музыкальные

ЖАНРЫ

предсказание...

ЗАДАЧА:

Определить музыкальный жанр

Вы сотрудник Отдела Data Science популярного музыкального стримингового сервиса. Сервис расширяет работу с новыми артистами и музыкантами, в связи с чем возникла задача - правильно классифицировать новые музыкальные треки, чтобы улучшить работу рекомендательной системы. Ваши коллеги из отдела работы со звуком подготовили датасет, в котором собраны некоторые характеристики музыкальных произведений и их жанры. Ваша задача - разработать модель, позволяющую классифицировать музыкальные произведения по жанрам.

(с)изображение сгенерировано нейросетью Kandinsky

ЭТАПЫ РАБОТЫ:

1. Обзор данных

- общая информация
- пропуски и дубликаты
- названия треков

2. Обработка данных

- замена пропусков
- удаление параметров

3. Новые признаки

- признак LAE
- высокая инструментальность
- не латинские символы

4. Обучение моделей

- CatBoostClassifier
- LGBM-Boost
- Random Forest
- XGBoost

5. Выводы

ОБЗОР ДАННЫХ:

1.1 Основная информация

Рассматриваемый датасет содержит следующую информацию:

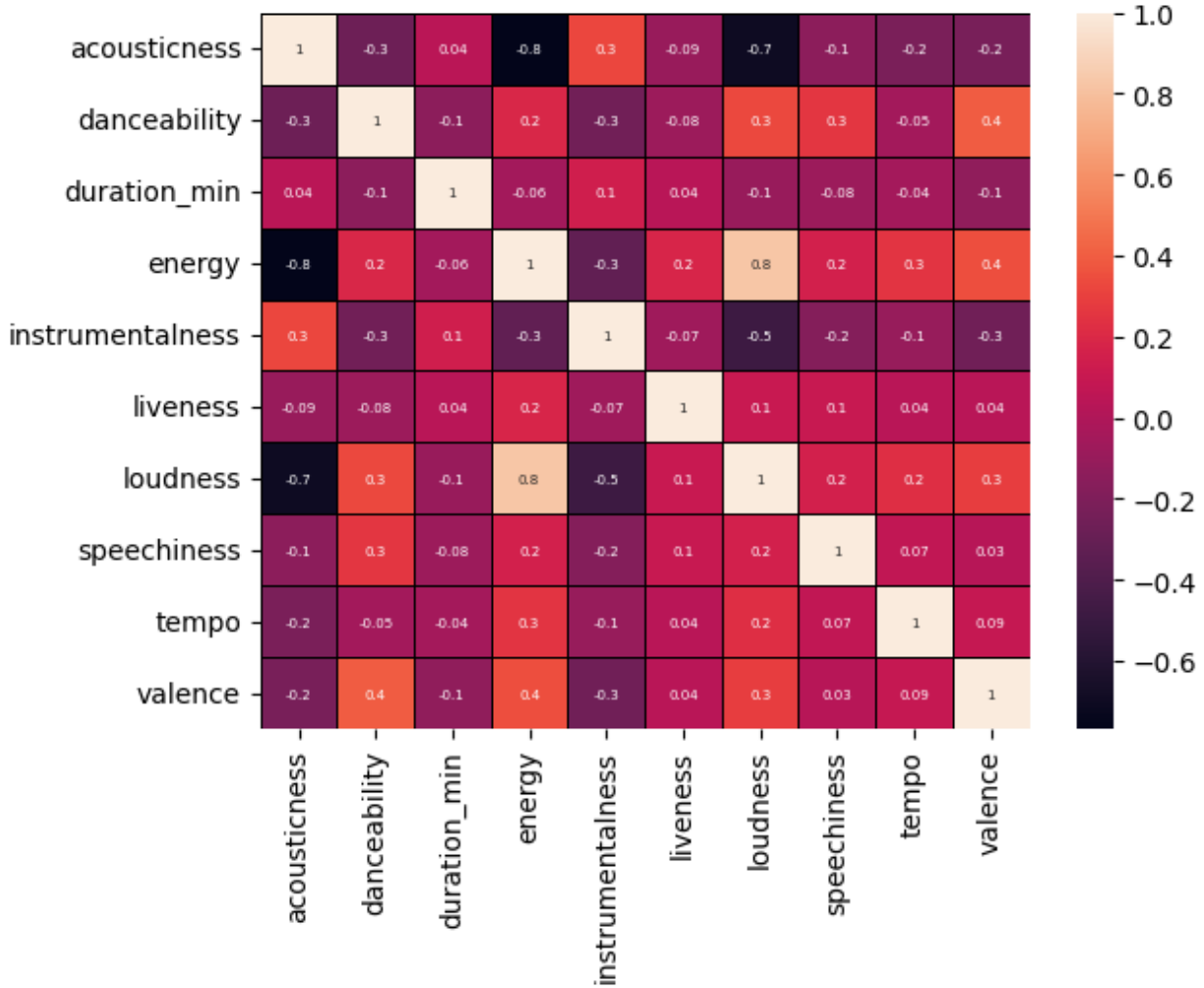
- instance_id – уникальный идентификатор трека
- track_name – название трека
- acousticness – акустичность
- danceability – танцевальность
- duration_ms – продолжительность в милисекундах
- energy – энергичность
- instrumentalness – инструментальность
- key – тональность
- liveness – привлекательность
- loudness – громкость
- mode – наклонение
- speechiness – выразительность
- tempo – темп
- obtained_date – дата загрузки в сервис
- valence – привлекательность для пользователей
- music_genre – музыкальный жанр

ОБЗОР ДАННЫХ:

1.1 Основная информация

признаки	типы	кол-во	уникальные
• instance_id	- float64	- 20394	
• track_name	- object	- 20394	- 18643
• acousticness	- float64	- 20394	
• danceability	- float64	- 20394	
• duration_ms	- float64	- 20394	
• energy	- float64	- 20394	
• instrumentalness	- float64	- 20394	
• key	- object	- 19659	- 13
• liveness	- float64	- 20394	
• loudness	- float64	- 20394	
• mode	- object	- 19888	- 3
• speechiness	- float64	- 20394	
• tempo	- float64	- 19952	
• obtained_date	- object	- 20394	- 4
• valence	- float64	- 20394	
• music_genre*	- object	- 20394	- 10

** целевая переменная*



ОБЗОР ДАННЫХ:

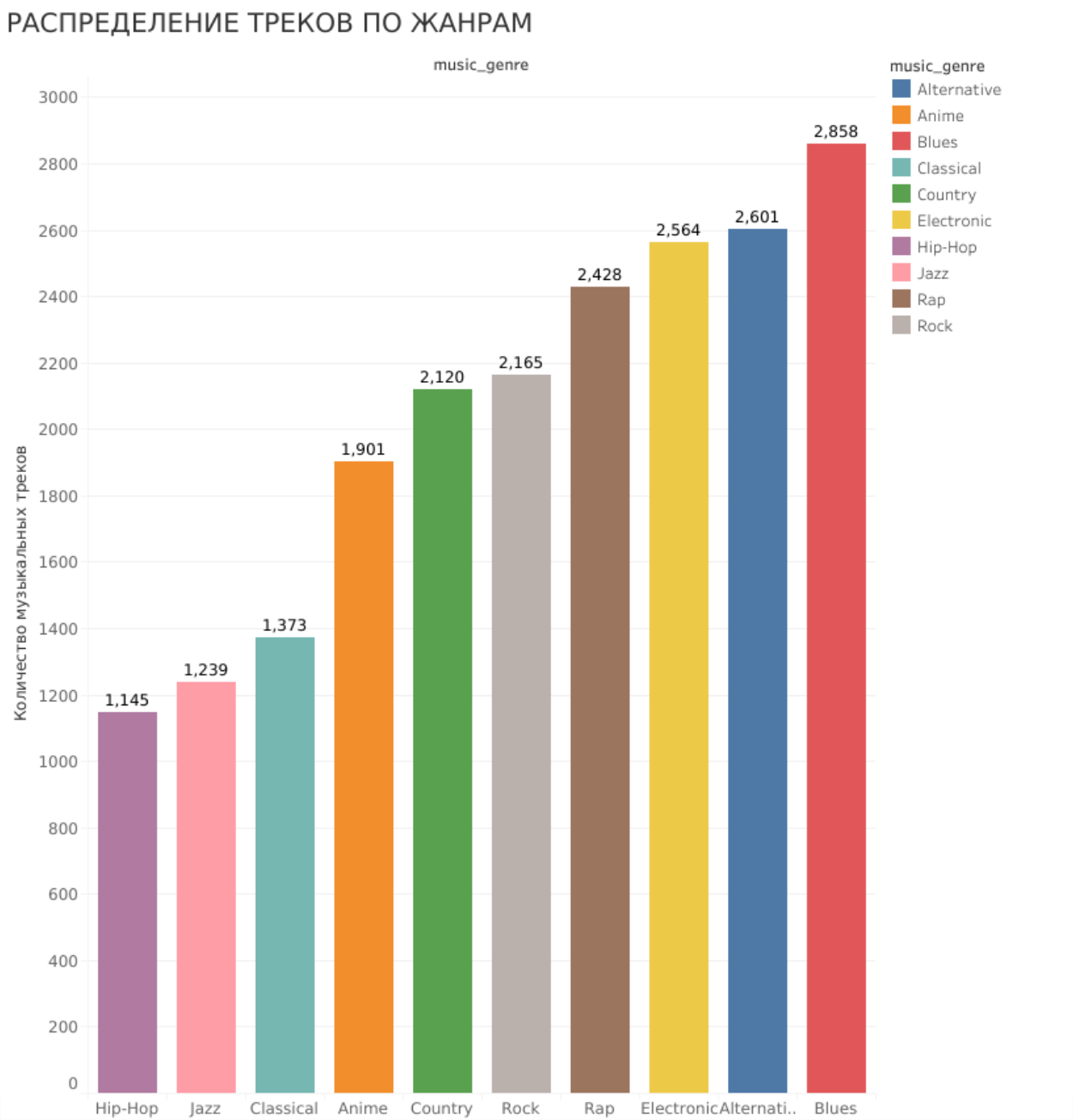
1.1 Основная информация

Всего в данных представлено **10** разных видов музыкальных жанров.

Наибольшее и наименьшее количество треков представлено в следующих жанрах по TOP-3:

- **Blues** – 2858 треков
- **Alternative** – 2610 треков
- **Electronic** – 2564 трека

- **Hip-Hop** – 1145 треков
- **Jazz** – 1239 треков
- **Classical** – 1373 трека



ОБЗОР ДАННЫХ:

1.1 Основная информация

В названиях треков встречают иероглифы и названия без использования латиницы.

жанр	лат.	не лат.	% не лат.
Alternative	2422	40	1.6%
Anime	1445	450	23.7%
Blues	2789	21	0.7%
Classical	1164	206	15%
Country	2060	12	0.6%
Electronic	2516	28	1.1%
Hip-Hop	1030	17	1.6%
Jazz	1200	24	2.0%
Rap	2265	13	0.6%
Rock	1992	25	1.2%

Наибольшее количество начертаний не на латинице встречается в жанрах anime и classical

ОБЗОР ДАННЫХ:

1.2 Пропуски в данных и дубликаты

Как мы увидели ранее из обзора датасета в наших данных отсутствуют показатели в 3-х столбца:

- **key** – пропущено **735** значения, что составляет 3.60 %
- **mode** – пропущено **506** значения, что составляет 2.48 %
- **tempo** – пропущено **442** значения, что составляет 2.17 %

Показатели **key** и **mode** – являются категорийными.
Показатель **tempo** – числовой.

Количество пропусков в целом незначительное.
Дубликаты присутствуют в названиях треков – всего **1751** шт.

Полностью совпадают по основным параметрам
(кроме жанра и индекса) - **635** треков.
Предположительно, это было отнесение трека к нескольким жанрам одновременно.

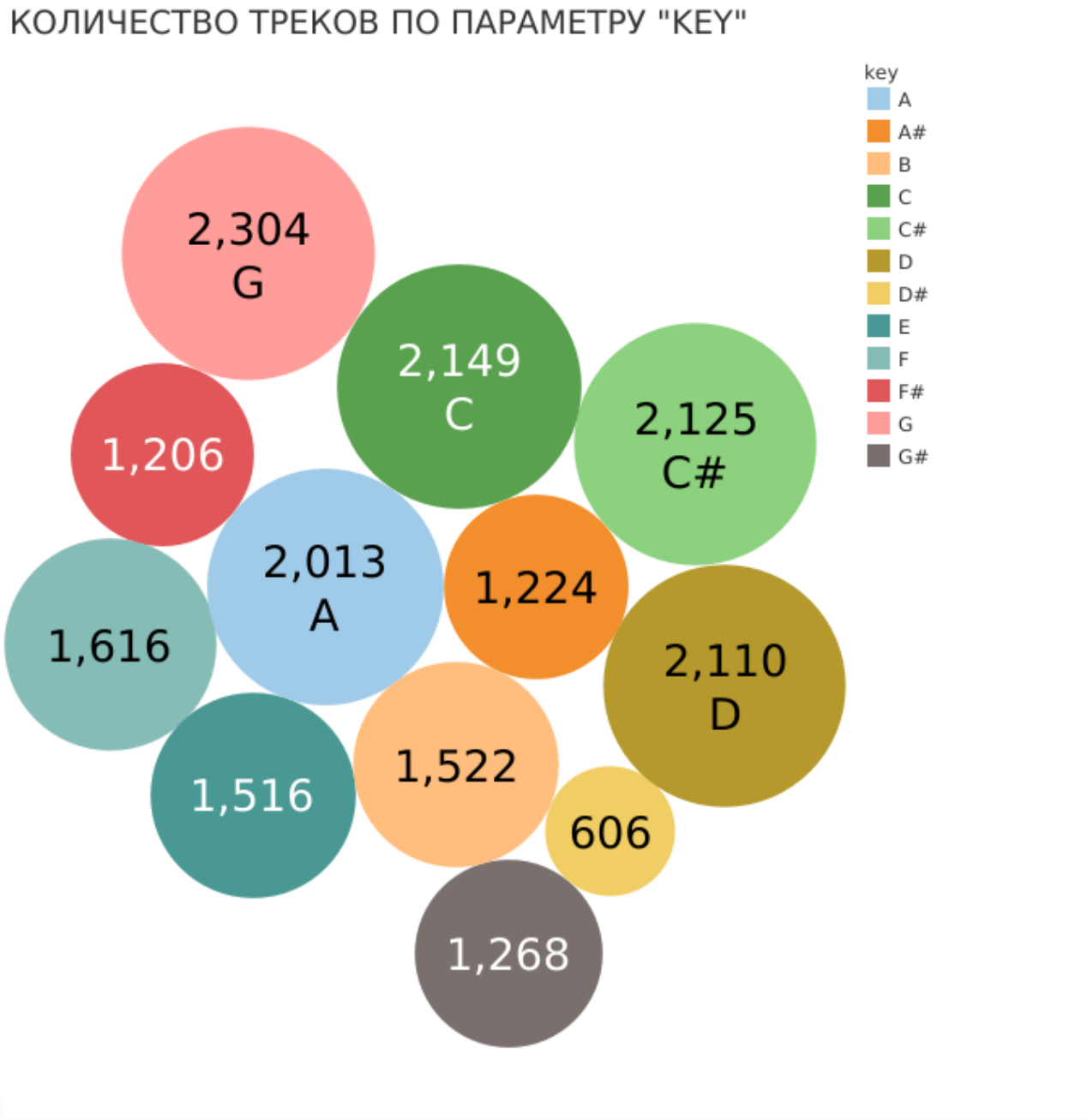
ОБРАБОТКА ДАННЫХ:

2.1 Замена пропусков

Вместо удаления строк датасета с пропущенными значениями, заполним их. Для разных признаков используем немного разные подходы.
Параметры **key** и **mode** заполним модой в зависимости от жанра.

• Alternative	G	• Major
• Anime	G	• Major
• Blues	G	• Major
• Classical	D	• Major
• Country	G	• Major
• Electronic	C#	• Major
• Hip-Hop	C#	• Major
• Jazz	F	• Major
• Rap	C#	• Major
• Rock	D	• Major

Параметр **tempo** заменим на среднее значение.



ОБРАБОТКА ДАННЫХ:

2.2 Удаление параметров

Некоторые характеристики не влияют на целевой показатель и оставлять их в датасете нет необходимости.
К таким параметрам относятся:

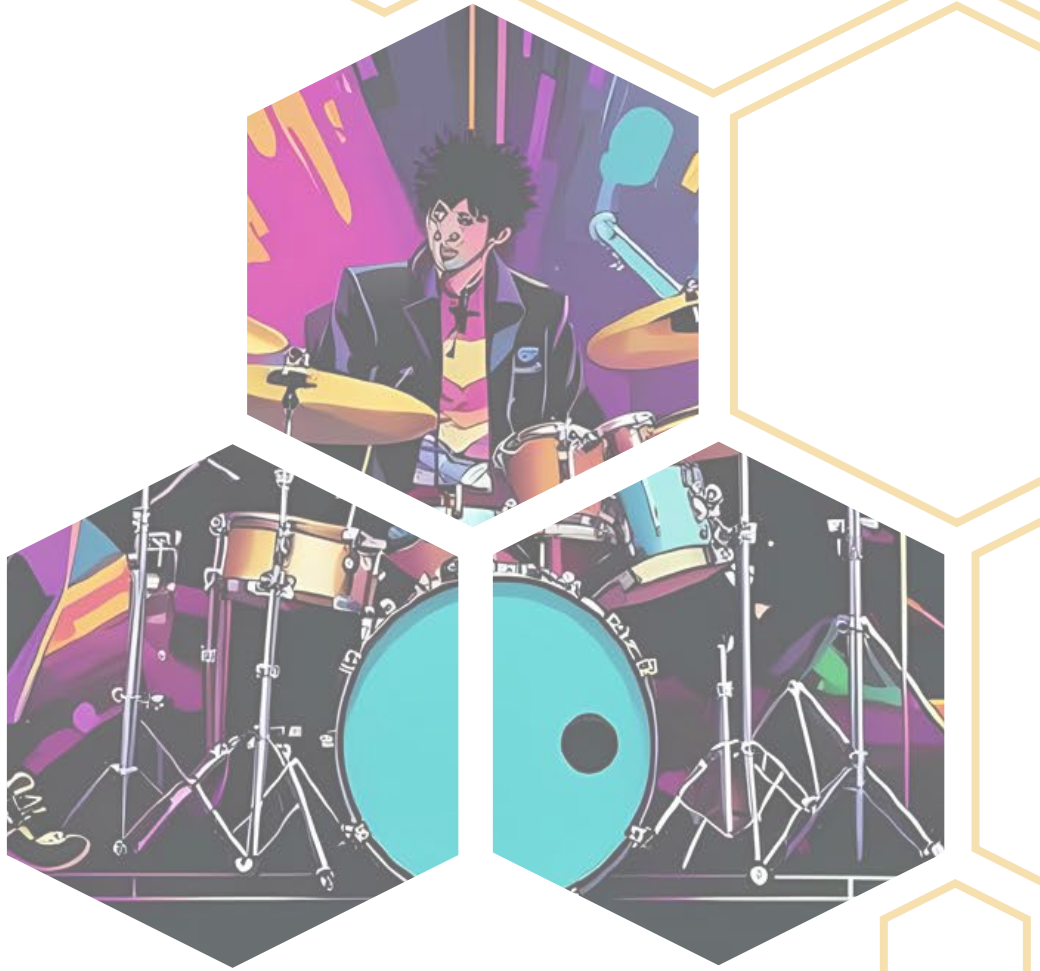
- **obtained_date** – в котором 4 уникальных значения (даты)
- **instance_id** – не влияет на предсказание жанра (индекс)

В перспективе нам будет не нужна колонка с названием треков. Названия часто содержат много «шума» (*Remaster Edition, Piano Version, Live Concert* и т.п.), качественно не влияющих (или не значительно влияющих) на целевую переменную.

- **track_name** – наименование композиции

Также были удалены **675** треков, у которых стояло несколько типов жанров. Они будут вносить путаницу в определение единого жанра.

Параметр **duration_ms** переведен в минуты, чтобы быть более сомасштабным с другими данными.



НОВЫЕ ПРИЗНАКИ:

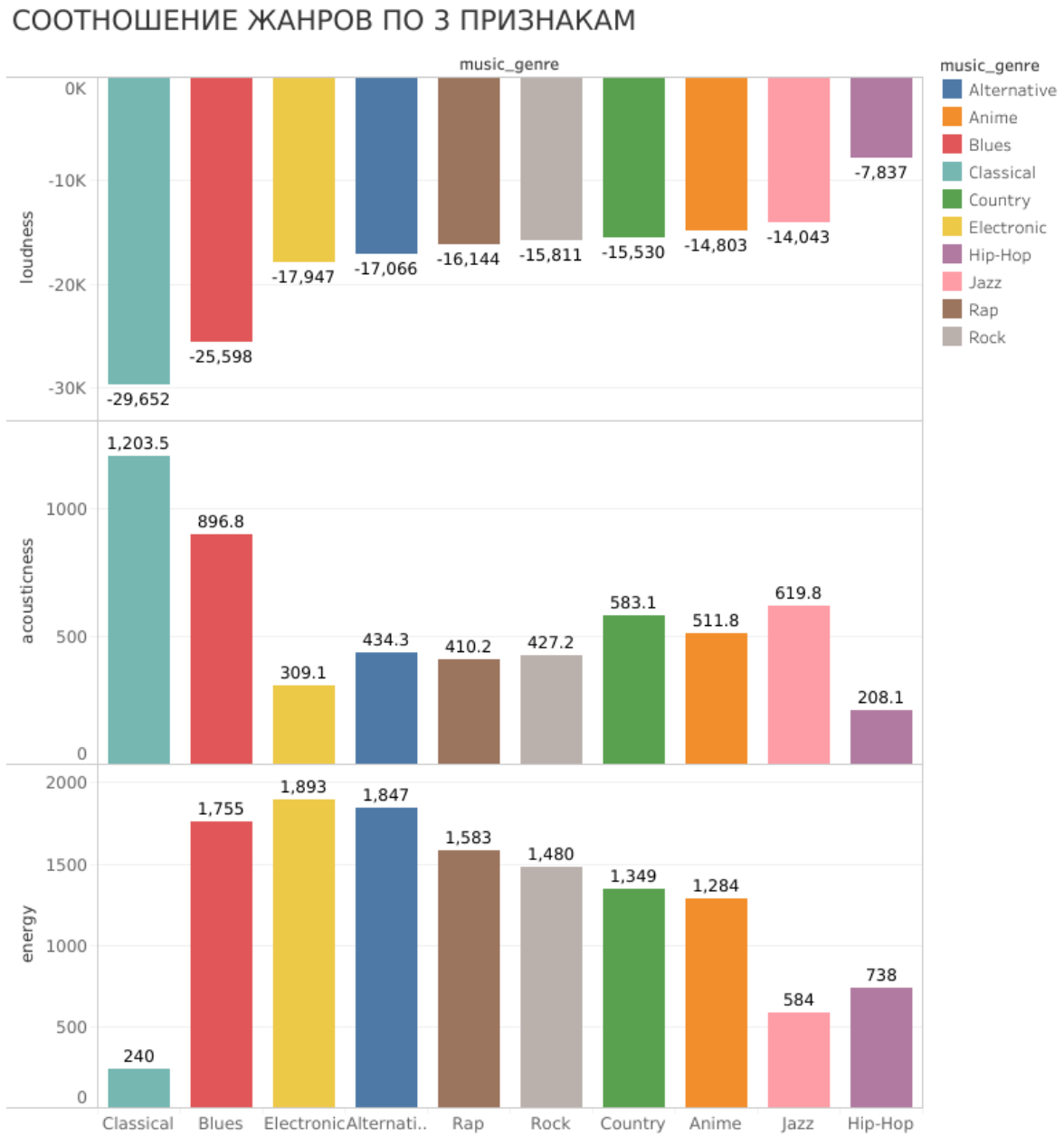
3.1 Признак LAE

При ознакомлении с данными можно обратить внимание, что признаки **loudness**, **acousticness** и **energy** сильнее характерны для композиций в жанрах:

- Blues
- Electronic
- Classical

Как мы уже выяснили ранее больше всего у нас композиций в жанре **Blues**, поэтому важно определять его более точно.
Несмотря на то, что для жанра **Classical** не характерна высотка энергичность, но зато она самая низкая.
Попробуем создать комбинированный признак **LAE**:

LAE = (loudness + acousticness) / energy



НОВЫЕ ПРИЗНАКИ:

3.2 Высока инструментальность

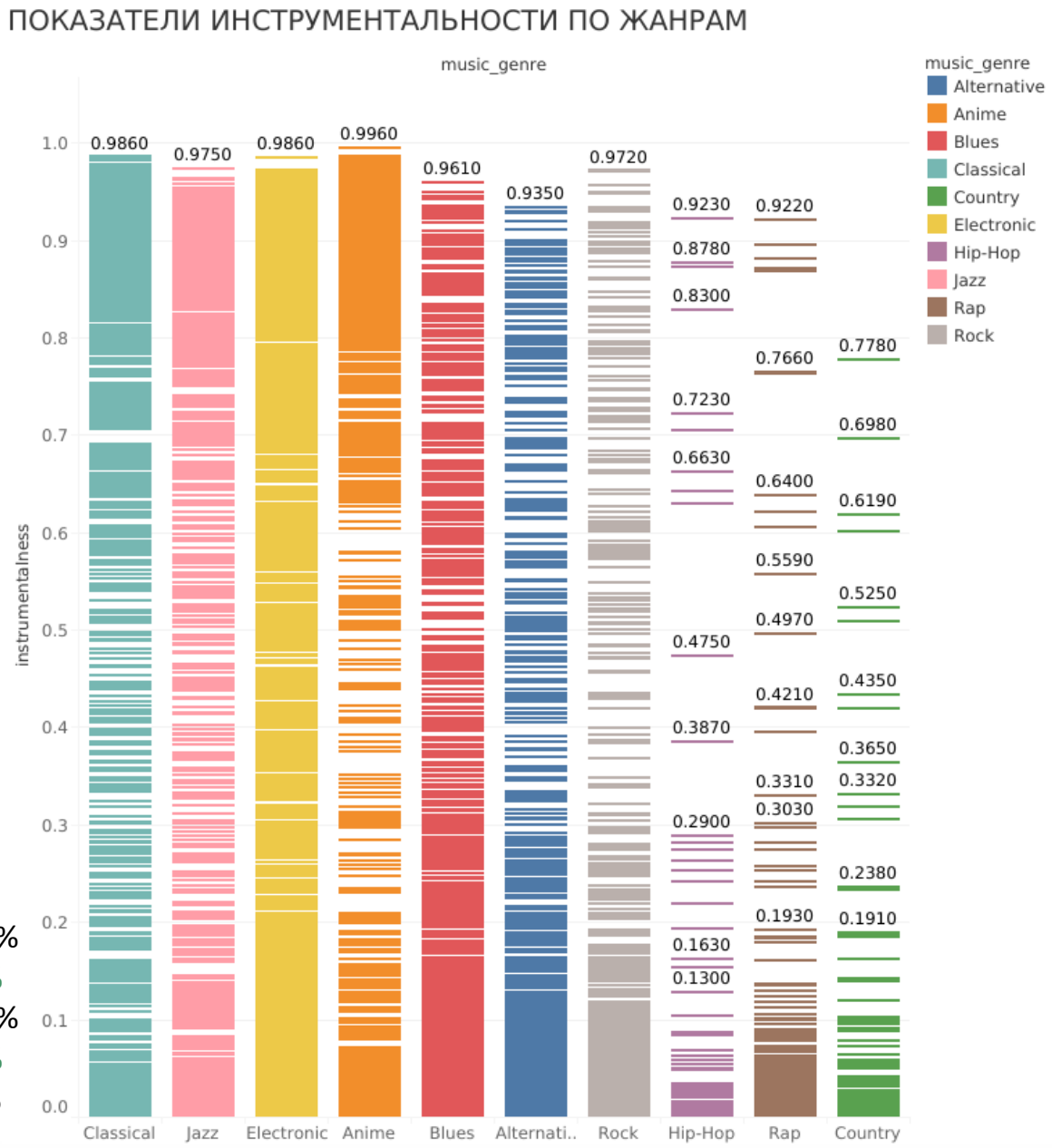
На основе параметра **instrumentalness** попробуем увеличить точность предсказания.

Добавляем новый параметр **high_instr** в котором показатели ниже порогового значения (0.60) будут обнуляться (не учитываться), а выше будут отображаться. Останется меньше треков в жанрах **Country, Rap, Hip-Hop.**

Посмотрим сколько процентов % жанров останется при:

high_instr >= 0.60

- Alternative 4.02%
- Anime 28.71%
- Blues 6.44%
- Classical 63.36%
- **Country 0.19%**
- Electronic 32.67%
- **Hip-Hop 0.86%**
- Jazz 36.67%
- **Rap 0.48%**
- Rock 3.72%



НОВЫЕ ПРИЗНАКИ:

КОРРЕЛЯЦИЯ ПО ИНСТРУМЕНТАЛЬНОСТИ

music_genre	instrumentalness
Alternative	161.9
Anime	512.2
Blues	261.3
Classical	819.3
Country	11.2
Electronic	894.6
Hip-Hop	11.5
Jazz	451.8
Rap	19.8
Rock	119.3

instrumentalness: 11.2 — 894.6

3.3 Не латинские названия

Добавляем новый признак non_latin, который помогает лучше идентифицировать музыкальные треки в жанре Anime и, как ни странно, Classical.

ОБУЧЕНИЕ МОДЕЛИ:

4.1 Результаты

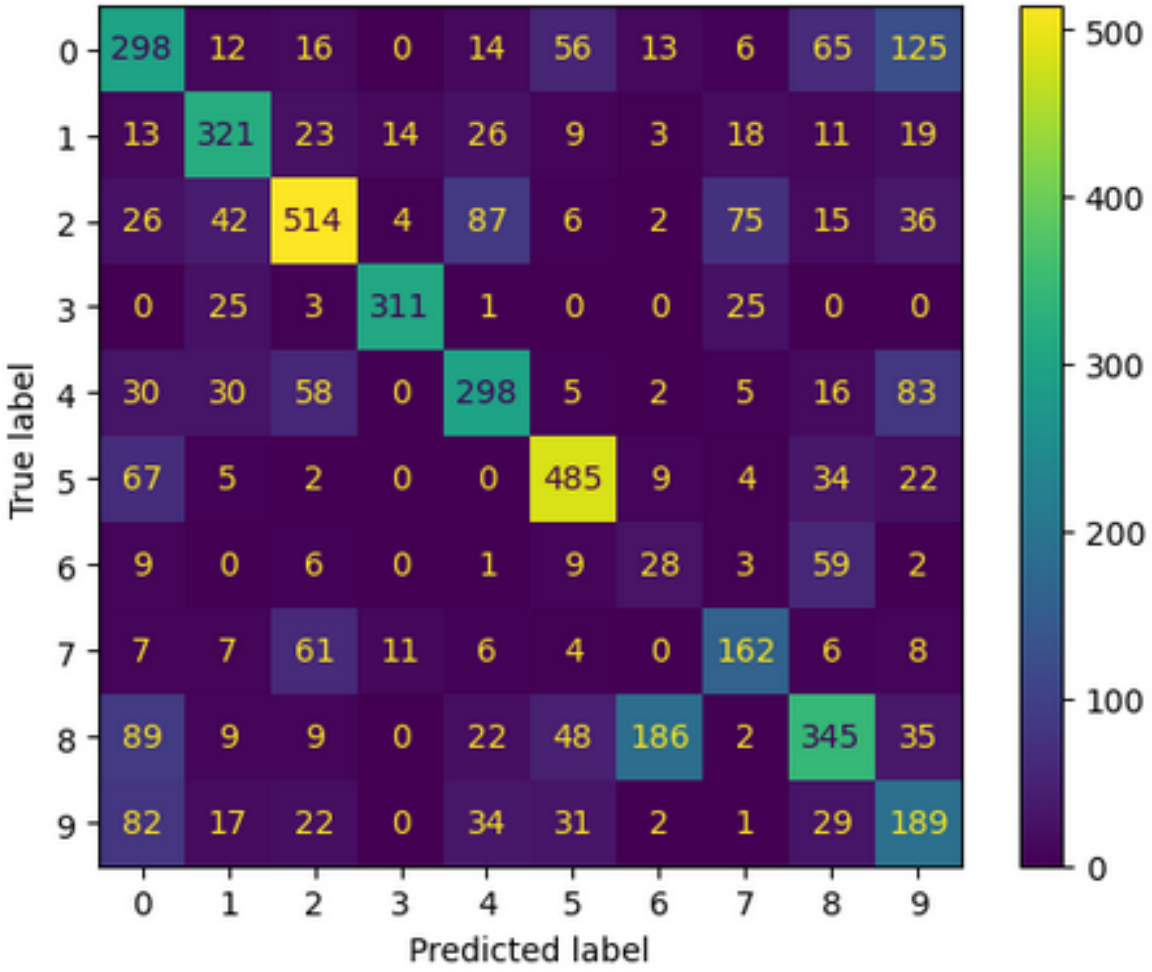
При работе использовались 4 модели классического машинного обучения: CatBoostClassifier, LGBM-Boost*, Random Forest, XGBoost.

Итоговые показатели:

	precision	recall	f1-score	support
Alternative	0.51	0.48	0.50	621
Anime	0.69	0.70	0.70	468
Blues	0.62	0.72	0.67	714
Classical	0.85	0.93	0.89	340
Country	0.58	0.62	0.60	489
Electronic	0.77	0.74	0.76	653
Hip-Hop	**0.28**	**0.12**	**0.17**	**245**
Jazz	0.63	0.54	0.58	301
Rap	0.47	0.62	0.53	580
Rock	0.47	0.35	0.40	519

Accuracy: 0.6046653144016227

F1-Score: 0.5954091521627362



ВЫВОДЫ:

В результате рассмотрения датасета были выявлены следующие зависимости:

- больше всего композиций в жанре **Blues**
- больше всего композиций с высокой акустичностью и громкостью в жанре **Classical**
- не латинское написание названий характерно для 23% треков в жанре **Anime**
- в данных содержатся треки, которые относятся одновременно к нескольким жанрам
- Самые трудноопределяемые композиции в жанре **Hip-Hop**

СПАСИБО!

за внимание ☺