Capstone Proposal
Spotify Song Popularity Prediction

Domain Background:

Music plays an integral part in our lives. Listening to music has never been so easy with a plethora of music apps streaming music with paid/unpaid membership offering any song with a simple click. Spotify is the largest music streaming service provider, and often the popularity of a song on Spotify reflects the popularity of the song in general. I've pondered on the question of why some songs become hits while some do not make it. While I or any individual may have a certain taste, it's interesting to find out what makes a song popular for all the audience.

Problem Statement

How can we determine the popularity of a song based on its features?

Datasets and Inputs

Dataset is obtained from Kaggle: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

The data for this capstone has been limited to 1990-2021.

Following are the features:

- id: Identifier of a song generated by Spotify
- name: Name of the song
- popularity: Popularity of the song (Ranges from 0 to 100)
- duration_ms: Duration of the song
- explicit: Determines whether the song contains explicit content or not (0/1)
- artists: Names of the artists
- id_artists: Identifiers for artists
- release_date: The release date of the song
- danceability: Determines how suitable is the song for dancing (ranges from 0 to 1)
- energy: Determined how energetic is the song (ranges from 0 to 1)

All the features are defined and provided by Spotify.

Solution Statement

Prediction of popularity based on other attributes. Since the target variable i.e., popularity varies 0-99 we will treat it as a regression problem.

Benchmark Model

Null Model: In case we do not observe any attributes, our best guess would be the average value of the popularity rating. Hence, the benchmark model or null model predicts the average value of the training set for the test set.

Evaluation Metric

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

Since the problem at hand is regression, we use a classical evaluation metric for regression tasks i.e., RMSE. The goal is to minimize RMSE.

Project Design

Machine Learning Workflow:

- Exploratory Data Analysis
  Visualization and statistical analysis of data to get more understanding of the features and target variables

- Feature Engineering
  Based on EDA and features selection techniques best features will be selected or generated for the regression task.

- Data Split
  Data will be randomly divided into three parts:
  1. Training set (80%)
  2. Validation set (10%)
  3. Test Set (10%)

  Test set will not be used for feature engineering or hyperparameter tuning tasks.

- Modeling Techniques
  Following modeling techniques will be implemented:
  1. Linear Models
  2. K Nearest Neighbors
  3. Tree Models
  4. Neural Network

  Hyperparameters of the models will be fine-tuned for the best metric

- Deployment: Best performing model will be deployed which is available for the test set. Also, it will be tested on new samples via API