KAGGLE – Analyzing and Predicting Top Songs on Spotify

Merilin Radvilavičius
Samuel Amankwaa
Sandra Schihalejev
Repository: https://github.com/sandraschih/IDS_H2

Business understanding (Task 2)

Identifying your business goals

Background

- The project aims to analyze and predict the popularity of songs on Spotify, based on daily updated data from 73 countries. This involves examining relationships between song features and their popularity and predicting future trends.
- In this digital age, data plays a pivotal role in understanding consumer preferences. Platforms like Spotify collect vast amounts of data on user behavior, including what, when, and how people listen to music. This data is invaluable for artists and record labels, as it provides insights into listener preferences, emerging trends, and potential markets.
- The characteristics of a song, such as its genre, tempo, mood, and even the instruments used, play a crucial role in determining its popularity on streaming platforms. Understanding the relationship between these features and song popularity can provide artists and producers with critical insights into what makes a song successful. This understanding can guide them in making creative decisions that align with listener preferences.

Business goals

- The primary business goal is to gain insights into what makes a song popular on Spotify. This can help artists, producers, and record labels tailor their music to audience preferences, increasing the likelihood of success on the platform. Additionally, the insights could aid Spotify in curating more engaging playlists and recommending songs that align with user preferences.

- Another crucial business goal is to develop predictive models that can forecast the future popularity of songs. This could be a game-changer for the industry, allowing stakeholders to anticipate market trends and adapt accordingly. Such models would be instrumental in guiding decisions related to album releases, concert planning, and even long-term artist development strategies.

Business success criteria

- Identification of key song features that correlate with high popularity.
- Development of a reliable predictive model for future song popularity.
- Providing actionable insights for music industry stakeholders.

Assessing your situation

Inventory of resources

- Dataset of top Spotify songs in 73 countries, including features and popularity metrics.
- Data analysis and machine learning tools (e.g., Python libraries).
- Expertise in data analysis and machine learning.

Requirements, assumptions, and constraints

- Assumption: The dataset accurately reflects song popularity and features.
- Constraint: Limited to data available in the dataset; real-time data might not be included.
- Requirement: Adherence to data privacy and intellectual property laws.

Risks and contingencies

- Risk of overfitting in predictive models.
- Contingency: Regular model evaluation and updates with new data.

Terminology

- Popularity: A metric provided by Spotify reflecting song success.

- Features: Characteristics of songs such as genre, tempo, instrumentalness,
  key: The key the track is in,
  loudness: The overall loudness of a track in decibels,
  mode: Modality of the track,
  speechiness: The presence of spoken words in a track,
  acousticness: A measure of the acoustics in the track,
  instrumentalness: Indicates the absence of vocals,
  liveness: Detects the presence of an audience in the recording,
  valence: A measure of the musical positiveness conveyed by a track,
  tempo: The overall tempo of the track in beats per minute,
  time_signature: The time signature of the track.

Costs and benefits

- Costs: Time and resources for data processing and analysis,
- Benefits: Improved decision-making in music production and marketing; enhanced user experience on Spotify. Enhanced strategic decision-making in the music industry, leading to potentially higher success rates for songs; improved user experience on Spotify through better song recommendations.

Defining your data-mining goals

Data-mining goals

- To identify significant relationships between song features and their popularity.
- To develop and validate a predictive model for song popularity.

Data-mining success criteria

- Identification of key features impacting song popularity with a high degree of statistical significance.
- Achievement of high accuracy in the predictive model for future song popularity.

The insights derived from this project are intended to benefit music industry stakeholders (artists, record labels, Spotify itself) by providing a deeper understanding of what drives song

popularity on the platform. This can lead to more effective music production, marketing strategies, and personalized user experiences. The project is not just a business endeavor but also aims to enhance cultural and entertainment value for Spotify users worldwide. By harnessing the power of data, the project not only aims to enhance business strategies but also to enrich the cultural landscape of music by aligning production with audience preferences, thus contributing to a vibrant and dynamic global music scene.

Data understanding (Task 3)

Gathering data

Data requirements

There should be

- decent amount of data over a long period of time, because new songs are released pretty frequently,
- some musical features describing the song (danceability, tempo, key etc) to adjudge the popularity of the songs,
- data about the ranking of the songs and
- of course some information abou the songs (ID, artist, song/album release date, duration etc).

Verify data availability

- The data that meets our needs is available on Kaggle: https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated/data.
- The dataset is in CSV-format is updated daily since October 18, 2023.

Define selection criteria

- We will use the dataset (1 table with total size of approx. 38 MB) as of November 30, 2023 so in total of 49 days of data.

    - We excluded the data from December 1 to 4 as it is a official start of the Christmas season and to minimize the impact of seasonal songs appearing in the ranking and affecting the results.
- From the table we will probably use all of the fields except the column *popularity* as there is not enough information about the feature, and also maybe column *album_name* as it does not seem to be important to predict the popularity of the song.

Describing & Exploring data

- 160 133 rows and 25 columns
- The overall memory consumption of the data is 102 626 617 bytes (102,6 MB).
- There are more than enough data and features to achieve our goals.
    - For example there is a chance to do predictions by region: Europe, Asia, North America and South America as there are countries named in the dataset.
    - Maybe excluding Australia & Oceania and Africa as there are only few countries represented.

Details (including descriptions provided by Kaggle and metadata) about the columns after removing the songs without proper metadata:

- **spotify_id** - the unique identifier for the song in the Spotify database (type: string)
    - 3882 unique songs
- **name** - the title of the song (type: string)
- **artists** - the name(s) of the artist(s) associated with the song (type: strstring
    - 2729 unique artists and collaborations, artists appearing the most as solo artists are Taylor Swift (5304 times) and Bad Bunny (4093 times).
- **daily_rank** - the daily rank of the song in the top 50 list (type: integer)
- **daily_movement** - the change in rankings compared to the previous day (type: integer)
    - min: -46
    - max: 49
    - median: 0
    - mean: 1,046161

- **weekly_movement** - the change in rankings compared to the previous week (type: integer)
    - min: -46
    - max: 49
    - median: 1
    - mean: 4,95763
- **country** - the ISO code of the country of the Top 50 Playlist (type: string)
    - 72 countries and NaN as 'Global Top 50' playlist
- **snapshot_date** - the date on which the data was collected from the Spotify API (type: string)
    - between October 18 and November 30, 2023
- ***popularity*** *- a measure of the song's current popularity on Spotify (type: integer)*
- **is_explicit** - indicates whether the song contains explicit lyrics (type: boolean)
    - 59 857 explicit songs and 100 255 non-explicit songs
- **duration_ms** - the duration of the song in milliseconds (type: integer) → **converted into seconds (s) and the column name to 'duration_s'**
    - min: 0,000000
        - as it is impossible that the song does not have length, then as we found out there are total of 21 rows, specifically 3 songs, which do not even have a song name and an artist
        - **corrected min: 34,285000**
    - max: 641,941000
    - median: 187,368000
    - mean: 194,143233
- ***album_name*** *- the title of the album the song belongs to (type: string)*
    - *2889 unique albums appearing in the rankings, as the most appearing ones are "nadie sabe lo que va a pasar mañana" (7640 times) and "1989 (Taylor's Version)" (3249 times).*
- **album_release_date** - the release date of the album the song belongs to (type: datetime)
    - The album release dates vary between January 1, 1942 and November 28, 2023 - that is almost 81 years of difference!

- **danceability** - a measure of how suitable the song is for dancing based on various musical elements (type: float)
    - min: 0,159000
    - max: 0,982000
    - median: 0,706000
    - mean: 0,687154
- **energy** - a measure of the intensity and activity level of the song (type: float)
    - min: 0,012400
    - max: 0,997000
    - median: 0,666000
    - mean: 0,645158
- **key** - the key of the song (type: integer)
    - min: 0
    - max: 11
    - median: 6
    - mean: 5,402549
- **loudness** - the overall loudness of the song in decibels (type: float)
    - min: -31,042000
    - max: 1,155000
    - median: -6,246000
    - mean: -6,633544
- **mode** - indicates whether the song is in a major or minor key (type: integer)
    - min: 0
    - max: 1
    - median: 1
    - mean: 0,507826
- **speechiness** - a measure of the presence of spoken words in the song (type: float)
    - min: 0,023000
    - max: 0,912000
    - median: 0,062000
    - mean: 0,104289
- **acousticness** - a measure of the acoustic quality of the song (type: float)
    - min: 0,000008

    - max: 0,996000
    - median: 0,192000
    - mean: 0,287244
- **instrumentalness** - a measure of the likelihood that the song does not contain vocals (type: float)
    - min: 0,000000
    - max: 0,970000
    - median: 0,000002
    - mean: 0,018159
- **liveness** - a measure of the presence of a live audience in the recording (type: float)
    - min: 0,015400
    - max: 0,968000
    - median: 0,120000
    - mean: 0,172808
- **valence** - a measure of the musical positiveness conveyed by the song (type: float)
    - min: 0,027100
    - max: 0,981000
    - median: 0,528000
    - mean: 0,533812
- **tempo** - the tempo of the song in beats per minute (type: float)
    - min: 47,914000
    - max: 217,969000
    - median: 119,935000
    - mean: 121,873081
- **time_signature** - the estimated overall time signature of the song (type: int)
    - min: 1
    - max: 5
    - median: 4
    - mean: 3,904561

Verifying data quality

- The data was collected since Oct 18 so it still overlaps the Christmas season.
- The date was given as a string, we converted it to **datetime**.

- There is maybe a need to convert duration to minutes as the attribute ‘tempo’ is given per minute. If so, then **duration_min** computational statistics will be:
    - min: 0,571417
    - max: 10,699017
    - median: 3,122800
    - mean: 3,235721
- There are 3 songs (total of 21 rows) with missing values, so after removing them the total number of songs is **160 112**.
- In conclusion the data quality is more than enough to satisfy our needs.

Planning (Task 4)

Make a detailed plan of your project with a list of tasks. There should be at least five tasks. Specify how many hours each team member will contribute to each task.

1. Examine the correlation between individual features and songs’ popularity and identify which features have a significant impact on popularity.
2. Create and train model for predicting popularity.
3. Evaluate the performance of our model, if necessary, make improvements or train other models.
4. Apply the trained model to predict the popularity of new or future songs
5. Analyze and interpret results.
6. Create a poster and prepare for the presentation.

We contribute equally to each step, get together and discuss how to approach the problems. Tasks should take 20-30 hours per person.

List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

- Python with libraries like Matplotlib and Seaborn for data visualization.
- Decision tree, random forest
- Linear regression
- If there is still time, then experiment with Principal Component Analysis (PCA)