ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ

“ЦЕНТР ОНЛАЙН-ОБУЧЕНИЯ НЕТОЛОГИЯ-ГРУПП”

(ООО “ЦОО НЕТОЛОГИЯ-ГРУПП”)

ДИПЛОМНЫЙ ПРОЕКТ

по профессии “Аналитик данных”

На тему: “Анализ данных Spotify по музыкальным композициям (исследование данных, выявление закономерностей, построение модели прогноза популярности трека)”

Выполнил: Жданов А.А.

Группа: DA-13

Москва, 2021

ОГЛАВЛЕНИЕ

1. ПОСТАНОВКА ЗАДАЧИ

В мире все чаще используются в качестве простого и удобного способа прослушивания музыки стриминговые сервисы. Из-за растущего доступа к сети Интернет и модой на более удобные сервисы популярность их неуклонно растет.
В данной работе будет производиться анализ данных такого музыкального стримингового сервиса как Spotify. Информация взята из открытого источника с сайта kaggle.com (ссылка https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks), автором является Yamac Eren Ay. Данные, которые будут использоваться при анализе (свыше 170 тысяч композиций) были собраны посредством Spotify Web API, а также Spotipy (модуль Python для Spotify web servers (API)). Треки были выборочно выбраны из всего представленного многообразия, полный список собрать не представляется возможным из-за ограничения в Spotify Web API.

Предполагаемым возможным заказчиком/стейкхолдером может выступать звукозаписывающая компания. Ее же можно отнести к стейкхолдерам группы артистов, продюсеров музыкальных групп, малоизвестных сольных артистов, диджеев и т.п. Также данное исследование будет интересно всем лицам, получающим доход от продажи или выпуска музыкальных композиций.
Основной бизнес-задачей будет являться построение модели прогноза популярности трека, основываясь на исторических фактах популярности музыкальных композиций с 1920-2021 гг. Результат можно будет использовать для того, чтобы предварительно узнать возможный денежный потенциал выпущенной композиции, а также для информирования стейкхолдеров, какие направления и качества в музыкальной композиции потенциально более выгодны.
Будем считать, что бизнес-заказчику приоритетнее получать более высокий доход, что это является бизнес-требованием по задаче, а для этого необходимо иметь максимально возможное количество прослушиваний трека, иметь максимально возможную популярность композиции. Также необходимо будет провести анализ по наиболее прослушиваемому жанру, чтобы узнать, какие группы по жанру исполнения заведомо более популярны.
Основными гипотезами для проверки будут являться:

- “Наиболее популярными треками являются те, которые имеют темп в 120 ударов в минуту (далее - BPM)”;
- “Продолжительность наиболее популярных треков - 3-3,5 минуты”;
- “Композиции в мажорном исполнении более популярны”.

Также будет произведен анализ на выявление наиболее часто встречающихся октав в популярных композициях.
При проверке гипотез и в построении модели будет использоваться основная метрика - популярность. Она уже входит в исследовательский датасет и рассчитывается внутренними алгоритмами сервиса Spotify, основана на том, сколько раз была воспроизведена музыкальная композиция и насколько недавно эти воспроизведения были.

2. АНАЛИЗ

Первым этапом загрузим датасет с сайта kaggle.com и произведем анализ основных показателей, а также посмотрим, какая информация предоставляется нам.
В датасете, на март 2021 года, представлено 174389 записей и 19 столбцов (показатели/параметры свойств трека). Имеются как числовые, так и текстовые значения..
Исходя из описания первоначального источника данные столбцы означают:

*id - идентификатор трека, созданного Spotify - текстовые значения;

*acousticness - акустичность (от 0 до 1);

*danceability - танцевальность (от 0 до 1);

*energy - энергичность (от 0 до 1);

*duration_ms - продолжительность, в мс (целое число, обычно от 200k до 300k);

*instrumentalness - инструментальность (от 0 до 1);

*valence - валентность (от 0 до 1) - чем выше значение, тем более позитивное настроение у песни;

*popularity - популярность (от 0 до 100) - наша основная метрика;

*tempo - темп (Float обычно от 50 до 150);

*liveness - живучесть (от 0 до 1);

*loudness - громкость (обычно от -60 до 0);

*speechiness - есть ли речь в треке (от 0 до 1);

*year - год (в диапазоне от 1920 до 2021);

*mode - тип музыки (0 = минор, 1 = мажор);

*explicit - содержание откровенного текста (0 = не содержит , 1 = содержит);

*key - ключ октавы (все клавиши на октаве закодированы как значения от 0 до 11, начиная с C как 0, C # как 1 и так далее...)

*artists - исполнители (Список упомянутых исполнителей) - текстовые значения;

*release_date - Дата выпуска в основном в формате гггг-мм-дд, однако точность даты может отличаться - текстовые значения;

*name - Название песни - текстовые значения.

Из всей массы данных можно выделить 172230 уникальных значений, а также 137013 уникальных песен и 36195 артистов (групп). Нулевые значения отсутствуют. Для исключения отклонений по анализу, в данной базе были исключены дубликаты, и были оставлены именно уникальные значения. Весь датасет был разделен на 2 таблицы, первая - очищенная от дубликатов, но с оставлением всех параметров, включая текстовые значения, и вторая, с удалением таких столбцов как 'artists', 'id', 'name', 'release_date'. Данные параметры носят информативный параметр (ID трека), а также текстовую информацию (группа, имя песни). В данном анализе мы будем ориентироваться именно на числовые значения, поэтому текст из анализа или прогноза мы исключаем. Также был исключен столбец 'release_date', т.к. значения были заполнены хаотично, где-то проставлялась полная дата релиза, а где-то только год, что дает большую разность и правильно проанализировать данный параметр будет невозможно, будут большие отклонения. И ко всему этому длительность трека перед построением модели по прогнозированию популярности будет переведен из миллисекунд в минуты для лучшего понимания.
При удалении ненужных столбцов формировались дополнительные дубликаты, это было вызвано тем, что без них остальные параметры некоторых треков мало чем отличались и были похоже друг на друга. Данные дубликаты также были исключены.

Если посмотреть внимательно на статистики параметров можно увидеть интересное наблюдение по нашей метрике, а именно то, что среднее значение по популярности 25,93, ¾ доли нашей выборки имеет 42 пункта по популярности из 100 возможных. Это говорит о том, что основная масса треков имеет низкие показатели популярности, только небольшая доля (менее ¼) выходит за рамки половины возможной шкалы по популярности.

Если посмотреть динамику популярности по годам с использованием среднего значения, то увидим ее рост с 1956 - 2000 гг. 00-10-е года идет спад популярности треков, но ближе к 2020-му немного стабилизируется на значении в 35 пунктов. В 2021 году этот показатель резко упал, что может характеризовать недостаточным заполнением данного параметра, т.к. прошло мало времени с релиза трека и до сбора данных, требуется больше времени для корректного вывода популярности треков данного диапазона в 2021 году.
Особой популярностью пользуются треки 90-х и 00-х годов. Это можно объяснить тем, что предполагаемая аудитория, наиболее полно использующая платные сервисы прослушивания музыки находится диапазоне 20-35 лет. Для них характерно прослушивание треков, которые захватывают их молодой период жизни. Данный факт необходимо

дополнительно проверять, чтобы подтвердить данную гипотезу, но в нашем случае у нас отсутствуют необходимые данные для ее подтверждения или опровержения. Хотелось бы отметить, что бизнес-заказчику необходимо смотреть на текущие тренды в музыке с ориентиром нашего предположения, возможно популярность согласно жанрам циклична с исторической точки зрения.
Если говорить про жанры, то из связанного датасета мы видим 3232 наименования уникального жанра. При выделении ТОП-10 самых популярных жанров сильно выделяются жанры Азиатского направления. Данный факт можно объяснить с демографической точки зрения, т.к. население из стран Азии занимает бОльшую долю в населении планеты.
Также стоит выделить то, что направлений достаточно много и выделить четкую грань или разницу между ними довольно сложно, часто какой-либо жанр включает в себя стили или поджанры других. Если их объединить по наиболее встречающимся направлениям, то можно выделить направления поп-музыки, рэп-поп, а также электроники и r&b, у которых наиболее высокие показатели популярности.
Если посмотреть на наиболее часто встречающиеся жанры в композициях, то наиболее встречающиеся жанры - музыка к фильмам/шоу, оркестровая музыка, а также подвиды танцевальных жанров. Большая доля (порядка 13563 треков) указана без типа жанра, без заполненного значения.

Если говорить, про взаимосвязь популярности и типа ключа октавы, то прямой взаимосвязи данный параметр не имеет, разница в используемых ключах по критерию популярности не имеет значения. Стоит только отметить о чуть сниженном уровне популярности у треков с использованием октавы под номером 3.

3. КОРРЕЛЯЦИЯ ДАННЫХ

Стоит более детально посмотреть на корреляцию, чтобы оценить взаимосвязь параметров музыкальных композиций к метрике.
Наиболее коррелируемые с популярностью будут такие характеристики (показатели) как: year, acousticness, loudness, energy, instrumentalness, speechiness, explicit, danceability, и вычисляемая их взаимосвязь к метрике находится от 0,12-0,54. Следовательно, взаимосвязь показателей является слабой или практически отсутствующей. Часть из них имеет как прямую связь с популярностью (year, loudness, energy, explicit, danceability), так и обратную (acousticness, instrumentalness, speechiness). Такие параметры, как tempo, valence, duration_ms, mode, key, liveness имеют практически отсутствующую взаимосвязь с метрикой, если не полностью отсутствующую.

Хорошо выделяется умеренная прямая корреляция с таким параметром, как year (год) - 0,54. Это говорит о том, что с возрастанием года, увеличивается и популярность треков.
Если рассматривать остальные наиболее коррелируемые показатели, то можно выявить следующие закономерности, которые необходимо учесть стейкхолдерам. Acousticness - чем ниже данный показатель, тем выше популярность, нам выгоднее, чтобы песни были менее акустическими. Loudness, energy - чем громче и энергичнее музыка, тем больше шансов трека на большую популярность. Также треки с меньшим количеством инструментальной музыки имеют больше шансов получить более высокие показатели нашей метрики.

Интересно проверить тот факт, насколько отличаются показатели употребления нецензурной лексики по нашей метрике. Если брать во внимание нашу выборку и закреплять все это визуальной составляющей, а именно диаграммой, то можно увидеть, что у треков с использованием ненормативной лексики выше показатель уровня популярности, чем у тех, которые не используют ненормативную лексику в музыкальных композициях. Разность в данных составляет около 15 пунктов, что довольно существенно.

4. ПРОВЕРКА ГИПОТЕЗ

Перейдем к проверке наших гипотез.

При проверке первой гипотезы, что “Композиции в мажорном исполнении более популярны, чем в минорном”, использовалась дополнительная проверка t-теста Стьюдента на момент статистического значимого различия между ними. Данный метод проверки гипотезы более точно будет давать результат для наших двух независимых выборок. При проставлении уровня значимости в пределах 5% данный тест дал отрицательный результат, также при построении диаграммы особых различий обнаружено не было.

Вывод: Гипотеза “Композиции в мажорном исполнении более популярны, чем в минорном” у нас не подтвердилась, следовательно на практике данный факт оказывает статистически незначимое влияние.

При проверке второй гипотезы "Продолжительность наиболее популярных треков - 3-3,5 минуты" использовалось визуальное представление. Для начала был произведен перевод миллисекунд в минуты для информативного ответа, выведены основные статистики. По исследуемым данным ¾ всех композиций имеет продолжительность до 4,5 минут

(бОльшая часть в диапазоне 2,5-4,5 минут), среднее значение - 3,8 минут, но хотелось бы отметить, что есть композиции с продолжительностью от часто используемых 10-12 минут до около 89 минут. Данные треки дают погрешность в наших статистиках и их можно принять за выбросы, популярность у данных треков в среднем менее 40 пунктов, исключаем их из проверки.
При исследовании треков с продолжительностью до 10 минут наблюдается восходящая линия регрессии, в теории это означает, что чем выше продолжительность, тем выше популярность, но опять же в пределах до 10 минут. Из распределения популярности в зависимости от продолжительности треков при построении диаграммы, видим, что наибольшая плотность популярности находится в нашем проверяемом диапазоне - 3-3,5 минут.

Вывод: Гипотеза "Продолжительность наиболее популярных треков - 3-3,5 минуты" подтвердилась, но ее также можно и расширить, указав, что продолжительность наиболее популярных треков находится в диапазоне 2,5-3,5 минут, так будет дан более точный ответ.

Третья гипотеза проверялась также с использованием визуальной составляющей. Для начала были исключены из выборки треки, с нулевым значением темпа, т.к. это невозможно. Заполнение было лишним, т.к. исключаемых данных было немного и на проверку это не влияет. Линия регрессии восходящая, что говорит о возможном незначительном увеличении популярности треков при большем количестве ударов в минуту (BPM). Распределение популярности на графике примерно одинаковое с 90-220 BPM, но большая плотность наблюдается в диапазоне 110-130 BPM. При выведении статистик видно, что разброс темпа находится в диапазоне 30-243 BPM, медиана - 117 BPM, среднее арифметическое значение на 115,7 BPM.

Исходя из информации взаимосвязи темпа музыки с популярностью точно ответить на нашу гипотезу не получится. Да, среднее (117 BPM) и медианное (115,7 BPM) значения наиболее распространенных треков находятся рядом с нашей гипотезой, косвенно ее подтверждая. Однако, сама плотность популярности не сильно выделяется и находится примерно на одном уровне с 75-160 BPM. Если брать количеством, тогда да, основной темп треков с максимальной популярностью находится в значениях 116 BPM, что близко к нашему предположению.

Вывод: данная гипотеза косвенно подтвердилась, но однозначного положительного ответа на нее нет.

5. ПОСТРОЕНИЕ МОДЕЛИ ПРЕДСКАЗАНИЯ ПОПУЛЯРНОСТИ ТРЕКА

Перед началом работы производилась очистка данных - удалялся столбец по данным длины трека в миллисекундах, т.к. у нас есть информация по длине трека в минутах. В миллисекундах было большое количество значений, что могло давать менее точный прогноз, по минутам эти данные будут более сжатые, без потерь в качестве информации.

Для построения модели предсказания трека будем использовать метод классификации (кодирования) данных get_dummies, а также метод машинного обучения, основанном на градиентном бустинге CatBoost.
Метод классификации был выбран по причине более простого разбиения данных по классам, другие методы, например, One-hot encoder, требуют больших вычислительных мощностей, т.к. у нас преобладают данные с плавающей точкой (float64). Также One-hot encoder выдает хорошие результаты при небольшом значении признаков, и негативно влияет на модель, если их много. Разбивка по классам необходима для обучения модели, которая работает только с целочисленными данными и категориальными значениями..

CatBoost - это современный метод на основе градиентного бустинга, использующий повышение градиента на деревьях решений, работает с числовыми признаками. Данный метод дает хорошие результаты по предсказаниям, рекомендательным системам, поэтому он был выбран как основа для нашей модели.

Для оценки результата работы модели использовался метод RMSE (среднеквадратическая ошибка), т.к. цель задачи - регрессия, т.е. разработка алгоритма для предсказания результата, в нашем случае - популярности. RMSE рассчитывается как квадратный корень из среднего квадрата разностей между фактическим результатом и прогнозом. В силу свойств этой метрики, усиливается влияние ошибок, по квадратуре от исходного значения. Чем сильнее ошибка, тем сильнее у нас будет отклонение от базовой модели. Также она оперирует меньшими величинами по абсолютному значению, что может быть полезно для вычисления на компьютере. Идеальным значением данного метода будет значение 0, что означает прогноз без каких-либо ошибок.

Данные разбивались на тестовые и тренировочные, результативная колонка с метрикой была выделена отдельно для сверки результатов, далее производилось обучение модели и сверка результатов.

В итоге, после проведения всех этапов наша модель показала результат по RMSE в 11,7 пунктов, что вполне неплохой результат.

6. ВЫВОД

Мы выполнили исследовательский анализ данных треков, частично выбранных автором датасета Yamac Eren Ay из стримингового сервиса Spotify с 1920-2021 гг.

Из набора данных мы нашли параметры, которые наиболее сильно влияют на популярность треков. Прямую взаимосвязь с популярностью имеют: год, громкость, энергичность, содержание ненормативной лексики, танцевальность. Обратную взаимосвязь с популярностью имеют: акустичность, инструментальность, использование речи в треке. Также выделили чаще встречающиеся и наиболее популярные жанры.

Выявили, что ключ октавы не влияет на популярность треков, однако стоить учесть о чуть сниженном уровне у ключа под маркировкой №3.

Из проверки гипотез было выявлено:

1. Композиции в мажорном или минорном исполнении оказывают статистически незначимое влияние на популярность;
2. Продолжительность наиболее популярных треков от 2,5 до 3,5 минут;
3. Гипотеза “Наиболее популярными треками являются те, которые имеют темп в 120 ударов в минуту” косвенно подтвердилась, но однозначного положительного ответа на нее нет. Да, среднее (117 BPM) и медианное (115,7 BPM) значения наиболее распространенных треков находятся рядом с нашей гипотезой, косвенно ее подтверждая. Однако, сама плотность популярности не сильно выделяется и находится примерно на одном уровне с 75-160 BPM. У треков в этом диапазоне приблизительно равные шансы на успех, скорее всего на результат уровня популярности влияют другие факторы.

Итоговым результатом было построение модели по прогнозированию популярности треков в зависимости от параметров. В качестве модели использовался алгоритм Catboost, а для оценки результата - RMSE, среднеквадратическая ошибка, как более чувствительная к отклонениям или ошибкам. При этом сама модель при использовании таких параметров как explicit, mode и классифицированного показателя key дала результат со среднеквадратической ошибкой в 11,7 пунктов.

СПИСОК ИСТОЧНИКОВ

1. «Современное»: Почему это слушают? Уловки популярной музыки - Юрий Волошин. https://concepture.club/post/infopoloz/pop-music-tricks
2. Spotify Dataset 1921-2020, 160k+ Tracks - Yamac Eren Ay. https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks
3. User Guide Pandas. https://pandas.pydata.org/docs.html
4. CatBoost is a high-performance open source library for gradient boosting on decision trees. https://catboost.ai/
5. Как правильно выбрать метрику оценки для моделей машинного обучения: часть 1 Регрессионные метрики. https://www.machinelearningmastery.ru/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0/
6. User Guide scikit-learn. https://scikit-learn.org
7. Выбор статистического критерия для тестирования гипотез. https://lit-review.ru/biostatistika/vybor-statisticheskogo-kriteriya/
8. 50 оттенков matplotlib — The Master Plots (с полным кодом на Python). https://habr.com/ru/post/468295/
9. KeyError Pandas – How To Fix. https://www.dataindependent.com/pandas/keyerror/
10. Tutorial SciPy. https://docs.scipy.org/doc/scipy/reference/index.html
11. Как выбрать метрики для валидации результата Machine Learning. http://blog.datalytica.ru/2018/05/blog-post.html
12. seaborn: statistical data visualization. https://seaborn.pydata.org/index.html