Project: TikTok Tracks Popularity Score Prediction

Group Name: Kebab to Pasta

By: Anurag Chatterjing, Jian Gan, Sampurna Gera, Vidhi Jain

The project focuses on the topic of regression analysis to predict the popularity score of the musical tracks used in TikTok videos. This project is inspired by a TikTok tracks dataset that is available on Kaggle, via this link. Each observation in the dataset corresponds to a trending musical track used on the TikTok platform.

Please read carefully the description of the challenge provided in the above link.

To obtain the data

Log on to kaggle.com to download the dataset :
https://www.kaggle.com/yamqwe/tiktok-trending-tracks

Pandas for handling the data

We plan to parse CV with pandas and perform exploratory data analysis to analyse the features impacting our end result. Along with Pandas, we plan to use Matplotlib for visualization.

If you open the dataset using Pandas, you will find that it contains following columns:

1. An integer ID for the track
2. Title of the track
3. Various musical features such as energy, danceability, liveness, valence, acousticness, speechiness, etc.
4. Duration of the track
5. Genre
6. Popularity score, which we will want to predict

Data preprocessing tasks before starting the project

1. Register yourself and download the dataset from kaggle.
2. Load the data using pandas and perform data wrangling techniques like -
    a. removing duplicates.
    b. removing/masking NA or null values
    c. Labeling categorical data using Encoders in sklearn library
    d. HotEncoding values to prepare for ML algorithms
3. After the above steps we can identify the importance of columns using PCA or SVD dimensionality reduction methods. In this stage we identify important features for our regression algorithms.
4. Engineer features from existing features in the dataset to predict scores more accurately.

Task for the project

The project guidelines are:

1. Apply all approaches taught in the course and practiced in lab sessions (Decision Trees, Bagging, Random forests, Boosting, Gradient Boosted Trees, AdaBoost, etc.) on this data set. The goal is to predict the target variable (popularity).
2. Use GridSearchCV/RandomSearchCV to identify the most optimal hyper parameter for the algorithm and improve the accuracy.
3. Compare performances of all these models (for example, in terms of the weighted-f1 scores or accuracy metric you can output).
4. Conclude about the most appropriate ensemble or combination of models on this data set for maximum boosted additive performance for the predictive task.
5. Write a report in .tex format that addresses all these guidelines with a maximal page number of 6 (including figures, tables and references). We will take into account the quality of writing and presentation of the report.