BLUEPRINT OF THE MODEL

Date: Oct 18, 2021

Using NLP Techniques to Predict Song Skips on Spotify based on Sequential User and Acoustic Data

1.Planning:

As streaming services like Spotify become the go-to providers for music, it is increasingly important that these platforms can recommend the right music to their users.

The focus of our project is to build a sequential skip prediction model to predict if a user will skip over a track based on the user's interactions with previous songs and the song's musical qualities in an individual music listening session.

In our project, the input will be an embedding of audio features and user behavior and the output will be predicting whether a track is 'skipped' or 'not skipped'. We can also learn about the relative importance of each input feature by examining a heat map of their weights on a simple logistic regression model. ***We will train a logistic regression model to predict the skip label***

We use ***LSTM's*** for feature analysis and other **NLP** techniques (like **Tokenization, Word2vec, Bag of words**, **word Embedding**) to deal with different EDA in the data. We use binary cross-entropy loss and experimental loss function to find get the maximum accuracy by the means of minimum loss.

2.Designing:

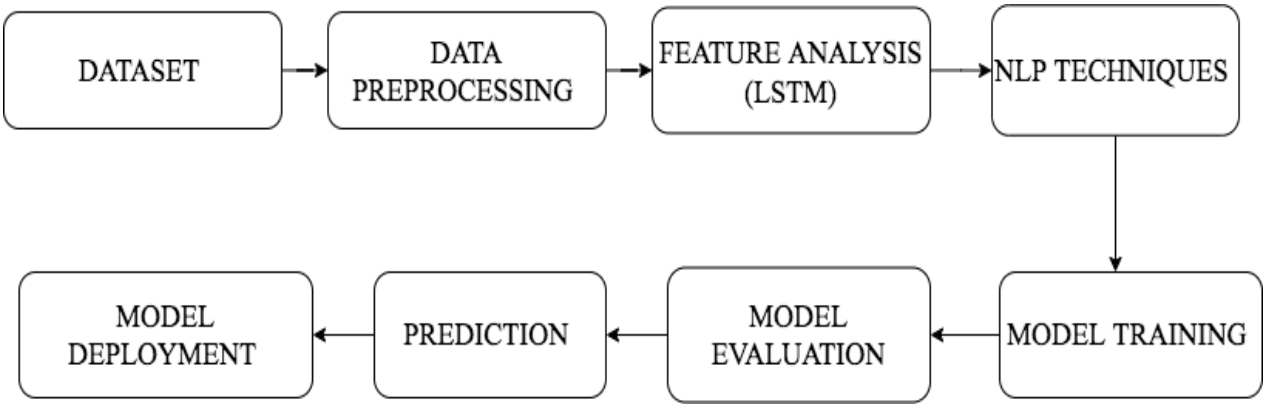

fig: Dataflow diagram

2.1 Datasets:

- Our datasets consist of two files, tf_mini.csv (50705, 30) and log_mini.csv (167881,21).
- We collected the Raw data from GitHub having the file format as CSV.

2.2 Data Preprocessing:

- Data Cleaning: Handling missing values, Handling Outliers.
- Feature Engineering: Categorical data encoding, merged user behavior and acoustic features.

2.3 Feature analysis:

- We will use LSTM and bi-LSTM for Feature analysis.
- LSTM is RNN architecture used in the field of deep learning.

2.4 NLP Techniques:

- We will use different NLP techniques for EDA like **Tokenization, Word2vec, Bag of words**, **word Embedding.**

2.5 Model training:

- We will train a logistic regression model to predict the skip label.

2.6 Model evaluation and prediction:

- Test Data: we will keep 20% of the whole Dataset. Testing will be done using test data on the model.
- The prediction will be done. Check for accuracy score using classification report.

2.7 Model deployment

- We will use python frameworks like Django, Flask for deployment.