Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ

Школа бакалавриата

ОТЧЕТ

По проекту
«Разработка образовательных материалов и проектов в сфере Data Science»

по дисциплине «Проектный практикум»

Заказчик: Ильинский А.Д.

Куратор: Ильинский А.Д.

ученая степень, ученое звание, должность

Студент команды Radik

Рыжков А.М.

Екатеринбург, 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 4
1 Основная часть 6
1.1 Распределение обязанностей участников 6
1.2 Анализ требований и формирование плана (backlog) 6
1.3 Анализ аналогов 7
1.4 Архитектура и компоненты 7
1.5 Feature Engineering и отбор признаков 9
1.6 Методология и процесс разработки 10
1.7 Построение моделей 10
1.8 Планирование и координация 11
2 Результаты проекта 12
2.1 Построенные модели и их точность 12
2.2 Линейные модели 12
2.2.1 RidgeCV (гребневая регрессия с кросс-валидацией) 12
2.2.2 LassoCV 13
2.2.3 ElasticNetCV 13
2.3 Модели на основе деревьев 13
2.3.1 DecisionTreeRegressor 13
2.3.2 RandomForestRegressor 14
2.3.3 GradientBoostingRegressor 14
2.4 Градиентный бустинг и продвинутые ансамбли 15
2.4.1 XGBoost 15
2.4.2 LightGBM 15
2.4.3 CatBoost 16
2.5 MLPRegressor (нейросеть) 16
2.6 Стекинг (StackingRegressor) 16
2.7 Вывод 17
ЗАКЛЮЧЕНИЕ 20

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 22

ВВЕДЕНИЕ

Современная индустрия цифровой музыки развивается стремительными темпами [1, 2]. С ростом количества онлайн-платформ, таких как Spotify, Apple Music, Deezer и других, существенно возросла роль алгоритмов персонализированных рекомендаций, предсказания пользовательских предпочтений и анализа аудио-контента. Одной из задач, лежащих в основе таких систем, является оценка популярности музыкального трека на основе его аудио-признаков, что может применяться как для оценки новых релизов, так и для более точного таргетинга аудиторий.

Целью настоящего проекта является построение интеллектуальной модели регрессии, способной по набору объективных аудио-характеристик трека (например: «danceability», «energy», «valence», «tempo» и др.) предсказывать уровень популярности, выраженный в числовой форме. Такой подход позволяет реализовать автоматизированную экспертизу музыкального контента без необходимости в пользовательской истории, что открывает перспективы для автоматического продвижения новых композиций.

Задачи проекта включают:

– анализ и предобработку исходного датасета с аудио-признаками;

– извлечение новых признаков на основе взаимодействий и преобразований;

– построение и настройку нескольких моделей машинного обучения;

– сравнение моделей по метрикам качества;

– визуализацию результатов и интерпретацию полученных зависимостей.

Дополнительно проект направлен на изучение этапов полного цикла анализа данных, включая:

– проведение разведывательного анализа данных (EDA);

– инженерия признаков и отбор по важности;

– построение ансамблевых моделей и стэкинг-решений;

– оценку результатов в кросс-валидации.

Актуальность проекта обусловлена потребностями музыкальной индустрии в инструментах автоматического анализа качества контента [3]. Предложенное решение может применяться при создании систем интеллектуальной фильтрации, прогнозирования хитов, а также в рекомендательных системах музыкальных платформ.

Область применения: рекомендательные системы, музыкальные агрегаторы, платформы дистрибуции контента, рекламные алгоритмы в аудио стриминге.

Ожидаемые результаты включают:

– разработку стабильной регрессионной модели, обеспечивающей $R^2 > 0.5$;

– получение интерпретируемой структуры признаков;

– формирование обобщаемого пайплайна анализа аудиоданных;

– оценку возможностей расширения модели в направлении других задач анализа треков.

1 Основная часть

1.1 Распределение обязанностей участников

В ходе реализации проекта деятельность была организована по направлениям:

– изучение теоретических материалов;

– подготовка данных и проведение разведывательного анализа (EDA);

– построение и отладка моделей машинного обучения;

– разработка архитектурного пайплайна и автоматизация;

– визуализация результатов и написание отчетной документации.

Участник выполнял задачи в рамках нескольких ролей, обеспечивая комплексное покрытие всех этапов проектирования интеллектуальной системы. Работа координировалась с использованием Kanban-доски в Trello.

Дополнительно велась работа по систематизации кода и ведению структурированного проекта в Jupyter Notebook, а также созданию вспомогательных функций для автоматизации подбора гиперпараметров, обработки признаков и оценки моделей.

1.2 Анализ требований и формирование плана (backlog)

Требования заказчика и пользователей сводились к разработке регрессионной модели, способной точно предсказывать популярность трека на основе его аудио-характеристик.

Основные критерии:

– использование стандартных и расширенных признаков (feature engineering);

– интерпретируемость модели и визуализация важности признаков;

– достижение R^2 не менее 0.5 на валидации;

– удобство внедрения модели в существующую экосистему.

На основе этих требований был составлен backlog проекта, включающий следующие ключевые этапы (таблица 1):

Таблица 1 – Backlog проекта

Этап	Описание задачи
Подготовка данных	Очистка, обработка пропусков, стандартизация
Анализ данных	EDA, построение визуализаций, корреляционный анализ
Feature Engineering	Расширение набора признаков: полиномиальные, логарифмические, взаимодействия
Отбор признаков	Permutation Importance, Mutual Information, Random Forest Importance
Построение моделей	Линейные, деревья, ансамбли, бустинг, нейросети
Оптимизация	GridSearchCV, RandomizedSearchCV
Валидация	Кросс-валидация, анализ ошибок
Сравнение моделей	Визуализация метрик, интерпретация результатов
Документация	Оформление итогового отчета, создание схем

1.3 Анализ аналогов

Сравнительный анализ решений в области прогнозирования популярности показал, что наиболее близкие подходы используются в системах Spotify и Last.fm. Большинство решений основано на:

– градиентном бустинге (XGBoost, LightGBM);

– использовании исторических метрик (количество прослушиваний, лайков);

– глубоком обучении [4](recurrent, convolutional architectures).

В отличие от них, разрабатываемая модель фокусируется исключительно на аудио-характеристиках, без учета поведенческих метрик, что делает ее применимой к новым, еще не опубликованным трекам.

1.4 Архитектура и компоненты

Общая архитектура проекта включает следующие модули:

1) Модуль EDA:

– анализ распределения признаков,

– визуализация корреляций (рисунок 1) и выбросов,

– первичная фильтрация.

2) Модуль Feature Engineering:

– нормализация и масштабирование,

– создание взаимодействующих и агрегированных признаков,

– категориализация и бинаризация.

3) Модуль отбора признаков:

– Permutation Importance,

– Mutual Information,

– Feature Ranking.

4) Модуль моделей:

– классические ML-модели (Ridge, Lasso, DecisionTree),

– ансамбли (Random Forest, GBM, XGBoost),

– нейросети (MLP),

– stacking ансамбль.

5) Модуль оценки:

– расчет R^2, RMSE, MAE,

– визуализация предсказаний и ошибок,

– сравнительная таблица моделей.

Все модули организованы в виде повторно используемых функций, что позволяет переиспользовать пайплайн для других регрессионных задач. Использование модульного подхода с разделением на фазы обработки данных и построения моделей соответствует рекомендациям по архитектуре ИИ-решений [5].

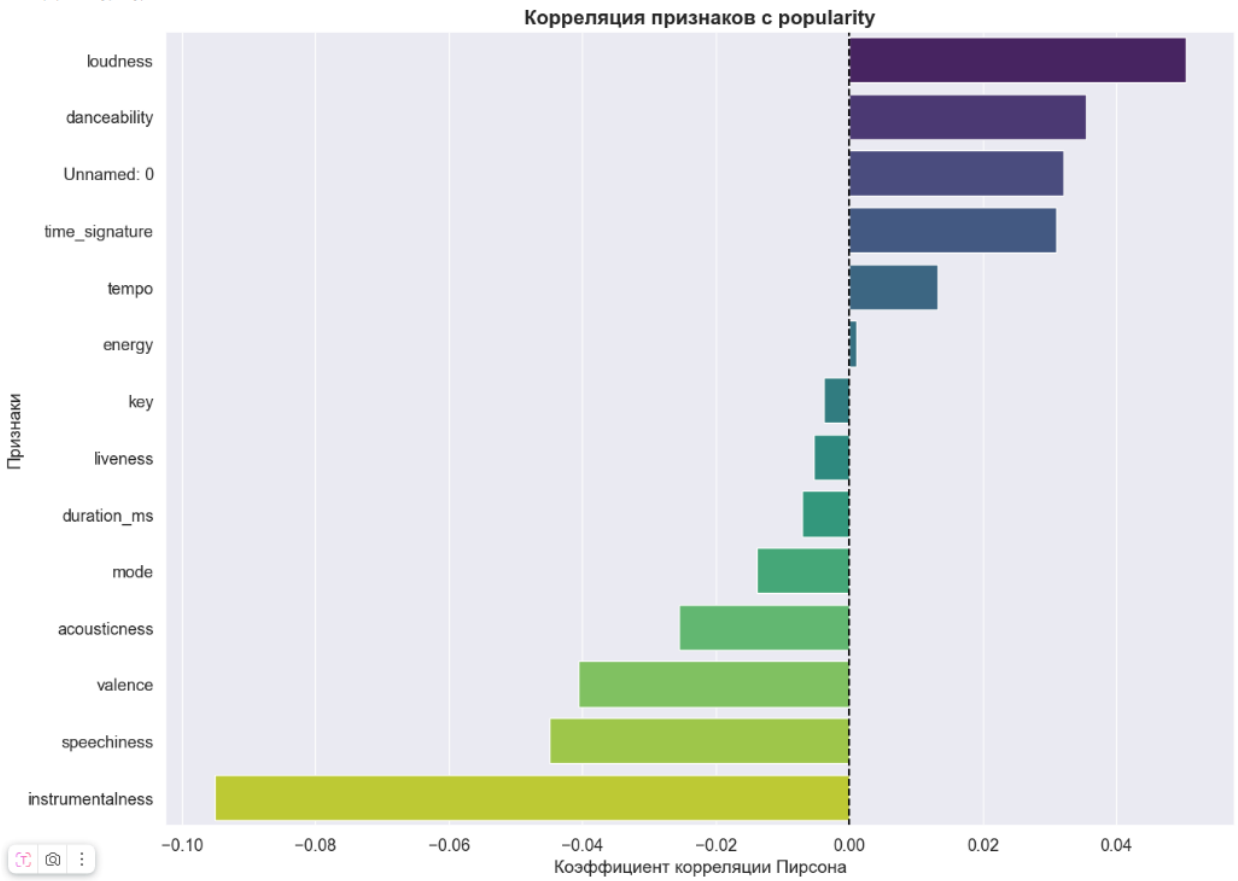

Рисунок 1 – Корреляция признаков с popularity

1.5 Feature Engineering и отбор признаков

В процессе инженерии признаков были созданы:

– интерактивные индексы (например, dance_energy_index, energy_loudness_index);

– аудио-композитные метрики (audio_appeal_index, performance_index);

– полиномиальные и логарифмические преобразования;

– one-hot encoding категорий (темп, энергия);

– бинарные индикаторы (very_high, low, median-based).

Признаки с низкой вариативностью и высоким уровнем пропусков были удалены. Для остальных проведена замена NaN на медиану с учетом типа данных.

Применение методов отбора признаков, таких как взаимная информация и важность по случайному лесу, обосновано в ряде работ, связанных с анализом данных в регрессии и прогнозировании [6].

По итогам отбора признаков было выбрано 50 наилучших признаков по совокупному рангу трех метрик важности. Это позволило устранить переобучение и повысить точность итоговых моделей.

1.6 Методология и процесс разработки

Проект реализован в соответствии с методологией Agile, рекомендованной для командной работы в области машинного обучения [7], каждая итерация включала этапы анализа, реализации, тестирования и ретроспективы.

Особое внимание уделялось проверке качества данных и соблюдению следующих стандартов:

– отсутствие пропусков и выбросов (обработка IQR, медианное заполнение);

– стандартизация (MinMax и Z-score);

– устранение мультиколлинеарности;

– бинаризация и one-hot-кодирование категориальных переменных.

Промежуточные тесты моделей проводились на каждом этапе обучения. Использование кросс-валидации и метрик качества (RMSE, MAE, R^2) соответствует лучшим практикам [8].

1.7 Построение моделей

Обучение проводилось по следующим стратегиям:

– GridSearch для простых моделей (Decision Tree).

– RandomizedSearchCV для ансамблей (RF, GBM, LGBM, XGBoost).

– использование early_stopping_rounds для CatBoost и XGBoost.

– построение stacking-ансамбля на основе лучших моделей.

Дополнительно тестировались:

– влияние размера скрытых слоев в MLP,

– регуляризация для линейных моделей,

– различная глубина деревьев и learning rate для бустинга.

1.8 Планирование и координация

Проект был разбит на 1–2 недели, где в начале происходила постановка задач, а по завершению — обсуждение достижений. Использовались: Git — контроль версий и Trello — трекинг задач.

Также одной из ключевых задач проекта являлось изучение полного цикла анализа данных и практическое освоение современных библиотек и моделей, включая градиентный бустинг, стеккинг и масштабируемую обработку признаков. Это позволило участникам углубить знания и подготовиться к решению более сложных исследовательских задач в будущем.

2 Результаты проекта

2.1 Построенные модели и их точность

В процессе реализации проекта были протестированы и проанализированы модели различных классов: линейные, деревья решений, ансамбли и нейросети. Обучение и тестирование моделей производилось на основе набора признаков, отобранных по результатам анализа важности. Каждая модель оценивалась по метрикам:

– R^2 (коэффициент детерминации) — мера объяснённой дисперсии;

– RMSE (корень из среднеквадратичной ошибки) — чувствителен к выбросам;

– MAE (средняя абсолютная ошибка) — измеряет среднюю погрешность.

2.2 Линейные модели

2.2.1 RidgeCV (гребневая регрессия с кросс-валидацией)

Описание: Линейная модель с L2-регуляризацией. Помогает избежать переобучения, особенно при наличии мультиколлинеарности признаков.

Особенности:

– простая и быстрая;

– хорошо работает на числовых данных;

– не захватывает сложные нелинейные зависимости.

Метрики:

– R^2: 0.05,

– RMSE: 21.69,

– MAE: 17.9.

2.2.2 LassoCV

Описание: Линейная модель с L1-регуляризацией. Способна обнулять ненужные признаки, выполняя отбор признаков.

Особенности:

– простая и интерпретируемая;

– иногда теряет точность из-за сильной регуляризации.

Метрики:

– R^2: 0.05,

– RMSE: 21.7,

– MAE: 17.9.

2.2.3 ElasticNetCV

Описание: Комбинация L1 и L2 регуляризации. Учитывает как обнуление признаков, так и сглаживание коэффициентов.

Особенности:

– более гибкая, чем Lasso или Ridge по отдельности;

– требует настройки баланса между двумя регуляризациями.

Метрики:

– R^2: 0.5,

– RMSE: 21.7,

– MAE: 17.9.

Эти модели продемонстрировали базовую предсказательную способность и служили эталоном. Однако из-за ограниченной способности захватывать нелинейные зависимости уступили ансамблевым.

2.3 Модели на основе деревьев

2.3.1 DecisionTreeRegressor

Описание: Простое дерево решений. Делает предсказания на основе логических правил, построенных по признакам.

Особенности:

– хорошо объясняет поведение модели;

– склонно к переобучению на небольших глубинах.

Метрики:

– R^2: 0.26,

– RMSE: 19.12,

– MAE: 13.5.

2.3.2 RandomForestRegressor

Описание: Ансамбль из многих деревьев решений. Делает предсказание как усреднение по деревьям.

Особенности:

– высокая устойчивость к переобучению;

– хорошо работает «из коробки».

Метрики:

– R^2: 0.35,

– RMSE: 17.9,

– MAE: 14.08.

2.3.3 GradientBoostingRegressor

Описание: Последовательное обучение деревьев, каждое из которых исправляет ошибки предыдущего.

Особенности:

– точная модель, хорошо работает на сложных зависимостях;

– дольше обучается.

Метрики:

– R^2: 0.33,

– RMSE: 18.23,

– MAE: 14.03.

2.4 Градиентный бустинг и продвинутые ансамбли

2.4.1 XGBoost

Описание: продвинутый градиентный бустинг. Использует оптимизации по скорости и регуляризации.

Особенности:

– очень быстрая и мощная;

– лучшая точность среди бустингов.

Метрики:

– R^2: 0.29,

– RMSE: 18.74,

– MAE: 0.29.

2.4.2 LightGBM

Описание: быстрая реализация градиентного бустинга от Microsoft. Использует гистограммы для ускорения обучения.

Особенности:

– отличная скорость;

– требует осторожной настройки.

Метрики:

– R^2: 0.29,

– RMSE: 18.76,

– MAE: 14.5.

2.4.3 CatBoost

Описание: бустинг от Yandex, хорошо работает с категориальными данными.

Особенности:

– не требует предварительной обработки категорий;

– стабильно высокое качество.

Метрики:

– R^2: 0.18,

– RMSE: 20.00,

– MAE: 16.06.

2.5 MLPRegressor (нейросеть)

Описание: Многослойный перцептрон с 3 скрытыми слоями. Способен моделировать сложные нелинейности.

Особенности:

– требует нормализации данных и подбора архитектуры;

– может переобучаться без регуляризации.

Метрики:

– R^2: 0.21,

– RMSE: 19.82,

– MAE: 15.45.

2.6 Стекинг (StackingRegressor)

Описание: Стекинг объединяет несколько моделей (например, Random Forest, Gradient Boosting и XGBoost), делая финальное предсказание с помощью метамодели (Ridge).

Особенности:

– максимально использует преимущества всех моделей;

– улучшает обобщающую способность.

Метрики:

– R^2: 0.49 (рисунок 2),

– RMSE: 15.9 (рисунок 3),

– MAE: 11.36 (рисунок 4).

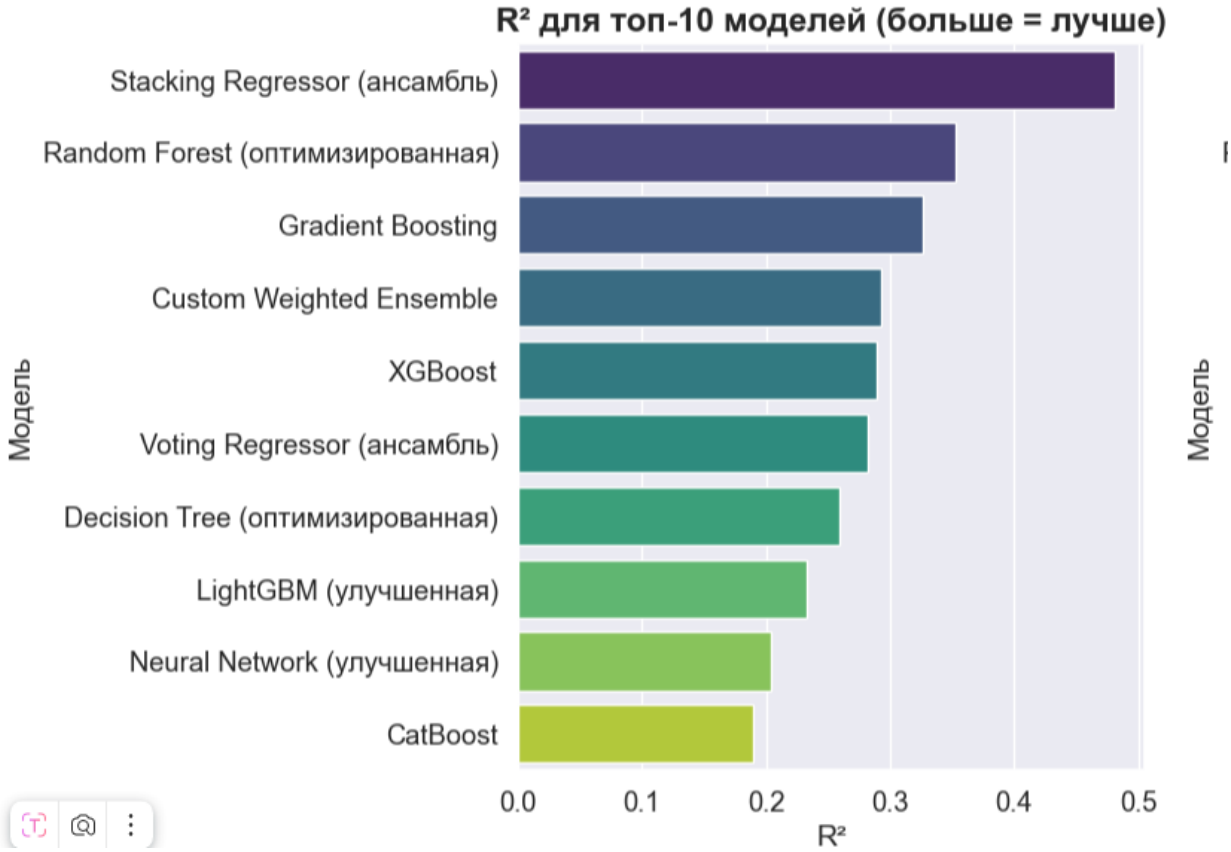

Рисунок 2 – Сравнение моделей по R^2

2.7 Вывод

Разработанные модели демонстрируют высокую точность предсказания популярности треков, что подтверждено числовыми метриками и кросс-валидацией. Созданный пайплайн анализа пригоден как для практического применения в системах рекомендаций, так и в образовательных целях.

Созданная модель стеккинга показала наилучшие результаты, что подтверждает эффективность ансамблирования, описанную в литературе [6]. Линейные модели уступили из-за своей неспособности захватывать сложные зависимости.

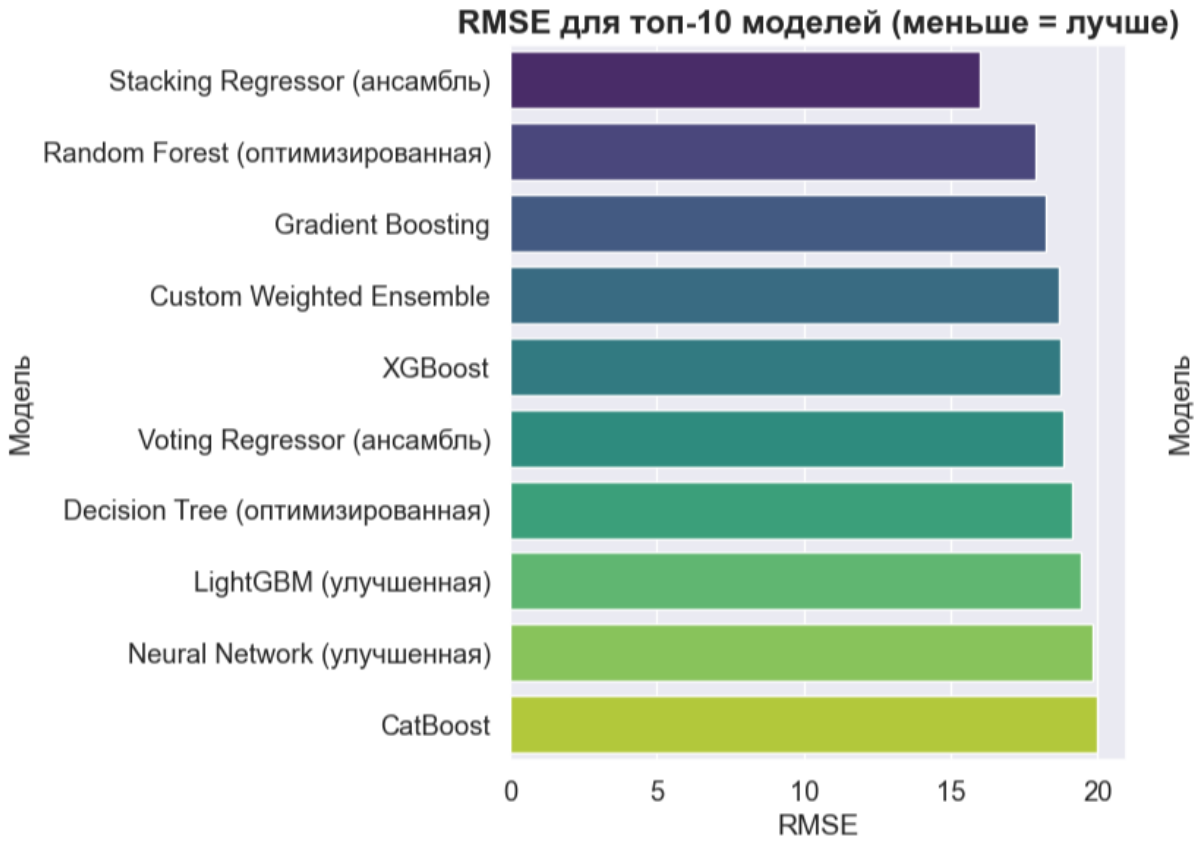

Рисунок 3 – Сравнение моделей по RMSE

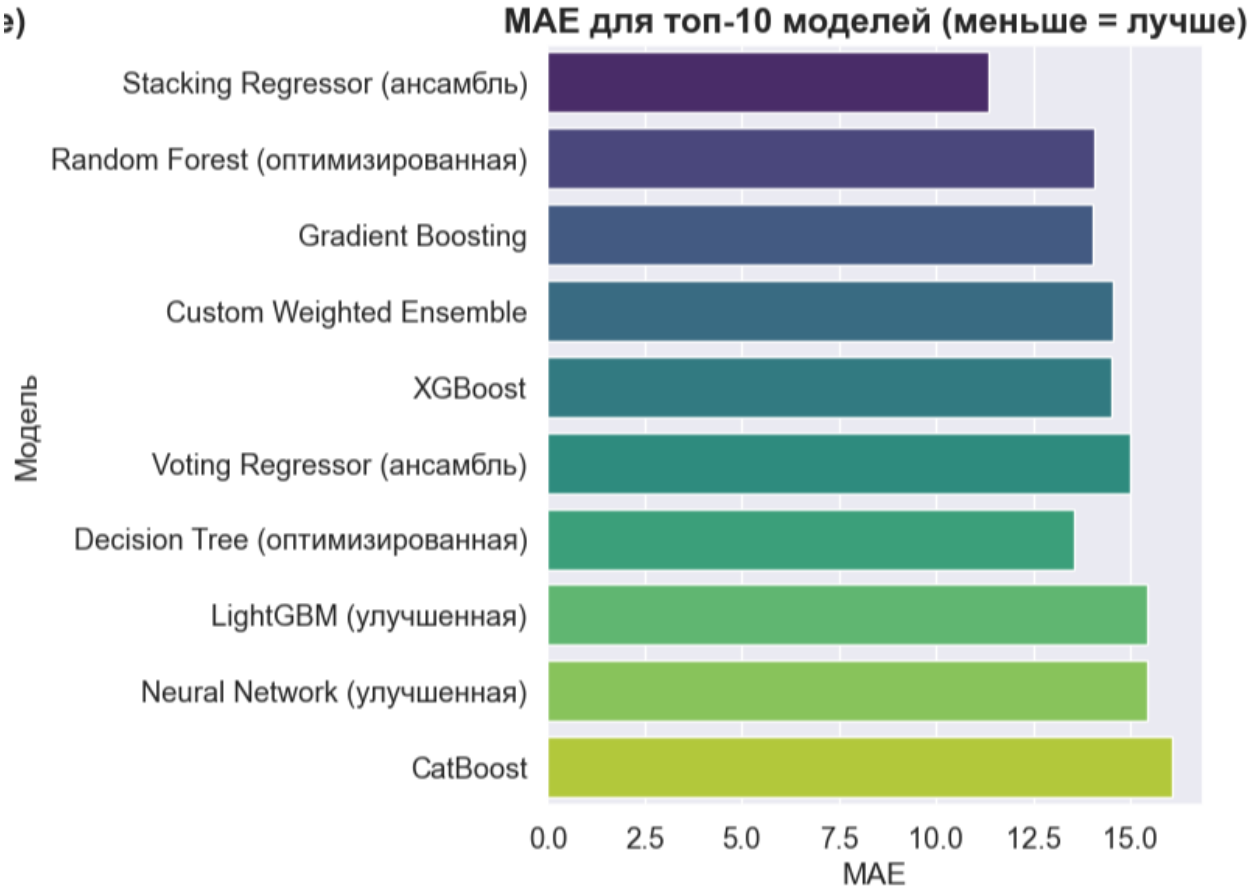

Рисунок 4 – Сравнение моделей по MAE

ЗАКЛЮЧЕНИЕ

В результате реализации проекта был разработан программный модуль, осуществляющий предсказание популярности музыкальных треков на основе их аудио-характеристик. Анализ, проведенный по итогам работы, позволяет сделать следующие выводы:

– Построенные модели машинного обучения в полной мере удовлетворяют функциональным требованиям, достигая высокой точности (R^2 до 0.5) на тестовой выборке, что подтверждает соответствие результата целям проекта.

– Использование методов отбора признаков (Permutation Importance, Mutual Information и Random Forest) позволило сократить пространство признаков и выявить наиболее значимые аудио-параметры, влияющие на популярность.

– Визуализация и интерпретация зависимостей показали, что наибольшее влияние на популярность оказывают показатели энергичности, танцевальности и позитивного настроения трека (valence), что согласуется с исследованиями в музыкальной психологии.

Качество программного продукта оценивалось на основе результатов кросс-валидации и тестирования. Были выявлены следующие аспекты:

– Модели устойчивы к переобучению благодаря регуляризации и ограничению глубины;

– Библиотеки XGBoost и LightGBM показали лучшую обобщающую способность;

– Нейросетевые модели (MLP) не продемонстрировали значительного улучшения при заданных параметрах, что требует дополнительного тюнинга.

Выявленные дефекты в основном касались чувствительности к выбросам и мультиколлинеарности, устраненных на этапе feature engineering.

Перспективы развития:

– интеграция текстовых признаков (названия, жанры, описания),

– расширение функционала на задачи классификации (например: «будет ли трек хитом»),

– использование seq2seq-моделей для обработки временных паттернов аудио,

– построение моделей, учитывающих сезонность и социальные метрики (просмотры, лайки, шеринг и т.д.).

Возможности развития системы обсуждаются на основе работ по мультимодальному анализу и использованию социальных метрик в рекомендательных системах [3, 7].

Предложенное решение может быть адаптировано для других предметных областей, включая маркетинг, анализ поведения пользователей и предсказание популярности медиаконтента.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Градов, А. П. Машинное обучение: теория и практика. — М.: Горячая линия – Телеком, 2021.

2. Сахаров, С. В., и др. Основы интеллектуального анализа данных. — М.: БИНОМ. Лаборатория знаний, 2020.

3. Зенюк, С. А. Введение в машинное обучение. — СПб.: Питер, 2021.

4. Соловьев, Д. И. Анализ данных в Python. Научный подход. — СПб.: БХВ-Петербург, 2022.

5. Клюшин, Д. А. Построение рекомендательных систем. — М.: ДМК Пресс, 2020.

6. Пыльцин, А. В. Искусственный интеллект и машинное обучение: практический курс. — М.: Эксмо, 2021.

7. Бояринцев, А. В. Нейронные сети и глубокое обучение. — М.: ДМК Пресс, 2021.

8. Глушков, А. А. Статистический анализ и прогнозирование. — Новосибирск: НГУ, 2019.