

Song Genre Prediction

I am proposing to work on automatically predicting genre for songs. This is very helpful for music/song streaming websites and apps. Usually such website want to be able to recommend new songs to users based on the user's listening history. Having genre information for all songs can help them improve such recommendations because many individuals have a few preferred genres that they listen to. Example, Spotify already recommends users genre specific lists for newly released songs. Building a machine learning model for this is very important because manually adding genre information for the millions of songs present is not practical.

The data that I am going to use for this task is the data from the Million song dataset. This dataset is public and has both lyrical and audio information about the tracks. A subset of the tracks have genre labels and these are the tracks I will use for my project.

The lyrical information, audio information and genre labels are all available in different file formats (SQL, HDF5, txt) so will utilize my data wrangling skills developed throughout the course to combine all the relevant information in a usable form. This will be a supervised classification problem. My approach would be to use data exploration and hypothesis testing techniques to see which features present in the data are most relevant to predicting genre. Then the dataset will be divided into train set, validation set and testing set. A classifier model will be trained using the selected features to predict genre labels for the songs present in test set.

The deliverables would be the code, a report and a slide deck which will all be present in my github repository.