CS5033 - SL Project Proposal

Airi Shimamura Khoi Trinh

For our SL project, we want to build a few models to predict a user's preference for a song on Spotify.

Each song on Spotify have their own set of 11 features. These features are: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo.

A user can request their streaming history from Spotify via the $Privacy$ section of their account. For the purpose of this project, Khoi's one-year streaming history from February 2022 to February 2023 will be used as the 1 class, for songs we like; and about 4200 random songs which Khoi has never heard will be used as the 0 class, for songs we do not like.

The dataset will have 56,210 songs in class 1, and 4,170 songs in class 0. However, the songs in class 1 have duplicates, due to the nature of streaming a song multiple times. We expect this number will be much lower once the duplicates are removed. We anticipate around 2000 songs in class 1 at the end of the day.

Our current hypothesis is that our supervised learning models will be able to classified these songs as likes or dislikes with at least 90% accuracy. To this end, Airi will implement Naive Bayes and Random Forests; while Khoi will implement Logistics Regression and Decision Trees.