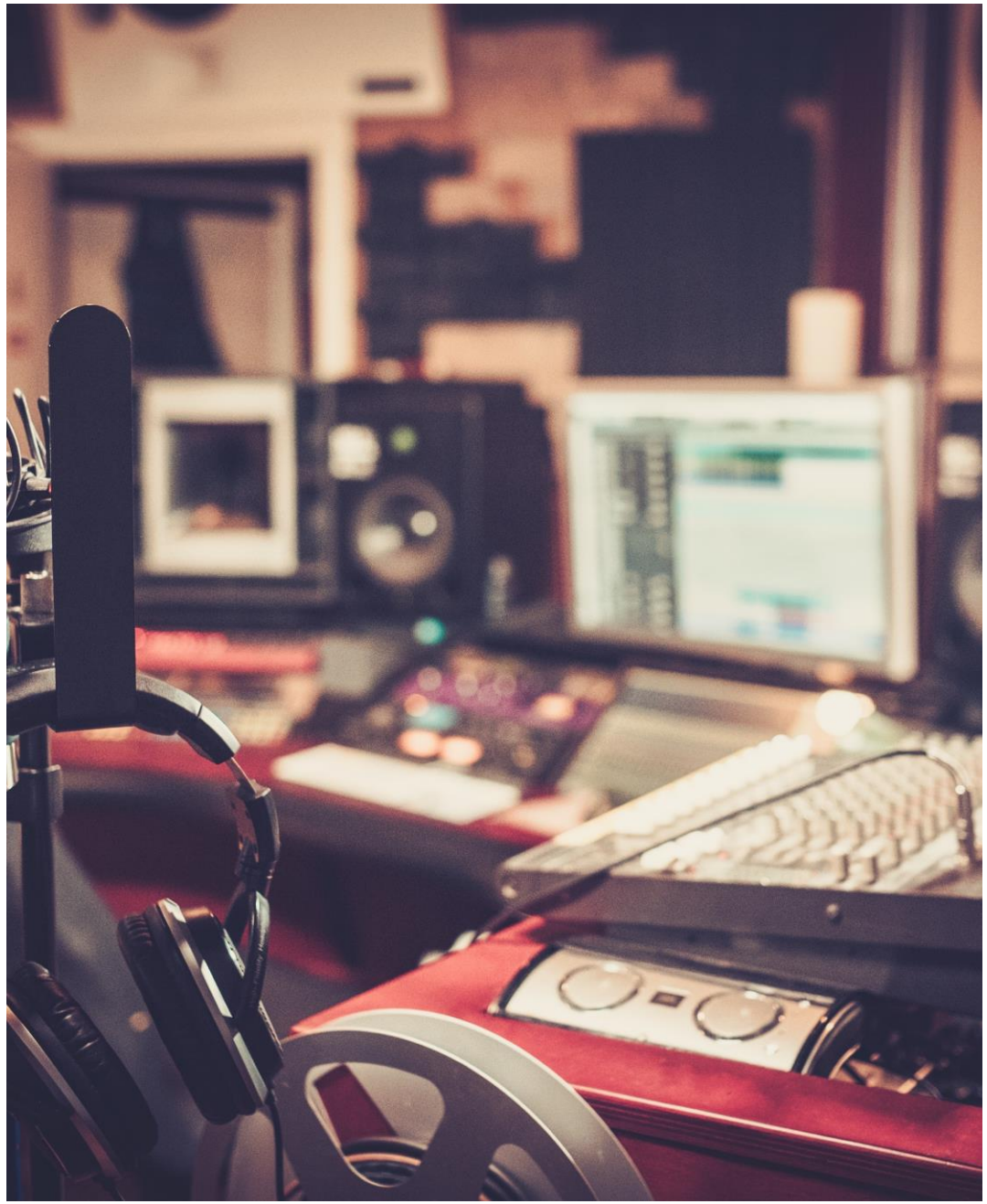
Групповой проект «Music genre prediction»

Анастасия Горевалова
Анна Булкина
Наталья Джога



ЗАДАЧА НА ПРОЕКТ

Разработать модель, позволяющую классифицировать музыкальные произведения по жанрам

ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

Качество предсказания лучшей модели, оцененное по метрике F1-меры, достигло уровня 0.5274, что говорит о хорошей предсказательной способности модели

Развернутое web-приложение (с использованием библиотеки Streamlit)

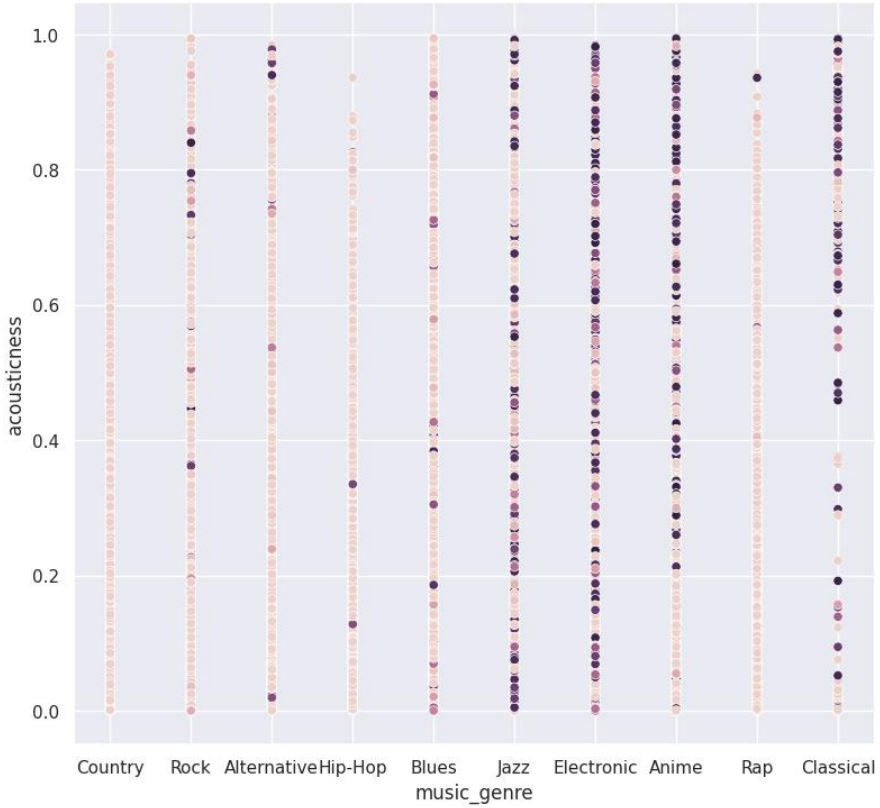


График зависимостей между "music_genre", "acousticness", "instrumentalness "

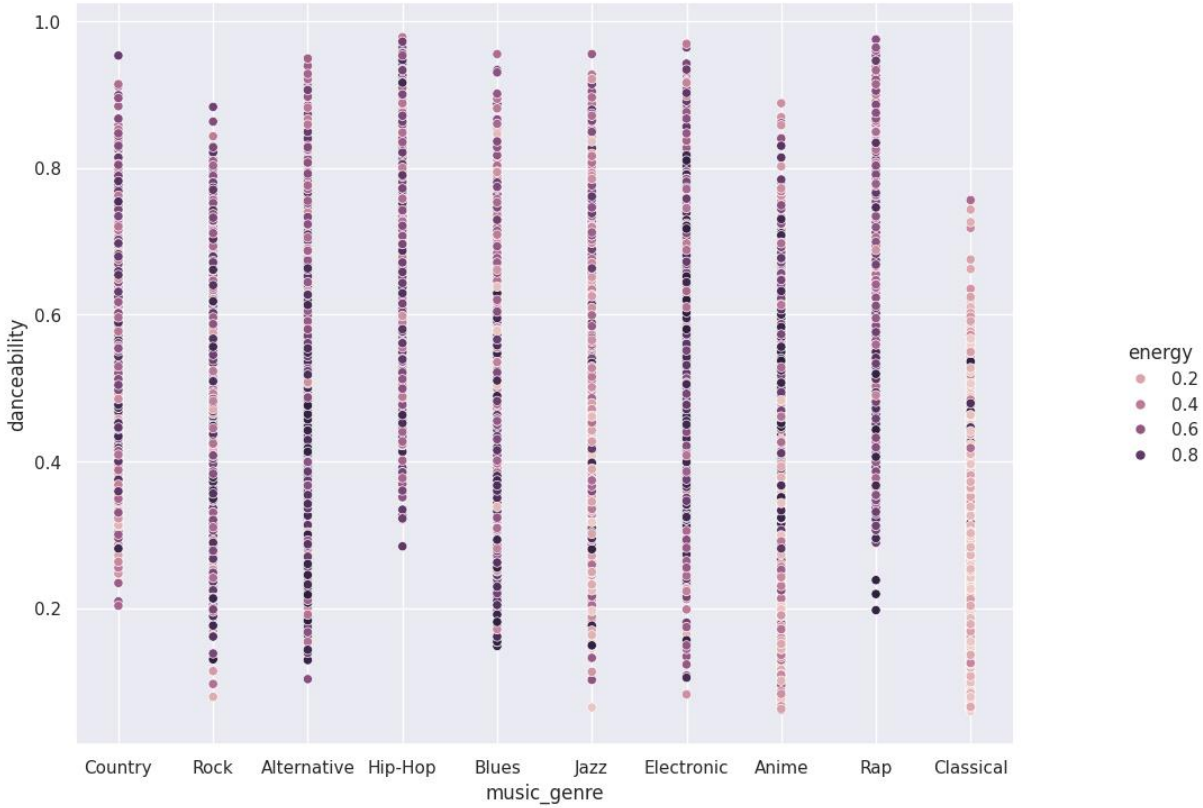


График зависимостей между "music_genre", "danceability", "energy"

После изучения различных методов обработки признаков и проведения анализа зависимостей, рассмотрели графики, иллюстрирующие взаимосвязь признаков.

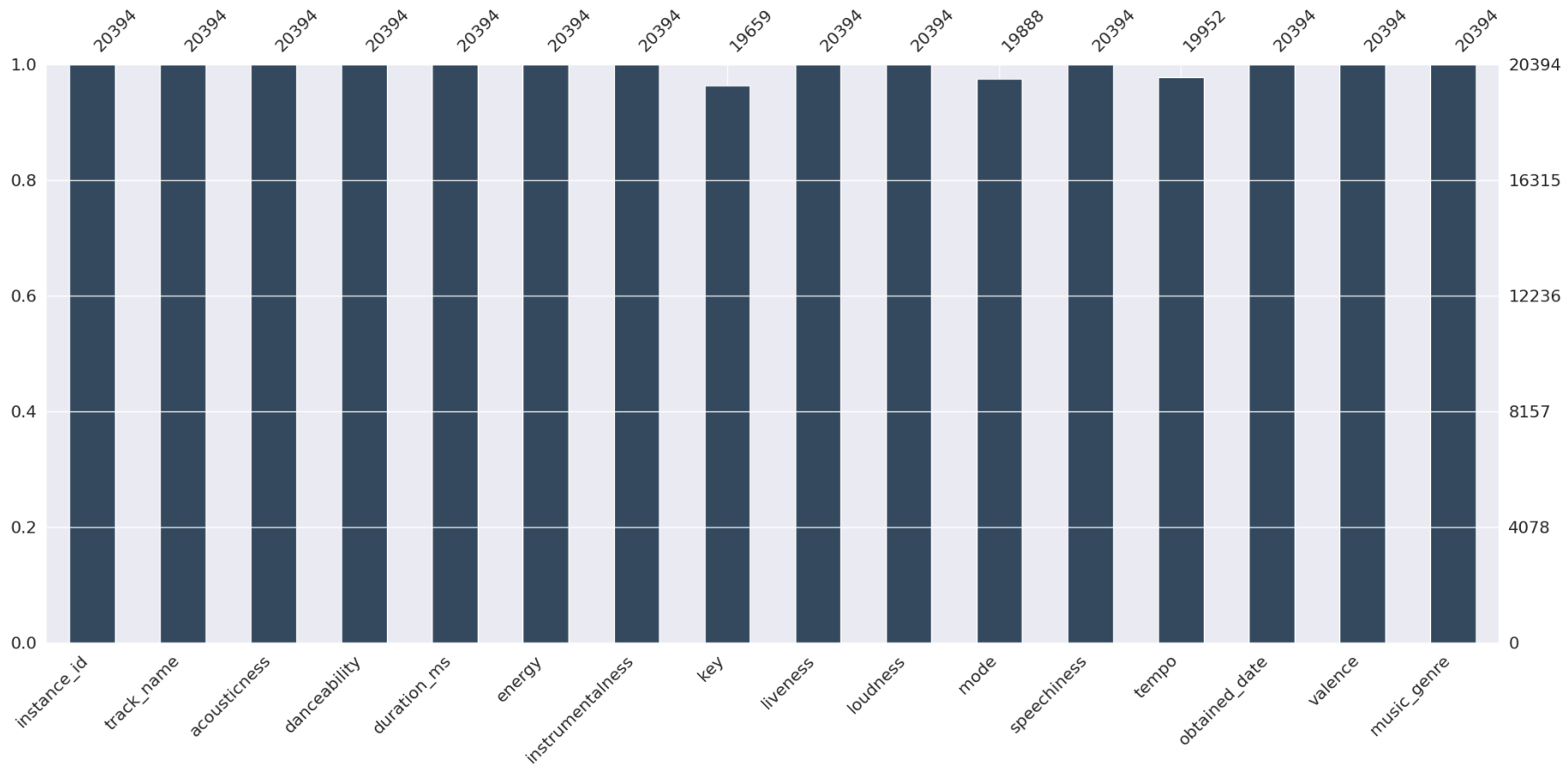
РАБОТА С ПРИЗНАКАМИ

energy*loudness	acousticness*instrumentalness
-5.911542	8.448000e-03
-5.622640	1.234440e-05
-4.200900	0.000000e+00
-4.194765	1.791180e-07
-3.191250	4.017900e-07
...	...
-6.773139	2.777600e-06
-3.749596	1.067800e-03
-3.676014	1.518900e-08
-5.116010	0.000000e+00
-1.587545	2.626850e-02

При проведении анализа данных, мы экспериментировали с добавлением новых признаков, также пытались анализировать тональность треков и использовать ее в числовом формате, однако это не принесло ожидаемых результатов и казалось нелогичным. Мы также пробовали удалить названия треков из рассмотрения, но выяснили, что это действие привело к улучшению результатов, как обнаружили у коллег.

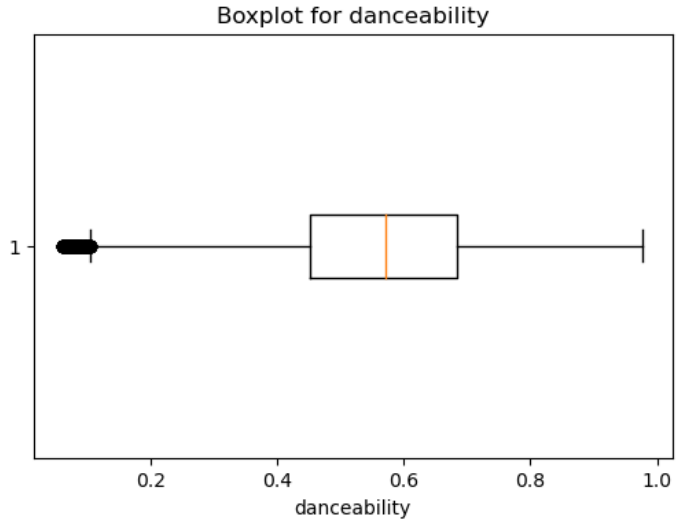
Также, пробовали убирать категориальные признаки из анализа, однако это не привело к желаемым изменениям. Опыт проведенных экспериментов нам позволил лучше понять, какие признаки влияют на результаты и какие лучше исключить для достижения лучшей модели.

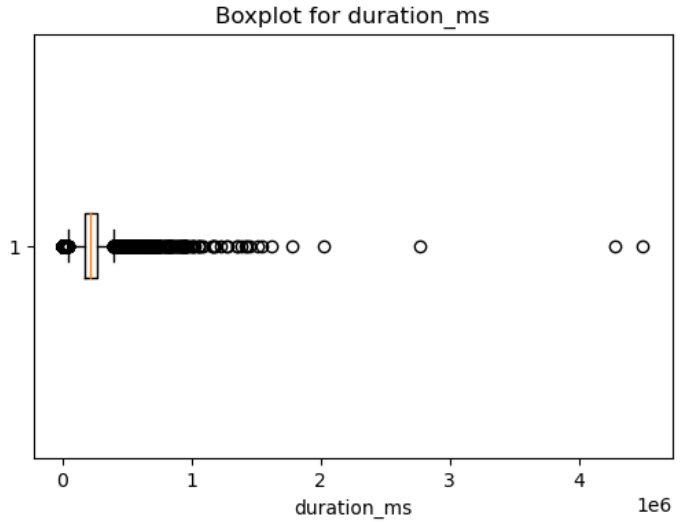
ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ



- Удаление столбцов 'instance_id' и 'obtained_date'.
- Замена значений в столбце 'track_name' на их длину.
- Удаление дубликатов.
- Замена 'Major' на 1 и 'Minor' на -1 в столбце 'mode'.
- Заполнение пропущенных значений в столбце 'mode' нулями.
- Заполнение пропущенных значений в 'tempo' 120.011.
- Заполнение пропущенных значений в 'key' значением 'Pusto'.
- Удаление строк с пропущенными значениями.

ВЫБОР МОДЕЛИ И ПАРАМЕТРОВ





В процессе анализа данных мы провели исследование с использованием ящиков с усами, удаляли выбросы, считая некоторые значения подозрительными, и заполняли их средними значениями по жанру. Затем мы создали пайплайн, в котором рассмотрели различные варианты обработки данных и выяснили, что CatBoost является одним из наиболее эффективных алгоритмов для нашей задачи.

Мы провели оптимизацию параметров CatBoost с помощью Optuna, не ограничиваясь пробегом в 100 прогонов, но результаты не улучшились. Затем, опираясь на подобранные параметры CatBoost, мы построили итоговую модель и экспортировали ее в формат pickle.

Наш исследовательский процесс позволил нам обнаружить наиболее эффективный алгоритм для задачи и построить оптимальную модель для дальнейшего использования.

РАЗВЕРНУТОЕ WEB-ПРИЛОЖЕНИЕ (С ИСПОЛЬЗОВАНИЕМ БИБЛИОТЕКИ STREAMLIT)

Наше web-приложение, разработанное с помощью библиотеки Streamlit, содержит четыре отдельные раздела, каждый из которых представляет собой уникальный функционал: «О проекте", «Графики", "Модель" и "Предсказание по песне".

СПАСИБО!

Анастасия Горевалова

Анна Булкина

Наталья Джога