Project Proposal

CDS 492
In Hoi Kim

I. Project Title

Predict Music Genre and Hit Song

II. Research question/hypothesis

- Do the sound features of music help distinguish the genre of music?
- Will the similarities of past hit songs also apply to songs that will be popular in 2023?

III. Motivation

As the information age develops further over time, people can easily get or find the information they need. Just by looking at Google, when people enter the information they want, related articles and pictures appear numerous times. Text can be determined by whether or not the searched information is included, but in the case of an image, it may appear because it has already been verified that the image matches the searched information through learning. As more and more information increases, the technology to find the necessary information effectively and accurately becomes important. In the past, people used to simply find and listen to songs they liked, but now people are more likely to find a wide range of songs. Then, what is the way to find the song people want? Since you cannot insert a song directly on the Internet, there is a way to easily find the information by the genre of the song. Genre is a characteristic of a particular song, and the same genre has a similar atmosphere. The genres of music can be largely divided into four categories. There is classical music, jazz, pop, and rock. If songs can be indexed and classified into genres, people will be able to easily find songs with the feeling they want. Also, classifying many songs is much easier if you have the standard genre. While listening to music often, I thought that there are many songs in the world, and among them, the songs that end up at the top of the chart are decided. And it was speculated that the singer's fame or name alone did not determine the ranking, but that the song had its own characteristics. Therefore, I would like to find out and analyze the commonalities of the songs through the songs in the rankings.

IV. Why researching this question is important to you

However, since genres are eventually created by humans, there will be standards for distinguishing them. What is used here are musical features. When you play music, you can feel the properties of the music such as tempo, liveliness, danceability, and acoustic. No matter how similar two people can't have the same DNA or body structure, no matter how similar a song has the same characteristics. And this same song is defined by the word plagiarism. To avoid this, each song was expressed in countless different ways and existed. People’s favorite foods vary, so their favorite music tastes vary. And big music companies such as Spotify, YouTube Music, and others try to provide playlists so that customers who use the service can listen to songs that suit their tastes. Therefore, companies often have to pick out the commonalities of songs, which become the genre of music. If it predicts the genre by analyzing musical characteristics that are already digitized, it can fill the customer's playlist with the same songs as the genre. Simply providing popular charts every day is less beneficial for consumers to use the music service. Therefore, it will become increasingly important to meet people's needs through accurate genre prediction.

Every month, the top songs keep changing. Obviously, it is a song by a famous singer, so it may attract more response, but there are cases where people listen to it a lot because the song is simply good. Then there must be a reason for the songs on the charts. With the trend of people who change frequently, it is very important for people in music to notice what kind of music is in trend now. This is because the fact that they are behind the trend alone will cause a major setback in their moves. If they analyze the songs in the rankings and derive common musical characteristics, they will be able to predict what form of songs will be popular next. Furthermore, marketing using the most similar singer who sings such songs could have a greater effect.

V. What do you hope to learn beyond more insight on the question

Once predicting the genre belongs to the classification, I think I can find out the skills of various verification methods to verify the classification. In addition, although it is a similar genre, it is also a good way to narrow down the number of genres divided into various genres. Trying methods such as statistical EDA, normalization, PCA, etc. will be able to see a clearer

relationship, and by visualizing it, I will be able to see a broader perspective, not just prediction. When creating a model, there are many classification models, such as simple logistic regression, decision trees, or random forest, and I am going to spin them around to find the best model.

VI. Data / Metadata

❖ **Name:** Top Spotify songs from 2010-2019 - BY YEAR
Credit: Leonardo Henrique
Source: http://organizeyourmusic.playlistmachinery.com/
Link: https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year

Format: CSV file, 15 columns, 603 rows

- ID, title, artist, top genre, year are all song identification/general information
- Bpm : the value denotes the beats per minute (tempo) of the song
- Nrgy : the value denotes how energetic the song is
- Dnce : the value denotes the danceability or how easy it is to dance to this song
- dB : the value denotes the loudness (dB) of a song
- Live : the value denotes the liveness, or the likelihood of the song being a live recording
- Val : the value denotes the valence, or how positive the mood of the song is
- Dur : the value denotes the duration of the song
- Acous : the value denotes how acoustic the song is
- Spch : the value denotes the speechiness, how much spoken words are in the song
- Pop : the value denotes the popularity of the song

❖ **Name:** Spotify - All Time Top 2000s Mega Dataset
Credit: Paul Lamere
Source: http://sortyourmusic.playlistmachinery.com/
Link: https://www.kaggle.com/datasets/iamsumat/spotify-top-2000s-mega-dataset

Format: CSV file, 15 columns, 1994 rows

- Genre - the genre of the track
- Year - the release year of the recording. Note that due to vagaries of releases, re-releases, re-issues, and general madness, sometimes the release years are not what you'd expect.
- Added - the earliest date you added the track to your collection.
- Beats Per Minute (BPM) - The tempo of the song.
- Energy - The energy of a song - the higher the value, the more energetic. song
- Danceability - The higher the value, the easier it is to dance to this song.
- Loudness (dB) - The higher the value, the louder the song.
- Liveness - The higher the value, the more likely the song is a live recording.

- Valence - The higher the value, the more positive mood for the song.
- Length - The duration of the song.
- Acousticness - The higher the value the more acoustic the song is.
- Speechiness - The higher the value the more spoken word the song contains.
- Popularity - The higher the value the more popular the song is.
- Duration - The length of the song.

❖ **Name:** Spotify Top 200 Charts (2020-2021)
Credit: Sashank Pillai
Source: https://spotifycharts.com
Link: https://www.kaggle.com/datasets/sashankpillai/spotify-top-200-charts-20202021

Format: CSV file, 23 columns, 1556 rows

- Highest Charting Position: The highest position that the song has been on in the Spotify Top 200 Weekly Global Charts in 2020 & 2021.
- Number of Times Charted: The number of times that the song has been on in the Spotify Top 200 Weekly Global Charts in 2020 & 2021.
- Week of Highest Charting: The week when the song had the Highest Position in the Spotify Top 200 Weekly Global Charts in 2020 & 2021.
- Song Name: Name of the song that has been on in the Spotify Top 200 Weekly Global Charts in 2020 & 2021.
- Song iD: The song ID provided by Spotify (unique to each song).
- Streams: Approximate number of streams the song has.
- Artist: The main artist/ artists involved in making the song.
- Artist Followers: The number of followers the main artist has on Spotify.
- Genre: The genres the song belongs to.
- Release Date: The initial date that the song was released.
- Weeks Charted: The weeks that the song has been on in the Spotify Top 200 Weekly Global Charts in 2020 & 2021.
- Popularity:The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.
- Danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- Acousticness: A measure from 0.0 to 1.0 of whether the track is acoustic.
- Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
- Instrumentalness: Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
- Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track. Values typical range between -60 and 0 db.
- Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.

- Tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- Chord: The main chord of the song instrumental.

VII. Method of analysis

For the first question)

To predict a genre is to classify it. Therefore, I will use Random Forest, Naive Bayes, and Support Vector Machine and logistic regression for the basic classification model. Logistic regression is designed for classification and is most useful for understanding the effects of multiple independent variables on a single outcome variable. However, since the predictor is not a binary variable, it will not result in effective results. The Random Forest Classifier is a meta-estimator that fits multiple decision trees for various subsamples of the dataset and uses means to improve the predictive accuracy of the model and control overfitting. Based on the Bayes theorem, which assumes independence between all feature pairs compared to complex random forests, the Naive Bayes classifier is widely used in complex situations such as document classification and spam filtering, so it will help predict many genres. Finally, the support vector machine represents the training data as points in the space as broad as possible. The new example is then mapped to the same space and is predicted to fall into the category depending on which side of that space it corresponds to.

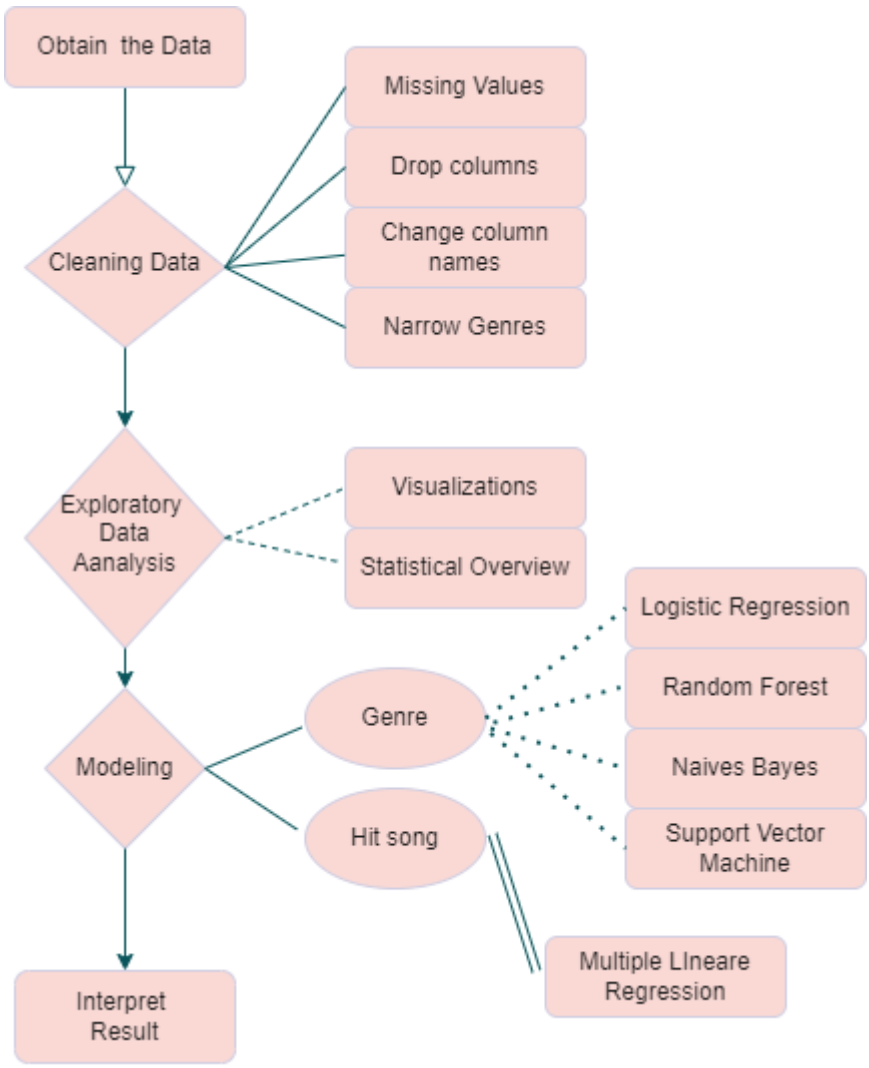
For the second question)

I will use the multiple linear regression model because predicting a song that will be a hit in 2023 is to learn several variables to find the optimal value. In addition, overfitting will be prevented by using K-fold Cross Validation rather than one-time learning.

VIII. Software to be used for analysis

Python using Jupyter Notebook

IX. Workflow diagram



X. Highlight progress to date and add tentative deadlines for tasks going forward

- 02/20 : Summarizing articles of related research
- 02/26 : Search Data and brief look / Cleaning Data
- 02/27 : Proposal
- 02/28 : Finalize Proposal
- 03/01 ~ 03/06 : EDA
- 03/07 ~ 03/14 : Modeling & Evaluation
- 03/14 ~ : Work on presentation and final report

XI. References

1) MUSICAL GENRE CLASSIFICATION USING SUPPORT VECTOR MACHINES

A multi-layer support vector machine-based classifier is applied to automatic musical genre classification. The class boundaries between different genres of music are optimized by learning from training data. The music dataset used in the musical genre classification experiment contains 100 music samples. It is 48.0kHz sample rate, stereo channels and 16 bits per sample. We use support vector machine learning to classify musical frames into relevant genres and then use the derived classification parameters to discriminate different musical genres. We select 60 music samples as training data and segment each sample into 2000 frames. 120,000 frames are used for training the SVM1 and SVM2 and 40,000 frames are used for training the SVM3 and the rest 40 samples are used as a test set. SVMs are used to separate classic, jazz, pop and rock music samples on the test set. The proposed approach can achieve an ideal result in musical genre classification.

Changsheng Xu, Maddage, N. C., Xi Shao, Fang Cao, & Qi Tian. (n.d.). Musical genre classification using support Vector Machines. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).* https://doi.org/10.1109/icassp.2003.1199998

2) Music Genre Classification using Machine Learning Techniques

They compare two classes of models for categorizing music files according to their genre. They train four traditional machine learning classifiers with hand-crafted features and compare their performance. In this work, they make use of Audio Set, a large-scale human annotated database of sounds, which was created by extracting 10-second sound clips from 2.1 million YouTube videos. The data provides the YouTubeID of the corresponding videos along with the start and end times. In this study, each audio signal was converted into a MEL spectrogram using STFT. The parameters used to generate the power spectrogram are listed below. The convolutional neural network model based on VGG-16 that uses only the spectrogram to predict the music genre performs best on all metrics. The baseline feed-forward neural network performs poorly on the test set. In this work, they propose two different approaches to solving the problem of music genre classification using the Audioset data. The CNN based deep learning models were shown to outperform the feature-engineered models, and ensembling the CNN and XGBoost models proved to be beneficial.

Bahuleyan, H. (2018, April 3). *Music genre classification using Machine Learning Techniques*. arXiv.org. Retrieved February 26, 2023, from https://arxiv.org/abs/1804.01149

3) Musical trends and predictability of success in contemporary songs in and out of the top charts

They analyzed 500 000 songs released in the UK between 1985 and 2015 to understand the dynamics of success, correlate success with acoustic features and explore the predictability of success. They found several multi-decadal trends, including a downward trend in happiness and brightness, and a slight upward trend in sadness. We used trade magazine charts to gauge a song's popularity over time and paired these data with metadata and musical feature descriptors from MusicBrainz and AcousticBrainz. Successful songs often differ systematically from average songs in several ways, including being happier, brighter, less sad, more party-like, and more danceable. Popular songs do not reflect the overall tendency of decreased happiness and brightness and increased sadness, with an underlying tendency of increased negativity.

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science*, *5*(5), 171274. https://doi.org/10.1098/rsos.171274