение сильно скошено: непопулярных треков много, а популярных мало). Были протестированы разные схемы на базовой модели, результаты приведены ниже.

Таблица 1: Сравнение схем кросс-валидации

Бины	Фолды	Средний RMSE	RMSE std	Мин. RMSE	Макс. RMSE
7	7	4.1207	0.1231	3.8119	4.3643
5	5	4.1329	0.0823	3.9911	4.2727
3	3	4.1699	0.0604	4.0597	4.2672

Выбрана схема 5×5 как некоторый компромисс.

3 Описание экспериментов

3.1 Базовая модель CatBoost

В качестве отправной точки использовалась модель CatBoost со стандартными параметрами. Результаты кросс-валидации представлены в таблице.

Таблица 2: Результаты базовой модели CatBoost

Метрика	Значение
RMSE	4.1329
RMSE std	0.0823

3.2 Добавление новых признаков и их обработка

На основе анализа распределений и основных статистик признаков train и test и таргета для улучшения качества модели были добавлены агрегированные признаки по альбому, композитору и их комбинации, а также статистики (среднее, стандартное отклонение, медиана и др.), бинарные признаки выбросов, логарифмы признаков. Произведена замена выбросов на границы квантилей. Результаты поэтапного улучшения представлены в таблице.

Для уменьшения рассогласования статистики агрегировались по объединённым данным train и test. Это применимо для одного предсказания на test для лидборда. Однако при обучении модели для многоразового применения к новым данным надо агрегировать статистики только по train.

Дальнейшее улучшение качества модели возможно за счёт применения более продвинутых методов статистики, анализа признаков и обработки данных. А также за счёт исследования и построения методов, чуть более устойчивых к небольшим рассогласованиям в распределениях, появлению новых категорий в данных и т.д.

Таблица 3: Результаты после добавления признаков и обработки

Описание	RMSE	RMSE std
Агрегаты по album, composer и album+composer	3.6141	0.0924
Агрегаты (mean, std, median) + бинарные признаки выбросов	3.6184	0.0887
Агрегаты + замена выбросов на границы квантилей	3.6075	0.0937
Агрегаты + замена выбросов + логарифмирование признаков	3.5904	0.0863
Агрегаты (mean, std, median, min, max, sum, count) + замена выбросов + логарифмирование	3.5234	0.0909
Агрегаты + замена выбросов + логарифмирование + разница/отношение к медиане	3.5774	0.0926
Агрегаты + замена выбросов + логарифмирование + ранги	3.5461	0.0960
Агрегаты (mean, std, median, min, max, sum, mad, count) + замена выбросов + логарифмирование	3.5046	0.0953
Агрегаты + замена выбросов + логарифмирование + dist_min, dist_max	3.5353	0.0950
Агрегаты + замена выбросов + логарифмирование + квантильные признаки (q25, q75, iqr)	**3.4954**	0.0916
Агрегаты + замена выбросов + логарифмирование + квантильные признаки (q25, q75, iqr) + полиномиальные признаки	3.5383	0.0952

3.3 Добавление признаков музыкальной стилистики

Были предприняты попытки добавить признаки, характеризующие музыкальную стилистику и выразительные особенности композиции. Среди них категории темпа (ритмический рисунок произведения), валентности (связано с эмоциональной окраской музыки), инструментальности (влияет на жанровую и стилистическую принадлежность музыки). Но добавление таких признаков (по одному) показывало ухудшение результатов кросс-валидации.

3.4 Оптимизация гиперпараметров CatBoost

Для улучшения качества модели была проведена минимальная настройка параметров. Результаты кросс-валидации на подобранных параметрах: RMSE: **3.4152**, RMSE std: 0.0859

Дальнейшее улучшение возможно с помощью байесовской оптимизации или случайного поиска гиперпараметров. Однако эти методы довольно требовательны в плане времени и ресурсов.

3.5 Отбор признаков

Для повышения эффективности модели был произведён отбор признаков на основе важности, вычисленной CatBoost. Рассматривались топ-100 и топ-200 признаков из 628 в контексте работы CatBoost.

Таблица 4: Результаты моделей с отобранными признаками

Набор признаков	RMSE	RMSE std
Топ-200 признаков	**3.3972**	0.0827
Топ-100 признаков	3.4075	0.0884

Дальнейшее улучшение качества модели возможно за счёт более продвинутых методов отбора признаков и более глубокого изучения их взаимодействий. Включая, в частности, байесовскую оптимизацию и генетические алгоритмы для поиска оптимального подмножества признаков.

3.6 Анализ важности признаков с помощью SHAP

Анализ топ-30 признаков методом SHAP показал следующее:

- Принадлежность трека к определённому альбому или композитору оказывает существенное влияние на предсказанную популярность.
- Агрегированные признаки по альбому и композитору входят в число наиболее значимых факторов.
- Некоторые индивидуальные признаки также оказывают влияние, однако по значимости уступают групповым агрегатам.
- Трансформации признаков оправданы, так как преобразованные признаки присутствуют среди топ-30 по важности.

Дальнейшее улучшение качества модели возможно за счёт более глубокого изучения признаков, их взаимодействий, а также особенностей конкретных треков, принадлежащих определённым альбомам или композиторам. Для этого можно применить такие методы интерпретации моделей как ICE и LIME, а также проанализировать наиболее типичные и нетипичные объекты выборки и т.д.

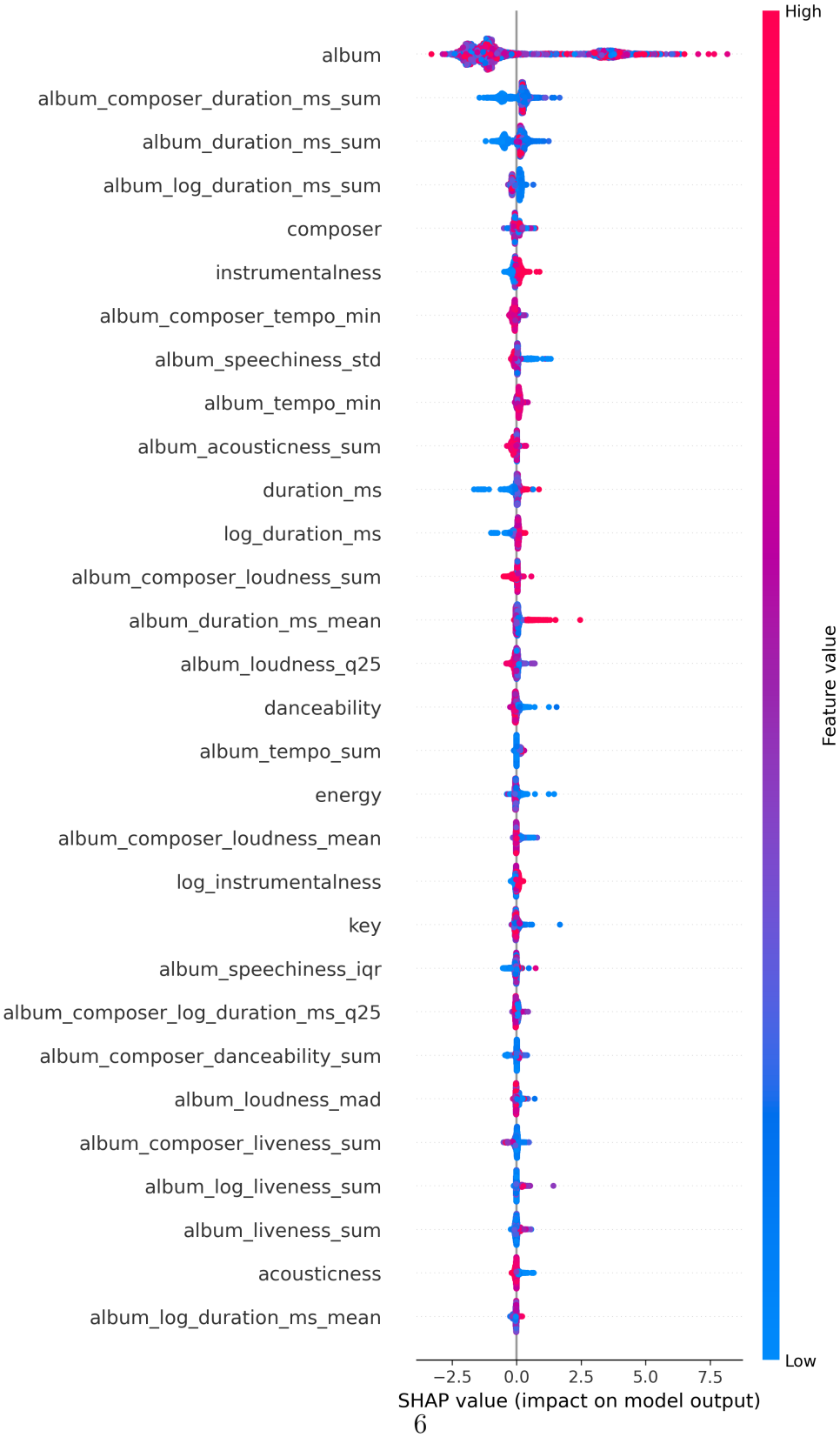

Рис. 1: SHAP summary plot: топ-30 признаков по важности для модели CatBoost. Цвет отражает значение признака: от низких (синий) к высоким (красный). По оси X - вклад признака в предсказание (SHAP value).

3.7 Модель случайного леса

Была обучена базовая модель случайного леса с достаточно стандартными параметрами. Числовые признаки использовались практически те же за небольшим исключением, категориальные кодировались таргетно с учётом стратификации по целевой переменной и использованием кросс-валидации. Композитор и альбом кодировались как единая пара. Для новых альбомов в тестовой выборке использовалось среднее по композитору.

3.8 Стэкинг

Из предыдущих моделей был собран ансамбль, в качестве мета-модели использовалась линейная регрессия с дефолтными параметрами.

Дальнейшее повышение качества ансамбля возможно за счёт расширения состава базовых моделей, тщательной настройки базовых моделей и мета-модели. В частности, можно было добавить различные алгоритмы бустинга и особо случайные деревья. Также улучшения могут быть достигнуты за счёт использования более разнообразных признаков, различных методов кодирования категориальных признаков и др.

3.9 Особенности таргета и предсказаний

Целевая переменная распределена почти экспоненциально. Для стабилизации вариации и улучшения обучения модели применялось логарифмирование таргета.

При предсказании модели иногда возникали отрицательные значения, что не имеет физического смысла в контексте популярности в данной задаче. Для устранения этой проблемы все отрицательные предсказания заменялись на ноль.

4 Заключение

При использовании выбранных конфигураций моделей CatBoost показал RMSE на тестовой выборке, равное 3.28406, случайный лес - 3.36234, а стэкинг позволил снизить ошибку до **3.27006**. В итоге был выбран стэкинг, поскольку он продемонстрировал наилучшее качество предсказаний за счёт объединения преимуществ обеих моделей.