СРАВНЕНИЕ НЕСКОЛЬКИХ МОДЕЛЕЙ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

COMPARISON OF MULTIPLE MODELS OF MACHINE LEARNING METHODS

УДК 004.5

Сенин С.С., студент, seninka11@yandex.ru, Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана, кафедра «Системы обработки информации и управления»
Яковлев Д.С., студент, cinhrofazatron8@yandex.ru, Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана, кафедра «Системы обработки информации и управления»

Senin S.S., student, seninka11@yandex.ru, Russia, 105005, Moscow, MSTU N.E. Bauman, Department of Information Processing and Management Systems'
Yakovlev D.S., student, cinhrofazatron8@yandex.ru, Russia, 105005, Moscow, MSTU N.E. Bauman, Department of Information Processing and Management Systems'

Аннотация

В настоящее время большой популярностью пользуются различные модели машинного обучения. В данной статье будут взяты несколько датасетов, для сравнения их в разных условиях. Будет произведена очистка данных и их визуализация, а также анализ.

Summary

Currently, various machine learning models are very popular. In this article, we will take several datasets to compare them in different conditions. Data will be cleaned up and visualized, as well as analyzed.

Ключевые слова: машинное обучение, датасет, очистка данных, дерево решений.

Key words: machine learning, dataset, data cleaning, decision tree.

План исследования

1. Поиск и выбор набора данных для построения моделей машинного обучения. На основе выбранного набора данных студент должен построить модели машинного обучения для решения или задачи классификации, или задачи регрессии.
2. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.
3. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
4. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения. В зависимости от набора данных, порядок выполнения пунктов 2, 3, 4 может быть изменен.
5. Выбор метрик для последующей оценки качества моделей. Необходимо выбрать не менее двух метрик и обосновать выбор.
6. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии. Необходимо использовать не менее трех моделей, хотя бы одна из которых должна быть ансамблевой.
7. Формирование обучающей и тестовой выборок на основе исходного набора данных.

1. Поиск и выбор набора данных для построения моделей машинного обучения. На основе выбранного набора данных построение модели машинного обучения для решения или задачи регрессии.

В качестве набора данных возьмем набор с данными о песнях и их характеристиках.

Набор содержит такие колонки как:

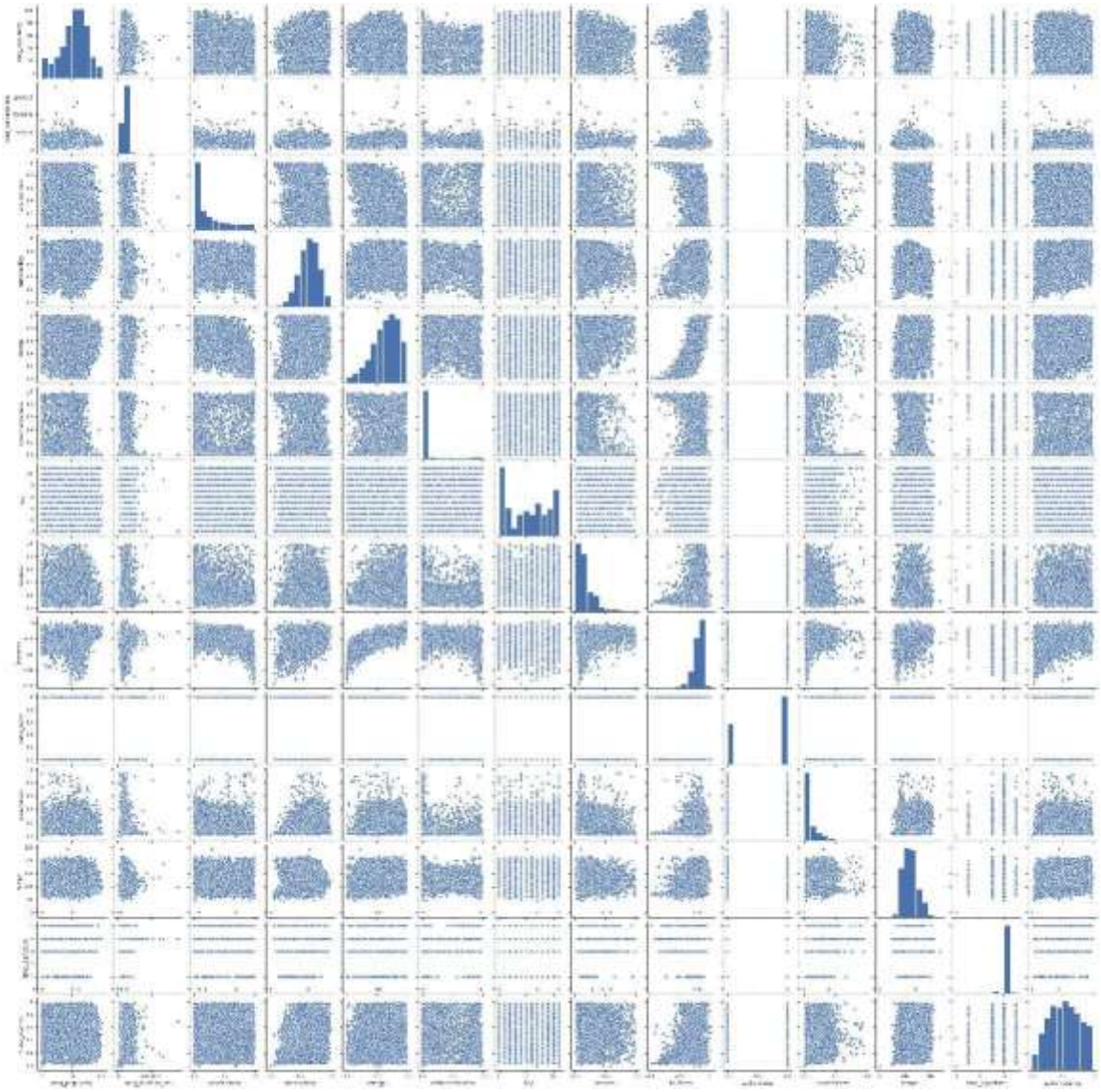
- song_name - название песни
- song_popularity - индекс популярности песни
- song_duration_ms - длительность в мс
- acousticness - индекс акустики
- danceability - индекс танцевальности
- energy - индекс энергичности
- instrumentalness - индекс инструментальности
- key - ключ
- liveness - индекс живости
- loudness - индекс громкости
- audio_mode - режим аудио
- speechiness - индекс разговорности
- tempo - темп
- time_signature - временная метка
- audio_valence

Поставим задачу предсказания популярности песни по данным характеристикам. Построим модель машинного обучения для данного набора и решим задачу регрессии.

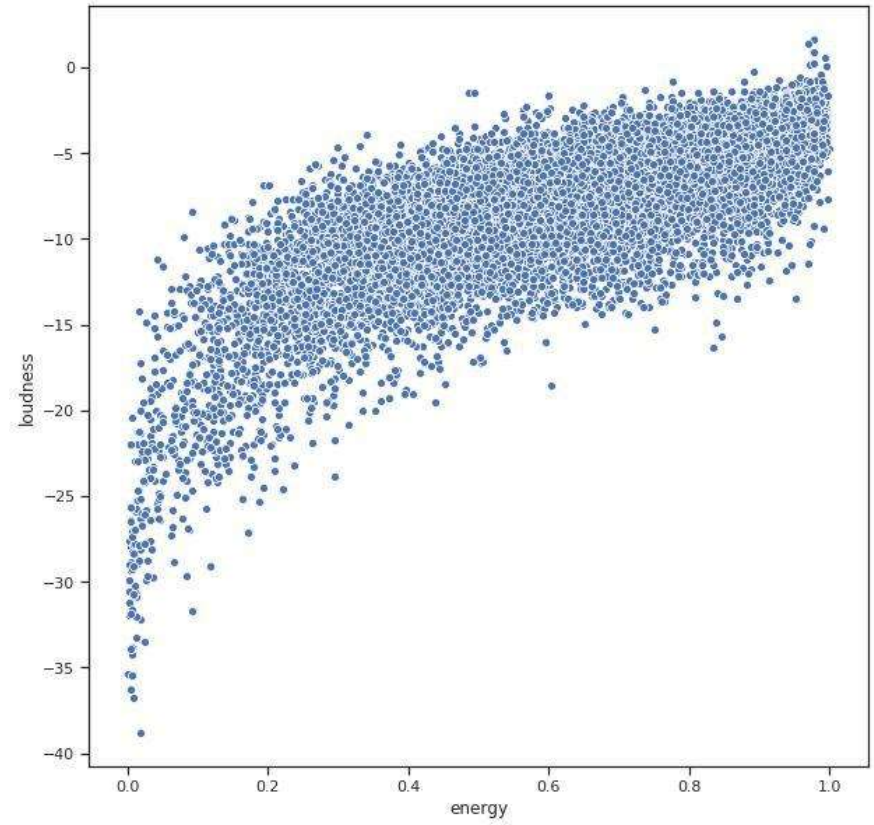
2. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.

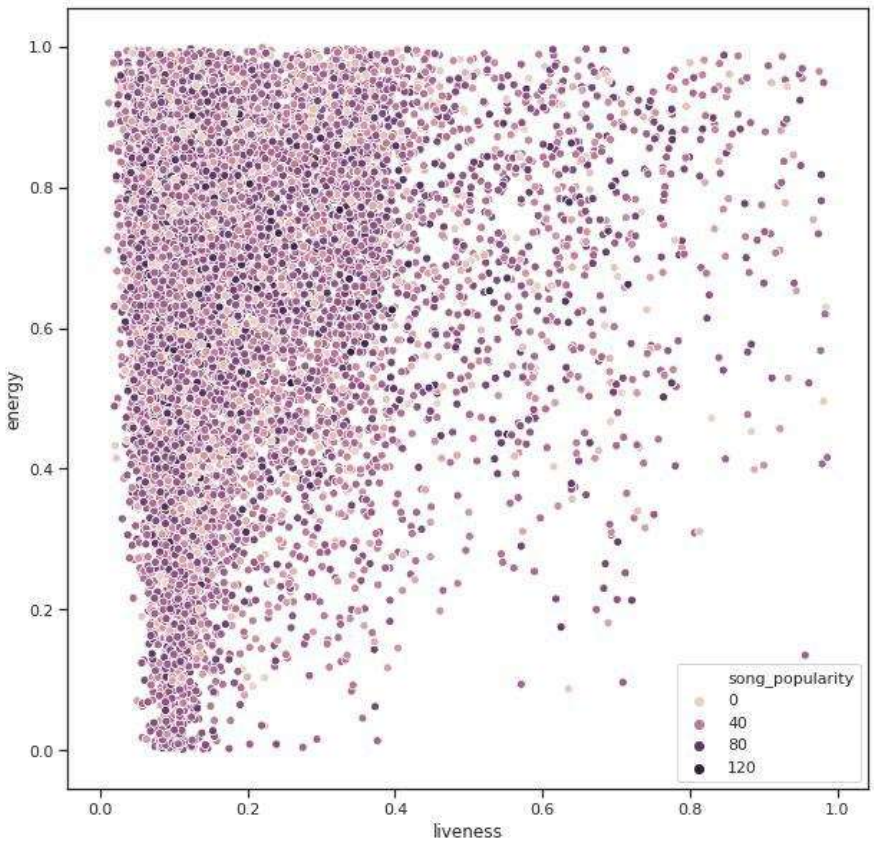
	song_popularity	song_duration_ms	acousticness	danceability	energy
count	18835.000000	1.883500e+04	18835.000000	18835.000000	18835.000000
mean	52.991877	2.182116e+05	0.258539	0.633348	0.644995
std	21.905654	5.988754e+04	0.288719	0.156723	0.214101

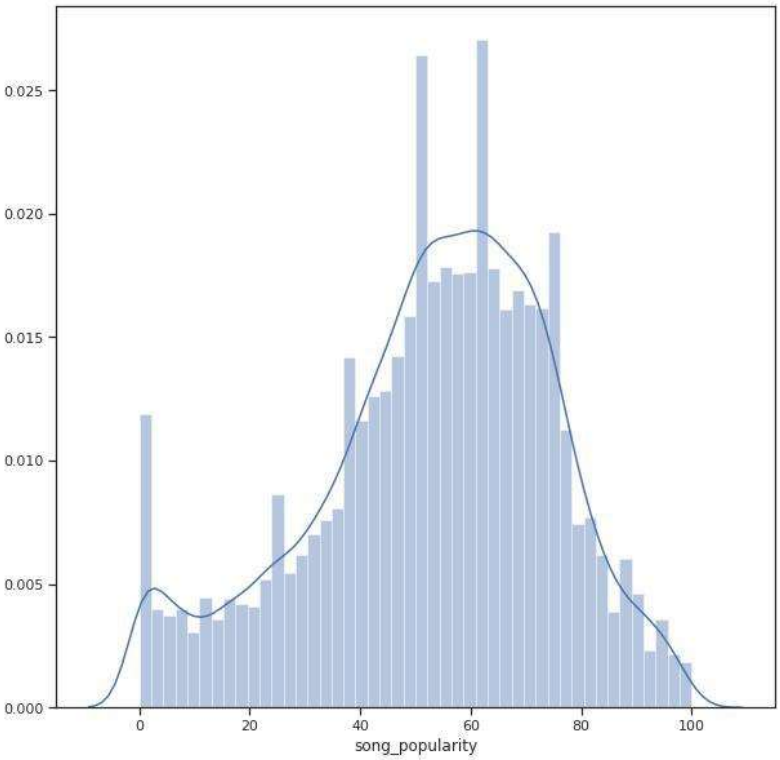
min	0.000000	1.200000e+04	0.000001	0.000000	0.001070
25%	40.000000	1.843395e+05	0.024100	0.533000	0.510000
50%	56.000000	2.113060e+05	0.132000	0.645000	0.674000
75%	69.000000	2.428440e+05	0.424000	0.748000	0.815000
max	100.000000	1.799346e+06	0.996000	0.987000	0.999000

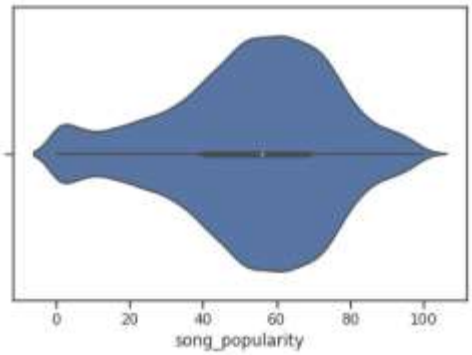


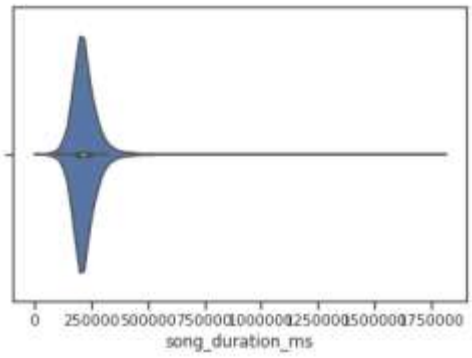
Видим, что наиболее заметна корреляция таких характеристик как громкость и энергичность. В остальных случаях зависимости не такие очевидные.

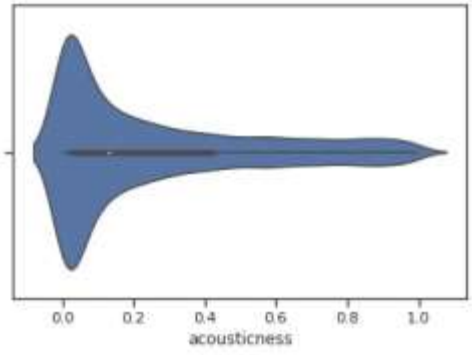

loudness
energy
0
−5
−10
−15
−20
−25
−30
−35
−40
0.0
0.2
0.4
0.6
0.8
1.0

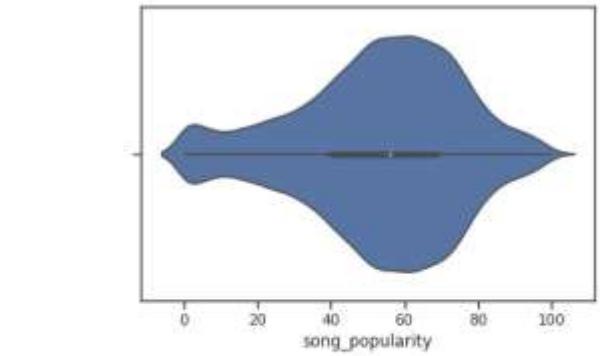

energy
liveness
song_popularity
0
40
80
120
0.0
0.2
0.4
0.6
0.8
1.0

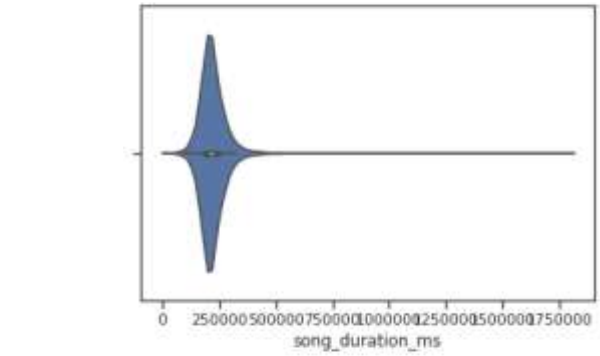

0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
song_popularity

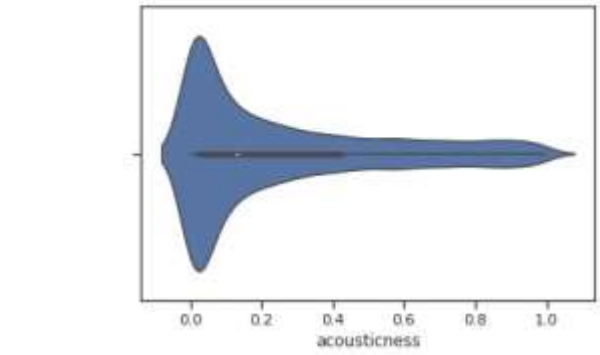

0
20
40
60
80
100
song_popularity

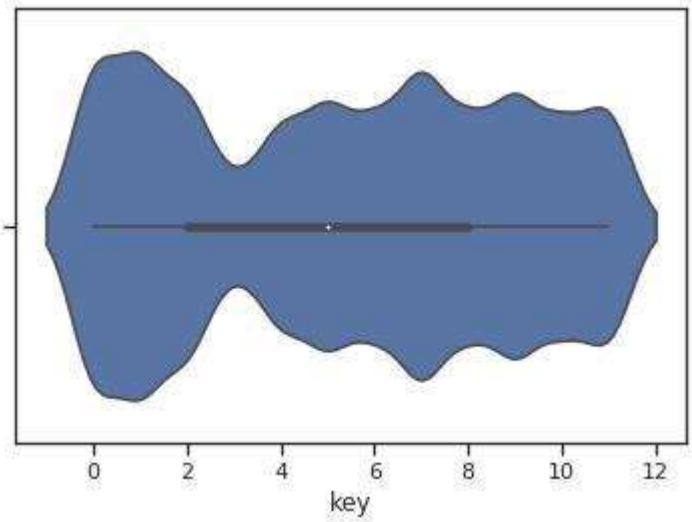

0
250000
500000
750000
1000000
1250000
1500000
1750000
song_duration_ms

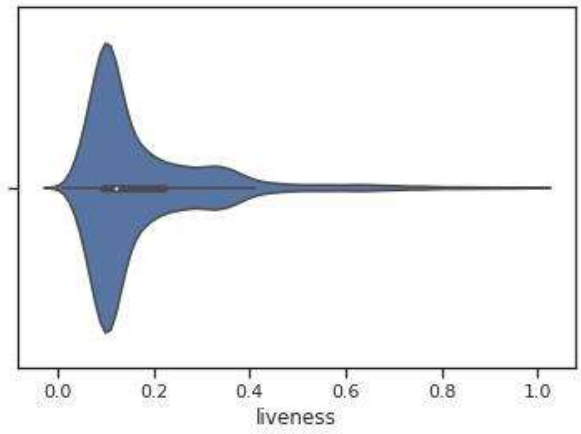

0.0
0.2
0.4
0.6
0.8
1.0
acousticness

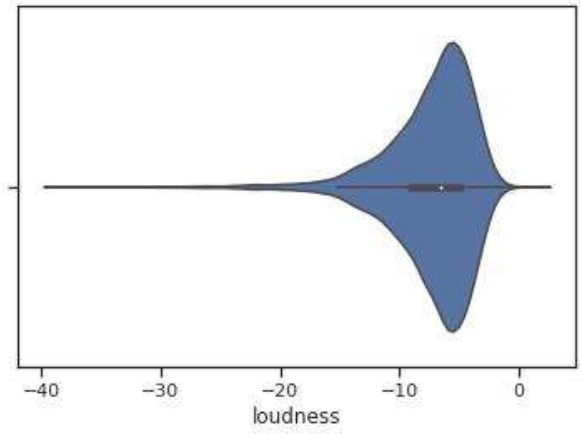

0
20
40
60
80
100
song_popularity

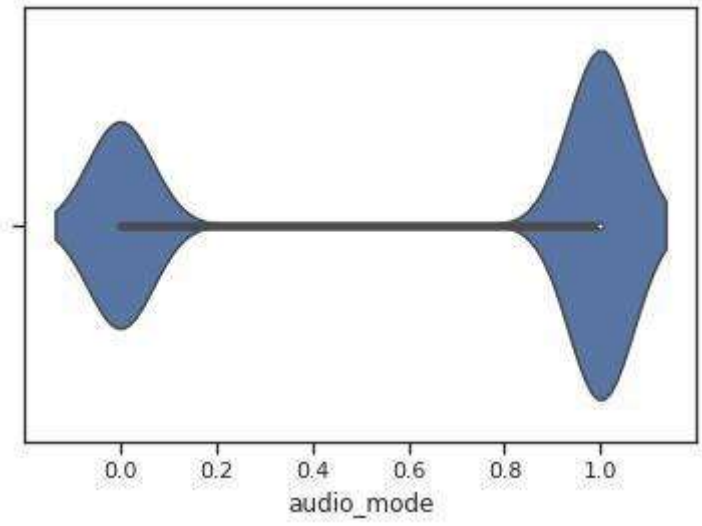

0
song_duration_ms

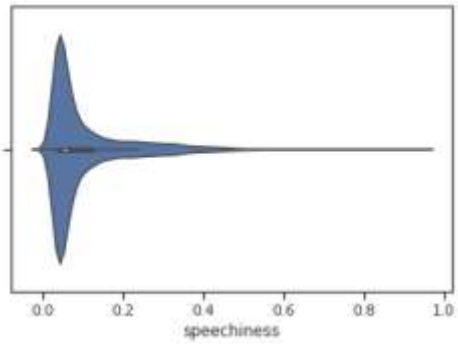

0.0
0.2
0.4
0.6
0.8
1.0
acousticness

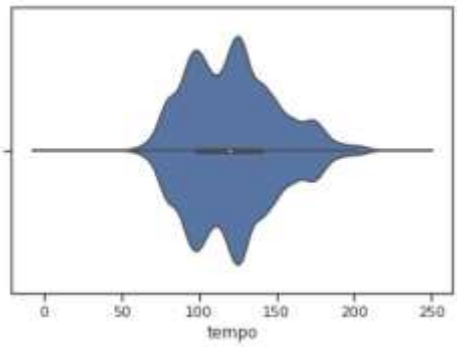

0
2
4
6
8
10
12
key

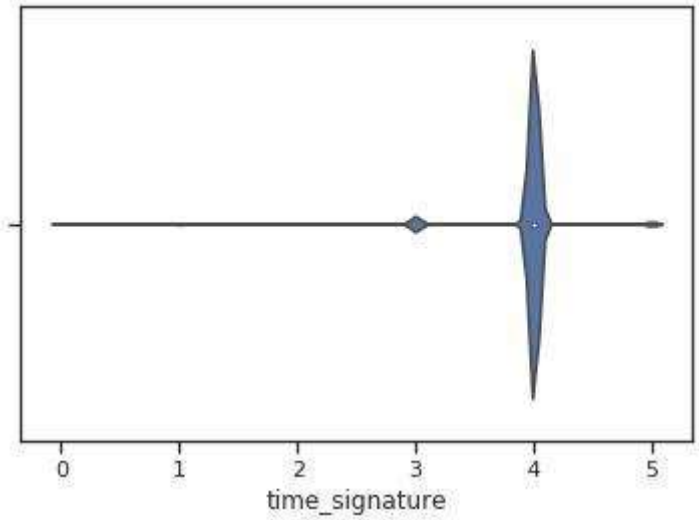

0.0
0.2
0.4
0.6
0.8
1.0
liveness

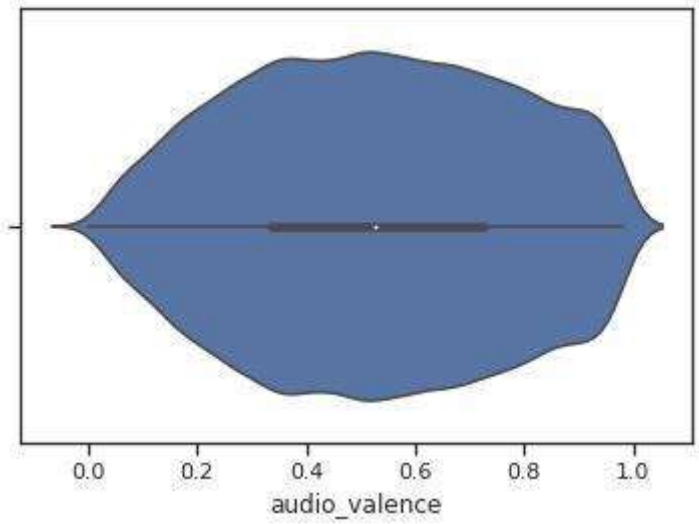

−40
−30
−20
−10
0
loudness


0.0
0.2
0.4
0.6
0.8
1.0
audio_mode


0.0
0.2
0.4
0.6
0.8
1.0
speechiness


0
50
100
150
200
250
tempo

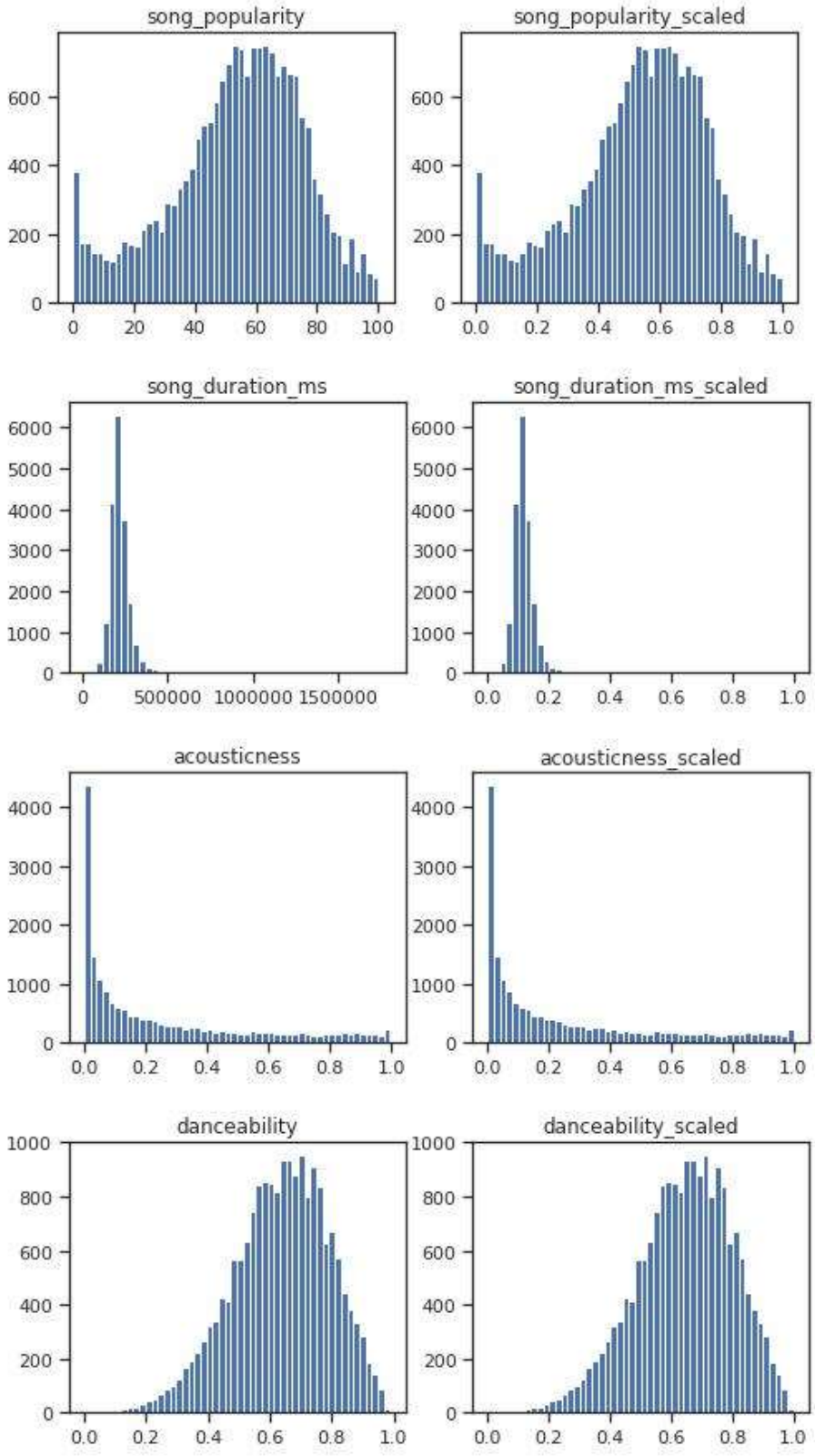

0
1
2
3
4
5
time_signature

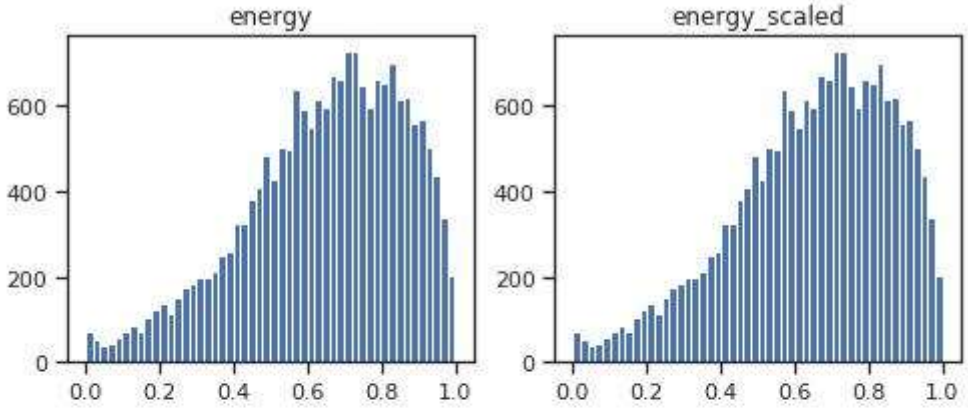

0.0
0.2
0.4
0.6
0.8
1.0
audio_valence

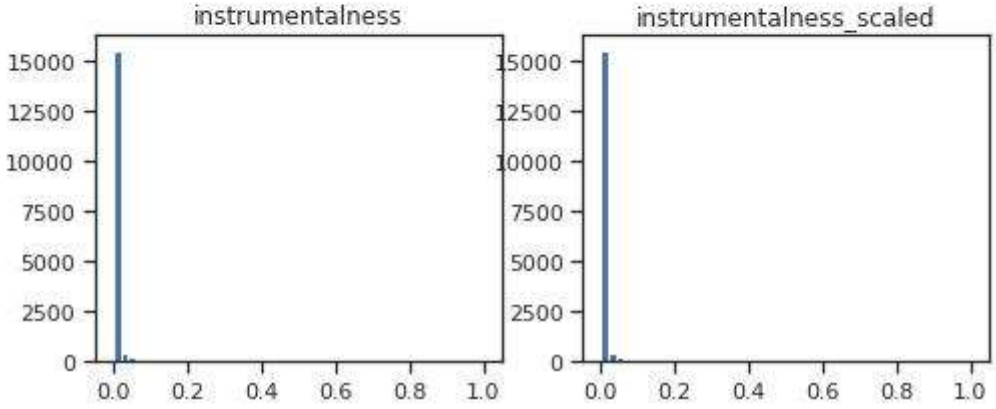
Анализ и заполнение пропусков в данных.

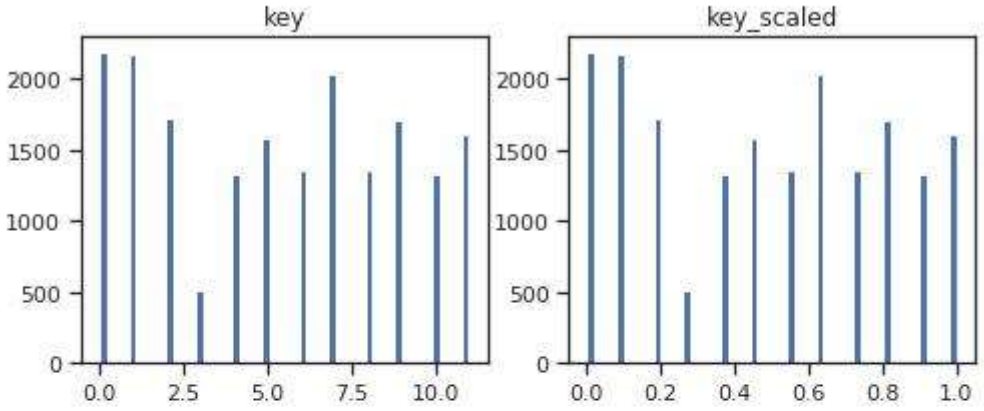
Поскольку в данном наборе пустых значений нет, пропустим данный пункт.

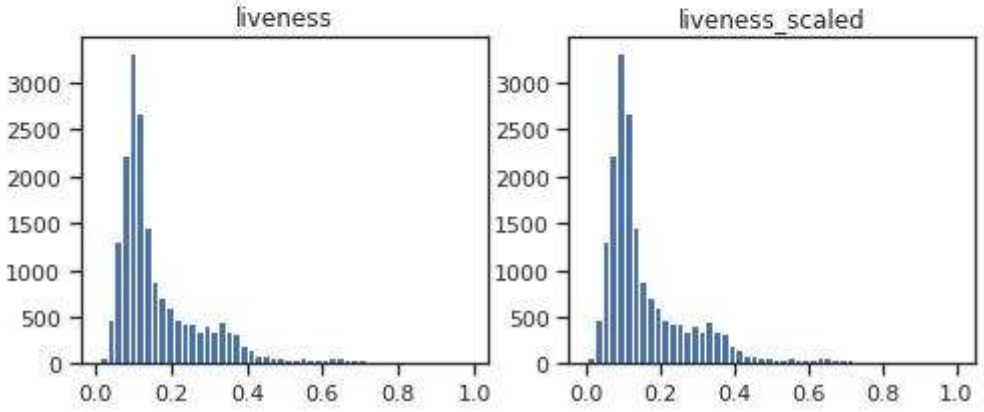
3. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.

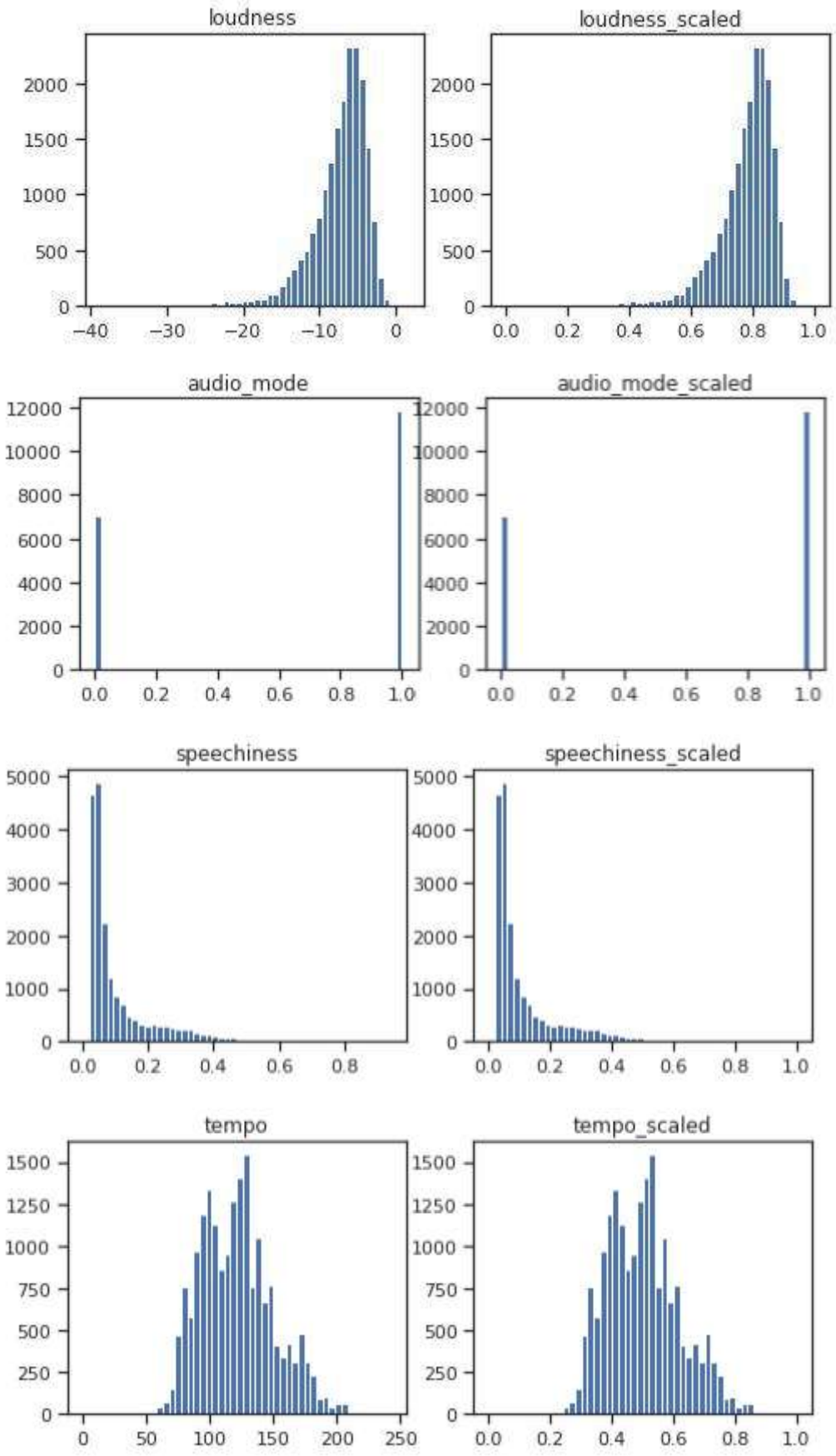


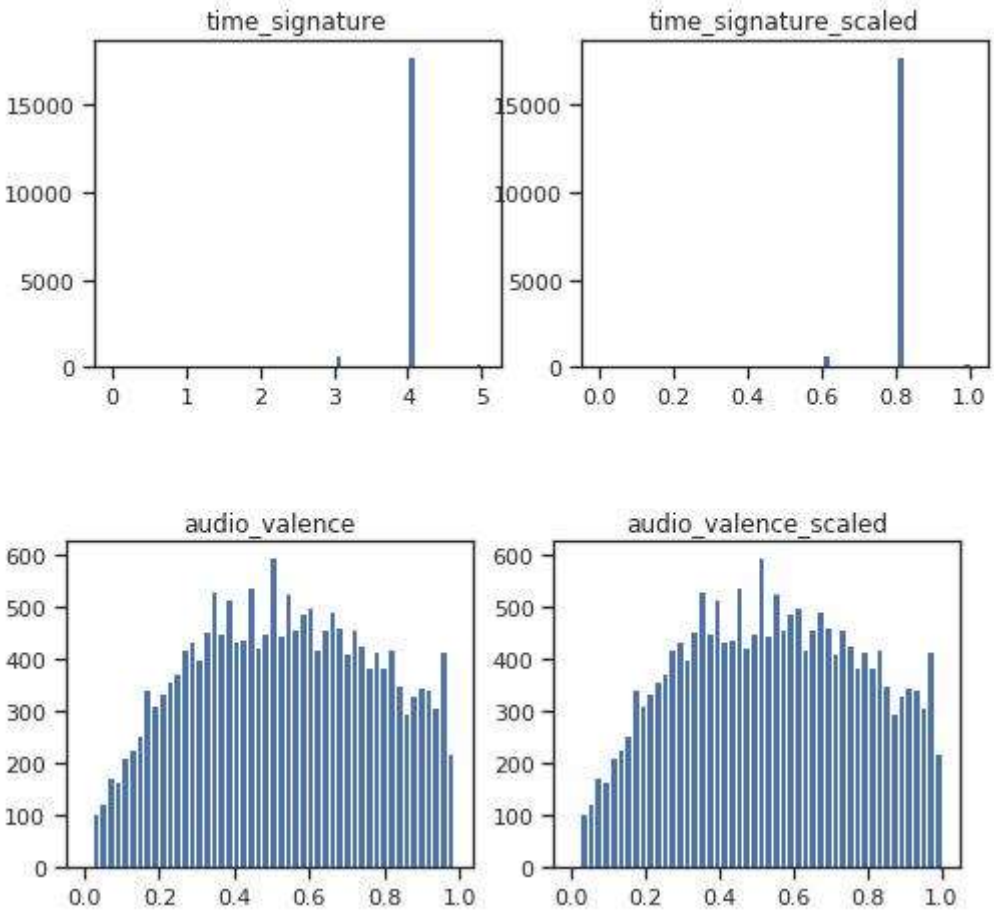

energy
energy_scaled
0
200
400
600
0.0
0.2
0.4
0.6
0.8
1.0


instrumentalness
instrumentalness_scaled
0
2500
5000
7500
10000
12500
15000
0.0
0.2
0.4
0.6
0.8
1.0

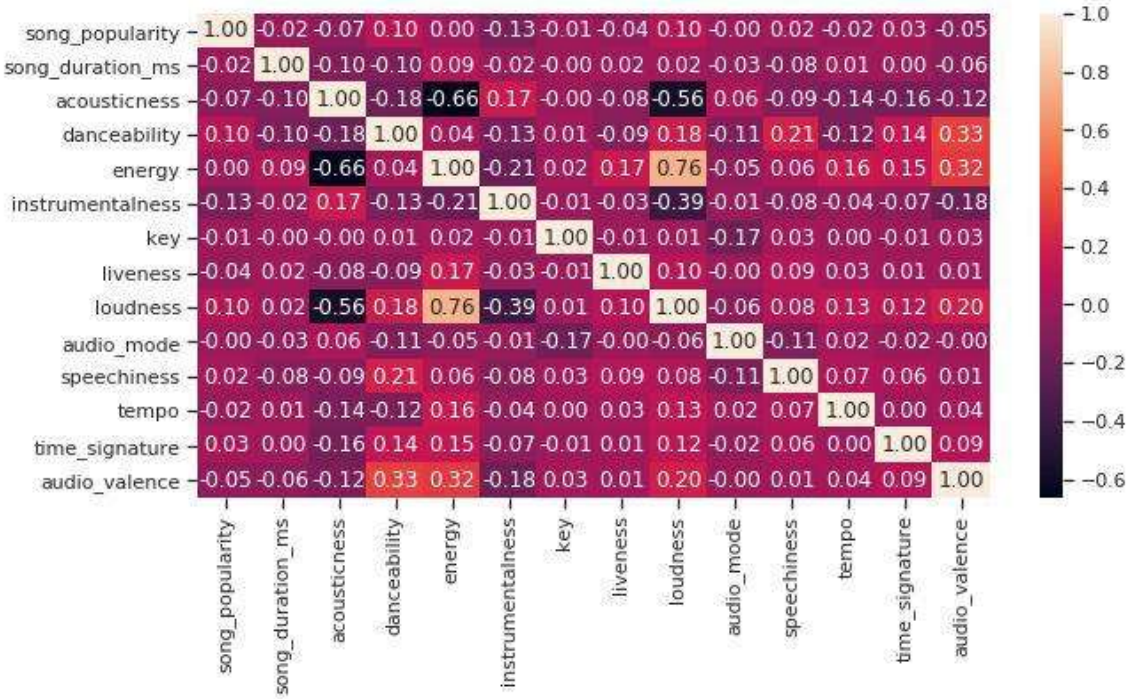

key
key_scaled
0
500
1000
1500
2000
0.0
2.5
5.0
7.5
10.0
0.2
0.4
0.6
0.8
1.0

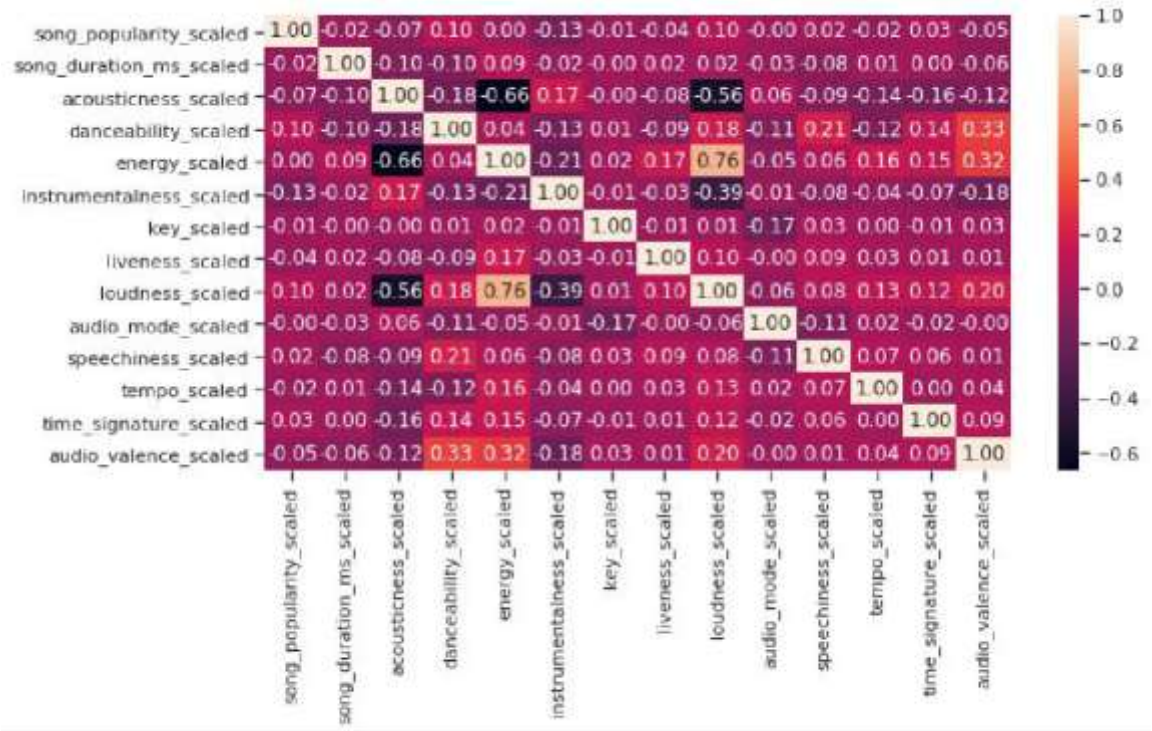

liveness
liveness_scaled
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0


loudness
loudness_scaled
audio_mode
audio_mode_scaled
speechiness
speechiness_scaled
tempo
tempo_scaled



4. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.





Видим, что популярность песни не сильно коррелирует с данными характеристиками. Наибольшее влияние на популярность оказывают такие признаки как танцевальность трека и громкость. Наибольшую корреляцию видим между громкостью и энергичностью трека, как и во 2 пункте.

5. Выбор метрик для последующей оценки качества моделей. Необходимо выбрать не менее двух метрик и обосновать выбор.

Возьмем метрики MAE, Median Absolute Error и R^2.

MAE (Mean Absolute Error) — это среднее абсолютное значение ошибки(среднее модуля ошибки). Данная метрика удобна, так как показывает среднюю ошибку, но при этом не так чувствительна к выбросам, как, например, MSE.

Медиана абсолютного отклонения(Median Absolute Error) - это альтернатива стандартного отклонения, но она менее чувствительна к воздействию промахов, чем среднее отклонение. Коэффициент детерминации, или R^2 покажет насколько модель соответствует или не соответствует данным.

6. Формирование обучающей и тестовой выборок на основе исходного набора данных.

	song_name	song_popularity	song_duration_ms	acousticness	danceability	en
5764	12054	50	204213	0.8570	0.543	0
16725	8262	33	160693	0.0027	0.700	0
5593	4559	100	214289	0.1910	0.687	0
2612	12841	63	246186	0.0133	0.546	0
2528	4667	7	231800	0.0193	0.646	0
...	...	...	...	...	...	
15148	45	54	228240	0.3800	0.745	0
7191	11719	72	233028	0.2000	0.667	0
12766	1204	47	119360	0.0387	0.647	0
3779	4837	66	307200	0.2250	0.932	0
3085	12698	24	161751	0.5160	0.738	0

18835 rows × 29 columns

Вывод:

Лучше всего показала себя модель случайный лес, на втором месте - ансамблевая модель, на третьем - дерево решений. Однако в другой задаче лучше показала себя ансамблевая модель, так что в дальнейшем можно использовать обе эти модели и проверять, какая будет работать лучше для конкретной задачи.

Список использованных источников

1. Надежность технических систем/ Под ред. И.А.Ушакова.–М.: Радио и связь, 1985.
2. Надежность автоматизированных производственных систем. под ред. Я.А.
3. Кузовлев В. И. «ИССЛЕДОВАНИЕ МЕТОДОВ РЕЗЕРВИРОВАНИЯ СИСТЕМ» по разделу «Модели и методы оценки надежности автоматизированных систем» курса «Модели оценки качества АСОиУ»
4. Глазунов Л. П. Основы теории надежности автоматизированных систем управления. 1984г.
5. Панфилов И. В., Половко А. М. Вычислительные системы. 1980г.

List of sources used

1. Reliability of technical systems / Ed. I.A.Ushakova. –M .: Radio and communication, 1985.
2. Reliability of automated production systems. ed. Ya.A.
3. Kuzovlev V. I. "RESEARCH METHODS OF BACKUP SYSTEMS" in the section "Models and methods for assessing the reliability of automated systems" of the course "Models for assessing the quality of automated control systems"
4. Glazunov LP Fundamentals of the theory of reliability of automated control systems. 1984
5. Panfilov IV, Polovko AM Computing systems. 1980