Capstone Project

Music Genre Classifier Final Report

1. **Define the Problem Statement:**
    What is a *genre* when it applies to music or the arts in general? Is it an agreed-upon classification scheme or a loose collection of groupings? I ask this question after seeing an award for 'Best Comedy Series' going to "The Bear". Anyone who has watched this series knows that this is a dramatic series, with very little comedy, if any. Undoubtedly, the genre was intentionally selected, perhaps to reach a target audience or demographic.

    Can we apply this type of marketing strategy to music? How does a recording artist decide what genre to classify their music[1]? This is important because as a listener, I generally find new music by searching a genre or listening to a genre-based mix (e.g., New Indie Rock Mix). Music streaming companies categorize songs by genre. Recording artists reach their audience by identifying or aligning with a specific genre.

    But genre is not a well-defined group of categories. How does a recording artist categorize their music? How do they select the genre that will result in reaching their target audience? This model can assist in that effort.

    Since definitions and categories are loose (see Appendix A), could using the characteristics of their songs help them define a better fit for 'genre' and allow them to reach their target audience?

2. **Model Outcomes or Predictions:** Identify the type of learning (classification or regression) and specify the expected output of your selected model. Determine whether supervised or unsupervised learning algorithms will be used.

    The data used for the model included numeric and text fields. Regression analysis and classification algorithms were needed to process and model the data. The 'genre' field had to be encoded as the value was one of a possible 14 genres.

    The model takes in specific song characteristics, characteristics that are available in multiple data sets and in many streaming services (e.g., Spotify). These characteristics are then used to predict the most likely genre.

    Since the actual performance of my model did not exceed 75%, it may not be a useful tool for the original, intended purpose. It can be used, however, to populate NaN values for genre that appeared in the kagglehub dataset. This is demonstrated in my jupiter notebook. The value is that it allows for over 28000 rows to be "saved" and included in the dataset, rather than dropped from training a model.

3. **Data Acquisition:** The deliverable at this step is to identify what data you plan to acquire and use with your model. For the best results, data should come from multiple sources and your analysis for including specific data should be clear.

---

[1] Spotify attaches 'genre' to an artist rather than to a song

Please provide a clear visualization to assess the data's potential to solve the problem as well.

Dataset: Million Song Dataset + Spotify + Last.fm.

Field Descriptions: Field list | Million Song Dataset

This data was also available on Kaggle, as can be seen in the jupiter notebook.

The tags ended up providing the most significant feature to the model, but other song features such as instrumentalness, loudness and duration had significant impacts as well (see picture below).

```
Correlation of features with genre_encoded:
instrumentalness    0.269725
loudness            0.138843
duration_ms         0.127678
valence             0.098733
danceability        0.097004
acousticness        0.085451
energy              0.063302
mode                0.061356
year                0.057659
speechiness         0.036797
tempo               0.031346
key                 0.011273
time_signature      0.003986
liveness            0.000572
dtype: float64
```

4. **Data Preprocessing/Preparation:** For this deliverable, you are tasked with detailing how you cleaned the data for your notebook.
    a. What techniques did you use to ensure your data was free of missing values, and inconsistencies?

The dataset was analyzed for missing values (NaNs) and showed missing values for 'tags' (1127) and for 'genre' (28335). Because the objective of this model is to guess 'genre', it didn't make sense for me to assign values here. I dropped these rows with null values from the dataset

    b. How did you split the data into training and test sets?

Both datasets were split into an 80/20 training/test set. The first dataset (X_train, y_train, X_test , y_test)included only the numeric features that had significant correlation to genre ('instrumentalness', 'loudness' and 'duration_ms').
The second dataset (X_train_all, y_train_all, X_test_all, y_test_all) included the larger feature set of features listed below.

My initial approach was to use exclusively song characteristics like beats per minute, valence, instrumentalness, etc., to predict a genre. As you can see from the definitions below for **genre** there is no scientific approach to the categorization. This approach yielded only a 54.5% accuracy using the best model, RandomForrestClassifier.

My next approach included additional data/features, specifically date of the song (my theory is that the older songs fell more neatly into a category) and tags which include several genres and characteristics included in the song.

- ❖ Tags
- ❖ Year
- ❖ Duration_ms
- ❖ Danceability
- ❖ Energy
- ❖ Key
- ❖ Loudness
- ❖ Mode
- ❖ Speechiness
- ❖ Acousticsness
- ❖ Instrumentalness
- ❖ Liveness
- ❖ Valence
- ❖ Tempo
- ❖ Time_signature

C. Please include any necessary analysis and encoding steps you took as well.

For the numeric features, I applied a standard scaler to the data before feeding it into a model to ensure that the scale of each feature was reasonable and not dominating the other features.

For the text features, specifically 'tags', I found that there were multiple comma-separated tags listed. I applied an encoder to create a binary response to each individual tag.

For all models, the y value (genre), or prediction needed to be encoded using OneHotEncoder or GetDummies since the values were 1 of 13 potential genres. These genres are 'RnB' 'Rock' 'Pop' 'Metal' 'Electronic' 'Jazz' 'Punk' 'Country' 'Folk' 'Reggae' 'Rap' 'Blues' 'New Age' 'Latin' 'World'

I also attempted to use PCA since the number of features was much larger, but found no marked improvement in the model.

5. **Modeling:** For this deliverable, please document your selection of machine learning algorithms that you selected for your problem statement from the first deliverable.

Initially I planned to use Linear Regression as the algorithm for my model. This was based on the fact that all fields were numeric. However, the target output of the model (y) is 'genre' which is a text field. So I had to change my approach to use Logistic Regression instead.

Even Logistic Regression did not perform so well, so I decided to use it for the baseline. I applied different algorithms that were introduced later in the course.

I also had planned to apply a standard scaler to the feature to ensure that the scale of one feature doesn't result in skewing or dominating the other features.

Finally, I needed to use an encoder for the model output ('genre', y) and for text fields like 'tags' to prepare the data for a model.

6. **Model Evaluation:** Share your model evaluation here. What types of models did you consider for your problem (classification, regression, unsupervised)?

Articulate the evaluation metrics you used and how you determined which model was most optimal for your problem.

I used different algorithms for each of the two datasets I created. Both datasets were split into training and test sets, and a baseline was created using LogisticRegression.

The initial (all numeric) dataset showed a baseline of 52.7% accuracy.

The baseline for the second dataset (all features listed above) scored a 73.6% accuracy. The most significant feature is 'tags'. A sample of the data contained in 'tags' is below. As you can see, there is some overlap with genre, but tags provide additional detail or "sub-genres" that can grow or change at any time.

```
['rock, alternative, indie, alternative_rock, indie_rock, 00s'
 'rock, alternative, indie, pop, alternative_rock, british, 90s,
love, britpop'
 'rock, alternative, alternative_rock, 90s, grunge' ...
 'rock, instrumental, japanese' 'electronic, indie_pop, love'
 'rock, alternative_rock, japanese, cover']
```

For both sets of data, I tried a number of algorithms including supporting vector classifier (SVC), logistic regression (since y was not numeric, i could not use linear regression), k-nearest neighbors (knn) classifier and the decisiontreeclassifier. I identified the best model and parameters by conducting a grid search. The results for the second set of data resulted in SVC showing the best performance using a svc_C=1 and svc_kernel of 'rbf' as the best parameters. The model showed an accuracy of 75.11%.

name	train score	test score	average fit time	best_params
svc	0.783539	0.751141	1029.279493	{'svc__C': 1, 'svc__kernel': 'rbf'}
logisticregression	0.743326	0.735630	14.147298	{'logisticregression__C': 0.1}
knn	0.755818	0.723312	3.732122	{'knn__n_neighbors': 7}
decisiontreeclassifier	0.795802	0.669480	1.076520	{'decisiontreeclassifier__max_depth': 15}

The best model was svc with a test score of 75% using the training and test scores. However, svc requires significant processing resources as evidenced by the average fit time in the table. SVC is far more resource intensive than all of the other models. For a lighter weight version of the Genre Classifier model I created, logistic regression can provide similar results without the need for significant computing resources.

After achieving the results I attempted additional improvements on the model. Ensemble techniques were applied, including bagging classifier and random forest classifier. However, this did not improve the model. The accuracy_score method was applied to a BaggingClassifier and oob_score_ method was applied to the RandomForestClassifier to determine accuracy.

Using a neural network also did not result in an improvement to the model.

Appendix 1 **Genre Reference Information**

'RnB' 'Rock' 'Pop' 'Metal' 'Electronic' 'Jazz' 'Punk' 'Country' 'Folk' 'Reggae' 'Rap' 'Blues' 'New Age' 'Latin' 'World'

These definitions are edited results produced by ChatGPT in response to a number of genre queries.

Music genres are categorized based on shared characteristics like rhythm, melody, instrumentation, cultural origin, and lyrical content. While there are many ways to classify music, here are some of the most common music genre categories, along with brief definitions:

1 R&B (Rhythm and Blues)-RNB

Definition: Originating in African American communities, R&B combines jazz, blues, and gospel influences with a strong emphasis on rhythm and soulful vocals.

Characteristics:
Smooth, emotional vocals.
Strong emphasis on rhythm, especially with bass and drums.
Lyrically focused on love, relationships, and emotional expression.

Subgenres: Contemporary R&B, soul, neo-soul

Examples: Beyoncé, Usher, Alicia Keys

2. Rock

Definition: Rock music is defined by its use of electric guitars, bass, drums, and strong rhythms. It originated in the late 1940s and early 1950s.

Characteristics:
Focus on guitars (electric and acoustic).
Drums and bass are central to the rhythm.
Lyrics often focus on personal or societal issues.

Subgenres: Classic rock, alternative rock, punk rock, indie rock.
Examples: The Beatles, Queen, Nirvana, Foo Fighters.

3. Pop

Definition: Short for "popular music," pop is characterized by catchy melodies, simple structures, and lyrics that often focus on love, relationships, and emotions.
Characteristics:
Strong hooks and choruses

Radio-friendly length (usually around 3-5 minutes).

A broad appeal to a wide audience.
Examples: Taylor Swift, Ariana Grande, Ed Sheeran

4. Metal

Definition: Metal is a subgenre of rock that emphasizes heavy, distorted guitars, complex structures, and intense or dark themes.
Characteristics:
Distorted guitar riffs, fast tempos, and heavy drums.
Often themes of fantasy, rebellion, or personal struggle.
Vocal styles can range from clean singing to harsh growling or screaming.
Subgenres: Heavy metal, thrash metal, death metal, black metal.
Examples: Metallica, Iron Maiden, Slipknot.

5. Electronic Dance Music (EDM)

Definition: A broad category that includes any music created primarily with electronic devices, often produced for dancing in clubs or festivals.
Characteristics:
Synthesizers, drum machines, and computer-based production.
Focus on beats and rhythm for dancing.
Build-ups, drops, and breakdowns are common.
Subgenres: House, techno, dubstep, trance, drum and bass.
Examples: Calvin Harris, Skrillex, Marshmello.

6. Jazz

Definition: Developed in the early 20th century, jazz combines elements of African rhythms with European harmonic structures. It emphasizes improvisation and intricate instrumentation.
Characteristics:
Complex rhythms and harmonies.
Improvisation is key.
Use of instruments like saxophones, trumpets, piano, and double bass.
Subgenres: Bebop, swing, smooth jazz, fusion.
Examples: Miles Davis, John Coltrane, Duke Ellington.

7. Punk

Definition: Punk is a fast, loud, and rebellious genre that emerged in the mid-1970s. It's defined by its DIY ethos and opposition to mainstream culture.
Characteristics:
Short, fast songs with simple structures.
Raw, distorted guitars and politically charged lyrics.

Emphasis on attitude and anti-authoritarian themes.
Examples: The Ramones, Sex Pistols, Green Day.

8. Country

Definition: Country music blends folk, Western, and blues traditions. It's deeply tied to rural American culture and often features storytelling lyrics.
Characteristics:
Use of guitars, banjos, fiddles, and harmonicas

Lyrics often reflect themes of love, loss, and life in rural areas.
Clear, simple song structures.
Subgenres: Traditional country, modern country, bluegrass, outlaw country.
Examples: Johnny Cash, Dolly Parton, Luke Bryan.

9. Folk

Definition: Folk music is a genre that includes traditional and contemporary music tied to specific cultures, often passed down orally.
Characteristics:
Acoustic instrumentation (guitar, banjo, fiddle, etc.).
Storytelling lyrics that reflect social issues, history, or personal experiences.
Examples: Bob Dylan, Joan Baez, Mumbo Gumbo.

10. Reggae

Definition: Emerging from Jamaica in the late 1960s, reggae blends ska, rocksteady, and African rhythms with a focus on social and political themes.
Characteristics:
Syncopated rhythms, often with a "one-drop" drum pattern.
Emphasis on bass lines and offbeat guitar or piano.
Lyrical themes often focus on love, unity, and resistance.
Examples: Bob Marley, Peter Tosh, Jimmy Cliff.

11. Hip Hop / Rap

Definition: Originating from African American and Latino communities in the 1970s, hip hop is defined by its rhythm-based, lyrical style known as rap, along with DJing, beatboxing, and breakdancing.
Characteristics:
Rhythmic vocal delivery (rap).
Sampling of other songs or sounds.
Heavy use of bass and drums.
Subgenres: Trap, boom bap, gangsta rap, conscious rap.

Examples: Kendrick Lamar, Kanye West, Drake, Tupac.

12. Blues

Definition: Originating in the Deep South, blues music is a genre that focuses on expressing emotional experiences, especially pain, through a distinctive 12-bar structure and "blue" notes.
Characteristics:
Simple, repetitive chord progressions.
Use of call-and-response vocals and guitar licks.
Lyrically focuses on struggle, pain, and heartache.
Examples: B.B. King, Muddy Waters, Robert Johnson.

13. New Age

Definition: New Age music is a genre intended to create artistic inspiration, relaxation, and optimism. It is often used for meditation, yoga, massage, and spiritual practices.

Characteristics:

Atmospheric and Ambient: Focuses on tone and mood rather than traditional song structure.
Electronic or Acoustic: May include synthesizers, flutes, piano, nature sounds, or world instruments.
Slow Tempo: Typically slow-paced, minimalistic, and non-percussive.
No Lyrics (Often): Many New Age tracks are instrumental or feature ethereal, wordless vocals.
Spiritual and Healing Intentions: Emphasizes inner peace, well-being, and transcendence.
Examples: Enya*,* Yanni, Kitarō*,* George Winston, Deuter

14. Latin

Definition: Latin music refers to a broad category of music originating from Latin America, Spain, and Portugal, often characterized by rhythms, instruments, and vocal styles rooted in Hispanic and Lusophone cultures.

Characteristics:

Common instruments include congas, bongos, timbales, claves, guitar, accordion, and brass instruments.
Typically sung in Spanish or Portuguese.
Most Latin genres are dance-oriented, with strong beats and infectious grooves.
Combines elements of European (melody/harmony), Indigenous, and African (rhythm) traditions.
Includes salsa, merengue, reggaeton, bachata, cumbia, tango, samba, flamenco, and more.

Examples: Marc Anthony, Daddy Yankee, Paco de Lucia

15. World Music

Definition: World music is an umbrella term for music that originates from non-Western cultures, incorporating traditional sounds and instruments from around the globe.

Characteristics:

Use of traditional instruments like sitar, djembe, or marimba.

Fusion of indigenous rhythms and melodies with contemporary styles.

Examples: Ravi Shankar, Youssou N'Dour, Cesária Évora.