HTML2023 Spring Final Project

b10902003, b10902067

June 13, 2023

1 Initial Settings

1.1 Data Inspection

The features in the dataset can be classified into three types:

- Ordinal: Energy, Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Duration_ms, Views, Likes, Comments
- Categorical: Album_type, Licensed, official_video, Album, Channel, Composer, Artist
- Text: Track, Uri, Uri_spotify, Url_youtube, Description, Title

Notice that we drop ID since it is a manually added feature in the dataset, it should provide no insight about the danceability of each track. We also drop text data, due to the fact that it is provided by the uploader and has no standardized format. We believe that in such cases, it contains more noise than the insight about danceability.

Diving into the dataset, we found out that there are only 11 composers and 97 artists. It seemed unnatural to us that there are such less unique composers and artists given that we have more than 17,000 tracks. We dug into several tracks and realized that both features are totally incorrect, and provide us nothing but misleading information. As a result, we also dropped those two features.

We also observed that there are some uncanny information in the testing dataset. For example, negative values in `Duration_ms` and decimals in `Views`! Those information appear after `id=19170`, which is the 2001-th test data. Moreover, aside from unreasonable values, some features (e.g. Loudness, Speechiness, etc.) have over 10 decimal places after the 2001-th test data and only 3 decimal places before. In light of those weirdness, we made a assumption that **Nijika** had fiddled those test data. To counter such mischief, we had different approaches toward the two sections of data. For the first 2000 test data, we believe they are more related to the training data and thus apply models with stronger power and less regulizations; as for the remaining data, we believe there are huge amounts of noise and thus apply stronger regulization on the models.

In conclusion, we mainly focus on the ordinal features. Inside each model, the feature selection is performed manually by permuation test or automatically by the model (e.g. ShapValues in CatBoost). The details are specified in each model subsection.

1.2 Validation Set

In the first stage, we use k-fold cross validation to tune the parameters and evaluate each model. Nonetheless, the public score is quite far away from the validation score. When the public score is around 1.9, selecting a model with lower validation score does not necessarily give a lower public score. It felt like the public score was randomly hovering between 1.8 to 2.0. We believe that this is due to the manual distortion in test data and thus the validation score can only provide a considerably vague approximation on the public / private score.

Blindly follow the validation score seemed to lead to an overfitting on the training dataset, and did not help us much on selecting the best model. Therefore, we utilized the 5 submissions per day to try out every possible model we had, and managed to reach a relatively lower public score by the end of the first competition. However, on the award ceremony, professor Lin revealed that there is a distribution shift between public and private dataset. Thus what we had done was overfitting on the public dataset, resulting in an appalling private score.

In the second stage, we are given a small subset of the private dataset. Considering the problematic validation score above, we decided to apply it on validating and give up on cross validation. We believed that we had had enough training data but lack in validating, thus apply it on validation may make the greatest use on it.

2 Individual Models

2.1 SVM

2.2 SGD Regressor

2.3 Cat Boost

2.4 Other Models

3 Blending

4 Final Result