**Анализ данных:**

**топ самых прослушиваемых песен**

**за январь-август 2023 года**

Морозова Евгения

*Технический университет Молдовы*

*Факультет Вычислительной Техники,*

*Информатики и Микроэлектроники*

e-mail: evghenia.morozova@iis.utm.md

**Резюме**

Был проведён исследовательский анализ датасета «Топ самых прослушиваемых песен за январь-август 2023 года». Информация представлена в различных графиках, таких как боксплот, барплот, круговая диаграмма, график рассеяния и проч. Далее следует описание трёх моделей: линейной регрессии, древа решений и случайного леса. Эти модели созданы для предсказания количества прослушиваний каждой песни на Spotify по коррелирующим колонкам «в плейлистах Spotify», «в плейлистах Apple Music», «в плейлистах Deezer».

**Введение**

Для проведения анализа был выбран датасет с Kaggle «Most Streamed Spotify Songs 2023» (версия за август).

Несмотря на название, датасет представляет собой список самых прослушиваемых песен не только в Spotify, но и в Deezer, Shazam и Apple Music с января по август 2023 года. Присутствуют следующие колонки:

* + Название песни;
  + Исполнитель;
  + Количество вовлечённых исполнителей;
  + Дата выпуска;
  + Количество прослушиваний песни на Spotify;
  + Количество плейлистов Spotify, Apple Music и Deezer, в которые была включена песня;
  + Наличие и позиция песни в чартах Spotify, Apple Music, Deezer, Shazam;
  + Лад: мажор или минор;
  + Различные аудио-характеристики: процент речи и инструментальной музыки, позитивность песни, её энергичность и т. д. Указано в процентах.

На момент обновления в датасете присутствовало около 950 строк, однако вследствие чистки N/A значений использовались лишь 902 из них.

**Методы**

Для предобработки данных, их фильтрации, создания всех представленных ниже графиков и моделей использовалась среда разработки RStudio, предназначенная для языка программирования R. Приложенный к статье исходный код также написан на R.

**Обсуждение**

Анализ датасета подразумевает работу с содержащейся в нём информацией и представление этой информации в удобном виде. Создание графиков позволяет выделить наиболее значимые колонки и, что важно, найти их корреляцию.

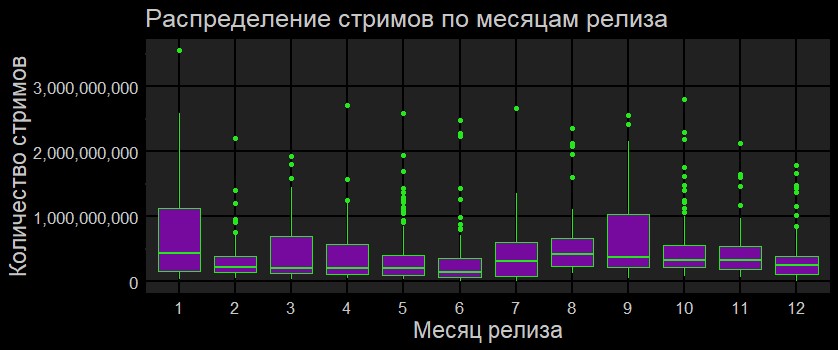
Для начала были построены графики, содержащие общие сведения о песнях:

* Год релиза (рис.1);
* Прослушивания по месяцам релиза (рис. 2);
* Распределение ладов (рис. 3);
* Диаграмма рассеяния (рис. 4).



(Рис. 1) Распределение песен по годам релиза

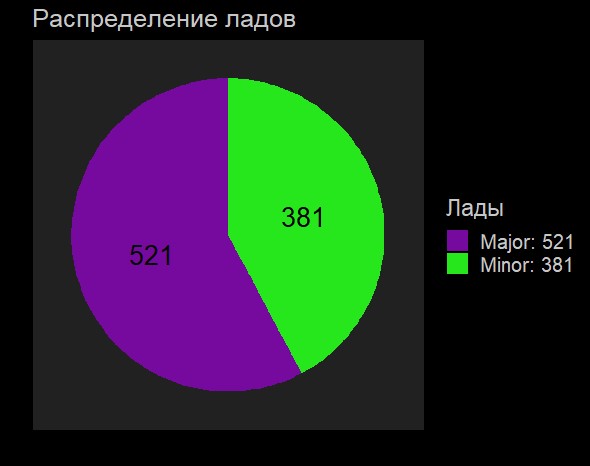
График распределения по годам релиза показывает, что в январе-августе 2023 года наибольшее количество популярных песен было выпущено в 2022 году. Далее следуют песни самого 2023 года, затем 2021. Песни, выпущенные после 2010 года, в 2023 популярнее, чем те, что вышли до него. Также есть несущественная, но вызывающая любопытство доля песен прошлого века: удивительно, как в наше время старые песни могут переживать «второе рождение» и вновь «врываться» в музыкальные чарты.



(Рис. 2) Распределение стримов по месяцам

Боксплот распределения прослушиваний по месяцам релиза показывает огромное количество выбросов. Каждый из этих выбросов – хит, «взорвавший» чарты Spotify. Можно заметить, как одна из песен, выпущенная в январе, была проиграна пользователями Spotify более 3 млрд. раз: вот она, единственная точка сверху в подгруппе «1».

Стоит отметить, что этот график большой смысловой нагрузки не несет. Месяц выпуска песни не должен влиять на её популярность; на популярность скорее влияют тренды, а также события, к которым эти песни приурочены. Именно поэтому в этом графике столько выбросов. Будь датасет больше, скажем, на 10 000 или даже 100 000 строк, разница между подгруппами стала бы незначительной.

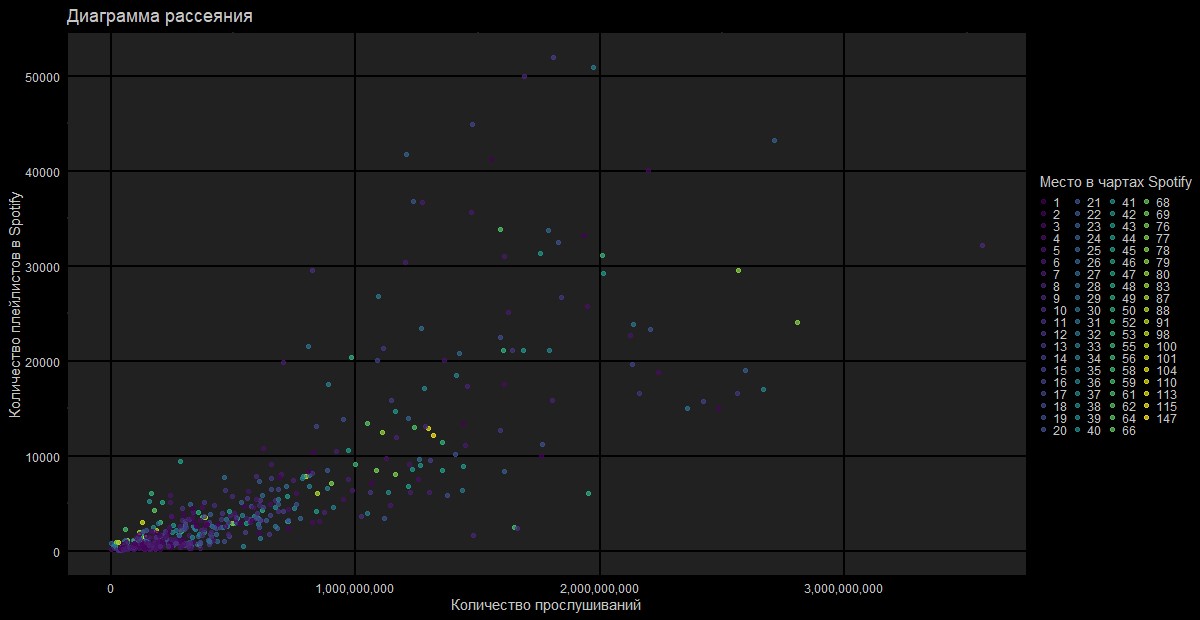


(Рис. 3) Распределение ладов

Другое дело – график распределения ладов. Эта небольшая, но ёмкая круговая диаграмма показывает разделение всех песен на две категории: мажор и минор.

Мажорный лад присущ весёлым, энергичным песням. Считается, что мажорный лад задаёт радостное настроение. Минорный же навевает слушателю грусть, он больше связан с меланхолией, спокойствием.

График показывает, что в промежутке с января по август 2023 года слушатели предпочитали мажорный лад.



(Рис. 4) Диаграмма рассеяния

Диаграмма рассеяния учитывает три колонки сразу. По вертикали – количество плейлистов в Spotify, в которые была включена песня. По горизонтали – количество прослушиваний этой песни в том же Spotify. На легенде разными цветами, в спектре от тёмно-фиолетового до жёлтого, обозначена наивысшая позиция данной песни в чартах, опять-таки, Spotify. Всего в легенде 79 элементов, наивысшая позиция – 1, низшая – 147.

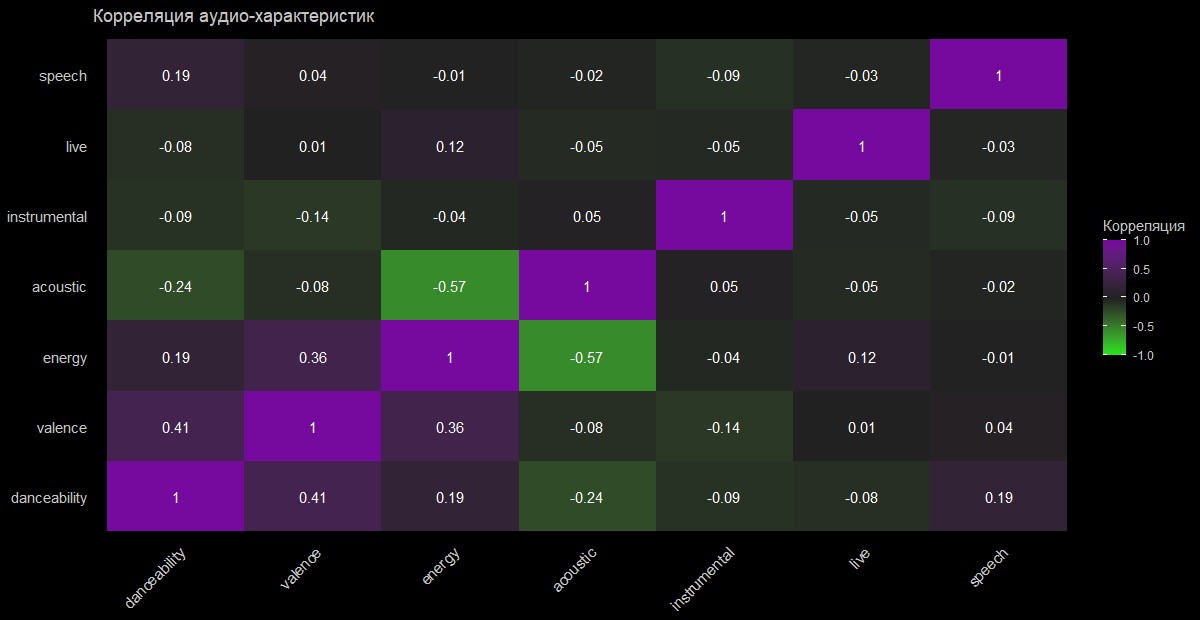
При составлении этой диаграммы были исключены все песни, где место в чарте было обозначено как «0», то есть песня никогда в чарте не присутствовала. Эти песни составляли значительное число от общей массы, однако рассматривались, как если бы они имени N/A в своём поле.

Какие выводы можно сделать по этому графику? Итак, большая часть популярнейших песен (1-20 места в чарте) оказалась только-только вышедшими хитами, на момент последнего обновления датасета набравшими не более миллиарда прослушиваний и включенные в менее чем 10 000 плейлистов. Примерно такая же картина и с песнями, занимающими в чарте позиции пониже (20-40). Существуют, однако, хиты, которые, видимо, оставались популярными дольше – та самая песня с более чем 3 миллиардами прослушиваний. Есть также немногочисленные экземпляры, занявшие высокую позицию и включенные в более чем 50 000 плейлистов.

И всё же большинство песен лишь набирают популярность. Со временем количество прослушиваний и список плейлистов возрастёт, но нет никаких гарантий, что это поможет песне удержаться в чарте. Всё-таки в чарты попадают лишь наиболее «модные», «трендовые» представители музыкальной индустрии, но держатся они там, особенно во главе рейтинга, зачастую недолго.

Далее представлены два хитмэпа корреляций:

* корреляция аудио-характеристик (рис. 5)
* и корреляция прослушиваний на Spotify, присутствия в чартах Apple Music, Deezer, Spotify, Shazam и в плейлистах Apple Music, Deezer, Spotify (рис. 6).



(Рис. 5) Корреляция аудио-характеристик

Хитмэп корреляции аудио-характеристик был создан с целью выявить некоторые закономерности в том, как расставляются аудио-характеристики песни. Но, к удивлению, выяснилось, что никаких закономерностей выявлено не было.

На графике видно, что существует связь между полем acoustic и полем energy. Акустичность музыки – это присутствие в ней классических, не электронных и не электрических инструментов. Пример классического инструмента – флейта. Высокая акустика характерна для песен в жанре фолк.

Утверждать, будто песня тем энергичней, чем меньше в ней чистого звучания классических инструментов, неправильно. Вероятно, такой график получился исключительно из-за малого размера выборки. Отдельный вопрос стоит задать, собственно, составителю датасета – как определялись такие колонки, как energy, valence, dance ability? По своей природе они сугубо субъективны. К слову, разница между energy и dance ability тоже не совсем понятна.



(Рис. 6) Корреляция прослушиваний, чартов и плейлистов

Зато следующий хитмэп позволяет сделать нам несколько вполне логичных выводов. Первое: чем выше позиция песни в одном из чартов (здесь рассматриваются чарты Spotify, Apple Music, Shazam и Deezer), тем выше она и в других чартах. Второе: чем больше количество плейлистов (Spotify, Apple Music и Deezer), в которые включена песня, на одной платформе, тем больше их количество и на другой. Третье: количество прослушиваний песни на Spotify коррелирует как раз с плейлистами, но не с чартами. Это соответствует утверждению, что присутствие в чартах краткосрочно. А вот плейлисты остаются в ходу ещё много лет, поскольку они собраны «под настроение» и популярность песни для них не столь важна.

Примечание. Можно заметить, что платформа Deezer коррелирует с другими платформами меньше, чем они между собой. Дело в том, что Apple, Shazam и Spotify – международны, а Deezer – исключительно французская платформа, поверженная влиянию местных предпочтений.

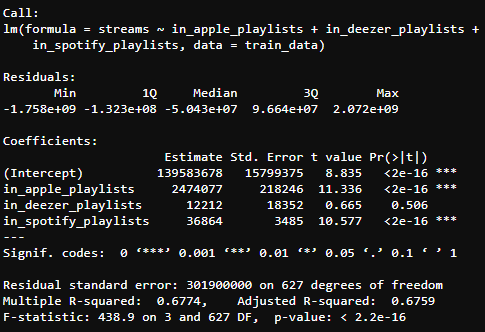
**Предсказания. Результаты**

Для создания моделей предсказания автор руководствовался последним графиком корреляции. Было решено предсказывать колонку streams (количество прослушиваний каждой песни на Spotify) по коррелирующим колонкам in\_spotify\_playlists, in\_apple\_playlists, in\_deezer\_playlists («в плейлистах Spotify», «в плейлистах Apple Music», «в плейлистах Deezer» соответственно). Ссылка на код находится в следующем разделе статьи.

1) Линейная регрессия. Это статистическая модель для анализа отношений между некой зависимой переменной (Y) и одной или несколькими независимыми переменными (X1, X2) путем приближения их линейной функцией.

В данном случае зависимая переменная – значение из streams, независимые – значения из in\_spotify\_playlists, in\_apple\_playlists, in\_deezer\_playlists.

Для обучения модели было взято 70% выборки, остальные 30% использовались для теста.

****

(Рис. 7) Вывод в RStudio для линейной регрессии

При создании модели линейной регрессии RStudio позволяет проанализировать некоторые интересные аспекты.

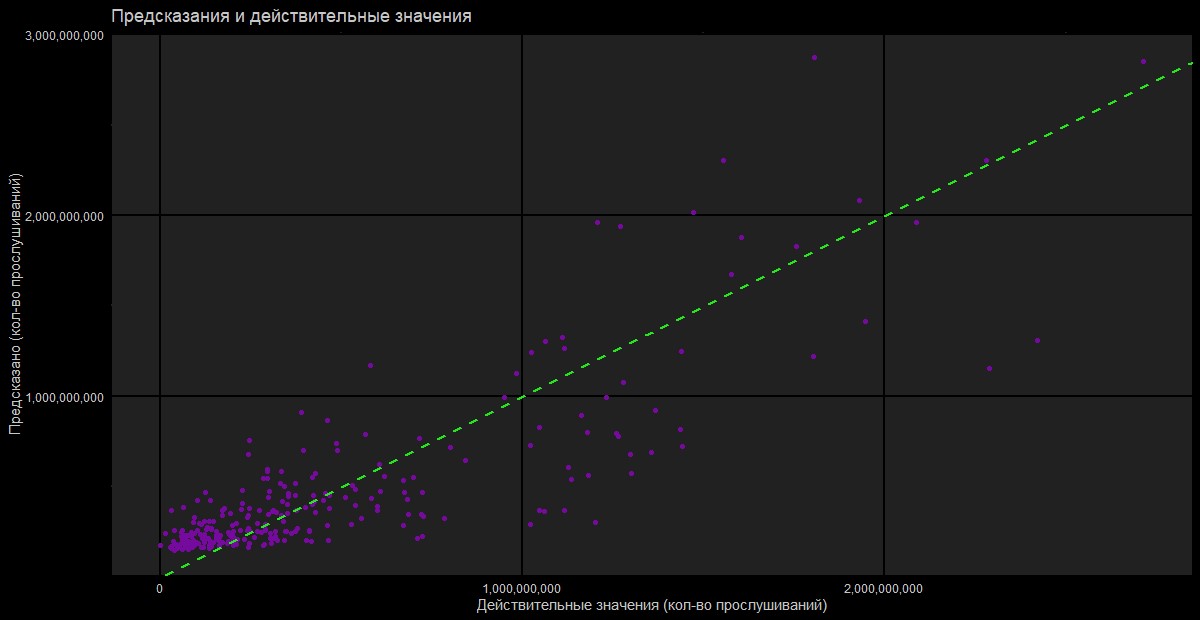
В матрице коэффициентов показано, как каждая из независимых переменных влияет на предсказание. По количеству звёздочек можно узнать, какие из иксов действительно повлияли на конечный результат. Чем больше звёздочек (от нуля до трёх), тем важнее переменная. Здесь видно, что переменная in\_deezer\_playlists оказалась бесполезна: её ошибка больше, чем estimate-значение. Другие же две переменные крайне полезны: для estimate-значений в 2,4 миллиона и 36 тысяч погрешности в 218 тысяч и 3,4 тысячи соответственно не так уж и заметны (менее 10%).

В итоге доля дисперсии streams, объяснённая моделью, составляет 67,74% - довольно неплохой результат.

MSE (среднеквадратическая ошибка) для линейной регрессии оказался равен 7.442487e+16; RMSE (корень из MSE) равен 272.809.213. Чем ниже значения MSE и RMSE, тем лучше.

Для модели линейной регрессии можно построить два графика:

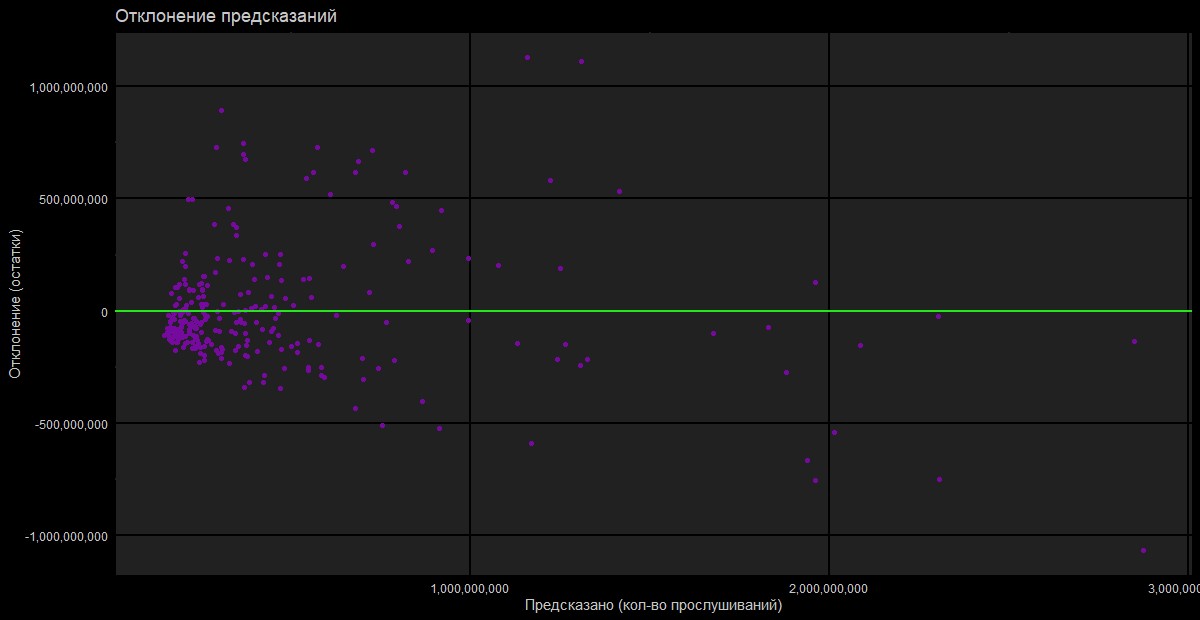
* график предсказаний (рис. 8)
* и график остатков (рис. 9)



(Рис. 8) Предсказания линейной регрессии

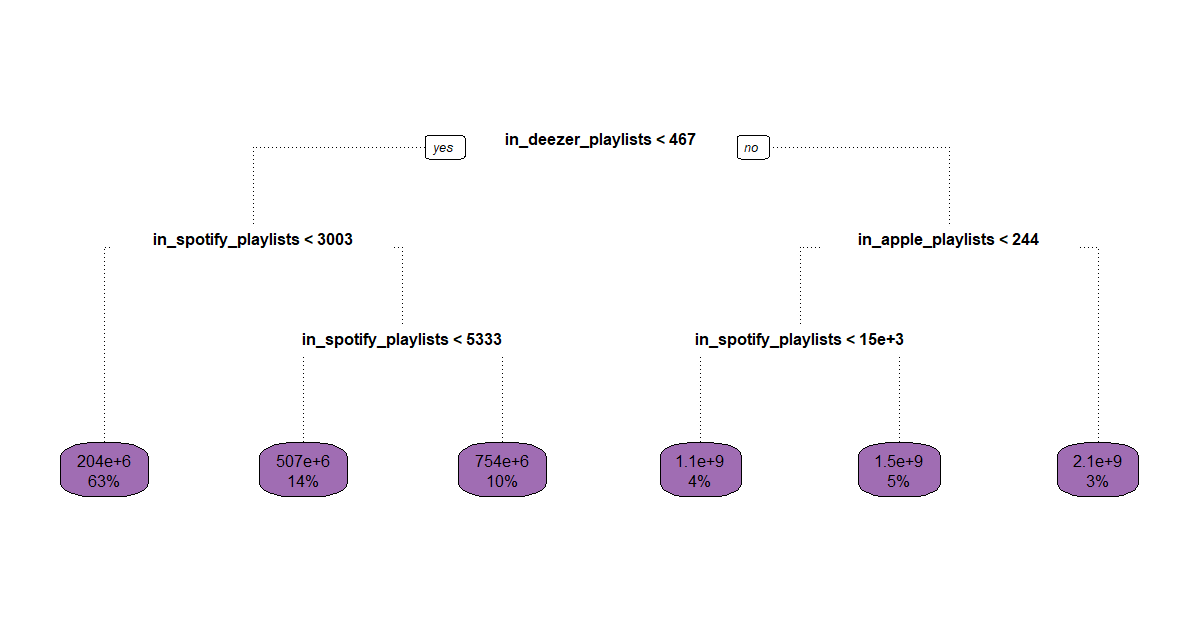
График предсказаний представляет собой точечный график с действительными значениями на одной оси и предсказанными значениями на другой, линия регрессии показывает приближение зависимости. Позволяет сравнить, как хорошо модель улавливает закономерность.

График остатков похож на график предсказаний, но сосредоточен на отклонениях вокруг нуля. Так как отклонения носят не систематический характер, то есть никаких паттернов не наблюдается и все остатки распределены случайно, можно сказать, что модель работает хорошо.



(Рис. 9) График остатков для линейной регрессии

2) Древо решений. Это модель, где узлы – решения относительно некоторого признака данных, каждое ребро представляет собой результат теста на этом признаке, а листья древа отражают прогнозы.



(Рис. 10) Древо решений

В RStudio древо решений строится автоматически, изменить значения узлов вручную нельзя. MSE для древа решений равно 7.265298e+16; RMSE составляет 269.542.161.

3) Случайный лес. Состоит из нескольких деревьев, объединённых для улучшения обобщающей способности модели. Это мощный инструмент, популярный в решении задач машинного обучения.

RStudio не предусматривает никаких графиков или подробностей для случайного леса, поэтому перейдём сразу к результатам. MSE здесь равняется 5.201319e+16; RMSE – 228.063.998. Это наилучшее значение среди всех моделей.

**Код и данные**

Датасет: <https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023>

Исходный код, графики и проч. можно посмотреть здесь: <https://github.com/yuiterute/DataScienceProject>